



**VYSOKÁ ŠKOLA KARLA ENGLIŠE a.s.**

# **Aplikovaná statistika**

Studijní materiály

Brno 2013

RNDr. Rudolf Schwarz, CSc.

Pro listování dokumentem **NE**používejte kolečko myši!  
 Nebo zvolte následující možnost: *Full Screen*

## Úvodem

se pokusme společně zodpovědět otázku,  
 kterou položil profesor Disman ve své knize [2, str.92]:



„Kolik vran musíme pozorovat,  
 abychom mohli spolehlivě říci,  
 že všechny vrány jsou černé?“

Odpověď na takovou stupidní otázku je strašně jednoduchá a zní:

**„Přece všechny!“**

Ovšem jak to provést, abychom mohli pozorovat všechny vrány na celém světě? Při řešení tohoto problému se vyskytne celá řada otázek. Zde je pouze několik málo z nich:





***Kolik pozorovatelů*** musíme vyslat a ***do jakých míst*** terénu?  
Stačí na tomto **místě** skutečně pouze jeden pozorovatel?



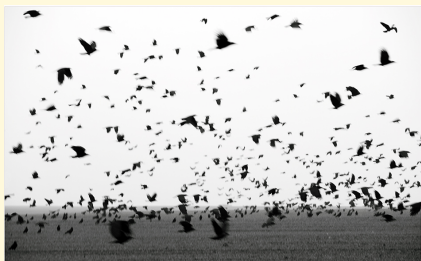
**Kolik pozorovatelů** musíme vyslat a **do jakých míst** terénu? Stačí na tomto **místě** skutečně pouze jeden pozorovatel?



Jak mají být vybaveni? Minimálně zápisníkem a tužkou, ale hodil by se i dalekohled, svačinka, ochrana před nepříznivým počasím, a kdo ví co ještě.

A co z toho lze vůbec realizovat pouze na základě nadšení dobrovolníků a co již my, jakožto zadavatelé výzkumu, musíme **zafinancovat**?





**Kolik pozorovatelů** musíme vyslat a **do jakých míst** terénu? Stačí na tomto **místě** skutečně pouze jeden pozorovatel?



Jak mají být vybaveni? Minimálně zápisníkem a tužkou, ale hodil by se i dalekohled, svačinka, ochrana před nepříznivým počasím, a kdo ví co ještě.

A co z toho lze vůbec realizovat pouze na základě nadšení dobrovolníků a co již my, jakožto zadavatelé výzkumu, musíme **zafinancovat**?



Jsou námi oslovení dobrovolní ornitologové vůbec schopni zjistit barvu kašdické vrány?

Nemůže se stát, že některá **vrána** (i více) přece jenom **unikne** ostrážím zrakům vyslaných pozorovatelů?



Zdravé lidské oko dokáže rozlišit více jak 16 miliónů barevných odstínů.

Kolik z nich ***budeme považovat*** za černou?

Je „antracitová“ ještě černá nebo již není?

A budou v tom všichni pozorovatelé jednotni?



Zdravé lidské oko dokáže rozlišit více jak 16 miliónů barevných odstínů.

Kolik z nich ***budeme považovat*** za černou?

Je „antracitová“ ještě černá nebo již není?

A budou v tom všichni pozorovatelé jednotni?

Z toho všeho co jsme uvedli, plyne následující závěr. **Asi nikdy nebudeme schopni získat údaje o barvě VŠECH vran.** Takže na základě dostupných informací nezbyvá než konstatovat, že „většina vran je černých“. To je ale tvrzení pravděpodobnostního charakteru!

Jak se závěry pravděpodobnostního charakteru nakládat, se dozvíte v první kapitole této příručky o aplikované statistice, která se zabývá **PRAVDĚPODOBNOSTÍ**. O pravděpodobnosti se někdy hovoří, jako o teoretickém základu statistiky.

A co prakticky provedeme s navrátilivšími se zápisníky dobrovolných ornitologů, je náplní kapitoly o **POPISNÉ STATISTICE**.

Na otázku, zda můžeme z těchto zápisníků (tedy z informací pouze o některých vranách), vyvozovat závěry, které platí pro celou populaci vran, například: *tolik a tolik procent vran má jinou barvu*, se pokusíme najít odpověď v kapitole zabývající se **STATISTICKOU INDUKCÍ**. Vždyť jak praví stará vinařská moudrost: *Chceme-li vědět, jak chutná víno v sudu, nemusíme vypít celý sud. Stačí jenom malý doušek a hned víme, na čem jsme.*

A neplatí náhodou, že u mladších vran je větší podíl jedinců s jinou barvou jak černou než u starších vran? Existuje vůbec nějaká souvislost mezi barevností a věkem u vran? Jaké lze činit závěry o vztazích mezi veličinami, neboli analyzovat závislosti, bude probíráno v kapitole zkoumající **REGRESI a KORELACI**.

Pokud se naše poznatky o vranách v čase vyvíjejí (například v zimě je jiné barevné složení jak v létě), dostáváme se do oblasti **ČASOVÝCH ŘAD**, což je další kapitola tohoto kurzu.

A abychom nezůstali pouze u vran, přidáme ještě kapitolu o **HOSPODÁŘSKÉ STATISTICE**, kde budeme pomocí indexů srovnávat ekonomické jevy.



Vzhledem ke skutečnosti, že získané hodnoty jednotlivých znaků (barva konkrétní vrány), nejsou v surové podobě (zápisníky pozorovatelů) ničím jiným než chaotickou a neuspořádanou horou údajů, nelze z nich bez dalšího zpracování vyčíst prakticky žádné užitečné informace.

**Statistika** si klade za cíl informace a zákonitosti, které případně existují mezi některými hodnotami (a na počátku mohou být skryty) odhalit. To znamená uspořádat proměnné (jejich pozorované hodnoty) do názornější grafické či tabulkové formy a popsat je případně několika málo hodnotami, které by obsahovaly co největší množství informací obsažených v původním souboru dat.

V praxi většinou nemáme tolik času, energie a financí (viz příklad o černých vranách), abychom mohli pro učinění kvalifikovaného rozhodnutí prozkoumat všechny údaje vztahující se k analyzovanému problému. V mnoha oborech se proto setkáme s průzkumy opírajícími se o relativně malou část (výběr, vzorek) základního souboru. Statistika pak na základě teorie pravděpodobnosti používá postupy, pomocí nichž můžeme, sice s určitým (odhadnutelným) rizikem, na základě vlastností vzorku usuzovat na vlastnosti celého základního souboru.

***Po zvládnutí této příručky byste měli být schopni popsat problémy, při kterých hraje roli náhoda. A dále je umět řešit pomocí prostředků a nástrojů teorie pravděpodobnosti.***

# Úvod do **Teorie pravděpodobnosti**

## Obsah kapitoly: Teorie pravděpodobnosti

<b>1. Pokusy a jevy</b>	<b>12</b>
1.1. Elementární jev	15
Operace s elementárními jevy	15
<b>2. Pravděpodobnost</b>	<b>17</b>
2.1. Statistická	18
2.2. Klasická	20
Kombinatorika	22
2.3. Geometrická	23
2.4. Axiomatická	25
2.5. Vlastnosti pravděpodobnosti	26
Úplná pravděpodobnost a Bayesův vzorec	30
<b>3. Náhodné veličiny</b>	<b>39</b>
3.1. Základní pojmy	39
3.2. Distribuční funkce $F(x)$	40
3.3. Náhodné veličiny diskrétního typu	42
Příklad	43
3.4. Náhodné veličiny spojitého typu	51
<b>4. Číselné charakteristiky náhodných veličin</b>	<b>55</b>
4.1. Střední hodnota $E(X)$	56
Příklad	57
4.2. Rozptyl $D(X)$ , směrodatná odchylka	59

<b>5. Používaná rozdělení náhodných veličin</b>	<b>60</b>
5.1. Základní pojmy	60
5.2. Diskrétní náhodná veličina — některá její rozdělení	60
Binomické rozdělení	61
Hypergeometrické rozdělení	63
5.3. Spojitá náhodná veličina — některá její rozdělení	65
Normální rozdělení	65
Rovnoměrné rozdělení	70
Exponenciální rozdělení	71
Intenzita poruch	72
<b>6. Náhodné vektory</b>	<b>75</b>
6.1. Sdružená distribuční funkce	75
6.2. Marginální distribuční funkce	76
6.3. Kontingenční tabulka	77
6.4. Číselné charakteristiky náhodného vektoru	78
Kovariance, korelační koeficient	78
Příklad: kontingenční tabulka a korelační koeficient	80
6.5. Příklad: $E(X)$ a $D(X)$ libovolného rozdělení	84
<b>7. Závěr kapitoly – Vztah pravděpodobnosti a statistiky</b>	<b>90</b>

# 1. Pokusy a jevy

**Pokusem** nazveme uskutečnění (výsledek<sup>1</sup>) přesně popsaného komplexu podmínek (např. hod mincí na rovnou desku, zhotovení daného výrobku předepsaným způsobem, provedení chirurgického zákroku, zahřívání vody, výskyt počtu hnízd na jednotlivých stromech apod.).

Předpokládá se, že pokus lze (alespoň teoreticky) za stejných podmínek neomezeně opakovat. Říkáme pak, že se provádí **hromadná stejnorodá operace**. Zákonitostmi, které lze při těchto (opakovaných) pokusech pozorovat, se zabývá teorie pravděpodobnosti.

Pokud není pokus za stejných podmínek opakovatelný — například počet narozených dětí v ČR v letošním roce je pokus, který je pozorovatelný pouze jednou — hovoříme o subjektivní pravděpodobnosti.

**Jevem** pak nazveme každý výsledek nebo důsledek pokusu.

Cílem pokusu (experimentu) je stanovení (správné určení) jevu. Tedy například změření správné a dostatečně přesné hodnoty hledané veličiny.

**Správností** výsledku rozumíme, že soubor experimentálních (získaných, změřených) hodnot je rozptýlen v blízkosti skutečné hodnoty, například obsahu dané látky v roztoku.

**Přesnost** pak vyjadřuje, jak veliké je rozptýlení získaných hodnot při opakování experimentu.

Při jakémkoliv měření se nikdy nevyhneme tomu, aby hodnoty (výsledek) byly zatíženy chybou. Obvykle se chyby dělí do tří skupin.

<sup>1</sup> Takovéto pozorování nazýváme pokusem, ačkoliv je z uvedených příkladů zřejmé, že nemusí jít o skutečný pokus, který je řízený pozorovatelem.

Například při ekonomických „pokusech“ si nemůžeme libovolně nastavovat hodnotu inflace, produktivity práce, úrokové míry, aj.



**Hrubé chyby** vznikají z řady příčin (závada na přístroji, chyba obsluhy, ...) a jsou zapříčiněny nejčastěji jednorázovým dějem.

**Systematické chyby** (soustavné) pravidelně a soustavně zatěžují výsledek pokusu a to vždy jedním směrem (hod falešnou hrací kostkou) a jsou kvantifikovatelné. Jsou zapříčiněny například chybnou kalibrací přístroje, nedodržením podmínek pokusu, ....

**Náhodné chyby** jimž se nikdy nevyhneme. Jsou zapříčiněny nejrůznějšími náhodnými vlivy a obvykle jde o chyby malé, které mají vliv na přesnost výsledků.

Některé další chyby mohou vzniknout při zpracování výsledků (například zaokrouhlovací chyby).

Poznamenejme ale, že jestliže musí být podmínky pokusu přesně vymezeny, neznamena to ještě, že musí být vyjmenovány vyčerpávajícím způsobem. Například při sériové výrobě daného produktu nemusí být vyjmenována teplota a vlhkost vzduchu, atmosférický tlak, kolísání kvality surovin v přípustných mezích, kolísání pozornosti pracovníka při práci, malé rozdíly v opotřebení strojního zařízení, atd.

**Deterministický pokus** končí jediným výsledkem (zahřejeme chemicky čistou vodu na 100 °C při normálním tlaku  $\Rightarrow$  voda vře).

**Náhodný pokus** (stochastický) končí jedním výsledkem z několika možných<sup>2</sup>, přičemž dopředu nevíme kterým.

V dalším se zaměříme pouze na náhodné pokusy, proto budeme často slovíčko „náhodný“ vynechávat a mluvit pouze o pokusu.

<sup>2</sup> Ani při opakování pokusu, jehož výsledek určujeme měřením, nezískáme vždy stejnou hodnotu. Získané výsledky jednotlivých měření se budou (v ideálním případě) lišit v důsledku náhodných chyb. Jednotlivá experimentální měření budou představována realizacemi **náhodné veličiny**. Při posuzování experimentálních dat vycházíme z představy, že signál měřené veličiny je zatížen náhodnou chybou (šumem), přičemž jedním z nejdůležitějších úkolů statistiky je najít vhodný model popisující chování šumu a odhadnout správnou hodnotu signálu. V tomto bodě pak nastává setkání experimentálního měření s matematickou statistikou a teorií pravděpodobnosti. [Otyepka, M., Banáš, P., Otyepková, E. *Základy zpracování dat*. Str. 2. Dostupné z: <http://fch.upol.cz/skripta/zzd/chemo/chemo.pdf>]

**Náhoda** jako pojem.

Když řekneme, že provedeme hod regulérní mincí, máme všeobecnou představu o tom, jak tento pokus provádíme. Neuvažujeme již ale třeba o tom, z jakého materiálu je zhotovena, z jaké výšky a jakým způsobem hod provedeme, neuvažujeme vlhkost vzduchu, tlak vzduchu a jeho proudění apod. Protože nemusíme znát všechny faktory, které výsledek pokusu ovlivňují, nebo je jich příliš mnoho, abychom je do svých úvah všechny zakomponovali, zahrnujeme jejich vliv pod pojem náhoda.

Jakmile byl pokus proveden, můžeme rozhodnout, zda jev o který se zajímáme (např. padnutí lícní strany při hodu mincí, kvalita zhotoveného výrobku, úspěšnost provedené operace, ...) nastal nebo nenastal.

Jevy, které mohou při realizaci pokusu nastat, dělíme na tři skupiny:

**Jistý jev** nastane při každém pokusu (při hodu klasickou kostkou padne číslo větší než NULA).

**Náhodný jev** může, ale také nemusí při realizaci pokusu nastat (při hodu klasickou kostkou padne číslo TŘI).

**Nemožný jev** při žádném pokusu nenastane (při hodu klasickou kostkou padne číslo DESET).

Dále někdy ještě potřebujeme různé jevy mezi sebou kombinovat. Například při jednom hodu uvedenou kostkou, kdy jako výsledek může být „hození“ pouze některého z těchto čísel {1, 2, 3, 4, 5, 6}:

Jev *A* — padne číslo TŘI **nebo** padne číslo PĚT (padne TROJKA nebo PĚTKA).

Jev *B* — padne číslo **sudé** (padne DVOJKA nebo ČTYŘKA nebo ŠESTKA).

Jev *C* — padne číslo sudé **a zároveň** padne číslo větší než čtyři (padne ŠESTKA).

Jev *D* — **nepadne** číslo JEDNA (padne DVOJKA nebo TROJKA nebo ČTYŘKA nebo ...).

Jev *E*<sub>1</sub> — nepadne JEDNIČKA **ani** nepadne DVOJKA (padne TROJKA nebo ČTYŘKA nebo ...).

⋮

Jevy (tak jako ve výše uvedeném příkladu) budeme označovat velkými písmeny latinské abecedy, případně opatřenými indexy. Výjimku má pouze  $\Omega$  jistý jev a  $\emptyset$  nemožný jev.

**Elementární jev** je takový jev, který nelze rozložit na menší částečné jevy, proto různé elementární jevy nemohou nastat současně (ani jeden z výše uvedených jevů  $A, B, C, D$  a  $E_1$  není elementární). Při hodu kostkou je například elementárním jevem padnutí ŠESTKY.

Všechny elementární jevy dohromady tvoří úplnou skupinu (soubor, množinu) **základního prostoru**, což jsou všechny možné výsledky uvažovaného pokusu. Pokud vezmeme vhodný systém  $\mathcal{A}$  podmnožin tohoto základního prostoru splňující následující podmínky (základní prostor je prvkem  $\mathcal{A}$ ; s libovolným jevem  $A$  patřícím do  $\mathcal{A}$  i jeho opačný jev  $\bar{A}$  musí patřit do  $\mathcal{A}$ ; s libovolnými jevy  $A$  a  $B$  i jejich sjednocení musí patřit do  $\mathcal{A}$  — *sjednocení jevů a opačný jev bude vysvětleno vzápětí*), nazveme tento systém  $\mathcal{A}$  polem<sup>3</sup> jevů  $\Rightarrow$  **jevovým polem**.

**Implikace jevů  $A \subset B$**  Říkáme, že jev  $A$  implikuje jev  $B$  (jev  $A$  má za důsledek jev  $B$ ), jestliže jev  $B$  nastane v realizaci pokusu vždy, když v realizaci pokusu nastane jev  $A$ .

**Rovnost jevů  $A = B$**  Říkáme, že jevy  $A$  a  $B$  jsou si rovny, jestliže  $A \subset B$  a zároveň  $B \subset A$ . Jinak řečeno, jestliže jev  $A$  nastane v realizaci pokusu vždy, když nastane v realizaci pokusu jev  $B$  a nikdy jindy.

**Průnik jevů  $A \cap B$**  (*společné nastoupení všech jevů*) je jev, který nastane právě tehdy, když v realizaci pokusu nastane jev  $A$  a zároveň jev  $B$ .

**Sjednocení jevů  $A \cup B$**  (*nastoupení alespoň jednoho z jevů*) je jev, který nastane právě tehdy, když v realizaci pokusu nastane jev  $A$  nebo jev  $B$  (nebo i oba společně).

<sup>3</sup> Termín **pole** má zde význam algebraické struktury (komutativního tělesa).

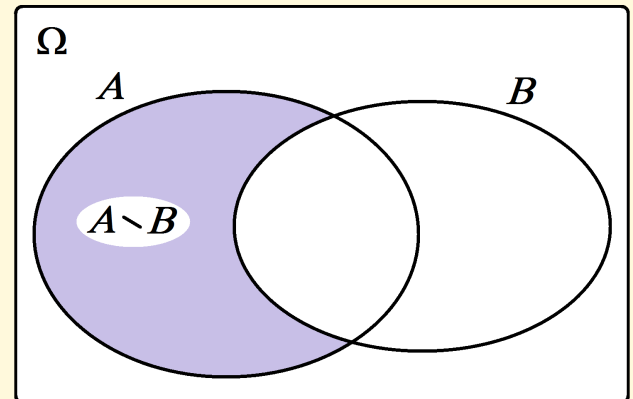
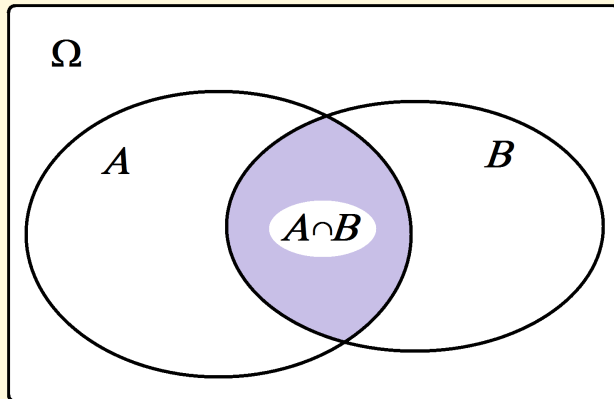
**Rozdíl jevů  $A \setminus B$**  je jev, který nastane právě tehdy, když v realizaci pokusu nastane jev  $A$  a v realizaci pokusu nenastane jev  $B$ .

**Opačný jev  $\bar{A}$**  (**komplementární**, někdy označujeme *non A*) je jev, který v realizaci pokusu nastane právě tehdy, když v realizaci pokusu nenastane jev  $A$ . Poznamenejme, že platí

$$A \setminus B = A \cap \bar{B}$$

$$\bar{A} = \Omega \setminus A$$

Často si výše uvedené vztahy znázorňujeme pomocí tak zvaných **Vennových diagramů**. Například takto můžeme zakreslit průnik (je vybarven) jevů  $A \cap B$  nebo rozdíl jevů  $A \setminus B$ .



## 2. Pravděpodobnost

Výsledek náhodného pokusu nelze s jistotou předpovědět. Některé výsledky však nastávají častěji, některé méně často, některé velmi zřídka. Při velkých sériích opakování však i tyto náhodné pokusy (přesněji jejich výsledky) vykazují určité zákonitosti a pravidelnosti.

**Cílem teorie pravděpodobnosti** je právě studium těchto zákonitostí, jejich popsání a vytvoření pravidel pro určení míry početnosti výskytů těchto jevů.

S těmito zákonitostmi se běžně setkáváme, aniž bychom si to mnohdy uvědomovali. Například každý ví, či intuitivně tuší, že při hodu mincí má stejnou šanci rub i líc a že tudíž při velkém počtu pokusů budou nejspíš padat stejně často (pokud není mince záměrně nějak upravená). Stejně tak ze statistických ročenek lze snadno zjistit, že podíl chlapců narozených v jednotlivých letech vzhledem k celkovému počtu narozených dětí se pohybuje okolo 51,5 %. Přestože v jednotlivých případech nelze pohlaví dítěte předpovědět, můžeme poměrně přesně odhadnout, kolik se narodí chlapců z celkového počtu 10 000 narozených dětí.

Z uvedených příkladů vyplývá, že relativní četnosti některých jevů se s rostoucím počtem opakování ustálí na určitých číslech. Tento úkaz budeme nazývat **stabilitou relativních četností**. Tato stabilita relativních četností je empirickým základem pojmu **pravděpodobnost jevu**.

Zabývejme se pokusem, při němž může nastat jev, který označíme písmenem **A**. Povedeme jednu sérii **n** opakování tohoto pokusu za stejných podmínek. Počet výskytů jevu **A**, který nám říká, kolikrát během série opakování pokusů jev **A** nastal, označíme **m**.<sup>4</sup>

Číslo **m** nazýváme (absolutní) **četností** jevu **A** a číslo  $\frac{m}{n}$  **relativní četností** jevu **A**.

<sup>4</sup>  $f(A) = m$  je vlastně funkcí, která jevu **A** přiděluje přirozené číslo vyjadřující počet výskytů jevu **A** při opakovaném provádění pokusu. Zobecnění této myšlenky vede na **axiomatické** zavedení pravděpodobnosti.

Jestliže provedeme několik sérií (první série měla  $n_1$  opakování a jev  $A$  se vyskytl v  $m_1$  z nich, ve druhé sérii se jev  $A$  vyskytl  $m_2$  krát z  $n_2$  opakování, výsledky třetí série označme  $m_3, n_3, \dots$ ) výše uvedených opakování pokusu, pak lze obvykle pozorovat, že relativní četnosti v jednotlivých sériích kolísají a ustalují se kolem jistého čísla, které nazýváme **pravděpodobností jevu**  $A$  a označujeme  $P(A)$ .

Tedy symbolicky můžeme psát  $\frac{m_i}{n_i} \rightarrow P(A)$ . Je zřejmé, že  $0 \leq \frac{m_i}{n_i} \leq 1$  pro jakýkoli sérii pokusů s pořadovým číslem  $i$ .

Potom zřejmě také  $0 \leq P(A) \leq 1$ .

Pojem **kolísání** lze v pojetí teorie pravděpodobnosti chápat tak, že odchylky (rozdíly) relativních četností od pravděpodobností závisí na náhodě. Číslo  $P(A)$  lze interpretovat tak, že při několika (mnoha) opakováních pokusů (přičemž v každém z těchto pokusů může nastat jev  $A$ ), jev  $A$  nastane asi ve  $P(A) \cdot 100\%$  těchto pokusů.

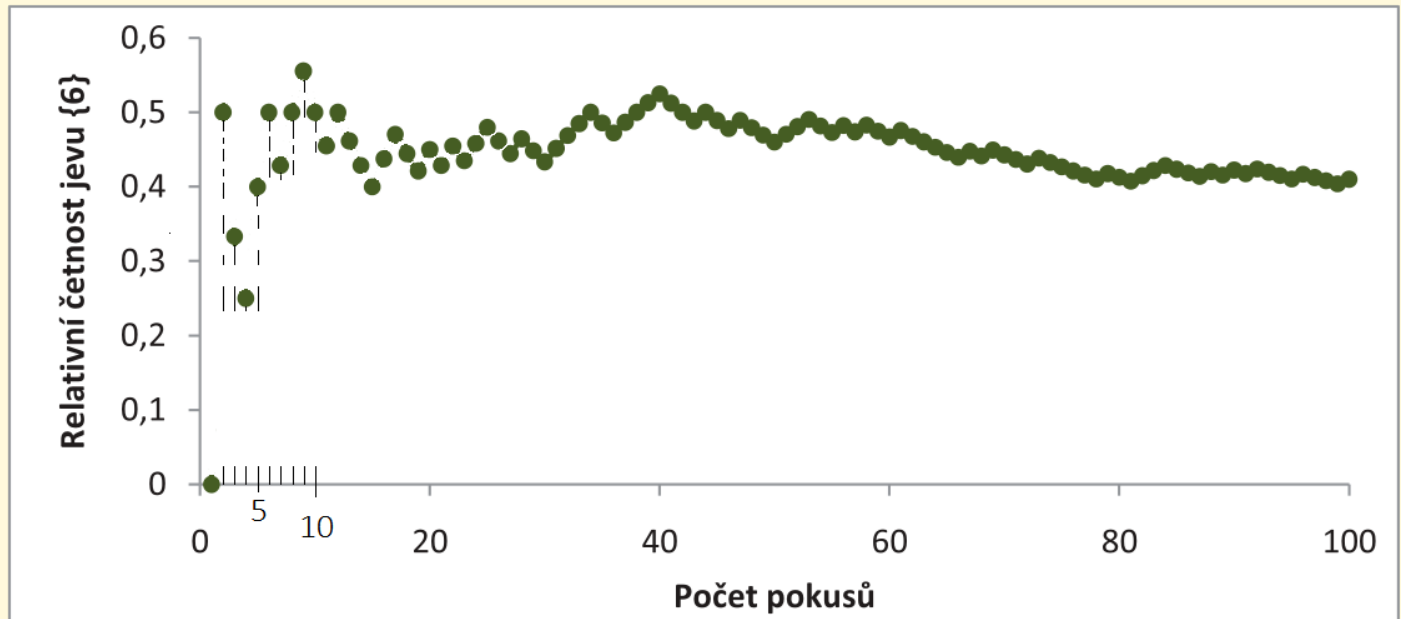
**Statistická** (von Misesova definice  $\Leftarrow$  způsob určení) pravděpodobnosti. Označíme-li jako v předchozích úvahách **relativní četnost** hromadného (pokus za stejných podmínek  $n$  krát opakujeme) jevu  $A$ , přičemž v této sérii nastal jev  $A$   $m$  krát, pak

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n} = \lim_{\text{počet pokusů} \rightarrow \infty} \frac{\text{kolikrát nastal jev } A}{\text{počet všech pokusů}} \quad (1)$$

Misesův přístup k pravděpodobnosti je založen na empirickém zkoumání, jež vede k pozorování „stability relativních četností“. Umožňuje určit pravděpodobnost jevu v případě, že není známo jeho bližší chování (tedy jaké jsou elementární jevy, při kterých zkoumaný jev nastává, a jejich pravděpodobnosti). Jestliže je náhodný pokus libovolně krát (alespoň teoreticky) opakovatelný za stejných statistických podmínek (například hod kostkou či mincí, ...), pak lze pravděpodobnost jevu odhadnout na základě počtu jevů příznivých výsledku pokusů.

Tento odhad je tím přesnější, čím je počet realizací náhodného pokusu ( $n$ ) vyšší. Statistická definice pravděpodobnosti nám například umožňuje odhadnout pravděpodobnost toho, že padne šestka na nepoctivé („cinknuté“) kostce.

Obrázek 1: Převzat z [9, str. 48]



$$\text{Závislost relativní četnosti „padnutí šestky“ na nepoctivé kostce} = \frac{\text{kolikrát padla ŠESTKA}}{\text{počet VŠECH pokusů}}$$

**Klasická** (Laplaceova definice) pravděpodobnosti. Pokud máme konečný počet ***m*** elementárních jevů a **všechny** tyto elementární jevy **jsou stejně možné**, pak pravděpodobnost jevu ***A***, který nastane při ***p*** těchto elementárních jevech určíme pomocí vzorce

$$P(A) = \frac{p}{m} = \frac{\text{příznivé případy}}{\text{všechny možné}} \quad (2)$$

Předpoklad, že všechny výsledky pokusu mají stejnou pravděpodobnost výskytu, je možná pochopitelný, ale v praxi málo obvyklý. Málokterá hrací kostka je totiž natolik ideální, aby na ní čísla padala se stejnou pravděpodobností. Proto jsme dříve uvedli i *statistický* způsob zavedení pravděpodobnosti.

Uvažujme nyní například jev ***A***, že na normální hrací kostce padne šestka. Jak bude (podle předchozích úvah) **hledání pravděpodobnosti** tohoto jevu  $P(A) = ?$ , tedy padnutí šestky ve skutečnosti vypadat?

**Statisticky zavedená pravděpodobnost** (empirická) vychází z experimentu. Kostkou **mnohokrát** hodíme a určíme relativní četnost jevu ***A***, kterou budeme považovat za nejlepší odhad pravděpodobnosti tohoto jevu  $P(A)$ . Viz obrázek 1, kde v prvním hodu šestka NEpadla, ve druhém PADLA, ve třetím NEpadla, ...

Pro „správnou“ kostku se dá očekávat, že se tento odhad bude blížit jedné šestině. Pro falešnou kostku na obrázku 1 je to přibližně 0,4.

**Klasicky zavedená pravděpodobnost** (teoretická) vychází z obecných vlastností dané situace. V případě kostky abstrahuje od její nedokonalosti a bude ji považovat za ideální, na které všechny hodnoty padají se stejnou pravděpodobností. Potom lze k výpočtu pravděpodobnosti využít klasické definice a výsledkem je známá hodnota **jedna šestina**.  $P(A) = 1/6$ .

Všimněte si, že oba pohledy jsou pouze přibližné. Ani jeden z nich neurčí pravděpodobnost naprosto přesně, ale pouze se k ní přiblíží. Jsou to tedy pouhé **modely skutečnosti**, skutečného chování zkoumané kostky.



U empirického přístupu **přesnost** výsledku **závisí na počtu** provedených **pokusů** (experimentů, hodů kostkou). Čím více pokusů, tím přesnější lze očekávat výsledek.

U teoretického přístupu **přesnost** výsledku **závisí na** zvolené **abstrakci** (idealizaci, zjednodušení) celého problému. Čím větší abstrakce, tím jednodušší výpočet, ale tím méně přesný výsledek.

Který přístup tedy zvolit? To vždy závisí na:

- Informacích, které máme k dispozici:
  - Známe všechny elementární jevy?
  - Jsou elementární jevy skutečně stejně možné?
- Možnostech provedení experimentu:
  - Dá se pokus opakovat?
  - Je provedení pokusu náročné na prostředky, na čas?
- A na dalších souvisejících faktorech.

Souvislost obou přístupů (jejichž výsledky se většinou od sebe příliš neliší) pak vede k následujícímu tvrzení: „*PRAVDĚPODOBNOST JE TEORIÍ STATISTIKY A STATISTIKA JE PRAXÍ TEORIE PRAVDĚPODOBNOSTI.*“ [3, str. 176]

**Kombinatorika** Jestliže je počet elementárních jevů (všechny možné) velký, je obtížné je vypisovat všechny. Pokud potřebujeme znát pouze jejich počet, pak ho lze často určit pomocí **kombinatorických schémat** (viz tabulka 1).

Nejdříve připomeňme, že výraz  **$k!$**  (čteme: „ká FAKTORIÁL“), kde  $k$  je přirozené číslo (1, 2, 3, ...), počítáme takto:

$$k! = k \cdot (k - 1)! , \quad \text{přičemž} \quad 0! = 1 .$$

**Určete 5!**

**Řešení:**  $5! = 5 \cdot 4! , \quad 4! = 4 \cdot 3! , \quad 3! = 3 \cdot 2! , \quad 2! = 2 \cdot 1! , \quad 1! = 1 \cdot 0! = 1 \cdot 1 = 1$

$$\begin{aligned} \text{Tedy: } 5! &= 5 \cdot 4! = 5 \cdot (4 \cdot 3!) = 5 \cdot [4 \cdot (3 \cdot 2!)] = 5 \cdot \{4 \cdot [3 \cdot (2 \cdot 1!)]\} = 5 \cdot \{4 \cdot \{3 \cdot [2 \cdot (1)]\}\} = \\ &= 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120. \end{aligned}$$

V kombinatorických schématech uvedených v následující tabulce předpokládáme, že je dáno  **$k$**  prvků, z nichž vytváříme skupiny po  **$r$**  prvcích (neboli výběr třídy  **$r$**  z  **$k$**  prvků). Počet takto vytvořených skupin závisí jednak na tom, jak jsou prvky ve skupině **uspořádány** (zda **je významné**, že jeden prvek stojí před druhým – číslo 12 je jiné než číslo 21 i když v obou jsou stejné cifry jednička a dvojka – nebo **není významné** – částku 7 Kč zaplatíme například tak, že na pult položíme dvoukorunu a pětikorunu, přičemž není významné, kterou položíme jako první, nebo dokonce dáme-li obě mince společně) a potom ještě na tom, zda se každý z prvků může libovolně krát **opakovat** nebo ne.

V prvním sloupci následující tabulky je podmínka, zda záleží na pořadí prvků ve skupině. Ve druhém sloupci je podmínka, zda se prvky ve skupině mohou libovolně krát opakovat. Ve třetím sloupci je název skupiny a ve čtvrtém její označení a počet těchto skupin. Když si shrneme, co znáte ze střední školy:

- **uspořádaný** výběr  $\Rightarrow$  **variace**:  $V_r(k)$
- **neuspořádaný** výběr  $\Rightarrow$  **kombinace**:  $C_r(k)$
- $V_k(k) = P(k) \Rightarrow$  **permutace**:  $P(k) = k!$

Tabulka 1: Kombinatorická schémata

POŘADÍ je podstatné	Prvky se <b>OPAKUJÍ</b>	Název skupiny	Označení skupiny <b>Počet</b> skupin
ano	ne	variace	$V_r(k) = \frac{k!}{(k-r)!}$
ano	ano	variace s opakováním	$V'_r(k) = k^r$
ne	ne	kombinace	$C_r(k) = \binom{k}{r} = \frac{k!}{r! \cdot (k-r)!}$
ne	ano	kombinace s opakováním	$C'_r(k) = \binom{k+r-1}{r} = \frac{(k+r-1)!}{r! \cdot (k-1)!}$

**Geometrická** (definice) pravděpodobnosti. Pokud existuje **nekonečně mnoho** stejně možných elementárních jevů (všechny tyto elementární jevy dohromady označujeme  $\Omega$ ), můžeme je znázornit jako část přímky, roviny, prostoru nebo času, přičemž jakýkoliv jev  $A$  je opět (menší – pokud to není jev jistý) částí takto znázorněné přímky, roviny, prostoru nebo času. Tyto části lze měřit (je to délka, plocha, objem, apod.) a tuto míru označme  $\mu$ . Potom pravděpodobnost, že nastane jev  $A$  je

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} \quad (3)$$

kde  $\mu(\Omega)$ , což je míra základního prostoru (všech elementárních jevů dohromady) je vždy větší než nula. Tedy nulou nikdy nedělíme!

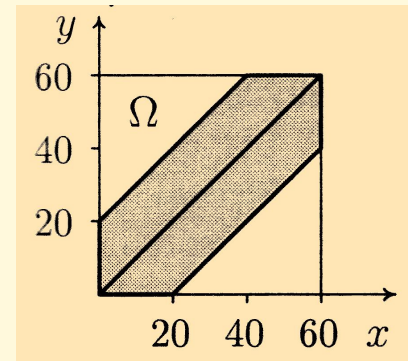
**Příklad:** Dva milenci se dohodli [5, str. 17], že se setkají na stanoveném místě v sobotu mezi druhou a třetí hodinou odpoledne. Po svém příchodu bude každý z milenců čekat na příchod druhého přesně 20 minut a když se nedočká, tak odejde. Předpokládá se, že příchod každého z milenců je v sobotu od 14 hodin do 15 hodin stejně možný. Jaká je pravděpodobnost, že milenci

- se dočkají jeden druhého;
- přijdou ve stejnou dobu.

**Řešení:** Pokusem je zjištění doby, kdy každý z milenců přišel na místo schůzky.

Označme dobu příchodu prvního z milenců (mezi 14. a 15. hodinou)  $x$  a dobu příchodu druhého milence  $y$ , kdy údaje jsou v minutách. Potom lze výsledky pokusu vyjádřit dvojicemi čísel  $[x; y]$  (můžeme si je představit jako body roviny – viz sousední obrázek, který byl převzat z [5, str. 17]), kde  $0 \leq x \leq 60$  a  $0 \leq y \leq 60$ . Počátek soustavy souřadnic je ve 14:00 hod.

Základní prostor  $\Omega$  lze znázornit čtvercem s délkou strany 60 minut. Potom míra základního prostoru je rovna obsahu čtverce, tedy  $\mu_2(\Omega) = 60 \cdot 60 = 3\,600$  jednotek<sup>2</sup>. Protože počítáme plochu (obsah části roviny), budeme míru označovat indexem dva.



**Jev A** – milenci se sejdou.

Tento jev nastane právě tehdy, když rozdíl v dobách příchodů milenců nepřesáhne 20 minut. Tedy platí:  $|x - y| \leq 20$ . Jev A je vyznačen stínovaným obrazcem ohraničeným přímkou  $y = x + 20$  a  $y = x - 20$ . Jeho míra je  $\mu_2(A) = 60^2 - 40^2 = 2\,000$  jednotek<sup>2</sup> (od plochy čtverce odečteme plochu dvou shodných trojúhelníků).

Dle vzorce (3) pro výpočet geometrické pravděpodobnosti:  $P(A) = \frac{\mu_2(A)}{\mu_2(\Omega)} = \frac{2\,000}{3\,600} \doteq 0,556$

**Jev  $B$**  – milenci přijdou ve stejnou dobu.

Tedy  $x = y$ , což je rovnice přímky, na obrázku úhlopříčka čtverce spojující body  $[0; 0]$  a  $[60; 60]$ .

Protože plošný obsah úsečky je roven nule, bude míra jevu  $B$  nula  $\Rightarrow \mu_2(B) = 0$ .

Pak dle vzorce (3) pro výpočet geometrické pravděpodobnosti: 
$$P(B) = \frac{\mu_2(B)}{\mu_2(\Omega)} = \frac{0}{3\,600} = 0$$

Vypočtené pravděpodobnosti lze interpretovat takto: *Během většího počtu sobot se asi v 55,6 % milenci setkají a prakticky žádnou sobotu nepřijdou přesně ve stejnou dobu, i když to není vyloučeno.*

Poznamenejme, že sice pravděpodobnost  $P(B) = 0$ , ale protože jev  $B$  může nastat, není to nemožný jev. Připomeňme, že *nemožným* nazýváme jev, který **nemůže** nastat a přiřazujeme mu nulovou pravděpodobnost.

**Axiomatická** (Kolmogorovova definice) pravděpodobnosti. Pravděpodobnost  $P$  je funkce (viz poznámka 4), která každému jevu  $A$  patřícímu do pole jevů přiřazuje reálné nezáporné číslo<sup>5</sup>

nejvýše rovné jedné, tedy  $0 \leq P(A) \leq 1$ , přičemž funkce  $P$  má následující vlastnosti:

<sup>5</sup> Takto stanovená pravděpodobnost (statistická, klasická i geometrická definice pravděpodobnosti představují pouze speciální, v praxi však často používané, případy axiomatické definice) je z našeho hlediska vhodná pro pochopení toho, jak se pravděpodobnost chová při výpočtech. Všimněte si, že axiomatický systém vymezuje vlastnosti pravděpodobnosti, neudává však žádný návod k jejímu určení (jak ji spočítat).

**Vlastnosti pravděpodobnosti** Pro jevy  $A$ ,  $B$  a  $C$  ze základního prostoru platí:

**Jistý jev**  $P(\Omega) = 1$

**Nemožný jev**  $P(\emptyset) = 0$

**Neslučitelné jevy** Pro libovolné jevy  $A$  a  $B$ , které nemají společný průnik (tedy platí  $A \cap B = \emptyset$  nebo jinak  $A \cap B = \emptyset$ ) je  $P(A \cup B) = P(A) + P(B)$

**Implikace jevů** když  $A \subset B$  pak  $P(A) \leq P(B)$

**Opačný jev**  $P(\bar{A}) = 1 - P(A)$

**Rozdíl jevů**  $P(A \setminus B) = P(A) - P(A \cap B)$

**Sjednocení jevů**  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

**Průnik jevů** (jejich společné nastoupení)  $P(A \cap B) =$  viz upravený vzorec (5)

**Bernoulliovo schéma** Jestliže při určitém pokusu může nastat jev  $A$  s pravděpodobností  $p$  /tedy  $P(A) = p/$  a při  $n$  opakování tohoto pokusu za stejných podmínek se tato pravděpodobnost  $p$  nemění, pak takové opakování pokusu nazýváme **Bernoulliovou posloupností nezávislých pokusů**<sup>6</sup>. Potom jev  $A_k$  (jev  $A$  nastane při tomto opakování přesně  $k$ -krát) bude mít následující pravděpodobnost:

$$P(A_k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n. \quad (4)$$

<sup>6</sup> Zobecníme-li úvahu tak, že budeme popisovat počet náhodných událostí v nějakém pevném časovém intervalu, tak při splnění určitých podmínek (viz [9, str. 160] — ordinarita, stacionarita, nezávislé přírůstky, beznáslednost) dostaneme tak zvaný **Poissonův** proces, kterým se v této příručce nebudeme zabývat.

**Podmíněná pravděpodobnost** Prozatím jsme rozebírali pokusy typu, že hodíme homogenní hrací kostkou tvaru krychle a zkoumáme pravděpodobnost, kdy padne například ŠESTKA (tento jev označíme  $A$ ). Nyní potřebuje zavést nějakou doplňující informaci. Například jaká je pravděpodobnost, že padla zmíněná šestka, když vím (za předpokladu), že padlo sudé číslo (tento jev označíme  $B$ ).

Nezajímáme se o pravděpodobnost, vztahující se k podmínkám původního pokusu, ale na „jinou pravděpodobnost“, vztahující se k podmínkám pokusu, které jsou doplněny o předpoklad, že nastal jev  $B$ . Tuto „jinou pravděpodobnost“ označíme  $P(A|B)$  a nazveme ji **podmíněnou pravděpodobností**. Je to pravděpodobnost, že nastane jev  $A$  za předpokladu, že jev  $B$  již nastal<sup>7</sup>.

Tento příklad, ve kterém se vyskytuje pouze několik málo možností, můžeme počítat přímo pomocí rozkladu na elementární jevy. Je jediná příznivá možnost, a to, že padla šestka. Když víme, že padlo sudé číslo, tak všechny možnosti jsou tři (dvojka, čtyřka, šestka). Tedy podle vzorce (2) pro výpočet klasické pravděpodobnosti  $P(A|B) = \frac{1}{3}$ . Když jej zevšeobecníme, pak z vícero podobných příkladů dostaneme následující vzorec:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{pokud } P(B) > 0. \quad (5)$$

**Pravděpodobnost průniku  $A \cap B$**  dvou jevů je po úpravě vzorce (5) rovna

$$P(A \cap B) = P(B) \cdot P(A|B)$$

a stejně tak  $P(A \cap B) = P(B \cap A) = P(A) \cdot P(B|A)$

<sup>7</sup> Například pravděpodobnost kolize za jakéhokoliv počasí coby nepodmíněná pravděpodobnost a pravděpodobnost kolize podmíněná výskytem náledí.

**Nezávislost dvou jevů  $A, B$ .** Jestliže pro dva jevy platí

$$P(A|B) = P(A) \quad \text{nebo} \quad P(B) = 0$$

pak říkáme, že jev  $A$  není závislý na jevu  $B$ .

Jestliže je jev  $A$  nezávislý na jevu  $B$ , pak je také jev  $B$  nezávislý na jevu  $A$ . Říkáme, že jevy  $A$  a  $B$  jsou **vzájemně nezávislé**.

Jsou-li jevy  $A$  a  $B$  vzájemně nezávislé, pak platí:

$$P(\bar{A}|B) = P(\bar{A}), \quad P(A|\bar{B}) = P(A), \quad P(\bar{A}|\bar{B}) = P(\bar{A})$$

Také můžeme říci, že dva jevy  $A$  a  $B$  jsou vzájemně nezávislé právě tehdy, když

$$P(A \cap B) = P(A) \cdot P(B) \quad (6)$$

Podle informací správce školní počítačové sítě víme, že během sta provozních **hodin** je počítačová síť **v průměru** nedostupná:

**6 minut** v důsledku výpadku serveru (kdy server nereaguje na požadavky klientů) a

**2 minuty** v důsledku poruchy (odstavení) elektrické rozvodné sítě 230 V (nefungují přípojný body sítě).

Serveru se to netýká, protože nepřetržitý zdroj napájení **UPS** udrží server v provozu nezávisle na stavu rozvodné elektrické sítě minimálně 10 minut.

Určete pravděpodobnost, že v daný okamžik (konkrétní minutu) nebudeme moci využívat školní počítačovou síť v důsledku její nedostupnosti.



**Řešení:** Nejdříve si označíme jednotlivé jevy: **V** ... Výpadek serveru

**E** ... odstavení rozvodné Elektrické sítě 230 V

Jev označující skutečnost, že se nebudeme moci připojit do školní počítačové sítě, ať již pro výpadek serveru nebo pro přerušení dodávky elektrické energie, je vlastně sjednocením uvedených jevů. Tedy se ptáme, jaká je pravděpodobnost  $P(V \cup E) = ?$

Pravděpodobnosti uvedených jevů (po převedení na společné jednotky — minuty) jsou následující:

$$P(V) = 6 \text{ minut ze sta hodin} = \frac{6}{100 \cdot 60} = 0,001$$

$$P(E) = 2 \text{ minuty ze sta hodin} = \frac{2}{100 \cdot 60} = 0,000 \bar{3}$$

$$P(V \cup E) = P(V) + P(E) - P(V \cap E) \quad / \text{dříve uvedená vlastnost pravděpodobnosti}/$$

Zbývá nám tedy určit pravděpodobnost  $P(V \cap E)$ .

Jinak řečeno: Zajímá nás, jaká je pravděpodobnost, že rozvodná elektrická síť bude odstavena právě v okamžiku (ve stejné minutě), kdy je server nedostupný v důsledku jeho výpadku. Tedy, kdy oba jevy nastoupí společně (současně  $\Rightarrow$  průnik jevů). Protože jevy **V** a **E** jsou vzájemně nezávislé (dodávka elektrické energie není podmíněna stavem serveru) podle vzorce (6) platí:

$$P(V \cap E) = P(V) \cdot P(E) = 0,001 \cdot 0,000 \bar{3} = 0,000 \, 000 \, \bar{3}$$

$$\text{A konečně: } P(V \cup E) = 0,001 + 0,000 \bar{3} - 0,000 \, 000 \, \bar{3} \doteq 0,001 \, 332 \, 666 \doteq 0,001 = 0,1 \, \%$$

V daný okamžik nebudeme moci využívat školní počítačovou síť s pravděpodobností rovnou desetíně procenta <sup>8</sup>.

<sup>8</sup> V praxi, pokud se nejedná o *bezpečnost jaderné elektrárny, lety do kosmu apod.* je výpočet pravděpodobnosti s přesností na desetiny procenta naprosto dostačující.

**Úplná pravděpodobnost** Pokud neslučitelné jevy  $H_1, H_2, \dots, H_n$  vyplňují **celý** základní prostor jevů (jevové pole), pak pro libovolný jev  $A$  platí

$$P(A) = \sum_{i=1}^n P(H_i) \cdot P(A|H_i) \quad (7)$$

což chápeme tak, že základní prostor je rozdělen mezi takzvané hypotézy  $H_i$  a sledovaný jev  $A$  (jeho část) může nastat společně vždy jen s jedinou z nich (obrázek převzat z [4]).

Úpravou vzorce (7) dostáváme následující

**Bayesova věta** Pokud neslučitelné jevy  $H_1, H_2, \dots, H_n$  vyplňují **celý** základní prostor jevů (jevové pole), pak pro libovolný jev  $A$  platí

$$P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{P(A)} \quad (8)$$

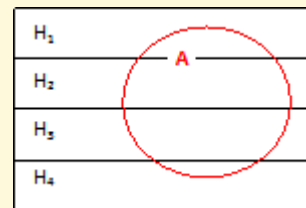
Bayesův vzorec používáme tehdy, chceme-li z výskytu jevu  $A$  při realizaci pokusu odhadnout, jak se jednotlivé hypotézy „podílely“ na výskytu jevu  $A$ .

Pravděpodobnosti  $P(H_i)$ ,  $i = 1, 2, \dots, n$ , nazýváme **apriorními** pravděpodobnostmi jevu  $H_i$ , tj. pravděpodobnostmi uskutečnění hypotézy  $H_i$  „před pokusem“.

Pravděpodobnosti  $P(H_i|A)$ ,  $i = 1, 2, \dots, n$ , nazýváme **aposteriorními** pravděpodobnostmi jevu  $H_i$ , tj. pravděpodobnostmi uskutečnění hypotézy  $H_i$  „po provedení pokusu“, při němž jev  $A$  nastal.

Můžeme tedy říci, že tento vzorec nám umožňuje dávat pozdější (aposteriorní) zkušenosti do souladu s původními (apriorními) předpoklady, případně jak takové zkušenosti změni současné hodnocení situace oproti původním předpokladům.

Využití Bayesova vzorce naznačíme na příkladu z medicínské praxe. [14, str.193]



- Představme si pacienta, který určitě trpí jednou z nemocí **A** či **B**. Na základě dosavadních znalostí (anamnéza, klinický stav, ...) víme, že nemoc **A** se vyskytuje s pravděpodobností 0,8; nemoc **B** s pravděpodobností 0,2. Lékař nechal u pacienta provést zkoušku enzymů v séru, o níž víme, že u nemoci **A** je pozitivní v 90 % případů a u nemoci **B** jen ve 20 % případů. Tento test měl negativní výsledek. Jak je tím ovlivněna lékařova diagnóza tohoto pacienta?

Jak v takových případech postupovat, ukazuje

(<http://mi21.vsb.cz/flash-animace/aplikace-bayesovy-vety-v-biomedicine>)

nebo následující dva příklady.

Doporučuji přečíst také rozbor [Monty Hallova problému](#) na Wikipedii.

Je samozřejmé, že také pro Bayesův vzorec platí, že závěry nemohou mít průkaznější vypovídací schopnost, než jim předpoklady (premisy) dovolí. Výsledek nemůže být spolehlivější než odhadované pravděpodobnosti předpokladů.

V praxi je to však bohužel často tak, že pro apriorní pravděpodobnost jsou k dispozici jen zcela nespolehlivé odhady nebo dokonce protichůdné údaje.

**Příklad 1.** [5, Příklad 1. 10., str. 30] Při automatickém vymývání lahví je dobře vymytých 98 % z nich. Po vymytí se všechny láhve ještě kontrolují v automatické prohlížečce, která propustí 3 % špatně vymytých lahví a vrátí k novému promytí 5 % dobře vymytých lahví.

Kolik procent lahví se znovu vymývá?  $\Rightarrow$   **$P(\text{láhev neprošla kontrolou}) = ?$**

A kolik procent lahví, z těch co neprošly kontrolou, bylo dobře vymyto?

$\Rightarrow$   **$P(\text{láhev byla dobře vymyta přestože neprošla kontrolou}) = ?$**

**Řešení:** Při popisu výsledků pokusu (*vymývání láhve a kontrolu vymytí dohromady*) použijeme následující označení:  $V$  — láhev je dokonale vymytá;  $P_r$  — kontrola vymytou láhev propustí. Další případy popíšeme pomocí opačných jevů, kde jev  $\bar{V}$  značí, že láhev nebyla dobře vymytá a  $\bar{P}_r$  označuje, že kontrola láhev nepropustí a vrátí ji k novému vymytí. Je zřejmé, že jevy  $V$  a  $\bar{V}$  vyplňují celý základní prostor jevů. Nic jiného, než že láhev je dobře nebo není dobře vymytá, nemůže nastat. Podle předchozího značení tedy máme  $i = 2$  a  $H_1 = V$ ,  $H_2 = \bar{V}$ . Vše je nejlepším zaznamenávat do přehledného schématu, kde na pomyslné spojnici mezi jednotlivými jevy (zleva doprava) budeme vypisovat pravděpodobnosti, s jakými nastal jev **vpravo**. A tohle bylo zadáno:

$P_r$  – **propuštěna**

$V$  – **vymytá  
dokonale**

$$P(V) = 0,98$$

$$P(\bar{P}_r|V) = 0,05$$

$\bar{P}_r$  – **nepropuštěna  
vrácena**

jednotlivá  
láhev

$P_r$  – **propuštěna**

$$P(P_r|\bar{V}) = 0,03$$

$\bar{V}$  – **nevymytá dokonale  
špatně vymytá**

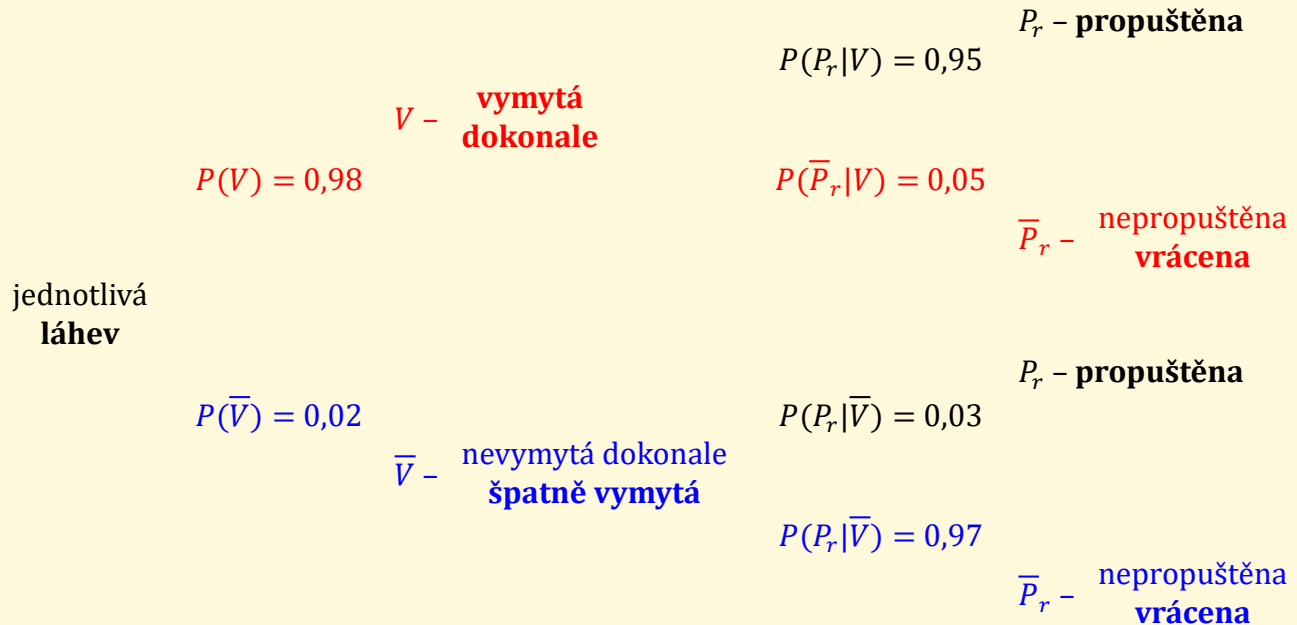
$\bar{P}_r$  – **nepropuštěna  
vrácena**

**Řešení:** Při popisu výsledků pokusu (vymývání láhve a kontrolu vymytí dohromady) použijeme následující označení:  $V$  — láhev je dokonale vymytá;  $P_r$  — kontrola vymytou láhev propustí. Další případy popíšeme pomocí opačných jevů, kde jev  $\bar{V}$  značí, že láhev nebyla dobře vymytá a  $\bar{P}_r$  označuje, že kontrola láhev nepropustí a vrátí ji k novému vymytí. Je zřejmé, že jevy  $V$  a  $\bar{V}$  vyplňují celý základní prostor jevů. Nic jiného, než že láhev je dobře nebo není dobře vymytá, nemůže nastat. Podle předchozího značení tedy máme  $i = 2$  a  $H_1 = V$ ,  $H_2 = \bar{V}$ . Vše je nejlepší zaznamenávat do přehledného schématu, kde na pomyslné spojnici mezi jednotlivými jevy (zleva doprava) budeme vypisovat pravděpodobnosti, s jakými nastal jev **vpravo**.

jednotlivá láhev			$P_r$ – propuštěna
		$V$ – vymytá dokonale	$P(P_r V) = 0,95$
	$P(V) = 0,98$		$P(P_r V) + P(\bar{P}_r V) = 1$
			$P(\bar{P}_r V) = 0,05$
			$\bar{P}_r$ – nepropuštěna vrácena
	$P(V) + P(\bar{V}) = 1$		
			$P_r$ – propuštěna
	$P(\bar{V}) = 0,02$		$P(P_r \bar{V}) = 0,03$
		$\bar{V}$ – nevymytá dokonale špatně vymytá	$P(P_r \bar{V}) + P(\bar{P}_r \bar{V}) = 1$
			$P(\bar{P}_r \bar{V}) = 0,97$
			$\bar{P}_r$ – nepropuštěna vrácena

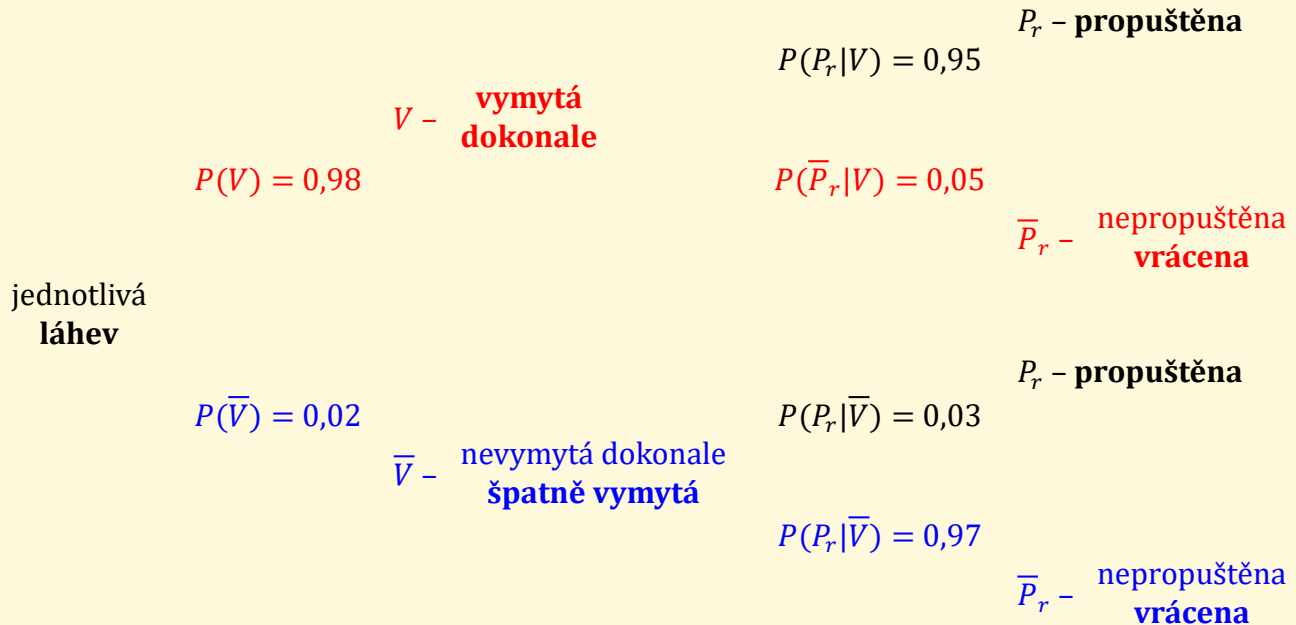
Protože součet pravděpodobností musí být jedna ...

**Řešení:** Při popisu výsledků pokusu (vymývání láhve a kontrolu vymytí dohromady) použijeme následující označení:  $V$  — láhev je dokonale vymytá;  $P_r$  — kontrola vymytou láhev propustí. Další případy popíšeme pomocí opačných jevů, kde jev  $\bar{V}$  značí, že láhev nebyla dobře vymytá a  $\bar{P}_r$  označuje, že kontrola láhev nepropustí a vrátí ji k novému vymytí. Je zřejmé, že jevy  $V$  a  $\bar{V}$  vyplňují celý základní prostor jevů. Nic jiného, než že láhev je dobře nebo není dobře vymytá, nemůže nastat. Podle předchozího značení tedy máme  $i = 2$  a  $H_1 = V$ ,  $H_2 = \bar{V}$ . Vše je nejlepším zaznamenávat do přehledného schématu, kde na pomyslné spojnici mezi jednotlivými jevy (zleva doprava) budeme vypisovat pravděpodobnosti, s jakými nastal jev **vpravo**.



Pak je vráceno  $P(\bar{P}_r)$

**Řešení:** Při popisu výsledků pokusu (vymývání láhve a kontrolu vymytí dohromady) použijeme následující označení:  $V$  — láhev je dokonale vymytá;  $P_r$  — kontrola vymytou láhev propustí. Další případy popíšeme pomocí opačných jevů, kde jev  $\bar{V}$  značí, že láhev nebyla dobře vymytá a  $\bar{P}_r$  označuje, že kontrola láhev nepropustí a vrátí ji k novému vymytí. Je zřejmé, že jevy  $V$  a  $\bar{V}$  vyplňují celý základní prostor jevů. Nic jiného, než že láhev je dobře nebo není dobře vymytá, nemůže nastat. Podle předchozího značení tedy máme  $i = 2$  a  $H_1 = V$ ,  $H_2 = \bar{V}$ . Vše je nejlepší zaznamenávat do přehledného schématu, kde na pomyslné spojnici mezi jednotlivými jevy (zleva doprava) budeme vypisovat pravděpodobnosti, s jakými nastal jev **vpravo**.



Pak je vráceno  $P(\bar{P}_r) = P(V) \cdot P(\bar{P}_r|V) + P(\bar{V}) \cdot P(\bar{P}_r|\bar{V}) = 0,98 \cdot 0,05 + 0,02 \cdot 0,97 = 0,0684$

což znamená, že asi necelých sedm procent (6,84 %) lahví se znovu vymývá.

A kolik procent lahví, z těch co neprošly kontrolou, bylo dobře vymyto?

Tedy:  $P(\text{láhev byla dobře vymyta} \mid \text{za podmínky, že neprošla kontrolou}) = ?$

To určíme podle Bayesova vzorce (8)

$$P(V|\bar{P}_r) = \frac{P(V) \cdot P(\bar{P}_r|V)}{P(\bar{P}_r)} = \frac{0,98 \cdot 0,05}{0,0684} = 0,716,$$

což znamená, že asi 72 % (přesněji 71,6) z nově vymývaných lahví se vymývá zbytečně.

A protože v matematice (a tím také v pravděpodobnosti) nemůže výsledek záviset na písmenech, která použijeme na označení něčeho, ukažme si podobný příklad ještě jednou.

**Příklad 2.** Banka má pro styk s klienty dvě pobočky, *VELKOU* a *malou*. „Velká“ pobočka poskytuje 70 % všech úvěrů této banky a mezi jejími smlouvami o poskytnutí úvěru je 5 %, které byly uzavřeny s právníckými osobami. „Malá“ pobočka poskytuje zbytek úvěrů a z tohoto zbytku činí smlouvy o úvěru uzavřené s právníckými osobami 15 %. Banka se rozhodla provést náhodnou kontrolu poskytnutých úvěrů. Při této kontrole je náhodně vybrána jedna úvěrová smlouva.

Určete pravděpodobnost, že:

- A) náhodně vybraná smlouva byla uzavřena s právníckou osobou;
- B) pokud byla vybrána smlouva uzavřená s právníckou osobou, pak poskytnutí tohoto úvěru realizovala „velká“ pobočka.



Označme si jevy a jejich pravděpodobnosti, které plynou přímo ze zadání:

$P_o$  ... úvěr byl poskytnut **P**rávnické osobě

$\overline{P_o}$  ... úvěr **NE**byl poskytnut právnické osobě  $\Rightarrow$  komukoliv jinému jak právnické osobě

$V$  ... úvěr realizovala „**V**elká“ pobočka

$M$  ... úvěr realizovala „**M**alá“ pobočka

$P(V) = 0,70$  protože 70 % úvěrů poskytuje „velká“ pobočka

$P(M) = 0,30$  protože zbytek  $(100 - 70 = 30)$  % poskytuje „malá“ pobočka

$P(P_o|V) = 0,05$  „velká“ pobočka uzavřela 5 % smluv s právnickými osobami

$P(\overline{P_o}|V) = 0,95$  „velká“ pobočka uzavřela  $(100 - 5 = 95)$  % jiných smluv

$P(P_o|M) = 0,15$  „malá“ pobočka uzavřela 15 % smluv s právnickými osobami

$P(\overline{P_o}|M) = 0,85$  „malá“ pobočka uzavřela  $(100 - 15 = 85)$  % jiných smluv

úvěry banky	$P(V) = 0,7$ ...„velká“ pobočka	$P(P_o V) = 0,05$ ...sml. s právnickou osobou
		$P(\overline{P_o} V) = 0,95$ ...jiná smlouva
	$P(M) = 0,3$ ...„malá“ pobočka	$P(P_o M) = 0,15$ ...sml. s právnickou osobou
		$P(\overline{P_o} M) = 0,85$ ...jiná smlouva

**A)** Určete pravděpodobnost, že náhodně vybraná smlouva byla uzavřena s právnickou osobou, v naší symbolice  $P(P_o) = ?$  Když projdeme všechny cesty (v grafu) na jejichž konci je jev  $P_o$ ,

- hodnoty **v každé z cest** (smlouvy jedné pobočky) mezi sebou **násobíme** (plyne z upraveného vztahu (5) na pravděpodobnost průniku /společné nastoupení/ jevů)

$$P(A \cap B) = P(A|B) \cdot P(B) \text{ a}$$

- hodnoty celých **cest mezi sebou sečítáme** (plyne z pravděpodobnosti sjednocení jevů)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

dostáváme vzorec (7) na úplnou pravděpodobnost (protože jevy **V** a **M** vyplňují celý základní prostor  $\Leftrightarrow$  další pobočka banka nemá). Jevy **V** a **M** jsou vzájemně neslučitelné (jednu smlouvu nemohly uzavřít obě pobočky společně)  $\Rightarrow P(V \cap M) = 0 \Rightarrow P(P_o \cap V) \cap (P_o \cap M) = 0$

$$\begin{aligned} P(P_o) &= P(P_o \cap V) \cup (P_o \cap M) = P(P_o \cap V) + P(P_o \cap M) - P(P_o \cap V) \cap (P_o \cap M) = \\ &= P(P_o \cap V) + P(P_o \cap M) - 0 = P(P_o|V) \cdot P(V) + P(P_o|M) \cdot P(M) = 0,05 \cdot 0,7 + 0,15 \cdot 0,3 = 0,08 \end{aligned}$$

Náhodně vybraná smlouva bude s pravděpodobností 8 % uzavřena s právnickou osobou.

Nebo jinak řečeno: Ze sta náhodně vybraných smluv jich osm bude pravděpodobně uzavřeno s právnickou osobou.

- B)** Byla vybrána smlouva uzavřená s právnickou osobou. Určete (aposteriorní) pravděpodobnost, že poskytnutí tohoto úvěru realizovala „velká“ pobočka, v naší symbolice  $P(V|P_o) = ?$

Podle Bayesova vzorce (8)

$$P(V|P_o) = \frac{P(V) \cdot P(P_o|V)}{P(P_o)} = \frac{0,7 \cdot 0,05}{0,08} = \frac{0,035}{0,08} = 0,4375$$

S pravděpodobností téměř 44 % náhodně vybranou smlouvu s právnickou osobou uzavírala „velká“ pobočka.

Nebo jinak: Nejpravděpodobněji čtyřicet čtyři smluv uzavřených s právnickou osobou (ze sta náhodně vybraných smluv) bylo realizováno na „velké“ pobočce.

## 3. Náhodné veličiny

### 3.1. Základní pojmy

Až doposud jsme se zabývali otázkou, zda při uvažovaném pokusu nastanou či nenastanou určité jevy a jak lze vypočítat jejich pravděpodobnost. Avšak ve většině pokusů se jejich výsledky vyjadřují čísly, jejichž hodnoty závisí na náhodě. Například:

- výška mužů v populaci,
- počet obdržených bodů při zkoušce,
- spotřeba pohonných hmot při ujetí 100 km,
- počet nemocných, kteří přijdou k lékaři během dne,
- doba bezporuchové funkce přístroje,
- počet zásahů při střelbě do terče,
- skutečná cena postaveného domu,
- atd.

Veličiny, které výsledkům pokusů jednoznačně přiřazují reálná čísla a jejichž hodnoty závisí na náhodě, se nazývají **náhodné veličiny**.

Pravděpodobnost, že náhodná veličina  $X$  nabyla hodnoty  $x$  — tedy nastal jev, který označujeme  $\{X = x\}$  — zapíšeme  $P(X = x)$ .

Sestavíme-li seznam všech možných dvojic  $[x_i; P(X = x_i)]$ , nazveme ho **rozdělením pravděpodobnosti** (jaké hodnoty a s jakou pravděpodobností může náhodná veličina nabývat).

Náhodné veličiny se tradičně označují velkými písmeny latinské abecedy, například  $X$ ,  $Y$ ,  $T$  a podobně. Hodnoty náhodných veličin jsou (reálná) čísla přiřazená určitým způsobem výsledkům uvažovaného pokusu.

My se budeme zabývat náhodnými veličinami pouze těchto typů:

**Diskrétního typu** — jejichž oborem hodnot jsou izolované body (například počet výrobků).

**Spojitého typu** — jejichž oborem hodnot jsou hodnoty z nějakého intervalu, přičemž každý bod z tohoto intervalu má nulovou pravděpodobnost (například vzdálenost, teplota).

Mimo výše uvedených typů náhodných veličin existují ještě další typy (zejména náhodná veličina *smíšeného typu*, jejíž hodnoty vytvoří jistý interval, přičemž některý bod z tohoto intervalu má nenulovou pravděpodobnost), těmi se však zabývat nebudeme.

Náhodná veličina tedy nabývá při daném pokusu určité hodnoty, přičemž předem nevíme, jaká hodnota to bude. Jestliže ale provedeme větší počet těchto pokusů, pak lze pozorovat, že výskyty jednotlivých hodnot náhodné veličiny vykazují jisté zákonitosti, (její pravděpodobnost je nějak **rozdělena**), což lze popsat pomocí tak zvaných **zákonů rozdělení** pravděpodobnosti. Ty určují pravděpodobnosti, s jakými náhodná proměnná nabude určitou hodnotu nebo nějaké hodnoty z určitého intervalu. Nejobecnějším z těchto zákonů rozdělení je distribuční funkce

**Distribuční funkce  $F(x)$**  náhodné veličiny  $X$  nazýváme (reálnou) funkcí, pro kterou platí

$$F(x) = P(X \leq x) . \quad (9)$$

Distribuční funkce  $F(x)$  tedy vyjadřuje pravděpodobnost, s jakou náhodná veličina  $X$  nabude hodnot z intervalu  $(-\infty; x]$ , nebo jinak řečeno, že náhodná veličina  $X$  **nepřekročí** „hraniční číslo“  $x$ .

**Příklad:** Značí-li náhodná veličina  $X$  výšku (v cm) mužů v populaci, pak hodnota  $F(170) = 0,45$  udává, že asi 45 % mužů v populaci má výšku do 170 cm včetně.

**Vlastnosti distribuční funkce  $F(x)$  náhodné veličiny  $X$ .**

1.  $0 \leq F(x) \leq 1$
2. Distribuční funkce je neklesající funkcí — pro každá dvě reálná čísla  $x_1 < x_2$  platí:  $F(x_1) \leq F(x_2)$
3. Distribuční funkce je spojitá zprava — pro každé reálné číslo  $x$  platí:  $\lim_{h \rightarrow 0^+} F(x + h) = F(x)$
4.  $\lim_{x \rightarrow -\infty} F(x) = 0$  a  $\lim_{x \rightarrow +\infty} F(x) = 1$

Poznamenejme, že tyto vlastnosti plně distribuční funkci<sup>9</sup> charakterizují.

Někdy se distribuční funkce definuje jako pravděpodobnost, že náhodná veličina  $X$  nabude hodnot **ostře menších než  $x$** , tj.  $F(x) = P(X < x)$ <sup>10</sup>. Pak se uvedené vlastnosti distribuční funkce až na třetí vlastnost nezmění. V případě, že nepřipouštíme rovnost, je funkce  $F(x)$  spojitá zleva.

Pravděpodobnost, že náhodná veličina  $X$  nabude některé hodnoty z intervalu  $\langle x_1; x_2 \rangle$ , lze určit následovně:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1). \quad (10)$$

<sup>9</sup> Inverzní funkce k distribuční funkci se nazývá **kvantilová funkce** a značí se  $Q = F^{-1}$ . Kvantil  $x_p$  je veličina pro kterou platí  $F(x_p) = p$ . Například  $x_{0,95}$  je 95% kvantil, tedy taková hodnota, pro kterou je distribuční funkce rovna 0,95 a kterou náhodná veličina překročí s 5% pravděpodobností.

<sup>10</sup> My připouštíme rovnost zejména kvůli analogii s **kumulativní** četností číselného statistického znaku.

## Náhodné veličiny diskrétního typu

Někdy zkráceně říkáme jen **diskrétní náhodné veličiny**. Jak jsme již dříve uvedli, jejich oborem hodnot jsou izolované body. Toto sice není exaktní definice, ale nám plně postačuje.

**Pravděpodobnostní funkci  $f(x)$  diskrétní** náhodné veličiny  $X$  nazýváme (reálnou) funkci, pro kterou platí

$$f(x_k) = P(X = x_k) .$$

Často při zápisu pravděpodobnostní funkce symbol  $f(x_k)$  vynecháváme a tuto funkci označujeme pouze  $P(X = x_k)$ . Čísla  $P(X = x_k)$  jsou hodnoty pravděpodobnostní funkce. Jejich význam je v tom, že kolem nich kolísají relativní četnosti hodnot náhodné veličiny  $X$ , vypočtené ze sérií pokusů.

Pro pravděpodobnostní funkci platí:

1.  $0 \leq P(X = x_k) \leq 1$  , protože pravděpodobnost nabývá hodnot pouze z intervalu  $\langle 0 ; 1 \rangle$
2. Pro všechny ostatní reálná čísla  $x$ , **nepatřící** do oboru hodnot veličiny  $X$ , je pravděpodobnostní funkce rovna **nule**.
3.  $\sum_{\forall k} P(X = x_k) = 1$  , pro všechna  $x_k$  z oboru hodnot náhodné veličiny  $X$ .

Ze vztahu (9) vyplývá, že

**Distribuční funkci diskrétní** náhodné veličiny  $X$  lze pro každé reálné číslo  $x$  vyjádřit předpisem

$$F(x) = \sum_{x_k \leq x} P(X = x_k) , \quad (11)$$

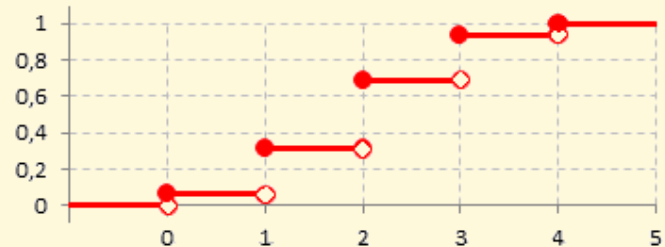
který vyjadřuje, že sčítáme pravděpodobnosti  $P(X = x_k)$  ve všech bodech  $x_k$ , ležících v intervalu  $(-\infty; x)$ . Spojením vzorců (10) a (11) dostáváme

$$P(X \in J) = \sum_{x_k \in J} P(X = x_k), \quad (12)$$

což vyjadřuje: pravděpodobnost, že diskrétní náhodná veličina  $X$  nabude některé hodnoty z intervalu  $J$  určíme tak, že sečteme pravděpodobnosti  $P(X = x_k)$  disjunktní jevů  $\{X = x_k\}$ , kde body  $x_k$  leží v intervalu  $J$ .

Distribuční funkci diskrétní náhodné veličiny lze znázornit stupňovitou funkcí, mající v bodech  $x_k$  skoky o velikostech  $P(X = x_k)$ . Mimo těchto bodů nabývá konstantních hodnot.

Známe-li hodnoty distribuční funkce, pak hodnoty pravděpodobnostní funkce jsou rovny velikostem „skoků“ distribuční funkce.



**Příklad:** Sportovní střelec má tři náboje. Na terč vystřelí postupně třikrát, přičemž střelbu ukončí buď zásahem terče (při němž je terč zničen a on nemá na CO střílet) nebo spotřebováním všech nábojů (již nemá ČÍM střílet). Pravděpodobnost zásahu prvním výstřelem je 0,6 a po každém výstřelu se zvýší o 0,1 (tak zvané se zastřeluje). Jaké jsou zákony rozdělení pro počet zbylých nábojů?

**Řešení:** Pokus je postupná střelba na terč končící prvním zásahem nebo spotřebováním všech nábojů. Jevy, které při pokusu mohou nastat uvedeme pro přehlednost v následující tabulce. Vyjádříme je pomocí elementárních jevů  $Z_i$  (pruhem nad písmenem budeme tak jako dříve označovat opačný jev).

Jevy	z			
$\overline{Z_1} \cap \overline{Z_2} \cap \overline{Z_3}$				
$\overline{Z_1} \cap \overline{Z_2} \cap Z_3$				
$\overline{Z_1} \cap Z_2$				
$Z_1$				

jev  $Z_i$ : terč je **Z**asažen výstřelem s pořadovým číslem  $i$  ( $i = 1, 2, 3$ ),

jev  $\overline{Z_i}$ : terč **není** zasažen výstřelem s pořadovým číslem  $i$ .

Do druhého sloupce tabulky (označeným písmenem **z**) zapíšeme počet zbylých (ne-spotřebovaných) nábojů.

zadání



Jevy	z	pravd.		
$\overline{Z_1} \cap \overline{Z_2} \cap \overline{Z_3}$	0			
$\overline{Z_1} \cap \overline{Z_2} \cap Z_3$	0			
$\overline{Z_1} \cap Z_2$	1			
$Z_1$	2			

jev  $Z_i$ : terč je **Z**asažen výstřelem s pořadovým číslem  $i$  ( $i = 1, 2, 3$ ),

jev  $\overline{Z_i}$ : terč **není** zasažen výstřelem s pořadovým číslem  $i$ .

