

**Slezská univerzita v Opavě  
Obchodně podnikatelská fakulta v Karviné**



**STATISTICKÉ METODY PRO EKONOMY**  
**distanční studijní opora**

**Jaroslav Ramík – Radmila Stoklasová – Josef Tošenovský**

**Karviná 2005**



## OBSAH MODULU STATISTICKÉ METODY PRO EKONOMY

<b>1 ZÁKLADNÍ POJMY A METODY ZE STATISTIKY</b> .....	<b>7</b>
<b>1.1 Statistický soubor s jedním znakem - argumentem</b> .....	<b>8</b>
1.1.1 Charakteristiky polohy .....	10
1.1.2 Charakteristiky variability .....	11
1.1.3 Šikmost (skewness, z angl. vyslov: skjúnis).....	12
1.1.4 Špičatost (kurtosis, z angl. vyslov: kærtosis) .....	13
<b>1.2 Statistický soubor se dvěma znaky</b> .....	<b>14</b>
<b>1.3 Testy statických hypotéz</b> .....	<b>16</b>
<b>2 TESTOVÁNÍ HYPOTÉZ V MARKETINGU</b> .....	<b>27</b>
<b>2.1 Marketingová případová studie</b> .....	<b>28</b>
<b>2.2 Co přináší testování statistických hypotéz v marketingu?</b> .....	<b>28</b>
<b>2.3 Jednovýběrový t-test: řešení příkladu 2-1 pomocí MS Excelu</b> .....	<b>30</b>
<b>2.4 Dvouvýběrový t-test – nepárový a párový: řešení příkladů ze Studie</b> .....	<b>31</b>
<b>2.5 Mediánový t-test (pro 1 výběr)</b> .....	<b>34</b>
<b>2.6 Chi-kvadrát test pro 1 výběr</b> .....	<b>35</b>
<b>2.7 Dvouvýběrové testy</b> .....	<b>38</b>
<b>2.8 Chi-kvadrát test pro 2 výběry</b> .....	<b>38</b>
<b>3 REGRESNÍ ANALÝZA</b> .....	<b>45</b>
<b>3.1 Podstata regresní analýzy</b> .....	<b>45</b>
<b>3.2 Odhad regresních koeficientů</b> .....	<b>48</b>
<b>3.3 Test významnosti regresních koeficientů</b> .....	<b>52</b>
<b>3.4 Intervaly spolehlivosti regresních koeficientů</b> .....	<b>53</b>
<b>3.5 Test vhodnosti regresního modelu</b> .....	<b>54</b>
<b>4 KORELAČNÍ ANALÝZA</b> .....	<b>63</b>
<b>4.1 Koeficient korelace</b> .....	<b>64</b>
<b>4.2 Index korelace</b> .....	<b>66</b>
<b>4.3 Spearmanův koeficient (pořadové) korelace</b> .....	<b>67</b>
<b>4.4 Vícenásobná lineární závislost – vztahy pro dvě vysvětlující proměnné</b> .....	<b>68</b>
<b>5 METODY PROGNÓZOVÁNÍ TRŽEB</b> .....	<b>75</b>
<b>5.1 Časové řady – prognózování tržeb</b> .....	<b>76</b>
<b>5.2 Analýza trendové složky</b> .....	<b>77</b>
<b>5.3 Analýza sezónní složky</b> .....	<b>80</b>
5.3.1 Model konstantní sezónnosti .....	81
<b>5.4 Analýza náhodné složky</b> .....	<b>83</b>
<b>5.5 Testování vlastností náhodné složky</b> .....	<b>84</b>
<b>5.6 Prognózování v ČR</b> .....	<b>86</b>
<b>5.7 Kauzální prognostické metody</b> .....	<b>89</b>
<b>6 ANALÝZA ROZPTYLU</b> .....	<b>92</b>
<b>6.1 Jednofaktorová ANOVA</b> .....	<b>93</b>
<b>6.2 Postup při analýze rozptylu s jedním faktorem</b> .....	<b>95</b>
<b>6.3 Míra těsnosti závislosti</b> .....	<b>98</b>
<b>7 ANALÝZA ROZPTYLU (ANOVA): DVOJNÉ TRÍDĚNÍ A LATINSKÉ ČTVERCE</b> .....	<b>104</b>
<b>7.1 Dvojné třídění</b> .....	<b>105</b>
<b>7.2 Trojné třídění (Latinské čtverce)</b> .....	<b>109</b>

<b>8 ÚPLNÉ FAKTOROVÉ PLÁNY .....</b>	<b>119</b>
<b>8.1 Základy experimentování a oblasti použití.....</b>	<b>119</b>
<b>8.2 Experimentální procedura .....</b>	<b>120</b>
<b>8.3 Efekt (vliv) faktoru, významnost efektu .....</b>	<b>124</b>
8.3.1 Efekt (vliv) faktoru .....	124
8.3.2 Test významnosti efektu.....	126
<b>8.4 Grafické hodnocení efektu faktoru, grafy interakcí .....</b>	<b>127</b>
8.4.1 Grafické hodnocení efektu faktorů .....	127
8.4.2 Grafy interakcí.....	128
<b>8.5 Model experimentu 2<sup>3</sup> .....</b>	<b>129</b>
<b>9 ČÁSTEČNÝ FAKTOROVÝ EXPERIMENT SE DVĚMA ÚROVNĚMI.....</b>	<b>136</b>
<b>9.1 Poloviční plány .....</b>	<b>137</b>
<b>9.2 Grafická metoda.....</b>	<b>140</b>
<b>10 TAGUCHIHO METODY: ZTRÁTOVÁ FUNKCE .....</b>	<b>150</b>
<b>10.1 Definice a vlastnosti ztrátové funkce .....</b>	<b>151</b>
<b>10.2 Ztrátová funkce pro různé typy tolerance .....</b>	<b>153</b>
<b>11 TAGUCHIHO METODY: CELKOVÉ NÁKLADY NA JAKOST .....</b>	<b>163</b>
<b>11.1 Monitorování nákladů na jakost .....</b>	<b>164</b>
<b>11.2 Provádí se 100% kontrola procesu .....</b>	<b>165</b>
<b>11.3 Kontrola procesu se provádí po n jednotkách.....</b>	<b>166</b>
<b>11.4 Regulační diagramy .....</b>	<b>168</b>
<b>12 HODNOCENÍ ZPŮSOBILOSTI VÝROBNÍCH PROCESŮ .....</b>	<b>177</b>
<b>12.1 Co je způsobilost procesu .....</b>	<b>178</b>
<b>12.2 Jaké jsou cíle hodnocení způsobilosti .....</b>	<b>178</b>
<b>12.3 Jak se provádí výběr vhodného ukazatele .....</b>	<b>178</b>
<b>12.4 Předpoklady hodnocení způsobilosti.....</b>	<b>179</b>
<b>12.5 Atributy .....</b>	<b>179</b>
<b>12.6 Měřitelné charakteristiky kvality .....</b>	<b>180</b>
<b>12.7 Index C<sub>p</sub>.....</b>	<b>181</b>
<b>12.8 Index C<sub>pK</sub> .....</b>	<b>182</b>

## ÚVODEM MODULU STATISTICKÉ METODY PRO EKONOMY

Tento text představuje studijní oporu pro kombinované a distanční studium všech akreditovaných studijních programů v bakalářském a magisterském studiu na Slezské univerzitě, Obchodně podnikatelské fakultě v Karviné. Předmět Statistické metody pro ekonomy, který navazuje na předmět Kvantitativní metody B obsahující základní statistické metody, klade důraz na uplatnění statistických metod v aplikovaných ekonomických disciplínách, jako jsou marketing, management, plánování výroby a řízení jakosti.

Samotný učební text, nebo jak se říká v terminologii distančního studia: *studijní opora* - umožňující distančnímu studentovi plnohodnotné a zároveň samostatné studium – je rozčleněn do 12 tématických bloků. Jednotlivé bloky odpovídají obvyklým 12 výukovým týdnům jednoho semestru a jsou přibližně stejně obsahově rozsáhlé a obtížné. Takový rozsah učiva odpovídá klasické dvouhodinové přednášce v prezenčním studiu na vysoké škole ekonomického zaměření. V prezenčním studiu je ovšem přednáška doplněna seminářem – cvičením, kde se probraná látka aplikuje na konkrétní číselné příklady, které se řeší až k požadovanému výsledku často pomocí počítače.

Distanční vysokoškolské studium je specifická forma, která v případě předmětu Statistické metody pro ekonomy vyžaduje enormní úsilí studenta zaměřené na pravidelnost a vytrvalost v samostudiu, schopnost koncentrace na předmět, aktivní přístup spočívající v samostatném řešení příkladů. V tom všem by vám tato studijní opora měla pomoci nahradit kvalitní prezenční výuku i úlohu učebnic a skript. Distanční opora je k tomu účelu vybavena určitými nástroji, specifickými právě pro distanční formu, o jejichž funkcích byste měli vědět a mohli je tudíž účelně využívat ve svůj prospěch. Pro lepší zvládnutí látky jsou vám v příslušných adresářích k dispozici ještě doplňkové elektronické opory. Dalšími podpůrnými zdroji ke studiu mohou být učebnice a skripta z prezenčního studia a další doporučená literatura.

Zajisté jste si již povšimli, že text je opatřen širšími okraji, obsahujícími marginálie: hesla a ikony. Volné místo slouží k tomu, abyste jej zaplnili svými poznámkami a značkami k lepšímu pochopení studované látky (samozřejmě u opory v tištěné podobě). Význam ikon používaných v textu naleznete na konci modulu v seznamu použitých značek, symbolů a zkratk.

Předpokladem pro úspěšné zvládnutí tohoto kurzu Statistické metody pro ekonomy je zvládnutí kurzu Kvantitativní metody B. Ne všechno, co jste se v kurzu Kvantitativní metody B naučili, zde využijete, řada věcí tam prezentovaných měla jiný účel. Rozhodně se vám však vyplatí nabytá schopnost přesného a logického uvažování, nezbytností je též zvládnutí matematické symboliky a základních partií teorie pravděpodobnosti a základů inferenční statistiky.

Nyní něco k obsahu předmětu Statistické metody pro ekonomy. Přesnější název předmětu by zněl: Vybrané statistické metody pro ekonomy, nebo ještě přesněji: Vybrané statistické metody a jejich použití v marketingu, managementu a kontrole jakosti. To jsou totiž tři významné oblasti uplatnění statistických metod, s nimiž se absolventi Obchodně podnikatelské fakulty SU často v praxi setkávají. Uplatnění statistických metod v marketingu a managementu je obsahem kapitol 2 až 6, přičemž kapitola 1 je věnována zopakování základních statistických pojmů a metod. Kapitoly 7 až 12 jsou věnovány aplikacím statistických metod při plánování výroby a kontrole jakosti. Během studia budete využívat k řešení úloh známého programu Excel, s nímž jste pracovali již v předmětu Kvantitativní metody B.

Jak již bylo řečeno na začátku, text je rozdělen do 12 kapitol. Průměrně by vám samostatné studium každé kapitoly mělo zabrat 4 až 6 hodin. Odměna, která vás na konci studia našeho předmětu očekává, však stojí za to: je to pocit, že jste překonali něco významného, že jste se přenesli

---

přes překážku, za níž se nachází svět profesionálů, kteří rozumějí odborným metodám a postupům, jež jsou obyčejným smrtelníkům nepřístupné. Získaný nadhled vám umožní snadněji pochopit a osvojit si praktické zásady analýzy informací, jimiž jsme všichni dnes zahlceni.

# 1 ZÁKLADNÍ POJMY A METODY ZE STATISTIKY

## RYCHLÝ NÁHLED DO PROBLEMATIKY KAPITOLY ZÁKLADNÍ POJMY A METODY ZE STATISTIKY

Tato kapitola je koncipována jako opakování a malé rozšíření základních pojmů a metod ze základního kurzu statistiky (Kvantitativní metody B).

*Rychlý  
náhled*

## CÍLE KAPITOLY ZÁKLADNÍ POJMY A METODY ZE STATISTIKY

### Po úspěšném a aktivním absolvování této KAPITOLY

#### Budete umět:

- objasnit pojmy kumulativní a relativní četnost
- popsat statistický soubor pomocí číselných charakteristik
- popsat statistický soubor se dvěma znaky

*Budete  
umět*

#### Získáte:

- zopakujete si a pochopíte význam pojmů základní statistický soubor, výběrový soubor, náhodný výběr
- připomenete si pojmy: střední hodnota, rozptyl, kovariance

*Získáte*

#### Budete schopni:

- formulovat nulovou a alternativní hypotézu a rozhodnout, zda nulovou hypotézu lze či nelze zamítnout.

*Budete  
schopni*

## ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je 4 hodiny.

## PRŮVODCE STUDIEM 1



Cílem statistiky je odhalit zákonitosti a analyzovat informace, které jsou obsaženy v datech. V této kapitole se nejprve seznámíme se statistickým souborem s jedním znakem a najdete zde také již známé pojmy, jako je variační řada, rozsah souboru, četnosti atd. Statistický soubor lze popsat dvěma způsoby: pomocí grafických metod a pomocí číselných charakteristik. V druhé části se budeme zabývat statistickým souborem se dvěma znaky. V tomto případě se dá zjistit také např. vzájemná závislost daných znaků. V této úvodní kapitole si také připomeneme, co je testování statistických hypotéz. S některými testy hypotéz jste se již seznámili v kurzech Kvantitativní metody B, Statistika A nebo Statistika B. Aplikací testování hypotéz v marketingu – parametrickými i neparametrickými testy – se zabývají také další kapitoly 2 a 3. V závěru této kapitoly jsou uvedeny řešené příklady, které si samostatně projděte a pak se pusťte do příkladů k procvičení, které najdete na konci kapitoly.

**DEFINICE 1-1**

Veličina, která je předmětem statistického sledování, se nazývá *sledovaný statistický znak*. Hodnoty, které může znak nabývat, tvoří *základní statistický soubor (ZS)*, někdy se mu říká také *populační soubor*, či *populace*. Každá část ZS je tzv. *výběrový soubor (VS)*, nebo též *výběr*, či *vzorek*. Ten může být získán různými způsoby, nejčastěji tzv. *náhodným výběrem*, tj. výběrem, kdy každý prvek ZS má stejnou pravděpodobnost, že bude vybrán do vzorku. Náhodný výběr o rozsahu  $n$  interpretujeme jako *náhodný vektor*  $(X_1, \dots, X_n)$ . Statistické metody a postupy, které umožňují činit z údajů získaných ve výběrovém souboru závěry o základním souboru, nazýváme *statistickou indukci*.

**Statistický znak**

**Statistický soubor – základní / výběrový**

## 1.1 Statistický soubor s jedním znakem - argumentem

Jestliže  $x_1$  je vybraná hodnota, např.  $x_1 = 7$ , říkáme, že náhodná proměnná  $X_1$  nabyla hodnotu  $x_1$ .

Hodnota  $x_i$  se může vyskytovat i vícekrát, obecně  $f_i$  krát. Číslo  $f_i$  nazýváme *četností* hodnoty  $x_i$ .

Seřadíme-li hodnoty  $x_i$  podle velikosti do tabulky i s příslušnými četnostmi, budeme je značit  $x_{(i)}$ . Vznikne *variační řada* daná následující tabulkou:

$x_{(i)}$	$x_{(1)}$	$x_{(f_1+1)}$	...	$x_{\left(\sum_{i=1}^{k-1} f_i+1\right)}$
$f_i$	$f_1$	$f_2$	...	$f_k$

Celkem je v souboru  $k$  **různých** hodnot, přičemž  $k \leq n$ , neboť některé hodnoty se mohou opakovat. Všimněte si, že v prvním řádku výše uvedené tabulky se vyskytují uspořádané (rostoucí) vzájemně různé hodnoty.

$x_{(i)}$  jsou tzv. *pořádkové statistiky*, pro které platí:  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . Rozlišujeme tedy hodnoty  $x_i$ , kde index  $i$  určuje *pořadí měření*, např.  $x_1$  bylo získáno jako první údaj,  $x_5$  jako pátý údaj a hodnoty  $x_{(i)}$ , kde index  $i$  udává *pořadí této hodnoty podle velikosti*. Pak  $x_{(1)}$  je nejmenší hodnota, tj.  $x_{(1)} = x_{\min}$ ,  $x_{(5)}$  je v pořadí pátá nejmenší hodnota,  $x_{(n)}$  je největší hodnota z uvažovaných  $n$  hodnot, tj.  $x_{(n)} = x_{\max}$ .

Součet  $\sum_{i=1}^k f_i = n$  je *rozsah souboru*. Dále značíme

$$x_{\max} - x_{\min} = x_{(n)} - x_{(1)} = R.$$

Přitom  $R$  nazýváme *variační rozpětí*.



**ŘEŠENÝ PŘÍKLAD 1-1**

Měřením byly získány tyto údaje:

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
15	10	5	20	20

$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$
10	5	20	20	10

Sestavte variační řadu a určete rozsah souboru.

**Řešení příkladu**

Uspořádáním  $x_i$  podle velikosti dostáváme:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$
5	5	10	10	10

$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
15	20	20	20	20

Príslušná variační řada bude dána touto tabulkou

$x_{(i)}$	5	10	15	20
$f_i$	2	3	1	4

Rozsah souboru:  $\sum f_i = 2 + 3 + 1 + 4 = 10 = n$ .

Kromě četnosti  $f_i$  se dále používají tyto pojmy:

**a. kumulativní četnost:**  $F_i = f_1 + f_2 + \dots + f_i$ ,

**b. relativní četnost:**  $f_i / n$ ,

**c. relativní kumulativní četnost:**  $F_i / n$ .

**Pokračování příkladu:** Vypočtěte kumulativní četnost, relativní četnost a relativní kumulativní četnost.

$x_i$	$f_i$	$F_i$	$f_i/n$	$F_i/n$
5	2	2	0,2	0,2
10	3	5	0,3	0,5
15	1	6	0,1	0,6
20	4	10	0,4	1,0
	četnost	kumulativní četnost	relativní četnost	relativní kumulativní četnost

**Variační řada**



**Četnosti**

Každý statistický soubor lze popsat v zásadě dvěma způsoby:

- a. pomocí grafických metod
- b. pomocí číselných charakteristik

Grafické metody představují nejrůznější typy grafů: bodové, čárové, sloupcové, koláčové a jiné.

Připomeňte si je v Excelu: menu **Vložit** → **Graf...**

\*

## K ZAPAMATOVÁNÍ 1



**Charakteristiky**

Číselné charakteristiky jsou numerickým vyjádřením nezákladnějších vlastností statistického souboru. Podle toho, které vlastnosti popisují, je lze rozdělit na charakteristiky:

1. polohy
2. variability
3. šikmosti
4. špičatosti.

### 1.1.1 Charakteristiky polohy

*Aritmetický průměr* obdržíme jako součet jednotlivých výsledků měření nebo zjišťování vydělený celkovým počtem výsledků. Rozlišujeme přitom průměr z celého souboru údajů, nebo jen z určitého vzorku - výběru. Ten první nazýváme *populačním průměrem* a označujeme jej řeckým písmenem  $\mu$  (mí), pro ten druhý používáme označení  $\bar{x}$  s horním pruhem a nazýváme jej *výběrovým průměrem*. Zda se jedná o výběrový nebo populační průměr, závisí na konkrétní situaci. Vybereme-li z daného souboru všechny prvky, pak je výběrový a populační průměr totožný. Matematické vyjádření je následující:

$$\text{populační průměr} \quad \mu = \frac{1}{N} \sum_{i=1}^N x_i ,$$

$$\text{výběrový průměr} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i .$$

Pomocí vah  $w_i$  – kladných čísel dostáváme

$$\text{vážený aritmetický průměr} \quad \bar{x}_w = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i x_i .$$

Přitom  $N$  představuje počet údajů celého souboru,  $n$  představuje počet údajů z příslušného výběru.

Jsou-li k dispozici různé hodnoty  $x_i$  a k nim příslušné četnosti  $f_i$  (tj. počty opakování), pak populační, resp. výběrový průměr obdržíte také takto

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i f_i , \text{ resp. } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i f_i$$

*Modus*  $\hat{x}$  představuje nejčetnější hodnotu, tedy takovou hodnotu, která se v souboru vyskytuje nejčastěji. Je zcela nezávislý na ostatních hodnotách, které se mohou libovolně měnit, aniž se

modus změny. Modus tedy neříká nic o krajnostech, neprozrazuje, ve kterém směru a jak jsou velké.

Naproti tomu *medián*  $\tilde{x}$  představuje prostřední hodnotu v souboru hodnot, to je takovou hodnotu, kdy existuje stejný počet menších (nebo stejných) a stejný počet větších (nebo stejných) hodnot. Při sudém počtu hodnot neexistuje pochopitelně žádná prostřední hodnota. Medián se pak definuje jako aritmetický průměr z nejvyšší hodnoty dolní poloviny a nejnižší hodnoty horní poloviny hodnot uspořádaných podle velikosti. Takto funguje např. statistická funkce Median v Excelu. Lze se setkat též s definicí mediánu coby 50% kvantilu. V tom případě je medián největší hodnotou v dolní polovině uspořádaných hodnot.

### 1.1.2 Charakteristiky variability

*Rozptyl* je aritmetickým průměrem kvadrátů odchylek od aritmetického průměru. Podle toho, zda se jedná o rozptyl z celého souboru - celé populace, nebo jen rozptyl z jistého vzorku - výběru - z této populace, rozlišujeme populační rozptyl, kterému říkáme jednoduše *rozptyl*, označujeme jej  $\sigma^2$  ("sigma na druhou"), a *výběrový rozptyl*, označujeme jej  $s^2$  ("es na druhou").

Vzorce vypadají následovně:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2,$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}.$$

Číslo  $n-1$  nazýváme *počet stupňů volnosti*. Ve vzorcích uvádíme vždy dva tvary, první vyplývá přímo z definice, druhý dostaneme z prvního algebraickou úpravou a tento tvar je vhodnější pro výpočet.

Jsou-li k dispozici různé hodnoty  $x_i$  a k nim příslušné četnosti  $f_i$  (tj. počty opakování stejné hodnoty), pak populační, resp. výběrový rozptyl obdržíte také takto

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^K (x_i - \mu)^2 f_i = \frac{1}{N} \sum_{i=1}^K x_i^2 f_i - \mu^2, \text{ resp.}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i - \bar{x})^2 f_i = \frac{\sum_{i=1}^k x_i^2 f_i - n\bar{x}^2}{n-1}.$$

Přitom  $K$ , resp.  $k$  značí počet různých hodnot v populačním, resp. výběrovém souboru.

*Směrodatná odchylka* je odmocninou z rozptylu. Ve shodě s předchozí terminologií rozlišujeme *populační směrodatnou odchylku*, označujeme ji  $\sigma$ , které říkáme prostě *směrodatná odchylka*, a *výběrovou směrodatnou odchylku*, která je odmocninou z výběrového rozptylu, označujeme s.

Je zřejmé, že průměry, zejména často používaný aritmetický průměr, potřebují k správnému zhodnocení svého významu a vypovídací schopnosti ještě alespoň jeden rozměr. Takovým nejjednodušším doplněním je, že se uvede, jak jsou extrémní od sebe vzdáleny. K tomu slouží *variální rozpětí*  $R$ :

$$R = x_{\max} - x_{\min}.$$

*Variační koeficient* poskytuje nástroj, který je nezávislý na měrných jednotkách a umožňuje srovnávání. Používá se často jako míra rizika, např. cenných papírů při investování. Definujeme jej jako podíl průměru a směrodatné odchylky a vyjadřujeme jej často v procentech:

$$V = \frac{\sigma}{\mu}, \text{ resp. } v = \frac{s}{\bar{x}},$$

podle toho, jedná-li se o populační, resp. výběrový variační koeficient. Pro vyjádření variačního koeficientu v procentech (z průměru) násobíme  $V$ , resp.  $v$  číslem 100.

### ŘEŠENÝ PŘÍKLAD 1-2



Vypočítejte průměr, rozptyl a variační koeficient z výběrového souboru 10 hodnot  $x_i$  uvedených v následujících tabulce:

$i$	$x_i$
1	2
2	4
3	4
4	6
5	4
6	8
7	4
8	2
9	4
10	8

#### Řešení příkladu

$$\bar{x} = 0,1(2 + 4 + \dots + 8) = 4,60$$

$$s^2 = 0,1(2^2 + 4^2 + \dots + 8^2) - 4,6^2 = 4,04$$

$$v = \frac{s}{\bar{x}} = \frac{\sqrt{4,04}}{4,60} = 0,437$$

Průměrná hodnota daného souboru dat je 4,6 rozptyl je 4,04 a variační koeficient nabývá hodnoty 0,437.

Je-li  $v > 0,5$ , považuje se rozptyl za velký.



\*

### 1.1.3 Šikmost (skewness, z angl. vyslov: skjúnis)

$$g_1 = Sk = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^3 f_i}{s^3}$$

Šikmost

je charakteristikou, popisující symetrii pravděpodobnostního rozdělení (četností  $f_i$ ) vzhledem k aritmetickému průměru. Přitom druhou část vzorce použijete výhodně tehdy, jsou-li k dispozici různé hodnoty  $x_i$  a k nim příslušné četnosti  $f_i$  (tj. počty opakování stejné hodnoty).

### 1.1.4 Špičatost (kurtosis, z angl. vyslov: kærtosis)

Špičatost

$$g_2 = Ku = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} = \frac{\frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^4 f_i}{s^4}.$$

Druhou část vzorce je výhodné použít tehdy, jsou-li k dispozici různé hodnoty  $x_i$  a k nim příslušné četnosti  $f_i$ .

Při výpočtu číselných charakteristik je možné zjednodušit výpočty pomocí transformace dat.

Jednoduchou transformaci ukážeme na příkladě výpočtu šikmosti a špičatosti  $g_1$  a  $g_2$ .

#### ŘEŠENÝ PŘÍKLAD 1-3



Ze souboru hodnot  $x_i$  uvedených v následující tabulce (sloupec 1 a 2), vypočítejte šikmost a špičatost.

1	2	3	4	5	6	7
$x_i$	$f_i$	$u_i$	$f_i u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
9,52	2	-4	-8	32	-128	512
9,72	1	-3	-3	9	-27	81
9,92	4	-2	-8	16	-32	64
10,12	4	-1	-4	4	-4	4
10,32	9	0	0	0	0	0
10,52	3	1	3	3	3	3
10,72	5	2	10	20	40	80
10,92	4	3	12	36	108	324
11,12	2	4	8	32	128	512
11,32	0	5	0	0	0	0
11,52	1	6	6	36	216	1296
Sumy	35		$S_1=16$	$S_2=188$	$S_3=304$	$S_4=2876$

#### Řešení příkladu

Transformace dat, kdy přecházíme od proměnné  $x$  k proměnné  $u$ , se provede takto

$$u_i = \frac{x_i - x_0}{h},$$

kde  $x_0$  je jedna z hodnot. Volíme  $x_0 = 10,32$ , přitom  $h$  je vzdálenost sousedních znaků (předpokládá se, že je stejná). Zde je  $h = 0,2$  a  $n = 35$ .

Součty ve sloupcích 4 až 7, označené  $S_1$  až  $S_4$  (index odpovídá mocnině  $u$ ) se použijí k výpočtu tzv. *obecných momentů*  $M_1 - M_4$ . Z nich se pak vypočítají potřebné charakteristiky

$$M_i = S_i/n, i = 1,2,3,4,$$

konkrétně

$$M1 = 16/35 = 0,457$$

$$M2 = 188/35 = 5,371$$

$$M3 = 304/35 = 8,686$$

$$M4 = 2876/35 = 82,171$$

Nyní platí:

a. průměr nové proměnné  $u$ :  $\bar{u} = M1 = 0,457$

b. rozptyl  $u$ :  $s_u^2 = M2 - M1^2 = 5,162$

c. šikmost  $u$ :  $g_{1u} = (M3 - 3M1.M2 + 2M1^3) / s^3 = 1,513/11,728 = 0,096$

d. špičatost  $u$ :  $g_{2u} = (M4 - 4M3.M1 + 6M2.M1^3 - 3M1^4) / s^4 = 72,890 / 26,646 = 2,735$

Na závěr je nutné provést zpětnou transformaci (vrátit se k původní proměnné)  $x$ :

a. průměr  $x$ :  $\bar{x} = \bar{u}.h + x_o = 0,457.0,2 + 10,32 = 10,411$

b. rozptyl  $x$ :  $s_x^2 = s_u^2.h^2 = 5,1622.0,2^2 = 0,206$

c. šikmost  $x =$  šikmost  $u$ .

d. špičatost  $x =$  špičatost  $u$ .

Zpětná transformace u šikmosti a špičatosti není zapotřebí.

\*

## 1.2 Statistický soubor se dvěma znaky

### DEFINICE 1-2



Máme-li statistický soubor o rozsahu  $n$  takový, že každému prvku tohoto souboru jsou přiřazeny hodnoty dvou znaků  $(x_i, y_j)$ , hovoříme o *statistickém souboru se dvěma znaky* (též *argumenty*). Počet stejných dvojic  $(x_i, y_j)$  se nazývá *četností* těchto dvojic a značí se  $f_{ij}$ . Veličina  $f_{ij}/n$  je *relativní četnost* dvojice  $(x_i, y_j)$ . Rozsah souboru  $n = \sum_{i,j} f_{ij}$ .

Tabulka rozdělení četností má tvar

$y$	$y_1$	$y_2$	...	$y_n$	
$x$					
$x_1$	$f_{11}$	$f_{12}$	...	$f_{1n}$	$f_{1\cdot}$
$x_2$	$f_{21}$	$f_{22}$	...	$f_{2n}$	$f_{2\cdot}$
...	...	...	...	...	...
$x_m$	$f_{m1}$	$f_{m2}$	...	$f_{mn}$	$f_{m\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$		$f_{\cdot n}$	$n$

Hovoříme o *tabulce se dvěma vstupy* nebo též *kontingenční tabulce*. V posledním řádku a po-

sledním sloupci jsou tzv. *marginální četnosti*  $f_{1\cdot}, f_{2\cdot}, \dots, f_{m\cdot}, f_{\cdot 1}, f_{\cdot 2}, \dots, f_{\cdot n}$ .

Pro rozsah souboru platí:  $\sum_{i,j} f_{ij} = \sum f_{\cdot j} = \sum f_{i\cdot} = n$ .

Při zpracování dvourozměrného souboru v tabulce postupujete prakticky stejně jako u jedno-rozměrného souboru: nejprve zjistíte četnosti jednotlivých hodnot  $x_i$  (resp.  $y_j$ ) bez ohledu na druhou proměnnou (marginální četnosti). V tabulce jsou označeny  $f_{\cdot i}$ , resp.  $g_{i\cdot}$ . Potom už postupujete pro každou proměnnou stejně jako u jednorozměrného souboru.

## Přehled vzorců pro statistický soubor se dvěma znaky:

**Důležité  
vztahy**

### 1. Střední hodnoty

$$\bar{x} = \frac{1}{n} \sum_j \sum_i x_i f_{ij}$$

$$\bar{y} = \frac{1}{n} \sum_i \sum_j y_j f_{ij}$$

### 2. Rozptyly

$$s_x^2 = \frac{1}{n} \sum_j \sum_i (x_i - \bar{x})^2 f_{ij} = \frac{1}{n} \sum_j \sum_i x_i^2 f_{ij} - \bar{x}^2$$

$$s_y^2 = \frac{1}{n} \sum_i \sum_j (y_j - \bar{y})^2 f_{ij} = \frac{1}{n} \sum_i \sum_j y_j^2 f_{ij} - \bar{y}^2$$

### 3. Kovariance

$$c_{xy} = \frac{1}{n} \sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = \frac{1}{n} \sum_i \sum_j x_i y_j f_{ij} - \bar{x}\bar{y}$$

## ŘEŠENÝ PŘÍKLAD 1-4



Pro ilustraci si ukážeme 3 příklady

**Kovariance**

### Řešení příkladu

**a.**

$x_i$	1	2	3	4	5
$y_j$	3	4	5	6	7

Data jsou volena tak, aby s rostoucími hodnotami znaku  $x$  rostly lineárně také hodnoty znaku  $y$ :

$$\bar{x} = 3, \bar{y} = 5, c_{xy} = (1/5) \cdot 0,85 - 3 \cdot 5 = 2$$

**b.**

$x_i$	1	2	3	4	5
$y_j$	5	4	3	2	1

Zde s rostoucím  $x$  klesá  $y$ :

$$\bar{x} = 3, \bar{y} = 5, c_{xy} = (1/5) \cdot 35 - 3 \cdot 3 = -2$$

c.

$x_i$	1	2	3	4	5
$y_j$	2	2	2	2	2

Při rostoucím  $x$  zůstává  $y$  konstantní:

$$\bar{x} = 3, \bar{y} = 5, c_{xy} = (1/5) \cdot 30 - 3 \cdot 2 = 0$$

Můžeme učinit následující závěry, které platí i obecně, platí totiž:

$c_{xy} > 0 \Leftrightarrow$  přímá (lineární) závislost mezi znaky  $x$  a  $y$ ,

$c_{xy} < 0 \Leftrightarrow$  nepřímá (lineární) závislost mezi znaky  $x$  a  $y$ ,

$c_{xy} = 0 \Leftrightarrow$  nezávislost znaků  $x$  a  $y$ .

Z hodnoty  $c_{xy}$  však nelze stanovit *velikost závislosti*  $x$  a  $y$ . Proto se častěji používá *koefficient korelace*  $r_{xy}$ , což je normovaná kovariance

$$r_{xy} = \frac{c_{xy}}{s_x s_y},$$

viz též kapitola 5.

U statistického souboru se dvěma argumenty jsou některé charakteristiky, které nebyly potřebné u jednoho argumentu: zde se sleduje navíc vzájemná *závislost* znaků  $x$  a  $y$ . Pod pojmem závislost se dále bude rozumět

- tvar* závislosti: jak se na základě změny  $x$  mění  $y$ . Hledáním tvaru závislosti se zabývá regresní analýza (viz kapitola 4)
- míra* závislosti, kde se kvantifikuje stupeň závislosti  $x$  a  $y$  pomocí různých charakteristik. Tímto problémem se zabývá korelační analýza (viz kapitola 5).

\*



**Míra  
závislosti**

## 1.3 Testy statických hypotéz

### PRŮVODCE STUDIEM 2



Testováním statistických hypotéz jste se zabývali v základním kurzu statistiky. Protože jde o jednu z nejčastěji používaných statistických metod jak v oblasti marketingového výzkumu, tak i v jiných aplikačních oblastech ekonomie, zopakujeme si zde základní pojmy, metody a postupy. Statistické hypotézy tvoří jen část vědeckých (nebo alespoň odborných) hypotéz. Týkají se náhodných veličin a rozdělujeme je do dvou velkých tříd na parametrické hypotézy a neparametrické hypotézy.

Parametrické hypotézy se vztahují na jeden nebo několik parametrů daného pravděpodobnostního rozdělení náhodné veličiny (neboli znaku populace). Neparametrické hypotézy se netýkají parametrů rozdělení náhodné veličiny, nýbrž jiných statistických vlastností, např. tvaru rozdělení (například binomické rozdělení nebo normální rozdělení).

Zopakujme si, že v každém testu hypotézy vystupují proti sobě dvě hypotézy: testovaná hypotéza, kterou nazýváme *nulová hypotéza* a značíme  $H_0$  a *alternativní hypotéza*, značí se  $H_1$ . Obecně  $H_1$  nemusí být negací (nebo doplňkem)  $H_0$ .

Při testování parametrické hypotézy máme k dispozici výsledky náhodného výběru – vzorku – např. data z marketingového výzkumu a na jejich základě rozhodujeme testovanou hypotézu buď přijmout, nebo zamítnout. Za tím účelem rozdělíme výběrový prostor na dvě části: *kritický*



obor  $C$  a obor přijetí  $A$ . Padne-li hodnota *testového kritéria*  $T$  pro získaný vzorek do kritického oboru, potom ji zamítáme. Naopak, padne-li hodnota kritéria pro získaný vzorek do oboru přijetí, pak nulovou hypotézu nezamítáme (neboli přijímáme). Kritický obor odděluje od oboru přijetí obvykle *kritická hodnota*  $K$ .

Pozor, testem statistické hypotézy nemůžeme dokázat její platnost nebo neplatnost, či správnost nebo nesprávnost! Testování hypotéz nemusí vždy vést ke správným rozhodnutím, což je přirozené, neboť jde o náhodný proces využívající omezené informace náhodného výběru. Při praktickém testování hypotéz navrhne 4 kroky, které povedou k rozhodnutí o zamítnutí, resp. přijetí nulové hypotézy.

### Princip testování hypotézy

Sestrojí se *testovací kritérium*  $T$ , (též testové kritérium) které je z hlediska teorie pravděpodobnosti náhodná veličina. Kritérium  $T$  se sestrojí tak, aby mělo některé ze známých tabelovaných rozdělení a to právě tehdy, když platí  $H_0$ . Vypočítanou hodnotu  $T$  chápeme jako realizaci této náhodné veličiny. Ta se porovná s kritickou hodnotou, neboli zjistí se, zda hodnota kritéria padne do kritického oboru.

**Princip  
testování  
hypotézy**

## K ZAPAMATOVÁNÍ 2



**Postup  
testování**

### Obecný postup testování

1. Formulace nulové hypotézy  $H_0$  a alternativní hypotézy  $H_1$ ,
2. výpočet testovacího kritéria  $T$ ,
3. nalezení kritické hodnoty  $K$ ,
4. porovnání  $K$  a  $T$ , přijetí nebo zamítnutí  $H_0$ .

Jak jsme již řekli, kritická hodnota  $K$  rozděluje obor hodnot náhodné veličiny  $T$  na dvě části:

- kritický obor  $C$ ,
- obor přijetí  $A$ .

Je-li  $T > K$ , neboli  $T \in C$ ,  $H_0$  se zamítá.

Je-li  $T \leq K$ , neboli  $T \in A$ ,  $H_0$  se přijímá - nezamítá.

Protože rozhodnutí přijmout nebo zamítnout hypotézu závisí na omezené informaci ze vzorku, můžete se při testování dopustit těchto chyb:

**a.** Zamítnete hypotézu, která platí, to je *chyba prvního druhu*.

Pravděpodobnost chyby prvního druhu se značí  $\alpha$  (=hladina významnosti) nebo  $p$  (někdy též  $p$ -value /v Excelu/ =  $p$ -hodnota, též Sig. = signifikance /v SPSS/) a její smysl vyjadřuje vztah

$$\alpha = P(T \in C | H_0).$$

**b.** Přijmete hypotézu, která neplatí, to je *chyba druhého druhu*.

Pravděpodobnost chyby druhého druhu se značí  $\beta$  a její smysl vyjadřuje vztah

$$\beta = P(T \in A | H_1).$$

Pravděpodobnost  $1 - \beta$  se nazývá *síla testu*. Je to pravděpodobnost, že test povede k oprávněnému zamítnutí testované hypotézy  $H_0$ .

V posledním odstavci této úvodní kapitoly naleznete vybrané základní parametrické testy oboustranných hypotéz, s nimiž jste se seznámili v základním kurzu statistiky:

- (A) Jednovýběrový  $t$  – test.
- (B) Dvouvýběrový  $t$  – test s rovností rozptylů.

**Základní  
parametrické  
testy  
oboustranných  
hypotéz**

(C) Dvouvýběrový t – test s nerovností rozptylů.

(D) Dvouvýběrový párový t-test.

(E) Dvouvýběrový F – test pro rozptyly.

Každý test je uveden ve čtyřkrokovém postupu jeho použití a lze pro něj též využít program **MS Excel**, kde v menu **Nástroje** → **Analýza dat...** zvolíte příslušné analytické nástroje.

**(A) Test předpokladu o střední hodnotě základního souboru (Jednovýběrový t – test střední hodnoty)**

*Jednovýběrový t-test*

Nechť  $X = (X_1, \dots, X_n)$  je náhodný výběr z rozdělení  $N(\mu, \sigma^2)$ , přitom  $\sigma^2$  není známo.

1. Testuje se nulová hypotéza

$$H_0: \mu = \mu_0 \text{ proti alternativě } H_1: \mu \neq \mu_0.$$

2. Testovací kritérium je

$$T = \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} \quad (1-1)$$

kde  $\bar{X}$  = výběrový průměr,  
 $S$  = výběrová směrodatná odchylka,  
 $\mu_0$  = předpoklad o  $\mu$ ,  
 $n$  = rozsah výběru.

3. Kritická hodnota  $K = t_{n-1}(\alpha/2)$  se určí z tabulek Studentova rozdělení pro  $n-1$  stupně volnosti a hladinu významnosti  $\alpha$ . Pro oboustranný test se použije  $\alpha/2$ , neboť obor přijetí  $A$  je symetrický interval  $A = [-K, K]$  a kritický obor  $C$  je jeho doplňkem,  $C = (-\infty, -K) \cup (K, +\infty)$ .

K výpočtu kritické hodnoty pomocí Excelu použijte funkci  $TINV(\alpha/2; n-1)$  v menu: **Vložit** → **Funkce** → **Statistické** → **TINV...**

*Výpočet kritické hodnoty v Excelu*

4. Je-li  $|T| > t_{n-1}(\alpha/2)$ , zamítá se  $H_0$  a přijímá  $H_1$ , jinak se přijímá  $H_0$ .

**(B) Test významnosti rozdílu mezi dvěma výběrovými průměry (Dvouvýběrový t-test s rovností rozptylů)**

*Dvouvýběrový t-test s rovností rozptylů*

Jsou dány dva náhodné výběry o rozsahu  $n_1$  a  $n_2$  z rozdělení  $N(\mu_1, \sigma_1^2)$  a  $N(\mu_2, \sigma_2^2)$  a výběrové průměry  $\bar{X}_1, \bar{X}_2$ . Rozptyly  $\sigma_1^2$  a  $\sigma_2^2$  jsou neznámé, předpokládá se, že  $\sigma_1^2 = \sigma_2^2$  (tj. rovnost rozptylů).

1. Testuje se nulová hypotéza

$$H_0: \mu_1 = \mu_2 \text{ proti alternativě } H_1: \mu_1 \neq \mu_2.$$

2. Použije se testovací kritérium  $T$ :

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2}}, \quad (1-2)$$

Kde  $\bar{X}_1, \bar{X}_2$  jsou výběrové průměry,  
 $S_1^2, S_2^2$  výběrové rozptyly,  
 $n_1, n_2$  rozsahy 1. a 2. výběru.

3. Kritická hodnota  $K = t_{n_1+n_2-2}(\alpha/2)$  se určí z tabulek Studentova rozdělení pro  $n_1+n_2-2$  stupně

volnosti a hladinu významnosti  $\alpha$ .

K výpočtu kritické hodnoty  $K = t_{n_1+n_2-2}(\alpha/2)$  pomocí Excelu použijte funkci

$$\text{TINV}(\alpha/2; n_1+n_2-2).$$

4. Je-li  $T > K$  zamítá se  $H_0$  a přijímá  $H_1$ , jinak se přijímá  $H_0$ .

### (C) Test významnosti rozdílu mezi dvěma výběrovými průměry (Dvouvýběrový t-test s nerovností rozptylů)

*Dvouvýběrový t-test s nerovností rozptylů*

Dány dva náhodné výběry o rozsahu  $n_1$  a  $n_2$  z rozdělení  $N(\mu_1, \sigma_1^2)$  a  $N(\mu_2, \sigma_2^2)$  a výběrové průměry  $\bar{X}_1, \bar{X}_2$ . Rozptyly  $\sigma_1^2$  a  $\sigma_2^2$  nejsou známy, předpokládá se, že  $\sigma_1^2 \neq \sigma_2^2$  (tj. nerovnost rozptylů).

1. Testuje se hypotéza

$$H_0: \mu_1 = \mu_2 \text{ proti alternativě } H_1: \mu_1 \neq \mu_2.$$

2. Použije se testovací kritérium:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{V_1 + V_2}}, \quad (1-3)$$

$$\text{kde } V_i = \frac{S_i^2}{n_i - 1}, \quad i = 1, 2.$$

3. Kritická hodnota  $K$  se stanoví podle vzorce:

$$K = \frac{V_1 \cdot t_{n_1-1}(\alpha/2) + V_2 \cdot t_{n_2-1}(\alpha/2)}{V_1 + V_2}, \quad (1-4)$$

kde  $t_{n_1-1}(\alpha/2), t_{n_2-1}(\alpha/2)$  se určí z tabulek Studentova rozdělení pro  $n_1-1$  a  $n_2-2$  stupně volnosti a hladinu významnosti  $\alpha$ .

K výpočtu kritické hodnoty  $K$  podle (1-4) pomocí Excelu použijte funkce

$$\text{TINV}(\alpha/2; n_1-1) \text{ a } \text{TINV}(\alpha/2; n_2-1).$$

4. Je-li  $T > K$ , zamítáme  $H_0$  a přijímáme  $H_1$ , jinak přijímáme  $H_0$ .

### K ZAPAMATOVÁNÍ 3



*Využití Excelu*

Všimněte si, že tento test se redukuje na test (A) v tom případě, kdy druhý výběr má stejný rozsah jako první, tj.  $n_1 = n_2 = n$ , a dále všechny hodnoty ve výběru jsou stejné tj.  $\mu_0$ . Potom je totiž rozptyl souboru složeného jen ze stejných hodnot roven nule a platí:  $S_2^2 = V_2 = 0$ . Dále je zřejmé, že  $\bar{X}_2 = \mu_0$  a jak vztah pro testové kritérium tak kritickou hodnotu se redukuje na analogické vztahy uvedené v testu (A). Tohoto faktu lze využít při použití Excelu k řešení testu (A) pomocí testu (C). Test (A) totiž není v **Analytických nástrojích** Excelu k dispozici, na rozdíl od testu (C), který k dispozici je.

### (D) Dvouvýběrový párový t-test

*Párový t-test*

Jsou dány dva spárované náhodné výběry  $X$  a  $Y$  o rozsahu  $n$  z rozdělení  $N(\mu_1, \sigma_1^2)$  a  $N(\mu_2, \sigma_2^2)$  s výběrovými průměry  $\bar{X}, \bar{Y}$ . Přitom spárování znamená, že všechny dvojice náhodných veličin

$X_i$  a  $Y_i$  spolu souvisejí.

1. Testuje se nulová hypotéza

$$H_0: \mu_1 = \mu_2 \text{ proti alternativě } H_1: \mu_1 \neq \mu_2.$$

2. Použije se testovací kritérium:

$$T = \frac{\bar{D}}{S_D} \sqrt{n}, \quad (1-5)$$

kde  $D_i = X_i - Y_i, i = 1, 2, \dots, n, \bar{D} = \bar{X} - \bar{Y}$  a

$$S_D = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2} \quad (1-6)$$

3. Kritická hodnota  $K = t_{n-1}(\alpha/2)$  se určí z tabulek Studentova rozdělení pro  $n-1$  stupňů volnosti a hladinu významnosti  $\alpha$ .

K výpočtu kritické hodnoty  $K = t_{n-1}(\alpha/2)$  pomocí Excelu použijte funkci

$$\text{TINV}(\alpha/2; n-1).$$

4. Je-li  $T > K$ , zamítá se  $H_0$  a přijímá  $H_1$ , jinak se přijímá  $H_0$ .

### (E) Dvouvýběrový F – test pro rozptyly

F-test

Jsou dány dva náhodné výběry z rozdělení  $N(\mu_1, \sigma_1^2)$  a  $N(\mu_2, \sigma_2^2)$  o rozsahu  $n_1$  a  $n_2$  a výběrové rozptyly  $S_1^2$  a  $S_2^2$ .

1. Testuje se hypotéza, že rozptyly obou rozdělení jsou stejné, tj.

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ proti alternativní hypotéze } H_1: \sigma_1^2 \neq \sigma_2^2$$

2. Testovací kritérium

$$T = \frac{S_1^2}{S_2^2} \quad (1-7)$$

se sestaví tak, aby větší z rozptylů  $S_1^2$  a  $S_2^2$  byl v čitateli.

3. Kritická hodnota  $K = F_{n_1-1, n_2-1}(\alpha/2)$  se určí z tabulek Fischerova rozdělení pro  $n_1-1$  a  $n_2-1$  stupně volnosti a hladinu významnosti  $\alpha$ .

K výpočtu kritické hodnoty  $K = F_{n_1-1, n_2-1}(\alpha/2)$  pomocí Excelu použijte funkci

$$\text{FINV}(\alpha/2; n_1-1; n_2-1).$$

4. Je-li  $T > K$ , zamítá se  $H_0$  a přijímá  $H_1$ .

### ŘEŠENÝ PŘÍKLAD 1-5



Každé ze dvou polí bylo rozděleno na 10 lánů a zaseto pšenice. Přitom na lánech prvního pole bylo použito fosfátové hnojivo. Výnosy z lánů prvního a druhého pole měly průměry  $\bar{x}_1 = 6, \bar{x}_2 = 5,7$  a rozptyly  $s_1^2 = 0,064, s_2^2 = 0,024$ .

Zjistěte na 5% hladině významnosti, zda hnojení mělo průkazný vliv na výnosy.

Výnosy  
pšenice

**Řešení příkladu**

Nejprve pomocí F – testu ověříte, zda  $\sigma_1^2 = \sigma_2^2$ , abyste mohli vybrat správnou variantu dvouvýběrového t – testu.

Testovací kritérium bude  $T = 0,064 / 0,024 = 2,67$ .

Kritická hodnota je  $K = F_{n_1-1, n_2-1}(\alpha/2) = F_{9,9}(0,025) = 4,03$ .

Protože je  $T < K$ , přijímáte  $H_0$ , tedy  $\sigma_1^2 = \sigma_2^2$ .

Při provádění požadovaného t – testu volíte tedy variantu (B).

Pomocí tohoto dvouvýběrového t – testu porovnáte výběrové průměry  $\bar{x}_1 = 6, \bar{x}_2 = 5,7$ , je-li  $s_1^2 = 0,064, s_2^2 = 0,024, n_1 = n_2 = 10$ .

Použijete testovací kritérium

$$T = \frac{6 - 5,7}{\sqrt{9 \cdot 0,064 + 9 \cdot 0,024}} \cdot \sqrt{\frac{10 \cdot 10 \cdot (10 + 10 - 2)}{10 + 10}} = 3,2.$$

Kritická hodnota  $K = t_{10+10-2}(0,05) = 2,101$ .

Protože  $T > K$ , zamítá se  $H_0$  a přijme  $H_1$ .

Konstatujete tedy, že mezi výběrovými průměry je významný rozdíl, tedy průměr výnosů z hnojeného pole je průkazně vyšší než výnos nehnojeného. To znamená, že hnojení má průkazný vliv na výnosy.

\*

**SHRNUTÍ KAPITOLY ZÁKLADNÍ POJMY A METODY ZE STATISTIKY**



**Shrnutí kapitoly**

V této první kapitole jste si spíše jen zopakovali základní pojmy statistiky. Věnovali jste se souboru s jedním a se dvěma znaky. Poznali jste pojmy jako: základní statistický soubor, výběrový soubor, charakteristiky polohy, charakteristiky variability, šikmost, špičatost, kovariance, princip testování hypotéz, kritický obor, obor přijetí, testové kritérium, hladina významnosti a chyby, které mohou při testování vzniknout.

V poslední části kapitoly byl uveden obecný 4-krokový postup testování hypotéz pro 5 základních parametrických testů.

**ŘEŠENÝ PŘÍKLAD 1-6**

V následující tabulce jsou uvedeny údaje o 40 panelech plynových hořáků, na nichž byla nalezena alespoň jedna vada. Vypočítejte relativní a kumulativní relativní četnosti.

*Plynové  
hořáky*

Místo výskytu vady	Absolutní četnost
Plynový přívod	14
Povrchová úprava	10
Ovladače	8
Termoizolace	6
Elektrické zapalování	2

**Řešení příkladu**

Tabulku doplňte o požadované dva sloupce. Relativní četnosti vypočítáte jako podíl absolutních četností k celkovému počtu údajů, kumulativní četnosti pak postupně načítáte.

Místo výskytu vady	Absolutní četnost	Relativní četnost	Kumulativní četnost
Plynový přívod	14	0,35	0,35
Povrchová úprava	10	0,25	0,60
Ovladače	8	0,20	0,80
Termoizolace	6	0,15	0,95
Elektrické zapalování	2	0,05	1,00

Z výsledků kumulativní četnosti lze usoudit, že například tři nejčastěji se vyskytující vady (plynový přívod, povrchová úprava a závada některého ovladače) tvoří 80% všech nalezených vad.

\*

**ŘEŠENÝ PŘÍKLAD 1-7**

Máte k dispozici údaje o počtech výrobků vyrobených za směnu v souboru 15 dělníků: 8, 5, 9, 8, 6, 6, 10, 8, 7, 6, 9, 7, 8, 6, 8. Určete modus, medián a průměrný počet výrobků vyrobených za směnu.

*Výrobky*

**Řešení příkladu**

Nejčetnější hodnota – modus:  $\hat{x} = 8$ . Tato hodnota je v souboru vyskytuje 5krát.

Prostřední hodnota – medián:  $\tilde{x} = 8$ . Jestliže seřadíme údaje vzestupně, pak na 8. místě se nachází hodnota 8.

$$\text{Průměrný počet výrobků: } \bar{x} = \frac{111}{15} = 7,4.$$

\*

**ŘEŠENÝ PŘÍKLAD 1-8**

Tabulka zachycuje výsledky studentů v předmětech Statistika B a Statistické metody pro ekonomy.

*Výsledky studentů*

- (a) Vypočítejte průměrné výsledky v obou předmětech a porovnejte variabilitu výsledků.  
 (b) Jsou výsledky v některém předmětu statisticky významně lepší?

Hodnocení	Statistika B	Statistické metody pro ekonomy
Výborně (1)	24	58
Velmi dobře (2)	51	45
Dobře (3)	68	35
Neprospěl (4)	21	2

**Řešení příkladu**

- (a) Nejprve vypočtete průměr a rozptyl v souboru výsledků předmětu Statistika B:

$$\bar{x}_1 = \frac{1 \cdot 24 + 2 \cdot 51 + 3 \cdot 68 + 4 \cdot 21}{164} = \frac{414}{164} = 2,52,$$

$$s_1^2 = \frac{1}{164} \left[ 24(1 - 2,52)^2 + 51(2 - 2,52)^2 + 68(3 - 2,52)^2 + 21(4 - 2,52)^2 \right] = 0,79.$$

Analogicky pro hodnocení předmětu Statistické metody pro ekonomy dostanete následující výsledky:

$$\bar{x}_2 = \frac{1 \cdot 58 + 2 \cdot 45 + 3 \cdot 35 + 4 \cdot 2}{140} = \frac{261}{140} = 1,86,$$

$$s_2^2 = \frac{1}{140} \left[ 58(1 - 1,86)^2 + 45(2 - 1,86)^2 + 35(3 - 1,86)^2 + 2(4 - 1,86)^2 \right] = 0,71.$$

Lze tedy konstatovat, že studenti mají lepší průměrné výsledky v kurzu Statistické metody pro ekonomy. V tomto kurzu byla také zjištěna menší variabilita souboru výsledků. Jsou však tyto hodnoty průkazné, tj. statisticky významné? To zjistíte pomocí dvouvýběrového *t*-testu. Abyste mohli vybrat ten správný: (B) nebo (C), musíte předtím testovat, zda rozdíly rozptylů jsou statisticky významné, tedy použít k tomu test (E). Pro uvedené testy použijte obvyklou hladinu významnosti  $\alpha = 0,05$ .

- (b) Máme dva náhodné výběry z rozdělení  $N(\mu_1, \sigma_1^2)$  a  $N(\mu_2, \sigma_2^2)$  o rozsahu  $n_1$  a  $n_2$  a výběrové rozptyly  $s_1^2$  a  $s_2^2$ .

**1. Testujete hypotézu**

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ proti alternativní hypotéze } H_1: \sigma_1^2 \neq \sigma_2^2.$$

## 2. Testovací kritérium

$$T = \frac{s_1^2}{s_2^2} = \frac{0,79}{0,71} = 1,113.$$

## 3. Kritická hodnota je

$$K = F_{n_1-1, n_2-1}(\alpha) = F_{163, 139}(0,05) = 1,311.$$

Hodnotu naleznete např. v Excelu pomocí statistické funkce FINV.

4. Protože  $T < K$ , přijímá se  $H_0: \sigma_1^2 = \sigma_2^2$ . Proto k testování rovnosti středních hodnot použijeme testu (B).

## 1. Testujeme hypotézu

$H_0: \mu_1 = \mu_2$  proti alternativě  $H_1: \mu_1 \neq \mu_2$ .

2. Použijeme testovací kritérium  $T$ :

$$T = \frac{2,52 - 1,86}{\sqrt{163 \cdot 0,79 + 139 \cdot 0,71}} \cdot \sqrt{\frac{164 \cdot 140 \cdot 302}{304}} = 7,531,$$

kde  $\bar{x}_1 = 2,52$ ,  $\bar{x}_2 = 1,86$  jsou výběrové průměry,

$s_1^2 = 0,79$ ,  $s_2^2 = 0,71$  jsou výběrové rozptyly,

$n_1 = 164$ ,  $n_2 = 140$  jsou rozsahy 1. a 2. výběru.

3. Kritická hodnota  $K = t_{302}(0,05) = 1,97$ .

4. Protože je  $T > K$  zamítáme  $H_0$  a přijímáme  $H_1: \mu_1 \neq \mu_2$ . Konkrétně průměrné výsledky (známky) z předmětu Statistické metody pro ekonomy jsou statisticky významně lepší.

\*

**KONTROLNÍ TEST 1**

1.1 Variační rozpětí  $R$  je dáno vztahem  $R = x_{\min} - x_{\max}$ .

1.2 Statistický soubor lze popsat pomocí grafických a výpočetních metod.

1.3 Z hodnoty kovariance lze zjistit velikost závislosti  $y$  na  $x$ .

1.4 Hledáním tvaru závislosti se zabývá regresní analýza.

1.5 Alternativní hypotéza je vždy negací nulové hypotézy.

1.6 Mezi charakteristiky variability patří: \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_.

1.7 Máme-li statistický soubor se dvěma argumenty, pak součty četností v jednotlivých řádcích či v jednotlivých sloupcích nazýváme \_\_\_\_\_.

1.8 Pokud kovariance  $c_{xy} > 0$ , jedná se o \_\_\_\_\_ závislost.

1.9 Pod pojmem závislost rozumíme: \_\_\_\_\_ a \_\_\_\_\_ závislosti.

1.10 Při testování hypotéz se testové kritérium porovnává s \_\_\_\_\_ hodnotou.

Ano či ne?

