Základy statistiky pro analýzu dat

doc. RNDr. Jan Řehák
RNDr. Irena Bártová

ACREA CR, spol. s r.o.

Krakovská 7, 110 00 Praha 1
tel./fax: 234 721 444
email: kurzy@acrea.cz
http://www.acrea.cz

Obsah

1. PŘEDNÁŠKA

1.1. MATICE DAT

Tabulka údajů pro **statistické jednotky** umístěné v řádcích tabulky a charakterizované sloupci tabulky se nazývá **matice dat.**

Data sloupce matice tvoří tzv. **statistickou řadu**. Jsou-li hodnoty číselné statistické řady uspořádány podle velikosti, tvoří uspořádanou statistickou řadu.

Datový soubor v počítači má své jméno, kterým je identifikován.

řádek = jednotka, objekt, vzorek, výrobek, případ
sloupec = proměnná, záznam informace o jedné vlastnosti jednotek

Typy proměnných (sloupce v datové matici):

a) číselné - spojité, počty, poměrové indexy
b) kategorizované - nominální, dichotomické, ordinální, kardinální
c) textové
d) datum a čas

Popis proměnných:

a) název sloupce - pro práci programu a pro identifikaci
b) popis sloupce - charakteristika proměnné (sloupce)
c) popis kódů, resp. popis čísel (lze jimi zaměnit kódy v matici)
d) chybějící hodnoty -„missing values“ - určení kódů, které se vynechávají z výpočtů
e) formát zápisu (počet desetinných míst, text, apod.)

Termínům matice dat a datový soubor používaných v oblasti počítačového zpracování odpovídá v statistické teorii termín **výběrový soubor** či výběr.

Příklady:

- **země** - textová proměnná
- **region** - nominální kategorizovaná proměnná/ 1= západní a severní Evropa, 2 = jižní Evropa, 3 = střední Evropa, 4 = SNS a Balkán
- **kojumr** - spojitá číselná proměnná poměrového typu (procento), „kojenecká úmrtnost“
- **sdzmuzi** - spojitá číselná proměnná, „střední délka života“
- **vzd** - ordinální kategorizovaná proměnná/ 1= základní a nedokončené základní, 2 = střední bez maturity, 3 = maturita, 4= vysokoškolské/ „vzdělání“
- **katvek** - kategorizovaná kardinální proměnná/ hodnota kategorie je střed věkového intervalu (19, 23,28, 33, 38, ...)/ „věkové kategorie“

2. PŘEDNÁŠKA

2.1. ROZLOŽENÍ ČETNOSTÍ

Vlastnosti rozložení dat v kategoriích (souboru četností) se hodnotí v závislosti na typu znaku (obdobně jako u číselných dat):

- ***poloha četností***
  Kde se soustřeďují jednotky? Ve které kategorii (ích)? Na které části škály?
- ***rozptýlení v kategoriích a podél škály***
  Jak se jednotky soustřeďují do jedné kategorie? Jak se polarizují na ordinální škále? Je rozložení rovnoměrné v kategoriích nebo se soustřeďuje do (kolem) jedné kategorie**.**
- ***symetrie rozložení na preferenční nebo znaménkové škále***
  Převažují preference jedné strany škály proti druhé? Převažují kladné hodnoty proti záporným na znaménkové škále? Které obsahově protipolné kategorie porušují vyváženost rozložení?

Charakteristiky se liší podle typu proměnné:

a) *míry polohy*: modus, mediánová kategorie, ordinální medián, průměr
b) *míry variability*: nomvar (Giniho míra), dorvar, rozptyl
c) *míry symetrie*: koeficient asymetrie, šikmost

MÍRY	**nominální**	**ordinální**	**kardinální**
POLOHA	módus	modus	módus
	variační koeficient	mediánová kategorie	
		ordinální medián	
VARIABILITA	nomvar	nomvar	variance/rozptyl
		dorvar	směrodatná odch.
SYMETRIE		koeficient asymetrie	šikmost

POLOHA:

modus = nejčetnější kategorie
mediánová kategorie = kategorie, v níž kumulativní četnost dosáhne 50%
ordinální medián =

$$\tilde{X} = Me - .5 + \frac{.5 - F_{Me-1}}{f_{Me}} = Me + .5 - \frac{F_{Me} - .5}{f_{Me}}$$

průměr =

$$\overline{X} = (1/N)\sum X_i$$

VARIABILITA:

Variabilitu měříme těmito mírami:
a) u nominálních proměnných

koeficient variability ***v*** $= 1 - f_{mod}$ *(f_{mod} – relativní modální četnost)*
nomvar $= 1 - \Sigma_{i=1...K}\ f_i^2$ *(f_i – i-tá relativní četnost; K – počet kategorií)*
normovaný nomvar $= K * nomvar / (K - 1)$

b) u ordinálních proměnných

dorvar $= 2 * \Sigma_{i=1...K}\ F_i\ (1 - F_i)$ *(F_i – i-tá kumulativní relativní četnost)*
normovaný dorvar $= 2 * dorvar / (K - 1)$

c) u kardinálních proměnných

rozptyl $var\ X = s^2 = \Sigma i_{=1...N}\ (x_i - X)^2 / (N - 1)$ *(N – počet měření; X – průměr x_i)*
směrodatná odchylka $s = (var\ X)^{½}$

Čím je příslušná míra variability větší, tím více variability daná proměnná vykazuje

Příklady*:*
a) Hodnocení značek na začátku a poté na konci rozhovoru po předložení karet.
Nominální proměnná, porovnání variabilit je možné přímo, neboť dotazy mají stejný počet kategorií odpovědi.

	Air Fresh	Stick Up	Bonaria	Ambi	Chevy Aer	Brise	žádná	nomvar
obliba značky osvěžovače	10,20%	1,00%	15,30%	21,10%	1,70%	***46,30%***	4,40%	0,295
obliba značky osvěžovače	9,10%	2,00%	12,50%	22,30%	3,40%	***47,30%***	3,40%	0,3
	DE	Eduscho	Jacobs	Meinl	Tchibo	jiná	žádná	
obliba značky kávy	29,20%	3,70%	18,60%	11,90%	***30,20%***	1,40%	5,10%	0,229
obliba značky kávy	***29,80%***	5,10%	18,30%	12,50%	28,80%	0,70%	4,70%	0,226

b) Dotaz na frekvenci sledování TV stanic (listopad 1995, neváženo) - ordinální znak.

rozložení odpovědí n=957	téměř denně	3-4x týdně	1-2x týdně	velmi zřídka	nikdy	nemá signál	nezná stanici
ČT 1	***34,30%***	22,10%	20,40%	21,40%	1,40%	0,20%	0,10%
ČT2	9,70%	10,80%	24,00%	***40,80%***	6,00%	8,60%	0,10%
Kabel Plus Film	1,10%	1,10%	1,80%	4,70%	17,50%	***66,80%***	7,00%
NOVA	***70,00%***	15,60%	8,70%	4,20%	1,20%	0,20%	0,10%
Premiéra TV	6,70%	8,40%	13,10%	15,00%	10,40%	***44,00%***	2,30%
TV ze satelitů	5,40%	3,70%	4,70%	10,40%	14,50%	***56,90%***	4,40%

kumulativní procenta a variabilita	téměř denně	3-4x týdně	1-2x týdně	velmi zřídka	nikdy	dorvar	norm. dorvar
ČT 1	34%	57%	77%	99%	100%	0,662	0,331
ČT2	11%	22%	49%	93%	100%	0,580	0,290
Kabel Plus Film	4%	8%	15%	33%	100%	0,468	0,234
NOVA	70%	86%	95%	99%	100%	0,394	0,197
Premiéra TV	13%	28%	53%	81%	100%	0,717	0,359
TV ze satelitů	14%	24%	36%	63%	100%	0,764	0,382

2.2. ROZLOŽENÍ ČETNOSTÍ - TABULKY A GRAFY

U kategorizovaných dat je první informaci souboru rozložení případů (jednotek) v kategoriích - absolutní a relativní (resp. procentní). Toto rozložení získáváme ve formě tabulek a grafů. Tabulky i grafy mohou mít různé tvary.

Kategorizovaná data - kategorie tvoří úplný disjunktní systém: vzájemně se vylučují (disjunktnost) a jejich sjednocení pokrývá všechny možnosti (úplnost).

Analýza kategorizovaných dat závisí na **typu kategorizované proměnné**:

a) **nominální** typ - kategorie vyjadřují různé kvality
b) **ordinální** typ - kategorie vyjadřují uspořádané kvality
c) **kardinální** typ - kategorie vyjadřují kvantifikované kvality

Specielním případem kategorizované proměnné je **dichotomie**, která má jen dvě hodnoty (např. ANO/NE, MÁ/NEMÁ VLASTNOST). Dichotomie lze tabelovat úsporně a lze vycházet i z toho, že při kódování 1=ANO, 0=NE je průměrný skór souboru roven relativní četnosti kategorie ANO, při kódování 100=ANO, 0=NE je průměr procentem. Proto lze použít pro tabelace těchto dat i postupy připravené pro číselné proměnné.

Hodnoty rozložení - typy četností:

a) *absolutní četnosti* - počty jednotek v kategoriích
b) *relativní četnosti* - podíl kategorie na celém souboru
c) *relativní četnosti z validních dat* - podíl kategorie na souboru validních dat (tj. po redukci těch jednotek, jejichž údaj chybí, je chybně zapsán, respondent odmítl odpovědět, nebo i neuměl odpovědět, tedy po redukci o údaje, které deklarujeme jako „vynechávané“ (missing values)
d) *procenta z celého souboru i z validních dat*
e) *kumulativní absolutní i relativní četnosti z validních dat* - aplikujeme jen pro ordinální a kardinální data, tj. pro proměnné, které mají uspořádané kategorie nebo kategoriím jsou přiřazeny číselné skóry.

Grafy:

a) histogram - vhodný pro číselné proměnné, jejichž hodnoty byly vytříděny do intervalů a kategorie číselně označeny; i pro kumulativní četnosti
b) *sloupkový graf* - bar chart vhodný pro jakýkoliv typ proměnné; i pro kumulativní četnosti
c) *kruhový graf, koláčový graf - pie chart* - vhodný jen pro nominální proměnné

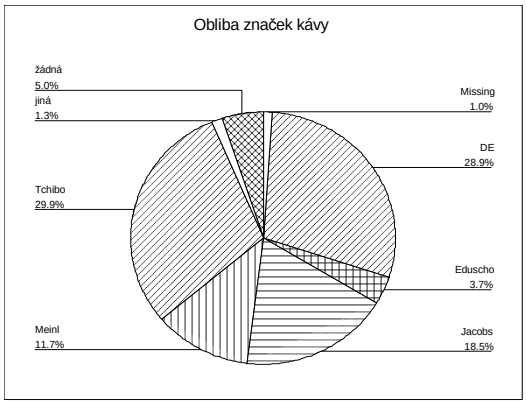
Tyto údaje a grafy zobrazují tvar koncentraci dat do jednotlivých kategorií, posunutí dat na škále, tvar rozptýlení dat, rozložení jako celek a úplnou informaci o něm.

2.3. ROZLOŽENÍ ČETNOSTÍ - TABULKY A GRAFY

Příklady: a) rozložení volby oblíbené značky kávy - nominální proměnná, sedm hodnot, pět konkrétních značek, jedna kategorie pro „jiné“, jedna kategorie pro „žádné“; chybějící pozorování.

obliba značky kávy

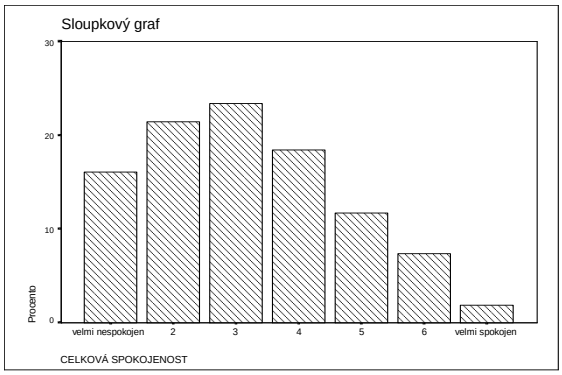
		Počet	%	Validní %
Valid	DE	86	28.9	29.2
	Eduscho	11	3.7	3.7
	Jacobs	55	18.5	18.6
	Meinl	35	11.7	11.9
	Tchibo	89	29.9	30.2
	jiná	4	1.3	1.4
	žádná	15	5.0	5.1
	Total	295	99.0	100.0
Missing	chybí	3	1.0	
	Total	3	1.0	
Total		298	100.0	



b) ordinální proměnná, jejíchž sedm kategorií je uspořádáno od krajní nespokojenosti po krajní spokojenost; tento typ proměnných je v praxi také považován za kardinální proměnnou, jejímiž číselnými hodnotami je obvykle číslo kategorie

CELKOVÁ SPOKOJENOST

	Frequency	Percent	Valid Percent	Cumulative Percent
1=velmi nespokojen	81	16.0	16.0	16.0
2	108	21.4	21.4	37.4
3	118	23.4	23.4	60.8
4	93	18.4	18.4	79.2
5	59	11.7	11.7	90.9
6	37	7.3	7.3	98.2
7=velmi spokojen	9	1.8	1.8	100.0
Total	505	100.0	100.0	



c) uspořádané četnosti podle velikosti se používají u nominálních znaků tam, kde chceme zvýraznit pořadí kategorií podle obsazení - např. volba značky, politika, alternativy pro budoucnost a pod.

obliba značky kávy

		Frequency	Percent	Valid Percent
Valid	DE	88	29.5	29.8
	Tchibo	85	28.5	28.8
	Jacobs	54	18.1	18.3
	Meinl	37	12.4	12.5
	Eduscho	15	5.0	5.1
	žádná	14	4.7	4.7
	jiná	2	.7	.7
	Total	295	99.0	100.0
Missing		3	1.0	
Total		298	100.0	

2.4. KONFIDENČNÍ INTERVALY PRO ČETNOSTI

***Úloha**:*
Jak přesně zjišťujeme procentní zastoupení v kategorii pomocí výběrových dat?

Zastoupení jevu v souboru je dáno jeho absolutní četností *m* a jeho relativní četností *f=m/n* (= absolutní četnost/velikost souboru).

