



MATEMATICKÁ STATISTIKA II.

P8**2006-11-20**

VÍCENÁSOBNÁ KORELACE A REGRESE:

Analýza kovariance:

- ✓ Analýza kovariance je statistická metoda, která kombinuje vlastnosti a principy analýzy rozptylu a rozšiřuje některé možnosti využití lineárních regresních modelů.
- ✓ Základní myšlenkou kovarianční analýzy je rozšíření nebo též modifikace modelu analýzy rozptylu s jedním nebo více kategoriálními faktory na model, který navíc obsahuje kontrolovatelné (nejlépe kvantitativní spojité, ale případně i další kategoriální) proměnné, které rovněž mají vliv na hodnoty vysvětlované či vysvětlovaných proměnných.
- ✓ Původním cílem analýzy kovariance je očištění studované závislosti vysvětlovaných proměnných na zvolených faktorech od zavádějícího působení doprovodných vlivů (označovaných za covariates).
- ✓ Působení doprovodných proměnných na vysvětlované proměnné je sice podstatné, ale není v dané úloze přímým předmětem zájmu.

Společné působení anebo smíchání vlivů:

- ✓ Regresní analýza má dva zásadně odlišné cíle. Prvním je předpověď průměrných nebo konkrétních hodnot vysvětlované proměnné pomocí skupiny vysvětlujících proměnných, zatímco druhým je kvantifikace individuálního vlivu vysvětlujících proměnných na vysvětlovanou proměnnou.
- ✓ Dobrá předpověď vyžaduje najít stabilní model, který odráží obecné rysy zkoumané závislosti a dobře vyhovuje výchozím pozorováním.
- ✓ Proti tomu úspěšná kvantifikace individuálního vlivu se opírá o kvalitní odhady regresní koeficientů nebo o jiné podobně interpretovatelné charakteristiky.
- ✓ Závislost dvou či více proměnných bývá zvykem posuzovat pomocí vhodných charakteristik, např. v regresní úloze to jsou především regresní koeficienty, ale mohou to být i jiné míry. Při snaze posuzovat význam dříve neuvažovaných proměnných (v souvislosti s analýzou kovariance se jim často říká doprovodné nebo kontrolní) je otázkou, jak je do analýzy zařadit a hodnotit.
- ✓ Pokud se hrubé charakteristiky (jednoduché regresní či korelační koeficienty), které neuvažují existenci mimo stojících (tedy dosud neuvažovaných) proměnných, z věcných hledisek velikostí **zásadně liší** od čistých charakteristik (dílní regresní nebo korelační koeficienty), uvažujících vliv dříve neuvažovaných proměnných, pak dochází k interpretačním potížím. Je zřejmé, že některá z těchto proměnných chybí a musí být do analýzy zařazena.
- ✓ V této situaci, kdy dochází k určitému promíchání vlivu, je obtížné až nemožné význam jednotlivých proměnných rozložit a smysluplně tak kvantifikovat podíl těchto proměnných na změnách hodnot vysvětlované nebo vysvětlovaných proměnných.
- ✓ Ve směsi významem nerozpoznatelných vlivů je obtížné rozhodnout, které proměnné jsou rozhodující a které je vhodné vypustit jako nepodstatné nebo duplicitní.
- ✓ Závažný je i jiný případ, kdy vztahy mezi vysvětlujícími a vysvětlovanými proměnnými se mění v závislosti na změnách hodnot nebo při různých úrovních (ne)uvažovaných proměnných (interakce dvou faktorů).
- ✓ Přitom předpoklad neexistence interakce mezi kvalitativními faktory a kvantitativními doprovodnými proměnnými má v analýze zásadní význam.
- ✓ Testovat existenci interakce proměnných je možné např. zařazením součinnových regresorů uvažovaných vysvětlujících proměnných.
- ✓ Třeba do lineární regresní rovnice se dvěma vysvětlujícími proměnnými ve formě $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ stačí přidat součinnový člen $\beta_3 X_1 X_2$ a po získání patřičných MNČ odhadů parametrů testovat hypotézu, že parametr β_3 je nulový.
- ✓ Zamítnutí této hypotézy lze považovat na zvolené hladině významnosti za statistický důkaz interakce, neboli za prokázání existence společného působení proměnných X_1 a X_2 na posuzovanou vysvětlovanou proměnnou.

Potřeba kontroly a modifikace nepřímých vlivů:

- ✓ Předchozí část naznačila důvody potřeby kontrolovat (hlídat) proměnné, které přímo nesouvisí s danou úlohou, ale jejichž vliv na vysvětlované proměnné je zjištěn, i když v dané úloze není hlavním předmětem zájmu.