Do druhého sloupce tabulky (označeným písmenem **z**) zapíšeme počet zbylých (ne-spotřebovaných) nábojů.

Dle **zadání**:  $P(Z_1) = 0,6$  ,  $P(Z_2) = 0,7$  ,  $P(Z_3) = 0,8$ .

Jevy	z	pravd.		
$\overline{Z_1} \cap \overline{Z_2} \cap \overline{Z_3}$	0			
$\overline{Z_1} \cap \overline{Z_2} \cap Z_3$	0			
$\overline{Z_1} \cap Z_2$	1			
$Z_1$	2	0,6		
$\Sigma$				

jev  $Z_i$ : terč je **Z**asažen výstřelem s pořadovým číslem  $i$  ( $i = 1, 2, 3$ ),

jev  $\overline{Z}_i$ : terč **není** zasažen výstřelem s pořadovým číslem  $i$ .

Do druhého sloupce tabulky (označeným písmenem **z**) zapíšeme počet zbylých (ne-spotřebovaných) nábojů.

Dle **zadání**:  $P(Z_1) = 0,6$ ,  $P(Z_2) = 0,7$ ,  $P(Z_3) = 0,8$ . Potom  $P(\overline{Z}_1) = 0,4$ ,  $P(\overline{Z}_2) = 0,3$ ,  $P(\overline{Z}_3) = 0,2$

Jevy	z	pravd.		
$\overline{Z_1} \cap \overline{Z_2} \cap \overline{Z_3}$	0			
$\overline{Z_1} \cap \overline{Z_2} \cap Z_3$	0			
$\overline{Z_1} \cap Z_2$	1			
$Z_1$	2	0,6		
$\Sigma$				

jev  $Z_i$ : terč je **Z**asažen výstřelem s pořadovým číslem  $i$  ( $i = 1, 2, 3$ ),

jev  $\overline{Z_i}$ : terč **není** zasažen výstřelem s pořadovým číslem  $i$ .

Do druhého sloupce tabulky (označeným písmenem **z**) zapíšeme počet zbylých (ne-spotřebovaných) nábojů.

Dle **zadání**:  $P(Z_1) = 0,6$ ,  $P(Z_2) = 0,7$ ,  $P(Z_3) = 0,8$ . Potom  $P(\overline{Z_1}) = 0,4$ ,  $P(\overline{Z_2}) = 0,3$ ,  $P(\overline{Z_3}) = 0,2$  a  $P(\{\overline{Z_1} \cap \overline{Z_2} \cap \overline{Z_3}\}) = P(\overline{Z_1}) \cdot P(\overline{Z_2}) \cdot P(\overline{Z_3})$  protože jde o nezávislé jevy (to jestli střelec druhým výstřelem terč zasáhne, není ovlivněno jeho prvním výstřelem, ...)

Jevy	$z$	pravd.	$P(X = z)$	
$\overline{Z}_1 \cap \overline{Z}_2 \cap \overline{Z}_3$	0	0,024		
$\overline{Z}_1 \cap \overline{Z}_2 \cap Z_3$	0	0,096		
$\overline{Z}_1 \cap Z_2$	1	0,28		
$Z_1$	2	0,6		
$\Sigma$	1			

jev  $Z_i$ : terč je **Z**asažen výstřelem s pořadovým číslem  $i$  ( $i = 1, 2, 3$ ),

jev  $\overline{Z}_i$ : terč **není** zasažen výstřelem s pořadovým číslem  $i$ .

Do druhého sloupce tabulky (označeným písmenem **z**) zapíšeme počet zbylých (ne-spotřebovaných) nábojů.

Dle **zadání**:  $P(Z_1) = 0,6$ ,  $P(Z_2) = 0,7$ ,  $P(Z_3) = 0,8$ . Potom  $P(\overline{Z}_1) = 0,4$ ,  $P(\overline{Z}_2) = 0,3$ ,  $P(\overline{Z}_3) = 0,2$  a  $P(\{\overline{Z}_1 \cap \overline{Z}_2 \cap \overline{Z}_3\}) = P(\overline{Z}_1) \cdot P(\overline{Z}_2) \cdot P(\overline{Z}_3)$  protože jde o nezávislé jevy (to jestli střelec druhým výstřelem terč zasáhne, není ovlivněno jeho prvním výstřelem, ...)

Jevy	$z$	pravd.	$P(X = z)$	$F(z)$
$\bar{Z}_1 \cap \bar{Z}_2 \cap \bar{Z}_3$	0	0,024	0,12	
$\bar{Z}_1 \cap \bar{Z}_2 \cap Z_3$	0	0,096		
$\bar{Z}_1 \cap Z_2$	1	0,28	0,28	
$Z_1$	2	0,6	0,6	
$\Sigma$	1			

jev  $Z_i$ : terč je **Z**asažen výstřelem s pořadovým číslem  $i$  ( $i = 1, 2, 3$ ),

jev  $\bar{Z}_i$ : terč **n**ení zasažen výstřelem s pořadovým číslem  $i$ .

Do druhého sloupce tabulky (označeným písmenem **z**) zapíšeme počet zbylých (ne-spotřebovaných) nábojů.

Dle **zadání**:  $P(Z_1) = 0,6$ ,  $P(Z_2) = 0,7$ ,  $P(Z_3) = 0,8$ . Potom  $P(\bar{Z}_1) = 0,4$ ,  $P(\bar{Z}_2) = 0,3$ ,  $P(\bar{Z}_3) = 0,2$  a  $P(\{\bar{Z}_1 \cap \bar{Z}_2 \cap \bar{Z}_3\}) = P(\bar{Z}_1) \cdot P(\bar{Z}_2) \cdot P(\bar{Z}_3)$  protože jde o nezávislé jevy (to jestli střelec druhým výstřelem terč zasáhne, není ovlivněno jeho prvním výstřelem, ...)

Všimněte si, že když náhodná veličina  $X$  přiřazuje výsledkům pokusu totéž číslo, je hodnota pravděpodobnostní funkce v tomto čísle rovna součtu pravděpodobností těchto výsledků.

Jevy	$z$	pravd.	$P(X = z)$	$F(z)$
$\overline{Z_1} \cap \overline{Z_2} \cap \overline{Z_3}$	0	0,024	0,12	0,12
$\overline{Z_1} \cap \overline{Z_2} \cap Z_3$	0	0,096		
$\overline{Z_1} \cap Z_2$	1	0,28	0,28	0,4
$Z_1$	2	0,6	0,6	1
$\Sigma$	1	1		

jev  $Z_i$ : terč je **Z**asažen výstřelem s pořadovým číslem  $i$  ( $i = 1, 2, 3$ ),

jev  $\overline{Z}_i$ : terč **n**ení zasažen výstřelem s pořadovým číslem  $i$ .

Do druhého sloupce tabulky (označeným písmenem **z**) zapíšeme počet zbylých (ne-spotřebovaných) nábojů.

Dle **zadání**:  $P(Z_1) = 0,6$ ,  $P(Z_2) = 0,7$ ,  $P(Z_3) = 0,8$ . Potom  $P(\overline{Z}_1) = 0,4$ ,  $P(\overline{Z}_2) = 0,3$ ,  $P(\overline{Z}_3) = 0,2$  a  $P(\{\overline{Z}_1 \cap \overline{Z}_2 \cap \overline{Z}_3\}) = P(\overline{Z}_1) \cdot P(\overline{Z}_2) \cdot P(\overline{Z}_3)$  protože jde o nezávislé jevy (to jestli střelec druhým výstřelem terč zasáhne, není ovlivněno jeho prvním výstřelem, ...)

Všimněte si, že když náhodná veličina  $X$  přiřazuje výsledkům pokusu totéž číslo, je hodnota pravděpodobnostní funkce v tomto čísle rovna součtu pravděpodobností těchto výsledků.

Ukažme si některé výsledky z tabulky:

Hodnota  $P(X = 2) = 0,6$  říká, že **pokud by se tento pokus opakoval vícekrát**, tak asi v 60 % těchto pokusů zůstanou střelci dva náboje. Číslo  $F(0) = P(X \leq 0) = 0,12$  značí, že asi ve 12 % těchto pokusů zůstane střelci žádný a méně  $\Rightarrow$  tedy žádný náboj.

Z tabulky lze získat i další informace. Třeba pravděpodobnost, že střelci zůstane alespoň jeden (jeden nebo dva) náboj. Tento jev označíme  $\{X \geq 1\}$  a jeho pravděpodobnost vypočteme podle vzorce (12), kdy:

$$P(X \geq 1) = P(X = 1) + P(X = 2) = 0,28 + 0,6 = 0,88.$$

To značí, že asi v 88 % pokusů zůstane střelci aspoň jeden náboj.

Jestliže se dále zajímáme o to, kolik procent z pokusů, v nichž zbyl střelci alespoň jeden náboj, připadá na jev, že střelci zůstane právě jeden náboj, pak tyto pravděpodobnosti vypočteme pomocí vzorce (5) pro **podmíněnou pravděpodobnost**.

$$P(X = 1 | X \geq 1) = \frac{P(\{X = 1\} \cap \{X \geq 1\})}{P(X \geq 1)} = \frac{P(X = 1)}{P(X \geq 1)} = \frac{0,28}{0,88} \doteq 0,318,$$

což lze interpretovat takto: V těch pokusech, v nichž zbyl střelci alespoň jeden náboj, je asi 31,8 % pokusů, v nichž mu zbyl právě jeden náboj.

## Náhodné veličiny spojitého typu

Někdy zkráceně říkáme jen **spojité náhodné veličiny** mohou (jak jsme uvedli na začátku kapitoly) nabývat libovolných hodnot z daného intervalu. Toto sice také není exaktní definice (stejně jako v případě diskrétní náhodné veličiny), ale nám opět plně postačuje.

Také u spojitě náhodné veličiny se užívá k jejímu popisu **distribuční funkce  $F(x)$** , kterou jsme zavedli vzorcem (9) a následně odvodili vzorec (10) pro výpočet pravděpodobnosti, že náhodná veličina  $X$  nabude nějaké hodnoty z daného intervalu.

A protože spojitá náhodná veličina může nabývat libovolné ( $\Leftarrow$  spojitá) hodnoty (na rozdíl od diskrétní veličiny, která může nabývat jen některých izolovaných hodnot), můžeme také uvažovaný interval stále zmenšovat, až bude mít **nekonečně malou** šířku ( $\Rightarrow$  limita). Tedy vzorec (10) můžeme také psát:

$$\lim_{h \rightarrow 0} P(x < X \leq x + h) = \lim_{h \rightarrow 0} [F(x + h) - F(x)]$$

Pokud dané limity budeme vyčíslovat pro  $h \rightarrow 0$ , tak se levá strana rovnice bude blížit k následující pravděpodobnosti  $P(x < X \leq x + h) \rightarrow P(X = x)$  a pravá strana rovnice se bude blížit k nule:

$F(x + h) - F(x) \rightarrow 0$  (pro  $h \rightarrow 0$ ). Tedy z toho plyne, že

$$P(X = x) = 0$$

což odpovídá skutečnosti, že oborem hodnot spojitých náhodných veličin je nějaký interval, přičemž každý bod z tohoto intervalu má nulovou pravděpodobnost<sup>11</sup>.

Proto nemá smysl počítat u spojitě náhodné veličiny pravděpodobnostní funkci, kterou jsme zavedli u diskrétních náhodných veličin, ale na základě pravděpodobnostní funkce zavádíme jinou funkci, kterou nazýváme **hustota pravděpodobnosti**<sup>12</sup> nebo také někdy *frekvenční funkce*.

**Hustota pravděpodobnosti spojitě** náhodné veličiny  $X$  na intervalu  $\langle a; b \rangle$  je následující funkce:

$$f(x) = \lim_{h \rightarrow 0} \frac{P(x < X \leq x + h)}{h} = F'(x)$$

kde

pro  $x \notin \langle a; b \rangle$  je  $f(x) = 0$ ;  $F(x)$  je distribuční funkce náhodné veličiny  $X$ , a

$x, x + h \in \langle a; b \rangle$ .

<sup>11</sup> Neznamená to však, že náhodná veličina  $X$  nemůže hodnotu  $x$  nikdy dosáhnout. Ale je to matematické vystižení faktu, že hodnot, kterých náhodná veličina  $X$  nabýt může, je tak velké množství, že **pravděpodobnost**, že nabyde právě jednu, konkrétně vybranou, **je příliš nepatrná, v limitě nulová**.

<sup>12</sup> Protože hustotu pravděpodobnosti zavádíme jako následující (speciální) limitu, která se nazývá (jak víme z kurzu o diferenciálním počtu) **derivate** distribuční funkce, může se stát, že pro některou hodnotu  $x$  je hodnota hustoty větší jak jedna:  $f(x) > 1$ . Uvedená limita však žádnou pravděpodobnost nevyjadřuje.

**Vždy** ale bude hustota pravděpodobnosti **nezáporná** ( $0 \leq f(x)$ ,  $\forall x$ ), protože distribuční funkce je neklesající (viz její **druhá** vlastnost).



**Distribuční funkci** u spojitě náhodné veličiny určujeme analogicky jako u diskrétní náhodné veličiny, kde jsme používali vzorec (11). Pouze si musíme uvědomit, že nyní místo pravděpodobnostní funkce  $P(X = x)$  máme k dispozici hustotu pravděpodobnosti  $f(x)$  a (sumou) vlastně sčítáme **nekonečně mnoho nekonečně malých** veličin, což vede na následující integrál<sup>13</sup>:

**Distribuční funkce spojitě** náhodné veličiny  $X$  je následující (primitivní) funkce

$$F(x) = \int_{-\infty}^x f(t) dt \quad (13)$$

kde  $f(t)$  je hustotou pravděpodobnosti této náhodné veličiny.

Spojením vzorců (10) a (13) dostáváme vzorec pro výpočet pravděpodobnosti, kdy spojitá náhodná veličina  $X$  nabude některé hodnoty z intervalu  $J = \langle x_1; x_2 \rangle$  ( $J = (x_1; x_2)$ ,  $J = \langle x_1; x_2 \rangle$ ,  $J = (x_1; x_2)$ )

$$P(X \in J) = \int_{x_1}^{x_2} f(x) dx = F(x_2) - F(x_1). \quad (14)$$

Z vlastností integrálu plyne, že vůbec nezáleží na tom, zda je interval  $J$  uzavřený, otevřený nebo polootevřený.

Protože hustotu pravděpodobnosti  $f(x)$  v bodě  $x$  získáme z distribuční funkce  $F(x)$  její derivací,

$$f(x) = F'(x) \quad (15)$$

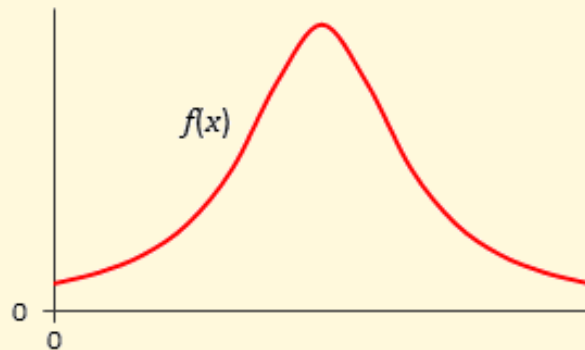
<sup>13</sup> Při praktickém výpočtu se dolní mez  $-\infty$  nahrazuje skutečnou dolní mezí, od které je náhodná veličina  $X$  definována.

spočívá význam hustoty pravděpodobnosti v tom, že vyjadřuje velikost okamžité změny distribuční funkce v daném bodě, tedy „okamžitou“ velikost nárůstu (či poklesu) pravděpodobnosti v tomto bodě. Nebo ještě jinak — jak **hustě** jsou ostatní hodnoty náhodné veličiny  $X$  rozmístěny okolo tohoto bodu.

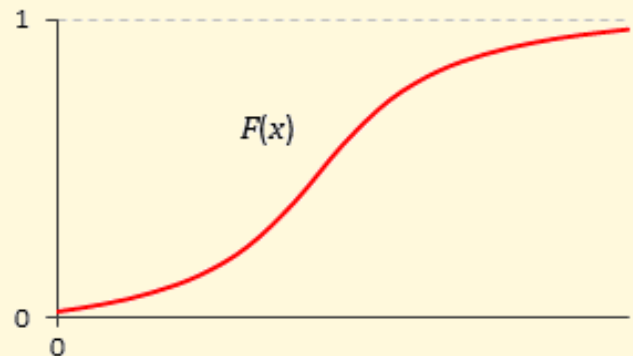
Jako příklad uveďme náhodnou veličinu  $X$ , která označuje výšku náhodně vybraného muže v populaci České republiky. Pokud bychom rozdělili všechny muže podle jejich výšek do intervalů po deseti centimetrech, pak do každého z těchto intervalů „patří“ velmi mnoho mužů, ale v intervalu (180 cm ; 190 cm) jich bude podstatně více jak například v intervalu (140 cm ; 150 cm).

Hustota pravděpodobnosti u spojitě náhodné veličiny je analogická pravděpodobnostní funkci u diskrétní náhodné veličiny. Ovšem teď to již nejsou izolované body, ale na nějakém intervalu spojitá křivka.

Podobně i distribuční funkce již nebude „rozkouskovaná“.



Hustota pravděpodobnosti  
spojitě náhodné veličiny



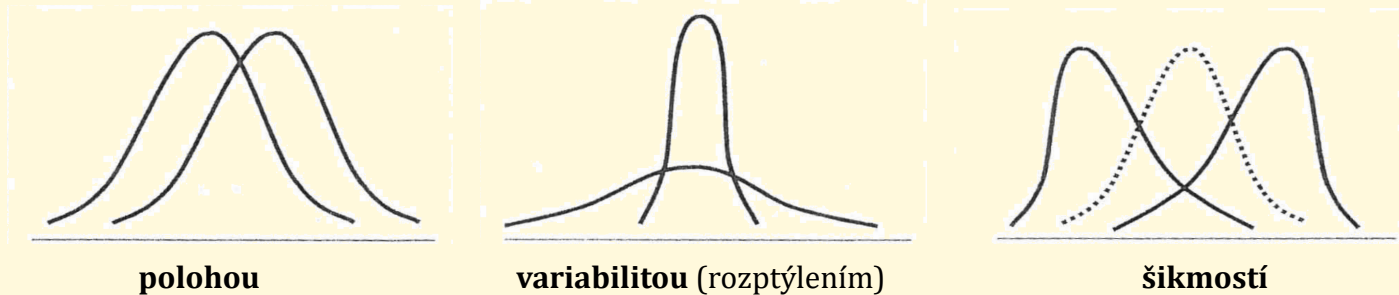
Distribuční funkce  
spojitě náhodné veličiny

Můžeme také říci, že náhodná veličina je spojitá, pokud má spojitou distribuční funkci.

## 4. Číselné charakteristiky náhodných veličin

Distribuční funkce  $F(x)$  s pravděpodobnostní funkcí  $P(X = x)$  (u spojité náhodné veličiny je to hustota pravděpodobnosti) popisují rozdělení pravděpodobností hodnot příslušné diskrétní náhodné veličiny  $X$  vyčerpávajícím způsobem. Tyto funkce jsou však často poměrně složité a jejich určení pracné. Proto je někdy výhodné shrnout celkovou informaci o náhodné veličině do několika čísel, která charakterizují další její vlastnosti a umožňují srovnávání různých náhodných veličin. Tato čísla se nazývají **charakteristikami náhodné veličiny**.

Obrázek 2: Rozdělení spojitých náhodných proměnných, které se odlišují



My si uvedeme pouze střední hodnotu, rozptyl a směrodatnou odchylku. Další charakteristiky, které charakterizují podrobnější vlastnosti náhodné veličiny (například koeficient šikmosti a koeficient špičatosti) uvádět nebudeme.

**Střední hodnota  $E(X)$**  (také očekávaná hodnota, *expected value*) je pro diskrétní náhodnou proměnnou definována vztahem

$$E(X) = \sum_{\forall k} x_k \cdot P(X = x_k) \quad (16)$$

a pro spojitou náhodnou proměnnou vztahem

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Předpokládáme, že jak suma tak integrál konvergují.

Střední hodnota charakterizuje polohu hodnot náhodné proměnné, podobně jako aritmetický průměr ve statistice nebo těžiště ve fyzice. Střední hodnotu si můžeme představit jako „pomyslný střed“ oboru hodnot náhodné veličiny  $X$ , kolem kterého „kolísají“ jednotlivé hodnoty této veličiny.

**Vlastnosti střední hodnoty  $E(X)$**  (pokud uvedené střední hodnoty existují) pro libovolné **konstanty**  $a$ ,  $b$ ,  $c$  a náhodné veličiny  $X$  a  $Y$  jsou tyto:

1.  $E(a) = a$
2.  $E(b \cdot X \pm c \cdot Y) = b \cdot E(X) \pm c \cdot E(Y)$
3.  $E(X \cdot Y) = E(X) \cdot E(Y)$ , pokud jsou  $X$  a  $Y$  **nezávislé**.

V následujícím příkladu ukážeme „užitečnost“ znalosti výpočtu střední hodnoty v hazardní hře.

**Příklad** Hráč vsadí částku  $a$  korun na učitě číslo na hrací kostce. Jinak řečeno zvolí si jedno číslo z následujících šesti:  $\{1, 2, 3, 4, 5, 6\}$ . Poté bankéř hodí **tři** kostky. Jestliže vsazené číslo nepadne na žádné kostce, vklad propadá. Když vsazené číslo padne na  $r$  kostkách, pak hráč dostane **výhru** ( $r \cdot a$ ) korun a **vsazenou částku zpět**. Je tato hra pro hráče výhodná?

**Řešení:** Hod kostkou považujeme za pokus. Padne-li na první kostce vsazené číslo, pak tento jev, který označíme  $A$  (na druhé  $B$  a na třetí  $C$ ), má pravděpodobnost  $\frac{1}{6}$ . Hod třemi kostkami, pokud se provádí regulérně, považujeme za **Bernoulliou** posloupnost nezávislých pokusů, kde  $n = 3$ .

Označíme-li  $D_k$  jev, že při hodu třemi kostkami je na  $k$  kostkách vsazené číslo ( $k = 0, 1, 2, 3$ ), lze pravděpodobnost tohoto jevu označenou  $P(D_k)$  spočítat pomocí vzorce (4).

Jevy  $D_k$  (složené z elementárních jevů  $A, B, C$ ) a jejich pravděpodobnosti  $P(D_k)$  jsou v prvních dvou sloupcích následující **tabulky**.

$k$	$D_k$	$P(D_k)$	$x_k$	$P(X = x_k)$	$x_k \cdot P(X = x_k)$
0	$D_0 = \bar{A} \cap \bar{B} \cap \bar{C}$	$\frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6}$	$0 \cdot a + 0$	$\frac{125}{216}$	$0 \cdot \frac{125}{216} = 0$
1	$D_1 = (A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)$	$3 \cdot \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{5}{6}$	$1 \cdot a + a$	$\frac{75}{216}$	$2a \cdot \frac{75}{216} = \frac{150a}{216}$
2	$D_2 = (A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)$	$3 \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6}$	$2 \cdot a + a$	$\frac{15}{216}$	$3a \cdot \frac{15}{216} = \frac{45a}{216}$
3	$D_3 = A \cap B \cap C$	$\frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6}$	$3 \cdot a + a$	$\frac{1}{216}$	$4a \cdot \frac{1}{216} = \frac{4a}{216}$
$\Sigma$		$\frac{125+75+15+1}{216} = 1$	výhra + vklad	1	$\frac{199a}{216}$

Náhodnou veličinou  $X$  (její jednotlivé možné hodnoty  $x_k$ ) označme částku, kterou hráč po každé hře obdrží. Její hodnoty přiřazené výsledkům pokusu  $D_k$  napíšeme do třetího sloupce **tabulky**. V případě prohry nic (nula), v případě, že uhodne, tak výhru a vklad. Ve čtvrtém sloupci jsou hodnoty pravděpodobnosti  $P(X = x_k)$  náhodné veličiny  $X$ , které odpovídají pravděpodobnosti výsledkům příslušných jevů.

Podle vzorce (16) je střední hodnota  $E(X) = \frac{199}{216} \cdot a$  rovna součtu hodnot v posledním sloupci.

Jako kritérium výhodnosti hry lze vzít rozdíl mezi střední hodnotou vyplacených částek a vsazenou částkou  $a$ . Podle tohoto kritéria dostaneme:

$$E(X) - a = \frac{199a}{216} - a \doteq -0,078\,704\,a.$$

Protože rozdíl mezi střední hodnotou vyplacených částek a vsazenou částkou  $a$  je přibližně  $-0,079 \cdot \text{částka } a$

je tato hra pro hráče nevýhodná (ale pro bankéře je naopak výhodná), protože ze vsazené částky v každé hře (usuzujeme z hodnoty  $E(X) \Rightarrow$  v průměru při mnoha opakováních) ztrácí hráč průměrně necelých 8 % svého vkladu.

**Rozptyl  $D(X)$**  (také variance  $\text{var}(X)$  či *disperze*) zavedeme jako  $D(X) = E\{[X - E(X)]^2\}$ .  
Z vlastností střední hodnoty plyne:

$$D(X) = E(X^2) - [E(X)]^2. \quad (17)$$

Pro diskrétní náhodnou proměnnou pak platí

$$D(X) = \sum_{\forall k} [x_k - E(x)]^2 \cdot P(X = x_k) = \sum_{\forall k} x_k^2 \cdot P(X = x_k) - [E(X)]^2$$

a pro spojitou náhodnou proměnnou platí

$$D(X) = \int_{-\infty}^{\infty} [x - E(x)]^2 \cdot f(x) dx = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - [E(X)]^2$$

Předpokládáme, že jak sumy tak integrály konvergují.

Rozptyl vyjadřuje, jak mnoho jsou hodnoty náhodné proměnné rozptýleny kolem střední hodnoty. Rozptyl vychází v „kvadratických“ jednotkách, přičemž zvýrazňuje extrémy (váhu těch bodů, které jsou více vzdáleny od střední hodnoty). Abychom srovnali tyto jednotky, počítáme ještě charakteristiku zvanou směrodatná odchylka. Ta má jednotky shodné s jednotkami  $E(X)$ .

**Směrodatná odchylka  $\sigma(X)$**  je definována jako druhá odmocnina z rozptylu.

$$\sigma(X) = \sqrt{D(X)}$$

## 5. Používaná rozdělení náhodných veličin

### 5.1. Opakování již dříve uvedených pojmů

Souhrn všech hodnot, kterých náhodná veličina může nabývat, se nazývá **obor hodnot náhodné veličiny**. Některé náhodné veličiny nabývají pouze izolovaných hodnot (například výsledek hodu kostkou). Takovou náhodnou veličinu nazýváme **diskrétní**. Jindy tvoří obor hodnot náhodné veličiny nějaký číselný interval (například kurs koruny vůči euru). V takovém případě hovoříme o **spojité** náhodné veličině. O diskrétní i spojitě náhodné veličině jsme již mluvili, ale opakování vůbec není na škodu.

Chceme-li popsat chování náhodné veličiny, nestačí pouze uvést obor hodnot, kterých může nabývat. Některé hodnoty z oboru se totiž mohou vyskytovat s větší, jiné s menší pravděpodobností. Pravidlo, kterým se tato pravděpodobnost řídí, se nazývá **zákon rozdělení** (rozložení) náhodné veličiny.

**Zákon rozdělení** je vlastně pravidlo (funkce, předpis), které každé hodnotě (nebo skupině hodnot) z oboru hodnot náhodné veličiny přiřazuje pravděpodobnost jejich výskytu.

V konkrétní statistické praxi se vychází z toho, že velké skupiny náhodných pokusů mají stejné pravděpodobnostní chování, které závisí na jejich charakteru. Probereme nyní postupně některé typy rozdělení pravděpodobnosti, které mají náhodné veličiny, popisující v jistém smyslu analogické náhodné pokusy. Na příkladech budeme vždy ilustrovat základní situace.

### 5.2. Diskrétní náhodná veličina — některá její rozdělení

Zákon rozdělení **diskrétní** náhodné veličiny  $X$  lze nejjednodušeji vyjádřit pomocí **pravděpodobnostní funkce**, o které jsme již mluvili. Druhou možností, jak vyjádřit rozložení pravděpodobnosti diskrétní náhodné veličiny  $X$ , je pomocí **distribuční funkce**  $F(x)$ , což jsme si již také říkali.



## Binomické rozdělení

Binomické rozdělení má náhodná veličina  $X$ , která představuje  $k$  výskytů daného jevu v posloupnosti  $n$  nezávislých pokusů, přičemž  $p$  je pravděpodobnost (stále stejná) nastoupení daného jevu v jediném pokusu.

Jeho **pravděpodobnostní funkce** je dána (viz vzorec (4)) vztahem

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} \quad (18)$$

a **charakteristiky** jsou

$$E(X) = n \cdot p \quad D(X) = n \cdot p \cdot (1 - p) \quad (19)$$

Binomické se nazývá proto, že hodnoty funkce  $P(X = k)$  určené podle vzorce (18) jsou členy v binomickém rozvoji výrazu  $[p + (1 - p)]^n$ .

Protože jsou výpočty hodnot  $\binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$  pro velká  $n$  a  $k$  početně značně náročné, lze k jejich výpočtům použít počítačové programy (např. *Excel 2010*), případně lze pro velmi velký rozsah pokusů ( $n$  je v řádu stovek a víc) toto rozdělení díky centrální limitní větě aproximovat **normálním rozdělením**  $N(E; D)$ , o kterém bude řeč dále.

**Příklad** skriptu [4, číslo 23].

V krabici jsou dvě zelené a tři černé koule. Náhodně vybereme jednu, zjistíme její barvu a **vrátíme** kouli do krabice. Toto provedeme ještě dvakrát. Náhodná veličina  $X$  představuje počet vybraných černých koulí.

**Příklad.** Je pravděpodobnější vyhrát v tenise se stejně silným soupeřem **tři zápasy ze čtyř** nebo je pravděpodobnější vyhrát **šest zápasů z osmi**?

**Řešení:** Tenisové zápasy jsou vlastně opakované nezávislé pokusy. Mohou nastat pouze dva výsledky v jednom utkání: buďto vyhraje nebo prohraje. Hrajeme-li se stejně silným soupeřem, je pravděpodobnost výhry v každém zápase  $p = 0,5$ . V jiném zápase je zase  $0,5 \Rightarrow$  nemění se. Tedy náhodná veličina  $X$ , která určuje počet vyhraných zápasů má binomické rozdělení.

**3 ze 4** Do vzorce (18):  $P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$  dosadíme  $k = 3$ ,  $n = 4$  a  $p = 0,5$ .

$$P(X = 3) = \binom{4}{3} \cdot 0,5^3 \cdot (1 - 0,5)^{4-3} = \frac{4!}{3! \cdot (4-3)!} \cdot 0,5^3 \cdot 0,5 = \frac{4 \cdot 3!}{3! \cdot 1} \cdot 0,5^4 = 4 \cdot 0,0625 = 0,25$$

**6 z 8** Do vzorce (18) dosadíme  $k = 6$ ,  $n = 8$  a  $p = 0,5$ .

$$P(X = 6) = \binom{8}{6} \cdot 0,5^6 \cdot (1 - 0,5)^{8-6} = \frac{8!}{6! \cdot (8-6)!} \cdot 0,5^6 \cdot 0,5^2 = \frac{8 \cdot 7 \cdot 6!}{6! \cdot 2!} \cdot 0,5^8 = 0,109375$$

Je tedy pravděpodobnější zvítězit ve třech zápasech ze čtyř.

V praxi se ale mimo případů — kdy můžeme rozhodnout naprosto přesně, kolikrát daný jev nastal a kolikrát daný jev nenastal (Například: Danou křižovatkou za daný čas projelo  $a$  automobilů se spalovacím motorem. Jestliže  $b$  z nich mělo benzínový motor, pak  $a - b$  mělo jiný typ motoru: *naftový, na LPG, na vodík, ...*) — vyskytují také případy typu:

- Při bouřce bylo XYZ blesků — a kolik blesků NEBYLO?
- V sobotu se v porodnici narodilo ZYX dětí — a kolik se jich NENARODILO?
- atd.

V těchto případech nemůžeme binomické rozdělení použít. Proto jsou známa i jiná rozdělení pravděpodobnosti (například Poissonovo), než my si uvádíme.

## Hypergeometrické rozdělení – kontrola jakosti

Hypergeometrické rozdělení má náhodná veličina  $X$ , která představuje počet  $k$  prvků s vlastností  $A$  ve skupině  $n$  prvků vybraných z množiny  $N$  prvků, z nichž  $M$  má vlastnost  $A$ .

Jeho **pravděpodobnostní funkce** je dána vztahem

$$P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}} \quad (20)$$

a **charakteristiky** jsou

$$E(X) = \frac{n \cdot M}{N} \quad D(X) = \frac{n \cdot M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1}$$

Hypergeometrické rozdělení (někdy používáme i termín statistický výběr bez opakování) se používá například ve statistické kontrole jakosti (hlavně při zkoumání jakosti malého počtu výrobků, nebo když kontrola má charakter destrukční zkoušky – při kontrole je výrobek zničen) a jako pravděpodobnostní model některých her (např. Sportka apod.).

A protože nemá smysl kontrolovat jeden výrobek třikrát (výběr bez opakování), jde vlastně o to, že náhodně vybrané prvky určené ke kontrole nevracíme zpět do základního souboru, který je tvořen všemi výrobky. Jednotlivé pokusy jsou pak závislé (pravděpodobnost výskytu vlastnosti  $A$  v určitém pokusu závisí na výsledcích v předcházejících pokusech).

**Poznámka** Jestliže rozsah  $N$  je velký a  $n$  a  $M/N$  se nemění, blíží se hypergeometrické rozdělení binomickému. To znamená, že pro velká  $N$  můžeme zanedbat rozdíl mezi výběrem bez vracení a s vracením. V praxi se rozhodujeme podle hodnoty tak zvaného **výběrového poměru** ( $n/N$ ). Je-li tento poměr menší než 0,05, lze hypergeometrické rozdělení nahradit binomickým s parametry  $n$  a  $p = M/N$ .

**Příklad** skripta [4, číslo 24].

Mezi devíti ( $N$ ) žárovkami určenými k pevnostním zkouškám jsou tři ( $M$ ) nižší jakosti, které zkoušky nevydrží. Tedy ostatní ( $N-M$ ) žárovky by pevnostní zkoušky měly vydržet. Jaká je pravděpodobnost, že mezi čtyřmi ( $n$ ) náhodně vybranými žárovkami nebude žádná ( $k$ ) nižší jakosti?

Vraťme se nyní (o jednu kapitolu moudřejší) opět k předchozímu příkladu č. 23 — který jsme řešili v souvislosti s binomickým rozdělením — a uvažujme jej ve dvou modifikacích (poněkud upravíme zadání ve skriptech):

**Příklad** skripta [4, číslo 23].

V krabici jsou dvě zelené a tři černé koule. Náhodně vybereme jednu, zjistíme její barvu a:

1. **vrátíme** ji do krabice  $\Rightarrow X$  má binomické rozdělení;
2. **NEvrátíme** ji do krabice  $\Rightarrow X$  má hypergeometrické rozdělení.

Toto provedeme ještě dvakrát. Náhodná veličina  $X$  představuje počet vybraných černých koulí.

## 5.3. Spojitá náhodná veličina — některá její rozdělení

### Normální rozdělení

Normální rozdělení  $N(\mu, \sigma^2)$  má náhodná veličina  $X$ , jejíž kolísání je způsobeno mnoha drobnými nezávislými vlivy, z nichž žádný samostatně není významný.

Jeho **hustota pravděpodobnosti** je dána vztahem

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{kde} \quad -\infty < x < \infty$$

a **charakteristiky** jsou

$$E(X) = \mu \quad D(X) = \sigma^2$$

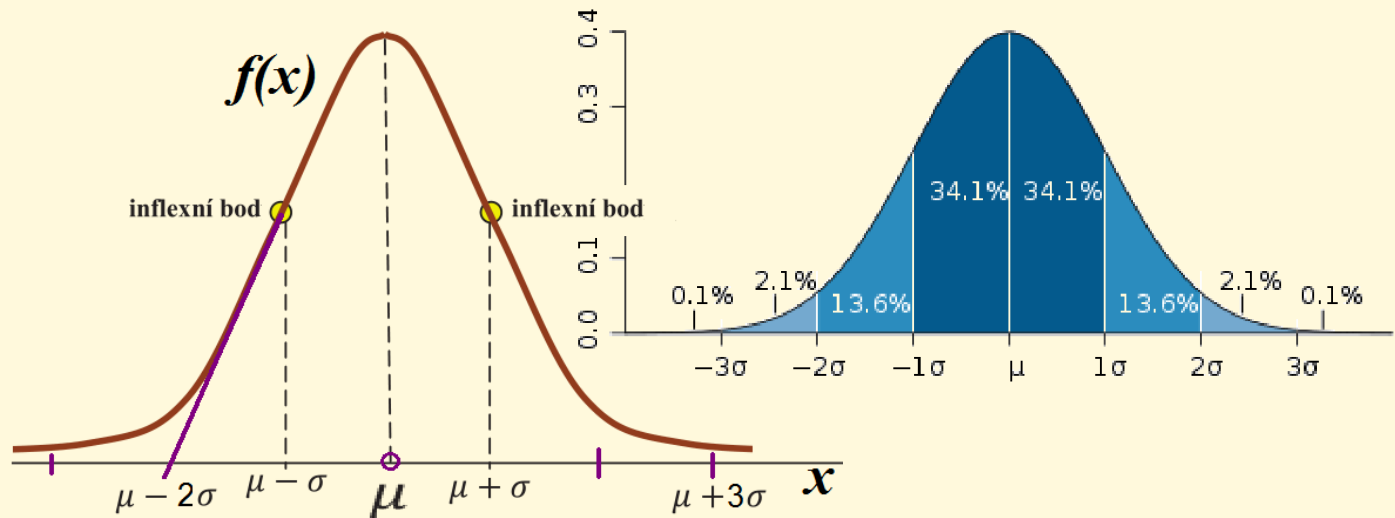
Normální rozdělení<sup>14</sup> mají mnohé náhodné veličiny — procentové změny v cenách akcií na dobře fungujících trzích, devizové výplatní poměry měn, chyby měření, rozměry výrobků při hromadné výrobě, rozptyl při střelbě a mnohé jevy ve fyzice, v biologii, v medicíně. Obecně lze říci, že je použitelné všude tam, kde hodnoty náhodné veličiny jsou ovlivněny působením velkého počtu nepatrných, vzájemně nezávislých nebo slabě závislých náhodných vlivů.

**Graf funkce  $f(x)$**  popisující hustotu pravděpodobnosti normálního rozdělení se nazývá **Gaussova křivka**<sup>15</sup> (Gaussův klobouk, zvonová funkce, angl. „bell curve“). Je charakteristická tím, že:

<sup>14</sup> Neznamená to, že by ostatní rozdělení byla **nenormální** či **abnormální**. Název pouze vyjadřuje skutečnost, že všechny soubory o velkém rozsahu, které byly zkoumány v době, kdy se tento název ujal, měly (alespoň přibližně) toto rozdělení (soubory o menších rozsazích se tehdy nezkoumaly). Bylo proto přirozené („normální“) očekávat, že i další v budoucnu studované soubory budou mít toto rozdělení.

<sup>15</sup> V roce 1733 uveřejnil **Abraham de Moivre** spisek, ve které popsal rovnici této křivky. Křivka (i její rovnice) upadla v zapomnutí a byla znovuobjevena jako „křivka chyb“ **Laplaceem** (se zápornými chybami se vypořádal pomocí absolutní hodnoty) a **Gaussem** (záporné znaménko u chyb odstranil umocněním *na druhou*) [14, str. 77–78].

- je symetrická kolem svislé přímky procházející bodem  $\mu$  v němž má funkce  $f(x)$  globální (absolutní) maximum;
- ve vzdálenostech  $\sigma$  vlevo a vpravo od bodu  $\mu$  má funkce  $f(x)$  inflexní body;
- tečny funkce  $f(x)$  sestrojené v bodech  $\mu \pm \sigma$  protínají vodorovnou osu v bodech  $\mu \pm 2\sigma$ ;
- ve vzdálenostech  $3\sigma$  se funkce  $f(x)$  téměř dotýká vodorovné osy.



Gaussova křivka

Parametr  $\sigma$  udává „horizontální“ vzdálenost inflexních bodů od střední hodnoty a tím i **šířku křivky**.

Pro normální rozdělení platí **pravidlo „tří sigma“**, kdy do intervalu  $\langle \mu - 3\sigma; \mu + 3\sigma \rangle$  padne přibližně 99,7 % všech hodnot náhodné proměnné  $X$ .

Tedy v uvedeném intervalu  $3\sigma$  ( $3\sigma$  na každou stranu od střední hodnoty  $\Rightarrow$  tento interval má délku rovnou  $6\sigma$ , proto se někdy můžete setkat i s jeho označováním „šest sigma“) jsou prakticky všechny hodnoty tohoto rozdělení. Toto pravidlo  $3\sigma$  je jedním ze základních principů, na nichž stojí kontrola kvality a jakosti (SPC — Statistics Process Control, ISO normy pro SPC).

Navíc:

do intervalu  $\langle \mu - 2\sigma; \mu + 2\sigma \rangle$  padne přibližně 95 % hodnot a

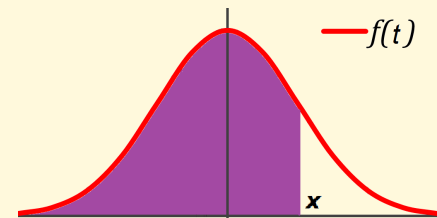
do intervalu  $\langle \mu - \sigma; \mu + \sigma \rangle$  přibližně 68,3 % hodnot.

Normální rozdělení je nejdůležitějším rozdělením spojitě náhodné proměnné. Jeho význam zvyšuje to, že se jím dají (za určitých podmínek) aproximovat i jiná rozdělení, ať spojitě či diskrétní náhodné proměnné (například binomické, chí-kvadrát, Poissonovo, Studentovo).

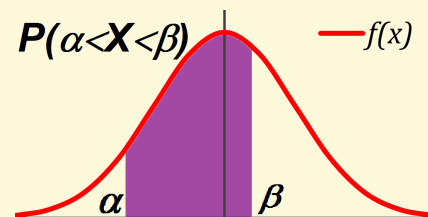
Jak jste si všimli, doposud jsme neuvedli distribuční funkci normálního rozdělení, kterýžto integrál neumíme analyticky vypočítat.

$$F(x) = \int_{-\infty}^x f(t) dt$$

Pokud si vzpomenete na aplikace určitého integrálu, které byly probírány v předmětu **Matematika**, tak určitým integrálem určíme velikost rovinné plochy (ve vedlejším obrázku vybarvené fialově) ohraničené zdola **souřadnou osou  $x$** , shora **hustotou pravděpodobnosti  $f(t)$** , zprava hodnotou  $x$  a vlevo jde plocha až do mínus nekonečna.



A proč se tolik zajímáme o hodnotu distribuční funkce? Protože pomocí ní a podle vzorce (14) dokážeme určit pravděpodobnost, že náhodná proměnná  $X$  patří do nějakého intervalu.



Například

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = F(\mu + 3\sigma) - F(\mu - 3\sigma) \doteq 0,997$$

což je pravidlo **tří sigma**, které jsme uvedli na předchozí straně.

Numerický výpočet integrálu uvedeného v předchozím příkladu bývá součástí nejrůznějších počítačových programů. A pro speciální případ normálního rozložení s nulovou střední hodnotou ( $\mu = 0$ ) a směrodatnou odchylkou rovnou jedné ( $\sigma = 1$ ) /takové rozdělení se nazývá **normované**/ existují statistické **tabulky** hodnot. Zavedeme-li substituci  $u = \frac{x-\mu}{\sigma}$ , která udává o kolik směrodatných odchylek je hodnota  $x$  vzdálena od střední hodnoty, převedeme libovolné normální rozdělení na normované, jehož distribuční funkci označujeme  $F_N(u)$  nebo  $\Phi(u)$ .

**Příklad.** Pro normální rozdělení s parametry  $\mu = 84,4$  a  $\sigma = 36$  požadujeme najít hodnotu distribuční funkce v čísle 77,5. Jinak řečeno:  $F(77,5) = P(X \leq 77,5) = ?$

Postup si ukážeme jak pomocí tabulek, tak i pomocí počítačového (tabulkového) programu.

**1. Statistické tabulky.**  $F_N\left(\frac{77,5-84,4}{36}\right) \doteq F_N(-0,191667) \doteq 1 - F_N(0,192) \doteq 1 - 0,5735 \doteq 0,426$

**2. Programové vybavení.** V tabulkovém programu *Excel 2010* firmy Microsoft do řádku vzorců zadáme hodnoty:  $\Rightarrow$   
=NORM.DIST( $x$ ;  $\mu$ ;  $\sigma$ ; PRAVDA)

$f_x$	=NORM.DIST(77,5;84,4;36;1)		
	C	D	E
		0,424001659	



Ještě pohodlněji  
nalezneme  
hodnotu  
distribuční  
funkce  $F(x)$   
normálního  
rozdělení  
pomocí nabídky  
funkcí  $\Rightarrow$

NORM.DIST

X	77,5	=	77,5
Střed_hodn	84,4	=	84,4
Sm_odch	36	=	36
Kumulativní	1	=	PRAVDA

= 0,424001659

Vrátí hodnotu normálního rozdělení pro zadanou střední hodnotu a směrodatnou odchylku.

**Kumulativní** je logická hodnota: kumulativní distribuční funkce = PRAVDA, funkce hustoty pravděpodobnosti = NEPRAVDA.

Výsledek = 0,424001659

[Nápověda k této funkci](#)

OK Storno

**Příklad** skriptu [4, číslo 28].

Obsah ampulky s lékem je náhodnou veličinou s rozdělením  $N(10; 0,1^2)$  v  $\text{cm}^3$ .

Většina jevů v přírodě (bohužel ne tak docela automaticky ve společenských vědách) má toto normální rozložení. Na stromě je nejméně hodně malých lístků. S přibývajícím velikostí stromových listů jejich frekvence narůstá a dosáhne maxima u listů střední velikosti. Když velikost listů překročí průměrnou hodnotu, jejich četnost ubývá a opět, jako tomu bylo s nejmenšími lístky, nejméně bude těch největších stromových listů. Podobné rozložení (distribuci) – i když ne tak soustavně – objevíme i u řady sociálních jevů: výše příjmu, počet dětí v rodině, léta školního vzdělání, ...

Někdy bývá normální rozdělení také označováno jako **zákon chyb**.

## Rovnoměrné rozdělení

Rovnoměrné rozdělení na intervalu  $\langle a; b \rangle$  má náhodná veličina  $X$ , jejíž hodnota je úměrná délce podintervalu, do něhož má padnout a nezávisí na umístění podintervalu v intervalu  $\langle a; b \rangle$ .

Jeho **hustota pravděpodobnosti** je dána vztahem

$$f(x) = \frac{1}{b-a} \quad \text{kde} \quad a < x < b$$

a **charakteristiky** jsou

$$E(X) = \frac{a+b}{2} \quad D(X) = \frac{(b-a)^2}{12}$$

Jde o rozdělení, jehož hustota pravděpodobnosti je konstantní na nějakém intervalu  $\langle a; b \rangle$  a všude jinde je nulová. Křivka popisující hustotu pravděpodobnosti je na intervalu  $\langle a; b \rangle$  úsečka rovnoběžná s osou  $x$ .

Rovnoměrné rozdělení popisuje například chyby při zaokrouhlování čísel, doby čekání na uskutečnění jevu opakujícího se v pravidelných časových intervalech apod.

**Příklad** skripta [4, číslo 25].

Tramvaje přijíždějí do zastávky ve 12 minutových intervalech. Doba čekání na příjezd tramvaje je náhodná proměnná  $X$ .

## Exponenciální rozdělení

Exponenciální rozdělení  $E(A; \sigma)$  má náhodná veličina  $X$ , která představuje dobu, během níž nastane sledovaný jev.

Jeho **hustota pravděpodobnosti** je dána vztahem

$$f(x) = \begin{cases} \frac{1}{\sigma} \cdot e^{-\frac{A-x}{\sigma}} & \text{pro } x > A \\ 0 & \text{jinak} \end{cases} \quad (21)$$

a **charakteristiky** jsou

$$E(X) = A + \sigma \quad D(X) = \sigma^2$$

Exponenciální rozdělení má širokou použitelnost, zejména v teorii hromadné obsluhy (teorie front)<sup>16</sup>, v teorii spolehlivosti, v teorii obnovy atd. Náhodnou veličinou  $X$  bývá obvykle doba, během níž nastane sledovaný jev (například porucha přístroje, příchod zákazníka do opravy, atd.). Číslo  $A$  značí počáteční

<sup>16</sup> Příchod cestujícího v daný čas na zastávku MHD lze považovat za náhodnou veličinu, která má exponenciální rozdělení. Zpočátku to vypadá tak, že autobusy (trolejbusy, tramvaje) jezdí podle jízdního řádu a na jednotlivých zastávkách přicházejí náhodně rozdělení cestující – jednou krátce několik po sobě a pak určitou dobu zase nikdo.

Ted' k tomu přistoupí další série náhodných jevů. Např.: hustota provozu, povětrnostní podmínky (v mlze se asi jezdí pomaleji), ... Později se zpočátku nezávislé jevy stanou navzájem závislými a můžeme se dostat do následující spirály. Např.: autobus zůstane stát na „červenou“ a tím přibývá čekajících na nejbližší zastávce. Jejich odbavení (nástup a později i výstup) trvá déle, doby stání autobusu v zastávce jsou nadprůměrné, jízdní řád již nelze dodržet, na dalších zastávkách se nahromadí ještě více čekajících cestujících atd.

**A co s tím?** Změníme napříště jízdní řád, nasadíme autobusy disponující širšími (případně vícero) dveřmi, nasadíme velkokapacitní autobusy, nebo více autobusů bude jezdit v kratších intervalech, ...?

dobu, až do které sledovaný jev nastat nemůže. Parametr  $A$  se často interpretuje jako **parametr posunutí** rozdělení na ose  $x$ .

Parametr  $\sigma$  se někdy nazývá **parametr měřítka** a jeho převrácená hodnota  $\frac{1}{\sigma} = \lambda$  se někdy nazývá (průměrná) **rychlost výskytu** dané události.

V některých případech (například čekání na poruchu zařízení) má náhodná veličina  $X$  význam „doby života“ zkoumaného zařízení, přičemž je „bez paměti“, neboť platí:

Pravděpodobnost toho, že jev  $X$  nastane po nějaké době je stejná, jako by se do té doby nic nedělo.

Exponenciální rozdělení je z těchto důvodů vhodné k popisu rozdělení doby života těch zařízení, u nichž dochází k poruše ze zcela náhodných (vnějších) příčin, nikoliv například vlivem stárnutí materiálu.

Doby života mnohých strojních součástí a jiných zařízení — zvláště takových, u nichž se projevuje mechanické opotřebování a únava materiálu — mají Weibullovo rozdělení (s pamětí).

**Příklad** skripta [4, číslo 29].

Doba do poruchy zařízení se řídí exponenciálním rozdělením se střední hodnotou 8 hodin.

## Intenzita poruch

Modelujeme-li dobu do výskytu události (životnost, dobu do poruchy, dobu do návratu onemocnění, dobu do příchodu zákazníka apod.), používáme kromě hustoty pravděpodobnosti a distribuční funkce také funkci známou pod názvem intenzita poruch (hazardní funkce, angl. „hazard function“).

**Intenzitu poruch**  $\lambda(t)$  zavádíme pro nezápornou náhodnou veličinu  $X$  se spojitým rozdělením popsaným distribuční funkcí  $F(t)$ , kde  $F(t) \neq 1 : \forall t$  (tedy  $F(t) < 1$ ) takto:

$$\lambda(t) = \frac{F'(t)}{1 - F(t)} \quad (22)$$

Představuje-li náhodná veličina  $X$  dobu do poruchy nějakého zařízení, pak pravděpodobnost, že pokud do času  $t$  nedošlo k žádné poruše, tak k ní dojde v následujícím krátkém úseku délky  $\Delta t$ , je přibližně rovna  $\lambda(t) \cdot \Delta t$ .

Speciálně pro náhodnou proměnnou s **exponenciálním rozdělením**, jejíž hustota pravděpodobnosti je dána vztahem (21) platí, že

$$\lambda(t) = \frac{1}{\sigma} = \text{konstanta}, \quad \text{pro } t > A,$$

což jednoduše ověříme tak, že vztah (21 – popisuje funkci hustoty) dosadíme do vztahu (22 – platí pro distribuční funkci) za využití vzorců (15) a (13):

$$\lambda(t) = \frac{f(t)}{1 - \int_0^t f(x) dx} = \frac{\frac{1}{\sigma} \cdot e^{\frac{A-t}{\sigma}}}{1 - \int_A^t \frac{1}{\sigma} \cdot e^{\frac{A-x}{\sigma}} dx} = \frac{\frac{1}{\sigma} \cdot e^{\frac{A-t}{\sigma}}}{1 + \left[ e^{\frac{A-x}{\sigma}} \right]_A^t} = \frac{\frac{1}{\sigma} \cdot e^{\frac{A-t}{\sigma}}}{1 + e^{\frac{A-t}{\sigma}} - e^{\frac{A-A}{\sigma}}} = \frac{\frac{1}{\sigma} \cdot e^{\frac{A-t}{\sigma}}}{1 + e^{\frac{A-t}{\sigma}} - 1} = \frac{1}{\sigma}$$

Má-li doba do výskytu události exponenciální rozdělení, pak je intenzita poruch konstantní. Což mimo jiné znamená, že není závislá na délce předcházejícího provozu sledovaného systému. Tedy jsme skutečně oprávněni, tak jako na předchozí straně, tvrdit, že čekání na poruchu zařízení je rozdělení „bez paměti“.

Pokud bychom skutečně sledovali četnost poruch nějakého druhu výrobku, nejspíše by zakreslená křivka intenzity poruch měla tři části:

- I. V prvním úseku křivka intenzity poruch klesá. Odpovídající časový interval se nazývá **období časných poruch** (období záběhu, počátečního provozu, osvojování, dětských nemocí). Příčinou zvýšené intenzity poruch v tomto období jsou poruchy v důsledku výrobních vad, nesprávné montáže, chyb při návrhu, při výrobě apod.
- II. Ve druhém úseku dochází k běžnému využívání zaběhnutého výrobku, k poruchám dochází většinou z vnějších příčin, nedochází k opotřebením, které by změnilo funkční vlastnosti výrobku. Příslušný časový interval se nazývá **období normálního užití**, či období stabilního života.  
*Intenzita poruch je v tomto období přibližně konstantní.*
- III. Ve třetím úseku procesy stárnutí a opotřebením mění funkční vlastnosti výrobku, projevují se nastřádané otřesy výrobku z období II, trhliny materiálu a intenzita poruch vzrůstá. Příslušný časový interval se nazývá **období poruch v důsledku stárnutí a opotřebením**.

Intenzitu poruch modelujeme v jednotlivých úsecích většinou pomocí různých rozdělení. Pouze ve druhém úseku používáme v této kapitole probírané exponenciální rozdělení. A pouze v tomto druhém úseku jde o „rozdělení bez paměti“. A to ještě ne u všech druhů výrobků.

Již zmiňované Weibullovo rozdělení je obecnější než exponenciální rozdělení a proto je mnohem flexibilnější. Umožňuje tak modelovat dobu do výskytu události i u systémů, které jsou v I. období časných poruch nebo ve III. období stárnutí (tedy tam, kde se projevuje mechanické opotřebením nebo únava materiálu).

Exponenciální rozdělení je speciálním typem Weibullova rozdělení.

## 6. Náhodné vektory

Až doposud jsme se zabývali náhodnou veličinou, která výsledku pokusu přiřazovala jedno reálné číslo. Jestliže je výsledek pokusu vyjádřen několika reálnými čísly, závislými na náhodě, chápeme tato čísla jako hodnoty jistého systému náhodných veličin a používáme pro ně pojem **náhodný vektor**.

Uveďme příklady náhodných vektorů:

- výška  $X_1$ , hmotnost  $X_2$ , věk  $X_3$  a inteligenční kvocient  $X_4$  studentů daného ročníku představují složky náhodného vektoru  $\mathbf{X} = (X_1; X_2; X_3; X_4)$ ;
- doba zaměstnání  $X$  a výška platu  $Y$  zaměstnanců daného podniku jsou složky náhodného vektoru  $\mathbf{Z} = (X; Y)$ ;
- známka  $X_1$ , kterou student získal z matematiky v prvním semestru a známka  $X_2$ , kterou student získal z matematiky ve druhém semestru jsou složky náhodného vektoru  $\mathbf{Y} = (X_1; X_2)$ .
- údaje zaznamenávané meteorologickou sondou (výška; tlak; teplota; rosný bod).

Jednotlivé náhodné veličiny v rámci náhodného vektoru mohou být naprosto nezávislé (například věk  $X_3$  a inteligenční kvocient  $X_4$  v prvním příkladu), mohou však také mít silnou vazbu (například výška a tlak v posledním příkladu).

Pro jednoduchost se v následujícím omezíme na dvousložkový náhodný vektor.

**Sdružená distribuční funkce** (simultánní distribuční funkce) náhodných veličin  $X$  a  $Y$  je vyjádřena vztahem

$$F(x, y) = P(X \leq x; Y \leq y)$$

Sdružená distribuční funkce<sup>17</sup> má obdobné vlastnosti jako distribuční funkce jedné proměnné

1.  $0 \leq F(x, y) \leq 1$
2. Distribuční funkce je neklesající funkcí v každé proměnné.
3. Distribuční funkce je spojitá zprava v každé proměnné.
4.  $\lim_{x, y \rightarrow -\infty} F(x, y) = 0$  a  $\lim_{x, y \rightarrow +\infty} F(x, y) = 1$

Chceme-li určit distribuční funkci složky  $X$  (případně složky  $Y$ ) náhodného vektoru, mluvíme o

**Marginální distribuční funkci** která má tvar

$$F_X(x) = P(X \leq x; Y \text{ libovolné}) = \lim_{y \rightarrow \infty} F(x, y)$$

$$F_Y(y) = P(X \text{ libovolné}; Y \leq y) = \lim_{x \rightarrow \infty} F(x, y)$$

Z tohoto vyjádření dále plyne, že v případě diskrétního náhodného vektoru s pravděpodobnostní funkcí  $P(x_i; y_j)$  můžeme získat následující vztahy pro **marginální pravděpodobnosti**

$$P_X(x) = \sum_{\forall y_j} P(X = x; Y = y_j)$$

$$P_Y(y) = \sum_{\forall x_i} P(X = x_i; Y = y)$$

<sup>17</sup> Poznamenejme, že ve výrazu  $P(X \leq x; Y \leq y)$  se podle tradice používá středník (čárka) ve významu průniku jevů. Správnější je tedy zápis:  $P(\{X \leq x\} \cap \{Y \leq y\})$  nebo  $P((X \leq x) \wedge (Y \leq y))$



Obdobně pro spojitý náhodný vektor s hustotou  $f(x, y)$  získáme vztahy pro marginální hustoty pravděpodobnosti

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

## Kontingenční (korelační) tabulka

V případě diskrétního dvousložkového náhodného vektoru s konečným počtem hodnot se sdružená pravděpodobnostní funkce často prezentuje prostřednictvím kontingenční tabulky<sup>18</sup> (viz následující [příklad](#)). V této tabulce se mimo sdružené pravděpodobnostní funkce (uprostřed tabulky) rovněž uvádí v posledním řádku a v posledním sloupci marginální pravděpodobnostní funkce.

Ve statistice takovou tabulku někdy nazýváme korelační. [3, str. 121]

<sup>18</sup> Slovo **kontingence** se do statistiky dostalo přes angličtinu z latiny [14, str. 310] – znamená téměř doslova *setkání, spojení*. V takové tabulce se tedy zaznamenávají výsledky, které vycházejí ze spojení dvou řad znaků.

## Číselné charakteristiky náhodného vektoru

Pokud bychom si jednotlivě všímali pouze složek náhodného vektoru, pak **pro každou složku** již umíme podle vzorce (16) určit **střední hodnotu** a podle vzorce (17) **rozptyl**. Nyní si k nim přidáme ještě další charakteristiky používané pro stanovení míry vazby<sup>19</sup> mezi náhodnými veličinami. A stejně jako u kontingenční tabulky se omezíme pouze na náhodné vektory, jejichž obě náhodné proměnné jsou diskrétního typu.

Marginální střední hodnoty a rozptyly popisují pouze charakteristiky rozdělení jednotlivých náhodných veličin, neříkají však nic o „těsnosti“ vztahu mezi oběma veličinami.

K charakteristikám, které měří těsnost (= míru) **lineární vazby** mezi náhodnými veličinami  $X$  a  $Y$  patří následující dvě charakteristiky: kovariance a koeficient korelace<sup>20</sup>.

Zdůrazněme, že ani jedna z charakteristik měřících těsnost vazby nic neříká o vztahu **příčina  $\Rightarrow$  účinek**. Jenom vypovídají, že mezi těmito proměnnými existuje tak a tak silná vazba. Potom si musí odborník v příslušné oblasti lámat hlavu, který důsledek je způsoben kterou příčinou.

## Kovariance $cov(X, Y)$

je střední hodnota součinu odchylek náhodných veličin  $X$  a  $Y$  od jejich středních hodnot:

$$cov(X, Y) = E \{ [X - E(X)] \cdot [Y - E(Y)] \} \stackrel{\text{vlastnosti } E(x)}{=} E(X \cdot Y) - E(X) \cdot E(Y)$$

<sup>19</sup> Představme si, že měříme výšku  $X$  a váhu  $Y$  dospělého člověka. Ze zkušenosti víme, že zhruba řečeno: čím je někdo vyšší, tím je těžší. Ale jistě známe i výjimky z tohoto pravidla. Jednak malé-tlusté a také vysoké-hubené lidi. Závislost mezi výškou a váhou tedy není přesná funkční závislost, jak ji známe z matematiky, ale je to závislost jiného druhu, tzv. **statistická**.

A pokud výšku a váhu více dospělých osob zaznamenáme do souřadné soustavy (osa  $x$  výška, osa  $y$  váha), kde každému člověku odpovídá jeden bod v rovině (zde prázdné kolečko), můžeme obdržet obrázek podobný [tomuto](#).

<sup>20</sup> Koeficient korelace (*korelační koeficient*) je pro měření těsnosti vztahu mezi  $X$  a  $Y$  vhodnější charakteristikou než kovariance, protože je jednak bezrozměrný a jednak je normován. Platí:  $|r(X, Y)| \leq 1$ .

## Korelační (Pearsonův) koeficient $\rho(X, Y)$

určuje míru (jak je silná závislost) **lineárních závislostí** náhodných veličin  $X$  a  $Y$ .

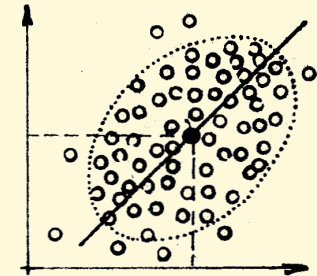
$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{D(X) \cdot D(Y)}} \quad \text{kde} \quad D(X) \cdot D(Y) \neq 0.$$

- $-1 \leq \rho(X, Y) \leq 1$
- Když  $\rho(X, Y) = 0$ , pak veličiny  $X$  a  $Y$  jsou nekorelované.  
Ovšem mohou být závislé (kvadraticky, exponenciálně či jinak), jen **hodnoty neleží na přímce**.
- Když  $\rho(X, Y) > 0$ , pak hovoříme o kladné (přímé, pozitivní) korelaci; **roste**-li  $X$ , tak  $Y$  nejspíše také **roste**.

Jinak: Pro velké hodnoty  $X$  lze očekávat spíše velké hodnoty  $Y$  a pro malé hodnoty  $X$  lze očekávat spíše malé hodnoty  $Y$ .

Když  $\rho(X, Y) < 0$ , pak hovoříme o záporné (nepřímé, negativní) korelaci; **roste**-li  $X$ , tak  $Y$  naopak spíše **klesá**.

Pro velké hodnoty  $X$  lze očekávat spíše malé hodnoty  $Y$  a pro malé hodnoty  $X$  lze očekávat spíše velké hodnoty  $Y$ .



- Hodnoty  $\rho(X, Y)$  blízké  $\pm 1$  znamenají **silnou lineární závislost**. Hodnoty veličin  $X$  a  $Y$  téměř leží na přímce.

Hodnoty  $\rho(X, Y)$  blízké 0 znamenají **slabou lineární závislost** mezi veličinami  $X$  a  $Y$ .

V mnoha případech však nelze na první pohled určit, zda hodnotu korelačního koeficientu už můžeme považovat za blízkou „1“ (nebo „-1“ či „0“), a potom je nutné významnost (blízkost „k“ něčemu) korelačního koeficientu testovat (viz kapitola o [testování hypotéz](#)).

**Příklad:**

Házíme jednou mincí **tříkrát po sobě**. Sestavte kontingenční tabulku a určete (Pearsonův) **korelační koeficient** (míru lineární závislosti) pro tyto náhodné veličiny:

$X$  ... počet pokusů, než padne první RUB;

$Y$  ... počet po sobě padlých RUBů.

Náhodný vektor  $V = (X, Y)$ .

**Řešení:** Házíme třikrát (zajímají nás trojice) mincí (v jednom hodu dva možné výsledky — Rub×Líc), přičemž klidně mohou padnout dva LÍCe po sobě (prvky se mohou opakovat). Shrnutí: jde o skupiny trojic ze dvou prvků, které se mohou opakovat a přitom záleží na pořadí, protože rozlišujeme, o jaký hod šlo. Tedy podle tabulky **kombinatorických skupin** jde o variace třetí třídy ( $r = 3$ ) ze dvou prvků ( $k = 2$ ) s opakováním, proto  $V_3'(2) = 2^3 = 8$ . Protože možností není tak mnoho, vypíšeme si schematicky všechny možné výsledky tří hodů

**3×Rub** – RRR;

**2×Rub** – RRL, RLR, LRR;

**1×Rub** – RLL, LRL, LLR;

**žádný Rub** – LLL

$X$  ... počet pokusů, než padne první RUB;

$Y$  ... počet po sobě padlých RUBů.

a určíme, které elementární jevy vyhovují daným hodnotám náhodných veličin  $X$  a  $Y$ .

$X = 0$	(již v prvním hoďu padl RUB) .....	<b>RRR, RRL, RLR, RLL</b>
$X = 1$	(až ve druhém hoďu padl RUB) .....	<b>LRR, LRL</b>
$X = 2$	(až ve třetím hoďu padl RUB) .....	<b>LLR</b>
$X = 3$	(vůbec nepadl RUB) .....	<b>LLL</b>
$Y = 0$	(vůbec nepadl RUB) .....	LLL
$Y = 1$	(po každém RUBu nepadl další RUB) .....	RLR, RLL, LRL, LLR
$Y = 2$	(po každých dvou RUBech nepadl další RUB)	RRL, LRR
$Y = 3$	(pokaždé padl RUB) .....	RRR

Uvědomme si, že jde o nezávislé pokusy (padnutí RUBu v prvním hoďu nijak neovlivní to, co padne v hoďu následujícím), kde  $p = 0,5$  (pravděpodobnost padnutí RUBu), můžeme tedy podle vzorce (6) přímo spočítat pravděpodobnosti jednotlivých elementárních jevů, například:

$P(\{RLR\}) = 0,5 \cdot (1 - 0,5) \cdot 0,5 = 0,125$ , podobně pro všechny ostatní.

Je zřejmé, že **všechny trojice mají stejnou pravděpodobnost**.

Dále například:

$$P(X = 0; Y = 1) = P(\{RRR, RRL, RLR, RLL\} \cap \{RLR, RLL, LRL, LLR\}) = P(\{RLR, RLL\})$$

A protože elementární jevy jsou navzájem neslučitelné (když padne RUB, nemůže ve stejném hoďu zároveň padnout LÍČ)

$$P(\{RLR, RLL\}) = P(\{RLR\}) + P(\{RLL\}) = 0,125 + 0,125 = 2 \cdot 0,125 = 0,25$$

Nyní již zkonstruujeme levou kontingenční tabulku, do které vypíšeme nejdříve elementární jevy, které vyhovují příslušným podmínkám. Pak do pravé tabulky doplníme patřičné pravděpodobnosti.

		Y			
		0	1	2	3
X	0	—	RLR, RLL	RRL	RRR
	1	—	LRL	LRR	—
	2	—	LLR	—	—
	3	LLL	—	—	—

X \ Y	0	1	2	3	$P_X(x)$
0	0	0,25	0,125	0,125	0,5
1	0	0,125	0,125	0	0,25
2	0	0,125	0	0	0,125
3	0,125	0	0	0	0,125
$P_Y(y)$	0,125	0,5	0,25	0,125	1

Máme určit **koeficient korelace**, na který potřebujeme znát **kovarianci** a marginální výběrové rozptyly. Pro výpočet rozptylu zase podle (17) je nutné znát střední hodnoty. Například pro  $E(X)$  využijeme podle (16) šedě označené hodnoty v prvním a posledním sloupci pravé tabulky, pro  $E(Y)$  zase žlutě označené hodnoty a pro  $E(X \cdot Y)$  neobarvené hodnoty v tabulce.

$$E(X) = 0 \cdot 0,5 + 1 \cdot 0,25 + 2 \cdot 0,125 + 3 \cdot 0,125 = 0 + 0,25 + 0,25 + 0,375 = 0,875$$

$$E(Y) = 0 \cdot 0,125 + 1 \cdot 0,5 + 2 \cdot 0,25 + 3 \cdot 0,125 = 0 + 0,5 + 0,5 + 0,375 = 1,375$$

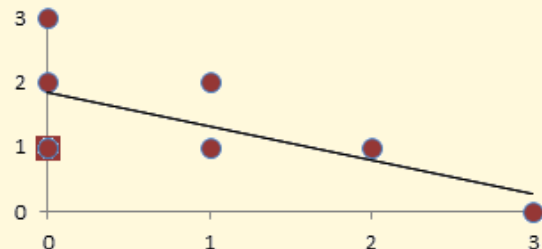
$$E(X \cdot Y) = 0 + 0 + 0 + 0 + 0 + 0 + 1 \cdot 1 \cdot 0,125 + 1 \cdot 2 \cdot 0,125 + 0 + 0 + 2 \cdot 1 \cdot 0,125 + 0 + 0 + 0 + 0 + 0 = 0,625$$

Pro **kovarianci** platí:  $cov(X, Y) = 0,625 - 0,875 \cdot 1,375 = -0,578125$

**Pro ostatní charakteristiky můžeme postupovat obdobně.**

**Práci strojům!** Podstatně méně pracné je využití skutečnosti, že některé počítačové programy umí počítat požadované charakteristiky. Pokud například výše uvedené hodnoty (souřadnice bodů) přepíšeme do *Excelu 2010*, můžeme si ušetřit další práci (s dosazováním do vzorců a jejich vyčíslováním) a nechat funkci **CORREL**, ať ukáže, co umí. Levou kontingenční tabulku přepíšeme do *Excelu 2010* (podle následujícího levého obrázku) ve tvaru, kolikrát se příslušný bod  $[X; Y]$  vyskytuje. Vidíme, že  $[0; 1]$  je dvakrát a body  $[0; 2]$ ,  $[0; 3]$ ,  $[1; 1]$ ,  $[1; 2]$ ,  $[2; 1]$  a  $[3; 0]$  jedenkrát.

	A	B	C	D	E	F	G	H	I	J
2		X	0	0	0	0	1	1	2	3
3		Y	1	1	2	3	1	2	1	0
5		(Pearsonův) Korelační koeficient								
6		$\rho(X;Y) = -0,641$ =PEARSON(C2:J2;C3:J3)								



Na základě hodnoty korelačního koeficientu  $\rho \doteq -0,6$  můžeme říci, že mezi náhodnými veličinami  $X$  a  $Y$  existuje středně silná negativní korelace. Je tedy pravděpodobné, že s růstem  $X$  bude  $Y$  klesat (lineárně).

Na očekávanou otázku:

*Umí Excel 2010 počítat i další charakteristiky?*

existuje také očekávaná odpověď: **UMÍ**

(viz vedlejší obrázek pro českou verzi Excelu 2010).

	A	B	C	D	E	F	G	H	I	J	K
2		X	0	0	0	0	1	1	2	3	
3		Y	1	1	2	3	1	2	1	0	
8		<b>CHARAKTERISTIKY</b>									
9		<b>Střední hodnoty</b>									
10		$E(X) = 0,875$ =PRŮMĚR(C2:J2)									
11		$E(Y) = 1,375$ =PRŮMĚR(C3:J3)									
12		<b>Rozptyly</b>									
13		$D(X) = 1,109$ =VAR.P(C2:J2)									
14		$D(Y) = 0,734$ =VAR.P(C3:J3)									
15		<b>Kovariance</b>									
16		$cov(X,Y) = -0,578$ =COVARIANCE.P(C2:J2;C3:J3)									

## Příklad

Zjistěte **střední hodnotu** a **směrodatnou odchylku** náhodné veličiny (skripta [4, příklad 19]), která popisuje počet padlých LÍCŮ při současném hodu čtyřmi **rozlišitelnými** mincemi (skripta [4, příklad 15]) – nebo házeme jednou mincí čtyřikrát po sobě.

## Řešení

Pravděpodobnost, že padne líc při hodu jednou mincí je 0,5. Totéž platí pro rub mince. Házeme-li čtyřmi mincemi, musí to platit pro každou z nich. Proto například padnutí lícu na všech čtyřech mincích má pravděpodobnost  $0,5 \cdot 0,5 \cdot 0,5 \cdot 0,5 = 0,0625$ . Všechny možnosti si můžeme schématicky znázornit, když označíme **L** jev, že padne líc a **R** jev, že padne rub.

				<b>RRLL</b>
			<b>RLRL</b>	
	<b>RRRL</b>	<b>RLLR</b>	<b>RLLL</b>	
	<b>RRLR</b>	<b>LRRL</b>	<b>LRLR</b>	
	<b>RLRR</b>	<b>LRLR</b>	<b>LLRL</b>	
<b>RRRR</b>	<b>LRRR</b>	<b>LLRR</b>	<b>LLLR</b>	<b>LLLL</b>

což můžeme zaznamenat v následující tabulce, kde střední hodnotu značíme  $E(X)$  a rozptyl  $D(X)$ .

**Poznámka:** Pokud si uvědomíme, že jde o **binomické rozdělení**, kde  $n = 4$  (házeme čtyřmi mincemi),  $p = 0,5$  (pravděpodobnost padnutí LÍCE), můžeme podle vzorce (19) přímo spočítat požadované charakteristiky.  $E(X) = 4 \cdot 0,5 = 2$ ,  $D(X) = 4 \cdot 0,5 \cdot (1 - 0,5) = 1$ .

My budeme postupovat tak, jako bychom to nevěděli. Alespoň víme, co nám má vyjít.



k	$x_k$	$P(X = x_k)$				
1	0	0,062 5				
2	1	0,25				
3	2	0,375				
4	3	0,25				
5	4	0,062 5				
$\Sigma$		<b>1</b>				

k	$x_k$	$P(X = x_k)$	$x_k \cdot P(X = x_k)$			
1	0	0,062 5	0			
2	1	0,25	0,25			
3	2	0,375	0,75			
4	3	0,25	0,75			
5	4	0,062 5	0,25			
$\Sigma$	<b>1</b>	<b>2</b>				

$E(X)$

k	$x_k$	$P(X = x_k)$	$x_k \cdot P(X = x_k)$	$x_k - E(X)$		
1	0	0,062 5	0	-2		
2	1	0,25	0,25	-1		
3	2	0,375	0,75	0		
4	3	0,25	0,75	1		
5	4	0,062 5	0,25	2		
$\Sigma$		<b>1</b>	<b>2</b>	<b>0</b>		

$E(X)$

k	$x_k$	$P(X = x_k)$	$x_k \cdot P(X = x_k)$	$x_k - E(X)$	$[x_k - E(X)]^2$	
1	0	0,062 5	0	-2	4	
2	1	0,25	0,25	-1	1	
3	2	0,375	0,75	0	0	
4	3	0,25	0,75	1	1	
5	4	0,062 5	0,25	2	4	
$\Sigma$		<b>1</b>	<b>2</b>	<b>0</b>		

$E(X)$

k	$x_k$	$P(X = x_k)$	$x_k \cdot P(X = x_k)$	$x_k - E(X)$	$[x_k - E(X)]^2$	$[x_k - E(X)]^2 \cdot P(X = x_k)$
1	0	0,062 5	0	-2	4	0,25
2	1	0,25	0,25	-1	1	0,25
3	2	0,375	0,75	0	0	0
4	3	0,25	0,75	1	1	0,25
5	4	0,062 5	0,25	2	4	0,25
$\Sigma$		<b>1</b>	<b>2</b>	<b>0</b>		<b>1</b>

$E(X)$

$D(X)$

k	$x_k$	$P(X = x_k)$	$x_k \cdot P(X = x_k)$	$x_k - E(X)$	$[x_k - E(X)]^2$	$[x_k - E(X)]^2 \cdot P(X = x_k)$
1	0	0,062 5	0	-2	4	0,25
2	1	0,25	0,25	-1	1	0,25
3	2	0,375	0,75	0	0	0
4	3	0,25	0,75	1	1	0,25
5	4	0,062 5	0,25	2	4	0,25
$\Sigma$		<b>1</b>	<b>2</b>	<b>0</b>		<b>1</b>

 $E(X)$ 
 $D(X)$ 

## (popisná) Statistika

Nyní vyjděme z předpokladu, že nám není známo, že výše uvedený příklad popisuje počet padlých „*líců*“ při současném hodu čtyřmi rozlišitelnými mincemi. Proto ani netušíme, že by mohlo jít o binomické rozdělení. **Máme pouze tato sesbíraná data:**

2   2   1   2   3   1   2   3   1   2   3   0   1   2   3   4

## Zpracování (statistického) materiálu

která poskládáme a zapíšeme do tabulky. Zajímá nás, jaké **charakteristiky** (příslušné vzorce uvedeme v **následující kapitole**) můžeme z takto sesbíraných hodnot (a zapsaných do tabulky) získat.

Pokud zobecníme poznatky z předchozího příkladu (a také to, co jsme se dozvěděli v této kapitole o pravděpodobnosti) můžeme říci, že okolo nás existuje spousta věcí, jevů a událostí, které nelze předvídat, protože jsou důsledkem náhody. Pod pojmem náhoda rozumíme působení faktorů, které se živelně mění. Otázkami náhody a náhodných dějů se zabývají dvě disciplíny: teorie pravděpodobnosti a matematická statistika.

