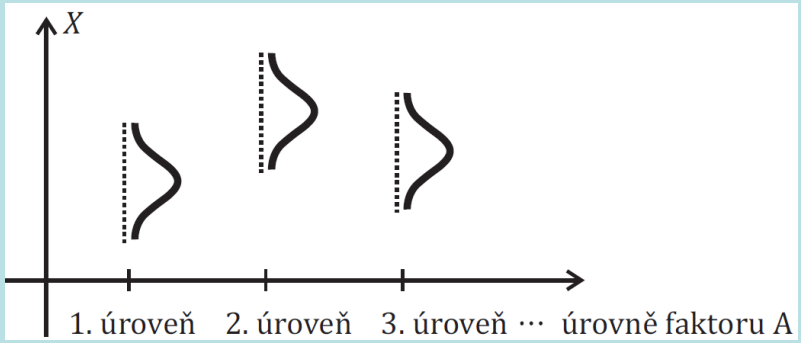
Analýza rozptylu jednoduchého třídění

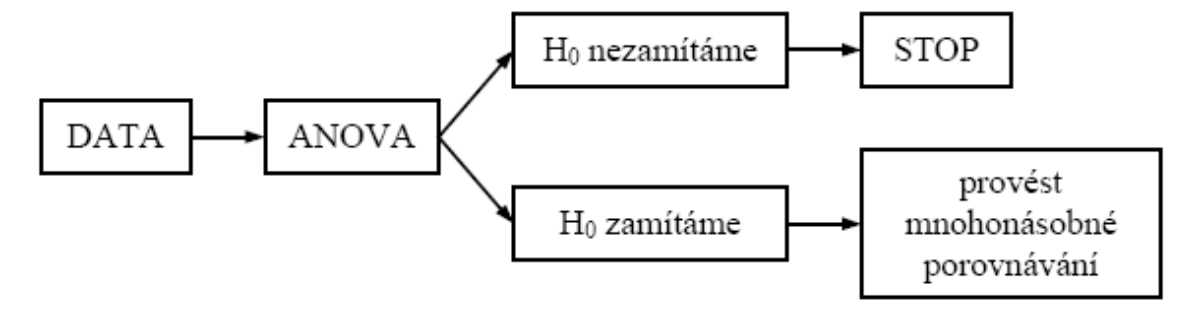
Motivace: Zajímáme se o problém, zda lze určitým faktorem (tj. nominální náhodnou veličinou A) vysvětlit variabilitu pozorovaných hodnot náhodné veličiny X, která je intervalového či poměrového typu. Např. zkoumáme, zda metoda výuky určitého předmětu (faktor A) ovlivňuje počet bodů dosažených studenty v závěrečném testu (náhodná veličina X).
Předpokládáme, že faktor A má $r \geq 3$ úrovní a přitom i-té úrovni odpovídá n_i pozorování $x_{i1}, \ldots, x_{in_i}$, které tvoří náhodný výběr z rozložení $N(\mu_i, \sigma^2)$, $i = 1, ..., r$ a jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy $X_{ij} = \mu_i + \varepsilon_{ij}$, kde ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, $i = 1, \ldots, r$, $j = 1, \ldots, n_i$.
Výsledky lze zapsat do tabulky

faktor A	výsledky
úroveň 1	$x_{11}, \ldots, x_{1n_1}$
úroveň 2	$x_{21}, \ldots, x_{2n_2}$
…	…
úroveň r	$x_{r1}, \ldots, x_{m_r}$

Ilustrace:



Na hladině významnosti α testujeme nulovou hypotézu, která tvrdí, že všechny střední hodnoty jsou stejné, tj.
H_0: $\mu_1 = \ldots = \mu_r$ proti alternativní hypotéze H_1, která tvrdí, že aspoň jedna dvojice středních hodnot se liší.
Jedná se tedy o zobecnění dvouvýběrového t-testu a na první pohled se zdá, že stačí utvořit $\binom{r}{2}$ dvojic náhodných výběrů a na každou dvojici aplikovat dvouvýběrový t-test. Hypotézu o shodě všech středních hodnot bychom pak zamítli, pokud aspoň v jednom případě z $\binom{r}{2}$ porovnávání se prokáže odlišnost středních hodnot. Odtud je vidět, že k neoprávněnému zamítnutí nulové hypotézy (tj. k chybě 1. druhu) může dojít s pravděpodobností větší než α. Proto ve 30. letech 20. století vytvořil R. A. Fisher metodu ANOVA (analýza rozptylu, v popsané situaci konkrétně analýza rozptylu jednoduchého třídění), která uvedenou podmínku splňuje.
Pokud na hladině významnosti α zamítneme nulovou hypotézu, zajímá nás, které dvojice středních hodnot se od sebe liší.
K řešení tohoto problému slouží metody mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.