Doplňte



Řešte

- 1.11** Směnárna v jisté bance je vybavena dvěma přepážkami pro odbavování klientů. První z nich každou hodinu smění průměrně cizí měny v protihodnotě 81 328 Kč, druhá 96 018 Kč. Celkový hodinový průměr za obě pracoviště je 87 222 Kč. Jak se která přepážka podílí v % na směnárenské činnosti?
- 1.12** Uvažujte soubor důchodců, jejichž průměrný měsíční důchod je 4 800 Kč a směrodatná odchylka měsíčních důchodů je 500 Kč. Při které úpravě důchodu se nezmění směrodatná odchylka a proč?
- každý důchodce dostane přidáno 150 Kč,
  - důchod každého důchodce se zvýší o 5%
- 1.13** V tabulce jsou údaje o týdenních mzdách ve dvou dílnách. Porovnejte variabilitu v obou dílnách výpočtem rozptylu.

Dílna 1	Dílna 2
2400	2613
2134	2496
2407	2736
2445	2676
2984	3093
3354	3537
3515	3622
3515	3561
3225	3385
3063	3155
2694	2788
2600	2838

- 1.14** Podnik vyrábějící bílé zboží si provedl průzkum výdajů domácností za tento druh průmyslového zboží u dvou typů domácností (se dvěma dětmi a se čtyřmi dětmi). Cílem tohoto průzkumu bylo zjistit, zda domácnosti s více dětmi skutečně nakupují vícekapacitní, a tedy dražší přístroje. U 12 náhodně vybraných rodin s dvěma dětmi byly zjištěny následující hodnoty ročních výdajů za bílé zboží ( v tis. Kč):
- 41,2 39,4 36,3 38,7 39,9 38,3  
40,6 41,5 37,4 43,1 35,7 35,8.
- Podobně u 6 náhodně vybraných rodin se čtyřmi dětmi byly tyto údaje následující:
- 39,2 43,8 38,9 44,3 41,2 44,1.
- Formulujte příslušnou nulovou hypotézu a testujte ji dvouvýběrovým vhodným testem.

## ŘEŠENÍ KONTROLNÍHO TESTU 1



Ano či ne?

- 1.1** ne  
**1.2** ne  
**1.3** ne  
**1.4** ano  
**1.5** ne

- 1.6** rozptyl, směrodatná odchylka, variační rozpětí, variační koeficient  
**1.7** marginální četnosti  
**1.8** přímou  
**1.9** tvar, míru  
**1.10** kritickou

**Doplňte**

- 1.11** podíly  $p_1 = 0,599 = 59,9\%$ ,  $p_2 = 0,401 = 40,1\%$   
**1.12** v případě a); průměr se zvýší také o 150,- Kč, odchylka hodnoty důchodu od průměru zůstane proto stejná, tedy i součet kvadrátů odchylek.  
**1.13**  $s_1^2 = 207375,708$ ,  $s_2^2 = 149971,694$   
**1.14** Hodnota testového kritéria  $T = 0,247$  leží v oboru přijetí, tudíž se neprokázal na hladině významnosti 5% statisticky významný rozdíl mezi rozptyly.

**Řešte**

## 2 TESTOVÁNÍ HYPOTÉZ V MARKETINGU

### CÍLE KAPITOLY TESTOVÁNÍ HYPOTÉZ V MARKETINGU

#### Po úspěšném a aktivním absolvování této KAPITOLY

##### Budete umět:

- co přináší testování statistických hypotéz
- rozlišovat jednostranné a oboustranné testy, parametrické a neparametrické testy
- Mediánový test
- Chi-kvadrát test pro jeden výběr
- Chi-kvadrát test pro dva výběry

*Budete  
umět*

##### Budete schopni:

- aplikovat výše uvedené testy v oblasti marketingu

*Budete  
schopni*

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je 6 hodin.

#### PRŮVODCE STUDIEM 3



Testováním statistických hypotéz jste se částečně zabývali již v předchozí kapitole a také jste se již s ním setkali v základním kurzu statistiky. Protože jde o jednu z nejčastěji používaných statistických metod jak v oblasti marketingového výzkumu, tak i v jiných aplikačních oblastech ekonomie, budeme se jím zabývat podrobněji ještě v této kapitole a také kapitole následující. Půjde nám především o aplikaci těchto metod v marketingu, nebo jak se vznešeněji říká - v marketingovém výzkumu. Jak již bylo řečeno, proces testování zahrnuje většinou vyšetřování předpokladů o hodnotách parametrů určitého rozdělení. Statistické hypotézy tvoří jen část vědeckých (nebo alespoň odborných) hypotéz. Týkají se náhodných veličin a rozdělujeme je do dvou velkých tříd na parametrické hypotézy a neparametrické hypotézy.

Parametrické hypotézy se vztahují na jeden nebo několik parametrů daného pravděpodobnostního rozdělení náhodné veličiny (neboli znaku populace). Neparametrické hypotézy se netýkají parametrů rozdělení náhodné veličiny, nýbrž jiných statistických vlastností, např. tvaru rozdělení (například binomické rozdělení nebo normální rozdělení).

Při testování parametrické hypotézy máme k dispozici výsledky náhodného výběru – vzorku – data z marketingového výzkumu a na jejich základě rozhodujeme testovanou hypotézu buď přijmout, nebo zamítnout.

Na rozdíl od parametrických testů, které vyžadují intervalová data, neparametrické testy vystačí s ordinálními nebo s nominálními daty. Typickým příkladem jsou dvourozměrná data uspořádaná do kontingenční tabulky. Představte si pro příklad jednoduchou kontingenční tabulku se dvěma znaky: Pohlaví (s hodnotami Muž a Žena) a Kuřáctví (s hodnotami Kouří, Nekouří). Neparametrickým testem se testuje hypotéza, že tyto dva znaky spolu souvisejí, jinak řečeno, že muži kouří v průměru (statisticky) významně více než ženy, nebo je to obráceně?

Testem statistické hypotézy nemůžeme dokázat její platnost nebo neplatnost, správnost nebo nesprávnost! Testování hypotéz nemusí vždy vést ke správným rozhodnutím, což je přirozené, neboť jde o náhodný proces využívající pouze informace náhodného výběru.

## 2.1 Marketingová případová studie

Pro výukové účely demonstrace statistických metod se v celém následujícím kurzu budete čas od času zabývat případovou marketingovou studií, kterou budeme zkráceně označovat jako [Studie](#):

*Jistá firma (nebudeme ji zde jmenovat) chce na trhu s nealkoholickými nápoji uplatnit svůj nový výrobek: syčený nealkoholický nápoj - limonáda typu "cola". Nový výrobek se uvádí na trh ve dvou verzích názvu: Kafola a Kofíkola (Ka/Ko). Obě limonády se liší pouze názvem a designem obalu, jinak mají totožný obsah.*

*Na jisté fakultě (nebudeme ji rovněž raději jmenovat) je bufet a několik prodejních automatů s nealko nápoji, kde se nový nápoj pokusně již nějakou dobu prodává. Fakulta má přibližně 4000 studentů a 400 zaměstnanců. Byl*

*vytvořen dotazník, na jehož základě byla získána data, tj. vzorek 200 respondentů - studentů a zaměstnanců, o spotřebě a oblibě nealkoholických nápojů včetně nového nápoje, viz [Dotazník](#).*

**Marketingová studie**

### Účel studie:

- prezentovat typickou studii marketingového výzkumu v malém měřítku,
- demonstrovat všechny důležité statistické metody tak, byla zřejmá jejich užitečnost pro marketing,
- porozumět podstatě těchto metod,
- používat tyto metody pomocí běžných programů – Excel a SPSS.

**Účel studie**

## 2.2 Co přináší testování statistických hypotéz v marketingu?

V případě intervalových dat, tj. číselných dat, která mohou nabývat libovolných hodnot z nějakého číselného intervalu hodnot, odpovídá testování statistických hypotéz v marketingovém výzkumu na specifické otázky:

- Existuje významný soulad dané charakteristiky (parametru) – například průměru vzorku se zadanou hodnotou?
- Existuje významný rozdíl dané charakteristiky (parametru) mezi 2 (nebo více) vzorky?

**Otázky k řešení pomocí testů hypotéz**

Na rozdíl od první kapitoly, kde jste si zopakovali oboustranné testy hypotéz, se v této kapitole budete zabývat i jednostrannými testy, kdy hodnota parametru má být **větší** (eventuálně **menší**) než předem daná hodnota. Připomeňte si, že u oboustranných testů měla být hodnota parametru **různá** od předem zadané hodnoty.

**Parametr** - v oblasti marketingu to bývá obvykle průměr  $\mu$ , směrodatná odchylka  $\sigma$ , poměr  $P$ .

**Parametr**

V případě ordinálních nebo nominálních dat odpovídá testování neparametrických hypotéz na tyto otázky:

1. Existuje významný soulad dané charakteristiky vzorku se zadanou charakteristikou nebo hodnotou?
2. Existuje významný rozdíl dané charakteristiky mezi 2 (nebo více) vzorky?

**Otázky, na které odpovídá neparametrický test**

Charakteristikou se míní např. medián, zadané pořadí, rozdělení pravděpodobnosti (četnosti) apod.

Neparametrické testy hypotéz dělíme na:

### 1. Jednovýběrové testy:

- Má medián populace s neznámým rozdělením stanovenou hodnotu? (na tuto otázku odpoví mediánový test).

- Pochází výběr z populace se zadaným (známým) rozdělením pravděpodobnosti? (na tuto otázku odpoví Chi-kvadrát test).

**Jednovýběrový test**

### 2. Dvouvýběrové testy:

- Mají dva výběry stejný medián? (Odpoví mediánový test).

- Pochází dva výběry ze stejné populace? (Lze použít Chi kvadrát test).

**Dvouvýběrové testy**

## ŘEŠENÝ PŘÍKLAD 2-1



Je průměrný počet vypitých limonád týdně menší než sedm, tedy  $\mu_0 = 7$ ? Zde se jedná o soulad průměru se zadanou hodnotou.

**Příklad ze Studie**

### Řešení příkladu

#### Obecný postup (tzv. Jednostranný jednovýběrový t-test):

Formulujte nulovou hypotézu  $H_0$ , tj.  $H_0: \mu < \mu_0$  a k ní alternativní hypotézu  $H_1: \mu \geq \mu_0$ . Použijeme test (A) z kapitoly 1, který modifikujeme pro případ jednostranné hypotézy.

#### 1. Testujete hypotézu

$$H_0: \mu < \mu_0 \text{ proti alternativě } H_1: \mu \geq \mu_0$$

#### 2. Použijete testovací kritérium

$$T = \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n}$$

kde  $\bar{X}$  = výběrový průměr,

$S$  = výběrová směrodatná odchylka,

$\mu_0 = 7$  předpoklad o střední hodnotě,

$n = 200$  rozsah výběru.

3. Kritická hodnota  $K = t_{n-1}(\alpha)$  se určí z tabulek Studentova rozdělení pro  $n-1$  stupně volnosti a hladinu významnosti  $\alpha$ . Pro jednostranný test se použije  $\alpha$ , nikoliv  $\alpha/2$ , jako v kap.1, kde se pravděpodobnost  $\alpha$  „rozdělila“ stejně na obě strany. Obor přijetí  $A$  je zde interval  $A = (-\infty, K]$  a kritický obor  $C$  je jeho doplňkem,  $C = (K, +\infty)$ . Pravděpodobnost  $\alpha$  „zůstává“ na jedné straně.

K výpočtu kritické hodnoty pomocí Excelu použijte funkci  $\text{TINV}(\alpha; n-1)$  v menu: **Vložit** → **Funkce** → **Statistické** → **TINV...**

4. Je-li  $T > K$ , potom zamítáme  $H_0$  a přijímáme  $H_1$ , jinak přijímáme  $H_0$ .

Protože jste měli k dispozici pouze ne příliš velký vzorek ( $n = 200$ ), může být přesto hypotéza správná. Uvědomte si, že svůj soud opíráte o vzorek 200 respondentů a populace má 4400 potenciálních respondentů – ti se mohou (teoreticky) chovat jinak než respondenti ve vzorku! Avšak pravděpodobnost, že jste zamítli správnou hypotézu, je  $\alpha$ , tj. rovná se zvolené hladině významnosti, tj. 5%!

**Jednovýběrový t-test  
Nulová hypotéza  
Hladina významnosti**

(1-1)

**Číselné řešení příkladu:**

$H_0: \mu < 7$ ,  $H_1: \mu \geq 7$ ,  $\alpha = 0,05$ ,  $n = 200$ ,  $\bar{X} = 8,15$ ,  $S = 11,39$

Hodnota testového kritéria je

$$T = \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} = \frac{8,15 - 7}{11,39} \sqrt{200} = 1,428$$

Kritická hodnota je  $K = t_{199}(0,05) = 1,972$ .

Při použití Excelu tuto hodnotu získáte s použitím funkce

TINV(0,05;199).

Protože  $T < K$ , nulovou hypotézu  $H_0$  nezamítáme, jinak řečeno, na hladině významnosti 5% ( $\alpha = 0,05$ ) přijímáme hypotézu, že průměrný počet vypitých limonád za týden je ve vyšetřované populaci menší než sedm. Jak jste se již dříve dozvěděli, tato hypotéza nemusí být správná, avšak pravděpodobnost, že jsme přijali nesprávnou hypotézu je rovna hladině významnosti, tj. je pouze 5%.

Přirozeně vzniká otázka, na jaké hladině významnosti je možné naši nulovou hypotézu zamítnout. Tato hladina (a rovněž tak pravděpodobnost omylu) bude přirozeně vyšší než 5%. Konkrétní hodnota takové hladiny  $K_z$  je dána vztahem  $T > K_z$ . Protože hodnota testovacího kritéria je  $T = 1,428$ , musí platit

$$K_z = t_{199}(\alpha_z) = 1,428,$$

kde  $\alpha_z$  je hledaná hladina významnosti. Tuto hodnotu naleznete buď v tabulkách Studentova rozdělení  $t$ , nebo pomocí Excelu použitím funkce

$$\text{TDIST}(1,428; 199; 1) = 0,077.$$

Hledaná hladina významnosti  $\alpha_z = 0,077$ , tj. 7,7%. Tato hladina se nazývá *p-hodnota testu* nebo také *signifikance*. Na této hladině významnosti lze tedy  $H_0$  zamítnout, avšak s pravděpodobností omylu 7,7%.

\*

## 2.3 Jednovýběrový t-test: řešení příkladu 2-1 pomocí MS Excelu

### Řešení příkladu 2-1 ze Studie pomocí MS Excel

#### *V Excelu lze řešit Příklad 2-1 třemi způsoby:*

S prvním způsobem jste se vlastně již seznámili v předchozím odstavci: Podle vzorce (1-1) jste vypočítali hodnotu testového kritéria  $T$  a taktéž s pomocí excelovské funkce TINV jste zjistili kritickou hodnotu a pomocí funkce TDIST jste stanovili  $p$ -hodnotu (signifikanci).

Druhý způsob spočívá ve využití Analytických nástrojů, které jsou standardní součástí Excelu v menu **Nástroje** → **Analýza dat...** Musíme upozornit, že po čerstvé instalaci Excelu se v nabídce Nástroje položka Analýza dat neobjeví, je ji zapotřebí ještě doinstalovat. Tato instalace je však velmi snadná, v menu Nástroje je položka Doplňky, kde se zaklikne položka Analytické nástroje a volba se potvrdí. V okamžiku se doinstaluje potřebná položka Analýza dat, která obsahuje 19 statistických metod, v tom i 5 testů statistických hypotéz. Hlavní výhoda použití Analýzy dat oproti prvnímu způsobu spočívá v tom, že není zapotřebí vytvářet příslušné vzorce a provádět dílčí výpočty. Každý test je integrován do jednoho okna a výsledek je prezentován v unifikované tabulce.

Jednovýběrový t-test se převede na dvouvýběrový, druhý výběr se vytvoří uměle z kon-

Číselné  
řešení  
příkladu



Řešení v  
Excelu

stantních hodnot  $\mu_0 = 7$ . Přípravu dat i vyplnění okna příslušného testu naleznete v souboru [Studie](#) na listu t-test\_1. Níže je ještě jednou uveden výsledek - formou tabulky, kterou poskytuje Excel.

Dvouvýběrový t-test s nerovností rozptylů

	Počet lahví/týden	pom
Stř. hodnota	8,15	7
Rozptyl	129,7763819	0
Pozorování	200	200
Hyp. rozdíl stř. hodnot	0	
Rozdíl	199	
t stat	1,427627836	
P(T<=t) (1)	0,077483183	
t krit (1)	1,652547326	
P(T<=t) (2)	0,154966365	
t krit (2)	1,971957317	

První 4 řádky není zapotřebí komentovat. V dalších řádcích znamená:

- Rozdíl = počet st. volnosti,  
t stat = hodnota testové statistiky  $T$ ,  
P(T<=t) (1) =  $p$ -hodnota testu, jednostranná varianta,  
t krit (1) = kritická hodnota, jednostranná varianta,  
P(T<=t) (2) =  $p$ -hodnota testu, dvoustranná varianta,  
t krit (2) = kritická hodnota, dvoustranná varianta.

Třetí způsob využití Excelu spočívá v použití funkce TTEST z menu:

**Vložit → Funkce → Statistické → TTEST...**

Poté je zapotřebí vyplnit zadávací okno. Podobně jako u použití Analytických nástrojů je zapotřebí připravit druhý umělý výběrový soubor z konstantních hodnot  $\mu_0 = 7$ , neboť funkce TTEST umožňuje provádět pouze dvouvýběrové testy. V položce **Strany** zadejte hodnotu 1 (jednostranný test), v položce **Typ** zadejte hodnotu 3 (Dvouvýběrový test s nerovností rozptylů). Výsledkem je jediné číslo –  $p$ -hodnota testu, v našem konkrétním případě je  $p$ -hodnota testu = 0,077. Toto je nejnižší hladina významnosti, na které se  $H_0$  zamítá. Na nižší hladině významnosti než 0,077 se  $H_0$  nezamítá (např. na hladině 0,05), naopak na vyšší hladině významnosti než 0,077 se  $H_0$  zamítá.

\*

## 2.4 Dvouvýběrový t-test – nepárový a párový: řešení příkladů ze Studie

Nejprve budete řešit Příklad 2-2. Z datového souboru "Počet lahví/týden" vyrobíte 2 samostatné výběry: pro muže (M) a pro ženy (Z) s rozsahy  $n_1, n_2$ .

U obou výběrů nejprve vypočítáte

- výběrové průměry:  $\bar{X}_M, \bar{X}_Z$
- výběr. směrodatné odchylky:  $S_M, S_Z$

Nejprve pomocí F – testu (test (E) z kapitoly 1) testujete, zda  $\sigma_1^2 = \sigma_2^2$ , abyste pak mohli vybrat

správnou variantu dvouvýběrového  $t$  – testu.  
Pak použijete buď test (B), nebo test (C) z kapitoly 1.

## ŘEŠENÝ PŘÍKLAD 2-2



Je průměrný počet limonád týdně u mužů stejný, jako je u žen? Zde jde o rozdíl dané charakteristiky (průměru) mezi 2 vzorky: muži a ženami.

**Příklad ze  
Studie**

### Řešení příkladu

#### Řešení příkladu v Excelu:

Podobně jako při řešení příkladu 2-1, můžete pro řešení v Excelu použít různé způsoby. Zde konkrétně zvolíte druhý způsob k testu rovnosti rozptylů a třetí způsob, totiž využití funkce TTEST k testu rovnosti středních hodnot.

V menu postupně zvolíte:

**Nástroje → Analýza dat → Dvouvýběrový F-test pro rozptyl ...**

Poté vyplníte zadávací okno: Vložíte umístění obou připravených souborů M a Z a potvrdíte OK. Obdržíte následující výstupní tabulku:

**Řešení v  
Excelu**

Dvouvýběrový F-test pro rozptyl

	<i>Soubor 1</i>	<i>Soubor 2</i>
Stř. hodnota	9,866667	5,575
Rozptyl	115,6123	141,5639
Pozorování	120	80
Rozdíl	119	79
F	0,816679	
P(F<=f) (1)	0,157707	
F krit (1)	0,717051	

První 3 řádky není zapotřebí komentovat. V dalších řádcích znamená:

Rozdíl = počet st. volnosti,

F = hodnota testové statistiky  $T$ ,

P(F<=f) (1) =  $p$ -hodnota testu,

F krit (1) = kritická hodnota.

Protože  $p$ -hodnota testu = 0,158 a to je hodnota větší než 0,05 ,  $H_0$  se nezamítá, neboli přijímá se hypotéza o rovnosti rozptylů:  $\sigma_1^2 = \sigma_2^2$ .

Dále v hlavním menu Excelu postupně zvolíte:

**Vložit → Funkce → Statistické → TTEST...**

Poté vyplníte zadávací okno: Vložíte umístění obou připravených souborů M a Z, v položce **Strany** zadejte hodnotu 2 (dvoustranný test), v položce **Typ** zadejte hodnotu 2 (Dvouvýběrový test s rovností rozptylů). Výsledkem je jediné číslo –  $p$ -hodnota testu, v našem konkrétním případě je  $p$ -hodnota testu = 0,0087. Toto je nejnižší hladina významnosti, na které se  $H_0$  zamítá. Na nižší hladině významnosti než 0,0087 se  $H_0$  nezamítá, naopak na vyšší hladině významnosti než 0,0087 , a tedy také na obvyklé hladině 0,05 , se  $H_0$  zamítá.



Konkrétně to znamená, že týdenní konzumace sycených nápojů se u mužů statisticky významně liší od žen, jmenovitě u mužů je konzumace významně vyšší než u žen, viz řádek Střední hodnota ve výše uvedené tabulce.

\*

## ŘEŠENÝ PŘÍKLAD 2-3



viz [Studie](#) – Dvouběžrový párový t-test

- Liší se významně průměrné hodnocení Ceny a Image?
- Je průměrné hodnocení Důležitosti image značky lepší než hodnocení Důležitosti ceny? (jednostranný test)

Dvouběžrový párový t-test

### Řešení příkladu

**Řešení pomocí výpočtů vzorců v Excelu – 1. způsob:**

Řešení příkladu (a)

#### Řešení příkladu (a)

Jedná se o dvoustranný test hypotézy o střední hodnotě. Přitom při hodnocení používáme škálu 1, 2 až 7 kterou pro účely tohoto testu považujeme za intervalovou škálu. Tento předpoklad neodpovídá skutečnosti, že Studie je zřejmé, že se jedná o ordinální hodnotící škálu, pro níž bychom však nemohli použít t-test. Abychom t-test použít mohli, a v praxi se to tak dělá, dopouštíme se této nepřesnosti. Vhodné testy pro ordinální data uvedeme v následující kapitole.

Prvky ve dvou výběrech jsou „spárovány“, tj. jednoznačně spolu souvisejí po dvou.

K řešení použijeme test (D) z kapitoly 1:

#### 1. Testujeme hypotézu

$$H_0: \mu_1 = \mu_2 \text{ proti alternativě } H_1: \mu_1 \neq \mu_2.$$

#### 2. Použije se testovací kritérium:

$$T = \frac{\bar{D}}{S_D} \sqrt{n} = \frac{0,190}{3,669} \sqrt{200} = 0,732,$$

#### 3. Kritická hodnota $K = t_{n-1}(\alpha/2) = \text{TINV}(0,025;199) = 2,258$ .

4. Protože je  $T < K$ ,  $H_0$  se přijímá, jinými slovy, průměrné hodnocení důležitosti image značky a důležitosti ceny se neliší.

#### Řešení příkladu (b): Jednostranný test

Analogický postup jako u oboustranného testu.

Řešení příkladu (b)

#### 1. Testujeme hypotézu

$$H_0: \mu_1 > \mu_2 \text{ proti alternativě } H_1: \mu_1 \leq \mu_2.$$

#### 2. Použije se testovací kritérium:

$$T = \frac{\bar{D}}{S_D} \sqrt{n} = \frac{0,190}{3,669} \sqrt{200} = 0,732,$$

#### 3. Kritická hodnota $K = t_{n-1}(\alpha) = \text{TINV}(0,05;199) = 1,960$ .

4. Protože je  $T < K$ ,  $H_0$  se opět přijímá. I když průměrné hodnocení Důležitosti image značky = 4,18 je větší než průměrné hodnocení Důležitosti ceny = 3,99 , není tento rozdíl statisticky významný. Přijímáme hypotézu, že se obě hodnocení vzájemně neliší.

**Řešení pomocí funkce Analýza dat v Excelu – 2. způsob**  
**Nástroje → Analýza dat → Dvouvýběrový párový t-test...**

**Řešení pomocí Excelu**

	<i>Důležitost ceny</i>	<i>Důležitost image</i>
S tř. hodnota	4,18	3,99
Rozptyl	5,173467337	4,090351759
Pozorování	200	200
Pears. korelace	-0,456223683	
Hyp. rozdíl s tř. hodnot	0	
Rozdíl	199	
t stat	0,732363368	
P(T<=t) (1)	0,232404248	
t krit (1)	1,652547326	
P(T<=t) (2)	0,464808495	
t krit (2)	1,971957317	

Výsledky řešení jsou patrné z tabulky a jsou analogické výsledkům předchozích výpočtů – 1. způsob.

\*

### SAMOSTATNÝ ÚKOL 1

Testujte hypotézu, že průměrný počet vypitých limonád **denně** je alespoň dvě, tedy  $\mu_0 = 2$ .



**Jednovýběrový t-test**

### SAMOSTATNÝ ÚKOL 2

Testujte hypotézu, že průměrný počet vypitých limonád denně u žen je alespoň takový, jako u mužů.



**Dvouvýběrový t-test**

### SAMOSTATNÝ ÚKOL 3

Liší se statisticky významně u mužů (!) průměrné hodnocení Ceny a Image?



**Párový t-test**

## 2.5 Mediánový t-test (pro 1 výběr)

Mediánový test lze použít nejen v případě intervalových dat, ale i v případě ordinálních dat. Pokud jsou data alespoň intervalová a k tomu normálně rozdělená, je výhodné použít parametrický t-test. Nevíte-li, zda má populace normální rozdělení, resp. víte, že rozdělení není normální, použijete mediánový test. Medián se označí  $\tilde{\mu}$ , rozsah vzorku  $n$ .

**Mediánový test**

1. Testujete oboustrannou nulovou hypotézu  $H_0: \tilde{\mu} = \tilde{\mu}_0$  proti alternativní hypotéze  $H_1: \tilde{\mu} \neq \tilde{\mu}_0$ . Jedná se tedy o dvoustranný test.

*Postup testu*

2. Za testové kritérium položíme:  $T = \frac{|2m - n|}{\sqrt{n}}$ , kde  $m$  je počet pozorování ve vzorku menších než  $\tilde{\mu}_0$ .

3. Kritická hodnota je  $K = z_{1-\alpha/2}$ .

Přitom  $z_{1-\alpha/2}$  je kritická hodnota normovaného normálního rozdělení pro zadanou hladinu významnosti  $\alpha$ . Tuto hodnotu naleznete buď v tabulkách normovaného normálního rozdělení  $N(0,1)$ , nebo pomocí Excelu použitím funkce  $\text{NORMSINV}(1-\alpha/2)$ .

4. Jestliže platí  $T > K$ , potom se  $H_0$  zamítá, jinak se  $H_0$  přijímá!

## ŘEŠENÝ PŘÍKLAD 2-4



Ze Studie testujte hypotézu, že průměrný (mediánový) věk zákazníků, tj. studentů a zaměstnanců fakulty, je 21 let.

*Test průměrného mediánového věku*

### Řešení příkladu

Budete analyzovat ve Studii znak Věk. Je zřejmé, že věk zaměstnanců a studentů fakulty nemá normální rozdělení. Proto použijete mediánový test.

1. Budete testovat nulovou hypotézu  $H_0: \tilde{\mu} = 21$  proti alternativní hypotéze  $H_1: \tilde{\mu} \neq 21$ .

Zvolíte obvyklou hladinu významnosti  $\alpha = 0,05$ .

2. Snadno po uspořádání respondentů podle věku zjistíte, že  $m = 96$ . Potom hodnota testového kritéria je  $T = \frac{|192 - 200|}{\sqrt{200}} = 0,566$ .

3. Kritická hodnota  $K = z_{0,975} = \text{NORMSINV}(0,975) = 1,96$ .

4. Platí  $T < K$  a proto se  $H_0$  přijímá. Pro populační soubor studentů a zaměstnanců fakulty (počet 4000) přijímáme na základě vzorku (počet 200 respondentů) hypotézu, že mediánový věk je 21 let.

\*

## 2.6 Chi-kvadrát test pro 1 výběr

Data mohou být nominální (nejslabší možný požadavek)!

Testuje se (nulová) hypotéza: výběr pochází z populace se zadaným rozdělením.

**Zadané rozdělení** je obvykle:

- diskrétní rozdělení se **stejnými** pravděpodobnostmi, pak Chi-kvadrát test nazývá **test nezávislosti**,
- diskrétní rozdělení s **rozdílnými** pravděpodobnostmi, pak se Chi-kvadrát test nazývá **test dobré shody**.

Následující dva příklady jsou představiteli obou typů testu Chi-kvadrát.

*Nominální data*

*Chi-kvadrát test*

**ŘEŠENÝ PŘÍKLAD 2-5**

*Styl designu obalu versus počet prodaných limonád*

**Příklad 1**

Nová limonáda Kofela ze Studie se prodávala za stejnou cenu jeden týden ve fakultním bufetu ve 3 různých formách designu obalu: A, B, C, počet prodaných limonád je uveden v následující tabulce:

Typ obalu	Prodané kusy
A	135
B	130
C	155
Celkem	420

Ovlivňuje styl designu obalu počet prodaných limonád? Testujte hypotézu, že styl obalu nemá na prodej vliv, jinak řečeno, že styl obalu je na prodeji nezávislý.

**Příklad 2**

V bufetu fakulty se prodávají sycené nápoje Coke, Fanta, Sprite. Na základě dlouhodobých zkušeností je známo, že se prodávají v poměru 60:30:10. Ve sledovaném týdnu se tyto nápoje prodaly v množství, které je uvedeno v následující tabulce.

Nápoj	Prodaných kusů
Coke	378
Fanta	149
Sprite	56
Celkem	583

Prodávají se uvedené nápoje v očekávaných poměrech? Testujte hypotézu, že prodaná množství odpovídají předpokládaným poměrům.

**Řešení příkladu**

K Chi-kvadrát testu pro oba typy testů použijeme opět známé 4 kroky:

1. Nulová hypotéza:

$H_0$ : Pozorované četnosti se významně neliší od očekávaných četností.

Přítom

Očekávané četnosti se označí:  $E_1, E_2, \dots, E_k$

Pozorované četnosti se označí:  $O_1, O_2, \dots, O_k$ .

K označení používáme  $E$  z angl. *Expected* (očekávaný),  $O$  z angl. *Observed* (pozorovaný).

Počet testovaných tříd (kategorií) označujeme  $k$ . V našich příkladech 2 a 3 je vždy  $k = 3$ .

2. Testové kritérium (nazývá se *Chi-kvadrát kritérium*):

$$T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

*Postup testování - algoritmus*

kde  $k$  - počet kategorií (např.  $k = 3$ ).

3. Kritická hodnota je  $K = \chi_{\alpha}(k - 1)$ .

Přitom  $\chi_{\alpha}(k - 1)$  je kritická hodnota rozdělení Chi-kvadrát (též Pearsonova rozdělení) s  $df = (k-1)$  stupni volnosti pro zadanou hladinu významnosti  $\alpha$ . Tuto hodnotu naleznete buď v tabulkách rozdělení Chi-kvadrát, nebo pomocí Excelu použitím funkce CHIINV( $\alpha$ ;  $df$ ).

**Poznámka:** Označení  $df$  je z angl. *degrese of freedom* (stupeň volnosti).

4. Jestliže platí  $T > K$ , potom se  $H_0$  zamítá, jinak se  $H_0$  přijímá.

**Podmínka:** K dosažení dostatečné přesnosti testu by v každé kategorii mělo být alespoň 5 hodnot!

**Alternativně:** Pro hodnotu testového  $T$  zjistíte  $p$ -hodnotu (signifikaci), tuto hodnotu porovnáte se zvolenou hladinou významnosti ( $\alpha = 0,05$ ), pokud  $p$ -hodnota  $> \alpha$ , pak  $H_0$  nezamítáte, v opačném případě  $H_0$  zamítáte. V Excelu můžete  $p$ -hodnotu vypočítat pomocí funkce CHIDIST.

**Řešení příkladu 1 pomocí Excelu:**

Ve worksheetu Excelu vyrobíte následující tabulku, použijete přitom vestavěné funkce SUMA, CHIINV, CHIDIST. Pokud prodej nápoje nezávisí na druhu obalu, pak všechny tři očekávané četnosti jsou stejné, tj.  $E_i = 420/3$ .

**Tabulka→Funkce: SUMA, CHIINV, CHIDIST...**

Typ obalu	O <sub>i</sub>	E <sub>i</sub>	(O <sub>i</sub> - E <sub>i</sub> ) <sup>2</sup> /E <sub>i</sub>
A	135	140	0,179
B	130	140	0,714
C	155	140	1,607
Sumy	420	420	2,500

T	=	2,5
alfa	=	0,05
k-1	=	2
CHIINV	=	5,991
CHIDIST	=	0,287

$p$ -hodnota = CHIDIST(2,5 ; 2) = 0,287 > 0,05 potom  $H_0$  nezamítáte. Interpretace výsledku: počet prodaných kusů nezávisí na typu obalu, rozdíly v prodeji u vzorku jsou pouze dílem náhody.

**Řešení příkladu 2 pomocí Excelu:**

Ve worksheetu Excelu vyrobíte následující tabulku, použijete přitom vestavěné funkce SUMA, CHIINV, CHIDIST. Tři očekávané četnosti prodeje jednotlivých druhů limonád vypočítáte z celkového počtu 583 prodaných kusů a známého poměru prodaných kusů 60:30:10 takto:

$$E_1 = 583 \cdot 0,60 = 349,8$$

$$E_2 = 583 \cdot 0,30 = 174,9$$

$$E_3 = 583 \cdot 0,10 = 58,3$$

**Tabulka→Funkce: SUMA, CHIINV, CHIDIST...**



Řešení  
v Excelu



Řešení  
v Excelu

Nápoj	O <sub>i</sub>	E <sub>i</sub>	(O <sub>i</sub> - E <sub>i</sub> ) <sup>2</sup> /E <sub>i</sub>
Coke	378	349,8	2,273
Fanta	149	174,9	3,835
Sprite	56	58,3	0,091
Celkem	583	583	6,200

T	=	6,2
alfa	=	0,05
k-1	=	2
CHIINV	=	5,991
CHIDIST	=	0,045

$p$ -hodnota = CHIDIST(6,2 ; 2) = 0,045 < 0,05 , proto  $H_0$  zamítáte. Interpretace výsledku: Zamítáte hypotézu, že ve sledovaném týdnu poměr prodeje 3 limonád odpovídal dlouhodobě známým poměrům. Přijímáte alternativní hypotézu, že ve sledovaném týdnu poměr prodaných limonád neodpovídal poměrům z minulosti.

\*

## 2.7 Dvouvýběrové testy

Často se v marketingovém výzkumu setkáte s těmito otázkami:

1. Existuje významný rozdíl dané charakteristiky mezi 2 vzorky?

Charakteristikou může být např. medián, zadané pořadí, typ rozdělení pravděpodobnosti (četnosti) apod.

2. Pocházejí 2 výběry ze stejné populace?

Na obě otázky svým způsobem odpovídají neparametrické testy hypotéz - tzv. *dvouvýběrové testy*. Budete se zabývat nejpoužívanějším z nich: Chi-kvadrát test.

*Dvouvýběrové testy*

## 2.8 Chi-kvadrát test pro 2 výběry

Samostatným a novým problémem je v tomto testu vytvoření speciální *kontingenční tabulky*. Uvažují se dva znaky: znak  $A$  (např. Pohlaví) a znak  $B$  (např. Způsob odměny). Znak  $A$  je *třídící znak*, který má 2 kategorie značené  $A_1$  a  $A_2$ , *sledovaný znak*  $B$  má  $s$  možných kategorií hodnot  $B_1, B_2, \dots, B_s$ . Kontingenční tabulka má tento tvar:

*Kontingenční tabulka*

Kategorie znaku $A / B$	$B_1$	$B_2$	$B_3$	...	$B_s$	Součet
$A_1$	$n_{11}$	$n_{12}$	$n_{13}$	...	$n_{1s}$	$n_{1.}$
$A_2$	$n_{21}$	$n_{22}$	$n_{23}$	...	$n_{2s}$	$n_{2.}$
Součet	$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.s}$	$n$

V tabulce značí  $n_{ij}$  četnost případů, při kterých znak  $A$  nabývá hodnoty (kategorie)  $A_j$  a znak  $B$

hodnoty  $B_j$ . Symbolem  $n_{i\bullet} = \sum_{j=1}^s n_{ij}$  značíme celkový počet opakování, při kterých se vyskytla  $i$ -tá kategorie znaku  $A$ , symbolem  $n_{\bullet j} = n_{1j} + n_{2j}$  značíme celkový počet opakování, při kterých se vyskytla  $j$ -tá kategorie znaku  $B$ .

Nominální (nebo ordinální) data jsou zachycena v kontingenční tabulce tak, že v řádcích jsou uvedeny četnosti pro oba výběry, ve sloupcích se nacházejí četnosti jednotlivých kategorií sledovaného znaku. Testuje se vzájemná nezávislost 2 výběrů v kontingenční tabulce. Pakliže jsou výběry nezávislé, v tom případě pocházejí ze stejné populace (tj. ze stejného základního souboru). V opačném případě pocházejí ze dvou rozdílných základních souborů.

Zadává se tato nulová hypotéza:

$H_0$  : Výběry jsou navzájem nezávislé a pocházejí z jediné populace (s daným rozdělením pravděpodobnosti).

Nulová hypotéza se na základě Chi-kvadrát testu buď přijme (tj. nezamítne), což znamená, že výběry pocházejí ze stejné populace, nebo se  $H_0$  zamítne a pak výběry pocházejí z různých populací, jinak řečeno, výběry jsou vzájemně závislé. Následuje pak analýza pomocí koeficientů kontingence a koeficientů asociace, které vyjadřují míru závislosti (viz další kapitola).

**Nulová  
hypotéza**

## ŘEŠENÝ PŘÍKLAD 2-6



### Studie

Jsou k dispozici 2 výběry: Muži a Ženy, sleduje se znak Způsob odměny, který má dvě hodnoty (kategorie): Peníze a Limonády.

**Problém 1:** Vytvořte kontingenční tabulku pro znaky: Pohlaví respondentů (řádky) versus Způsob odměny (sloupce).

**Problém 2:** Souvisí Pohlaví respondentů se Způsobem odměny?

Použijte Chi-kvadrát test.

**Poznámka:** Všimněte si, že jednotlivé kategorie (tj. Muži, Ženy) určují výběrové soubory, na nichž se sleduje znak Způsob odměny.

Naučíte se řešit daný test s pomocí Excelu.

**Testování  
závislosti  
způsobu  
odměny  
a pohlaví  
respondentů**

### Řešení příkladu

**Řešení příkladu pomocí Excelu:**

**Problém 1.** Kontingenční tabulka.

**Data** → **Kontingenční tabulka a graf** → **výběr dat ... studie.xls**

Počet z Výběr	Výběr		
Pohlaví	Peníze	Limonády	Celkový součet
Muži	78	42	120
Ženy	46	34	80
Celkový součet	124	76	200

Výstupem z Excelu je uvedená tabulka, která obsahuje příslušné četnosti, např. peníze si jako odměnu vybralo 78 mužů atd.

**Kontin-  
genční  
tabulka**

**Problém 2.** Nezávislost výběrů: Chi-kvadrát test**Postup Chi – kvadrát testu (4 kroky):****1.** Nulová hypotéza  $H_0$ :

Četnosti kategorií znaku A nezávisí na četnostech kategorií znaku B (tj. rozdíly četností jsou pouze dílem náhody).

Očekávané četnosti se vypočítají z údajů v kontingenční tabulce následovně:

relat. četnosti kategorie  $A_i = n_{i\cdot}/n$ ,

relat. četnosti kategorie  $B_j = n_{\cdot j}/n$ ,

$E_{ij}$  = očekávaná četnosti kategorie  $B_j$  v  $i$ -tém výběru:

$$E_{ij} = (n_{i\cdot}/n) * (n_{\cdot j}/n) * n = (n_{i\cdot} * n_{\cdot j})/n$$

Pozorované četnosti jsou přímo hodnoty z kontingenční tabulky:  $O_{ij}$ ,  $i=1, j=1, 2, \dots, s$ .

**2.** Testové kritérium  $T$  je analogické jako v jednovýběrovém testu:

$$T = \sum_{i=1}^2 \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

přitom  $s$  značí počet kategorií sledovaného znaku, tj. počet sloupců kontingenční tabulky. V našem příkladu je  $s = 2$ .

**3.** Kritická hodnota je  $K = \chi_{\alpha}(s-1)$ .

Přitom  $\chi_{\alpha}(s-1)$  je kritická hodnota rozdělení Chi-kvadrát se stupněm volnosti  $df = s-1$  a pro zadanou hladinu významnosti  $\alpha$ . Tuto hodnotu naleznete buď v tabulkách rozdělení Chi-kvadrát, nebo pomocí Excelu použitím funkce CHIINV( $\alpha$ ;  $df$ ).

**4.** Jestliže platí  $T > K$ , potom se  $H_0$  zamítá, jinak se  $H_0$  přijímá.

**Podmínka:** K dosažení dostatečné přesnosti testu by v každé kategorii mělo být alespoň 5 hodnot!

**Alternativně:** Pro hodnotu testového  $T$  zjistíte  $p$ -hodnotu (signifikaci), tuto hodnotu porovnáte se zvolenou hladinou významnosti ( $\alpha = 0,05$ ), pokud  $p$ -hodnota  $> \alpha$ , pak  $H_0$  nezamítáte, v opačném případě  $H_0$  zamítáte. V Excelu můžete  $p$ -hodnotu vypočítat pomocí funkce CHIDIST( $T$ ;  $df$ ).

Pozorované čet.	Peníze	Limonády	Celkový součet
Muži	78	42	120
Ženy	46	34	80
Celkový součet	124	76	200

Tab. 2.1.

Očekávané četnosti v Tab. 2.2. vypočítáte z kontingenční tabulky Tab. 2.1. jako součin příslušného řádkového a příslušného sloupcového součtu vyděleného celkovým součtem. V Tab. 2.3. jsou uvedeny příslušné sčítance testového kritéria, jejich součet - hodnota testového kritéria  $T$  je pak uvedena v prvním řádku Tab. 2.4. jako CHI-SQUARE.

Očekávané čet.	Peníze	Limonády
Muži	74,4	45,6
Ženy	49,6	30,4

Tab. 2.2.

$(E_{ij} - O_{ij})^2/E_{ij}$	
0,174	0,284
0,261	0,426

Tab. 2.3.

Nezávislost

Algoritmus

Testové kritérium



Řešení v Excelu



CHISQUARE =	1,14601
alfa =	0,05
df =	1
CHIINV =	3,84146
CHIDIST =	0,28439

Tab. 2.4.

Kritickou hodnotu pro danou hladinu významnosti a stupeň volnosti vypočítáte v Excelu pomocí funkce CHIINV. Protože platí  $df = 2 - 1 = 1$ , obdržíte

$$\text{CHI-SQUARE} = 1,146 < \text{CHIINV}(0,05 ; 1) = 3,841 ,$$

nulovou hypotézu o nezávislosti na hladině významnosti 0,05 nezamítáte! K témuž závěru lze dojít porovnáním  $p$ -hodnoty, kterou v Excelu vypočítáte pomocí funkce CHIDIST. Konkrétně  $\text{CHIDIST} = 0,284 > 0,05$  , proto  $H_0$  nezamítáte.

**Jinak řečeno:** Způsob vybrané odměny je nezávislý na pohlaví respondentů.

**Ještě jinak:** Pokud jde o výběr odměny soubory Muži a Ženy jsou navzájem nezávislé.

\*

### ÚKOL K ZAMYŠLENÍ 1



Předpokládejte, že třídící znak  $A$  má více než 2 kategorie – konkrétně  $r$  kategorií, např.  $r = 5$  pro věkové třídy 18, 19, 20, 21, 22 a více. Pak třídící znak Věková třída (je její třeba nově vytvořit ze znaku Věk!) člení data na 5 výběrových souborů. Testem Chi-kvadrát lze testovat  $H_0$ , že všech 5 výběrů pochází ze stejné populace, nebo-li že všechny soubory jsou vzájemně nezávislé. Náš 4-krokový postup řešení testu by pak byl analogický s výše uvedeným postupem pouze s tím rozdílem, že namísto 2 výběrů uvažujete  $r = 5 (> 2)$  výběrů. Navíc platí:  $df = (r-1) \cdot (s-1)$ .

Kde všude se ještě změny oproti případu  $r = 2$  projeví? Připravte si data ze studie a testujte nezávislost výběru typu odměny podle výše zmíněných 5 věkových tříd.

### PRŮVODCE TEXTEM



Následující samostatné úkoly řešte nejprve pomocí Excelu a potom pomocí SPSS, viz [Studie](#).

Rozumíš?

### SAMOSTATNÝ ÚKOL 4



Testujte hypotézu, že průměrné (mediánové) hodnocení "důležitosti image" je 3, tedy  $\tilde{\mu}_0 = 3$ .

Mediánový test

### SAMOSTATNÝ ÚKOL 5



Je rozdělení hodnocení důležitosti image jednotlivými respondenty rozloženo náhodně, tj. 7 hodnotících stupňů (a tudíž „image“ není významná - nulová hypotéza), nebo je některé hodnocení významně četnější?

Chi-kvadrát test, 1 výběr

**SAMOSTATNÝ ÚKOL 6**

Vytvořte kontingenční tabulku pro znaky: Důležitost ceny versus Důležitost image. Souvisí Důležitost ceny s Důležitostí image? Použijte Chi-kvadrát test.



*Chi-kvadrát test, 2 výběry*

**SHRnutí KAPITOLY TESTOVÁNÍ HYPOTÉZ V MARKETINGU - NEPARAMETRICKÉ TESTY**

*Shrnutí kapitoly*

Parametrické hypotézy se vztahují na jeden nebo několik parametrů daného pravděpodobnostního rozdělení náhodné veličiny (neboli znaku populace). Naproti tomu neparametrické hypotézy se netýkají parametrů rozdělení náhodné veličiny, nýbrž jiných statistických vlastností, např. tvaru rozdělení (například binomické rozdělení nebo normální rozdělení). Při testování neparametrické hypotézy máme k dispozici výsledky náhodného výběru – vzorku, tedy data z marketingového výzkumu a na jejich základě rozhodujeme testovanou hypotézu buď přijmout, nebo zamítnout. Padne-li hodnota statistiky pro získaný vzorek do kritického oboru (tj. překročili kritickou hodnotu), potom  $H_0$  zamítáme. Naopak, padne-li hodnota statistiky do oboru přijetí, pak nulovou hypotézu nezamítáme (neboli přijímáme). Znovu si prosím uvědomte, že testem statistické hypotézy nemůžeme dokázat její pravdivost nebo nepravdivost, správnost nebo nesprávnost! Testování neparametrických hypotéz nemusí vždy vést ke správným rozhodnutím, neboť jde o náhodný proces využívající omezené informace z náhodného výběru. V této kapitole jste se naučili tyto neparametrické testy: Mediánový test, Chi-kvadrát test pro jeden výběr, Chi-kvadrát test pro dva výběry. Vše bylo ilustrováno na příkladech ze Studie.

**KONTROLNÍ TEST 2**

- 2.1** Neparametrický test statistické hypotézy slouží k (označte správnou odpověď):
- výpočtu rozdělení četnosti
  - odhadu parametru hledaného rozdělení
  - zjištění typu rozdělení pravděpodobnosti
  - zjištění nezávislosti statistických znaků
- 2.2** Při testu nezávislosti:
- testuje se nulová hypotéza, že hodnoty znaků jsou stejné,
  - testuje se nulová hypotéza, že dva statistické znaky jsou vzájemně závislé,
  - testuje se nulová hypotéza, že hodnota znaku se liší od zadané hodnoty
  - testuje se nulová hypotéza, že hodnota znaku je rovna zadané hodnotě,
  - testuje se nulová hypotéza, že dva statistické znaky jsou vzájemně nezávislé.
- 2.3** Při Chi-kvadrát testu statistické hypotézy lze využít kritické hodnoty:
- Studentova t-rozdělení pravděpodobnosti,
  - Pearsonova Chi-kvadrát rozdělení pravděpodobnosti,
  - Fisherova F-rozdělení pravděpodobnosti,

- d. Gaussova normálního rozdělení pravděpodobnosti,
- e. exponenciálního rozdělení pravděpodobnosti.

2.4 Při použití Excelu pro Chi-kvadrát test mohou být data alespoň typu:

- a. nominální
- b. ordinální
- c. intervalová
- d. číselná
- e. nečíselná

2.5 Ke každé vestavěné funkci Excelu z levého sloupce přiřaďte správné vysvětlení fungování z pravého sloupce:

(1)	NORMSINV	(A)	Nalezne pravděpodobnost, že náhod. veličina s Chi-kvadrát rozdělením s daným stupněm volnosti $df$ nabude nejvýše zadané hodnoty $X^2$
(2)	NORMDIST	(B)	Nalezne kvantil normovaného normálního rozdělení k zadané pravděpodobnosti $\alpha$
(3)	CHIINV	(C)	Nalezne pravděpodobnost že náhod. veličina s normovaným normálním rozdělením nabude nejvýše zadané hodnoty $Z$
(4)	CHIDIST	(D)	Nalezne kvantil Chi-kvadrát rozdělení s daným stupněm volnosti $df$ k zadané pravděpodobnosti $\alpha$

2.6 Uspořádejte posloupnost kroků po sobě u dvouvýběrového Chi-kvadrát testu:

Krok ( ). Formulujte nulovou hypotézu  $H_0$ : Hodnoty znaku  $B$  jsou u dvou výběrů vzájemně nezávislé.

Krok ( ). Vypočítejte pozorované a očekávané četnosti hodnota  $B$  pro 1. a 2. výběr.

Krok ( ). Stanovte hladinu významnosti  $\alpha$ .

Krok ( ). Stanovte kritický obor  $C$  pomocí kritické hodnoty.

Krok ( ). Stanovte počet stupňů volnosti Chi-kvadrát rozdělení

Krok ( ). Vypočítejte hodnotu testového kritéria

Krok ( ). Zjistěte, zda hodnota testového kritéria padne do  $C$  a učiňte závěr testu.

2.7 Určete, zda následující výroky jsou správné (vepište A) nebo nesprávné (vepište N):

- a. U testu nezávislosti dvou souborů, kde sledovaný znak nabývá pouze dvou hodnot, je počet stupňů volnosti  $df = 2$ .
- b. Pokud hodnota testového kritéria padne do kritického oboru  $C$ , pak nulovou hypotézu zamítnete.
- c. Víte-li, že analyzovaná veličina nemá normální rozdělení, použijete pro testování její střední hodnoty mediánový test.

2.8 Doplňte správný výraz:

- a. To, že hypotézu zamítáte nebo \_\_\_\_\_ ještě neznamená, že hypotéza je správná nebo nesprávná.
- b. P-hodnota testu je \_\_\_\_\_ hladina významnosti, na níž ještě nulovou hypoté-

zu zamítáme.

- c. V Excelu k Chi-kvadrát testu použijete vestavěnou funkci \_\_\_\_\_.

**2.9** Doplňte správné sousloví:

- a. Jestliže analyzovaná veličina nemá normální rozdělení, použijete pro testování její střední hodnoty \_\_\_\_\_.
- b. Padne-li hodnota testového kritéria do \_\_\_\_\_ pak nulovou hypotézu přijímáte.
- c. K výpočtu kritického oboru jednovýběrového Chi-kvadrát testu potřebujete znát - \_\_\_\_\_ a \_\_\_\_\_.

## ŘEŠENÍ SAMOSTATNÝCH ÚKOLŮ



### Samostatný úkol 4

Sig. 0,0539

-výstup z Excelu.

Nulovou hypotézu přijímáme. Průměrný počet limonád vypitých denně je roven 2.

### Samostatný úkol 5

Sig. 0,00764

-výstup z Excelu.

Nulovou hypotézu zamítáme. Průměrný počet limonád vypitých denně u mužů je statisticky významně vyšší než u žen.

### Samostatný úkol 6

Sig. 0,28549

Nulovou hypotézu přijímáme. Průměrné hodnocení důležitosti Ceny a důležitosti Image se u mužů statisticky významně neliší.

Samostatný úkol 4

Samostatný úkol 5

Samostatný úkol 6

## ŘEŠENÍ KONTROLNÍHO TESTU 2



2.1 d.

2.2 d.

2.3 b.

2.4 a.

2.5 (1) – (B), (2) – (A), (3) – (C)

2.6 (1), (3), (2), (5), (4), (6), (7)

2.7 N, A, A, A, A

2.8 a. přijímáte, c. minimální (nejmenší), d. CHIINV

2.9 a. mediánový test, b. oboru přijetí, c. stupeň volnosti a hladina významnosti

Kontrolní test

## 3 REGRESNÍ ANALÝZA

### CÍLE KAPITOLY REGRESNÍ ANALÝZA

Po úspěšném a aktivním absolvování této KAPITOLY

#### Budete umět:

- rozhodnout, zda jsou příslušné regresní koeficienty statisticky významné

**Budete umět**

#### Budete schopni:

- objasnit princip stochastické závislosti mezi dvěma i více kvantitativními znaky
- vypočítat vektor regresních koeficientů,
- dokázali testovat vhodnost volby lineárního regresního modelu.

**Budete schopni**

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny.

#### PRŮVODCE STUDIEM 4



Regresní analýza se zabývá závislostí kvantitativního znaku na kvantitativním znaku (nebo více kvantitativních znacích). V případě závislosti jednoho znaku na jednom znaku mluvíme o jednoduché regresi, u závislosti jednoho znaku na více kvantitativních znacích hovoříme o vícenásobné (nebo mnohonásobné) regresi. V této kapitole byste si měli prohloubit znalosti získané o regresní analýze v základním kurzu statistiky zejména o vícenásobné regresi. Nejprve jsou zopakovány základní pojmy a uvedeny předpoklady regresní analýzy. Dále je zde odvozen vztah pro výpočet vektoru regresních koeficientů a testování statistické významnosti těchto regresních koeficientů. Výklad kapitoly uzavírá test vhodnosti regresního modelu. Také v této kapitole najdete jak příklady řešené tak i neřešené.

### 3.1 Podstata regresní analýzy

#### DEFINICE 3-1



Regresní analýza hledá matematický vztah - rovnici, jež nejlépe vyjadřuje změny proměnné (znaku)  $y$  v závislosti na změnách proměnné (znaku)  $x$ . V dalším se bude tato otázka řešit obecně pro případ, že proměnná  $y$  – tzv. *závisle proměnná* závisí na  $k$  tzv. *nezávisle proměnných*, tj. vektoru proměnných  $\vec{x} = (x_1, \dots, x_k)$ . Tvar závislosti  $y$  na  $\vec{x} = (x_1, \dots, x_k)$  je v nejjednodušším případě vyjádřen funkcí  $y = f(\vec{x})$ . Funkční předpis  $f$  může mít nejrůznější tvar, např.  $y = \beta_1 + \beta_2 x$ ,  $y = \beta_0 + \beta_1 x + \beta_2 x^2$ ,  $y = \ln(x)$  nebo jiný.

**Regresní analýza**

Při závislosti  $y$  na  $k$  proměnných máme rovnici  $y = f(x_1, \dots, x_k) = f(\bar{x})$ . Je-li tato funkce uvažována ve tvaru

$$y = \beta_1 f_1(\bar{x}) + \beta_2 f_2(\bar{x}) + \dots + \beta_k f_k(\bar{x}) \quad (3-1)$$

hovoříme o *lineární regresi* (lineární podle parametrů), nebo *lineárním regresním modelu*. Nejčastěji uvažujeme model (3-1) ve tvaru:



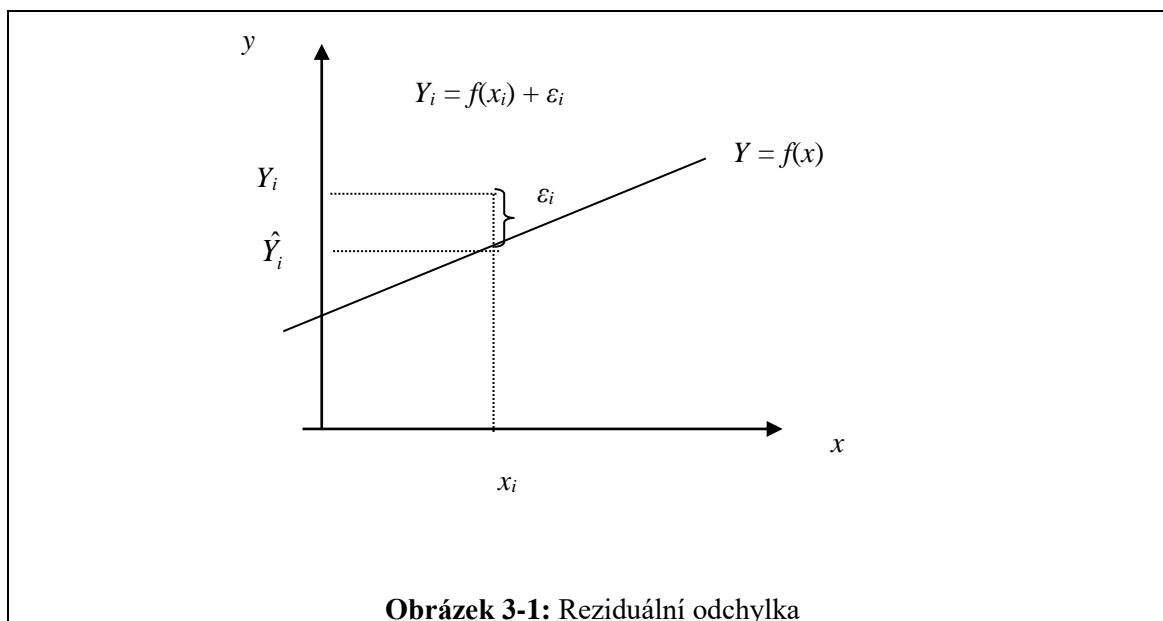
**Lineární regrese**

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (3-2)$$

Jiné tvary než (3-2) jsou pak případy nelineární regrese. Začneme nejjednodušším případem, kdy  $y$  je lineární funkcí

$$x: y = \beta_1 + \beta_2 x. \quad (3-3)$$

K dispozici necht' jsou například experimentálně získané body  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ .