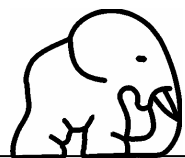
- ✓ Prvním důvodem je snaha identifikovat a hodnotit případnou interakci vlivů; druhým důvodem je hledání možností, jak řešit problém obtížné či nemožné separace vzájemně závislých vlivů, a třetím důvodem je obecný požadavek co největší přesnosti odhadů všech relevantních charakteristik zkoumané závislosti.
- ✓ V regresních úlohách se potřeba kontroly řeší přidáním sporných vysvětlujících proměnných k nesporným (přímo vyplývají ze zadání úlohy).
- ✓ Pozornost je pak soustředěna na modifikaci hodnot odhadnutých regresních koeficientů po zařazení nových proměnných a na změny, ke kterým došlo.
- ✓ Modifikace pomocí kovarianční analýzy se při použití regresního přístupu zabezpečuje současným zařazením jak studovaných faktorů (dominálních proměnných) ve formě umělých nula-jedničkových veličin, tak i kontrolovaných doprovodných proměnných.
- ✓ Při tomto postupu se předpokládá, že z hlediska jejich simultánního působení na vysvětlovanou proměnnou neexistuje interakce mezi nominálními a doprovodnými proměnnými.

Příklad:

- ✓ Vysvětlovaná proměnná Y – systolický krevní tlak
- ✓ Vysvětlující proměnná – věk náhodně vybraných mužů a žen
- ✓ Předpokládá se, že dobrým modelem závislosti krevního tlaku na věku je přímka. Nejprve uvažujme dvě otázky:
 - Vyjadřuje závislost krevního tlaku na věku pro muže a ženy stejná regresní rovnice přímky?
 - Je průměrný krevní tlak mužů a žen stejný, vezmeme-li v úvahu (neboli po modifikaci, resp. kontrolujeme-li) možné zavádějící důsledky rozdílných věkových rozdělení mužů a žen?
- ✓ Pro odpovědi na tyto otázky nemůžeme použít stejné statistické nástroje.
- ✓ Odpověď na první otázku vyžaduje porovnat dvě regresní přímky, zatímco druhá musí zhodnotit rozdíly mezi průměry ve skupinách.
- ✓ První otázku lze řešit pomocí regresního modelu $\beta_0 + \beta_1 X + \beta_2 A + \beta_3 XA + \varepsilon$, kde X je věk a A je pohlaví ($a_1 = 0$ pro muže, $a_2 = 1$ pro ženy).
- ✓ Podle provedených testů o parametrech regresní přímky výsledků výpočtů je možné učinit některý z následujících závěrů:
 - Přímky jsou shodné (koincidentní), neboli $\beta_2 = \beta_3 = 0$.
 - Přímky jsou rovnoběžné (paralelní), neboli $\beta_2 \neq 0$, ale $\beta_3 = 0$.
 - Přímky nejsou rovnoběžné ani shodné, neboli $\beta_2 \neq 0$, $\beta_3 \neq 0$.
- ✓ Tyto závěry úzce souvisí i s odpovědí na druhou otázku.
- ✓ Jsou-li shodné přímky, pak se ani neliší průměrný krevní tlak mužů a žen.
- ✓ Jsou-li přímky rovnoběžné, pak přímka s vyšší hodnotou absolutního členu má (při stejné směrnici přímky) i vyšší průměr.
- ✓ Nejsou-li přímky rovnoběžné, je třeba se jimi důkladněji zabývat. Mají-li průsečík mimo zajímavou oblast věku, nic se nemění proti předchozímu případu.
- ✓ Mají-li průsečík v zajímavé oblasti věku, pak lze říci, že existuje interakce mezi věkem a pohlavím, takže do určitého věku má jedna skupina nižší průměrný tlak a od tohoto věku má tato skupina vyšší průměrný tlak.
- ✓ Pochopitelně pro vyšší kvalitu těchto úsudků bychom provedli příslušné výpočty a testy o shodě dvou přímek, resp. o shodě dvou průměrů na základě údajů pocházejících ze dvou nezávislých výběrů.

Typy proměnných v analýze kovariance:

- ✓ Analýzu kovariance lze považovat za rozšíření metod analýzy rozptylu a regresní analýzy. Jde o zkoumání závislosti v poměrně složitém souboru proměnných.
- ✓ Uplatňují se v něm:
 - Jedna nebo několik vysvětlujících proměnných – faktorů A_1, A_2, \dots, A_s , přičemž stejně jako v analýze rozptylu jde o obvykle o nominální nebo alternativní proměnné, ale mohou to být i jiné kategoriální proměnné.
 - Jedna nebo více vysvětlovaných proměnných Y_1, Y_2, \dots, Y_p , na něž je při analýze soustředěna pozornost v tom smyslu, že chceme prokázat jejich závislost na faktoru či faktorech.
- ✓ Jedna nebo více doprovodných proměnných (kontrolovaných proměnných) X_1, X_2, \dots, X_q , které zahrnujeme do modelu a počítáme s nimi zejména proto, abychom závislost vysvětlovaných proměnných na faktorech očistili od jejich vlivu.