Přesnost informace o relativní četnosti *f* zjišťujeme pomocí *konfidenčního intervalu*.

$$p = f \pm z_\alpha \sqrt{\frac{f(1-f)}{n}}$$

p je neznámá populační hodnota, která je pokryta intervalem

Obdobně u kategorizované proměnné jsou zastoupení v jednotlivých kategoriích: absolutní četnosti n_1, n_2, n_3, ..., n_K, a relativní četnosti f_1, f_2, f_3, ..., f_K, kde $f_k = n_k/n$, n je velikost souboru, který je vzat za základ.

$$p_k = f_k \pm z_\alpha \sqrt{\frac{f_k(1-f_k)}{n}}$$

pro $k = 1,2, ..., K$

Vyjádření obliby politika

	ne	ano	ne	ano	sterr	dolní mez	horní mez
Dlouhý	598	365	62,1%	37,9%	2,0%	***33,9%***	***41,9%***
Havel	677	286	70,3%	29,7%	1,0%	***27,7%***	***31,7%***
Klaus	719	244	74,7%	25,3%	1,0%	***23,3%***	***27,3%***
Zeman	734	229	76,2%	23,8%	1,0%	***21,8%***	***25,8%***
Dienstbier	776	187	80,6%	19,4%	1,0%	***17,4%***	***21,4%***
Buzková	812	151	84,3%	15,7%	1,0%	***13,7%***	***17,7%***
Stráský	822	141	85,4%	14,6%	1,0%	***12,6%***	***16,6%***
Kočárník	836	127	86,8%	13,2%	1,0%	***11,2%***	***15,2%***
Kalvoda	841	122	87,3%	12,7%	1,0%	***10,7%***	***14,7%***
Falber	843	120	87,5%	12,5%	1,0%	***10,5%***	***14,5%***

Výzkum SC&C, listopad 1995

Předchozí tabulka vyjadřuje informaci z celé baterie otázek (na každého politika se ptáme zvlášť. Každý řádek má svých 100%.

Další příklad ukazuje volbu alternativy z několika možností v jedné otázce.
100% je součtem odpovědí v tabulce.

obliba značky kávy

	Četnost	Procento	sterr	dolní mez	horní mez
DE	86	28,9	2,6	***23,7***	***34,1***
Eduscho	11	3,7	1,1	***1,5***	***5,9***
Jacobs	55	18,5	2,3	***13,9***	***23,1***
Meinl	35	11,7	1,9	***7,9***	***15,5***
Tchibo	89	29,9	2,7	***24,5***	***35,3***
jiná	4	1,3	0,7	***-0,1***	***2,7***
žádná	15	5	1,3	***2,4***	***7,6***
Total	295				

Intervaly spolehlivosti platí jednotlivě pro jednotlivé hodnoty - *ne*vyjadřují společnou spolehlivost pro celou tabulku současně. Kdybychom chtěli zkonstruovat intervaly spolehlivosti, které s 95%ní spolehlivostí zahrnují všechna procenta, t.j. riziko pro nepokrytí kteréhokoliv parametru (procenta) je v souhrnu jen 5%, byly by jednotlivé intervaly podstatně širší. Aplikovali bychom na ně místo původních skórů *z* tak zvané Bonferroniho skóry *z*, jež jsou odvozeny na stejném principu, pouze vycházejí z rizika 0.5/(počet zahrnutých parametrů). V příkladu je to 0.5/ 7 = 0.00714, skór *z* je roven 2.69.

3. PŘEDNÁŠKA

3.1. TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

Testování statistických hypotéz je rozhodovací problém, v němž proti sobě stavíme dva výroky – dvě hypotézy: H_0 (**nulovou hypotézu***)* a H_A (**alternativní hypotézu**). Neyman-Pearsonův princip testování je založen na ověřování modelu H_0 proti modelu H_A.

Výsledkem může být jedno ze dvou rozhodnutí:

- není důvod zamítnout H_0,
- data nulové hypotéze odporují, H_0 tedy neplatí, přijímáme H_A.

O tom, zda data nulové hypotéze odpovídají, či zda indikují H_A, vypovídá vždy vhodně zvolená statistická funkce dat (**testová statistika**), která charakterizuje stupeň vzdálenosti dat od H_0 směrem k H_A, a tím stupeň platnosti H_0.

Test je statistické rozhodovací pravidlo, které stanoví, zda testová statistika nabývá takové hodnoty, aby nulová hypotéza, ze které vycházíme, byla odmítnuta.

Při testování hypotéz poznatky zjištěné na konkrétním výběrovém souboru zobecňujeme na základní – hypotetický soubor (někdy se označuje jako **populace**).

Možné chyby v procesu:

Chyby se můžeme dopustit již při samotné formulaci statistické hypotézy (nulová a/nebo alternativní hypotéza neodpovídají řešenému problému).

Při samotném rozhodování se lze dopustit těchto chyb:
1. statistické chyby rozhodování:

		výsledek rozhodnutí	
		H_0	**H_A**
skutečnost	**H_0**	O.K.	chyba 1. druhu (α)
	H_A	chyba 2. druhu (β)	O.K.

2. nesprávně zvolená testová statistika
3. nesprávně určené rozhodovací pravidlo

Pravděpodobnost chyby 1. druhu α je pravděpodobnost, že se rozhodneme pro H_A a ve skutečnosti platí H_0. Je menší nebo rovna předem dané hodnotě α (v praxi většinou volená jako 0,05 nebo 0,01). Mluvíme také o riziku zamítnutí nulové hypotézy, když tato platí.

Pravděpodobnost chyby 2. druhu β je pravděpodobnost, že se rozhodneme pro H_0 a ve skutečnosti platí H_A. Pravděpodobnost této chyby značíme β. Její doplněk do jedné se nazývá **síla testu**.

Rozhodovací pravidlo určujeme tak, abychom nepřekročili zvolené riziko neoprávněného zamítnutí nulové hypotézy a zároveň pokud možno minimalizovali její chybné přijetí. (Není možno minimalizovat obě rizika zároveň.)
Vzhledem k této asymetrii je třeba zformulovat nulovou hypotézu tak, abychom se jejím zamítnutím dostali k tomu, co chceme ukázat.

Při zvolené hodnotě α říkáme, že testujeme hypotézu na **hladině významnosti** α nebo na **hladině spolehlivosti** (1-α) x 100 (%).

Postup testování hypotéz:

- zformulují se hypotézy H_0 a H_A
- zvolí se hladina významnosti α
- vybere se vhodný test a příslušná testová statistika – rozhodovací funkce dat
- do testové statistiky se dosadí hodnoty z dat
- provede se vlastní test hypotézy
- a) *manuální postup*: hodnota statistiky se porovná s kritickou hodnotou zjištěnou v tabulce příslušné danému testu (základní statistické tabulky jsou přílohou většiny učebnic, pro speciální testy je třeba použít samostatné publikace statistických tabulek). Při překročení kritické hodnoty se zamítne nulová hypotéza ve prospěch alternativní; nepřekročí-li testová statistika kritickou, je možno se domnívat, že odchylka od nulové hypotézy byla způsobena náhodnými vlivy a chybami
- b) *počítač*: zjistí se tzv. dosažená hladina významnosti, která znamená vypočtené empirické riziko odmítnutí nulové hypotézy za předpokladu, že H_0 platí (je to odhad pravděpodobnosti chyby prvního druhu); je-li toto riziko menší než předem zvolená hranice α, rozhodujeme se pro alternativní hypotézu, je-li riziko větší než pro nás přijatelná hranice, nezamítáme nulovou hypotézu. (Na výstupech z počítače se označuje většinou jako *P*, *tail probability* nebo *Sig* = significance.)

3.2. ROZLOŽENÍ ČETNOSTÍ - HYPOTÉZA DOBRÉ SHODY

Úlohy:
A) Komparace rozložení s hypotetickým resp. normativním, stabilizovaným stavem.
B) Jsou výzkumná data reprezentativní?
C) Pokrývá trh výrobku proporcionálně jednotlivé sociální a demograficky definované skupiny?

Tyto úlohy mají společný rys: existuje buď objektivní, normativní nebo hypotetické rozložení četností, ke kterému se poměřuje rozložení dat. Jsou to tři typy úloh:

- *Kontrola reprezentativity výběrového šetření* - porovnáváme rozložení kategorizovaných proměnných s dostupnými statistickými daty, například věk, příjem, povolání ap. (ovšem jen takové proměnné, které se neúčastní výběru jako stratifikační, kvótní, řízené).

- *Odchylky od standardu a normy* - některé vlastnosti jsou stabilizované a po dlouhou dobu se opakují, jejich rozložení se stává standardem a vychází se z něj např. při plánování. To může být stabilní podíl značek na trhu, stabilně distribuovaný zájem o typy výrobků, stabilní poptávka po předmětech dlouhodobé spotřeby apod. Četnosti (i výběrové, při dostatečném opakování) se používají jako populační parametr. Testujeme v novém výzkumu, zda nedošlo ke změně.

- *Ověření modelu chování nebo vzniku dat* - naše představa o chování populace může být formulována jako váhy zastoupení v kategoriích. Testujeme platnost naší představy/hypotézy/modelu a odchylky od ní.

***Úloha dobré shody**:*

Statistické hypotézy:

H_0: $p_k = \pi_k$ pro všechna *k*; H_A: $p_k \neq \pi_k$ alespoň pro jedno *k*

p_k jsou četnosti, které reprezentuje náš výběr
π_k jsou četnosti (proporce) které předpokládá hypotéza (stav, standard, model)

Testová statistika:

$$X^2 = \sum_k \frac{(n_k - n\pi_k)^2}{n\pi_k} \qquad df = K - 1$$

$$X^2 = \sum_k \frac{n_k^{\,2}}{n\pi_k} - n \qquad df = K - 1$$

Podmínky pro aplikaci testu: n>=30
Další podmínky pro situaci kde některé očekávané četnosti $n\pi_k$ jsou menší než 5 jsou popsány v knize Řehák, Řeháková (1986, str. 125)

Hodnoty $n\pi_k$ se nazývají *očekávané četnosti*.

Pro rozhodnutí proti nulové hypotéze shody se použije *tabulka kritických hodnot:*

Tabulka kritických hodnot testu chí-kvadrát

df=K-1	alfa=0.05	alfa=0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21

Po odmítnutí hypotézy si klademe další otázky, neboť výsledkem předchozího kroku je prosté negování shody, tedy závěrem je neshoda. Tu ale chceme specifikovat: ve které kategorii nastal významný rozdíl?

Úlohy:

A) Je některá skupina typická v zájmu o daný výrobek? Ve kterých skupinách není výrobek přijímán? Je značka přijímána rovnoměrně v populaci nebo se její akceptace diferencuje?

B) Jaké je uspořádání kategorií podle typičnosti, tj. četnostního nadhodnocení/podhodnocení oproti očekávání?

Úlohy se řeší testem pro odchylky v jednotlivých kategoriích.

Hypotézy:

H_{0k}: $p_k = \pi_k$; H_{Ak}: $p_k \neq \pi_k$ postupně pro $k = 1, \ldots K$.

Každá z těchto hypotéz se testuje pomocí *z*-testu:

$$z_k = \frac{n_k - n\pi_k}{\sqrt{n\pi_k(1-\pi_k)}}$$

Podmínky pro použití: n je alespoň 30
Očekávaná četnost $n\pi_k$ je alespoň 5

Kritické hodnoty pro *alfa* = 0.05,0.01,0.001 jsou postupně:
1.96, 2.58, 3.29
Shodu zamítáme je-li $|z_k|$ >= zvolená kritická hodnota

Znaménkové schéma:

Pro znázornění odchylek můžeme použít znaménkové schéma.

Postup pro vytvoření znaménkového schématu:

1. Provedeme postupně z-testy pro jednotlivá pole tabulky.
2. Určíme zda hodnoty statistik překračují kritické hodnoty a každému poli přiřadíme znaménko(a) plus nebo minus podle toho, jak silně je odchylka signifikantní a zda je rozdíl skutečné a očekávané četnosti kladný nebo záporný; to provedeme podle tří zvolených haldin významnosti (např. 0.05, 0.01, 0.001) :

je-li $\|z_k\| < 1.96$	přiřadíme kategorii 0 nebo znaménko shody
je-li $\|z_k\| >= 1.96$	přiřadíme kategorii znaménko + nebo -
je-li $\|z_k\| >= 2.58$	přiřadíme kategorii znaménko ++ nebo --
je-li $\|z_k\| >= 3.21$	přiřadíme kategorii znaménko +++ nebo ---

Toto schema platí, chceme-li hodnotit každou katagorii zvlášť. Chceme-li mít souhrnný závěr o celé struktuře znamének se zvolenými spolehlivostmi 0.05, 0.01, 0.001, požijeme Holmovu sekvenční metodu simultánní inference.

Úloha:
Jsou respondenti v kategoriích rozloženi rovnoměrně?

Pro řešení takové úlohy použijeme test dobré shody s rovnoměrným rozložením:

H_0: $p_k = 1/K$ pro všechna k; H_A: $p_k \neq 1/K$ alespoň pro jedno k

Pro charakterizaci stupně v neshody je možné použít také *koeficienty neshody*.