**Teorie pravděpodobnosti** je matematická disciplína, zabývající se studiem zákonitostí v náhodných pokusech a jejich modelováním matematickými prostředky  $\Rightarrow$  *matematika náhody* (číselně popisuje míru naděje, že „náhodou“ nějaký jev/výsledek nastane). Její logická struktura je budována axiomaticky. To znamená, že její základ tvoří několik tvrzení (tak zvaných axiomů), která vyjadřují základní vlastnosti axiomatizované veličiny a všechna další tvrzení jsou z nich odvozena deduktivně. Systém axiomů vzniká abstrakcí z pozorovaných skutečností reálného světa. Axiomy se nedokazují, považují se za prověřené dlouhou lidskou zkušeností.

Představme si to tak, že máme perfektně **popsán model** (v minulé případě to bylo současné házení čtyřmi rozlišitelnými mincemi). Ptáme se: ***Jak dopadne následující pokus – hod?*** Kolik padne LÍČŮ? ...

**Statistika** (matematická) je naproti tomu věda, která zahrnuje studium dat vykazujících náhodná kolísání, ať už jde o data získaná pečlivě připraveným pokusem provedeným pod stálou kontrolou experimentálních podmínek v laboratoři, či o data provozní. Statistika jako věda se dále zabývá otázkami získávání dat, jejich analýzou a formulováním závěrů o pokusech a experimentech, nebo závěrů při rozhodování založeném na datech.

Takže nyní máme několik (dostatek) výsledků realizace nějakého děje (tolikrát padl například LÍČ) a ptáme se: ***Jaké vlastnosti má model, který co nejlépe popisuje daný děj?*** Můžeme z dat usoudit, že házeme rozlišitelnými mincemi (závisí na pořadí  $\Rightarrow$  variace) nebo stejnými mincemi (nezávisí na pořadí  $\Rightarrow$  kombinace)? A co ještě můžeme usoudit?

Obecně se matematická statistika snaží formulovat závěry a tvrzení o pozorovaných veličinách, které plynou z výsledků pokusů, měření nebo pozorování, které vykazují jisté náhodné chování.

Zatímco teorie pravděpodobnosti usuzuje z vytvořeného pravděpodobnostního modelu zkoumaného děje na výsledky jeho jednotlivých realizací, statistika odhaduje vlastnosti zkoumaného děje, u kterého neznáme model na základě dat, zjištěných z jeho jednotlivých realizací.

Dávno před prvními elementárními úvahami o počtu pravděpodobnosti (hazardní hry) a ještě před prvním (statistickým) zkoumáním údajů o obyvatelstvu, byly známy dva jevy, které vlastně představují syntézu teorie pravděpodobnosti a statistiky.

**Sázky** a později **loterie**, kde hlavně u velkých loterií podnikatel (většinou stát) zprostředkuje bez vlastního rizika vyrovnání mezi množstvím sázejících.

Početné malé dílčí příspěvky (sázky, cena losu, ...) jsou po srážce nákladů a daní odevzdány do rukou těch **několika málo**, kteří měli štěstí.

**Pojištění** pracuje na stejném principu. Četné malé dílčí částky (pojistné) jsou po srážce nákladů a zisku odevzdány těm **několika málo**, kdo mají dostat náhradu za utrpěnou škodu.

Jistý rozdíl tady ale je. Zatímco u loterií se mezi výherce rozdělí pouze tolik, kolik se vybralo (navíc ponížené o náklady a daně), u pojištění se při vzniku pojistné události vyplácí předem pevně stanovené odškodné. Proto si musejí pojišťovací společnosti velmi dobře rozvážit, jak velký kapitál musejí mít k dispozici. Jen na základě „**mlhavých**“ představ o četnosti škod, (data, která jsou k dispozici – viz předchozí příklad), můžeme očekávat dva stejně nepříjemné omyly:

- Bud' podceníme četnost škod, požadujeme nízké pojistné, ale přitom musíme v případě škodní události hodně vyplácet  $\Rightarrow$  úpadek firmy.
- Nebo z opatrnosti nasadíme pojistné příliš vysoko a z počátku vyděláváme více než dost. Brzy však ztratíme zákazníky, kteří přejdou ke správněji kalkulující a tím lacinější konkurenci.



A vystavuji se úpadku v ještě větší míře, protože předpokladem fungujícího pojištění je pokud možno velký počet pojištěných.

Proto musejí pojišťovny více „**kalkulovat**“ než firmy provozující loterie. Snahou pojišťoven je, co nejvíce konkretizovat svoje představy o škodných událostech tak, aby tyto představy co nejvěrněji odpovídaly realitě. A tomu následně přizpůsobit svůj podnikatelský záměr.

Předpokladem pro vznik pojištění bylo poznání, že jisté škodné události se vyskytují s přibližně odhadnutelnou četností. Pak přišel další logický krok. Když víme, že v průměru například **každá desátá loď** ztroskotá, je možno škodu vyrovnat tak, že každý vlastník lodi zaplatí desetinu hodnoty (lodí a zboží při každé plavbě) jako pojistné.

Již ve čtvrtém století před naším letopočtem [14, str. 255], kdy ostrov Rhodos ovládl lodní plavbu ve východním Středomoří a vytvořil počátky obchodního a námořního práva, vznikla první úprava rozdělení ztráty při vyhazování zboží přes palubu v případě nebezpečí na moři. Úprava, která byla později jako **lex Rhodia de iactu** (rhódský zákon o odlehčování lodi potopením zboží) převzat do římského práva.

Uvedený zákon se zakládal na této situaci. Obchodní loď je naložena zbožím, které patří více obchodníkům. Dostane se do bouře a musí se zbavit (alespoň části) nákladu, aby se nepotopila. Lodní posádka popadne, co jí právě přijde pod ruku a co se dá zvlášť snadno (nebo co je zvlášť těžké) hodit přes palubu a pokračuje (i když s případnými obtížemi) v plavbě do přístavu. Záchrana lodi, mužstva a často i většiny zboží byla možná jen za podmínky, že bylo obětováno (část nebo všechno) zboží jednoho (nebo více) obchodníků. A měli by být právě oni poškozeni, aby ostatní nepřišli k újmě? „Lex Rhodia de iactu“ rozhodl tak, že se škoda rovnoměrně rozdělí na všechny, kdo měli zájem na záchraně lodi a nákladu.

Od tohoto zákonem upraveného dělení škody po havárii je pouze malý krok k dobrovolnému **předchozímu** placení pojistného za dopravované zboží. Náklady přitom velmi podstatně klesnou, protože se pojistné platí i za ty lodní přepravy, které skončí beze ztrát. Musíme ale rozlišovat dvě věci, které se velmi lehce směšují: matematicky objektivně očekávanou hodnotu (každá desátá loď ztroskotá) a subjektivní osobní riziko (co z toho pro mne plyne, pokud to bude moje loď?).

# Úvod do **Popisné statistiky**

## Obsah kapitoly: Popisná statistika

<b>1. Co je to statistika?</b>	<b>96</b>
1.1. Základní pojmy . . . . .	97
<b>2. Číselné charakteristiky statistických souborů</b>	<b>101</b>
2.1. Charakteristiky <b>polohy</b> . . . . .	102
Modus, medián . . . . .	102
Aritmetický, geometrický, harmonický a chronologický průměr . . . . .	103
2.2. Charakteristiky <b>rozptylu</b> (variability) . . . . .	109
Rozptyl (výběrový), směrodatná odchylka . . . . .	109
Příklad . . . . .	114
Včetně odlehklých (extrémních) hodnot . . . . .	114
Očištěná data . . . . .	124
<b>3. Zpracování statistického materiálu</b>	<b>134</b>
3.1. Menší vzorek . . . . .	134
3.2. Rozsáhlý vzorek . . . . .	143
3.2.1. Třídění dat – tabulka . . . . .	145
3.2.2. Další sloupce tabulky . . . . .	151
3.2.3. Určení číselných charakteristik . . . . .	156
<b>4. Využití programu Excel 2010</b>	<b>157</b>
<b>5. Základy zpracování kvalitativních dat</b>	<b>162</b>
<b>6. Závěr kapitoly – Etapy statistické práce</b>	<b>168</b>

# 1. Co je to statistika?

**Popisná statistika**<sup>21</sup> bývá prvním krokem k odhalení informací skrytých ve velkém množství proměnných a jejich variant.

Statistika (jako vědní disciplína) si klade za cíl informace a zákonitosti, které případně existují mezi některými hodnotami (a na počátku mohou být skryty) odhalit. To znamená uspořádat proměnné (jejich pozorované hodnoty) do názornější formy (**graf**×**tabulka**) a popsat je několika málo hodnotami (proto proměnné podle potřeby sdružujeme do tříd – viz poznámka pod **obrázkem 3**), které by obsahovaly co největší množství informací obsažených v původním souboru.

Nyní si na příkladu ukážeme některé úlohy statistiky a přístup k jejich řešení. Výrobce součástek změnil technologii výroby. Chce zjistit, jaká je životnost součástek vyráběných touto novou technologií a zda se tato životnost významně liší od životnosti součástek vyráběných dřívějším způsobem.

Je zřejmé, že nemá smysl zjišťovat životnost každé vyrobené součástky. Trvalo by to jednak dlouho a po provedení zkoušek by nebylo co prodávat. Výrobce proto volí následující postup:

- Ze série vyráběných součástek vybere určitý počet součástek a na takto vybraných součástkách provede zkoušky životnosti.
- Ze získaných hodnot životnosti pak určí parametry, které nejlépe charakterizují životnost vybraného souboru součástek.
- Tyto charakteristiky pak slouží jako podklad pro závěry týkající se životnosti celé vyrobené série.

Stěžejním úkolem je najít postup, aby výsledky které získá na vzorku, byly co nejvíce podobné těm, které by získal po prozkoumání všech vyrobených součástek. První věc, která nás s otázkou přesnosti napadne,

<sup>21</sup> Vyvinula se z původních starověkých sčítání obyvatel a majetku.

je mít vzorek co největší. Ale tento postup má svá úskalí, z nichž na některá jsme již poukázali (například pokud je při testování součástka zničena, nelze ji prodat).

Vyvstávají pak například následující otázky:

- Jaká je životnost součástek vyráběných změněnou technologií?
- Je výrazný rozdíl mezi životnostmi součástek vyráběných oběma způsoby?
- Jaký je pravděpodobnostní zákon pro rozdělení doby životnosti součástek?

Vhodným matematickým nástrojem pro řešení těchto a dalších otázek je (matematická) statistika, jejímž hlavním úkolem je rozbor dat (získaných z vyšetřování skupiny prvků) a rozšíření závěrů získaných z tohoto vyšetřování na celý soubor (populaci). **Statistika – to je sběr a zpracování dat.**

## 1.1. Základní pojmy

**Znak (náhodná veličina).** Prvky (*statistické jednotky*), na nichž provádíme statistická šetření, mají některé vlastnosti (znaky) společné a liší se v jednom nebo více znacích, o jejichž vlastnosti se zajímáme.

**V našem příkladě** ke společným znakům výše zmíněných součástek počítáme to, že jsou vyrobeny ze stejného materiálu, v určité továrně, danou technologií, atd. Znak v němž se liší je například jejich životnost.

Statistickou jednotkou je v tomto případě vyrobená součástka. Pojmem *zpravodajská jednotka* (firma, obec, domácnost, ...) označuje státní statistika subjekty, které v souladu s příslušnou legislativou mají vůči státu takzvanou zpravodajskou povinnost (musejí něco hlásit).

**Základní soubor (populace)** <sup>22</sup> obsahuje všechny objekty, které chceme poznat. Jinak řečeno, je to soubor jednotek, o kterém předpokládáme, že jsou pro něj naše závěry platné.

**V našem příkladu** tvoří základní soubor všechny součástky, které byly nebo ještě budou vyrobeny.

**Výběrový soubor (vzorek)** obsahuje pouze objekty skutečně vyšetřené, neboli skupinu jednotek, které skutečně pozorujeme.

**V našem příkladu** je výběrový soubor tvořen součástkami, na nichž proběhly zkoušky.

Abychom byli schopni z chování vzorku předpovídat chování populace, musí struktura vzorku imitovat (napodobovat) složení populace tak přesně, jak je to jen možné <sup>23</sup>.

Lze předpokládat, že s rostoucí velikostí vzorku se rozdíl mezi strukturou populace a vzorku zmenšuje. Skutečně; nejdříve rychle, pak pomaleji a pomaleji. Úplné shody mezi strukturou populace a vzorku dosáhneme teprve tehdy, když jsme zahrnuli všechny elementy populace do vzorku.

**Datový soubor** je tvořen šetřením získanými údaji, kterým říkáme **hromadná** data nebo jenom **data**.

**V našem příkladu** zjištěné hodnoty životnosti na vybraných součástkách tvoří datový soubor.

<sup>22</sup> Název **populace** se tradičně používá proto, že prapůvodně se statistikou rozuměla činnost, spočívající ve zjišťování stavu nějakého území a spíše stavu obyvatelstva na tomto území — aby měla „vrchnost“ představu, kolik prostředků například získá na daních, kolik mužů si může dovolit povolání zbraně apod. Za příklad takového statistického zjišťování může sloužit sčítání lidu, které v roce Kristova narození nechal provést císař Augustus (viz Bible, Druhá kniha Samuelova, kapitola 24 a Lukášovo evangelium, kapitola 2). A protože to, co se tehdy zkoumalo bylo obyvatelstvo daného území, zaužíval se název populace, který nyní stále používáme pro základní soubor, i když v hledáčku pozornosti **námi popisovaného příkladu** jsou vyráběné součástky.

<sup>23</sup> Jen si zkuste představit, jaké hodnoty o čase stráveném na internetu získáte v domovech pro seniory nebo na vysokoškolských kolejkách.

**Poznámka.** Aby se při řešení úloh statistiky mohlo využít metod teorie pravděpodobnosti, vychází se z následujících úvah:

**Princip realizace** pravděpodobnostního modelu statistického zkoumání, to je získání statistický dat a vhodných charakteristik.

Protože hodnoty znaku nabývají působením náhodných vlivů na jednotlivých objektech různých hodnot, považujeme znak za náhodnou veličinu, kterou označíme  $X$ . Proto předpokládáme, že získaná data jsou realizacemi této náhodné veličiny  $X$  (vyšetřovaného znaku), která má distribuční funkci  $F(x)$ , kterou ovšem neznáme. Abychom získali informace o rozdělení této náhodné veličiny v celém základním souboru (populaci), provedeme několik (tím vlastně sestojíme vzorek – uskutečňujeme výběr) vzájemně nezávislých pokusů (měření, pozorování, ...) při nichž sledujeme realizace této náhodné veličiny (jaké jsou výsledky jednotlivých pokusů). Z hodnot získaných ze vzorku (datový soubor) vypočteme **empirické charakteristiky** (my známe střední hodnotu  $E(X)$  a rozptyl  $D(X)$ ) a **empirické zákony rozdělení** (například distribuční funkci  $F_n(x)$ ). Pomocí nich pak odhadujeme hledané charakteristiky a zákony rozdělení náhodné veličiny  $X$ .

Například průměrný plat 20 občanů ČR je náhodná veličina, kterou označme  $X$ . Výpočtem průměrného platu (stanovením střední hodnoty  $E(X)$  z 20 platů) **konkrétních** 20 občanů (Ferda, Marie, ...) získáme jednu **realizaci** tohoto průměru. Výpočtem průměrného platu jiného vzorku 20 občanů ČR (Lojzicka, Josef, ...) získáme jinou realizaci průměru.

**Princip pravděpodobnostního modelu** použitého pro vyvození závěrů vyplývajících ze získaných statistických údajů a charakteristik.

Má-li ale datový soubor podávat dobrou informaci o vlastnostech základního souboru, musí být výběr objektů prováděn náhodně, přičemž má mít každý objekt v základním souboru stejnou možnost být vybrán. Protože objekty ve výběrovém souboru byly vybrány náhodně, lze očekávat, že při

jiných výběrech dostaneme jiný datový soubor. A ten bude mít jiné empirické charakteristiky a jiné empirické zákony rozdělení, i když charakteristiky a zákon rozložení celé populace (základního souboru) jsou stále stejné. Získané hodnoty vzorku  $(x_1, x_2, \dots, x_n)$  lze tedy považovat za realizace náhodného vektoru  $(X_1, X_2, \dots, X_n)$ , jehož složky  $X_i$  jsou vzájemně nezávislé náhodné veličiny.

Empirické charakteristiky (střední hodnota, rozptyl), obecně označené  **$b$** , které jsou funkcemi hodnot vzorku, pak považujeme za realizace jistých náhodných veličin  **$B$** .

Protože  $b = f(x_1, x_2, \dots, x_n)$ , bude  $B = f(X_1, X_2, \dots, X_n)$ . Takto sestrojené náhodné veličiny  **$B$**  nazýváme obecně **statistikami** (nebo **výběrovými charakteristikami**) a jejich hodnoty, které nabývají na statistickém souboru nazýváme **pozorované hodnoty statistiky** nebo **empirickými charakteristikami**.

S některými statistikami (výběrovými charakteristikami) se nyní seznámíme.



## 2. Číselné charakteristiky statistických souborů

Představte si situaci, že máte k dispozici statistický soubor o poměrně velkém rozsahu a stojíte před otázkou co s ním, jak jej co nejvýstižněji popsat. Číselné hodnoty, kterými takovýto rozsáhlý soubor „nahradíme“, postihují základní vlastnosti tohoto souboru a my jim budeme říkat statistické charakteristiky (statistiky).

Jsou to jednočíselné charakteristiky, které charakterizují všechny hodnoty zkoumané veličiny v celém souboru jediným číslem.

Jde zejména o průměrnou hodnotu veličiny v celém souboru — například průměrnou výšku studenta ve třídě. Kromě průměrné hodnoty veličiny se používají i další obdobné **míry polohy** (míry úrovně) veličiny v daném souboru, například prostřední hodnota z naměřených hodnot uspořádaných podle velikosti apod.

Vedle určení nějaké míry polohy je dalším základním úkolem při zpracování naměřených hodnot získání alespoň hrubé informace o tom, jak jsou hodnoty zkoumané veličiny rozděleny mezi jednotlivé objekty souboru, jak mnoho se tyto hodnoty na jednotlivých objektech od sebe navzájem liší, jak mnoho jsou rozptýleny kolem hodnoty průměrné. Aby bylo možné tuto rozptýlenost či variabilitu veličiny charakterizovat jednou hodnotou, jedním číslem, byly vyvinuty různé **míry variability** zkoumané veličiny v daném souboru. Všechny nějakým způsobem zhruba udávají průměrnou odchylku hodnot náhodné veličiny naměřených na jednotlivých objektech od průměrné hodnoty této veličiny v celém souboru. Například se zjišťuje, o kolik se průměrně liší výška studenta ze třídy od průměrné výšky všech studentů z dané třídy.

## Charakteristiky polohy — data: 1 2 2 2 4

Tyto charakteristiky vyjadřují pomyslný střed proměnné.

**Modus:**  $\bar{x}_{mo}$  u diskrétní proměnné je **nejčastější** hodnota (nejčastěji se vyskytující; ta, která má nejvyšší četnost) = **2**. Dvojka se v daných datech vyskytuje třikrát.

Pouze tato charakteristika je použitelná u **jmenných – nominálních** (názvových, alfabetických) proměnných, které nabývají rovnocenných variant. Proto je nelze ani porovnávat, ani seřadit. O dvou hodnotách lze pouze konstatovat, že jsou buď stejné, nebo že jsou různé.

Například: pohlaví, národnost, značka hodinek, barva svetru, ...

V tomto případě modus představuje typického reprezentanta (hodnotu proměnné), který chování souboru ovlivňuje nejvíce, protože se vyskytuje nejvíce krát.

U spojitě proměnné nelze modus takto určovat, ale v této příručce se tím nebudeme trápit.

Existence dvou a více modů ve výběru obvykle signalizuje nesourodost (heterogenitu) hodnot proměnné. Tuto nesourodost bývá možné odstranit rozdělením souboru na podsoubory — roztříděním podle některého jiného znaku (například dvoumodální znak **výška člověka** lze roztřídit podle pohlaví na dva unimodální (jsou určeny jednoznačně) znaky – výška žen a výška mužů).

**Medián:**  $\bar{x}_{me}$  je **prostřední** hodnota z naměřených hodnot **uspořádaných podle velikosti**. Přesněji:

- **prostřední** hodnota při lichém počtu prvků;
- **jakákoliv** hodnota mezi prostředními hodnotami (i včetně nich) při sudém počtu prvků.

Nejčastěji (pokud má smysl ho určovat) bereme **aritmetický průměr** z těchto prostředních hodnot. O něm si více řekneme za chvíli.

Tedy pro naše zadaná data opět **2**.

Medián lze použít u **pořadových – ordinálních** proměnných, u kterých lze stanovit pořadí a tím je vzájemně porovnávat (pouze na základě pořadí) nebo seřadit.

Například: známka ve škole, velikost oděvů (S, M, L, XL), medaile ve sportovních soutěžích (zlatá, stříbrná, bronzová), ...

Někdy ovšem můžeme mít problém s aritmetickým průměrem prostředních hodnot.

Následující čtyři charakteristiky s názvem *nějaký průměr* používáme pouze u (kvantitativních, měřitelných) proměnných, které lze vyjádřit čísly a pak je pomocí těchto čísel porovnávat. Tedy má smysl se ptát **O KOLIK** je jeden prvek lepší než druhý, případně KOLIKRÁT je jeden prvek lepší než druhý, ...

**Data:** 1   2   2   2   4   průměr všem hodnotám proměnné přiřazuje vhodné číslo  $\Rightarrow$  je **funkcí**

**Aritmetický průměr:** 
$$\bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{n} \cdot \sum_{i=1}^k n_i \cdot x_i = \frac{1 + 2 + 2 + 2 + 4}{5} = \frac{1 \cdot 1 + 3 \cdot 2 + 1 \cdot 4}{5} = \frac{11}{5} = 2,2$$

**Geometrický průměr:** 
$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{\prod_{i=1}^k x_i^{n_i}} = \prod_{i=1}^k x_i^{\frac{n_i}{n}} = \sqrt[5]{1 \cdot 2 \cdot 2 \cdot 2 \cdot 4} = \sqrt[5]{1^1 \cdot 2^3 \cdot 4^1} = \sqrt[5]{32} = 2$$

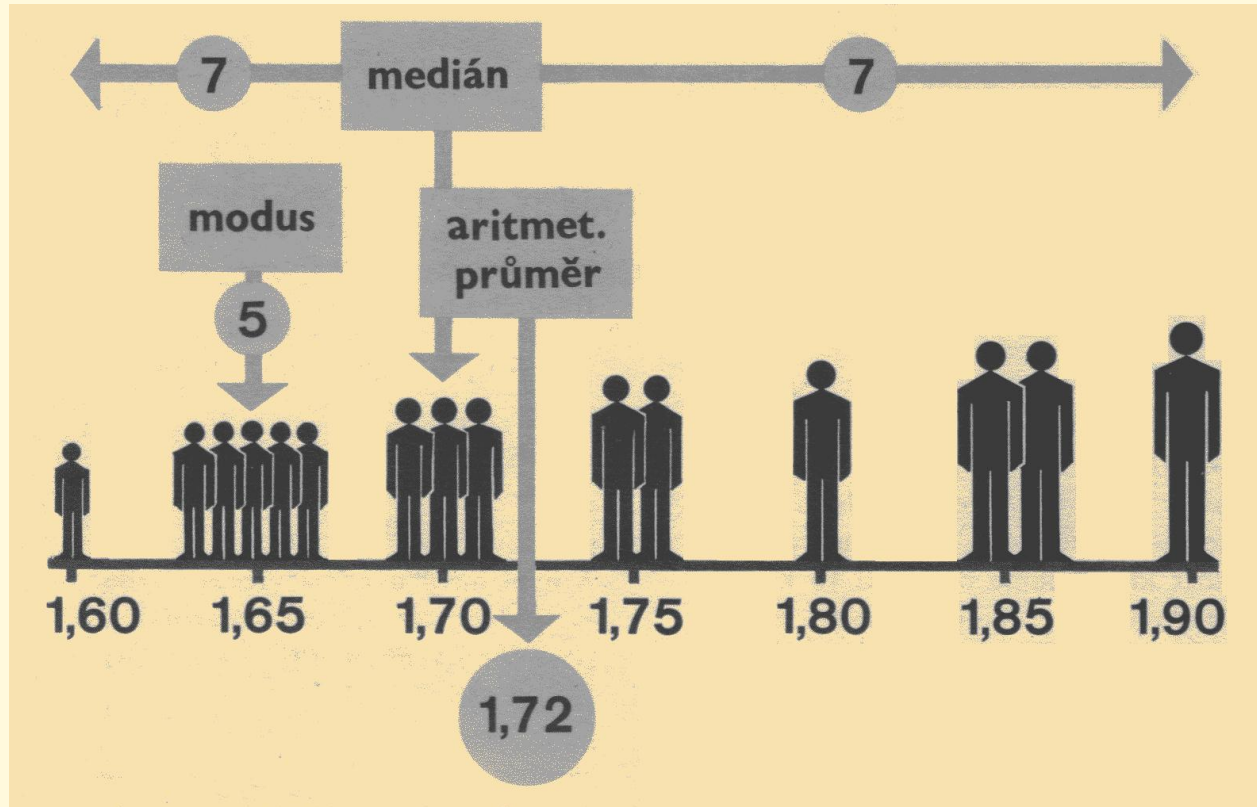
pro  $x_i > 0$

**Harmonický průměr:** 
$$\bar{x}_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^k \frac{n_i}{x_i}} = \frac{5}{\frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{4}} = \frac{5}{\frac{1}{1} + \frac{3}{2} + \frac{1}{4}} = \frac{20}{11} \doteq 1,818$$

pro  $x_i > 0$

Ze vzorce  $\frac{1}{\bar{x}_H} = \frac{\sum_{i=1}^n \frac{1}{x_i}}{n}$  je zřejmé, že převrácená hodnota harmonického průměru je aritmetickým průměrem převrácených hodnot proměnných.

Obrázek 3: Převzat z [14]



Zkoumané osoby byly zařazeny do **tříd** (skupin) podle jejich velikosti (v metrech)!  
 Například pro druhou skupinu zleva: vyšších jak 162,5 cm a nižších jak 167,5 cm jich bylo pět.

**Chronologický průměr:** 
$$\bar{x}_{ch} = \frac{1}{2 \cdot (n - 1)} \cdot (x_1 + 2x_2 + \dots + 2x_{n-1} + x_n)$$

$$= \frac{\frac{x_1 + x_2}{2} + \frac{x_2 + x_3}{2} + \dots + \frac{x_{n-1} + x_n}{2}}{n - 1} \quad \text{kde: } x_1 \leq x_2 \leq \dots \leq x_{n-1} \leq x_n$$

Jde vlastně o aritmetický průměr aritmetických průměrů sousedních hodnot / data: 1 2 2 2 4

$$\bar{x}_{ch} = \frac{\frac{1+2}{2} + \frac{2+2}{2} + \frac{2+2}{2} + \frac{2+4}{2}}{4} = \frac{(1+2) + (2+2) + (2+2) + (2+4)}{2 \cdot 4} = \frac{17}{8} = 2,125$$

### Několik poznámek

1. Uvědomte si, že požadavek, aby mělo smysl se ptát „o kolik, kolikrát, ...“, je oprávněný (významný, důležitý, právníci používají termín relevantní). Formálně sice můžeme například modré barvě přiřadit jedničku a červené barvě dvojku. Ovšem již nemůžeme pro jeden svetr barvy hodnoty 1 a pro druhý barvy hodnoty 2 tvrdit, že v průměru máme dva svetry v barvě 1,5.

Toto tvrzení však postrádá smysl.

2. Na rozdíl od obecné představy, aritmetický průměr není vždy pro výpočet **průměru** výběrového souboru nejvhodnější. Pracujeme-li, například, s proměnnou představující relativní změny (růstové indexy, cenové indexy, ...), používáme geometrický průměr. Pro výpočet průměru v případech, kdy proměnná má charakter části z celku (úlohy o společné práci, některé úlohy o pohybu, ...), používáme průměr harmonický.
3. **Formálně** bychom sice mohli i pro záporné hodnoty použít v **určitých případech** vzorec pro geometrický průměr (musí být definována odmocnina) a stejně tak vzorec pro harmonický průměr.

Například pro hodnoty  $-4$ ,  $-2$ ,  $1$  dostáváme:

$\bar{x}_G = \sqrt[3]{(-4) \cdot (-2) \cdot (1)} = \sqrt[3]{8} = 2$  ,      což je naprosto mimo zadané hodnoty,  
proto tento výsledek nemůže představovat „**průměr**“ zadaných hodnot.

$$\bar{x}_H = \frac{1}{\frac{1}{-4} + \frac{1}{-2} + \frac{1}{1}} = \frac{1}{\frac{-1-2+4}{4}} = \frac{4}{1} = 4 \quad \text{což je opět naprosto mimo zadané hodnoty, ...}$$

Obdobně pro hodnoty       $-1, \quad 2, \quad 4$

$$\bar{x}_G = \sqrt[3]{(-1) \cdot (2) \cdot (4)} = \sqrt[3]{-8} = -2$$

$$\bar{x}_H = \frac{1}{\frac{1}{-1} + \frac{1}{2} + \frac{1}{4}} = \frac{1}{\frac{-4+2+1}{4}} = -\frac{4}{1} = -4$$

Proto se přidržíme obecně uznávané zásady, že jak geometrický průměr, tak harmonický průměr budeme počítat pouze pro **kladné** hodnoty sledované proměnné, což je jak v případě **indexů** (budou probírány v kapitole o [hospodářské statistice](#)) tak v případě **společné práce** automaticky splněno.

Kvantitativní (měřitelná) proměnná, která nabývá v určitém statistickém souboru pouze kladných číselných hodnot, se někdy nazývá **kardinální proměnná**.

4. Vzhledem k tomu, že každý z průměrů se stanovuje ze všech hodnot proměnné, nese maximum informací o výběrovém souboru. Na druhé straně je však chronologický, ale hlavně aritmetický průměr velmi citlivý na tak zvaná **odlehlá pozorování**, což jsou hodnoty, které se mimořádně liší od ostatních a dokáží proto vychýlit aritmetický průměr natolik, že přestává daný výběr dobře reprezentovat. Viz [následující](#) příklad.
5. Vzpomenete-li si například na [normální rozdělení](#), můžeme jej nyní přesněji charakterizovat a říci o něm, že normální rozdělení je jednomodální rozdělení, symetrické kolem střední hodnoty  $\mu$ , přičemž tato střední hodnota je rovna modu a mediánu.

## Kvantily

Kvantily (srovnej s již dříve uvedenou **kvantilovou funkcí**) jsou statistiky, které charakterizují polohu jednotlivých hodnot v rámci proměnné. Podobně jako modus, jsou i kvantily rezistentní (odolné) vůči odlehlým pozorováním. Obecně je kvantil definován jako hodnota, která rozděluje výběrový soubor **uspořádaný podle velikosti** na dvě části:

1. část obsahuje hodnoty, které jsou menší než daný kvantil anebo stejné;
2. část obsahuje hodnoty, které jsou větší nebo rovny danému kvantilu.

Pro určení kvantilu je proto nutné výběr uspořádat od nejmenší hodnoty k největší.

Kvantil proměnné  $x$ , který odděluje  $100p$  % menších hodnot od zbytku souboru, tedy od  $100(1 - p)$  % hodnot, nazýváme **100p % kvantilem** a značíme jej  $x_p$ .

Zejména v souvislosti s hodnocením normovaných testů (SCIO testy, biometrické normy, ...) se často setkáváme s vyjádřením: „**Patříte do xyz. percentilu**“ [8, str. 43], přičemž **xyz** je celé číslo od jedné do sta. Například „**Patříte do 80. percentilu**“ znamená, že nejméně 79 % a nejvýše **80 %** účastníků testu dosáhlo **nižšího** výsledku než vy.

$x_{0,5}$  **kvantil** již známe. Jmenuje se **medián**, kdy polovina (50 %) všech hodnot je menších nebo stejných jako  $x_{0,5}$  a polovina je větších anebo se rovná tomuto mediánu.

Aritmetický průměr (stejně jako jiné podobné reprezentace středních hodnot) nebo údaje v procentech<sup>24</sup> redukují informaci o mnoha prvcích vzorku do jednoho jediného údaje. A to je pěkně silná redukce, při které můžeme ztratit důležitý druh informace. Jakákoliv charakteristika polohy proto potře-

<sup>24</sup> [2, str. 186] „Po aplikaci preparátu B se 33,3 % kuřat uzdravilo, 33,3 % uhynulo a o zbývajících 33,3 % nejsme schopni poskytnout uspokojující informaci, protože se nám dosud nepodařilo to **třetí** kuře chytit.“

buje ke správnému vyhodnocení konkrétní situace ještě jeden rozměr (údaj). Alespoň hrubou informaci o tom, jak jsou hodnoty zkoumané veličiny rozděleny mezi jednotlivé objekty souboru, jak mnoho se tyto hodnoty na jednotlivých objektech od sebe navzájem liší, jak mnoho jsou **rozptýleny kolem hodnoty průměrné**. Aby bylo možné tuto rozptýlenost či variabilitu veličiny charakterizovat jednou hodnotou, jedním číslem, byly vyvinuty různé míry variability zkoumané veličiny v daném souboru. Všechny nějakým způsobem zhruba udávají průměrnou odchylku hodnot náhodné veličiny naměřených na jednotlivých objektech od průměrné hodnoty této veličiny v celém souboru. Například se zjišťuje, o kolik se průměrně liší hmotnost kapra vyloveného v rybníku od průměrné váhy všech kaprů z tohoto rybníku.

Variabilitu výběrových charakteristik přitom ovlivňují tři faktory [8, str. 106]:

1. rozsah populace  $N$ ;
2. rozsah výběru  $n$ ;
3. způsob získání náhodného výběru.

Míry variability charakterizují měřenou veličinu v celém daném souboru objektů jedním číslem z hlediska velikosti kolísání hodnot této veličiny. Je možno z nich ihned usoudit, jak mnoho jsou tyto hodnoty v souboru rozptýlené, jsou-li v průměru hodně či málo vzdálené od průměrné hodnoty veličiny v souboru.



## Charakteristiky rozptylu (variability) — data: 1 2 2 2 4

Víme, že pro tyto hodnoty platí:  $\bar{x} = 2,2$ , což je **aritmetický průměr**. Ten nám však nic neříká o rozložení jednotlivých hodnot proměnné kolem tohoto středu, tj. o **variabilitě proměnné**. Je zřejmé, že čím větší je rozptýlenost hodnot proměnné kolem jejího pomyslného středu, tím menší je schopnost tohoto středu reprezentovat celou proměnnou (viz [pivní hrdina](#)).

**Rozptyl (výběrový):**  $S^2 = \frac{1}{n-1} \left[ \sum x_i^2 - n \cdot \bar{x}^2 \right]$       nebo:  $S^2 = \frac{1}{n-1} \cdot \sum [x_i - \bar{x}]^2$

$$\frac{1}{5-1} \cdot [(1^2 + 2^2 + 2^2 + 2^2 + 4^2) - 5 \cdot 2,2^2] = \frac{1+4+4+4+16-5 \cdot 4,84}{4} = \frac{29-24,2}{4} = \frac{4,8}{4} = 1,2$$

$$\begin{aligned} & \frac{1}{5-1} \cdot [(1-2,2)^2 + (2-2,2)^2 + (2-2,2)^2 + (2-2,2)^2 + (4-2,2)^2] = \\ & = \frac{(-1,2)^2 + (-0,2)^2 + (-0,2)^2 + (-0,2)^2 + (1,8)^2}{4} = \frac{1,44+3 \cdot 0,04+3,24}{4} = \frac{4,8}{4} = 1,2 \end{aligned}$$

Nevýhodou použití (výběrového) rozptylu jakožto míry variability je to, že rozměr této charakteristiky je druhou mocninou rozměru proměnné.

Například je-li proměnnou denní tržba uvedena v Kč, bude výběrový rozptyl této proměnné vyjádřen v Kč<sup>2</sup>. Tento nedostatek odstraňuje další míra variability, a tou je:

**Směrodatná odchylka (výběrová):**  $S = \sqrt{S^2} = \sqrt{1,2} \doteq 1,095$

Nevýhodou (výběrového) rozptylu i (výběrové) směrodatné odchylky je ta skutečnost, že neumožňují porovnávat variabilitu proměnných vyjádřených v různých jednotkách.

Která proměnná má větší variabilitu — výška nebo hmotnost dospělého jedince? Na tuto otázku nám dá odpověď následující charakteristika, a tou je:

**Variační koeficient:**  $V = \frac{S}{|\bar{x}|} = \frac{1,095}{2,2} \doteq 0,498 \doteq 50 \%$

Variační koeficient je bezrozměrný, uvádíme jej často v procentech. Zhruba udává, jakou část aritmetického průměru představuje směrodatná odchylka.

**Variační rozpětí:**  $R = x_{\max} - x_{\min} = 4 - 1 = 3$

Toto variační rozpětí však z důvodu jeho přílišné citlivosti k případným ojedinělým extrémním hodnotám není moc dobrým odhadem variability. Proto někdy používáme i kvartilové (mezikvartilové) rozpětí, které je rozdílem horního a dolního kvartilu, tedy rozdílem 75% a 25% **kvantilu**:

$$R_q = x_{0,75} - x_{0,25}.$$

Následující charakteristiku (průměrnou absolutní odchylku), uvádíme pouze pro úplnost, abychom si ukázali, že se v praxi využívají dvě metody, jak zajistit kladný výsledek. U (výběrového) rozptylu rozdíl umocníme **na druhou**, u absolutní (výběrové) odchylky použijeme absolutní hodnotu a protože to je „průměrná“ odchylka, určíme její aritmetický průměr. A protože je to „výběrová“ odchylka, dělíme o jedničku zmenšeným počtem prvků.

**Průměrná absolutní odchylka (výběrová):**  $d = \frac{1}{n-1} \cdot \sum |x_i - \bar{x}|$

$$\frac{1}{4} \cdot (|1 - 2,2| + |2 - 2,2| + |2 - 2,2| + |2 - 2,2| + |4 - 2,2|) = \frac{1}{4} \cdot (1,2 + 3 \cdot 0,2 + 1,8) = \frac{1}{4} \cdot 3,6 = 0,9$$

**Další charakteristiky** (například šikmost, špičatost a další) si nebudeme uvádět. Vzorce podle nichž se určují tyto charakteristiky jsou poměrně složité, a proto se podle nich „ručně“ většinou nepočítá. Bývají součástí programů pro zpracování statistických dat. Například *Excel 2010*:

**Šikmost** (koeficient *skosu*; angl. „skewness“) =  $\text{SKEW}(\text{data})$  označuje stupeň asymetričnosti rozdělení veličiny kolem její střední hodnoty.

**Špičatost** (koeficient *excesu*; angl. „kurtois“) =  $\text{KURT}(\text{data})$  určuje relativní strmost nebo plochost rozdělení v porovnání s normovaným normálním rozdělením.

Význam některých empirických (spočítaných z hodnot vzorku, výběru) charakteristik pro celý základní soubor (populaci) je následující:

- aritmetický průměr  $\bar{x}_A$  vzorku je (nejlepším) číselným odhadem střední hodnoty  $E(X)$  základního souboru (populace),
- výběrový rozptyl  $S^2$  vzorku je (nejlepším) číselným odhadem rozptylu  $D(X)$  základního souboru,

jak si později ukážeme.

Otázky spojené s přesností těchto odhadů (co je vlastně nejlepším odhadem), pokud má základní soubor normální rozdělení, budou řešeny v kapitole o [intervalových odhadech](#).

**[14, str.92]:** „Statistika bez použití rozumu dává nesmysly — a to neplatí jen o statistice.“

Mám-li hodnoty proměnné hmotnosti zaokrouhlované na kilogramy, asi nemá smysl jakýkoliv průměr této proměnné počítat na gramy.

Směrodatnou odchylku jakožto míru nejistoty měření zaokrouhlujeme nahoru na maximálně dvě (většinou) až tři platné cifry.

**Rozšíření poznatků o rozptylu.** ANOVA<sup>25</sup> (analysis of variance), kterou se podrobněji nebudeme zabývat, je založena na představě, že variabilita (proměnlivost, rozptýlení, disperze), se kterou kolísají hodnoty sledované náhodné veličiny kolem střední hodnoty jejího rozdělení, vzniká jako důsledek různých vlivů, z nichž každý přispívá k této celkové variabilitě určitým podílem. Celkový rozptyl (kvadrát směrodatné odchylky) jako míru variability lze pak rozčlenit na dílčí rozptyly náležející těmto jednotlivým vlivům — faktorům.

Například nás zajímá variabilita měsíčních platů pobíraných ve státě. Platy jsou rozptýleny kolem střední hodnoty rozdělení a rozptýlení je vyvoláváno (nebo naše představa je, že by mohlo být vyvoláváno) mnoha vlivy — faktory. Jeden z nich (který nás enormně zajímá) je ekonomická sféra, v níž jsou platy vypláceny. V rámci tohoto faktoru můžeme např. rozlišovat zaměstnance ze zemědělství, státní zaměstnance, zaměstnance z oblasti peněžnictví, z oblasti služeb, z potravinářského průmyslu atd. Existují další faktory, které ovlivňují hodnotu platu a jejich změny přispívají k proměnlivosti platů. Faktor vzdělání zaměstnance (základní, středoškolské a vysokoškolské) nebo různá doba zaměstnání, umístění podniku podle krajů, podle velikosti obcí, faktor pohlaví zaměstnance a další.

Analýza rozptylu v průmyslových aplikacích umožňuje posoudit vliv různých faktorů na výrobní proces, hodnotit vliv použití různých druhů surovin na jakost produkce apod. V ekonomických aplikacích pak umožňuje posoudit vliv různých faktorů na hospodářský proces, hodnotit účinky různých přijatých opatření apod.

<sup>25</sup> Analýza rozptylu — byla vyvinuta [R. A. Fisherem](#) pro oblast zemědělství na počátku 20. století.

## Charakteristiky polohy (průměrů $\bar{x}$ ) a rozptylu setříděného vzorku

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200

**Variační rozpětí**

## Charakteristiky polohy (průměrů $\bar{x}$ ) a rozptylu seříděného vzorku

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200

**Modus**

**Variační rozpětí:**  $R = 1\,200 - 10 = 1\,190$

## Charakteristiky polohy (průměru $\bar{x}$ ) a rozptylu setříděného vzorku

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200

**Modus:**  $\bar{x}_{mo} = 18$

**Variační rozpětí:**  $R = 1\,200 - 10 = 1\,190$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200

3

**Medián**

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200

**Variační rozpětí:**  $R = 1\,200 - 10 = 1\,190$

**Medián:**  $\bar{x}_{me} = 29$

## Aritmetický průměr



## Charakteristiky polohy (průměrů $\bar{x}$ ) a rozptylu setříděného vzorku

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200

**Modus:**  $\bar{x}_{mo} = 18$

**Variační rozpětí:**  $R = 1200 - 10 = 1190$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200  
3

**Medián:**  $\bar{x}_{me} = 29$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200  
12 12

**Aritmetický průměr:**  $\bar{x}_A = 82,32$

$$\frac{1}{25} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150 + 1200)$$

**Geometrický průměr**

## Charakteristiky polohy (průměrů $\bar{x}$ ) a rozptylu setříděného vzorku

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200

**Modus:**  $\bar{x}_{mo} = 18$

**Variační rozpětí:**  $R = 1\,200 - 10 = 1\,190$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200  
3

**Medián:**  $\bar{x}_{me} = 29$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200  
12 12

**Aritmetický průměr:**  $\bar{x}_A = 82,32$

$$\frac{1}{25} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150 + 1\,200)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 32,73$

$$\sqrt[25]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150 \cdot 1\,200}$$

**Harmonický průměr**

**Charakteristiky polohy (průměrů  $\bar{x}$ ) a rozptylu setříděného vzorku**

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200

**Modus:**  $\bar{x}_{mo} = 18$ **Variační rozpětí:**  $R = 1\,200 - 10 = 1\,190$ 10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200  
3**Medián:**  $\bar{x}_{me} = 29$ 10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200  
12 12**Aritmetický průměr:**  $\bar{x}_A = 82,32$ 

$$\frac{1}{25} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150 + 1\,200)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 32,73$ 

$$\sqrt[25]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150 \cdot 1\,200}$$

**Harmonický průměr:**  $\bar{x}_H \doteq 24,26$ 

$$\frac{25}{\frac{1}{10} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{2}{16} + \frac{3}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{26} + \frac{1}{29} + \frac{1}{32} + \frac{1}{35} + \frac{1}{37} + \frac{1}{38} + \frac{1}{42} + \frac{1}{45} + \frac{1}{49} + \frac{1}{51} + \frac{1}{60} + \frac{1}{86} + \frac{1}{150} + \frac{1}{1\,200}}$$

**Chronologický průměr**

## Charakteristiky polohy (průměrů $\bar{x}$ ) a rozptylu setříděného vzorku

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200

**Modus:**  $\bar{x}_{mo} = 18$

**Variační rozpětí:**  $R = 1200 - 10 = 1190$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200  
3

**Medián:**  $\bar{x}_{me} = 29$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200  
12 12

**Aritmetický průměr:**  $\bar{x}_A = 82,32$

$$\frac{1}{25} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150 + 1200)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 32,73$

$$\sqrt[25]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150 \cdot 1200}$$

**Harmonický průměr:**  $\bar{x}_H \doteq 24,26$

$$\frac{25}{\frac{1}{10} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{2}{16} + \frac{3}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{26} + \frac{1}{29} + \frac{1}{32} + \frac{1}{35} + \frac{1}{37} + \frac{1}{38} + \frac{1}{42} + \frac{1}{45} + \frac{1}{49} + \frac{1}{51} + \frac{1}{60} + \frac{1}{86} + \frac{1}{150} + \frac{1}{1200}}$$

**Chronologický průměr:**  $\bar{x}_{ch} \doteq 60,54$

$$\frac{1}{2 \cdot (25-1)} \cdot (10 + 2 \cdot 11 + 2 \cdot 13 + 2 \cdot 14 + 4 \cdot 16 + 6 \cdot 18 + 2 \cdot 20 + 2 \cdot 24 + 2 \cdot 26 + 2 \cdot 29 + 2 \cdot 32 + 2 \cdot 35 + 2 \cdot 37 + 2 \cdot 38 + 2 \cdot 42 + 2 \cdot 45 + 2 \cdot 49 + 2 \cdot 51 + 2 \cdot 60 + 2 \cdot 86 + 2 \cdot 150 + 1200)$$

**Výběrový rozptyl**

## Charakteristiky polohy (průměrů $\bar{x}$ ) a rozptylu setříděného vzorku

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200

**Modus:**  $\bar{x}_{mo} = 18$

**Variační rozpětí:**  $R = 1200 - 10 = 1190$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200  
3

**Medián:**  $\bar{x}_{me} = 29$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200  
12 12

**Aritmetický průměr:**  $\bar{x}_A = 82,32$

$$\frac{1}{25} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150 + 1200)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 32,73$

$$\sqrt[25]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150 \cdot 1200}$$

**Harmonický průměr:**  $\bar{x}_H \doteq 24,26$

$$\frac{25}{\frac{1}{10} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{2}{16} + \frac{3}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{26} + \frac{1}{29} + \frac{1}{32} + \frac{1}{35} + \frac{1}{37} + \frac{1}{38} + \frac{1}{42} + \frac{1}{45} + \frac{1}{49} + \frac{1}{51} + \frac{1}{60} + \frac{1}{86} + \frac{1}{150} + \frac{1}{1200}}$$

**Chronologický průměr:**  $\bar{x}_{ch} \doteq 60,54$

$$\frac{1}{2 \cdot (25-1)} \cdot (10 + 2 \cdot 11 + 2 \cdot 13 + 2 \cdot 14 + 4 \cdot 16 + 6 \cdot 18 + 2 \cdot 20 + 2 \cdot 24 + 2 \cdot 26 + 2 \cdot 29 + 2 \cdot 32 + 2 \cdot 35 + 2 \cdot 37 + 2 \cdot 38 + 2 \cdot 42 + 2 \cdot 45 + 2 \cdot 49 + 2 \cdot 51 + 2 \cdot 60 + 2 \cdot 86 + 2 \cdot 150 + 1200)$$

**Výběrový rozptyl:**  $S^2 \doteq 55104$

**Směrodatná odchylka**

$$\frac{1}{25-1} \cdot [(10^2 + 11^2 + 13^2 + 14^2 + 2 \cdot 16^2 + 3 \cdot 18^2 + 20^2 + 24^2 + 26^2 + 29^2 + 32^2 + 35^2 + 37^2 + 38^2 + 42^2 + 45^2 + 49^2 + 51^2 + 60^2 + 86^2 + 150^2 + 1200^2) - 25 \cdot 82,32^2]$$

## Charakteristiky polohy (průměrů $\bar{x}$ ) a rozptylu setříděného vzorku

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200

**Modus:**  $\bar{x}_{mo} = 18$

**Variační rozpětí:**  $R = 1\,200 - 10 = 1\,190$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200  
3

**Medián:**  $\bar{x}_{me} = 29$

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200  
12 12

**Aritmetický průměr:**  $\bar{x}_A = 82,32$

$$\frac{1}{25} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150 + 1\,200)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 32,73$

$$\sqrt[25]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150 \cdot 1\,200}$$

**Harmonický průměr:**  $\bar{x}_H \doteq 24,26$

$$\frac{25}{\frac{1}{10} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{2}{16} + \frac{3}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{26} + \frac{1}{29} + \frac{1}{32} + \frac{1}{35} + \frac{1}{37} + \frac{1}{38} + \frac{1}{42} + \frac{1}{45} + \frac{1}{49} + \frac{1}{51} + \frac{1}{60} + \frac{1}{86} + \frac{1}{150} + \frac{1}{1\,200}}$$

**Chronologický průměr:**  $\bar{x}_{ch} \doteq 60,54$

$$\frac{1}{2 \cdot (25-1)} \cdot (10 + 2 \cdot 11 + 2 \cdot 13 + 2 \cdot 14 + 4 \cdot 16 + 6 \cdot 18 + 2 \cdot 20 + 2 \cdot 24 + 2 \cdot 26 + 2 \cdot 29 + 2 \cdot 32 + 2 \cdot 35 + 2 \cdot 37 + 2 \cdot 38 + 2 \cdot 42 + 2 \cdot 45 + 2 \cdot 49 + 2 \cdot 51 + 2 \cdot 60 + 2 \cdot 86 + 2 \cdot 150 + 1\,200)$$

**Výběrový rozptyl:**  $S^2 \doteq 55\,104$

**Směrodatná odchylka:**  $S = \sqrt{55\,104} \doteq 235$

$$\frac{1}{25-1} \cdot [(10^2 + 11^2 + 13^2 + 14^2 + 2 \cdot 16^2 + 3 \cdot 18^2 + 20^2 + 24^2 + 26^2 + 29^2 + 32^2 + 35^2 + 37^2 + 38^2 + 42^2 + 45^2 + 49^2 + 51^2 + 60^2 + 86^2 + 150^2 + 1\,200^2) - 25 \cdot 82,32^2]$$

## Stejný vzorek s vynechanou poslední (extrémní) hodnotou

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200

**Variační rozpětí**

## Stejný vzorek s vynechanou poslední (extrémní) hodnotou

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1 200

**Modus**

**Variační rozpětí:**  $R = 150 - 10 = 140$  (dříve 1 190)



10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 1200

**Variační rozpětí:**  $R = 150 - 10 = 140$  (dříve 1 190)

## Medián



## Stejný vzorek s vynechanou poslední (extrémní) hodnotou

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150   1 200

**Modus:**  $\bar{x}_{mo} = 18$  (dříve 18)

**Variační rozpětí:**  $R = 150 - 10 = 140$  (dříve 1 190)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
3

**Medián:**  $\bar{x}_{me} = 27,5$  (dříve 29)

10 11 13 14 16 16 18 18 18 20 24  $\frac{26+29}{2}$  32 35 37 38 42 45 49 51 60 86 150  
11 11

**Aritmetický průměr:**  $\bar{x}_A = 35,75$  (dříve 82,32)

$$\frac{1}{24} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150)$$

**Geometrický průměr**

## Stejný vzorek s vynechanou poslední (extrémní) hodnotou

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150   1 200

**Modus:**  $\bar{x}_{mo} = 18$  (dříve 18)

**Variační rozpětí:**  $R = 150 - 10 = 140$  (dříve 1 190)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
3

**Medián:**  $\bar{x}_{me} = 27,5$  (dříve 29)

10 11 13 14 16 16 18 18 18 20 24  $\frac{26+29}{2}$  32 35 37 38 42 45 49 51 60 86 150  
11 11

**Aritmetický průměr:**  $\bar{x}_A = 35,75$  (dříve 82,32)

$$\frac{1}{24} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 28,17$  (dříve 32,73)

$$\sqrt[24]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150}$$

**Harmonický průměr**

## Stejný vzorek s vynechanou poslední (extrémní) hodnotou

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150   1 200

**Modus:**  $\bar{x}_{mo} = 18$  (dříve 18)

**Variační rozpětí:**  $R = 150 - 10 = 140$  (dříve 1 190)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
3

**Medián:**  $\bar{x}_{me} = 27,5$  (dříve 29)

10 11 13 14 16 16 18 18 18 20 24  $\frac{26+29}{2}$  32 35 37 38 42 45 49 51 60 86 150  
11 11

**Aritmetický průměr:**  $\bar{x}_A = 35,75$  (dříve 82,32)

$$\frac{1}{24} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 28,17$  (dříve 32,73)

$$\sqrt[24]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150}$$

**Harmonický průměr:**  $\bar{x}_H \doteq 23,31$  (dříve 24,26)

$$\frac{24}{\frac{1}{10} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{2}{16} + \frac{3}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{26} + \frac{1}{29} + \frac{1}{32} + \frac{1}{35} + \frac{1}{37} + \frac{1}{38} + \frac{1}{42} + \frac{1}{45} + \frac{1}{49} + \frac{1}{51} + \frac{1}{60} + \frac{1}{86} + \frac{1}{150}}$$

**Chronologický průměr**

## Stejný vzorek s vynechanou poslední (extrémní) hodnotou

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150   1 200

**Modus:**  $\bar{x}_{mo} = 18$  (dříve 18)

**Variační rozpětí:**  $R = 150 - 10 = 140$  (dříve 1 190)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
3

**Medián:**  $\bar{x}_{me} = 27,5$  (dříve 29)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
11  $\frac{26+29}{2}$  11

**Aritmetický průměr:**  $\bar{x}_A = 35,75$  (dříve 82,32)

$$\frac{1}{24} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 28,17$  (dříve 32,73)

$$\sqrt[24]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150}$$

**Harmonický průměr:**  $\bar{x}_H \doteq 23,31$  (dříve 24,26)

$$\frac{24}{\frac{1}{10} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{2}{16} + \frac{3}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{26} + \frac{1}{29} + \frac{1}{32} + \frac{1}{35} + \frac{1}{37} + \frac{1}{38} + \frac{1}{42} + \frac{1}{45} + \frac{1}{49} + \frac{1}{51} + \frac{1}{60} + \frac{1}{86} + \frac{1}{150}}$$

**Chronologický průměr:**  $\bar{x}_{Ch} \doteq 33,83$  (dříve 60,54)

$$\frac{1}{2 \cdot (24-1)} \cdot (10 + 2.11 + 2.13 + 2.14 + 4.16 + 6.18 + 2.20 + 2.24 + 2.26 + 2.29 + 2.32 + 2.35 + 2.37 + 2.38 + 2.42 + 2.45 + 2.49 + 2.51 + 2.60 + 2.86 + 150)$$

**Výběrový rozptyl**

## Stejný vzorek s vynechanou poslední (extrémní) hodnotou

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 **1 200**

**Modus:**  $\bar{x}_{mo} = 18$  (dříve 18)

**Variační rozpětí:**  $R = 150 - 10 = 140$  (dříve 1 190)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
3

**Medián:**  $\bar{x}_{me} = 27,5$  (dříve 29)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
11 11

**Aritmetický průměr:**  $\bar{x}_A = 35,75$  (dříve 82,32)

$$\frac{1}{24} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 28,17$  (dříve 32,73)

$$\sqrt[24]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150}$$

**Harmonický průměr:**  $\bar{x}_H \doteq 23,31$  (dříve 24,26)

$$\frac{24}{\frac{1}{10} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{2}{16} + \frac{3}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{26} + \frac{1}{29} + \frac{1}{32} + \frac{1}{35} + \frac{1}{37} + \frac{1}{38} + \frac{1}{42} + \frac{1}{45} + \frac{1}{49} + \frac{1}{51} + \frac{1}{60} + \frac{1}{86} + \frac{1}{150}}$$

**Chronologický průměr:**  $\bar{x}_{Ch} \doteq 33,83$  (dříve 60,54)

$$\frac{1}{2 \cdot (24-1)} \cdot (10 + 2 \cdot 11 + 2 \cdot 13 + 2 \cdot 14 + 4 \cdot 16 + 6 \cdot 18 + 2 \cdot 20 + 2 \cdot 24 + 2 \cdot 26 + 2 \cdot 29 + 2 \cdot 32 + 2 \cdot 35 + 2 \cdot 37 + 2 \cdot 38 + 2 \cdot 42 + 2 \cdot 45 + 2 \cdot 49 + 2 \cdot 51 + 2 \cdot 60 + 2 \cdot 86 + 150)$$

**Výběrový rozptyl:**  $S^2 \doteq 923$  (dříve 55 104)

**Směrodatná odchylka**

$$\frac{1}{24-1} \cdot [(10^2 + 11^2 + 13^2 + 14^2 + 2 \cdot 16^2 + 3 \cdot 18^2 + 20^2 + 24^2 + 26^2 + 29^2 + 32^2 + 35^2 + 37^2 + 38^2 + 42^2 + 45^2 + 49^2 + 51^2 + 60^2 + 86^2 + 150^2) - 24 \cdot 35,75^2]$$

## Stejný vzorek s vynechanou poslední (extrémní) hodnotou

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150 **1 200**

**Modus:**  $\bar{x}_{mo} = 18$  (dříve 18)

**Variační rozpětí:**  $R = 150 - 10 = 140$  (dříve 1 190)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
3

**Medián:**  $\bar{x}_{me} = 27,5$  (dříve 29)

10 11 13 14 16 16 18 18 18 20 24 26 29 32 35 37 38 42 45 49 51 60 86 150  
11 11

**Aritmetický průměr:**  $\bar{x}_A = 35,75$  (dříve 82,32)

$$\frac{1}{24} \cdot (10 + 11 + 13 + 14 + 2 \cdot 16 + 3 \cdot 18 + 20 + 24 + 26 + 29 + 32 + 35 + 37 + 38 + 42 + 45 + 49 + 51 + 60 + 86 + 150)$$

**Geometrický průměr:**  $\bar{x}_G \doteq 28,17$  (dříve 32,73)

$$\sqrt[24]{10 \cdot 11 \cdot 13 \cdot 14 \cdot 16^2 \cdot 18^3 \cdot 20 \cdot 24 \cdot 26 \cdot 29 \cdot 32 \cdot 35 \cdot 37 \cdot 38 \cdot 42 \cdot 45 \cdot 49 \cdot 51 \cdot 60 \cdot 86 \cdot 150}$$

**Harmonický průměr:**  $\bar{x}_H \doteq 23,31$  (dříve 24,26)

$$\frac{24}{\frac{1}{10} + \frac{1}{11} + \frac{1}{13} + \frac{1}{14} + \frac{2}{16} + \frac{3}{18} + \frac{1}{20} + \frac{1}{24} + \frac{1}{26} + \frac{1}{29} + \frac{1}{32} + \frac{1}{35} + \frac{1}{37} + \frac{1}{38} + \frac{1}{42} + \frac{1}{45} + \frac{1}{49} + \frac{1}{51} + \frac{1}{60} + \frac{1}{86} + \frac{1}{150}}$$

**Chronologický průměr:**  $\bar{x}_{Ch} \doteq 33,83$  (dříve 60,54)

$$\frac{1}{2 \cdot (24-1)} \cdot (10 + 2 \cdot 11 + 2 \cdot 13 + 2 \cdot 14 + 4 \cdot 16 + 6 \cdot 18 + 2 \cdot 20 + 2 \cdot 24 + 2 \cdot 26 + 2 \cdot 29 + 2 \cdot 32 + 2 \cdot 35 + 2 \cdot 37 + 2 \cdot 38 + 2 \cdot 42 + 2 \cdot 45 + 2 \cdot 49 + 2 \cdot 51 + 2 \cdot 60 + 2 \cdot 86 + 150)$$

**Výběrový rozptyl:**  $S^2 \doteq 923$  (dříve 55 104)

**Směrodatná odchylka:**  $S = \sqrt{923} \doteq 31$  (235)

$$\frac{1}{24-1} \cdot [(10^2 + 11^2 + 13^2 + 14^2 + 2 \cdot 16^2 + 3 \cdot 18^2 + 20^2 + 24^2 + 26^2 + 29^2 + 32^2 + 35^2 + 37^2 + 38^2 + 42^2 + 45^2 + 49^2 + 51^2 + 60^2 + 86^2 + 150^2) - 24 \cdot 35,75^2]$$



## Robustnost charakteristik polohy vůči extrémním hodnotám

Za velmi dobré míry polohy se právem považují modus ( $\bar{x}_{mo}$ , nejčtenější hodnota) a medián ( $\bar{x}_{me}$ , prostřední hodnota), protože nejsou přímo ovlivněny velikostí všech hodnot. To má výhodu zejména tehdy, když se ve výběru (tak jako v předchozím příkladě) vyskytuje náhodně jedna nebo několik málo mimořádně extrémních hodnot (vzhledem k ostatním hodnotám příliš velkých nebo příliš malých). V těchto případech nejsou modus ani medián ovlivněny těmito odlehlými hodnotami a poskytují tak dobrou představu o objektivní poloze nejčastější a prostřední hodnoty a tím i o úrovni (poloze) hodnot sledované proměnné.

Někdy se však necitlivost (robustnost) těchto měr považuje za jistou nevýhodu. Tuto nevýhodu překonávají některé **průměry**, což jsou střední hodnoty definované tak, že jsou funkcí všech hodnot dané proměnné, takže jsou více citlivé na odlehlé hodnoty (hodnoty, které se mimořádně liší od ostatních a dokáží proto průměr vychýlit natolik, že přestává daný výběr reprezentovat):

- hlavně aritmetický  $\bar{x}_A$  a chronologický  $\bar{x}_{ch}$  (z těch, které jsme si uváděli),
- dále pak kvadratický  $\bar{x}_k$  (ten jsme si neuváděli).

Naopak geometrický průměr  $\bar{x}_G$  a harmonický průměr  $\bar{x}_H$  nejsou příliš citlivé vůči několika málo extrémním hodnotám, jak jsme demonstrovali na předchozích dvou příkladech.

Pokud o některé hodnotě proměnné **rozhodneme**, že je odlehlým pozorováním (například analogií s **pravidlem 3  $\sigma$** , kdy za odlehlé pozorování považujeme to, které je od aritmetického průměru vzdáleno více jak trojnásobek směrodatné odchylky), je nutné ještě určit, proč je toto pozorování odlehlé.

- V případě, že známe příčinu a předpokládáme, že tato již nenastane (překlep v zápisu, prokazatelné selhání lidí či techniky, technologické chyby), jsme oprávněni tato pozorování vyloučit z dalšího zpracování, takzvaně „**očistit data**“.
- V ostatních případech je nutné zvážit, zda se vyloučením odlehlých pozorování nepřipravíme o důležité informace o jevech vyskytujících se s nízkou četností.

### 3. Zpracování statistického materiálu

Jak jsme již uvedli na začátku této kapitoly, úkolem popisné statistiky je: uspořádat pozorované hodnoty proměnné do názornější formy (tabulka) a popsat je několika málo hodnotami (číselnými charakteristikami), které by obsahovaly co největší množství informací obsažených v původním souboru. Jak se to provádí prakticky, si nyní ukážeme.

#### 3.1. Menší vzorek

Máme k dispozici následující data (údaje), o kterých nám není známo, že pocházejí z příkladu popisujícího počet padlých „*líců*“ při současném hodu čtyřmi rozlišitelnými mincemi. Proto ani netušíme, že by mohlo jít o binomické rozdělení. ***Máme pouze tato sesbíraná data:***

2   2   1   2   3   1   2   3   1   2   3   0   1   2   3   4

která setřídíme a zapíšeme do následující tabulky.

Každá v datech vyskytující se cifra bude mít svůj vlastní sloupeček.

cifra	0	1	2	3	4
počet výskytů	1	4	6	4	1

Zajímá nás, jak můžeme určit požadované číselné **charakteristiky** z takto získaných a do tabulky zapsaných hodnot. Protože bychom přidávali další řádky s mezivýsledky, je lépe psát tabulku svisle a potom můžeme přidávat sloupce dle libosti.

**Poznámka** Pokud by případných řádků v tabulce mělo být více (viz následující **příklad**) a tabulka by se stávala nepřehlednou, zařadíme vždy podobné hodnoty do jedné třídy (viz **obrázek**)

2 2 1 2 3 1 2 3 1 2 3 0 1 2 3 4

třída index $k$	reprezentant $x_k$					
1	0					
2	1					
3	2					
4	3					
5	4					
$\Sigma$						

$k$  označuje číslo řádku tabulky, navíc jej nazveme **třídou**.

$x_k$  nazveme **reprezentantem** této třídy.

aritmetický průměr

rozptyl

2 2 1 2 3 1 2 3 1 2 3 0 1 2 3 4

třída index $k$	reprezentant $x_k$	četnost $n_k$				
1	0	1				
2	1	4				
3	2	6				
4	3	4				
5	4	1				
$\Sigma$		<b>n=16</b>				

$k$  označuje číslo řádku tabulky, navíc jej nazveme **třídou**.

$x_k$  nazveme **reprezentantem** této třídy.

**Četnost**  $n_k$  udává, kolikrát se daný reprezentant  $x_k$  v souboru dat vyskytuje.

aritmetický průměr

rozptyl

2 2 1 2 3 1 2 3 1 2 3 0 1 2 3 4

třída index $k$	reprezentant $x_k$	četnost $n_k$				
1	0	1				
2	1	4				
3	2	6				
4	3	4				
5	4	1				
$\Sigma$		<b>n=16</b>				

$k$  označuje číslo řádku tabulky, navíc jej nazveme **třídou**.

$x_k$  nazveme **reprezentantem** této třídy.

**Četnost**  $n_k$  udává, kolikrát se daný reprezentant  $x_k$  v souboru dat vyskytuje. Pokud bychom četnost podělili počtem prvků ( $n_k/n$ ), dostaneme **relativní četnost** (v procentech). Srovnej s „klasickou“ pravděpodobností:

příznivé případy  
všechny **možné**

aritmetický průměr

rozptyl

$$2 \quad 2 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 0 \quad 1 \quad 2 \quad 3 \quad 4$$

$$2 + 2 + 1 + 2 + 3 + 1 + 2 + 3 + 1 + 2 + 3 + 0 + 1 + 2 + 3 + 4 = 32$$

třída index $k$	reprezentant $x_k$	četnost $n_k$	$x_k \cdot n_k$			
1	0	1	0			
2	1	4	4			
3	2	6	12			
4	3	4	12			
5	4	1	4			
$\Sigma$		<b>n=16</b>	<b>32</b>			

$k$  označuje číslo řádku tabulky, navíc jej nazveme **třídou**.

$x_k$  nazveme **reprezentantem** této třídy.

**Četnost**  $n_k$  udává, kolikrát se daný reprezentant  $x_k$  v souboru dat vyskytuje. Pokud bychom četnost podělili počtem prvků ( $n_k/n$ ), dostaneme **relativní četnost** (v procentech). Srovnej s „klasickou“ pravděpodobností:  
příznivé případy  
 všechny **možné**

aritmetický průměr

$$\bar{x}_A = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{n} \cdot \sum_{i=1}^k n_i \cdot x_i = \frac{1}{16} \cdot 32 = 2$$

rozptyl

$$2 \quad 2 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 0 \quad 1 \quad 2 \quad 3 \quad 4$$

$$2 + 2 + 1 + 2 + 3 + 1 + 2 + 3 + 1 + 2 + 3 + 0 + 1 + 2 + 3 + 4 = 32$$

třída index $k$	reprezentant $x_k$	četnost $n_k$	$x_k \cdot n_k$	$x_k - \bar{x}$		
1	0	1	0	-2		
2	1	4	4	-1		
3	2	6	12	0		
4	3	4	12	1		
5	4	1	4	2		
$\Sigma$		<b>n=16</b>	<b>32</b>	<b>0</b>		

$k$  označuje číslo řádku tabulky, navíc jej nazveme **třídou**.

$x_k$  nazveme **reprezentantem** této třídy.

**Četnost**  $n_k$  udává, kolikrát se daný reprezentant  $x_k$  v souboru dat vyskytuje. Pokud bychom četnost podělili počtem prvků ( $n_k/n$ ), dostaneme **relativní četnost** (v procentech). Srovnej s „klasickou“ pravděpodobností:

příznivé případy  
všechny **možné**

aritmetický průměr

$$\bar{x}_A = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{n} \cdot \sum_{i=1}^k n_i \cdot x_i = \frac{1}{16} \cdot 32 = 2$$

rozptyl

$$2 \quad 2 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 0 \quad 1 \quad 2 \quad 3 \quad 4$$

$$2 + 2 + 1 + 2 + 3 + 1 + 2 + 3 + 1 + 2 + 3 + 0 + 1 + 2 + 3 + 4 = 32$$

třída index $k$	reprezentant $x_k$	četnost $n_k$	$x_k \cdot n_k$	$x_k - \bar{x}$	$(x_k - \bar{x})^2$	
1	0	1	0	-2	4	
2	1	4	4	-1	1	
3	2	6	12	0	0	
4	3	4	12	1	1	
5	4	1	4	2	4	
$\Sigma$		<b>n=16</b>	<b>32</b>	<b>0</b>		

$k$  označuje číslo řádku tabulky, navíc jej nazveme **třídou**.

$x_k$  nazveme **reprezentantem** této třídy.

**Četnost**  $n_k$  udává, kolikrát se daný reprezentant  $x_k$  v souboru dat vyskytuje. Pokud bychom četnost podělili počtem prvků  $(n_k/n)$ , dostaneme **relativní četnost** (v procentech). Srovnej s „klasickou“ pravděpodobností:

příznivé případy  
všechny **možné**

aritmetický průměr

$$\bar{x}_A = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{n} \cdot \sum_{i=1}^k n_i \cdot x_i = \frac{1}{16} \cdot 32 = 2$$

rozptyl



$$2 \quad 2 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 0 \quad 1 \quad 2 \quad 3 \quad 4$$

$$2 + 2 + 1 + 2 + 3 + 1 + 2 + 3 + 1 + 2 + 3 + 0 + 1 + 2 + 3 + 4 = 32$$

třída index $k$	reprezentant $x_k$	četnost $n_k$	$x_k \cdot n_k$	$x_k - \bar{x}$	$(x_k - \bar{x})^2$	$n_k \cdot (x_k - \bar{x})^2$
1	0	1	0	-2	4	4
2	1	4	4	-1	1	4
3	2	6	12	0	0	0
4	3	4	12	1	1	4
5	4	1	4	2	4	4
$\Sigma$		<b>n=16</b>	<b>32</b>	<b>0</b>		<b>16</b>

$k$  označuje číslo řádku tabulky, navíc jej nazveme **třídou**.

$x_k$  nazveme **reprezentantem** této třídy.

**Četnost**  $n_k$  udává, kolikrát se daný reprezentant  $x_k$  v souboru dat vyskytuje. Pokud bychom četnost podělili počtem prvků  $(n_k/n)$ , dostaneme **relativní četnost** (v procentech). Srovnej s „klasickou“ pravděpodobností:  
příznivé případy  
 všechny **možné**

aritmetický průměr 
$$\bar{x}_A = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{n} \cdot \sum_{i=1}^k n_i \cdot x_i = \frac{1}{16} \cdot 32 = 2$$

rozptyl

$$2 \quad 2 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 0 \quad 1 \quad 2 \quad 3 \quad 4$$

$$2 + 2 + 1 + 2 + 3 + 1 + 2 + 3 + 1 + 2 + 3 + 0 + 1 + 2 + 3 + 4 = 32$$

třída index $k$	reprezentant $x_k$	četnost $n_k$	$x_k \cdot n_k$	$x_k - \bar{x}$	$(x_k - \bar{x})^2$	$n_k \cdot (x_k - \bar{x})^2$
1	0	1	0	-2	4	4
2	1	4	4	-1	1	4
3	2	6	12	0	0	0
4	3	4	12	1	1	4
5	4	1	4	2	4	4
$\Sigma$		<b>n=16</b>	<b>32</b>	<b>0</b>		<b>16</b>

$k$  označuje číslo řádku tabulky, navíc jej nazveme **třídou**.

$x_k$  nazveme **reprezentantem** této třídy.

**Četnost**  $n_k$  udává, kolikrát se daný reprezentant  $x_k$  v souboru dat vyskytuje. Pokud bychom četnost podělili počtem prvků  $(n_k/n)$ , dostaneme **relativní četnost** (v procentech). Srovnej s „klasickou“ pravděpodobností:

příznivé případy  
všechny **možné**

aritmetický průměr  $\bar{x}_A = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{n} \cdot \sum_{i=1}^k n_i \cdot x_i = \frac{1}{16} \cdot 32 = 2$

(výběrový) rozptyl  $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \sum_{i=1}^k n_i \cdot (x_i - \bar{x})^2 = \frac{1}{16-1} \cdot 16 \doteq 1,067$

### 3.2. Rozsáhlý vzorek

Níže uvedená data zařadte do tříd a poté vypočítejte aritmetický průměr, geometrický průměr, harmonický průměr, (výběrový) rozptyl, směrodatnou odchylku, variační koeficient a sestavte interval  $3\sigma$ .

60	154	122	90	105	38	100	82	82	15	125	90	160	76
73	40	50	48	12	120	90	70	70	80	132	50	87	49
92	148	85	98	140	110	70	48	48	151	48	80	52	98

**Postup** při „ručním“ zpracování:

1. Nalezneme nejmenší a největší prvek a určíme **variační rozpětí vzorku**.
2. Rozhodneme se, **do kolika** (minimum je pět tříd a maximum 20 tříd; nejčastěji 8 až 13) jak **velkých tříd** (doporučuje se, aby třídy měly stejnou délku) budeme data zařazovat.
  - Pokud se zvolí malý počet tříd, dojde při třídění k výrazné ztrátě informace o průběhu původního znaku. Pokud se naopak zvolí příliš velký počet tříd (s malými četnostmi), bude vzniklá tabulka nepřehledná.
  - Délku intervalu (třídy) volíme tak, aby hranice intervalů byla dobře zapamatovatelná (případně zaokrouhlená) čísla<sup>26</sup>, intervaly jednoznačně pokrývaly celý obor hodnot popisovaného znaku (nesmí se stát, že by některá hodnota nepatřila do žádné třídy) a oba krajní intervaly rozdělení měly nenulové četnosti.
3. Začneme vyplňovat následující **tabulku** rozdělení četností, kterou doplníme o další sloupce hodnot, pomocí kterých pak určíme požadované číselné charakteristiky.

<sup>26</sup> Jindy zase raději požadujeme, aby reprezentanti jednotlivých tříd (většinou středy těchto tříd) byla dobře zapamatovatelná (případně zaokrouhlená) čísla (viz [obrázek](#)).

## Třídění dat

Nejdříve data zařadíte do **devíti** tříd.

60	154	122	90	105	38	100	82	85	15	125	90	<b>160</b>	76
73	40	50	48	<b>12</b>	120	90	70	55	80	132	50	87	49
92	148	85	98	140	110	70	48	149	151	48	80	52	98

**ad 1.** Variační rozpětí:  $R = x_{max} - x_{min} = 160 - 12 = 148$ .

**ad 2.** Chceme-li data rozdělit do **9 tříd** ( $148 : 9 = 16,4$ ), volíme **šířku třídy 17**.

Pak:  $9 \cdot 17 - R = 5$ , což rozdělíme na obě strany:  $5 : 2 = 2,5$ .

První třída bude potom mít **počátek**:  $x_{min} - 2,5 = 9,5$     a **konec**:  $9,5 + \text{šířka třídy} = 9,5 + 17 = 26,5$ .

**ad 3.** Vše budeme zapisovat do tabulky.

k	třída — interval <b>šířky 17</b> ( počátek ; konec=počátek+17 )
1	( $x_{min} - 2,5$ ; $26,5 = /12 - 2,5/ + 17$ )
2	( $26,5$ ; $43,5 = 26,5 + 17$ )
3	( $43,5$ ; $43,5 + 17$ )
4	( $60,5$ ; ... )
5	( ... ; ... )
6	( ... ; ... )
7	( ... ; ... )
8	( ... ; ... )
9	( ... ; $x_{max} + 2,5$ )

## Třídění dat

Po zařazení dat do devíti tříd vypočítejte nejprve **aritmetický** průměr a (výběrový) **rozptyl**.

60   154   122   90   105   38   100   82   85   **15**   125   90   160   76  
 73   40   50   48   **12**   120   90   70   55   80   132   50   87   49  
 92   148   85   98   140   110   70   48   149   151   48   80   52   98

Kolik prvků do každé třídy patří?

k	třída	četnost				
1	(9,5 ; 26,5)	12; 15				
2	(26,5 ; 43,5)					
3	(43,5 ; 60,5)					
4	(60,5 ; 77,5)					
5	(77,5 ; 94,5)					
6	(94,5 ; 111,5)					
7	(111,5 ; 128,5)					
8	(128,5 ; 145,5)					
9	(145,5 ; 162,5)					
Σ						

## Třídění dat

Po zařazení dat do devíti tříd vypočítejte nejprve **aritmetický** průměr a (výběrový) **rozptyl**.

60   154   122   90   105   38   100   82   85   **15**   125   90   160   76  
 73   40   50   48   **12**   120   90   70   55   80   132   50   87   49  
 92   148   85   98   140   110   70   48   149   151   48   80   52   98

Kolik prvků do každé třídy patří?

Četnost dané třídy si označíme  $n_k$ .

k	třída	četnost				
1	(9,5 ; 26,5)	12; 15 ⇒				
2	(26,5 ; 43,5)					
3	(43,5 ; 60,5)					
4	(60,5 ; 77,5)					
5	(77,5 ; 94,5)					
6	(94,5 ; 111,5)					
7	(111,5 ; 128,5)					
8	(128,5 ; 145,5)					
9	(145,5 ; 162,5)					
Σ						

## Třídění dat

Po zařazení dat do devíti tříd vypočítejte nejprve **aritmetický** průměr a (výběrový) **rozptyl**.