**Obr 3-1:**  
**Reziduální odchylka**

**Obrázek 3-1:** Reziduální odchylka

$Y_i$  = empirická (měřená) hodnota závisle proměnné

$\hat{Y}_i$  = teoretická hodnota závisle proměnné

$\varepsilon_i$  = reziduální odchylka

Vztah  $Y_i$  a  $\hat{Y}_i$  je vyjádřen rovnicí:

$$Y_i = \hat{Y}_i + \varepsilon_i.$$

O regresním modelu, přesněji o reziduálních odchylkách, které z pohledu teorie pravděpodobnosti představují náhodné veličiny, se činí tyto **předpoklady**:

**K ZAPAMATOVÁNÍ 4**

**Předpo-  
klady  
reziduí**

1. Střední hodnota  $\varepsilon_i$  je nula, tj.  $E(\varepsilon_i) = 0$ .
2. Rozptyl je konstantní, nezávislý na  $i$ , tj.  $Var(\varepsilon_i) = \sigma^2$ .
3. Veličiny  $\varepsilon_i, \varepsilon_j$  jsou nezávislé, tj.  $Cov(\varepsilon_i, \varepsilon_j) = 0$  pro  $i \neq j$ .
4. Veličiny  $\varepsilon_i$  mají normální rozdělení, tj.  $\varepsilon_i \sim N(0, \sigma^2)$

Přitom je zvykem značit střední hodnotu symbolem  $E$ , rozptyl (varianci) symbolem  $Var$  a kovarianci symbolem  $Cov$ . Pokud jste zapomněli tyto pojmy, zopakujte si je v základním kurzu statistiky, viz doporučená literatura.

Cílem je odhad koeficientů  $\beta_1, \beta_2$  nejprve v jednoduchém lineárním regresním modelu (3-3). Později se věnujeme vícenásobnému lineárnímu regresnímu modelu (3-2).

Závisí-li konkrétně  $y$  na dvou proměnných  $x_1, x_2$ , budou body (získané například experimentálně) reprezentovány trojicemi:

$$\begin{aligned} &(x_{11}, x_{12}, Y_1) \\ &(x_{21}, x_{22}, Y_2) \\ &\dots \\ &(x_{n1}, x_{n2}, Y_n) \end{aligned}$$

První index v každé trojici představuje pořadové číslo bodu (první, druhý, ...,  $n$ -tý bod). Tyto body v trojrozměrném prostoru budou aproximovány funkcí – rovinou o rovnici  $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3$ .

Obecněji, v případě závislosti  $y$  na  $k$  proměnných  $x_1, x_2, \dots, x_k$ , předpokládáme znalost bodů v  $k+1$ -rozměrném prostoru:

$$\begin{aligned} &(x_{11}, x_{12}, \dots, x_{1k}, Y_1) \\ &(x_{21}, x_{22}, \dots, x_{2k}, Y_2) \\ &\dots \\ &(x_{n1}, x_{n2}, \dots, x_{nk}, Y_n) \end{aligned}$$

(3-4)

Regresní funkci předpokládáme ve tvaru:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Body (3-4) získané například tak, že ke zvolené  $k$ -tici  $(x_{i1}, x_{i2}, \dots, x_{ik})$  nalezneme experimentálně  $Y_i$ , jsou jediným a rozhodujícím východiskem určujícím kvalitu odhadu parametrů  $\beta_i$ . O odhadu hovoříme proto, že pracujeme s omezeným počtem bodů (výběrem). Vektor parametrů  $\vec{\beta} = (\beta_1, \dots, \beta_k)$  přísluší celému populačnímu souboru všech možných  $k+1$ -tic typu (3-4). Odhadem vektoru parametrů  $(\beta_1, \dots, \beta_k)$  bude vektor  $\vec{b} = (b_1, \dots, b_k)$ , který zjistíme pouze z výběru  $n$  bodů (3-4). Příslušnou regresní funkci budeme značit:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

(3-5)

## 3.2 Odhad regresních koeficientů

Vektor regresních koeficientů získáme z vektorové rovnice

$$\vec{b} = (X^T \cdot X)^{-1} X^T \cdot Y, \quad (3-6)$$

kde  $X$  je matice

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad (3-7)$$



vytvořená z hodnot bodů (3-4), stejně jako  $Y^T = (Y_1, Y_2, \dots, Y_n)$ .

Symboly  $Z^T$  značí transpozici matice  $Z$  a symbolem  $Z^{-1}$  se značí inverze matice  $Z$ . Matice  $X$  a  $Y$  jsou vždy známy, neboť, jak již bylo řečeno, výchozími daty jsou body (3-4). První sloupec matice  $X$  je tvořen ze samých jedniček proto, abychom mohli v regresní rovnici uvažovat absolutní člen  $b_0$ .

*Vztah pro výpočet regresních koeficientů*

**POZOR:**  $n$ -rozměrné vektory se zde považují za sloupcové matice s  $n$  řádky a 1 sloupcem!



### ŘEŠENÝ PŘÍKLAD 3-1



Určete závislost spotřeby elektrické energie ( $y$ ) na délce elektrického vedení ( $x_1$ ) a odběru energie ( $x_2$ ). Regresní funkce se předpokládá ve tvaru

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2.$$

Jsou k dispozici následující data

$x_1$	$x_2$	$Y$
1,2	3,6	3,2
1,3	3,7	3,3
1,3	3,8	3,4
1,4	3,8	3,5
1,4	3,9	3,6
1,5	3,9	3,6
1,5	4	3,7
1,6	4	3,8
1,6	4,1	3,9
1,7	4,2	4

*Spotřeba elektrické energie*



**Řešení příkladu**

Tabulka představuje body, odkud získáme potřebné matice  $X$  a  $Y$ .

$$X = \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} \quad Y = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4 \end{bmatrix}$$

Vektor odhadu regresních koeficientů  $\vec{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$  obdržíme z (3-6) a (3-7) postupně:

$$X^T \cdot X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,3 & 1,4 & 1,4 & 1,5 & 1,5 & 1,6 & 1,6 & 1,7 \\ 3,6 & 3,7 & 3,8 & 3,8 & 3,9 & 3,9 & 4 & 4 & 4,1 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} = \begin{bmatrix} 10 & 14,5 & 39 \\ 14,5 & 21,25 & 56,8 \\ 39 & 56,8 & 152,4 \end{bmatrix}$$

$$(X^T \cdot X)^{-1} = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix}$$

$$X^T \cdot Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,3 & 1,4 & 1,4 & 1,5 & 1,5 & 1,6 & 1,6 & 1,7 \\ 3,6 & 3,7 & 3,8 & 3,8 & 3,9 & 3,9 & 4 & 4 & 4,1 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4 \end{bmatrix} = \begin{bmatrix} 36 \\ 52,56 \\ 140,82 \end{bmatrix}$$

Nyní můžete podle (3-6) vypočítat:

$$\bar{b} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & -45 \end{bmatrix} \cdot \begin{bmatrix} 36 \\ 52,56 \\ 140,82 \end{bmatrix} = \begin{bmatrix} -0,78 \\ 0,60 \\ 0,90 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

Hledaná regresní funkce má tedy rovnici  $\hat{Y} = -0,78 + 0,60x_1 + 0,90x_2$ .

#### Teoretické hodnoty:

Teoretické hodnoty  $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n$  obdržíte dosazením do (3-5) za  $x_1$  a  $x_2$  postupně z tabulku vstupních dat:

$$\hat{Y}_1 = -0,78 + 0,60 \cdot 1,2 + 0,90 \cdot 3,6 = 3,18$$

$$\hat{Y}_2 = -0,78 + 0,60 \cdot 1,3 + 0,90 \cdot 3,7 = 3,33$$

...

$$\hat{Y}_{10} = -0,78 + 0,60 \cdot 1,7 + 0,90 \cdot 4,2 = 4,02$$

Tyto hodnoty lze získat pomocí maticového násobení najednou takto:

$$\hat{Y} = X \cdot \bar{b} = \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} -0,78 \\ 0,6 \\ 0,9 \end{bmatrix} = \begin{bmatrix} 3,18 \\ 3,33 \\ 3,42 \\ 3,48 \\ 3,57 \\ 3,63 \\ 3,72 \\ 3,78 \\ 3,87 \\ 4,02 \end{bmatrix} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \hat{Y}_{10} \end{bmatrix}$$



Výpočet  
teoretic-  
kých  
hodnot

**Vektor reziduálních odchylek:**

Rozdíl teoretické a skutečné hodnoty, tj. rozdíl vektorů  $\vec{e} = Y - \hat{Y}$  představuje *vektor reziduálních odchylek*. V příkladu 3-1 jej obdržíte následovně:

$$\vec{e} = Y - \hat{Y} = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,9 \\ 3,9 \\ 4 \end{bmatrix} - \begin{bmatrix} 3,18 \\ 3,33 \\ 3,42 \\ 3,48 \\ 3,57 \\ 3,63 \\ 3,72 \\ 3,78 \\ 3,87 \\ 4,02 \end{bmatrix} = \begin{bmatrix} 0,02 \\ -0,03 \\ -0,02 \\ 0,02 \\ 0,03 \\ -0,03 \\ 0,02 \\ 0,02 \\ 0,03 \\ -0,02 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ e_{10} \end{bmatrix}$$



**Výpočet  
reziduí**

**Rozptyl odhadu regresních koeficientů:**

Protože při výpočtu regresních koeficientů se jedná o odhady, je účelné také nalézt rozptyly těchto odhadů. Získáme je jako prvky hlavní diagonály matice:



**Rozptyl  
odhadu  
regresních  
koeficientů**

$$\text{Var}(\vec{b}) = s^2 \cdot (X^T \cdot X)^{-1}$$

(3-8)

kde

$$s^2 = \frac{\vec{e}^T \cdot \vec{e}}{n - k}$$

(3-9)

je tzv. *reziduální rozptyl*, přitom

$\vec{e}$  = vektor reziduálních odchylek,

$\vec{e}^T$  = vektor transponovaný k  $\vec{e}$  ,

$n$  = počet bodů,

$k$  = počet parametrů .

V příkladu 3-1 obdržíme:

$$\vec{e}^T \cdot \vec{e} = [0,02 \quad -0,03 \quad -0,02 \quad 0,02 \quad 0,03 \quad 0,03 \quad -0,02 \quad 0,02 \quad 0,03 \quad -0,02] \cdot \begin{bmatrix} 0,02 \\ -0,03 \\ -0,02 \\ 0,02 \\ 0,03 \\ 0,03 \\ -0,02 \\ 0,02 \\ 0,03 \\ -0,02 \end{bmatrix} = 0,006$$

$$s^2 = \frac{\vec{e}^T \cdot \vec{e}}{n - k} = \frac{0,006}{10 - 3} = 0,0008571$$

$$\text{Var}(\vec{b}) = s^2 \cdot (X^T \cdot X)^{-1} = 0,0008571 \cdot \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix} = \begin{bmatrix} 0,2102 & 0,0926 & -0,0883 \\ 0,0926 & 0,0514 & -0,0429 \\ 0,0883 & -0,0429 & 0,0386 \end{bmatrix}$$

Diagonálu poslední matice tvoří rozptyly jednotlivých regresních koeficientů:

$$s^2(b_0) = 0,2102 \text{ odtud směrodatná odchylka je } s(b_0) = 0,4584.$$

$$s^2(b_1) = 0,0514 \text{ odtud směrodatná odchylka je } s(b_1) = 0,2267.$$

$$s^2(b_2) = 0,0386 \text{ odtud směrodatná odchylka je } s(b_2) = 0,1965.$$

Po nalezení regresního modelu a rozptylu odhadů regresních koeficientů píšeme obvykle výsledné řešení tak, že pod regresní koeficienty do závorek uvádíme příslušné směrodatné odchylky. V našem případě máme:

$$\hat{Y} = -0,78 + 0,60x_2 + 0,90x_3 \\ (0,4584) \quad (0,2267) \quad (0,1965)$$

Při výpočtu regresních koeficientů  $b_1, b_2, \dots, b_k$  se stává, že mezi koeficienty jsou až řádové rozdíly, např.  $b_1 = 200$  a  $b_2 = 0,02$ . V takových případech stojíme před problémem, zda má smysl zařadit např. zde  $b_2$  do regresní funkce. K objektivnímu posouzení významnosti regresních koeficientů lze použít následující statistický test.

\*

### 3.3 Test významnosti regresních koeficientů

1. Nulová hypotéza:

$$H_0: \beta_i = 0,$$

$$H_1: \beta_i \neq 0.$$

2. Testovací kritérium:

$$T = \frac{b_i}{s(b_i)}$$



Spotřeba  
el. energie



Testování  
regresních  
koeficientů

(3-10)

kde  $b_i$  je odhad parametru  $\beta_i$ ,  $s(b_i)$  je směrodatná odchylka odhadu tohoto parametru.

3. Kritická hodnota  $t_{n-k}(\alpha)$ , kde  $\alpha$  je zvolená hladina významnosti (obvykle je  $\alpha = 0,05$ ).

4. Porovnáme  $T$  a  $t_{n-k}(\alpha)$

Je-li  $|T| > t_{n-k}(\alpha)$ , zamítá se  $H_0$  a přijme se alternativní hypotézu  $H_1$ , podle které vypočítaný koeficient je možné považovat za nenulový, neboli statisticky významný a je proto důvod pro jeho zařazení do regresní funkce.

V příkladu 3-1 dostáváte konkrétně

$$T_1 = \frac{b_1}{s(b_1)} = \frac{0,60}{\sqrt{0,0514}} = 2,65,$$

$$T_2 = \frac{b_2}{s(b_2)} = 4,58,$$

přičemž  $t_{n-k}(\alpha) = t_{10-3}(0,05) = 2,365$ .

Protože  $T_1 > 2,365$  a také  $T_2 > 2,365$ , jsou oba regresní koeficienty statisticky významné a nenulové a proto zařadíme oba koeficienty do regresní funkce.

Jak již bylo řečeno, výchozím předpokladem pro nalezení odhadu  $\vec{b} = (b_1, \dots, b_k)$  regresních koeficientů  $\vec{\beta} = (\beta_1, \dots, \beta_k)$  je znalost matic  $X$  a  $Y$ , což znamená znalost bodů (3-4). Praktik, stojící před úlohou formulovanou v úvodu této kapitoly, však musí tyto hodnoty získat. V této souvislosti musí rozhodnout, jak volit hodnoty  $x_{ij}$  a kolik bodů je potřeba pro seriózní odhad regresních koeficientů.

Uvážíme-li, že jak volba bodů  $x_{ij}$ , tak jejich počet mají rozhodující význam pro kvalitu odhadu, jsou obě otázky klíčové. Další zpracování těchto údajů je už rutinní záležitostí. Uvedenou problematikou se zabývá plánování experimentů, které je náplní následujících kapitol.



Spotřeba  
elektrické  
energie

### 3.4 Intervaly spolehlivosti regresních koeficientů

#### DEFINICE 3-2

Intervaly spolehlivosti pro parametry  $\beta_1, \dots, \beta_k$ , tj. intervaly, ve kterých lze očekávat tyto parametry s pravděpodobností  $1-\alpha$ , získáme pomocí vztahu:

$$[b_i - t_{n-k}(\alpha) \cdot s(b_i), b_i + t_{n-k}(\alpha) \cdot s(b_i)]$$

kde

$b_i$  = odhad parametru  $\beta_i$ ,

$s(b_i)$  = směrodatná odchylka odhadu  $b_i$ ,

$t_{n-k}(\alpha)$  = kritická hodnota Studentova rozdělení,

$n$  = počet bodů,

$k$  = počet parametrů,

$\alpha$  = hladina významnosti,

Při hladině významnosti  $\alpha$  je stupeň spolehlivosti  $1-\alpha$ . S touto pravděpodobností se nachází neznámý parametr  $\beta_i$  v intervalu (3-11).



Intervaly  
spolehli-  
vosti pro  
parametry  
(3-11)

### 3.5 Test vhodnosti regresního modelu

#### DEFINICE 3-3



Testování  
modelu

Vhodnost volby regresního modelu (tj. volby nezávisle proměnných) se ověří testem.

1. Nulová hypotéza:  $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ ,

resp. ve tvaru  $H_0 : \vec{\beta} = \vec{0}$ .

Alternativní hypotéza:  $H_1 : \vec{\beta} \neq \vec{0}$ .

2. Testovací kritérium:

$$T = \frac{S_{\hat{Y}} / (k-1)}{S_e / (n-k)},$$

(3-12)

kde

$$S_{\hat{Y}} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2,$$

$$S_e = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

3. Kritická hodnota:

$$K = F_{k-1, n-k}(\alpha),$$

kde  $F_{k-1, n-k}(\alpha)$  je kritická hodnota Fischerova rozdělení  $F$  se stupni volnosti  $df_1 = k-1$  a  $df_2 = n-k$ . V Excelu se kritická hodnota vypočítá pomocí funkce  $\text{FINV}(\alpha; df_1; df_2)$ .

4. Je-li  $T > K$ , pak se  $H_0$  zamítá, jinak se  $H_0$  nezamítá.

Použijete-li test na příklad 3-1, obdržíte:

$$T = \frac{0,594 / (3-1)}{0,006 / (10-3)} = 346,5,$$

$$K = F_{3-1, 10-3}(0,05) = 6,54.$$

Protože  $T$  překročilo kritickou hodnotu  $K$ , zamítá se  $H_0$  a model se považuje za vyhovující, tj. zamítá se hypotéza o nulovosti všech regresních koeficientů (s výjimkou  $\beta_0$ !).



#### SHRnutí KAPITOLY REGRESNÍ ANALÝZA



Shrnutí  
kapitoly

V této kapitole jste se zabývali regresní analýzou. Regresní analýza řeší vztah mezi kvantitativním znakem  $Y$  nazývaným vysvětlovaná (závislá) proměnná a mezi kvantitativními znaky nazývanými vysvětlující (nezávislé) proměnné. Zabývali jste se lineárním regresním modelem (lineárním podle parametrů). V závěru kapitoly jste se seznámili s testy, pomocí nichž lze posoudit statistickou významnost regresních koeficientů či statistickou významnost celého modelu. Text

je proložen řešením jednoho příkladu a doplněn o řešené i neřešené příklady.

Pracovali jste s těmito pojmy: lineární regrese, reziduální odchylky, odhad regresních koeficientů, teoretická hodnota, rozptyl odhadu regresních koeficientů, testování regresních koeficientů, testování vhodnosti regresního modelu.

### ŘEŠENÝ PŘÍKLAD 3-2



*Jen data*

- odhadněte regresní koeficienty v regresní rovnici  $\hat{Y} = b_0 + b_1x_1 + b_2x_2$ ,
- vypočtete teoretické hodnoty,
- vypočtete reziduální odchylky,
- vypočtete rozptyl regresních koeficientů,
- testujte regresní koeficienty,
- pro vstupní hodnoty  $X_0 = \begin{bmatrix} 1 & 11 & 3 \\ 1 & 12 & 5 \end{bmatrix}$  proveďte predikci  $Y_0$ .

$Y$	$x_1$	$x_2$
10	1	0
25	3	-1
32	4	0
43	5	1
58	7	-1
62	8	0
67	10	-1
71	10	2

#### Řešení příkladu

- Odhad regresních koeficientů

$$X^T Y = \begin{bmatrix} 368 \\ 2710 \\ 35 \end{bmatrix}$$

$$\vec{b} = \frac{1}{4664} \begin{bmatrix} 2887 & -384 & 240 \\ -384 & 64 & -40 \\ -240 & -40 & 608 \end{bmatrix} \begin{bmatrix} 368 \\ 2710 \\ 35 \end{bmatrix} = \begin{bmatrix} 6,47 \\ 6,59 \\ 0,26 \end{bmatrix}$$

- Teoretické hodnoty

$$\hat{Y} = (13,06, 25,98, 32,83, 39,68, 52,34, 59,19, 72,11, 72,89)$$

- Rezidua

$$\vec{e}^T = (-3,06, -0,98, -0,83, 3,32, 5,66, 2,81, -5,11, -1,89)$$

d. Rozptyl koeficientů  $\sum_i e_i^2 = 91,65$ .

$$\text{Var}(\vec{b}) = \frac{91,65}{8-3} \cdot \frac{1}{4664} \begin{bmatrix} 2887 & -384 & 240 \\ -384 & 64 & -40 \\ 240 & -40 & 608 \end{bmatrix} = \begin{bmatrix} 11,35 & \dots & \dots \\ \dots & 0,25 & \dots \\ \dots & \dots & 2,39 \end{bmatrix}$$

e. Test koeficientů:

$$T_0 = \frac{b_0}{s(b_0)} = \frac{6,47}{3,37} = 1,92, \quad T_1 = \frac{6,59}{0,5} = 13,18, \quad T_2 = 0,17.$$

$$K = t_{n-k}(\alpha) = t_{8-3}(0,05) = 2,571.$$

Statisticky významný je pouze koeficient  $\beta_1$ , neboť pouze  $T_1 > K$ .

f. Test vhodnosti regresního modelu:

$$T = \frac{3318/(3-1)}{91,65/(8-3)} = \frac{16,59}{18,33} = 90,51.$$

$$K = F_{k-1, n-k}(\alpha) = F_{2,5}(0,05) = 5,79.$$

Model byl zvolen správně.

g. Predikce  $Y_0$ :

$$X_0 = \begin{bmatrix} 1 & 11 & 3 \\ 1 & 12 & 5 \end{bmatrix}$$

$$Y_0 = \begin{bmatrix} 1 & 11 & 3 \\ 1 & 12 & 5 \end{bmatrix} \cdot \begin{bmatrix} 6,47 \\ 6,59 \\ 0,26 \end{bmatrix} = \begin{bmatrix} 79,74 \\ 86,85 \end{bmatrix}.$$

\*

### ŘEŠENÝ PŘÍKLAD 3-3



Zjistěte závislost hodnoty produkce na výši investic u 12 vybraných firem lehkého strojírenství.

Přitom víte, že regresní koeficient  $b_1 = 2,1622$ . Dále víte, že směrodatná odchylka tohoto regresního koeficientu  $s(b_1) = 0,615516$ . Na základě těchto údajů otestujete na 5% hladině významnosti hypotézu o nulové hodnotě regresního koeficientu základního souboru.

**Produkce firem**

#### Řešení příkladu

Dosadíte-li do testového kritéria  $T = \frac{b_j}{s_{b_j}}$ ,

dostanete  $T = \frac{2,1622}{0,615516} = 3,513$ .



V tabulce Studentova rozdělení naleznete pro  $12 - 2 = 10$  stupňů volnosti kritickou hodnotu (při oboustranné alternativě)  $t_{10}(0,05) = 2,228$ . Protože  $3,513 > 2,228$ , zamítáte na zvolené hladině významnosti hypotézu o nulové hodnotě regresního koeficientu. Tento regresní koeficient je tedy na hladině významnosti 5% statisticky významný.

\*

### ŘEŠENÝ PŘÍKLAD 3-4



Nalezněte lineární regresní rovnici, která popisuje závislost celkové měsíční tržby – proměnná  $y$  (v 10 000 Kč) na tržbě ve stravovacích úsecích – proměnná  $x_1$  (v 10 000 Kč) a na počtu „lůžkonocí“ – proměnná  $x_2$ . Měsíční kapacita hotelů se udává často tímto ukazatelem a je dána součinem celkového počtu lůžek a počtu dnů v měsíci. Výchozí údaje obsahuje tabulka.

Hotely

#### Řešení příkladu

Výchozí údaje:

$Y$	$x_1$	$x_2$
12,0	2,0	150
8,0	1,2	94
76,4	14,8	811
17,0	8,3	254
21,3	8,4	399
10,0	3,0	95
12,5	4,8	149
97,3	15,6	312
88,0	16,1	952
25,0	11,5	247
38,6	14,2	400
47,3	14,0	312

Vektor vysvětlované proměnné a matice vysvětlujících proměnných mají tvar:

$$Y = \begin{bmatrix} 12,0 \\ 8,0 \\ 76,4 \\ 17,0 \\ 21,3 \\ 10,0 \\ 12,5 \\ 97,3 \\ 88,0 \\ 25,0 \\ 38,6 \\ 47,3 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 2,0 & 150 \\ 1 & 1,2 & 94 \\ 1 & 14,8 & 811 \\ 1 & 8,3 & 254 \\ 1 & 8,4 & 399 \\ 1 & 3,0 & 95 \\ 1 & 4,8 & 149 \\ 1 & 15,6 & 312 \\ 1 & 16,1 & 952 \\ 1 & 11,5 & 247 \\ 1 & 14,2 & 400 \\ 1 & 14,0 & 312 \end{bmatrix}.$$

Odtud

$$X^T X = \begin{bmatrix} 12,0 & 113,90 & 4175,0 \\ 113,9 & 1428,43 & 51958,5 \\ 4175,0 & 51958,50 & 2266001 \end{bmatrix}, \quad X = \begin{bmatrix} 453,4 \\ 6006,8 \\ 230647,8 \end{bmatrix},$$

$$(X^T X)^{-1} = \begin{bmatrix} 0,343 & -0,02629 & -0,0003 \\ -0,026 & 0,006234 & -0,000094 \\ -0,0003 & -0,000094 & 0,00000266 \end{bmatrix},$$

$$(X^T X)^{-1} \cdot (X^T Y) = \vec{b} = \begin{bmatrix} -9,126450 \\ 3,729273 \\ 0,033091 \end{bmatrix}.$$

Vícenásobná regresní rovnice má tedy tvar

$$\hat{Y} = -9,126450 + 3,729273x_1 + 0,033091x_2.$$

\*

### ŘEŠENÝ PŘÍKLAD 3-5



Testujte statistickou významnost regresních koeficientů z předchozího příkladu. Test hypotézy o nulovosti těchto regresních koeficientů proveďte na hladině významnosti 5%.

*Hotel*

**Řešení příkladu**

Nejprve samostatně vypočítejte směrodatné odchylky regresních koeficientů. Měli byste dospět k těmto hodnotám:

$$s(b_1) = 1,371, \quad s(b_2) = 0,0283.$$

Hodnoty testového kritéria podle  $T = \frac{b_j}{s(b_j)}$  jsou

$$T_1 = \frac{3,729273}{1,371} = 2,7201, \quad T_2 = \frac{0,033091}{0,0283} = 1,1693.$$

Provedeme-li test na hladině významnosti 5%, nalezneme v tabulce Studentova rozdělení tentokrát pro  $12 - 3 = 9$  stupňů volnosti kritickou hodnotu  $t_{0,975} = 2,262$ . Porovnáme-li obě hodnoty testového kritéria  $T_1$  a  $T_2$  s uvedenou kritickou hodnotou, vidíme, že v případě proměnné  $x_1$  zamítáme na zvolené hladině významnosti hypotézu  $H_0$  o nulové hodnotě regresního parametru  $\beta_1$ . V případě proměnné  $x_2$  ovšem hodnota testového kritéria nepřekročila kritickou mez a zůstává tedy v oboru přijetí. Jinými slovy řečeno, proměnná  $x_2$  na rozdíl od  $x_1$  významným způsobem neovlivňuje v modelu hodnoty závisle proměnné a její zařazení do vícenásobného regresního modelu oprávněné nebylo.

\*

**KONTROLNÍ TEST 3**

Ano či ne?

- 3.1** Regresní analýza zkoumá závislost kvantitativních znaků.  
**3.2** Odchylku naměřených hodnot od teoretických hodnot nazýváme reziduální odchylka.  
**3.3** Regresní analýza se zabývá pouze lineární regresí.  
**3.4** Kritická hodnota  $t$  při testování regresních koeficientů je kritickou hodnotou normálního rozdělení.  
**3.5** Nulová hypotéza při testování o vhodnosti volby modelu je tvaru:  
 $H_0 : \beta_1 \neq \beta_2 \neq \dots \neq \beta_k \neq 0$ .

Doplňte

- 3.6** Klasický regresní model předpokládá, že náhodné složky mají \_\_\_\_\_ střední hodnotu a \_\_\_\_\_ rozptyl.  
**3.7** Pro testování hypotézy o nulovosti individuálních regresních koeficientů se používá \_\_\_\_\_  
**3.8** Je-li v rovnici  $\hat{Y} = b_0 + b_1x_1 + \dots + b_kx_k$  absolutní člen  $b_0$ , jsou v matici  $X$  v prvním sloupci \_\_\_\_\_  
**3.9** Regresní analýza zkoumá závislost \_\_\_\_\_ znaků.  
**3.10** Rozptyly odhadů regresních koeficientů najdeme na \_\_\_\_\_ maticе  $Var(\vec{b}) = s^2(X^T X)^{-1}$ .

- 3.11** Personální oddělení shromáždilo údaje o věku ( $X$ ) a době pracovní neschopnosti ( $Y$ ) dvaceti náhodně vybraných zaměstnanců. Zjištěné údaje jsou zaznamenány v tabulce.

Řešte

$x$	$Y$	$x$	$Y$
20	4	58	20
35	14	46	13
35	15	43	16
34	10	33	10
32	10	29	10
28	9	36	11
25	12	48	14
46	15	55	15
38	15	36	14
50	16	19	6

Odhadněte regresní koeficienty v rovnici regresní funkce  $\hat{Y} = b_0 + b_1x$ .

- 3.12** Statistický úřad zkoumal souvislost ročních úspor s ročními příjmy rodin s dvěma dětmi školou povinnými. Výsledky studie jsou zachyceny v tabulce.

<b>Příjem (tis. Kč)</b>	104	125	146	167	111	135	189	196	205	210	170	230
<b>Úspory (tis. Kč)</b>	6	5,6	9,2	14	8	9,1	20,5	29	23,2	38,5	25	40

Najděte lineární regresní model popisující závislost úspor na příjmech a na základě tohoto modelu odhadněte úspory rodiny, bude-li její roční příjem 205 tis. Kč?

- 3.13** Ze statistiky rodinných účtů jsme náhodně vybrali osm čtyřčlenných domácností se dvěma nezaopatřenými dětmi. Z podkladů jsme zjišťovali celkový hrubý roční příjem domácnosti v Kč (vysvětlující proměnná  $x$ ) a celkové roční výdaje těchto domácností za průmyslové zboží (Kč, vysvětlovaná proměnná  $Y$ ) – viz tabulka:

$x$	211399	306502	250251	264138	274060	297046	328645	249987
$Y$	42276	72341	49852	53827	54914	60409	71729	47997

- odhadněte parametry lineární regresní rovnice vystihující závislost roční výše výdajů domácností za průmyslové zboží na celkovém hrubém ročním příjmu těchto domácností,
- naleznete odhad teoretické výše těchto výdajů v domácnosti, která by měla roční příjem přes 300 000Kč,

**3.14**

- Stanovte z dat kontrolní otázky 4.13 směrodatnou odchylku odhadu parametru  $b_i, i = 0, 1$ .
- Určete pro odhad parametru  $b_1$  hodnotu testové statistiky  $T$  používané v testu hypotézy  $H_0$  o nezávislosti mezi příjmy a výdaji a tento test proveďte proti oboustranné alternativě  $H_1$  na hladině významnosti  $\alpha = 0,05$ . Údaje opět převeďte z kontrolní otázky 4.13.

**3.15** Z údajů o objemu produkce (mil. FRF, stálé ceny roku 1980, proměnná  $Y$ ), o hrubé tvorbě fixního kapitálu (mil. FRF, stálé ceny roku 1980, proměnná  $x_1$ ) a o zaměstnanosti (tis. osob, stav k 31.12.1998, proměnná  $x_2$ ) v roce 1998 v jednotlivých odvětvích národního hospodářství Francie, které jsou v tabulce, odhadněte parametry vícenásobné regresní funkce typu

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Odvětví	$Y_i$	$x_{1i}$	$x_{2i}$
Zemědělství	288443	18781	1055
Potravinářství	393828	13990	551
Energetika	330300	33813	223
Výroba polotovarů	602182	32022	1101
Výroba výrobních zařízení	426720	19520	965
Výroba zařízení pro domácnosti	34008	1258	49
Výroba dopravních prostředků	185887	10462	358
Výroba spotřebních předmětů	427766	16392	1030
Stavebnictví a veřejné práce	436926	19828	1472
Obchod	495319	36354	2691
Doprava a spoje	417147	58196	1268
Tržní služby	1002132	116083	4617
Pojišťovací služby	61827	2053	158
Finanční služby	709297	6908	441
Netržní služby	840622	136923	6148
Celkem	6652404	522583	22127

### ŘEŠENÍ KONTROLNÍHO TESTU 3



- 3.1 ano  
3.2 ano  
3.3 ne  
3.4 ne  
3.5 ne

Ano či ne?

- 3.6 nulovou, konstantní  
3.7  $t$ -test  
3.8 jedničky  
3.9 kvantitativních  
3.10 hlavní diagonále

Doplňte

- 3.11  $\hat{Y} = 1,394 + 0,296x$   
3.12  $\hat{Y} = -26,399 + 0,274x$ ; 29 711Kč

Řešte

**3.13**

- a.  $\hat{Y} = -19599,4 + 0,2796x_i$ ,  
b. pro  $x_i = 300000$  Kč je  $\hat{y}_i = 64298$  Kč,

**3.14**

- a.  $s_{b_1} = 0,03375$ ,  
b. Testová statistika  $T = 8,284$ , k tomu krit. hodnota  $t_6(0,05) = 2,447 \Rightarrow$  přijímáme  $H_1$  o existenci závislosti roční výše výdajů domácností za průmyslové zboží na celkovém hrubém ročním příjmu těchto domácností.

**3.15**  $\hat{Y} = 263684,7 + 2,2331x_1 + 66,7912x_2$

## 4 KORELAČNÍ ANALÝZA

### CÍLE KAPITOLY KORELAČNÍ ANALÝZA

Po úspěšném a aktivním absolvování této KAPITOLY

#### Budete umět:

- stanovit Spearmanův koeficient (pořadové) korelace a použít jej v praxi

**Budete  
umět**

#### Získáte:

- pochopíte vzájemnou souvislost mezi koeficientem korelace a indexem korelace

**Získáte**

#### Budete schopni:

- vypočítat koeficient korelace a index korelace,
- rozhodnout o statistické významnosti korelačních koeficientů na dané hladině významnosti.

**Budete  
schopni**

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny.

#### PRŮVODCE STUDIEM 5



V předcházející kapitole jsme u regresních modelů předpokládali, že hodnoty vysvětlujících proměnných ( $x$ ) jsou předem zadané, zatímco hodnoty vysvětlované proměnné ( $Y$ ) jsou náhodné veličiny. Často se však vyskytují případy, kdy máme k dispozici více náhodných veličin a není dopředu známo, které jsou vysvětlující a které vysvětlované. Z korelačních modelů jsou opracovány modely, které předpokládají, že pozorovaná data jsou hodnotami vícerozměrné náhodné veličiny s vícerozměrným normálním rozdělením. Korelační analýza hledá číselné charakteristiky k výjádření míry závislosti veličin  $X$  a  $Y$ , resp. obecně vektorů  $\vec{X} = (x_1, x_2, \dots, x_n)$  a  $\vec{Y} = (y_1, y_2, \dots, y_n)$ .

Kapitola je doplněna řešenými příklady, které by vám měly pomoci danou problematiku pochopit. Po prostudování této kapitoly si ověřte, zda jste schopni samostatně vypočítat příklady na konci kapitoly. Pokud budete studovat důkladně určitě to zvládnete.

## 4.1 Koeficient korelace

### DEFINICE 4-1



Koeficient korelace

V nejjednodušším případě, kdy se

- sleduje závislost veličiny  $Y$  na veličině  $X$ ,
- $X$  a  $Y$  jsou měřitelné,  $X = (x_1, \dots, x_n)$ ,  $Y = (y_1, \dots, y_n)$ ,
- regresní funkcí je přímka,

lze použít jako míru jejich lineární závislosti (*párový*) koeficient korelace  $\rho_{xy}$ , resp. výběrový koeficient korelace  $r_{xy}$ , který se vypočítá pomocí druhého vzorce (první vzorec je definiční, druhý výpočtový):

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)}, \quad r_{xy} = \frac{n \cdot \sum x_i \cdot y_i - \sum x_i \cdot \sum y_i}{\sqrt{[n \cdot \sum x_i^2 - (\sum x_i)^2][n \cdot \sum y_i^2 - (\sum y_i)^2]}} \quad (4-1)$$

Výběrový koeficient korelace  $r_{xy}$  je bodovým odhadem  $\rho_{xy}$ . Koeficient korelace nabývá hodnoty v intervalu  $[-1, 1]$ . V krajních mezích  $+1$  (resp.  $-1$ ) se jedná o *lineární funkční závislost*  $Y$  na  $X$ , přičemž přímka

$$y = b_0 + b_1 x,$$

vyjadřující tuto závislost je při hodnotě koef. korelace  $+1$  rostoucí a při hodnotě  $-1$  klesající.

Při hodnotě blízké  $0$  není mezi  $X$  a  $Y$  významná lineární závislost. Nelze však říci, že  $Y$  je nezávislé na  $X$ , neboť mezi dvěma proměnnými může být také *jiná než lineární* závislost (např. kvadratická, viz následující příklad)!

### ŘEŠENÝ PŘÍKLAD 4-1



Výpočet koeficientu korelace

Vypočítejte koeficient korelace  $r_{xy}$ , je-li dáno:

$x$	-2	-1	0	1	2
$y$	4	1	0	1	4

#### Řešení příkladu

K ručnímu výpočtu  $r_{xy}$  je výhodné uspořádat potřebné výpočty do tabulky

$x_i$	$y_i$	$x_i \cdot y_i$
-2	4	-8
-1	1	-1
0	0	0
1	1	1
2	4	8
$\sum x_i = 0$	$\sum y_i = 10$	$\sum x_i \cdot y_i = 0$



Další údaje vyskytující se ve vzorci (4-1) není potřeba počítat (tj.  $\sum x_i^2$ ,  $\sum y_i^2$ ), protože už z čitatele je vidět  $\sum x_i \cdot y_i = \sum x_i = 0$  a tedy  $r_{xy} = 0$ .

Pokud bychom z tohoto výsledku usuzovali, že  $X$  a  $Y$  jsou nezávislé, dopouštíme se hrubé chyby, neboť se jedná v tomto případě o funkční závislost  $y = x^2$ . Na základě  $r_{xy} = 0$  lze tedy pouze konstatovat:  $Y$  je **lineárně** nezávislé na  $X$ . Většinou však nevychází  $r_{xy} = 0$ , ale „blízké nule“. Je ovšem diskutabilní, od kdy máme  $r_{xy}$  považovat za blízké nule, např. zda  $r_{xy} = 0,4$ , nebo až  $r_{xy} = 0,004$ ?

Abychom objektivně rozhodli o míře závislosti (či nezávislosti)  $Y$  na  $X$ , testujeme hypotézu  $\rho_{xy} = 0$  pomocí *testu hypotézy o nulovosti koeficientu korelace* pomocí výběrového koeficientu korelace  $r_{xy}$  takto:

**1.** Nulová hypotéza (= výchozí předpoklad)

$$H_0: Y \text{ nezávisí na } X, \text{ tj. } \rho_{xy} = 0$$

$\rho_{xy}$  = teoretická hodnota (populačního) koeficientu korelace,

Alternativní hypotéza (pro případ neplatnosti  $H_0$ )

$$H_1: \rho_{xy} \neq 0$$

**2.** Testovací kritérium:

$$T = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

(4-2)

$n$  = počet dvojic  $(x_i, y_i)$ .

**3.** Kritická hodnota  $K = t_{n-2}(\alpha/2)$  z tabulek Studentova rozdělení pro daný rozsah souboru  $n$  a zvolenou hladinu významnosti  $\alpha$ ,  $df = n-2$  je stupeň volnosti.

**4.** Je-li  $|T| \leq K$ , pak se  $H_0$  přijímá, tj.  $Y$  není lineárně závislé na  $X$ , jinak přijímáme  $H_1$ , což znamená, že  $Y$  je lineárně závislé na  $X$ .

\*

## ŘEŠENÝ PŘÍKLAD 4-2

Mějme hodnoty  $x_i$  a  $y_i$ :

$x_i$	$y_i$	$x_i y_i$	$x_i^2$	$y_i^2$
-2	-5	10	4	25
-1	-3	3	1	9
0	0	0	0	0
1	1	1	1	1
2	4	8	4	16
$\Sigma = 0$	$\Sigma = -3$	$\Sigma = 22$	$\Sigma = 10$	$\Sigma = 51$



Testování  
koeficientu  
korelace



Tabulky  
údajů

### Řešení příkladu

S pomocí vzorce (4-1) obdržíme

$$r_{xy} = \frac{5.22 - 0.(-3)}{\sqrt{(5.10 - 0).(5.51 - (-3))^2}} = 0,9918.$$

Taková hodnota je bez testování jasným znakem vysoké závislosti na  $x$ .

Přesto provedeme test, podle (4-2) vypočítáme:

$$T = \frac{0,9918 \cdot \sqrt{5-2}}{\sqrt{1-0,99812}} = \frac{1,718}{\sqrt{0,126}} = 13,63 ,$$

$\alpha = 0,1$  ,  $K = t_{5-2}(0,05) = 3,182$  ( =TINV(0,05; 3) ).

Protože  $T = 13,63 > K = t_3(0,05) = 3,182$ , je závislost  $Y$  na  $X$  statisticky významná. Potvrzuje to také  $p$ -hodnota testu = TDIST(13,63; 3; 2) = 0,00085 , to znamená, že hodnota korelačního koeficientu je statisticky významná počínaje na hladině 0,085%, tj. na každé „rozumné“ hladině významnosti.

\*

## 4.2 Index korelace

### DEFINICE 4-2



Není-li regresní funkcí přímka, ale jiná **nelineární** funkce, je možné k vyjádření závislosti  $X$  a  $Y$  použít *index korelace*:

$$I_{xy} = \sqrt{\frac{S_{\hat{Y}}}{S_Y}} ,$$

kde vystupují součty čtverců

$$S_{\hat{Y}} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 ,$$

$$S_Y = \sum_{i=1}^n (Y_i - \bar{Y})^2 .$$

Výpočet  $I_{xy}$  je pracnější než u  $r_{xy}$  z toho důvodu, že je potřeba nejprve najít rovnici regresní funkce, aby mohly být vypočítány teoretické hodnoty  $\hat{Y}_i$  a z nich potřebný součet čtverců.

Index korelace  $I_{xy}$  nabývá hodnot  $0 < I_{xy} < 1$ . Diskuse k výsledkům je stejná jako u  $r_{xy}$ , testování významnosti se obvykle neprovádí.  $I_{xy}$  je použitelný též pro lineární regresní funkci, pak ovšem je totožný s korelačním koeficientem.

(4-3)

### 4.3 Spearmanův koeficient (pořadové) korelace

#### DEFINICE 4-3

Jsou-li hodnoty veličin  $X, Y$  zadány **pořadím**, používá se k vyjádření míry jejich závislosti *Spearmanův koeficient (pořadové) korelace*, který se vypočítá

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

$d_i$  je diference pořadí znaků  $X$  a  $Y$ .



**Spearmanův koeficient (4-4)**

#### ŘEŠENÝ PŘÍKLAD 4-3



Výrobky byly seřazeny odbornou a laickou komisí podle jakosti. Rozhodněte, zda se výsledky hodnocení obou komisí shodují.

**Komise laiků a odborníků**

Výrobek	Laické hodnocení	Odborné hodnocení	$d_i$	$d_i^2$
1	7	8	-1	1
2	9	9	0	0
3	8	7	1	1
4	10	10	0	0
5	6	6	0	0
6	5	4	1	1
7	3	5	-2	4
8	4	3	1	1
9	2	2	0	0
10	1	1	0	0

#### Řešení příkladu

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6.8}{10.9} = 0,95.$$

Testování významnosti  $r_s$  lze provést jako u párového koeficientu korelace.

\*

## 4.4 Vícenásobná lineární závislost – vztahy pro dvě vysvětlující proměnné

### DEFINICE 4-4



Vícenásobná lineární závislost

Chceme-li zjistit závislost proměnné  $Y$  na větším počtu vysvětlujících proměnných  $X_1, X_2, \dots, X_p$  používáme k měření těsnosti závislosti buď:

- koeficienty dílčí (parciální) korelace,
- koeficient vícenásobné korelace.

**Ad a.** *Koeficient dílčí (parciální) korelace*  $r_{y x_1 \cdot x_2, \dots, x_p}$  měří intenzitu lineární závislosti proměnné  $Y$  na vysvětlující proměnné  $X_1$ , za předpokladu, že všechny ostatní proměnné **za tečkou** jsou konstantní.

koeficient dílčí (parciální) korelace

Vztahy pro výpočet v případě dvou vysvětlujících proměnných:

$$r_{y x_1 \cdot x_2} = \frac{r_{y x_1} - r_{y x_2} r_{x_1 x_2}}{\sqrt{(1 - r_{y x_2}^2)(1 - r_{x_1 x_2}^2)}} \quad (4-5)$$

$$r_{y x_2 \cdot x_1} = \frac{r_{y x_2} - r_{y x_1} r_{x_1 x_2}}{\sqrt{(1 - r_{y x_1}^2)(1 - r_{x_1 x_2}^2)}} \quad (4-6)$$

**Test statistické významnosti koeficientu parciální korelace:**

Testování parciálních koeficientů

- $H_0: \rho_{y x_1 \cdot x_2, \dots, x_p} = 0$  (není korelač. závislost),  $H_1: \rho_{y x_1 \cdot x_2, \dots, x_p} \neq 0$

- Testové kritérium:  $T = \frac{r_{y x_1 \cdot x_2, \dots, x_p} \sqrt{n - p - 1}}{\sqrt{1 - r_{y x_1 \cdot x_2, \dots, x_p}^2}}$ ,

$p$  je počet vysvětlujících proměnných.

- Kritická hodnota Studentova rozdělení:  $t_{n-p-1}(\alpha)$  (=TINV( $\alpha$ ;  $n-p-1$ )).

Pokud  $|T| > t_{n-p-1}(\alpha)$ , pak je koeficient parciální korelace statisticky významný.

**Ad b.** *Koeficient vícenásobné korelace* měří těsnost proměnné  $Y$  na všech vysvětlujících proměnných  $X_1, X_2, \dots, X_p$  dohromady. **Pro 2 vysvětlující proměnné** obdržíte:



$$r_{y \cdot x_1 x_2} = \sqrt{\frac{r_{y x_1}^2 - 2r_{y x_1} r_{y x_2} r_{x_1 x_2} + r_{y x_2}^2}{1 - r_{x_1 x_2}^2}}; \quad 0 \leq r_{y \cdot x_1 x_2} \leq 1. \quad (4-7)$$

**Test statistické významnosti koeficientu vícenásobné korelace:**

Testování koeficientu vícenásobné korelace

- $H_0: \rho_{y \cdot x_1 x_2, \dots, x_p} = 0$  (není závislost),

$$H_1: \rho_{y \cdot x_1 x_2, \dots, x_p} \neq 0$$

- Testové kritérium:  $T = \frac{r_{y \cdot x_1 x_2, \dots, x_p}^2 (n - p - 1)}{(1 - r_{y \cdot x_1 x_2, \dots, x_p}^2) p}$ ,

$p$  je počet vysvětlujících proměnných.

Kritická hodnota Fisherova rozdělení:  $F_{(p,n-p-1)}(\alpha) = \text{FINV}(\alpha;p;n-p-1)$ . Pokud je  $|T| > F_{(p,n-p-1)}(\alpha)$ , pak je koeficient vícenásobné korelace statisticky významný na dané hladině významnosti.

## SHRnutí KAPITOLY KORELAČNÍ ANALÝZA



Shrnutí kapitoly

V této kapitole jste se seznámili s dalším důležitým tématem: korelační analýzou. Naučili jste se stanovit koeficient korelace, index korelace a Spearmanův index pořadové korelace. V závěru kapitoly jsou uvedeny koeficienty parciální a vícenásobné korelace, ale protože výpočet pro případ, že  $Y$  závisí na proměnných  $X_1, X_2, \dots, X_p$ , kde  $p > 2$ , je pracný, je v řešených příkladech uveden tvar pro výpočet vícenásobného koeficientu korelace pro případ  $p = 2$ .

## ŘEŠENÝ PŘÍKLAD 4-4



Ubytovací zařízení

Zjistěte, zda existuje mezi následujícími proměnnými závislost. Vypočítejte parciální korelační koeficienty, proveďte test významnosti menšího z nich (hladina významnosti 5%), vypočítejte koeficient vícenásobné korelace a testujte jeho významnosti na hladině významnosti 5%.

Data představují celkové měsíční tržby  $Y$  (v mil. Kč), tržby stravovacích úseků  $X_1$  (v mil. Kč), počet „lůžkonocí“  $X_2$ . (Počet lůžkonocí je dán součinem počtu lůžek a počtu dnů v měsíci.

$Y$	12	8	76,4	17	21,3	10
$X_1$	2	1,2	14,8	8,3	8,4	3
$X_2$	150	94	811	254	399	95

$Y$	12,5	97,3	88	25	38,6	47,3
$X_1$	4,8	15,6	16,1	11,5	14,2	14
$X_2$	149	312	952	247	400	312

### Řešení příkladu

Nejprve vypočteme jednoduché korelační koeficienty:

$$r_{yx_1} = 0,85 \quad r_{yx_2} = 0,75 \quad r_{x_1x_2} = 0,73.$$

Dosadíme do vztahu pro parciální korelační koeficienty (4-5), (4-6):

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1x_2}^2)}},$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{x_1x_2}^2)}}.$$

Po dosazení dostáváme následující hodnoty parciálních korelačních koeficientů:

$$r_{yx_1 \cdot x_2} = 0,67 \quad r_{yx_2 \cdot x_1} = 0,36.$$

Dalším úkolem je testovat parciální korelační koeficient  $r_{yx_2 \cdot x_1} = 0,36$ :

$$1. \quad H_0 : \rho_{yx_2 \cdot x_1} = 0 \text{ (není závislost)}, \quad H_1 : \rho_{yx_2 \cdot x_1} \neq 0.$$

$$2. \quad \text{Testové kritérium: } T = \frac{r_{yx_2 \cdot x_1} \sqrt{12-2-1}}{\sqrt{1-r_{yx_2 \cdot x_1}^2}} = 1,16.$$

$$3. \quad \text{Kritická hodnota Studentova rozdělení: } t_{12-2-1}(0,05) = 2,262.$$

Protože hodnota testového kritéria leží v oboru přijetí, tj.  $|T| < t_9(0,05)$ , přijímáme nulovou hypotézu a parciální korelační koeficient není statisticky významný. Jinak řečeno, nebylo prokázáno, že by celkové měsíční tržby závisely na počtu lůžkonocí.

Nakonec vypočítáme koeficient vícenásobné korelace a testujme jeho statistickou významnost. Po dosazení do vztahu (4-7) dostaneme:

$$r_{y \cdot x_1 x_2} = \sqrt{\frac{r_{yx_1}^2 - 2r_{yx_1} r_{yx_2} r_{x_1 x_2} + r_{yx_2}^2}{1 - r_{x_1 x_2}^2}} = 0,87$$

Jak vidíte, je jeho hodnota je vždy větší než největší z jednoduchých korelačních koeficientů. Nyní následuje test:

$$1. \quad H_0 : \rho_{y \cdot x_1 x_2} = 0 \text{ (není závislost)}, \quad H_1 : \rho_{y \cdot x_1 x_2} \neq 0.$$

$$2. \quad \text{Testové kritérium: } T = \frac{r_{y \cdot x_1 x_2}^2 (12-2-1)}{(1-r_{y \cdot x_1 x_2}^2)^2} = 14,54.$$

3. Kritická hodnota Fisherova rozdělení:

$$F_{(2,12-2-1)}(0,05) = F_{(2,9)} = 4,26 \text{ (=FINV(0,05;2;9))}.$$

Protože hodnota testového kritéria leží v kritickém oboru, zamítáme nulovou hypotézu a lze tvrdit, že celkové měsíční tržby jsou **společně** ovlivněny jak tržbami stravovacích úseků, tak současně počtem lůžkonocí.

\*

## ŘEŠENÝ PŘÍKLAD 4-5



V náhodném výběru o rozsahu 25 pozorování byl vypočítán koeficient korelace  $r_{xy} = 0,23$ .  
Ověřte testem, zda z tohoto výsledku lze usuzovat na lineární nezávislost mezi proměnnými  $X$  a  $Y$  v základním souboru.

### Řešení příkladu

Vypočítáme hodnotu testového kritéria podle

$$T = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0,23 \sqrt{25-2}}{\sqrt{1-0,23^2}} = 1,133.$$

V tabulce kritických hodnot Studentova rozdělení  $t$  naleznete při 5% hladině významnosti pro 23 stupňů volnosti kritickou hodnotu

$$t_{23}(0,05) = 2,069.$$

Protože je  $1,133 < 2,069$ , tj. hodnota testového kritéria náleží do oboru přijetí, nelze na 5% hladině významnosti zamítnout hypotézu  $H_0$ . Existenci lineární závislosti mezi veličinami v základním souboru považujeme tedy za neprokázanou.

\*

### ŘEŠENÝ PŘÍKLAD 4-6



V podniku Canard se při hodnocení údajů vycházejících z vnitropodnikového účetnictví sledovala závislost vlastních nákladů připadajících na jednotku produkce (proměnná  $Y$ ) na objemu produkce v 1 000 ks (proměnná  $X$ ).

**Náklady**

$x_i$	60	71	92	144	192	306
$y_i$	5157	2620	1986	1582	1100	954

$x_i$	437	481	747	989	1383
$y_i$	729	456	200	196	110

Vypočítejte index korelace za předpokladu hyperbolické regresní závislosti.

#### Řešení příkladu

Nejprve je zapotřebí stanovit metodou nejmenších čtverců regresní koeficienty  $a$  a  $b$  hyperbolické závislosti  $Y = \frac{a}{X} + b$ , poté sem dosadit  $x_i$  a vypočítat  $\hat{y}_i$  a také  $\bar{y}$ .

Dosadíte-li příslušné součty řádků do (5-3):

$$I_{yx} = \sqrt{\frac{\frac{1}{n} \sum (\hat{y}_i - \bar{y})^2}{\frac{1}{n} \sum (y_i - \bar{y})^2}} = \sqrt{\frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}},$$

dostanete

$$I_{yx} = \sqrt{\frac{19813814}{22155242}} = 0,945.$$

Výsledná hodnota indexu ukazuje na vysokou těsnost závislosti mezi vlastními náklady na jednotku produkce a objemem produkce. Lze tedy říci, že použití regresní hyperboly tuto závislost velmi dobře vystihuje a případné regresní odhady prováděné na jejím základě možno považovat za kvalitní.

\*

**KONTROLNÍ TEST 4**

Ano či ne?

- 4.1 Koeficient korelace vyjadřuje lineární závislost  $Y$  na  $X$ .  
 4.2 Koeficient korelace nabývá hodnot z intervalu  $[-1,1]$ .  
 4.3 Nulová hypotéza při testování statistické významnosti korelačního koeficientu předpokládá, že  $Y$  nezávisí na  $X$ .  
 4.4 Výpočet indexu korelace  $I$  je mnohem snadnější než výpočet koeficientu korelace.  
 4.5 Spearmanův koeficient nabývá hodnot z intervalu  $[-1,1]$ .

Doplňte

- 4.6 Korelační analýza hledá číselné charakteristiky k vyjádření \_\_\_\_\_  
 4.7 Je-li hodnota koeficientu korelace  $r_{xy} = 1$ , pak je přímka, která vyjadřuje tuto závislost \_\_\_\_\_  
 4.8 Index korelace  $I$  nabývá hodnot z intervalu \_\_\_\_\_  
 4.9 Pokud jsou hodnoty veličin  $X, Y$  zadány pořadím, používá se k vyjádření míry jejich závislosti \_\_\_\_\_ koeficient korelace.  
 4.10 Je-li veličina  $Y$  závislá na vektoru  $X = (X_1, X_2, \dots, X_m)$  používá se k výpočtu míry jejich závislosti koeficient \_\_\_\_\_

Řešte

- 4.11 Vypočítejte koeficient korelace mezi těžbou uhlí v 1000t a náklady na vytěženou tunu v Kč. Výchozí údaje potřebné k výpočtu jsou uvedeny v tabulce.

Důl č.	$x$ .	$y$ .
1	350	37
2	351	38
3	329	38
4	329	38,5
5	327	37,5
6	322	39,1
7	321	39,6
8	316	42,1
9	298	42,9
10	286	43,5
$\Sigma$	3229	396,2



- 4.12** Ze statistiky rodinných účtů jsme náhodně vybrali osm čtyřčlenných domácností se dvěma nezaopatřenými dětmi. Z podkladů jsme zjišťovali celkový hrubý roční příjem domácnosti v Kč (proměnná  $X$ ) a celkové roční výdaje těchto domácností za průmyslové zboží (proměnná  $Y$ ). Údaje jsou uvedeny v tabulce. Vypočítejte koeficient determinace a korelační koeficient.

$x_i$	211399	306502	250251	264138
$y_i$	42276	72341	49852	53827

$x_i$	274060	297046	328645	249987
$y_i$	59914	60409	71729	47997

- 4.13** Filmový festival v kategorii hudebních filmů představil festivalové porotě 10 snímků, které současně oceňovala v anketě i divácká obec. Pořadí hodnocení děl (ta označíme písmeny  $A, B, \dots, J$ ) shrnuje tabulka.

Film	$A$	$B$	$C$	$D$	$E$	$F$	$G$	$H$	$I$	$J$
Pořadí odb. poroty	5	7-9	7-9	1	2-3	7-9	2-3	4	6	10
Pořadí v anketě	1	6	4	3	8	7	2	5	10	9

- 4.14** Z údajů o objemu produkce (mil. FRF, stálé ceny roku 2000, proměnná  $Y$ ), o hrubé tvorbě fixního kapitálu (mil. FRF, stálé ceny roku 2000, proměnná  $X_1$ ) a o zaměstnanosti (tis. osob, stav k 31. 12. 2002, proměnná  $X_2$ ) v roce 2002 v jednotlivých odvětvích národního hospodářství Francie, které jsou v tabulce, vypočítejte koeficient vícenásobné korelace.