**Předpoklady analýzy kovariance:**

- ✓ Obvyklé algoritmy v analýze kovariance lze uplatnit při splnění řady podmínek, z nichž některé jsou stejné jako v analýze rozptylu:
 - Náhodnost výběru
 - Nezávislost výběrů (skupin), do nichž se výběrový soubor rozpadá. Obecně se nezávislé výběry většinou týkají různých skupin (účelově definovaných částí) sledované populace, ale též to mohou být výběry z různých porovnávaných (nezávislých) populací.
 - Normální rozdělení Y , popř. vícerozměrné normální rozdělení y , ve všech populacích (skupinách populace).
 - Homoskedasticita, tedy stejné rozptyly, popř. kovarianční matice, ve všech populacích (skupinách populace).
 - Lineární závislost Y na X , popř. Y_1, Y_2, \dots, Y_p na X_1, X_2, \dots, X_q , ve všech populacích (skupinách populace).
 - Shoda regresních koeficientů, neboli rovnoběžnost regresních přímek, popř. rovin nebo nadrovin, ve všech populacích (skupinách populace).
- ✓ Jako další podmínky se někdy uvádí nenáhodný charakter doprovodné proměnné X , popř. doprovodných veličin X_1, X_2, \dots, X_q , a nepřítomnost interakce mezi doprovodnou proměnnou X a faktorem A , popř. mezi q doprovodnými proměnnými a několika faktory. Tyto požadavky lze však těžko striktně dodržet.

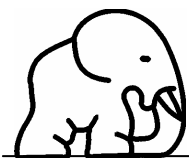
Modelový příklad – komparace účinku dvou intervenčních postupů:

- ✓ Ptáme se, zda se liší efekt terapie zachycený hodnotu testu úzkosti (Y) u dvou náhodně sestavených skupin jedinců, které jsou léčeny dvěma odlišnými postupy.
- ✓ Proměnná Y se měří součtem skóre z vhodného psychologického dotazníku. Pro lepší kontrolu výsledů experimentu se zaznamenávaly také počáteční úzkosti (X_1) před experimentem a obecná vegetativní labilita (X_2).
- ✓ Předpokládáme, že kovarianty (rušivé nezávisle proměnné) mají v obou skupinách stejný vliv na závisle proměnnou.
- ✓ Proměnná Z – indikátorová proměnná – měření patří osobě z experimentální nebo kontrolní skupiny.

Osoby	Y	X_1	X_2	$X_3 = Z$	Osoby	Y	X_1	X_2	$X_3 = Z$
1	6	7	5	0	11	2	4	1	1
2	4	4	1	0	12	5	6	2	1
3	4	5	1	0	13	3	6	2	1
4	7	8	5	0	14	1	3	1	1
5	5	3	1	0	15	6	9	4	1
6	4	3	4	0	16	4	5	2	1
7	7	6	4	0	17	5	8	5	1
8	5	6	1	0	18	3	4	3	1
9	3	5	2	0	19	2	7	3	1
10	3	4	1	0	20	3	8	2	1

	1. skupina		2. skupina		Dohromady	
	Průměr m	Odchylka s	Průměr m	Odchylka s	Průměr m	Odchylka s
Y	4,80	1,47573	3,40	1,57762	4,10	1,65116
X_1	5,10	1,66333	6,00	2,00000	5,55	1,84890
X_2	2,50	1,77951	2,50	1,26930	2,50	1,50438

- ✓ Zkoumáme-li velikost rozdílů průměrů odrážející odlišnost působení obou postupů, pak t-testem zjistíme, že není důvodu přiklonit se k alternativní hypotéze: terapie působí rozdílně. Ekvivalentní výsledek indikuje jednoduchá analýza rozptylu.



T-testy

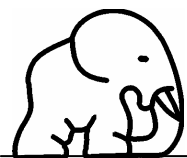
Proměnná	Metoda	Rozptyl	DF	t hodnota	Pr > t
body	Pooled	Equal	18	2.05	0.0553
body	Satterthwaite	Unequal	17.9	2.05	0.0554

Rovnost variancí

Proměnná	Metoda	DF čit	DF jmen	F hodnota	Pr > F
body	Folded F	9	9	1.14	0.8456

- ✓ Jestliže však vezmeme v úvahu okolnost, že na začátku experimentu měla první skupina menší průměrnou úzkostnost a zároveň že obě proměnné mohou uvnitř skupin navzájem korelovat, pak bychom při rovnosti účinku spíše očekávali, že první skupina bude mít svůj průměr po experimentu také menší než druhá skupina. Naopak rozdíl v průměrech proměnné Y by byl pravděpodobně větší, kdyby ve skupinách byly průměry proměnné X_1 stejné. Dosavadním postupem jsme ale nerespektovali informaci obsaženou v X_1 . Na základě vztahu mezi Y a X_1 by se pravděpodobně část rozdílnosti mezi skupinami pro proměnnou Y dala předpovědět pomocí X_1 a tak eliminovat z pozorovaných hodnot. Pro zbytkové hodnoty by pak analýza rozptylu byla relevantnější. Totéž platí i pro proměnnou X_2 .
- ✓ Uvedený problém analýzy kovariance se dá také zpracovat pomocí regresní analýzy. K tomu je zapotřebí vytvořit jednu kódovací proměnnou (X_3), která popisuje zařazení jedinců do obou skupin. Její hodnoty jsou doplněny do matice měření X.
- ✓ Zkoumáme nyní ovlivnění Y proměnnou X_3 . Chceme zodpovědět otázku, zda zavedení proměnné X_3 do regresní rovnice, jež zachycuje vztah mezi Y a X_1 , X_2 , povede ke statisticky významnému zlepšení predikce Y.
- ✓ Použijeme tedy kritérium F pro hodnocení významného zlepšení mnohonásobného korelačního koeficientu