60 154 122 90 105 38 100 82 85 **15** 125 90 160 76  
 73 40 50 48 **12** 120 90 70 55 80 132 50 87 49  
 92 148 85 98 140 110 70 48 149 151 48 80 52 98

Kolik prvků do každé třídy patří?

Jakého bude mít třída reprezentanta (my si zvolíme střed)?

Četnost dané třídy si označíme  $n_k$ .

k	třída	četnost	$x_k$	$n_k$		
1	(9,5 ; 26,5)	12; 15 $\Rightarrow$	18			
2	(26,5 ; 43,5)		35			
3	(43,5 ; 60,5)		52			
4	(60,5 ; 77,5)		69			
5	(77,5 ; 94,5)		86			
6	(94,5 ; 111,5)		103			
7	(111,5 ; 128,5)		120			
8	(128,5 ; 145,5)		137			
9	(145,5 ; 162,5)		154			
$\Sigma$						

## Třídění dat

Po zařazení dat do devíti tříd vypočítejte nejprve **aritmetický** průměr a (výběrový) **rozptyl**.

60 154 122 90 105 38 100 82 85 **15** 125 90 160 76  
 73 40 50 48 **12** 120 90 70 55 80 132 50 87 49  
 92 148 85 98 140 110 70 48 149 151 48 80 52 98

Kolik prvků do každé třídy patří?

Jakého bude mít třída reprezentanta (my si zvolíme střed)?

Četnost dané třídy si označíme  $n_k$ .

Do tabulky doplníme další potřebné sloupce.

k	třída	četnost	$x_k$	$n_k$	$x_k \cdot n_k$	
1	(9,5 ; 26,5)	12; 15 $\Rightarrow$	18	2		
2	(26,5 ; 43,5)		35	2		
3	(43,5 ; 60,5)		52	9		
4	(60,5 ; 77,5)		69	5		
5	(77,5 ; 94,5)		86	10		
6	(94,5 ; 111,5)		103	5		
7	(111,5 ; 128,5)		120	3		
8	(128,5 ; 145,5)		137	2		
9	(145,5 ; 162,5)		154	4		
$\Sigma$				<b>n=42</b>		



## Třídění dat

Po zařazení dat do devíti tříd vypočítejte nejprve **aritmetický** průměr a (výběrový) **rozptyl**.

60 154 122 90 105 38 100 82 85 **15** 125 90 160 76  
 73 40 50 48 **12** 120 90 70 55 80 132 50 87 49  
 92 148 85 98 140 110 70 48 149 151 48 80 52 98

Kolik prvků do každé třídy patří?

Jakého bude mít třída reprezentanta (my si zvolíme střed)?

Četnost dané třídy si označíme  $n_k$ .

Do tabulky doplníme další potřebné sloupce.

k	třída	četnost	$x_k$	$n_k$	$x_k \cdot n_k$	$(x_k - \bar{x}_A)^2 \cdot n_k$
1	(9,5 ; 26,5)	12; 15 $\Rightarrow$	18	2	36	
2	(26,5 ; 43,5)		35	2	70	
3	(43,5 ; 60,5)		52	9	468	
4	(60,5 ; 77,5)		69	5	345	
5	(77,5 ; 94,5)		86	10	860	
6	(94,5 ; 111,5)		103	5	515	
7	(111,5 ; 128,5)		120	3	360	
8	(128,5 ; 145,5)		137	2	274	
9	(145,5 ; 162,5)		154	4	616	
$\Sigma$				<b>n=42</b>	<b>3 544</b>	

### Aritmetický průměr

$$\bar{x}_A = \frac{1}{n} \cdot \sum_{\forall k} x_k \cdot n_k =$$

$$\frac{1}{42} \cdot 3\,544 \doteq 84,4$$

## Třídění dat

Po zařazení dat do devíti tříd vypočítejte nejprve **aritmetický** průměr a (výběrový) **rozptyl**.

60 154 122 90 105 38 100 82 85 **15** 125 90 160 76  
 73 40 50 48 **12** 120 90 70 55 80 132 50 87 49  
 92 148 85 98 140 110 70 48 149 151 48 80 52 98

Kolik prvků do každé třídy patří?

Jakého bude mít třída reprezentanta (my si zvolíme střed)?

Četnost dané třídy si označíme  $n_k$ .

Do tabulky doplníme další potřebné sloupce.

k	třída	četnost	$x_k$	$n_k$	$x_k \cdot n_k$	$(x_k - \bar{x}_A)^2 \cdot n_k$
1	(9,5 ; 26,5)	12; 15 $\Rightarrow$	18	2	36	8 817,92
2	(26,5 ; 43,5)		35	2	70	4 880,72
3	(43,5 ; 60,5)		52	9	468	9 447,84
4	(60,5 ; 77,5)		69	5	345	1 185,80
5	(77,5 ; 94,5)		86	10	860	25,60
6	(94,5 ; 111,5)		103	5	515	1 729,80
7	(111,5 ; 128,5)		120	3	360	3 802,08
8	(128,5 ; 145,5)		137	2	274	5 533,52
9	(145,5 ; 162,5)		154	4	616	19 376,64
$\Sigma$				<b>n=42</b>	<b>3 544</b>	<b>54 799,92</b>

### Aritmetický průměr

$$\bar{x}_A = \frac{1}{n} \cdot \sum_{\forall k} x_k \cdot n_k =$$

$$\frac{1}{42} \cdot 3\,544 \doteq 84,4$$

## Třídění dat

Po zařazení dat do devíti tříd vypočítejte nejprve **aritmetický** průměr a (výběrový) **rozptyl**.

60 154 122 90 105 38 100 82 85 **15** 125 90 160 76  
 73 40 50 48 **12** 120 90 70 55 80 132 50 87 49  
 92 148 85 98 140 110 70 48 149 151 48 80 52 98

Kolik prvků do každé třídy patří?

Jakého bude mít třída reprezentanta (my si zvolíme střed)?

Četnost dané třídy si označíme  $n_k$ .

Do tabulky doplníme další potřebné sloupce.

k	třída	četnost	$x_k$	$n_k$	$x_k \cdot n_k$	$(x_k - \bar{x}_A)^2 \cdot n_k$
1	(9,5 ; 26,5)	12; 15 $\Rightarrow$	18	2	36	8 817,92
2	(26,5 ; 43,5)		35	2	70	4 880,72
3	(43,5 ; 60,5)		52	9	468	9 447,84
4	(60,5 ; 77,5)		69	5	345	1 185,80
5	(77,5 ; 94,5)		86	10	860	25,60
6	(94,5 ; 111,5)		103	5	515	1 729,80
7	(111,5 ; 128,5)		120	3	360	3 802,08
8	(128,5 ; 145,5)		137	2	274	5 533,52
9	(145,5 ; 162,5)		154	4	616	19 376,64
$\Sigma$				<b>n=42</b>	<b>3 544</b>	<b>54 799,92</b>

### Aritmetický průměr

$$\bar{x}_A = \frac{1}{n} \cdot \sum_{\forall k} x_k \cdot n_k =$$

$$\frac{1}{42} \cdot 3\,544 \doteq 84,4$$

### Výběrový rozptyl

$$S^2 = \frac{\sum_{\forall k} (x_k - \bar{x}_A)^2 \cdot n_k}{n - 1} =$$

$$= \frac{54\,799,92}{42 - 1} \doteq 1\,337$$

Budeme-li požadovat i další číselné charakteristiky, doplníme tabulku o další potřebné sloupce.

## Určení číselných charakteristik

Vypočítejte **aritmetický** průměr, **geometrický** průměr, **harmonický** průměr, (výběrový) **rozptyl**, směrodatnou **odchylku**, **variační koeficient** a sestavte **interval  $3\sigma$** .

k	třída	$x_k$	četnost	$n_k$	$x_k \cdot n_k$			
1	(9,5 ; 26,5)	18		2	36			
2	(26,5 ; 43,5)	35		2	70			
3	(43,5 ; 60,5)	52		9	468			
4	(60,5 ; 77,5)	69		5	345			
5	(77,5 ; 94,5)	86		10	860			
6	(94,5 ; 111,5)	103		5	515			
7	(111,5 ; 128,5)	120		3	360			
8	(128,5 ; 145,5)	137		2	274			
9	(145,5 ; 162,5)	154		4	616			
<b><math>\Sigma</math></b>				<b>n=42</b>	<b>3 544</b>			

## Určení číselných charakteristik

Vypočítejte **aritmetický** průměr, **geometrický** průměr, **harmonický** průměr, (výběrový) **rozptyl**, směrodatnou **odchylku**, **variační koeficient** a sestavte **interval  $3\sigma$** .

k	třída	$x_k$	četnost	$n_k$	$x_k \cdot n_k$	$x_k^2 \cdot n_k$		
1	(9,5 ; 26,5)	18		2	36	648		
2	(26,5 ; 43,5)	35		2	70	2 450		
3	(43,5 ; 60,5)	52		9	468	24 336		
4	(60,5 ; 77,5)	69		5	345	23 805		
5	(77,5 ; 94,5)	86		10	860	73 960		
6	(94,5 ; 111,5)	103		5	515	53 045		
7	(111,5 ; 128,5)	120		3	360	43 200		
8	(128,5 ; 145,5)	137		2	274	37 538		
9	(145,5 ; 162,5)	154		4	616	94 864		
<b><math>\Sigma</math></b>				<b>n=42</b>	<b>3 544</b>	<b>353 846</b>		

## Určení číselných charakteristik

Vypočítejte **aritmetický** průměr, **geometrický** průměr, **harmonický** průměr, (výběrový) **rozptyl**, směrodatnou **odchylku**, **variační koeficient** a sestavte **interval  $3\sigma$** .

k	třída	$x_k$	četnost	$n_k$	$x_k \cdot n_k$	$x_k^2 \cdot n_k$	$x_k^{\frac{n_k}{n}}$	
1	(9,5 ; 26,5)	18		2	36	648	1,148	
2	(26,5 ; 43,5)	35		2	70	2 450	1,184	
3	(43,5 ; 60,5)	52		9	468	24 336	2,332	
4	(60,5 ; 77,5)	69		5	345	23 805	1,655	
5	(77,5 ; 94,5)	86		10	860	73 960	2,888	
6	(94,5 ; 111,5)	103		5	515	53 045	1,736	
7	(111,5 ; 128,5)	120		3	360	43 200	1,408	
8	(128,5 ; 145,5)	137		2	274	37 538	1,264	
9	(145,5 ; 162,5)	154		4	616	94 864	1,616	
<b><math>\Sigma</math></b>				<b>n=42</b>	<b>3 544</b>	<b>353 846</b>	<b><math>\Pi</math></b>	

**75,638**

## Určení číselných charakteristik

Vypočítejte **aritmetický** průměr, **geometrický** průměr, **harmonický** průměr, (výběrový) **rozptyl**, směrodatnou **odchylku**, **variační koeficient** a sestavte **interval**  $3\sigma$ .

k	třída	$x_k$	četnost	$n_k$	$x_k \cdot n_k$	$x_k^2 \cdot n_k$	$x_k^{\frac{n_k}{n}}$	$\frac{n_k}{x_k}$
1	(9,5 ; 26,5)	18		2	36	648	1,148	0,111
2	(26,5 ; 43,5)	35		2	70	2 450	1,184	0,057
3	(43,5 ; 60,5)	52		9	468	24 336	2,332	0,173
4	(60,5 ; 77,5)	69		5	345	23 805	1,655	0,072
5	(77,5 ; 94,5)	86		10	860	73 960	2,888	0,116
6	(94,5 ; 111,5)	103		5	515	53 045	1,736	0,049
7	(111,5 ; 128,5)	120		3	360	43 200	1,408	0,025
8	(128,5 ; 145,5)	137		2	274	37 538	1,264	0,015
9	(145,5 ; 162,5)	154		4	616	94 864	1,616	0,026
<b>Σ</b>				<b>n=42</b>	<b>3 544</b>	<b>353 846</b>	<b>Π</b>	<b>0,644</b>

**75,638**

Tedy:  $\sum_{k=1}^9 n_k = n = 42$     $\sum_k x_k \cdot n_k = 3544$     $\sum_k x_k^2 \cdot n_k = 353846$     $\prod_k x_k^{\frac{n_k}{n}} = 75,638$     $\sum_k \frac{n_k}{x_k} = 0,644$

Tedy:  $\sum_{k=1}^9 n_k = n = 42$     $\sum_k x_k \cdot n_k = 3544$     $\sum_k x_k^2 \cdot n_k = 353846$     $\prod_k x_k^{\frac{n_k}{n}} = 75,638$     $\sum_k \frac{n_k}{x_k} = 0,644$

Vypočítejte **aritmetický** průměr, **geometrický** průměr, **harmonický** průměr, (výběrový) **rozptyl**, směrodatnou **odchylku**, **variační koeficient** a sestavte **interval  $3\sigma$** .

## Určení dalších charakteristik

**Geometrický průměr:**  $\bar{x}_G = \prod_{k=1}^9 x_k^{\frac{n_k}{n}} \doteq 75,6$

**Harmonický průměr:**  $\bar{x}_H = \frac{n}{\sum_{k=1}^9 \frac{n_k}{x_k}} = \frac{42}{0,644} \doteq 65,2$

**Rozptyl:**  $S^2 = \frac{1}{n-1} \cdot \left[ \sum_{k=1}^9 x_k^2 \cdot n_k - n \cdot \bar{x}_A^2 \right] = \frac{1}{42-1} \cdot (353846 - 42 \cdot 84,381^2) \doteq \frac{54799}{41} \doteq 1337$

**Směrodatná odchylka:**  $S = \sqrt{S^2} = \sqrt{1337} \doteq 37$  ( $\doteq 40$ )

**Variační koeficient:**  $V = \frac{S}{\bar{x}_A} = \frac{37}{84,4} \doteq 0,44$

**Interval  $3\sigma$**  (pouze pro normální rozdělení!)  $= \langle \bar{x}_A - 3 \cdot S ; \bar{x}_A + 3 \cdot S \rangle = \langle -25 ; 194 \rangle$

**Poznámka:** Mohli jsme také volit například 10 tříd o rozpětí 16. Tím bychom sice měli hranice celočíselné, ale měli bychom třídy ( 74 ; 90 ) a ( 90 ; 106 ). A do které z nich potom zařadíme číslo **90**, které se jenom v prvním řádku zadaných dat vyskytuje dvakrát a potom ještě jednou ve druhém řádku? Tomuto problému jsme se díky neceločíselným hranicím vyhnuli.

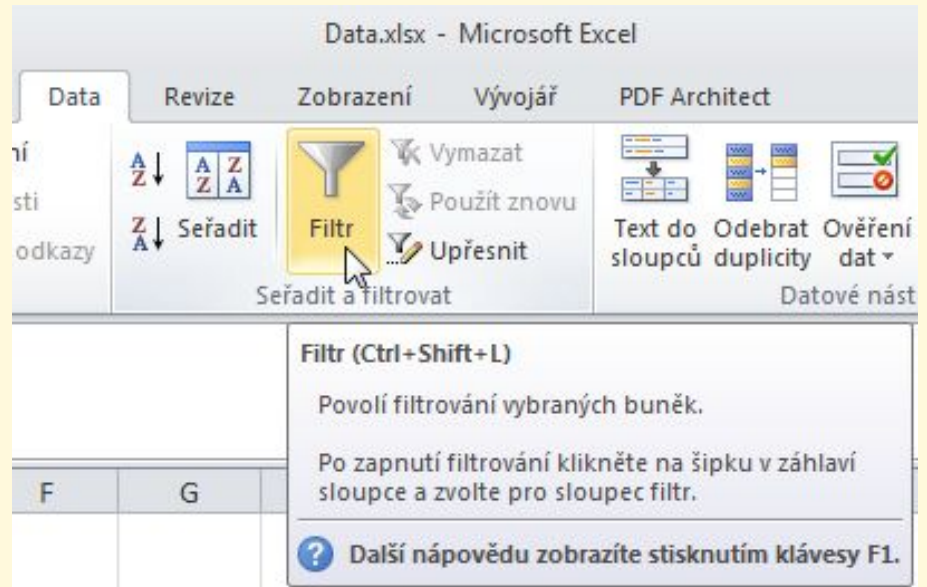


## 4. Využití programu Excel 2010

Velký význam pro rozvoj a využití statistických metod měl nástup výpočetní techniky, zejména osobních počítačů. Počítač vítězí nad člověkem především v těch úkonech, které jsou pro člověka tradičně nejzdlouhavější — při třídění, vyhledávání a výpočtech s velkým množstvím dat.

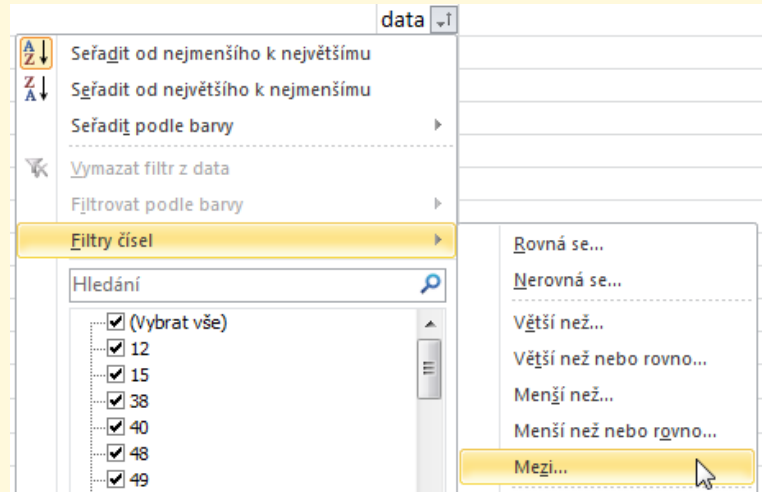
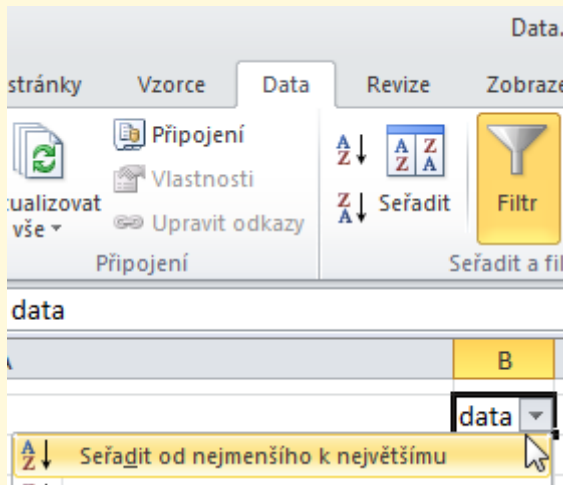
Takže například vyplňování předchozí tabulky bychom zvládli za použití **Excelu** s poněkud menším úsilím.

Stačí napsat všechny hodnoty do sloupce pod sebe, na kartě [Data] v záložce [Seřadit a filtrovat] zvolit nabídku [Filtr],



rozbalit nabídku pod objevivším se [trojúhelníkem],

vybrat si vhodnou funkci (například)



a nechat si seřadit data podle velikosti. Tím lehce určíme nejmenší a největší prvek a můžeme stanovit třídy.

Pokud nás zajímá **četnost** konkrétní třídy (tedy kolik a jakých konkrétně je v ní prvků) — například první třídy ( 9,5 ; 26,5 ) — naprosto stejným postupem si vybereme pouze jinou vhodnou funkci.

Vlastní automatický filtr

Zobrazit řádky:  
data

Je větší než nebo rovno 9,5

☒ A ☐ Nebo

Je menší než nebo rovno 26,5

Znak ? zastupuje jeden znak.  
Znak \* zastupuje posloupnost znaků.

OK Storno

2	data
3	12
4	15
45	
46	

Mohli bychom postupovat i jinak. Vedle dat (mohou být seřazená podle velikosti anebo v původním pořadí tak, jak byla zadána – na další postup to nemá naprosto vliv) do jiného sloupce napíšeme horní hranice jednotlivých tříd.

Potom volíme následující položky menu: [Data] [Analýza] [Analýza dat] [Histogram] <sup>27</sup>

horní hranice

26,5
43,5
60,5
77,5
94,5
111,5
128,5
145,5
162,5

Analýza dat

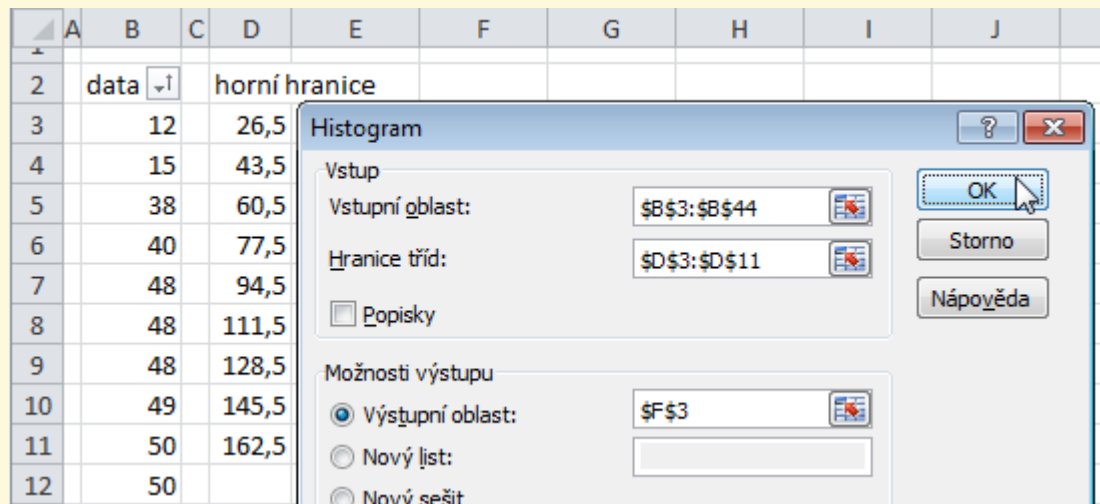
Analytické nástroje:

- Anova: dva faktory bez opakování
- Korelace
- Kovariance
- Popisná statistika
- Exponenciální vyrovnání
- Dvouvýběrový F-test pro rozptyl
- Fourierova analýza
- Histogram**
- Klouzavý průměr
- Generátor pseudonáhodných čísel

OK Storno Nápoředa

<sup>27</sup> Pokud výše uvedenou nabídku [Histogram] nemůžeme najít, pravděpodobně tento doplněk na konkrétním počítači není nainstalován. V tom případě postupujeme následovně: [Soubor] [Možnosti] [Doplňky] [Spravovat] [Doplňky aplikace Excel] [Přejít] a přidáme [Analytické nástroje] [OK].

a doplníme patřičné parametry (nejlépe označováním oblastí pomocí myši):



- [Vstupní oblast] — sloupcový vektor, ve kterém jsou zadaná data;
- [Hranice tříd] — sloupcový vektor, do kterého jsme zadali horní hranice všech tříd.  
**Poznámka:** Pokud bychom nezadali horní hranici poslední třídy, četnost této poslední třídy by se objevila v řádku označeném Další. Takhle je tam uvedena NULA.
- [Výstupní oblast] — označuje levou horní buňku, od které program *Excel* začne vypisovat tabulku četností jednotlivých tříd (viz následující obrázek).

	A	B	C	D	E	F	G
1							
2		data	↓1	horní hranice			
3		12		26,5		<i>Třídy</i>	<i>Četnost</i>
4		15		43,5		26,5	2
5		38		60,5		43,5	2
6		40		77,5		60,5	9
7		48		94,5		77,5	5
8		48		111,5		94,5	10
9		48		128,5		111,5	5
10		49		145,5		128,5	3
11		50		162,5		145,5	2
12		50				162,5	4
13		52				Další	0

Do nově vzniklé tabulky pak stačí dopsat případně počátky, ale hlavně **reprezentanty** jednotlivých tříd a do dalších sloupců pak doplnit další hodnoty podle vztahů tak, jak jsme je vyplňovali „ručně“.

**Poznámka:** Účelem tohoto kurzu není čtenáře naučit bravurně ovládat konkrétní statistický software, ale umožnit mu pochopení a zvládnutí dané problematiky tak, aby byl schopen si poradit i v případech, kdy v dosahu nemá příslušné počítačové vybavení, na které byl zaučen. To znamená, že v dalším nebudeme příliš často uvádět jednotlivé statistické funkce<sup>28</sup>, ale zmíníme se o nich pouze tam, kde to bude z didaktického hlediska vhodné (například náhrada statistických tabulek). Přednost budeme dávat běžným funkcím tabulkových kalkulátorů při dosazování do uvedených vzorců.

<sup>28</sup> Jen například pro výpočet rozptylu uvádí *Excel 2010* tyto 4 možnosti: VAR.P, VAR.S, VARA a VARPA. A kdo si není jist tím, co vlastně chce počítat, má tedy pouze 25% pravděpodobnost, že zvolí tu správnou z nich.

Navíc *Excel 2007* disponuje pouze dvěma funkcemi pro výpočet rozptylu. Tedy s každou novou verzí nějakého programu to znamená neustálou kontrolu toho, co vlastně počítám a nové „učení se“ obsluhy programu.

## 5. Základy zpracování kvalitativních dat

Doposud jsme se zabývali pouze náhodnými veličinami, jejichž hodnoty lze „smysluplně“ vyjádřit číselně, přičemž číselné hodnoty těchto veličin mají skutečně význam čísel–hodnot, nikoliv pouze číslic, symbolů, znaků nebo pouze pořadí či uspořádání. Takovéto veličiny se většinou nazývají **kvantitativní** (číselné, numerické, někdy též kardinální – pro kladné hodnoty).

Přesněji řečeno: ***Kvantitativní se nazývají ty veličiny, u nichž rozdíl a případně i podíl (poměr) dvou změřených hodnot těchto veličin má reálný význam.*** [12, str. 137]

Ne všechny náhodné veličiny jsou kvantitativní. V sociologických průzkumech velmi často převažují **kvalitativní** veličiny, jak nejčastěji označujeme nekvantitativní veličiny. My jsme na tento pojem narazili již při povídání o **charakteristikách polohy**, kde jsme mimo jiné říkali, že:

**modus** je použitelný u **jmenných – nominálních** proměnných, které nabývají rovnocenných variant. Proto je nelze ani porovnávat, ani seřadit.

Například: pohlaví, národnost, značka hodinek, barva svetru, ...

**medián** lze použít u **pořadových – ordinálních** proměnných, u kterých lze stanovit pořadí a tím je vzájemně porovnávat (pouze na základě pořadí) nebo seřadit.

Například: známka ve škole, velikost oděvů (S, M, L, XL), medaile ve sportovních soutěžích (zlatá, stříbrná, bronzová), ...

Kvalitativní náhodné veličiny jsou ze své podstaty chápány jako diskrétní náhodné veličiny. Mnoho statistických metod vypracovaných pro kvantitativní veličiny (kvantitativní data) nelze použít pro veličiny kvalitativní (například u nominálních veličin nemá žádný smysl i zcela běžný pojem střední hodnoty). Pro analýzu nominálních a ordinálních náhodných veličin se používají buď upravené metody pro veličiny kvantitativní, nebo metody zcela speciální.

Poměrně často se vyskytující statistickou úlohou je rozhodnout, zda dvě náhodné veličiny, které nejsou kvantitativní, spolu nějak významně souvisí, zda jsou či nejsou vzájemně závislé. Přitom může jít jak

o veličiny nominální (jmenné), tak i ordinální (pořadové). Rozhodnutí o závislosti či nezávislosti dvou kvalitativních náhodných veličin je možné učinit pomocí **testu nezávislosti v kontingenční tabulce**.

**Příklad [12, str. 138]:** Máme rozhodnout, zda je chuť určitého druhu vína nějak ovlivněna **materiálem** nádoby (sudu, tanku, demižonu, ...), ve které bylo víno **skladováno**. Označíme **X** materiál nádoby s hodnotami **dřevo**, **sklo**, **kov** a **plast**. Dále označíme **Y** chuť daného druhu vína, hodnocenou znalcem na tříhodnotové škále hodnotami **pP**–podPrůměrná, **Pr**–Průměrná a **nP**–nadPrůměrná. **X** a **Y** jsou zřejmě kvalitativní náhodné veličiny, přičemž chuť **Y** je veličina ordinální (pořadová) a materiál sudu **X** je veličina pouze nominální (jmenná).

Pro posouzení závislosti těchto dvou veličin expert posuzoval chuť vína celkem v 1097 nádobách z různých materiálů. Výsledky jsou uvedeny v následující kontingenční tabulce, v níž jsou již počteny řádkové a sloupcové součty (srovnej s **levou tabulkou**).

		chuť vína stanovená expertem			$\Sigma$
		pP–podPrůměrná	Pr–Průměrná	nP–nadPrůměrná	
materiál nádoby	dřevo	100	118	51	269
	sklo	32	73	16	121
	kov	151	159	103	413
	plast	124	130	40	294
$\Sigma$		407	480	210	1 097

Z tabulky lze ihned zjistit, že například skleněných nádob s vínem nadprůměrné chuti bylo 16, plastových s průměrnou chutí 130 atd. Z řádkových a sloupcových součtů (na pravém a spodním okraji tabulky) můžeme zjistit například, že všech kovových nádob bylo 413, všech nádob s podprůměrnou chutí vína bylo 407 atd. Ovšem zda je chuť vína ovlivněna materiálem nádoby z tabulky přímo nevyčteme.

A naším úkolem je rozhodnout (= otestovat – viz **testy statistických hypotéz** v další kapitole), zda je chuť vína ovlivněna materiálem nádoby, tedy **otestovat** na hladině významnosti např.  $\alpha = 1\%$

**nulovou** hypotézu  $H_0$ : chuť vína nezávisí na materiálu nádoby, ve které bylo víno skladováno,

**proti alternativě**  $H_A$ : tyto dvě veličiny nejsou nezávislé.

Je velmi vhodné [12, str. 144] určit nejprve číselně hypotetické četnosti jednotlivých políček, tedy hodnoty  $\hat{n}_{r,s} = \frac{n_{r \cdot} \cdot n_{\cdot s}}{n} \quad \forall r, s$  (kde  $r$  je řádkový a  $s$  sloupcový index) a tyto hypotetické četnosti (v závorce) vepsat přímo do kontingenční tabulky pod příslušné četnosti skutečně napozorované.

$$\hat{n}_{d,pP} = \frac{n_d \cdot n_{pP}}{n} = \frac{269 \cdot 407}{1097} \doteq 99,80 \quad \dots$$

$$\hat{n}_{k,nP} = \frac{n_k \cdot n_{nP}}{n} = \frac{413 \cdot 210}{1097} \doteq 79,06 \quad \dots$$

		chuť vína			$n_r = \Sigma$
		pP	Pr	nP	
materiál nádoby	d	100 (99,80)	118 (117,70)	51 (51,49)	269
	s	32 (44,89)	73 (52,94)	16 (23,16)	121
	k	151 (153,23)	159 (180,71)	103 (79,06)	413
	p	124 (109,08)	130 (128,64)	40 (56,28)	294
$n_s = \Sigma$		407	480	210	1097

V našem případě jde o dvourozměrný test dobré shody, který je analogií později uváděného testu „**chi kvadrát**“. Testové kritérium potom je

$$\chi^2 = \sum_{r=1}^4 \sum_{s=1}^3 \frac{(n_{r,s} - \hat{n}_{r,s})^2}{\hat{n}_{r,s}} = \frac{(100 - 99,8)^2}{99,8} +$$

$$+ \frac{(118 - 117,7)^2}{117,7} + \frac{(51 - 51,49)^2}{51,49} + \dots + \frac{(40 - 56,28)^2}{56,28} \doteq$$

$\doteq 30,176$  . Obor přijetí hypotézy je:

$$I_\alpha = \langle 0 ; \chi_{1-\alpha}^2[(r-1) \cdot (s-1)] \rangle =$$

$$= I_{0,01} = \langle 0 ; \chi_{1-0,01}^2(3 \cdot 2) \rangle = \langle 0 ; \chi_{0,99}^2(6) \rangle =$$

$$= \langle 0 ; 16,8119 \rangle \quad \text{Protože } \chi^2 \notin I_{0,01}$$

( $30,176 \notin \langle 0 ; 16,8119 \rangle$ ) můžeme s velkou spolehlivostí [na 99 % (= 100 % –  $\alpha$ )] prohlásit, že chuť vína (podle experta) závisí (statisticky) významně na materiálu nádoby.



V uvedeném příkladu jsme mohli spolehlivě tvrdit, že chuť vína závisí na materiálu nádoby, ve které bylo víno delší dobu uskladněno. Další přirozenou otázkou by po prokázání závislosti samozřejmě mělo být, **jakým způsobem, závisí chuť vína na materiálu nádoby?** jaký materiál působí na chuť příznivě, jaký nepříznivě, případně neutrálně?

Jinými slovy, v jakých políčkách kontingenční tabulky je pozorovaná četnost výrazně menší či výrazně větší než by měla být v případě nezávislosti. Které kombinace chuti a materiálu jsou výrazně méně četné, než kdyby chuť nezávisela na materiálu? Které jsou naopak výrazně čtenější?

Ještě jinak: která políčka jsou „zodpovědná“ za zamítnutí hypotézy nezávislosti? Aniž bychom uváděli přesné statistické postupy pro rozhodování, kdy je v jednotlivých políčkách **pozorovaná** četnost **výrazně jiná** než četnost hypotetická (**předpokládaná**), naznačíme zde myšlenku takové analýzy závislosti, spolehlivě předtím testem nezávislosti prokázané (někdy se také mluví o analýze políček kontingenční tabulky).

Z porovnání skutečně pozorovaných a hypotetických četností políček kontingenční tabulky můžeme často učinit alespoň zhruba nějaké závěry o typu prokázané závislosti [12, str. 145]. Pokusme se o to pro situaci z předchozího příkladu a porovnejme napozorované (skutečné) a hypotetické (předpokládané) četnosti políček v **tabulce** (budeme postupovat po řádcích odspodu):

**Ze 4. řádku** je vidět, že bylo pozorováno výrazně více (než kdyby byla chuť nezávislá na materiálu) plastových nádob s podprůměrnou chutí vína (124 místo zhruba 109), naopak výrazně méně než při nezávislosti bylo plastových nádob s nadprůměrnou chutí (40 místo zhruba 56). Počty plastových nádob s průměrnou chutí se výrazně neliší. Z toho můžeme usoudit, že zřejmě plastový materiál zvyšuje počet vzorků s podprůměrnou chutí na úkor vzorků s chutí nadprůměrnou. Tedy že plast zřejmě zhoršuje chuť vína.

**Ze 3. řádku** je obdobně vidět, že je výrazně méně než při nezávislosti kovových nádob s průměrnou chutí vína a naopak výrazně více těchto nádob s nadprůměrnou chutí. Kovových nádob s podprůměrnou

chutí je přibližně stejně. Kov tedy zřejmě zvyšuje počet nadprůměrných vzorků na úkor průměrných  $\Rightarrow$  „zlepšuje“ průměrnou chuť.

**Ze 2. řádku** se dá usoudit, že sklo zvyšuje počet průměrných vzorků na úkor vzorků podprůměrných i nadprůměrných  $\Rightarrow$  „zprůměrnňuje“ chuť.

**Z 1. řádku** je vidět, že u dřevěných nádob se pozorované počty nádob s jednotlivými chutěmi vína téměř neliší od počtů předpokládaných, očekávaných při nezávislosti. Lze usuzovat, že dřevo neovlivňuje výrazně chuť vína.

Přesných postupů (obdobných výše jen zhruba naznačené interpretaci výsledků) v případě zamítnutí hypotézy nezávislosti je v literatuře celá řada. Některé z nich v podstatě pouze určují, co znamená pojem „výrazná odlišnost“ pozorované a předpokládané četnosti v políčku tabulky.

Nejpoužívanější je asi tak zvané **znaménkové schéma**, které doplňuje znaménko **PLUS** a **MÍNUS** do těch políček tabulky, u kterých se příslušným speciálním testem (na zadané hladině významnosti  $\alpha$ ) spolehlivě prokáže, že pozorovaná četnost políčka je větší, případně menší, než by měla být při hypotéze nezávislosti.

## Pro zájemce.

Uvědomme si, že v uvedeném příkladu chceme prokázat **nezávislost dvou veličin**  $X$  a  $Y$ , neboli odmítnout jejich závislost. Pokud se to nepodaří, budeme konstatovat, že veličiny jsou závislé jedna na druhé.

Převedeme-li naši úvahu do terminologie **pravděpodobnosti jevů**, pak jev

**X** znamená, že u náhodně vybraného vzorku vína se budeme zajímat o materiál nádoby, v jaké bylo dané víno uskladněno;

**Y** znamená, že u náhodně vybraného vzorku vína se budeme zajímat o to, jakou má chuť

a nás potom zajímá, zda jsou tyto dva **jevy vzájemně nezávislé**.

Za předpokladu, že jsou tyto jevy vzájemně nezávislé, můžeme podle vzorce (6) přímo spočítat pravděpodobnosti jednotlivých políček v tabulce (srovnej s [pravou tabulkou](#)). Ke stanovení jednotlivých pravděpodobností využijeme vzorce (1) pro statistickou definici pravděpodobnosti.

Potom například

$$P(d \cap pP) \stackrel{(6)}{=} P(d) \cdot P(pP) \stackrel{(1)}{=} \frac{n_d}{n} \cdot \frac{n_{pP}}{n} = \frac{n_d \cdot n_{pP}}{n^2}$$

a teoretická (očekávaná, hypotetická) četnost vypočtená podle upraveného vzorce (1) při platnosti předpokladu nezávislosti je:

$$\hat{n}_{d,pP} = n \cdot P(d \cap pP) = n \cdot \frac{n_d \cdot n_{pP}}{n^2} = \frac{n_d \cdot n_{pP}}{n}$$

**Podmínky použití testu.** Uvádí se [12, str. 143], že popsany test nezávislosti v kontingenční tabulce, se dá bez větších chyb použít jen v těch případech, kdy je **hypotetická četnost každého políčka alespoň 1** a alespoň pro 80 % políček je tento odhad hypotetické četnosti alespoň 5. K dosažení těchto požadavků lze často (má-li to nějaký reálný důvod) sloučit některé dvě i více sousedních hodnot veličiny **X** nebo **Y** do hodnoty jediné, případně některou málo četnou hodnotu zcela vypustit. Tím vznikne menší kontingenční tabulka s obecně většími hypotetickými četnostmi políček.

## Etapy statistické práce

Při statistické práci se většinou rozlišují čtyři kroky:

**Formulace problému.** Co chceme zjistit, koho (případně čeho) se daný problém týká.

**Šetření** (sběr dat). V předchozím příkladu jsme data obdrželi  $\Rightarrow$  šlo o sestavení druhotné statistiky.

**Zpracování** bylo podstatou předchozího příkladu — sestavení tabulky a určení číselných charakteristik. Tedy *analýza* shromážděných dat vedoucí k získání potřebné informace.

**Vyhodnocení** získané informace — bude probíráno v následujících kapitolách.

Daleko nejdůležitější částí práce se zdá být vyhodnocení — tím se zpravidla zabývají učebnice statistiky nejpodrobněji. Nesmíme však zapomenout na elementární pravdu: **žádná statistika nemůže být lepší než její surovina** [14, str. 133], tak jako nemůže být správný úsudek, jsou-li nesprávné předpoklady (z „nepravdy“ klidně může vyplývat „pravda“, jak se učí ve výrokové logice).

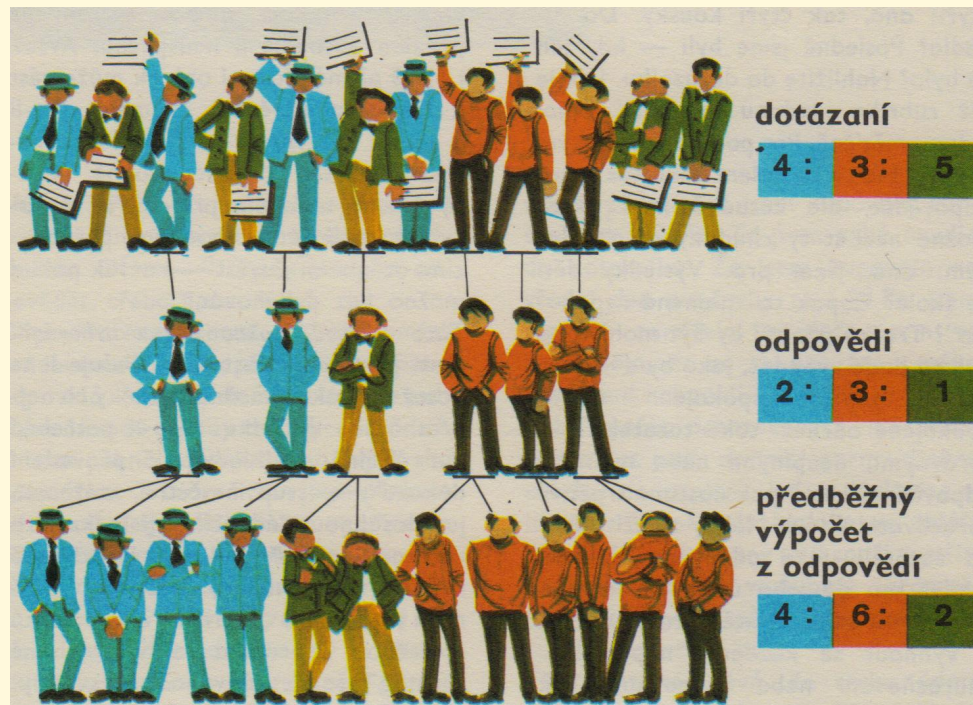
Stejně tak jsou k ničemu nejobtížnější početní operace, když číselný materiál je hned od počátku nesprávný nebo nedostačující. A co je ještě horší — početní chyby lze opravit, nevhodné metody zpracování mohou být nahrazeny lepšími. Ovšem pokud je prvotní záznam údajů chybný, většinou již s tím nejde nic dělat.

Co tedy můžeme udělat, abychom tomu v praxi předešli? První otázkou je, zda víme, co vlastně chceme. Není tomu tak vždy, protože mnohé statistické údaje úřady shromažďují v očekávání, že později mohou posloužit jako cenné podklady, aniž se v daném okamžiku dá přesně říci, co se vlastně hledá.

Když je možné použít bohatého materiálu úředních statistik a pouze je zapotřebí jej uspořádat z jiného hlediska, mluvíme o **sestavení druhotné statistiky**. Protikladem tomu je **zhotovení prvotní statistiky**, kde musíme nejprve údaje získat zjišťováním a provést jejich členění (třídění) a kdy se nepoužívá údajů, které jsou k dispozici.

Pozorováním, sčítáním a měřením mohou statistikové získat jen zlomek číselného materiálu. Stále znovu se ukazuje jako nezbytné použít anketního šetření (dotazování). Pro rozlehlé oblasti průzkumu trhu a veřejného mínění je to samozřejmé, neboť trh se skládá z více nebo méně koupěchtivých lidí, jejichž ochota kupovat se má prozkoumat. A mínění (zcela subjektivní představa) je měřitelné pouze tak, když se projeví nějakou akcí (například odevzdáním hlasu při volbách). **Ovšem pozor!** Zpracovávané

Obrázek 4: Převzat z [14]



(ne každý je odevzdá) anketní odpovědi nemusejí představovat výběr přesně odpovídající základnímu souboru, ze kterého byl vybrán. Takže můžeme obdržet zkreslené výsledky.

Cílem tohoto kurzu je naučit se statistiky číst, kriticky je posuzovat a pokusit se odhalovat statistiky chybné nebo vědomě zkreslené.

**Poznámka.** Slovu statistika bývá dáván nejrůznější význam. Jednou jsou to vyplněné statistické výkazy či dotazníky, příště tak nazveme nejrůznější číselné údaje uveřejněné ve sdělovacích prostředcích.

Oficiálně lze slovo statistika používat nejméně ve třech pojetích:

**Číselné údaje** o hromadných jevech.

**Praktická činnost** spočívající ve sběru, zpracování a vyhodnocování dat.

**Teoretická disciplína** zabývající se metodami vyhodnocení hromadných jevů.

A to je ta složitá matematika, kterou přenecháme profesionálním statistikům. My si dopřejeme toho přepychu, že můžeme výsledky jejich práce (za splnění předpokladů) s důvěrou využívat.

A proč jsme v celé této kapitole mluvili o základním souboru (populaci), výběrových souborech, empirických charakteristikách či empirických zákonech rozdělení? Výběrové šetření (nezkoumáme celou populaci, nýbrž pouze její vybranou část) samozřejmě nedosahuje přesnosti, jakou by nám přineslo zkoumání celé populace. Proč tedy dávat přednost použití výběru?

**Úspora času i finančních prostředků** — a fyzická proveditelnost vůbec, zejména u rozsáhlé populace.

**Destruktivní testy** — měření pevnosti, životnosti, ... Odpovězte si sami, k čemu by vedlo testování celé populace.

**Základní soubor nemusí být vždy dostupný** — například předvolební průzkumy.

A tady narážíme i na limity statistiky. Ty nejsou, paradoxně, v matematických metodách, nýbrž především ve sběru dat. Největším problémem bývá sestavení výběrového souboru tak, aby co nejlépe promítal vlastnosti celé populace (volby, test integrovaných obvodů na jedné desce, výběr výrobků pro přejímku — pohodlnost ...), a pak také lidský faktor (placené dotazníky, snaha upravit údaje tak, aby odpovídaly požadavkům nadřízeného).

# Úvod do **Statistické indukce**



## Obsah kapitoly: Statistická indukce

<b>1. Bodové odhady parametrů</b>	<b>175</b>
<b>2. Intervalové odhady parametrů</b>	<b>178</b>
Střední hodnota $\mu$ populace s normálním rozdělením . . . . .	180
Rozptyl populace s normálním rozdělením . . . . .	181
<b>3. Testy statistických hypotéz</b>	<b>182</b>
3.1. Postup při testování hypotéz . . . . .	191
Klasické testování . . . . .	194
3.2. Vybrané parametrické testy . . . . .	197
Test o střední hodnotě $\mu$ normálního rozdělení . . . . .	198
Test o rozptylu normálního rozdělení . . . . .	200
3.3. Vybrané testy shody . . . . .	201
Příklad: $\chi$ kvadrát – test dobré shody (Pearsonův) . . . . .	204
Příklad: Kolmogorovův–Smirnovův jednovýběrový test shody . . . . .	212
<b>4. Příklad – bodový a intervalový odhad střední hodnoty, test velikosti střední hodnoty</b>	<b>221</b>
Volba testových kritérií . . . . .	221
Použití testových kritérií . . . . .	246
<b>5. Závěr kapitoly – Čistý test významnosti</b>	<b>248</b>

## Úvod kapitoly

Základní úlohou matematické statistiky je zobecnění (zvané v tomto oboru statistická indukce či statistické usuzování): zkoumá se, jak informace zjištěné o prvcích výběru zobecnit na celou populaci<sup>29</sup>.

Za účelem, abychom získali představu o vlastnostech základního souboru (*populace*) a nemuseli vyšetřovat všechny jeho prvky<sup>30</sup>, vybereme náhodným způsobem, vzájemně nezávisle  $n$  prvků ze základního souboru. Dostaneme tak **vzorek**  $n$  prvků  $(x_1, x_2, \dots, x_n)$ , pro něhož hodnoty zkoumaného znaku zjistíme<sup>31</sup>. Neboli vypočteme **empirické charakteristiky** (statistiky). Podle výsledků výběrových zkoumání si pomáháme při rozhodování typu: tato nerovnoměrnost výroby nemůže být jen nahodilá, tento lék se zdá být významně účinnější, tuto zdánlivou shodu dvou jevů je možno vysvětlit působením náhody.

Používané metody se opírají o zákon velkých čísel a příbuzné věty teorie pravděpodobnosti (což přesahuje rámec této *příručky*); ty ukazují, že při rostoucím rozsahu reprezentativního výběru se empirické charakteristiky výběru (**bodový odhad**) obvykle limitně blíží skutečným hodnotám na celé populaci. Matematická statistika zároveň stanovuje, jak přesný tento odhad pro daná data je (**intervalový odhad**), anebo testuje, zda vlastnosti vzorku jsou slučitelné s předpoklady o chování celé populace (**testování statistických hypotéz**).

Uvědomme si, že na rozdíl od charakteristik základního souboru, které jsou konstanty, jsou empirické výběrové charakteristiky (neboli charakteristiky vypočtené z provedeného výběru) **náhodnými veličinami**, protože jejich hodnoty mohou být pro každý výběr rozdílné.

<sup>29</sup> HENDL, JAN. *Přehled statistických metod zpracování dat*. Praha : Portál, 2004, str. 18. ISBN 80-7178-820-1.

<sup>30</sup> Důvody, proč nezkoumáme celý základní soubor, jsme uváděli na **závěr** předchozí kapitoly nazvané **Popisná statistika**.

<sup>31</sup> Ovšem takových výběrů z jednoho základního souboru můžeme provést více a pokaždé dostaneme jiný vzorek. Takže empirické charakteristiky různých vzorků nemusejí být stejné. V této kapitole budeme zkoumat vztahy mezi rozdělením pravděpodobnosti konkrétního **znaku** základního souboru a rozdělením pravděpodobnosti stejného **znaku** v jednotlivých výběrech „vytahovaných“ ze základního souboru.

## Bodové odhady (vybraných) parametrů

Bodovým odhadem charakteristiky základního statistického souboru rozumíme takové **číslo**, které hodnotě toho parametru odpovídá.

Bodový odhad (odhad jedinou hodnotou) nás zdánlivě naplňuje jistotou přesně stanoveného čísla, které nám umožňuje bez problémů s tímto odhadem pracovat; například jej srovnávat s nějakým předepsaným limitem. Opak je ovšem pravdou! Protože bodový odhad se prakticky nikdy nemůže „strefit“ do odhadované hodnoty a při opakovaném určení odhadu s jiným výběrem dostaneme téměř vždy jinou hodnotu odhadu.

Zpravidla lze z výběrového souboru vypočítat několik různých (výběrových) charakteristik, pomocí nichž můžeme odhadovat neznámý parametr základního souboru (populace). Například střední hodnotu symetrického základního souboru můžeme odhadnout tak, že ze vzorku (výběru) určíme aritmetický průměr (případně jiný průměr), modus nebo medián. Tyto výběrové charakteristiky ale neposkytují stejně kvalitní odhady. **Vhodná výběrová charakteristika** (k provedení odhadu příslušného parametru základního souboru) splňuje následující kritéria (má vhodné vlastnosti).

*Je:*

**Konzistentní** — pro velký počet dat ve vzorku je málo pravděpodobné, že se odhad významně liší od zkoumané charakteristiky.

**Nestranná** (nevychýlená, nezkreslená) — vybereme-li jiný vzorek, odhady se sice budou lišit, ale jejich průměr je velmi blízký zkoumané charakteristice. Jinak řečeno: použitá charakteristika systematicky nenadhodnocuje ani nepodhodnocuje odhadovaný parametr.

**Vydatná** (eficientní) — nestranný odhad, jehož rozptyl je nejmenší mezi všemi nestrannými odhady příslušného parametru.

**Dostatečná** — obsahuje veškerou informaci o sledovaném parametru, kterou může výběrový soubor poskytnout. Znamená to, že žádný jiný parametr neobsahuje větší množství informace o výběrovém souboru.

Existuje řada metod, pomocí nichž lze získávat bodové odhady. Mezi nejznámější patří metoda nejmenších čtverců, momentová metoda nebo metoda maximální věrohodnosti. Bližší informace o teorii odhadu lze získat v příslušné literatuře.

Ukazuje se (zákon velkých čísel a spol.), že pro základní soubor s **normálním rozdělením** platí:

$\mu$ : **Aritmetický průměr  $\bar{x}_A$  vzorku** (výběru) je nejlepší (ve smyslu výše zmíněných vlastností) bodový **odhad střední hodnoty**  $\mu$  základního souboru (populace) s **normálním rozdělením**.

$\sigma$ : Podobně ze vzorku (výběru) zjištěná empirická charakteristika **výběrová směrodatná odchylka**  $S$  je nejlepší odhad směrodatné odchylky  $\sigma$  základního souboru s **normálním rozdělením**.

Analogické závěry platí i pro jiné charakteristiky, případně i pro jiná rozdělení základního souboru.

Jak jsme již uvedli, bodovým odhadem se nikdy nemůže přesně „strefit“ do správné hodnoty hledaného parametru. Můžeme jen předpokládat, že se odhadované „správné“ hodnotě více či méně přiblížil. Je tedy vhodnější pokusit se „zachytit“ odhadovanou hodnotu v určitém rozmezí (intervalu, který hledaný parametr pokrývá) kolem bodového odhadu, protože bodový odhad obvykle neposkytuje žádnou představu o přesnosti (spolehlivosti) získané aproximace. Přitom termín **spolehlivost** 90 % většinou odpovídá představě, že výpověď bodového odhadu je z 90 % správná. Přitom vědomě připouštíme 10% **chybu**.

Celé to ale můžeme říci také jinak: Výsledek (bodový odhad) je **statisticky významný na hladině** 10 %, protože by jen čistou náhodou nenastal v 90 % případů. To je pravděpodobnost jistoty, která vymezuje také interval spolehlivosti.

Proto se počítá takzvaný intervalový odhad, jehož výsledkem je interval spolehlivosti (konfidenční interval), tedy interval, v němž se s jistou předem zadanou pravděpodobností nachází hodnota hledané statistiky základní populace.

Přitom stanovení intervalu spolehlivosti (jeho „šířka“ – rozpětí) vůbec nezávisí na velikosti populace. Jedině velikost vzorku a jeho homogenita ovlivňují velikost chyby.

Představme si následující případ. Zkoumáme týdenní konzumaci piva ve dvou skupinách studentů.

	skupina I	skupina II
	počet piv	počet piv
1. student	8	0
2. student	8	0
3. student	8	0
4. student	8	0
5. student	8	40
Součet:	40	40
Aritmetický průměr:	8	8

Pro obě skupiny jsme obdrželi zcela shodný průměr, 8 piv za týden. Je zřejmé, že průměr 8 reprezentuje skupinu I perfektně. Ale skupina **II je vlastně skupina abstinentů, do které se vloudil jediný pivní hrdina**, který se snaží udržet průměrnou konzumaci piva na úrovni srovnatelné se skupinou I.

Je nesporné, že rozdíl mezi dvěma průměry signalizuje přítomnost souvislosti mezi proměnnou, podle které byli jedinci rozděleni do dvou výběrů, a proměnnou popsanou jako průměr. Problém je jenom v tom, jak zjistit, že ten rozdíl mezi dvěma průměry je dostatečně významný. Ted' již víme, že nestačí vzít v úvahu jen velikost vzorku, ale i to, jak je vzorek (a potažmo celá populace) homogenní.

My jsme již některé intervalové odhady pro soubory s normálním rozdělením zmiňovali, když jsme v pravidle **tří sigma** uváděli, že se v tomto intervalu nachází přibližně 99,7 % všech hodnot náhodné proměnné. Lze to interpretovat také tak, že se spolehlivostí 99,7 % padne střední hodnota  $\mu$  do tohoto intervalu.

Podobně pravidlo **dvou sigma** určuje interval, který přibližně s 95% spolehlivostí vymezuje střední hodnotu  $\mu$  daného souboru.

Další intervalové odhady si ukážeme nyní.

## Intervalové odhady (vybraných) parametrů

Intervalovým odhadem charakteristiky (parametru) základního statistického souboru rozumíme **interval spolehlivosti** (konfidenční interval), který tuto charakteristiku (s velkou pravděpodobností  $\Rightarrow$  **spolehlivost odhadu**) pokrývá (charakteristika v tomto intervalu leží).

95% spolehlivost znamená <sup>32</sup>, že skutečná proporce zkoumaného znaku existující v populaci (základním souboru), se nalézá s pravděpodobností 95 % uvnitř stanoveného **intervalu spolehlivosti** (konfidenčního intervalu). Kdybychom vytvořili 100 vzorků obdobné velikosti, pravděpodobně jen v pěti vzorcích by bylo možné, že skutečná proporce zkoumaného znaku by ležela POD nebo NAD (prostě vně) vypočítaným konfidenčním intervalem (intervalem spolehlivosti).

<sup>32</sup> Zda postačí 95 % jistoty či nikoliv, nelze říci všeobecně. O tom je zapotřebí rozhodnout v každém jednotlivém případě samostatně. Jestliže chceme dosáhnout vyšší spolehlivosti, je nutné zkoumat větší výběrové soubory. Jestliže se nemá vydat moc peněz a postačí-li přibližný přehled (jako například u mnohých otázek průzkumu trhu), je postačující 90 % či ještě nižší hodnota. Jde-li o zvláště velmi závažné rozhodnutí (medicína, letectví, atd.), bude snaha dosáhnout i vyšší pravděpodobnosti jak 99 %.

Je zřejmé, že čím vyšší spolehlivost odhadu požadujeme, tím širší interval spolehlivosti bude (hledaná hodnota se v něm musí nacházet s vyšší pravděpodobností). Bohužel to však ubírá na jeho vypovídací schopnosti, jeho **významnost** klesá. *Uvědomte si, jaká je vypovídací schopnost informace, že průměrný věk všech lidí na zemi leží se 100% spolehlivostí v intervalu od 0 do 195 let.* Proto v praxi vždy hledáme kompromis mezi spolehlivostí a významností (vypovídací schopností).

Označíme-li **spolehlivost odhadu**  $(1 - \alpha)$ , pak  $\alpha$  se nazývá **hladinou významnosti**. Je zřejmé, že s rostoucí spolehlivostí odhadu klesá hladina významnosti.

Intervaly spolehlivosti konstruujeme jako **jednostranné** (důležitá je pouze jedna mez; odhadujeme-li například délku života nějakého zařízení, je pro nás důležitá pouze dolní mez — pak mluvíme o levostranném intervalu spolehlivosti / v případě horní meze pak pravostranném) nebo **dvoustranné**.

Zajímají-li nás obě meze odhadu (dolní i horní), konstruujeme oboustranný interval spolehlivosti. Většinou tyto meze určujeme tak, aby pravděpodobnost, že parametr populace (základního souboru) leží pod dolní mezí, byla stejná jako pravděpodobnost, že leží nad horní mezí a byla rovna  $\frac{\alpha}{2}$ .

**Pozor! Dolní (horní) mez dvoustranného intervalu spolehlivosti není stejná jako mez u levostranného (pravostranného) intervalu spolehlivosti.**

Obecné metody konstrukce intervalů spolehlivosti jsou značně náročné. Pro naše účely se omezíme na **dvoustranné intervaly spolehlivosti pro parametry normálního rozdělení**, které jsou dobře prozkoumané (i proto se tak často setkáme s požadavkem na normalitu zpracovávaných dat). **Normalita** (předpoklad, že data pocházejí z normálního rozdělení) je hlavním předpokladem o datech v drtivé většině analýz a testů. **Ověření normality** si za chvíli ukážeme pomocí testů shody (přílehlavosti).

Zopakujeme, že 90% interval spolehlivosti odhadu střední hodnoty bude s pravděpodobností 90 % obsahovat skutečnou střední hodnotu základního souboru  $\mu$ .

**Střední hodnota  $\mu$ .** Máme náhodný výběr z populace s **normálním rozdělením**  $N(\mu; \sigma^2)$ , u kterého neznáme ani střední hodnotu  $\mu$ , ani rozptyl  $\sigma^2$ . Potom střední hodnota  $\mu$  se  $100(1 - \alpha)\%$  spolehlivostí padne do intervalu

$$\left( \bar{x} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1); \bar{x} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1) \right)$$

kde:

$\bar{x}$  je aritmetický průměr daného vzorku,

$S$  je výběrová směrodatná odchylka (druhá odmocnina výběrového rozptylu) vzorku,

$n$  je rozsah vzorku (počet získaných dat, která máme k dispozici) a

$t$  je kvantil **Studentova rozdělení**,

který najdeme ve statistických **tabulkách**,

nebo pomocí Excelu 2010: `=T.INV.2T( $\alpha$ ;  $n$ )`.

Například pro hladinu významnosti  $\alpha = 5\%$  a  $n = 16$  takto:

Tedy  $\alpha$  je číslo, které je *kladné a blízke nule*<sup>33</sup>.

$f_x$	<code>=T.INV.2T(0,05;16)</code>
	1
	<b>2,119905299</b>

Všimněte si, že při konstantním rozsahu výběru se s rostoucí spolehlivostí ( $\alpha$  se zmenšuje  $\Rightarrow$  hodnota kvantilu  $t$  roste) šířka intervalu zvětšuje. Naopak, s rostoucím rozsahem náhodného výběru  $n$  šířka intervalu klesá (dělíme větším číslem a také hodnota kvantilu  $t$  klesá), takže se odhad zpřesňuje (při

<sup>33</sup> Proč volíme parametr  $\alpha$  blízky nule? Představme si následující situaci: Před trestním senátem stojí obviněný, což ovšem může být jak zločinec, který projednávaný trestný čin skutečně spáchal, tak bezúhonný člověk, který s projednávaným trestným činem nemá naprosto nic společného.

Vyнесенý rozsudek může dopadnout čtyřmi způsoby. Dva z nich jsou správné a tudíž očekávané (1. potrestání zločince, 2. osvobození nevinného) a ve dvou soud pochybil (3. osvobození zločince, 4. potrestání nevinného). Zvláště poslední (čtvrtý) případ má z hlediska odsouzeného fatální důsledky.

Proto se v praxi snažíme co nejvíce omezit výskyt tohoto druhu chyb.

Více o **chybách** a jejich druzích uvedeme v kapitole **Testy statistických hypotéz**.



konstantní spolehlivosti). Dále pokud je rozsah výběru velký (v řádu stovek a víc), lze místo kritických hodnot Studentova rozdělení (díky centrální limitní větě) použít kritické hodnoty normálního rozdělení.

**Rozptyl  $\sigma^2$ .** Máme náhodný výběr z populace s **normálním rozdělením**  $N(\mu; \sigma^2)$ , u kterého neznáme ani střední hodnotu  $\mu$ , ani rozptyl  $\sigma^2$ . Potom rozptyl  $\sigma^2$  se  $100(1 - \alpha)\%$  spolehlivostí ( $\alpha$  kladné blízké nule) padne do intervalu

$$\left( \frac{(n-1) \cdot S^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)}; \frac{(n-1) \cdot S^2}{\chi_{\frac{\alpha}{2}}^2(n-1)} \right)$$

kde:

$S^2$  je výběrový rozptyl zkoumaného vzorku,

$n$  je rozsah vzorku (počet získaných dat, která máme k dispozici) a

$\chi^2$  je kvantil **rozdělení chí-kvadrát**, který najdeme ve statistických **tabulkách**,

nebo pomocí Excelu 2010: =CHISQ.INV.RT( $\alpha$ ;  $n$ ).

Například pro hladinu významnosti  $\alpha = 1\%$  a  $n = 5$  takto:

$f_x$	=CHISQ.INV.RT(0,01;5)
	P
	15,08627247

V případě jiných charakteristik nebo charakteristik pro jiná rozložení základního statistického souboru (populace) odkazujeme zájemce na příslušnou literaturu.

### 3. Testy statistických hypotéz

Jako ukázkou typického statistického uvažování uvedeme na úvod bez přesných statistických formulací následující

**příklad:** [12, str. 72] Máme minci, o které chceme rozhodnout, zda je či není **férová** (symetrická, homogenní, ...). Statistickou metodou to lze provést následujícím způsobem.

Hodíme ***n***krát touto mincí a zjistíme, kolikrát z těchto ***n*** hodů padne **Líc**. Z odstavce o binomickém rozdělení víme, že počet líců v ***n*** hodech mincí je náhodná veličina ***X*** s binomickým rozdělením pravděpodobnosti s parametry ***n*** a ***p***, kde ***p*** je pravděpodobnost, že v jednom hodu padne líc. Jestliže je mince **férová**, je  $p = \frac{1}{2}$ . Není-li mince férová, je  $p \neq \frac{1}{2}$ .

Řekněme, že jsme provedli pokus a mincí hodili 10 000 *krát*, přičemž líc padl 5 101 *krát*. Nyní uvažujeme následovně. Jestliže je mince férová, má náhodná veličina ***X*** (počet líců v 10 000 hodech) **binomické rozdělení** s parametry  $n = 10\,000$ ,  $p = 0,5$ .

Nás zajímá hodnota  $P(X > 5\,100)$ . Proč, to vysvětlíme v závěru příkladu.

**Jedno řešení:**  $P(X > 5\,100) = 1 - P(X \leq 5\,100) = 1 - F(5\,100)$ , kde ***F(5 100)*** je hodnota **distribuční funkce** binomického rozdělení. Ovšem distribuční funkce pro binomické rozdělení není v této příručce uvedena.

Hodnotu  $F(5\,100)$  můžeme (**za předpokladu, že je mince férová**) určit například takto:

1. Spočítáme všech 10 001 členů binomického rozvoje  $\binom{10\,000}{k} \cdot 0,5^k \cdot (1 - 0,5)^{10\,000-k}$  podle vzorce (18) pro každé ***k*** od nuly do deseti tisíc.
2. Získané hodnoty poskládáme podle velikosti a budeme hledat příslušný **kvantil**.

**Další řešení:** Nebo můžeme zkusit spočítat hledanou pravděpodobnost přímo podle vzorce (18)

$$P(X > 5\,100) = \sum_{k=5\,101}^{10\,000} P(X = k) = \sum_{k=5\,101}^{10\,000} \binom{10\,000}{k} \cdot 0,5^k \cdot (1 - 0,5)^{10\,000-k} = \dots$$

Je více než zřejmé, že oba uvedené postupy jsou velmi pracné. Proto zkusíme následující

**vylepšené první řešení** podle [12]. Protože jde v našem případě o velmi velký počet pokusů (řádově desetitisíce), můžeme podle poznámky pod **binomickým rozdělením** a s využitím vzorců (19) použít **distribuční funkci normálního rozdělení** s parametry  $N(10\,000 \cdot 0,5 ; 10\,000 \cdot 0,5 \cdot 0,5) = N(5\,000; 2\,500)$  nebo také  $N(5\,000; 50^2)$  a potom

$$P(X > 5\,100) = 1 - P(X \leq 5\,100) = 1 - F(5\,100) = 1 - F_N\left(\frac{5\,100 - 5\,000}{50}\right) = 1 - F_N(2) =$$

zde využijeme **tabelované hodnoty** distribuční funkce normovaného normálního rozdělení nebo bez provedení normování například *Excel 2010*: `=NORM.DIST(5 100;5 000;50;1)`

$$= 1 - 0,977\,25 \text{ (Excel = 0,977 249 866)} \doteq 0,023 = 2,3 \%$$

Posledním postupem nám po zaokrouhlení vyšla hodnota pravděpodobnosti  $P(X > 5\,100) = 2,3 \%$ . Jinými slovy. Pravděpodobnost, že počet líců v 10 000 hodech se liší od průměrné hodnoty 5 000 o více než o 100, je pouze 4,6 %, protože

$$P(|X - 5\,000| > 100) = P(X > 5\,100) + P(X < 4\,900) = 2,3 \% + 2,3 \% = 4,6 \%$$

Tedy za předpokladu, že mince je férová a počet líců v našich 10 000 hodech byl 5 101 (tedy počet, který se lišil o více než o 100 od očekávané hodnoty 5 000), **nastal jeden z těch výsledků našeho pokusu**,

**kteřé byly před pokusem velmi nepravděpodobné** (měly pravděpodobnost pouze 4,6 %). **Předpoklad**, že mince je férová, **tedy asi neplatí** a proto rozhodneme, že mince férová není. Ve statistice se říká, že spolehlivost tohoto zamítavého rozhodnutí je velká, v tomto případě konkrétně 95,4 % (což určíme:  $100 \% - 4,6 \%$ ).

Tento způsob uvažování je typický pro mnoho statistických metod, speciálně pro tzv. testování hypotéz. V tomto příkladu jsme stanovili hypotézu **mince je férová** a na základě výsledků pokusu (10 000 hodů mincí) jsme tuto hypotézu dostatečně spolehlivě (95,6 %) zamítli.

Mezi další významné otázky při zpracování dat patří úvahy typu:

- Splňují data charakter normálního rozdělení?
- Liší se hodnoty naměřené technikem **A** a technikem **B**?
- Liší se hodnoty získané v různých časových intervalech?
- Liší se hodnoty získané v místech **A** a **B**?
- Liší se obsah účinné látky v léčivu od deklarované hodnoty?
- Liší se výsledky získané metodami **A** a **B**?

K řešení těchto (a jim podobných) otázek využíváme metody **testování statistických hypotéz**, s jejichž pomocí lze hledat odpovědi a činit závěry. V dalším nebudeme vytvářet takové testy, ale naučíme se používat některé z existujících. Tvorba testu se pak změnila vlastně ve výběr vhodného používaného testu při řešení daného problému a aplikování vybraného testu na daný případ.

**Statistickou hypotézou** rozumíme každý (jakýkoliv) předpoklad<sup>34</sup> o neznámé vlastnosti rozložení náhodné proměnné celého základního statistického souboru. Pravdivost předpokladu můžeme ověřovat pomocí výběru pořízeného z uvažovaného základního souboru. Toto **ověřování** nazýváme **testováním hypotézy**. Většinou nás zajímá, zda (z výběru) získané empirické charakteristiky dostatečně přesně (pravdivě) popisují odpovídající charakteristiky základního souboru.

V praxi často požadujeme určit, jak má být rozsáhlý výběr (vzorek), který by zabezpečil, abychom přípustnou chybu odhadu určili s danou spolehlivostí.

**Příklad.** Kupující nechce platit za „zajíce v pytli“ a chce dojednat *přejímací kontrolu*.

**Kupující** převezme zboží jen tehdy,

*jestliže v náhodném výběru určitého rozsahu nepřekročí počet nevyhovujících kusů dohodnutý počet.*

**Prodávající** by naproti tomu měl vědět, na jaký druh přejímací (odběratelské) kontroly může přistoupit a kdy se má lákavé objednávky raději vzdát. Měl by totiž být (na základě běžné výrobní<sup>35</sup> kontroly) schopen posoudit, do jaké míry asi jeho zboží odpovídá požadavkům kupujícího.

Zcela bez problémů je ideální případ absolutně bezvadných sérií, protože není-li v základním souboru ani jediný vadný kus, nemůže se objevit ani ve vzorku. Spíše je ale nutno vycházet z realistického předpokladu, že veškeré vyráběné zboží nemůže být opravdu dokonalé. Proto se hledá způsob, jak nalézt takový zkušební postup, který by vyhovoval jak *odběrateli* tak *dodavateli* a především **nebyl příliš nákladný**. Jinými slovy: je potřeba vytvořit takovou přejímací kontrolu, která by pokud možno při malém výběrovém souboru poskytla záruku, že odběratel (pokud výrobce dodrží svůj výrobní standard) dosta-

<sup>34</sup> Hypotéza znamená doslovně předpoklad či domněnku, že něco **by mohlo být** tak a tak či **vysvětleno** tak a tak. Je to domněnka, která může vzniknout z okamžitého nápadu nebo může být vypracována po dlouhých úvahách z určité pokusné řady: „Bylo by přece docela dobře možné, že ...“ nebo: „Předpokládejme, že je možná souvislost ...“.

<sup>35</sup> Výrobní kontrola se provádí běžně a má umožnit včas poznat a odstranit výrobní závady – například seřazením určitého stroje nebo vyřazením zvláště nepozorného pracovníka z výrobního procesu.

Odběratelská kontrola se naproti tomu provádí pouze u těch výrobků, které již prošly sítí výrobní kontroly a o kterých se výrobce domnívá, že plně odpovídají jeho normám kvality.

ne s největší pravděpodobností uspokojivou jakost a prodávající se s vysokou pravděpodobností dočká přejímky bez závad.

Oba, odběratel i dodavatel, musí podstoupit jediné riziko:

**Kupujícímu** se může stát, že náhodný vzorek je podstatně lepší než skutečná jakost celé dodávky;

Proto může odebrat a zaplatit (o takovýchto chybách viz dále) dodávku zboží, která obsahuje více zmetků, než je ochoten připustit.

**Výrobci** se může přihodit, že sice vyrábí převážně dobré zboží, ale že (téměř) všechny vadné exempláře proklouznou do výběrového souboru.

Proto odběratel může odmítnout převzít dodávku zboží, která splňuje jeho předpoklady.

**Příklad** [14, str.173]

Odběratel je ochoten akceptovat 2 % zmetků, zatímco výrobce ví, že jeho výroba jich má asi 1 %. V dodávaném množství (základní soubor) 1 000 ( $N$ ) kusů je tedy asi 10 ( $M$ ) kusů vadných. Vzorek o rozsahu 100 ( $n$ ) kusů byl stanoven dohodou. Může výrobce klidně očekávat přejímací zkoušku?

**Ne tak docela!**

**Hypergeometrické rozdělení.** Protože jde o statistický výběr bez opakování (každý výrobek kontrolujeme pouze jednou, tedy ve vzorku bude skutečně 100 různých výrobků), **můžeme** (viz kapitola *Rozdělení diskrétní náhodné veličiny*) říci

$$E(X) = \frac{n \cdot M}{N} = \frac{100 \cdot 10}{1\,000} = 1$$

že jakýkoliv vzorek 100 výrobků určených k přejímací kontrole bude v průměru obsahovat pouze **jeden** vadný výrobek.

Je však docela dobře možné, že ve vzorku (výběru) budou 3 (**k**) vadné kusy, případně ještě více. Tři zmetky ve výběru se mohou vyskytnout s pravděpodobností stanovenou podle **vzorce**

$$P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}} = \frac{\binom{10}{3} \cdot \binom{1\,000-10}{100-3}}{\binom{1\,000}{100}} = \frac{\binom{10}{3} \cdot \binom{990}{97}}{\binom{1\,000}{100}} =$$

$$\frac{\frac{10!}{3! \cdot (10-3)!} \cdot \frac{990!}{97! \cdot (990-97)!}}{\frac{1\,000!}{100! \cdot (1\,000-100)!}} = \frac{\frac{10!}{3! \cdot 7!} \cdot \frac{990!}{97! \cdot 893!}}{\frac{1\,000!}{100! \cdot 900!}} \doteq 0,056\,909$$

Tedy přibližně šest ze stovky uskutečněných výběrů bude obsahovat tři zmetky a přinejmenším dva výběry dokonce budou obsahovat čtyři nebo více zmetků.

**Poznámka.** Pokud se výše uvedené faktoriály pokusíte spočítat pomocí své kalkulačky, může se vám stát, že u vyšších čísel obdržíte hlášení *Result too large* nebo *Out of range* či něco podobného. Proto je lépe využít služeb *Excelu 2010*:






=KOMBINACE(M;k)

=KOMBINACE(10;3)*KOMBINACE(990;97)/KOMBINACE(1000;100)						
E	F	G	H	I	J	K
			0,05690986			

nebo jednodušeji

=HYPGEOM.DIST(k;n;M;N;0)

HYPGEOM.DIST

Úspěch	3		= 3
Celkem	100		= 100
Základ_úspěch	10		= 10
Základ_celkem	1000		= 1000
Kumulativní	0		= NEPRAVDA

= 0,056909857

Vrátí hodnotu hypergeometrického rozdělení.

**Kumulativní** je logická hodnota: kumulativní distribuční funkce = PRAVDA, funkce hustoty pravděpodobnosti = NEPRAVDA.

Výsledek = 0,056909857

[Nápověda k této funkci](#)

OK Storno

nebo postupovat následovně:



**Binomické rozdělení.** Protože dodávka má velký rozsah  $N$  (1 000 výrobků),  $n$  (rozsah kontrolního vzorku je pevně stanoven na 100 kusů) a  $M/N$  (procento vyrobených zmetků je 10/1 000) se nemění, **můžeme** (podle poznámky pod hypergeometrickým rozdělením) toto hypergeometrické rozdělení nahradit **binomickým** s následujícími parametry:  $n = 100$ ,  $p = M/N = 0,01$ .

Tři ( $k$ ) zmetky ve výběru se mohou vyskytnout s pravděpodobností podle **vzorce**

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} = \binom{100}{3} \cdot \left(\frac{10}{1000}\right)^3 \cdot \left(1 - \frac{10}{1000}\right)^{100-3} =$$

$$= \frac{100 \cdot 99 \cdot 98}{3 \cdot 2 \cdot 1} \cdot (0,01)^3 \cdot (0,99)^{97} \doteq 0,060\,999$$

Excel 2010: =BINOM.DIST(3;100;0,01;0)

BINOM.DIST

Počet_úspěchů	3		= 3
Pokusy	100		= 100
Pravděpodobnost_úspěchu	10/1000		= 0,01
Kumulativní	0		= NEPRAVDA

= 0,060999166

Vrátí hodnotu binomického rozdělení pravděpodobnosti jednotlivých veličin.

**Kumulativní** je logická hodnota: kumulativní distribuční funkce = PRAVDA, hromadná pravděpodobnostní funkce = NEPRAVDA.

Výsledek = 0,060999166

[Nápověda k této funkci](#)

OK

Storno



### 3.1. Postup při testování hypotéz

**1. krok** Při testování hypotéz vždy klademe proti sobě dvě hypotézy (tvrzení), z nichž jedna něco tvrdí (předpokládá), druhá to popírá. V **klasické** teorii testování se vychází z toho, že platí předpokládaná vlastnost zkoumaných náhodných veličin. Tento předpoklad se označuje **nulová (testovaná) hypotéza**<sup>36</sup> a značí  $H_0$  (nebo jenom  $H$ ).

Jelikož data jsou náhodná a náhoda může *pracovat proti nám*, nelze obvykle závěry testování vyslovit s naprostou jistotou<sup>37</sup>. Proto se zároveň předem stanoví **hladina významnosti**  $\alpha$ , což určuje míru rizika (pravděpodobnost) toho, že hypotézu  $H_0$  **zamítneme, ačkoliv ve skutečnosti platí** (omyl označovaný jako **chyba prvního druhu** — viz následující [tabulka](#)).

SKUTEČNOST	ROZHODNUTÍ	
	$H_0$ přijmeme	$H_0$ zamítneme
$H_0$ platí	<b>správné rozhodnutí</b> s pravděpodobností $1 - \alpha$ <b>spolehlivost</b>	<b>chyba PRVNÍHO druhu</b> s pravděpodobností $\alpha$ <b>hladina významnosti</b>
$H_0$ NEplatí	<b>chyba DRUHÉHO druhu</b> s pravděpodobností $\beta$	<b>správné rozhodnutí</b> s pravděpodobností $1 - \beta$ <b>síla testu</b>

<sup>36</sup> Proč nulová viz [14, str. 182] — „Testujeme hypotézu: Vše zůstane při starém, nový postup (lék, ...) není ani lepší, ani horší než starý. (Zde je také etymologický základ **nulové hypotézy**, která říká, že **změna se rovná nule**.)“

<sup>37</sup> Protože při rozhodování o nulové hypotéze vycházíme z výběrového souboru, který nemusí dostatečně přesně odpovídat vlastnostem základního souboru, můžeme se při rozhodování dopustit chyby.