Označení:

V analýze rozptylu jednoduchého třídění se používá tzv. tečková notace.

$n = \sum_{i=1}^{r} n_i$ … celkový rozsah všech r výběrů

$X_{i.} = \sum_{j=1}^{n_i} X_{ij}$ … součet hodnot v i-tém výběru

$M_{i.} = \frac{1}{n_i} X_{i.}$ … výběrový průměr v i-tém výběru

$X_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij}$ … součet hodnot všech výběrů

$M_{..} = \frac{1}{n} X_{..}$ … celkový průměr všech r výběrů

Zavedeme součty čtverců

$S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{..})^2$ … celkový součet čtverců (charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru),

počet stupňů volnosti $f_T = n - 1$,

$S_A = \sum_{i=1}^{r} n_i (M_{i.} - M_{..})^2$ … skupinový součet čtverců (charakterizuje variabilitu mezi jednotlivými náhodnými výběry),

počet stupňů volnosti $f_A = r - 1$.

Sčítanec $(M_{i.} - M_{..})$ představuje bodový odhad efektu α_i.

$S_E = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{i.})^2$ … reziduální součet čtverců (charakterizuje variabilitu uvnitř jednotlivých výběrů),

počet stupňů volnosti $f_E = n - r$.

Lze dokázat, že $S_T = S_A + S_E$.

(Důkaz je proveden např. ve skriptech Budíková, Mikoláš, Osecký: Popisná statistika v poznámce 5.20.)

Testování hypotézy o shodě středních hodnot

Náhodné veličiny X_{ij} se řídí modelem
M0: $X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
pro $i = 1, \ldots, r$, $j = 1, \ldots, n_i$, přičemž
ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$,
μ je společná část střední hodnoty závisle proměnné veličiny,
α_i je efekt faktoru A na úrovni i.
Parametry μ, α_i neznáme.

Požadujeme, aby platila tzv. reparametrizační rovnice: $\sum_{i=} \alpha_i = 0$.

(Pokud je třídění vyvážené, tj. pokud mají všechny výběry stejný rozsah: $n_1 = n_2 = \ldots = n_r$, pak lze použít zjednodušenou podmínku $\sum_{i=} \alpha_i = 0$.)

Kdyby nezáleželo na faktoru A, platila by hypotéza $\alpha_1 = \ldots = \alpha_r = 0$ a dostali bychom model
M1: $X_{ij} = \mu + \varepsilon_{ij}$.
Během analýzy rozptylu tedy zkoumáme, zda výběrové průměry $M_1, \ldots, M_r$ se od sebe liší pouze v mezích náhodného kolísání kolem celkového průměru M nebo zda se projevuje vliv faktoru A.

Rozdíl mezi modely M0 a M1 ověřujeme pomocí testové statistiky
$F_A = \frac{S_A / f_A}{S_E / f_E}$, která se řídí rozložením F(r-1,n-r), je-li model M1 správný. Hypotézu o nevýznamnosti faktoru A tedy zamítneme na hladině významnosti α, když platí: $F_A \geq F_{1-\alpha}(r-1,n-r)$.

Výsledky výpočtů zapisujeme do tabulky analýzy rozptylu jednoduchého třídění.

Zdroj variability	součet čtverců	stupně volnosti	podíl	F_A
skupiny	S_A	$f_A = r - 1$	S_A/f_A	$\frac{S_A/f_A}{S_E/f_E}$
reziduální	S_E	$f_E = n - r$	S_E/f_E	-
celkový	S_T	$f_T = n - 1$	-	-

Sílu závislosti náhodné veličiny X na faktoru A můžeme měřit pomocí poměru determinace: $P^2 = \frac{S_A}{S_T}$. Nabývá hodnot z intervalu $\langle 0,1 \rangle$.