$$F = \frac{(n-k-1)(r_{y,x_1x_2x_3}^2 - r_{y,x_1x_2}^2)}{(k-2)(1-r_{y,x_1x_2x_3}^2)} = \frac{(20-4)(0,6368-0,3771)}{1-0,6368} = 11,4386$$
- ✓ Toto F srovnáme s kritickou hodnotou F-rozdělení o (1; 16) stupních volnosti, která má na 1% hladině významnosti hodnotu 8,53. Prokázali jsme, že při uvážení vlivu doprovodných proměnných X_1 a X_2 je účinek obou terapií odlišný. Rovnice pro odhad cílové proměnné má tvar $y = 1,99 + 0,36 x_1 + 0,39 x_2 - 1,73 x_3$.
- ✓ Ovlivnění cílové proměnné proměnnými X_1 a X_2 se modeluje v použitém regresním modelu stejně v obou skupinách.
- ✓ Provedení regrese uvnitř obou skupin však může prokázat, že ve skutečnosti tomu tak není – působení proměnných X_1 a X_2 je při uvážení rozdílnosti terapií jiné.
- ✓ Tuto okolnost zkoumáme tak, že do regrese na proměnných X_1 , X_2 a X_3 přidáme proměnné $X_4 = X_2X_1$ a $X_5 = X_3X_2$, které odpovídají interakci doprovodných proměnných s intervencemi v obou skupinách. Příspěvek nových proměnných k regresi testujeme opět pomocí F kritéria. Jestliže testovací statistika F není významná, nemůžeme zamítnout hypotézu homogenity regresní uvnitř skupin.
- ✓ **Poznámka:** koeficient determinace $R^2 = 0,638$ není nestranným odhadem teoretické hodnoty – má systematicky větší hodnotu, protože nezohledňuje počet proměnných a počet měřených objektů. Vhodnější je tedy použití korigované hodnoty adjusted R^2 .



Multiple Regression Analysis

Dependent variable: Y3

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	1,81598	0,998736	1,81828	0,0867
X31	0,176441	0,212892	0,828783	0,4187
X32	0,521909	0,261646	1,99471	0,0624

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	19,5362	2	9,76809	5,15	0,0179
Residual	32,2638	17	1,89787		
Total (Corr.)	51,8	19			

R-squared = 37,7146 percent

R-squared (adjusted for d.f.) = 30,3869 percent

Standard Error of Est. = 1,37763

Multiple Regression Analysis

Dependent variable: Y3

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	1,98921	0,787797	2,52503	0,0225
X31	0,361892	0,176316	2,05252	0,0569
X32	0,386056	0,209829	1,83986	0,0844
skupina	-1,7257	0,510247	-3,38209	0,0038

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	32,9863	3	10,9954	9,35	0,0008
Residual	18,8137	16	1,17586		
Total (Corr.)	51,8	19			

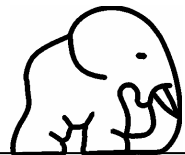
R-squared = 63,6801 percent

R-squared (adjusted for d.f.) = 56,8701 percent

Standard Error of Est. = 1,08437

ANALÝZA KATEGORIÁLNÍCH DAT

- ✓ Kategoriální data – jedná se především o znaky kvalitativní, např. zaměstnání, pohlaví, typ automobilu, vkus zákazníka.
- ✓ Získaná data zachycujeme pomocí jedno-, dvou- nebo vícerozměrných tabulek četností nebo relativních četností. Každý rozměr (dimenze) tabulky odpovídá klasifikaci do kategorií podle určité proměnné.
- ✓ Některé proměnné mají podle úlohy charakter závisle proměnné (cílové proměnné), jiné považujeme za nezávislé.
- ✓ Proměnné jsou často nominálního, resp. kvalitativního typu. Také však mohou mít nějaké přirozené řazení (např. vedlejší reakce na lék mohou být žádné, mírné nebo silné) – jsou ordinálního typu.
- ✓ Četnostní tabulky vznikají i zařazením jinak spojitých metrických údajů do kategorií, který byly navrženy jako intervaly pokrývající rozsah hodnot sledované proměnné.
- ✓ Při zkoumání četností dat stojíme před podobnými úkoly jako v případě dat metrických.