4. PŘEDNÁŠKA

4.1. KOMPARAČNÍ TABULKA - ZNAMÉNKOVÉ SCHÉMA

Úlohy: A) Ve kterých kategoriích se liší část populace (skupina) od zbytku?
B) Která kategorie je pro danou skupinu typická?
C) Která kategorie v dané skupině významně absentuje? Je atypická?

Rozdíly mezi očekávanými četnostmi (*Ers=Nrns/N)* a skutečně získanými četnostmi *nrs* v jednotlivých polích se nazývají *rezidua* nebo *odchylky od modelu*.

Jejich významnost se testuje pomocí jejich *standardizace*, tz. *adjustovaných standardizovaných reziduí* nebo *z-skórů*, které mají pro *N*≥30 a *Ers*≥ 5 standardní normální rozdělení. K inferenci použijeme kritické hodnoty normálního rozdělení.

Testy pro jednotlivá pole:

$$z_{rs} = \sqrt{N}\,\frac{Nn_{rs} - N_r n_s}{\sqrt{N_r n_s (N - N_r)(N - n_s)}}$$

Test se provádí porovnáním *zrs* s kritickými hodnotami $z\alpha$; v polích můžeme označit takovou významnost znaménky + a - a tak zobrazit strukturu významných reziduí:

alfa = 0.05	z = 1.96	|*zrs*| ≥ 1.96	znaménko	+ nebo -
alfa = 0.01	z = 2.58	|*zrs*| ≥ 2.58	znaménko	++ nebo --
alfa = 0.001	z = 3.29	|*zrs*| ≥ 3.29	znaménko	+++ nebo ---

Toto grafické znázornění významných odchylek se nazývá *(komparační) znaménkové schéma.*

Pro přijetí celé struktury vztahů určené se zadanou spolehlivostí se provádí simultánní testování všech RxS polí Holmovou metodou.

Příklad:

Preference značek kávy podle věkových skupin žen

		značka kávy							
		DE	**Eduscho**	**Jacobs**	**Meinl**	**Tchibo**	**jiná**	**žádná**	*Total*
odchylky	15-19 let	-9.30%	6.30%	-8.70%	-1.90%	20.10%	-1.40%	-5.10%	0.00%
% v řádcích	20-24 let	-7.40%	-3.70%	0.10%	-8.80%	1.40%	1.70%	16.80%	0.00%
věková	25-29 let	-0.10%	-3.70%	-2.00%	-3.60%	3.40%	-1.40%	7.40%	0.00%
kategorie	30-39 let	-14.50%	0.00%	12.80%	-0.80%	5.30%	0.50%	-3.20%	0.00%
	40-49 let	9.00%	3.70%	-7.60%	0.40%	-2.70%	-0.20%	-2.60%	0.00%
	50-59 let	9.50%	-3.70%	-0.80%	3.00%	-4.50%	-1.40%	-2.10%	0.00%
		29.30%	*3.70%*	*18.70%*	*11.90%*	*29.90%*	*1.40%*	*5.10%*	*100%*
adjustovaná residua	15-19 let	-0.7	1.1	-0.7	-0.2	1.4	-0.4	-0.7	
z-skóry	20-24 let	-1	-1.2	0	-1.6	0.2	0.9	4.6	
věková	25-29 let	0	-1	-0.3	-0.6	0.4	-0.6	1.7	
kategorie	30-39 let	-2.6	0	2.7	-0.2	0.9	0.3	-1.2	
	40-49 let	2.1	2	-2.1	0.1	-0.6	-0.1	-1.3	
	50-59 let	2	-1.8	-0.2	0.9	-0.9	-1.1	-0.9	
	60 let a v.	-1.2	1.1	0.6	1.2	-0.4	1.1	-1.2	
znaménkové schéma		**DE**	**Eduscho**	**Jacobs**	**Meinl**	**Tchibo**	**jiná**	**žádná**	
	15-19 let	0	0	0	0	0	0	0	
věková	20-24 let	0	0	0	0	0	0	+++	
kategorie	25-29 let	0	0	0	0	0	0	0	
	30-39 let	--	0	++	0	0	0	0	
	40-49 let	+	+	-	0	0	0	0	
	50-59 let	+	0	0	0	0	0	0	
	60 let a v.	0	0	0	0	0	0	0	

5. PŘEDNÁŠKA

5.1. KVANTILOVÝ POPIS ŘADY

Vlastnosti statistické řady - složky statistické informace, které lze o souboru získat:

- ***poloha na stupnici proměnné***
  Jaká úroveň? Jaký stupeň? Jaká intenzita? Jak mnoho? Jak často?
- ***rozptýlení na stupnici proměnné***
  Jaká je rozptýlenost dat? Jak homogenní je soubor? Jaký stupeň heterogenity je v datech? Jak jsou si případy podobné/nepodobné? Chovají se jednotky souboru soudržně nebo se polarizují a extremizují?
- ***symetrie dat na stupnici proměnné***
  Jsou údaje rozloženy symetricky nebo na jedné straně se rozbíhají více než na druhé?
- ***směs dvou nebo několika homogenních souborů***
  Chovají se všechny případy podle stejných pravidel (principů, zákonů) nebo se soubor rozděluje na několik segmentů s různými vlastnostmi?
- ***cizí pozorování nepatřící do souboru***
  Patří krajní hodnoty organicky do souboru? Nejsou extrémní hodnoty souboru vychýleny nějakými specifickými faktory, které u ostatních případů nepůsobí? Není nutno tyto případy z analýzy vyloučit, aby nezkreslovaly informaci?

Charakteristiky uspořádané statistické řady:

- *minimum* a *maximum*; *rozpětí* = max - min
- *medián*=prostřední člen řady resp. průměr dvou prostředních členů
- *kvartily*=oddělují čtvrtinu nejnižších čísel a čtvrtinu nevyšších čísel řady
- *hradby*=oddělují pozorování, které patří k souboru jen s nepatrnou pravděpodobností; *vnitřní h.* oddělují *vnější pozorování* (outliers), *vnější hradby* oddělují *vzdálená pozorování* (extremes)
- *přilehlá pozorování*= pozorování, která přiléhají k vnitřním hradbám ale nepřekročí je, tj. jsou to poslední pozorování, která ještě nejsou indikována k vynechání ze souboru.
- *kvantily*=hodnoty na stupnici proměnné, které oddělují určitá zvolená procenta počtu pozorování; např. decily oddělují v řadě stejně početné skupiny, které tvoří desetiny souboru, kvartily tvoří čtvrtiny, tercily třetiny, kvintily pětiny apod.
- *kvantilová rozpětí*=rozdíl mezi posledním a prvním (t.j. nejvyšším a nejnižším) kvantilem, např. mezi třetím a prvním kvartilem (*kvartilové rozpětí*), mezi devátým a prvním decilem (*decilové rozpětí*)
- *kvantilová odchylka*=kvantilové rozpětí dělené počtem kvantilových intervalů mezi prvním a posledním kvantilem, např. kvartilová odchylka=kvartilové rozpětí/2, decilová odchylka=decilové rozpětí/8.

5.2. KVANTILOVÝ GRAF ROZPTÝLENÍ - BOX PLOT

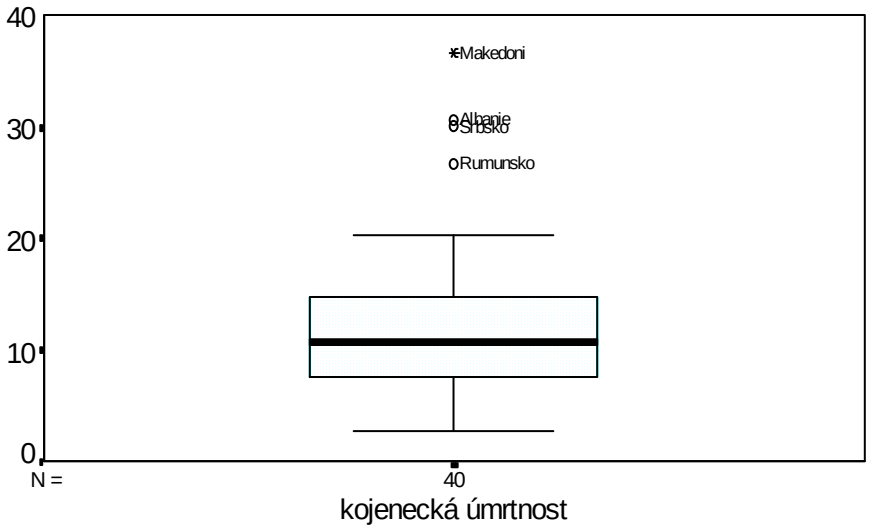
Úloha:
Charakterizujte názorně rozptýlení statistické řady.

		Statistic
kojenecká úmrtnost	Median	10.750
	Minimum	2.7
	Maximum	36.7
	Range	34.0
	Interquartile Range	7.250

Percentiles

		Percentiles		
		25	**50**	**75**
Tukey's Hinges	**kojenecka umrtnost**	**7.550**	**10.750**	**14.750**

Graf rozptýlení - Box plot:



Contents

Graf vyjadřuje rozložení dat na svislé ose. Obdélník je shora a zdola ohraničen kvartily, uprostřed obdélníku je značka mediánu. Úsečky jdoucí od kvartilových hodnot končí u přilehlých pozorování. Body (a popisem) jsou označeny vnější (kroužky) a vzdálená (hvězdička) pozorování. Z grafu je vidět, že čtyři země jsou diagnostikovány jako netypické pro soubor, jejich testové označení okamžitě naznačuje interpretaci plynoucí ze společných rysů všech čtyř případů. To vede k novým otázkám jako: Liší se balkánské země od ostatních? Existují nějaké další regionální rozdíly? Je nehomogennost zemí z hlediska kojenecké úmrtnosti významně vysvětlena faktorem „region Evropy“?

5.3. CIFROVÝ HISTOGRAM - STEM AND LEAF

Úloha:
Charakterizujte rozložení dat na stupnici přehledně a schematicky, avšak co možná v nejvyšším stupni detailu (informační přesnosti o jednotlivých pozorováních).

Metoda: Cifrový diagram - Stem and leaf

Výpis dat v tabulce a současně grafické vyjádření hustoty bodů na stupnici. Informuje o poloze i rozptýlení. Pod tabulkou/grafem se uvádějí vnější a vzdálená pozorování.

U čísel se definuje lodyha (stem) a listy (leaves) tak, že se určí část ciferného zápisu jako první a další část jako druhé hledisko. První (lodyha) znamená třídění, druhá (listy) je zápis čísla. Prakticky jde o informaci o a) rozložení a hustotě pozorování podél stupnice ve skupinách určených lodyhou, a o b) uspořádání hodnot. Výsledkem je zápis uspořádané statistické řady a to ve skupinách.

Postup je nejlépe vidět z příkladu:

Kojenecká úmrtnost - Stem-and-Leaf Plot

Frequency	Stem & Leaf
2.00	0 . 24
16.00	0 . 5577777777777888
13.00	1 . 0001111123444
4.00	1 . 5678
1.00	2 . 0

4.00 Extremes (>=27)

Stem width: 10.0
Each leaf: 1 case(s)

Stem = desítková cifra - určuje interval třídění
Leaf = jednotková cifra - určuje zápis každé jednotky

Šířka lodyhy je 10%, v tabulce je ale rozdělena na dvě části (0-4) a (5-9), aby byl graf podrobnější a aby lépe charakterizoval hustotu bodů a tím pozici řady na škále. Toto rozdělení provádí počítač automaticky. Pro větší počet bodů také může rozdělit šířku lodyhy na pět částí 0-1,2-3,4-5, atd.

Zápis se provádí v tomto případě s přesností na jedno procento
původní data byla zaznamenána s přesností na jednu desetinu procenta.

Tabulka resp. tříděný seznam má také roli grafického zobrazení (z toho důvodu počítač netiskne tento výstup v proporcionálním písmu). Délka řádku je přímo proporcionální počtu pozorování v dané skupině. V tabulce/grafu nejsou zahrnuty vnější a vzdálená pozorování.

Extreme Values	Case		ZEME	Value Number
kojenecká úmrtnost	1	29	Makedonie	36.7
	2	30	Albanie	30.8
	3	31	Srbsko	30.2
	4	32	Rumunsko	26.9

6. PŘEDNÁŠKA

6.1. PRŮMĚR

Průměr je mírou polohy, vyjadřuje střed datové řady ve smyslu těžiště bodů umístěných na stupnici proměnné, číselná hodnota charakterizuje pozici skupiny dat (souboru, podsouboru) na škále.

Výhody průměru:

a) je to míra široce používaná a obecně přijatá,
b) je vhodná pro statistickou práci, protože pro ni platí rozsáhlá statistická teorie, která poskytuje mnoho užitečných metod,
c) platí pro ní zákony velkých čísel v jednoduchém tvaru (tzv. centrální limitní věty)
d) má vhodné vlastnosti pro aplikace: mění se stejně s posunutím počátku škály proměnné i s násobkem měřítka škály
e) využívá všech dat, tj. veškeré informace, která je dostupná.

Nevýhodou průměru je, že jeho hodnota je velmi citlivá na extrémní hodnoty, které jej vychylují, a je nestabilní u polarizovaných rozložení.

Velkou předností je relativně snadné zjištění přesnosti měření průměru ve výběrových souborech.

Úloha: Určete míru polohy na stupnici proměnné.