Odvětví	$y_i$	$x_{1i}$	$x_{2i}$
Zemědělství	288443	18781	1055
Potravinářství	393828	13990	551
Energetika	330300	33813	223
Výroba polotovarů	602182	32022	1101
Výroba výrobních zařízení	426720	19520	965
Výroba zařízení pro domácnosti	34008	1258	49
Výroba dopravních prostředků	185887	10462	358
Výroba spotřebních předmětů	427766	16392	1030
Stavebnictví a veřejné práce	436926	19828	1472

**ŘEŠENÍ KONTROLNÍHO TESTU 4**

- 4.1 ano  
4.2 ne  
4.3 ano  
4.4 ne  
4.5 ano

**Ano či ne?**

- 4.6 míry závislosti  
4.7 rostoucí  
4.8 (0,1)  
4.9 Spearmanův  
4.10 vícenásobné korelace

**Doplňte**

4.11  $r_{yx} = -0,8967$

**Řešte**

4.12  $r_{yx}^2 = 0,9196 = 91,96\%$ ,  $r_{yx} = 0,959$

4.13 Spearmanův koeficient pořadové korelace  $r_{i_x i_y} = 0,4493$

4.14  $r_{y \cdot x_1 x_2}^2 = 0,6069 = 60,69\%$ ,  $r_{y \cdot x_1 x_2} = 0,779$

## 5 METODY PROGNÓZOVÁNÍ TRŽEB

### CÍLE KAPITOLY METODY PROGNÓZOVÁNÍ TRŽEB

#### Po úspěšném a aktivním absolvování této KAPITOLY

##### Budete umět:

- Prognózovat časové řady tržeb.
- Znat podstatu dekompozičních modelů časových řad.
- Rozlišit hlavní složky modelu časových řad:
  - trendová složka (lineární trend, logistický trend),
  - sezónní složka,
  - náhodná složka (předpoklady, které má splňovat).
- Pomocí Excelu a SPSS provést dekompozici dané časové řady tržeb a na jejím základě prognózovat hodnoty časové řady

**Budete  
umět**

##### Budete schopni:

- Vysvětlit cíle prognózování časových řad tržeb eventuálně jiných ekonomických ukazatelů.
- Správně použít dekompoziční modely časových řad: aditivní a multiplikační.
- Dekomponovat danou časovou řadu a identifikovat v ní její hlavní složky: trend, sezónní složku a náhodnou složku (reziduum).
- Pomocí Excelu a SPSS provést analýzu dané časové řady a na jejím základě prognózovat hodnoty časové řady v budoucnosti.

**Budete  
schopni**

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny.

#### PRŮVODCE STUDIEM 6



Analýza časových řad je vedena snahou po vysvětlení minulosti a předvídání budoucnosti, v ekonomické oblasti se jedná o vývojové trendy ukazatelů hospodářské činnosti. Analýza časových řad jako soubor metod a postupů nabízí širokou škálu nástrojů a technik. Ke klasickým analytickým postupům založeným na regresi z předchozích kapitol a syntetickým přístupům založeným na technikách vyrovnání časových řad, přistupuje moderní, výpočetně náročnější harmonická analýza a Box-Jenkinsova metodologie, modely ARIMA, ARCH a GARCH využívající současného mohutného rozvoje výpočetní techniky. Naše ambice však až k těmto metodám nedosáhnou. Zájemce o hlubší poznání metod časových řad odkazujeme na literaturu, nebo jim doporučujeme si zvolit ve studijním programu některý ze specializovaných kurzů, např. Ekonometrie nebo přímo Analýza časových řad. Podobně jako v předchozích kapitolách, sledujte řešené příklady v textu a potom zkuste samostatně vyřešit neřešené příklady. Výsledky si zkontrolujte se správnými výsledky ze závěru kapitoly. Pokud něco samostatně nezvládáte, vraťte se nazpět k řešeným příkladům a projděte je ještě jednou.

## 5.1 Časové řady – prognózování tržeb

Časová řada (ČŘ) je posloupnost prostorově a věcně srovnatelných číselných údajů uspořádaných v čase od minulosti přes přítomnost do budoucnosti. Zde nás budou zajímat zejména ČŘ ekonomických veličin, speciálně tržeb tzv. **ekonomické časové řady**. Rozlišují se **okamžikové ČŘ** (stavové hodnoty v daných časových okamžicích, např. počet zaměstnanců podniku vždy k 1. dni měsíce) a **intervalové ČŘ** (hodnoty ukazatele za časové intervaly, např. hodnoty tržeb podniku za jednotlivé měsíce). V časové řadě se obvykle předpokládá, že:

- jediným (agregovaným) faktorem změny je čas (označuje se  $t$ ),
- údaje jsou uvedeny za ekvidistantní (tj. stejně dlouhé) časové intervaly.

**Prognózování** (též predikování, předvídaní apod.) **tržeb** představuje odhadování velikostí tržeb na jeden nebo více časových intervalů do budoucna. Rozlišujeme bodové a intervalové prognózy.

### Dekompoziční modely ČŘ

Každá časová řada může obsahovat až 4 složky, které vyjadřují různé druhy pohybu analyzovaného ukazatele:

- trendovou složku (trend)  $T_t$ ,
- sezónní složku  $S_t$ ,
- cyklickou složku  $C_t$ ,
- náhodnou složku  $\varepsilon_t$ .

Trendová složka vyjadřuje základní směřování hodnot ČŘ (růst, pokles a jejich eventuální zesílení nebo tlumení). Sezónní a cyklická složka souhrnně nazývané **periodická složka** zachycuje pravidelné kolísání hodnot ČŘ.

Trendová, sezónní a cyklická složka tvoří společně **deterministickou složku**, která se značí  $Y_t$ , tj.  $Y_t = T_t + S_t + C_t$ . Zpravidla se uvažuje, že složky  $Y_t$  jsou v aditivním vztahu, takže model časové řady můžete zapsat ve tvaru

$$y_t = T_t + S_t + C_t + \varepsilon_t \quad (5-1)$$

V tomto případě se hovoří o **aditivním modelu** časové řady. V ekonomických časových řadách se nejčastěji setkáte se dvěma speciálními případy modelu (5-1). U střednědobých modelů (s roční periodicitou) se obvykle předpokládá  $S_t = C_t = 0$ , pak model časové řady (5-1) má tvar:

$$y_t = T_t + \varepsilon_t \quad (5-2)$$

U krátkodobých modelů časových řad (s čtvrtletní nebo měsíční periodicitou) se předpokládá, že  $C_t = 0$ , a tedy model (5-1) má tvar:

$$y_t = T_t + S_t + \varepsilon_t \quad (5-3)$$

jedná se pak o **časovou řadu se sezónní složkou**.

Vedle aditivního modelu (5-1) je **multiplikativní model** založen na předpokladu, že vzájemný vztah jednotlivých složek obsažených v modelu je dán vzájemným násobením:

$$y_t = T_t \cdot S_t \cdot C_t \cdot \varepsilon_t, \quad \text{resp.} \quad y_t = T_t \cdot S_t \cdot C_t + \varepsilon_t. \quad (5-4)$$

Časová  
řada – ČŘ  
Ekonomická  
ČŘ

Prognózo-  
vání tržeb

4 složky  
ČŘ

Determi-  
nistická  
složka

Aditivní  
model

Sezónní  
složka

Multiplika-  
tivní model

*Popis a kvantifikace jednotlivých složek modelu časové řady patří k hlavním úkolům analýzy časových řad.*

**Metody dekompozice**, které slouží ke stanovení jednotlivých složek ČŘ, se rozdělují do dvou skupin, na:

**Metody dekompozice**

1 - **analytické**: regresní analýza (metoda nejmenších čtverců),

2 - **syntetické**: klouzavé průměry a exponenciální vyrovnání (jednoduché, Holtovo, Wintersovo).

### PRŮVODCE TEXTEM



V dalších odstavcích se budete postupně zabývat jednotlivými složkami ČŘ s ohledem na metody jejich dekompozice a možnosti prognózování tržeb

## 5.2 Analýza trendové složky

Jak již bylo řečeno, jediným faktorem vývoje dynamiky analyzovaného ukazatele je čas  $t$ . Tato úvaha, která sice často výrazně zjednodušuje skutečnost, umožňuje jednoduše modelovat časovou řadu a oddělit od sebe její jednotlivé složky, z nichž trendová složka představuje nejdůležitější komponentu analyzované časové řady.

**Jediný faktor**

V předešlém odstavci jste zaznamenali dva obecné přístupy:

(A) analytický a (B) syntetický.

Trendová složka je analyzována každým z těchto přístupů specifickými prostředky, se kterými se nyní podrobněji seznámíte, neboť tvoří základ analýzy ČŘ.

**ad (A):** Analytický přístup - nejdůležitější trendy používané při prognózování tržeb:

### 1. Lineární trend

**Lineární trend**

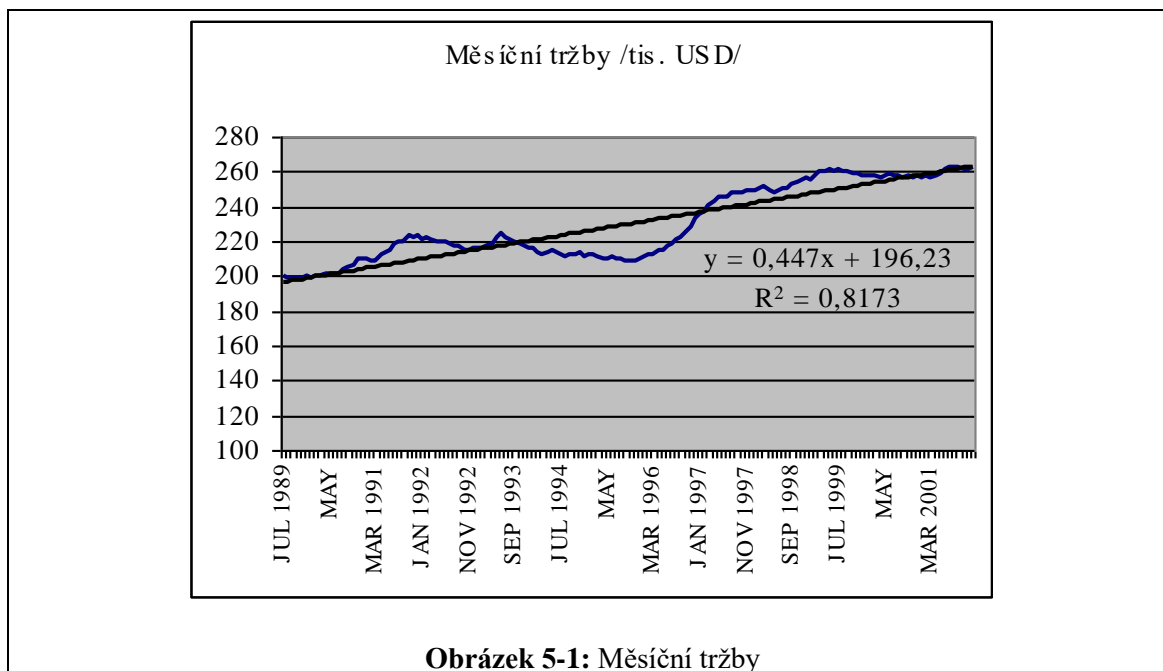
**Příklad 1.** Tržby obchodní organizace ( v USA), viz Obr. 5-1.

Na Obr. 5-1. vidíte zobrazenou ČŘ měsíčních tržeb jisté obchodní organizace za období červenec 1989 až červen 2001, tedy za období 12 let, tj. 144 měsíců.

K zobrazení pomocí Excelu byl použit postup volby z menu:

**Vložit → Graf → XY bodový → Bodový s danými body spojenými pomocí úseček**

Tento postup zobrazení časových řad je obvyklý i v jiných programech, např. také v SPSS. Po klepnutím myši na graf vyvoláte další nabídku, z níž volbu: **Přidat spojnicí trendu** již znáte z jednoduché regresní analýzy, viz kapitola 4. Výsledkem je zobrazení regresní přímky, její rovnice a koeficient determinace  $R^2$ .

Obr. 5-1:  
Měsíční  
tržby

## 2. Logistický trend.

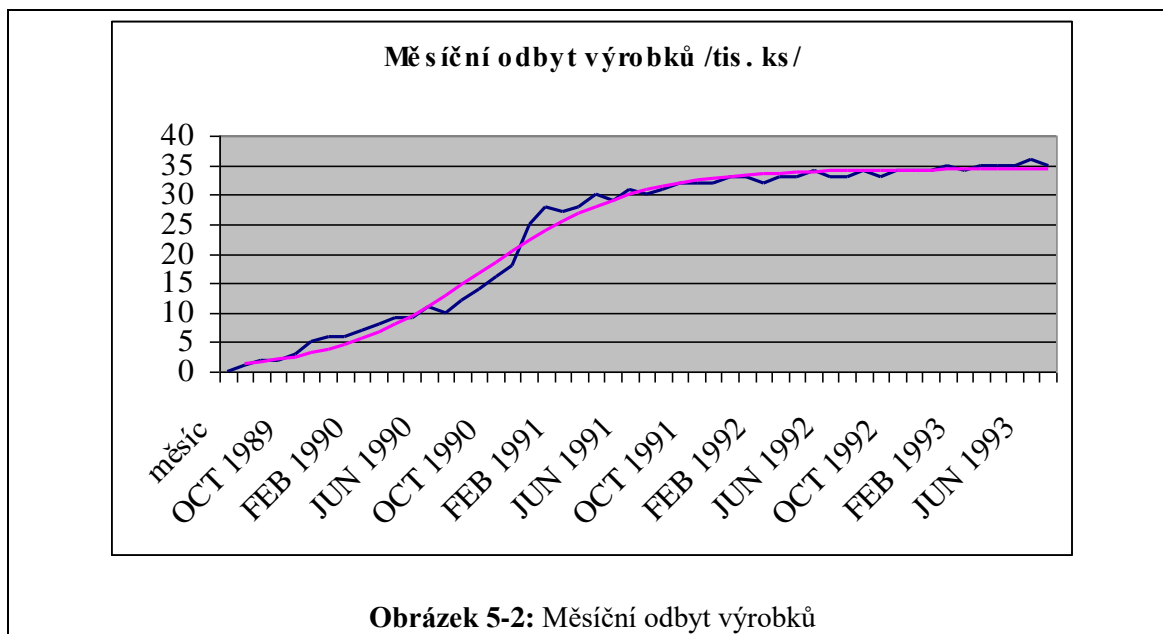
Lineární trend může být vhodným modelem ČR jen v omezeném časovém intervalu. Dříve nebo později se růst nebo pokles ukazatele ČR začne zpomalovat - tlumit. K vyjádření tohoto modelu slouží logistický trend vyjádřený vztahem:

$$T_t = \frac{\kappa}{1 + \beta_0 \beta_1^t}, t = 1, 2, \dots, n.$$

(5-5)

Index  $t$  představuje označení časových intervalů.

**Příklad 2.** Měsíční tržby z prodeje výrobků

Obr. 5-2:  
Měsíční  
odbyt  
výrobků

Výpočet neznámých parametrů  $\kappa, \beta_0, \beta_1$  v definici logistického trendu (5-5) se provede pomocí metody nejmenších čtverců (MNC). V Excelu to lze provést např. pomocí zabudované funkce Řešitel, podobně jako řešení příkladu 4 v této kapitole.

V SPSS je možné k zobrazení a predikci ČR použít menu:

**Analyze → Regression → Curve Estimation...**

s následným výběrem vhodného trendu z 11 nabízených možností.

#### Postup při volbě vhodného analytického modelu trendu:

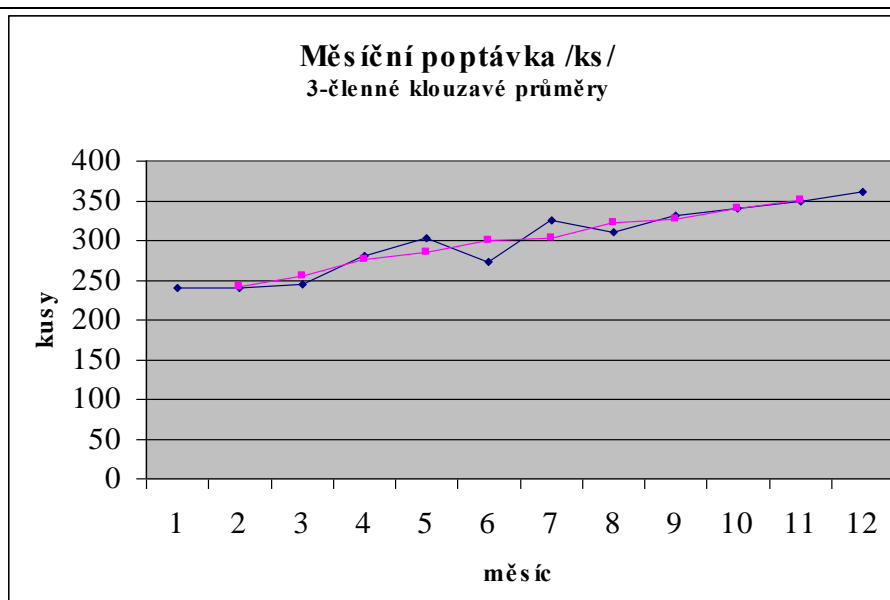
1. Stanovení konkrétního typu trendové funkce.
2. Výpočet přiléhavosti dat k trendové (regresní) křivce pomocí koeficientu determinace  $R^2$  a reziduálního rozptylu  $s_R^2$ .

Nevýhoda koeficientu determinace: více parametrů regresní funkce znamená vyšší hodnotu  $R^2$ , proto je často vhodnější mírou reziduální rozptyl  $s_R^2 = \frac{S_R}{n-p}$ , kde  $n$  je počet dat,  $p$  je počet parametrů,  $S_R$  je reziduální součet čtverců, viz kapitola 4.

**ad (B):** Syntetické modely trendu ČR

**Syntetický přístup stanovení trendu** spočívá ve vyhlazení a vyrovnání odchylek daného ukazatele v časové řadě takovým způsobem, že získané vyrovnané hodnoty vyjadřují trendový faktor obsažený pouze v ČR, nikoliv však vnější faktory. Nemusíte proto znát předem typ trendové funkce, což je předností syntetického přístupu oproti přístupu analytickému, kde jste typ trendové funkce museli stanovit předem. Jeho nevýhodou je naopak obtížnější využití pro prognózování hodnot časové řady. Z existujících metod syntetického přístupu se seznámíte s metodami **klouzavého průměru** a **exponenciální vyrovnání**. Trendy **nejso** zadány explicitně vzorcem, nýbrž jsou zadány hodnotami nové ČR (tj. syntetického trendu).

**Příklad 3.** Prostý 3-členný klouzavý průměr



**Obrázek 5-3:** Měsíční poptávka (ks)

**Volba trendu**

**Syntetické přístupy k trendu**

**Obr. 5-3: Měsíční poptávka**

**Klouzavé průměry** jsou ve skutečnosti posouvané aritmetické průměry několika hodnot ČŘ. Volba rozsahu klouzavé části závisí na věcném (ekonomickém) charakteru časové řady a nelze ji obvykle stanovit na podkladě exaktních statistických metod. V praxi jsou u ekonomických neperiodických časových řad voleny většinou klouzavé části menší liché délky, např. 3, 5 nebo 7 časových jednotek, což vede ke snadnějšímu porozumění výsledkům. Pak totiž můžete hodnotu klouzavého průměru přiřadit prostřednímu časovému okamžiku klouzavé části. U periodických časových řad se volí délka klouzavých částí totožná s délkou periody (sezóny, cyklu).

*Klouzavé  
průměry*

**Exponenciální vyrovnání** představuje posouvané vážené průměry hodnot ČŘ, kdy váhy exponenciálně s časem ubývají. Při něm se nová vyrovnaná hodnota stanoví na základě exponenciálně váženého průměru současné hodnoty a všech předchozích hodnot časové řady. Přitom se používá koeficient **exponenciálního zapomínání**, kdy novější hodnota má vždy větší váhu (tj. důležitost), než hodnota starší.

*Exponen-  
ciální  
vyrovnání*

Nechť  $y_t$  značí pozorovanou hodnotu v časovém okamžiku  $t$ ,  $\alpha$  je koeficient exponenciálního zapomínání  $0 < \alpha < 1$ ,  $\tilde{y}_t$  je vyrovnaná hodnota v čase  $t$ . Procedura exponenciálního vyrovnání začíná tím, že první vyrovnanou hodnotu časové řady  $\tilde{y}_1$  (v čase 1) položíte rovnu pozorované hodnotě  $y_1$ , tedy:  $\tilde{y}_1 = y_1$ .

*Koeficient  
exponen-  
ciálního  
zapomíná-  
ní*

Další vyrovnané hodnoty se definují rekurentním vztahem:

$$\tilde{y}_t = \alpha y_t + (1 - \alpha)\tilde{y}_{t-1}, t = 2, 3, \dots, n,$$

(5-6)

který umožňuje postupně vypočítat všechny vyrovnané hodnoty dané časové řady. Vypočítané vyrovnané hodnoty pak tvoří syntetický trend dané ČŘ.

Výpočet podle vzorce (5-6) je možné snadno provést v Excelu jen za pomoci vzorců bez použití zabudovaných funkcí, alternativně s využitím menu:

*Řešení  
v Excelu*

**Nástroje → Analýza dat → Exponenciální vyrovnání...**

kde po zadání oblasti (sloupce) s uloženými hodnotami ČŘ obdržíme vyrovnanou ČŘ - syntetický trend. (**Pozor:** koeficient zapomínání se tu nazývá koeficient útlumu).

Program SPSS je pro exponenciální vyrovnání vybaven mnohem lépe. V na-bídce menu:

*Řešení  
v SPSS*

**Analyze → Time Series → Exponential smoothing...**

je k dispozici kromě výše popsané jednoduché metody stanovení syntetického trendu (metoda Simple) ještě metoda Holtova. Rozšíření jednoduché metody na sezónní ČŘ poskytuje metoda Wintersova a k dispozici je ještě položka Custom. Touto problematikou se budete zabývat v následující subkapitole.

### 5.3 Analýza sezónní složky

Při analýze časových řad s periodicitou kratší než jeden rok se setkáte téměř vždy s existencí sezónních vlivů, reprezentovaných v modelu časové řady **sezónní složkou**. **Sezónní vlivy** tvoří soubor příčin, které se pravidelně opakují v důsledku koloběhu přírody. Důsledkem působení sezónních vlivů na analyzovanou časovou řadu jsou pravidelné výkyvy nahoru a dolů vůči určitému normálnímu vývoji. Pokud se u časových řad vyskytují podobné vlivy v delším časovém horizontu, hovoříme o **cyklické složce** ČŘ. Souhrnně se sezónní a cyklické složky označují jako **periodické složky** časové řady, takže model časové řady (5-1) můžete vyjádřit ve tvaru:

*Sezónní  
vlivy*

*Analýza  
periodické  
složky*



$$y_t = T_t + P_t + e_t.$$

(5-7)

kde  $P_t$  je periodická složka  $P_t = S_t + C_t$ ,  $S_t$  - sezónní složka,  $C_t$  - cyklická složka. Úkolem popisu periodické složky časové řady je nalézt její vhodné modelové vyjádření, které by umožnilo periodickou (nejčastěji sezónní) složku nejen vhodně identifikovat, ale i následně použít k predikci chování časové řady v budoucnu.

Metody, které se k tomu používají, jsou zejména:

- harmonická analýza
- exponenciální vyrovnání (Wintersův model),
- modely sezónnosti (konstantní, proporcionální).

### PRŮVODCE TEXTEM



Metodou harmonické analýzy a použitím Wintersova sezónního modelu exponenciálního vyrovnání se zde nebudete zabývat, zájemce odkazujeme na doplňkovou literaturu, např. [Seger-Hindls, 1995].



*Dále se však budete zabývat modely konstantní sezónnosti, které je možno řešit v Excelu*

#### 5.3.1 Model konstantní sezónnosti

Při popisu trendové složky i periodické složky v předchozím odstavci jste používali posloupnost časové proměnné  $t = 1, 2, \dots, n$ , nyní se bude toto označení používat pro označení časových intervalů (např. roků), které se člení na dalších  $r$  dílčích časových obdobích, které se nazývají **sezóny** (např. měsíce) a označují se  $j = 1, 2, \dots, r$ . V praktických aplikacích lze vystačit s předpokladem, že  $r$  je sudé, protože např. pro řady s čtvrtletní periodicitou je  $r = 4$ , pro řady s měsíční periodicitou během roku je  $r = 6$ . Model (5-7) lze s použitím uvedené symboliky zapsat ve tvaru:

**Model konstantní sezónnosti**

$$y_{tj} = T_{tj} + P_{tj} + \varepsilon_{tj}, \quad t = 1, 2, \dots, n, \quad j = 1, 2, \dots, r.$$

(5-7\*)

U modelu konstantní sezónnosti se vychází z předpokladu, že:

$$P_{tj} = \gamma_j \quad \text{pro sezónu } j \text{ v letech } t = 1, 2, \dots, n,$$

kde  $\gamma_j$  jsou neznámé sezónní parametry, o nichž se dále předpokládá, že splňují rovnost:

$$\sum_{j=1}^r \gamma_j = 0$$

Předchozí dva poklady vycházejí z představy, že v důsledku pravidelného (ročního) koloběhu sezónních vlivů se v  $j$ -té sezóně opakují sezónní výkyvy  $\gamma_j$ , které se mezi jednotlivými léty neliší a dále se tyto vlivy během roku ( $r$  sezón) vykompenzují, takže jejich roční součet je nulový.

**Podmínka pro sezónní parametry**

**Nejprve budeme předpokládat, že trendová složka  $T_{tj}$  v roku  $t$  nabývá ve všech sezónách hodnotu  $\alpha_t$ , takže posloupnost těchto hodnot v letech  $t = 1, 2, \dots, n$  představuje schodovitý trend. Model pak bude mít tvar:**

$$y_{tj} = \alpha_t + \gamma_j + \varepsilon_{tj}, \quad t = 1, 2, \dots, n, \quad j = 1, 2, \dots, r.$$

(5-8)

Odhady  $a_t, c_j$  tedy  $n + r$  neznámých parametrů  $\alpha_t, \gamma_j$  tohoto modelu obdržíme po aplikaci metody nejmenších čtverců.

### ŘEŠENÝ PŘÍKLAD 5-1



V souboru [madona](#) jsou ve sloupci Poč.Ubyt. uvedena data - tržby v Hotelu Madona v jednotlivých měsících od ledna 1998 až do prosince 2002, celkem tedy 60 hodnot. Úkolem je vytvořit model konstantní sezónnosti a na jeho základě prognózovat hodnoty ubytovaných ve 12 měsících roku 2003.

Tržby  
v Hotelu  
Madona

#### Řešení příkladu

Máme  $n = 5, r = 6$ . Podle vztahů (5-8) vypočítáme odhady neznámých koeficientů modelu konstantní sezónnosti:

Řešení  
v Excelu

$t$	$a_t$	$j$	$c_j$	$y_{tj}$
1	403,8	1	-29,8	285,3
2	353,0	2	-125,2	189,9
3	337,9	3	-86,2	228,9
4	323,0	4	42,9	357,9
5	319,1	5	75,1	390,1
6	315,0	6	132,5	447,5
		7	153,7	468,7
		8	127,9	442,9
		9	-19,8	295,3
		10	-112,8	202,3
		11	-164,6	150,5
		12	6,3	321,3

Tab. 5.1

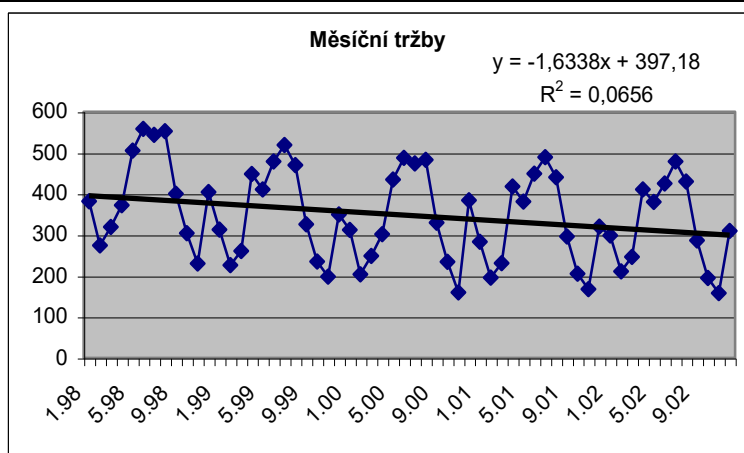
V Tab. 5.1. jsou hodnoty  $a_t$  pro  $t = 1, 2, 3, 4, 5$  vypočítány podle prvního vztahu ve (5-9), pro  $t = 6$  je hodnota stanovena odhadem jako prognóza průměrů na rok 2003 z předchozích hodnot. Hodnoty  $c_j$  pro  $j = 1, 2, 3, \dots, 12$  jsou vypočítány podle druhého vztahu (5-9). Prognózované hodnoty tržeb v jednotlivých měsících roku 2003 jsou vypočítány v posledním sloupci Tab. 5.1. ze vztahu (5-8).

V příloženém souboru [madona](#) je kromě sloupce dat - časové řady tržeb v jednotlivých měsících - uvedeno také **Řešení**: celkový průměr (zelená barva), odhady koeficientů modelu:  $\alpha_t$  (žlutá barva), koeficienty  $\gamma_j$  (modrá barva).

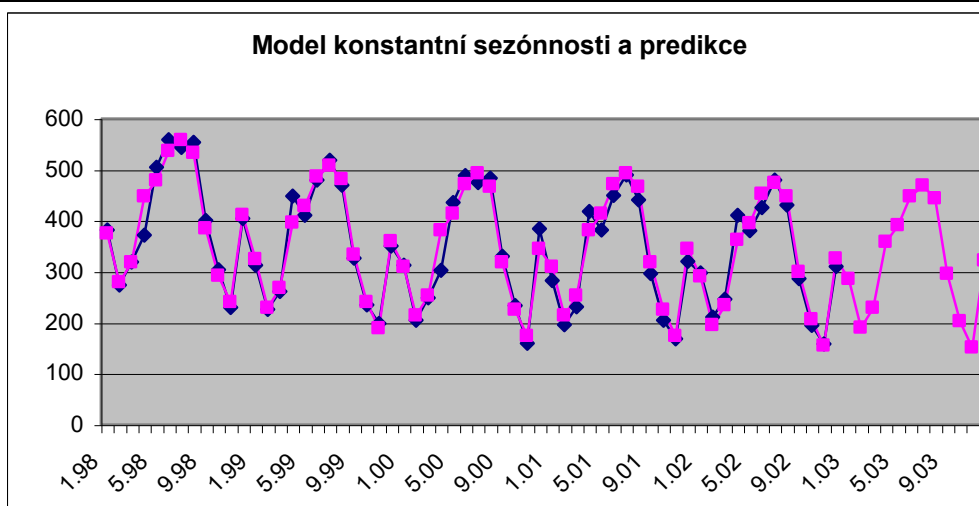
Na Obr. 5-4. je zobrazen graf časové řady a na Obr. 5-5. společně tentýž graf a graf modelu konstantní sezónnosti (fialová barva) včetně prognózované části na rok 2003.

$$a_t = \frac{1}{r} \sum_{j=1}^r y_{tj} = \bar{y}_t, \quad c_j = \frac{1}{n} \sum_{t=1}^n y_{tj} - \frac{1}{m} \sum_{t=1}^n \sum_{j=1}^r y_{tj}.$$

(5-9)

Obr. 5-4:  
Měsíční  
tržby

Obrázek 5-4: Měsíční tržby

Obr. 5-5:  
Model  
konstantní  
sezónnosti  
a predikce

Obrázek 5-5: Model konstantní sezónnosti a predikce

\*

## 5.4 Analýza náhodné složky

Náhodnou složku  $\varepsilon_t$  lze z modelu (5-7) vyjádřit ve tvaru:

$$\varepsilon_t = y_t - Y_t, \quad t = 1, 2, \dots, n,$$

(5-10)

kde  $Y_t = T_t + P_t$ . Jedná se zde o vyjádření blíže nespécifikovaných náhodných vlivů. Zdrojem této složky jsou obvykle nepodchycené drobné vzájemně nezávislé náhodné vlivy. Realizace náhodné složky pro konkrétní časovou řadu, trend a sezónní složku nazývá reziduum. K zajištění spolehlivé předpovědi na základě modelu časové řady je třeba mít splněny některé předpoklady

o náhodné složce. Jedná se tedy o tyto předpoklady:

1. Náhodné složky  $\varepsilon_t$  v modelu (5-10) mají nulovou střední hodnotou a normální rozdělení pravděpodobnosti.
2. Náhodné složky  $\varepsilon_t$  v modelu (5-10) mají konstantní rozptyl  $\sigma^2$ . Konstantnost rozptylu nazýváme **homoskedasticita**.
3. Náhodné složky  $\varepsilon_t$  jsou **nekorelovatelné**, tj.  $r(\varepsilon_t, \varepsilon_{t'}) = 0$  pro každé  $t \neq t'$ ,  $t, t' = 1, 2, \dots, n$ . ( $r$  značí korelační koeficient, viz kapitola 5).

V praxi jsou podmínky klasického modelu často splněny. Nejste-li si však jejich platností jisti, můžete provést testy příslušné testy hypotéz.

## 5.5 Testování vlastností náhodné složky

K získání dobré prognózy na základě sestaveného modelu je nutné prověřit, zda náhodná složka – rezidua, vyhovuje výše uvedeným 3 předpokladům:

1. nulová střední hodnota rezidua a normalita, 2. homoskedasticita, 3. nekorelovanost

**ad 1.** Jednoduchým nástrojem, kterým lze ověřit nulovou střední hodnotu reziduí, je **jednovýběrový  $t$ -test** střední hodnoty, viz kapitola 2. Použití tohoto testu si procvičíte níže na Příkladu 4. Normalitu lze testovat Chi-kvadrát testem, viz kapitolu 2, ale to se prakticky neprovádí.

**ad 2.** Konstantnost rozptylu lze testovat s pomocí Excelu jednoduše tak, že ČŘ rozdělíme na 2 poloviny (začátek a konec) a provedeme dvouvýběrový  $F$ -test pro rozptyl, který je k dispozici v nabídce:

**Nástroje → Analýza dat → Dvouvýběrový  $F$ -test pro rozptyl...**

Použití si procvičíte níže na Příkladu 4.

**ad 3.** K testování nekorelovanosti slouží známý Durbin - Watsonův test autokorelace. V Excelu se bohužel žádný vhodný nástroj nevyskytuje, proto využijete SPSS, kde takové nástroje jsou k dispozici. Použití demonstrujeme na Příkladu 4. Budete k tomu potřebovat pojem autokorelace.

S pojmem korelace dvou náhodných veličin jste se seznámili již dříve. Každou ČŘ lze považovat za posloupnost realizací náhodné veličiny, pak odhad korelačního koeficientu dvou časových řad můžete vypočítat podle vztahu (6-7). Představte si 2 časové řady: první ČŘ má hodnoty  $y_1, y_2, y_3, y_4, \dots, y_n, \dots$  druhá ČŘ vznikne posunutím o 1 časovou jednotku, tedy  $y_2, y_3, y_4, \dots, y_n, \dots$ . Korelační koeficient těchto 2 časových řad se nazývá autokorelační koeficient 1. řádu. Posunete-li druhou časovou řadu o  $k = 2, 3, \dots$  atd. časových jednotek, obdržíte analogicky autokorelační koeficient 2. řádu, 3. řádu, atd., obecně **autokorelační koeficient  $k$ -tého řádu**, označuje se  $\rho(k)$ .

Posloupnost autokorelačních koeficientů  $\rho(0), \rho(1), \rho(2), \rho(3), \dots$  se nazývá **autokorelační funkce** (zkráceně z angličtiny **ACF**). První hodnota ACF je vždy rovna 1, neboť se jedná o korelační koeficient 2 totožných ČŘ (posun 0), tedy  $\rho(0) = 1$ . Ostatní hodnoty ACF jsou čísla z intervalu  $[-1; 1]$ . Pokud hodnota ACF padne do 95%-ního intervalu spolehlivosti, považujeme ji za nulovou. Padnou-li proto všechny hodnoty ACF do pásu 95%-ního intervalu spolehlivosti (ovšem s výjimkou  $\rho(0)$ , která je vždy rovna 1), jsou původní hodnoty ČŘ vzájemně nekorelované. V případě, že je tato ČŘ řadou reziduí, splňuje podmínku 3.

Použití ACF si procvičíte na Příkladu 5-1.

*Vlastnosti náhodné složky*

*Test střední hodnoty*

*Test homoskedasticity*

*Test nekorelovanosti*

*Autokorelační funkce*

## Řešení příkladu

Časová řada reziduí je uvedena ve sloupci Reziduum v souboru [madona](#).

**ad 1.** Nejprve budete testovat nulovost reziduí, tj. nulovou hypotézu

$H_0$ : Střední hodnota reziduí je 0.

Jak již bylo řečeno, v Excelu použijete jednovýběrový  $t$ -test střední hodnoty, podrobnosti viz kapitola 5.4.

Zjistíte, že  $p$ -hodnota testu je  $0,324 > 0,05$  a proto nulovou hypotézu přijímáme a tedy je splněna podmínka 1.

**ad 2.** K testování heteroskedasticity použijete opět Excel. Sloupec s ČŘ Reziduum rozdělíte "na dvě poloviny" - dvě ČŘ, každá s 30 hodnotami.

Nulovou hypotézu formulujete takto:

$H_0$ : Rozptyly obou ČŘ jsou stejné.

K testování použijete menu:

**Nástroje** → **Analýza dat** → **Dvouvýběrový F-test pro rozptyl...**,

kde po zadání oblastí umístění obou ČŘ obdržíte tento výstup:

Dvouvýběrový F-test pro rozptyl		
	Soubor 1	Soubor 2
Stř. hodnota	-1,56	-4,40667
Rozptyl	646,4026	452,6232
Pozorování	30	30
Rozdíl	29	29
F	1,428125	
P(F<=f) (1)	0,171355	
F krit (1)	1,860812	

**Pokračování řešení příkladu 5-1**

**Obr. 5-6: Dvouvýběrový f-test pro rozptyl**

**Obrázek 5-6:** Dvouvýběrový f-test pro rozptyl

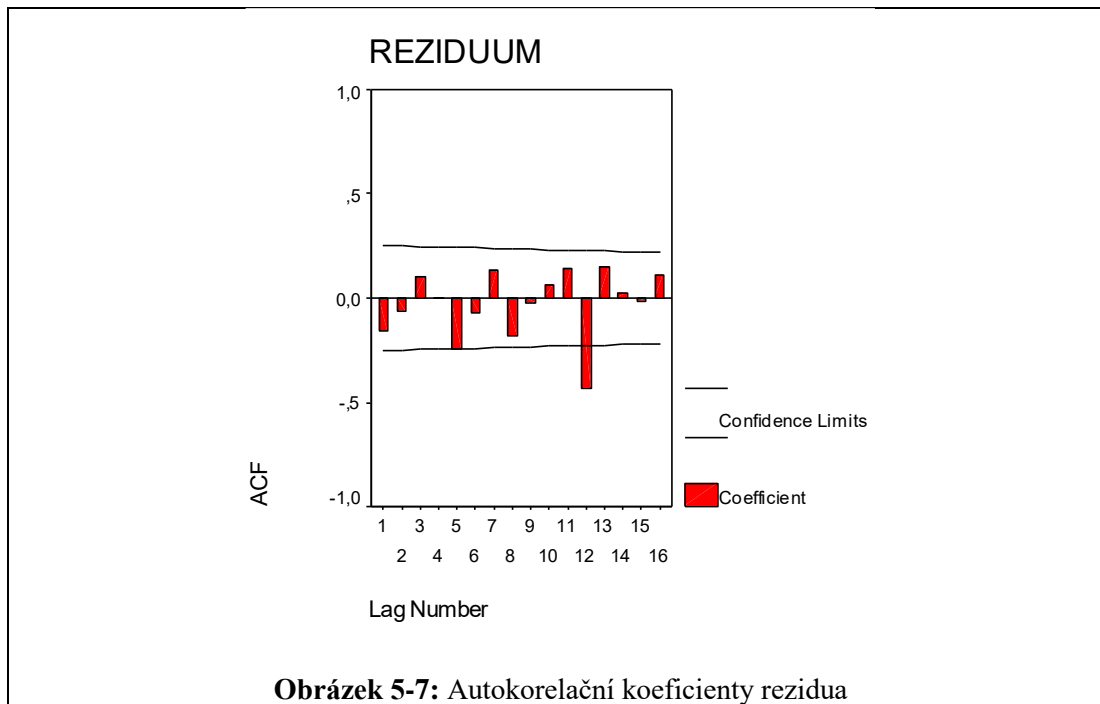
$P$ -hodnota testu je uvedena v předposledním řádku tabulky jako 0,171. Tato hodnota je větší než porovnatelná hladina významnosti 0,05 a proto nulovou hypotézu přijímáme, rozptyly se tedy statisticky významně neliší - jsou stejné. Předpoklad 2. je tedy rovněž splněn.

**ad 3.** K testu nekorelovanosti reziduí použijete pro soubor [madona](#) v Excelu menu:

**Nástroje** → **Analýza dat** → **Korelace...**

Vypočítáte jednotlivé autokorelační koeficienty a použijete testu jejich statistické významnosti z kapitoly 4.1. Na Obr. 5-6., jsou uvedeny sloupečky autokorelačních koeficientů všech (zobrazených) řádů, s výjimkou 6., leží v pásu 95%-ního intervalu spolehlivosti kolem 0, tedy jsou statisticky nevýznamné - nulové. Nenulovost 6. hodnoty souvisí s tím, že původní řada vykazuje sezónní charakter právě s periodou sezónnosti 12 měsíců. Z tohoto důvodu můžete vypustit tuto hodnotu z dalších úvah a konstatovat, že hodnoty reziduí jsou vzájemně nekorelované a splňují proto předpoklad 3.

Vzhledem k tomu, že všechny 3 předpoklady o náhodné (reziduální) složce jsou splněny, je vytvořený model ČŘ s konstantní sezónností a schodovým trendem statisticky validní a prognózované hodnoty jsou použitelné.



Obr. 5-7: Autokorelační koeficienty rezidua

\*

## 5.6 Prognózování v ČŘ

Prognózování v ČŘ se někdy nazývá predikování, předpovídání, předvídaní, extrapolace, věštění apod. Mezi prognostickými metodami hrají významnou roli statistické prognostické metody. Do této skupiny patří také metody používající při konstrukci prognóz extrapolaci časových řad využívající regresní analýzy. Podstata extrapoláčnických metod spočívá ve studiu minulosti prognózovaného jevu a v přenosu zákonitostí vývoje z minulosti a přítomnosti do budoucnosti.

Prognózování

Tyto metody jsou konstruovány na základě předpokladu relativní stability již existujících tendencí vývoje zkoumaného jevu. Latinsky se tento předpoklad nazývá *ceteris paribus*, což znamená za jinak stejných okolností shodně s vývojem minulým. U procesů, které jsou v čase stabilní, lze tento princip s úspěchem použít. Naopak v případě, kdy během prognózovaného období probíhají podstatné kvalitativní změny, je použití extrapoláčnických modelů problematické.

Ceteris paribus

Uvažujme ČŘ  $y_t = T_t + S_t + \varepsilon_t$ , (přitom  $t$  je současný okamžik). Bodový odhad  $\tilde{y}_{t+h}$  neznámé veličiny ČŘ v čase  $t + h$ , kde  $h$  je zadaný horizont **bodové prognózy** (předpovědi,...) lze stanovit takto:

Bodová prognóza

$$\tilde{y}_{t+h} = T_{t+h} + S_{t+h},$$

kde  $T$  - analytická trendová funkce,  $S$  - sezónní funkce (podle typu modelu),  $\varepsilon = 0$ , tj. náhodnou složku pokládáme rovnu její střední hodnotě.

Bodová předpověď umožňuje pomocí jednoho čísla stanovit hodnotu předvídané veličiny. Intervaly spolehlivosti (konfidenční intervaly) dovolují stanovit příslušné intervalové předpovědi.

**Intervalová prognóza** vytvořená v čase  $n$  na časový horizont  $i$  je definována jako oboustranný interval spolehlivosti:

Intervalová prognóza

$$[ \tilde{y}_{n+i} - t_{1-\alpha/2}(n-2) s_R \sqrt{Q_n(i)}, \tilde{y}_{n+i} + t_{1-\alpha/2}(n-2) s_R \sqrt{Q_n(i)} ],$$

(5-11)

kde  $s_R^2$  je reziduální rozptyl definovaný vztahem:

$$s_R^2 = \frac{1}{n-p} \sum_{t=1}^n (y_t - \tilde{y}_t)^2,$$

$p$  je počet odhadnutých parametrů modelu a

$$Q_n(i) = \sqrt{(1-R^2) \frac{n(n^2-1)+12i^2}{(n^2-1)(n-2)}},$$

přičemž  $R^2$  je koeficient determinace.

## ŘEŠENÝ PŘÍKLAD 5-2



V souboru [sabena](#) je v listu Data uložena ČR tržeb za jednotlivá čtvrtletí od roku 1995 do konce 2. čtvrtletí 2003.

Sestrojte vhodný model uvedené ČR a na jeho základě prognózuje tržby ve III. a IV. čtvrtletí 2003. Stanovte bodovou a intervalovou prognózu s 95% spolehlivostí.

Čtvrtletní  
tržby le-  
tecké  
společnos-  
ti SABENA

### Řešení příkladu

Z věcné podstaty problému a zejména z průběhu časové řady zobrazené graficky na Obr. 5-7 lze usuzovat na sezónní charakter ČR se zvětšujícími se sezónními vlivy v souvislosti se zvyšujícími se tržbami. Z tohoto důvodu je vhodné uvažovat nikoliv aditivní, ale multiplikativní typ modelu, v němž se trendová a sezónní složka vzájemně násobí, tedy:

$$y_t = T_t \cdot S_t + \varepsilon_t$$

Dále je zřejmé, že se jedná o lineární průběh trendu pro roky  $t = 1, 2, \dots, 8$  a v rámci každého roku jsou 4 sezóny (čtvrtletí) označované  $j = 1, 2, 3, 4$ . Model lze proto upřesnit takto:

$$y_{tj} = T_t \cdot S_j + \varepsilon_{tj}, \quad t = 1, 2, \dots, 8, j = 1, 2, 3, 4,$$

(5-12)

přítom je trend lineární funkce (tj. přímka), tedy:

$$T_t = \alpha t + \beta, \quad t = 1, 2, \dots, 8,$$

a  $S_j$  představují sezónní koeficienty, které jsou ve všech letech stejné - konstantní. Výsledný multiplikativní model lze upravit na konečný tvar:

$$y_{tj} = (t + \gamma) \cdot s_j + \varepsilon_{tj}, \quad t = 1, 2, \dots, 8, j = 1, 2, 3, 4,$$

(5-13)

Přítom byla použita substituce koeficientů:  $\gamma = \frac{\beta}{\alpha}$ ,  $s_j = \alpha S_j$ , čímž se snížil počet neznámých koeficientů o jeden. V modelu je 5 neznámých koeficientů:  $\gamma, s_1, s_2, s_3, s_4$ . Neznámé koeficienty se stanoví metodou nejmenších čtverců tak, aby se minimalizoval součet kvadrátů odchylek:  $\sum_{t,j} \varepsilon_{tj}^2$ , tj. po dosažení ze vztahu (5-13)

$$S = \sum_{t,j} (y_{tj} - (t + \gamma) s_j)^2$$

(5-14)

K tomu účelu využijete z Excelu vestavěnou funkci **Řešitel**. Ještě předtím je zapotřebí připravit tzv. měněné buňky pro 5 neznámých koeficientů. Souboru Sabena.xls jsou tyto buňky v listu Sabena označeny žlutou barvou. Pomocí vzorců podle (5-14) vytvoříme sloupec  $T.S$ , vypočítáme pak reziduální součet čtverců (buňka je označena modrou barvou). Tuto buňku určíme v Řešiteli pro Nastavit buňku a zvolíme tlačítko Min. Po stisknutí tlačítka Řešit a následném potvrzení se ve žlutých buňkách zobrazí vypočítané odhady neznámých parametrů, viz list Sabena v souboru Sabena.xls. Konkrétní hodnoty parametrů jsou:

**Řešitel**

gamma=	1,71
s1=	162,38
s2=	163,13
s3=	166,23
s4=	180,41

Bodové odhady predikce  $\tilde{y}_{tj}$  v posledních 2 kvartálech ( $j = 3,4$ ) roku 2003 ( $t = 9$ ) vypočteme jednoduše dosazením do vztahu:

$$\tilde{y}_{tj} = (t + \gamma) \cdot s_j.$$

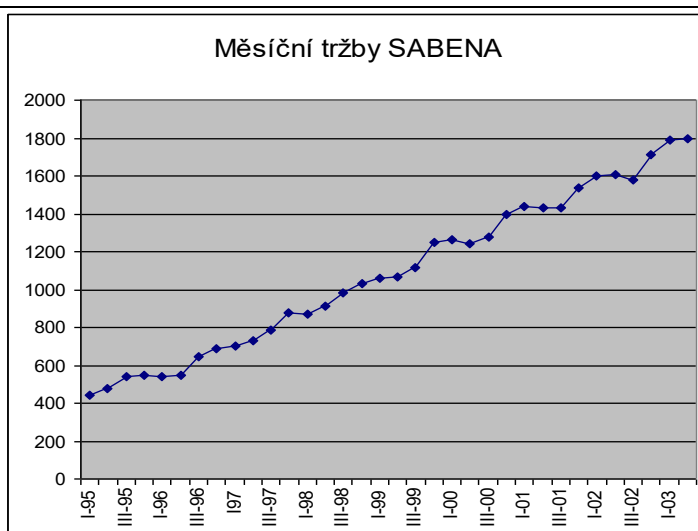
Bodové predikce jsou následující, viz Obr. 5-8.:

III-03		1781
IV-03		1932

Intervalové odhady vypočítáme podle vztahu (5-11), jsou to ve skutečnosti intervaly "bodový odhad plus/mínus kousek". V našem případě pokládáme ve vztahu (5-11):  $n = 8.4 = 32$ .

95%-ní interval spolehlivosti pro 3. čtvrtletí 2003 je interval [1752;1809].

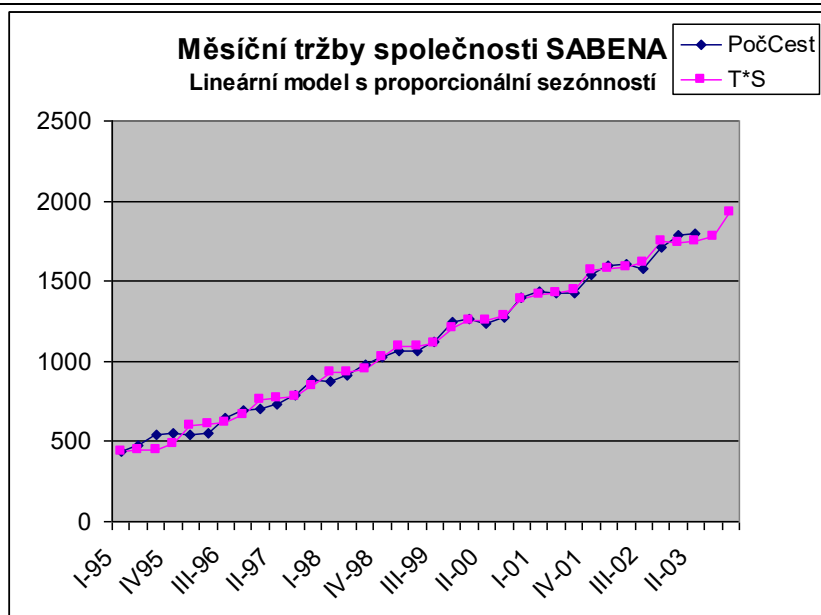
95%-ní interval spolehlivosti pro 4. čtvrtletí 2003 je interval [1904;1961].



**Obr. 5-8:**  
Měsíční  
tržby  
SABENA

**Obrázek 5-8:** Měsíční tržby SABENA





Obr. 5-9:  
Měsíční  
tržby spo-  
lečnosti  
SABENA

Obrázek 5-9: Měsíční tržby společnosti SABENA

\*

## 5.7 Kauzální prognostické metody

Kauzální prognostické metody (KPM) ve srovnání s dekompozičními metodami ČR využívají k prognózám tržeb kromě faktoru času ještě další faktory (ekonomické, sociální aj.). Těmito složitějšími metodami se zde nebudete zabývat, částečně jste se s nimi seznámili v příkladech předchozí kapitoly pojednávající o vícerozměrné regresní analýze.

KPM

### Příklady.

1. Tržby zmrzlinového stánku jsou funkcí více proměnných:

tržby =  $f(\text{denní doba, počasí, druh zmrzliny, počet kolemjdoucích})$

2. Tržby prodejny nábytku jsou funkcí více proměnných :

tržby =  $f(\text{čas, průměrný příjem rodiny, demografická struktura regionu, výdaje na reklamu})$ .

### SAMOSTATNÝ ÚKOL 7

Uvažujte časovou řadu měsíčních prodejů oděvů firmy Boutique za 17 let (tj. 204 měsíců od ledna 1987 do prosince 2003), data viz [Oděvy](#).

Prognózuje měsíční prodeje oděvů firmy na 12 měsíců roku 2004. Stanovte bodové a 95% intervalové odhady tržeb. Využijte lineární model proporcionalní sezónnosti s pomocí Řešitele v Excelu.



Lineární  
model  
proporcio-  
nální se-  
zónnosti

**SHRnutí KAPITOLY METODY PROGNOZOVÁNÍ TRŽEB****Shrnutí kapitoly**

V této kapitole jste studovali problematiku časových řad zejména časových řad tržeb a možnosti jejich prognózování. Nejprve jste vyšetřovali trendovou složku ČŘ, dále sezónní složku a objasnili si význam náhodné složky v časové řadě a předpoklady, kterým by tato složka měla vyhovovat, aby byl model validní pro prognózování. Byly uvedeny základní přístupy a metody prognózování v ČŘ, které byly ilustrovány na příkladech a řešeny pomocí Excelu a SPSS.

**KONTROLNÍ TEST 5**

- 5.1** Deterministická složka časové řady představuje (označte správnou odpověď, může jich být i více):
- trendovou složku
  - trendovou a sezónní složku
  - trendovou, sezónní a cyklickou složku
  - sezónní a cyklickou složku
- 5.2** Periodická složka ČŘ je složena z:
- sezónní složky
  - trendové a sezónní složky
  - trendové, sezónní a cyklické složky
  - sezónní a cyklické složky
- 5.3** Při použití Excelu, Graf, přidat spojnicí trendu je předvolen typ trendu
- lineární,
  - logaritmický
  - exponenciální
  - žádný
- 5.4** Ke každému pojmu z levého sloupce přiřaďte správné charakteristiky z pravého sloupce vztahující se k použití měř asociace:
- |                              |     |  |
|------------------------------|-----|--|
| (1) Aditivní model ČŘ        | (A) | Jednotlivé složky se mezi sebou sečítají |
| (2) Multiplikativní model ČŘ | (B) | Trend tvoří lineární funkce              |
| (3) Lineární model trendu ČŘ | (C) | Jednotlivé složky se mezi sebou násobí   |
- 5.5** Určete, zda následující výroky jsou správné (vepište A) nebo nesprávné (vepište N):
- Mezi nejdůležitější trendové funkce patří logistická funkce.
  - Exponenciální vyrovnání patří mezi analytické metody stanovení trendu.
  - Klouzavé průměry slouží k nalezení syntetického trendu.

- d. Intervalová prognóza je definována jako jednostranný interval spolehlivosti.
- e. Ceteris paribus znamená „Za nezměněných podmínek“.

**5.6** Doplňte správný výraz:

- a. Nevýhoda koeficientu determinace: více parametrů regresní funkce znamená \_\_\_\_\_ hodnotu  $R^2$ .
- b. Konstantnost rozptylu v modelu ČŘ nazýváme \_\_\_\_\_.
- c. Náhodné složky  $\varepsilon$  v modelu ČŘ jsou \_\_\_\_\_.
- d. V intervalovém odhadu prognózy tržeb na základě modelu ČŘ se používá \_\_\_\_\_ rozptyl.

**5.7** Doplňte správné sousloví:

- a. Použitím koeficientu \_\_\_\_\_ zapomínání má novější hodnota vždy \_\_\_\_\_ váhu (tj. důležitost), než hodnota starší.
- b. V důsledku pravidelného (ročního) koloběhu sezónních vlivů se v každé sezóně opakují \_\_\_\_\_ výkyvy, které se mezi jednotlivými léty \_\_\_\_\_.
- c. Konstantnost rozptylu lze testovat s pomocí Excelu tak, že ČŘ rozdělíme na 2 poloviny a provedeme \_\_\_\_\_ pro rozptyl, který je k dispozici v nabídce:  
Nástroje → Analýza dat → \_\_\_\_\_.
- d. Kauzální prognostické metody ve srovnání s dekompozičními metodami ČŘ využívají k prognózám tržeb kromě faktoru času ještě \_\_\_\_\_.

## ŘEŠENÍ SAMOSTATNÉHO ÚKOLU



### Samostatný úkol 7

Sig. 0,539

-výstup z Excelu.

Nulovou hypotézu přijímáme. Průměrný počet limonád vypitých denně je roven 2.

Samostatný úkol 7

## ŘEŠENÍ KONTROLNÍHO TESTU 5



**5.1** c.

**5.2** d.

**5.3** a.

**5.4** (1) – (A), (2) – (C), (3) – (B)

**5.5** A, N, A, N, A

**5.6** a. vyšší, b. homoskedasticita, c. nekorelované, d. reziduální

**5.7** a. exponenciálního - větší, b. dvouvýběrový F-test, c. Dvouvýběrový F-test pro rozptyl..., d. další faktory

## 6 ANALÝZA ROZPTYLU

### CÍLE KAPITOLY ANALÝZA ROZPTYLU

#### Po úspěšném a aktivním absolvování této KAPITOLY

##### Budete umět:

- význam jednotlivých dílčích rozptylů studovaného kvantitativního statistického znaku,
- předpoklady jednofaktorové ANOVA,
- 3 kroky algoritmu výpočtu jednofaktorové ANOVA,
- míry těsnosti a závislosti mezi faktory,
- rozdíl mezi jednofaktorovou a vícefaktorovou ANOVA.

*Budete umět*

##### Budete schopni:

- objasnit význam jednotlivých dílčích rozptylů studovaného kvantitativního statistického znaku,
- vysvětlit předpoklady jednofaktorové ANOVA,
- vyjmenovat 3 kroky algoritmu výpočtu jednofaktorové ANOVA,
- aplikovat míry těsnosti a závislosti,
- uplatnit rozdíl mezi jednofaktorovou a vícefaktorovou ANOVA.

*Budete schopni*

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny.