- ✓ Porovnáváme náhodné chování proměnné s pravděpodobnostním rozdělením, jež je předem přesně specifikované, nebo srovnáváme rozdělení sledované proměnné ve dvou nebo více populacích, aniž bychom předem specifikovali tvar jejich rozdělení.
- ✓ Také nás zajímá síla asociace jednotlivých proměnných mezi sebou.

Porovnání relativní četnosti s teoretickou hodnotou:

- ✓ Posuzujeme relativní četnost přítomnosti určité vlastnosti v ZS pomocí náhodného výběru o rozsahu n .
- ✓ Předpokládejme hodnotu relativní četnosti výskytu sledované vlastnosti p_0 .
- ✓ Testujeme nulovou hypotézu $H_0: p = p_0$ proti alternativní hypotéze $H_1: p \neq p_0$.

- ✓ Testové kritérium má tvar
$$u = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$
.

- ✓ Kritický obor pro zamítnutí H_0 je vymezen následovně:

Alternativa	Kritický obor
$H_1: p \neq p_0$	$K = \{ u > u_\alpha \}$
$H_1: p > p_0$	$K = \{ u > u_{2\alpha} \}$
$H_1: p < p_0$	$K = \{ u < -u_{2\alpha} \}$

- ✓ Je možné v rámci hodnocení stanovit také intervalový odhad relativní četnosti, kdy dvoustranný interval spolehlivosti pro spolehlivost $1 - \alpha$ má tvar:
$$P\left(f_i - u_\alpha \sqrt{\frac{f_i(1-f_i)}{n}} < p < f_i + u_\alpha \sqrt{\frac{f_i(1-f_i)}{n}} \right) = 1 - \alpha,$$
- ✓ Uvedené vztahy lze ale použít za předpokladu normální aproximace rozdělení relativní četnosti a jsou vhodné pouze pro větší rozsahy výběru.

Porovnání dvou relativních četností:

- ✓ Zajímá nás porovnání dvou pravděpodobností p_1 a p_2 výskytu nějaké vlastnosti ve dvou ZS.
- ✓ Na základě náhodných výběrů o velkých rozsazích n_1 a n_2 ($n_1 > 100$; $n_2 > 100$) je třeba ověřit hypotézu $H_0: p_1 = p_2$.

- ✓ Test je založen na statistice
$$u = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\bar{p} \cdot (1 - \bar{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$
.

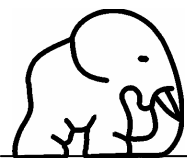
- ✓ Pokud $|u| > u_\alpha \Rightarrow H_0$ zamítáme.
- ✓ Cílem analýzy může také být testovat a odhadovat velikost jejich rozdílu $\Delta = p_1 - p_2$.
- ✓ Testová statistika se opírá o standardizovanou odchylku rozdílu empirických četností p_1 a p_2 od předpokládané hodnoty Δ .
- ✓ Počet prvků se sledovanou vlastností ve výběrových souborech o rozsahu n_1 a n_2 je m_1 a m_2 .
- ✓ Teoretické hodnoty p_i potom odhadujeme pomocí relativních četností $f_i = m_i/n_i$.
- ✓ Nulovou a alternativní hypotézu lze zapsat jako:
 - $H_0: (p_1 - p_2) = \Delta$, příp. $= 0$
 - $H_1: (p_1 - p_2) \neq \Delta$, příp. $\neq 0$

- ✓ Testové kritérium má tvar
$$u = \frac{(p_1 - p_2) - \Delta}{s_{(p_1 - p_2)}}$$
.

- ✓ Výpočet odhadu směrodatné odchylky $s(p_1 - p_2)$ závisí na hodnotě Δ . Jestliže $\Delta \neq 0$, pak

$$s_{(p_1 - p_2)} = \sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}.$$

- ✓ Nulová hypotéza se zamítá, pokud $|u| > u_\alpha \Rightarrow H_0$.



- ✓ V případě, že $\Delta = 0$, má $s_{(p_1 - p_2)}$ hodnotu $s_{(p_1 - p_2)} = \sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$, kde $p = \frac{m_1 + m_2}{n_1 + n_2}$ je spojený odhad teoretické relativní četnosti a $q = 1 - p$.
- ✓ Rozsahy obou výběrů musí být dostatečně veliké, abychom mohli pro výběrové rozdělení rozdílů hodnot $p_1 - p_2$ uplatnit centrální limitní teorém.
- ✓ Dvoustranný interval spolehlivosti má tvar $(p_1 - p_2) \in (f_1 - f_2) \pm u_\alpha \cdot s_{(p_1 - p_2)}$
- ✓ Jestliže podmínka o rozsazích výběru není splněna, ale počty jsou větší než 20, uplatňuje se arcussinová transformace na druhou mocninu odhadů pravděpodobností $\varphi(p) = \arcsin \sqrt{p}$.
- ✓ Hypotézu o rovnosti pravděpodobností pak testujeme pomocí statistiky $z = \frac{\varphi(p_1) - \varphi(p_2)}{28,648 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