Testování hypotézy o shodě rozptylů

Před provedením analýzy rozptylu je zapotřebí ověřit předpoklad o shodě rozptylů v daných r výběrech.

a) Levenův test: Položme $Z_{ij} = |X_{ij} - M_{i.}|$. Označíme

$$M_{Zi} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij},$$

$$M_{Z} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Z_{ij},$$

$$S_{ZE} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} (Z_{ij} - M_{Zi})^2,$$

$$S_{ZA} = \sum_{i=1}^{r} n_i (M_{Zi} - M_{Z})^2$$

Platí-li hypotéza o shodě rozptylů, pak statistika

$$F_{ZA} = \frac{S_{ZA}/(r-1)}{S_{ZE}/(n-r)} \approx F(r-1, n-r).$$

Hypotézu o shodě rozptylů tedy zamítáme na asymptotické hladině významnosti α, když $F_{ZA} \geq F_{1-\alpha}(r-1, n-r)$.
(Levenův test je vlastně založen na analýze rozptylu absolutních hodnot centrovaných pozorování. Vzhledem k tomu, že náhodné veličiny $X_{ij} - M_i$ nejsou stochasticky nezávislé a absolutní hodnoty těchto veličin nemají normální rozložení, je Levenův test pouze aproximativní.)

b) Brownův – Forsytheův test je modifikací Levenova testu. Modifikace spočívá v tom, že místo výběrového průměru i-tého výběru se při výpočtu veličiny z_{ij} používá medián i-tého výběru.

c) Bartlettův test: Platí-li hypotéza o shodě rozptylů a rozsahy všech výběrů jsou větší než 6, pak statistika $B = \frac{1}{C}\left[(n-r)\ln S_*^2 - \sum_{i=1}^{r} (n_i - 1)\ln S_i^2 \right]$ se asymptoticky řídí rozložením $\chi^2(r-1)$. Přitom konstanta $C = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r} \frac{1}{n_i - 1} - \frac{1}{n-r} \right)$ a S_*^2 je vážený průměr výběrových rozptylů.
H_0 zamítáme na asymptotické hladině významnosti α, když B se realizuje v kritickém oboru $W = \langle \chi^2_{1-\alpha}(r-1), \infty)$.

Post – hoc metody mnohonásobného porovnávání

Zamítneme-li na hladině významnosti α hypotézu o shodě středních hodnot, chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti α, tj. na hladině významnosti α testujeme H_0: $\mu_l = \mu_k$ proti H_1: $\mu_l \neq \mu_k$ pro všechna l, k = 1, .., r, $l \neq k$.

a) Mají-li všechny výběry týž rozsah p (říkáme, že třídění je vyvážené), použijeme Tukeyovu metodu.

Testová statistika má tvar $\frac{|M_{k.} - M_{l.}|}{\frac{S_*}{\sqrt{p}}}$. Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$\frac{|M_{k.} - M_{l.}|}{\frac{S_*}{\sqrt{p}}} \geq q_{1-\alpha}(r, n-r)$, kde hodnoty $q_{1-\alpha}(r, n-r)$ jsou kvantily studentizovaného rozpětí a najdeme je ve statistických tabulkách. (Studentizované rozpětí je náhodná veličina $Q = \frac{X_{(n)} - X_{(1)}}{S}$.)

Existuje modifikace Tukeyovy metody pro nestejné rozsahy výběrů, nazývá se Tukeyova HSD metoda. V tomto případě má testová statistika tvar $\frac{|M_{k.} - M_{l.}|}{S_* \sqrt{\frac{1}{2}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}}$. Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$\frac{|M_{k.} - M_{l.}|}{S_* \sqrt{\frac{1}{2}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}} \geq q_{1-\alpha}(r, n-r)$.

b) Nemají-li všechny výběry stejný rozsah, použijeme Scheffého metodu: rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$$|M_{k.} - M_{l.}| \geq S_* \sqrt{(r-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right) F_{1-\alpha}(r-1, n-r)}.$$

Výhodou Scheffého testu je, že k jeho provedení nepotřebujeme speciální statistické tabulky s hodnotami kvantilů studentizovaného rozpětí, ale stačí běžné statistické tabulky s kvantily Fisherova – Snedecorova rozložení.