#### PRŮVODCE STUDIEM 7



V kapitole 3 jste se mimo jiné věnovali nezávislosti dvou statistických znaků, konkrétně jste tento fakt nezávislosti testovali neparametrickým statistickým testem (Chi-kvadrát test nezávislosti). Rozšíření této závažné problematiky na více než 2 výběry je obsahem této kapitoly, která se zabývá tzv. analýzou rozptylu - ANOVA a patří k nejčastěji používaným metodám v marketingovém výzkumu. Analýza rozptylu umožňuje ověřit významnost rozdílu mezi výběrovými průměry dvou nebo více náhodných výběrů. Metoda umožňuje posoudit vliv různých kategorií nějakého kvalitativního znaku (faktoru) na proces charakterizovaný kvantitativním statistickým znakem. Například dovoluje hodnotit účinky různých reklamních kampaní. Základní myšlenka analýzy rozptylu spočívá v rozkladu celkového rozptylu kvantitativního znaku na dílčí rozptyly příslušející jednotlivým vlivům, podle nichž jsou data roztržena. Kromě dílčích rozptylů je složkou celkového rozptylu také reziduální rozptyl, způsobený nepostiženými vlivy. Podle počtu analyzovaných faktorů rozlišujeme jednofaktorovou, dvoufaktorovou a více-faktorovou analýzu rozptylu. Seznámíte se především s jedno-faktorovou a dvoufaktorovou ANOVA. Všeobecně používané označení ANOVA je akronymem anglických slov „ANalysis Of Variance“ (doslovný překlad: analýza rozptylu). Projděte podrobně řešené příklady v textu a potom zkuste samostatně vyřešit 2 neřešené příklady, dosažené výsledky zkontrolujte se správnými výsledky ze závěru kapitoly. Pokud vás výsledek neuspokojí, vraťte se nazpět k řešeným příkladům a postup opakujte.

## 6.1 Jednofaktorová ANOVA

Často se vyskytuje situace, kdy máte  $k$  (např.  $k=5$ ) nezávislých náhodných výběrů, které obecně nemusí pocházet z jednoho základního souboru, s příslušnými rozsahy  $n_1, n_2, \dots, n_k$ . Přitom  $k$  může být 2, 3, ... Tyto rozsahy výběrů rovněž nemusí být stejné, v každém z nich je znám průměr  $\bar{x}_i$ , a také rozptyl  $s_i^2$ ,  $i = 1, 2, \dots, k$ . V praktických situacích obvykle tyto výběry vzniknou tak, že základní soubor rozdělíme podle určitého obvykle kvalitativního - nečíselného třídícího statistického znaku  $X$  do  $k$  skupin, v každé z nich pak vybíráme samostatně  $n_i$  prvků.

Více nezávislých náhodných výběrů

Znak  $X$  se pak označuje jako **faktor**, jehož hodnoty jsou předem stanoveny a hovoří se proto často o **faktoru kontrolovaném**, nebo **faktoru pozorovaném**, např. věková skupina, druh výrobku, typ reklamy, typ služby apod. Hodnoty faktoru  $X$  se označují  $x_1, x_2, \dots, x_k$

Faktor kontrolovaný a pozorovaný

Faktor  $X$  má  $k$  úrovní – kategorií a ovlivňuje statistický znak  $Y$ , jež má **kvantitativní** - intervalovou nebo podílovou (tedy číselnou) povahu. Hodnoty znaku  $Y$  příslušné hodnotě  $x_i$  faktoru  $X$  označujeme  $y_{i1}, y_{i2}, \dots, y_{in_i}$

Pro analýzu rozptylu je výhodné uspořádat výchozí údaje do přehledné tabulky, viz Tab.6.1.

Schéma - tabulka

Číslo výběru	Zjištěné hodnoty sledovaného znaku	Počet prvků	Průměr	Rozptyl
1	$y_{11}, y_{12}, \dots, y_{1j}, \dots, y_{1n_1}$	$n_1$	$\bar{y}_1$	$s_1^2$
2	$y_{21}, y_{22}, \dots, y_{2j}, \dots, y_{2n_2}$	$n_2$	$\bar{y}_2$	$s_2^2$
⋮	⋮	⋮	⋮	⋮
$i$	$y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{in_i}$	$n_i$	$\bar{y}_i$	$s_i^2$
⋮	⋮	⋮	⋮	⋮
$k$	$y_{k1}, y_{k2}, \dots, y_{kj}, \dots, y_{kn_k}$	$n_k$	$\bar{y}_k$	$s_k^2$
Celkem		$N$	$\bar{y}$	$s^2$

Tab. 6.1. Schéma výchozí tabulky analýzy rozptylu pro jeden faktor

Metoda analýzy rozptylu ANOVA spočívá v tom, že se celková variabilita měřena součtem čtverců odchylek od celkového průměru rozdělí na variabilitu uvnitř jednotlivých výběrů a na variabilitu mezi jednotlivými výběry. V následujícím odstavci si tento postup ujasníte.

### ŘEŠENÝ PŘÍKLAD 6-1



Ve Studii (viz kapitola 2.1) uvažujte znak  $X$  (tj. faktor) = Věk (18 až 22 let, počet kategorií -roků je tedy  $k = 5$ , počet takových respondentů - studentů je 142). Jako kvantitativní znak  $Y$  se označí Počet lahví/týden.

**Problém:** Má věk (= stáří) respondentů (= studentů) vliv na konzumaci limonád?

## Řešení příkladu

Nulová hypotéza se v tomto případě formuluje takto:

$H_0$ : Věk respondentů (spotřebitelů) nemá na konzumaci limonád vliv.

Jinak řečeno: skupiny spotřebitelů vytvořené podle věku pocházejí z jediné populace.

Definuje se alternativní hypotéza:  $H_1$ : „opak  $H_0$ “

Všimněte si, že nulovou hypotézu formulujeme jako "nulová změna". To je totiž obvyklý princip formulování nulové hypotézy: Nic se neděje, žádná změna, "nula" - proto pojmenování "nulová hypotéza".

Idea postupu řešení: Čím větší variabilita (tj. rozptyl) mezi skupinami spotřebitelů limonád, tím větší vliv faktoru.

Metoda ANOVA je svojí podstatou testem hypotézy, viz kapitola 1 a další. Musí se proto nejprve formulovat nulová a alternativní hypotéza, pak testové kritérium a nakonec kritický obor testu.

**Test  
hypotézy**

Předpokladem je, že faktor  $X$  má  $k$  úrovně (kategorií = hodnot  $x_i$ ), s účinkem na znak  $Y$ , který lze vyjádřit vztahem:

**Základní  
předpoklad**

$$\mu_i = \mu + \alpha_i, i = 1, 2, \dots, k,$$

kde  $\mu_i$  je průměr znaku  $Y$  v  $i$ -té skupině (příslušné k hodnotě faktoru  $x_i$ ),

**Efekt  
hodnoty  
faktoru**

$\mu$  je celkový průměr znaku  $Y$ ,

$\alpha_i$  je efekt hodnoty faktoru  $x_i$  na znak  $Y$ .

Přeformulujte nyní nulovou hypotézu tak, že všechny výběry pocházejí ze stejné základní populace (základního souboru), jinak řečeno, že hodnoty faktoru  $x_i$  nemají na hodnoty znaku  $Y$  žádný efekt (vliv). Bude se tedy v nulové hypotéze předpokládat, že  $\alpha_i$  pocházejí z **normálně rozdělené populace s nulovou střední hodnotou a konstantním rozptylem**  $\sigma^2$ , tedy formuluje se nulová hypotéza takto:

**Reformu-  
lace  
nulové  
hypotézy**

$$H_0 : E(\alpha_1) = E(\alpha_2) = \dots = E(\alpha_k) = 0,$$

proti alternativní hypotéze, že  $H_0$  neplatí, tudíž alespoň pro dvě hodnoty, např.  $i$  a  $j$ , platí:

$$H_1 : E(\alpha_i) \neq E(\alpha_j).$$

**Alternativ-  
ní  
hypotéza**

Symbolem  $E(\alpha_i)$  se tu označuje **střední hodnota** náhodné veličiny  $\alpha_i$ . Předpoklad konstantního rozptylu pro všechny veličiny  $\alpha_i$  je podstatný pro platnost testu a je ho možno ověřit jiným statistickým testem, např. t-testem z kapitoly 1. Normalitu rozdělení veličin  $\alpha_i$  lze taktéž testovat příslušným testem, např. Chi-kvadrát testem dobré shody, s nímž jste se seznámili v kapitole 2. V praxi obvykle předpokládáme (na podkladě věcné znalosti dat), že zmíněné dva předpoklady jsou automaticky splněny a při konkrétní aplikaci ANOVA tyto předpoklady již neověřujeme.

Cílem, k němuž směřujeme, je přijmout nulovou hypotézu  $H_0$ , eventuálně  $H_0$  zamítnout (na zvolené hladině významnosti).

\*

## 6.2 Postup při analýze rozptylu s jedním faktorem

Celkovou variabilitu znaku (souboru) se měří obvykle výběrovým rozptylem:

$$s^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y})^2}{N - 1} \quad (6-1)$$

V souvislosti s analýzou rozptylu nás bude zajímat pouze čitatelem výše uvedeného zlomku, totiž součet čtverců odchylek zjištěných hodnot  $y_{ij}$  od celkového průměru  $\bar{y}$ , přičemž tento průměr je dán vztahem:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{N_i} y_{ij} \quad (6-2)$$

Tento **celkový součet čtverců** se označuje symbolem  $S_y$ , tedy:

$$S_y = \sum_{i=1}^k \sum_{j=1}^{N_i} (y_{ij} - \bar{y})^2 \quad (6-3)$$

Celkovému součtu čtverců přísluší počet stupňů volnosti  $df_c = N - 1$ .

Variabilitu uvnitř skupin označujeme jako vnitroskupinovou, nebo také reziduální a používáme přitom označení  $S_{y,v}$ , přičemž definujeme **vnitroskupinový (reziduální) součet čtverců** takto:

$$S_{y,v} = \sum_{i=1}^k \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_i)^2 \quad (6-4)$$

Vnitroskupinovému součtu čtverců přísluší počet stupňů volnosti  $df_v = N - k$ . Aritmetickými úpravami výše uvedených vzorců lze snadno dokázat **základní vztah analýzy rozptylu**, totiž, že celkový součet čtverců je roven sumě meziskupinového a vnitroskupinového součtu čtverců, symbolicky:

$$S_y = S_{y,m} + S_{y,v} \quad (6-5)$$

V anglické literatuře nebo SW se můžete také setkat s následujícím označením:

$$\begin{aligned} S_y &= S_D \quad (D \text{ z angl. Diference}) \\ S_{y,m} &= S_T \quad (T \text{ z angl. Treatment}), \\ S_{y,v} &= S_R \quad (R \text{ z angl. Residual}). \end{aligned}$$

Základní vztah ANOVA (6-5) má potom tvar

$$S_D = S_T + S_R.$$

V této kapitole budeme nadále používat původní značení, v následující kapitole se pak přidržíme nového označení.

Pro ověření nulové hypotézy  $H_0$  použijeme testové kritérium

**Celkový součet čtverců**

**Počet stupňů volnosti**

**Základní vztah analýzy rozptylu**

$$T = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{N-k}} \quad (6-6)$$

kteřá má při platnosti nulové hypotézy **Fisherovo rozdělení** pravděpodobnosti  $F_{(k-1, N-k)}(\alpha)$ . **Kritické hodnoty** Fisherova rozdělení  $F_{(df_1, df_2)}(\alpha)$  jsou tabelovány pro různé hodnoty hladiny významnosti  $\alpha$  a různé hodnoty parametrů (stupňů volnosti: **degree of freedom**)  $df_1$  a  $df_2$ . V Excelu se používá funkce FINV( $\alpha$ ;  $df_1$ ;  $df_2$ ). Někdy se namísto kritických

**Fisherovo rozdělení**

hodnot tabelují *kvantily* Fisherova rozdělení  $F_{1-\alpha}(df_1, df_2)$ . Vztah mezi kritickými hodnotami a kvantily je jednoduchý:

**Kvantily versus kritické hodnoty**

$$F_{(df_1, df_2)}(\alpha) = F_{1-\alpha}(df_1, df_2).$$

Např. 5-ti procentní (jednostranná) kritická hodnota je rovna 95-ti procentnímu kvantilu při stejných hodnotách parametrů  $df_1$  a  $df_2$ .

Postup testování hypotézy  $H_0$  je charakterizován následujícími 3 kroky:

**Krok 1.** Zvolte hladinu významnosti  $\alpha$ , (kteřá představuje chybu 1. druhu, tj. pravděpodobnost zamítnutí správné hypotézy). Praktické hodnoty hladiny významnosti  $\alpha$  jsou obvykle: 0,1, 0,05, 0,01, nebo-li v procentech: 10%, 5%, 1%.

**Krok 2.** Vypočítejte hodnotu testového kritéria  $T$  podle vzorce (6-6), přičemž pro hodnoty meziskupinového součtu čtverců  $S_{y,m}$  a pro výpočet vnitroskupinového součtu čtverců  $S_{y,v}$  použijte vzorce (6-3) a (6-4). Výpočetně výhodnější, např. pro výpočet na kalkulačce, jsou následující vzorce:

$$S_y = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{N} \left( \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \right)^2 \quad (6-7)$$

$$S_{y,m} = \sum_{i=1}^k n_i \bar{y}_i^2 - \frac{1}{n} \left( \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \right)^2 \quad (6-8)$$

K výpočtu  $S_{y,v}$  lze využít základního vztahu (6-5) a právě uvedených vztahů (6-7) a (6-8):

$$S_{y,v} = S_y - S_{y,m}.$$

**Krok 3.** Porovnejte hodnotu kritéria  $T$  z Kroku 2 s kritickou hodnotou  $F_{(k-1, n-k)}(\alpha)$ . Výsledek tohoto porovnání může být dvojnásobný. Platí

$$F \leq F_{(k-1, N-k)}(\alpha).$$

Potom se nulová hypotéza  $H_0$  přijímá (nezamítá) a tudíž konstatujeme, že hodnoty faktoru  $X$  **nemají** na hodnoty znaku  $Y$  **statisticky významný vliv**.

Jinak řečeno, faktor  $X$  je **neúčinný**.

Platí-li:

$$F > F_{(k-1, N-k)}(\alpha),$$



potom nulovou hypotézu  $H_0$  zamítáme, přijímáme hypotézu alternativní  $H_1$ , a tudíž konstatujeme, že hodnoty faktoru  $X$  **mají** na hodnoty znaku  $Y$  statisticky významný vliv.

$H_0$  zamítáme

Podáří-li se výše uvedeným testem prokázat, že hodnoty faktoru  $X$  mají na hodnoty znaku  $Y$  statisticky významný vliv, můžete se dále zajímat, které skupiny se významně odlišují od průměru, eventuálně jak skupinové průměry seřadit, případně zařadit do společných celků. V krajním případě by se totiž mohlo stát, že významnost rozdílnosti  $k$  skupin způsobuje výlučně skupina jediná a ostatní skupiny se navzájem neliší. Touto problematikou se zabývají metody tzv. **simultánního testování**, z nichž nejznámější je metoda Shaffeho. Avšak touto problematikou se zde nezabýváme, zájemce odkazujeme na literaturu, viz např. [Anděl].

### Řešení příkladu

#### Použití Excelu:

Příklad 6-1,  
dokončení

#### Nástroje→Analýza dat →ANOVA: Jeden faktor...

Pro použití Excelu je třeba nejprve data upravit. Předně musíme v souboru Studie - Data vybrat 2 znaky:  $X$  = Věk (faktor) a  $Y$  = Počet lahví/týden. Dále je vhodné data ve sloupcích uspořádat podle faktoru  $X$  a vyloučit z analýzy ty respondenty, kteří mají Věk > 22, takových respondentů je celkem 142. Data je třeba mít uspořádána ve sloupcích podle hodnot faktoru. Vytvoří se proto záhlaví sloupců z 5 hodnot faktorů, např. v18, v19,...,v22. Do každého sloupce se pak nakopírují hodnoty všech respondentů s příslušným faktorem.

Jako vstupní oblast se pak označí obdélníková oblast obsahující všech pět sloupců upravených dat. Zaklikněte popisky a zvolte výstupní oblast, kde se po potvrzení OK objeví následující výsledek:

Anova: jeden faktor						
Faktor						
<i>Výběr</i>	<i>Počet</i>	<i>Součet</i>	<i>Průměr</i>	<i>Rozptyl</i>		
v18	18	92	5,111111	50,9281		
v19	36	262	7,277778	68,77778		
v20	42	298	7,095238	35,99071		
v21	30	162	5,4	43,28276		
v22	16	142	8,875	207,85		
ANOVA						
<i>Zdroj variability</i>	<i>SS</i>	<i>Rozdíl</i>	<i>MS</i>	<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
Mezi výběry	190,2619	4	47,56548	0,714402	0,583447	2,437751
Všechny výběry	9121,569	137	66,5808			
Celkem	9311,831	141				

Jak je zřejmé  $p$ -hodnota (*Hodnota P*) testu = 0,583 je větší než  $\alpha = 0,05$ , proto nulovou hypotézu nezamítáme (přijímáme), tedy faktor věku je neúčinný na počet vypitých limonád. Jinými slovy, počet vypitých limonád nesouvisí s věkem studentů.

Ve výstupu (výše uvedená tabulka) mají uvedené termíny (nepřesné překlady z angličtiny) následující význam:

*SS* = součet čtverců (Sum of Squares),  
*Rozdíl* = počet stupňů volnosti,  
*MS* = průměrný součet čtverců (Mean of Squares),  
*F* = hodnota testového kritéria *T*,  
*Hodnota P* = *p*-hodnota (signifikance),  
*F krit* = kritická hodnota Fisherova rozdělení,  
*Mezi výběry* = meziskupinový součet čtverců,  
*Všechny výběry* = vnitroskupinový součet čtverců,  
*Celkem* = celkový součet čtverců, resp. počet st. volnosti,

#### **Použití SPSS:**

#### **Analyze → Compare Means → One-Way ANOVA...**

Pro použití SPSS je rovněž třeba data předem upravit. Musíme však také v souboru Studie - Data vybrat 2 znaky: *X* = Věk (faktor) a *Y* = Počet lahví/týden.

Úprava dat - vypuštění respondentů s věkem nad 22 let:

**Data → Select Cases → If → vek =< 22** (vyškrtají se „nepotřebné“ řádky)

**Použití  
v SPSS**

Výstupem je následující tabulka:

**ANOVA**

		Sum of Squares	df	Mean Square	F	Sig.
Pocet lahvi/tyden	Between Groups	190,262	4	47,565	,714	,583
	Within Groups	9121,569	137	66,581		
	Total	9311,831	141			

Odpovídající si hodnoty obou výstupů jsou stejné.

\*

## 6.3 Míra těsnosti závislosti

Variabilita podmíněných (skupinových) průměrů  $\bar{y}_i$  kolem celkového průměru  $\bar{y}$  je způsobena závislostí znaku *Y* na znaku *X*. Tuto variabilitu jsme vyjádřili meziskupinovým součtem čtverců  $S_{y,m}$ . Variabilita znaku *Y* uvnitř jednotlivých skupin – vyjádřena vnitroskupinovým (reziduálním) součtem čtverců  $S_{y,v}$ , je způsobena jinými (neuvažovanými) činiteli.

**Variabilita**

Čím větší je  $S_{y,m}$ , tím větší je těsnost závislosti znaků *X* a *Y*. Protože však jsou jednotlivé součty čtverců vzájemně vázány vztahem (6-5), lze míru těsnosti závislosti vyjádřit jako podíl meziskupinového a celkového součtu čtverců. Zavádíme proto jako míru těsnosti závislosti znaku *Y* na znaku *X* *poměr determinace*  $P^2$  takto:

**Poměr  
determinace**

$$P^2 = \frac{S_{y,m}}{S_y}$$

(6-9)

Odmocninu z poměru determinace  $P$  nazýváme *poměr korelace*.

**Poměr korelace**

Poměr determinace nabývá hodnot z intervalu  $[0,1]$ . Čím těsnější je závislost  $Y$  na  $X$ , tím více se hodnota poměru determinace blíží k 1, tím více se také vnitroskupinový součet čtverců blíží k celkovému součtu čtverců, přičemž meziskupinový součet čtverců se blíží k nule. Naopak, čím více se poměr determinace blíží k 0, tím menší část z celkového součtu čtverců tvoří meziskupinový součet čtverců (na úkor vnitroskupinového), a tím menší je těsnost závislosti znaku  $Y$  na  $X$ .

**Těsnost závislosti**

## ŘEŠENÝ PŘÍKLAD 6-2



### Studie

Ve Studii uvažujte znak  $X$  (faktor) = Pořadí K/K (1 až 5, počet kategorií je tedy  $k = 5$ ). Jako kvantitativní znak  $Y$  je Pravděpodobnost nákupu K/K.

**Problém:** Má pořadí nového nápoje K/K mezi tradičními limonádami vliv na pravděpodobnost nákupu nového nápoje?

### Řešení příkladu

Nulová hypotéza se v tomto případě formuluje takto:

$H_0$ : Pořadí nového nápoje K/K mezi tradičními limonádami nemá vliv na pravděpodobnost nákupu K/K.

Alternativní hypotéza je negací (opakem) hypotézy nulové:  $H_1$ : „opak  $H_0$ “

Pro použití Excelu je třeba nejprve data upravit. Předně musíte v souboru Studie - Data vybrat 2 znaky:  $X$  = Pořadí K/K (faktor) a  $Y$  = Pravděpodobnost nákupu K/K. Data je třeba mít uspořádána ve sloupcích podle hodnot faktoru. Vytvoříte proto záhlaví sloupců z 5 hodnot faktorů, např. P1, P2, ..., P5. Do každého sloupce pak nakopírujete hodnoty všech respondentů s příslušným faktorem.

Jako vstupní oblast pak označíte obdélníkovou oblast obsahující všech pět sloupců upravených dat. Zaškrtnete popisky a zvolíte výstupní oblast, kde se po potvrzení OK objeví následující výstup:

Anova: jeden faktor

Faktor	Výběr	Počet	Součet	Průměr	Rozptyl
P1		35	2190	62,57143	1252,017
P2		41	2030	49,5122	1363,506
P3		44	1790	40,68182	1185,571
P4		40	1740	43,5	1554,103
P5		40	1690	42,25	1462,756

ANOVA

Zdroj variability	SS	Rozdíl	MS	F	Hodnota P	F krit
Mezi výběry	11886,14	4	2971,535	2,180464	0,072619	1,973909
Všechny výběry	265745,9	195	1362,799			
Celkem	277632	199				

Hladina významnosti je  $\alpha = 0,10$  (tj. 10%). Kritická hodnota je

$$F_{(k-1, n-k)}(\alpha) = F_{(4, 195)}(0,1) = 1,97.$$

Kritický obor je

$$C = [1,97; +\infty).$$

Vidíte, že  $2,18 > 1,97$ , tzn.  $T \in C$ . Proto nulovou hypotézu  $H_0$  zamítáte!

Znamená to, že faktor  $X$  - Pořadí K/K ovlivňuje pravděpodobnost nákupu, neboli, že pravděpodobnost nákupu nového nápoje je statisticky významně ovlivněna hodnocením (pořadím) tohoto nápoje.

Na závěr ještě vypočítáme těsnost této závislosti pomocí poměru korelace:

$$P = \sqrt{\frac{11886}{277632}} \approx 0,21.$$

Hladina významnosti je zvolena poměrně vysoká (10%) a hodnota poměru korelace (těsnost závislosti) vyšla poměrně nízká. Tento výsledek demonstruje nepřesvědčivou závislost pravděpodobnosti prodeje nového nápoje na jeho hodnocení ve skupině tradičních limonád.

**Kritická  
hodnota**

**Kritický  
obor**

\*

### PRŮVODCE TEXTEM



Následující samostatné úkoly řešte nejprve pomocí Excelu a potom pomocí SPSS, viz [Studie](#).

### SAMOSTATNÝ ÚKOL 8



Ve Studii uvažujte znak  $X$  (tzv. faktor) = Věk (18 až 22 let, počet kategorií -roků je tedy  $k = 5$ , počet takových respondentů - studentů je  $N=142$ ). Jako kvantitativní znak  $Y$  označíme Počet lahví/den. Má věk - stáří respondentů - studentů vliv na konzumaci limonád?

ANOVA

### SAMOSTATNÝ ÚKOL 9



Ve Studii uvažujte znak  $X$  (tzv. faktor) = Pořadí K/K (1 až 5, počet kategorií je tedy  $k = 5$ ). Jako kvantitativní znak  $Y$  označíme Hodnocení chuti K/K. Má pořadí nového nápoje K/K mezi tradičními limonádami vliv na hodnocení chuti nového nápoje?

ANOVA

### SHRNUTÍ KAPITOLY ANALÝZA ROZPTYLU



Statistickou významnost rozdílů mezi výběrovými průměry dvou náhodných výběrů umožňují testy hypotéz, např. t-test, F-test apod. Analýza rozptylu – ANOVA umožňuje ověřit významnost rozdílů mezi výběrovými průměry většího počtu náhodných výběrů, umožňuje posoudit vliv různých faktorů na proces charakterizovaný kvantitativním statistickým znakem. Základní myšlenka analýzy rozptylu spočívá v rozkladu celkového rozptylu na dílčí rozptyly příslušející

**Shrnutí  
kapitoly**

jednotlivým vlivům, podle nichž jsou data roztríděna. Kromě dílčích rozptylů je jednou složkou celkového rozptylu tzv. reziduální rozptyl, způsobený nepostiženými vlivy. Podle počtu analyzovaných faktorů rozlišujeme jednofaktorovou, dvoufaktorovou a vícefaktorovou analýzu rozptylu. V této kapitole jsme se věnovali prvním dvěma jmenovaným typům ANOVA.

Metoda jednofaktorové analýzy rozptylu spočívá v tom, že celkovou variabilitu měřenou součtem čtverců odchylek od celkového průměru rozdělíme na variabilitu uvnitř jednotlivých výběrů a na variabilitu mezi jednotlivými výběry. Předpokládáme, že faktor  $X$  má  $k$  úrovní s efektem na znak  $Y$  a formulujeme nulovou hypotézu, že všechny výběry pocházejí ze stejné základní populace. Pro ověření nulové hypotézy použijeme statistiku (6-6), která má při platnosti nulové hypotézy Fisherovo rozdělení. Kritické hodnoty Fisherova rozdělení jsou tabelovány pro různé hodnoty hladiny významnosti  $\alpha$  a různé hodnoty stupňů volnosti. Nulovou hypotézu buď přijímáme (nezamítáme) a tudíž konstatujeme, že hodnoty faktoru  $X$  nemají na hodnoty znaku  $Y$  statisticky významný vliv, anebo nulovou hypotézu zamítáme, přijímáme hypotézu alternativní a tudíž konstatujeme, že hodnoty faktoru  $X$  mají na hodnoty znaku  $Y$  statisticky významný vliv. Míru těsnosti závislosti vyjadřujeme jako podíl meziskupinového a celkového součtu čtverců a zavádíme proto jako míru těsnosti závislosti znaku  $Y$  na znaku  $X$  poměr determinace  $P^2$ , který nabývá hodnot mezi 0 a 1. Čím těsnější je závislost  $Y$  na  $X$ , tím více se hodnota poměru determinace blíží k 1. Odmocninu z poměru determinace  $P$  nazýváme poměr korelace.

U analýzy rozptylu s jedním faktorem jsme uvažovali výsledky tříděné podle jistého kvalitativního znaku  $X$  do několika (konkrétně do  $k$ ) skupin. Proto v tomto případě se hovoří o ANOVA při jednoduchém třídění. Kromě třídění do skupin se vyskytuje další faktor, podle něhož se výsledky (tj. hodnoty znaku  $Y$ ) třídí do bloků. Jedná se o ANOVA při dvojném třídění.

Na příkladech ze studie se metoda ANOVA demonstruje s použitím Excelu a alternativně SPSS.

## KONTROLNÍ TEST 6



- 6.1** Jednofaktorová analýza rozptylu - ANOVA slouží k (označte správnou odpověď, může jich být i více):
- výpočtu rozdělení četnosti jednotlivých znaků
  - testování vlivu faktoru na kvantitativní znak
  - zjištění typu rozdělení pravděpodobnosti
  - zjištění nezávislosti statistických znaků
- 6.2** Při testu hypotézy se v ANOVA:
- testuje se nulová hypotéza, že hodnoty obou znaků jsou stejné,
  - testuje se nulová hypotéza, že dva statistické znaky jsou vzájemně závislé,
  - testuje se nulová hypotéza, že hodnota znaku se liší od zadané hodnoty
  - testuje se nulová hypotéza, že pro jednotlivé kategorie faktoru jsou hodnoty znaků stejné,
  - testuje se nulová hypotéza, že dva statistické znaky jsou vzájemně nezávislé.
- 6.3** Při ANOVA se využívá kritické hodnoty:
- Studentova t-rozdělení pravděpodobnosti,
  - Pearsonova Chi-kvadrát rozdělení pravděpodobnosti,
  - Fisherova F-rozdělení pravděpodobnosti,
  - Gaussova normálního rozdělení pravděpodobnosti,

**6.4** Při použití Excelu pro ANOVA: jeden faktor musí být hodnoty kvantitativního znaku uspořádány:

- a. ve sloupcích podle kategorií faktoru,
- b. ve sloupcích podle velikosti kvantitativního znaku,
- c. v řádcích podle kategorií faktoru,
- d. v řádcích podle velikosti kvantitativního znaku.

**6.5** Ke každému pojmu z levého sloupce přiřaďte správné charakteristiky z pravého sloupce vztahující se k dvouvýběrovým neparametrickým testům:

- |     |                     |     |   |
|-----|---------------------|-----|---|
| (1) | Chi-kvadrát test    | (A) | Testuje stejnou hodnotu průměru nezávislou na kategoriích znaku |
| (2) | Mann-Whitneyův test | (B) | Testuje nezávislost hodnot ve dvou výběrech                     |
| (3) | ANOVA               | (C) | Testuje nezávislost pořadí ve dvou výběrech                     |

**6.6** Určete, zda následující výroky jsou správné (vepište A) nebo nesprávné (vepište N):

- a. K testu neúčinnosti faktorů se v analýze rozptylu používá  $F$ -test.
- b. Determinační poměr nabývá hodnot z intervalu  $[0;1]$ .
- c. Závislost mezi znaky  $X$  a  $Y$  je tím větší, čím menší je meziskupinová variabilita.
- d. Obor přijetí je při analýze rozptylu oboustranný.
- e. Rozptyl podmíněných průměrů odráží vnitroskupinovou variabilitu znaku  $Y$ .

**6.7** Doplňte správný výraz:

- a. Leží-li hodnota testového kritéria  $F$  v kritickém oboru, lze na dané hladině významnosti považovat znak  $Y$  za \_\_\_\_\_ na znaku  $X$ .
- b. Analýza rozptylu, kdy celkový počet pozorování je  $n$  a faktor má  $k$  kategorií, předpokládá nalezení příslušného kvantilu  $F$  rozdělení o stupních volnosti \_\_\_\_\_ a \_\_\_\_\_.
- c. Hodnota testového kritéria  $F$  nabývá vždy \_\_\_\_\_ hodnot.
- d. Jednofaktorová analýza rozptylu zkoumá závislost číselných hodnot znaku  $Y$  na \_\_\_\_\_ znaku  $X$ .

**6.8** Doplňte správné sousloví:

- a. Odmocnina z determinačního poměru se nazývá \_\_\_\_\_.
- b. Padne-li hodnota testového kritéria  $F$  do \_\_\_\_\_ pak nulovou hypotézu přijímáme.
- c. K výpočtu kritického oboru jednofaktorové ANOVA potřebujeme znát \_\_\_\_\_, \_\_\_\_\_ a \_\_\_\_\_.
- d. V SPSS naleznete výsledek ANOVA tak, že ve výstupní tabulce naleznete hodnotu Signifikance a porovnáte ji s uvažovanou \_\_\_\_\_.

**ŘEŠENÍ SAMOSTATNÝCH ÚKOLŮ****Samostatný úkol 8**

Sig. 0,0539

-výstup z Excelu.

Nulovou hypotézu přijímáme. Průměrný počet limonád vypitých denně je roven 2.

**Samostatný úkol 9**

Sig. 0,00764

-výstup z Excelu.

Nulovou hypotézu zamítáme. Průměrný počet limonád vypitých denně u mužů je statisticky významně vyšší než u žen.

**Samostatný úkol 8****Samostatný úkol 9****ŘEŠENÍ KONTROLNÍHO TESTU 6****6.1** d.**6.2** d.**6.3** c.**6.4** a., c.**6.5** (1) – (B), (2) – (C), (3) – (A)**6.6** A, A, N, N, A**6.7** a. závislý, b.  $k-1$  a  $N-k$ , c. kladných, d. kategoriích**6.8** a. korelační poměr, b. oboru přijetí, c. stupně volnosti  $k-1$ ,  $N-k$  a hladina významnosti, d. hladina významnosti.

## 7 ANALÝZA ROZPTYLU (ANOVA): DVOJNÉ TŘÍDĚNÍ A LATINSKÉ ČTVERCE

### CÍLE KAPITOLY ANALÝZA ROZPTYLU (ANOVA): DVOJNÉ TŘÍDĚNÍ A LATINSKÉ ČTVERCE

Po úspěšném a aktivním absolvování této KAPITOLY

#### Budete umět:

- sestavit ANOVA tabulku pro dvojné a trojné třídění
- správně interpretovat výsledky získané pomocí statistického softwaru

*Budete umět*

#### Budete schopni:

- porozumět pojmu vícefaktorová analýza rozptylu,
- stanovit nulové hypotézy

*Budete schopni*

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny.

#### PRŮVODCE STUDIEM 8



V předešlé kapitole jste se seznámili s jednofaktorovou analýzou rozptylu. V této kapitole se dozvíte o vícefaktorové analýze rozptylu. Jedná se o situaci, kdy se zkoumá, zda kvantitativní znak  $Y$  je ovlivňován dvěma nebo třemi kvalitativními proměnnými. Vícefaktorová analýza rozptylu se používá k vyhodnocení závislostí (vlivu) pomocí tzv. experimentu, kdy nejsou dána všechna vícerozměrná data, nýbrž část z nich lze vhodným způsobem zvolit tak, aby výsledek byl co nejpřesvědčivější.

Tato kapitola je rozdělena do dvou částí. V části první se hovoří o dvojném třídění a v části druhé o třídění trojném – neboli o latinských čtvercích.

Sleduje-li se vliv dvou (kvalitativních) znaků - faktorů, které mohou ovlivnit hodnotu sledovaného (kvantitativního) znaku, hovoříme o dvojném třídění. Jeden ze znaků – nazývaný faktor lze aktivně ovlivnit v experimentu, který slouží k zjištění vlivu faktoru na sledovaný znak. Vliv druhého faktoru – tzv. blokového znaku lze v experimentu eliminovat znáhodněním hodnot tohoto znaku.

U trojného třídění pomocí tzv. latinských čtverců se opět vyskytuje jeden faktor se dvěma blokovými znaky, jejichž vliv na sledovaný znak lze eliminovat vhodnou organizací experimentu.

Kapitola je proložena řešenými příklady, které si pečlivě prostudujte tak, abyste byli schopni odpovědět na kontrolní otázky v závěru této kapitoly.



## 7.1 Dvojné třídění

### DEFINICE 7-1



Rozklad  
součtu  
čtverců

Je-li sledovaný znak ovlivňován dvěma faktory, hovoříme o dvojném třídění. Jeden ze znaků – faktor lze aktivně ovlivnit v experimentu, který slouží k zjištění vlivu faktoru na sledovaný znak. Vliv druhého znaku – tzv. *blokového znaku* lze v experimentu eliminovat znáhodněním hodnot tohoto znaku.

Rozklad celkového součtu čtverců  $S_D$  se provede analogicky jako v případě jednoduchého třídění (viz ANOVA z předešlé kapitoly), pouze přibude nový sčítanec. Označíme jej  $S_B$ , (symbol  $B$  je z anglického Block) a přináleží *blokovému faktoru*. Součet čtverců  $S_T$  přináleží *meziskupinovému faktoru* (angl. Treatment factor), součet čtverců  $S_R$  přináleží *vnitroskupinovému faktoru* (angl. Residual factor). Rozklad celkového součtu čtverců  $S$  potom bude

$$S_D = S_B + S_T + S_R \quad (7-1)$$

kde  $n$  značí počet bloků,  $k$  značí počet skupin a platí

$$S_B = k \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 \quad (7-2)$$

$$S_T = n \sum_{j=1}^k (\bar{y}_j - \bar{y})^2 \quad (7-3)$$

Pro zjednodušení výpočtů dále označíme

$$S = \sum_{j=1}^k \sum_{i=1}^n y_{ij}^2 \quad (7-4)$$

$$S_A = nk\bar{y}^2 \quad (7-5)$$

potom je

$$S_D = S - S_A \quad (7-6)$$

a reziduální součet čtverců, který je nejpracnější, se vypočítá ze vztahů (7-1) a (7-6) takto

$$S_R = S - S_A - S_B - S_T. \quad (7-1a)$$

### PRŮVODCE TEXTEM



Uvedené vztahy objasníme v následujících příkladech.

### ŘEŠENÝ PŘÍKLAD 7-1



Sledují se emise výfukových plynů ( $Y$ ) v závislosti na dvou faktorech. Jedná se o typ přísady (A,B,C,D), což představuje první faktor ( $X_1$ ), který ovlivňuje emise výfukových plynů, druhým faktorem ( $X_2$ ) je vliv řidiče (I,II,III,IV). Celkem byly provedeny 4 pokusy s každým typem přísady. Naměřené hodnoty emise jsou v Tabulce 7.1, počet hodnot dat je  $N = 16$ .

Proveďte test na hladině významnosti 5%, kterým ověříte, zda jsou emise výfukových plynů statisticky významně ovlivněny prvním faktorem (typ přísady), nebo druhým faktorem (vliv řidiče), popř. oběma faktory současně.

Emise  
výfuko-  
vých plynů

Řidič	Přísada			
	A	B	C	D
I	21	26	25	20
II	20	27	26	23
III	16	15	16	13
IV	15	20	17	20

Tab.7.1 Plán a výsledky experimentu

Analýza rozptylu je provedena oběma výše uvedenými způsoby.

Rozklad na základě vztahů (7-1a) a (7-6):

### Řešení příkladu

Analýza rozptylu je provedena oběma výše uvedenými způsoby.

Rozklad na základě vztahů (7-1a) a (7-6):

Součet čtverců	$df$	Rozptyl	Testová kritéria	Kritická hodnota
$S_A = 6400$	1		$\frac{s_B^2}{s_R^2} = 16,3$	$F_{(3;9)}(0,05) = 3,86$
$S_B = 216$	$n - 1 = 3$	$s_B^2 = 72$		
$S_T = 40$	$k - 1 = 3$	$s_T^2 = 13,3$	$\frac{s_T^2}{s_R^2} = 3,0$	
$S_R = 40$	9	$s_R^2 = 4,4$		
$S = 2696$	16	Pozn.: Pro $S_R$ je $df = (n - 1)(k - 1)$		

Tab. 7.2 ANOVA se 2 faktory – ruční způsob

Protože jen u druhého faktoru, kterým je vliv řidiče, překračuje testovací kritérium kritickou hodnotu, je pouze tento faktor statisticky významný. Lze tedy tvrdit, že emise výfukových plynů jsou ovlivňovány řidičem.

Dále provedeme rozklad na základě vztahu (7-1):

Součet čtverců	d.f.	Rozptyl	Testovací kritéria	Sig. level
$S_B = 216$	3	$s_B^2 = \frac{216}{3} = 72$	$\frac{s_B^2}{s_R^2} = 16,36$	0,0006
$S_T = 40$	3	$s_T^2 = \frac{40}{3} = 13,3$	$\frac{s_T^2}{s_R^2} = 3,02$	0,0877
$S_R = 40$	9	$s_R^2 = \frac{40}{9} = 4,4$		
$S_D = 296$	15			

Tab. 7.3 ANOVA se 2 faktory - způsob pro PC

Jak vidět, pouze u faktoru B, kterým je vliv řidiče, je hodnota signifikance menší než 0,05 (což je hladina významnosti, na které test provádíme). To znamená, že pouze tento faktor je statisticky významný a je dokonce významný i na hladině významnosti 0,01.

**Použití Excelu:**

V hlavním Menu v Excelu zvolíme postupně:

**Nástroje**→**Analýza dat**→**ANOVA: Dva faktory bez opakování...**

Data musejí být uspořádána v tabulce obdobné Tab. 7.1, tj. skupiny v řádcích a bloky ve sloupcích, případně obráceně (s popiskami nebo bez nich), viz Tab. 7.4.

	26	25	20
20	27	26	23
16	15	16	13
15	20	17	20

Tab. 7.4 ANOVA se 2. faktory – data

Jako vstupní oblast se pak označí obdélníková (v tomto případě čtvercová) oblast obsahující všechny řádky a sloupce upravených dat. Pokud máte data i s popisky řádků a sloupců, zaklikněte popisky a zvolte výstupní oblast, kde se po potvrzení OK objeví následující výsledek:



**Řešení  
v Excelu**

Anova: dva faktory bez opakování

Faktor	Počet	Součet	Průměr	Rozptyl
I	4	92	23	8,666667
II	4	96	24	10
III	4	60	15	2
IV	4	72	18	6
A	4	72	18	8,666667
B	4	88	22	31,333333
C	4	84	21	27,333333
D	4	76	19	18

ANOVA

Zdroj variability	SS	Rozdíl	MS	F	Hodnota P	F krit
Řádky	216	3	72	16,2	0,000569	3,862539
Sloupce	40	3	13,333333	3	0,087713	3,862539
Chyba	40	9	4,444444			
Celkem	296	15				

Výsledek (*Hodnota P*) je stejný jako v Tab. 7.3.

\*

**ŘEŠENÝ PŘÍKLAD 7-2**

Zjistěte, zda počet vyrobených výrobků závisí na stroji nebo na operátorovi, popř. zda oba tyto faktory mají vliv na počet vyrobených výrobků. Test proveďte na hladině významnosti  $\alpha = 0,05$ . Počet vyrobených výrobků zachycuje Tab. 7.5.

Výkony  
strojů

Operátor	Stroj		
	A	B	C
I	53	61	51
II	47	55	51
III	46	52	49
IV	50	58	54
V	49	54	50

Tab.7.5 Plán a výsledky experimentu

**Řešení příkladu**

Řešte tento příklad na základě vztahu (7-1a):

Součet čtverců	d.f.	Rozptyl	Testovací kritéria	Sig. level
$S_B = 72$	4	$s_B^2 = \frac{72}{4} = 18$	$\frac{s_B^2}{s_R^2} = 6,5$	0,012
$S_T = 130$	2	$s_T^2 = \frac{130}{2} = 65$	$\frac{s_T^2}{s_R^2} = 23,6$	0,0004
$S_R = 22$	8	$s_R^2 = \frac{22}{8} = 2,75$		
$S_D = 224$	14			

Tab.7. 6 ANOVA – 2 faktory, příklad 2

Z tabulky vidíte, že obě hodnoty v posledním sloupci jsou menší než zvolená hladina významnosti  $\alpha = 0,05$ , proto lze nulové hypotézy o nezávislosti obou faktorů na počet vyrobených výrobků zamítnout. Můžete tedy z 95% tvrdit, že počet vyrobených výrobků je ovlivněn jak strojem tak i operátorem.

Při použití vztahu (7-1) byste v tabulkách našli tyto kritické hodnoty

$$F_{(4;8)}(0,05) = 3,84 \quad F_{(2;8)}(0,05) = 4,46.$$

Vidíte tedy, že obě testovací kritéria leží za těmito kritickými hodnotami, tj.  $6,5 > 3,84$  a  $23,6 > 4,46$ , tudíž obě nulové hypotézy o nezávislosti zamítáme: oba faktory mají na počet výrobků statisticky významný vliv.

\*

## 7.2 Trojné třídění (Latinské čtverce)

### PRŮVODCE STUDIEM 9



Latinské čtverce patří mezi klasické metody plánování experimentu. Historicky pochází tento pojem z 18. století, kdy L. Euler (1707 – 1783) předložil petrohradské akademii úlohu o 36 důstojnících: Sestavte 36 důstojníků 6 různých hodnotí ze 6 různých pluků do čtverce tak, aby v každé řadě a v každém zástupu byli důstojníci všech hodnotí a všech pluků.

Obecněji se tento problém dá zformulovat takto:

Uvažujte objekty, které mají 2 vlastnosti: A a B (např. A=hodnota, B=pluk). Každá vlastnost nabývá n různých hodnot ( $n = 6$ , 6 různých hodnotí: podporučík, poručík, nadporučík, kapitán, major, plukovník, 6 různých pluků: pluk 1, pluk 2, ..., pluk 6). Úkolem je postavit  $n^2$  objektů do čtverce tak, aby v každé vodorovné řadě i v každé svislé řadě byly vždy objekty všech hodnotí vlastnosti A a zároveň všech hodnotí vlastnosti B (např. v první řadě stojí podporučík z pluku 6,

poručík z pluku 5,..., plukovník z pluku 1). Takovéto schéma objektů se nazývá latinský čtverec řádu  $n$ .

Známy výsledek, který pochází od samotného L. Eulera, pak tvrdí, že pro každé přirozené číslo  $n$  existuje alespoň jeden latinský čtverec řádu  $n$ .

V této subkapitole se tedy budeme zabývat latinskými čtverci, které slouží k vyhodnocování experimentů.

### DEFINICE 7-2



**Rozklad  
součtu  
čtverců  
(7-7)**

Je-li sledovaný znak  $Y$  ovlivňován třemi faktory, hovoříme o tzv. latinských čtvercích a jedná se o trojné třídění. Rozklad součtu čtverců odchylek zde bude

$$S = S_A + S_B + S_C + S_T + S_R,$$

resp. při použití statistického programu

$$S_D = S_B + S_C + S_T + S_R,$$

kde

$$S_A = k^2 \bar{y}^2,$$

(7-8)

$k$  je rozměr tabulky s výsledky experimentu, součty čtverců  $S_B, S_C, S_T$  se počítají podle vztahů

$$k \sum_{i=1}^k (\bar{y}_i - \bar{y})^2,$$

(7-9)

kde  $\bar{y}_i$  je průměr v příslušné (testované) skupině.

$$S_R = S - S_A - S_B - S_C - S_T,$$

(7-10)

$$S = \sum_{i=1}^k \sum_{j=1}^k \bar{y}_{ij}^2,$$

(7-11)

$$S_D = \sum_{i=1}^k \sum_{j=1}^k (y_{ij} - \bar{y})^2,$$

(7-12)

### ŘEŠENÝ PŘÍKLAD 7-3



**Emise  
výfuko-  
vých plynů**

Uvažujte opět příklad 7-1, kde jste se zabývali sledováním množství emisí výfukových plynů ( $Y$ ) v závislosti na těchto třech faktorech:

- typu přísady do benzínu (A,B,C,D) = faktor,
- řidič vozidla (I, II, III, IV) = blok B.

Nyní budeme navíc uvažovat blokový znak

- použité vozidlo (1,2,3,4) = blok C.

Výsledky experimentu jsou uvedeny v Tab. 6. Protože 1. řádek a 1. sloupec jsou již obsazené, bude třetí faktor zaznamenáván do tabulky. Tato filosofie zápisu se uplatňuje i u čtyř a více faktorů.

Řidič\vozidlo:Y	1	2	3	4
I	A : 21	B : 26	D : 20	C : 25
II	D : 23	C : 26	A : 20	B : 27
III	B : 15	D : 13	C : 16	A : 16
IV	C : 17	A : 15	B : 20	D : 20

Tab. 7.7 Výsledky experimentu - 3 faktory

**Řešení příkladu**

Výpočty průměrů pro jednotlivá třídění jsou v pomocné tabulce 7.8

Řidiči	Vozidla	Přísady
I:23	1:19	A:18
II:24	2:20	B:22
III:15	3:19	C:21
IV:18	4:22	D:19

Tab. 7.8 Výpočty průměrů pro jednotlivá třídění

Dále vypočteme na základě vztahů (7-7) – (7-13) jednotlivé součty čtverců:

$$S = 21^2 + 26^2 + 20^2 + \dots + 20^2 = 6696,$$

$$S_B = 4[(23-20)^2 + (24-20)^2 + (15-20)^2 + (18-20)^2] = 216,$$

$$S_C = 4[(19-20)^2 + (20-20)^2 + (19-20)^2 + (22-20)^2] = 24,$$

$$S_T = 4[(18-20)^2 + (22-20)^2 + (21-20)^2 + (19-20)^2] = 40,$$

$$S_A = 4 \cdot 4 \cdot 20^2 = 6400,$$

$$S_R = 6696 - 6400 - 216 - 24 - 40 = 16.$$

Uvedené výsledky zaznamenejme do tabulek pro ruční výpočet a pro výstup statistického programu.

	d.f.	Rozptyl	Testová kritéria	Kritická hodnota
$S_A = 6400$	1	$s_A^2 = 6400$		
$S_B = 216$	$k - 1 = 3$	$s_B^2 = \frac{216}{3} = 72$	$\frac{s_B^2}{s_R^2} \approx 27$	$F_{(3;6)}(0,05) = 4,76$
$S_C = 24$	$k - 1 = 3$	$s_C^2 = \frac{24}{3} = 8$	$\frac{s_C^2}{s_R^2} \approx 3$	
$S_T = 40$	$k - 1 = 3$	$s_T^2 = \frac{40}{3} = 13,3$	$\frac{s_T^2}{s_R^2} \approx 5$	
$S_R = 16$	$(k-1)(k-2) = 6$	$s_R^2 = \frac{16}{6} = 2,67$		
$S = 6696$	16			

Tab. 7.9 ANOVA (ruční výpočet)

Testová kritéria pro blokovou proměnnou B (vliv řidiče) a pro faktor (přísady do benzínu) leží za kritickou hodnotou, tedy v kritickém oboru. Proto lze na 95% tvrdit, že na hodnoty emisí výfukových plynů jsou nejvíce ovlivněny vlivem řidiče a vlivem přísad do benzínu.

	d.f.	Rozptyl	Testová kritéria	Sig.level
$S_B = 216$	$k - 1 = 3$	$s_B^2 = \frac{216}{3} = 72$	$\frac{s_B^2}{s_R^2} \approx 27$	0,0007
$S_C = 24$	$k - 1 = 3$	$s_C^2 = \frac{24}{3} = 8$	$\frac{s_C^2}{s_R^2} \approx 3$	0,1170
$S_T = 40$	$k - 1 = 3$	$s_T^2 = \frac{40}{3} = 13,3$	$\frac{s_T^2}{s_R^2} \approx 5$	0,0452
$S_R = 16$	$(k - 1)(k - 2) = 6$	$s_R^2 = \frac{16}{6} = 2,67$		
$S = 6696$	16			

Tab. 7.10 ANOVA (výpočet na PC)

Test je prováděn na hladině významnosti  $\alpha = 0,05$  a statisticky významné faktory jsou ty, u nichž je hodnota signifikance menší než 0,05. Jsou to tedy faktory vliv řidiče a přísady do benzínu.

\*

### SHRNUTÍ KAPITOLY ANALÝZA ROZPTYLU (ANOVA): DVOJNÉ TŘÍDĚNÍ A LATINSKÉ ČTVERCE



Shnutí kapitoly

V této kapitole jsme se seznámili s vícefaktorovou analýzou rozptylu. Byly zde uvedeny dvě možnosti techniky výpočtu a také jsme se seznámili s počítačovými výstupy vícefaktorové analýzy rozptylu. V řešených příkladech, které najdete v následující subkapitole si zopakujte jednofaktorovou analýzu rozptylu a pojem korelační koeficient.

Poznali jsme pojmy jako: dvojné třídění, Latinské čtverce, rozklad součtu čtverců, ANOVA tabulka, faktor, bloková proměnná.

### PRŮVODCE STUDIEM 10



Dále následují řešené příklady. Prvním příkladem si zopakujete učivo předcházející kapitoly, protože tento příklad se týká jednofaktorové analýzy rozptylu.



## ŘEŠENÝ PŘÍKLAD 7-4



Na hladině významnosti 0,05 testujte, zda výnosy petržele závisí na použitém druhu hnojiva.

Výnosy  
petržele

Hnojivo	Výnosy (1kg/10m <sup>2</sup> )					
A	40	42	45	40	44	47
B	76	75	82	68		
C	60	58	62	64	70	

## Řešení příkladu

Faktor typ hnojiva má 3 skupiny, tzn.  $k = 3$ , s následujícími četnostmi ve skupinách  $n_1 = 6$ ,  $n_2 = 4$ ,  $n_3 = 5$ . Celkový počet měření  $N = 15$ .

Testujeme nulovou hypotézu  $H_0: \mu_1 = \mu_2 = \mu_3$ , tj., že **výnosy petržele nezávisí na typu hnojiva**.

Před vypočtením testového kritéria musíme zjistit hodnoty následujících veličin:

- *Podmíněné průměry*  $\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ , pro  $i = 1, 2, \dots, k$ , kde:  $y_{ij}$  jsou zjištěné hodnoty.
- *Celkový průměr*  $\bar{y} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^k n_i \bar{y}_i$ , kde:  $n$  je celkový rozsah souboru.
- *Meziskupinový součet čtverců*  $S_{y,m} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$ , kde:  $n_i$  je počet měření v jednotlivých skupinách,  $\bar{y}_i$  je výběrový průměr v jednotlivých skupinách.
- *Vnitroskupinový součet čtverců*  $S_{y,v} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ .
- *Celkový součet čtverců*  $S_y = S_{y,m} + S_{y,v}$ .

Hodnoty výše uvedených veličin jsou:

$$\bar{y}_1 = \frac{40 + 42 + \dots + 47}{6} = 43$$

$$\bar{y}_2 = \frac{76 + 75 + \dots + 68}{4} = 75,25$$

$$\bar{y}_3 = 62,8$$

$$\bar{y} = \frac{43 \cdot 6 + 75,25 \cdot 4 + 62,8 \cdot 5}{15} = 58,2$$

$$S_{y,m} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = 6(43 - 58,2)^2 + 4(75,25 - 58,2)^2 + 5(62,8 - 58,2)^2 = 2654,85$$

$$S_{y,v} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = (40 - 43)^2 + (42 - 43)^2 + \dots + (47 - 43)^2 + \\ + (76 - 75,25)^2 + (75 - 75,25)^2 + \dots + (68 - 75,25)^2 + \\ + (60 - 62,8)^2 + (58 - 62,8)^2 + \dots + (70 - 62,8)^2 = 223,55$$

$$S_y = 2878,4$$

Zdroj variability	Součty čtverců odchylek	Počty stupňů volnosti	Průměrné čtverce	Testové kritérium F
<b>Faktor x</b> (meziskupinová variabilita)	2654,85	$k - 1 = 2$	1327,42	71,25
<b>Reziduální</b> (vnitroskupinová variabilita)	223,55	$N - k = 12$	18,63	
<b>Celkový</b>	2878,4	$N - 1 = 14$		

Hodnota statistiky  $T = 71,25$ , kritická hodnota  $F_{(2;12)}(0,05) = 3,89$ , kritický obor je  $C = (3,89; +\infty)$ . Hodnota  $T$  leží v kritickém oboru, proto zamítáme na hladině významnosti 5% nulovou hypotézu. **Výnosy petržele závisí na zvoleném typu hnojiva.**

\*

## ŘEŠENÝ PŘÍKLAD 7-5



Pomocí determinačního poměru zjistěte těsnost závislost výnosů petržele na použitém druhu hnojiva.

### Řešení příkladu

Na otázku „Jak silná je vazba mezi nezávislou nominální proměnnou (typ hnojiva) a proměnnou číselnou (výnosy petržele)??“, odpovídá hodnota *korelačního poměru*.

**Korelační poměr** 
$$P = \sqrt{\frac{S_{y,m}}{S_y}},$$

kde:  $S_{y,m}$  je meziskupinový součet čtverců,

$S_y$  je celkový součet čtverců.

Dosadíme-li do vztahu dostaneme 
$$P = \sqrt{\frac{2654,85}{2878,4}} = \sqrt{0,922} = 0,96.$$



**Míra závislosti**

Pokud hodnotu korelačního poměru umocníme, dostáváme poměr determinace  $P^2 = 0,922$ . Hodnoty determinačního poměru blízké 1 svědčí o *vysoké závislosti výnosů petržele na použitém typu hnojiva*.

\*

### ŘEŠENÝ PŘÍKLAD 7-6



Na hladině významnosti 0,05 (0,01) testujte, zda názory respondentů na spokojenost s nakupováním v obchodním domě závisí na jejich příjmech. K dispozici máte počítačový výstup statistického softwaru SPSS.

*Názory respondentů*

			Sum of Squares	df	Mean Square	F	Sig.
H.CELK* PRIJEM	Between (Combined)		87,955	4	21,989	12,697	,000
	Within Groups		238,982	138	1,732		
	Total		326,937	142			

### Řešení příkladu

Protože signifikantní hodnota je 0,000...; což je menší než 0,05 (než 0,01); zamítáme nulovou hypotézu o nezávislosti příjmu a hodnocením spokojenosti s nakupováním v obchodním domě. Lze tedy tvrdit, že názory respondentů na spokojenost s nakupováním jsou závislé na jejich příjmech.

\*

### ŘEŠENÝ PŘÍKLAD 7-7



Bylo vybráno 6 řidičů, z nichž každý absolvoval s každým typem benzínu jednu jízdu. Na hladině významnosti 0,05 testujte, je-li průměrná spotřeba paliva závislá na typu použitého benzínu a na tom, který řidič s vozem jel.

*Spotřeba paliva*

Typ benzínu	Řidiči						Průměry
	A	B	C	D	E	F	
Aral	7,5	6,9	7,9	7,3	6,9	7,8	<b>7,38</b>
Shell	7,6	7,2	7,5	8	7,3	8,2	<b>7,63</b>
Benzina	7,2	8,1	7,8	7,6	7,8	6,9	<b>7,57</b>
Slovnaft	7	7,3	7,2	7,5	8,2	7,7	<b>7,48</b>
<b>Průměry</b>	<b>7,33</b>	<b>7,38</b>	<b>7,6</b>	<b>7,6</b>	<b>7,55</b>	<b>7,65</b>	<b>7,5</b>

## Řešení příkladu

Zkoumáme tedy závislost průměrné spotřeby (znak  $Y$ ) na typu použitého benzínu (znak  $X_1$ ) a na řidiči (znak  $X_2$ ). (V tabulce jsou již doplněny podmíněné průměry a celkový průměr.)

Znak  $X_1$  má  $k = 4$  skupiny, znak  $X_2$  má  $r = 6$  bloků.

Pro faktor  $X_1$  formulujeme hypotézu:

$H_0$  faktor  $X_1$  je neúčinný,

$H_1$  faktor  $X_1$  je účinný; tj. průměrná spotřeba závisí na použitém druhu benzínu.

Pro faktor  $X_2$  formulujeme hypotézu:

$H_0$  faktor  $X_2$  neúčinkuje,

$H_1$  faktor  $X_2$  účinkuje; tj. průměrná spotřeba závisí na řidiči vozu.