Příklad:

U 500 náhodně vybraných domácností bylo prováděno v roce 1997 zjišťování, zda mají ve svém jídelníčku zařazenu cereální výživu. Kladně odpovědělo 67 domácností. U stejného počtu domácností bylo provedeno zjišťování v roce 1998. V tomto roce kladně odpovědělo 202 domácností. Vypočtete 95 % interval spolehlivosti pro změnu podílu domácností.

$$n_1 = 500 \quad m_1 = 67 \quad f_1 = 67/500 = 0,134$$

$$n_2 = 500 \quad m_2 = 202 \quad f_2 = 202/500 = 0,404$$

$$s_{(p_1 - p_2)} = \sqrt{\frac{0,134 \cdot 0,866}{500} + \frac{0,404 \cdot 0,596}{500}} = 0,0267$$

$$(p_1 - p_2) \in (0,134 - 0,404) \pm 1,96 \cdot 0,0267$$

$$(p_1 - p_2) = (-0,3224; -0,21764)$$

- ✓ Protože daný interval nepokrývá 0, můžeme na hladině významnosti 0,05 zamítnout nulovou hypotézu, že v obou skupinách domácností mají zařazeny v jídelníčku cereální potraviny.
- ✓ Chceme testovat hypotézu, že podíl domácností v roce 1998 není větší o více než 30 % ve srovnání s podílem domácností v roce 1997. Použijeme jednostranný test na 5% hladině významnosti (kritická hodnota je 1,6448)

$$u = \frac{(0,134 - 0,404) - 0,3}{0,0267} = -21,334$$

- ✓ Výsledek svědčí ve prospěch alternativní hypotézy.

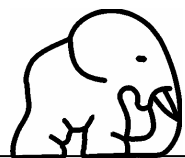
 χ^2 – test dobré shody:

- ✓ Přezkoušujeme, zda tvar pravděpodobnostního rozdělení kategoriální proměnné X má specifickou podobu. Při pozorování proměnné X se zjistily četnosti n_j jednotlivých kategorií. Předpokládáme, že pravděpodobnostní rozdělení proměnné je určené pravděpodobnostmi p_j .
- ✓ Testem dobré shody testujeme hypotézu $H_0: F(x) = F_0(x)$ proti alternativě $H_1: F(x) \neq F_0(x)$.
- ✓ Předpokládáme, že $F_0(x)$ je pevně daná hypotetická distribuční funkce, v níž nefigurují žádné neznámé parametry. Nulová hypotéza udává pouze typ rozdělení, nikoli jeho parametry.
- ✓ Rozdíl mezi pozorovanými a očekávanými četnostmi zachycuje testovací statistika, která má tvar

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - np_j)^2}{np_j}, \text{ kde } k \text{ je počet možných hodnot kategoriální proměnné, } n_j \text{ jsou empirické (skutečné)}$$

četnosti v intervalu j , np_j jsou teoretické (očekávané) četnosti v intervalu j vypočítané za předpokladu platnosti H_0 , přičemž n označuje rozsah výběru a p_j teoretickou pravděpodobnost kategorie j .

- ✓ Za platnosti H_0 má statistika asymptoticky χ^2 - rozdělení o $k-1$ stupních volnosti.
- ✓ Jestliže hodnota statistiky χ^2 překročí kritickou mez, signalizuje to špatnou shodu dat s teoretickým rozdělením.

**Příklad:**

- ✓ V n nezávislých náhodných pokusech očekáváme, že četnosti náhodných jevů A_1, A_2, A_3 , které v pokusu vůbec mohou nastat, jsou v poměru 1 : 2 : 1. V 80 pokusech jsme získali jejich četnosti 14, 50 a 16. Máme naši hypotézu zamítnout? Pro vypočtení testovací statistiky vytvoříme následující tabulku.

n_j	np_j	$n_j - np_j$	$(n_j - np_j)^2$	$(n_j - np_j)^2 / np_j$
14	20	-6	36	1,8
50	40	10	100	2,5
16	20	-4	16	0,8
80	80	$\chi^2 = 5,10$		

- ✓ χ^2_{α} pro 2 stupně volnosti má kritickou hodnotu 5,991. Protože $5,1 < 5,991$, nemůžeme nulovou hypotézu zamítnout.

Závislost kategoriálních proměnných:

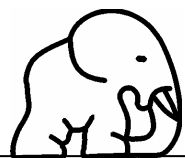
- ✓ Zabývá se statistickou analýzou četnostních tabulek, které vznikají, když popisujeme a analyzujeme vztah kategoriálních proměnných.
- ✓ Jedná se o analogii korelační analýzy spojitých proměnných nebo o podobnost s analýzou rozptylu.
- ✓ Rozdíl mezi oběma metodami spočívá v tom, že v případě analýzy četnostních tabulek obě kategoriální proměnné považujeme za náhodné, zatímco v analýze rozptylu posuzujeme vliv faktoru (kategoriální proměnné) s určitým počtem hladin jako nezávisle proměnné na chování náhodné závisle proměnné, jež má kvantitativní charakter.