Chyba **prvního druhu  $\alpha$  se tradičně** v ekonomické praxi (sociologii apod.) **volí** 0,05 a v technických oblastech stanovuje 0,05 nebo 0,01. Pouze ve speciálních případech (lékařské účely, kosmonautika, ...) požadavek na pravděpodobnost chyby I. druhu dále stupňujeme (volíme ještě nižší  $\alpha$ ).

Chybu **II. druhu  $\beta$  snižujeme volbou vhodného testu** (pokud máme možnost výběru z více testů, dáváme přednost takovému testu, který má větší sílu testu ( $1 - \beta$ ) při stejné hladině významnosti ( $\alpha$ )) **popřípadě zvětšením rozsahu výběrového souboru** (což je jediný způsob jak snížit  $\beta$ , aniž bychom tím zvýšili  $\alpha$  — bohužel však je rozsah výběru téměř vždy limitován praktickými omezeními / přílišné finanční nebo časové náklady, přílišná pracnost, případně fakt, že výběr je již proveden, nemohli jsme jej ovlivnit a nelze jej opakovat). Pravděpodobnost chyby II. druhu závisí na přesné hodnotě alternativní hypotézy. Dokážeme tedy určit  $\beta$  pro případ, že alternativní hypotéza je přesně specifikovaná.

**2. krok** Dále se z dat vypočítá takzvané **testovací kritérium**, jehož rozdělení podmíněné předpokládanou platností nulové hypotézy je známo. Vyjde-li hodnota testovacího kritéria **typická** pro toto známé rozdělení, nulovou hypotézu akceptujeme či přesněji řečeno **nezamítáme** na základě známých dat. Naopak vyjde-li hodnota extrémní, tedy v oblasti hodnot, do níž realizace předpokládaného rozdělení padají s pravděpodobností menší než  $\alpha$  (tj. hodnota testovacího kritéria překročí kritickou mez), usoudíme, že testovací kritérium nejspíše nepochází z předpokládaného rozdělení a nulovou hypotézu zamítneme ve prospěch opačné tzv. alternativní hypotézy, označované  $H_A$  (nebo  $\bar{H}$ , nebo  $H_1$ ).

Vždyť co se děje při přejímce zboží?

- Odběratel vlastně testuje tuto svoji hypotézu: ***Chtějí mne doběhnout, zboží je špatné.*** A rozhodne se pro převzetí zboží pouze tehdy, je-li tato jeho hypotéza **vyvrácena**.

Ideální (v tomto případě z hlediska dodavatele) je výsledek, že nulová hypotéza se **zamítá** ve prospěch alternativní hypotézy. Statistické ověřování hypotéz není ve své podstatě ničím jiným než pokusem zamítnout nulovou hypotézu.

Tedy ***tvrzení, které chceme dokázat, volíme za alternativní hypotézu.***

**Proč nepoužíváme pojem „přijímáme nulovou hypotézu“?** Testování hypotéz můžeme provádět různými způsoby. Při každém z nich může být testovaná hypotéza zamítnuta. Nezamítneme-li ji, znamená to, že prováděným testem jsme ji nemohli zamítnout. Nikoliv to, že je správná.

Je možné, že nějakým jiným testem se ji zamítnout podaří. Pokud používáme stále přesnější testy a stále docházíme ke stejnému závěru o nezamítnutí nulové hypotézy, můžeme jednat tak, jako by nulová hypotéza byla správná. Nikdy to však nevíme jistě. *»Podobá se to dostihovému závodu s neomezeným trváním. Na každém skoku může kůň padnout, a tím by byl konec jeho závodění. Nepadne-li však, zbývá jen jedno — pokračovat v závodě.«* (prof. Dr. Ragnar Frisch, nositel Nobelovy ceny za ekonomii) Převzato z [3, str. 214]

Výsledkem testování platnosti nějakého předpokladu o vlastnosti zkoumaného znaku tedy mohou být následující dvě rozhodnutí:

- Neprokázali jsme žádný přesvědčivý důvod pro zamítnutí nulové hypotézy.
- Hodnoty sledovaného znaku ve výběrovém souboru odporují původnímu předpokladu natolik, že jej zamítáme<sup>38</sup> a přijímáme alternativní hypotézu.

Test statistické hypotézy je ověřování učiněných předpokladů o neznámé vlastnosti rozložení náhodné proměnné celého základního statistického souboru pomocí údajů získaných z náhodného výběru.

<sup>38</sup> Zamítneme-li nulovou hypotézu, tak to neznámá, že tato hypotéza neplatí (viz chyby [prvního](#) a [druhého druhu](#)). Jen dáváme najevo, že jí nedůvěřujeme na základě výsledků objektivního vyšetřování údajů, které máme k dispozici.

## Postup při klasickém testu (máme výběrový soubor):

1. Zformulujeme (testovanou) **nulovou hypotézu**  $H$  nebo  $H_0$  (představuje tvrzení, že sledovaný efekt je nulový), která se má ověřit. Bývá vyjádřena rovností mezi testovaným parametrem a jeho očekávanou hodnotou. Proti ní postavíme *alternativní hypotézu*  $\bar{H}$  nebo  $H_1$  případně  $H_A$ , která vyjadřuje tu možnost, se kterou najisto počítáme v případě, že testovaná nulová hypotéza neplatí.

Nulová hypotéza  $H$  bývá stanovena jednoznačně, například  $\mu = 55$ . Pro stanovení alternativní hypotézy bývá více možností, v našem případě tři:  $\mu < 55$ ,  $\mu > 55$  a  $\mu \neq 55$ . Obsahuje-li zadání problému vedoucího na testování hypotéz vztah jednostranné nerovnosti, volí se jako alternativní hypotéza příslušná jednostranná hypotéza. V ostatních případech volíme oboustrannou alternativní hypotézu.

Alternativní hypotéza by měla být v souladu s výběrovým souborem. Pokud tomu tak není, přizpůsobujeme alternativní hypotézu závěrům získaným z výběrového souboru.

2. Zvolíme **hladinu významnosti** (úroveň, velikost) akceptovatelné chyby prvního druhu  $\alpha$ . Potom číslo  $1 - \alpha$  určuje koeficient spolehlivosti.

Jinými slovy: Pravděpodobnost, že hodnota testové statistiky bude ležet v oblasti svědčící pro zamítnutí nulové hypotézy, přestože je nulová hypotéza platná, má být rovna předem zvolené hodnotě  $\alpha$ .

3. Zvolíme **testové kritérium** (testovou statistiku), tj. statistiku  $B = f(X_1, X_2, \dots, X_n)$ , která má vztah k nulové hypotéze a jejíž pozorovanou hodnotu (získanou ze vzorku) označíme  $b$ .

Jde o funkci výběru, která vyjadřuje sílu platnosti nulové hypotézy ve srovnání s hypotézou alternativní. Pro další krok testu musíme znát rovněž rozdělení testové statistiky při platnosti nulové hypotézy  $H$  (**nulové rozdělení**).

4. Vypočítáme pozorovanou hodnotu ***b*** testové statistiky ***B*** z výběrového souboru. Při tomto výpočtu předpokládáme platnost nulové hypotézy.
5. Určíme **kritický obor** (obor přijetí hypotézy)  $W_\alpha$  hodnot statistiky ***B***, do níž hodnoty ***B*** za platnosti hypotézy  $H_0$  padnou s pravděpodobností  $\alpha$ , tj.  $P(B \in W_\alpha | H_0) = \alpha$ .

Jde o rozdělení prostoru všech možných hodnot testové statistiky ***S*** na dva podprostory: **obor přijetí *A*** obsahující hodnoty testové statistiky svědčící pro nezamítnutí nulové hypotézy a **kritický obor *C*** obsahující hodnoty testové statistiky svědčící pro zamítnutí nulové hypotézy. Je zřejmé, že:  $A \cup C = S$ ;  $A \cap C = \emptyset$ . Hranice mezi kritickým oborem a oborem přijetí se nazývá **kritická hodnota testu**.

Známe-li nulové rozdělení testové statistiky ***B***, není obtížné pro dané  $\alpha$  stanovit kritický obor: Je-li **alternativní hypotéza** ve tvaru

- < (ve prospěch alternativy svědčí extrémně nízké hodnoty testové statistiky),  
pak je kritický obor vymezen jako:  $C \leq W_\alpha$
- > (ve prospěch alternativy svědčí extrémně vysoké hodnoty testové statistiky),  
pak je kritický obor vymezen jako:  $W_{1-\alpha} \leq C$
- $\neq$  (ve prospěch alternativy svědčí nízké nebo vysoké hodnoty testové statistiky),  
pak je kritický obor vymezen jako:  $\left(C \leq W_{\frac{\alpha}{2}}\right) \vee \left(W_{1-\frac{\alpha}{2}} \leq C\right)$

6. Formulujeme závěr:

- a) Leží-li testová statistika ***b*** v **kritickém oboru *C*** ( $b \in C$ ), pak **zamítáme nulovou hypotézu ve prospěch alternativní hypotézy**;
- b) Leží-li testová statistika ***b*** v oboru **přijetí** (neleží v kritickém oboru  $\Rightarrow b \notin C$ ), pak **nulovou hypotézu NEzamítáme**.

Jestliže výsledek testování umožňuje závěr, že testová statistika je například mimo 95% konfidenční interval (koeficient spolehlivosti) testované nulové hypotézy, mohu si být na „95 % jist“, že hypotéza není správná. Tím se pět dostáváme k našemu dřívějšímu zjištění, že **hypotéza nemůže být přímo dokázána, nýbrž může být jen zamítnuta jí odporující (nulová) hypotéza**.

V praxi ověřování hypotéz jde tedy většinou o používání takové sestavy testu (především volby výběrového souboru), aby zamýšlené úrovně odmítnutí bylo pokud možno přesně dosaženo. Jinak vznikají zbytečné náklady.

Při testování statistických hypotéz se můžeme dopustit několika chyb:

1. Volba nevhodné dvojice hypotéz (nulová hypotéza *versus* alternativní). K této chybě dochází, pokud si důkladně nerozmyslíme, co vlastně chceme testovat. Důležitý je především výběr vhodné alternativy (jednostranná, dvoustranná).
2. Chybně určená testová statistika.
3. Chybně určený obor přijetí nebo kritický obor.
4. Chyby při rozhodování (již dříve diskutované **chyby** prvního a druhého druhu).

První tři uvedené chyby lze eliminovat dobrou přípravou testu. Jde tedy o chyby, které lze ovlivnit, případně jim zcela zabránit. Jinými slovy: „*I při testování hypotéz platí pravidlo **dvakrát měř a jednou řeš**.*“ [3, str. 212]

I sebelépe připravený test však nemusí vést ke správným rozhodnutím, neboť využívá pouze omezené informace náhodného výběru. Může se stát, že náhodný výběr nebude dostatečně kopírovat vlastnosti základního souboru a při rozhodování bude zvolena opačná hypotéza, než odpovídá skutečnosti.

A jsme opět u již známých **chyb prvního a druhého druhu**.



Pamatujte si, že <sup>39</sup>:

**Hladina významnosti** (chyba I. druhu, statistická významnost) *je pravděpodobnost, s jakou bychom — za předpokladu pravdivosti nulové hypotézy — mohli obdržet data odporující nulové hypotéze stejně či ještě více než pozorovaná data.* (str. 80)

**Síla testu** *je pravděpodobnost (hodnota pohybující se mezi 0 a 1) správného přijetí alternativní hypotézy za předpokladu, že je tato v základním souboru platná.* (str. 90)

Častou statistickou úlohou je rozhodnout, zda neznámý **parametr** rozdělení populace (nejčastěji střední hodnota, rozptyl nebo relativní četnost) **je roven** nějaké konkrétní číselné **hodnotě**, případně zda je neznámý parametr rozdělení populace větší či menší než nějaká konkrétní číselná hodnota. Rozhodovací proces, který je pro řešení těchto úloh používán, bývá označován jako **jednovýběrový test**.

Jak lze z celého předchozího povídání usoudit, **střední hodnota** je základní charakteristikou každého statistického znaku. Není proto divu, že většina výběrových šetření se zabývá právě zkoumáním této veličiny. Odhady a testy průměrných příjmů, průměrných výkonů, průměrné životnosti výrobku, střední hmotnosti výrobku, atd. jsou nejběžnějšími úlohami statistiky.

## Nejpoužívanější parametrické testy

Parametrickými testy prověřujeme hypotézy o parametrech základního souboru a oceňujeme rozdíly mezi teoretickými (které má základní soubor) a empirickými (vypočtenými ze vzorku) charakteristikami. K jejich odvození je nutné pro daný výběr specifikovat typ rozdělení a v některých případech i některé parametry tohoto rozdělení.

<sup>39</sup> SOUKUP, Petr. Nesprávná užívání statistické významnosti a jejich možná řešení. In: *Data a výzkum — SDA Info* [[online](#)]. 2010, roč. 4, čís. 2 [cit. 25. 6. 2013], str. 80 a str. 90. ISSN 1802–8152.

## Test o střední hodnotě $\mu$ normálního rozdělení

Předpokládejme, že máme **normálně rozdělenou** populaci (základní soubor) s **neznámou** střední hodnotou  $\mu$  a **neznámým** rozptylem  $\sigma^2$ . Na základě výběru  $X_1, X_2, \dots, X_n$  z dané populace chceme ověřit předpoklad, jestli se střední hodnota populace  $\mu$  rovná hodnotě  $\mu_0$ .

Neznámou střední hodnotu  $\mu$  odhadneme výběrovým aritmetickým průměrem  $\bar{x}_A$ , který určíme z pozorovaných výběrových hodnot  $x_1, x_2, \dots, x_n$ . Je zřejmé, že vypočtená ( $\bar{x}_A$ ) a předpokládaná střední hodnota ( $\mu_0$ ) se mohou od sebe lišit. Rozdíl může být pouze nevýznamný a lze ho přičíst účinku náhodných vlivů, působících při výběru. Tento rozdíl však může být i nenáhodný (říkáme také statisticky významný nebo signifikantní). Test o střední hodnotě tak představuje ověření, zda se výběrový aritmetický průměr  $\bar{x}_A$  a předpokládaná střední hodnota  $\mu_0$  liší statisticky významně nebo pouze náhodně.

Nulovou hypotézu  $H_0$  volíme ve tvaru  $\mu = \mu_0$ . Zatímco volba nulové hypotézy je zřejmá, u alternativní hypotézy  $H_A$  můžeme volit ze tří možností:  $\mu < \mu_0$ ,  $\mu > \mu_0$ ,  $\mu \neq \mu_0$ .

Tedy, když to včetně testového kritéria a oboru přijetí hypotézy shrneme:

### Parametrický test o střední hodnotě normálního rozdělení

Předpoklad:  $\{X_1, X_2, \dots, X_n\}$  je náhodný výběr z  $N(\mu; \sigma^2)$

Hypotéza  $H_0$ :  $\mu = \mu_0$ , kde  $\mu_0$  je dané číslo

Hypotéza  $H_A$ :  $\mu < \mu_0$   $\mu > \mu_0$

$\mu \neq \mu_0$

Testové kritérium:  $T = \frac{(\bar{x} - \mu_0)}{S} \cdot \sqrt{n}$

Obor přijetí hypotézy:  $I_\alpha = \langle -t_{1-\alpha}(n-1); \infty \rangle$   $I_\alpha = (-\infty; t_{1-\alpha}(n-1))$

$$I_\alpha = \left\langle -t_{1-\frac{\alpha}{2}}(n-1); t_{1-\frac{\alpha}{2}}(n-1) \right\rangle$$

kde  $T$  má **Studentovo rozdělení** s  $n - 1$  stupni volnosti a  $t$  je kvantil *Studentova rozdělení*, který najdeme ve statistických **tabulkách**, nebo **pro oboustrannou** (červenou) **alternativu** pomocí *Excelu 2010*  $=T.INV.2T(\alpha; n - 1)$  Například na vedlejším obrázku je hodnota kvantilu  $t$  pro  $\alpha = 5\%$  a  $(n - 1) = 16$ .

$f_x$	$=T.INV.2T(0,05;16)$
	1
	<b>2,119905299</b>

Předpoklad, že výběr pochází z normálního rozdělení  $N(\mu; \sigma^2)$ , nemusí být za každou cenu dodržen. Test totiž pracuje s průměrem výběru, a tento výběr již při rozsahu v řádu desítek má přibližně normální rozdělení díky centrální limitní větě. Proto pokud je rozsah výběru velký (v řádu stovek a víc), lze místo kritických hodnot Studentova rozdělení použít kritické hodnoty normálního rozdělení.

**Příklad:** Podle údajů na obalu<sup>40</sup> čokolády by její čistá hmotnost měla být **125 g**. Výrobce dostal několik stížností, že hmotnost prodaných čokolád byla nižší. Z tohoto důvodu oddělení kontroly náhodně vybralo 50 čokolád určených k expedici a zjistilo, že jejich průměrná hmotnost je **122 g** a směrodatná odchylka činí **8,6 g**. Za předpokladu, že hmotnost čokolád se řídí **normálním rozložením**, můžeme na hladině významnosti 0,01 považovat stížnosti spotřebitelů za oprávněné?

**Řešení:** Použijeme parametrický **test o střední hodnotě normálního rozdělení**, kdy testujeme **nulovou hypotézu**  $H_0 : \mu = 125$  proti **levostranné** alternativě  $H_A : \mu < 125$  s (**černě** stanoveným) **oborem přijetí hypotézy**  $I_\alpha = \langle -t_{1-0,01}(50 - 1); \infty \rangle = \langle -t_{0,99}(49); \infty \rangle \doteq \langle -2,405; \infty \rangle$  kde  $t_{0,99}(49)$  určíme pro levostrannou alternativu pomocí *Excelu 2010* takto:  $=T.INV(0,99;49)$

$$\text{Testové kritérium: } T = \frac{(\bar{x} - \mu_0)}{S} \cdot \sqrt{n} = \frac{122 - 125}{8,6} \cdot \sqrt{50} \doteq -2,467 \Rightarrow T \notin I_\alpha$$

**Závěr:** Protože hodnota testového kritéria nespadá do oboru přijetí hypotézy na hladině významnosti 1 %, můžeme usoudit, že stížnosti spotřebitelů jsou oprávněné (dostávají méně čokolády).

<sup>40</sup> Řezáč, M., Budíková, M. *Statistika II*. Brno : Masarykova univerzita 2013, str. 142

## Test o rozptylu $\sigma^2$ normálního rozdělení

Předpokládejme, že máme **normálně rozdělenou** populaci (základní soubor) s **neznámou** střední hodnotou  $\mu$  a **neznámým** rozptylem  $\sigma^2$ . Na základě výběru  $X_1, X_2, \dots, X_n$  z dané populace chceme ověřit předpoklad, jestli se rozptyl populace  $\sigma^2$  rovná hodnotě  $\sigma_0^2$ .

Neznámý rozptyl  $\sigma^2$  odhadneme výběrovým rozptylem  $S^2$ , který určíme z pozorovaných výběrových hodnot  $x_1, x_2, \dots, x_n$ . Je zřejmé, že se vypočtený výběrový rozptyl ( $S^2$ ) a předpokládaná hodnota rozptylu ( $\sigma_0^2$ ) mohou od sebe lišit. A to statisticky významně nebo pouze náhodně.

### Parametrický test o rozptylu normálního rozdělení

Předpoklad:  $\{X_1, X_2, \dots, X_n\}$  je náhodný výběr z  $N(\mu; \sigma^2)$

Hypotéza  $H_0$ :  $\sigma^2 = \sigma_0^2$ , kde  $\sigma_0^2$  je dané číslo

Hypotéza  $H_A$ :  $\sigma^2 \neq \sigma_0^2$

Testové kritérium:  $T = \frac{S^2}{\sigma_0^2} \cdot (n - 1)$

Obor přijetí hypotézy:  $I_\alpha = \left\langle \frac{(n - 1) \cdot S^2}{\chi_{1-\frac{\alpha}{2}}^2(n - 1)}; \frac{(n - 1) \cdot S^2}{\chi_{\frac{\alpha}{2}}^2(n - 1)} \right\rangle$

kde  $T$  má **rozdělení „CHI kvadrát“** s  $n - 1$  stupni volnosti a  $\chi^2$  je kvantil rozdělení chí kvadrát (někdy též *Pearsonovo rozdělení*), který najdeme ve statistických **tabulkách**, nebo pomocí *Excelu 2010*:  $=\text{CHISQ.INV.RT}(\alpha; n - 1)$  Například na vedlejším obrázku je hodnota kvantilu  $\chi^2$  pro  $\alpha = 1\%$  a  $(n - 1) = 5$ .

Pro tvary dalších testových kritérií a způsoby určení intervalu spolehlivosti odkazujeme zájemce na příslušnou literaturu.

$f_x$	<code>=CHISQ.INV.RT(0,01;5)</code>
	P
	15,08627247

## Nejpoužívanější testy shody (přiléhavosti):

Domněnka o tom, že studovaná data (výběr) pocházejí z určitého teoretického (očekávaného) rozdělení, bývá podložena buď informacemi o sledovaném jevu, nebo odhadem teoretického rozdělení na základě grafického zobrazení výběrového rozdělení. Náš odhad však nemusí být správný. Proto jej v praxi ověřujeme testy shody, zda se **shoduje** teoretické (očekávané, předpokládané) a empirické (pozorované, výběrové) rozdělení. Nulovou ( $H$  nebo  $H_0$ ) a alternativní ( $\bar{H}$  nebo  $H_1$ ,  $H_A$ ) hypotézu můžeme v tomto případě formulovat:

$H_0$  — teoretické a empirické rozdělení se **shoduje**.

$H_A$  — teoretické a empirické rozdělení se **Neshoduje**.

## $\chi^2$ („chí kvadrát“ – Pearsonův) jednovýběrový test dobré shody – absolutní četnosti

Nejznámější z testů dobré shody ověřuje, zda se empirické (pozorované) absolutní četnosti  $O_i$  (anglicky „observed“) jednotlivých variant náhodné veličiny shodují s očekávanými absolutními četnostmi  $E_i$  (angl. „expected“). Tedy s četnostmi, které bychom očekávali v případě platnosti nulové hypotézy.

Hypotéza  $H_0$ : testovaný **výběr** pochází z **teoretického rozdělení** (značíme **stříškou**)

Hypotéza  $H_A$ : náhodný výběr  **$n$**  prvků pochází z **jiného** rozdělení

Testové kritérium:  $\chi^2 = \sum_{i=1}^k \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \sum_{i=1}^k \frac{n_i^2}{\hat{n}_i} - n$  což je součet čtverců rozdílů skutečných a očekávaných četností vážených očekávanými (teoretickými) četnostmi

Obor přijetí  
hypotézy:  $I_\alpha = \langle 0; \chi_{1-\alpha}^2(k-1-L) \rangle$

- kde  $k$  počet **tříd**, na které byl rozdělen interval pozorovaných hodnot náhodné proměnné;  
 $n_i$  **pozorovaná** (zjištěná na základě pokusu) třídní četnost intervalu  $\langle a_i; b_i \rangle$ ,  
 $\hat{n}_i$  **teoretická** (to co očekáváme, platí-li  $H_0$ ) četnost intervalu  $\langle a_i; b_i \rangle$ :  $\hat{n}_i = n \cdot P(a_i \leq X \leq b_i)$ .  
 Není-li splněna podmínka na velikost výběru:  $\hat{n}_i > 5$   
 — buď výběr rozšíříme tak, aby podmínka byla splněna  
 — nebo třídy s malou četností sdružujeme (týká se to zpravidla krajních tříd);  
 $L$  počet **stupňů volnosti**, tj. neznámých parametrů (modus, rozptyl, ...) teoretického rozdělení, které je nutno (z hodnot výběru) počítat. Pro  $N(\mu, \sigma^2)$  je  $L = 2, \dots$

## Kolmogorovův–Smirnovův **jednovýběrový test** – kumulativní četnosti

Používáme jej při hodnocení rozdílů mezi **kumulativními četnostmi**.

Toto je jedna z variant testů autorů **Andreje Nikolajeviče Kolmogorova** a Vladimira Ivanoviče Smirnova, která ověřuje, zda se rozdělení náhodné veličiny v populaci liší od určitého teoretického rozdělení.

Nulová hypotéza: testovaný **výběr** pochází z **teoretického rozdělení** (značíme **stříškou**)

Alternativní hypotéza: náhodný výběr  $n$  prvků pochází z **jiného** rozdělení

Testové kritérium:  $D = \frac{1}{n} \cdot \max_i |N_i - \hat{N}_i|$  kde  $N_i$  a  $\hat{N}_i$  jsou kumulativní četnosti

Obor přijetí hypotézy:  $I_\alpha = \langle 0; D_\alpha(n) \rangle$  kde,  $D_\alpha(n)$  je **tabelována** a pro  $n > 40$  pak **platí**:

$$D_{20\%}(n) \doteq \frac{1,07}{\sqrt{n}} \quad D_{10\%}(n) \doteq \frac{1,22}{\sqrt{n}} \quad D_5\%(n) \doteq \frac{1,36}{\sqrt{n}} \quad D_2\%(n) \doteq \frac{1,52}{\sqrt{n}} \quad D_1\%(n) \doteq \frac{1,63}{\sqrt{n}}$$

Vstupem této varianty testu je  $k$  tříd testovaného výběru a předpokládané (například normální) teoretické rozdělení, které se rozdělí do stejného počtu tříd.

Pro každou třídu  $i$  ( $i = 1, \dots, k$ ) testovaného výběru se spočítají **četnosti**  $n_i$  zjištěné ve výběru a pro každou třídu teoretického rozdělení se spočítají předpokládané četnosti  $\hat{n}_i$ .

Dále spočítáme

**kumulativní četnosti** pro výběr  $N_i = \sum_{j=1}^i n_j$  a pro testované rozdělení  $\hat{N}_i = \sum_{j=1}^i \hat{n}_j$ .

(<http://mi21.vsb.cz/flash-animace/kolmogorovuv-smirnovuv-test-reseny-priklad>)

Pokud máme k dispozici pouze výběr malého rozsahu, dáváme při ověřování dobré shody mezi empirickým a teoretickým rozdělením přednost tomuto testu před předchozím testem.

Výhody Kolmogorovova – Smirnovova testu oproti Pearsonovu testu dobré shody [8, str. 348]:

- větší síla testu ( $1 - \beta$ ) ;
- nemá omezující podmínky;
- pokud navíc použijeme jinou variantu testu (než jsme si uvedli), která pracuje přímo s distribučními funkcemi výběru a předpokládaného rozdělení (namísto jejich kumulativních četností), tedy vychází z jednotlivých pozorování a nikoliv z údajů seřazených do skupin, lze ji použít i na výběry skutečně malého rozsahu a nedochází ke ztrátě informace obsažené ve výběru.

## Příklad: $\chi^2$ – Test dobré shody

Při opakovaném házení kostkou (60 hodů) padla jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×. Ptáme se, zda je kostka regulérní (férová) či zda je falešná (upravená, cinknutá), a to na hladině významnosti 0,01 (= 1 %).

**Řešení:** Hrací kostka je „v pořádku“, když je pravděpodobnost padnutí každého čísla na kostce stejná. Nebo jinak: každé ze šesti čísel bude mít shodné zastoupení při větším počtu pokusů.

Při 60 pokusech  $\Rightarrow 60 : 6 = 10$ .

Budeme tedy testovat, zda rozdělení „počtu padlých ok“ je takové, že má stejné pravděpodobnosti pro všechny možné varianty. Jestliže lze základní soubor (ze kterého pochází výběr, který máme k dispozici) roztřídit podle nějakého znaku do  $k$  disjunktních skupin a my chceme na základě náhodného výběru ověřit, zda jsou relativní četnosti jednotlivých variant rovny číslům  $\pi_1, \dots, \pi_k$ , můžeme použít  $\chi^2$  – test dobré shody (Pearsonův).

### Volba nulové a alternativní hypotézy:

$H_0$ : **Kostka je v pořádku**, když výběr pochází ze základního souboru, kde jsou pravděpodobnosti jednotlivých variant rovny  $\frac{1}{6}$ .

$H_A$ : Kostka není v pořádku (**je „falešná“**), když platí cokoliv jiného.

**Testové kritérium:** Jako testové kritérium používáme náhodnou veličinu

$$T(X) = \chi^2 = \sum_{i=1}^k \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

která má za předpokladu, že provádíme **dostatečně velký výběr** (každá třída má aspoň **pět** prvků), přibližně  $\chi^2$  rozdělení s  $k - 1$  stupni volnosti. My očekáváme v každé třídě deset prvků.



60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .







První sloupec označuje číslo řádku – index  $i$ . Do druhého sloupce tabulky zapíšeme číslo, které padlo

index řádek $i$	třída				
1	•				
2	••				
3	•••				
4	••••				
5	•••••				
6	••••••				
$\Sigma$					

$$T(X) = \sum_{i=1}^6 \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .







První sloupec označuje číslo řádku – index  $i$ . Do druhého sloupce tabulky zapíšeme číslo, které padlo, do třetího  $n_i$  kolikrát padlo (pozorovaná četnost)

index řádek $i$	třída	$n_i$			
1					
2					
3					
4					
5					
6					
$\Sigma$					

$$T(X) = \sum_{i=1}^6 \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .







První sloupec označuje číslo řádku – index  $i$ . Do druhého sloupce tabulky zapíšeme číslo, které padlo, do třetího  $n_i$  kolikrát padlo (pozorovaná četnost) a do čtvrtého  $\hat{n}_i$  teoretickou (očekávanou) četnost.

index řádek $i$	třída	$n_i$	$\hat{n}_i$		
1		7			
2		9			
3		10			
4		6			
5		15			
6		13			
				$\Sigma$	

$$T(X) = \sum_{i=1}^6 \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .







První sloupec označuje číslo řádku – index  $i$ . Do druhého sloupce tabulky zapíšeme číslo, které padlo, do třetího  $n_i$  kolikrát padlo (pozorovaná četnost) a do čtvrtého  $\hat{n}_i$  teoretickou (očekávanou) četnost.

index řádek $i$	třída	$n_i$	$\hat{n}_i$		
1		7	10		
2		9	10		
3		10	10		
4		6	10		
5		15	10		
6		13	10		
$\Sigma$					

$$T(X) = \sum_{i=1}^6 \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

První sloupec označuje číslo řádku – index  $i$ . Do druhého sloupce tabulky zapíšeme číslo, které padlo, do třetího  $n_i$  kolikrát padlo (pozorovaná četnost) a do čtvrtého  $\hat{n}_i$  teoretickou (očekávanou) četnost.

index řádek $i$	třída	$n_i$	$\hat{n}_i$	$n_i - \hat{n}_i$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
1		7	10	-3	
2		9	10	-1	
3		10	10	0	
4		6	10	-4	
5		15	10	5	
6		13	10	3	
$\Sigma$					

$$T(X) = \sum_{i=1}^6 \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

První sloupec označuje číslo řádku – index  $i$ . Do druhého sloupce tabulky zapíšeme číslo, které padlo, do třetího  $n_i$  kolikrát padlo (pozorovaná četnost) a do čtvrtého  $\hat{n}_i$  teoretickou (očekávanou) četnost.

index řádek $i$	třída	$n_i$	$\hat{n}_i$	$n_i - \hat{n}_i$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
1	•	7	10	−3	0,9
2	••	9	10	−1	0,1
3	•••	10	10	0	0
4	••••	6	10	−4	1,6
5	•••••	15	10	5	2,5
6	••••••	13	10	3	0,9
$\Sigma$					6

$$T(X) = \sum_{i=1}^6 \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = 6$$

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

První sloupec označuje číslo řádku – index  $i$ . Do druhého sloupce tabulky zapíšeme číslo, které padlo, do třetího  $n_i$  kolikrát padlo (pozorovaná četnost) a do čtvrtého  $\hat{n}_i$  teoretickou (očekávanou) četnost.

index řádek $i$	třída	$n_i$	$\hat{n}_i$	$n_i - \hat{n}_i$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
1	•	7	10	-3	0,9
2	••	9	10	-1	0,1
3	•••	10	10	0	0
4	••••	6	10	-4	1,6
5	•••••	15	10	5	2,5
6	••••••	13	10	3	0,9
$\Sigma$					6

$$T(X) = \sum_{i=1}^6 \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = 6$$

Protože žádný parametr  
(průměr, modus, rozptyl, ...)   
nepočítáme, za  $L$  dosazujeme **nulu**.

Je zřejmé, že čím větší jsou odchylky pozorovaných a očekávaných četností, tím vyšší pozorovanou hodnotu testové statistiky  $T(X)$  dostáváme a tím silnější je výpověď vůči nulové hypotéze.

**Obor pro přijetí nulové hypotézy** je:  $I_\alpha = \langle 0; \chi_{1-\alpha}^2(k-1-L) \rangle$

$$\chi_{1-0,01}^2(6-1-0) = \chi_{0,99}^2(5) \doteq 15,086 \Rightarrow I_{0,01} = \langle 0; 15,086 \rangle$$

Pro stanovení hodnoty  $\chi_{0,99}^2(5)$  využijeme *Excel 2010*: =CHISQ.INV.RT( $\alpha$ ;  $n$ )

$f_x$	=CHISQ.INV.RT(0,01;5)
	P
	15,08627247

**Závěr:** Protože v našem případě  $T(X) \in I_{0,01}$ , **nezamítáme** nulovou hypotézu.

**Nelze tedy tvrdit, že kostka je falešná.**

A teď si na stejném příkladu zkusme otestovat předpoklad o „*NEfalešnosti*“ kostky druhým z testů.

### Příklad: Kolmogorovův–Smirnovův jednovýběrový test shody

Při opakovaném házení kostkou (60 hodů) padla jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×. Ptáme se zda je kostka regulérní (férová) či zda je falešná (upravená, cinknutá), a to na hladině významnosti 0,01 (= 1 %).

**Řešení:** *Odsud:* / Hrací kostka je „v pořádku“, když je pravděpodobnost padnutí každého čísla na kostce stejná. Nebo jinak: každé ze šesti čísel bude mít shodné zastoupení při větším počtu pokusů. Při 60 pokusech  $\Rightarrow 60 : 6 = 10$ .

Budeme tedy testovat, zda rozdělení „počtu padlých ok“ je takové, že má stejné pravděpodobnosti pro všechny možné varianty. Jestliže lze základní soubor (ze kterého pochází výběr, který máme k dispozici) roztrdit podle nějakého znaku do  $k$  disjunktních skupin ... , můžeme použít Kolmogorovův–Smirnovův test. / **až sem** (kromě názvu použitého testu) je to naprosto shodné s předchozím testem, a to včetně volby hypotéz. Lišit se bude až testové kritérium.

**Volba nulové a alternativní hypotézy:**

$H_0$ : **Kostka je v pořádku**, když výběr pochází ze základního souboru, kde jsou pravděpodobnosti jednotlivých variant rovny  $\frac{1}{6}$ .

$H_A$ : Kostka není v pořádku (**je „falešná“**), když platí cokoliv jiného.







**Testové kritérium:** Jako testové kritérium používáme náhodnou veličinu

$$D(X) = \frac{1}{n} \cdot \max_{\forall i} |N_i - \hat{N}_i|$$

Pozorované i předpokládané četnosti (včetně jejich kumulativních četností) zase zapíšeme do tabulky.









60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

<i>i</i>	tř. <i>i</i>					
1						
2						
3						
4						
5						
6						

**První** sloupec je číslo řádku, neboli index.







Do **druhého** sloupce tabulky zapíšeme číslo, které padlo. Zároveň to bude představovat třídu *i*.

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

<i>i</i>	tř. <i>i</i>	<i>n<sub>i</sub></i>				
1						
2						
3						
4						
5						
6						

**První** sloupec je číslo řádku, neboli index.  
 Do **druhého** sloupce tabulky zapíšeme číslo, které padlo. Zároveň to bude představovat třídu *i*.  
 Do **třetího** sloupce označeného *n<sub>i</sub>* kolikrát padlo (**pozorovaná četnost**) toto číslo.

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

$i$	tř. $i$	$n_i$	$\hat{n}_i$			
1		7				
2		9				
3		10				
4		6				
5		15				
6		13				

**První** sloupec je číslo řádku, neboli index.

Do **druhého** sloupce tabulky zapíšeme číslo, které padlo. Zároveň to bude představovat třídu  $i$ .

Do **třetího** sloupce označeného  $n_i$  kolikrát padlo (**pozorovaná četnost**) toto číslo.







Do **čtvrtého** sloupce označeného  $\hat{n}_i$  teoretickou (tu, kterou očekáváme) četnost.

V **pátém** sloupci označeném  $N_i$  jsou **kumulativní** pozorované četnosti. Tedy například ve druhém řádku je četnost výsledků, že padlo číslo menší nebo rovno **2**. Jinak řečeno, kolikrát padla **jednička** nebo **dvojka**.

A v **šestém** sloupci jsou **kumulativní** předpokládané četnosti.

Do **sedmého** sloupce zapíšeme hodnoty testového kritéria pro každou třídu.

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

$i$	tř. $i$	$n_i$	$\hat{n}_i$	$N_i = \sum_{j=1}^i n_j$		
1		7	10			
2		9	10			
3		10	10			
4		6	10			
5		15	10			
6		13	10			

**První** sloupec je číslo řádku, neboli index.

Do **druhého** sloupce tabulky zapíšeme číslo, které padlo. Zároveň to bude představovat třídu  $i$ .

Do **třetího** sloupce označeného  $n_i$  kolikrát padlo (**pozorovaná četnost**) toto číslo.







Do **čtvrtého** sloupce označeného  $\hat{n}_i$  teoretickou (tu, kterou očekáváme) četnost.

V **pátém** sloupci označeném  $N_i$  jsou **kumulativní** pozorované četnosti. Tedy například ve druhém řádku je četnost výsledků, že padlo číslo menší nebo rovno **2**. Jinak řečeno, kolikrát padla **jednička** nebo **dvojka**.

A v **šestém** sloupci jsou **kumulativní** předpokládané četnosti.

Do **sedmého** sloupce zapíšeme hodnoty testového kritéria pro každou třídu.

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

$i$	tř. $i$	$n_i$	$\hat{n}_i$	$N_i = \sum_{j=1}^i n_j$	$\hat{N}_i = \sum_{j=1}^i \hat{n}_j$	
1		7	10	7 = 7		
2		9	10	16 = 7+9		
3		10	10	26 = 7+9+10		
4		6	10	32 = 7+9+10+6		
5		15	10	47 = 7+9+10+6+15		
6		13	10	60 = 7+9+10+6+15+13		

**První** sloupec je číslo řádku, neboli index.

Do **druhého** sloupce tabulky zapíšeme číslo, které padlo. Zároveň to bude představovat třídu  $i$ .

Do **třetího** sloupce označeného  $n_i$  kolikrát padlo (**pozorovaná četnost**) toto číslo.







Do **čtvrtého** sloupce označeného  $\hat{n}_i$  teoretickou (tu, kterou očekáváme) četnost.

V **pátém** sloupci označeném  $N_i$  jsou **kumulativní** pozorované četnosti. Tedy například ve druhém řádku je četnost výsledků, že padlo číslo menší nebo rovno **2**. Jinak řečeno, kolikrát padla **jednička** nebo **dvojka**.

A v **šestém** sloupci jsou **kumulativní** předpokládané četnosti.

Do **sedmého** sloupce zapíšeme hodnoty testového kritéria pro každou třídu.

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

$i$	tř. $i$	$n_i$	$\hat{n}_i$	$N_i = \sum_{j=1}^i n_j$	$\hat{N}_i = \sum_{j=1}^i \hat{n}_j$	$\frac{ N_i - \hat{N}_i }{n}$
1		7	10	7 = 7	10 = 10	
2		9	10	16 = 7 + 9	20 = 10 + 10	
3		10	10	26 = 7 + 9 + 10	30 = 10 + 10 + 10	
4		6	10	32 = 7 + 9 + 10 + 6	40 = 10 + 10 + 10 + 10	
5		15	10	47 = 7 + 9 + 10 + 6 + 15	50 = 10 + 10 + 10 + 10 + 10	
6		13	10	60 = 7 + 9 + 10 + 6 + 15 + 13	60 = 10 + 10 + 10 + 10 + 10 + 10	

$$60 \text{ hodů} \Rightarrow n = 60$$

$$D = \max_{\forall i} \frac{|N_i - \hat{N}_i|}{n}$$

**První** sloupec je číslo řádku, neboli index.

Do **druhého** sloupce tabulky zapíšeme číslo, které padlo. Zároveň to bude představovat třídu  $i$ .

Do **třetího** sloupce označeného  $n_i$  kolikrát padlo (**pozorovaná četnost**) toto číslo.







Do **čtvrtého** sloupce označeného  $\hat{n}_i$  teoretickou (tu, kterou očekáváme) četnost.

V **pátém** sloupci označeném  $N_i$  jsou **kumulativní** pozorované četnosti. Tedy například ve druhém řádku je četnost výsledků, že padlo číslo menší nebo rovno **2**. Jinak řečeno, kolikrát padla **jednička** nebo **dvojka**.

A v **šestém** sloupci jsou **kumulativní** předpokládané četnosti.

Do **sedmého** sloupce zapíšeme hodnoty testového kritéria pro každou třídu.

60 hodů — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

$i$	tř. $i$	$n_i$	$\hat{n}_i$	$N_i = \sum_{j=1}^i n_j$	$\hat{N}_i = \sum_{j=1}^i \hat{n}_j$	$\frac{ N_i - \hat{N}_i }{n}$
1		7	10	7 = 7	10 = 10	0,050 = $ 7 - 10  : 60$
2		9	10	16 = 7+9	20 = 10+10	0,067
3		10	10	26 = 7+9+10	30 = 10+10+10	0,067
4		6	10	32 = 7+9+10+6	40 = 10+10+10+10	<b>0,133</b>
5		15	10	47 = 7+9+10+6+15	50 = 10+10+10+10+10	0,050
6		13	10	60 = 7+9+10+6+15+13	60 = 10+10+10+10+10+10	0 = $ 60 - 60  : 60$

$$60 \text{ hodů} \Rightarrow n = 60$$

$$D = \max_{\forall i} \frac{|N_i - \hat{N}_i|}{n} = 0,133$$

**První** sloupec je číslo řádku, neboli index.

Do **druhého** sloupce tabulky zapíšeme číslo, které padlo. Zároveň to bude představovat třídu  $i$ .

Do **třetího** sloupce označeného  $n_i$  kolikrát padlo (**pozorovaná četnost**) toto číslo.

Do **čtvrtého** sloupce označeného  $\hat{n}_i$  teoretickou (tu, kterou očekáváme) četnost.







V **pátém** sloupci označeném  $N_i$  jsou **kumulativní** pozorované četnosti. Tedy například ve druhém řádku je četnost výsledků, že padlo číslo menší nebo rovno **2**. Jinak řečeno, kolikrát padla **jednička** nebo **dvojka**.

A v **šestém** sloupci jsou **kumulativní** předpokládané četnosti.

Do **sedmého** sloupce zapíšeme hodnoty testového kritéria pro každou třídu.

Pokud nechceme zbytečně šestkrát dělit, můžeme tabulku vyplnit následovně:

60 hodů ( $n$ ) — jednička 7×, dvojka 9×, trojka 10×, čtyřka 6×, pětka 15× a šestka 13×;  $\alpha = 0,01$ .

index	třída $i$	$n_i$	$\hat{n}_i$	$N_i = \sum_{j=1}^i n_j$	$\hat{N}_i = \sum_{j=1}^i \hat{n}_j$	$ N_i - \hat{N}_i $
1		7	10	7 = 7	10 = 10	3 =  7 - 10
2		9	10	16 = 7 + 9	20 = 10 + 10	4 =  16 - 20
3		10	10	26 = 7 + 9 + 10	30 = 10 + 10 + 10	4 =  26 - 30
4		6	10	32 = 7 + 9 + 10 + 6	40 = 10 + 10 + 10 + 10	8 =  32 - 40
5		15	10	47 = 7 + 9 + 10 + 6 + 15	50 = 10 + 10 + 10 + 10 + 10	3 =  47 - 50
6		13	10	60 = 7 + 9 + 10 + 6 + 15 + 13	60 = 10 + 10 + 10 + 10 + 10 + 10	0 =  60 - 60

$n_i$  ...pozorovaná četnost

$\hat{n}_i$  ...očekávaná (teoretická) četnost

$N_i$  ...kumulativní pozorovaná četnost

$\hat{N}_i$  ...kumulativní teoretická četnost

60 hodů  $\Rightarrow n = 60$

**Testové kritérium:**  $D = \frac{1}{n} \cdot \max_{\forall i} |N_i - \hat{N}_i| = \frac{1}{60} \cdot 8 \doteq 0,133$

**Obor pro přijetí nulové hypotézy** je:  $I_\alpha = \langle 0; D_\alpha(n) \rangle \doteq \langle 0; 0,210 \rangle$ ,  $D_{0,01}(60) \doteq \frac{1,63}{\sqrt{60}}$

**Závěr:** Protože v našem případě  $D \in I_{0,01}$ , **nezamítáme** nulovou hypotézu.

**Na hladině významnosti 1 % nelze tvrdit, že kostka je falešná.**

Dokonce ani na hladině významnosti 20 %, protože  $D_{20\%}(60) \doteq 0,138$ .



## Příklad

K dispozici máme následující datový vzorek, který již byl dříve **zpracován** ve formě vodorovné tabulky (data vzorku jsme zařadili do devíti tříd).

$k$	1	2	3	4	5	6	7	8	9
$x_k$	18	35	52	69	86	103	120	137	154
$n_k$	2	2	9	5	10	5	3	2	4

Pro celý statistický soubor, ze kterého byl vzorek vybrán, **určete**:

1. **bodový odhad střední hodnoty**  $\mu$ .
2. **intervalový odhad střední hodnoty**  $\mu$ , kde volte významnost 5 %.
3. s 95% **spolehlivostí**, zda **hypotézu**  $H_0 : \mu = 85$  **přijmout** či nikoliv.

**Řešení — 1. volba testových kritérií** ; 2. **aplikace** těchto kritérií

Intervalové odhady (druhý bod zadání) jsme si uváděli pouze pro soubory mající **normální rozdělení**. Proto budeme nejprve zkoumat, zda náš vzorek pochází ze souboru s normálním rozdělením. K tomu využijeme napřed první test shody – který jsme si uvedli – Pearsonův test  $\chi^2$ .

Vidíme, že asi nebude (poznáme to ale až určením teoretických četností  $\hat{n}_i$ ) splněna podmínka minimální třídní četnosti. Proto spojíme **první** a **druhou** třídu do jedné, která bude mít za reprezentanta hodnotu  $\frac{18+35}{2} = 26,5$ . Stejně tak i **šestou** se **sedmou** a **osmou** s **devátou**.

Protože testujeme, zda se jedná o normální rozdělení, upravíme i krajní meze hraničních intervalů. Ostatní hranice intervalů a všechny reprezentanty nespojovaných tříd ponecháme tak, jak byly.

Vše zase zapíšeme (svisle) do tabulky.

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$				
1	$(-\infty; 43,5)$	26,5	4					
2	$(43,5; 60,5)$	52	9					
3	$(60,5; 77,5)$	69	5					
4	$(77,5; 94,5)$	86	10					
5	$(94,5; 128,5)$	111,5	8					
6	$(128,5; \infty)$	145,5	6					
$\Sigma$			42					

Původní tabulka

statistických tabulkách

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$				
1	$(-\infty; 43,5)$	26,5	4					
2	$(43,5; 60,5)$	52	9					
3	$(60,5; 77,5)$	69	5					
4	$(77,5; 94,5)$	86	10					
5	$(94,5; 128,5)$	111,5	8					
6	$(128,5; \infty)$	145,5	6					
$\Sigma$			42					

### Původní tabulka

kde  $\hat{n}_r$  je teoretická třídní četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$			
1	$(-\infty; 43,5)$	26,5	4					
2	$(43,5; 60,5)$	52	9					
3	$(60,5; 77,5)$	69	5					
4	$(77,5; 94,5)$	86	10					
5	$(94,5; 128,5)$	111,5	8					
6	$(128,5; \infty)$	145,5	6					
$\Sigma$			42					

### Původní tabulka

kde  $\hat{n}_r$  je teoretická třídň četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \Sigma(n_r \cdot x_r)$

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$		
1	$(-\infty; 43,5)$	26,5	4		106			
2	$(43,5; 60,5)$	52	9		468			
3	$(60,5; 77,5)$	69	5		345			
4	$(77,5; 94,5)$	86	10		860			
5	$(94,5; 128,5)$	111,5	8		892			
6	$(128,5; \infty)$	145,5	6		873			
$\Sigma$			42		3 544			

### Původní tabulka

kde  $\hat{n}_r$  je teoretická třídň četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \doteq 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right]$$

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$		
1	$(-\infty; 43,5)$	26,5	4		106	2 809		
2	$(43,5; 60,5)$	52	9		468	24 336		
3	$(60,5; 77,5)$	69	5		345	23 805		
4	$(77,5; 94,5)$	86	10		860	73 960		
5	$(94,5; 128,5)$	111,5	8		892	99 458		
6	$(128,5; \infty)$	145,5	6		873	127 021,5		
$\Sigma$			42		3 544	351 389,5		

## Původní tabulka

kde  $\hat{n}_r$  je teoretická třídní četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$		
1	$(-\infty; 43,5)$	26,5	4		106	2 809		
2	$(43,5; 60,5)$	52	9		468	24 336		
3	$(60,5; 77,5)$	69	5		345	23 805		
4	$(77,5; 94,5)$	86	10		860	73 960		
5	$(94,5; 128,5)$	111,5	8		892	99 458		
6	$(128,5; \infty)$	145,5	6		873	127 021,5		
$\Sigma$			42		3 544	351 389,5		

### Původní tabulka

kde  $\hat{n}_r$  je teoretická třídní četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

Pak:  $\sigma = \sqrt{\sigma^2} = \sqrt{1\,277} \hat{=} 36,1$

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$	$A = \frac{a_r - \mu}{\sigma}$	
1	$(-\infty; 43,5)$	26,5	4		106	2 809		
2	$(43,5; 60,5)$	52	9		468	24 336		
3	$(60,5; 77,5)$	69	5		345	23 805		
4	$(77,5; 94,5)$	86	10		860	73 960		
5	$(94,5; 128,5)$	111,5	8		892	99 458		
6	$(128,5; \infty)$	145,5	6		873	127 021,5		
$\Sigma$			42		3 544	351 389,5		

### Původní tabulka

kde  $\hat{n}_r$  je teoretická třídní četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

Pak:  $\sigma = \sqrt{\sigma^2} = \sqrt{1\,277} \hat{=} 36,1$



r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$	$A = \frac{a_r - \mu}{\sigma}$	
1	$(-\infty; 43,5)$	26,5	4		106	2 809	$-\infty$	
2	$(43,5; 60,5)$	52	9		468	24 336	-1,14	
3	$(60,5; 77,5)$	69	5		345	23 805	-0,67	
4	$(77,5; 94,5)$	86	10		860	73 960	-0,19	
5	$(94,5; 128,5)$	111,5	8		892	99 458	0,28	
6	$(128,5; \infty)$	145,5	6		873	127 021,5	1,23	
$\Sigma$			42		3 544	351 389,5	—	

### Původní tabulka

kde  $\hat{n}_r$  je teoretická třídň četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

$$\text{Pak: } \sigma = \sqrt{\sigma^2} = \sqrt{1\,277} \hat{=} 36,1 \quad \text{a} \quad F(77,5) = F_N\left(\frac{77,5-84,4}{36,1}\right) \hat{=} F_N(-0,19) = 1 - F_N(0,19)$$

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$	$A = \frac{a_r - \mu}{\sigma}$	$F_N(A)$
1	$(-\infty; 43,5)$	26,5	4		106	2 809	$-\infty$	
2	$(43,5; 60,5)$	52	9		468	24 336	-1,14	
3	$(60,5; 77,5)$	69	5		345	23 805	-0,67	
4	$(77,5; 94,5)$	86	10		860	73 960	-0,19	0,424 65
5	$(94,5; 128,5)$	111,5	8		892	99 458	0,28	
6	$(128,5; \infty)$	145,5	6		873	127 021,5	1,23	
$\Sigma$			42		3 544	351 389,5	—	

## Původní tabulka

kde  $\hat{n}_r$  je teoretická třídní četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

$$\text{Pak: } \sigma = \sqrt{\sigma^2} = \sqrt{1\,277} \hat{=} 36,1 \quad \text{a} \quad F(77,5) = F_N\left(\frac{77,5-84,4}{36,1}\right) \hat{=} F_N(-0,19) = 1 - F_N(0,19) = 0,424\,65$$

Tab 1. Distribuční funkce  $F_N(u)$  normovaného normálního rozdělení  $N(0, 1)$ 

x	0	1	2	3	4	5	6	7	8	9
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$	$A = \frac{a_r - \mu}{\sigma}$	$F_N(A)$
1	$(-\infty; 43,5)$	26,5	4		106	2 809	$-\infty$	
2	$(43,5; 60,5)$	52	9		468	24 336	-1,14	
3	$(60,5; 77,5)$	69	5		345	23 805	-0,67	
4	$(77,5; 94,5)$	86	10		860	73 960	-0,19	0,424 65
5	$(94,5; 128,5)$	111,5	8		892	99 458	0,28	
6	$(128,5; \infty)$	145,5	6		873	127 021,5	1,23	
$\Sigma$			42		3 544	351 389,5	—	

## Původní tabulka

kde  $\hat{n}_r$  je teoretická třídní četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

Pak:  $\sigma = \sqrt{\sigma^2} = \sqrt{1\,277} \hat{=} 36,1$  a  $F(77,5) = F_N\left(\frac{77,5-84,4}{36,1}\right) \hat{=} F_N(-0,19) = 1 - F_N(0,19) = 0,424\,65$

$f_x$	=NORM.DIST(77,5;84,4;36;1)	
C	D	E
	0,424001659	

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$	$A = \frac{a_r - \mu}{\sigma}$	$F_N(A)$
1	$(-\infty; 43,5)$	26,5	4		106	2 809	$-\infty$	0
2	$(43,5; 60,5)$	52	9		468	24 336	-1,14	0,127 14
3	$(60,5; 77,5)$	69	5		345	23 805	-0,67	0,251 43
4	$(77,5; 94,5)$	86	10		860	73 960	-0,19	0,424 65
5	$(94,5; 128,5)$	111,5	8		892	99 458	0,28	0,610 26
6	$(128,5; \infty)$	145,5	6		873	127 021,5	1,23	0,890 65
$\Sigma$			42		3 544	351 389,5	—	—

## Původní tabulka

kde  $\hat{n}_r$  je teoretická třídň četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \Sigma(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \Sigma(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

$$\text{Pak: } \sigma = \sqrt{\sigma^2} = \sqrt{1\,277} \hat{=} 36,1 \quad \text{a} \quad F(77,5) = F_N\left(\frac{77,5-84,4}{36,1}\right) \hat{=} F_N(-0,19) = 1 - F_N(0,19) = 0,424\,65$$

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$	$A = \frac{a_r - \mu}{\sigma}$	$F_N(A)$
1	$(-\infty; 43,5)$	26,5	4	5,340	106	2 809	$-\infty$	0
2	$(43,5; 60,5)$	52	9	5,220	468	24 336	-1,14	0,127 14
3	$(60,5; 77,5)$	69	5	7,275	345	23 805	-0,67	0,251 43
4	$(77,5; 94,5)$	86	10	7,796	860	73 960	-0,19	0,424 65
5	$(94,5; 128,5)$	111,5	8	11,776	892	99 458	0,28	0,610 26
6	$(128,5; \infty)$	145,5	6	4,593	873	127 021,5	1,23	0,890 65
$\Sigma$			42	—	3 544	351 389,5	—	—

### Původní tabulka

kde  $\hat{n}_r$  je teoretická třídní četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \Sigma(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \Sigma(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

$$\text{Pak: } \sigma = \sqrt{\sigma^2} = \sqrt{1\,277} \hat{=} 36,1 \quad \text{a} \quad F(77,5) = F_N\left(\frac{77,5-84,4}{36,1}\right) \hat{=} F_N(-0,19) = 1 - F_N(0,19) = 0,424\,65$$

r	$(a_r; b_r)$	$x_r$	$n_r$	$\hat{n}_r$	$n_r \cdot x_r$	$n_r \cdot x_r^2$	$A = \frac{a_r - \mu}{\sigma}$	$F_N(A)$
1	$(-\infty; 43,5)$	26,5	4	5,340	106	2 809	$-\infty$	0
2	$(43,5; 60,5)$	52	9	5,220	468	24 336	-1,14	0,127 14
3	$(60,5; 77,5)$	69	5	7,275	345	23 805	-0,67	0,251 43
4	$(77,5; 94,5)$	86	10	7,796	860	73 960	-0,19	0,424 65
5	$(94,5; 128,5)$	111,5	8	11,776	892	99 458	0,28	0,610 26
6	$(128,5; \infty)$	145,5	6	4,593	873	127 021,5	1,23	0,890 65
$\Sigma$			42	—	3 544	351 389,5	—	—

### Původní tabulka

kde  $\hat{n}_r$  je teoretická třídní četnost pro normální rozdělení.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) = n \cdot [F(b_r) - F(a_r)] = n \cdot \left[ F_N\left(\frac{b_r - \mu}{\sigma}\right) - F_N\left(\frac{a_r - \mu}{\sigma}\right) \right] \text{ a } F_N(-u) = 1 - F_N(u).$$

Chceme-li určit např.  $F(77,5) = ?$ , potřebujeme znát  $\mu$  a  $\sigma$ , abychom ve [statistických tabulkách](#) nebo pomocí Excelu (či jinak) našli hodnotu distribuční funkce  $F_N$  normovaného normálního rozložení  $N(0; 1)$ .

Proto provedeme **bodové odhady**:  $\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \Sigma(n_r \cdot x_r) = \frac{1}{42} \cdot 3\,544 \hat{=} 84,4$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \Sigma(n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (351\,389,5 - 42 \cdot 84,4^2) \hat{=} \frac{52\,343,4}{41} \hat{=} 1\,277$$

$$\text{Pak: } \sigma = \sqrt{\sigma^2} = \sqrt{1\,277} \hat{=} 36,1 \quad \text{a} \quad F(77,5) = F_N\left(\frac{77,5-84,4}{36,1}\right) \hat{=} F_N(-0,19) = 1 - F_N(0,19) = 0,424\,65$$

Vidíme, že není splněna nutná podmínka v šesté třídě, protože očekávaná četnost  $\hat{n}_6 \hat{=} 4,6$  není větší než 5. Proto necháme původní šestou třídu tak jak byla a spojíme **sedmou** a **osmou** s **devátou** (všechny tři třídy dohromady) třídou. A celý výpočet provedeme znovu! Asi se změní  $\mu$  a  $\sigma$ .

r	$a_r ; b_r$	$x_r$	$n_r$	$\hat{n}_r$			
1	$(-\infty ; 43,5)$	26,5	4				
2	$\langle 43,5 ; 60,5)$	52	9				
3	$\langle 60,5 ; 77,5)$	69	5				
4	$\langle 77,5 ; 94,5)$	86	10				
5	$\langle 94,5 ; 111,5)$	103	5				
6	$\langle 111,5 ; \infty)$	137	9				
$\Sigma$			<b>42</b>				

$$P(a_r \leq X \leq b_r) = F(b_r) - F(a_r)$$

kde  $F(x)$  je distribuční funkce ověřovaného **normálního** rozdělení

Původní tabulka

r	$a_r ; b_r$	$x_r$	$n_r$	$\hat{n}_r$			
1	$(-\infty ; 43,5)$	26,5	4				
2	$\langle 43,5 ; 60,5)$	52	9				
3	$\langle 60,5 ; 77,5)$	69	5				
4	$\langle 77,5 ; 94,5)$	86	10				
5	$\langle 94,5 ; 111,5)$	103	5				
6	$\langle 111,5 ; \infty)$	137	9				
$\Sigma$			42				

$$P(a_r \leq X \leq b_r) = \\ = F(b_r) - F(a_r)$$

kde  $F(x)$  je distribuční  
funkce ověřovaného  
**normálního** rozdělení

### Původní tabulka

a teoretickou třídní četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:



r	$a_r ; b_r$	$x_r$	$n_r$	$\hat{n}_r$	$x_r \cdot n_r$		
1	$(-\infty ; 43,5)$	26,5	4				
2	$\langle 43,5 ; 60,5)$	52	9				
3	$\langle 60,5 ; 77,5)$	69	5				
4	$\langle 77,5 ; 94,5)$	86	10				
5	$\langle 94,5 ; 111,5)$	103	5				
6	$\langle 111,5 ; \infty)$	137	9				
$\Sigma$			<b>42</b>				

$$P(a_r \leq X \leq b_r) = F(b_r) - F(a_r)$$

kde  $F(x)$  je distribuční funkce ověřovaného **normálního** rozdělení

### Původní tabulka

a teoretickou třídni četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:

$$\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \Sigma(n_r \cdot x_r)$$

r	$a_r ; b_r$	$x_r$	$n_r$	$\hat{n}_r$	$x_r \cdot n_r$	$x_r^2 \cdot n_r$	
1	$(-\infty ; 43,5)$	26,5	4		106		
2	$\langle 43,5 ; 60,5)$	52	9		468		
3	$\langle 60,5 ; 77,5)$	69	5		345		
4	$\langle 77,5 ; 94,5)$	86	10		860		
5	$\langle 94,5 ; 111,5)$	103	5		515		
6	$\langle 111,5 ; \infty)$	137	9		1 233		
$\Sigma$			<b>42</b>		<b>3 527</b>		

$P(a_r \leq X \leq b_r) =$   
 $= F(b_r) - F(a_r)$   
 kde  $F(x)$  je distribuční  
 funkce ověřovaného  
**normálního** rozdělení

### Původní tabulka

a teoretickou třídní četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:

$$\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum (n_r \cdot x_r) = \frac{1}{42} \cdot 3\,527 = 83,976\,190 \doteq 84,0 \quad (\text{dříve } 84,4)$$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum (n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right]$$

r	$a_r; b_r$	$x_r$	$n_r$	$\hat{n}_r$	$x_r \cdot n_r$	$x_r^2 \cdot n_r$	
1	$(-\infty; 43,5)$	26,5	4		106	2 809	
2	$\langle 43,5; 60,5)$	52	9		468	24 336	
3	$\langle 60,5; 77,5)$	69	5		345	23 805	
4	$\langle 77,5; 94,5)$	86	10		860	73 960	
5	$\langle 94,5; 111,5)$	103	5		515	53 045	
6	$\langle 111,5; \infty)$	137	9		1 233	168 921	
$\Sigma$			<b>42</b>		<b>3 527</b>	<b>346 876</b>	

$P(a_r \leq X \leq b_r) =$   
 $= F(b_r) - F(a_r)$   
 kde  $F(x)$  je distribuční  
 funkce ověřovaného  
**normálního** rozdělení

### Původní tabulka

a teoretickou třídni četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:

$$\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum (n_r \cdot x_r) = \frac{1}{42} \cdot 3\,527 = 83,976\,190 \doteq 84,0 \quad (\text{dříve } 84,4)$$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum (n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (346\,876 - 42 \cdot 84^2) = \frac{50\,524}{41} \doteq 1\,232$$

r	$a_r; b_r$	$x_r$	$n_r$	$\hat{n}_r$	$x_r \cdot n_r$	$x_r^2 \cdot n_r$	
1	$(-\infty; 43,5)$	26,5	4		106	2 809	
2	$\langle 43,5; 60,5 \rangle$	52	9		468	24 336	
3	$\langle 60,5; 77,5 \rangle$	69	5		345	23 805	
4	$\langle 77,5; 94,5 \rangle$	86	10		860	73 960	
5	$\langle 94,5; 111,5 \rangle$	103	5		515	53 045	
6	$\langle 111,5; \infty \rangle$	137	9		1 233	168 921	
$\Sigma$			<b>42</b>		<b>3 527</b>	<b>346 876</b>	

$P(a_r \leq X \leq b_r) =$   
 $= F(b_r) - F(a_r)$   
 kde  $F(x)$  je distribuční  
 funkce ověřovaného  
**normálního** rozdělení

### Původní tabulka

a teoretickou třídni četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:

$$\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum (n_r \cdot x_r) = \frac{1}{42} \cdot 3\,527 = 83,976\,190 \doteq 84,0 \quad (\text{dříve } 84,4)$$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum (n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (346\,876 - 42 \cdot 84^2) = \frac{50\,524}{41} \doteq 1\,232$$

$$\text{Potom } \sigma = \sqrt{\sigma^2} = \sqrt{1\,232} \doteq 35,1 \quad (\text{dříve } 36,1)$$

r	$a_r ; b_r$	$x_r$	$n_r$	$\hat{n}_r$	$x_r \cdot n_r$	$x_r^2 \cdot n_r$	
1	$(-\infty ; 43,5)$	26,5	4	8,024	106	2 809	
2	$\langle 43,5 ; 60,5)$	52	9		468	24 336	
3	$\langle 60,5 ; 77,5)$	69	5		345	23 805	
4	$\langle 77,5 ; 94,5)$	86	10		860	73 960	
5	$\langle 94,5 ; 111,5)$	103	5		515	53 045	
6	$\langle 111,5 ; \infty)$	137	9		1 233	168 921	
$\Sigma$			<b>42</b>		<b>3 527</b>	<b>346 876</b>	

$P(a_r \leq X \leq b_r) =$   
 $= F(b_r) - F(a_r)$   
 kde  $F(x)$  je distribuční  
 funkce ověřovaného  
**normálního** rozdělení

### Původní tabulka

a teoretickou třídní četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:

$$\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum (n_r \cdot x_r) = \frac{1}{42} \cdot 3\,527 = 83,976\,190 \doteq 84,0 \quad (\text{dříve } 84,4)$$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum (n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (346\,876 - 42 \cdot 84^2) = \frac{50\,524}{41} \doteq 1\,232$$

$$\text{Potom } \sigma = \sqrt{\sigma^2} = \sqrt{1\,232} \doteq 35,1 \quad (\text{dříve } 36,1)$$

$\hat{f}_x$	=42*(NORM.DIST(94,5;84;35,1;1)-NORM.DIST(77,5;84;35,1;1))					
C	D	E	F	G	H	
			8,02382757			

r	$a_r ; b_r$	$x_r$	$n_r$	$\hat{n}_r$	$x_r \cdot n_r$	$x_r^2 \cdot n_r$	
1	$(-\infty ; 43,5)$	26,5	4	5,220	106	2 809	
2	$\langle 43,5 ; 60,5)$	52	9	5,347	468	24 336	
3	$\langle 60,5 ; 77,5)$	69	5	7,348	345	23 805	
4	$\langle 77,5 ; 94,5)$	86	10	8,024	860	73 960	
5	$\langle 94,5 ; 111,5)$	103	5	6,961	515	53 045	
6	$\langle 111,5 ; \infty)$	137	9	9,100	1 233	168 921	
$\Sigma$			<b>42</b>	42,000	<b>3 527</b>	<b>346 876</b>	

$P(a_r \leq X \leq b_r) =$   
 $= F(b_r) - F(a_r)$   
 kde  $F(x)$  je distribuční  
 funkce ověřovaného  
**normálního** rozdělení

### Původní tabulka

a teoretickou třídni četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:

$$\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum (n_r \cdot x_r) = \frac{1}{42} \cdot 3\,527 = 83,976\,190 \doteq 84,0 \quad (\text{dříve } 84,4)$$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum (n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (346\,876 - 42 \cdot 84^2) = \frac{50\,524}{41} \doteq 1\,232$$

$$\text{Potom } \sigma = \sqrt{\sigma^2} = \sqrt{1\,232} \doteq 35,1 \quad (\text{dříve } 36,1)$$

$$\text{Testové kritérium: } \chi^2 = \sum \frac{(n_r - \hat{n}_r)^2}{\hat{n}_r}$$

r	$a_r ; b_r$	$x_r$	$n_r$	$\hat{n}_r$	$x_r \cdot n_r$	$x_r^2 \cdot n_r$	$\frac{(n_r - \hat{n}_r)^2}{\hat{n}_r}$
1	$(-\infty ; 43,5)$	26,5	4	5,220	106	2 809	
2	$\langle 43,5 ; 60,5)$	52	9	5,347	468	24 336	
3	$\langle 60,5 ; 77,5)$	69	5	7,348	345	23 805	
4	$\langle 77,5 ; 94,5)$	86	10	8,024	860	73 960	
5	$\langle 94,5 ; 111,5)$	103	5	6,961	515	53 045	
6	$\langle 111,5 ; \infty)$	137	9	9,100	1 233	168 921	
$\Sigma$			<b>42</b>	42,000	<b>3 527</b>	<b>346 876</b>	

$P(a_r \leq X \leq b_r) =$   
 $= F(b_r) - F(a_r)$   
 kde  $F(x)$  je distribuční  
 funkce ověřovaného  
**normálního** rozdělení

### Původní tabulka

a teoretickou třídni četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:

$$\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum (n_r \cdot x_r) = \frac{1}{42} \cdot 3\,527 = 83,976\,190 \doteq 84,0 \quad (\text{dříve } 84,4)$$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum (n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (346\,876 - 42 \cdot 84^2) = \frac{50\,524}{41} \doteq 1\,232$$

$$\text{Potom } \sigma = \sqrt{\sigma^2} = \sqrt{1\,232} \doteq 35,1 \quad (\text{dříve } 36,1)$$

$$\text{Testové kritérium: } \chi^2 = \sum \frac{(n_r - \hat{n}_r)^2}{\hat{n}_r}$$

r	$a_r; b_r$	$x_r$	$n_r$	$\hat{n}_r$	$x_r \cdot n_r$	$x_r^2 \cdot n_r$	$\frac{(n_r - \hat{n}_r)^2}{\hat{n}_r}$
1	$(-\infty; 43,5)$	26,5	4	5,220	106	2 809	0,295
2	$\langle 43,5; 60,5)$	52	9	5,347	468	24 336	2,494
3	$\langle 60,5; 77,5)$	69	5	7,348	345	23 805	0,746
4	$\langle 77,5; 94,5)$	86	10	8,024	860	73 960	0,495
5	$\langle 94,5; 111,5)$	103	5	6,961	515	53 045	0,547
6	$\langle 111,5; \infty)$	137	9	9,100	1 233	168 921	0,001
$\Sigma$			<b>42</b>	42,000	<b>3 527</b>	<b>346 876</b>	<b>4,577</b>

$P(a_r \leq X \leq b_r) =$   
 $= F(b_r) - F(a_r)$   
 kde  $F(x)$  je distribuční  
 funkce ověřovaného  
**normálního** rozdělení

### Původní tabulka

a teoretickou třídní četnost  $\hat{n}_r$  pro normální rozdělení určíme na počítači.

$$\hat{n}_r = n \cdot P(a_r \leq X \leq b_r) [=42 \cdot (\text{NORM.DIST}(b_r; \mu; \sigma; 1) - \text{NORM.DIST}(a_r; \mu; \sigma; 1))] \Leftarrow [\text{Excel 2010}]$$

K tomu potřebujeme znát  $\mu$  a  $\sigma$ . Proto provedeme **bodové odhady**:

$$\mu \hat{=} \bar{x}_A = \bar{x} = \frac{1}{n} \cdot \sum (n_r \cdot x_r) = \frac{1}{42} \cdot 3\,527 = 83,976\,190 \doteq 84,0 \quad (\text{dříve } 84,4)$$

$$\sigma^2 \hat{=} S^2 = \frac{1}{n-1} \left[ \sum (n_r \cdot x_r^2) - n \cdot \bar{x}^2 \right] = \frac{1}{42-1} \cdot (346\,876 - 42 \cdot 84^2) = \frac{50\,524}{41} \doteq 1\,232$$

$$\text{Potom } \sigma = \sqrt{\sigma^2} = \sqrt{1\,232} \doteq 35,1 \quad (\text{dříve } 36,1)$$

$$\text{Testové kritérium: } \chi^2 = \sum \frac{(n_r - \hat{n}_r)^2}{\hat{n}_r} = 4,577$$

$$\text{Obor přijetí hypotézy: } I_{5\%} = \langle 0; \chi_{1-0,05}^2(6-1-2) \rangle = \langle 0; \chi_{0,95}^2(3) \rangle = \langle 0; 7,815 \rangle \Rightarrow \chi^2 \in I_{0,05}$$

Protože hodnota testového kritéria patří do oboru přijetí hypotézy, nelze v tomto případě vyloučit, že vzorek pochází za základního souboru, který je rozložen normálně (má normální rozložení).



## Kolmogorovův–Smirnovův test

- V **původní tabulce** jsme upravili spodní mez první třídy a horní mez poslední třídy (normální rozdělení je od  $-\infty$  do  $\infty$ ); protože Excel (jako většina programů) neumí pracovat se symbolem nekonečno, nahradíme jej hodnotami z několika devítek. Dále využijeme již dříve spočítaný **aritmetický průměr**  $\bar{x}_A \doteq 84,4$  a **směrodatnou odchylku**  $S \doteq 37$ .
- Hodnoty  $\tilde{N}_k$  (kumulované četnosti normálního rozdělení pro horní hranici  $b_k$  dané třídy) získáme pomocí *Excelu 2010*: =NORM.DIST(**b**; 84, 4; 37; 1)\*42

Zde jsme třídy označili indexem **k** a ne **r** jako v předchozím příkladu, ale to doufám příliš nevadí.

k	(a <sub>k</sub> ; b <sub>k</sub> )		n <sub>k</sub>	N <sub>k</sub>	$\tilde{N}_k$	$ N_k - \tilde{N}_k /n$
1	-999999	26,5	2	2	2,470	0,011
2	26,5	43,5	2	4	5,649	0,039
3	43,5	60,5	9	13	10,885	0,050
4	60,5	77,5	5	18	17,893	0,003
5	77,5	94,5	10	28	25,518	<b>0,059</b>
6	94,5	111,5	5	33	32,258	0,018
7	111,5	128,5	3	36	37,101	0,026
8	128,5	145,5	2	38	39,928	0,046
9	145,5	999999	4	42	42	0

$$n = 42$$

$$D = \max_{\forall k} \frac{|N_k - \tilde{N}_k|}{n} \doteq 0,059$$

$$I_\alpha = \langle 0 ; D_\alpha(n) \rangle \doteq \langle 0 ; 0,210 \rangle$$

$$D_{0,05}(42) \doteq \frac{1,36}{\sqrt{42}}$$

$$D(X) = \max. \quad 0,0591$$

Protože hodnota testového kritéria patří do oboru přijetí hypotézy, nelze na hladině významnosti 5 % vyloučit, že vzorek pochází za základního souboru, který je rozložen normálně (má normální rozložení). Tedy nulovou hypotézu, že: *vzorek pochází z populace mající normální rozdělení*, **NEzamítáme**.

Zopakujme si, co jsme zatím (u vzorku, který má 42 hodnot) vyřešili:

Vzorek byl „asi“ vybrán ze souboru s **normálním rozdělením** a má tyto charakteristiky:

$\bar{x} = 84,4$      $S = 37$     A cože máme vlastně pro celý základní statistický soubor **vypočítat**?

**Bodový odhad střední hodnoty** základního souboru (populace)  $\mu \doteq \bar{x} = 84,4$  jsme již vyřešili.

**Intervalový odhad střední hodnoty** na hladině významnosti 5 %

$$\begin{aligned} & \left( \bar{x} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1); \bar{x} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1) \right) = \\ & = \left( 84,4 - \frac{37}{\sqrt{42}} \cdot t_{1-\frac{0,05}{2}}(42-1); 84,4 + \frac{37}{\sqrt{42}} \cdot t_{1-\frac{0,05}{2}}(42-1) \right) = \\ & = (84,4 - 5,709 \cdot t_{0,975}(41); 84,4 + 5,709 \cdot t_{0,975}(41)) = (84,4 - 11,538; 84,4 + 11,538) = (72,862; 95,938) \end{aligned}$$

Intervalový odhad střední hodnoty populace na hladině významnosti 5 % je: **(72,862 ; 95,938)**.

Hodnotu  $t_{0,975}(41) \doteq 2,021$  najdeme v **tabulkách**, nebo využijeme *Excel 2010*: =T.INV.2T(0,05;41)

K určení hodnoty  $\frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1) \doteq 11,538$  můžeme také využít *Excel*: =CONFIDENCE.T( $\alpha$ ;  $\sigma$ ;  $n$ ),

kde  $\sigma$  nahradíme  $S$ .

Tím se sice dopustíme chyby (která není až tak velká), protože správně máme zadat směrodatnou odchylku základního souboru  $\sigma$ , ale tu neznáme. Proto místo ní použijeme její bodový odhad – výběrovou směrodatnou odchylku vzorku  $S$ .

CONFIDENCE.T

Alfa 0,05 = 0,05

Sm\_odch 37 = 37

Velikost 42 = 42

= 11,53001167

Vrátí interval spolehlivosti pro střední hodnotu základního souboru pomocí Studentova t-rozdělení.

Velikost je velikost výběru.

Výsledek = 11,53001167

[Nápověda k této funkci](#) OK Storno

Zopakujme si, co jsme zatím (u vzorku, který má 42 hodnot) vyřešili:

Vzorek byl „asi“ vybrán ze souboru s **normálním rozdělením** a má tyto charakteristiky:

$\bar{x} = 84,4$      $S = 37$     A cože máme vlastně pro celý základní statistický soubor **vypočítat**?

**Bodový odhad střední hodnoty** základního souboru (populace)  $\mu \doteq \bar{x} = 84,4$  jsme již vyřešili.

**Intervalový odhad střední hodnoty** na hladině významnosti 5 %

$$\begin{aligned} & \left( \bar{x} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1); \bar{x} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1) \right) = \\ & = \left( 84,4 - \frac{37}{\sqrt{42}} \cdot t_{1-\frac{0,05}{2}}(42-1); 84,4 + \frac{37}{\sqrt{42}} \cdot t_{1-\frac{0,05}{2}}(42-1) \right) = \\ & = (84,4 - 5,709 \cdot t_{0,975}(41); 84,4 + 5,709 \cdot t_{0,975}(41)) = (84,4 - 11,538; 84,4 + 11,538) = (72,862; 95,938) \end{aligned}$$

Intervalový odhad střední hodnoty populace na hladině významnosti 5 % je: **(72,862 ; 95,938)**.