Metoda: Výpočet průměru:

$$\overline{X} = \frac{1}{n}\sum X_i$$

Příklad:

Průměrná spokojenost se službami v obchodní síti (měřeno na sedmibodové stupnici):

SPOKOJENOST	N	Průměr
umístění prodejen	503	4,36
čistota prodejen	502	3,99
informace o novém zboží	501	3,54
šíře sortimentu	504	3,36
prodejní doba	504	3,27
orientace ve zboží	502	3,26
CELKOVÁ SPOKOJENOST	505	3,17
příjemná obsluha	505	2,91
kvalita potravin	503	2,86
ceny proti ostatním	503	2,64
prostornost	503	2,6

6.2. PRŮMĚRY - INTERVALY SPOLEHLIVOSTI

Interval spolehlivosti (konfidenční interval) vyjadřuje přesnost měření průměru ve výběrovém souboru. Přesněji: neurčitost závěru o průměru, která plyne z dat, chceme-li provést závěr se zvolenou spolehlivostí (v praxi obvykle 95%, 99%). Jde o intervalový odhad průměru.

Úloha: Určete míru polohy na stupnici proměnné a interval spolehlivosti pro něj.

Metoda: Výpočet vychází ze vzorců:

$$\overline{X} = \frac{1}{n}\sum X_i$$

$$\mu = \overline{X} \pm z_\alpha * s/\sqrt{n}$$

$$\mu = \overline{X} \pm z_\alpha * sterr$$

α = hladina spolehlivosti
s = směrodatná odchylka
sterr= standardní chyba
z = skór spolehlivosti

Pro běžně požadovanou hladinu spolehlivosti 95% je z=1.96, v praxi se běžně používá z=2.0.

Interval spolehlivosti, a tím i neurčitost závěrů o průměru, jsou závislé na:

a) zvoleném stupni spolehlivosti - čím vyšší spolehlivost závěru, tím širší interval; vyšší požadavek spolehlivosti je reprezentován vyšší hodnotou skóru *z*;

b) na heterogenitě souboru, která je reprezentovaná směrodatnou odchylkou *s*, čím větší je rozrůzněnost hodnot, tím nižší je přesnost zjišťování průměru;

c) na velikosti souboru - šířka konfidenčního intervalu klesá s odmocninou z počtu pozorování *(POZOR! Dvojnásobný výběr a s tím spojené dvojnásobné přímé náklady na jednotku vedou pouze k 1.4x kratšímu intervalu spolehlivosti)*.

Vzorec platí pro normálně rozložená data (Gaussova křivka) libovolným N alespoň rovným dvěma, ale podle zákona velkých čísel i pro jakékoliv rozložení dat, stejnou přesnost vzorce dosáhneme s větším počtem pozorování (v tom případě záleží přesnost vzorce na tvaru rozložení, např. pro rovnoměrné rozložení platí vzorec dobře už od 12 pozorování, u polaritních, velmi nesymetrických rozložení a při existenci extrémních dat je zapotřebí pozorování podstatně více).

Vzorec lze numericky zlepšit nahrazením skórů *z* Studentovými *t*-skóry.

Příklad:

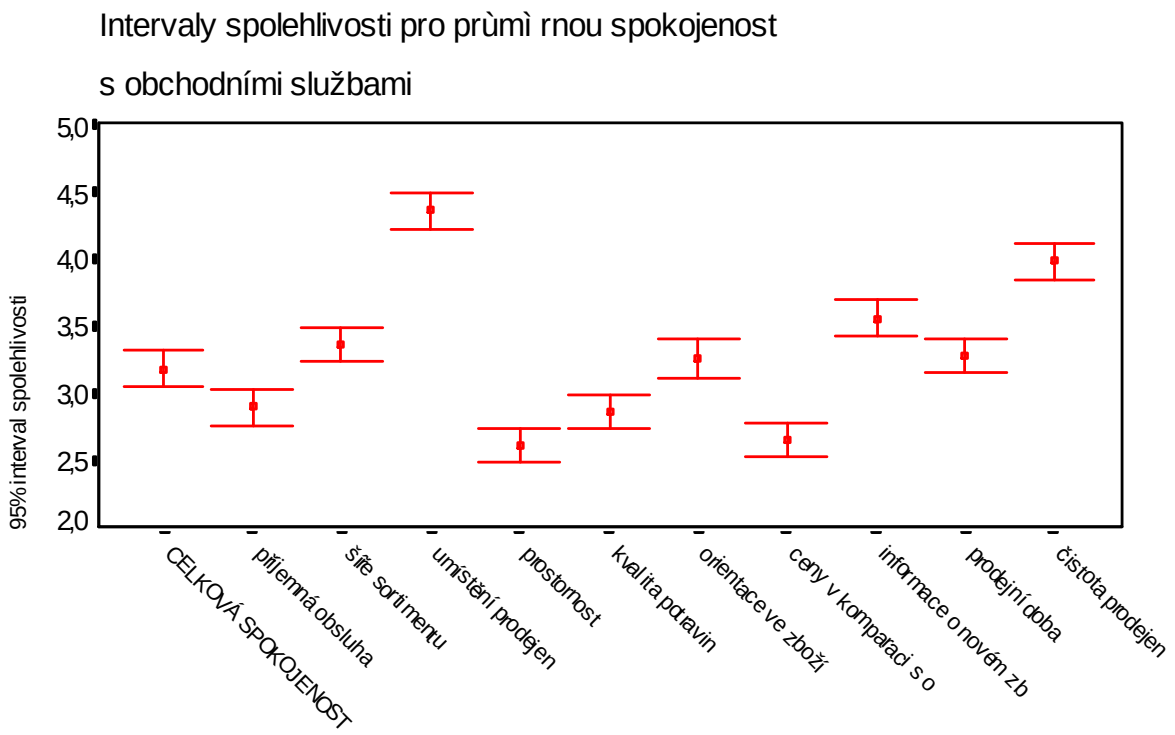
SPOKOJENOST	N	Průměr	Standardní chyba	Interval spolehlivosti 95%	
umístění prodejen	503	4,36	0,07	4,22	4,5
čistota prodejen	502	3,99	0,07	3,85	4,13
informace o novém zboží	501	3,54	0,07	3,4	3,68
šíře sortimentu	504	3,36	0,06	3,24	3,48
prodejní doba	504	3,27	0,07	3,13	3,41
orientace ve zboží	502	3,26	0,07	3,12	3,4
CELKOVÁ SPOKOJENOST	505	3,17	0,07	3,03	3,31
příjemná obsluha	505	2,91	0,07	2,77	3,05
kvalita potravin	503	2,86	0,06	2,74	2,98
ceny proti ostatním	503	2,64	0,06	2,52	2,76
prostornost	503	2,6	0,07	2,46	2,74

6.3. PRŮMĚRY - ZOBRAZENÍ INTERVALŮ SPOLEHLIVOSTI

Úloha: Vyjádřete přesnost průměru v grafu.

Metoda: Graf typu “error bar”

Graf zobrazuje průměr a interval pro každou proměnnou.



6.4. ROZPTYL A SMĚRODATNÁ ODCHYLKA

Rozptyl měří rozptýlenost, heterogenitu, vnitřní nepodobnost a rozmanitost údajů. Všechny údaje poměřuje vzhledem k průměru a charakterizuje odlišnost jednotky od průměru čtvercem rozdílu. V případě, že všechny údaje jsou číselně stejné (úplná homogenita) rozptyl je roven nule. Čím jsou hodnoty dekoncentrovanější, tj. vzdálenější od průměru, tím větší je hodnota rozptylu.
Rozptyl je základním pojmem pro *statistickou explikaci*, která je založena na určení, jak se podílí různé faktory na rozptylu závislé proměnné.

Úloha: Charakterizujte heterogenitu/homogenitu souboru (podsouboru) z hlediska proměnné.

Metoda: Heterogenita dat na číselné ose je nejčastěji charakterizována průměrnou čtvercovou odchylkou jednotlivých hodnot od průměru, nebo její odmocninou.

$$varX = s^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$$

$$s = \sqrt{varX}$$

varX=rozptyl proměnné X
s =směrodatná odchylka

Vzorce lze zapsat i v jiném tvaru (výpočetním).
Uvedený vzorec pro rozptyl není přesným průměrem, neboť faktor, kterým dělíme je (*N* -1). Vyskytují se i vzorce s dělením *N*. Uvedená definice má řadu předností a proto je běžně přijímána.

Příklad: Hodnocení služeb

	N	Průměr	sm. chyba	směr. odch.	rozptyl
umístění prodejen	503	4.36	.07	1.50	2.25
čistota prodejen	502	3.99	.07	1.56	2.42
informace o novém zboží	501	3.54	.07	1.49	2.21
šíře sortimentu	504	3.36	.06	1.38	1.89
prodejní doba	504	3.27	.07	1.47	2.16
orientace ve zboží	502	3.26	.07	1.55	2.41
CELKOVÁ SPOKOJENOST	505	3.17	.07	1.56	2.43
příjemná obsluha	505	2.91	.07	1.59	2.52
kvalita potravin	503	2.86	.06	1.38	1.92
ceny v komparaci s ostatními	503	2.64	.06	1.40	1.95
prostornost	503	2.60	.07	1.47	2.15

Pořadí položek podle nejednotnosti názorů

položka	s	rozptyl	poměr heterogenity	%
příjemná obsluha	1.59	2.52	1.33	33%
CELKOVÁ SPOKKOJENOST	1.56	2.43	1.29	29%
čistota prodejen	1.56	2.42	1.28	28%
orientace ve zboží	1.55	2.41	1.28	28%
umístění prodejen	1.5	2.25	1.19	19%
informace o novém zboží	1.49	2.21	1.17	17%
prodejní doba	1.47	2.16	1.14	14%
prostornost	1.47	2.15	1.14	14%
ceny v komparaci s ostatními	1.4	1.95	1.03	3%
kvalita potravin	1.38	1.92	1.02	2%
šíře sortimentu	1.38	1.89	***1.00***	0%

7. PŘEDNÁŠKA

7.1. POROVNÁNÍ PRŮMĚRU S NOMINÁLNÍ HODNOTOU

Úloha: Porovnání průměru s předem stanovenou hodnotou, standardem, normou, prahem apod.

Komparace průměru s hypotetickou hodnotou se provádí **Studentovým t-testem** pro jeden výběr:

Situace: Jedna populace a jedna proměnná; testujeme hypotetickou hodnotu průměru této proměnné; všechna individuální měření jsou získávána nezávisle na sobě, tj. jedno měření neovlivňuje druhé ani v rámci jednoho souboru;

Hypotézy:
A) dvojstranná alternativa: $H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$
B) jednostranná alternativa: $H_0: \mu = \mu_0$ $H_A: \mu > \mu_0$
C) jednostranná alternativa: $H_0: \mu = \mu_0$ $H_A: \mu < \mu_0$

Testová statistika:

$$t = \sqrt{n}\frac{\overline{X} - \mu_0}{S}, \qquad df = n - 1$$

kritické hodnoty se získají v tabulkách jednostranného a dvojstranného t-testu
pro dvojstranný test se používá statistika $|t|$.

Poznámka: tabulky dvojstranných a jednostranných kritických hodnot jsou na sebe převoditelné tak, že dvojstranná kritická hodnota pro stejný počet stupňů volnosti df a hodnotu rizika α je je stejná jako jednostranná kritická hodnota pro riziko $\alpha/2$.

Příklad: Spokojenost se službami se měří na stupnici 1 až 7. Středem stupnice je tedy hodnota 4. Otázka: je daný rys služeb hodnocen významně nad nebo pod tímto středem? Které položky jsou nadprůměrné a které jsou podprůměrné?

One-Sample Statistics

	N	Mean	Std. Dev.	St. Error
CELKOVÁ SPOKOJENOST	505	3.17	1.56	.069
příjemná obsluha	505	2.91	1.59	.071
šíře sortimentu	504	3.36	1.38	.061
umístění prodejen	503	4.36	1.50	.067
prostornost	503	2.60	1.47	.065
kvalita potravin	503	2.86	1.38	.062
orientace ve zboží	502	3.26	1.55	.069
ceny v komparaci s ostatními	503	2.64	1.40	.062
informace o novém zboží	501	3.54	1.49	.066
prodejní doba	504	3.27	1.47	.066
čistota prodejen	502	3.99	1.56	.069

One-Sample Test

	Test Value = 4					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tail)	Mean Diff.	Lower	Upper
CELKOVÁ SPOKOJENOST	-11.914	504	.000	-.83	-.96	-.69
příjemná obsluha	-15.435	504	.000	-1.09	-1.23	-.95
šíře sortimentu	-10.526	503	.000	-.64	-.77	-.52
umístění prodejen	5.379	502	.000	**.36**	.23	.49
prostornost	-21.394	502	.000	-1.40	-1.53	-1.27
kvalita potravin	-18.523	502	.000	-1.14	-1.26	-1.02
orientace ve zboží	-10.722	501	.000	-.74	-.88	-.61
ceny v komparaci s ostatními	-21.772	502	.000	-1.36	-1.48	-1.23
informace o novém zboží	-6.854	500	.000	-.46	-.59	-.32
prodejní doba	-11.172	503	.000	-.73	-.86	-.60
čistota prodejen	-.143	501	.886	***-.01***	-.15	.13

7.2. POROVNÁNÍ PRŮMĚRŮ DVOU SKUPIN

Úloha: Jsou průměrné hodnoty dvou skupin stejné nebo se liší?

Komparace průměrů dvou skupin se provádí **Studentovým t-testem**:

Situace: Dvě populace nebo dvě části jedné populace, kterým odpovídají dva datové výběrové soubory nebo dvě nepřekrývající se části jednoho souboru; data se sbírají nezávisle v obou souborech, tj. výběr v jednom souboru (části) neovlivňuje výběr v druhém souboru (části); všechna individuální měření uvnitř souboru (části) jsou získávána nezávisle na sobě, tj. jedno měření neovlivňuje druhé ani v rámci jednoho souboru;
úlohou je komparovat průměry obou populací resp. subpopulací.