Výpočet jednotlivých součtů:

$$S_{ym} = r \sum_{i=1}^4 (\bar{y}_{i.} - \bar{y})^2 = 6[(7,38 - 7,5)^2 + \dots + (7,48 - 7,5)^2] = 0,21$$

$$S_{yb} = k \sum_{j=1}^6 (\bar{y}_{.j} - \bar{y})^2 = 4[(7,33 - 7,5)^2 + \dots + (7,38 - 7,5)^2] = 0,36$$

Potřebujeme znát i hodnotu součtu  $S_{yv}$ , z praktického hlediska je však výhodnější vypočítat hodnotu součtu  $S_y$ . Součet  $S_{yv}$  pak dopočteme ze vztahu  $S_y = S_{ym} + S_{yv} + S_{yb}$ .

$$S_y = \sum_{i=1}^4 \sum_{j=1}^6 (y_{ij} - \bar{y})^2 = (7,5 - 7,5)^2 + (6,9 - 7,5)^2 + \dots + (8,2 - 7,5)^2 + (7,7 - 7,5)^2 = 3,79$$

Po dosazení  $S_{yv} = 3,22$ .

$$\text{Testové kritérium pro 1. hypotézu: } F = \frac{\frac{S_{ym}}{k-1}}{\frac{S_{yv}}{(k-1)(r-1)}} = \frac{\frac{0,21}{3}}{\frac{3,22}{3 \cdot 5}} = 0,33$$

V tabulce kritických hodnot  $F$ -rozdělení najdeme  $F_{(3,15)}(0,05) = 3,29$ . Protože  $0,33 < 3,29$ , nelze zamítnout  $H_0$ , což znamená, že **použitý typ benzínu nemá na průměrnou spotřebu statisticky významný vliv**.

$$\text{Testové kritérium pro 2. hypotézu: } T = \frac{\frac{S_{yb}}{r-1}}{\frac{S_{yv}}{(k-1)(r-1)}} = \frac{\frac{0,36}{5}}{\frac{3,22}{3 \cdot 5}} = 0,34$$

V tabulce kritických hodnot  $F$ -rozdělení najdeme  $F_{(5,15)}(0,05) = 2,9$ . Protože  $0,34 < 2,9$ , nelze zamítnout  $H_0$ , což znamená, že **volba řidiče nemá na průměrnou spotřebu vliv**.

\*

**KONTROLNÍ TEST 7**

Ano či ne?

- 7.1** Je-li sledovaný znak ovlivňován dvěma faktory, hovoříme o dvojném třídění.
- 7.2** Jestliže hodnota testovacího kritéria pro faktor A leží v oboru přijetí, pak hodnota testovacího kritéria pro faktor B leží vždy v kritickém oboru.
- 7.3** V případě dvojného třídění hovoříme o tzv. Latinských čtvercích.
- 7.4** V případě trojného třídění má kritická hodnota Fischerova rozdělení tři stupně volnosti.
- 7.5** Analýza rozptylu zkoumá závislost kvantitativního znaku na jednom nebo více kvantitativních znacích.

Doplňte

- 7.6** Hodnotu testovacího kritéria porovnááme s kritickou hodnotou, která má \_\_\_\_\_ rozdělení.
- 7.7** Je-li sledovaný znak ovlivňován třemi faktory, hovoříme o tzv. \_\_\_\_\_.
- 7.8** Tabulka zachycuje výsledek analýzy rozptylu dvojného třídění zpracovaný v Excelu. Doplňte chybějící hodnoty A, B, C.

Zdroj variability	SS	Rozdíl	MS	F	F krit
Řádky	0,25	1	0,25	1	<b>C</b>
Sloupce	20,25	1	20,25	<b>B</b>	161,4462
Chyba	<b>A</b>	1	0,25		
Celkem	20,75	3			

- 7.9** Provedte analýzu rozptylu u dvojného třídění.

Řešte

Faktor /Blok	1	2
<b>A</b>	7	13
<b>B</b>	36	44
<b>C</b>	2	18

- 7.10** Testujte hypotézu na hladině významnosti  $\alpha = 0,05$ , že úspěšnost testu není ovlivněna typem testu ani tím, zda test skládá muž nebo žena. (Dvojné třídění)

	muž	žena
test 1	18	20
test 2	16	15

- 7.11** Tři dělníci mají naprosto stejnou pracovní náplň – balení výrobků. Počet zabalených výrobků za každou vybranou hodinu je zaznamenán v tabulce. Vypočítejte tabulku analýzy rozptylu a rozhodněte na hladině významnosti  $\alpha = 0,05$ , která z blokových proměnných (hodina, muž) ovlivňuje počet zabalených výrobků?

**7.12** Zkoumá se kvalita povrchu na ořez v závislosti na

- a) typu látky (A, B, C, D),
- b) čísel vzorku (1, ..., 4),
- c) místě kontroly (I, ..., IV).

Vypočítejte  $S, S_T, S_B, S_C, S_D$ , je-li dáno

	I	II	III	IV
1	A 261	D 200	C 324	B 207
2	B 223	A 269	D 215	C 350
3	C 290	B 224	A 263	D 228
4	D 306	C 288	B 252	A 264

### ŘEŠENÍ KONTROLNÍHO TESTU 7



Ano či ne?

**7.1** ano

**7.2** ne

**7.3** ne

**7.4** ne

**7.5** ne

**7.6** Fischerovo

**7.7** Latinských čtvercích

**7.8** A=0,25; B=81; C=161,4

Doplňte

**7.9** Hodnota  $Y$  závisí na typu faktoru, ale nezávisí na typu bloku (pro 0,05)

**7.10** Nebyla prokázána závislost úspěšnosti na typu testu ( $5,4 < 161,4$ ) ani na pohlaví ( $0,11 < 161,4$ ), neboť obě hodnoty testových kritérií leží v oboru přijetí.

Řešte

**7.11** Počet zabalených výrobků závisí tom, kdo výrobky balí.

Zdroj variability	Součty čtverců	Stupně volnosti	Průměrné čtverce	$F$	Kritická hodnota
Hodiny	18	2	9	3,6	6,94
Dělníci	78	2	39	15,6	6,94
Reziduál	10	4	2,5		
Celkem	106	8			

**7.12**  $S_B = 2178,5$   $S_C = 1338,5$   $S_T = 17866,5$   $S_R = 6905,5$   $S_D = 28289$

## 8 ÚPLNÉ FAKTOROVÉ PLÁNY

### CÍLE KAPITOLY ÚPLNÉ FAKTOROVÉ PLÁNY

Po úspěšném a aktivním absolvování této KAPITOLY

#### Budete umět:

- sestavit úplný a částečný faktorový plán
- vypočítat efekt faktorů a dokázat testovat statistickou významnost tohoto efektu,
- sestavit rovnici modelu experimentu.

*Budete  
umět*

#### Budete schopni:

- definovat pojem plán experimentu,
- graficky vyhodnotit efekt faktorů

*Budete  
schopni*

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny.

#### PRŮVODCE STUDIEM 11



Pod pojmem experimentovat se dále rozumí měnit obvyklé pracovní podmínky s cílem nalézt nejlepší pracovní postupy a současně získat hlubší poznatky o vlastnostech výrobku a výrobního procesu.

Formulaci „nejlepší pracovní postupy“ lze interpretovat takto: Označíme-li sledovaný ukazatel kvality  $Y$  (resp. ukazatele kvality  $Y_1, Y_2, \dots, Y_k$ ) a faktory, které jej ovlivňují  $A, B, C, \dots$  se mohou pohybovat na různých úrovních  $A_1, A_2, A_3, \dots$  pro faktor  $A$ ;  $B_1, B_2, B_3, \dots$  pro faktor  $B$  atd. pak cílem plánování experimentů je rozhodnout, které z faktorů  $A, B, C, \dots$  významně ovlivňují ukazatel kvality  $Y$  a určit optimální úrovně významných faktorů. Kapitola je opět doplněna řešenými příklady.

### 8.1 Základy experimentování a oblasti použití

Experimentování představuje vlastně testování kombinací různých hodnot (úrovní) faktorů, o nichž si myslíme, že mají vliv na odezvu (charakteristiku jakosti). Odezva je veličina, pomocí které vyjadřujeme výsledky experimentu. Protože testování všech variant představuje neúměrně velký počet zkoušek, používají se pro experimentování zkrácené metody, které zkoumají pouze určitou frakci všech možných kombinací.

**DEFINICE 8-1****Faktor**

*Faktor* neboli parametr je nezávislá návrhová proměnná, která ovlivňuje charakteristiku jakosti. Jak již bylo uvedeno v úvodu, symbolicky označujeme faktory velkými tiskacími písmeny, tj.  $A, B, C$  atd. a jejich úrovně pro experiment označujeme jako  $A_1$  (faktor  $A$  na první úrovni),  $A_2$  (faktor  $A$  na druhé úrovni), atd.

Rozlišujeme dva základní druhy faktorů:

- regulovaný faktor – je návrhová proměnná, o které si myslíme, že ovlivňuje odezvu a je přitom začleněna do experimentu. Hodnotu proměnné můžeme a zároveň chceme nastavit a udržovat.
- šumový faktor – je faktor, který negativně ovlivňuje odezvu. Takový faktor nemůžeme nebo nechceme při vlastní aplikaci nastavit a udržovat na požadované hodnotě, ale můžeme to provádět během experimentu.

**DEFINICE 8-2****Interakce faktorů**

Pod pojmem *interakce faktorů* si představme kombinovaný účinek dvou faktorů, tzn. Účinek jednoho faktoru je závislý na hodnotě nastavení druhého faktoru. Interakci dvou faktorů zapisujeme symbolicky jako  $AB$ .

**DEFINICE 8-3****Oblasti použití experimentování**

Uvedme oblasti, ve kterých se experimentování používá. Jsou to tyto oblasti: analytická simulace, návrh a vývoj výrobku, návrh a vývoj procesu, zlepšování procesu, testování a validace, řešení problémů s jakostí ve výrobě a pro analýzu a zlepšování systému měření.

## 8.2 Experimentální procedura

Experimentální proceduru můžeme rozčlenit do čtyř kroků:

plánování experimentů (brainstorming), návrh experimentů, provedení experimentů a jejich analýza. Nyní se věnujme jednotlivým částem experimentální procedury.

**DEFINICE 8-4****Plánování experimentů**

Prvním krokem při *plánování experimentů* je ustanovení experimentálního týmu. Do týmu by měli být zahrnuti zástupci všech oddělení, která ovlivňují produkt nebo proces. Velikost týmu by však neměla přesahovat rozumnou míru a měla by se pohybovat v rozmezí 2 – 15 lidí.

Brainstormingová sezení jsou věnována velkému množství otázek a na každou z nich musí experimentální tým nalézt správnou odpověď. Jedná se o určení cíle experimentu, definování charakteristiky jakosti a výběr faktorů a jejich úrovní.



**DEFINICE 8-5**

Výsledkem plánování experimentů je definovaný cíl, kterého chceme dosáhnout, a charakteristika jakosti, která je měřítkem pro posouzení, zda cíle bylo dosaženo. Kromě toho známe potenciální faktory, které nepravděpodobněji ovlivňují danou charakteristiku jakosti, jejich úrovně. Při návrhu experimentů musíme přiřadit faktory do tabulky. Vlastní postup při návrhu experimentů budeme demonstrovat na konkrétním příkladu *Pružina*.

**Návrh  
experimentů**

**DEFINICE 8-6**

Experimenty můžeme provádět buďto v laboratorních nebo přímo v provozních podmínkách. Při experimentování ve výrobě můžeme narazit na „střety zájmů“ mezi potřebným množstvím produkce na jedné straně a mezi potřebným časem na experimenty, který snižuje vlastní produktivní čas výroby, na straně druhé. V praxi je obvyklé řešit tento problém tak, že se experimenty provádějí mimo pracovní dobu, např. na zvláštních nočních směnách, o sobotách a nedělích, apod. Kdykoli je to možné, měli bychom experimenty provádět v náhodném pořadí.

**Provádění  
experimentů**

**DEFINICE 8-7**

Analýza výsledků experimentů spočívá především v nalezení kombinace faktorů, která dává nejlepší výsledek a dále v určení relativního podílu jednotlivých faktorů na jakosti výstupu. Na závěr provádíme odhad výsledku při optimálních podmínkách, který musíme verifikovat ověřovacími experimenty.

**Analýza  
experimentů**

**PRŮVODCE TEXTEM**

V následujícím příkladu, který je proložen výkladem, se seznámíme se sestavením tabulky faktorů, s úplným faktorovým plánem, s pojmem kódovaná proměnná.

**ŘEŠENÝ PŘÍKLAD 8-1**

Sleduje se, kolik stlačení ( $Y$ ) vydrží pružina až do zničení v závislosti na těchto faktorech:

$L$  = délka pružiny,

$G$  = tloušťka drátu,

$T$  = typ materiálu.

Má se zjistit, které faktory jsou rozhodující pro životnost pružiny. Pro test významnosti těchto faktorů popř. interakce faktorů použijme hladinu významnosti  $\alpha = 0,05$ .

**Pružina**

## Řešení příkladu

Sestavíme tabulku faktorů a jejich uvažovaných úrovní (tab.1) :

faktor	označení	dolní úroveň	horní úroveň
		-	+
délka pružiny	$L$	10 cm	15 cm
tloušťka drátu	$G$	5 mm	7 mm
materiál	$T$	$A$	$i$

Tab. 1 Seznam faktorů a jejich úrovní

Existuje více způsobů jak sestavit plán, podle kterého se budou provádět jednotlivé pokusy. Mezi nejpoužívanější plány patří úplný faktorový plán, který v daném případě vypadá takto:

pokus	$L$	$G$	$T$	$Y$
1	10	5	$A$	
2	15	5	$A$	
3	10	7	$A$	
4	15	7	$A$	
5	10	5	$B$	
6	15	5	$B$	
7	10	7	$B$	
8	15	7	$B$	

Tab.2 Plán experimentu

V uvedené tabulce je  $Y$  je výsledek pokusu.

Plán experimentu je výhodnější psát pomocí této symboliky: Je-li každý z faktorů uvažován na dvou úrovních, pak dolní úroveň bude značena -1 (resp. jen „-“) a horní úroveň +1 (resp. „+“). Tabulka 2 potom bude mít tvar

pokus	$L$	$G$	$T$	$Y$
1	-1	-1	-1	
2	+1	-1	-1	
3	-1	+1	-1	
4	+1	+1	-1	
5	-1	-1	+1	
6	+1	-1	+1	
7	-1	+1	+1	
8	+1	+1	+1	

Tab.3 Plán experimentu v kódovaných proměnných

Přepočtení původních proměnných na tzv. *kódované proměnné* se může provést nejen pro krajní hodnoty  $x_{\max}$  ( $= +1$ ) a  $x_{\min}$  ( $= -1$ ) takto:



**Přepočtení  
původních  
proměnných  
na  
kódované**

$$x_c = \frac{x_0 - \frac{x_{\max} + x_{\min}}{2}}{\frac{x_{\max} - x_{\min}}{2}}$$

(8-1)

kde  $x_0$  je proměnná v původních jednotkách,

$x_c$  je kódovaná proměnná,

$x_{\max}$  horní úroveň  $x$ ,

$x_{\min}$  dolní úroveň  $x$ .

Například přepočtení  $L$  pro dolní hodnotu 10 je

$$L_c = \frac{10 - \frac{15 + 10}{2}}{\frac{15 - 10}{2}} = -1,$$

a pro horní úroveň  $G = 7$  bude kódovaná hodnota

$$G_c = \frac{7 - \frac{7 + 5}{2}}{\frac{7 - 5}{2}} = +1$$

Počet pokusů, ze kterých je sestaven úplný experiment, se vypočítá při  $k$  faktorech pomocí vztahu

$$n = 2^k,$$

takže zde, při  $k = 3$  faktorech, je počet pokusů (řádků)  $n = 2^3 = 8$ .

Proto má tabulka 8 řádků.

Jestliže je plánem experimentu stanoveno, za jakých podmínek se provádí jednotlivé pokusy, je možné provést celý experiment a zaznamenat hodnoty sledovaného ukazatele  $Y$ . V našem případě byl každý pokus opakován dvakrát. Výsledky jsou v tabulce 4:

pokus	faktor	faktor	faktor	výsledek	výsledek	průměr
	$L$	$G$	$T$	$Y_1$	$Y_2$	$\bar{Y}$
1	-	-	-	77	81	79
2	+	-	-	98	96	97
3	-	+	-	76	74	75
4	+	+	-	90	94	92
5	-	-	+	63	65	64
6	+	-	+	82	86	84



**Počet  
pokusů  
úplného  
experimentu**



**Příklad  
Pružina -  
pokračování**

7	-	+	+	72	74	73
8	+	+	+	92	88	90

Tab.4 Výsledky opakovaných pokusů

Sestavením tabulky 4 skončily přípravné a experimentální práce. Dále následují výpočty, jejichž cílem bude stanovit, které z faktorů ovlivňují významným způsobem životnost pružiny  $Y$ . Vzhledem k tomu, že pro určení optimální úrovně faktorů a pro sestavení modelu je důležité také vědět, které dvojice faktorů mají vzájemně významnou interakci, počítá se také vliv interakcí na  $Y$ .

Znaménka ve sloupcích  $LG$ ,  $LT$ ,  $GT$ ,  $LGT$  se získají jako součin znamének v odpovídajících sloupcích, jak je uvedeno v následující tabulce 5.

pokus	$L$	$G$	$T$	$LG$	$LT$	$GT$	$LGT$
1	-	-	-	+	+	+	-
2	+	-	-	-	-	+	+
3	-	+	-	-	+	-	+
4	+	+	-	+	-	-	-
5	-	-	+	+	-	-	+
6	+	-	+	-	+	-	-
7	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+

Tab.5 Interakce faktorů

\*

## 8.3 Efekt (vliv) faktoru, významnost efektu

### 8.3.1 Efekt (vliv) faktoru

#### DEFINICE 8-8

*Efekt faktorů* se rozumí změna ukazatele kvality  $Y$ , kterou způsobí přechod tohoto faktoru z dolní úrovně (-) na horní úroveň (+).

Princip *znaménkové metody*, kterou budeme používat při výpočtu efektu faktoru, spočívá v tom, že se sečtou se hodnoty ve sloupci  $Y$ , přičemž každá hodnota má znaménko, odpovídající znaménku u příslušného faktoru v odpovídajícím řádku. Součet se vydělí  $\frac{n}{2}$ , kde  $n$  je počet pokusů.

Například pro faktor  $L$  bude

$$\text{efekt}(L) = \frac{1}{4}(-79 + 97 - 75 + 92 - 64 + 84 - 73 + 90) = 18,$$

$$\text{efekt}(T) = \frac{1}{4}(-79 - 97 - 75 - 92 + 64 + 84 + 73 + 90) = -8.$$



Znamén-  
ková me-  
toda



Podobně se postupuje u interakcí, například pro interakci  $LG$  znaménkovou metodou bude

$$\text{efekt}(LG) = \frac{1}{4}(79 - 97 - 75 + 92 + 64 - 84 - 73 + 90) = -1.$$

Posloupnosti znamének u interakcí jsou uvedeny v tabulce 5.

Dále vypočteme efekty jednotlivých faktorů a jejich interakcí a doplníme je do tabulky 6.

číslo	$L$	$G$	$T$	$LG$	$LT$	$GT$	$LGT$	$\bar{y}$
1	-	-	-	+	+	+	-	79
2	+	-	-	-	-	+	+	97
3	-	+	-	-	+	-	+	75
4	+	+	-	+	-	-	-	92
5	-	-	+	+	-	-	+	64
6	+	-	+	-	+	-	-	84
7	-	+	+	-	-	+	-	73
8	+	+	+	+	+	+	+	90
<b>efekt</b>	<b>18</b>	<b>1,5</b>	<b>-1</b>	<b>-8</b>	<b>0,5</b>	<b>6</b>	<b>-0,5</b>	

Tab.6 Efekt faktorů

Abychom zjistili, zda faktory resp. interakce faktorů jsou statisticky významné, potřebujeme ještě vypočítat rozptyl efektu faktoru.

### DEFINICE 8-9

Odhad rozptylu efektu faktoru  $\sigma_e^2$ , který je stejný pro všechny faktory, bude:

$$s_e^2 = \frac{4\sigma^2}{N}$$

kde  $N$  je celkový počet pokusů (včetně opakování), tzn. v našem případě  $N = 16$ .

V případě opakovaných pokusů se  $\sigma^2$  odhadne pomocí veličiny  $s^2$ , která se vypočítá

$$s^2 = \frac{v_1 s_1^2 + \dots + v_k s_k^2}{v_1 + \dots + v_k}$$

kde  $v_i = n_i - 1$ ,

$n_i$  je počet opakování  $i$ -tého pokusu,

$s_i^2$  je rozptyl v  $i$ -tém pokusu.

**Efekty  
faktorů  
a jejich  
interakcí**



**Rozptyl  
efektu  
faktoru  
(8-2)**

**(8-3)**

### 8.3.2 Test významnosti efektu

#### DEFINICE 8-10

Rozptyl odhadu efektu se použije k testování významnosti efektu faktorů.

Celý test vypadá takto:

1. Nulová hypotéza  $H_0$ : efekt faktoru je bezvýznamný,  
alternativní hypotéza  $H_1$ : efekt faktoru je významný.

2. Testovací kritérium

$$t = \frac{\text{efekt}}{s_e}$$

3. Kritická hodnota

$$t_{n_1+n_2+\dots+n_k-n}(\alpha),$$

kde  $n_1, \dots, n_k$  jsou počty opakování pokusů, zde  $n_i = 2$ ,  $n$  je počet pokusů bez opakování ( $n = 8$ ).

4. Závěr testu: pro

$$|t| > t_{n_1+n_2+\dots+n_k-n}(\alpha),$$

se zamítá nulová hypotéza, což znamená, že efekt (a tedy faktor) je významný.



Testování  
význam-  
nosti efek-  
tu

Vraťme se opět k příkladu *Pružina* a vypočteme potřebné hodnoty:

$$t_{n_1+n_2+\dots+n_k-n}(\alpha) = t_{16-8}(0,05) = 2,306.$$

Hodnotu veličiny  $s^2$  vypočteme ze vztahu (8-3)

$$s^2 = \frac{8+2+2+8+2+8+2+8}{8} = 5,$$

a výsledek dosadíme do vztahu (8-2) a dostáváme:

$$s_e^2 = \frac{4s^2}{N} = \frac{4 \cdot 5}{16} = 1,25,$$

$$s_e = 1,12.$$

Výpočty jsou uspořádány v tabulce.

číslo	$Y_1$	$Y_2$	$Y_1 - Y_2$	$s_i^2$	efekt	$t = \frac{\text{efekt}}{s_e}$
1	77	81	-4	8		
2	98	96	2	2	$L = 18$	16,07
3	76	74	2	2	$G = 1,5$	1,34
4	90	94	-4	8	$LG = -1,0$	-0,89
5	63	65	-2	2	$T = -8,0$	-7,14
6	82	86	-4	8	$LT = 0,5$	0,45
7	72	74	-2	2	$GT = 6,0$	5,36
8	92	88	4	8	$LGT = -0,5$	-0,45

Tab.7 Testování významnosti efektu



Příklad  
Pružina -  
pokračo-  
vání

Rozptyl v jednotlivých pokusech se v případě dvou opakování vypočítá pomocí vztahu

$$s_i^2 = \frac{(Y_1 - Y_2)^2}{2}.$$

Kritickou hodnotu 2,306 převyšuje v absolutní hodnotě testovací kritérium faktorů  $L$ ,  $T$  a interakce  $GT$ . To jsou tedy významné faktory a interakce, ostatní faktory vliv na životnost pružiny nemají.

## 8.4 Grafické hodnocení efektu faktoru, grafy interakcí

### 8.4.1 Grafické hodnocení efektu faktorů

#### DEFINICE 8-11

Pokud se neprovádí opakování jednotlivých pokusů nelze použít výše uvedenou metodu a používá se grafická metoda k určování významných faktorů. V grafu se na vodorovnou osu vynáší efekt a na svislou osu pravděpodobnost

$$P_i = \frac{100(i - 0,5)}{m}$$

kde  $i = 1, 2, \dots, m$ ,  $m$  je počet faktorů a interakcí.

Za významné se považují ty faktory, které se nacházejí výrazně **mimo hlavní linii**.

Při použití grafické metody je užitečné sestavit do pomocné tabulky, ve které jsou efekty ve druhém řádku uspořádány vzestupně.

Číslo	1	2	3	4	5	6	7
Efekt	-8,0	-1,0	-0,5	0,5	1,5	6,0	18
Faktor	$T$	$LG$	$LGT$	$LT$	$G$	$GT$	$L$
$P_i$	7,14	21,42	35,71	50	64,29	78,57	92,86

Tab. 8 Grafické hodnocení významnosti efektu (pružina)

Graf, sestavený z údajů ve druhém řádku (vodorovná osa) a čtvrtého řádku (svislá osa) je na obr.8-1.

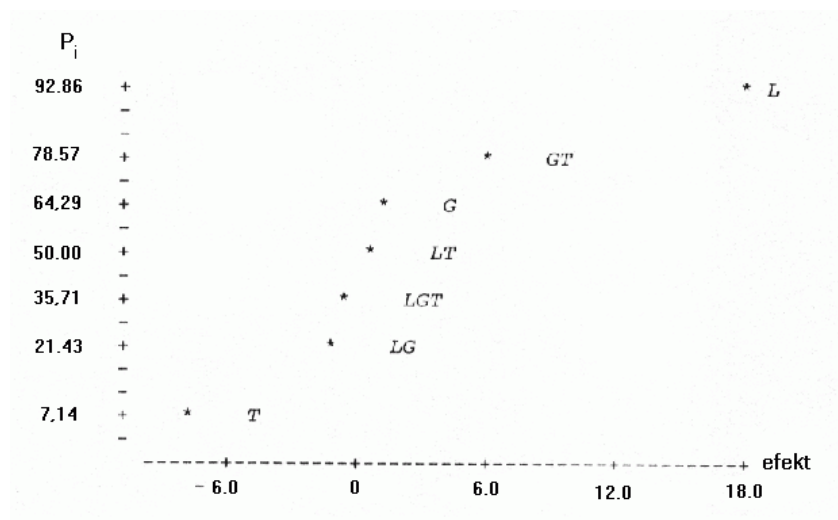
V grafu je vidět, že mimo hlavní linii jsou ty faktory, u kterých testovací kritérium překročilo kritickou hodnotu. Jsou to faktory  $L$  (nejvýrazněji),  $T$  a interakce  $GT$ .



(8-4)



**Příklad**  
Pružina –  
Grafické  
hodnocení  
význam-  
nosti efek-  
tu

Obr. 8-1:  
Grafické  
určení  
vlivných  
faktorů  
(pružina)

Obrázek 8-1: Grafické určení vlivných faktorů (pružina)

### 8.4.2 Grafy interakcí

#### DEFINICE 8-12



Pro významné interakce se sestavují grafy, umožňující diskusi o optimální úrovni jednotlivých faktorů.

Tak například pro významnou interakci  $GT$  sestavíme graf vlivu  $G$  na ukazatel kvality  $Y$  v závislosti na faktoru  $T$ . Z tabulky 6 vybereme údaje, které odpovídají příslušným úrovním faktorů  $G$  a  $T$ . Hodnoty zaznamenáme do tabulky 9, kterou doplníme o průměry.

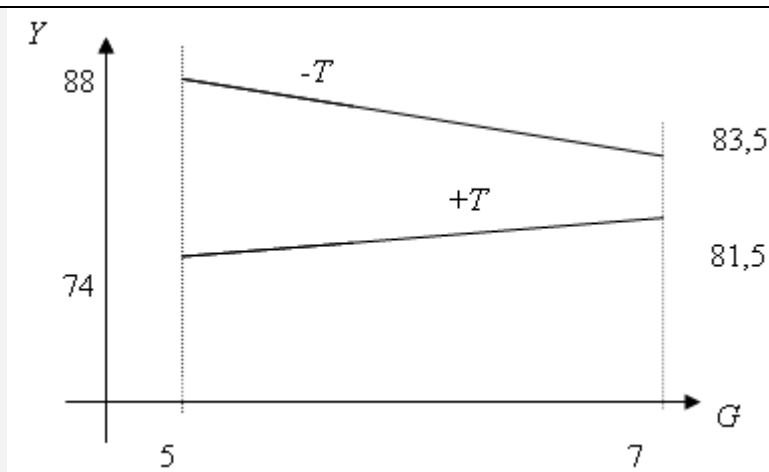
$G$	$T$			průměr
-	-	79	97	88
+	-	75	92	83,5

$G$	$T$			průměr
-	+	64	84	74
+	+	73	90	81,5

Tab.9 Určení krajních bodů grafu interakce

Spojením bodů  $[5; 88]$ ,  $[7; 83,5]$  dostaneme situaci pro dolní úroveň faktoru  $T$ , spojením bodů  $[5; 74]$ ,  $[7; 81,5]$  znázorníme horní úroveň faktoru  $T$ .





Obrázek 8-2: Vliv G na Y v závislosti na T

Obr. 8-2:  
Vliv G na Y  
v závislosti  
na T

Z grafu je vidět, že pro maximalizaci Y je nejlepší T na dolní úrovni (- T) a že změna G neovlivňuje výrazně Y, protože úsečky jsou přibližně rovnoběžné s osou x.

## 8.5 Model experimentu $2^3$

### DEFINICE 8-13

Jakmile je stanoven efekt faktorů a jejich interakcí, je možné sestavit model experimentu. Neúplný kvadratický model experimentu  $2^3$  s faktory A, B, C má tvar

$$Y = b_0 + b_1A + b_2B + b_3C + b_{12}AB + b_{13}AC + b_{23}BC + b_{123}ABC.$$

Koeficienty  $b_1, b_2, \dots, b_{123}$  se vypočítají jako polovina příslušného efektu (resp. interakce), u kterého jsou a absolutní člen  $b_0 = \bar{Y}$ . V modelu experimentu jsou zařazeny jen vlivné faktory a interakce. Zde např.

$$Y = 81,75 + 9L - 4T + 3GT.$$

Přepočteme kódované proměnné na původní proměnné pomocí vztahu (8-1) a dostáváme

$$Y = 81,75 + 9 \frac{L_0 - 12,5}{2,5} - 4T + 3 \frac{G_0 - 6}{1} T.$$

Po úpravě vychází

$$Y = 36,75 + 3,6L_0 - 22T + 3G_0T$$

Model experimentu má mnohostranné použití. Mezi nejvýznamnější patří:

1. určení lokálně optimálních hodnot faktorů,
2. stanovení směru dynamického plánování,
3. predikce ukazatele kvality Y.



Rovnice  
modelu  
experimentu  
(8-5)

Použití  
modelu  
experimentu

**SHRnutí KAPITOLY ÚPLNÉ FAKTOROVÉ PLÁNY****Shrnutí kapitoly**

Tato kapitola nás seznámila s plánováním experimentu. Lze sestavit úplný faktorový plán nebo, pokud jsme okolnostmi omezeni, částečný faktorový plán. Každý faktor nějakým způsobem ovlivňuje ukazatele kvality  $Y$ , jedná se o efekt faktoru, kterou způsobí přechod tohoto faktoru z dolní úrovně (-) na horní úroveň (+). Lze testovat statistickou významnost tohoto efektu, jak bylo uvedeno na konkrétním příkladu.

Plánované experimenty se řídí *plánem experimentu*.

Plán experimentu stanovuje

- počet pokusů, ze kterých se experiment skládá,
- podmínky, za kterých se jednotlivé pokusy uskuteční,
- pořadí pokusů.

Z uvedeného je zřejmé, že se zde rozlišuje význam pojmů

*pokus* = zjištění hodnoty ukazatele kvality za určitých, předem plánovaných, podmínek výroby a

*experiment* = systém všech pokusů.

Cílem plánování experimentů je

**a.** rozhodnout, které z faktorů ( $A, B, C, D, \dots$ ) významným způsobem ovlivňují ukazatel kvality  $Y$ ,

**b.** určit optimální úroveň významných faktorů s ohledem na optimalizaci a stabilitu  $Y$ .

Stabilitou v bodě b) se rozumí, že  $Y$  je optimální i při různých vnějších podmínkách (prostředí, zacházení apod.). Hovoří se o tzv. *robustnosti* výrobku.

Poznali jsme pojmy jako: plán experimentu, pokus, experiment, částečný faktorový plán, efekt faktorů, test významnosti efektu, grafické hodnocení efektu faktorů, model experimentu.

**Cíl plánování****ŘEŠENÝ PŘÍKLAD 8-2**

Pro faktory  $A, B$  byl sestaven úplný plán. Každý pokus se opakoval 2x. Výsledky jsou uvedeny v tabulce

$A$	$B$	$Y_1$	$Y_2$
-	-	5	6
+	-	5	5
-	+	7	6
+	+	5	4

Vypočtete:

- efekty faktorů  $A, B$  a efekt interakce faktorů  $AB$ ,
- napište rovnici modelu experimentu,
- rozptýl odhadu efektu faktorů,
- testujte, zda jsou faktory  $A, B, AB$  statisticky významné.

**Řešení příkladu**

a. Pro výpočet efektu faktorů  $A, B$  a efekt interakce faktorů  $AB$  doplníme tabulku o průměry v jednotlivých pokusech:

$A$	$B$	$Y_1$	$Y_2$	$\bar{Y}$
-	-	5	6	5,5
+	-	5	5	5
-	+	7	6	6,5
+	+	5	4	4,5

Pro efekty faktorů  $A, B$  a efekt interakce  $AB$  dostáváme:

$$e_A = \frac{1}{2}(-5,5 + 5 - 6,5 + 4,5) = -1,25$$

$$e_B = \frac{1}{2}(-5,5 - 5 + 6,5 + 4,5) = 0,25$$

$$e_{AB} = \frac{1}{2}(5,5 - 5 - 6,5 + 4,5) = -0,75.$$

b. Rovnice modelu experimentu

$$Y = 5,375 - \frac{1,25}{2}A + \frac{0,25}{2}B - \frac{0,75}{2}AB$$

c. Rozptyl odhadu efektu faktoru:  $s_e^2 = \frac{4\sigma^2}{N}$ ,

$$s^2 = \frac{1,0,5 + 1,0 + 1,0,5 + 1,0,5}{4} = 0,375$$

$$s_e^2 = \frac{4 \cdot 0,375}{8} = 0,1875$$

Směrodatná odchylka odhadu efektu faktoru je  $s_e = 0,433$ .

d. Testujeme nulovou hypotézu:

$H_0$  : efekt faktoru (interakce) není statisticky významný,  
proti alternativní hypotéze:

$H_1$  : efekt faktoru (interakce) je statisticky významný.

Testové kritérium  $t = \frac{\text{efekt}}{s_e}$ .

Pro jednotlivé faktory dostáváme:

$$t_A = 2,88 \quad t_B = 0,577 \quad t_{AB} = -1,73$$

Hodnoty porovnááme s kritickou hodnotou Studentova rozdělení

$$t_{N-n}(0,05) = t_{8-4}(0,05) = 2,776$$

( $N$  označuje počet provedených pokusů včetně opakování,  $n$  počet pokusů bez opakování)

Protože  $|-2,88| > 2,776$      $|0,577| < 2,776$      $|-1,73| < 2,776$   
je statisticky významný pouze faktor A.

\*

**ŘEŠENÝ PŘÍKLAD 8-3**

Pomocí grafické metody zjistěte, které z faktorů jsou významné, jestliže jsou dostupné následující hodnoty, které jsou uvedeny v tabulce.

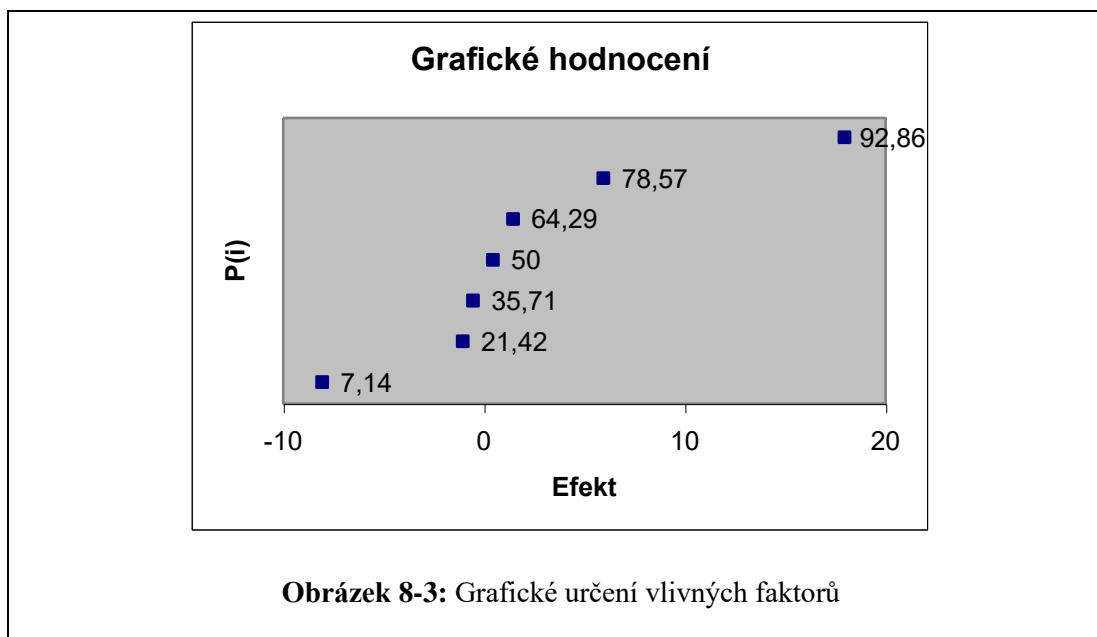
**Významnost faktorů**

$i$	1	2	3	4	5	6	7
<b>Efekt</b>	<b>-8</b>	<b>-1</b>	<b>-0,5</b>	<b>0,5</b>	<b>1,5</b>	<b>6</b>	<b>18</b>
Faktor	C	AB	ABC	AC	B	BC	A
$P_i$							

**Řešení příkladu**

Dopočteme poslední řádek v tabulce podle vztahu:  $P_i = \frac{100(i-0,5)}{m}$ ,  $i = 1, 2, \dots, m$   $m$  je počet faktorů a interakcí, zde  $m = 7$ .

$i$	1	2	3	4	5	6	7
<b>Efekt</b>	<b>-8</b>	<b>-1</b>	<b>-0,5</b>	<b>0,5</b>	<b>1,5</b>	<b>6</b>	<b>18</b>
Faktor	C	AB	ABC	AC	B	BC	A
$P_i$	<b>7,14</b>	<b>21,42</b>	<b>35,71</b>	<b>50</b>	<b>64,29</b>	<b>78,57</b>	<b>92,86</b>



**Obr. 8-3:**  
Grafické určení vlivných faktorů

Na hlavní linii neleží hodnoty: 92,86; 7,14; 78,57. Což odpovídá těm faktorům, které jsou nejvýznamnější:  $A, C, BC$ .

\*

## KONTROLNÍ TEST 8



Ano či ne?

- 8.1 Plán experimentu stanovuje jen pořadí pokusů.  
 8.2 Úplný experiment, který má 4 faktory, je sestaven z 8 pokusů.  
 8.3 Efekt faktoru nabývá jen kladných hodnot.  
 8.4 Při testování významnosti efektu faktorů se porovnává testové kritérium s kritickou hodnotou Fischerova rozdělení.  
 8.5 Při grafické metodě určování významnosti faktorů se za významné považují ty faktory, které se nacházejí významně mimo hlavní linii.

Doplňte

- 8.6 Experiment je systém všech \_\_\_\_\_  
 8.7 Počet pokusů, ze kterých je sestaven úplný experiment (při 3 faktorech), je roven \_\_\_\_\_  
 8.8 Nulová hypotéza při testování významnosti efektu faktorů zní: Efekt faktoru je \_\_\_\_\_  
 8.9 Grafická metoda určování významných faktorů bývá používána tehdy, pokud se neprovádí \_\_\_\_\_  
 8.10 Při grafické metodě určování významných faktorů se v grafu na vodorovnou osu vynášší efekt faktoru a na svislou osu pravděpodobnost, která je dána vztahem \_\_\_\_\_

Řešte

- 8.11 Doplňte do tabulky kódované hodnoty pro úplný plán:

Pokus	$A$	$B$
1		
2		
3		
4		

- 8.12 Pro faktory  $A, B$  byl sestaven úplný plán. Každý pokus se opakoval dvakrát. Výsledky jsou v tabulce.

$A$	$B$	$Y_1$	$Y_2$
-	-	2,3	2,6
+	-	3,1	2,9
-	+	3	3,5
+	+	1,9	2,2

Vypočtěte:

- efekt faktoru  $A$ ,  $B$  a  $AB$ ,
- model experimentu,
- rozptyl odhadu efektů faktorů

- 8.13** Testujte na hladině významnosti 0,05, zda jsou efekty faktorů  $A$ ,  $B$  a efekt interakce  $AB$  statisticky významné.
- 8.14** Nakreslete graf interakce faktorů  $A$  a  $B$  z příkladu 8.14. Znázorněte vliv faktoru  $A$  na  $Y$  v závislosti na  $B$ .

### ŘEŠENÍ KONTROLNÍHO TESTU 8



Ano či ne?

**8.1** ne

**8.2** ne

**8.3** ne

**8.4** ne

**8.5** ano

**8.6** pokusů

**8.7**  $2^3 = 8$

**8.8** bezvýznamný

**8.9** opakování jednotlivých pokusů

**8.10**  $P_i = \frac{100(i-0,5)}{m}$ , kde  $i = 1, 2, \dots, m$  a  $m$  je počet faktorů a interakcí

Doplňte

**8.11**

Pokus	$A$	$B$
1	-	-
2	+	-
3	-	+
4	+	+

Řešte

**8.12**

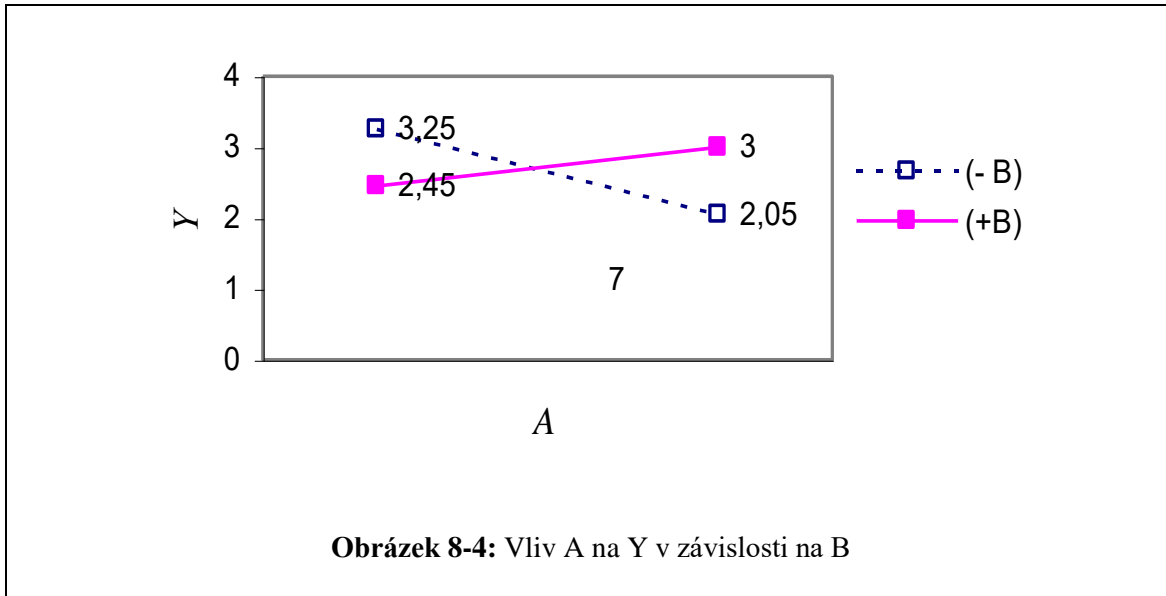
a. efekt( $A$ ) = -0,325; efekt( $B$ ) = -0,075; efekt( $AB$ ) = -0,875

b.  $Y = 2,69 - 0,1625 A - 0,0375 B - 0,4375 AB$

c.  $s^2 = 0,0587$ ;  $s_e = 0,0294$

**8.13** Statisticky významný je pouze efekt interakce  $AB$ , protože  $|-5,11| > 2,776$

**8.14** Z grafu je vidět, že pro maximální hodnotu  $Y$  je nejlepší  $B$  na horní úrovni a změna faktoru  $A$  ovlivňuje  $Y$ , protože úsečky nejsou rovnoběžné.



Obr. 8-4:  
Vliv A na Y  
v závislosti  
na B

## 9 ČÁSTEČNÝ FAKTOROVÝ EXPERIMENT SE DVĚMA ÚROVNĚMI

### CÍLE KAPITOLY ČÁSTEČNÝ FAKTOROVÝ EXPERIMENT SE DVĚMA ÚROVNĚMI

Po úspěšném a aktivním absolvování této KAPITOLY

#### Budete umět:

- definovat pojem stupeň snížení,
- rozdělit částečné faktorové plány podle stupně snížení
- vlastnosti operací s faktory
- definovat pojem zaměnitelné dvojice

**Budete umět**

#### Budete schopni:

- byli schopni určit významné faktory v polovičním plánu grafickou metodou

**Budete schopni**

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny.

#### PRŮVODCE STUDIEM 12



V předcházející kapitole jsme se seznámili s úplnými faktorovými plány. Ne vždy je možné sestavit úplný faktorový plán, ať už z důvodů finančních, časových nebo jiných. V takových případech se přistupuje k částečnému plánování. V této kapitole se dovíte, že podle stupně snížení se částečné faktorové plány dělí na plány s nejnižším, nejvyšším snížením a na středové plány. V této kapitole se budeme zabývat problematikou polovičních plánů.

#### DEFINICE 9-1



V úplném faktorovém experimentu se sestavuje plán experimentu pro každý faktor.

U částečného faktorového experimentu se plán sestaví jen pro několik faktorů (*hlavní faktory*) a ostatní (*vedlejší faktory*) se vyjádří jako

jejich kombinace. Tím se dosáhne snížení počtu pokusů. Je-li  $2^k$  označení pro úplný experiment, kde

$2$  = počet úrovní faktorů,

$k$  = počet faktorů,

pak  $2^{k-p}$  je částečný faktorový experiment,  $p$  označuje stupeň snížení.

**Částečný faktorový experiment**



Chceme-li například v plánu  $2^7$ , který představuje  $n = 128$  pokusů, snížit počet pokusů na polovinu, tj.

$$\frac{2^7}{2} = 2^{7-1},$$

dostáváme částečný faktorový experiment, který představuje

$$n = 2^{7-1} = 64$$

pokusů, tedy p o l o v i n u. Je to nejmenší možné snížení počtu pokusů. Plány se snížením počtu pokusů na polovinu se nazývají *poloviční plány*.

Stupeň snížení  $p$  může být i vyšší než 1, například

$$2^{7-4},$$

kde bude jen  $n = 8$  pokusů. To je pro  $k = 7$  faktorů největší možné snížení. Největší možné snížení vychází z pravidla, podle kterého počet pokusů nesmí být menší než počet faktorů. Musí tedy platit  $n \geq k$ . V uvedeném případě  $k = 7$  a  $n = 2^{7-4} = 8$ . Například pro  $k = 15$  faktorů je nejvyšší možné snížení  $2^{15-11}$ , neboť  $n = 16$  a  $k = 15$ . Pokud bychom provedli ještě vyšší snížení, např.  $2^{15-12}$ , potom  $k = 15$ , ale  $n = 8$ , takže  $n < k$ . Mezi plány s nejmenším (polovičním) a největším snížením počtu pokusů může být ještě řada možností snížení. Takové plány se nazývají *středové*.

Například mezi  $2^{7-1}$  a  $2^{7-4}$  jsou plány  $2^{7-2}$ ,  $2^{7-3}$ .

Částečné faktorové plány lze tedy rozdělit na

- plány s nejnižším snížením, tzv. poloviční plány,
- plány s nejvyšším snížením,
- plány se snížením mezi a) a b), tzv. středové plány.

**Poloviční plány**



## PRŮVODCE TEXTEM



Mezi nejvýznamnější patří poloviční plány, kterými se budeme zabývat v další subkapitole.

## 9.1 Poloviční plány

Nejprve budeme definovat pojmy, se kterými budeme dále pracovat.

Tyto pojmy pak budou osvětleny na konkrétním příkladu *Barvivo*, který je řešen v této kapitole.

### DEFINICE 9-2



Označme  $I$  faktor, obsahující jen „+“. Takový faktor se nazývá *jednotkový*.

$$A.A = I$$

$$A.I = I.A = A$$

$$(A.B).C = A.(B.C)$$

$$A.B = B.A$$

Předpokládejme, že  $A, B, C, D, E$  jsou faktory, pro které se má sestavit poloviční plán. Je třeba určit 4 *hlavní faktory* (například  $A, B, C, D$ ), pro které se sestaví úplný plán a zbývající (*vedlejší*) faktor  $E$  se vyjádří jako jejich kombinace, například  $E = ABCD$ .

**Operace s faktory  
Vlastnosti operací s faktory**

**DEFINICE 9-3**

Každá kombinace faktorů tvoří *slovo*. Slovo se skládá z písmen (faktorů). Počet písmen ve slově je *délka slova*.

**Slovo****DEFINICE 9-4**

Vztah  $E = ABCD$  se nazývá *generátor plánu*. V plánu  $2^{k-p}$  je  $p$  generátorů. Vynásobením generátoru levou stranou  $E.E = E.ABCD$  a s využitím vlastností operací s faktory dostáváme  $I = ABCDE$ .

**Generátor plánu****DEFINICE 9-5**

Slova, která jsou rovna jednotkovému faktoru  $I$ , se nazývají *definiční rovnice*. Definičních rovnic může být i více. Nejkratší slovo v definičních rovnicích je tzv. *řešení plánu* a zapisuje se k typu plánu římským číslem jako index. Zde např.  $2_V^{5-1}$ .

**Definiční rovnice****DEFINICE 9-6**

Řešení plánu je  $V$  proto, že slovo v definiční rovnici má 5 písmen (faktorů). Pomocí definiční rovnice lze najít dvojice faktorů (resp. interakcí), které tvoří stejné posloupnosti znamének a které se nazývají *zaměnitelné dvojice*. Je-li například generátor plánu ve tvaru  $E = ABCD$ , pak definiční rovnice bude  $I = ABCDE$

Zaměnitelnou interakci k interakci  $DE$  nalezneme vynásobením definiční rovnice touto interakcí

$$I = ABCDE \quad / . DE$$

$$DE.I = DE.ABCDE.$$

Odtud dostáváme

$$DE = ABC.$$

**Zaměnitelné dvojice****ŘEŠENÝ PŘÍKLAD 9-1**

Zkoumejme množství barviva  $Y$ , které zůstane na látce po absolvování testů (ve srovnání se standardním vzorkem) v závislosti těchto pěti faktorech:

$$A = \text{pH}, \quad B = \text{teplota}, \quad C = \text{koncentrace},$$

$$D = \text{dokončovací teplota}, \quad E = \text{dokončovací čas}.$$

Sestavme poloviční plán experimentu a najděme nevýznamnější faktory, které ovlivňují množství barviva na látce. Potřebné údaje jsou uvedeny v tabulce 1.

**Barvivo**

Faktor	Symbol	-	+
pH	<i>A</i>	4,5	5,5
teplota	<i>B</i>	70 <sup>0</sup> C	80 <sup>0</sup> C
koncentrace	<i>C</i>	1 g/l	3 g/l
dok.teplota	<i>D</i>	170 <sup>0</sup> C	190 <sup>0</sup> C
dok.čas	<i>E</i>	50s.	70s.

Tab.1 Faktory a jejich úrovně

**Řešení příkladu**

Sestavme tedy poloviční plán experimentu, kde *A*, *B*, *C*, *D* jsou hlavní faktory a faktor *E* je vedlejším faktorem. Výsledky pokusů pro jednotlivé úrovně faktorů jsou zaznamenány v tabulce 2.

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E = ABCD</i>	<i>Y</i>
1	-	-	-	-	+	6,4
2	+	-	-	-	-	9,9
3	-	+	-	-	-	8,1
4	+	+	-	-	+	6,6
5		-	+	-	-	9,0
6	+	-	+	-	+	5,3
7	-	+	+	-	+	-5,1
8	+	+	+	-	-	-1,0
9	-	-	-	+	-	10,6
10	+	-	-	+	+	12,7
11	-	+	-	+	+	12,9
12	+	+	-	+	-	11,2
13	-	-	+	+	+	12,4
14	+	-	+	+	-	9,7
15	-	+	+	+	-	4,1
16	+	+	+	+	+	4,0

Tab.2 Poloviční plán

Výpočet efektu faktorů a interakcí se provede stejně, jako u úplného plánu. Např. pro efekt faktoru *D* dostáváme

$$\text{efekt}(D) = \frac{1}{8}(-6,4 - 9,9 - \dots - (-1) + 10,6 + 12,7 + \dots + 4) = 4,8.$$

Ostatní hodnoty efektu faktorů a jejich interakcí jsou uvedeny v tabulce 3.

Faktor	Efekt
$A + BCDE$	0,0
$B + ACDE$	-4,4
$C + ABDE$	-5,0
$D + ABDE$	4,8
$E + ABCD$	-0,8
$AB + CDE$	0,2
$AC + BDE$	-0,6
$AD + BCE$	-0,6
$AE + BCD$	0,5
$BC + ADE$	-4,2
$BD + ACD$	1,1
$BE + ACD$	-0,2
$CD + ABE$	0,7
$CE + ABD$	-0,5
$DE + ABC$	2,4

Tab. 3 Efekt faktorů v obou polovičních plánech

Proč jsou u efektů uváděny dvojice a nikoli sólové faktory, vysvětlíme dále.

\*

## 9.2 Grafická metoda

### Řešení příkladu

Použití grafické metody je opět podobné jako u úplných plánů. Nejprve vypočteme hodnotu pravděpodobnosti ze vztahu (8-4).

$i$	1	2	3	4	5	6
Faktor	$C + ABDE$	$B + ACDE$	$BC + ADE$	$E + ABCD$	$AD + BCE$	$AC + BDE$
Efekt	-5	-4,4	-4,2	-0,8	-0,6	-0,6
$P_i$	3,3	10	16,6	23,3	30	36,6

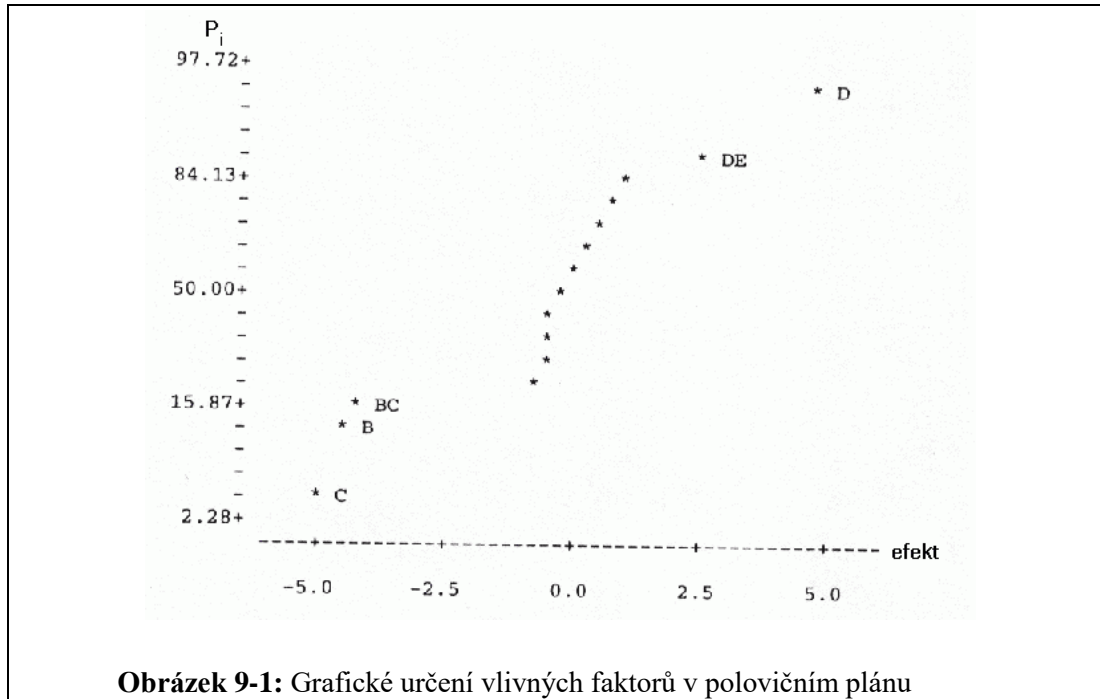
$i$	7	8	9	10	11	12
Faktor	$CE + ABD$	$BE + ACD$	$A + BCDE$	$AB + CDE$	$AE + BCD$	$CD + ABE$
Efekt	-0,5	-0,2	-0,0	0,2	0,5	0,7
$P_i$	43,3	50	56	63,3	70	76,6

Příklad Barvivo - pokračování

$i$	13	14	15
Faktor	$BD + ACD$	$DE + ABC$	$D + ABCE$
Efekt	1,1	2,4	4,8
$P_i$	83,3	90	96,6

Tab. 4 Výpočet pravděpodobností  $P_i$

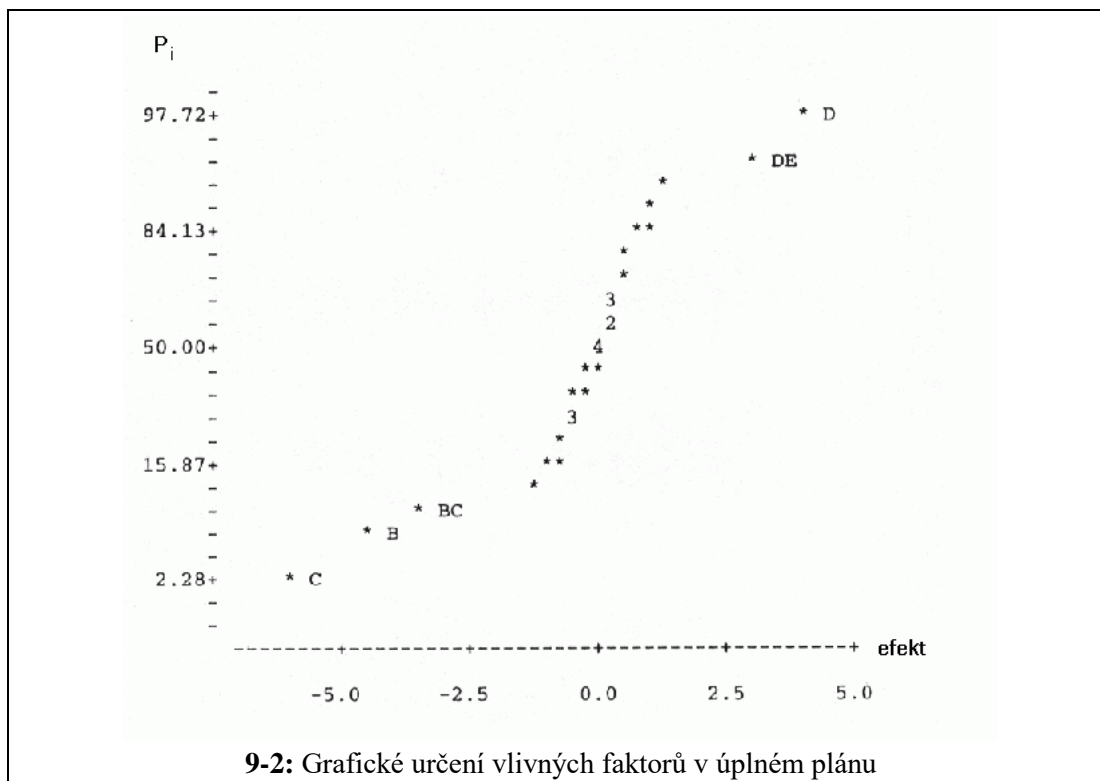
Z těchto hodnot sestrojíme graf, kde na osu  $x$  nanášíme efekt a na osu  $y$  hodnotu pravděpodobnosti  $P_i$ .