Hodnotu  $t_{0,975}(41) \doteq 2,021$  najdeme v **tabulkách**, nebo využijeme *Excel 2010*: =T.INV.2T(0,05;41)

**Hypotézu o střední hodnotě**  $H_0 : \mu = 85$  s 95% spolehlivostí přijmout či odmítnout?

Alternativní hypotéza:  $H_A : \mu \neq 85$

Testové kritérium: 
$$T = \frac{(\bar{x} - \mu) \cdot \sqrt{n}}{S} = \frac{(84,4 - 85) \cdot \sqrt{42}}{37} \doteq \frac{(-0,6) \cdot 6,481}{37} = \frac{-3,888}{37} \doteq -0,105$$

Obor přijetí hypotézy: 
$$I_\alpha = \left\langle -t_{1-\frac{\alpha}{2}}(n-1); t_{1-\frac{\alpha}{2}}(n-1) \right\rangle = \langle -t_{0,975}(41); t_{0,975}(41) \rangle = \langle -2,021; 2,021 \rangle \Rightarrow T \in I_{0,05},$$

hypotézu  $H_0$ , že střední hodnota  $\mu = 85$  s 95% spolehlivostí **nezamítáme**.

## Poznámka o strojovém zpracování.

Zatímco při **klasickém testování** v předchozím příkladu bylo třeba hledat kritické meze příslušného testovacího kritéria, *každý slušnější* statistický software vypisuje takzvanou **hodnotu významnosti** <sup>41</sup> (též zvanou signifikance nebo  $p$ -hodnota, jejíž velikost vůbec nezávisí na zvolené hladině spolehlivosti  $\alpha$ ). Tato hodnota udává pravděpodobnost, že při platnosti nulové hypotézy vyjde testová statistika rovna naměřené nebo ještě extrémnější. Hodnota významnosti  **$p$**  ( $p$ -hodnota,  $p$ -value, significance level) tedy představuje minimální hladinu významnosti, na které je možno zamítnout nulovou hypotézu.

Test se vyhodnocuje takto:

- Je-li hodnota významnosti menší než hladina spolehlivosti ( $p < \alpha$ ), pak zamítneme nulovou hypotézu a přijmeme alternativní hypotézu. Riskujeme chybu prvního druhu (že zamítneme správnou hypotézu) s pravděpodobností nanejvýš  $\alpha$ .
- Je-li hodnota významnosti větší nebo rovna hladině spolehlivosti ( $p \geq \alpha$ ), pak nulovou hypotézu nezamítneme.

Tento postup využívá početně (většinou) náročnějšího **Čistého testu významnosti**.

<sup>41</sup> Je to hodnota hladiny významnosti, kterou bychom museli volit, aby vypočtená hodnota testovací statistiky se rovnala právě kritické hodnotě. Tedy aby hodnota testovací statistiky ležela právě na hranici mezi oborem přijetí hypotézy a kritickým oborem, ve kterém hypotézu zamítáme. Nebo ještě jinak řečeno: **Moderní statistické programy při výpočtech předkládají přímo pravděpodobnost chyby I. řádu.**

# Úvod do

# **Regresní a korelační analýzy**

# Obsah kapitoly: Regresní a korelační analýza

<b>1. Souvislosti mezi jevy</b>	<b>251</b>
<b>2. Regresní analýza</b>	<b>253</b>
2.1. Regresní přímka — lineární regrese	257
Poznámka k metodě nejmenších čtverců	259
2.2. Regresní parabola — kvadratická regrese	261
2.3. Volba regresní funkce	262
2.3.1 Lineární závislost:	262
2.3.2 Kvadratická závislost:	265
<b>3. Korelační analýza — výběrový korelační koeficient</b>	<b>266</b>
3.1. Příklady	268
3.1.1. Odlehlé pozorování a původní nekorelovaný vzorek	268
3.1.2. Vzorek téměř nekorelovaný, jeho části perfektně korelované	269
3.1.3. Vzorek pozitivně korelovaný, jeho části negativně korelované	270
3.2. Test významnosti hodnoty korelačního koeficientu $r$	271
<b>4. Příklad</b>	<b>272</b>
Lineární regrese	272
Excel	272
Kovariance	273
Soustavy normálních rovnic	279
Výběrový korelační koeficient	287
Kvadratická regrese	288
<b>5. Závěr kapitoly – Radíte se s rozumem</b>	<b>295</b>

# 1. Souvislosti mezi jevy

Zkoumání souvislostí (zkoumání tzv. korelace mezi jevy)

- vztah mezi průměrnou rychlostí auta a průměrnou spotřebou pohonných hmot,
- vztah mezi spotřebou hnojiva a výnosem,
- vztah mezi rychlostí auta a délkou dráhy, kterou auto urazí za stejný čas,
- a další a další (viz třeba příklad o [víně](#))

je jedním z nejdůležitějších úkolů statistiky. Snažíme se o (matematický) popis systematických okolností, které provází první dvě zmíněné **volné**<sup>42</sup> (tzv. **stochastické**) závislosti. Třetí uvedená závislost je **pevná** (**funkční**), protože vzdálenost závisí pouze na čase a rychlosti.

Východiskem k popisu statistických závislostí jsou statistické údaje. První informace o průběhu závislosti dvou proměnných (znaků) získáme již tak, že údaje uspořádáme do tabulky. Například takto:

	muž	žena	$\Sigma$
rtěnku POUŽÍVÁ	50	950	1 000
NEpoužívá rtěnku	850	50	900
$\Sigma$	900	1 000	1 900

A proč si všímáme závislostí mezi proměnnými? Protože žádný jev v přírodě ani ve společnosti nevniká ani neprobíhá libovolně, ale je ve vztahu k jiným jevům a nemůže být pochopen správně, je-li

<sup>42</sup> Není zaručeno, že když na jeden ar aplikujeme dané množství hnojiva, tak ze sousedního aru při stejném množství hnojiva budeme mít naprosto stejný výnos. Tedy určité hodnotě  $x$  (hnojivo) neodpovídá jediná hodnota  $y$  (výnos), ale celé rozdělení hodnot  $y$ , které kolísají s určitým rozptylem kolem určité střední hodnoty. Podobně jako v případě vzdělání versus plat v [pravém](#) obrázku.

z těchto vztahů a souvislostí vytržen. S nejjednoduššími formami příčinných souvislostí (závislostí veličin) se setkáváme u některých přírodních jevů. Se složitými formami se setkáváme u jevů společenských (ekonomických).

Soubor postupů a metod, dovolujících řešení závislosti veličin, se nazývá **regresní** (termín regrese »krok zpět« naprosto nevystihuje podstatu problému; vznikl historicky a nadále se používá) **a korelační analýza**. Tato analýza umožňuje řešit dvě základní úlohy. A to:

**Regresní úlohu** — zjistit **formu závislosti** a vyjádřit ji **matematickou** (tzv. regresní) **funkcí**.

Jedna veličina je považovaná za závislou (**vysvětlovanou**), obvykle ji značíme **y**. Další proměnná nebo proměnné jsou považovány za nezávislé (**vysvětlující**). Statistika neurčí, která veličina je příčinou a která následkem, tedy která je nezávislá a která je závislá. To rozhodne (pokud je to vůbec možné) specifická věda, která se vztahem zabývá. Může to být například dáno tím, jak je veden pokus – pozorování (jednu veličinu vnějším zásahem měníme, druhá se dle toho mění).

Statistika sleduje pouze, zda existuje mezi veličinami vztah, že když se mění jedna veličina, mění se i druhá, a to takovým způsobem, že to nelze vysvětlit pouze náhodnými změnami této druhé veličiny. Proto se také používají raději pojmy vysvětlující veličina a vysvětlované veličina.

**Korelační úlohu** — určit **stupeň síly**, nebo také **průkaznost závislosti**, s jakou se předpokládaná závislost projevuje. Tedy zda změna vysvětlované (závislé) proměnné vyvolaná změnou proměnné vysvětlující (případně změnami více vysvětlujících proměnných – nezávislých) se prosadí proti změnám vysvětlované proměnné vzniklým náhodně (jsou způsobeny jinými, nesledovanými a náhodně se měnícími jevy), či nikoliv. To pochopitelně závisí nejen na chování vlastní závislosti, ale i na počtu naměřených výsledků a případně rozmezí měřených hodnot.



## 2. Regresní analýza

My se budeme zabývat pouze **jednoduchou regresí**, kdy hledáme předpokládaný vztah pouze mezi **dvěma veličinami**, obecně obvykle značenými  **$x$**  a  **$y$** . Jinak bychom museli použít maticový počet.

Provedeme pozorování obou veličin — změříme výsledky pokusu. Při něm volíme hodnoty jedné veličiny (nezávisle proměnné) označované obvykle  **$x$**  (ve statistice nazývané častěji jako vysvětlující veličina). Často nejde o volbu libovolných hodnot, ale o změření hodnot, které se v praxi vyskytly.

K těmto hodnotám proměřujeme objevující se hodnoty druhé (závislé proměnné) veličiny  **$y$**  (statisticky je to veličina vysvětlovaná). Tak získáme určitý počet (výběr z dvourozměrného rozdělení) spárovaných hodnot  $[x_i; y_i]$ , což jsou body v rovině.

Hodnoty veličiny nezávislé (vysvětlující) známe obvykle velmi přesně, což je jedna z podmínek klasické regrese.

Hodnoty naměřené veličiny (vysvětlující) jsou nahodilými vlivy vychýleny více či méně od závislosti, kterou předpokládáme. Tyto nahodilé výchylky mohou být vyvolány tím, že hodnoty  **$y$**  mohou být ovlivňovány dalšími faktory (nejen veličinou  **$x$** ), které se během měření náhodně měnily (například teplota vzduchu, sluneční záření, síla větru, apod.).

Pokud jsme korelační analýzou prokázali, že závislost mezi veličinami je statisticky významná, tedy že změny veličiny  **$y$**  svázané (sledovanou závislostí) se změnou veličiny  **$x$**  jsou tak velké, že se neztrácejí ve změnách vyvolaných náhodnými faktory, má smysl metodami regresní analýzy hledat matematické vyjádření této závislosti. Zvolený matematický tvar (**regresní funkce**) sledované závislosti však obsahuje neznámé parametry. Úkolem regresní analýzy je stanovení hodnot parametrů této závislosti.

Regresní metody se snaží odstranit vliv náhodných výchylek naměřených hodnot  **$y_i$**  a získanými body proložit regresní funkci tak, aby došlo k vyrovnání těchto nahodilých chyb měření.

Statistická indukce nás vede k představě, že existují „jediné skutečné“ hodnoty konstant regresní funkce, které platí pro základní soubor (populaci), tedy pro všechny možné naměřené páry hodnot. To jsou hledané parametry regresní funkce — regresní koeficienty. My však můžeme určit pouze **výběrové re-**

**gresní koeficienty**, kterými tyto parametry odhadujeme. Tyto výběrové regresní koeficienty budou pro opakované výběry nabývat různých hodnot, které jsou náhodně rozloženy kolem hledaných parametrů základního souboru. Existuje tedy pravděpodobnostní rozdělení možných hodnot výběrového regresního koeficientu s určitou střední hodnotou a určitou směrodatnou odchylkou tohoto parametru, kterou nazýváme také **standardní chyba**.

Odchyly naměřených hodnot od prokládané regresní funkce ale nemusejí být způsobeny jen chybami měření veličiny  $y$ . Podílí se na nich i naše případná chybná volba regresní funkce (chyba modelu), která nemusí plně odpovídat skutečnému (přirozenému) průběhu závislosti. Například zkoumaná závislost je vyjádřena hyperbolou namísto námi prokládané přímky.

**Nejčastěji používané regrese** (rovnice stochastického vztahu mezi veličinami):

- lineární (přímková) regrese:  $f(x) \equiv y = a + b \cdot x$
- kvadratická (parabolická) regrese:  $f(x) \equiv y = a + b \cdot x + c \cdot x^2$
- polynomiální stupně  $p$ :  $f(x) \equiv y = a + b_1 \cdot x + b_2 \cdot x^2 + \dots + b_p \cdot x^p$
- hyperbolická regrese:  $f(x) \equiv y = a + \frac{b}{x}$
- logaritmická regrese:  $f(x) \equiv y = a + b \cdot \log x$
- exponenciální regrese:  $f(x) \equiv y = a \cdot b^x$

Uvedené parametry ( $a$ ,  $b$ ,  $c$ ,  $b_i$ ), neboli výběrové regresní koeficienty jak jsme již uvedli výše, jsou střední hodnoty pravděpodobnostních rozdělení všech možných hodnot určených z výběrů. Jsou to tedy

konstanty, (**střední hodnota** je číslo) neměnná čísla, které ovšem nemůžeme nikdy určit přesně. Můžeme pouze z hodnot výběru určit jejich **bodové** odhady, případně určit **intervalové** odhady tak, jak jsme si ukazovali v kapitole o **statistické indukci**. Ze získaného náhodného výběru dvojic pak určíme (empirickou) **výběrovou regresní funkci**  $f(x) = E(Y(x))$ , která je jedním z možných odhadů hledané regresní funkce. Pro každou hodnotu  $x_i$  tak budeme mít dvě hodnoty (konkrétní čísla) závisle proměnné  $Y$ :

- jednak získanou (empirickou) hodnotu  $y_i$ ,
- jednak vyrovnanou hodnotu  $f(x_i)$ , což je odhad (teoretické) střední hodnoty  $E(Y)$  /kterou ovšem neznáme/ celého základního souboru.

Jejich rozdíly  $[f(x_i) - y_i]$  nazýváme odchylky (rezidua). Jsou to vlastně odhady chyb.

Bodové odhady regresních koeficientů nejčastěji získáváme **metodou nejmenších čtverců**<sup>43</sup>. Tato metoda nejmenších čtverců vychází z požadavku, aby **součet čtverců** (druhých mocnin) odchylek pozorovaných hodnot  $y_1, y_2, \dots, y_n$  od odhadované regresní funkce  $f(x)$  **byl minimální** (veškeré chyby modelu přeneseme do svislého směru osy  $y$ ), tedy:

$$S = \sum_{i=1}^n [f(x_i) - y_i]^2 \rightarrow \min. \quad (23)$$

Z kurzu matematické analýzy (konkrétně z kapitoly o diferenciálním počtu) víme, že extrém funkce (a minimum je extrém) může nastat pouze tam, kde:

první derivace dané funkce neexistuje,

nebo první derivace dané funkce existuje a je rovna NULE.

Budeme tedy vztah (23) derivovat:

<sup>43</sup> Metodu zavedl francouzský matematik **Adrien-Marie Legendre** již počátkem 19. století. Vyžaduje znalost diferenciálního počtu, který je náplní předmětu matematika.

$$S' = \left\{ \sum_{i=1}^n [f(x_i) - y_i]^2 \right\}' \quad \text{derivace součtu se rovná součtu derivací}$$

$$S' = \sum_{i=1}^n \{ [f(x_i) - y_i]^2 \}' \quad \text{derivujeme složenou funkci; } y_i \text{ je daná hodnota – konstanta}$$

$$\begin{aligned} S' &= \sum_{i=1}^n 2 \cdot [f(x_i) - y_i]^{2-1} \cdot [f(x_i) - y_i]' = 2 \cdot \sum_{i=1}^n [f(x_i) - y_i] \cdot [f'(x_i) - y_i'] = \\ &= 2 \cdot \sum_{i=1}^n [f(x_i) - y_i] \cdot [f'(x_i) - 0] = 2 \cdot \sum_{i=1}^n [f(x_i) - y_i] \cdot f'(x_i) \end{aligned}$$

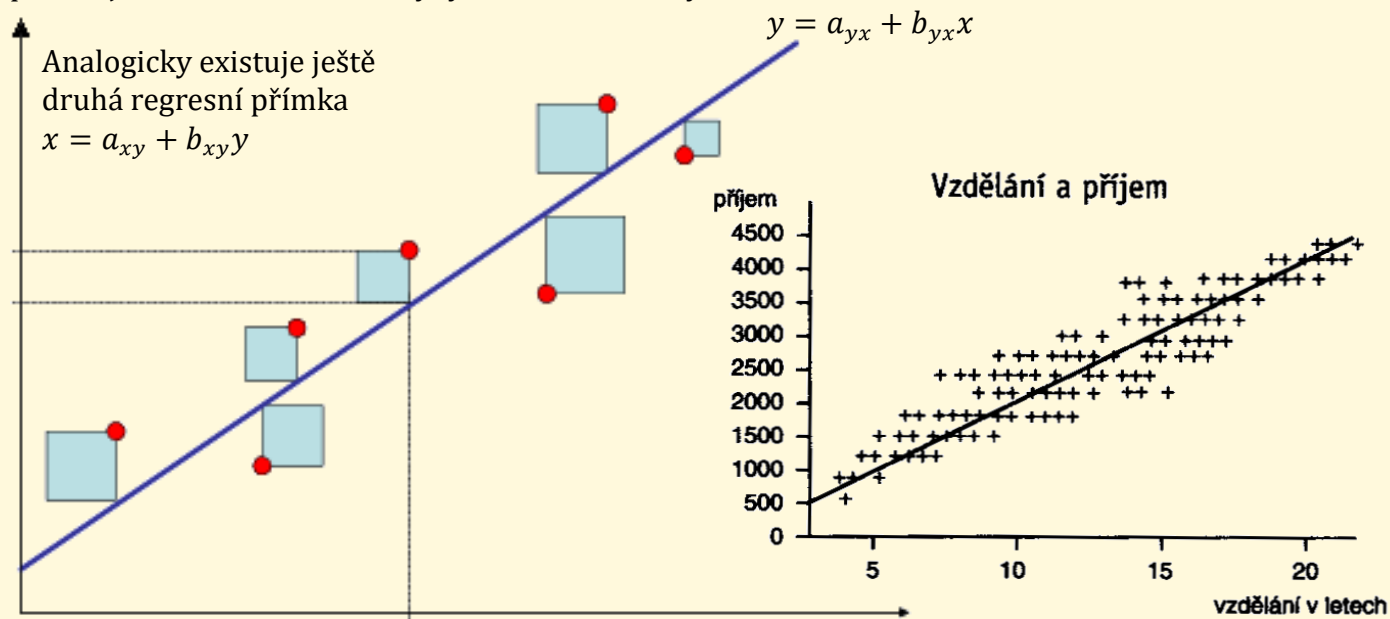
Další postup derivování závisí na tvaru regresní funkce  $f(x)$ . Výslednou derivaci (v případě *parciálních derivací* je jich více a dostáváme systém rovnic) pak položíme rovnu nule a hledáme řešení dané rovnice.

### Předpoklady metody nejmenších čtverců

- Chyby nezávislé veličiny  $X$  mají být relativně menší než chyby závislé veličiny  $Y$ . V opačném případě je pro správný odhad potřeba použít jinou metodu.
- Chyby hodnot veličiny  $Y$  mají mít **normální rozdělení** s nulovou střední hodnotou a s konstantním rozptylem (a tedy i konstantní směrodatnou odchylkou). To znamená, že se rozptýlení hodnot nesmí měnit podle velikosti hodnot  $y_i$  (např. u malých hodnot  $y$  nemají být chyby menší než u hodnot velkých). Dále tyto chyby nemají být vzájemně závislé. Na grafu mají být tedy naměřené body rovnoměrně rozptýleny kolem proložené regresní křivky bez zjevných tendencí (například v růstu) a se zhruba stejným počtem bodů nad a pod křivkou.
- Přítomnost jediného vychýleného bodu v datech může způsobit překvapivě velké vychýlení odhadů při použití metody nejmenších čtverců. Takovýto bod strhává proložení regresní křivky výrazně na svoji stranu (viz **obrázek**) a je třeba jej případně vyloučit.

## 2.1. Regresní přímka — lineární regrese $f(x) : y = a + b x$

Parametr ***b*** se také nazývá **regresní koeficient** a říká o kolik jednotek průměrně vzroste příjem (pravý obrázek), když vzdělání vzroste o jeden rok. Z pohledu geometrie je to směrnice regresní přímky, u které požadujeme minimalizovat chyby náhodné veličiny *Y*.



Hledáme minimum (23) funkce  $\sum_{i=1}^n [a + b \cdot x_i - y_i]^2$  tak, že **parciální derivace podle** proměnných ***a*, *b*** (různé přímky se odlišují právě jenom proměnnými parametry *a*, *b* a my hledáme takové hodnoty těchto parametrů/proměnných, aby součet čtverců chyb byl minimální) **položíme rovny nule** (zadané

body  $[x_i; y_i]$  jsou v levém obrázku označeny červenými kolečky; jejich souřadnice jsou tedy čísla – a pro derivování jsou to konstanty)

$$2 \cdot \sum_{i=1}^n [(a + b \cdot x_i - y_i) \cdot 1] = 0$$

$$2 \cdot \sum_{i=1}^n [(a + b \cdot x_i - y_i) \cdot x_i] = 0$$

což vede na následující **soustavu normálních rovnic** (kde:  $y = a + b \cdot x = a \cdot 1 + b \cdot x = a \cdot x^0 + b \cdot x = y$

a  $\sum_{i=1}^n x_i^0 = \sum_{i=1}^n 1 = n$ ) a sumační meze kvůli přehlednosti již vynecháme:

$$\begin{aligned} a \cdot \sum x_i^0 + b \cdot \sum x_i &= \sum y_i \\ a \cdot \sum x_i + b \cdot \sum x_i^2 &= \sum (x_i \cdot y_i) \end{aligned}$$

Tuto soustavu můžeme řešit mnoha způsoby (např. **Cramerovým pravidlem**), protože má jediné řešení. Po obecném vyřešení (a náročnějších úpravách) dostáváme tuto podobu **rovnice regresní přímky**:

$$y - \bar{y} = \frac{\text{cov}(X, Y)}{(x^2) - (\bar{x})^2} \cdot (x - \bar{x}) \quad \text{nebo jinak} \quad f(x) : y = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{(x^2) - (\bar{x})^2} \cdot (x - \bar{x}) + \bar{y} \quad (24)$$

kde: **pruh** označuje **aritmetický průměr** a  $\text{cov}(X, Y)$  je **výběrová kovariance** náhodných veličin **X** a **Y**.

Použijeme-li „S“kové varianty vestavěných funkcí, můžeme pomocí /Excelu 2010/ rovnicí regresní přímky sestavit následovně:

$$f(x) : y = \frac{/=COVARIANCE.S(X;Y)/}{/=VAR.S(X)/} \cdot (x - /=PRŮMĚR(X)/) + /=PRŮMĚR(Y)/ \quad (25)$$

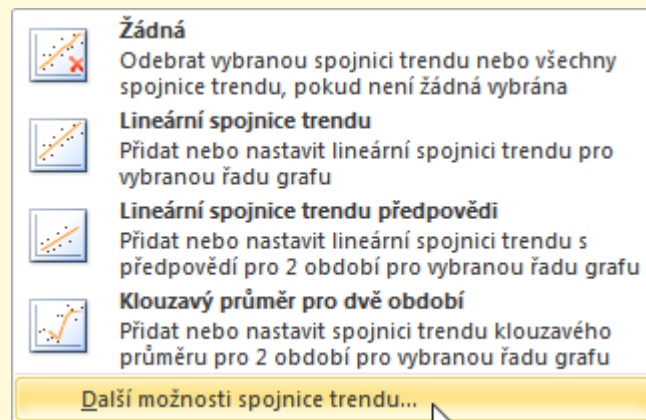
## Poznámka k metodě nejmenších čtverců

Uvedená metoda je v praxi natolik používána, že jak některé komerční programy (například **Excel**, Mathematica, Matlab, MathCad, ...) tak jejich freewarové alternativy (například GNUplot) hledají aproximační funkce pouze na základě námi zadaných diskrétních bodů. Vše ostatní již provádějí samostatně, bez našeho přičinění.

Konkrétně v programu **Excel 2010** postupujeme následovně:

1. Zadané hodnoty označíme jako blok.
2. Potom na kartě [**Vložení**] v oblasti „**Grafy**“ vybereme <**Bodový**>
3. Nakonec na kartě [**Nástroje grafu**] v záložce „**Rozložení**“ v oblasti <**Analýza**> a položce „**Spojnice trendu**“ vybereme [**Další možnosti spojnice trendu**]

	A	B	C
1			
2		-1	1
3		0	2
4		2	-2
5		3	-7
6			




Po případném dalším upřesnění (například jaká má být barva čar, zda požadujeme v grafu vypisovat výslednou rovnici /v levém obrázku druhá volba od spodu/, ...) se již vykreslí požadovaný graf.


Možnosti spojnice trendu


Barva čáry
Styl čáry
Stín
Záře a měkké okraje


### Možnosti spojnice trendu


Typ trendu a regrese



☐ Exponenciální


☒ Lineární


☐ Logaritmický


☐ Polynomický Pořadí: 2


☐ Mocninový


☐ Kluzavý průměr Období: 2

Název spojnice trendu

☒ Automaticky: Lineární (Řady1)

☐ Vlastní:

Odhad

Vpřed: 0,0 období

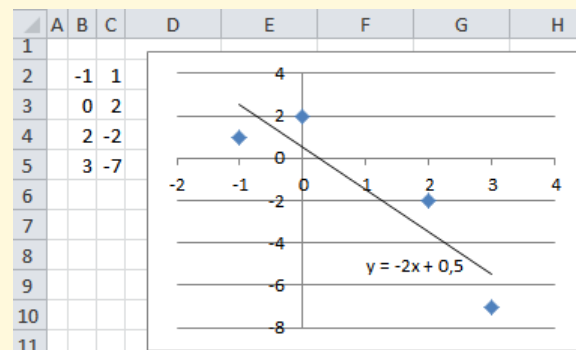
Nazpět: 0,0 období

☐ Hodnota  $\bar{y}$  = 0,0

☒ Zobrazit rovnici v grafu

☐ Zobrazit hodnotu spolehlivosti R

Zavřít





## 2.2. Regresní parabola — kvadratická regrese $f(x) : y = a + b x + c x^2$

má následující **soustavu normálních rovnic**:

$$\begin{aligned} a \cdot \sum x_i^0 + b \cdot \sum x_i + c \cdot \sum x_i^2 &= \sum y_i \\ a \cdot \sum x_i + b \cdot \sum x_i^2 + c \cdot \sum x_i^3 &= \sum (x_i \cdot y_i) \\ a \cdot \sum x_i^2 + b \cdot \sum x_i^3 + c \cdot \sum x_i^4 &= \sum (x_i^2 \cdot y_i) \end{aligned}$$

### Poznámka

- Uvedená soustava normálních rovnic má vždy regulární matici soustavy, to znamená, že vždy existuje jediné řešení dané soustavy. Proto můžeme využít libovolnou metodu pro hledání řešení soustavy rovnic. Třeba Crammerovo pravidlo, kdy si jednotlivé determinanty necháme spočítat například *Excelem 2010*: =DETERMINANT(matice).
- Uvedenou soustavu normálních rovnic můžeme **formálně** sestavit také tak, že si vezmeme rovnici paraboly (z nadpisu) a jenom ji napíšeme v jiném pořadí a s indexy ( $1 = x_i^0 \Rightarrow$  **první** rovnice).

$$a \cdot x_i^0 + b \cdot x_i^1 + c \cdot x_i^2 = y_i \quad | \cdot ( ) \quad \Rightarrow \quad \sum_{i=1}^n [a \cdot x_i^0 + b \cdot x_i^1 + c \cdot x_i^2] = \sum_{i=1}^n y_i$$

Když tuto rovnici vynásobíme výrazem  $x_i$ , dostaneme **druhou** rovnici.

A když ji vynásobíme výrazem  $x_i^2$ , dostaneme **třetí** rovnici.

Pak přidáme vždy k oběma stranám rovnic sumační symboly. Dále na levé strany aplikujeme asociativní zákon (o sdružování sčítanců) pro sčítání a konstanty vytkneme před dané sumy.

První člen první rovnice můžeme ještě upravit na jednodušší tvar.

$$\sum_{i=1}^n a \cdot x_i^0 = a \cdot \sum_{i=1}^n x_i^0 = a \cdot \sum_{i=1}^n 1 = a \cdot n$$

- Také kvadratické vyrovnaní *Excelem 2010* umí provést samostatně. Postup je stejný jako v [poznámce](#) k metodě nejmenších čtverců, jen na [obrázku](#) zvolíme trend **Polynomický** (Pořadí: 2).

## 2.3. Volba regresní funkce

Jak ale pouze ze zadaných dat poznat, kterou regresní funkci (ze dvou, které jsme si před chvílí uvedli) máme zvolit?

Někdy stačí nakreslit bodový graf (**korelační pole**), v němž je každá dvojice údajů graficky znázorněna jedním bodem v rovině (například [tyto dva grafy](#) a další dva následující grafy). A z polohy jednotlivých bodů se nám (někdy) povede určit vhodný typ regresní funkce. Jiné dvě možnosti určení vyhovující funkce si nyní ukážeme.

### 2.3.1. Lineární závislost:

Z rovnice přímky  $y = k \cdot x + q$  plyne, že pro stejné přírůstky (diference) nezávisle proměnné (jednoho znaku)  $X$  ( $x_i - x_{i-1} = \text{konst.}$ ) bychom měli mít (alespoň přibližně) **stejné přírůstky** (druhého znaku) závisle proměnné  $Y$  ( $\Delta_i^{(1)} = y_i - y_{i-1} = \text{konst.}$ ).

**Příklad.** Máme dáno těchto devět bodů: [1 ; -1], [2 ; 0,9], [3 ; 3], [4 ; 4,9], [5 ; 7], [6 ; 9,1], [7 ; 11], [8 ; 13], [9 ; 15,1]. Hodnoty si přepíšeme do následující tabulky, kterou doplníme o příslušné výpočty, včetně již spočítané regresní přímky.

$i$	1	2	3	4	5	6	7	8	9
$x_i$	1	2	3	4	5	6	7	8	9
$y_i$	-1	0,9	3	4,9	7	9,1	11	13	15,1
$\Delta_i^{(1)} = y_i - y_{i-1}$	/	1,9	2,1	1,9	2,1	2,1	1,9	2	2,1
$y_{P_i} = 2,015 \cdot x_i - 3,075$	-1,06	0,955	2,97	4,985	7	9,015	11,03	13,045	15,06
$\Delta_{y_i} = y_i - y_{P_i}$	0,06	-0,055	0,03	-0,085	0	0,085	-0,03	-0,045	-0,04

Vidíme, že  $\Delta_{y_i} \in \langle -0,085; 0,085 \rangle$ , tedy že zadané body skutečně „téměř perfektně“ leží na regresní přímce  $y = 2,015 \cdot x - 3,075$  a přitom námi zjištěné „přírůstky“  $\Delta_i^{(1)} \in \langle 1,9; 2,1 \rangle$ .

**První problém.** Uvedené tvrzení však skutečně platí pouze za předpokladu, že jednotlivé hodnoty  $x_i$  jsou **ekvidistantní** (následující hodnota je vždy „stejně“ vzdálena od předchozí hodnoty).

Protože, když z předchozích devíti bodů, které leží „téměř“ na přímce  $y = 2,015x - 3,075$  vynecháme dva body (například **třetí** a **šestý**), polohu ostatních bodů tím nezměníme. Tedy zbylých sedm bodů musí opět „téměř“ ležet na stejné přímce. Nám ale, jak plyne z následující tabulky, „téměř konstantní“  $\Delta_i^{(1)}$  nevychází.

$i$	1	2	3	4	5	6	7
$x_i$	1	2	4	5	7	8	9
$y_i$	-1	0,9	4,9	7	11	13	15,1
$\Delta_i^{(1)} = y_i - y_{i-1}$	/	1,9	4	2,1	4	2	2,1

Zkusme rozdíl  $\Delta_i^{(1)}$  uvažovat s vahou rovnou velikosti rozdílu  $x_i - x_{i-1}$ , tedy  $\Delta_i^{(1)} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$ .

Výpočty opět zapíšeme do následující tabulky:

$x_i$	1	2	4	5	7	8	9
$y_i$	-1	0,9	4,9	7	11	13	15,1
$\Delta_i^{(1)} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$	/	1,9	2	2,1	2	2	2,1

Vidíme, že nyní je opět  $\Delta_i^{(1)} \doteq 2$ .

**Druhý problém.** A co se stane, když bude dáno těchto devět bodů:  $[1; -1]$ ,  $[2; 0,9]$ ,  $[3; 3]$ ,  $[4; 4,9]$ ,  $[5; 7]$ ,  $[6; 9,1]$ ,  $[7,5; 13]$ ,  $[7,5; 11]$ ,  $[9; 15,1]$ , kde sedmý a osmý bod mají stejnou hodnotu  $x$ ?

Jaký bude rozdíl  $\Delta_7^{(1)}$  od předchozího (šestého) bodu?

Bude to  $\Delta_7^{(1)} = 13 - 9,1$  nebo  $\Delta_7^{(1)} = 11 - 9,1$ ?

A co když budeme chtít určit **vážený** rozdíl  $\Delta_8^{(1)} = \frac{y_8 - y_7}{x_8 - x_7}$ ? Ve jmenovateli zlomku by byla NULA a my víme, že nulou dělit NELZE!

V tomto případě sedmý a osmý bod nahradíme jedním bodem, jehož hodnota  $y$  je „**někde mezi**“ hodnotou sedmého a osmého bodu, tedy je to nějaký z průměrů hodnot. Vhodným kandidátem je aritmetický průměr, takže dostáváme následující tabulku:

$i$	1	2	3	4	5	6	7	8
$x_i$	1	2	3	4	5	6	7,5	9
$y_i$	-1	0,9	3	4,9	7	9,1	12	15,1
$\Delta_i^{(1)} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$	/	1,9	2,1	1,9	2,1	2,1	1,933	2,067

 $\Delta_i^{(1)} \doteq 2$ 

Pro námi zjištěné **vážené přírůstky** (a když jsme příslušné body, které pro stejná  $x$  mají různá  $y$  vhodným způsobem „zprůměrovali“) platí, že:  $\Delta_i^{(1)} \in \langle 1,9; 2,1 \rangle$ .

### 2.3.2. Kvadratická závislost:

Pro stejné přírůstky nezávisle proměnné  $X$  ( $x_i - x_{i-1} = \text{konst.}$ ) bychom měli mít stejné přírůstky přírůstků  $\Delta_i^{(2)}$  závisle proměnné  $Y$  ( $\Delta_i^{(2)} = \Delta_i^{(1)} - \Delta_{i-1}^{(1)} = \text{konst.}$ , kde  $\Delta_i^{(1)} = y_i - y_{i-1}$ )

#### Příklad.

$x_i$	1	2	3	4	5	6	7	8	9
$y_i$	16,1	9	4,1	1	0,1	1,1	4	9,1	16
$\Delta_i^{(1)} = y_i - y_{i-1}$	/	-7,1	-4,9	-3,1	-0,9	1	2,9	5,1	6,9
$\Delta_i^{(2)} = \Delta_i^{(1)} - \Delta_{i-1}^{(1)}$	/	/	2,2	1,8	2,2	1,9	1,9	2,1	1,8

$$y = 0,054 x^2 + 0,446 x + 40,693$$

**Poznámka.** I pro určení, zda se jedná o kvadratickou závislost platí analogické podmínky jako jsme ukázali u lineární závislosti: **ekvidistantní**  $x_i$ , kde **pro každé**  $x_i$  **je dáno jediné**  $y_i$ . Pokud tyto podmínky nejsou splněny a my chceme použít předchozí postup, musíme nějak zajistit, aby tvrzení platilo (jako jsme to naznačili při řešení předchozích dvou problémů).

### 3. Korelační analýza — výběrový korelační koeficient

Druhým základním úkolem statistické analýzy vztahů mezi náhodnými veličinami je určení těsnosti závislosti – korelace (souvztažnosti). Zatímco regresní analýza se zaměřuje na formu vztahu mezi sledovanými veličinami, korelační analýza ukazuje, jak je tento vztah silný.

Východiskem pro měření těsnosti závislosti je příslušný regresní model. Znalost intenzity závislosti mezi analyzovanými veličinami je užitečná zejména z těchto důvodů:

- Je zřejmé, že čím jsou určité veličiny těsněji vázány, s tím větší pravděpodobností lze očekávat, že změny jedné veličiny budou mít za následek změny veličiny s ní statisticky vázané.
- Stupeň vázanosti náhodných veličin charakterizuje, jaká je vypovídací schopnost užitého regresního modelu. Čím bude rozptýl empirických hodnot závisle proměnné kolem příslušné regrese menší (a tedy závislost těsnější), tím budou regresní odhady, založené na dané regresní funkci, přesnější.

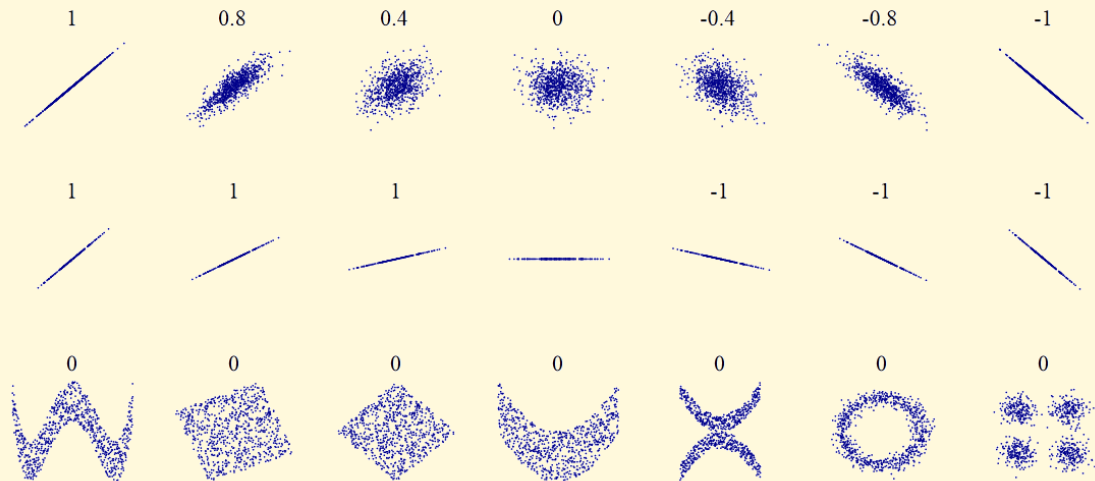
Těsnost závislosti je možno měřit pomocí řady charakteristik [13]. My si uvedeme jedinou – **výběrový korelační koeficient** pro případ **lineární závislosti** mezi dvěma proměnnými, kdy  $S(X) \cdot S(Y) \neq 0$ . S **korelačním koeficientem**  $\rho$  jsme se setkali u náhodných vektorů.

$$r = \frac{\sum (x_i \cdot y_i) - \frac{1}{n} \cdot \sum x_i \cdot \sum y_i}{\sqrt{\left[ \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \right] \cdot \left[ \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \right]}} = \frac{S(X; Y)}{\sqrt{S^2(X) \cdot S^2(Y)}} = \frac{S(X; Y)}{S(X) \cdot S(Y)} = \sqrt{b_{yx} \cdot b_{xy}}$$

Zatímco **regresní koeficient**  $b$  (což je vlastně směrnice regresní přímky) nám naznačuje, CO máme hádat, korelační koeficient  $r$  nám říká, JAK DOBŘE budeme schopni hádat. Pokud vyjdeme z **(menšího) pravého** obrázku, můžeme říci, že výběrový korelační koeficient (pro přímku) umocněný na druhou ( $r^2$  nazýváme koeficientem determinace, který je roven součinu směrnic *sdužených přímek*, kdy jedna je metodou

nejmenších čtverců stanovena pro minimální odchylky ve vodorovném směru osy  $x$  a druhá pro minimální odchylky ve svislém směru osy  $y$ ) poskytuje informaci, jaké procento rozdílů existujících v příjmu se zdá být vysvětlitelné rozdíly, které existují ve vzdělání.

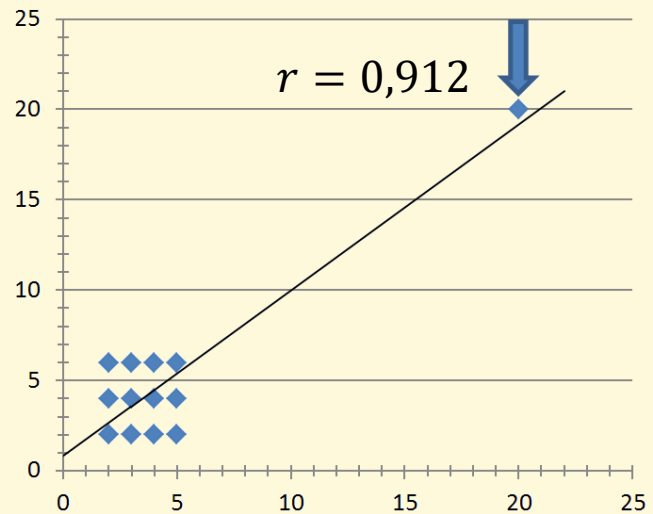
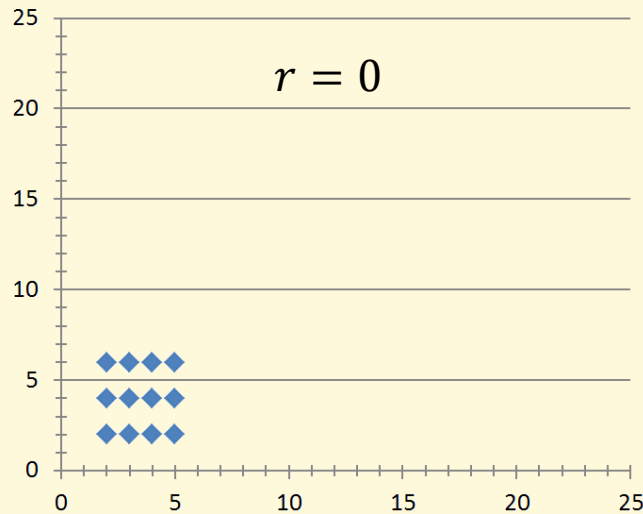
Obrázek 5: Zdroj WIKIPEDIE



Několik příkladů grafického zobrazení naměřených dat a jejich koeficienty korelace  $r$ .

***I při nulovém korelačním koeficientu ( $r = 0$ ) na sobě veličiny mohou záviset***, pouze tento vztah nelze vyjádřit lineární funkcí, a to ani přibližně (spodní řada obrázků). Stanovit stupnici oceňující závislost (*slabá*, *střední*, *silná*) není úkol pro matematiku, ale pro profesního odborníka. Podobné stupnice bývají součástí oborových norem.

## Příklad 1.



V levém grafu vidíme jednoduché seskupení 12 pozorování  $[(2; 2), (2; 4), (2; 6), (3; 2), \dots, (5; 6)]$ . Je zřejmé, že symbolizují perfektní nezávislost ( $\Rightarrow r = 0$ ), protože bez ohledu na hodnotu proměnné  $X$  může proměnná  $Y$  nabývat pouze hodnot **2, 4** nebo **6**.

A teď se podívejme, co se stane, když k našim 12 pozorováním přidáme jedno další  $[20; 20]$ , s vysokými hodnotami obou proměnných. V pravém grafu je toto přidané pozorování označeno (pro přehlednost) tlustou šipkou. Podívejte se teď na novou hodnotu korelačního koeficientu. Korelace je téměř perfektní.

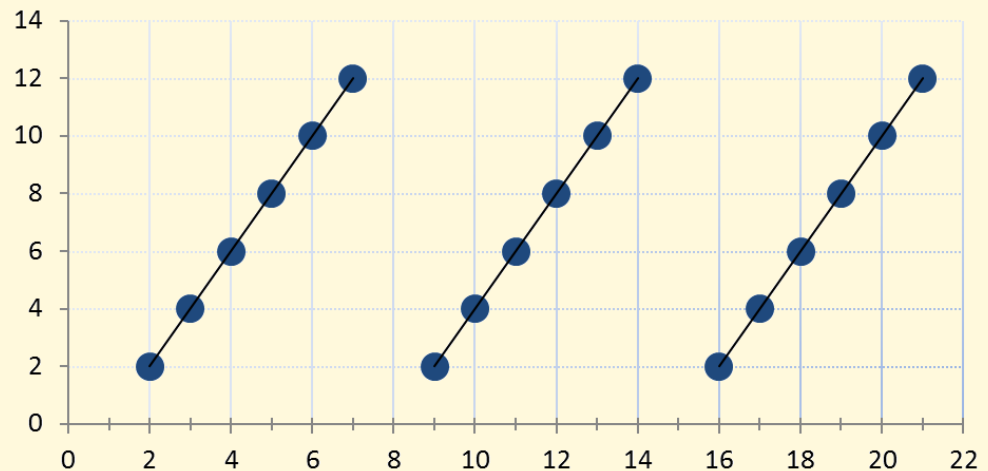
Co vlastně způsobilo toto jedno jediné další pozorování? Prostě velice podstatně zvětšil rozptyl našeho vzorku. Matematicky je tu všechno v pořádku. Víme, že kvadrát korelačního koeficientu odpovídá proporci rozptylu závisle proměnné, kterou je možné vysvětlit rozdíly hodnot druhé proměnné.



Ne tak docela v pořádku je interpretace dat. Téměř všechny rozptyly byl vnesen do našeho vzorku tímto jediným, novým pozorováním. Ta velká vysvětlující síla  $r^2$  se týká jenom tohoto nového pozorování ve vztahu ke zbytku pozorování. Vůbec nám nepomůže k lepšímu porozumění vztahu v jádru našeho vzorku, v původních našich 12 pozorování.

## Příklad 2.

A co těchto 18 pozorování,  
pro které  $r = 0,286$ ?



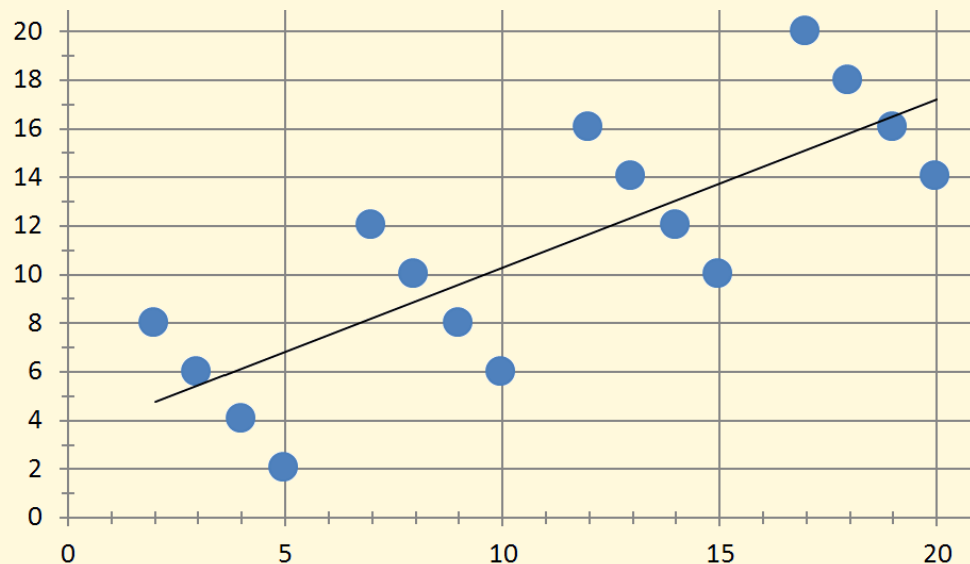
Jistě jste si všimli, že data mají zajímavou konfiguraci, kterou můžeme dobře využít. Rozdělíme prostě náš původní vzorek podle hodnot nezávisle proměnné  $X$  do tří částečných vzorků.

V prvním částečném vzorku budou všechna pozorování, která mají hodnotu  $X$  z intervalu  $\langle 2; 7 \rangle$ ;  
ve druhém vzorku budou všechna pozorování, která mají hodnoty  $X$  z intervalu  $\langle 9; 14 \rangle$ ;  
a ve třetím budou všechna pozorování, která mají hodnoty  $X$  z intervalu  $\langle 16; 21 \rangle$ .

Na první pohled vidíme, že v každém částečném vzorku leží všechna pozorování přesně na přímkce, tedy že v každém částečném vzorku existuje perfektní souvislost mezi  $X$  a  $Y$ . Tím jsme si ukázali jednu velice důležitou věc, Korelační koeficient je lineární a jeho hodnota udává, jak moc je vhodné charakterizovat všechny pozorované hodnoty jedinou přímkou. V některých případech (**částečné vzorky** v předchozím [grafu](#)) je lineární reprezentace výborná. Jindy (**celý vzorek** v předchozím [grafu](#)) může takový lineární model ztratit důležitou část informace.

### Příklad 3.

A co těchto 16 pozorování,  
pro která  $r = 0,789$ ?



Konfigurace dat ukazuje, že v celém souboru existuje celkem dosti **silný pozitivní** (kladný) **vztah** mezi proměnnými  $X$  a  $Y$ . Naproti tomu v každém podsouboru můžeme pozorovat **perfektní negativní** (zápornou) **souvislost**.

### 3.2. Test významnosti hodnoty korelačního koeficientu $r$

Jak již víme, korelační koeficient základního souboru  $\varrho$  má hodnotu nula, když není mezi veličinami **lineární** závislost. Jestliže tedy statisticky prokážeme, že se vypočtená hodnota výběrového korelačního koeficientu  $r$  významně liší od nuly, prokážeme tím, že mezi veličinami  $X$  a  $Y$  je lineární závislost.

Tedy podle **postupu**, který byl uveden v kapitole zabývající se testováním hypotéz, testujeme nulovou hypotézu  $H_0 : \varrho = 0$  – mezi zkoumanými veličinami neexistuje lineární závislost proti alternativní hypotéze  $H_A : \varrho \neq 0$  – lineární závislost existuje.

Pro danou hladinu významnosti zvolíme testové kritérium a pro naměřené dvojice  $[x_i ; y_i]$  vypočítáme pozorovanou hodnotu testové statistiky. Poté určíme kritický obor (obor přijetí hypotézy) a rozhodneme, zda testová statistika leží v kritickém oboru nebo v oboru přijetí.

V literatuře jsou pro prokazování významnosti  $r$  předepisovány různé testovací statistiky.

#### Poznámky ke korelační analýze

1. S rostoucím počtem sledovaných bodů sice většinou klesá hodnota výběrového korelačního koeficientu  $r$ , ale stále se (limitně) přibližuje hodnotě korelačního koeficientu populace  $\varrho$ .

Máme-li pouze dvě pozorování, najdeme vždy přímku (přímka je určena dvěma body), která oběma body prochází a to bez nevysvětlitelných odchylek. Ve vzorku tedy dostáváme perfektní lineární závislost, i když v celé populaci mezi zkoumanými veličinami **žádná** (a tím pádem ani lineární) závislost nemusí vůbec existovat.

Když přidáme další (třetí) pozorování, přímka již nemusí všemi třemi body procházet, takže korelační koeficient se již nerovná nule, ale je stále vysoký.

Čím větší bude počet naměřených bodů, tím větší bude možnost nalezení případné závislosti, i třeba v bodech široce rozptýlených kolem přímky, kdy závislost je slabá, tedy i pro případy nízkých (blízkých nule) korelačních koeficientů.

2. Při korelační analýze (hledání, zda existuje významná přímková závislost) jediný bod vzdálený (odlehlý) od ostatních může zajistit nalezení významné korelace, ač zbylé body (bez tohoto odlehlého) mohou vykazovat naprostou nezávislost mezi sledovanými veličinami — viz [obrázek](#).

Jediný vzdálený (možná problematický) bod zajistí hodnotu korelačního koeficientu překračující kritickou hodnotu. V takovém případě nelze brát výsledek testu významnosti hodnoty korelačního koeficientu příliš vážně, protože rozdělení bodů zřejmě nevyhovuje předpokladům nutným pro platnost použitého testu.

## 4. Příklad

K dispozici jsou tato data o prodeji (druhý řádek), jak je ovlivňovaly náklady na reklamu (první řádek):

x	0	1	2	3	4	5
y	40	42	43	41	43	44,8

Určete rovnici [lineární regrese](#), rovnici [kvadratické regrese](#) a [výběrový korelační koeficient](#) (těsnost vztahu pro lineární regresi) pro těchto šest dvojic hodnot  $[x_i; y_i]$ , kde  $i = 1, 2, \dots, 6$ .

**Lineární regrese 1.** Pro lineární regresi vyjdeme ze vztahu (25) a nejprve necháme *Excel 2010* spočítat všechny potřebné hodnoty. Uvedenou tabulku přepíšeme do Excelu a vyvoláme příslušné vestavěné statistické funkce.

	B	C	D	E	F	G	H
2	x	0	1	2	3	4	5
3	y	40	42	43	41	43	44,8
4							
5	2,50	=COVARIANCE.S(C2:H2;C3:H3)					
6	3,50	=VAR.S(C2:H2)					
7	2,50	=PRŮMĚR(C2:H2)					
8	42,30	=PRŮMĚR(C3:H3)					

Potom již stačí dosadit získané hodnoty do vztahu (25) a obdržíme hledanou rovnici regresní přímky.

$$y = \frac{\text{cov}(X, Y)}{S_x^2} \cdot (x - \bar{x}_A) + \bar{y}_A = \frac{2,5}{3,5} \cdot (x - 2,5) + 42,3 \Rightarrow y \doteq 0,714x + 40,514$$

**Lineární regrese 2.** A co v situaci, kdy nemáme po ruce vhodný softwarový nástroj? Nezbyvá nám, než si příslušné charakteristiky spočítat. Regresní přímka potom bude mít (podle 24) rovnici:

$$y = \frac{\overline{x \cdot y}_A - \bar{x}_A \cdot \bar{y}_A}{\overline{(x^2)}_A - (\bar{x}_A)^2} \cdot (x - \bar{x}_A) + \bar{y}_A$$

Vidíme, že potřebujeme

$$\overline{(x \cdot y)}_A \quad \bar{x}_A \quad \bar{y}_A \quad \overline{(x^2)}_A$$

což určíme tak, že tabulku přepíšeme svisle a doplníme vhodnými sloupci.

$$y = \frac{\overline{x \cdot y_A} - \bar{x}_A \cdot \bar{y}_A}{(\overline{x^2})_A - (\bar{x}_A)^2} \cdot (x - \bar{x}_A) + \bar{y}_A$$

i	$x_i$	$y_i$	$x_i \cdot y_i$	
1	0	40		
2	1	42		
:	2	43		
n=6	3	41		
:	4	43		
6	5	44,8		
$\Sigma$				

Po dosazení:

$$\Rightarrow y \doteq$$

$$y = \frac{\overline{x \cdot y}_A - \bar{x}_A \cdot \bar{y}_A}{(\overline{x^2})_A - (\bar{x}_A)^2} \cdot (x - \bar{x}_A) + \bar{y}_A$$

i	$x_i$	$y_i$	$x_i \cdot y_i$	$x_i^2$
1	0	40	0	
2	1	42	42	
:	2	43	86	
n=6	3	41	123	
:	4	43	172	
6	5	44,8	224	
$\Sigma$	15	253,8		

$$\bar{x}_A = \frac{15}{6} = 2,5$$

$$\bar{y}_A = \frac{253,8}{6} = 42,3$$

Po dosazení:

$$\Rightarrow y \doteq$$

$$y = \frac{\overline{x \cdot y}_A - \bar{x}_A \cdot \bar{y}_A}{(\overline{x^2})_A - (\bar{x}_A)^2} \cdot (x - \bar{x}_A) + \bar{y}_A$$

i	$x_i$	$y_i$	$x_i \cdot y_i$	$x_i^2$
1	0	40	0	0
2	1	42	42	1
:	2	43	86	4
n=6	3	41	123	9
:	4	43	172	16
6	5	44,8	224	25
$\Sigma$	15	253,8	647	

$$\bar{x}_A = \frac{15}{6} = 2,5$$

$$\bar{y}_A = \frac{253,8}{6} = 42,3$$

$$\overline{(x \cdot y)}_A = \frac{647}{6} \doteq 107,833$$

Po dosazení:

$$\Rightarrow y \doteq$$



$$y = \frac{\overline{x \cdot y}_A - \bar{x}_A \cdot \bar{y}_A}{(\overline{x^2})_A - (\bar{x}_A)^2} \cdot (x - \bar{x}_A) + \bar{y}_A$$

i	$x_i$	$y_i$	$x_i \cdot y_i$	$x_i^2$
1	0	40	0	0
2	1	42	42	1
:	2	43	86	4
n=6	3	41	123	9
:	4	43	172	16
6	5	44,8	224	25
$\Sigma$	15	253,8	647	55

$$\bar{x}_A = \frac{15}{6} = 2,5$$

$$\bar{y}_A = \frac{253,8}{6} = 42,3$$

$$\overline{(x \cdot y)}_A = \frac{647}{6} \doteq 107,833$$

$$\overline{(x^2)}_A = \frac{55}{6} \doteq 9,167$$

Po dosazení:

$$\Rightarrow y \doteq$$

$$y = \frac{\overline{x \cdot y}_A - \bar{x}_A \cdot \bar{y}_A}{(\overline{x^2})_A - (\bar{x}_A)^2} \cdot (x - \bar{x}_A) + \bar{y}_A$$

i	$x_i$	$y_i$	$x_i \cdot y_i$	$x_i^2$
1	0	40	0	0
2	1	42	42	1
:	2	43	86	4
n=6	3	41	123	9
:	4	43	172	16
6	5	44,8	224	25
$\Sigma$	15	253,8	647	55

$$\bar{x}_A = \frac{15}{6} = 2,5$$

$$\bar{y}_A = \frac{253,8}{6} = 42,3$$

$$\overline{(x \cdot y)}_A = \frac{647}{6} \doteq 107,833$$

$$(\overline{x^2})_A = \frac{55}{6} \doteq 9,167$$

Po dosazení: 
$$y = \frac{107,833 - 2,5 \cdot 42,3}{9,167 - 2,5^2} \cdot (x - 2,5) + 42,3 \quad \Rightarrow \quad y \doteq 0,714x + 40,514$$

$$y = \frac{\overline{x \cdot y}_A - \bar{x}_A \cdot \bar{y}_A}{(\overline{x^2})_A - (\bar{x}_A)^2} \cdot (x - \bar{x}_A) + \bar{y}_A$$

i	$x_i$	$y_i$	$x_i \cdot y_i$	$x_i^2$
1	0	40	0	0
2	1	42	42	1
:	2	43	86	4
n=6	3	41	123	9
:	4	43	172	16
6	5	44,8	224	25
$\Sigma$	15	253,8	647	55

$$\bar{x}_A = \frac{15}{6} = 2,5$$

$$\bar{y}_A = \frac{253,8}{6} = 42,3$$

$$\overline{(x \cdot y)}_A = \frac{647}{6} \doteq 107,833$$

$$(\overline{x^2})_A = \frac{55}{6} \doteq 9,167$$

Po dosazení: 
$$y = \frac{107,833 - 2,5 \cdot 42,3}{9,167 - 2,5^2} \cdot (x - 2,5) + 42,3 \quad \Rightarrow \quad y \doteq 0,714x + 40,514$$

**Lineární regrese 3.** A co v případě, že si na vzorec (24) nevzpomeneme? Anebo (jako v tomto případě) kdy požadujeme i kvadratickou regresi? Potom je vhodnější využít [soustavy normálních rovnic](#). Opět přepíšeme tabulku, tentokrát svisle a doplníme ji vhodnými sloupci tak, abychom mohli sestavit příslušné soustavy normálních rovnic.

$i$	$x_i$	$y_i$	$x_i^2$					
1	0	40						
2	1	42						
:	2	43						
	3	41						
:	4	43						
6	5	44,8						
$\Sigma$								

## Lineární regrese

## Výběrový korelační koeficient

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$				
1	0	40	0					
2	1	42	1					
:	2	43	4					
$n = 6$	3	41	9					
:	4	43	16					
6	5	44,8	25					
$\Sigma$	15	253,8						

## Lineární regrese

## Výběrový korelační koeficient

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$				
1	0	40	0	0				
2	1	42	1	42				
:	2	43	4	86				
$n = 6$	3	41	9	123				
:	4	43	16	172				
6	5	44,8	25	224				
$\Sigma$	15	253,8	55					

## Lineární regrese

## Výběrový korelační koeficient

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$				
1	0	40	0	0				
2	1	42	1	42				
:	2	43	4	86				
$n = 6$	3	41	9	123				
:	4	43	16	172				
6	5	44,8	25	224				
$\Sigma$	15	253,8	55	647				

## Lineární regrese

Soustava normálních rovnic

## Výběrový korelační koeficient

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$				
1	0	40	0	0				
2	1	42	1	42				
:	2	43	4	86				
$n = 6$	3	41	9	123				
:	4	43	16	172				
6	5	44,8	25	224				
$\Sigma$	15	253,8	55	647				

## Lineární regrese

Soustava normálních rovnic

$$\begin{aligned} 6a + 15b &= 253,8 \\ 15a + 55b &= 647 \end{aligned}$$

## Výběrový korelační koeficient

## Kvadratická regrese



$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$				
1	0	40	0	0				
2	1	42	1	42				
:	2	43	4	86				
$n = 6$	3	41	9	123				
:	4	43	16	172				
6	5	44,8	25	224				
$\Sigma$	15	253,8	55	647				

## Lineární regrese

Soustava normálních rovnic 
$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) \\ 15a + 55b = 647 & | \cdot (-6) \end{array}$$
 má řešení: 
$$\begin{array}{l} a \doteq 40,514 \\ b \doteq 0,714 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

## Výběrový korelační koeficient

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$			
1	0	40	0	0				
2	1	42	1	42				
:	2	43	4	86				
$n = 6$	3	41	9	123				
:	4	43	16	172				
6	5	44,8	25	224				
$\Sigma$	15	253,8	55	647				

## Lineární regrese

Soustava normálních rovnic 
$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) \\ 15a + 55b = 647 & | \cdot (-6) \end{array}$$
 má řešení: 
$$\begin{array}{l} a \doteq 40,514 \\ b \doteq 0,714 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

## Výběrový korelační koeficient

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$			
1	0	40	0	0	1600			
2	1	42	1	42	1764			
:	2	43	4	86	1849			
$n = 6$	3	41	9	123	1681			
:	4	43	16	172	1849			
6	5	44,8	25	224	2 007,04			
$\Sigma$	15	253,8	55	647	10 750,04			

## Lineární regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) & \\ 15a + 55b = 647 & | \cdot (-6) & \end{array}$$

má řešení:  $a \doteq 40,514$   
 $b \doteq 0,714$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

## Výběrový korelační koeficient

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$	$x_i^3$		
1	0	40	0	0	1600			
2	1	42	1	42	1764			
:	2	43	4	86	1849			
$n = 6$	3	41	9	123	1681			
:	4	43	16	172	1849			
6	5	44,8	25	224	2 007,04			
$\Sigma$	15	253,8	55	647	10 750,04			

## Lineární regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) & \\ 15a + 55b = 647 & | \cdot (-6) & \end{array} \quad \begin{array}{l} \text{má řešení:} \\ a \doteq 40,514 \\ b \doteq 0,714 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

**Výběrový korelační koeficient**  $r = \frac{647 - \frac{1}{6} \cdot 15 \cdot 253,8}{\sqrt{(55 - \frac{1}{6} \cdot 15^2) \cdot (10\,750,04 - \frac{1}{6} \cdot 253,8^2)}} \doteq 0,790$

Tedy korelace (lineární závislost) je prokázána.

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$	$x_i^3$	$x_i^4$	
1	0	40	0	0	1600	0		
2	1	42	1	42	1764	1		
:	2	43	4	86	1849	8		
$n = 6$	3	41	9	123	1681	27		
:	4	43	16	172	1849	64		
6	5	44,8	25	224	2 007,04	125		
$\Sigma$	15	253,8	55	647	10 750,04	225		

## Lineární regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) & \\ 15a + 55b = 647 & | \cdot (-6) & \end{array} \quad \begin{array}{l} \text{má řešení:} \\ a \doteq 40,514 \\ b \doteq 0,714 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

**Výběrový korelační koeficient**  $r = \frac{647 - \frac{1}{6} \cdot 15 \cdot 253,8}{\sqrt{(55 - \frac{1}{6} \cdot 15^2) \cdot (10\,750,04 - \frac{1}{6} \cdot 253,8^2)}} \doteq 0,790$

Tedy korelace (lineární závislost) je prokázána.

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$	$x_i^3$	$x_i^4$	$x_i^2 \cdot y_i$
1	0	40	0	0	1600	0	0	
2	1	42	1	42	1764	1	1	
:	2	43	4	86	1849	8	16	
$n = 6$	3	41	9	123	1681	27	81	
:	4	43	16	172	1849	64	256	
6	5	44,8	25	224	2 007,04	125	625	
$\Sigma$	15	253,8	55	647	10 750,04	225	979	

## Lineární regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) & \\ 15a + 55b = 647 & | \cdot (-6) & \end{array} \quad \begin{array}{l} \text{má řešení:} \\ a \doteq 40,514 \\ b \doteq 0,714 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

**Výběrový korelační koeficient**  $r = \frac{647 - \frac{1}{6} \cdot 15 \cdot 253,8}{\sqrt{(55 - \frac{1}{6} \cdot 15^2) \cdot (10\,750,04 - \frac{1}{6} \cdot 253,8^2)}} \doteq 0,790$

Tedy korelace (lineární závislost) je prokázána.

## Kvadratická regrese

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$	$x_i^3$	$x_i^4$	$x_i^2 \cdot y_i$
1	0	40	0	0	1600	0	0	0
2	1	42	1	42	1764	1	1	42
:	2	43	4	86	1849	8	16	172
$n = 6$	3	41	9	123	1681	27	81	369
:	4	43	16	172	1849	64	256	688
6	5	44,8	25	224	2 007,04	125	625	1 120
$\Sigma$	15	253,8	55	647	10 750,04	225	979	2 391

## Lineární regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) & \\ 15a + 55b = 647 & | \cdot (-6) & \end{array} \quad \begin{array}{l} \text{má řešení:} \\ a \doteq 40,514 \\ b \doteq 0,714 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

**Výběrový korelační koeficient**  $r = \frac{647 - \frac{1}{6} \cdot 15 \cdot 253,8}{\sqrt{(55 - \frac{1}{6} \cdot 15^2) \cdot (10\,750,04 - \frac{1}{6} \cdot 253,8^2)}} \doteq 0,790$

Tedy korelace (lineární závislost) je prokázána.

## Kvadratická regrese

Soustava normálních rovnic

$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$	$x_i^3$	$x_i^4$	$x_i^2 \cdot y_i$
1	0	40	0	0	1600	0	0	0
2	1	42	1	42	1764	1	1	42
:	2	43	4	86	1849	8	16	172
$n = 6$	3	41	9	123	1681	27	81	369
:	4	43	16	172	1849	64	256	688
6	5	44,8	25	224	2 007,04	125	625	1 120
$\Sigma$	15	253,8	55	647	10 750,04	225	979	2 391

## Lineární regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) & \\ 15a + 55b = 647 & | \cdot (-6) & \end{array} \quad \begin{array}{l} \text{má řešení:} \\ a \doteq 40,514 \\ b \doteq 0,714 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

**Výběrový korelační koeficient**  $r = \frac{647 - \frac{1}{6} \cdot 15 \cdot 253,8}{\sqrt{(55 - \frac{1}{6} \cdot 15^2) \cdot (10\,750,04 - \frac{1}{6} \cdot 253,8^2)}} \doteq 0,790$

Tedy korelace (lineární závislost) je prokázána.

## Kvadratická regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b + 55c = 253,8 \\ 15a + 55b + 225c = 647 \\ 55a + 225b + 979c = 2\,391 \end{array}$$



$i$	$x_i$	$y_i$	$x_i^2$	$x_i \cdot y_i$	$y_i^2$	$x_i^3$	$x_i^4$	$x_i^2 \cdot y_i$
1	0	40	0	0	1600	0	0	0
2	1	42	1	42	1764	1	1	42
:	2	43	4	86	1849	8	16	172
$n = 6$	3	41	9	123	1681	27	81	369
:	4	43	16	172	1849	64	256	688
6	5	44,8	25	224	2 007,04	125	625	1 120
$\Sigma$	15	253,8	55	647	10 750,04	225	979	2 391

## Lineární regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b = 253,8 & | \cdot (15) & \\ 15a + 55b = 647 & | \cdot (-6) & \end{array} \quad \begin{array}{l} \text{má řešení:} \\ a \doteq 40,514 \\ b \doteq 0,714 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,714x + 40,514$

**Výběrový korelační koeficient**  $r = \frac{647 - \frac{1}{6} \cdot 15 \cdot 253,8}{\sqrt{(55 - \frac{1}{6} \cdot 15^2) \cdot (10\,750,04 - \frac{1}{6} \cdot 253,8^2)}} \doteq 0,790$

Tedy korelace (lineární závislost) je prokázána.

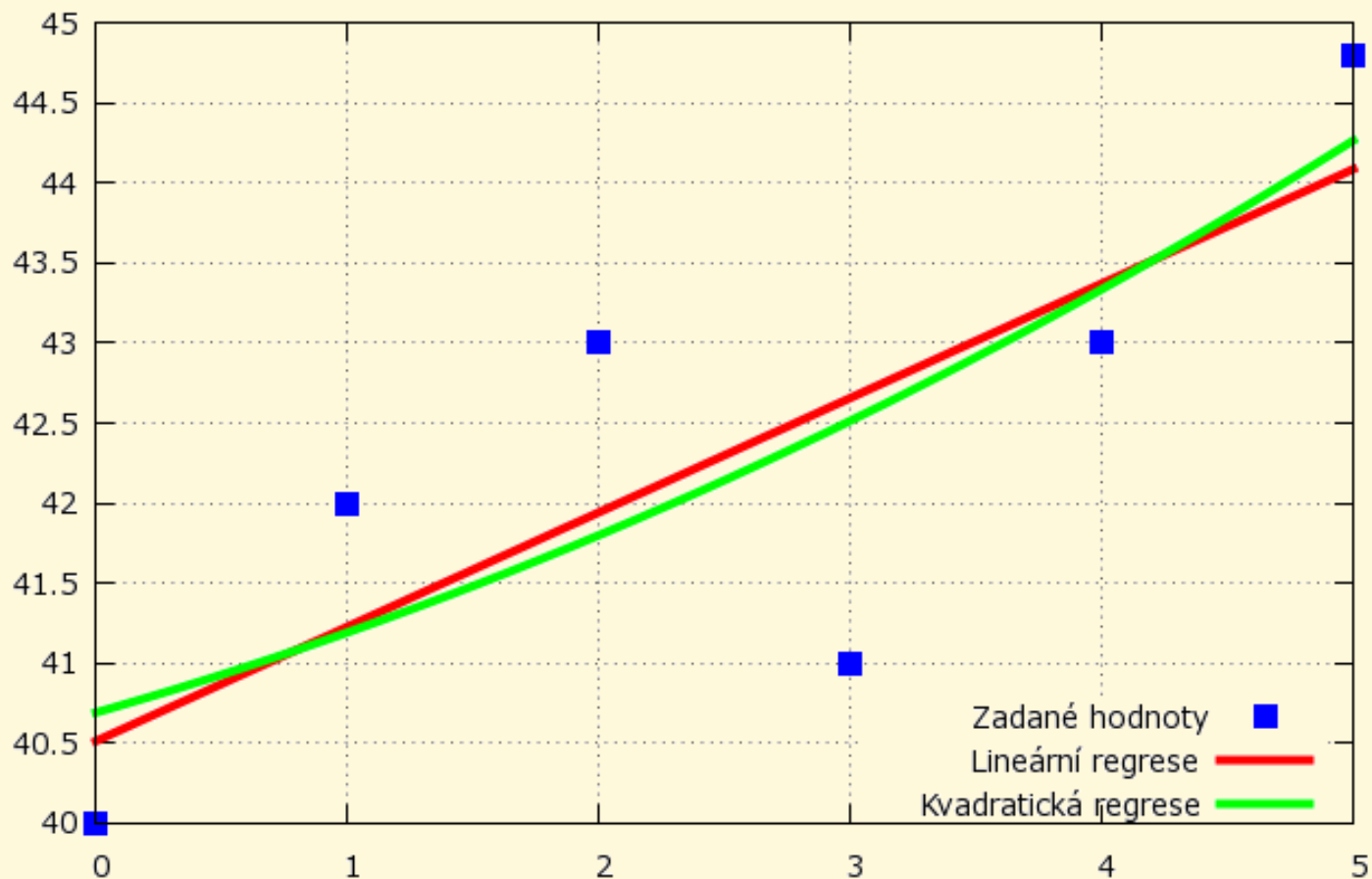
## Kvadratická regrese

Soustava normálních rovnic

$$\begin{array}{rcl} 6a + 15b + 55c = 253,8 & & \\ 15a + 55b + 225c = 647 & & \\ 55a + 225b + 979c = 2\,391 & & \end{array} \quad \begin{array}{l} \text{má řešení:} \\ a \doteq 40,693 \\ b \doteq 0,446 \\ c \doteq 0,054 \end{array}$$

**regresní funkce**  $f(x) : y \doteq 0,054x^2 + 0,446x + 40,693$

## REGRESE - Metoda nejmenších čtverců



## Rad'te se s rozumem

**Nevěřte slepě všemu, co je podloženo čísly!** Například:

V řadě evropských regionů bylo zjištěno, že čím více čápů žije v určité krajině, tím vyšší je tam porodnost. Korelační koeficienty byly tak významné, že je velice nepravděpodobné, že zjištěná souvislost je náhodná. Jsme tedy ochotni přijmout hypotézu, že čápi přece jen nosí děti? Asi sotva. Ale pak bychom měli navrhnout hypotézu, která by uspokojivě vysvětlovala naměřenou souvislost. [2, str. 21]

Abychom mohli prohlásit, že vztah mezi dvěma proměnnými (náhodnými veličinami) má charakter *příčina*  $\Rightarrow$  *důsledek*, musí být splněny všechny následující podmínky:

- Musí existovat souběžné změny u obou proměnných,
- musíme vyloučit existenci nějaké další, vnější příčiny a
- změny v obou proměnných se musí objevit v logickém časovém pořadí.

Jsou-li prokázány závislosti, zbývá většinou ještě otázka o příčině a účinku, o náhodě nebo hlubším významu, o přímé závislosti, společném třetím (co způsobuje vyšší porodnost a současně výskyt většího počtu čápů) nebo klamném zdání. Co například toto: nepoměrně více lidí umírá v posteli, než na ulici, při sportu, zábavě, ...  $\Rightarrow$  **postel je nejnebezpečnější místo pobytu.**

Při korelaci platí mnohem více než v jiné oblasti statistické práce následující [14, str. 307]:  
**„číselné výsledky, i když byly vypočteny z bezvadných podkladů, ještě nic nedokazují, nýbrž ukazují, upozorňují, vyzývají k vytváření hypotéz a jejich následnému testování.“**

Viz též článek „[Potrhlá astrologie](#)“ Frederika Velinského z prosince 2009.

# Úvod do **Hospodářské statistiky**

## Obsah kapitoly: Hospodářská statistika

<b>1. Statistika a ekonomie</b>	<b>298</b>
1.1. Základní pojmy . . . . .	298
<b>2. Individuální indexy</b>	<b>306</b>
2.1. Jednoduché individuální indexy . . . . .	306
Příklad: tržby . . . . .	308
2.1.1. Poznámka k veličině s názvem „průměrný koeficient vývoje“ . . . . .	314
Příklad: přeprava cestujících . . . . .	314
2.2. Složené individuální indexy . . . . .	317
Příklad . . . . .	318
<b>3. Souhrnné (agregátní) indexy</b>	<b>325</b>
Příklad . . . . .	326
<b>4. Závěr kapitoly – Shrnutí</b>	<b>333</b>
4.1. Příklady používaných indexů v praxi . . . . .	335

# 1. Statistika a ekonomie

Statistika byla zpočátku využívána spíše ve vědách přírodních (fyzika, chemie), v posledních letech však zaznamenává úspěch také v disciplínách humanitního charakteru, například v psychologii, sociologii, ale také v ekonomii. K výraznějšímu rozvoji statistických metod v ekonomii došlo na přelomu 19. a 20. století, a to zejména díky novým objevům ve statistice (zejména nástupu metod matematické statistiky).

*V současné době patří statistika stejně jako informatika nebo operační výzkum ke standardnímu vybavení moderního ekonoma. Proto je nutné, aby ekonomové znali základy statistiky a měli alespoň základní představu o možnostech a nástrojích této disciplíny. [3, str. 35]*

Aplikací statistických metod na ekonomická a sociálně ekonomická data vznikla samostatná statistická disciplína, hospodářská (ekonomická) statistika.

**Předmětem** ekonomické statistiky je analýza stavu a vývoje jevů v hospodářské oblasti.

**Cílem** hospodářské statistiky je nalezení způsobu měření a vyhodnocení ekonomické skutečnosti jako východiska k hospodářskému rozhodování či stanovení hospodářské politiky.

## 1.1. Základní pojmy

**Ukazatelé** jsou veličiny, se kterými se denně setkáváme. Ať již v denním tisku, v rozhlase, či v televizi. Seznamujeme se s takovými pojmy jako hrubý domácí produkt (HDP), dovoz, vývoz, produktivita práce, průměrná mzda, výsledky voleb, apod. Tyto pojmy jsou vždy doprovázeny čísly, která charakterizují velikost odpovídajícího (ekonomického, společenského, ...) jevu, případně vývoj daného jevu. Dovídáme se, že například HDP vzrostl o  $xy$  %, saldo zahraničního obchodu dosáhlo výše  $yz$  mld. Kč, roční míra inflace byla  $xz$  %. Zároveň se zpravidla seznamujeme s tím, zda tyto výsledky máme hodnotit kladně či záporně, v jakých souvislostech a za jakých podmínek.

V praxi však zpravidla nepracujeme s jednotlivými izolovanými hodnotami určitého ukazatele, ale snažíme se zjistit, zda ekonomická skutečnost (vyjádřená hodnotou určitého ukazatele) znamená určitou změnu oproti téže skutečnosti v minulém období nebo v jiné územní či organizační jednotce. Nejjednodušší a často používanou metodou statistického rozboru je porovnávání takových statistických údajů.

Jednou z možností, jak vzájemně porovnat dvě hodnoty, je zkoumání, kolikrát je jedna hodnota větší jak druhá. To provedeme matematickou operací **dělení**, jejímž výsledkem je podíl. Druhou možností je zkoumat, o kolik je jedna hodnota větší jak druhá. To provedeme matematickou operací **odčítání**, jejímž výsledkem je rozdíl. Obě tyto míry jsou rovnocenné a nezastupitelné a vzájemně se doplňují.

**Statistický ukazatel** je číslo, které v daném prostoru a čase charakterizuje určitou skutečnost (určitý jev).

Přesněji řečeno je *funkcí hodnot znaku statistických jednotek* (funkcí charakteristik znaku). Je to kvantitativní popis určité sociálně–ekonomické skutečnosti.