Příklad:

- ✓ V roce 1912 se na své první plavbě srazil luxusní zámořský parník Titanic s plovoucí ledovou krou a potopil se. Někteří cestující se dostali na záchrané čluny, ostatní zemřeli. Představme si, že zkáza Titaniku je experimentem, jak se lidé chovají tváří v tvář smrti, když jenom někteří mohou uniknout. Předpokládáme, že pasažéři jsou nestranným vzorkem z populace stratifikované podle majetkových poměrů. V následující tabulce uvádíme data zvláště pro muže a ženy (Lord, 1998 – nejsou zachyceni cestující, u nichž není znám jejich sociální status). Při popisné analýze takovýchto dat se doporučuje uvést údaje v tabulkách jako procenta z řádkových nebo sloupcových součtů. Tím se lépe prezentují rozdílnosti rozdělení v jednotlivých kategoriích. Procenta nebo absolutní četnosti také zobrazujeme pomocí sloupcových grafů.
- ✓ Pro jednoduchou inferenční analýzu lze použít metody pro srovnání procent. Snadno lze spočítat, že celkově zemřelo 680 mužů a 168 se jich zachránilo. Žen zemřelo 126, uniknout smrti se podařilo 317. Existuje evidence, že muži v této situaci více umírají? Jaké jsou pro to důvody? Můžeme se však také zeptat, zda existují statisticky významné rozdíly v procentuálních podílech zemřelých žen mezi jednotlivými třídami. Nechceme však srovnávat páry tříd, ale vyhodnotit globální hypotézu, zda vůbec existuje nějaký rozdíl. Stejně hodnocení můžeme provést pro muže. Zajímáme se, zda existuje stochastický vztah mezi proměnnou třída cestujícího a proměnnou, která popisuje status přežití cestujícího (ANO, NE). Jinak řečeno, ptáme se, zda ovlivňuje proměnná třída cestujícího pravděpodobnost přežití cestujícího.
- ✓ Poznámka: Tento příklad pracuje dohromady se třemi proměnnými (pohlaví, třída cestujícího a status přežití).
- ✓ Data o cestujících při ztroskotání Titaniku.

Status	Muži		Ženy	
	zemřeli	přežili	zemřely	přežily
I. třída	111	61	6	126
II. třída	150	22	13	90
III. třída	419	85	107	101

Status	Muži			Ženy		
	zemřeli	přežili	počet celkem	zemřely	přežily	počet celkem
I. třída	64,5 %	35,5 %	172	4,4 %	95,6 %	135
II. třída	84,7 %	15,3 %	177	12,6 %	87,4 %	103
III. třída	83,1 %	16,9 %	504	51,4 %	48,6 %	208

**Kontingence:**

- ✓ Kontingence se zabývá zkoumáním vztahu mezi množnými znaky, které mají větší počet obměn. V tomto případě hodnotíme tabulky dvoudimenzionální, což jsou tabulky vzniklé tříděním podle dvou proměnných – jde o tzv. **kontingenční tabulky**.
- ✓ Předpokládáme přitom, že každá jednotka může být klasifikována podle dvou proměnných (kritérií) A a B. proměnná A má r kategorií (úrovní) a proměnná B má s kategorií (úrovní). Označme n_{ij} počet prvků z výběru o rozsahu n , které podle proměnné A patří do kategorie A_i a podle proměnné B do kategorie B_j . Dále označme $n_{i.}$ počet prvků z výběru, které patří do kategorie A_i (bez ohledu na hodnotu proměnné B), a podobně $n_{.j}$ počet prvků patřících do kategorie B_j .

- ✓ Platí tedy vztahy $\sum_{i=1}^r n_{ij} = n_{.j}$, $\sum_{j=1}^s n_{ij} = n_{i.}$, $\sum_{j=1}^s n_{ij} = n_{i.}$, $\sum_{i=1}^r n_{ij} = n_{.j}$.

- ✓ Kontingenční tabulka typu $r \times s$ pak vypadá následovně:

Znak B Znak A	b_1	b_2	b_j	b_s	celkem
a_1	n_{11}	n_{12}	n_{1j}	n_{1s}	$n_{1.}$
a_2	n_{21}	n_{22}	n_{2j}	n_{2s}	$n_{2.}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
a_i	n_{i1}	n_{i2}	n_{ij}	n_{is}	$n_{i.}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
a_r	n_{r1}	n_{r2}	n_{rj}	n_{rs}	$n_{r.}$
celkem	$n_{.1}$	$n_{.2}$	$n_{.j}$	$n_{.s}$	n