Hypotézy: A) dvojstranná alternativa: $H_0: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$

B) jednostranná alternativa: $H_0: \mu_1 = \mu_2$ $H_A: \mu_1 > \mu_2$

C) jednostranná alternativa: $H_0: \mu_1 = \mu_2$ $H_A: \mu_1 < \mu_2$

Testová statistika se používá ve dvou variantách podle toho, zda v obou komparovaných souborech jsou či nejsou stejné rozptyly.

Pro případ stejných rozptylů platí známý vzorec

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} * \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}, \qquad df = n_1 + n_2 - 2$$

kritické hodnoty se získají v tabulkách jednostranného a dvojstranného t-testu
pro dvojstranný test se používá statistika $|t|$.

Poznámka: tabulky dvojstranných a jednostranných kritických hodnot jsou na sebe převoditelné tak, že dvojstranná kritická hodnota pro stejný počet stupňů volnosti df a hodnotu rizika α je stejná jako jednostranná kritická hodnota pro riziko 2α.

Pro nestejné rozptyly se používá složitější vzorec.

V případě, že je nulová hypotéza zamítnuta a přijímáme rozhodnutí o různých průměrech, zajímá nás přesnost rozdílu:

Konfidenční interval pro rozdíl $\delta = \mu_1 - \mu_2$

$$\delta = (\overline{X}_1 - \overline{X}_2) \pm t_{(df,\alpha)} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} * \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$t_{(df,\alpha)}$ = *kritická hodnota dvoustranného* t *– testu*

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Mean	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
CELKOVÁ SPOKOJENOST	Stejné rozptyly	.203	.653	.150	502	.881	.02	.14	-.25	.29
	Nestejné rozptyly			.149	500.623	.881	.02	.14	-.25	.29
příjemná obsluha	Stejné rozptyly	.104	.748	.168	502	.867	.02	.14	-.25	.30
	Nestejné rozptyly			.168	501.906	.867	.02	.14	-.25	.30
šíře sortimentu	Stejné rozptyly	.765	.382	-1.231	501	.219	-.15	.12	-.39	.09
	Nestejné rozptyly			-1.229	496.671	.219	-.15	.12	-.39	.09
umístění prodejen	Stejné rozptyly	2.843	.092	-.756	500	.450	-.10	.13	-.36	.16
	Nestejné rozptyly			-.758	497.960	.449	-.10	.13	-.36	.16
prostornost	Stejné rozptyly	3.106	.079	.972	500	.331	.13	.13	-.13	.39
	Nestejné rozptyly			.974	499.183	.331	.13	.13	-.13	.38
kvalita potravin	Stejné rozptyly	.428	.513	.151	500	.880	.02	.12	-.22	.26
	Nestejné rozptyly			.151	496.301	.880	.02	.12	-.22	.26
orientace ve zboží	Stejné rozptyly	3.511	.062	-.196	499	.845	-.03	.14	-.30	.25
	Nestejné rozptyly			-.196	498.573	.845	-.03	.14	-.30	.25
ceny v komparaci s ostatními	Stejné rozptyly	3.464	.063	2.482	500	.013	.31	.12	.06	.55
	Nestejné rozptyly			2.488	498.899	.013	.31	.12	.06	.55
informace o novém zboží	Stejné rozptyly	.145	.704	1.308	498	.191	.17	.13	-.09	.44
	Nestejné rozptyly			1.309	497.833	.191	.17	.13	-.09	.44
prodejní doba	Stejné rozptyly	.100	.752	1.366	501	.173	.18	.13	-.08	.44
	Nestejné rozptyly			1.365	499.649	.173	.18	.13	-.08	.44
čistota prodejen	Stejné rozptyly	5.709	.017	.605	499	.545	.08	.14	-.19	.36
	Nestejné rozptyly			.607	497.247	.544	.08	.14	-.19	.36

7.3. POROVNÁNÍ PRŮMĚRŮ DVOU PROMĚNNÝCH - 1 SOUBOR

Úloha: Jsou průměrné úrovně odpovědí na dvě otázky stejné?

Komparace průměrů dvou proměnných na jednom souboru se provádí jednovýběrovým **Studentovým t-testem**:

Situace: Jedna populace nebo jedna vybraná skupina odpovídá na dvě různé otázky, které mají stejnou škálu odpovědí;
individuální odpovědi na dotazník jsou vzájemně nezávislé, tj. jedno měření neovlivňuje v rámci souboru druhé;
úlohou je komparovat průměry obou proměnných a tak zjistit jejich vzájemnou pozici na škále.

Hypotézy:
A) dvojstranná alternativa: $H_0: \mu_X = \mu_Y$ $H_A: \mu_X \neq \mu_Y$
B) jednostranná alternativa: $H_0: \mu_X = \mu_Y$ $H_A: \mu_X > \mu_Y$
C) jednostranná alternativa: $H_0: \mu_X = \mu_Y$ $H_A: \mu_X < \mu_Y$

Testová statistika: používá *t*-test pro komparaci s předem určenou hodnotou (nominálem) aplikovaný na rozdíl *d=X-Y*.

Proto lze hypotézy přeformulovat (*reparametrizace modelu*):

*Hypotézy pro rozdíl **d=X-Y** s očekávanou hodnotou* δ*:*

A) dvojstranná alternativa: $H_0: \delta = 0$ $H_A: \delta \neq 0$
B) jednostranná alternativa: $H_0: \delta = 0$ $H_A: \delta > 0$
C) jednostranná alternativa: $H_0: \delta = 0$ $H_A: \delta < 0$

Vzorce lze ekvivalentně formulovat pro obě parametrizace (s_d=směrodatná odchylka rozdílu):

$$t = \sqrt{n}\frac{\bar{d}}{s_d}, \qquad df = n-1$$

$$t = \sqrt{n}\frac{\bar{X}-\bar{Y}}{\sqrt{(s_X^2 - 2s_X s_Y r_{XY} + s_Y^2)}}, \quad df = n-1, \quad r = korelační\,koeficient$$

kritické hodnoty se získají v tabulkách jednostranného a dvojstranného t-testu
pro dvojstranný test se používá statistika $|t|$.

Poznámka: tabulky dvojstranných a jednostranných kritických hodnot jsou na sebe převoditelné tak, že dvojstranná kritická hodnota pro stejný počet stupňů volnosti df a hodnotu rizika α je stejná jako jednostranná kritická hodnota pro riziko $\alpha/2$.

V případě, že je nulová hypotéza zamítnuta a přijímáme rozhodnutí o různých průměrech, zajímá nás přesnost rozdílu:

Konfidenční interval pro rozdíl $\delta = \mu_X - \mu_Y$

$$\delta = d \pm t_{\alpha,df} \frac{s_d}{\sqrt{n}}$$

Příklad: Porovnáme průměrnou spokojenost (na škálach 1-7) pro různé aspekty služeb a to vzájemně po dvojicich.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	prostornost	2.60	502	1.47	.07
	kvalita potravin	2.86	502	1.38	.06
Pair 2	kvalita potravin	2.86	502	1.38	.06
	orientace ve zboží	3.26	502	1.55	.07
Pair 3	šíře sortimentu	3.36	503	1.38	.06
	umístění prodejen	4.36	503	1.50	.07

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tail)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	prostornost - kvalita potravin	-.26	1.58	.07	-.40	-.12	-3.648	501	.000
Pair 2	kvalita potravin - orientace ve zboží	-.40	1.74	.08	-.55	-.24	-5.116	501	.000
Pair 3	šíře sortimentu - umístění prodejen	-1.00	1.54	.07	-1.14	-.87	-14.577	502	.000

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	prostornost & kvalita potravin	502	.388	.000
Pair 2	kvalita potravin & orientace ve zboží	502	.305	.000
Pair 3	šíře sortimentu & umístění prodejen	503	.428	.000

7.4. POROVNÁNÍ ROZPTYLŮ DVOU SKUPIN

Úloha: Je variabilita ve dvou skupinách (souborech) stejná?

Komparace rozptylů ve dvou skupinách se provádí **Fisherovým F- testem**.

Situace: Dvě populace nebo dvě části jedné populace, kterým odpovídají dva datové výběrové soubory nebo dvě nepřekrývající se části jednoho souboru;
data se sbírají nezávisle v obou souborech, tj. výběr v jednom souboru (části) neovlivňuje výběr v druhém souboru (části); všechna individuální měření uvnitř souboru (části) jsou získávána nezávisle na sobě, tj. jedno měření neovlivňuje druhé ani v rámci jednoho souboru; úlohou je komparovat průměry obou populací resp. subpopulací.

Hypotézy:
A) dvoustranná alternativa: $H_0: \sigma_1 = \sigma_2$ $H_A: \sigma_1 \neq \sigma_2$
B) jednostranná alternativa: $H_0: \sigma_1 = \sigma_2$ $H_A: \sigma_1 > \sigma_2$
C) jednostranná alternativa: $H_0: \sigma_1 = \sigma_2$ $H_A: \sigma_1 < \sigma_2$

Testová statistika: vyjadřuje podíl většího a menšího z obou rozptylů u dvoustranné alternativy, a podíl odpovídající poměru většího a menšího hypotetického rozptylu u jednostranné alternativy:

$$F = \frac{s_1^2}{s_2^2} \qquad s_1^2 \geq s_2^2$$

$$df = \left[(n_1 - 1), (n_2 - 1)\right]$$

kritické hodnoty se získají v tabulkách jednostranného a dvojstranného F-testu; u jednostranné alternativy musí ovšem být příslušný poměr s ní ve shodě.

Poznámka: tabulky dvojstranných a jednostranných kritických hodnot jsou na sebe převoditelné tak, že dvojstranná kritická hodnota pro stejný počet stupňů volnosti df a hodnotu rizika α je stejná jako jednostranná kritická hodnota pro riziko $\alpha/2$.

V případě, že je nulová hypotéza zamítnuta a přijímáme rozhodnutí o různých rozptylech, můžeme zjistit v dalším kroku zjistit *konfidenční interval* pro $F=\sigma_1^2/\sigma_2^2$

$$\frac{s_1^2}{s_2^2} F_{(df_d, df_n, \alpha/2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \frac{1}{F_{(df_n, df_d, \alpha/2)}}$$

$$F_{(df_n, df_d, \alpha/2)} = kritická\ hodnota\ F-testu$$

$$df_n = n_1 - 1 = stupně\ volnosti\ "nahoře"$$

$$df_d = n_2 - 1 = stupně\ volnosti\ "dole"$$

8. PŘEDNÁŠKA

8.1. JEDNODUCHÁ ANALÝZA ROZPTYLU - KOMPARACE PRŮMĚRŮ

Úloha:
Porovnejte průměry v K skupinách! Jsou průměry v populačních skupinách reprezentovaných daty stejné či se od sebe výběrové průměry nenáhodně liší a tyto rozdíly prokazují i rozdíly v populaci?

Úloha porovnání průměrů v několika skupinách je rozšířením případu *t*-testu pro dvě nezávislé skupiny na případ více nezávislých skupin.

Situace: Měříme hodnoty číselné proměnné u jednotek, které jsou klasifikovány do K nepřekrývajících se skupin, při čemž výběr jednotek v jedné skupině neovlivňuje výběr v žádné jiné (jedotky nejsou párovány ani jinak k sobě vzájemně mezi skupinami přiřazovány). Chyby měření se vzájemně neovlivňují, jsou nezávislé.

Testujeme hypotézy:

$H_0: \mu_i = \mu_j$ pro všechna *i* a *j* $H_A: \mu_i \neq \mu_j$ pro alespoň pro jednu dvojici

Alternativní formulace:

$H_0: \mu_i = \mu$ pro všechna *i* a nějakou konstantu μ
$H_A: \mu_i \neq \mu$ alespoň pro jednu skupinu *i*, tj. společná hodnota μ neexistuje

Další alternativní formulace pro jiné parametry (reparametrizace):

neznámé průměry zapíšeme
$\mu_i = \mu + \delta_i$ = společný průměr + efekt skupiny *i*

$H_0: \delta_i = 0$ pro všechna *i* $H_A: \delta_i \neq 0$ pro alespoň jednu skupinu

tedy: žádná skupina nevykazuje nenáhodně vzniklý systematický efekt vs. alespoň jedna skupina vykazuje nenulový efekt

Jednoduchá ANOVA je založena na jednoduché algebraické vlastnosti pro součty čtverců (mnohorozměrná analogie Pythagorovy věty):

součet čtverců rozdílů všech měření od společného průměru =
součet čtverců rozdílů všech měření od průměrů svých skupin
+ součet čtverců rozdílů všech měření zaměněných skupinovými průměry od společného průměru

$$\sum (X_{ik} - \overline{X})^2 = \sum (X_{ik} - \overline{X}_k)^2 + \sum n_k (\overline{X}_k - \overline{X})^2$$

X_{ik} = i – té pozorování k – té skupiny

$\overline{X}_k$ = průměr k – té skupiny

TSS = WSS + BSS

Protože součty čtverců charakterizují variabilitu (rozptýlení), můžeme tento vztah vyjádřit jinými

slovy:

variabilita celé řady = variabilita uvnitř skupin + variabilita mezi skupinami

Na tomto principu je konstruován Fisherův *F*-test, který je sumarizován v tabulce ANOVA:

Testová statistika:

$$F = \frac{\sum n_k (\overline{X}_k - \overline{X})^2 / (k-1)}{\sum (X_{ik} - \overline{X}_k)^2 / (n-k)}, \qquad df = (k-1, n-k)$$

= průměrná čtvecová odchylka mezi skupinami/
průměrná čtvercová odchylka uvnitř skupin

Kritickou hodnotu nalezneme v tabulkách Fisherova F-testu pro dané *alfa* a danou dvojici stupňů volnosti.