Vezmeme-li například ukazatel „odpracovaná doba“, pak tento ukazatel je v metodických předpisech vymezen jako úhrn pracovní doby odpracované dělníky (pracovníky) daného podniku (závodu, provozovny) v měsíci (čtvrtletí, roce). Jde tedy o popis ukazatele, kde je obecně definován **čas** (měsíc) a **prostor** (podnik). Jestliže přesně definujeme tento čas a prostor (například únor 1997, podnik E.ON), dostaneme konkrétní hodnotu ukazatele nazývanou **údaj**.

**Poměrný ukazatel** vznikne jako podíl (poměr) dvou číselných hodnot.

Mohou být podílem stejnorodých údajů, které jsou stejného obsahu a rozměru. Potom je poměrné číslo bezrozměrné a často ho vyjadřujeme v procentech. Příkladem může být ukazatel podílu žen v celkovém počtu pracovníků firmy.

Pokud je v čitateli poměrného ukazatele hodnota jiného obsahu a rozměru než ve jmenovateli, jedná se o podíl nestejnorodých ukazatelů a poměrný ukazatel je rozměrový. Například počet obyvatel na jednoho zubaře, produktivita práce podniku apod.

Při srovnávání ukazatelů z časového hlediska hovoříme o **základním období**, které označujeme indexem **0** a **běžném období**, které označujeme indexem **1**.

**Poměrné ukazatele struktury** (neboli složení) vyjadřují podíl určité části vzhledem k celku.

**Indexy** jsou poměrné hodnoty, které umožňují srovnání shodně vymezených ukazatelů (stejného druhu a obsahu).

**Index** je podíl hospodářských ukazatelů, indikátor pokroku či neúspěchu. Je to bezrozměrné číslo (čas-to se uvádí v procentech), které nám ukazuje průběh nějakého vývoje tím, že zaznamenává změny oproti dřívějšímu období. Musí charakterizovat celkovou situaci, nejen situaci jednotlivého výrobku.

Dříve jsme uvedli, že hodnota statistického ukazatele vzniká jeho konkrétním

- časovým
- prostorovým
- druhovým (charakterizuje určitou skutečnost)

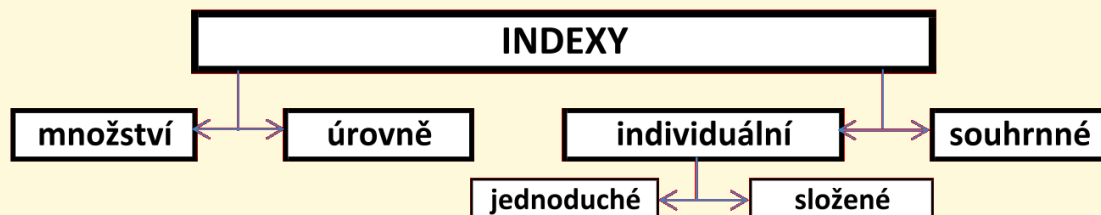
vymezením. Pak index, jakožto podíl dvou hodnot téhož ukazatele, se může lišit pouze v jednom z uvedených hledisek a zbývající dvě hlediska musí být vymezena stejně. Budeme-li například srovnávat:

- zisky podniku **A** ve dvou následujících letech, dostaneme **časový** index;
- zisk podniku **A** se ziskem podniku **B** ve stejném roce dostaneme **prostorový** index;
- zisk při výrobě produktu **X** a produktu **Y** v jednom podniku a daném roce dostaneme **druhový** index.



Indexů existuje velké množství a záleží na více hlediscích, který druh indexu použijeme.

Obrázek 6: Podle [11, str. 109]



Dělení indexů podle ukazatele, jehož dynamiku máme charakterizovat

Členění na indexy množství a úrovně vychází z typu ukazatele. Říkáme, že rozdělujeme indexy podle **extenzitních** a **intenzitních** ukazatelů.

Druhým kritériem členění je **stejnorodost** nebo nestejnorodost ukazatele. **Souhrnné** indexy jsou indexy nestejnorodých (extenzitních i intenzitních) ukazatelů, **individuální** indexy jsou indexy stejnorodých (extenzitních i intenzitních) ukazatelů. Indexy **stejnorodých** ukazatelů třídíme dále na indexy jednoduché a složené. **Jednoduché** indexy jsou takové, u nichž neprovádíme shrnování. U **složených** indexů shrnujeme dílčí hodnoty sledovaného ukazatele.

**Extenzitní ukazatel  $q$**  udává **množství**, **objem**, **rozsah** nebo **počet** sledovaného jevu (například výroba, prodej, počet pracovníků, zboží v kusech apod.) v nějaké jednotce (Kč, kg, m<sup>2</sup>, ...) a vyjadřuje tak nějakou (ekonomickou) skutečnost; je vyjádřen číslem. Obvykle jej označujeme  $q$ .

*Extenzitní (stejnorodé) ukazatele shrnujeme* (určujeme celkovou hodnotu ukazatele na základě jeho dílčích hodnot) **součtem**. Můžeme například sečíst množství prodaných akcií téže firmy u několika makléřů. Nebo součet produkcí (v kusech) jednoho druhu zboží za jednotlivé měsíce roku dává roční produkci tohoto druhu zboží.

Nestejnorodé extenzitní veličiny sčítat nelze. Například nemá smysl sčítat prodané vkladové listy a množství poskytnutých úvěrů, i když byly realizovány v jedné bance.

**Intenzitní ukazatel  $p$**  dává do poměru (**podílu**) dva extenzitní ukazatele, které mají logickou souvislost a jsou vyjádřeny každý v jiných jednotkách (Kč/m, t/ha, ...). Tedy vyjadřuje **úroveň** (například cena je podíl tržeb a prodaného množství). Obvykle jej označujeme  $p$ .

*Intenzitní (stejnorodé) ukazatele shrnujeme* (určujeme celkovou hodnotu ukazatele na základě jeho dílčích hodnot) **váženým průměrem**. Různorodé intenzitní veličiny vznikají jako podíl nestejnorodých extenzitních veličin (například ceny elektřiny a plynu). Takové veličiny nelze ani sčítat ani průměrovat.

Intenzitní a extenzitní veličiny se často vyskytují ve dvojici, kde určují intenzitu (úroveň) a kvantitu (množství) daného jevu (například: cenu  $\times$  prodané množství, produktivitu práce  $\times$  odpracovaný počet hodin, ...). Odpovídající hodnotu veličiny intenzitní  $p$  a extenzitní  $q$  lze násobit, přičemž vznikne nová souhrnná extenzitní veličina, kterou obvykle označujeme  $Q$  ( $Q = p \cdot q$ ). Tuto veličinu lze opět sčítat, a to i v případě nestejnorodých veličin  $q$ . Třeba sečtením tržeb za jednotlivé výrobky dostaneme celkovou tržbu prodejny.

Chceme-li vědět, **kolikrát** (o kolik %) je jedna hodnota ukazatele menší/větší než jiná, budeme obě hodnoty srovnávat **podílem**. Budeme-li chtít vědět *o kolik jednotek* je jedna hodnota ukazatele menší/větší než jiná, budeme obě hodnoty srovnávat **rozdílem**. Podílem dvou hodnot téhož ukazatele získáme (jak jsme již uvedli) **index**, rozdílem pak absolutní přírůstek. Obě tyto míry rozdílnosti jsou rovnocenné a nezastupitelné, ale vzájemně se doplňují.

**Poměrná čísla rozměrová** jsou tvořena jako podíl ukazatelů různého obsahu a rozměru. Pokud označíme poměrný ukazatel  $z = \frac{y}{x}$ , pak můžeme průměrnou hodnotu poměrného ukazatele (indexu) vypočítat různými způsoby.

Tak jako mnohokrát v této příručce budeme psát (kvůli úspoře místa) pouze prostý symbol **sumy**,

u které vynecháme sčítací index<sup>44</sup>.

- Průměrnou hodnotu poměrného ukazatele vypočítáme jako **prostý aritmetický průměr**, tedy jako podíl součtu všech hodnot čitatele a součtu hodnot jmenovatele poměrného ukazatele:

$$\bar{z} = \frac{\sum y_i}{\sum x_i}$$

- Průměrnou hodnotu poměrného ukazatele vypočítáme jako **vážený aritmetický průměr** hodnot poměrného ukazatele, kde vahami bude jmenovatel poměrného ukazatele:

$$\bar{z} = \frac{\sum z_i \cdot x_i}{\sum x_i}$$

- Průměrnou hodnotu poměrného ukazatele vypočítáme jako **vážený harmonický průměr** hodnot poměrného ukazatele, kde vahami bude číselný jmenovatel poměrného ukazatele:

$$\bar{z} = \frac{\sum y_i}{\sum \frac{y_i}{z_i}}$$

V hospodářské praxi je časté použití váženého harmonického průměru například při výpočtu průměrné produktivity práce ve firmě složené z několika filiálék.

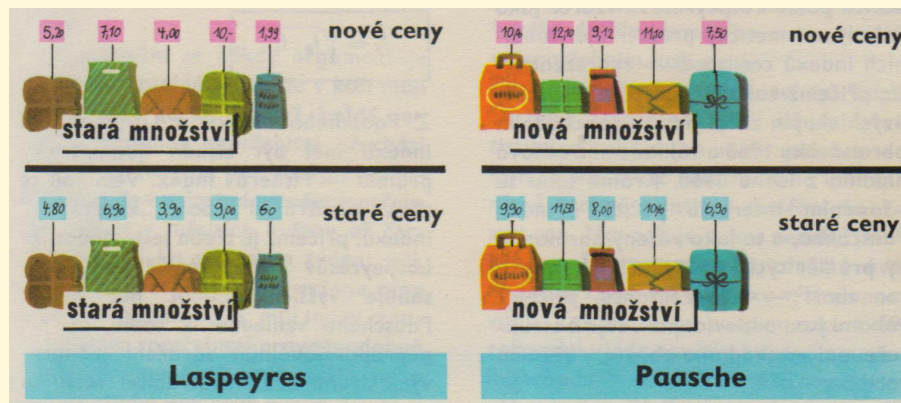
Na celém světě nejznámější a také nejvíce napadaný je **index spotřebitelských cen**, kterému také někdy říkáme **index životních nákladů**. Proti tomuto indexu se často namítá, že se v něm skutečná změna

<sup>44</sup> Správně by mělo být například  $\sum_i x_i$  nebo (pokud  $i = 1, 2, \dots, n$ )  $\sum_{i=1}^n x_i$

životních nákladů zrcadlí jen nedostatečně, protože spotřební zvyklosti se mění a navíc je zkonstruován na základě spotřebního schématu — **spotřebního koše**, který přesně neodpovídá snad pro žádného spotřebitele.

Srovnání dnešního indexu životních nákladů s rokem 1989 je již skoro k ničemu a jestliže se dalším zřetězováním počítá zpět až do roku 1900, je to sice matematicky zcela možné, ale jinak zcela nesmyslné. Cituji: „Tím se zabývají jen historikové — podivíni, kteří nám ještě dnes pečlivě a přesně vypočítají, jakou hodnotu měl sestericius ve starém Římě.“ [14, str. 111]

Obrázek 7: Převzat z [14]



Mezi četnými cenovými indexy nabyly zvláštního významu dva: Laspeyresův index (porovnání cen na základě původně spotřebovaného množství — stará množství jako základna) a Paascheho index (porovnání cen na základě nové spotřeby)<sup>45</sup>. Laspeyresův index téměř vždy dosahuje vyšších hodnot jako

<sup>45</sup> Paasche a Laspeyres byli němečtí národohospodáři z konce 19. století.

index Paascheho vzhledem k tomu, že při neproporcionálním zdražení jednotlivých druhů zboží spotřebitel většinou přechází na jiné (lacinější) druhy, takže „nová zboží“ zachytí část zdražení.

Musíme si ovšem uvědomit, že stoupne-li nějaký index (stanovovaný například pomocí koše – pak jde /jak již víme/ o bodový odhad charakteristiky) z hodnoty 108,6 v jednom měsíci na 108,8 v následujícím měsíci, neříká to nic jiného, než toto: **Pravděpodobnost, že hodnoty (které jsou základem výpočtu) stoupily, je nepatrně větší než pravděpodobnost, že se nezměnily nebo klesly.** Protože i když budeme předpokládat „směrodatnou odchylku přesnosti  $\sigma$ “ jen ve výši 3 ‰ (a to je i při pečlivé práci **nereálně** málo), musíme říci: **Údaj prvního měsíce s bodovým odhadem 108,6 leží s 95 % pravděpodobností mezi 108,0 a 109,2** (pravidlo dvou  $\sigma$  dává 95% pravděpodobnost). Údaj 108,8, který byl určen za nový měsíc, leží (má intervalový odhad) mezi 108,2 a 109,4. Není tady vůbec vyloučeno, že správný index za předchozí měsíc je 108,8 a za nový měsíc jen 108,5 nebo také že oba jsou si přesně rovny.

Jestliže však naproti tomu delší řada takových indexů vykazuje stále stejný vývoj, stává se správnost pozorování stále pravděpodobnější. Následují-li například po hodnotách 108,6 a 108,8 jako další čísla v řadě 109,1 a 109,5, můžeme právem — nikoli však s absolutní jistotou — předpokládat, že vývoj indexu za dané čtyři měsíce vyjadřuje skutečně existující vzestupný vývoj.

**Žádný index není zcela přesný!** To však není argument proti indexu nebo proti jakémukoliv jinému statistickému šetření. Není-li možno získat žádnou dokonalou informaci, musíme se spokojit s pokud možno nejpresnějšími odhady. A i ten nejpresnější odhad je stále jen odhad — ale je nepoměrně cennější než nevědomost, prázdná domněnka nebo „věštění z křišťálové koule“. Každý koš zboží je konec konců jen výběrový soubor a již v samé podstatě výběru je, že nemůže zprostředkovat absolutní jistotu o celém základním souboru.

**Potřebujeme vždy výpočty na zlomky procent?** Naše myšlení je většinou příliš ovládáno utkvělou představou, že číslo vypočítané až na poslední platné místo je vrcholem přesnosti a pravdivosti. Ve skutečnosti je tomu často naopak. Jen zřídka kdy je možno na otázku „*Kolik je hodin?*“ odpovědět naprosto přesně

ve tvaru („gong oznámí“) „**15 hodin, 32 minuty, 40 sekund**“. Stejně užitečná a nepříliš lživá je odpověď „**půl čtvrté**“.

## 2. Individuální indexy

**Individuální indexy** jsou nejjednoduššími veličinami, které bezprostředně srovnávají dvě hodnoty téhož ukazatele (*podíl stejnorodých veličin*).

Pokud porovnáváme údaj o úrovni jedné veličiny, který jsme získali bez shrnování součtem nebo průměrem, hovoříme o **jednoduchých individuálních indexech**. Pokud jsou údaje sumarizovány nebo průměrovány z více zdrojů (například z více prodejen) hovoříme o **složených individuálních indexech**.

### 2.1. Jednoduché individuální indexy

Tyto jednoduché individuální indexy nejsou nijak podrobněji členěny ani shrnovány. Budeme-li srovnávat hodnotu intenzitního ukazatele ***p*** v situaci **1** (v časovém srovnání nazývané *běžným obdobím b. o.*) a v situaci **0** (v časovém srovnání nazývané *základním obdobím z. o.*), obdržíme  $I_p$  (někdy se též označuje  $i_p$ ). Analogicky můžeme konstruovat jednoduché indexy i pro extenzitní ukazatele ***q*** a ***Q***. Tedy

$$I_p = \frac{p_1}{p_0} \quad I_q = \frac{q_1}{q_0} \quad I_Q = \frac{Q_1}{Q_0} \quad (26)$$

Ze vztahu  $Q = p \cdot q$  plyne, že

$$I_Q = I_q \cdot I_p$$

Individuální jednoduché indexy (zde výlučně časové  $\Rightarrow$  zjišťujeme hodnotu jednoho ukazatele v daném prostoru, ale v různém čase) se často vyskytují sdružené do delších časových řad. Tehdy mohou být příslušné indexy počítány

**ke stejnému základu – bázi** například (26) k nejstarší hodnotě (bází může být jakékoliv období, nikoliv nutně první) v časové řadě původních pozorování  $\Rightarrow$  tzv. **bazické indexy**  $S_i = \frac{x_i}{x_0}$

**k proměnlivému základu** k bezprostředně předcházejícímu pozorování v časové řadě původních hodnot  $\Rightarrow$  tzv. **řetězové indexy**  $T_i = \frac{x_i}{x_{i-1}}$

- řetězový index vyjádřený v procentech se nazývá **tempo růstu**;
- geometrický průměr řetězových indexů se nazývá **průměrný koeficient vývoje**.

K posouzení téže změny u všech jednotek (prodej ve všech filiálkách daného obchodního řetězce apod.) musíme použít **složené individuální indexy**.

## Jednoduché (individuální) indexy $\Rightarrow$ jeden ukazatel jednoho střediska

Tržby: (základní období má VŽDY index NULA) **40** 42 43 41 43 44,8 (tedy  $n = 5$ )

Určete vhodné indexy

$i$	$x_i$		
0	40		
1	42		
2	43		
3	41		
4	43		
5	44,8		



## Jednoduché (individuální) indexy $\Rightarrow$ jeden ukazatel jednoho střediska

Tržby: (základní období má VŽDY index NULA) **40** 42 43 41 43 44,8 (tedy  $n = 5$ )

Určete vhodné indexy

$i$	$x_i$	$S_i = \frac{x_i}{x_0}$	
<b>0</b>	<b>40</b>	[1]	
1	42	1,05	
2	43	1,075	
3	41	1,025	
4	43	1,075	
5	44,8	1,12	

## Jednoduché (individuální) indexy $\Rightarrow$ jeden ukazatel jednoho střediska

Tržby: (základní období má VŽDY index NULA) **40** 42 43 41 43 44,8 (tedy  $n = 5$ )

Určete vhodné indexy, průměrný koeficient vývoje a odhadněte tržby v následujícím měsíci.

$i$	$x_i$	$S_i = \frac{x_i}{x_0}$	$T_i = \frac{x_i}{x_{i-1}}$
<b>0</b>	<b>40</b>	[1]	/
1	42	1,05	1,05
2	43	1,075	1,024
3	41	1,025	0,953
4	43	1,075	1,049
5	44,8	1,12	1,042

## Jednoduché (individuální) indexy $\Rightarrow$ jeden ukazatel jednoho střediska

Tržby: (základní období má VŽDY index NULA) **40** 42 43 41 43 44,8 (tedy  $n = 5$ )

Určete vhodné **indexy**, průměrný koeficient vývoje a odhadněte tržby v následujícím měsíci.

$i$	$x_i$	$S_i = \frac{x_i}{x_0}$	$T_i = \frac{x_i}{x_{i-1}}$
<b>0</b>	<b>40</b>	[1]	/
1	42	1,05	1,05
2	43	1,075	1,024
3	41	1,025	0,953
4	43	1,075	1,049
5	44,8	1,12	1,042

$$T_1 \cdot T_2 \cdot T_3 \cdot T_4 \cdot T_5 = \frac{x_1}{x_0} \cdot \frac{x_2}{x_1} \cdot \frac{x_3}{x_2} \cdot \frac{x_4}{x_3} \cdot \frac{x_5}{x_4} = S_5$$

## Jednoduché (individuální) indexy $\Rightarrow$ jeden ukazatel jednoho střediska

Tržby: (základní období má VŽDY index NULA) **40** 42 43 41 43 44,8 (tedy  $n = 5$ )

Určete vhodné indexy, průměrný koeficient vývoje a odhadněte tržby v následujícím měsíci.

$i$	$x_i$	$S_i = \frac{x_i}{x_0}$	$T_i = \frac{x_i}{x_{i-1}}$
0	40	[1]	/
1	42	1,05	1,05
2	43	1,075	1,024
3	41	1,025	0,953
4	43	1,075	1,049
5	44,8	1,12	1,042

$$T_1 \cdot T_2 \cdot T_3 \cdot T_4 \cdot T_5 = \frac{x_1}{x_0} \cdot \frac{x_2}{x_1} \cdot \frac{x_3}{x_2} \cdot \frac{x_4}{x_3} \cdot \frac{x_5}{x_4} = S_5$$

$$\bar{x}_G = \sqrt[5]{T_1 \cdot T_2 \cdot T_3 \cdot T_4 \cdot T_5} = \sqrt[5]{1,05 \cdot 1,024 \cdot 0,953 \cdot 1,049 \cdot 1,042} = \sqrt[5]{S_5} = \sqrt[5]{1,12} = 1,022$$

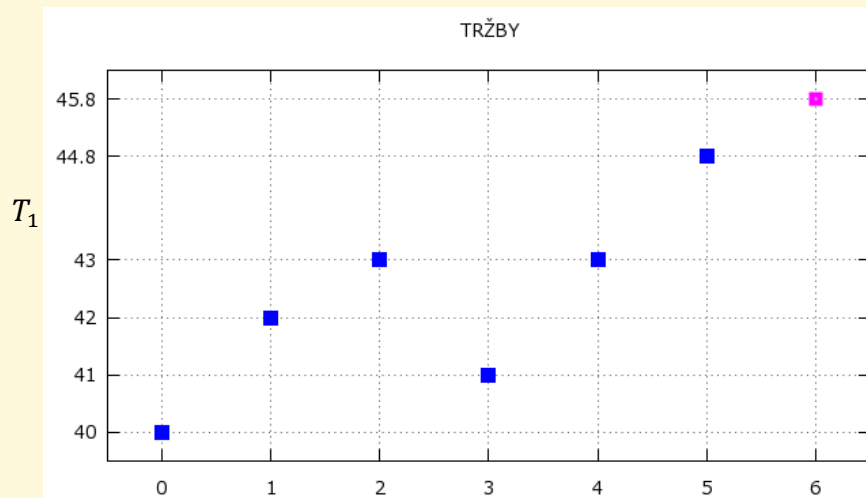
V každém období tedy tržby rostly 1,022 krát.

## Jednoduché (individuální) indexy $\Rightarrow$ jeden ukazatel jednoho střediska

Tržby: (základní období má VŽDY index NULA) **40** 42 43 41 43 44,8 (tedy  $n = 5$ )

Určete vhodné **indexy**, průměrný koeficient vývoje a odhadněte tržby v následujícím měsíci. Graficky znázorněte řadu tržeb (čísel) v čase.

$i$	$x_i$	$S_i = \frac{x_i}{x_0}$	$T_i = \frac{x_i}{x_{i-1}}$
0	<b>40</b>	[1]	/
1	42	1,05	1,05
2	43	1,075	1,024
3	41	1,025	0,953
4	43	1,075	1,049
5	44,8	1,12	1,042



$$\bar{x}_G = \sqrt[5]{T_1 \cdot T_2 \cdot T_3 \cdot T_4 \cdot T_5} = \sqrt[5]{1,05 \cdot 1,024 \cdot 0,953 \cdot 1,049 \cdot 1,042} = \sqrt[5]{S_5} = \sqrt[5]{1,12} = 1,022$$

V každém období tedy tržby rostly 1,022 krát.

Předpokládané tržby pro šesté období odhadneme nejsnadněji tak, že hodnotu pátého období vynásobíme koeficientem 1,022. **V šestém období** budou tržby pravděpodobně:  $44,8 \times 1,022 = 45,786$ .

Přesnější odhad pravděpodobných tržeb v šestém období získáme např. pomocí **regresní analýzy** (regresní přímku a regresní parabolu jsme zkoumali v předchozí kapitole) nebo pomocí **trendu** (lineární a kvadratický trend bude probírán v kapitole Modelování časových řad).

### 2.1.1. Poznámka k veličině s názvem „průměrný koeficient vývoje“

Na předchozím příkladu jsme si ukázali, že pro  $k$  zadaných hodnot nemusíme počítat  $k-1$  řetězových indexů a určovat jejich geometrický průměr, ale stačí vypočítat  $k-1$  odmocninu bazického indexu  $S_{k-1}$ . Jinými slovy náš odhad vývoje pomocí průměrného koeficientu vývoje je založen pouze na **první** a **poslední** zadané hodnotě. Ostatní zadané údaje nemají na náš odhad vývoje naprosto žádný vliv.

Průměrný koeficient vývoje můžeme ještě určit také tak, že ponecháme beze změny první a poslední zadanou hodnotu a zbylé údaje upravíme tak, aby všechny dohromady tvořily **geometrickou posloupnost**.

Průměrný koeficient vývoje je potom roven kvocientu této geometrické řady.

Vše si ukážeme na následujícím příkladu, kde jsou číselné údaje zaokrouhlené na **stovky**.

Dlouhodobým pozorováním bylo zjištěno, že autobusová linka xyz přepraví ve čtvrtek **4 tisíce** cestujících (tedy ve čtvrtek je přepraveno od 3 951 do 4 049 osob), v pátek je to také **4 tisíce**, zatímco v sobotu a v neděli pouze **1 tisíc**.

Nyní si představme, že máme k dispozici následující řadu údajů: (Čt) **4 000** ; (Pá) **4 000** ; (So) **1 000** a máme odhadnout, jaké číslo bude následovat. Tedy určit, *kolik asi pasažérů je přepravováno v neděli*. Sice se nejedná o typický případ, protože k dispozici máme příliš malý vzorek, ale to snad v tomto případě příliš nevadí. Alespoň si proto připomeňme, že **každý závěr a tím spíše také rozhodnutí by mělo být dostatečně podloženo**.

- Využití „*zdravého selského rozumu*“. Pokud víme, co jednotlivé zkratky znamenají a na základě této znalosti usoudíme, že nás zajímá počet pasažérů o víkendovém dnu, můžeme důvodně předpokládat, že to bude stejné jako jiný víkendový den. Tedy opět **jeden** tisíc.

- Odhad pomocí **průměrného koeficientu vývoje**.

$i$	den	$x_i$	$S_i = \frac{x_i}{x_0}$	$T_i = \frac{x_i}{x_{i-1}}$
0	Čt	4 000	[1]	/
1	Pá	4 000	1	1
2	So	1 000	0,25	0,25
	Ne	?	$\Leftarrow 1\,000 \cdot 0,5$	

Průměrný koeficient vývoje:

$$\sqrt{T_1 \cdot T_2} = \sqrt{1 \cdot 0,25} = \sqrt{S_2} = \sqrt{0,25} = 0,5$$

což je stejné jako kvocient  $q = \frac{1}{2}$  geometrické řady 4 000 ; 2 000 ; 1 000 ve které jsme vhodně upravili prostřední (páteční) hodnotu.

Tedy na základě průměrného koeficientu vývoje bychom pro neděli odhadovali **500** pasažérů a to, jak víme z první odrážky, nebude asi až tak moc přesné, ale v zásadě je to možné.

A už jsme zase u problému, který jsme diskutovali již dříve. A to u rozdělování původního vzorku na **častečné vzorky**, zde na pracovní dny a víkendové dny.

- V kapitole **regrese** jsme data vyrovnávali přímkou podle vzorce (25). Tento můžeme aplikovat i na náš případ, použijeme-li k označení dnů místo zkratk například jejich pořadové číslo. Pomocí souřadnic bodu ležícího na přímce pak odhadneme požadovaný údaj.

$x$	1	2	3	4
$y = f(x)$	4 000	4 000	1 000	?

Potom

$$f(x) = \frac{-1\,500}{1} \cdot (x - 2) + 3\,000$$

$$f(4) = -1\,500 \cdot (4 - 2) + 3\,000 = 0$$

Tedy na základě lineární regrese bychom pro neděli odhadovali do **50** pasažérů a to je velmi nepravděpodobné.

- V předmětu *Matematika* jsme zadanými body prokládali polynom, a to Lagrangeův interpolační mnohočlen. Pro náš případ:

$$f(x) = 4\,000 \cdot \frac{(x-2) \cdot (x-3)}{(1-2) \cdot (1-3)} + 4\,000 \cdot \frac{(x-1) \cdot (x-3)}{(2-1) \cdot (2-3)} + 1\,000 \cdot \frac{(x-1) \cdot (x-2)}{(3-1) \cdot (3-2)} = -1\,500x^2 + 4\,500x + 1\,000$$

$$f(4) = -1\,500 \cdot 4^2 + 4\,500 \cdot 4 + 1\,000 = -5\,000$$

Tedy na základě interpolačního mnohočlenu bychom pro neděli odhadovali **MÍNUS pět tisíc** pasažérů a to je nemožné.

- A co když bude zadáno: (Pá) **4 000** ; (So) **1 000** ; (Ne) **1 000** ?  
A chtěli bychom odhadnout, *kolik asi pasažérů je přepravováno v pondělí?*

$i$	den	$x_i$	$S_i = \frac{x_i}{x_0}$	$T_i = \frac{x_i}{x_{i-1}}$
0	Pá	4 000	[1]	/
1	So	1 000	0,25	0,25
2	Ne	1 000	0,25	1
	Po	?	$\Leftarrow 1\,000 \cdot 0,5$	

Průměrný koeficient vývoje:

$$\sqrt{T_1 \cdot T_2} = \sqrt{0,25 \cdot 1} = \sqrt{S_2} = \sqrt{0,25} = 0,5$$

Tedy na základě průměrného koeficientu vývoje bychom pro pondělí odhadovali **500** pasažérů, ovšem *selský rozum říká*, že pondělí je pracovní den a tedy bychom měli očekávat spíše **čtyři tisíce** přepravovaných osob.

Jak tedy dělat smysluplné odhady závislé na čase si ukážeme v **následující** kapitole.



## 2.2. Složené individuální indexy

Složené individuální indexy jsou indexy stejnorodého<sup>46</sup> extenzitního nebo intenzitního ukazatele, které používáme za situace, kdy hodnoty daného ukazatele jsou členěny na dílčí a v rámci výpočtu indexu provádíme shrnování dílčích hodnot. Tedy **porovnáváme údaje** (o množství, ceně, ...), které **vznikly součtem**. Vzhledem k poznámce 46 pak platí (sčítací index  $i$  z pohodlnosti opět uvedeme pouze u prvního výrazu):

$$I_{\Sigma Q} = I_{Q_s} = \frac{\sum Q_{1;i}}{\sum Q_{0;i}} \qquad I_{\Sigma q} = I_{q_s} = \frac{\sum q_1}{\sum q_0} \qquad (27)$$

$$I_{\bar{p}} = I_{p_s} = \frac{\bar{p}_1}{\bar{p}_0} = \frac{\frac{\sum Q_1}{\sum q_1}}{\frac{\sum Q_0}{\sum q_0}} = \frac{\frac{\sum (p_1 \cdot q_1)}{\sum q_1}}{\frac{\sum (p_0 \cdot q_0)}{\sum q_0}} = \frac{\sum (p_1 \cdot q_1) \cdot \sum q_0}{\sum (p_0 \cdot q_0) \cdot \sum q_1} \qquad (28)$$

Index  $I_{p_s}$  nazýváme *indexem proměnlivého složení*, protože na jeho velikost mají vliv jak změny intenzitní veličiny  $p$  (například ceny zboží v jednotlivých prodejnách), tak i změny extenzitní veličiny  $q$  (například množství prodaného zboží na jednotlivých prodejnách).

<sup>46</sup> Obecně lze říci, že: [11, str. 111]

- Ukazatel vyjadřující velikost určitého jevu bez vztahu k jinému jevu (časové průměry, zisk, přidaná hodnota apod.) je stejnorodý, má-li věcný smysl shrnovat jeho dílčí hodnoty součtem.
- Ukazatel vyjadřující velikost jednoho jevu na měrnou jednotku jiného jevu je stejnorodý tehdy,
  - když jsou stejnorodé ukazatele obou jevů, z nichž se skládá.
  - nebo když můžeme jeho dílčí hodnoty shrnovat průměrem.

Pokud toto neplatí, není ukazatel stejnorodý.

## Individuální indexy složené $\Rightarrow$ jeden ukazatel ve více střediscích

6 prodejen nabízí stejné zboží. K určitému datu každá prodejna upravila cenu tohoto konkrétního zboží, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

prodejna	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]			
	před	po	před	po		
	$p_0$	$p_1$	$q_0$	$q_1$		
A	2 621	2 622	2 705	2 702		
B	2 618	2 619	2 822	2 808		
C	2 960	2 955	2 658	2 670		
D	3 833	3 833	2 640	2 650		
E	2 682	2 690	2 720	2 695		
F	2 644	2 646	3 650	3 750		
$\Sigma$			17 195	17 275		

## Individuální indexy složené $\Rightarrow$ jeden ukazatel ve více střediscích

6 prodejen nabízí stejné zboží. K určitému datu každá prodejna upravila cenu tohoto konkrétního zboží, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

prodejna	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]		
	před	po	před	po	před	po	
	$p_0$	$p_1$	$q_0$	$q_1$	$Q_0 = p_0 \cdot q_0$	$Q_1 = p_1 \cdot q_1$	
A	2 621	2 622	2 705	2 702			
B	2 618	2 619	2 822	2 808			
C	2 960	2 955	2 658	2 670			
D	3 833	3 833	2 640	2 650			
E	2 682	2 690	2 720	2 695			
F	2 644	2 646	3 650	3 750			
$\Sigma$			17 195	17 275			

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky.

## Individuální indexy složené $\Rightarrow$ jeden ukazatel ve více střediscích

6 prodejen nabízí stejné zboží. K určitému datu každá prodejna upravila cenu tohoto konkrétního zboží, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

prodejna	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]		
	před	po	před	po	před	po	
	$p_0$	$p_1$	$q_0$	$q_1$	$Q_0 = p_0 \cdot q_0$	$Q_1 = p_1 \cdot q_1$	
A	2 621	2 622	2 705	2 702	7 089 805	7 084 644	
B	2 618	2 619	2 822	2 808	7 387 996	7 354 152	
C	2 960	2 955	2 658	2 670	7 867 680	7 889 850	
D	3 833	3 833	2 640	2 650	10 119 120	10 157 450	
E	2 682	2 690	2 720	2 695	7 295 040	7 249 550	
F	2 644	2 646	3 650	3 750	9 650 600	9 922 500	
$\Sigma$			17 195	17 275	49 410 241	49 658 146	

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky.

## Individuální indexy složené $\Rightarrow$ jeden ukazatel ve více střediscích

6 prodejen nabízí stejné zboží. K určitému datu každá prodejna upravila cenu tohoto konkrétního zboží, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

prodejna	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]		$p_1 \cdot q_0$
	před	po	před	po	před	po	
	$p_0$	$p_1$	$q_0$	$q_1$	$Q_0 = p_0 \cdot q_0$	$Q_1 = p_1 \cdot q_1$	
A	2 621	2 622	2 705	2 702	7 089 805	7 084 644	
B	2 618	2 619	2 822	2 808	7 387 996	7 354 152	
C	2 960	2 955	2 658	2 670	7 867 680	7 889 850	
D	3 833	3 833	2 640	2 650	10 119 120	10 157 450	
E	2 682	2 690	2 720	2 695	7 295 040	7 249 550	
F	2 644	2 646	3 650	3 750	9 650 600	9 922 500	
$\Sigma$			17 195	17 275	49 410 241	49 658 146	

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky. Potom ještě vyplníme pomocný sloupec  $p_1 \cdot q_0$  (pokud nás zajímá index **stálého složení**  $I_{ss}$  nebo index **struktury**  $I_{str}$ ).

## Individuální indexy složené $\Rightarrow$ jeden ukazatel ve více střediscích

6 prodejen nabízí stejné zboží. K určitému datu každá prodejna upravila cenu tohoto konkrétního zboží, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

prodejna	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]		$p_1 \cdot q_0$
	před	po	před	po	před	po	
	$p_0$	$p_1$	$q_0$	$q_1$	$Q_0 = p_0 \cdot q_0$	$Q_1 = p_1 \cdot q_1$	
A	2 621	2 622	2 705	2 702	7 089 805	7 084 644	7 092 510
B	2 618	2 619	2 822	2 808	7 387 996	7 354 152	7 390 818
C	2 960	2 955	2 658	2 670	7 867 680	7 889 850	7 854 390
D	3 833	3 833	2 640	2 650	10 119 120	10 157 450	10 119 120
E	2 682	2 690	2 720	2 695	7 295 040	7 249 550	7 316 800
F	2 644	2 646	3 650	3 750	9 650 600	9 922 500	9 657 900
$\Sigma$			17 195	17 275	49 410 241	49 658 146	49 431 538

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky. Potom ještě vyplníme pomocný sloupec  $p_1 \cdot q_0$  (pokud nás zajímá index **stálého složení**  $I_{ss}$  nebo index **struktury**  $I_{str}$ ).

## Individuální indexy složené $\Rightarrow$ jeden ukazatel ve více střediscích

6 prodejen nabízí stejné zboží. K určitému datu každá prodejna upravila cenu tohoto konkrétního zboží, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

prodejna	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]		$p_1 \cdot q_0$
	před	po	před	po	před	po	
	$p_0$	$p_1$	$q_0$	$q_1$	$Q_0 = p_0 \cdot q_0$	$Q_1 = p_1 \cdot q_1$	
A	2 621	2 622	2 705	2 702	7 089 805	7 084 644	7 092 510
B	2 618	2 619	2 822	2 808	7 387 996	7 354 152	7 390 818
C	2 960	2 955	2 658	2 670	7 867 680	7 889 850	7 854 390
D	3 833	3 833	2 640	2 650	10 119 120	10 157 450	10 119 120
E	2 682	2 690	2 720	2 695	7 295 040	7 249 550	7 316 800
F	2 644	2 646	3 650	3 750	9 650 600	9 922 500	9 657 900
$\Sigma$			17 195	17 275	49 410 241	49 658 146	49 431 538

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky. Potom ještě vyplníme pomocný sloupec  $p_1 \cdot q_0$  (pokud nás zajímá index **stálého složení**  $I_{ss}$  nebo index **struktury**  $I_{str}$ ).

$$\text{Index hodnoty: } I_{Q_s} = \frac{\sum Q_1}{\sum Q_0} = \frac{49\,658\,146}{49\,410\,241} \doteq 1,005 \quad \text{Index množství: } I_{q_s} = \frac{\sum q_1}{\sum q_0} = \frac{17\,275}{17\,195} \doteq 1,005$$

$$I_{ss} = \frac{\sum(p_1 \cdot q_0)}{\sum(p_0 \cdot q_0)} = \frac{49\,431\,538}{49\,410\,241} \doteq 1,001 \quad I_{str} = \frac{\sum(p_1 \cdot q_1) \cdot \sum q_0}{\sum(p_1 \cdot q_0) \cdot \sum q_1} = \frac{49\,658\,146 \cdot 17\,195}{49\,431\,538 \cdot 17\,275} \doteq 0,995$$

$$\text{Index proměnlivého složení: } I_{ps} = \frac{\sum(p_1 \cdot q_1) \cdot \sum q_0}{\sum(p_0 \cdot q_0) \cdot \sum q_1} = I_{ss} \cdot I_{str} = \frac{49\,658\,146 \cdot 17\,195}{49\,410\,241 \cdot 17\,275} \doteq 0,996$$

Pro jednotlivé prodejny máme:

$$I_p = \frac{p_1(E)}{p_0(E)} = \frac{2\,690}{2\,682} \doteq 1,003 \quad \Rightarrow \quad \text{cena vzrostla o tři desetiny procenta}$$

$$\text{např. E} \quad I_q = \frac{q_1(E)}{q_0(E)} = \frac{2\,695}{2\,720} \doteq 0,991 \quad \Rightarrow \quad \text{prodej klesl o devět desetin procenta}$$

$$I_Q = \frac{Q_1(E)}{Q_0(E)} = \frac{7\,249\,550}{7\,295\,040} \doteq 0,994 \quad \Rightarrow \quad \text{tržby klesly o šest desetin procenta}$$

### A celkově

$$I_{ps} \doteq 0,996 \quad \Rightarrow \quad \text{průměrná cena jednoho výrobku klesla o čtyři desetiny procenta}$$

$$I_{qs} \doteq 1,005 \quad \Rightarrow \quad \text{prodej v celé firmě vzrostl o pět desetin procenta}$$

$$I_{Qs} \doteq 1,005 \quad \Rightarrow \quad \text{objem tržeb celé firmy vrostl o pět desetin procenta, z toho: v důsledku změn ceny daného výrobku na jednotlivých pobočkách sice poklesl (kdy průměrná cena výrobku } I_{ps} \text{ klesla ve firmě přibližně o čtyři desetiny procenta), ale v důsledku změn v prodeji (počtu prodaných kusů) na pobočkách celkově vzrostl (kdy prodej v celé firmě } I_{qs} \text{ vzrostl přibližně o pět desetin procenta).}$$



### 3. Souhrnné (agregátní) indexy

**Souhrnné indexy** množství a úrovně jsou indexy **nestejnorodých** extenzitních a intenzitních veličin.

Pro nestejnorodé veličiny je charakteristické, že je nelze sčítat (ani když jsou vyjádřené ve stejných měrných jednotkách), ale nelze je ani průměrovat.

Používají se za situace, kdy nelze sestavit indexy extenzitních ukazatelů (27), případně index proměnlivého složení (28) z důvodu nemožnosti sestavit veličinu  $q$  nebo  $Q$  (například nelze určit průměrnou cenu pro skupinu různých výrobků).

Základem koncepce souhrnných indexů je myšlenka průměrování změn (vyjádřených jednoduchými indexy) dílčích hodnot sledovaného ukazatele. V případě cenových indexů se zřejmě jedná o průměrování indexů cen jednotlivých výrobků s tím, že jako váhy vystupuje hodnota produkce ze základního období (situace 0), nebo z běžného období (situace 1).

Jednou z možností je použití váženého aritmetického průměru individuálních jednoduchých indexů cen, kde jako váhy použijeme strukturu produkce ze základního období. Obdržíme pak průměrovaný tvar již dříve zmiňovaného **Laspeyresova indexu**  ${}_L I_p$  [11, 115], který po úpravě také nazýváme **Laspeyresův cenový index** a označujeme  $I_c$ .

$${}_L I_p = \frac{\sum(I_p \cdot p_0 \cdot q_0)}{\sum(p_0 \cdot q_0)} \stackrel{(26)}{=} \frac{\sum(\frac{p_1}{p_0} \cdot p_0 \cdot q_0)}{\sum(p_0 \cdot q_0)} = \frac{\sum(p_1 \cdot q_0)}{\sum(p_0 \cdot q_0)} = I_c$$

Budeme-li analogicky postupovat při změnách objemu různorodé produkce, dostaneme **Laspeyresův objemový index**  $I_o$ .

$$\text{Laspeyresův objemový index } I_o = \frac{\sum(p_0 \cdot q_1)}{\sum(p_0 \cdot q_0)} \quad \text{Souhrnný hodnotový index } I_h = \frac{\sum(p_1 \cdot q_1)}{\sum(p_0 \cdot q_0)}$$

## Souhrnné indexy $\Rightarrow$ více ukazatelů

Prodejna nabízí *stejně* (= srovnatelné) zboží od 6 výrobců. K určitému datu prodejna upravila ceny, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

výrobce	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]					
	před	po	před	po				
	$p_0$	$p_1$	$q_0$	$q_1$				
A	2 621	2 622	2 705	2 702				
B	2 618	2 619	2 822	2 808				
C	2 960	2 955	2 658	2 670				
D	3 833	3 833	2 640	2 650				
E	2 682	2 690	2 720	2 695				
F	2 644	2 646	3 650	3 750				
$\Sigma$			17 195	17 275				

## Souhrnné indexy $\Rightarrow$ více ukazatelů

Prodejna nabízí *stejně* (= srovnatelné) zboží od 6 výrobců. K určitému datu prodejna upravila ceny, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

výrobce	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]			
	před	po	před	po	před	po		
	$p_0$	$p_1$	$q_0$	$q_1$	$p_0 \cdot q_0$	$p_1 \cdot q_1$		
A	2 621	2 622	2 705	2 702				
B	2 618	2 619	2 822	2 808				
C	2 960	2 955	2 658	2 670				
D	3 833	3 833	2 640	2 650				
E	2 682	2 690	2 720	2 695				
F	2 644	2 646	3 650	3 750				
$\Sigma$			17 195	17 275				

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky.

## Souhrnné indexy $\Rightarrow$ více ukazatelů

Prodejna nabízí *stejně* (= srovnatelné) zboží od 6 výrobců. K určitému datu prodejna upravila ceny, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

výrobce	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]			
	před	po	před	po	před	po		
	$p_0$	$p_1$	$q_0$	$q_1$	$p_0 \cdot q_0$	$p_1 \cdot q_1$		
A	2 621	2 622	2 705	2 702	7 089 805	7 084 644		
B	2 618	2 619	2 822	2 808	7 387 996	7 354 152		
C	2 960	2 955	2 658	2 670	7 867 680	7 889 850		
D	3 833	3 833	2 640	2 650	10 119 120	10 157 450		
E	2 682	2 690	2 720	2 695	7 295 040	7 249 550		
F	2 644	2 646	3 650	3 750	9 650 600	9 922 500		
$\Sigma$			17 195	17 275	49 410 241	49 658 146		

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky.

## Souhrnné indexy $\Rightarrow$ více ukazatelů

Prodejna nabízí *stejně* (= srovnatelné) zboží od 6 výrobců. K určitému datu prodejna upravila ceny, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

výrobce	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]			
	před	po	před	po	před	po		
	$p_0$	$p_1$	$q_0$	$q_1$	$p_0 \cdot q_0$	$p_1 \cdot q_1$		
A	2 621	2 622	2 705	2 702	7 089 805	7 084 644		
B	2 618	2 619	2 822	2 808	7 387 996	7 354 152		
C	2 960	2 955	2 658	2 670	7 867 680	7 889 850		
D	3 833	3 833	2 640	2 650	10 119 120	10 157 450		
E	2 682	2 690	2 720	2 695	7 295 040	7 249 550		
F	2 644	2 646	3 650	3 750	9 650 600	9 922 500		
$\Sigma$			17 195	17 275	49 410 241	49 658 146		

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky. Potom ještě vyplníme pomocný sloupec  $p_1 \cdot q_0$

## Souhrnné indexy $\Rightarrow$ více ukazatelů

Prodejna nabízí *stejně* (= srovnatelné) zboží od 6 výrobců. K určitému datu prodejna upravila ceny, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

výrobce	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]			
	před	po	před	po	před	po		
	$p_0$	$p_1$	$q_0$	$q_1$	$p_0 \cdot q_0$	$p_1 \cdot q_1$		
A	2 621	2 622	2 705	2 702	7 089 805	7 084 644	7 092 510	7 081 942
B	2 618	2 619	2 822	2 808	7 387 996	7 354 152	7 390 818	7 351 344
C	2 960	2 955	2 658	2 670	7 867 680	7 889 850	7 854 390	7 903 200
D	3 833	3 833	2 640	2 650	10 119 120	10 157 450	10 119 120	10 157 450
E	2 682	2 690	2 720	2 695	7 295 040	7 249 550	7 316 800	7 227 990
F	2 644	2 646	3 650	3 750	9 650 600	9 922 500	9 657 900	9 915 000
$\Sigma$			17 195	17 275	49 410 241	49 658 146	49 431 538	49 636 926

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky. Potom ještě vyplníme pomocný sloupec  $p_1 \cdot q_0$

## Souhrnné indexy $\Rightarrow$ více ukazatelů

Prodejna nabízí *stejně* (= srovnatelné) zboží od 6 výrobců. K určitému datu prodejna upravila ceny, což se projevilo na počtu prodaných kusů. Spočítejte vhodné **indexy**.

Následující údaje máme k dispozici za stejný časový úsek PŘED a PO úpravě ceny.

výrobce	Cena $p$ [Kč/kus]		Prodej $q$ [kusy]		Tržby $Q$ [Kč]			
	před	po	před	po	před	po		
	$p_0$	$p_1$	$q_0$	$q_1$	$p_0 \cdot q_0$	$p_1 \cdot q_1$		
A	2 621	2 622	2 705	2 702	7 089 805	7 084 644	7 092 510	7 081 942
B	2 618	2 619	2 822	2 808	7 387 996	7 354 152	7 390 818	7 351 344
C	2 960	2 955	2 658	2 670	7 867 680	7 889 850	7 854 390	7 903 200
D	3 833	3 833	2 640	2 650	10 119 120	10 157 450	10 119 120	10 157 450
E	2 682	2 690	2 720	2 695	7 295 040	7 249 550	7 316 800	7 227 990
F	2 644	2 646	3 650	3 750	9 650 600	9 922 500	9 657 900	9 915 000
$\Sigma$			17 195	17 275	49 410 241	49 658 146	49 431 538	49 636 926

Dopočítáme tržby ( $Q = p \cdot q$ ) a zapíšeme je do tabulky. Potom ještě vyplníme pomocný sloupec  $p_1 \cdot q_0$  a určíme požadované indexy.

$$\text{Cenový i.: } I_c = \frac{\Sigma(p_1 \cdot q_0)}{\Sigma(p_0 \cdot q_0)} = \frac{49\,431\,538}{49\,410\,241} \doteq 1,001 \quad \text{Objemový i.: } I_o = \frac{\Sigma(p_0 \cdot q_1)}{\Sigma(p_0 \cdot q_0)} = \frac{49\,636\,926}{49\,410\,241} \doteq 1,005$$

$$\text{Hodnotový index: } I_h = \frac{\Sigma(p_1 \cdot q_1)}{\Sigma(p_0 \cdot q_0)} = \frac{49\,658\,146}{49\,410\,241} \doteq 1,005$$

Pro jednotlivé výrobce máme:

$$I_p = \frac{p_1(F)}{p_0(F)} = \frac{2\,646}{2\,644} \doteq 1,001 \quad \Rightarrow \quad \text{cena vzrostla o jednu desetinu procenta}$$

např. **F**    $I_q = \frac{q_1(F)}{q_0(F)} = \frac{3\,750}{3\,650} \doteq 1,027 \quad \Rightarrow \quad \text{prodej vzrostl o dvě celé sedm desetin procenta}$

$$I_Q = \frac{Q_1(F)}{Q_0(F)} = \frac{9\,922\,500}{9\,650\,600} \doteq 1,028 \quad \Rightarrow \quad \text{tržby vzrostly o dvě celé osm desetin procenta}$$

A celkově

$$I_c \doteq 1,001 \quad \Rightarrow \quad \text{vlivem změn v úrovni jednotkových cen celkové tržby vzrostly přibližně o jednu desetinu procenta}$$

$$I_o \doteq 1,005 \quad \Rightarrow \quad \text{vlivem změn v prodaném množství celkové tržby vzrostly přibližně o pět desetin procenta}$$

$$I_h \doteq 1,005 \quad \Rightarrow \quad \text{vlivem obou příčin celkové tržby vzrostly přibližně o pět desetin procenta}$$



## Přehledné uspořádání pojmů

**Statistický ukazatel** je číslo, které v daném prostoru a čase charakterizuje určitou skutečnost (určitý jev).

**Bazický index** porovnává konkrétní ukazatel (například tržby v jednom období) vždy se zvoleným (nultým) ukazatelem (většinou za bázi, tj. nultý ukazatel, volíme počáteční ukazatel, který je k dispozici). Pro  $i = 1, 2, \dots, n$  jej můžeme vyjádřit ve tvaru:

$$S_i = \frac{x_i}{x_0}$$

**Řetězový index** porovnává vždy dva sousední ukazatele. Pro  $i = 1, 2, \dots, n$  jej můžeme vyjádřit ve tvaru:

$$T_i = \frac{x_i}{x_{i-1}}$$

**Průměrný koeficient vývoje** je vývoj sledovaného ukazatele v čase vyjádřený **geometrickým průměrem řetězových indexů**. Pro  $i = 1, 2, \dots, n$  jej můžeme vyjádřit ve tvaru:

$$\bar{x}_G = \sqrt[n]{T_1 \cdot T_2 \cdot \dots \cdot T_n} = \sqrt[n]{\frac{x_1}{x_0} \cdot \frac{x_2}{x_1} \cdot \dots \cdot \frac{x_n}{x_{n-1}}} = \sqrt[n]{\frac{x_n}{x_0}} = \sqrt[n]{S_n}$$

**Index hodnoty**  $I_Q$  (například tržby).

**Index množství**  $I_q$  (jednoho konkrétního ukazatele).

**Index úrovně**  $I_p$  (například ceny jednoho konkrétního zboží).

Pro sumy kvůli přehlednosti opět použijeme stručný zápis s vynecháním symbolů, přes které sčítáme. Tedy například místo  $\sum_{i=1}^n Q_{1;i}$  budeme psát jen  $\sum Q_1$ .

	index <b>hodnoty</b>	index <b>množství</b>	index <b>úrovně</b>	
Individ. jednoduché i.	$I_Q = \frac{Q_1}{Q_0} = \frac{p_1 \cdot q_1}{p_0 \cdot q_0}$	$I_q = \frac{q_1}{q_0}$	$I_p = \frac{p_1}{p_0}$	$I_Q = I_q \cdot I_p$
Individuální			$I_{ps} = \frac{\sum(p_1 \cdot q_1) \cdot \sum q_0}{\sum(p_0 \cdot q_0) \cdot \sum q_1}$	index proměnlivého složení
složené	$I_{Qs} = \frac{\sum Q_1}{\sum Q_0}$	$I_{qs} = \frac{\sum q_1}{\sum q_0}$	$I_{ss} = \frac{\sum(p_1 \cdot q_0)}{\sum(p_0 \cdot q_0)}$	index stálého složení
indexy			$I_{str} = \frac{\sum(p_1 \cdot q_1) \cdot \sum q_0}{\sum(p_1 \cdot q_0) \cdot \sum q_1}$	index struktury $I_{ps} = I_{ss} \cdot I_{str}$
(dle Laspeyrese)	<b>hodnotový index</b>	<b>L. cenový index</b>	<b>L. objemový index</b>	
Souhrnný index	$I_h = \frac{\sum(p_1 \cdot q_1)}{\sum(p_0 \cdot q_0)}$	$I_c = \frac{\sum(p_1 \cdot q_0)}{\sum(p_0 \cdot q_0)}$	$I_o = \frac{\sum(p_0 \cdot q_1)}{\sum(p_0 \cdot q_0)}$	

## 4.1. Příklady používaných indexů v praxi

Cenové indexy patří k nejstarším oficiálně sledovaným indexům. Potřeba zachytit cenový vývoj v různých případech vedla nakonec k vytvoření tak zvané cenové statistiky. V České republice se v oblasti cenové statistiky používají souhrnné Laspeyeresovy cenové indexy s vahami, které jsou stále po celou dobu mezi revizemi cen. Soubor reprezentantů a váhový systém tvoří tak zvaný **spotřební koš**.

**Index spotřebitelských cen** je v současné době počítán na základě souboru 775 reprezentantů. Počet reprezentantů je kompromisem mezi přesností a náklady na průzkum. Nový revidovaný spotřební koš je založen na souboru vybraných druhů zboží a služeb, které se významně podílejí na výdajích obyvatelstva a svým rozsahem pokrývají celou sféru spotřeby s vahami roku 1999. Zpravodajskými jednotkami jsou rozdílné typy prodejen a provozoven služeb z hlediska velikosti, druhu, vlastnictví apod. — zhruba 10 tisíc.

Index spotřebitelských cen je konstruován ve tvaru:

$$I_p = \frac{\sum \left[ \frac{p_1}{p_0} \cdot (p_0 \cdot q_0) \right]}{\sum (p_0 \cdot q_0)}$$

kde výraz  $p_0 \cdot q_0$  představuje stálé váhy — výdaje domácností za zboží (službu) v základním období.

**Index životních nákladů** vyjadřuje, jak se index spotřebitelských cen promítá do výdajů domácností. Index životních nákladů je počítán pro následující sociální skupiny:

- domácnosti celkem;
- domácnosti zaměstnanců;

- domácnosti důchodců;
- domácnosti s dětmi v nízkém příjmovém pásmu;
- domácnosti žijící v hlavním městě Praze.

**Měření inflace** je založeno na indexu spotřebitelských cen. Základní mírou inflace je roční míra inflace, která klouzavě srovnává průměr posledních 12 měsíců s průměrem předcházejících měsíců.

V České republice jsou publikovány tyto míry inflace:

**Měsíční tempo inflace** (což je cenový index) srovnává úroveň cen v hodnoceném měsíci a v měsíci předcházejícím:

$$M_t = \left( \frac{I_t}{I_{t-1}} - 1 \right) \cdot 100$$

kde  $I_t$  je bazický index spotřebitelských cen ve sledovaném měsíci a  $I_{t-1}$  je bazický index spotřebitelských cen v měsíci předcházejícím. Báze je cena v prosinci roku 1999.

**Meziroční tempo inflace** srovnává úroveň cen v hodnoceném měsíci a ve stejném měsíci předcházejícího roku:

$$M_t = \left( \frac{I_t}{I_{t-12}} - 1 \right) \cdot 100$$

**Roční tempo inflace** srovnává úroveň cen v posledních 12 měsících a ve 12 měsících předcházejících:

$$M_t = \left( \frac{\sum_{i=t-11}^t I_i}{\sum_{j=t-23}^{t-12} I_j} - 1 \right) \cdot 100$$

**Jádrová inflace** vyjadřuje měsíční přírůstek indexu spotřebitelských cen počítaný na celém spotřebním koši po vyloučení vlivu změn ovlivněných regulovanými cenami, daňovými úpravami a jinými administrativními opatřeními.

**Čistá inflace** je počítána na neúplném spotřebním koši, z něhož jsou vyloučeny položky s regulovanými cenami a cenami ovlivněnými administrativními opatřeními, ale položky, u nichž jsou změny cen způsobené daňovými úpravami, zůstávají ve spotřebním koši. Pouze je eliminován vliv daňových úprav.

**Indexy kurzů akcií** představují zvláštní typ cenových indexů. Ne každý index kurzu akcií, se kterým se můžete setkat v praxi jednotlivých zemí, je indexem konstruovaným ve výše uvedené smyslu. V praxi používané indexy kurzů akcií se liší svou konstrukcí, ale i trhem, pro který jsou sestavovány.

Z hlediska konstrukce se používají buď jako aritmetický nebo harmonický či geometrický průměr. Ať již jako prostý průměr nebo jako vážený průměr indexů kurzů akcií. Jedná se buď o

**kurzově** (cenově) vážené průměry — sledují stav a vývoj průměrné ceny titulu akcie;  
nebo o

**tržně** vážené průměry — sledují průměrnou cenu akcie z celkového objemu emitovaných akcií.

# Úvod do Časových řad

## Obsah kapitoly: Časové řady

<b>1. Základní pojmy</b>	<b>340</b>
1.1. Základní charakteristiky dynamiky vývoje časových řad . . . . .	346
<b>2. Vyrovnání časových řad</b>	<b>348</b>
2.1. Problémy při analýze časových řad . . . . .	351
<b>3. Modelování časových řad — trend</b>	<b>351</b>
3.1 Lineární trend . . . . .	353
3.2 Kvadratický trend . . . . .	354
Trendy — příklady . . . . .	355
<b>4. Závěr kapitoly – Využití programového vybavení</b>	<b>374</b>

# 1. Základní pojmy

**Časovou řadou** (dynamickou ř., vývojovou ř.) rozumíme posloupnost věcně a prostorově srovnatelných dat, která jsou jednoznačně uspořádána z hlediska času ve směru „minulost → přítomnost“.

Časové řady upoutávají více než poměrná čísla nebo nehybná rozdělení četností, protože vnášejí dimenzi času. Ukazují několika čarami nebo čísly vývoj, který jsme zpravidla jen nejasně tušili.

Přesto není rozdíl mezi časovou řadou a jednotlivými statistickými výběry nebo vyčerpávajícím šetřením. Stejně jako se film skládá z jednotlivých nehybných obrázků, je i časová řada složena z takových jednotlivých snímků.

Časové řady v zásadě vytvářejí spojení mezi **stejnorodými**<sup>47</sup> údaji (zjištěními, výpověďmi) z různých dob, avšak stejného věcného obsahu. Může jít nejen o plynulá porovnávání (roční dovozy a vývozy za posledních  $x$  let) ale i o porovnání jednotlivých vybraných údajů, jako je například struktura povolání ve Švýcarsku v letech 1888, 1900, 1910, 1920, 1930, 1941, 1950 a 1960 (obrázek 8).

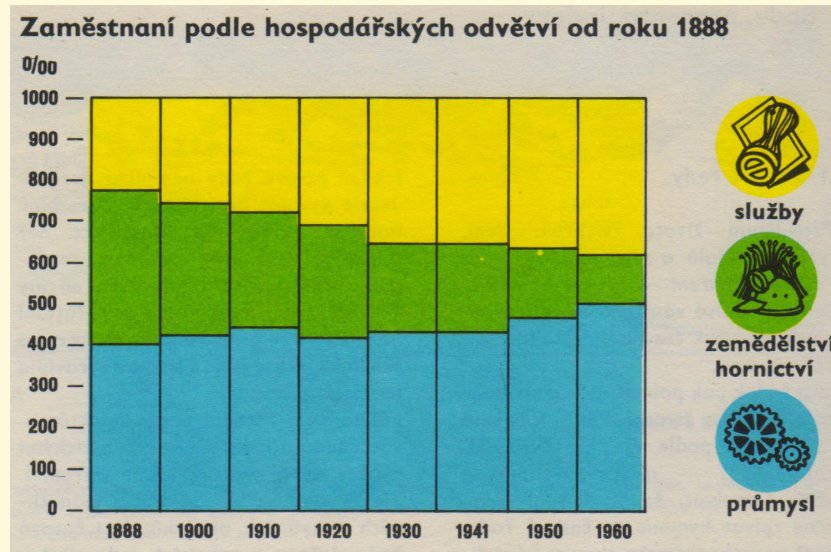
Plynulá pozorování nejsou však často vůbec možná. Časová řada z výsledků hromadných sčítání lidu nejenže přeskakuje velká (většinou desetiletá) období, ale kromě toho poskytuje jen bodové údaje, které přísně vzato platily jen v okamžiku odevzdání sčítacího lístku. Proto je třeba rozlišovat mezi časovými řadami **okamžikovými**, kdy se hodnoty ukazatele (statistického znaku) vztahují k určitému okamžiku, a časovými řadami **intervalovými**, kdy hodnoty ukazatele jsou sledovány za určité období (v určitém časovém intervalu) a jsou proto délkou tohoto období ovlivněny.

Zatímco údaje okamžikových řad lze zjišťovat pouze k rozhodnému dni (při posledním sčítání bylo tolik mužů a tolik žen, tolik rodin, tolik nezletilých dětí apod.), údaje intervalových řad musejí být naproti tomu zjišťovány a srovnávány za určité období. Pokud bychom sčítali sňatky minulou sobotu úderem

<sup>47</sup> Sledujeme-li například počty krádeží v dané oblasti (okres, kraj) za delší časový úsek, je možné, že v určitém období zaregistrujeme jejich náhlou změnu. Ta ovšem může způsobena jen tím, že zákonem byla změněna hodnota minimální způsobené škody nutné k zahrnutí mezi krádeže. Nebo mohlo dojít v rámci reformy státní správy ke změně rozsahu (sloučení či rozdělení) sledované oblasti.



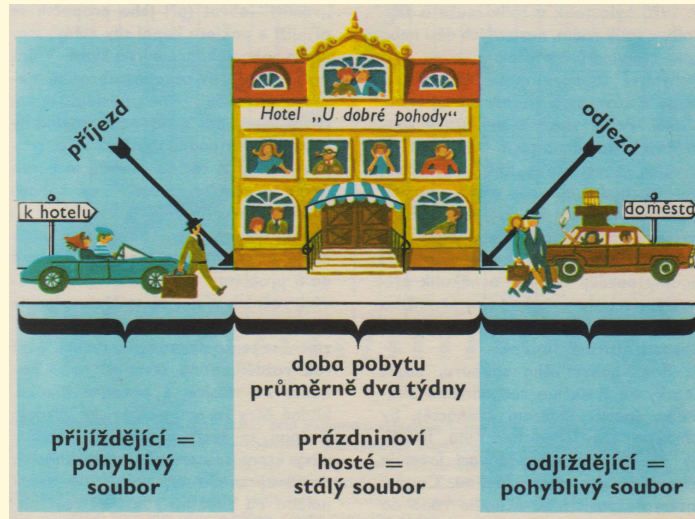
Obrázek 8: Převzat z [14]



dvanácté, dostali bychom téměř jistě nulu, ledaže by někde připadla oddávající formule přesně na poledne. Ovšem časový interval od 8 hodin ráno do 20 hodin večer poskytuje celkem rozumnou srovnávací hodnotu pro počet sňatků za jeden den.

**Okamžikové řady (*stálé soubory*)** jsou takové, jejichž prvky (hodnoty ukazatele) se plynule mění v čase a mají určitou dobu trvání. Například obyvatelstvo nějakého území. Jednotlivci se rodí a umírají — celek obyvatelstva je tím z dlouhodobého hlediska dotčen jen tehdy, když trvá zřejmá převaha narození či úmrtí. Nebo počet automobilů, které firma vlastní k určitému datu. Každý prvek stálého souboru má určitou dobu „setrvání“ — u lidí je to individuální délka života, u náhradního dílu setrvání ve skladu dílny a u hosta na dovolené doba pobytu v prázdninovém hotelu (obrázek 9).

Obrázek 9: Převzat z [14]



Častěji než „pohyblivý soubor“ používáme termín **intervalová časová řada**

a místo „stálý soubor“ říkáme **okamžiková časová řada**.

V případě okamžikových časových řad nemá součet hodnot znaku věcný smysl (například nemá význam sčítat počty zaměstnanců zjištěné vždy v první středě kalendářního měsíce). Ovšem má smysl vyjádřit průměrnou úroveň hodnot. K tomu využíváme **chronologický průměr**<sup>48</sup>. Tímto jediným číslem pak charakterizujeme úroveň ukazatele za celé období. Je ale zřejmé, že tím dochází ke značnému zjednodušování reality.

Oblíbenější jsou proto různé druhy klouzavých ukazatelů, které jsou schopny částečně eliminovat vliv náhodných vlivů na sledovaný ukazatel a tím časovou řadu „vyhladit“. Používají se jak **klouzavé**

<sup>48</sup> **Není-li krok** (délka mezi jednotlivými časovými okamžiky  $t_i$ ) **ekvidistantní** (délky  $d_i$  jednotlivých časových intervalů nejsou stejné), používáme **vážený chronologický průměr**:

$$\bar{x}_{Ch} = \frac{\frac{x_1+x_2}{2} \cdot (t_2 - t_1) + \frac{x_2+x_3}{2} \cdot (t_3 - t_2) + \dots + \frac{x_{n-1}+x_n}{2} \cdot (t_n - t_{n-1})}{t_n - t_1} = \frac{\frac{x_1+x_2}{2} \cdot d_1 + \frac{x_2+x_3}{2} \cdot d_2 + \dots + \frac{x_{n-1}+x_n}{2} \cdot d_{n-1}}{d_1 + d_2 + \dots + d_{n-1}}$$

**mediány**, tak **klouzavé průměry**. Vždy se postupuje tak, že údaj časové řady nahradíme zvoleným ukazatelem z okolních časově předcházejících a následujících údajů.

Jaký to má smysl? Například při sledování prodeje je pravidelně údaj za daný měsíc v některých obdobích zvlášť velký (nápoje či mraženého zboží v letních měsících) a v jiných zase pravidelně menší. Objevují se sezónní výkyvy. Vzájemným srovnáváním údajů pro různé měsíce nezískáme pak přehled o tom, zda dochází ke skutečné změně nebo jenom změně vyvolané sezónním výkyvem. Jestliže však srovnáváme pro dané měsíce součty vždy za posledních 12 měsíců, má v sobě každý tento klouzavý roční úhrn zahrnutý všechny sezónní výkyvy v roce a můžeme pak na nich pozorovat skutečné nárůsty či poklesy prodeje.

**Příklad:** V jistém podniku v lednu (má 31 kalendářních dnů) pracovalo 280 zaměstnanců, v únoru (28 dnů) 270, v březnu (31 dnů) 280, v dubnu (30 dnů) 250 a v květnu (31 dnů) 240. Určete průměrný stav zaměstnanců v tomto podniku za uvedených 5 měsíců když víte, že podnik propouští k poslednímu dni v měsíci a přijímá k prvnímu dni daného měsíce.

Přepišme napřed údaje do přehledné tabulky

měsíc	index $i$	zaměstnanci $x_i$	počet dnů $d_i$
leden	1	280	31
únor	2	270	28
březen	3	280	31
duben	4	250	30
květen	5	240	31

a protože měsíce nemají stejný počet dnů  
zadaná data poté dosadíme do vzorce pro  
**vážený chronologický průměr:**

$$\bar{x}_{Ch} = \frac{\frac{x_1+x_2}{2} \cdot d_1 + \frac{x_2+x_3}{2} \cdot d_2 + \frac{x_3+x_4}{2} \cdot d_3 + \frac{x_4+x_5}{2} \cdot d_4}{d_1 + d_2 + d_3 + d_4} =$$

$$= \frac{\frac{280+270}{2} \cdot 31 + \frac{270+280}{2} \cdot 28 + \frac{280+250}{2} \cdot 31 + \frac{250+240}{2} \cdot 30}{31 + 28 + 31 + 30} = 265$$

Můžeme tedy říci, že v daném podniku od ledna do května pracovalo každý měsíc průměrně 265 zaměstnanců.

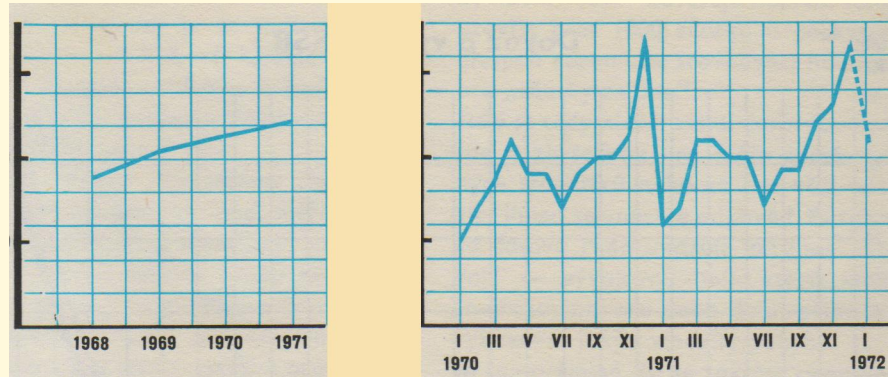
**Intervalové řady (*pohyblivé soubory*)** jsou soubory událostí. Vznikající události se dají měřit jen tím způsobem, že se sčítají jevy vzniklé během daného období. Například počet kusů zboží vyrobeného za daný měsíc. Také oba jevy, které vymezují dobu setrvání, lze pokládat za pohyblivé soubory: narození a smrt, příjem na sklad a výdej ze skladu, příjezd a odjezd návštěvníků (viz obrázek 9), nákup auta do firmy a odprodej auta, atd.

V případě intervalových řad již má smysl jejich sčítání (sečteme-li tržby od pondělí do neděle, získáme týdenní tržbu) a význam má i průměrná hodnota, většinou vyjádřená pomocí aritmetického průměru.

Pro intervalové řady ovšem musíme zajistit jejich **srovnatelnost** a to jak časovou (intervaly musejí být stejně dlouhé), tak prostorovou (údaje – data musejí pocházet ze „stejně velkých území“). V případě, že tomu tak není a údaje v sobě nesou zkreslení, provádíme tzv. **vyrovnaní** či očištění časové řady.

Časová řada ukazuje většinou vývoj – vývojovou linii. Ovšem pokud vezmeme obrat obchodního domu za posledních deset let, bude asi vykazovat stoupající tendenci, což ale nemusí mnoho znamenat. Je přece docela dobře myslitelné, že růst bude způsoben inflací, zatímco „reálný“ obrat (při jeho propočtu se přihlíží k poklesu kupní síly měny) stagnuje nebo dokonce mírně klesá. Z těchto důvodů je nutno údaje o obratu „očistit“ od tohoto rušivého faktoru. Zatímco křivka celkových ročních obrátů směřuje plynule (alespoň jak jsme se bez uvažování inflace domnívali) „nahoru“, změni se tento obraz velmi rychle, když rok rozdělíme na čtvrtletí nebo dokonce měsíce. Z původně hladké křivky se stane „divoce“ lomená čára (viz obrázek 10).

Obrázek 10: Převzat z [14]



Začnou se totiž projevovat všechny vlivy, které jsou pro obchody zpravidla typické: mdlá kupní nálada na počátku roku, sezónní výprodeje, předvánoční obchodní ruch apod.

Svátky a vliv počasí zavádějí vliv nepravidelnosti nejen do obratu obchodních řetězců, ale mohou vyvolat velký zmatek především v číslech statistik cestovního ruchu, které mohou od března jednoho roku k březnu druhého roku stejně jako od dubna jednoho roku k dubnu druhého roku vykazovat podivuhodné skoky díky velikonočním a jarním prázdninám, což obojí je pohyblivá událost.

Jestliže jsou k dispozici pozorování za dostatečně dlouhá časová období, je možné v některých případech postřehnout **cyklus**. Jak dlouhé musí být časové období, aby se dal určitý cyklus postřehnout, nelze obecně říci. To musí vyplynout ze získaných údajů. Podle periodicity (délky cyklu) lze časové řady dělit na **krátkodobé**, kdy perioda je kratší než jeden rok (počet smluv uzavřených během týdne, slapová dmoutí moře, ...), a **dlouhodobé**, kdy perioda je alespoň jeden rok (roční zisk firmy).

Zřetelný cyklus v průběhu přibližně 24 hodin představuje mořský příliv a odliv, který je možné zjistit řadou hodinových nebo dvouhodinových intervalů. Kdybychom však měřili stav vody na pobřeží každý čtvrtek přesně v pravé poledne, trvalo by asi velmi dlouho, než bychom z těchto měření mohli učinit správný závěr.

## 1.1. Základní charakteristiky dynamiky vývoje časových řad

Dynamikou vývoje časové řady rozumíme změny hodnot sledovaného ukazatele v čase. Nutnou podmínkou pro správnou interpretaci charakteristik jsou **ekvidistantní** časové intervaly (mají stejnou délku).

**Absolutní přírůstek**  $\Delta_t^{(1)}$  (někdy též **1. difference**) je rozdíl mezi hodnotou znaku v čase  $t$  a v čase předcházejícím:

$$\Delta_t^{(1)} = y_t - y_{t-1} \quad \text{kde} \quad t = 2, 3, 4, \dots$$

Hodnoty prvních diferencí nějakého ukazatele jsou nositelem důležité informace. Pokud se totiž jednotlivé členy této posloupnosti systematicky ani nezvětšují ani nezmenšují (můžeme říci, že jejich hodnoty pouze náhodně a „ne příliš“ kolísají), lze u původní časové řady předpokládat **lineární trend**. Hodnoty ukazatele  $Y$  časové řady budou ležet téměř na přímce, neboli jsou lineárně závislé v čase. Viz povídání o **lineární závislosti** v kapitole zabývající se **regresními vztahy** mezi dvourozměrnými daty. Tehdy jsme pouze nepoužívali slovíčko *absolutní*.

**Relativní přírůstek**  $\delta_t$  je podíl, kdy absolutní přírůstek dělíme hodnotou znaku v čase předcházejícím:

$$\delta_t = \frac{\Delta_t^{(1)}}{y_{t-1}} = \frac{y_t - y_{t-1}}{y_{t-1}} \quad \text{kde} \quad t = 2, 3, 4, \dots$$

Z hodnot relativních přírůstků můžeme usuzovat (proč si ukážeme u dalších charakteristiky  $I_t$ ) na tempo růstu sledovaného ukazatele  $Y$  v původní časové řadě. Rostou-li hodnoty  $\delta_t$ , vykazuje ukazatel rostoucí tempo růstu (a naopak). Pokud je posloupnost relativních přírůstků zhruba konstantní, lze usuzovat i na konstantní tempo růstu sledovaného ukazatele.



**Druhá diference  $\Delta_t^{(2)}$**  je absolutní diference prvních diferencí:

$$\Delta_t^{(2)} = \Delta_t^{(1)} - \Delta_{t-1}^{(1)} \quad \text{kde} \quad t = 3, 4, 5, \dots$$

S tímto pojmem jsme se již také setkali při povídání o **kvadratické závislosti** v kapitole zabývající se regresními vztahy mezi dvourozměrnými daty. Tehdy jsme jej nazývali *přírůstkem přírůstků*. Takže již víme, že pokud se jednotlivé členy posloupnosti druhých diferencí systematicky ani nezvětšují ani nezmenšují (jejich hodnoty oscilují pouze náhodně a „ne příliš“), lze u původní časové řady předpokládat **kvadratický trend**. Hodnoty ukazatele  $Y$  časové řady budou ležet téměř na parabole.

A následující charakteristiky také známe, a to z kapitoly o **hospodářské statistice**.

**Koeficient růstu  $I_t$  (řetězový index)** neboli individuální jednoduchý index o proměnlivém základu (vztažený k bezprostředně předcházejícímu pozorování v časové řadě původních hodnot):

$$I_t = \frac{y_t}{y_{t-1}} = \frac{y_t - y_{t-1} + y_{t-1}}{y_{t-1}} = \sigma_t + 1 \quad \text{kde} \quad t = 2, 3, 4, \dots$$

Koeficient růstu vyjádřený v procentech se nazývá **tempo růstu**.

Pokud hodnoty v posloupnosti koeficientů růstu „příliš“ neoscilují, lze předpokládat, že původní časová řada má **exponenciální trend**.

**Průměrný koeficient růstu  $T$**  je geometrickým průměrem koeficientů růstu:

$$T = \sqrt[n-1]{I_2 \cdot I_3 \cdot I_4 \cdot \dots \cdot I_n} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

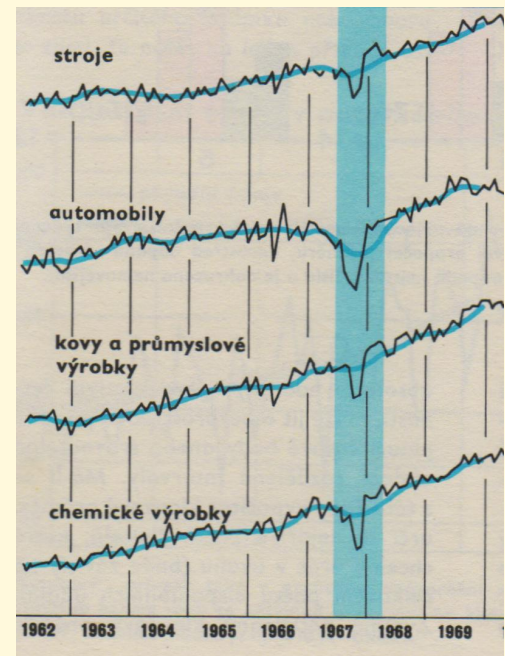
Protože průměrný koeficient růstu závisí pouze na krajních hodnotách řady, lze získat zcela stejný průměrný koeficient růstu pro řady, které se shodují pouze ve svých krajních úrovních, ale jinak mají zcela rozdílný průběh. Proto je nutné před výpočtem pečlivě analyzovat příslušnou časovou řadu a je-li to nutné, rozdělit ji na několik částí tak, aby v každé z těchto částí sledovaný ukazatel vykazoval v podstatě monotónní vývoj. A pro každou z těchto částí pak stanovit průměrné koeficienty růstu (podobně jako jsme to dělali v případě **lineárních** regresních funkcí).