Obr. 9-1: Grafické určení poloviční plán

Obrázek 9-1: Grafické určení vlivných faktorů v polovičním plánu

Pokud bychom sestavili úplný plán a sestrojili graf pro vyhodnocení vlivných faktorů, dostali bychom stejný výsledek, jako u polovičního plánu, jak je patrné z obrázku 9-2:



Obr. 9-2: Grafické určení – úplný plán

Porovnáme-li grafy v obr. 9-1 a obr. 9-2 je zřejmé, že poloviční plán dává stejné hodnocení jako úplný plán. Znamená to tedy, že snížením počtu pokusů na polovinu nedošlo ke ztrátě informací a tím ke změně výsledků. Kdybychom však porovnali numerické hodnoty efektů faktorů a jejich interakcí v úplném a v polovičním plánu, zjistíme, že nejsou stejné. Navíc jsou v tabulce uváděny efekty pro s o u č e t, například  $A + BCDE$ ,  $B + ACDE$  apod. Ve sloupci „faktor“ v tabulce 3 jsou uvedeny součty proto, že snížením počtu pokusů dochází k tomu, že některé posloupnosti znamének u interakcí faktorů jsou stejné. Například (podle tab.2)



<i>ABC</i>	-	+	+	-	+	-	-	+	-	+	+	-	+	-	-	+
<i>DE</i>	-	+	+	-	+	-	-	+	-	+	+	-	+	-	-	+

Znaménková metoda

Jsou-li u interakcí  $ABC$  a  $DE$  stejné posloupnosti znamének je jasné, že také efekt (zde počítaný znaménkovou metodou) je stejný. Součty efektů v tab. 3 jsou tedy uváděny proto, že vypočítaný efekt patří oběma interakcím. Neznamená to ale, že na každou připadá polovina!

Kolik celkového efektu připadá na jednotlivé sčítance nelze zjistit. Využívá se ale poznatek, že čím delší je „slovo“, tím menší má vliv. Snahou proto je dostat do kombinací krátké slovo (sólóvý faktor) s co nejdelším. To lze ovlivnit volbou definičních rovnic.



\*

**ŘEŠENÝ PŘÍKLAD 9-2**

Mějme pět faktorů  $A, B, C, D, E$ . Faktor  $E$  je vedlejším faktorem. Faktor  $E$  je možné vyjádřit mnoha způsoby. Porovnejme tyto dva:

- a.  $E = AB$ ,
- b.  $E = ABCD$ .

**Řešení příkladu**

Příslušné definiční rovnice jsou

- a.  $I = ABE$
- b.  $I = ABCDE$

V případě a) máme plán  $2_{III}^{5-1}$ , v případě b) plán  $2_V^{5-1}$ . Tento druhý plán je lepší, neboť má řešení plánu  $V$ , takže při hledání zaměnitelných dvojic, například k  $A$ , máme

- a.  $A = BE$
- b.  $A = BCDE$

V případě b) tvoří zaměnitelnou dvojici s  $A$  interakce více faktorů, která má proto menší podíl na celkovém efektu, u více než dvou faktorů dokonce tak malý, že se zanedbává a pracuje se jen (zde) s faktorem  $A$ . Tím se usnadňuje diskuse k vypočítanému efektu: přesto, že efekt náleží dvěma faktorům, podstatná část připadá na „čistý“ faktor a prakticky zanedbatelná část na interakci.

\*

**SHRNUTÍ KAPITOLY ČÁSTEČNÝ FAKTOROVÝ EXPERIMENT SE DVĚMA ÚROVNĚMI**

**Shrnutí kapitoly**

Po prostudování této kapitoly jste se dověděli, že u částečného faktorového experimentu se plán sestaví jen pro několik faktorů – to jsou hlavní faktory a ostatní – vedlejší faktory se vyjádří jako jejich kombinace, čímž se dosáhne snížení počtu pokusů. Seznámili jste se také s grafickou metodou pro určení vlivných faktorů v polovičním plánu a také víte, že volbou definiční rovnice lze ovlivnit kombinaci daných „slov“, protože se využívá poznatku, že čím delší je „slovo“, tím menší má vliv.

Poznali jsme pojmy jako: hlavní faktory, vedlejší faktory, poloviční plány, zaměnitelné dvojice, definiční rovnice.

**ŘEŠENÝ PŘÍKLAD 9-3**

Pro faktory  $A, B, C, D$  byl sestaven poloviční plán.

- a. doplňte chybějící údaj v tabulce,

**Doplnění hodnoty**

b. pomocí grafické metody rozhodněte, který faktor(y) je statisticky významný.

Vypočtené hodnoty jsou v následující tabulce 5.

	efekty	$i$	$P_i$
$A + BCD$	1	3	35,7143
$B + ACD$	-0,5	2	21,4286
$C + ABD$	-4	1	
$D + ABC$	3	4	50
$AB + CD$	9	6	78,5714
$AC + BD$	6	5	64,2857
$AD + BC$	17	7	92,8571

Tab. 5 Hodnoty efektů a pravděpodobností

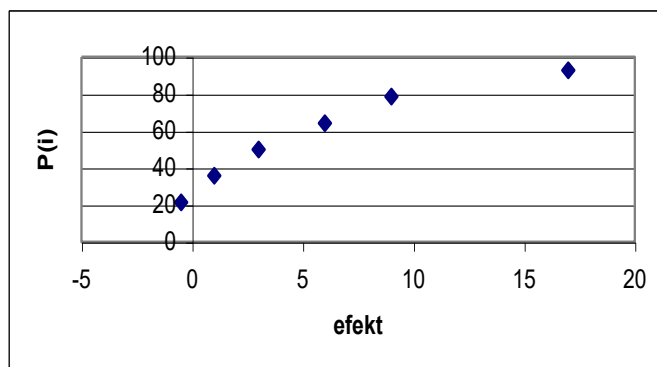
### Řešení příkladu

a. Chybějící hodnotu v tabulce vypočteme na základě vztahu (8-4)

$$P_i = \frac{100(i - 0,5)}{m}$$

Dostáváme 
$$P_1 = \frac{100(1 - 0,5)}{7} = 7,14.$$

b. Z grafu na obr. 9-3 vyplývá, že statisticky významné jsou interakce  $AD$ ,  $BC$  a faktor  $C$ .



Obr. 9-3: Grafické hodnocení významnosti faktorů

Obrázek 9-3: Grafické hodnocení významnosti faktorů

\*



**ŘEŠENÝ PŘÍKLAD 9-4**

Víme-li, že dvojice  $AB$  je zaměnitelná s  $CD$ , co platí pro efekty interakcí  $AB$  a  $CD$  ?

**Řešení příkladu**

Pro efekty interakcí  $AB$  a  $CD$  platí v případě zaměnitelných dvojic rovnost.

\*

**ŘEŠENÝ PŘÍKLAD 9-5**

Pro faktory  $A, B, C, D$  byl sestaven částečný (poloviční) plán s generátorem  $D = ABC$ . Výsledky experimentu jsou v tabulce 6.

*Poloviční plán*

- Vypočtete efekt jednotlivých faktorů a interakce trojic a čtveřice.
- Pomocí interakcí 3 a 4 odhadněte rozptyl efektu faktorů.
- Napište definiční rovnici a najděte všechny zaměnitelné dvojice.
- Proveďte grafické hodnocení efektu faktoru.

$A$	$B$	$C$	$D = ABC$	$Y$	$ABC$	$ABD$	$BCD$	$ACD$	$ABCD$
-	-	-	-	77					
+	-	-	+	67					
-	+	-	+	64					
+	+	-	-	51					
-	-	+	+	64					
+	-	+	-	53					
-	+	+	-	73					
+	+	+	+	67					

Tab.6 Výsledky experimentu v polovičním plánu

**Řešení příkladu**

- V tabulce doplňte znaménka a vypočtete efekty znaménkovou metodou:

$$e_A = \frac{1}{4}(-77 + 67 - 64 + 51 - 64 + 53 - 73 + 67) = -10 = e_{BCD};$$

$$e_B = -1,5 = e_{ACD}; \quad e_C = -0,5 = e_{ABD}; \quad e_D = 2 = e_{ABC};$$

$$e_{ABCD} = 129$$

- Pokud se pokusy neopakují, pak  $s^2$  se vypočte jako průměr druhých mocnin efektu nejvyšších interakcí.



$$s^2 = \frac{2^2 + (-0,5)^2 + (-10)^2 + (-1,5)^2 + 129^2}{5} = 3349,5$$

$$s_e^2 = \frac{4 \cdot s^2}{n} = 1674,75 \quad \Rightarrow \quad s_e = 40,9$$

c. Po vyjádření faktoru  $D = ABC$  má definiční rovnice tvar:

$$I = ABCD.$$

Zaměnitelné dvojice jsou  $AB, CD; AC, BD; AD, BC$ .

d. Sestavíme tabulku 7 a pak na osu  $x$  naneseme efekty faktorů a na osu  $y$  pravděpodobnosti dané vztahem:

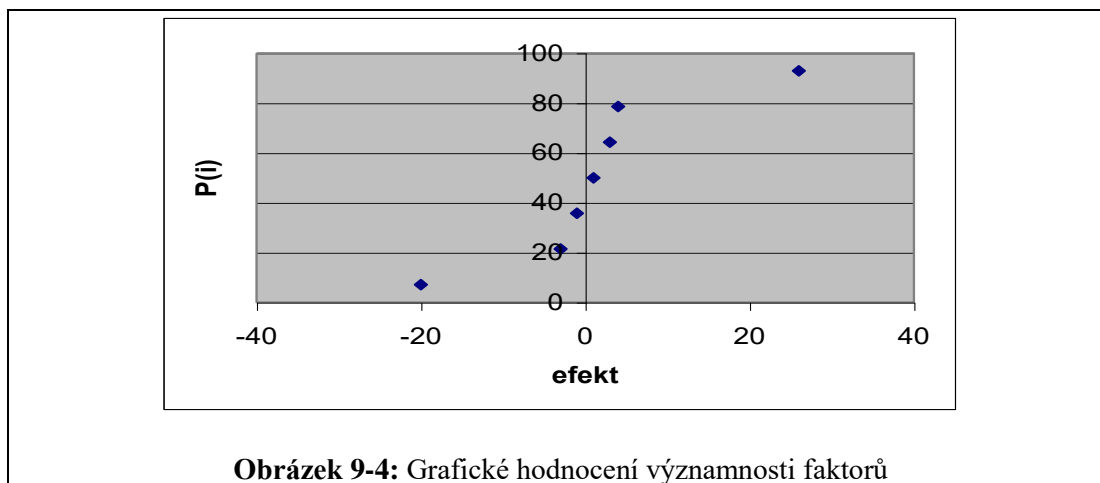
$$P_i = \frac{100(i - 0,5)}{m}, \quad i = 1, 2, \dots, m$$

Faktor	Efekt	$i$	$P_i$
A + BCD	-20	1	7,14
B + ACD	-3	2	21,42
C + ABD	-1	3	35,71
D + ABC	4	6	78,57
AB + CD	1	4	50
AC + BD	3	5	64,28
AD + BC	26	7	92,85

Tab.7 Efekty faktorů v obou polovičních plánech

Připomínáme, že v tabulce jsou uváděny efekty pro součet. Ve sloupci „faktor“ v tabulce jsou uvedeny součty proto, že snížením počtu pokusů dochází k tomu, že některé posloupnosti znamének u interakci faktorů jsou stejné.

Obecně platí, že čím delší je slovo, tím menší má efekt.



Obr. 9-4: Grafické hodnocení významnosti faktorů

Z grafu vidíme, že mimo hlavní linii leží body zobrazující faktory interakcí  $AD$ ,  $BC$  a faktor  $A$  s interakcí trojice  $BCD$ . Protože platí, že čím delší slovo, tím menší efekt faktor má, je statisticky významný faktor  $A$  a interakce dvojic  $AD$ ,  $BC$ .

\*

**KONTROLNÍ TEST 9**

Ano či ne?

- 9.1** U částečného faktorového experimentu se plán sestaví pro každý faktor.  
**9.2** Pro operace s faktory neplatí pro násobení komutativní zákon.  
**9.3** Je-li generátor plánu ve tvaru  $D=BC$ , pak definiční rovnice bude  $I=ABCD$ .  
**9.4** Jsou-li u interakcí  $ABC$  a  $DE$  stejné posloupnosti znamének, pak také efekt (počítaný znaménkovou metodou) je stejný.  
**9.5** Při sestavování polovičních plánů se jeden z faktorů vyjádří jako interakce všech ostatních.

Doplňte

- 9.6** Vedlejší faktory se vyjadřují jako \_\_\_\_\_ hlavních faktorů.  
**9.7** Částečné faktorové plány lze rozdělit na plány s \_\_\_\_\_ snížením, plány s \_\_\_\_\_ snížením a plány \_\_\_\_\_.  
**9.8** Pro faktor  $A$  platí:  $AI = IA = \underline{\quad}$ , kde  $I$  je jednotkový faktor.  
**9.9** Dvojice faktorů, které tvoří stejné posloupnosti znamének se nazývají \_\_\_\_\_.  
**9.10** Numerické hodnoty efektů faktorů a jejich interakcí v úplném a v polovičním plánu \_\_\_\_\_ stejné.

Řešte

- 9.11** Napište poloviční plán pro faktory  $A, B, C, D$ . Generátor volte  $B=ACD$ . Vypočítejte efekt faktoru  $C$ , jsou-li výsledky pokusů, které se dvakrát opakovaly tyto:

výsledky první série: 10,11,14,12,12,10,13,14

výsledky druhé série: 11,12,12,8,14,12,13,14.

- 9.12** Pro faktory  $A, B, C, D$  se má sestavit plán  $2^{4-1}$ . Generátor volte podle vlastního uvážení. Vypočítejte odhad  $s^2$ ,  $s_e^2$ , jestliže se prováděly dvě série (opakování) pokusů s výsledky:

v první sérii: 10,11,11,8,9,10,11,10

ve druhé sérii: 9,10,10,9,8,11,12,9.

- 9.13** Pro faktory  $A, B, C, D$  byl sestaven poloviční plán.
- doplňte chybějící údaje v tabulce 8,
  - pomocí grafické metody rozhodněte, který faktor je statisticky významný

	efekty	$i$	$P_i$
$A + BCD$	1		
$B + ACD$	-8		
$C + ABD$	-10		
$D + ABC$	4		
$AB + CD$	9		
$AC + BD$	7		
$AD + BC$	5		

### ŘEŠENÍ KONTROLNÍHO TESTU 9



Ano či ne?

9.1 ne

9.2 ne

9.3 ano

9.4 ano

9.5 ne

9.6 kombinace

9.7 nejnižším, nejvyšším, středové

9.8 A

9.9 zaměnitelné

9.10 nejsou

9.11  $e_C = 1$

9.12  $s^2 = \frac{8,0,5}{8} = 0,5$        $s_e^2 = \frac{4,0,5}{16} = 0,125$

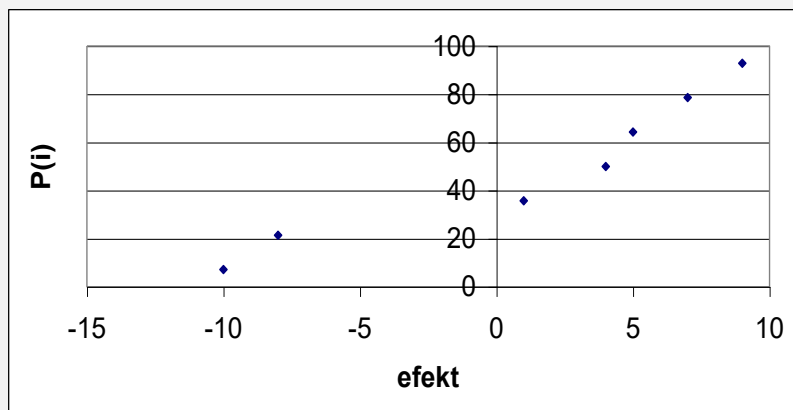
9.13 a.

Doplňte

Řešte

	efekty	$i$	$P_i$
$A + BCD$	1	3	35,71
$B + ACD$	-8	2	21,42
$C + ABD$	-10	1	7,14
$D + ABC$	4	4	50
$AB + CD$	9	7	92,85
$AC + BD$	7	6	78,57
$AD + BC$	5	5	64,28

b. Významné jsou faktory B, C.



## 10 TAGUCHIHO METODY: ZTRÁTOVÁ FUNKCE

### CÍLE KAPITOLY TAGUCHIHO METODY: ZTRÁTOVÁ FUNKCE

Po úspěšném a aktivním absolvování této KAPITOLY

#### Budete umět:

- definovat ztrátovou funkci,
- správně zvolit ztrátovou funkci pro různé typy tolerance,
- graficky znázornit průběh ztrátové funkce

**Budete  
umět**

#### Získáte:

- znalost vlastností ztrátové funkce

**Získáte**

#### Budete schopni:

- vypočítat prakticky zaměřené úlohy, ve kterých se ztrátová funkce využívá.

**Budete  
schopni**

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny

#### PRŮVODCE STUDIEM 13



Taguchiho metody (autorem je dr. Genichi Taguchi) lze rozdělit na metody používané přímo ve výrobním procesu (on line) a v předvýrobních etapách (off line). V kapitolách 10, 11 budeme hovořit pouze o metodách on line. Jejich základním prvkem je tzv. *ztrátová funkce* (loss function). V této kapitole vysvětlíme její smysl, konstrukci a použití. V první části kapitoly je uvedena definice a vlastnosti ztrátové funkce, druhá část se zabývá ztrátovou funkcí pro různé typy tolerance. V závěru kapitoly jsou uvedeny řešené příklady a samozřejmě Vás čekají i otázky typu „Ano či ne?“, otázky doplňovací a neřešené příklady. Vše jistě hravě zvládnete, pokud pečlivě prostudujete tuto kapitolu.

Taguchiho metody jsou založeny na principu plánování experimentů za účelem dosažení bližších optimálních kvalitativních charakteristik pro specifické cíle. Jejich význam je často snižován akademickou obcí pro jejich technické nedostatky, které však mohou být zlepšeny užitím metodologie odpovědného povrchu – Response surface methodology.

Taguchiho metody nejsou pouze statistickou aplikací plánování experimentů, ale zahrnují také integraci statistického plánování experimentů do inženýrského procesu. Skutečná síla Taguchiho metod přichází z jednoduchosti jejich uskutečnění. Často jsou aplikovány techniky v japonských výrobních k zlepšení jejich produktu a procesů. Cílem není pouze optimalizovat rozhodující předmětnou funkci, tak, jak jsou často užívány v USA, ale též redukce citlivosti inženýrských plánů na nekontrolovatelné faktory či hluky.

Spojení jakosti s náklady pomocí Taguchiho ztrátové funkce (Taguchi loss function) bylo hlavní výhodou v jakostním inženýrství, stejně tak při schopnosti plánovat náklady. Tvrdí se, že Taguchiho metody přispěly až 80% japonským ziskům na jakosti. Taguchiho metodám se také říká

mohutné (silné) plánování.

## 10.1 Definice a vlastnosti ztrátové funkce

### DEFINICE 10-1

Dříve, než uvedeme definiční rovnici, graf a vlastnosti této funkce, si musíme důkladně uvědomit předpoklady, které by měly být splněny, chceme-li ztrátovou funkci použít.

Vyslovíme je v jakýchsi čtyřech výchozích postulátech a je na každém, kdo chce dále uvedené metody používat aby se zamyslel, zda jsou v jeho podmínkách přijatelné:

1. U každého výrobku je sledována jeho určitá charakteristika (např. rozměr, váha, mechanické, chemické, estetické nebo jiné vlastnosti), podle které posuzujeme jeho kvalitu.
2. Tato charakteristika má stanovenou jistou optimální hodnotu  $T$ , tzv. cílovou hodnotu (*target value*).
3. Nekvalita se projevuje odchylkami od  $T$ .
4. Jakákoliv odchylka od  $T$  představuje určitou finanční ztrátu, která se projeví u *odběratele* zvýšenými náklady na provoz, údržbu, opravy, ekologii a pod.

Prvé tři předpoklady nejsou neobvyklé. Novým prvkem však je, že podle čtvrtého předpokladu nepovažuje Taguchi výrobky pohybující se v mezích tolerance za stejně kvalitní a bezztrátové. Naopak, jakákoliv odchylka od  $T$  přináší ztráty. Jejich vyčíslení je smyslem ztrátové funkce.

Matematické vyjádření ztrátové funkce má tvar

$$L(Y) = k(Y - T)^2$$

$T$  = cílová hodnota charakteristiky kvality,

$Y$  = skutečně dosažená úroveň charakteristiky kvality,

$L(Y)$  = ztráta způsobená odchylkou od  $T$ ,

$k$  = konstanta.

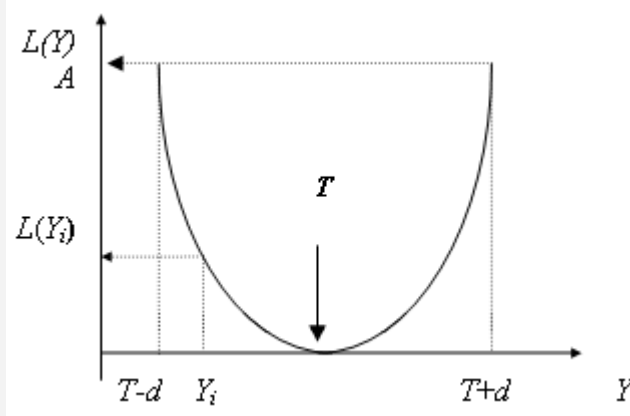
Grafem této funkce je parabola podle obr.10-1, hodnota parametru  $d$  představuje funkční toleranci.



Předpoklady ztrátové funkce



Rovnice ztrátové funkce (10-1)



Obr. 10-1: Ztrátová funkce

Obrázek 10-1: Ztrátová funkce

Z obr.10-1 je vidět, že jestliže  $Y$  dosáhne nebo překročí dolní toleranční hranici, tj. je-li

$$Y \leq T - d \quad \text{resp.} \quad Y \geq T + d \quad \text{tzn.} \quad d \leq |Y - T|,$$

pak  $L(Y) = A$ .

Můžeme proto psát:

$$A = kd^2$$

(10-2)

Protože parametry

$d$  (funkční tolerance)

$A$  (mezní ztráta = ztráta při překročení  $d$ )

jsou obvykle známy, lze použít rovnici (10-2) k výpočtu neznámé konstanty

$$k = \frac{A}{d^2}.$$

### ŘEŠENÝ PŘÍKLAD 10-1

Napište rovnici ztrátové funkce, je-li funkční tolerance  $d = 5$  a mezní ztráta  $A = 2$

#### Řešení příkladu

Podle vztahu (10-2) dostáváme  $k = \frac{2}{5^2} = 0,08$

a rovnice ztrátové funkce  $L(Y) = 0,08(Y - T)^2$ .

Z hlediska teorie pravděpodobnosti je  $Y$  (skutečně dosažená hodnota ukazatele jakosti) náhodná proměnná, která má nejčastěji normální rozdělení  $N(\bar{Y}, s^2)$ .

Často nás více než ztráta  $L(Y)$ , odpovídající konkrétní odchylce  $Y$  od  $T$ , zajímá **průměrná ztráta**, kterou označíme  $E(L)$ . Vzorec pro zjištění  $E(L)$  můžeme získat z rovnice (10-1) tak, že vypočítáme střední hodnotu

$$E(L) = E|k(Y - T)^2| = kE(Y - T)^2 = ks^2,$$

(10-3)

za předpokladu, že  $E(Y) = T$ , což prakticky znamená, že průměr ze skutečně dosažených hodnot sledovaného ukazatele kvality  $Y$  je roven žádané hodnotě  $T$ .

Je-li však

$$E(Y) \neq T \quad \Rightarrow \quad E(L) = ks^2 + k(\bar{Y} - T)^2$$

(10-4)

Máme tedy celkem tři modifikace ztrátové funkce:

**a.** definiční rovnice

$$L(Y) = k(Y - T)^2,$$

**b.** rovnice pro výpočet konstanty  $k$

$$A = kd^2,$$

Modifikace  
ztrátové  
funkce



c. rovnice pro určení průměrné ztráty

$$E(L) = ks^2; \quad E(L) = ks^2 + k(\bar{Y} - T)^2.$$

Ztrátovou funkci lze využít k mnohem komplexnějšímu vyčíslení nákladů na jakost tím, že uvažíme **všechny** ztráty související se zajišťováním jakosti, kde je zahrnuta cena za opravy, náklady na kontrolu, ztráty za zmetky, za nepřesnosti měření a nepřesnosti ve výrobě.

Této problematice je věnována kapitola 11.

Ztrátová funkce může být při řešení uvedených problémů aplikována nejen pro případ jediného ukazatele jakosti, ale i ve složitějších situacích, kdy  $Y$  je závislé na mnoha dalších faktorech  $X_1, X_2, \dots, X_k$  a dokonce i tehdy, když ukazatel jakosti nelze měřit, např. při posuzování estetických vlastností apod.

\*

## 10.2 Ztrátová funkce pro různé tolerance

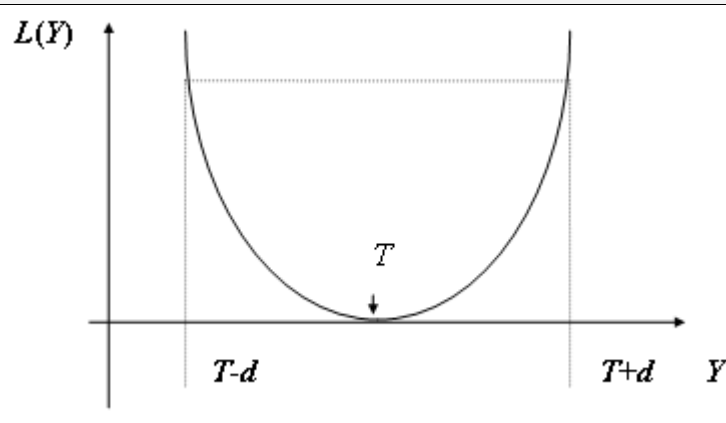
Protože se v následujících příkladech vyskytují různé typy tolerancí, uvedeme nejprve jejich klasifikaci. U různých typů tolerance jsou také příslušné grafy ztrátových funkcí.

Podle toho, co je v dané situaci považováno za optimální cílovou hodnotu  $T$ , rozlišujeme tyto typy tolerance:

### DEFINICE 10-2



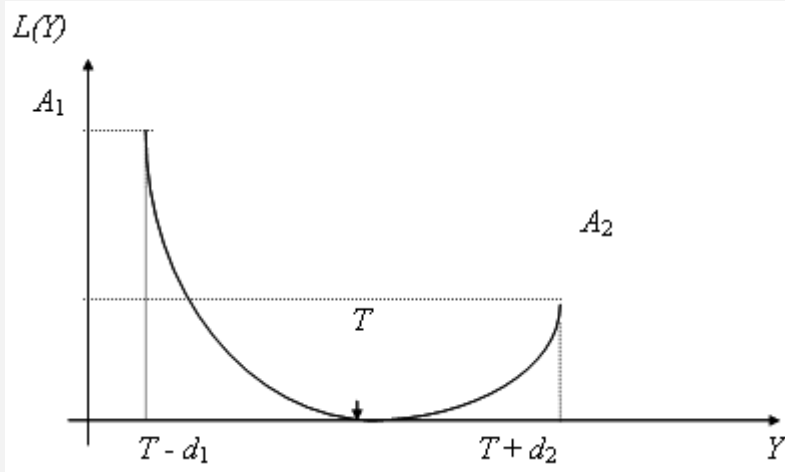
Ideálem je dosažení cílové hodnoty  $T$ .



Obrázek 10-2: Symetrická N-tolerance

1. tolerance typu N (Nominal)  
Obr. 10-2: Symetrická N-tolerance

Píšeme  $T \pm d$ ,  $d$  = tolerance,  $(T-d, T+d)$  = toleranční interval. Tato tolerance nemusí být symetrická, jak ukazuje obr.10-3.



Obrázek 10-3: Nesymetrická N-tolerance

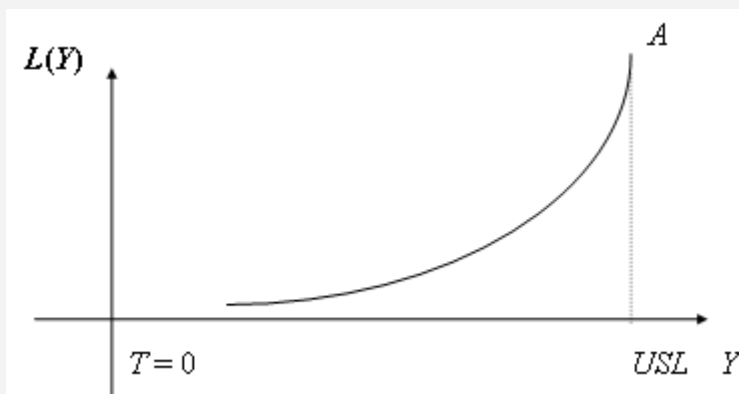
Zde je toleranční interval  $(T - d_1, T + d_2)$ . Z grafu vidíme, že tolerance  $d_1$  a  $d_2$  jsou různé, stejně jako  $A_1$  a  $A_2$ .

Například nedodržení průměru nad  $d_2$  lze opravit, kdežto pod  $d_1$  nikoliv, takže ztráta  $A_1$  je větší než  $A_2$ .

Obr. 10-3:  
Nesymetrická N-tolerance

### DEFINICE 10-3

Tolerance typu S (smaller):  $Y$  je tím lepší, čím je menší. Ideálem je  $T = 0$ .



Obrázek 10-4: S-tolerance

Příkladem veličiny  $Y$  s tolerancí S může být například drsnost povrchu, nebo nečistota v ovzduší, kde je stanovena jen horní přípustná hranice a samozřejmě platí, že čím menší hodnoty je dosaženo, tím lépe.



Tolerance  
typu S  
(Smaller)  
Obr. 10-4:  
S-tolerance

**DEFINICE 10-4**

Tolerance typu  $L$  (larger):  $T$  je tím lepší, čím je větší. Ideálem je  $T = \infty$ .

V případě  $L$  tolerance se průměrná ztráta počítá podle vzorce

$$E(L) = A \cdot d_y^2 \cdot s_z^2,$$

kde

$$s_z^2 = \frac{1}{n} \sum \frac{1}{Y^2}.$$

Při konstrukci ztrátové funkce jsme vycházeli ze čtyř postulátů, jejichž splnění je nutnou podmínkou pro aplikace této funkce. Připomeňme, že podle čtvrtého postulátu přináší nedodržení cílové hodnoty  $T$  odběrateli ztráty. Nyní se zaměříme na finanční vyjádření těchto ztrát s použitím ztrátové funkce. Jedná se o její nezákladnější použití. Další možné aplikace budou uvedeny v následujících kapitolách.



Tolerance  
typu  $L$   
(Larger)

(10-5)

(10-6)

**ŘEŠENÝ PŘÍKLAD 10-2**

Při výrobě hřídelí je předepsaný rozměr 150mm a tolerance je 4mm. Nedodržení tolerance způsobí ztrátu 40 Kč. Určete průměrnou ztrátu a porovnejte ztráty za nekvalitu u dvou výrobců: první se spokojí s dodržением tolerance, druhý usiluje o maximální přiblížení k optimální hodnotě  $T$ .



Hřídel –  
dva výrobc  
ci

**Řešení příkladu**

Jsou známy tyto parametry:

$A$  (ztráta při překročení  $d$ ) = 40 Kč,

$d$  (funkční tolerance) = 4mm,

$T$  (cílová hodnota) = 150mm.

Pro ztrátovou funkci určíme nejprve konstantu  $k$  ze vztahu (10-2):

$$k = \frac{A}{d^2} = \frac{40}{4^2} = 2,5.$$

Průměrná ztráta za nedodržení  $T$  podle (10-3) bude:

$$E(L) = k s^2 = 2,5 s^2$$

Protože nemáme k dispozici výsledky měření při kontrole hřídelí, je potřeba rozptýl odhadnout.

Usiluje-li druhý výrobce o to, aby co nejčastěji dosahoval hodnoty  $T=150$  znamená to, že odchylky  $Y$  od této hodnoty budou pravděpodobně rozděleny podle Gaussovy křivky. Tedy nejčastější je hodnota  $T = 150$  a čím je odchylka  $Y$  od  $T$  větší, tím je hodnota méně četná.

Směrodatnou odchylku zde odhadneme pomocí vztahu:

$$s = \frac{2 \cdot \text{tolerance}}{6} = \frac{2 \cdot 4}{6} = 1,33.$$

**Poznámka.**

Předpokládá se, že u normálního rozdělení je toleranční interval roven šestinásobku smě-



rodatné odchylky.

První výrobce se spokojuje s dodržáním tolerance což znamená, že v tolerančním intervalu  $(T - 4, T + 4) = (154, 146)$  může mít vyrobenou hodnotu v kterémkoliv místě se stejnými četnostmi. Říkáme, že  $Y$  má rovnoměrné rozdělení. Směrodatnou odchylku v tomto případě určíme ze vztahu

$$s = \frac{154 - 146}{\sqrt{12}} = \frac{8}{\sqrt{12}} = 2,31.$$

**Poznámka.**

U rovnoměrného rozdělení  $R(a, b)$  je rozptyl dán vztahem

$$s^2 = \frac{(b - a)^2}{12}.$$

Průměrné ztráty tedy budou:

1. výrobce:  $E(L) = ks^2 = 2,5 \cdot 2,31^2 = 13,34,$

2. výrobce:  $E(L) = ks^2 = 2,5 \cdot 1,33^2 = 4,42.$

Je vidět, že filozofie „stačí dodržovat toleranci“ není správná, neboť ztráty za nekvalitu jsou dokonce třikrát větší.

Z rovnice

$$E(L) = \frac{A}{d^2} s^2$$

je zřejmé, že průměrné ztráty za nekvalitu závisí nejen na rozptylu, ale také na  $A$  a  $d$ . Jestliže velikost rozptylu je určena **dělníkem**, pak parametry  $A$  a  $d$  stanoví **konstruktér**. Ten by měl navrhnout výrobek tak, aby byl robustní, tj. odolný vůči nepřesnostem výroby. Znamená to, že

- a. nedodržení  $T$  by mělo způsobit co nejmenší ztrátu, jinými slovy,  $A$  by mělo být co nejmenší. Je-li zde  $A = 40$  a sníží se o 50%, bude při stejném  $s = 1,33$

$$k = \frac{A}{d^2} = \frac{20}{4^2} = 1,25$$

$$E(L) = k \cdot s^2 = 1,25 \cdot 1,33^2 = 2,21 \text{ Kč/ks.}$$

- b. výrobek by měl být funkční v široké toleranci, tedy mělo by být dosti velké  $d$ . Je-li  $d = 8$ , pak konstanta  $k = \frac{40}{8^2}$  a při  $s = 1,33$  bude

$$E(L) = k \cdot s^2 = 0,625 \cdot 1,33^2 = 1,1 \text{ Kč/ks.}$$

Otázka správného stanovení tolerance je důležitým předpokladem použití ztrátové funkce. Rozptyl lze ovšem snížit i v případě, že odchylky  $Y$  od  $T$  mají žádané normální rozdělení.

Uvažujme například o snížení rozptylu  $s^2 = 1,33^2$  o 30% tj. na  $1,33^2 \cdot 0,70$ , přičemž toto zlepšení stojí 50 haléřů na kus. Snížení rozptylu přinese menší ztráty za nekvalitu.

Pro  $k = 2,5$ :

$$E(L) = 2,5 \cdot 1,33^2 \cdot 0,70 = 3,096 \text{ Kč/ks.}$$

K tomu je nutno připočítat náklady na snížení rozptylu, takže celkově bude

$$E(L) = 3,096 + 0,50 = 3,596 \text{ Kč/ks.}$$

To je stále méně než při rozptylu  $s^2 = 1,33^2$  takže lze tuto úpravu doporučit.

Rovno-  
měrné  
rozdělení



Co nej-  
menší  
hodnota  
parametru  
 $A$

Co nej-  
menší  
hodnota  
parametru  
 $d$



\*

**SHRnutí KAPITOLY TAGUCHIHO METODY: ZTRÁTOVÁ FUNKCE****Shrnutí kapitoly**

Po prostudování této kapitoly jste získali informace o ztrátové funkci. Také víte, že u každého výrobku je sledována jeho určitá charakteristika, podle které posuzujeme jeho kvalitu. Tato charakteristika má jistou optimální hodnotu a nekvalita se projevuje odchylkami od této hodnoty. Jakákoliv odchylka od této hodnoty představuje určitou ztrátu, která se projeví u odběratele zvýšenými náklady na provoz apod.

Podle toho, co je v dané situaci považováno za optimální cílovou hodnotu  $T$ , rozlišujeme tři typy tolerance:

1. Tolerance typu  $N$  (nominal): ideálem je dosažení cílové hodnoty  $T$ .
2. Tolerance typu  $S$  (smaller):  $Y$  je tím lepší, čím je menší. Ideálem je  $T=0$ .
3. Tolerance typu  $L$  (larger):  $T$  je tím lepší, čím je větší. Ideálem je  $T = \infty$ .

**ŘEŠENÝ PŘÍKLAD 10-3****Hřídel**

U vyráběných hřídelí se sleduje průměr a délka, přičemž průměr má být  $T_1 = 25 \pm 1$  a délka  $T_2 = 100 \pm 2$ . Nedodržení tolerance pro průměr stojí 40 Kč, pro délku 30 Kč. K dispozici jsou výsledky kontroly deseti hřídelí.

Výsledky kontroly pro průměr:

25.1, 25, 25, 24.9, 25.1, 25, 24.9, 25, 25.1, 24.9

Výsledky kontroly pro délku:

99.9, 99.9, 99.8, 100.2, 100, 100, 100.1, 98, 99.9, 100.2

Porovnejte kvalitu výroby při dodržování sledovaných rozměrů.

**Řešení příkladu**

Průměr:  $T_1 = 25 \pm 1$ ,  $A_1 = 40$ ,  $d_1 = 1$

$$s^2 = \frac{1}{10} \left[ (25,1 - 25)^2 + (25 - 25)^2 + \dots + (24,9 - 25)^2 \right] = 0,006$$

$$E(L) = \frac{40}{1^2} 0,006 = 0,24 \text{ Kč/ks}$$

Délka:  $T_2 = 100 \pm 2$ ,  $A_2 = 30$ ,  $d_2 = 2$

$$s^2 = \frac{1}{10} \left[ (99,9 - 100)^2 + \dots + (100,2 - 100)^2 \right] = 0,02$$

$$E(L) = \frac{30}{2^2} 0,02 = 0,15 \text{ Kč/ks}$$

Na základě výsledků lze konstatovat, že lepší kvalita je při výrobě délek. Celkové ztráty za nekvalitu jsou  $0,24 + 0,15 = 0,39$  Kč/ks.

\*

**ŘEŠENÝ PŘÍKLAD 10-4**

**Automa-  
tické prač-  
ky**

Při výrobě bubnů automatických praček byla předepsána tolerance  $30 -1, 30+4$ . Nedodržení dolní tolerance stojí 50 Kč, u horní tolerance 100 Kč, neboť buben nelze namontovat. Vypočítejte ztrátu za nekvalitu u pracovníků A a B, máte-li k dispozici tyto výsledky kontroly:

Pracovník	odchylka
A	0, 0, -1, 3, 0, 4, 2, -1, 0, 1, 2, 4
B	-1,-1, 0, 0, 0, 3, 2, -1, 1, 2, 0

**Řešení příkladu**

Parametry:  $A_1 = 50$ ,  $A_2 = 100$ ,  $d_1 = -1$ ,  $d_2 = 4$ .

Pracovník A:

$$E(L_1) = \frac{1}{12} \left\{ \frac{50}{(-1)^2} [(-1)^2 + (-1)^2] + \frac{100}{4^2} (3^2 + 4^2 + 2^2 + 1^2 + 2^2 + 4^2) \right\}$$

$$E(L_1) = 34,375 \text{ Kč/ks}$$

Pracovník B:

$$E(L_2) = \frac{1}{11} \left\{ \frac{50}{(-1)^2} [(-1)^2 + (-1)^2 + (-1)^2] + \frac{100}{4^2} (3^2 + 2^2 + 1^2 + 2^2) \right\}$$

$$E(L_2) = 23,864 \text{ Kč/ks}$$

Pracovník B odvádí kvalitnější práci než pracovník A.

\*

**ŘEŠENÝ PŘÍKLAD 10-5**

**Ložiska**

Kuličky do ložisek se vyrábějí na dvou pracovištích: Pracoviště A zajišťuje průměr kuliček  $T \pm 0,4$ . Pracoviště B zodpovídá za požadovanou tvrdost kuliček  $T \pm 1$ . Denně se vyrobí 50000 kuliček, cena jedné je 0,60 Kč. Při nedodržení kterékoliv tolerance se kulička vyřadí. Při kontrole byly získány tyto odchylky od předepsaných hodnot:

Odchylky od průměru:

-0.3, 0.1, 0.2, 0, 0, -0.2, -0.1, 0, 0.4, 0.1, -0.1, 0, 0, 0.1, -0.2

Výsledky kontroly pro délku:

0, 0, 1, -0.8, -0.8, 0, 0.6, 0.7, 0, -0.3, -0.2, 0, 0, 1, 0.2

Posuďte úroveň kvality těchto dvou pracovišť.

**Řešení příkladu****1. Průměr kuliček**Parametry:  $A = 0,6$ ;  $d = 0,4$ .

Rozptyl při kontrole průměrů

$$s^2 = \frac{1}{15} [(-0,3)^2 + 0,1^2 + \dots + (-0,2)^2] = \frac{0,42}{15} = 0,028$$

Průměrné ztráty za nekvalitu (Kč/ks):

$$E(L) = \frac{A}{d^2} s^2 = \frac{0,6}{0,4^2} \cdot 0,028 = 0,105 \text{ Kč/ks}$$

Denní ztráta za nekvalitu je  $50\,000 \cdot 0,105 = 5\,250$  Kč.**2. Tvrdost kuliček.**Parametry:  $A = 0,6$ ;  $d = 1$ 

Rozptyl při kontrole tvrdosti (počítá se s odchylkami)

$$s^2 = \frac{1}{15} [0^2 + \dots + 0,2^2] = 0,287$$

Průměrné ztráty za nekvalitu (Kč/ks)

$$E(L) = \frac{0,6}{1^2} 0,287 = 0,172 \text{ Kč/ks}$$

Denní ztráty jsou  $50\,000 \cdot 0,172 = 8\,600$  Kč.

\*

**ŘEŠENÝ PŘÍKLAD 10-6**

Při povrchové úpravě pístů je maximální povolená drsnost povrchu 10 tis.mm. Při překročení této tolerance je povrch znovu upraven nákladem 200 Kč. Povrch upravují dva pracovníci. Posuďte úroveň jejich práce na základě výsledků kontroly:

*Písty*

Pracovník	povrch
1	0, 1, 9, 6, 10, 2, 3, 0, 9
2	3, 2, 4, 4, 5, 2, 4, 6, 5, 3

**Řešení příkladu**Parametry:  $A = 200$ ,  $d = 10$  (S-typ tolerance)**1. pracovník:**

$$s^2 = \frac{1}{9} (0^2 + 1^2 + 9^2 + 6^2 + \dots + 9^2) = 34,67$$

$$E(L) = \frac{A}{d^2} \cdot s^2 = \frac{200}{10^2} 34,67 = 69,34$$

$$E(L) = 69,34 \text{ Kč/ks.}$$

2. pracovník:

$$s^2 = \frac{1}{10} (3^2 + 2^2 + 4^2 + \dots + 5^2 + 3^2) = 16$$

$$E(L) = \frac{200}{10^2} \cdot 16 = 32 \text{ Kč/ks.}$$

Druhý pracovník odvádí kvalitnější práci.

\*

### ŘEŠENÝ PŘÍKLAD 10-7



Výrobci horolezeckých lan je stanovena dolní hranice pevnosti lana v tahu na 300 kg. Ztráta při překročení této hranice je 50 Kč na metr. Týdně se vyrobí 100 000 m. Porovnejte dvě technologie výroby, máte-li k dispozici tyto údaje:

*Horolezecká lana*

technologie	pevnost lana
1	305, 350, 350, 410, 310, 300, 350, 400
2	305, 301, 308, 306, 300, 320, 310, 310, 320, 325

#### Řešení příkladu

Parametry:  $A = 50$ ,  $d = 300$ . Jedná se o  $L$  - typ tolerance.

1. technologie

Rozptyl na základě vztahu (10-6):

$$s_1^2 = \frac{1}{8} \left( \frac{1}{305^2} + \frac{1}{350^2} + \dots + \frac{1}{400^2} \right) = 8,62 \cdot 10^{-6}.$$

Průměrná ztráta je dána vztahem (10-5):

$$E(L_1) = 50 \cdot 300^2 \cdot 8,62 \cdot 10^{-6} = 38,79 \text{ Kč/m.}$$

2. technologie

Rozptyl:

$$s_2^2 = \frac{1}{10} \left( \frac{1}{305^2} + \frac{1}{301^2} + \dots + \frac{1}{400^2} \right) = 1,03 \cdot 10^{-5}.$$

Průměrná ztráta:

$$E(L_2) = 50 \cdot 300^2 \cdot 1,03 \cdot 10^{-5} = 46,76 \text{ Kč/m.}$$

Můžeme tedy říci, že první technologií se vyrábí lana s pevností v průměru větší

$$E_1 = 346,88 \text{ kg}, E_2 = 310,5 \text{ kg}$$

a s menším rozptylem

$$s_1^2 = 8,62 \cdot 10^{-6}, \quad s_2^2 = 1,03 \cdot 10^{-5},$$

proto také u první technologie jsou menší ztráty za nekvalitu

$$E(L_1) = 38,79 \text{ Kč/m}, \quad E(L_2) = 46,76 \text{ Kč/m.}$$

\*



**KONTROLNÍ TEST 10****Ano či ne?**

- 10.1** Matematické vyjádření ztrátové funkce má tvar  $L(Y) = k(Y - T)^2$ .
- 10.2** Příkladem veličiny  $Y$  s tolerancí typu  $S$  může být pevnost lana.
- 10.3** U každého výrobku je dána určitá charakteristika, která má jistou optimální hodnotu a nekvalita se projevuje odchylkami od této hodnoty.
- 10.4** U tolerance typu  $N$ , je ideálem dosažení menší hodnoty než je cílová hodnota  $T$ .
- 10.5** Grafem ztrátové funkce je parabola.

**Doplňte**

- 10.6** Jakákoliv odchylka od cílové hodnoty  $T$  přináší \_\_\_\_\_.
- 10.7** U každého výrobku je sledována jeho určitá \_\_\_\_\_, podle které posuzujeme jeho kvalitu.
- 10.8** Matematické vyjádření ztrátové funkce má tvar \_\_\_\_\_.
- 10.9** Podle toho,co je v dané situaci považováno za optimální cílovou hodnotu  $T$ , rozlišujeme tyto typy tolerance: \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_.
- 10.10** U tolerance typu  $S$ , je ideálem dosažení cílové hodnoty  $T =$  \_\_\_\_\_.

**Řešte**

- 10.11** U určitého výrobku se sleduje průměr a hmotnost, přičemž průměr má být  $T_1 = 20\text{cm} \pm 1$  a hmotnost  $T_2 = 100\text{g} \pm 2$ . Nedodržení tolerance pro průměr stojí 20 Kč, pro hmotnost 30 Kč. K dispozici jsou výsledky kontroly deseti výrobků.

Výsledky kontroly pro průměr:

20.1, 20, 20, 19.9, 20.1, 20, 19.9, 20, 20.1, 19.9.

Výsledky kontroly pro hmotnost:

99.9, 99.9, 99.8, 100.2, 100, 100, 100.1, 9.8, 99.9, 100.2

Porovnejte kvalitu výroby při dodržování sledovaných rozměrů.

- 10.12** Při výrobě odlučovačích filtrů je stanoveno maximální možné procento propustnosti 10%.

Kontrola filtrů u 2 výrobců přinesla tyto výsledky:

Podnik	% propustnosti
A	3, 9, 9, 7, 1
B	8, 8, 1, 1, 2, 5.

Překročení tolerance stojí u podniku A 600 Kč a u podniku B 700 Kč. Který výrobce je kvalitnější?

**ŘEŠENÍ KONTROLNÍHO TESTU 10****Ano či ne?****10. 1** ano**10. 2** ne**10. 3** ano**10. 4** ne**10. 5** ano**10. 6** ztráty**10. 7** charakteristika**10. 8**  $L(Y) = k(Y - T)^2$ **10. 9**  $N$  (nominal),  $S$  (smaller),  $L$  (larger)**10. 10** 0**Doplňte****10. 11** Průměr:  $E(L) = 0,12$  Kč/ks; Hmotnost:  $E(L) = 0,15$  Kč/ks**10. 12** Podnik A:  $E(L) = 265,2$  Kč/ks; Podnik B:  $E(L) = 185,5$  Kč/ks**Řešte**

# 11 TAGUCHIHO METODY: CELKOVÉ NÁKLADY NA JAKOST

## CÍLE KAPITOLY TAGUCHIHO METODY: CELKOVÉ NÁKLADY NA JAKOST

Po úspěšném a aktivním absolvování této KAPITOLY

### Budete umět:

- definovat vztahy pro výpočet celkových nákladů na jakost,
- odpovědět na otázku, jak často se má provádět kontrola, jestliže se neprovádí 100% kontrola

**Budete umět**

### Získáte:

- znalost postupu při konstrukci regulačního diagramu

**Získáte**

### Budete schopni:

- správně odhadnout průměrný počet výrobků mezi dvěma poruchami,
- definovat pojem regulační diagram.

**Budete schopni**

### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 4 hodiny.

### PRŮVODCE STUDIEM 14



V předcházející kapitole jsme se seznámili s Taguchiho metodami a s pojmem ztrátová funkce. V této kapitole se dovíme něco o celkových nákladech na jakost a regulačních diagramech. Kapitola je členěna do tří částí. První část je věnována monitorování nákladů na jakost, ve druhé části je uveden vztah pro výpočet celkových nákladů na jakost, jestliže se provádí 100% kontrola procesu, část třetí je věnována vztahu pro celkové náklady na jakost, jestliže se provádí kontrola procesu po  $n$  výrobcích a v části čtvrté jsou definovány regulační diagramy.

Základní vztah nebyl Taguchim odvozen, ale navržen. Zachycuje autorovy zkušenosti, které vyjádřil matematickými prostředky. V této kapitole opět najdete řešené i neřešené příklady.

### DEFINICE 11-1



Zabezpečování jakosti není pouze ryze technickým a organizačním problémem, ale i důležitou ekonomickou otázkou, je tedy klíčovým prvkem managementu jakosti.

Každý výrobce produkuje své výrobky za spotřeby určitých výrobních nákladů, které posléze umístí na trh za cenu, která mu tyto náklady nejen pokryje, ale zabezpečí i určitý zisk. Pomíne-li mezičlánek různých prodejních organizací, pořizovací cena výroku se pro uživatele stává první investicí, jejíž vynaložení je nutné, aby výrobek uživateli přinášel pozitivní efekty. Použijí

**Podstata a význam ekonomiky jakost**

i

Používání výrobku s sebou přináší také další výdaje – provozní náklady, např. pohonné hmoty, elektřina, jednotlivé části výrobku, které je třeba v průběhu života vyměňovat. Patří sem také ztráty z nedisponibility výrobku a náklady na servis. Součet těchto položek představuje tzv. náklady na životní cyklus.

V nákladech na výrobek se dále uvažují náklady na jakost. Vhodnou péčí o jakost můžeme dosáhnout při stejné ceně snížení nákladů na výrobek a tím také růst zisku. Zvyšování jakosti bude z pohledu zákazníků tak atraktivní, že budou ochotni akceptovat i vyšší ceny, a snížení provozních nákladů a zejména ztrát z nedisponibility každý zákazník rád uvítá.

Rámcem ekonomiky jakosti je především monitorování nákladů na jakost, sledování přínosů zlepšování jakosti a také tvorba cen produktů v závislosti na jejich jakosti. Pomocí monitoringu jsme schopni zjistit velikost ztrát vyvolaných nedostatky v jakosti, sledovat vliv jakosti na výsledky podnikání a odhalovat ty produkty podniku, které jsou díky jakosti nositeli prosperity.

## 11.1 Monitorování nákladů na jakost

Pojem „náklady na jakost“ může mít několik významů, především výdaje spojené přímo se zajišťováním, případně zlepšováním jakosti, např. výdaje na nákup měřicí techniky, nebo také může zahrnovat neproduktivní výdaje, jakými jsou ztráty z neshodných výrobků. Z praktického hlediska je vhodné náklady na jakost rozdělit do tří skupin:

- náklady na jakost u výrobce,
- náklady na jakost u uživatele,
- společenské náklady na jakost.

### PRŮVODCE TEXTEM



V této studijní opoře se budeme věnovat nákladům na jakost u výrobce.

### DEFINICE 11-2



*Náklady na jakost u výrobce* jsou výdaje vynaložené výrobcem a spojené s prevencí, hodnocením a vadami, aby bylo dosaženo požadavků jakosti v průběhu marketingu, vývoje, zásobování, výroby, instalace a užití. Jejich monitoring představuje velice účinný nástroj managementu jakosti, protože má možnosti odkrývání všech příležitostí ke zlepšování. Popíšme nyní způsoby monitoringu s využitím

*Náklady na jakost u výrobce*

- 1) tzv. PAF modelů,
- 2) aplikaci modelu tzv. procesních nákladů,
- 3) využití Taguchiho ztrátové funkce.

Ad1) PAF modely (Prevention, Appraisal, Failure)

*PAF modely*

Tento model je založený na tom, že se v podniku všechny nákladové položky spojené s jakostí rozdělují do čtyř skupin:

*náklady na interní vady* (vznikají uvnitř firmy v důsledku vad při plnění požadavků na jakost, nedostatky jsou odhaleny ještě před odesláním k zákazníkovi),

*náklady na externí vady* (zde patří reklamace, garanční servis, manipulační náklady, slevy z ceny soudní spory, ztráta trhu a důvěry zákazníků)

*náklady na hodnocení* (jsou to zejména náklady na měření spokojenosti zákazníků, mě-

řící techniku, software, certifikace, náklady na provoz podnikových i externích zkušeben a laboratoří)

*náklady na prevenci* (tato skupina nákladů na jakost by měla vykazovat trvalý růst, patří sem zjišťování požadavků zákazníka, rozvoj systému jakosti, školení, poradenské firmy, plánování jakosti).

#### Ad2) Model procesních nákladů

Jedná se o vyšší stupeň monitoringu, který je založen na tom, že se jednotlivé skupiny nákladů nesledují podle konkrétních výrobků, ale výhradně pro určité procesy. Za proces se považuje soubor činností, které transformují hmotné resp. informační vstupy na hmotné a informační výstupy. Tyto náklady se dělí na *náklady na shodu*, což jsou skutečné náklady na přeměnu vstupů na výstupy, a na *náklady na neshodu* jako náklady na nevyužitý čas, materiál a kapacity, spojené se vznikem neshod uvnitř procesu.

**Model  
procesních  
nákladů**

#### Ad 3) Taguchiho metoda

Otázka zvyšování jakosti výrobků je důležitá, ale musíme také brát v úvahu druhou stránku věci, tj. náklady spojené se zajišťováním této jakosti. Jedním ze způsobů minimalizace nákladů na jakost je Taguchiho metoda. Touto problematikou se zabývají kapitoly 11.2 a 11.3.

**Taguchiho  
metoda**

## 11.2 Provádí se 100% kontrola procesu

### DEFINICE 11-3

Provádí-li se 100 % kontrola, pak celkové náklady na jakost určíme ze vzorce

$$L = \frac{Q}{R} + \frac{A}{d^2} s_0^2$$

$Q$  = roční náklady na 100 % kontrolu,

$R$  = roční produkce v kusech,

$d$  = funkční tolerance, vymezující přípustné odchylky od jisté ideální hodnoty, ve které je výrobek ještě vyhovující,

$A$  = ztráta při překročení tolerance  $d$ ,

$$s_0^2 = \frac{1}{n-1} [(y_2 - y_1)^2 + (y_3 - y_2)^2 + \dots + (y_n - y_{n-1})^2].$$



**100%  
kontrola  
procesu  
(11-1)**

### PRŮVODCE TEXTEM

Aplikace vztahu (11-1) je v následujícím příkladu.



### ŘEŠENÝ PŘÍKLAD 11-1

Náklady na 100% automatickou kontrolu jsou 25 000 Kč za rok. Roční produkce je 4 000 000 kusů, tolerance je 9 a její překročení stojí 5 Kč. Určete celkové náklady na jakost.



**Automa-  
tická kon-  
trola**

**Řešení příkladu**

Pomocí výše zavedeného označení máme:

$$Q = 25\,000 \text{ Kč}$$

$$R = 4\,000\,000 \text{ ks,}$$

$$d = 9,$$

$$A = 5 \text{ Kč,}$$

$$s_0^2 = 1.$$

Dosadíme do vztahu (11-1) a dostáváme:

$$L = \frac{25000}{4000000} + \frac{5}{9^2} \cdot 1 = 0,068 \text{ Kč/ks.}$$

Roční náklady na jakost jsou  $4\,000\,000 \cdot 0,068 = 272\,000 \text{ Kč}$ .