V případě vyváženého třídění, kdy lze aplikovat Tukeyovu i Scheffého metodu, použijeme tu, která je citlivější. Tukeyova metoda tedy bude výhodnější, když
$q_{1-\alpha}^2(r, n-r) < 2(r-1)F_{1-\alpha}(r-1, n-r)$.

Metody mnohonásobného porovnávání mají obecně menší sílu než ANOVA.
Může nastat situace, kdy při zamítnutí H_0 nenajdeme metodami mnohonásobného porovnávání významný rozdíl u žádné dvojice středních hodnot. K tomu dochází zvláště tehdy, když p-hodnota pro ANOVU je jen o málo nižší než zvolená hladina významnosti. Pak slabší test patřící do skupiny metod mnohonásobného porovnávání nemusí odhalit žádný rozdíl.

Doporučený postup při provádění analýzy rozptylu:

a) Ověření normality daných r náhodných výběrů (grafické metody - NP plot, Q-Q plot, histogram, testy hypotéz o normálním rozložení - Lilieforsova varianta Kolmogorovova – Smirnovova testu nebo Shapirův – Wilkův test).
Doporučuje se kombinace obou způsobů. Závěry učiníme až na základě posouzení obou výsledků.
Obecně lze říci, že analýza rozptylu není příliš citlivá na porušení předpokladu normality, zvláště při větších rozsazích výběrů (nad 20), což je důsledek působení centrální limitní věty. Mírné porušení normality tedy není na závadu, při větším porušení použijeme např. Kruskalův – Wallisův test jako neparametrickou obdobu analýzy rozptylu jednoduchého třídění.

b) Po ověření normality se testuje homogenitu rozptylů, tj. předpoklad, že všechny náhodné výběry pocházejí z normálních rozložení s týmž rozpylem. Graficky ověřujeme shodu rozptylů pomocí krabicových diagramů, kdy sledujeme, zda je šířka krabic stejná. Numericky testujeme homogenitu rozptylů pomocí Levenova testu, Brownova – Forsytheova testu (oba jsou implementovány ve STATISTICE, Brownův – Forsytheův test v MINITABu) či Bartlettova testu (je k dispozici v MINITABu).
Slabé porušení homogenity rozptylů nevadí, při větším se doporučuje mediánový test.

c) Pokud jsou splněny předpoklady normality a homogenity rozptylů, můžeme přistoupit k testování shody středních hodnot. Předtím je samozřejmě vhodné vypočítat průměry a směrodatné odchylky či rozptyly v jednotlivých skupinách.

d) Dojde-li na zvolené hladině významnosti k zamítnutí hypotézy o shodě středních hodnot, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží post-hoc metody mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.

Příklad: U čtyř odrůd brambor (označených symboly A, B, C, D) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky (v kg):

odrůda	hmotnost
A	0,9 0,8 0,6 0,9
B	1,3 1,0 1,3
C	1,3 1,5 1,6 1,1 1,5
D	1,1 1,2 1,0

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

Řešení:

Data považujeme za realizace čtyř nezávislých náhodných výběrů ze čtyř normálních rozložení se stejným rozptylem. Testujeme hypotézu, že všechny čtyři střední hodnoty jsou stejné.