## 2. Vyrovnání časových řad

Mimo již dříve zmiňovaný celkový vývoj, vlivy ročních období a cykly mohou na časovou řadu působit ještě jednorázové mimořádné jevy. Tyto jevy mohou být rozeznatelné již předem, například devalvace měny (vedlejší obrázek převzatý z [14], na kterém je sezónně vyrovnaný export Velké Británie, zřetelně zachycuje projev snížení hodnoty libry z podzimu roku 1967). Jestliže jev, který vznikl jednorázově, působí trvale, lze mluvit o jakémsi „zlomu struktury“, který vede ke změně dalšího vývoje. Například vynález syntetických vláken postavil textilní průmysl před zcela novou situací.

Jednorázový jev a vývoj jsou tedy někdy v úzkém spojení, a cykly proto mohou mít pochybnou vypovídací hodnotu. Sezónní vlivy také nejsou vždy tak jasně prokazatelné, jako je tomu například při prodeji zmrzliny nebo ve stavebnictví či v cizineckém ruchu.

Proto je každý pokus o **vyrovnání** (očistění řady) prvotních údajů provázen nebezpečím, že může dojít ke novému zkreslení.





Nejméně škodlivé je, když se určí srovnatelné období. Tak například lze účelně srovnat údaje o cizineckém ruchu v září jednoho roku jen s údaji z měsíců září v ostatních letech. Ale i takové relativně jednoduché porovnání se „srovnatelným měsícem“ může být zavádějící, jestliže v loňském roce bylo září nádherné a teplé a následovalo po deštivém a chladném srpnu. V jiných letech to nebude platit.

Abychom se všeobecně vyhnuli takovým nahodilostem, pak v zájmu **opravy od rušivých**, ale nepodstatných vlivů, postupujeme většinou následovně:

**intervalové** časové řady (hodnoty za určitý časový interval) transformujeme na stejně dlouhý časový úsek. Protože běžný rok má 365 dní, tak za délku běžného měsíce bereme  $365 : 12 \doteq 30,42$ . Potom například lednovou hodnotu budeme násobit číslem  $\frac{30,42}{31}$ , únorovou  $\frac{30,42}{28}$ , atd.

Pokud více zaokrouhlíme, můžeme za průměrný měsíc považovat ten, který má 30 dnů. Podobně i pro jiné časové intervaly, než je zde zmiňovaný měsíc.

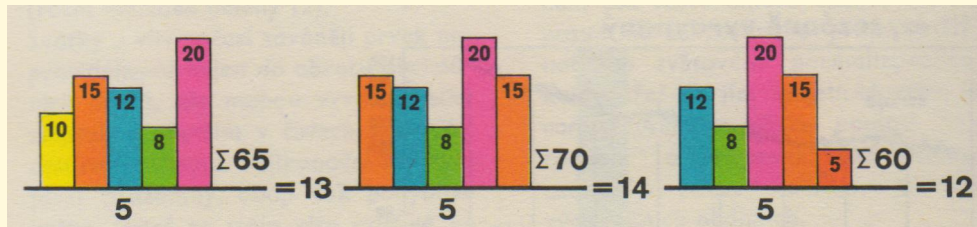
Měsíc	Dny	Množství vytěženého uhlí v prvním pololetí roku 2001			
		reálné h.	reálné hodnoty / den	výpočet	očištěné hodnoty
leden	31	1 180	$1\,180 : 31 = 38,065$	$38,065 \cdot 30,42 =$	1 158
únor	28	1 010	$1\,010 : 28 = 36,071$	$36,071 \cdot 30,42 =$	1 097
březen	31	1 200	$1\,200 : 31 = 38,710$	$38,710 \cdot 30,42 =$	1 178
duben	30	1 090	$1\,090 : 30 = 36,333$	$36,333 \cdot 30,42 =$	1 105
květen	31	1 180	$1\,180 : 31 = 38,065$	$38,065 \cdot 30,42 =$	1 158
červen	30	1 130	$1\,130 : 30 = 37,667$	$37,667 \cdot 30,42 =$	1 146
<b>součet</b>	181	6 790	$6\,790 : 181 = 37,514$	$37,514 \cdot 6 \cdot 30,42$	$= 6\,847 \mid \Sigma = 6\,842$

Pokud nás však zajímá pouze průměrná hodnota za určité období, stačí využít **váženého chronologického průměru**, podobně jako v **tomto** příkladu.

**okamžikové** časové řady (hodnoty vždy k danému datu) vyrovnáváme nejčastěji **metodou klouzavých průměrů**. V případě **prostého** klouzavého průměru každý úsek nahradíme jeho (prostým) aritmetickým průměrem (viz obrázek 11), kde: 10, 15, 12, 8, 20, 15, 5

$$\bar{x}=13$$

Obrázek 11: Schéma prostého klouzavého průměru (převzato z [14])



Vždy se připojuje nejnovější měsíční nebo roční či týdenní výsledek a nejstarší se vypouští. Vlevo je první propočet průměru, uprostřed odpadá *starých* 10 a *nových* 15 se přidává, vpravo rovněž odpadá nejstarší číslo a místo něj přidáváme nejnovější.

U **váženého klouzavého průměru** nahradíme každý úsek jeho váženým aritmetickým průměrem. Stanovováním vah se zde nebudeme zabývat. Postup lze nalézt např. v [11, str. 99].

Důležitou otázkou, kterou je nutno vyřešit, je stanovení počtu pozorování, ze kterých jsou jednotlivé klouzavé průměry počítány. V uvedeném příkladu je to pět pozorování. „*Nutno konstatovat, že volba rozsahu klouzavé části období interpolace je obtížná ... V praxi jsou většinou zvoleny klouzavé části menší délky ...*“ [11, str. 97]

Existuje ještě řada jiných postupů vyrovnání časových řad. Žádný z nich však není docela bez problémů, protože v samé podstatě vyrovnávání je obsažena nutnost odchylky od daných údajů tím, že se posuzují (s nutně subjektivním zabarvením) faktory, které se mohou koneckonců jen odhadnout.

## 2.1. Problémy při analýze časových řad

Při zpracování dat ve formě časové řady se potýkáme s množstvím problémů (na některé jsme upozornili v předchozím textu), které jsou právě pro časové řady specifické. Jedná se především o problémy:

- s volbou časových bodů pozorování;
- s kalendářem
  - různá délka měsíců,
  - různý počet víkendů v měsíci,
  - různý počet pracovních dnů v měsíci,
  - pohyblivé svátky;
- s délkou časových řad;
- nesrovnatelností dat.

## 3. Modelování časových řad — trend

Časovou řadu zkoumáme proto, abychom mohli „odhalit“ mechanismus působení času na utváření hodnot sledovaného statistického ukazatele  $Y$ . Nebo jinak, abychom pochopili příčiny, které na tyto jevy působily a ovlivňovaly jejich chování v minulosti. A následně abychom získané poznatky využili k **prognóze do budoucna**.

Předpokládáme, že model (který popisujeme časovou řadou) obsahuje následující složky:

**Trendovou**  $T_t$  — hlavní (obecná) tendence dlouhodobého vývoje zkoumaného jevu za dlouhé období. Je výsledkem dlouhodobých a stálých procesů. Trend může být rostoucí, klesající nebo někdy mohou hodnoty ukazatele dané časové řady kolísat (oscilovat) kolem určité úrovně. Pak se jedná o časovou řadu s konstantním trendem (někdy se nesprávně říká, že řada je bez trendu).

**Sezónní**  $S_t$  — pravidelně se opakující výkyvy (odchyly od trendové složky) s periodou kratší jak jeden rok;

Cyklickou — dlouhodobé kolísání kolem trendu<sup>49</sup>; v důsledku dlouhodobého cyklického vývoje (používá se spíše v makroekonomických úvahách).

**Náhodnou**  $\varepsilon_t$  — souhrn drobných nezávislých příčin, které se nedají popsat žádnou funkcí času. Je to „zbytek“ po vyloučení trendu, sezónní a cyklické složky.

Za (aditivní) model časové řady pak můžeme považovat vztah

$$y_t = T_t + S_t + \varepsilon_t$$

kde  $y_t$  je hodnota proměnné závislá na čase  $t$ , což je nezávislá (časová) proměnná a můžeme ji celkem libovolně vyjádřit v jakýchkoliv časových jednotkách s libovolným počátkem.

Když proměnnou  $t$  volíme tak, aby byla

- **ekvidistantní** (pravidelně rostla o stejný krok),
- **malá** (a raději pouze celočíselná; to kvůli zjednodušení výpočtů)
- a její aritmetický průměr byl **NULA**,

provedeme **centrovanou** časovou transformaci. Tím se výpočty pro klasickou metodu nejmenších čtverců zjednoduší.

<sup>49</sup> Přikloníme se k častému názoru, že cyklickou složku lze považovat za součást trendu.

**Model trendu** (vhodnou funkci, která nejlépe popisuje trend) si ukážeme pouze pro případ rovnice přímky nebo paraboly. Pro jiné typy křivek odkazujeme na příslušnou literaturu (např. [13]).

$$\text{Lineární trend } L(t) : y = a + b \cdot t \quad \Rightarrow \quad \text{Platí-li } \bar{t}_A = 0, \text{ pak } a = \frac{\sum_{\forall i} y_i}{n}, \quad b = \frac{\sum_{\forall i} (y_i \cdot t_i)}{\sum t_i^2}$$

**Lineární bodová předpověď**  $L_p = L(t_p)$  hodnoty časové řady v čase  $t_p$  se získá dosazením  $t_p$  za  $t$  do rovnice lineárního trendu.

**Lineární intervalová předpověď** hodnoty časové řady v čase  $t_p$  s  $\alpha\%$  spolehlivostí je interval  $(L_p - \Delta; L_p + \Delta)$ , kde  $\Delta = s \cdot h_p \cdot t_{1-\frac{\alpha}{2}}(n-2)$  nazýváme přípustná chyba a

$$s = \sqrt{\frac{\sum y^2 - \sum L^2(t)}{n-2}} \quad h_p = \sqrt{1 - \frac{1}{n} + \frac{t_p^2}{\sum t^2}}$$

Výraz  $t_{1-\frac{\alpha}{2}}(n-2)$  je kvantil Studentova rozdělení, který najdeme ve statistických [tabulkách](#) (Excel).

Po vhodné substituci indexu  $i$ <sup>50</sup> na index  $t$  (s průměrem NULA), jdou koeficienty  $a, b$  snadno určit. Pozor! Koeficient  $a$  z lineárního trendu má jinou hodnotu jak stejně označený koeficient  $a$  z kvadratického trendu.

<sup>50</sup> Protože například místo  $\sum y$  bychom správně měli psát  $\sum_{\forall i} y_i$

**Kvadratický trend**  $K(t) : y = a + b \cdot t + c \cdot t^2 \Rightarrow$  Platí-li  $\bar{t}_a = 0$ , pak

$$a = \frac{\sum y \cdot \sum t^4 - \sum t^2 \cdot \sum (y \cdot t^2)}{n \cdot \sum t^4 - (\sum t^2)^2}, \quad b = \frac{\sum (y \cdot t)}{\sum t^2}, \quad c = \frac{n \cdot \sum (y \cdot t^2) - \sum y \cdot \sum t^2}{n \cdot \sum t^4 - (\sum t^2)^2}$$

**Bodová předpověď**  $K_p = K(t_p)$  hodnoty časové řady v čase  $t_p$  se (stejně jako v případě lineárního trendu) získá dosazením  $t_p$  za  $t$  do rovnice kvadratického trendu.

**Intervalová předpověď** hodnoty časové řady v čase  $t_p$  je opět interval  $(K_p - \Delta; K_p + \Delta)$ , kde přípustná chyba  $\Delta = s \cdot g_p \cdot t_{1-\frac{\alpha}{2}}(n-3)$  ovšem není identická jako v případě lineárního trendu a výraz  $t_{1-\frac{\alpha}{2}}(n-3)$  je opět kvantil Studentova rozdělení, tentokrát s jiným argumentem než u lineárního trendu, stejně tak jako první koeficient  $s$ . Platí:

$$s = \sqrt{\frac{\sum y^2 - \sum K^2(t)}{n-3}} \quad \text{a} \quad g_p = \sqrt{1 + [1 \ t_p \ t_p^2] \cdot [X^T \cdot X]^{-1} \cdot [1 \ t_p \ t_p^2]^T}$$

kde symbol  $[...]^{-1}$  označuje inverzní matici, symbol  $\bullet$  součin matic a symbol  $X^T$  transponovanou (má zaměněny řádky za sloupce) matici k matici  $X$ , která je definována následovně:

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ \vdots & \vdots & \vdots \\ 1 & n & n^2 \end{bmatrix}$$

Podobně  $[1 \ t_p \ t_p^2]^T$  je vlastně sloupcová matice (matice, která má tři řádky a jeden sloupec).

**Autokorelace časových řad** Na závěr se zmíníme o typickém jevu, který je spojen s časovými řadami a komplikuje předpověď hodnot řady pomocí regrese. Hodnoty ukazatele v řadě za sebou bývají často vzájemně závislé. Jev se nazývá autokorelace.

Například

- dnešní teplota vzduchu je závislá na teplotě včerejší;
- dnešní cena akcie se odvíjí od ceny včerejší;
- nadbytečný nákup zásob v daném období způsobuje snížení nákupu v období příštím a naopak.

Podrobnější rozbor tohoto problému včetně testování významnosti autokorelace (například Durbinův–Watsonův test) lze nalézt v literatuře.

## Trendy — příklady

K dispozici jsou **následující** data

**40; 42; 43; 41; 43; 44,8**

(stejná jako ta, ze kterých jsme v kapitole o [hospodářské statistice](#) (indexech) počítali [průměrný koeficient vývoje](#) a odhadovali tržby v následujícím období a stejná jako ta, u kterých jsme v kapitole o [regresních závislostech](#) určovali [regresní funkce](#))

**o tržbách** (nákladech, obracech, ...) za šest po sobě jdoucích období (dny, týdny, měsíce, ...), kdy jednotkou mohou být tisíce (statisíce, milióny, ...) a měnovou jednotkou **Kč, \$, ...**

Graficky znázorněte časovou řadu a odhadněte hodnotu tržeb v následujícím období pomocí lineárního a kvadratického trendu, včetně 95% intervalů spolehlivosti.

Vše zapíšeme do tabulky

a určíme hodnotu následujícího období:  $i = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5\%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i$ ,  $y_i \cdot t_i$ ,  $t_i^2$  pro bodový odhad

období index $i$	$y_i$						
1	40						
2	42						
3	43						
4	41						
5	43						
6	44,8						
$\Sigma$	253,8						

**Lineární trend:**  $L(t) = a + b \cdot t$

Bodový odhad:

Intervalový odhad:  $t_{0,975}(6 - 2) =$

$s =$

$h =$   $\Rightarrow$   $\Delta =$



Vše zapíšeme do tabulky. Aby platilo  $\bar{t}_A = 0$ , volíme v našem případě za  $t$  tyto hodnoty:  $t_1 = -5$ ;  $t_2 = -3$ ;  $t_3 = -1$ ;  $t_4 = 1$ ;  $t_5 = 3$ ;  $t_6 = 5$  a určujeme hodnotu následujícího období:  $i = 7 \Rightarrow t_7 = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5 \%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i$ ,  $y_i \cdot t_i$ ,  $t_i^2$  pro bodový odhad

období index $i$	$y_i$	$t_i$					
1	40	-5					
2	42	-3					
3	43	-1					
4	41	1					
5	43	3					
6	44,8	5					
$\Sigma$	253,8	0					

Určíme **medián** indexu  $i$  a přiřadíme mu **NULU**.

V našem případě bude medián mezi řádkem 3 a 4.

Nejbližšímu nižšímu indexu než medián (3. řádek) přiřadíme hodnotu **-1** a nejbližšímu vyššímu indexu (4. řádek) **1**.

A protože časová proměnná  $t$  musí být *ekvidistantní*, můžeme doplnit zbylé hodnoty této proměnné.

**Lineární trend:**  $L(t) = a + b \cdot t$

Bodový odhad:  $L(t = 7) =$

Intervalový odhad:  $t_{0,975}(6 - 2) =$

$s =$

$h = \Rightarrow \Delta =$

Vše zapíšeme do tabulky. Aby platilo  $\bar{t}_A = 0$ , volíme v našem případě za  $t$  tyto hodnoty:  $t_1 = -5$ ;  $t_2 = -3$ ;  $t_3 = -1$ ;  $t_4 = 1$ ;  $t_5 = 3$ ;  $t_6 = 5$  a určujeme hodnotu následujícího období:  $i = 7 \Rightarrow t_7 = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5\%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i$ ,  $y_i \cdot t_i$ ,  $t_i^2$  pro bodový odhad

období index $i$	$y_i$	$t_i$	$y_i \cdot t_i$				
1	40	-5	-200				
2	42	-3	-126				
3	43	-1	-43				
4	41	1	41				
5	43	3	129				
6	44,8	5	224				
$\Sigma$	253,8	0	25				

Určíme **medián** indexu  $i$  a přiřadíme mu **NULU**.

V našem případě bude medián mezi řádkem 3 a 4.

Nejbližšímu nižšímu indexu než medián (3. řádek) přiřadíme hodnotu **-1** a nejbližšímu vyššímu indexu (4. řádek) **1**.

A protože časová proměnná  $t$  musí být *ekvidistantní*, můžeme doplnit zbylé hodnoty této proměnné.

**Lineární trend:**  $L(t) = a + b \cdot t$

Bodový odhad:  $L(t = 7) =$

Intervalový odhad:  $t_{0,975}(6 - 2) =$

$s =$

$h = \Rightarrow \Delta =$

Vše zapíšeme do tabulky. Aby platilo  $\bar{t}_A = 0$ , volíme v našem případě za  $t$  tyto hodnoty:  $t_1 = -5$ ;  $t_2 = -3$ ;  $t_3 = -1$ ;  $t_4 = 1$ ;  $t_5 = 3$ ;  $t_6 = 5$  a určujeme hodnotu následujícího období:  $i = 7 \Rightarrow t_7 = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5 \%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i, y_i \cdot t_i, t_i^2$  pro bodový odhad

období index $i$	$y_i$	$t_i$	$y_i \cdot t_i$	$t_i^2$			
1	40	-5	-200	25			
2	42	-3	-126	9			
3	43	-1	-43	1			
4	41	1	41	1			
5	43	3	129	9			
6	44,8	5	224	25			
$\Sigma$	253,8	0	25	70			

Určíme **medián** indexu  $i$  a přiřadíme mu **NULU**.

V našem případě bude medián mezi řádkem 3 a 4.

Nejbližšímu nižšímu indexu než medián (3. řádek) přiřadíme hodnotu **-1** a nejbližšímu vyššímu indexu (4. řádek) **1**.

A protože časová proměnná  $t$  musí být *ekvidistantní*, můžeme doplnit zbylé hodnoty této proměnné.

**Lineární trend:**  $L(t) = a + b \cdot t = \frac{253,8}{6} + \frac{25}{70} \cdot t \doteq 42,3 + 0,357 \cdot t$

Bodový odhad:  $L(t = 7) =$

Intervalový odhad:  $t_{0,975}(6 - 2) =$

$s =$

$h = \Rightarrow \Delta =$

Vše zapíšeme do tabulky. Aby platilo  $\bar{t}_A = 0$ , volíme v našem případě za  $t$  tyto hodnoty:  $t_1 = -5$ ;  $t_2 = -3$ ;  $t_3 = -1$ ;  $t_4 = 1$ ;  $t_5 = 3$ ;  $t_6 = 5$  a určujeme hodnotu následujícího období:  $i = 7 \Rightarrow t_7 = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5\%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i$ ,  $y_i \cdot t_i$ ,  $t_i^2$  pro bodový odhad

období index $i$	$y_i$	$t_i$	$y_i \cdot t_i$	$t_i^2$	$y_i^2$		
1	40	-5	-200	25			
2	42	-3	-126	9			
3	43	-1	-43	1			
4	41	1	41	1			
5	43	3	129	9			
6	44,8	5	224	25			
$\Sigma$	253,8	0	25	70			

Určíme **medián** indexu  $i$  a přiřadíme mu **NULU**.

V našem případě bude medián mezi řádkem 3 a 4. Nejbližšímu nižšímu indexu než medián (3. řádek) přiřadíme hodnotu **-1** a nejbližšímu vyššímu indexu (4. řádek) **1**.

A protože časová proměnná  $t$  musí být *ekvidistantní*, můžeme doplnit zbylé hodnoty této proměnné.

**Lineární trend:**  $L(t) = a + b \cdot t = \frac{253,8}{6} + \frac{25}{70} \cdot t \doteq 42,3 + 0,357 \cdot t$

Bodový odhad:  $L(t = 7) = 42,3 + 0,357 \cdot 7 = 44,799 \doteq 44,8$

Intervalový odhad:  $t_{0,975}(6 - 2) =$

$s =$

$h = \Rightarrow \Delta =$

Vše zapíšeme do tabulky. Aby platilo  $\bar{t}_A = 0$ , volíme v našem případě za  $t$  tyto hodnoty:  $t_1 = -5$ ;  $t_2 = -3$ ;  $t_3 = -1$ ;  $t_4 = 1$ ;  $t_5 = 3$ ;  $t_6 = 5$  a určujeme hodnotu následujícího období:  $i = 7 \Rightarrow t_7 = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5 \%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i, y_i \cdot t_i, t_i^2$  pro bodový odhad a  $y_i^2, L(t_i), L^2(t_i)$  pro intervalový.

období index $i$	$y_i$	$t_i$	$y_i \cdot t_i$	$t_i^2$	$y_i^2$	$L(t_i)$	
1	40	-5	-200	25	1 600		
2	42	-3	-126	9	1 764		
3	43	-1	-43	1	1 849		
4	41	1	41	1	1 681		
5	43	3	129	9	1 849		
6	44,8	5	224	25	2 007,04		
$\Sigma$	253,8	0	25	70	10 750,04		

Určíme **medián** indexu  $i$  a přiřadíme mu **NULU**.

V našem případě bude medián mezi řádkem 3 a 4.

Nejbližšímu nižšímu indexu než medián (3. řádek) přiřadíme hodnotu **-1** a nejbližšímu vyššímu indexu (4. řádek) **1**.

A protože časová proměnná  $t$  musí být *ekvidistantní*, můžeme doplnit zbylé hodnoty této proměnné.

**Lineární trend:**  $L(t) = a + b \cdot t = \frac{253,8}{6} + \frac{25}{70} \cdot t \doteq 42,3 + 0,357 \cdot t$

Bodový odhad:  $L(t = 7) = 42,3 + 0,357 \cdot 7 = 44,799 \doteq 44,8$

Intervalový odhad:  $t_{0,975}(6 - 2) =$

$s =$

$h = \Rightarrow \Delta =$

Vše zapíšeme do tabulky. Aby platilo  $\bar{t}_A = 0$ , volíme v našem případě za  $t$  tyto hodnoty:  $t_1 = -5$ ;  $t_2 = -3$ ;  $t_3 = -1$ ;  $t_4 = 1$ ;  $t_5 = 3$ ;  $t_6 = 5$  a určujeme hodnotu následujícího období:  $i = 7 \Rightarrow t_7 = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5 \%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i, y_i \cdot t_i, t_i^2$  pro bodový odhad a  $y_i^2, L(t_i), L^2(t_i)$  pro intervalový.

období index $i$	$y_i$	$t_i$	$y_i \cdot t_i$	$t_i^2$	$y_i^2$	$L(t_i)$	
1	40	-5	-200	25	1 600	40,514	
2	42	-3	-126	9	1 764	41,229	
3	43	-1	-43	1	1 849	41,943	
4	41	1	41	1	1 681	42,657	
5	43	3	129	9	1 849	43,371	
6	44,8	5	224	25	2 007,04	44,086	
$\Sigma$	253,8	0	25	70	10 750,04		

Určíme **medián** indexu  $i$  a přiřadíme mu **NULU**.

V našem případě bude medián mezi řádkem 3 a 4. Nejbližšímu nižšímu indexu než medián (3. řádek) přiřadíme hodnotu **-1** a nejbližšímu vyššímu indexu (4. řádek) **1**.

A protože časová proměnná  $t$  musí být *ekvidistantní*, můžeme doplnit zbylé hodnoty této proměnné.

**Lineární trend:**  $L(t) = a + b \cdot t = \frac{253,8}{6} + \frac{25}{70} \cdot t \doteq 42,3 + 0,357 \cdot t$

Bodový odhad:  $L(t = 7) = 42,3 + 0,357 \cdot 7 = 44,799 \doteq 44,8$

Intervalový odhad:  $t_{0,975}(6 - 2) =$

$s =$

$h = \Rightarrow \Delta =$

Vše zapíšeme do tabulky. Aby platilo  $\bar{t}_A = 0$ , volíme v našem případě za  $t$  tyto hodnoty:  $t_1 = -5$ ;  $t_2 = -3$ ;  $t_3 = -1$ ;  $t_4 = 1$ ;  $t_5 = 3$ ;  $t_6 = 5$  a určujeme hodnotu následujícího období:  $i = 7 \Rightarrow t_7 = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5\%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i$ ,  $y_i \cdot t_i$ ,  $t_i^2$  pro bodový odhad a  $y_i^2$ ,  $L(t_i)$ ,  $L^2(t_i)$  pro intervalový.

období index $i$	$y_i$	$t_i$	$y_i \cdot t_i$	$t_i^2$	$y_i^2$	$L(t_i)$	$L^2(t_i)$
1	40	-5	-200	25	1 600	40,514	1 641,384
2	42	-3	-126	9	1 764	41,229	1 699,830
3	43	-1	-43	1	1 849	41,943	1 759,215
4	41	1	41	1	1 681	42,657	1 819,620
5	43	3	129	9	1 849	43,371	1 881,044
6	44,8	5	224	25	2 007,04	44,086	1 943,575
$\Sigma$	253,8	0	25	70	10 750,04		10 744,668

Určíme **medián** indexu  $i$  a přiřadíme mu **NULU**.

V našem případě bude medián mezi řádkem 3 a 4.

Nejbližšímu nižšímu indexu než medián (3. řádek) přiřadíme hodnotu **-1** a nejbližšímu vyššímu indexu (4. řádek) **1**.

A protože časová proměnná  $t$  musí být *ekvidistantní*, můžeme doplnit zbylé hodnoty této proměnné.

**Lineární trend:**  $L(t) = a + b \cdot t = \frac{253,8}{6} + \frac{25}{70} \cdot t \doteq 42,3 + 0,357 \cdot t$

Bodový odhad:  $L(t = 7) = 42,3 + 0,357 \cdot 7 = 44,799 \doteq 44,8$

Intervalový odhad:  $t_{0,975}(6 - 2) =$

$s =$

$h = \Rightarrow \Delta =$

Vše zapíšeme do tabulky. Aby platilo  $\bar{t}_A = 0$ , volíme v našem případě za  $t$  tyto hodnoty:  $t_1 = -5$ ;  $t_2 = -3$ ;  $t_3 = -1$ ;  $t_4 = 1$ ;  $t_5 = 3$ ;  $t_6 = 5$  a určujeme hodnotu následujícího období:  $i = 7 \Rightarrow t_7 = 7$

Chceme počítat s 95% spolehlivostí, tedy chyba  $\alpha = 5\%$  a potom  $1 - \frac{\alpha}{2} = 0,975$ .

Tabulku doplníme o další sloupce:  $t_i$ ,  $y_i \cdot t_i$ ,  $t_i^2$  pro bodový odhad a  $y_i^2$ ,  $L(t_i)$ ,  $L^2(t_i)$  pro intervalový.

období index $i$	$y_i$	$t_i$	$y_i \cdot t_i$	$t_i^2$	$y_i^2$	$L(t_i)$	$L^2(t_i)$
1	40	-5	-200	25	1 600	40,514	1 641,384
2	42	-3	-126	9	1 764	41,229	1 699,830
3	43	-1	-43	1	1 849	41,943	1 759,215
4	41	1	41	1	1 681	42,657	1 819,620
5	43	3	129	9	1 849	43,371	1 881,044
6	44,8	5	224	25	2 007,04	44,086	1 943,575
$\Sigma$	253,8	0	25	70	10 750,04		10 744,668

Určíme **medián** indexu  $i$  a přiřadíme mu **NULU**.

V našem případě bude medián mezi řádkem 3 a 4. Nejbližšímu nižšímu indexu než medián (3. řádek) přiřadíme hodnotu **-1** a nejbližšímu vyššímu indexu (4. řádek) **1**.

A protože časová proměnná  $t$  musí být *ekvidistantní*, můžeme doplnit zbylé hodnoty této proměnné.

**Lineární trend:**  $L(t) = a + b \cdot t = \frac{253,8}{6} + \frac{25}{70} \cdot t \doteq 42,3 + 0,357 \cdot t$

Bodový odhad:  $L(t = 7) = 42,3 + 0,357 \cdot 7 = 44,799 \doteq 44,8$

Intervalový odhad:  $t_{0,975}(6 - 2) = 2,776 45$

$$s = \sqrt{\frac{10 750,04 - 10 744,668}{6 - 2}} = 1,159$$

$$h = \sqrt{1 - \frac{1}{6} + \frac{7^2}{70}} = 1,238 \Rightarrow \Delta = 1,159 \cdot 1,238 \cdot 2,776 45 = 3,983 \doteq 4$$

$$(44,8 - 4; 44,8 + 4) = (40,8; 48,8)$$

$n \backslash \alpha$	0.9	0.95	0.975	0.99	0.995	0.999
1	3.07768	6.31375	12.7062	31.8205	63.6567	318.309
2	1.88562	2.91999	4.30265	6.96456	9.92484	22.3271
3	1.63774	2.35336	3.18245	4.54070	5.84091	10.2145
4	1.53321	2.13185	2.77645	3.74695	4.60409	7.17318
5	1.47588	2.01505	2.57058	3.36493	4.03214	5.89343



V případě **kvadratického trendu** postupujeme **analogicky**. Chceme opět počítat s 95% spolehlivostí. Přípustná chyba  $\alpha = 5\%$  a  $1 - \frac{\alpha}{2} = 0,975$ . Nejdříve doplníme naši tabulku o sloupce  $y \cdot t^2, t^4$

$i$	$t$	$y$	$t^2$	$y \cdot t$	$y^2$	$L, L^2$		$t^4$	$K(t)$	$K^2(t)$
1	-5	40	25	-200	1 600					
2	-3	42	9	-126	1 764					
3	-1	43	1	-43	1 849					
4	1	41	1	41	1 681					
5	3	43	9	129	1 849					
6	5	44,8	25	224	2 007,04					
$\Sigma$	0	253,8	70	25	10 750,04					

$$K(t) = a + b \cdot t + c \cdot t^2$$

Bodová předpověď:  $K(7) =$

kdy  $t_7 = 7$

Intervalová předpověď:  $t_{0,975}(6 - 3) =$

$$g_p = \sqrt{1 + 3,2}$$

$s =$

$\Delta =$

V případě **kvadratického trendu** postupujeme **analogicky**. Chceme opět počítat s 95% spolehlivostí. Přípustná chyba  $\alpha = 5\%$  a  $1 - \frac{\alpha}{2} = 0,975$ . Nejdříve doplníme naši tabulku o sloupce  $y \cdot t^2, t^4$

$i$	$t$	$y$	$t^2$	$y \cdot t$	$y^2$	$L, L^2$	$y \cdot t^2$	$t^4$	$K(t)$	$K^2(t)$
1	-5	40	25	-200	1 600		1 000	625		
2	-3	42	9	-126	1 764		378	81		
3	-1	43	1	-43	1 849		43	1		
4	1	41	1	41	1 681		41	1		
5	3	43	9	129	1 849		387	81		
6	5	44,8	25	224	2 007,04		1 120	625		
$\Sigma$	0	253,8	70	25	10 750,04		2 969	1 414		

$$K(t) = a + b \cdot t + c \cdot t^2$$

Bodová předpověď:  $K(7) =$

kdy  $t_7 = 7$

Intervalová předpověď:  $t_{0,975}(6 - 3) =$

$$g_p = \sqrt{1 + 3,2}$$

$s =$

$\Delta =$

V případě **kvadratického trendu** postupujeme **analogicky**. Chceme opět počítat s 95% spolehlivostí. Přípustná chyba  $\alpha = 5\%$  a  $1 - \frac{\alpha}{2} = 0,975$ . Nejdříve doplníme naši tabulku o sloupce  $y \cdot t^2, t^4$

$i$	$t$	$y$	$t^2$	$y \cdot t$	$y^2$	$L, L^2$	$y \cdot t^2$	$t^4$	$K(t)$	$K^2(t)$
1	-5	40	25	-200	1 600		1 000	625		
2	-3	42	9	-126	1 764		378	81		
3	-1	43	1	-43	1 849		43	1		
4	1	41	1	41	1 681		41	1		
5	3	43	9	129	1 849		387	81		
6	5	44,8	25	224	2 007,04		1 120	625		
$\Sigma$	0	253,8	70	25	10 750,04		2 969	1 414		

$$K(t) = a + b \cdot t + c \cdot t^2 = \frac{253,8 \cdot 1\,414 - 70 \cdot 2\,969}{6 \cdot 1\,414 - 70^2} + \frac{25}{70} \cdot t + \frac{6 \cdot 2\,969 - 253,8 \cdot 70}{6 \cdot 1\,414 - 70^2} \cdot t^2$$

$$K(t) = 42,144 + 0,357 \cdot t + 0,013 \cdot t^2$$

Bodová předpověď:  $K(7) =$

kdy  $t_7 = 7$

Intervalová předpověď:  $t_{0,975}(6 - 3) =$

$$g_p = \sqrt{1 + 3,2}$$

$s =$

$\Delta =$

V případě **kvadratického trendu** postupujeme **analogicky**. Chceme opět počítat s 95% spolehlivostí. Přípustná chyba  $\alpha = 5\%$  a  $1 - \frac{\alpha}{2} = 0,975$ . Nejdříve doplníme naši tabulku o sloupce  $y \cdot t^2, t^4$

$i$	$t$	$y$	$t^2$	$y \cdot t$	$y^2$	$L, L^2$	$y \cdot t^2$	$t^4$	$K(t)$	$K^2(t)$
1	-5	40	25	-200	1 600		1 000	625		
2	-3	42	9	-126	1 764		378	81		
3	-1	43	1	-43	1 849		43	1		
4	1	41	1	41	1 681		41	1		
5	3	43	9	129	1 849		387	81		
6	5	44,8	25	224	2 007,04		1 120	625		
$\Sigma$	0	253,8	70	25	10 750,04		2 969	1 414		

$$K(t) = a + b \cdot t + c \cdot t^2 = \frac{253,8 \cdot 1\,414 - 70 \cdot 2\,969}{6 \cdot 1\,414 - 70^2} + \frac{25}{70} \cdot t + \frac{6 \cdot 2\,969 - 253,8 \cdot 70}{6 \cdot 1\,414 - 70^2} \cdot t^2$$

$$K(t) = 42,144 + 0,357 \cdot t + 0,013 \cdot t^2$$

Bodová předpověď:  $K(7) = 42,144 + 0,357 \cdot 7 + 0,013 \cdot 7^2 = 45,28 \doteq 45,3$  kdy  $t_7 = 7$

Intervalová předpověď:  $t_{0,975}(6 - 3) =$   $g_p = \sqrt{1 + 3,2}$

$s =$

$\Delta =$

V případě **kvadratického trendu** postupujeme **analogicky**. Chceme opět počítat s 95% spolehlivostí. Přípustná chyba  $\alpha = 5\%$  a  $1 - \frac{\alpha}{2} = 0,975$ . Nejdříve doplníme naši tabulku o sloupce  $y \cdot t^2$ ,  $t^4$ ,  $K$ ,  $K^2$ .

$i$	$t$	$y$	$t^2$	$y \cdot t$	$y^2$	$L, L^2$	$y \cdot t^2$	$t^4$	$K(t)$	$K^2(t)$
1	-5	40	25	-200	1 600		1 000	625		
2	-3	42	9	-126	1 764		378	81		
3	-1	43	1	-43	1 849		43	1		
4	1	41	1	41	1 681		41	1		
5	3	43	9	129	1 849		387	81		
6	5	44,8	25	224	2 007,04		1 120	625		
$\Sigma$	0	253,8	70	25	10 750,04		2 969	1 414		

$$K(t) = a + b \cdot t + c \cdot t^2 = \frac{253,8 \cdot 1\,414 - 70 \cdot 2\,969}{6 \cdot 1\,414 - 70^2} + \frac{25}{70} \cdot t + \frac{6 \cdot 2\,969 - 253,8 \cdot 70}{6 \cdot 1\,414 - 70^2} \cdot t^2$$

$$K(t) = 42,144 + 0,357 \cdot t + 0,013 \cdot t^2$$

Bodová předpověď:  $K(7) = 42,144 + 0,357 \cdot 7 + 0,013 \cdot 7^2 = 45,28 \doteq 45,3$  kdy  $t_7 = 7$

Intervalová předpověď:  $t_{0,975}(6 - 3) = 3,182$   $g_p = \sqrt{1 + 3,2}$

$s =$

$\Delta =$

V případě **kvadratického trendu** postupujeme **analogicky**. Chceme opět počítat s 95% spolehlivostí. Přípustná chyba  $\alpha = 5\%$  a  $1 - \frac{\alpha}{2} = 0,975$ . Nejdříve doplníme naši tabulku o sloupce  $y \cdot t^2, t^4, K, K^2$ .

$i$	$t$	$y$	$t^2$	$y \cdot t$	$y^2$	$L, L^2$	$y \cdot t^2$	$t^4$	$K(t)$	$K^2(t)$
1	-5	40	25	-200	1 600		1 000	625	40,684	1 655,188
2	-3	42	9	-126	1 764		378	81	41,190	1 696,616
3	-1	43	1	-43	1 849		43	1	41,800	1 747,240
4	1	41	1	41	1 681		41	1	42,514	1 807,440
5	3	43	9	129	1 849		387	81	43,332	1 877,662
6	5	44,8	25	224	2 007,04		1 120	625	44,254	1 958,417
$\Sigma$	0	253,8	70	25	10 750,04		2 969	1 414		10 742,563

$$K(t) = a + b \cdot t + c \cdot t^2 = \frac{253,8 \cdot 1\,414 - 70 \cdot 2\,969}{6 \cdot 1\,414 - 70^2} + \frac{25}{70} \cdot t + \frac{6 \cdot 2\,969 - 253,8 \cdot 70}{6 \cdot 1\,414 - 70^2} \cdot t^2$$

$$K(t) = 42,144 + 0,357 \cdot t + 0,013 \cdot t^2$$

Bodová předpověď:  $K(7) = 42,144 + 0,357 \cdot 7 + 0,013 \cdot 7^2 = 45,28 \doteq 45,3$  kdy  $t_7 = 7$

Intervalová předpověď:  $t_{0,975}(6 - 3) = 3,182$   $g_p = \sqrt{1 + 3,2}$

$$s = \sqrt{\frac{10\,750,04 - 10\,742,563}{6 - 3}} = 1,579 \quad \Delta =$$

V případě **kvadratického trendu** postupujeme **analogicky**. Chceme opět počítat s 95% spolehlivostí. Přípustná chyba  $\alpha = 5\%$  a  $1 - \frac{\alpha}{2} = 0,975$ . Nejdříve doplníme naši tabulku o sloupce  $y \cdot t^2$ ,  $t^4$ ,  $K$ ,  $K^2$ .

$i$	$t$	$y$	$t^2$	$y \cdot t$	$y^2$	$L, L^2$	$y \cdot t^2$	$t^4$	$K(t)$	$K^2(t)$
1	-5	40	25	-200	1 600		1 000	625	40,684	1 655,188
2	-3	42	9	-126	1 764		378	81	41,190	1 696,616
3	-1	43	1	-43	1 849		43	1	41,800	1 747,240
4	1	41	1	41	1 681		41	1	42,514	1 807,440
5	3	43	9	129	1 849		387	81	43,332	1 877,662
6	5	44,8	25	224	2 007,04		1 120	625	44,254	1 958,417
$\Sigma$	0	253,8	70	25	10 750,04		2 969	1 414		10 742,563

$$K(t) = a + b \cdot t + c \cdot t^2 = \frac{253,8 \cdot 1\,414 - 70 \cdot 2\,969}{6 \cdot 1\,414 - 70^2} + \frac{25}{70} \cdot t + \frac{6 \cdot 2\,969 - 253,8 \cdot 70}{6 \cdot 1\,414 - 70^2} \cdot t^2$$

$$K(t) = 42,144 + 0,357 \cdot t + 0,013 \cdot t^2$$

Bodová předpověď:  $K(7) = 42,144 + 0,357 \cdot 7 + 0,013 \cdot 7^2 = 45,28 \doteq 45,3$  kdy  $t_7 = 7$

Intervalová předpověď:  $t_{0,975}(6 - 3) = 3,182$   $g_p = \sqrt{1 + 3,2} = 2,049$

$$s = \sqrt{\frac{10\,750,04 - 10\,742,563}{6 - 3}} = 1,579 \quad \Delta = 1,579 \cdot 2,049 \cdot 3,182 = 10,295 \doteq 10,3$$

$$(45,3 - 10,3; 45,3 + 10,3) = (35; 55,6)$$

$$\text{kdy: } [1 \ 7 \ 7^2] \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 4 & 9 & 16 & 25 & 36 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 \\ 7 \\ 49 \end{bmatrix} =$$

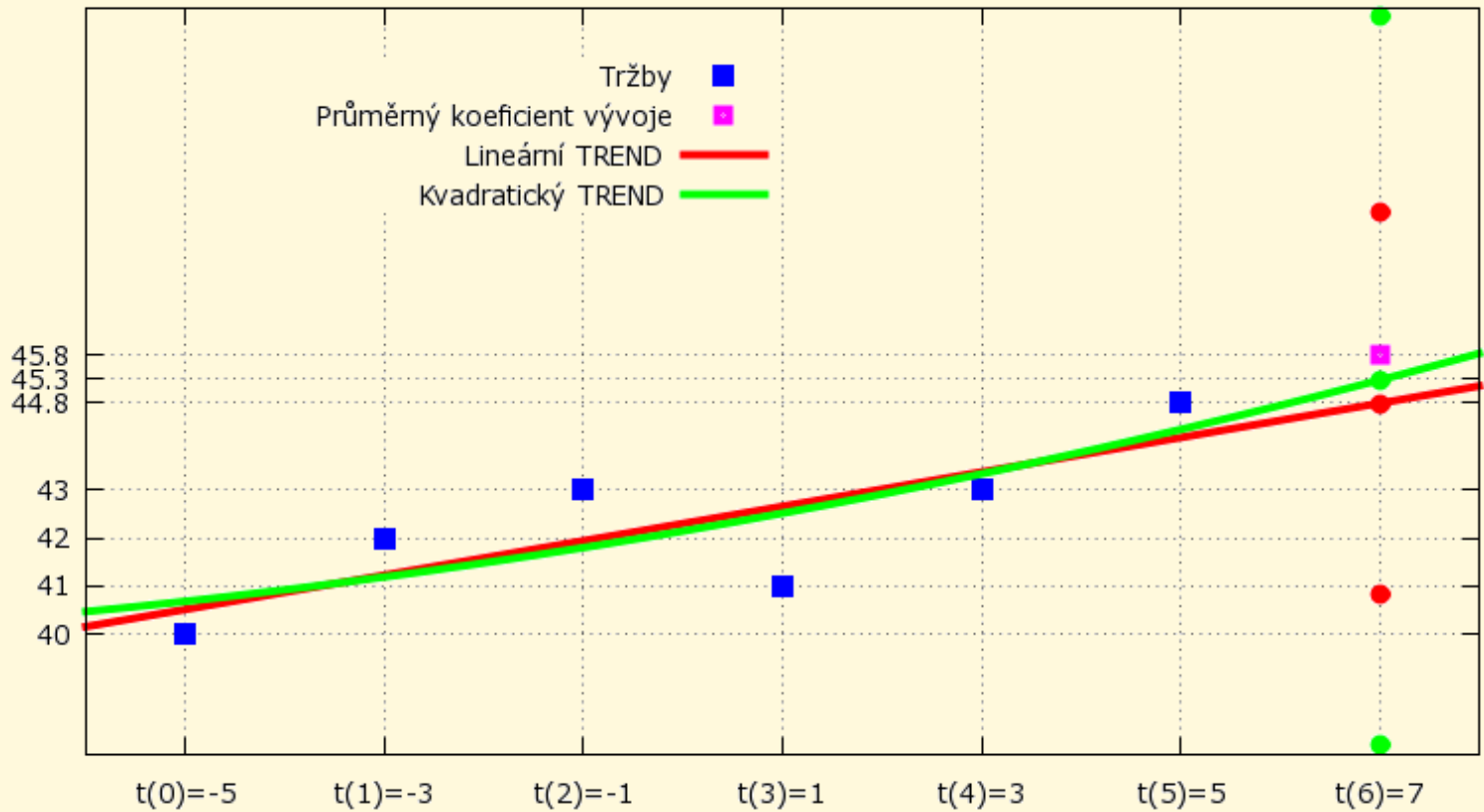
$$= [1 \ 7 \ 49] \cdot \begin{bmatrix} 6 & 21 & 91 \\ 21 & 91 & 441 \\ 91 & 441 & 2275 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 \\ 7 \\ 49 \end{bmatrix} = [1 \ 7 \ 49] \cdot \begin{bmatrix} 3,200 & -1,950 & 0,250 \\ -1,950 & 1,370 & -0,188 \\ 0,250 & -0,188 & 0,027 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 7 \\ 49 \end{bmatrix} =$$

$$= [1,8 \ -1,55 \ 0,25] \cdot \begin{bmatrix} 1 \\ 7 \\ 49 \end{bmatrix} = 3,2$$

Výsledek celého případu zobrazíme graficky na následující stránce.



# TRŽBY



## 4. Využití programového vybavení

V kapitole o [regresních závislostech](#) jsme pro zadaná data (představující [tržby](#)) určovali [lineární regresní funkci](#) a [kvadratickou regresní funkci](#) pomocí metody nejmenších čtverců.

V této kapitole jsme pro [stejná data](#) určovali [lineární trend](#) časové řady a [kvadratický trend](#). Zároveň jsme si (na příkladu lineárního trendu) ukázali, že v obou případech vycházejí stejné výsledky.

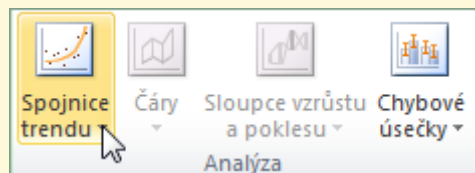
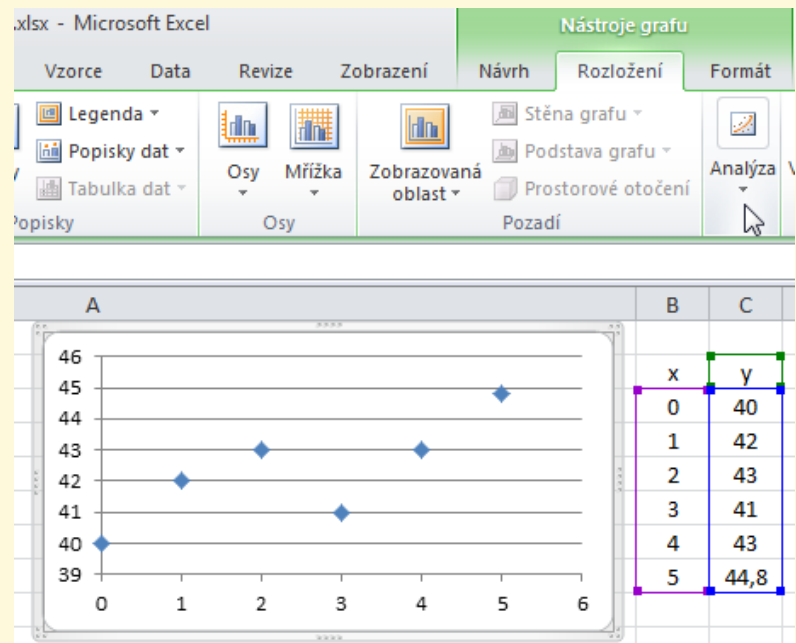
Uvedený postup (ať již se v nějakém oboru nazývá metoda nejmenších čtverců, v jiné oblasti regrese či spojnice trendu) je v praxi natolik používán, že jak některé komerční programy (například **Excel**, Mathematica, Matlab, MathCad, ...) tak jejich freewarové alternativy (například GNUplot) hledají aproximační funkce (funkce, které procházejí **co nejbliže** zadaným bodům) samostatně, bez našeho přičinění. Tedy kromě toho, že jim musíme v jimi požadovaném formátu sdělit, jaké body mají vzít v úvahu.

Konkrétně v programu **Excel 2010** postupujeme následovně:

1. Zadané hodnoty označíme jako blok.
2. Potom na kartě [**Vložení**] v oblasti „**Grafy**“ vybereme <**Bodový**> (první a druhý obrázek) a vedle zadaných dat Excel vloží jejich grafické znázornění (třetí obrázek). Můžeme měnit velikost zobrazení, upravovat popisy, barvy, atd.
3. Nakonec na kartě [**Nástroje grafu**] v záložce „**Rozložení**“ v oblasti <**Analýza**> (třetí obrázek) a položce „**Spojnice trendu**“ (čtvrtý obrázek) vybereme [**Další možnosti spojnice trendu**] (pátý obrázek).

Plošný  
Bodový  
Další grafy

B	C
x	y
0	40
1	42
2	43
3	41
4	43
5	44,8



**Žádná**  
Odebrat vybranou spojnici trendu nebo všechny spojnice trendu, pokud není žádná vybrána

**Lineární spojnice trendu**  
Přidat nebo nastavit lineární spojnici trendu pro vybranou řadu grafu

**Exponenciální spojnice trendu**  
Přidat nebo nastavit exponenciální spojnici trendu pro vybranou řadu grafu

**Lineární spojnice trendu předpovědi**  
Přidat nebo nastavit lineární spojnici trendu s předpovědí pro 2 období pro vybranou řadu grafu

**Klouzavý průměr pro dvě období**  
Přidat nebo nastavit spojnici trendu klouzavého průměru pro 2 období pro vybranou řadu grafu

**Další možnosti spojnice trendu...**

Po upřesnění, že chceme v grafu vypisovat výslednou rovnici (v levém obrázku druhá volba od spodu) včetně výběrového korelačního koeficientu  $r$  (nejspodnější volba — ovšem Excel zobrazuje hodnotu spolehlivosti  $R^2$ , což je druhá mocnina námi požadovaného koeficientu, tedy:  $r = \sqrt{R^2}$ ), se již vykreslí výsledný graf (body i aproximační funkce) včetně potřebných údajů.

Možnosti spojnice trendu

Barva čáry  
Styl čáry  
Stín  
Záře a měkké okraje

**Možnosti spojnice trendu**  
**Typ trendu a regrese**  

☐ Exponenciální

☒ Lineární

☐ Logaritmický

☐ Polynomický Pořadí: 2

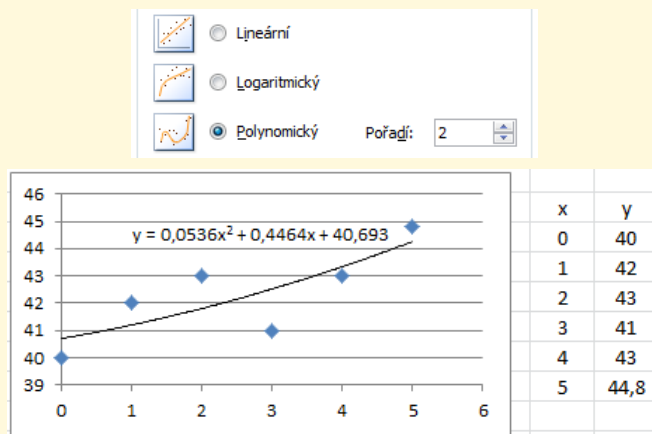
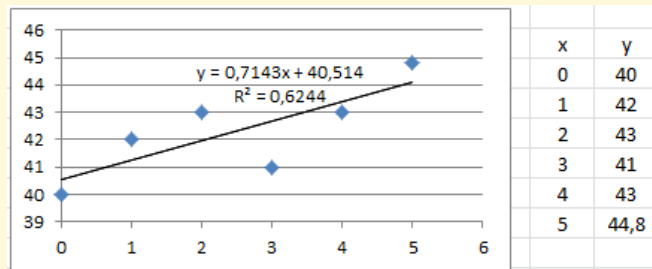
☐ Mocninový

☐ Klouzávy průměr Období: 2

**Název spojnice trendu**  
☒ Automaticky: Lineární (y)  
☐ Vlastní:

**Odhad**  
Vpřed: 0,0 období  
Nazpět: 0,0 období  
☐ Hodnota  $\bar{y}$  = 0,0  
☒ Zobrazit rovnici v grafu  
☒ Zobrazit hodnotu spolehlivosti  $R$

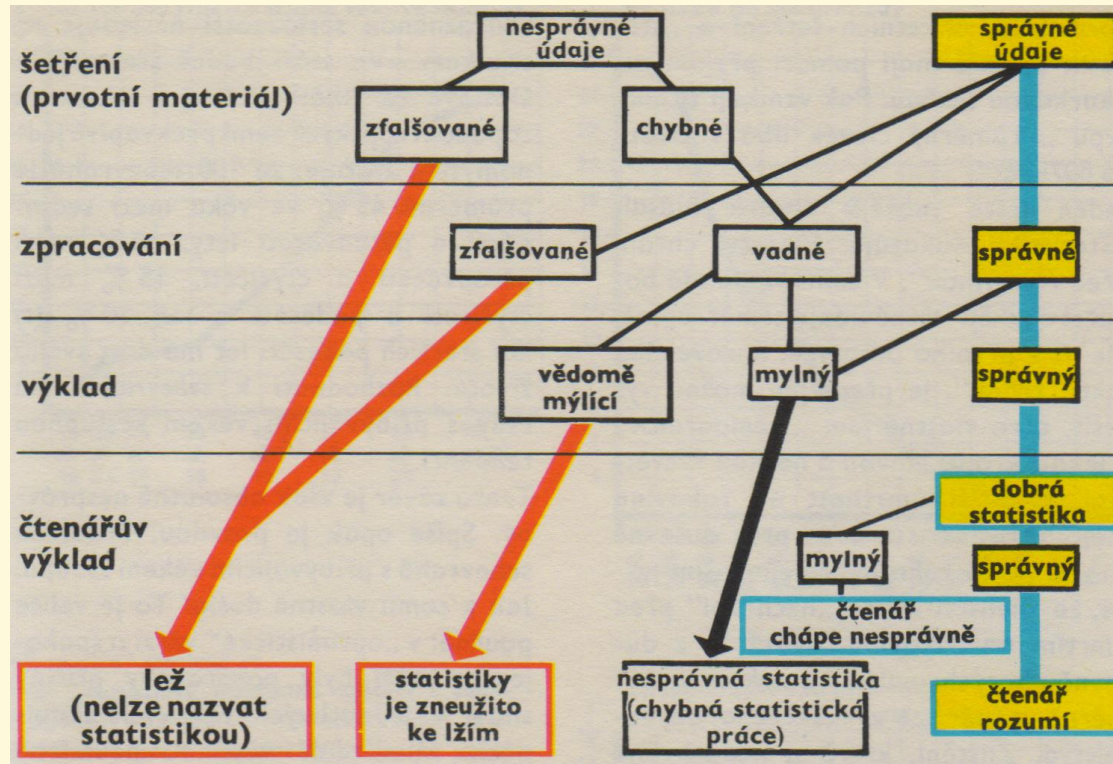
Zavřít



Pro aproximační parabolu volíme na kartě **Typ trendu a regrese** položku **Polynomický** Pořadí 2 (prostřední obrázek v pravém sloupci).

## Závěrečná poznámka

Obrázek 12: Převzat z [14]



Proč jsou tak oblíbené a časté — a bohužel také často tak úspěšné — lži pomocí statistik? Je tomu tak proto [14, str. 210], že průměrný člověk vyrostl v uctivé plachosti před čísly, která jsou obklopena posvátnou, ale nenapadnutelnou přesností matematiky.

Vzhledem k tomu, že statistika pracuje převážně s čísly, přenáší důvěřivý občan svůj vztah k počtům také na čísla statistiky — ačkoli vedle toho může docela dobře obstát ze zkušenosti získané přesvědčení, že **statistiky lžou**. Ve skutečnosti je obojí správné. Statistika používá matematických metod a matematické přesnosti a — statistika lže.

První pozitivní představa mimochodem převládá, jinak by také nebylo tolik pokusů lhát pomocí statistik. Představa, že **čísla dokazují**, není přes veškeré špatné zkušenosti překonána.

Jestliže je statistika (jako metodika nebo jako vědní obor) často posuzována s pochybnostmi a odmítavě, můžeme za to děkovat především statistikám, které ve skutečnosti statistikami nejsou. Je to stejné [14, 205] jako kdyby nemocného člověka léčil mastičkář, zřízenec nebo kuchyňský personál kliniky a nemocný pak mrzutě konstatoval: „Medicína není vůbec žádná věda; všichni lékaři jsou šarlatáni.“

Obrázek 13 názorně ukazuje, že stejnou věc je možné pozorovat z různých hledisek a podle toho statisticky různě vyjádřit.

Jestliže se tedy mluví o lži ve statistice, je nutno vždy zjistit, o jaký druh lži se jedná.

- Existuje především zdánlivá lež, která není v podstatě nic jiného než nesprávně pojatá přesná statistika. Je ovšem docela dobře možné, že je lstivě zaměřena na oklamání naivních lidí, ale sama o sobě (svými údaji a tvrzeními) je nenapadnutelná.
- Dále existuje odvozená lež, charakterizovaná tím, že se manipulátor „zmocní“ v podstatě správných čísel a buduje kolem nich konstrukce **lži** (vyhledává k nim vhodné *příčiny* či *následky*), která je nespornými čísly znamenitě udržována a posilována.
- Konečně existuje forma lži, při níž lze postupovat statisticky korektně jak při zpracování, tak při výkladu. Ovšem pracujeme se zfalšovaným prvotním materiálem.

Použitím nesprávných výchozích údajů (nebo dokonce vědomým falšováním „prvotního záznamu“, například — Irák vyvíjí nebo dokonce již vlastní jaderné zbraně  $\Rightarrow$  operace Pouštní bouře) je mož-

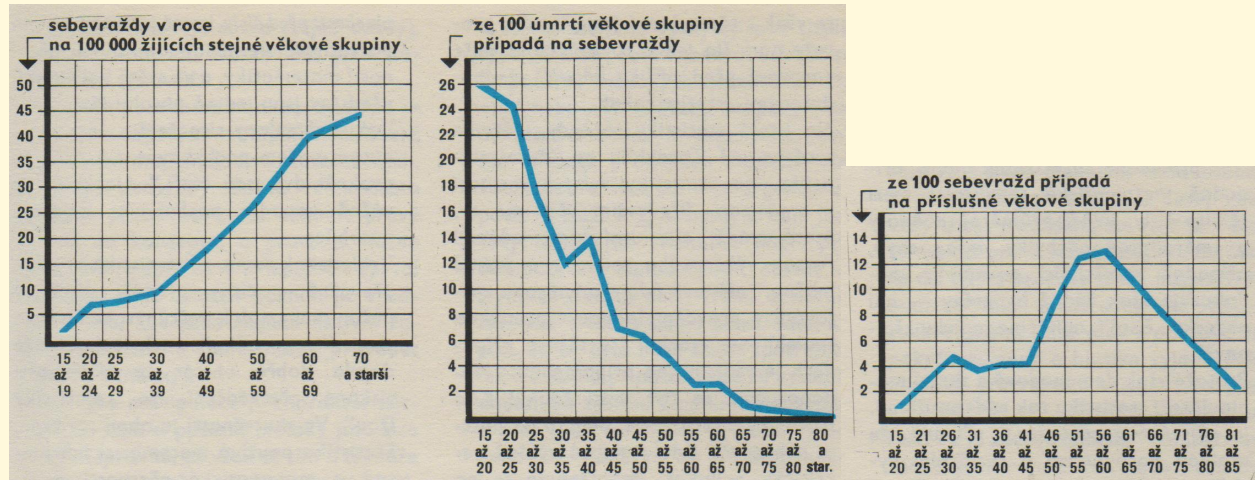


Obrázek 13: Převzat z [14]

**Vlevo** je případ, kdy jsme vzali žijící obyvatelstvo a sledovali počet sebevražd pro věkové kategorie.

**Uprostřed** je případ, kdy jsme vzali počty úmrtí v jednotlivých věkových skupinách a sledovali, kolik z nich připadá na sebevraždy.

**Vpravo** jsme vzali „úspěšné“ sebevrahy a sledovali, kolik jich je v jednotlivých věkových skupinách.



V závislosti na volbě základu vzniká zcela rozdílný obraz.

no dokázat všechno. I nekritickému čtenáři nebo posluchači bude obtížné namluvit, že **čtyři plus pět se rovná šest**, ale jestliže nejdříve zfalšujeme *pětku* na *dvojkou*, mohu potom plným právem tvrdit, že  $2 + 4 = 6$ .

Proto se doporučuje vždy nejdříve zjistit, zda se (v nemocnici) setkáváme s vrátným, ošetřovatelkou nebo primářem, zda s pseudostatistikou nebo se statistikou.

Rozlišení není na první pohled vždy snadné. I vrátný může v bílém plášti a v brýlích prohodit pár latinských slov a laikovi předstírat lékaře. Ještě snadnější je pro (většinou mladší) pracovní sílu osobního oddělení vypočítat na kalkulačce procenta s přesností na několik desetinných míst a tak vyrobit zdánlivě „pravou“ statistiku. A jak to poznat, když vůbec nemusí jít o úmysl?

Jak jednoduché je ze správných statistických údajů vyvodit nesmyslné závěry, můžeme dokumentovat na následujícím příkladě: *Je statisticky dokázáno, že každé čtvrté dítě, které se narodí, je Číňan.* Znamená to však něco při plánování počtu dětí pro průměrnou českou rodinu? Většina čtenářů asi tuší, že nikoliv. Jsme však schopni takový rozpor vždy odhalit?

Vždyť běžná praxe je, že původní číselný materiál byl někým (správně) interpretován a v této podobě předán do tisku. Potom nějaký novinář nikoli ve zlém úmyslu, nýbrž aby nerozzlobil čtenáře bojícího se čísel, část údajů vynechá a z komentovaného textu zvýrazní to, co působí alespoň trochu senzačně.

**W. J. Reichmann** [14, 206] komentuje například zprávu vytištěnou tučným písmem v jedněch anglických novinách: **„Každá druhá žena si stěžuje na bolesti v zádech“**, a uvádí pak, že již původní statistika obsahovala několik slabin.

Předně nešlo o základní soubor „ženy“ (mimo jiné by bylo zapotřebí vyjasnit, zda např. patnácti- či šestnáctileté slečny mají být zahrnuty či nikoliv atd.), nýbrž o pacientky. Ženy, které navštíví lékaře, jsou bezpochyby v průměru méně zdravé a trpí více bolestmi než ty, které nejsou v čekárnách ordinací. Tedy správně mělo být jen: **„Každá druhá PACIENTKA si stěžuje na bolesti v zádech.“**

Dále se ukázalo, že tento výsledek nebyl získán z reprezentativního anketního šetření mezi praktickými lékaři, nýbrž byl výsledkem soukromé statistiky jediného lékaře. Správně: **„Každá druhá pacientka DOKTORA X si stěžuje na bolesti v zádech.“**



Reichmann k tomu již zlomyslně poznamenává, že dotyčný lékař provozuje svoji praxi buď ve velmi vlhké krajině nebo má v čekárně dost nepohodlné židle. Ale to zdaleka není všechno ... (a původní citát pokračuje dále).

*Tak se scvrkává „statisticky podložené“ tvrzení, podle něhož si každá druhá žena stěžuje na bolesti v zádech na mnohem méně působivou skutečnost, že někde v Anglii je nějaký lékař, polovina jehož pacientek na otázku, zda také mají bolesti v zádech, odpovídá „ano“.*

V tomto případě bylo alespoň možné vystopovat na základě původní zprávy všechny zdroje chyb. Ale co má dělat čtenář, kterému se předkládá pod uvedeným titulkem hustý text, než se domnívat, že opravdu každá druhá žena v Anglii si stěžuje na bolesti v zádech?

Nyní si na jiném příkladu ukážeme manipulaci, která nemá demagogický záměr a přesto je značně matoucí. V roce 1966 sdělilo vídeňské letiště [14, str. 125]: „Mezi 37 západoevropskými letišti se Vídeň řadí ... sice ještě mezi menší letiště, pokud však jde o přírůstky dopravy, je Vídeň již na čtvrtém místě. V roce 1964 bylo při 22 818 startech odbaveno 725 049 cestujících ... V nejsilnějších dnech je registrováno až 5 000 cestujících.“

Zde je v několika málo slovech téměř vše, na co je nutno brát zřetel, chceme-li se naučit zacházet se statistikami. Jádrem výpovědi (**řadí se mezi menší**) je odsunuto stranou slůvkem „sice“ a pak se vynáší trumf: již na čtvrtém místě v přírůstcích. Toto „již“ je ale zcela nemístné, protože přírůstky jsou vysoké téměř vždy, jestliže je výchozí základna malá. Pak následují absolutní čísla pro určená pro laika, který nemá možnost porovnání: 22 818 letů a 725 049 cestujících — to je přece ohromné!

Absolutně ano, relativně nikoliv. Letiště Rýn–Mohan odbavilo ve stejném období téměř 4 miliony cestujících, nemluvě ani o Paříži, Londýně či amerických letištích.

A nakonec jako zvláštní pozoruhodnost poukáz na nejsilnější dny a v nich dosahované absolutní nejvyšší („až“) hodnoty. Věcně je jistě naprosto správné, že v jednom takovém „nejsilnějším dnu“ bylo jednou zaregistrováno až 5 000 cestujících. Protože se však současně neuvádí žádný denní průměr, utkví čtenáři

v mysli představa: „denně 5 000 cestujících“, i když toto tvrzení není ve zprávě výslovně řečeno (zcela jistě ne!). Čtení statistik se ještě nestalo všeobecně ovládaným uměním.» [konec citátu]

Příkladem statistiky vídeňského letiště jsme se podrobněji nezabývali proto, že by byla obzvlášť rafinovaná, záludná či demagogická, nýbrž proto, že umožňuje zřetelně ukázat, čeho se při čtení nějaké statistiky vyvarovat. *Nesnažit se vyčíst více, než je uvedeno.* „V nejsilnější dny až ...“ **neznamená** „denně“.

Téměř všechna čísla — a proto i všechny statistiky — je možno zneužít. Kdo nechce padnout za obět' takovému zneužití, kdo se nechce nechat od demagogů nebo přehorlivých novinářů vehnat do úzkých, bude se vždy s pochybností ptát: *Co se s čím srovnává? Má toto porovnání smysl a je oprávněné?* A především: *Netvrdí se v průvodním textu více, než dovolují čísla sama poznat? A konečně nikdy nemůže škodit, jestliže se zeptáme jako u soudu: Komu to slouží? Kdo se pomocí těchto čísel jeví ve zvlášť příznivém světle?*

**A na které statistiky se tedy můžeme spolehnout?** Zpravidla na úřední statistiky, na statistiky velkých institucí a organizací. Především však na ty, které uvádějí absolutní údaje, udávají rozsah výběrového souboru a pokud možno i některé údaje o způsobu zjišťování a pravděpodobnou teoretickou spolehlivost vzorku. Dobrá statistika poskytuje přehledně zpracované údaje, případně matematické souvislosti mezi těmito čísly, uvádí průměrné hodnoty a směrodatné odchylky, meze chyb, případně vysvětlující poznámky. Nedokazuje však žádné hypotézy — ani vědecky „čisté“, ani demagogicky „špinavé“.

Protože však demagogové a skrytí manipulátoři statistiku tak rádi používají, je užitečné zeptat se v případě jakýchkoliv pochybností o seriózním vyjádření ještě jednou nedůvěřivě **cui bono?** (Komu to prospívá?) Tato otázka pomáhá již po staletí odhalovat zločiny a osvědčuje se často jako velmi užitečná i při odhalování statistických podvodů.

## Použitá literatura

- [1] Český statistický úřad, www: <http://www.czso.cz/>
- [2] DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha : Univerzita Karlova v Praze – Karolinum, 4. nezměněné vydání, 2011. 372 stran. ISBN 978-80-246-1966-8
- [3] FRIEDRICH, V. *Statistika pro ekonomy*. Ostrava : VŠB-TUO, [skripta]. 2006. 241 stran.
- [4] KOVAŘÍK, P. *Aplikovaná statistika*. Brno : VŠKE, a. s. [skripta], 2011. 181 stran.  
ISBN 978-80-86710-28-0
- [5] KROPÁČ, J. *Statistika A. / Náhodné jevy, Náhodné veličiny, Náhodné vektory, Indexní analýza, Rozhodování za rizika*. Brno : Vysoké učení technické v Brně, Fakulta podnikatelská, druhé opravené vydání, 2007. 157 stran. ISBN 978-80-214-3194-6
- [6] KROPÁČ, J. *Statistika B. / Jednorozměrné a dvourozměrné datové soubory, Regresní analýza, Časové řady*. Brno : Vysoké učení technické v Brně, Fakulta podnikatelská, 2007. 155 stran.  
ISBN 80-214-3295-0
- [7] KROPÁČ, J. *Statistika C. / Statistická regulace, Indexy způsobilosti, Řízení zásob, Statistické přejímky*. Brno : Vysoké učení technické v Brně, Fakulta podnikatelská, 2008. 103 stran.  
ISBN 978-80-214-3591-9
- [8] LITSCHMANNOVÁ, M. *Úvod do statistiky*. [interaktivní učební text] Vysoká škola báňská — Technická univerzita Ostrava & Západočeská univerzita v Plzni, 2012. Dostupné z:  
[http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/interaktivni\\_uvod\\_do\\_statistiky.pdf](http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/interaktivni_uvod_do_statistiky.pdf)

- [9] LITSCHMANNOVÁ, M. *Vybrané kapitoly z pravděpodobnosti*. [interaktivní učební text] Vysoká škola báňská — Technická univerzita Ostrava & Západočeská univerzita v Plzni, 2012. Dostupné z:  
[http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/interaktivni\\_vybrane\\_kapitoly\\_pravdepodobnost.pdf](http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/interaktivni_vybrane_kapitoly_pravdepodobnost.pdf)
- [10] OTIPKA, P., ŠMAJSTRLA, P. *Pravděpodobnost a statistika*.  
<http://homen.vsb.cz/~oti73/cdpast1/>
- [11] PAVELKA, F., KLÍMEK, J. *Aplikovaná statistika*. Zlín : VUT–FAME 2000. 132 stran.  
ISBN 80–214–1545–2.
- [12] PAVLÍK J. A KOL. *Aplikovaná statistika*. Praha : Vysoká škola chemicko–technologická v Praze. 2005, 1. vydání, 173 stran. ISBN 80–7080–569–2
- [13] ŘEZANKOVÁ, H., MAREK, L., VRABEC, M. *IATEST — Interaktivní učebnice statistiky*.  
<http://iastat.vse.cz/>
- [14] SWOBODA, H. *Moderní statistika*. Praha : Svoboda, 1977. 352 stran

## Vybrané statistické tabulky

Na následujících stranách jsou uvedeny některé statistické tabulky:

**Distribuční funkce  $F_N(u)$  normovaného normálního rozdělení  $N(0, 1)$** , kdy využijeme poznatek, že každé rozdělení  $N(\mu, \sigma^2)$  lze transformací  $U = \frac{x-\mu}{\sigma}$  převést na normované  $N(0, 1)$ .

Hodnoty lze také získat pomocí *Excelu 2010* prostřednictvím funkce:

$$=\text{NORM.DIST}(x;\mu;\sigma;1)$$

**Kvantily rozdělení  $\chi^2(n)$**  používané například při Pearsonově testu shody (zda množina dat vyhovuje dané distribuční funkci). Platí, že rozdělení  $\chi^2(n)$  se s rostoucím  $n$  blíží **normálnímu rozdělení** se střední hodnotou  $n$  a rozptylem  $2n$ .

Hodnoty lze také získat pomocí *Excelu 2010* prostřednictvím funkce:

$$=\text{CHISQ.INV.RT}(\alpha;n)$$

**Kvantily Studentova rozdělení** Irský chemik a statistik W. S. Gosset roku 1908 poprvé publikoval toto rozdělení pod pseudonymem „*Student*“, protože jeho zaměstnavatel, pivovar Guinness v Dublinu, zakázal svým zaměstnancům publikovat pod svým vlastním jménem z obavy, že konkurence by odhalila tajemství jejich excelentního piva.

Pro vysoký počet stupňů volnosti (v praxi pro  $n > 30$ ) se Studentovo rozdělení blíží **normovanému normálnímu rozdělení**.

Hodnoty lze také získat pomocí *Excelu 2010* prostřednictvím funkce:

$$=\text{T.INV.2T}(\alpha;n)$$

# Distribuční funkce $F_N(u)$ normovaného normálního rozdělení $N(0, 1)$

x	0	1	2	3	4	5	6	7	8	9
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169

# Kvantily rozdělení $\chi^2_\alpha(n)$ — Excel 2010: „[=CHISQ.INV.RT(1 - $\alpha$ ; n)]“

$n \backslash \alpha$	0.005	0.01	0.025	0.05	0.1
1	0.00003	0.00015	0.00098	0.00393	0.01579
2	0.01002	0.02010	0.05063	0.10258	0.21072
3	0.07172	0.11483	0.21579	0.35184	0.58437
4	0.20698	0.29710	0.48441	0.71072	1.06362
5	0.41174	0.55429	0.83121	1.14548	1.61031
6	0.67572	0.87209	1.23734	1.63538	2.20413
7	0.98925	1.23904	1.68987	2.16735	2.83311
8	1.34441	1.64650	2.17973	2.73264	3.48954
9	1.73493	2.08790	2.70039	3.32511	4.16816
10	2.15586	2.55821	3.24697	3.94030	4.86518
11	2.60322	3.05348	3.81575	4.57481	5.57778
12	3.07382	3.57057	4.40379	5.22603	6.30380
13	3.56503	4.10692	5.00875	5.89186	7.04150
14	4.07467	4.66043	5.62873	6.57063	7.78953
15	4.60092	5.22935	6.26214	7.26094	8.54676
16	5.14221	5.81221	6.90766	7.96165	9.31224
17	5.69722	6.40776	7.56419	8.67176	10.0852
18	6.26480	7.01491	8.23075	9.39046	10.8649
19	6.84397	7.63273	8.90652	10.1170	11.6509
20	7.43384	8.26040	9.59078	10.8508	12.4426

# Pokračování: Kvantily rozdělení $\chi^2_\alpha(n)$ — Excel 2010: „[=CHISQ.INV.RT(1 – $\alpha$ ; n)]“

$n \backslash \alpha$	0.9	0.95	0.975	0.99	0.995
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2103	10.5966
3	6.2514	7.8147	9.3484	11.3449	12.8382
4	7.7794	9.4877	11.1433	13.2767	14.8603
5	9.2364	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5345	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5894
10	15.9872	18.3070	20.4832	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7568
12	18.5493	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3620	24.7356	27.6882	29.8195
14	21.0641	23.6848	26.1189	29.1412	31.3193
15	22.3071	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1565
19	27.2036	30.1435	32.8523	36.1909	38.5823
20	28.4120	31.4104	34.1696	37.5662	39.9968



# Kvantily $T_p(k)$ Studentova rozdělení — Excel 2010: „[=T.INV.2T(2 · (1 – $\alpha$ ); n)]“

n \ $\alpha$	0.9	0.95	0.975	0.99	0.995	0.999	n \ $\alpha$	0.9	0.95	0.975	0.99	0.995	0.999
1	3.07768	6.31375	12.7062	31.8205	63.6567	318.309	19	1.32773	1.72913	2.09302	2.53948	2.86093	3.57940
2	1.88562	2.91999	4.30265	6.96456	9.92484	22.3271	20	1.32534	1.72472	2.08596	2.52798	2.84534	3.55181
3	1.63774	2.35336	3.18245	4.54070	5.84091	10.2145	21	1.32319	1.72074	2.07961	2.51765	2.83136	3.52715
4	1.53321	2.13185	2.77645	3.74695	4.60409	7.17318	22	1.32124	1.71714	2.07387	2.50832	2.81876	3.50499
5	1.47588	2.01505	2.57058	3.36493	4.03214	5.89343	23	1.31946	1.71387	2.06866	2.49987	2.80734	3.48496
6	1.43976	1.94318	2.44691	3.14267	3.70743	5.20763	24	1.31784	1.71088	2.06390	2.49216	2.79694	3.46678
7	1.41492	1.89458	2.36462	2.99795	3.49948	4.78529	25	1.31635	1.70814	2.05954	2.48511	2.78744	3.45019
8	1.39682	1.85955	2.30600	2.89646	3.35539	4.50079	30	1.31042	1.69726	2.04227	2.45726	2.75000	3.38518
9	1.38303	1.83311	2.26216	2.82144	3.24984	4.29681	35	1.30621	1.68957	2.03011	2.43772	2.72381	3.34005
10	1.37218	1.81246	2.22814	2.76377	3.16927	4.14370	40	1.30308	1.68385	2.02108	2.42326	2.70446	3.30688
11	1.36343	1.79588	2.20099	2.71808	3.10581	4.02470	45	1.30065	1.67943	2.01410	2.41212	2.68959	3.28148
12	1.35622	1.78229	2.17881	2.68100	3.05454	3.92963	50	1.29871	1.67591	2.00856	2.40327	2.67779	3.26141
13	1.35017	1.77093	2.16037	2.65031	3.01228	3.85198	60	1.29582	1.67065	2.00030	2.39012	2.66028	3.23171
14	1.34503	1.76131	2.14479	2.62449	2.97684	3.78739	70	1.29376	1.66691	1.99444	2.38081	2.64790	3.21079
15	1.34061	1.75305	2.13145	2.60248	2.94671	3.73283	80	1.29222	1.66412	1.99006	2.37387	2.63869	3.19526
16	1.33676	1.74588	2.11991	2.58349	2.92078	3.68615	90	1.29103	1.66196	1.98667	2.36850	2.63157	3.18327
17	1.33338	1.73961	2.10982	2.56693	2.89823	3.64577	100	1.29007	1.66023	1.98397	2.36422	2.62589	3.17374
18	1.33039	1.73406	2.10092	2.55238	2.87844	3.61048	1000	1.28240	1.64638	1.96233	2.33008	2.58075	3.09840