Komparace průměrů v krajích - test rovnosti průměrů:

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
šíře sortimentu * kraj respondenta	Between Groups	29.566	7	4.224	2.275	.027
	Within Groups	919.150	495	1.857		
	Total	948.716	502			
umístění prodejen * kraj respondenta	Between Groups	43.518	7	6.217	2.827	.007
	Within Groups	1086.221	494	2.199		
	Total	1129.739	501			

Spokojenost zákazníků se diferencuje v krajích - to znamená buď nerovnoměrnost kvality služeb v regionech a tedy různá kvalita regionálních manažerů nebo fakt, že jednotná politika není úspěšná vzhledem různým představám zákazníků.

8.2. JEDNODUCHÁ ANALÝZA ROZPTYLU - KORELAČNÍ POMĚR

Úlohy:
A) Jak silná je vazba mezi nezávislou nominální proměnnou a proměnnou číselnou?
B) Jak silně (jak dobře) vysvětluje rozdělení souboru do zvolených skupin variabilitu zkoumané číselné proměnné?
C) Která ze zvolených demografických, geografických či jiných segmentací je nejvýraznější?

Idea měření síly vztahu vychází z rovnice

variabilita celé řady = variabilita uvnitř skupin + variabilita mezi skupinami

Korelační poměr je definován jako poměr:

$$\eta^2 = \frac{\text{variabilita mezi skupinami}}{\text{variabilita celé řady}}$$

$$\eta^2 = \frac{\sum n_k (\overline{X}_k - \overline{X})^2}{\sum (X_{ik} - \overline{X})^2} = \frac{BSS}{TSS} = 1 - \frac{WSS}{TSS}$$

Vlastnosti korelačního poměru:

1) je roven nule, jestliže jsou všechny průměry stejné (BSS=0)
2) je roven jedné, jestliže jsou všechna data uvnitř každé skupiny stejná, tj. všechna jsou rovna společné hodnotě, která je zároveň průměrem a alespoň dvě skupiny se od sebe liší.

Stonásobek korelačního poměru se vyjadřuje v procentech jako $100\eta^2$%;
je to procento variability proměnné X, vysvětlené pomocí dané klasifikace (daného rozdělení do skupin).

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
šíře sortimentu * kraj respondenta	Between Groups	29.566	7	4.224	2.275	.027
	Within Groups	919.150	495	1.857		
	Total	948.716	502			

Z tabulky ANOVA spočteme $100\eta^2$% jako (29.566/948.716)x100%=3.12%.

Korelační poměr patří do skupiny měr, které mají obecnou vlastnost poměru vysvětlené variance zvoleným modelem, nazvaných *koeficienty determinace*.

8.3. JEDNODUCHÁ ANALÝZA ROZPTYLU - TEST ROZPTYLŮ

Úloha:
Jsou rozptyly v K skupinách stejné?

Pro aplikaci analýzy rozptylu na porovnání průměrů v *K* skupinách předpokládáme:

a) všechna pozorování jsou provedena nezávisle na sobě
b) rozložení dat ve skupinách je normální (odpovídá Gaussově křivce)
c) rozptyly ve skupinách jsou stejné.

Předpoklad (a) musí být zajištěn při sběru dat resp.při měření. Předpoklad (b) lze ověřovat testen dobré shody k normálnímu rozdělení. Předpoklad (c) se ověřuje různými testy (Bartlett, Box, Levene).

Simulační studie a zkušenost z aplikací ukazují, že kritickým předpokladem je (a), jehož nedodržení velmi silně ovlivňuje aplikabilitu. Předpoklady (b) a (c) nemají na výsledky rozhodující vliv. Metoda je proti jejich porušení značně *robustní*.

Přesto předpoklad rovnosti rozptylů testujeme, neboť jeho porušení má i meritorní interpretaci a přijetí hypotézy o nerovnosti může mít závažné praktické aplikační důsledky. Při nerovnosti rozptylů také volíme jiný přístup při testování kontrastů.

Hypotéza:

$H_0: \sigma^2_i = \sigma^2_j$ pro všechna *i* a *j*
$H_A: \sigma^2_i \neq \sigma^2_j$ pro alespoň pro jednu dvojici

Testy jsou založeny na různých principech a předpokládají výpočet na počítači.

Příklad: Porovnání krajů vzhledem k hodnocení čistoty prodejen.

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
čistota prodejen	3.315	7	493	.002

Rozptyly se od sebe významně liší, metoda však neurčuje, které se od sebe liší a které ne. Proto použijeme popisné statistiky, abychom se orientovali.

čistota prodejen

			N	Mean	Std. Dev.	Std. Err.	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
	kraj	Praha	53	4.55	1.42	.20	4.16	4.94
		Středočeský kraj	63	4.37	1.71	.22	3.94	4.79
		Jihočeský kraj	34	4.32	1.34	.23	3.86	4.79
		Západočeský	49	3.96	1.66	.24	3.48	4.44
		Severočeský	56	3.64	1.48	.20	3.25	4.04
		Východočeský	51	4.02	1.19	.17	3.68	4.35
		Jihomoravský	93	4.01	1.46	.15	3.71	4.31
		Severomoravský	102	3.53	1.70	.17	3.20	3.86
		Total	501	3.99	1.56	.07	3.85	4.13

Popisně je vidět, které směrodatné odchylky se od sebe liší a které ne. Dalším krokem analýzy by mohle být postupné párové porovnávání rozptylů ve skupinách a to buď separovaně nebo simultánně pomocí Bonferroniho nebo (lépe) Holmovy metody.

8.4. JEDNODUCHÁ ANALÝZA ROZPTYLU - KONTRASTY

Úlohy:
A) Ověření vztahu mezi skupinovými průměry.
B) Porovnání vážených průměrů skupin mezi sebou.
C) Hypotézy o kombinované segmentaci skupin.

Kontrast je lineární funkce průměrů, která vyjadřuje hypotézu složitější komparace:

$$Y = \sum c_i \overline{X}_i, \qquad platí \ \sum c_i = 0$$

Kontrast je komparační funkcí, v níž průměry vystupují váženě.

komparace dvou průměrů:

$$\overline{X}_1 - \overline{X}_2 = 0$$

komparace dvou rozdílů:

$$(\overline{X}_1 - \overline{X}_2) - (\overline{X}_3 - \overline{X}_4) = 0$$

$$\overline{X}_1 - \overline{X}_2 - \overline{X}_3 + \overline{X}_4 = 0$$

komparace dvou skupin:

$$\tfrac{1}{2}(\overline{X}_1 + \overline{X}_2) - \tfrac{1}{3}(\overline{X}_3 + \overline{X}_4 + \overline{X}_5) = 0$$

ekvivalentně(*násobeno* 6*ti*):

$$3(\overline{X}_1 + \overline{X}_2) - 2(\overline{X}_3 + \overline{X}_4 + \overline{X}_5) = 0$$

Příklad: Porovnání krajů vzhledem k hodnocení prodejen

čistota prodejen

			N	Mean	Std. Dev.	Std. Err.	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
	kraj	Praha	53	4.55	1.42	.20	4.16	4.94
		Středočeský kraj	63	4.37	1.71	.22	3.94	4.79
		Jihočeský kraj	34	4.32	1.34	.23	3.86	4.79
		Západočeský	49	3.96	1.66	.24	3.48	4.44
		Severočeský	56	3.64	1.48	.20	3.25	4.04
		Východočeský	51	4.02	1.19	.17	3.68	4.35
		Jihomoravský	93	4.01	1.46	.15	3.71	4.31
		Severomoravský	102	3.53	1.70	.17	3.20	3.86
		Total	501	3.99	1.56	.07	3.85	4.13

Pro testování průměrů prověříme rozptyly

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
čistota prodejen	3.315	7	493	.002

Tabulka ANOVA pro testování homogenity průměrů

		Sum of Squares	df	Mean Square	F	Sig.
čistota prodejen	Between Groups	57.617	7	8.231	3.512	.001
	Within Groups	1155.333	493	2.343		
	Total	1212.950	500			

Průměry nejsou shodné.
Ověření hypotéz o komparaci Prahy s moravskými kraji (kontrast 1) a se středočeským krajem (kontrast 2).

Contrast Coefficients

	kraj respondenta							
Contrast	Praha	StČ	JČ	ZČ	SČ	VČ	JM	SM
1 Pha vs. mor. kraje	2	0	0	0	0	0	-1	-1
2 Pha vs. stř.kraj	1	-1	0	0	0	0	0	0

Testy kontrastů

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tail)
čistota prodejen	Assume equal variances	1	1.55	.47	3.276	493	.001
		2	.18	.29	.638	493	.524
	Does not assume equal variances	1	1.55	.45	3.444	89.915	.001
		2	.18	.29	.627	113.992	.532

Výsledek: Praha se odlišuje od průměru moravských krajů, ale neodlišuje se od středočeského kraje vzhledem k hodnocení čistoty prodejen.

8.5. JEDNODUCHÁ ANALÝZA ROZPTYLU - SROVNÁNÍ PRŮMĚRŮ S REFERENČNÍ KATEGORIÍ

Úlohy:
A) Jak se liší průměry ve skupinách od referenční (kontrolní) kategorie.
Jsou stejné jako hodnota referenční skupiny nebo je prokázán rozdíl?
B) Jsou nové výrobky přijímány lépe než původní výrobek?
C) Je naše značka hodnocena lépe než ostatní značky na trhu?
D) Které skupiny splývají s referenční kategorií?

Úlohy komparace průměrů s referenční kategorií se vyskytují jednak při testování výrobků (jak 'in-hall¨, tak 'product placement'), při specielně designovaných nebo přirozeně podle používání výrobků a značek stratifikovaných výzkumných souborů.

Postupně porovnáváme každý průměr s referenčním pomocí testu hypotézy:

$H_{0_i}: \mu_i = \mu_1$ $H_{A_i}: \mu_i \neq \mu_1$ pro všechna $i \neq 1$, kde první skupina je referenční

Je to (K-1) hypotéz, které testujeme simultánně.

Porovnání (K-1) průměrů s referenční kategorií můžeme provést:

a) *opakováním (K-1) t-testů na konvenční hladině* významnosti bez opravy na simultánní inferenci;

b) opakováním (K-1) *t*-testů na konvenční hladině významnosti s *Bonferroniho opravou* na simultánní inferenci (aplikace významnosti s hladinou $\alpha/(K\text{-}1)$;

c) opakováním (K-1) *t*-testů na konvenční hladině významnosti s *Holmovou sekvenční opravou* na simultánní inferenci (aplikace významnosti s postupnými hladinami $\alpha/(K\text{-}1), \alpha/(K\text{-}2), \alpha/(K\text{-}3), ..., \alpha$); tento postup je silnější než Bonferroniho test.

d) *Dunnettové párový vícenásobný t-test* - je možno volit jednostranný nebo dvoustranný test.

Příklad: Sedm porovnání všech krajů s Prahou (referenční kategorie) vzhledem k hodnocení čistoty prodejen.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
čistota prodejen	Between Groups	57.617	7	8.231	3.512	.001
	Within Groups	1155.333	493	2.343		
	Total	1212.950	500			

Multiple Comparisons

Dependent Variable: čistota prodejen

Dunnett (2-sided)[a]

(I) kraj respondenta	(J) kraj respondenta	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Středočeský kraj	Praha	-.18	.285	.981	-.93	.56
Jihočeský kraj	Praha	-.22	.336	.977	-1.10	.65
Západočeský	Praha	-.59	.303	.234	-1.38	.20
Severočeský	Praha	-.90	.293	.013	-1.67	-.14
Východočeský	Praha	-.53	.300	.326	-1.31	.26
Jihomoravský	Praha	-.54	.263	.193	-1.22	.15
Severomoravský	Praha	-1.02	.259	.001	-1.69	-.34

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Závěr simultánního testování je: od Prahy se významně liší Severočeský a Severomoravský kraj, kde je spokojenost s čistotou prodejen významně nižší.
U ostatních regionů se rozdíl v rámci této komplexní komparační hypotézy neprokázal.

9. PŘEDNÁŠKA

9.1. KORELAČNÍ KOEFICIENT (LINEÁRNÍ)

Úlohy:

A) Souvisí spolu výskyt proměnné X a proměnné Y tak, že s vyššími hodnotami X se pojí vyšší hodnoty Y a (a nižšími nižší), či naopak s vyššími hodnotami X se pojí nižší hodnoty Y (a s nižšími X vyšší Y)?

B) Můžeme v datech zjisti souběžnost resp. protiběžnost hodnot dvou číselných proměnných?

C) Je hodnota Y důsledkem hodnoty X? Reprezentuje proměnná X příčinu pro důsledek Y?

D) Jsou X a Y nositeli (částečně) stejné informace?

E) Vylučují se (resp. doplňují se) X a Y nebo naopak jedno předpokládá druhé?

Na tyto otázky odpovídá Pearsonův lineární korelační koeficient, který je mírou souběžnosti/protiběžnosti hodnot dvou proměnných podél lineárního trendu (podél přímky). Korelovanost *přímá* znamená, že čím vyšší je X, tím vyšší je Y a čím nižší je X tím nižší je Y. Korelovanost *nepřímá* znamená, že čím vyšší je X, tím nižší je Y a čím nižší je X tím vyšší je Y. Intenzita korelace je určena tím, čím těsněji přiléhají dvojice (X,Y) k nějakému přímkovému trendu (kromě kolmic k oběma osám).