Vypočítáme výběrové průměry v jednotlivých výběrech: $M_{1.} = 0{,}8$, $M_{2.} = 1{,}2$, $M_{3.} = 1{,}4$, $M_{4.} = 1{,}1$,
celkový průměr: $M_{..} = 1{,}14$,
výběrové rozptyly: $S_1^2 = 0{,}02$, $S_2^2 = 0{,}03$, $S_3^2 = 0{,}04$, $S_4^2 = 0{,}01$,

vážený průměr výběrových rozptylů: $S_*^2 = \frac{\sum_{i=} \; - \; S_i^2}{n - \cdot} = \frac{3 \cdot 0{,}02 + 2 \cdot 0{,}03 + 4 \cdot 0{,}04 + 2 \cdot 0{,}01}{11} = \frac{3}{110} = 0{,}02\overline{7}$,

reziduální součet čtverců: $S_E = \left(n - \cdot \right) S_*^2 = 11 \cdot \frac{3}{110} = 0{,}3$,

skupinový součet čtverců: $S_A = \sum_{i=} \; \left(M_{i.} - M_{..} \right)^2 = 4 \cdot \left(0{,}8 - {,}14 \right)^2 + 3 \cdot \left(1{.}2 - {,}14 \right)^2 + 5 \cdot \left(1{,}4 - {,}14 \right)^2 + 3 \cdot \left(1{,}1 - {,}14 \right)^2 = 0{,}816$

celkový součet čtverců: $S_T = S_A + S_E = 0{,}816 + 0{,}3 = 1{,}116$,

testová statistika $F_A = \frac{S_A / f_A}{S_E / f_E} = \frac{0{,}816 / 3}{0{,}3 / 11} = 9{,}97$,

Kritický obor $W = \left\langle F_{0{,}95} \left({,}11 \right) , \infty \right) = \left\langle 3{,}59 , \infty \right)$. Protože testová statistika se realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,05.

Vypočteme poměr determinace: $P^2 = \frac{S_A}{S_T} = \frac{0{,}816}{1{,}116} = 0{,}7312$

Výsledky zapíšeme do tabulky ANOVA:

Zdroj variability	Součet čtverců	Stupně volnosti	podíl	F_A
skupiny	$S_A = 0{,}816$	3	$S_A/3 = 0{,}272$	$\frac{S_A/(\quad - \quad)}{S_E/(\quad - \quad)} = 9{,}97$
reziduální	$S_E = 0{,}3$	11	$S_E/11 = 0{,}02727$	-
celkový	$S_T = 1{,}116$	14	-	-

Nyní pomocí Scheffého metody zjistíme, které dvojice odrůd se liší na hladině významnosti 0,05.

Srovnávané odrůdy	Rozdíly $\lvert M_{k.} - M_{l.} \rvert$	Pravá strana vzorce
A, B	0,4	0,41
A, C	0,67	0,36
A, D	0,3	0,41
B, C	0,2	0,40
B, D	0,1	0,44
C, D	0,3	0,40

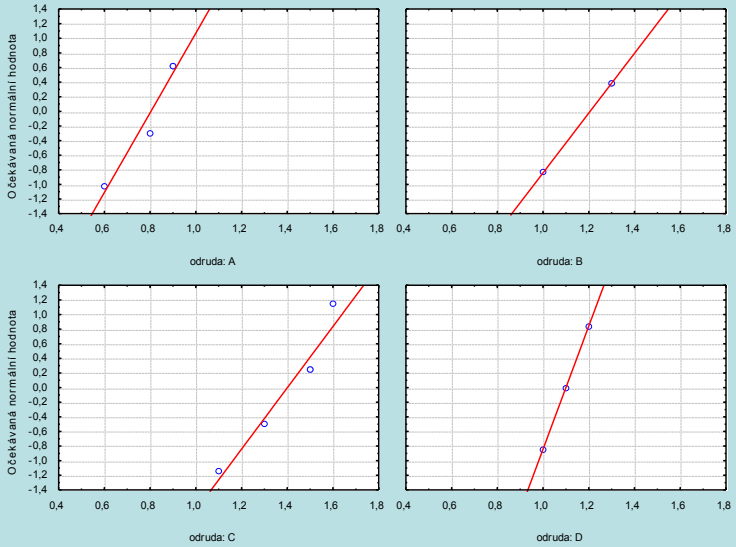
Na hladině významnosti 0,05 se liší odrůdy A a C.

Řešení pomocí systému STATISTICA

Otevřeme nový datový soubor o dvou proměnných X a odrůda a 15 případech. Do proměnné X zapíšeme zjištěné hmotnosti, do proměnné odrůda kódy pro dané odrůdy (1 pro A, 2 pro B, 3 pro C a 4 pro D).