\*

**11.3 Kontrola procesu se provádí po n jednotkách****DEFINICE 11-4**

Pokud se neprovádí 100% kontrola a mezi dvěma kontrolami je  $n$ -výrobků, určíme celkové náklady pomocí vzorce

$$L = \frac{B}{n} + \frac{C}{u} + \frac{A}{d^2} \frac{D^2}{3} + \frac{A}{d^2} \frac{D^2}{u} \left( \frac{n+1}{2} + z \right) + \frac{A}{d^2} s_m^2$$

(11-2)

$A =$  ztráta při překročení tolerance  $d$ ,

$B =$  cena kontroly výrobku,

$C =$  cena opravy stroje (linky),

$n =$  kontrolní interval,

$u =$  průměrný počet výrobků mezi opravami (poruchami),

$d =$  funkční tolerance, vymežující přípustné odchylky od jisté ideální hodnoty, ve které je výrobek ještě vyhovující,

$D =$  výrobní tolerance, která je obvykle podnikovým zpřísněním funkční tolerance,

$z =$  počet výrobků zhotovených během kontroly,

$\frac{B}{n} =$  cena kontroly na kus,

$\frac{C}{u} =$  cena opravy na kus,

$\frac{A}{d^2} \frac{D^2}{3} =$  ztráty způsobené nepřesností výroby,

$\frac{A}{d^2} \frac{D^2}{u} \left( \frac{n+1}{2} + z \right) =$  ztráty za zmetky,

$$\frac{A}{d^2} s_m^2 =$$

ztráty způsobené nepřesností měření.

Je třeba říci, že základní vzorec (11-1) nebyl Taguchim **odvozen**, ale **navržen**. Můžeme říci, že vzorec je matematickým vyjádřením dlouholetých zkušeností G. Taguchiho, představuje autorovy zkušenosti, vyjádřené matematickými prostředky. Umožňuje také po derivaci podle určité proměnné a jejím položením nule najít optimální hodnoty některých parametrů. Jeho jednotlivé části však mají matematické jádro: tři z pěti sčítanců vycházejí ze ztrátové funkce.

Diskutovat lze jistě k celkové skladbě vztahu (11-1) i ke způsobu výpočtu jednotlivých položek. Dále uváděné vzorce (11-3) – (11-5) jsou však již odvozené z těchto výchozích a diskuse k jejich tvaru je bezpředmětná.

V případě, že se neprovádí 100% kontrola, vyvstávají tyto přirozené otázky:

1. Jak často kontrolovat?
2. S jakou přesností kontrolovat?



### DEFINICE 11-5

Uvedené otázky jsou řešeny v následujících vzorcích. Ze vztahu (11-1), který parciálně derivujeme podle  $n$  resp.  $D$  a derivaci položíme rovnu nule, dostaneme optimální kontrolní interval

$$n^* = \sqrt{\frac{2u_0 B}{A} \frac{d}{D_0}}$$



Optimální  
kontrolní  
interval  
(11-3)

### DEFINICE 11-6

a optimální provozní toleranci

$$D^* = \sqrt[4]{\frac{3CD_0^2 d^2}{Au_0}}$$



Optimální  
provozní  
tolerance  
(11-4)

### DEFINICE 11-7

Průměrný počet výrobků mezi dvěma poruchami (opravami) **odhadujeme** vztahem

$$u = \frac{D^{*2}}{D_0^2} u_0$$



Průměrný  
počet  
výrobků  
(11-5)

## 11.4 Regulační diagramy

### PRŮVODCE STUDIEM 15



Hlavním nástrojem statistické regulace výrobního procesu jsou regulační diagramy (Steward, 1926). Jejich posláním je nepřetržitě monitorovat sledované ukazatele a včas upozornit na zhoršující se stav. To umožní předcházet defektům ve výrobě. I když sledovaným ukazatelem může být podle povahy procesu téměř cokoliv, jsou jisté požadavky, které by měl splňovat. Základním požadavkem je (alespoň pro zde uváděné typy RD), že má normální rozdělení.

### DEFINICE 11-8



U sledovaného technologického procesu se hodnotí:

- a. schopnost udržet předepsanou (cílovou) hodnotu,
- b. míra kolísání (variabilita) kolem cílové hodnoty.

Proto se vždy kombinují dva regulační diagramy, každý pro jeden z uvedených atributů.

Nejběžnější jsou RD pro dvojice (průměr  $\bar{x}$ , rozpětí  $R$ ), označovaný dále  $RD(\bar{x}, R)$  a (průměr, směrodatná odchylka), tj.  $RD(\bar{x}, s)$ .

V tomto textu uvádíme regulační diagram pro dvojice (průměr  $\bar{x}$ , rozpětí  $R$ ).

*Při konstrukci RD se postupuje v těchto krocích:*

- a. Shromažďují se data (vzorky) v časech  $t = 1, 2, \dots, m$  (první sloupec tab.1).
- b. Z každého vzorku se vypočítají potřebné charakteristiky, např. u  $RD(\bar{x}, R)$  to bude průměr  $\bar{x}$  a rozpětí  $R = x_{\max} - x_{\min}$ .
- c. Vypočítají se a v grafu znázorní tři údaje: hodnota, která je považována za optimální a také meze, které by sledovaný znak neměl překročit.
- d. V grafu se znázorní průměry (svislá osa) odpovídající jednotlivým časům (vodorovná osa). Tím je sestrojen  $RD(\bar{x})$ . Podobně se sestrojí  $RD(s)$ .
- e. Sleduje se vývoj parametrů a podle zásad o vyhodnocování RD se přijímají potřebná opatření. Jejich podstatu lze charakterizovat takto: parametr by měl vykazovat náhodné kolísání kolem předepsané hodnoty ve vymezených mezích. Jak se počítají tyto meze a předepsaná hodnota, je uvedeno dále.



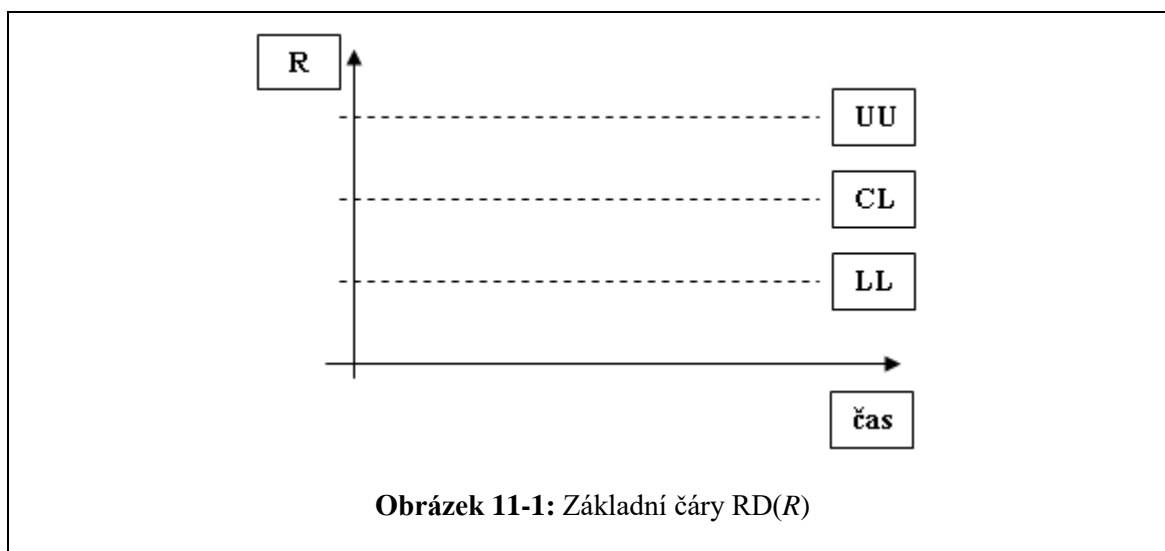
**Konstrukce regulačních diagramů**



	Osa $x$	Osa $y$	
	$i = \text{čas}$	Průměr $\bar{x}_i$	Rozpětí $R_i$
$x_{11}, x_{12}, \dots, x_{1n}$	1	$x_1$	$R_1$
$x_{21}, x_{22}, \dots, x_{2n}$	2	$x_2$	$R_2$
$x_{31}, x_{32}, \dots, x_{3n}$	3	$x_3$	$R_3$
$x_{m1}, x_{m2}, \dots, x_{mn}$	$m$	$x_m$	$R_m$

Tab.1 Vzorky a charakteristiky pro konstrukci  $RD(\bar{x}, R)$ 

Následující obrázek zachycuje základní čáry  $RD(R)$ .

Obr. 11-1:  
Základní  
čáry  $RD$   
( $R$ )

$t = 1$	$t = 2$	...	$t = m$
$x_{11}$	$x_{21}$		$x_{m1}$
$x_{12}$	$x_{22}$		$x_{m2}$
...	...		...
$x_{1n}$	$x_{2n}$		$x_{mn}$
$R_1$	$R_2$		$R_m$

Tab.2 Vzorky pro konstrukci  $RD(R)$ 

Označení v grafu:

UCL = horní regulační mez (upper control limit),

LCL = dolní regulační mez (lower control limit),

CL = střední přímka (central line).

Při použití softwaru musíme vědět:

1. jaký typ RD použít,

2. jak volit parametry RD,

- a. velikost vzorku ( $n$ ),
- b. frekvence vzorkování ( $h$ ),
- c. šířka regulačního pásma ( $L$ ),

3. jak vyhodnotit průběh RD.

Výpočet regulačních mezí provádět nebudeme. Přesto je dobré vědět, jak se počítají a to zejména proto, že z výpočtu je zřejmé, za jakých podmínek lze daný RD použít.

### DEFINICE 11-9



Regulační  
diagram  
pro průměr

#### Regulační diagram RD( $\bar{x}$ , $R$ )

U RD( $R$ ) se počítá rozpětí  $R$ . Proto se při výpočtu regulačních mezí RD( $\bar{x}$ ) využije znalost této hodnoty.

a. Pro regulační diagram RD( $\bar{x}$ ) je

$$LCL = \bar{\bar{x}} - A_2 \bar{R},$$

$$UCL = \bar{\bar{x}} + A_2 \bar{R},$$

$$CL = \bar{\bar{x}}$$

kde  $\bar{\bar{x}} = \frac{1}{m} \sum_{i=1}^m \bar{x}_i$  je průměr z průměrů vzorků,

$$\bar{R} = \frac{1}{m} \sum_{i=1}^m R_i \text{ je průměr z rozpětí vzorků.}$$

b. Pro regulační diagram RD( $R$ ) je

$$UCL = D_4 \bar{R},$$

$$LCL = D_3 \bar{R},$$

$$CL = \bar{R}.$$

Konstanty  $A_2$ ,  $D_3$ ,  $D_4$  jsou v tabulce 3.

Regulační  
diagram  
pro rozpětí

$n$	$A_2$	$A_3$	$B_3$	$B_4$	$D_3$	$D_4$	$d_2$
2	1,88	2,550		3,267		3,267	1,128
3	1,023	1,954		2,568		2,574	1,693
4	0,720	1,628		2,265		2,282	2,039
5	0,577	1,427		2,089		2,114	2,326
6	0,483	1,267	0,330	1,970		2,004	2,534
7	0,419	1,182	0,118	1,892	0,076	1,924	2,704
8	0,373	1,029	0,185	1,815	0,136	1,864	2,847
9	0,337	1,032	0,239	1,761	0,184	1,816	2,970

10	0,308	0,975	0,284	1,716	0,223	1,777	3,078
11	0,285	0,927	0,321	1,679	0,256	1,744	3,173
12	0,265	0,886	0,354	1,646	0,283	1,717	3,258
13	0,249	0,850	0,382	1,618	0,307	1,693	3,336
14	0,235	0,817	0,406	1,594	0,328	1,672	3,407
15	0,223	0,789	0,428	1,572	0,347	1,653	3,472
16	0,212	0,763	0,448	1,552	0,363	1,637	3,532
17	0,203	0,739	0,465	1,534	0,378	1,622	3,558
18	0,194	0,718	0,482	1,518	0,391	1,608	3,640
19	0,187	0,698	0,497	1,503	0,403	1,597	3,689
20	0,080	0,680	0,510	1,490	0,415	1,585	3,735
21	0,173	0,663	0,523	1,477	0,425	1,575	3,778
22	0,167	0,647	0,534	1,466	0,434	1,566	3,719
23	0,162	0,633	0,545	1,455	0,443	1,557	3,858
24	0,157	0,619	0,555	1,445	0,451	1,548	3,859
25	0,153	0,606	0,565	1,435	0,459	1,541	3,931

Tab.3 Konstanty pro výpočet LCL a UCL

**ŘEŠENÝ PŘÍKLAD 11-2**

Vypočítejte UCL, LCL a CL pro  $RD(\bar{x})$  a  $RD(R)$ , je-li dáno:

$i$					$\bar{x}$	$R$
1	9,9	9,9	11	9,8	10,15	1,2
2	9,1	9,8	9,9	11,2	10	2,1
3	9,6	9,4	10,7	9,9	9,9	1,3
4	10,4	9,4	9,2	9,9	9,725	1,2
5	9,9	10,6	9,6	10	10,03	1
6	10,3	9,8	9,7	9,9	9,925	0,6
7	10,2	11,1	9,6	10	10,23	1,5
suma					69,96	8,9

Tab.4 Vzorok pro konstrukci  $RD(\bar{x}, R)$ **Řešení příkladu**

a. Regulační diagram  $RD(\bar{x})$ :

$$\bar{\bar{x}} = \frac{1}{m} \sum_{i=1}^m \bar{x}_i = 9,994$$

$$\bar{R} = \frac{1}{m} \sum_{i=1}^m R_i = 1,271$$

dolní regulační mez:  $LCL = \bar{x} - A_2 \bar{R} = 9,994 - 0,720 \cdot 1,271 = 0,915$

horní regulační mez:  $UCL = \bar{x} + A_2 \bar{R} = 9,994 + 0,720 \cdot 1,271 = 10,909$

střední přímka:  $CL = \bar{x} = 9,994$

**b.** Regulační diagram  $RD(R)$ :

horní regulační mez:  $UCL = D_4 \bar{R} = 2,282 \cdot 1,271 = 2,900$

dolní regulační mez:  $LCL = D_3 \bar{R} = 0,1 \cdot 1,271 = 0$

střední přímka:  $CL = \bar{R} = 1,271$

\*

### SHRnutí KAPITOLY TAGUCHIHO METODY: CELKOVÉ NÁKLADY NA JAKOST



Shrnutí kapitoly

Tato kapitola nás provedla problematikou týkající se celkových nákladů na jakost a byly zde také uvedeny základní informace týkající se regulačních diagramů. Základní vzorec pro celkové náklady na jakost při 100% kontrole procesu nebyl Taguchim odvozen, ale navržen. V tomto vztahu jsou zachyceny autorovy zkušenosti, které vyjádřil matematickými prostředky.

V případě, že se neprovádí 100% kontrola procesu, pak je třeba zjistit jak často se má kontrola provádět a s jakou přesností.

Regulační diagramy jsou hlavním nástrojem statistické regulace výrobního procesu. Jejich hlavním úkolem je sledovat dané ukazatele a včas upozornit na případně problémy či zhoršující se stav výrobního procesu.

### ŘEŠENÝ PŘÍKLAD 11-3



Lis

Lis vyrábí při jednom zdvihu 8 ks výlisků najednou. Cena jednoho je 0,5 Kč. Kontrola se provádí jednou za hodinu vždy u jednoho výrobku dané série tak, že je-li jeden výrobek vadný, vyřadí se všech 8, zastaví se lis a provede se jeho seřízení nákladem 70 Kč. Funkční tolerance je 10 a počáteční výrobní tolerance 4. Za hodinu se vyrobí 480 ks, počet pracovních hodin za rok je 2000. Kontrola trvá 2 min a její cena je 10 Kč. Chybu měření neuvažujeme, počáteční průměrný interval mezi opravami je 4 hodiny.

Vypočítejte celkové náklady na jakost, stanovte optimální kontrolní režim a jeho finanční přínos.

**Řešení příkladu**

Parametry:

$$A = 8.0,5 = 4 \text{ Kč}$$

$$B = 10 \text{ Kč}$$

$$C = 70 \text{ Kč}$$

$$D_o = 4 \text{ tis. mm}$$

$$d = 10 \text{ tis. mm}$$

$$n_o = 480 \text{ ks}$$

$$z = \frac{480}{60} \cdot 2 = 16$$

$$u_o = 4 \cdot 480 = 1920 \text{ ks}$$

Dosažením  $u_o$ ,  $n_o$  a  $D_o$  do vztahu (11-2) vypočítáme celkové náklady na jakost  $L_o$  pro **vstupní** parametry:

$$L_o = \frac{10}{480} + \frac{70}{1920} + \frac{4}{10^2} \cdot \frac{4^2}{3} + \frac{4}{10^2} \left( \frac{480+1}{2} + 16 \right) \cdot \frac{4^2}{1920} = 0,356$$

$$L_o = 0,356 \text{ Kč/ks.}$$

Nyní určíme optimální parametry kontroly

**a.** Podle vztahu (11-3) je optimální kontrolní interval

$$n^* = \sqrt{\frac{2 \cdot u_o \cdot B}{A}} \cdot \frac{d}{D_o} = \sqrt{\frac{2 \cdot 1920 \cdot 10}{4}} \cdot \frac{10}{4} = 244,9 \approx 240 \text{ výrobků,}$$

tzn. kontrolu přibližně každou  $\frac{1}{2}$  hodinu.

**b.** Podle vztahu (11-4) je optimální provozní tolerance

$$D^* = \sqrt[4]{\frac{3 \cdot C \cdot D_o^2 \cdot d^2}{A \cdot u_o}} = \sqrt[4]{\frac{3 \cdot 70 \cdot 4^2 \cdot 10^2}{4 \cdot 1920}} = 2,57 \approx 2$$

**c.** Podle vztahu (11-5) vypočteme průměrný počet výrobků mezi dvěma poruchami

$$u = u_o \frac{D^{*2}}{D_o^2} = 1920 \frac{2^2}{4^2} = 480 \text{ výrobků.}$$

**d.** Náklady na jakost při optimálních parametrech jsou dány vztahem (11-1)

$$L = \frac{B}{n^*} + \frac{C}{u} + \frac{A}{d^2} \frac{D^{*2}}{3} + \frac{A}{d^2} \left( \frac{n^* + 1}{2} + z \right) \cdot \frac{D^{*2}}{u}$$

$$L = \frac{10}{240} + \frac{70}{480} + \frac{4}{10^2} \cdot \frac{2^2}{3} + \frac{4}{10^2} \left( \frac{240+1}{2} + 16 \right) \cdot \frac{2^2}{480} = 0,287$$

Finanční přínos při optimálních parametrech:

rozdíl  $L_0 - L = 0,356 - 0,287 = 0,069$  Kč na 1 výrobku,

za 1 rok dostáváme úsporu  $0,069 \cdot 480 \cdot 2000 = 66\,240$  Kč.

\*

### ŘEŠENÝ PŘÍKLAD 11-4



Síla nátěru

Při nástřiku karosérií automobilů je kontrolována síla nátěru. Předepsaná hodnota je  $m_0 \pm 0,010$  mm. Síla však závisí na hustotě barvy, kde je tolerance 4 poise.

Při nedodržení síly nátěru se lak přebrušuje nebo přestříkává, což v průměru stojí 70 Kč. Je zjištěno, že změna viskozity o jednotku mění sílu nátěru o 0,5tis.mm, takže  $b = 0,5$ . Průměrný interval mezi opravami je  $10\,000\text{m}^2$ , cena opravy stříkacího automatu je 400 Kč,  $z = 50\text{m}^2$ . Počáteční kontrolní interval je  $5000\text{m}^2$ , cena kontroly je 100 Kč.

Stanovte optimální kontrolní interval.

#### Řešení příkladu

Parametry:

$A$  (ztráta při nedodržení tolerance) = 70 Kč,

$B$  (cena kontroly) = 100 Kč,

$C$  (cena opravy) = 400 Kč,

$d_0$  (funkční tolerance nátěru) = 0,010 mm,

$D_0$  (počáteční výrobní tolerance viskozity) = 4 poise,

$n_0$  (počáteční kontrolní interval viskozity) = 5000 m<sup>2</sup>,

$u_0$  (počáteční průměrná nastříkaná plocha mezi opravami),

$u_0 = 10000\text{m}^2$ ,

$z$  (nastříkaná plocha během kontroly) = 50 m<sup>2</sup>,

$s_m$  se neuvažuje.

Funkční tolerance pro  $X$  (viskozita):

$$d = \frac{d_0}{b} = \frac{0,010}{0,5} = 20\ \mu\text{m}.$$

Náklady na jakost pro dané parametry vypočteme ze vztahu (11-2)

$$L = \frac{B}{n} + \frac{C}{u} + \frac{A}{d^2} \frac{D^2}{3} + \frac{A}{d^2} \frac{D^2}{u} \left( \frac{n+1}{2} + z \right) + \frac{A}{d^2} s_m^2$$

$$L_0 = \frac{100}{5000} + \frac{400}{10000} + \frac{70}{20^2} \frac{4^2}{3} + \frac{70}{20^2} \frac{4^2}{10000} \left( \frac{5000+1}{2} + 50 \right) = 1,707$$

$$L_0 = 1,707\ \text{Kč/m}^2$$

Optimální kontrolní interval:

$$n^* = \sqrt{\frac{2u_0 B}{A} \frac{d}{D_0}},$$

$$n^* = \sqrt{\frac{2 \cdot 10000 \cdot 100}{70} \cdot \frac{20}{4}} = 845.$$

Optimální výrobní tolerance:

$$D^* = \sqrt[4]{\frac{3CD_0^2 d^2}{Au_0}},$$

$$D^* = \sqrt[4]{\frac{3 \cdot 400 \cdot 4^2 \cdot 20^2}{70 \cdot 10000}} = 1,82 \approx 2$$

Očekávaná nastříkaná plocha mezi opravami:

$$u = \frac{D^{*2}}{D_0^2} u_0,$$

$$u = \frac{2^2}{4^2} \cdot 10000 = 2500 \text{ m}^2.$$

Náklady na jakost při optimálních parametrech kontroly:

$$L = \frac{B}{n^*} + \frac{C}{u} + \frac{A}{d^2} \frac{D^{*2}}{3} + \frac{A}{d^2} \frac{D^{*2}}{u} \left( \frac{n^* + 1}{2} + z \right) + \frac{A}{d^2} s_m^2,$$

$$L = \frac{100}{845} + \frac{400}{2500} + \frac{70}{20^2} \cdot \frac{2^2}{3} + \frac{70}{20^2} \frac{2^2}{2500} \left( \frac{845 + 1}{2} + 50 \right) = 0,644$$

Úspora v tomto případě je:  $L_0 - L = 1,707 - 0,644 = 1,063 \text{ Kč/m}^2$ .

Jestliže se denně nastříká  $8000 \text{ m}^2$ , je úspora za den  $8000 \cdot 1,063 = 8504 \text{ Kč}$ .

\*

## KONTROLNÍ TEST 11



Ano či ne?

- 11.1** Jestliže se provádí 100% kontrola, pak celkové náklady na jakost jsou dány vztahem  $L = \frac{Q}{R} + \frac{A}{d^2} s_0^2$ .
- 11.2** Základní vzorec pro výpočet celkových nákladů na jakost byl Taguchim odvozen a matematicky dokázán.
- 11.3** U sledovaného technologického procesu se hodnotí jeho schopnost udržet cílovou hodnotu a míra kolísání kolem cílové hodnoty.
- 11.4** Rozpětí daného souboru dat je dáno vztahem:  $R = x_{\max} - \bar{x}$ .
- 11.5** Regulační diagram RD  $(\bar{x}, R)$  označuje regulační diagram pro průměr a rozpětí.

**Doplňte**

- 11.6** Ve vztahu pro celkové náklady na jakost  $L = \frac{Q}{R} + \frac{A}{d^2} s_0^2$ , se  $s_0^2 =$  \_\_\_\_\_.
- 11.7** V případě, že se neprovádí 100% kontroly výrobního procesu, pak se zabýváme tím, jak často \_\_\_\_\_ a s jakou \_\_\_\_\_.
- 11.8** Průměrný počet výrobků mezi dvěma opravami odhadujeme vztahem \_\_\_\_\_.
- 11.9** Hlavním nástrojem statistické regulace výrobního procesu jsou \_\_\_\_\_.
- 11.10** Nejběžnější regulační diagramy jsou pro dvojice: průměr a \_\_\_\_\_; a průměr a \_\_\_\_\_.

**ŘEŠENÍ A ODPOVĚDI****Ano či ne?**

- 11.1** ano  
**11.2** ne  
**11.3** ano  
**11.4** ne  
**11.5** ano

**Doplňte**

- 11.6**  $s_0^2 = \frac{1}{n-1} [(y_2 - y_1)^2 + (y_3 - y_2)^2 + \dots + (y_n - y_{n-1})^2]$
- 11.7** kontrolovat, přesností
- 11.8**  $u = \frac{D^{*2}}{D_0^2} u_0$
- 11.9** regulační diagramy
- 11.10** rozpětí, směrodatná odchylka



## 12 HODNOCENÍ ZPŮSOBILOSTI VÝROBNÍCH PROCESŮ

### CÍLE KAPITOLY HODNOCENÍ ZPŮSOBILOSTI VÝROBNÍCH PROCESŮ

Po úspěšném a aktivním absolvování této KAPITOLY

#### Budete umět:

- Pojem způsobilosti výrobního procesu.
- Jaké jsou cíle hodnocení způsobilosti.
- Jak se provádí výběr vhodného ukazatele.
- Předpoklady hodnocení způsobilosti.
- Atributy způsobilosti procesu.
- Měřitelné charakteristiky kvality.
- Indexy způsobilosti.

*Budete  
umět*

#### Budete schopni:

- Vysvětlit, co je způsobilost procesu a jaké jsou cíle hodnocení způsobilosti.
- Správně vybrat vhodné ukazatele způsobilosti.
- Vyhodnotit předpoklady hodnocení způsobilosti.
- Použít měřitelné charakteristiky kvality a indexy způsobilosti.

*Budete  
schopni*

#### ČAS POTŘEBNÝ KE STUDIU



Celkový doporučený čas k prostudování KAPITOLY je cca 3 hodiny.

#### PRŮVODCE STUDIEM 16



Způsobilostí výrobního (též technologického) procesu (anglicky „process capability“) se rozumí jeho schopnost trvale dosahovat předem stanovená kritéria kvality. Je žádoucí vyjadřovat způsobilost kvantitativně, tedy nějakým číselným ukazatelem. Při konstrukci těchto ukazatelů máme určité představy o jejich vlastnostech. Z hlediska výpočtového mezi ně patří zejména jednoduchost, dále srozumitelnost, dobrá vypovídací schopnost, názornost. Dalším požadavkem je široká upotřebitelnost, což znamená co nejméně omezujících podmínek pro jejich použití, stručně řečeno, univerzálnost. Z tohoto hlediska je potřeba říci, že takový univerzální ukazatel se jen těžce hledá a i když existuje poměrně mnoho ukazatelů pro posuzování způsobilosti procesu, každý z nich je použitelný pouze tehdy, jsou-li splněny určité konkrétní předpoklady.

Ukazatel způsobilosti musí být samozřejmě konstruován tak, aby hodnotil všechny stránky způsobilosti.

## 12.1 Co je způsobilost procesu

Způsobilostí výrobního procesu (process capability) se rozumí jeho schopnost trvale dosahovat předem stanovená kritéria kvality. Je žádoucí vyjadřovat způsobilost kvantitativně, tedy nějakým číselným ukazatelem. Při konstrukci těchto ukazatelů máme určité představy o jejich vlastnostech. Z hlediska výpočtového mezi ně patří zejména jednoduchost, dále srozumitelnost, dobrá vypovídací schopnost, názornost. Dalším požadavkem je široká upotřebitelnost, což znamená co nejméně omezujících podmínek pro jejich použití, stručně řečeno, univerzálnost. Z tohoto hlediska je potřeba říci, že takový univerzální ukazatel se jen těžce hledá a i když existuje poměrně mnoho ukazatelů pro posuzování způsobilosti procesu, každý z nich je použitelný pouze tehdy, jsou-li splněny určité konkrétní předpoklady.

**Způsobilost procesu**

Ukazatel způsobilosti musí být samozřejmě konstruován tak, aby hodnotil všechny stránky způsobilosti.

**Ukazatel způsobilosti**

## 12.2 Jaké jsou cíle hodnocení způsobilosti

Při kvantitativním hodnocení způsobilosti technologického procesu se obvykle sledují tyto dva cíle :

**Cíle hodnocení**

1. Schopnost procesu udržet cílovou hodnotu  $T$  (anglicky „target value“) ukazatele kvality.

**Udržet cílovou hodnotu**

2. Přiměřená variabilita dosahovaných hodnot kolem cílové hodnoty.

**Přiměřená variabilita**

### PRŮVODCE TEXTEM



V dalším textu se pokusíme ukázat dva nejpoužívanější ukazatele způsobilosti s jejich klady a zápory a s podmínkami jejich použití.

## 12.3 Jak se provádí výběr vhodného ukazatele

Prvním kritériem, podle kterého se rozhodujeme a které dělí ukazatele způsobilosti na dvě skupiny je skutečnost, zda se sleduje jeden nebo více ukazatelů kvality. U jednoho pak, je-li to atribut, tedy neměřitelná charakteristika, nebo měřitelná veličina.

**První kritérium**

Měřitelné charakteristiky jsou reprezentovány výsledky měření (dále jen měření) výrobku nebo výrobního procesu. Tato měření, která jsou ze statistického hlediska výběrovým souborem, mají ve většině případů normální rozdělení. Předpoklad normality je, jak uvidíme dále, zásadní a musí být ověřen. Podle toho, zda je či není splněn, se pak volí vhodný ukazatel ze skupiny měřitelných. Schematicky můžeme rozdělit ukazatele kvality takto:

**Měřitelné charakteristiky**

Nejprve se vybírá index podle počtu znaků kvality, které se sledují:

Počet ukazatelů kvality	
Jeden znak	Více než jeden znak
	- normálně rozdělené, i vzájemně závislé znaky
	- nezávislé znaky, rozdělení libovolné,

Počet  
ukazatelů  
kvality

Jeden znak kvality	
Měřitelné znaky	Atributy
Normální rozdělení	Nenormální rozdělení

Jeden znak  
kvality

## 12.4 Předpoklady hodnocení způsobilosti

Každý z dále uváděných ukazatelů způsobilosti je spolehlivě použitelný pouze při splnění určitých předpokladů. Tyto předpoklady lze rozdělit na

Předpo-  
klady  
hodnocení

- obecné
- specifické

*Obecné předpoklady* jsou ty, které musí být splněny vždy, u všech používaných indexů způsobilosti. Jsou uvedeny v této kapitole.

Obecné  
předpokla-  
dy

*Specifické předpoklady* jsou ty, které jsou požadovány u jednotlivých indexů způsobilosti, vedle obecných předpokladů, navíc. Budou uvedeny současně s definicí každého ukazatele způsobilosti i se zdůvodněním, proč jsou požadovány.

Specifické  
předpokla-  
dy

Splnění obecných i specifických předpokladů je nutné ověřit, nejčastěji testem.

Obecné předpoklady, které musí být splněny při výpočtu jakéhokoliv indexu způsobilosti, jsou (zejména) tyto:

- proces je stabilizován
- data jsou nezávislá, bez odlehlých pozorování a v dostatečném rozsahu
- je stanovena správně tolerance

Jestliže kterákoliv z těchto podmínek není splněna, neměl by být počítán žádný z indexů způsobilosti. Při nedodržení této zásady jsou vypočítané hodnoty indexů buď velmi nespolehlivé (nahodnocené resp. podhodnocené - podle toho, co nebylo splněno) nebo i nesmyslné (např. extrémně velké při špatné toleranci, záporné a pod.).

*Proces je stabilizován* (je pod statistickou kontrolou) tehdy, jestliže všechny hodnoty sledovaného ukazatele kvality leží uvnitř regulačních mezí příslušného regulačního diagramu. Z uvedeného je zřejmé, že dříve, než se začne s hodnocením způsobilosti, měl by být zaveden do provozu (je-li to technicky možné) regulační diagram. Ten sleduje nepřetržitě vývoj klíčových parametrů v čase.

## 12.5 Atributy

Pro atributy se způsobilost procesu vyjadřuje procentem výrobků, které vyhovují požadovanému ukazateli kvality. Označíme-li relativní četnost špatných výrobků

$$V = \frac{\text{pocet\_nevyhovujících\_vyrobku\_mezi\_sledovanými}}{\text{celkový\_pocet\_sledovaných\_vyrobku}}$$

pak způsobilost  $C$  bude procento vyhovujících výrobků  $C = 100(1-V)$ . U tohoto ukazatele není stanovena obecně platná minimální hodnota  $C$ . Donedávna přijímaná úroveň byla 98-99%, dnes začíná převažovat přísný požadavek 3ppm (3 špatné výrobky na milion kusů).

## 12.6 Měřitelné charakteristiky kvality

V současné době se prakticky výhradně používá při posuzování způsobilosti procesů třída ukazatelů, nazývaných *indexy způsobilosti* („capability index“).

*Indexy  
způsobi-  
losti*

O třídě hovoříme proto, že existuje mnoho typů indexů způsobilosti, které se liší způsobem výpočtu, vlastnostmi i podmínkami použitelnosti. Princip jejich konstrukce je však stejný: vždy jde o poměr předepsané přesnosti a skutečně dosahované přesnosti výroby. Předepsaná přesnost je dána tolerancí a cílovou hodnotou.

Označme:

USL = horní toleranční hranice (Upper Specification Limit),

USL

LSL = dolní toleranční hranice (Lower Specification Limit),

LSL

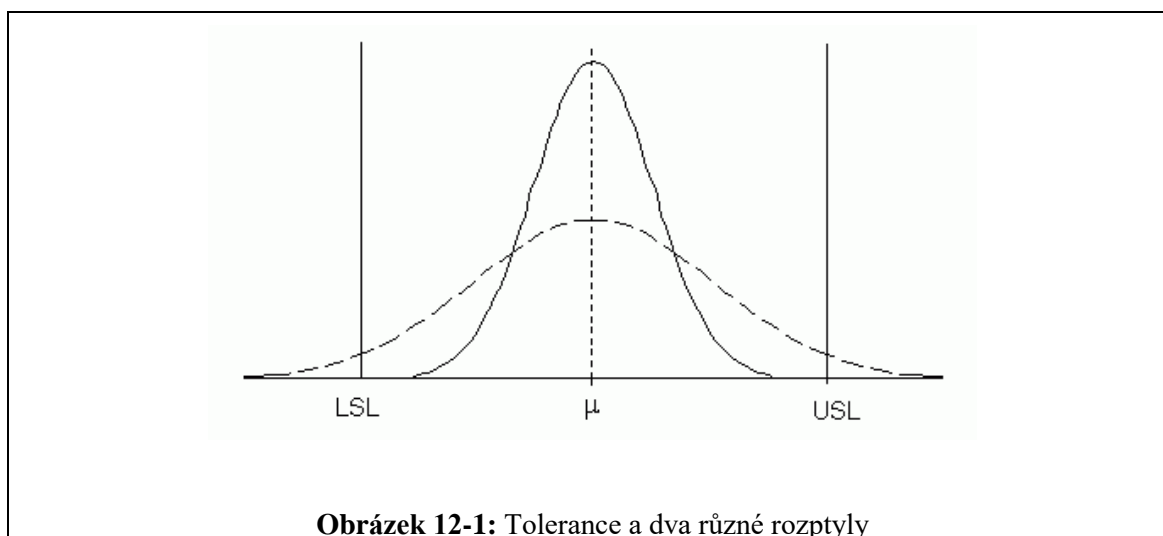
T = cílová hodnota (Target Value),

T

Pak toleranční interval bude (LSL,USL), jeho délka  $USL - LSL$  a střed intervalu  $MSL$  je:  $MSL = 1/2(USL + LSL)$ . *Specifikace výrobního procesu* je určena trojicí (LSL,T, USL). Skutečně dosahovaná přesnost je vyjádřena rozptylem. Je známo, že má-li soubor normální rozdělení  $N(\mu, \sigma^2)$ , pak podle „pravidla 3 sigma“ leží v intervalu  $(\mu - 3\sigma, \mu + 3\sigma)$  99,73 % hodnot, tj. prakticky všechny. Délka tohoto intervalu je  $6\sigma$ . Na obr.1 jsou znázorněny dva případy normálního rozdělení s různými rozptyly. Pro jeden jsou všechny hodnoty v toleranci, kdežto pro druhý je část mimo. Porovnáním délky tolerančního intervalu (LSL, USL) a intervalu  $6\sigma$ , získáme představu o poměru předepsané a skutečně dosahované přesnosti. Na tomto principu jsou také konstruovány indexy způsobilosti:

*Specifika-  
ce výro-  
ního pro-  
cesu*

$$\text{způsobilost} = \frac{\text{délka intervalu, kde mají být všechny hodnoty}}{\text{délka intervalu kde jsou všechny hodnoty}}$$



**Obr. 12-1:**  
Tolerance  
a dva  
různé  
rozptyly

**Obrázek 12-1:** Tolerance a dva různé rozptyly

## 12.7 Index $C_p$

Nejstarší index způsobilosti je označován  $C_p$  a definován vztahem

*Index  $C_p$*

$$C_p = \frac{USL - LSL}{6 \cdot \sigma}$$

(12-1)

Vzhledem k tomu, že směrodatná odchylka  $\sigma$  základního souboru většinou není k dispozici, je nahrazena výběrovou směrodatnou odchylkou  $s$ . Tím místo indexu  $C_p$  podle (12-1) máme jeho odhad

$$\hat{C}_p = \frac{USL - LSL}{6 \cdot s}$$

(12-2)

Připomeňme, že pravidlo  $3\sigma$ , na kterém je vzorec (12-1) postaven, platí pouze pro normální rozdělení. To je závažný poznatek, který prakticky znamená, že pokud měření nemají normální rozdělení, nelze hodnotit způsobilost podle (12-1). Je proto velmi důležité rozhodnout pokud možno spolehlivě, má-li soubor normální rozdělení či nikoliv. K tomu je nutné

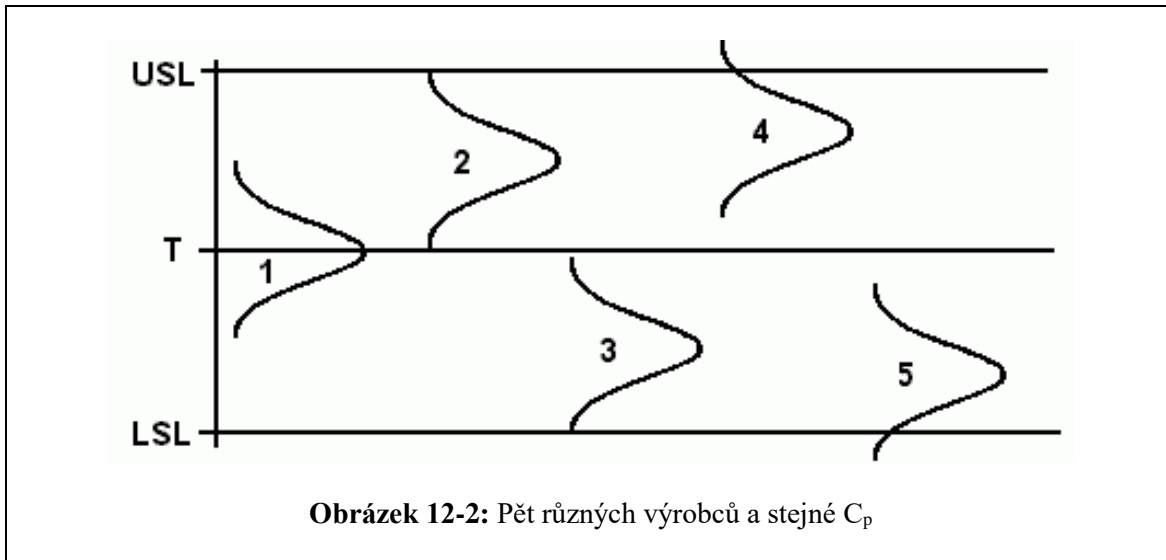
- mít dostatečně velký soubor
- použít spolehlivý test na normalitu
- odstranit ze souboru odlehlé hodnoty

Není-li normalita potvrzena, přirozeně se nabízí otázka, jak postupovat dále a existují -li vůbec vhodné nástroje. Zde pouze konstatujeme, že nelze použít klasické prostředky, avšak existují speciální metody pro tuto situaci. V případě, že je normalita potvrzena, je možné  $C_p$  počítat.

Vážným nedostatkem indexu  $C_p$  je, že vůbec nehodnotí, jak je proces **centrován**, tedy jaký je vztah předepsané hodnoty  $T$  a průměrné hodnoty  $\mu$ . Velmi názorně je to vidět z Obr.3, kde je pět různých výrobců:

1. výrobce je nejlepší, neboť nejčastěji vyrábí předepsanou hodnotu  $T$  a nepřekračuje toleranci.
2. a 3. výrobce sice nikdy nedosahuje hodnotu  $T$  (vychyluje se vlevo resp. vpravo), ale alespoň nepřekračuje toleranci.
4. a 5. výrobce nejenže nedosahuje nikdy  $T$ , ale navíc překračuje toleranci (vlevo resp. vpravo).

Je tedy každý výrobce jiný a přesto, vypočítáme-li index  $C_p$ , bude ve všech případech stejný, neboť délka tolerančního intervalu  $USL - LSL$  je stejná pro všechny a směrodatná odchylka  $s$  ve jmenovateli vzorce (12-1) bude také stejná, protože zákonem rozdělení je ve všech případech tatáž Gaussova křivka, jen s jinou polohou vrcholu. Můžeme tedy konstatovat, že  $C_p$  posuzuje pouze jeden ze dvou cílů hodnocení způsobilosti a to míru využití tolerance.

Obr. 12-2:  
Pět různých výrobců a stejné  $C_p$ 

Shrneme li dosud uvedené poznatky, dostáváme tyto *specifické* podmínky pro použití  $C_p$ :

- a) měření má rozdělení  $N(\mu, \sigma^2)$
- b)  $\mu = T$  (proces je centrován)

## 12.8 Index $C_{pK}$

Snaha o zavedení obecnější charakteristiky způsobilosti než  $C_p$  vedla ke konstrukci indexů

KPM

$$C_{pU} = \frac{USL - \mu}{3 \cdot \sigma}$$

(12-3)

$$C_{pL} = \frac{\mu - LSL}{3 \cdot \sigma}$$

(12-4)

$$C_{pK} = \min\{C_{pU}, C_{pL}\}$$

(12-5)

Tento index má jen jednu specifickou podmínku použití a to normální rozdělení sledovaného ukazatele kvality.

### ŘEŠENÝ PŘÍKLAD 12-1



Ověřte si, že vzdaluje-li se  $\mu$  od  $T$ , zhoršuje se  $C_{pK}$ , pokud se nemění. Parametry procesu jsou:  $LSL = 10$ ,  $USL = 18$ ,  $T = 14$  (symetrická tolerance; píšeme také  $T = MSL$ ).

Symetrická tolerance

**Řešení příkladu**a)  $\mu = T = 14$ :

$$C_{pU} = \frac{USL - \mu}{3\sigma} = \frac{18 - 14}{3 \cdot \frac{2}{3}} = 2$$

také

$$C_{pL} = \frac{\mu - LSL}{3\sigma} = \frac{14 - 10}{3 \cdot \frac{2}{3}} = 2$$

Je tedy pro  $\mu = T$ :  $C_{pU} = C_{pL} = C_{pK} = 2$ b)  $\mu = 15$ :

$$C_{pU} = \frac{USL - \mu}{3\sigma} = \frac{18 - 15}{3 \cdot (2/3)} = 1,5$$

$$C_{pL} = \frac{\mu - LSL}{3\sigma} = \frac{15 - 10}{3 \cdot (2/3)} = 2,5$$

Zde, protože  $\mu \neq T$ , je  $C_{pL} \neq C_{pU}$  a  $C_{pK} = \min\{1,5; 2,5\} = 1,5$ c)  $\mu = 16$ :

$$C_{pU} = \frac{USL - \mu}{3\sigma} = \frac{18 - 16}{3 \cdot (2/3)} = 1$$

$$C_{pL} = \frac{\mu - LSL}{3\sigma} = \frac{16 - 10}{3 \cdot (2/3)} = 3$$

$$C_{pK} = 1$$

d)  $\mu = 17$ ,  $\sigma = 0,25$  (snížení směrodatné odchylky)

$$C_{pU} = \frac{USL - \mu}{3\sigma} = \frac{18 - 17}{3 \cdot 0,25} = 1,33 \quad !!!$$

Pokud by  $C_{pU}$  resp.  $C_{pL}$  vycházelo záporné, pokládá se  $C_{pK} = 0$ . To nastane v případě, že průměrná hodnota  $\mu$  je mimo toleranční interval.

\*

**ŘEŠENÝ PŘÍKLAD 12-2**Ověřte si, že je-li  $\mu$  mimo toleranci, vychází  $C_{pU}$  (resp.  $C_{pL}$ ) záporné. Parametry procesu jsou:  $LSL = 10$ ,  $USL = 18$ ,  $T = 14$ ,  $\mu = 20$ ,  $\sigma = 2/3$ .**Indexy pro  $\mu$  mimo toleranci**

**Řešení příkladu**

Zde  $\mu = 20$ , tedy mimo horní toleranci, takže  $C_{pU} < 0$ .

Skutečně

$$C_{pU} = \frac{USL - \mu}{3\sigma} = \frac{18 - 20}{3 \cdot \frac{2}{3}} = -1$$

Proto se pokládá  $C_{pK} = 0$ , i když

$$C_{pL} = \frac{20 - 10}{3 \cdot \frac{2}{3}} = 5.$$

\*

**SHRNUTÍ KAPITOLY HODNOCENÍ ZPŮSOBILOSTI VÝROBNÍCH PROCESŮ**

*Shrnutí kapitoly*

V této kapitole jste studovali problematiku způsobilosti výrobního procesu (process capability). Tou se rozumí jeho schopnost trvale dosahovat předem stanovená kritéria kvality. Naučili jste se vyjadřovat způsobilost kvantitativně, tedy příslušným číselným ukazatelem. Při konstrukci těchto ukazatelů jsme měli určité představy o jejich vlastnostech. Z hlediska výpočtového mezi ně patří zejména jednoduchost, dále srozumitelnost, dobrá vypovídací schopnost, názornost. Dalším požadavkem je široká upotřebitelnost, což znamená co nejméně omezujících podmínek pro jejich použití, stručně řečeno, univerzálnost. Ukazatele způsobilosti jsou samozřejmě konstruovány tak, aby hodnotily všechny stránky způsobilosti výrobního procesu.

**ŘEŠENÝ PŘÍKLAD 12-3**

Ověřte si, že vzdaluje li se  $\mu$  od T (nabývá hodnot 50, 57.5 resp. 61) a současně se zmenšuje  $\sigma$  (nabývá hodnot 5, 2.5 resp. 1,33), zůstává  $C_{pK}$  nezměněno. Specifikace procesu je :  $USL = 65$ ,  $LSL = 35$ ,  $T = 50$ .

**Řešení příkladu**

a)  $\mu = 50, \sigma = 5$

$$C_{pU} = \frac{65 - 50}{3.5} = 1, \quad C_{pL} = \frac{50 - 35}{3.5} = 1; \quad C_{pK} = 1$$

b)  $\mu = 57.5, \sigma = 2.5$

$$C_{pU} = \frac{65 - 57.5}{3.2,5} = 1, \quad C_{pL} = \frac{57.5 - 35}{3.2,5} = 3; \quad C_{pK} = 1$$



$$c) \mu = 61, \sigma = 4/3$$

$$C_{pU} = \frac{65-61}{3.1,33} = 1, C_{pL} = \frac{61-35}{3.1,33} = 6,5; C_{pk} = 1$$

\*

**ŘEŠENÝ PŘÍKLAD 12-4**

Vypočítejte index C<sub>pm</sub>, který je definován vztahem

$$C_{pm} = \frac{USL - LSL}{6\tau}$$

pro parametry procesu jsou:  $\sigma = 2/3$ ,  $\mu = T = 14$ . USL = 18, LSL = 10.

Parametr

$$\tau^2 = \sigma^2 + (\mu - T)^2$$

takže C<sub>pm</sub>

$$\hat{C}_{pm} = \frac{USL - LSL}{6\tau} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - T)^2}}$$

**Řešení příkladu**

$$C_{pm} = \frac{18-10}{6.(2/3)} = 2$$

**Poznámka:**

Odhad  $\hat{\tau}^2$  je možné počít přímo z dat

$$\hat{\tau}_1^2 = \frac{1}{n} \sum (x_i - T)^2$$

\*

**ŘEŠENÝ PŘÍKLAD 12-5**

Je dána specifikace procesu : LSL = 10, USL = 18, T = 14. Směrodatná odchylka je  $\sigma = 2/3$ . Vypočítat C<sub>pm</sub> pro  $\mu = T$  a  $\mu \neq T$  a porovnat s C<sub>p</sub>.

**Řešení příkladu**

$$a) \mu = 14 = T$$

$$C_{pm} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - T)^2}} = \frac{18 - 10}{6\sqrt{(2/3)^2 + (14 - 14)^2}} = 2$$

$$C_p = \frac{USL - LSL}{6\sigma} = \frac{18 - 10}{6 \cdot (2/3)} = 2$$

Pro  $\mu = T$  je tedy  $C_p = C_{pm}$ .

b)  $\mu = 15 \neq T$

$$C_{pm} = \frac{USL - LSL}{6 \cdot \sqrt{\sigma^2 + (\mu - T)^2}} = \frac{18 - 10}{6 \sqrt{(2/3)^2 + (15 - 14)^2}} = 1,11$$

$C_p = 2$  (beze změny). Pro  $\mu \neq T$  je tedy  $C_{pm} \neq C_p$ .

\*

### ŘEŠENÝ PŘÍKLAD 12-6



Počítejte  $C_{pm}$  opět pro zhoršující se  $\mu$  a zlepšující se  $\sigma$ . Porovnejte  $C_{pm}$  a  $C_{pk}$ . Specifikace procesu:  $USL = 65$ ,  $LSL = 35$ ,  $T = 50$ .

#### Řešení příkladu

a)  $\mu = 50, \sigma = 5$ :

$$C_{pm} = \frac{USL - LSL}{6 \cdot \sqrt{\sigma^2 + (\mu - T)^2}} = \frac{65 - 35}{6 \sqrt{5^2 + (50 - 50)^2}} = 1$$

b)  $\mu = 57.5, \sigma = 2.5$

$$C_{pm} = \frac{USL - LSL}{6 \cdot \sqrt{\sigma^2 + (\mu - T)^2}} = \frac{65 - 35}{6 \sqrt{2,5^2 + (57,5 - 50)^2}} = 0,632$$

c)  $\mu = 61, \sigma = 4/3$

$$C_{pm} = \frac{USL - LSL}{6 \cdot \sqrt{\sigma^2 + (\mu - T)^2}} = \frac{65 - 35}{6 \sqrt{(4/3)^2 + (61 - 50)^2}} = 0,448$$

Z předchozích příkladů je vidět, že zatímco  $C_{pk}$  se při zhoršujícím se  $\mu$  nemění „díky“ zmenšování  $\sigma$ , index  $C_{pm}$  zhoršení průměru  $\mu$  zaznamenal.

\*

**ŘEŠENÝ PŘÍKLAD 12-7**

Vypočítejte  $C_{pK}$  a  $C_{pm}$ , je-li  
 $\mu = 17, \sigma = 2/3, T = 17$ . Toleranční meze jsou  $LSL = 10, USL = 18$ .

**Řešení příkladu**

$C_{pK} = 0.5; C_{pm} = 2$  (!).

\*

**ŘEŠENÝ PŘÍKLAD 12-8**

Vypočítejte  $C_{pK}$  a  $C_{pm}$  pro dva odlišné výrobní procesy:

- Malý rozptyl, ale špatné centrování procesu.  
Specifikace:  $LSL = 35, T = 50, USL = 65$ .  
Charakteristiky:  $\mu = 57, \sigma = 2.5$
- Velký rozptyl, ale dobré centrování.  
Specifikace:  $LSL = 12, T = 14, USL = 18$ .  
Charakteristiky:  $\mu = 14, \sigma = 0.75$   
Porovnejte v obou případech  $C_{pK}$  a  $C_{pm}$ .

**Řešení příkladu**

- Indexy vychází :  $C_{pK} = 1.06$  a  $C_{pm} = 0.67$ .

Uspokojivá hodnota  $C_{pK}$  je výsledkem malého rozptylu. Skutečnost, že proces není centrován, se však u tohoto indexu výrazněji neprojevila.

- Indexy vychází:  $C_{pK} = 0.88$  a  $C_{pm} = 1.33$ .

Zde obráceně velký rozptyl se neprojevil u indexu  $C_{pm}$ , ale naopak výrazně poznamenal  $C_{pK}$ .

\*

**ŘEŠENÝ PŘÍKLAD 12-9**

Pro dané výsledky měření vypočítejte indexy  $C_p, C_{pK}, C_{pm}$ :  
143, 200, 160, 181, 148, 178, 162, 215, 161, 141.  
Specifikace procesu:  $USL = 240, LSL = 100$  a  $T = 170$ .

**Řešení příkladu**

$$\mu = 168.9, \sigma = 24.49, \tau = 24.51, C_p = 0.953, C_{pK} = 0.938, C_{pm} = 0.952$$

\*

**KONTROLNÍ OTÁZKY****Kontrolní  
otázky**

1. Jaká hodnota indexu způsobilosti je vždy nevyhovující? (menší než 1)
2. Jaké dva atributy hodnotí každý index způsobilosti? (centrování a variabilitu procesu)
3. Jaká je společná podmínka pro použití indexů  $C_p$  a  $C_{pK}$ ? (normality sledovaného znaku)
4. Co je to způsobilost procesu? (schopnost trvale dosahovat předepsané ukazatele kvality)
5. Vyjmenujte dva typy podmínek pro hodnocení způsobilosti. (obecné a specifické)
6. Vyjmenujte obecné podmínky hodnocení způsobilosti. (proces je stabilní, data jsou nezávislá, bez odlehklých pozorování a v dostatečném rozsahu, tolerance je stanovena správně)
7. Jak se hodnotí způsobilost u atributů? (podíl nevyhovujících ku všem výrobkům)
8. Čím je určena specifikace výrobního procesu? (specifikace výrobního procesu je určena trojicí LSL, T, USL)
9. Co je pravidlo 3 sigma? (podle „pravidla 3 sigma“ leží v intervalu  $(\mu - 3\sigma, \mu + 3\sigma)$  99,73 % hodnot náhodné veličiny, která má normální rozdělení)
10. Napište vzorec pro výpočet indexů  $C_p$ ,  $C_{pK}$  a  $C_{pm}$ .
11. Kdy se index  $C_{pm}$  rovná indexu  $C_p$ ? (je-li  $\tau = \mu$ )
12. Nakreslete situaci, kdy je proces vycentrován, ale má příliš velkou variabilitu.
13. Nakreslete situaci, kdy má proces vyhovující variabilitu, ale není vycentrován.
14. Nakreslete situaci, kdy je proces nevyhovující z důvodu variability i centrování.

**SHRNUTÍ MODULU STATISTICKÉ METODY PRO EKONOMY**

Shrnutí modulu, stejně jako shrnutí každé kapitoly, by mělo být hutné, výstižné, přehledně strukturované.

**Shrnutí modulu**

Mělo by korespondovat s cíli, které byly uvedeny.

Shrnutí by jasně mělo deklarovat: Umím to, to a to.

**KLÍČOVÁ SLOVA MODULU ...**

..., ..., ... (kopírujte ze všech kapitol)

**Klíčová slova****DALŠÍ ZDROJE**

BISSEL, B.: *Statistical methods for SPC and TQM*. London: Chapman and Hall, 1994.

HÁTLE, J., LIKEŠ, J.: *Základy počtu pravděpodobnosti a matematické statistiky*. 2.vyd. Praha: SNTL, 1974.

HINDLS, R., HRONOVÁ, S., SEGER, J.: *Statistika pro ekonomy*. 1.vyd. Praha: Professional Publishing, 2002. ISBN 80-86419-26-6.

KAŇKA, M. *Vybrané partie z matematiky pro ekonomy*. 1.vyd. Praha:VŠE, 1998. ISBN 80-7079-537-9.

RAMÍK, J., ČEMERKOVÁ, Š.: *Statistika B*. 2. rozš. a upr. vyd. Karviná: SU OPF, 2000. ISBN 80-7248-099-5.

RAMÍK, J., ČEMERKOVÁ, Š.: *Kvantitativní metody B: statistika:distanční studijní opora*. 1.vyd. Karviná: SU OPF, 2003. ISBN 80-7248-198-3.

SEGER, J., HINDLS, R.: *Statistické metody v ekonomii*. 1.vyd. Jinočany: H&H, 1993. ISBN 80-85787-26-1.

SEGER, J., HINDLS, R., HRONOVÁ, S.: *Statistika v hospodářství*. 1.vyd. Praha: ETC Publishing, 1998. ISBN 80-86006-56-5.

TOŠENOVSKÝ, J., NOSKIEVIČOVÁ, D.: *Statistické metody pro zlepšování jakosti*.1.vyd. Ostrava: Montanex, a.s., 2001. ISBN 80-7225-040-X.

TOŠENOVSKÝ, J., DUDEK, M.: *Základy statistického zpracování dat*.1.vyd. Ostrava: VŠB, 2001. ISBN 80-248-0006-3.

**Další zdroje**

## SEZNAM POUŽITÝCH ZNAČEK, SYMBOLŮ A ZKRATEK

### INFORMATIVNÍ, NAVIGAČNÍ, ORIENTAČNÍ



Průvodce studiem



Průvodce textem, podnět, otázka, úkol



Shrnutí



Tutoriál



Čas potřebný k prostudování



Nezapomeň na odměnu a odpočinek

### KE SPLNĚNÍ, KONTROLNÍ, PRACOVNÍ



Kontrolní otázka



Samostatný úkol



Test a otázka



Řešení a odpovědi, návody



Korespondenční úkoly

### VÝKLADOVÉ



K zapamatování



Řešený příklad



Definice



Věta

### NÁMĚTY K ZAMYŠLENÍ, MYŠLENKOVÉ, PRO DALŠÍ STUDIUM



Úkol k zamyšlení



Část pro zájemce



Další zdroje