Korelační koeficient *r* je určen vzorcem:

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

$$r = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}\, X * \mathrm{var}\, Y}}$$

$$\mathrm{cov}(X,Y) = \frac{1}{n-1}\sum (X_i - \overline{X})(Y_i - \overline{Y})$$

Vlastnosti:

a) koeficient je definován vždy když X i Y nemají nulový rozptyl (nejsou to konstanty)
b) hodnoty korelačního koeficientu jsou kladné, když se v datech projevuje přímá úměra (kladný trend, relace „čím vyšší X tím vyšší Y“); jsou záporné, když se v datech projevuje nepřímá úměra (záporný trend, relace „čím vyšší X tím nižší Y a naopak“);
c) r=1, když dvojice (X,Y) leží na stoupající přímce; r=-1, když dvojice (X,Y) leží na klesající přímce; v obou případech lze jednoznačně určit jednu z proměnných pomocí druhé za použití lineárního převodu; *hovoříme o lineární závislosti;*
d) r=0 když u dvojic (X,Y) nelze nalézt žádnou stopu lineárního trendu;
e) čím více se blíží r k 1 nebo k -1, tím silnější lineární vazbu koeficient indikuje, tj. tím soustředěnější jsou body kolem svého lineárního trendu.

Statistické hypotézy:

H_0: $r = 0$; H_A: $r \neq 0$,

tj. H_0: lineární trend v datech neexistuje H_A: v datech lineární trend existuje

Test lze provést několika jednoduchými způsoby: Fisherovou z-transformací, Studentovým t-testem, pomocí tabulek pro malé soubory či přímého asymptotického vzorce pro velké soubory.

Pro analýzu více vztahů současně vytváříme matice korelačních koeficientů mezi zvolenými proměnnými: *čtvercové korelační matice* obsahují vztahy mezi všemi zvolenými proměnnými, *obdélníkové korelační matice* obsahují vztahy mezi dvěma množinami proměnných.

Příklad: 13 zemí je charakterizováno profilem proměnných, který vyjadřuje, jak dalece hodnotí respondenti v dané zemi jednotlivé aspekty jako podmínky pro získání bohatství.

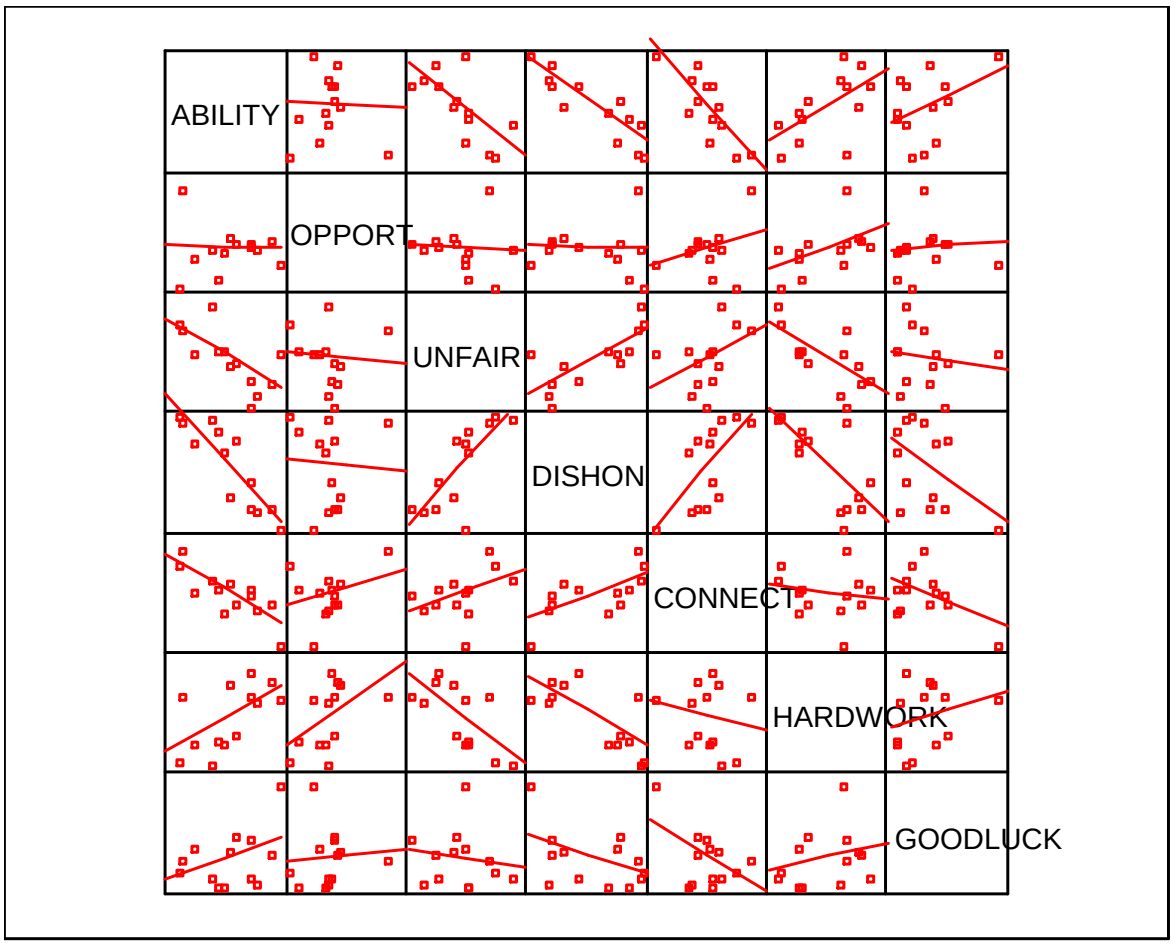
	Region	Schopnosti	Příležitost	Nestejné podmínky	Nečestnost	Konexe	Tvrdá práce	Štěstí
1	2 východ	3.29	4.27	4.08	4.10	4.46	3.54	3.09
2	2 východ	3.51	3.52	3.73	4.01	4.06	3.01	2.87
3	1 západ	3.83	3.83	3.24	3.10	3.91	3.69	3.15
4	1 západ	3.75	3.76	3.08	3.05	3.86	3.45	2.90
5	2 východ	3.48	3.76	4.42	4.16	4.16	2.72	2.95
6	1 západ	3.58	3.87	3.50	3.23	4.12	3.67	3.17
7	2 východ	3.62	3.81	3.57	3.91	3.90	3.07	3.30
8	1 západ	3.89	3.65	3.69	2.85	3.49	3.48	3.73
9	2 východ	3.37	3.68	3.70	3.86	4.02	2.98	3.20
10	2 východ	3.28	3.43	4.18	4.19	4.31	2.75	3.00
11	2 východ	3.55	3.74	3.77	3.76	3.83	2.97	2.86
12	1 západ	3.72	3.79	3.30	3.42	4.05	3.82	2.93
13	1 západ	3.72	3.82	2.87	3.11	3.99	3.52	3.28
	Průměr	3.5844	3.7651	3.6243	3.5947	4.0126	3.2818	3.1100

Korelační matice

		REG	AB	OPP	UNFAIR	DISHON	CONNECT	HW	GL
Pearson Correlation	REGION	1.000	-.806**	-.103	.754**	.938**	.442	-.826**	-.333
	ABILITY	-.806**	1.000	-.040	-.697**	-.879**	-.803**	.591*	.414
	OPPORTUN	-.103	-.040	1.000	-.083	-.056	.315	.520	.064
	UNFAIR	.754**	-.697**	-.083	1.000	.778**	.440	-.676*	-.157
	DISHON	.938**	-.879**	-.056	.778**	1.000	.677*	-.737**	-.476
	CONNECT	.442	-.803**	.315	.440	.677*	1.000	-.170	-.508
	HARDWORK	-.826**	.591*	.520	-.676*	-.737**	-.170	1.000	.268
	GOODLUCK	-.333	.414	.064	-.157	-.476	-.508	.268	1.000
Sig. (2-tailed)	REGION	.	.001	.737	.003	.000	.131	.000	.266
	ABILITY	.001	.	.897	.008	.000	.001	.033	.160
	OPPORTUN	.737	.897	.	.787	.856	.294	.068	.836
	UNFAIR	.003	.008	.787	.	.002	.132	.011	.608
	DISHON	.000	.000	.856	.002	.	.011	.004	.100
	CONNECT	.131	.001	.294	.132	.011	.	.579	.076
	HARDWORK	.000	.033	.068	.011	.004	.579	.	.375
	GOODLUCK	.266	.160	.836	.608	.100	.076	.375	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).



10. PŘEDNÁŠKA

10.1. REGRESNÍ ANALÝZA - JEDNODUCHÝ LINEÁRNÍ VZTAH

Úlohy:
A) Lze nalézt lineární vztah mezi nezávislou proměnnou X a závislou proměnnou Y?
B) Lze proměnnou Y predikovat pomocí hodnot X?
C) Projevuje se ve vztahu X a Y lineární trend?
D) Je možné charakterizovat kauzální hypotézu 'X ovlivňuje Y' lineární rovnicí? Platí taková hypotéza nebo jí data odporují?

Lineární vztah a studium lineárních trendů se provádí pomocí přímkového modelu. V modelu si klademe několik otázek?

a) je model relevantní (tj. obsahuje v sobě nějaké informace o vztahu vstupních proměnných)?
b) je model platný (tj. není zavádějící a zachycující nevhodnou část reality, která zkresluje celkový obrázek? jinak: neexistuje jiný, přesnější a vhodnější popis vztahu, jiný než lineární?
c) jak silný (těsný) je lineární vztah mezi X a Y? jak přesný je model, jak přesná je predikce? jak vhodný je lineární popis vztahu pro danou situaci?
d) jaké jsou hodnoty parametrů modelu? jaké jsou jejich odhady z dat?
e) lze identifikovat důvody snížené přesnosti modelu? které případy neodpovídají modelu a jak jej ovlivňují?

K tomu můžeme také položit otázky metodologického charakteru: na jakém principu a jakými postupy budeme hledat model a odhadovat jeho parametry a jaké hledisko optimality přijmeme pro určení nejlépe vyhovující rovnice.

Model jednoduché lineární regrese (pro dvě číselné proměnné X a Y):
(metoda nejmenších čtverců)

Regresní lineární rovnice a rovnice přímky vyjadřující vztah graficky:

$$Y = a + bX + \varepsilon$$

Y závislá proměnná, predikant, následek
X nezávislá proměnná, prediktor, příčina
ε chyba rovnice, chyba modelu, šum, chyba měření, souhrn nezahrnutých faktorů, odchylka rovnice, residuum

a regresní konstanta, parametr posunutí (hodnota Y pro X=0)
b regresní koeficient, je parametrem převodu X na Y, spád přímky (b=tg(φ), φ je úhel přímky s osou x)

$$Y = \tilde{Y} + \varepsilon$$

$$\tilde{Y} = a + bX$$

skutečná hodnota Y = hodnota modelu + chyba rovnice

O modelu předpokládáme, že residua splňují podmínku $\Sigma\varepsilon = 0$ pro všechna pozorování, z nichž je rovnice odvozena.

Model lineárního vztahu je tak vyjádřením naší představy o fungování vztahu, je to abstrakce, zachycení podstatné složky relace mezi X a Y, nebo je to zjednodušený pohled na vztah, který zachycuje jeho dominantní trend.

Kvalita regresní rovnice se posuzuje pomocí různých charakteristik:

a) *F-test významnosti rovnice* - tabulka ANOVA, která vychází z rozkladu celkového součtu čtverců pro proměnnou Y:

$$TSS = MSS + ESS$$

celkový součet čtverců = součet čtverců připadající na modelové hodnoty
+ součet čtverců připadající na chyby

F-test vyjadřuje poměr MSS a ESS upravený podle počtu stupňů volnosti. Významnost F-testu vyjadřuje významnost modelu: lze přijmout závěr, že model vyjadřuje část reality.

2. *Residuální rozptyl* - je odhad rozptylu residuí:

residuální rozptyl:

$$s_r^2 = \Sigma\varepsilon^2/(n-2)$$

3. *Koeficient determinace* je poměr MSS na TSS, tj. je to podíl vysvětlené variance Y pomocí modelu $Y=a+b_X$ na celkové variabilitě Y. Vyjadřuje se také v procentech a znamená procento vysvětlené variability Y pomocí zvoleného modelu. Tento princip je známý i z analýzy rozptylu (korelační poměr η^2) a je univerzální i pro jiné modely založené na *principu nejmenších čtverců.*

Koeficient determinace:

$$R^2 = MSS/TSS = 1 - ESS/TSS$$

resp.:

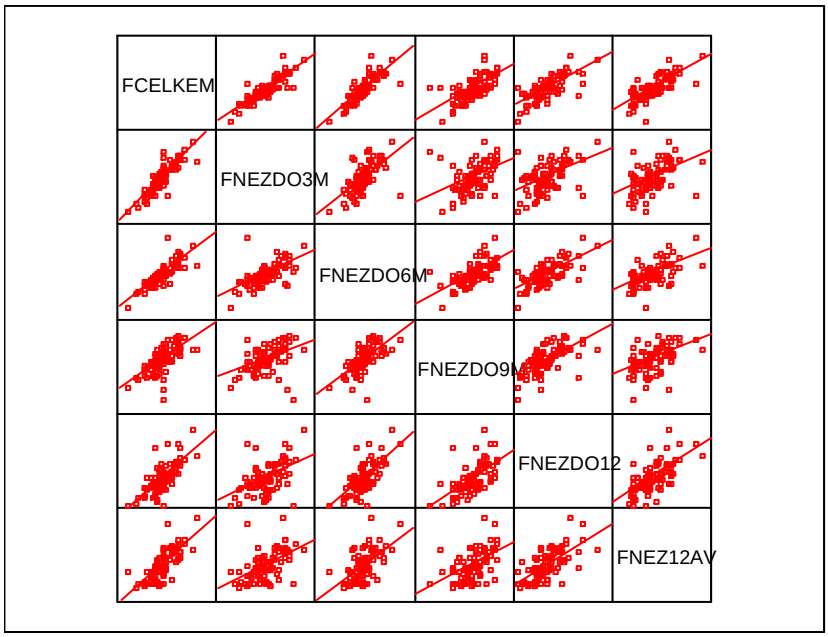
$100R^2\%$ v procentech

Residuální analýza:

Analýza residuí $\varepsilon = Y - a - b_X$ ukáže, zda se některá pozorování vymykají rovnici, a tudíž pro ně model neplatí a jejich přítomnost ve výpočtech zkresluje odhad modelových parametrů i charakteristik. Jejich analýza odhaluje také, zda jsou kolem přímky odchýlena náhodně, či zda je nutno hledat systematické vysvětlení jejich pravidelné struktury.