	1 X	2 odruda
1	0,9	A
2	0,8	A
3	0,6	A
4	0,9	A
5	1,3	B
6	1	B
7	1,3	B
8	1,3	C
9	1,5	C
10	1,6	C
11	1,1	C
12	1,5	C
13	1,1	D
14	1,2	D
15	1	D

Ověříme normalitu daných čtyř náhodných výběrů pomocí N-P plotu:



Odchylky od normality jsou jen nepatrné.

Vypočteme výběrové průměry a výběrové rozptyly:
Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – X, Grupovací - odrůda – OK – Skupiny tabulek - zaškrtneme Rozptyly - Výpočet.

Rozkladová tabulka popisných statistik (priklad8301)
N=15 (V seznamu záv. prom. nejsou ChD)

odruda	X průměr	X N	X Sm.odch.	X Rozptyl
A	0,800000	4	0,141421	0,020000
B	1,200000	3	0,173205	0,030000
C	1,400000	5	0,200000	0,040000
D	1,100000	3	0,100000	0,010000
Vš.skup.	1,140000	15	0,282337	0,079714

Nyní ověříme předpoklad shody rozptylů.
Na záložce Skupiny tabulek zaškrtneme Levenův test – Výpočet.

Leveneův test homogenity rozpylů (priklad8301)
Označ. efekty jsou význ. na hlad. p < ,05000

Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
X	0,018667	3	0,006222	0,065333	11	0,005939	1,047619	0,410027

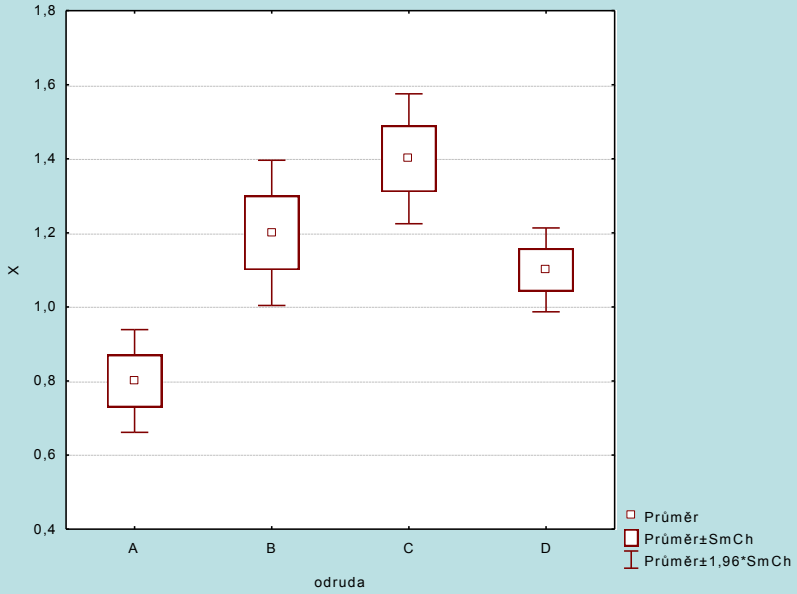
Vidíme, že p-hodnota Levenova testu je 0,41, tedy větší než hladina významnosti 0,05. Hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.

Přistoupíme k testu hypotézy o shodě středních hodnot.
Na záložce Skupiny tabulek zaškrtneme Analýza rozptylu – Výpočet.

	Analýza rozptylu (priklad8301) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
X	0,816000	3	0,272000	0,300000	11	0,027273	9,973333	0,001805

Jelikož p-hodnota = 0,001805 je menší než hladina významnosti 0,05, hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05.

Výpočet doplníme krabicovými diagramy:



Nyní aplikujeme Scheffého metodu mnohonásobného porovnávání, abychom zjistili, které dvojice odrůd se liší na hladině významnosti 0,05. Na záložce Post – hoc zvolíme Schefféův test.