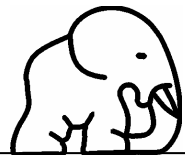
- ✓ Po vytvoření tabulky začínáme zkoumat vzájemný vztah obou proměnných A a B – nejdříve pomocí vhodného zobrazení, později lze testovat různé hypotézy.
- ✓ Hypotézy pro kontingenční tabulky se obvykle definují v pojmech stochastické nezávislosti, a to pomocí určitých podmínek.
- ✓ V kontextu stochastické nezávislosti proměnných A a B tyto podmínky indukují, že čísla $n_{ij}/n_{i.}$, resp. $n_{ij}/n_{.j}$ (řádkové, resp. sloupcové relativní četnosti) jsou pro všechna čísla i , resp. j až na náhodné odchylky konstantní.
- ✓ Jestliže jednu z proměnných kontrolujeme během výběru – třeba proměnnou A, nazýváme ji faktor. Tato proměnná vlastně určuje r disjunktních subpopulací W_1, W_2, \dots, W_r z populace W . V tomto případě se může hypotéza nezávislosti popsat jako hypotéza homogenity chování proměnné B vzhledem k faktoru A.

Hypotéza homogenity:

- ✓ Tato hypotéza předpokládá, že pravděpodobnostní rozdělení kategoriální proměnné B je stejné v různých populacích, které jsou identifikovány faktorem A.
- ✓ Příslušné statistické testy nazýváme někdy testy dobré shody, kdy nám jde o shodu rozdělení kategoriální proměnné.
- ✓ Úrovně faktoru A stratifikují v tomto případě celou populaci W do r disjunktních subpopulací W_1, W_2, \dots, W_r a každý prvek z W_i je klasifikován do jedné z kategorií proměnné B.
- ✓ Nechť P_{ij} je relativní četnost prvků subpopulace W_i , jež jsou v j -té kategorii proměnné B.
- ✓ Potom se hypotéza homogenity může vyjádřit jako $P_{1j} = P_{2j} = \dots = P_{rj}$ pro všechna $j = 1, 2, \dots, s$, což znamená, že pro každou kategorii má být relativní četnost prvků v dané subpopulaci stejná pro všechny subpopulace.
- ✓ Hypotézu homogenity můžeme provádět tehdy, jestliže máme k dispozici prostý náhodný výběr z každé subpopulace určené faktorem A nebo jsme provedli přiřazení objektů do jednotlivých skupin namátkově.

Příklad:

- ✓ Populace W studentů je stratifikována podle pohlaví a proměnná B je určena tím, zda má student zájem o účast ve školním sportovním oddíle. Je zřejmé, že proměnná B je kategoriální. Dotazování se provádí tak, že zvlášť se provede náhodný výběr 66 chlapců a 74 dívek.



- ✓ Z chlapců, resp. dívek mělo zájem 30, resp. 11 jedinců. Zařazením osob podle zájmu dostaneme tabulku typu 2x2.

	Zájem o sport		Celkem
	ano	ne	
Chlapci	30	36	66
Dívky	11	63	74
Celkem	41	99	140

- ✓ Jestliže P_{11} je relativní část chlapců se zájmem o sport a P_{21} je relativní část dívek se zájmem o sport, pak hypotéza homogenity má tvar $P_{11} = P_{21}$ (z toho plyne také $P_{12} = P_{22}$). V pojmech nezávislosti H_0 vyjadřuje, že relativní četnost jedinců zajímajících se o účast ve sportovním oddíle je nezávislá na pohlaví.

Hypotéza nezávislosti:

- ✓ V hypotéze nezávislosti se považují obě proměnné A a B za náhodné proměnné, přičemž předpokládáme jejich úplnou nezávislost. To znamená, že hodnota proměnné A neovlivňuje podmíněné rozdělení proměnné B a naopak.
- ✓ Uvažujeme populaci W, přičemž každý prvek této populace je klasifikován podle dvou kategoriálních proměnných A a B. Zkoumáme, zda hodnoty proměnné A neovlivňují rozdělení proměnné B a naopak.
- ✓ **Nulová hypotéza** zní, že obě proměnné jsou na sobě stochasticky nezávislé.
- ✓ Tuto hypotézu lze vyjádřit podmínkami pro pravděpodobnosti p_{ij} , což jsou pravděpodobnosti, že na osobě zjistíme hodnotu proměnné A v kategorii i a hodnotu proměnné B v kategorii j.
- ✓ Necht' $p_{i\cdot}$, resp. $p_{\cdot j}$ je pravděpodobnost v populaci W, že proměnná A nabude hodnoty i, resp. proměnná B nabude hodnoty j. Pak hypotézu nezávislosti obou proměnných můžeme vyjádřit rovnicemi $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$,

$p_{i\cdot} = \sum_{j=1}^s p_{ij\cdot}$, $p_{\cdot j} = \sum_{i=1}^r p_{ij\cdot}$, které platí pro všechna $i = 1, 2, \dots, r$ a $j = 1, 2, \dots, s$. Uvedené vyjádření vyplývá ze vzorce pro výpočet pravděpodobnosti současného výskytu dvou nezávislých jevů.

- ✓ Poznámka: Má-li platit nezávislost, pak pro všechna i a j musí být splněna podmínka $n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$.