Residua mohou být také pro všechna pozorování uložena jako nová proměnná a může být dále analyzována dalšími metodami. Jejich význam je „část Y nevysvětlená pomocí X“ nebo „souhrn faktorů působících na Y nezahrnutý v X (ovšem včetně chyby měření)“.

Příklad: Do jaké míry závisí podíl žen mezi registrovanými nezaměstnanými na délce nezaměstnanosti? Lze odvodit celkový (podíl) nezaměstnaných žen z délky v registru?



Correlations

		FCELKEM	FNEZDO3M	FNEZDO6M	FNEZDO9M	FNEZDO12	FNEZ12AV
Pearson Correlation	FCELKEM	1.000	.868**	.833**	.674**	.741**	.760**
	FNEZDO3M	.868**	1.000	.631**	.424**	.478**	.495**
	FNEZDO6M	.833**	.631**	1.000	.654**	.698**	.596**
	FNEZDO9M	.674**	.424**	.654**	1.000	.629**	.491**
	FNEZDO12	.741**	.478**	.698**	.629**	1.000	.673**
	FNEZ12AV	.760**	.495**	.596**	.491**	.673**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

Závislost FCELKEM na jednotlivých časových kategoriích lze vyjádřit jako

FCELKEM = a + b*FNEZx + ‘chyba rovnice’,
FNEZx postupně znamená podíl žen na nezaměstnaných v kategoriích do 3 měsíců, do 6 měsíců, do 9měsíců, do 12 měsíců, nad 12 měsíců.

Postupně dostaneme rovnice, které jsou sumarizovány v tabulce:

Rovnice pro závislou proměnnou FCELKEM				
nez.prom.	**a**	**b**	**koef.det**	sign.
do 3 měs	0,088	0,893	0,75	.000
do 6 měs	0,215	0,636	0,69	.000
do 9 měs	0,29	0,514	0,46	.000
do 12 měs	0,365	0,416	0,55	.000
nad 12 měs	0,298	0,512	0,58	.000

11. PŘEDNÁŠKA

11.1. REGRESNÍ ANALÝZA-VÍCEROZMĚRNÁ

Úlohy:

A) Lze nalézt vztah mezi několika nezávislými proměnnými $X_1, X_2, \ldots, X_K$ a závislou proměnnou Y? Jak silný je takový vztah?

B) Lze proměnnou Y predikovat pomocí hodnot $X_1, X_2, \ldots, X_K$? Jak dobrá je taková predikce?

C) Je možné charakterizovat kauzální hypotézu '$X_1, X_2, \ldots, X_K$ ovlivňují Y' lineární rovnicí? Platí taková hypotéza nebo jí data odporují?

D) Jaké procento variability Y vysvětlují proměnné $X_1, X_2, \ldots, X_K$?

Vícenásobný lineární regresní model je rozšířením jednoduché lineární regrese na případ skupiny nezávislých proměnných $X_1, X_2, \ldots, X_K$, která ovlivňuje závislou proměnnou *Y* způsobem vyjádřeným regresní rovnicí. Otázky spojené s modelem jsou obdobné jako u jednoduché regrese, analyticky jich však můžeme položit více (model je bohatší).

a) je model relevantní (tj. obsahuje v sobě nějaké informace o vztahu skupiny $X_1, X_2, \ldots, X_K$ a *Y*)?
b) je model platný (tj. není zavádějící a zachycující nevhodnou část reality, která zkresluje celkový obrázek? jinak: neexistuje jiný, přesnější a vhodnější popis vztahu, jiný než lineární nebo model s jinými nezávislými proměnnými?
c) jak silný (těsný) je lineární vztah mezi $X_1, X_2, \ldots, X_K$ a *Y*? jak přesný je model, jak přesná je predikce?
d) jaké odhady parametrů modelu dostáváme z dat?
e) lze identifikovat důvody snížené přesnosti modelu? které případy neodpovídají modelu a jak jej ovlivňují?
f) které proměnné z $X_1, X_2, \ldots, X_K$ jsou v modelu zbytečné?
g) můžeme porovnat intenzitu vlivů jednotlivých proměnných $X_1, X_2, \ldots, X_K$ na *Y* mezi sebou?

Model vícenásobné lineární regrese (pro číselné proměnné $X_1, X_2, \ldots, X_K$ a Y):
(metoda nejmenších čtverců)

Regresní lineární rovnice:

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots + b_K X_K + \varepsilon$$

Y závislá proměnná, predikant, následek
X_k nezávislé proměnné, prediktory, příčiny
ε chyba rovnice, chyba modelu, šum, chyba měření, souhrn nezahrnutých faktorů, odchylka rovnice, residuum
a regresní konstanta, parametr posunutí (hodnota Y pro X=0)
b_k regresní koeficienty, je parametrem převodu X_k na Y,

$$Y = \tilde{Y} + \varepsilon$$
$$\tilde{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_K X_K$$

skutečná hodnota Y = hodnota modelu + chyba rovnice

O modelu předpokládáme, že residua splňují podmínku $\Sigma\varepsilon = 0$ pro všechna pozorování, z nichž je rovnice odvozena.

Model je vyjádřením naší představy o fungování vztahů mezi $X_1, X_2, \ldots, X_K$, je to abstrakce, zachycení podstatné složky relace mezi X_k a Y, nebo je to zjednodušený pohled na vztah, který zachycuje jeho dominantní trend.

Kvalita regresní rovnice se posuzuje pomocí obdobných charakteristik jako u jednoduché regrese:

a) *F-test významnosti rovnice* - tabulka ANOVA, která vychází z rozkladu celkového součtu čtverců pro proměnnou Y:

$$TSS = MSS + ESS$$

celkový součet čtverců = součet čtverců připadající na modelové hodnoty
+ součet čtverců připadající na chyby

F-test vyjadřuje poměr MSS a ESS upravený podle počtu stupňů volnosti. Významnost F-testu vyjadřuje významnost modelu: lze přijmout závěr, že model vyjadřuje část reality.

2. *Residuální rozptyl* - je odhad rozptylu reziduí:

residuální rozptyl:

$$s_r^2 = \Sigma\varepsilon^2/(n-2)$$

3. *Koeficient determinace* je poměr MSS na TSS, tj. je to podíl vysvětlené variance Y pomocí modelu Y=a+bX na celkové variabilitě Y. Vyjadřuje se také v procentech a znamená procento vysvětlené variability Y pomocí zvoleného modelu. Tento princip je známý i z analýzy rozptylu (korelační poměr η^2) a je univerzální i pro jiné modely založené na *principu nejmenších čtverců.*

Koeficient determinace:

$$R^2 = MSS/TSS = 1 - ESS/TSS$$

resp.:

$100R^2\%$ v procentech

Residuální analýza:

Analýza reziduí $\varepsilon = Y - (a+b_1X_1 + b_2X_2+ ...+ b_KX_K)$ ukáže, zda se některá pozorování vymykají rovnici, a tudíž pro ně model neplatí a jejich přítomnost ve výpočtech zkresluje odhad modelových parametrů i charakteristik. Jejich analýza odhaluje také, zda jsou odchylky náhodné, či zda je nutno hledat systematické vysvětlení jejich pravidelné struktury.

Residua mohou být také pro všechna pozorování uložena jako nová proměnná a může být dále analyzována dalšími metodami. Jejich význam je „část Y nevysvětlená pomocí X“ nebo „souhrn faktorů působících na Y nezahrnutý v $X_1, X_2, ... , X_K$ včetně chyby měření“.

Porovnání a existence vlivu proměnných:

Regresní rovnice vyjadřuje vztah mezi nezávislými proměnnými jako celkem a závislou proměnnou, ale také příspěvek každé jednotlivé proměnné samostatně, její čistý vliv v rámci celého seskupení $X_1, X_2, ... , X_K$. Každý člen $b_k X_K$ je tou částí Y, kterou na sebe váže X_K, kterou X_K vysvětluje nebo vytváří; může to být také čistý kauzální příspěvek této nezávislé proměnné v kauzálním modelu.

Každý koeficient b_k je možné testovat a zjišťovat, zda je významně rozdílný od nuly:

H_0: $b_k = 0$ H_A: $b_k \neq 0$

H_0 znamená, že X_K z rovnice vypadává, H_A vyjadřuje, že X_K má v v rovnici statisticky prokázaný vliv.

Přímé srovnání regresních koeficientů lze provádět jen tehdy, když jsou všechna X_K měřena na stejné stupnici.

Mají-li X_K různé škály měření, regresní koeficienty mají jednotlivě smysl vzhledem k proměnné X_K, ale mezi sebou srovnatelné nejsou (jde o převodní koeficienty nestejných entit na proměnnou Y). Proto v takovém případě používáme regresní rovnici mezi standardizovanými proměnnými. Koeficienty takové rovnice se nazývají beta.

Rovnice mezi z-skóry příslušných proměnných s koeficienty beta:

$$Z_Y = \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_K Z_K + \varepsilon$$

Vzhledem k posunutí počátků všech proměnných do nuly vymizí konstanta *a*.

Příklad: Jak se vytváří celková spokojenost z dílčích spokojeností? Jaké jsou váhy jednotlivých složek spokojenosti při vytváření celkové spokojenosti? Do regresní rovnice vstupuje celková spokojenost jako závislá proměnná a dílčí spokojenosti jako proměnné X_K.

Celková spokojenost =
a + b1*příjemná obsluha + b2*šíře sortimentu + b3*umístění prodejen + b4*prostornost + b5*kvalita potravin + b6*orientace ve zboží + b7*ceny + b8*informace + b9*prodejní doba + b10*čistota prodejen + 'chyba rovnice'

1. Test hypotézy, zda model vcelku přináší relevantní informaci - alespoň jedna z nezávislých

proměnných se významně projevuje (její koeficient je nenulový):

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	491.881	10	49.188	33.469	.000[b]
	Residual	718.655	489	1.470		
	Total	1210.536	499			

a. Dependent Variable: OT.7.1 CELKOVÁ SPOKOJENOST

b. Independent Variables: (Constant), OT.7.11 čistota prodejen, OT.7.6 kvalita potravin, OT.7.2 příjemná obsluha, OT.7.3 šíře sortimentu, OT.7.5 prostornost, OT.7.7 orientace ve zboží, OT.7.4 umístění prodejen, OT.7.10 prodejní doba, OT.7.8 ceny v komparaci s ostatními, OT.7.9 informace o novém zboží

Koeficienty rovnice pro jednotlivé nezávislé proměnné a jejich významnost (v tabulce jsou také uvedeny koeficienty beta přes to, že zde nebudou používány: všechny proměnné jsou měřeny na stejné stupnici 1-7):

Celková spokojenost =
.198 + .218* příjemná obsluha + .360*šíře sortimentu + .087*umístění prodejen + .007*prostornost - .103*kvalita potravin + .084*orientace ve zboží + .032*ceny + .157*informace + .116*prodejní doba - .066*čistota prodejen
+ 'chyba rovnice'

Podtržené koeficienty jsou významné, nepodtržené mohou být chápány jako nahodilé a nepodstatné. Sloupec *t* v tabulce ukazuje na hodnotu statistiky Studentova *t*-testu, vedle je dosažená významnost. V posledních dvou sloupcích jsou hranice konfidenčních intervalů pro koeficienty *b*.

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.198	.207		.958	.338	-.208	.604
	příjemná obsluha	.218	.037	.222	5.827	.000	.144	.291
	šíře sortimentu	.360	.048	.318	7.424	.000	.265	.455
	umístění prodejen	.087	.045	.084	1.947	.052	-.001	.175
	prostornost	.007	.045	.007	.162	.872	-.081	.095
	kvalita potravin	-.103	.044	-.092	-2.336	.020	-.191	-.016
	informace o novém zboží	.084	.045	.084	1.879	.061	-.004	.173
	ceny	.032	.054	.029	.597	.551	-.074	.139
	informace o novém zboží	.157	.055	.150	2.856	.004	.049	.266
	prodejní doba	.116	.050	.110	2.326	.020	.018	.214
	čistota prodejen	-.066	.051	-.066	-1.301	.194	-.167	.034

a. Dependent Variable: OT.7.1 CELKOVÁ SPOKOJENOST

Přehled případů, které vykazují extrémní odchylky chyby rovnice a přehled statistik o residuích a modelových (predikovaných) hodnotách.

Casewise Diagnostics[a]

Case Number	Std. Residual	OT.7.1 CELKOVÁ SPOKOJENOST
25	3.240	7
464	3.393	6
484	3.582	7

a. Dependent Variable: OT.7.1 CELKOVÁ SPOKOJENOST

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.47	6.55	3.18	.99	497
Residual	-3.39	4.34	.00	1.20	497
Std. Predicted Value	-2.724	3.396	.00	.997	497
Std. Residual	-2.794	3.582	.00	.988	497

a. Dependent Variable: OT.7.1 CELKOVÁ SPOKOJENOST

Graf vyjadřuje, jak si vzájemně odpovídají skutečná a predikovaná hodnota.