	Scheffeho test; proměn.:X (priklad8301) Označ. rozdíly jsou významné na hlad. p < ,05000			
odruda	{1} M=,80000	{2} M=1,2000	{3} M=1,4000	{4} M=1,1000
A {1}		0,059165	0,001950	0,190463
B {2}	0,059165		0,464537	0,905502
C {3}	0,001950	0,464537		0,163499
D {4}	0,190463	0,905502	0,163499	

Tabulka obsahuje p-hodnoty pro vzájemné porovnání středních hodnot hmotnosti všech čtyř odrůd. Vidíme, že na hladině významnosti 0,05 se liší odrůdy A, C.

Význam předpokladů v analýze rozptylu

a) Nezávislost jednotlivých náhodných výběrů – velmi důležitý předpoklad, musí být splněn, jinak dostaneme nesmyslné výsledky.

b) Normalita – ANOVA není příliš citlivá na porušení normality, zvlášť pokud mají všechny výběry rozsah nad 20 (důsledek centrální limitní věty). Při výraznějším porušení normality se doporučuje Kruskalův – Wallisův test.

c) Shoda rozptylů – mírné porušení nevadí, při větším se doporučuje Kruskalův – Wallisův test. Test shody rozptylů má smysl provádět až po ověření předpokladu normality.

Neparametrické testy o mediánech

Motivace: Při aplikaci t-testů či analýzy rozptylu by měly být splněny určité předpoklady:

- normalita dat (pro výběry větších rozsahů ($n \geq 30$) nemá mírné porušení normality závažný dopad na výsledky)
- homogenita rozptylů
- intervalový či poměrový charakter dat

Pokud nejsou tyto předpoklady splněny, použijeme tzv. neparametrické testy, které nevyžadují předpoklad o konkrétním typu rozložení (např. normálním), stačí např. předpokládat, že distribuční funkce rozložení, z něhož náhodný výběr pochází, je spojitá.

Nevýhoda - ve srovnání s klasickými parametrickými testy jsou neparametrické testy slabší, tzn., že nepravdivou hypotézu zamítají s menší pravděpodobností než testy parametrické.

V této kapitole se omezíme na ty neparametrické testy, které se týkají mediánů.

Jednovýběrové testy (Jde o neparametrické obdoby jednovýběrového t-testu a párového t-testu.)

Znaménkový test a jeho asymptotická varianta

Nechť $X_1,\dots,X_n$ je náhodný výběr ze spojitého rozložení. Nechť $x_{0,50}$ je mediánem tohoto rozložení a c je reálná konstanta. Testujeme hypotézu $H_0 : x_{0,50} = c$ proti oboustranné alternativě $H_1 : x_{0,50} \neq c$ (resp. proti levostranné alternativě $H_1 : x_{0,50} < c$ resp. proti pravostranné alternativě $H_1 : x_{0,50} > c$).

Znaménkový test se nejčastěji používá jako párový test, kdy máme náhodný výběr ze spojitého dvourozměrného rozložení $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix},\dots,\begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ a testujeme hypotézu o rozdílu mediánů, tj. $H_0 : x_{0,50} - y_{0,50} = c$ proti $H_1 : x_{0,50} - y_{0,50} \neq c$ (resp. proti jednostranným alternativám). Přejdeme k rozdílům $Z_1 = X_1 - Y_1,\dots,Z_n = X_n - Y_n$ a testujeme hypotézu o mediánu těchto rozdílů, tj. $H_0 : z_{0,50} = c$.

a) Utvoříme rozdíly $D_i = X_i - c$ pro jednovýběrový test resp. $D_i = Z_i - c$ pro párový test, $i = 1,\dots,n$. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)

b) Zavedeme statistiku S_Z^+, která udává počet těch rozdílů D_i, které jsou kladné. S_Z^+ je součtem náhodných veličin s alternativním rozložením (i-tá veličina nabývá hodnoty 1, když i-tý rozdíl je kladný a hodnoty 0, když je záporný). Platí-li H_0, pak pravděpodobnost kladného i záporného rozdílu je stejná, tedy $S_Z^+ \sim Bi\left(n, \frac{1}{2}\right)$. Z vlastností binomického rozložení plyne, že $E\left(S_Z^+\right) = \frac{n}{2}$, $D\left(S_Z^+\right) = \frac{n}{4}$.

c) Stanovíme kritický obor.

Pro oboustrannou alternativu: $W = \langle 0, k_1 \rangle \cup \langle k_2, n \rangle$, pro levostrannou alternativu: $W = \langle 0, k_1 \rangle$, pro pravostrannou alternativu: $W = \langle k_2, n \rangle$.

(Nezáporná celá čísla k_1, k_2 pro oboustranný test i pro jednostranné testy lze najít v tabulkové příloze. Pozor – čísla k_1, k_2 pro oboustrannou alternativu jsou jiná než pro jednostranné alternativy!)

d) H_0 zamítáme na hladině významnosti α, když $S_Z^+ \in W$.

Asymptotická varianta testu

Pro velká n (prakticky $n > 20$) lze využít asymptotické normality statistiky S_Z^+.

Testová statistika $U_0 = \frac{S_Z^+ - E(S_Z^+)}{\sqrt{D(S_Z^+)}} = \frac{S_Z^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$ má za platnosti H_0 asymptoticky rozložení $N(0,1)$.

Kritický obor pro oboustranný test: $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$.

Kritický obor pro levostranný test: $W = (-\infty, -u_{1-\alpha})$.

Kritický obor pro pravostranný test: $W = (u_{1-\alpha}, \infty)$.

Aproximace rozložením $N(0,1)$ se zlepší, když použijeme tzv. korekci na nespojitost. Testová statistika pak má tvar $U_0 = \frac{S_Z^+ - \frac{n}{2} \pm \frac{1}{2}}{\sqrt{\frac{n}{4}}}$, přičemž $\frac{1}{2}$ přičteme, když $S_Z^+ < \frac{n}{2}$ a odečteme v opačném případě.

Příklad

U 9 náhodně vybraných manželských párů byl zjištěn průměrný roční příjem (v tisících Kč).

číslo páru	1	2	3	4	5	6	7	8	9
příjem manžela	216	336	384	432	456	528	552	600	1872
příjem manželky	336	240	192	336	384	288	960	312	576

Na hladině významnosti 0,05 testujte hypotézu, že mediány příjmů manželů a manželek jsou stejné.

Řešení:

Jedná se o párový test. Vypočteme rozdíly mezi příjmy manželů a manželek, čímž úlohu převedeme na jednovýběrový test.

Testujeme $H_0 : z_{0,50} = 0$ proti oboustranné alternativě $H_1 : z_{0,50} \neq 0$, kde $z_{0,50}$ je medián rozložení, z něhož pochází rozdílový náhodný výběr $Z_1 = X_1 - Y_1, \ldots, Z_9 = X_9 - Y_9$.

Vypočtené rozdíly $x_i - y_i$: -120 96 192 96 72 240 -408 288 1296

Testová statistika $S_Z^+ = 7$.

Ve statistických tabulkách najdeme pro $n = 9$ a $\alpha = 0{,}05$ kritické hodnoty $k_1 = 1$, $k_2 = 8$.

Protože kritický obor $W = \langle 0,1 \rangle \cup \langle 8,9 \rangle$ neobsahuje hodnotu 7, nemůžeme H_0 zamítnout na hladině významnosti 0,05.

Neprokázaly se tedy významné rozdíly v mediánech příjmů manželů a manželek.

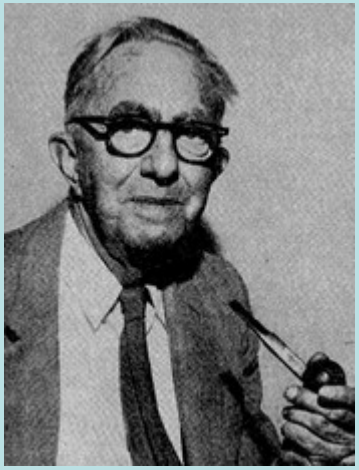
Výpočet pomocí systému STATISTICA:
Vytvoříme nový datový soubor se dvěma proměnnými a 9 případy. Do proměnné X napíšeme příjmy manželů, do proměnné Y příjmy manželek.
Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných X, 2. seznam proměnných Y – OK – Znaménkový test.

Dvojice proměnných	Počet různých	procent v < V	Z	Úroveň p
X & Y	9	22,22222	1,333333	0,182422

Vidíme, že nenulových hodnot n = 9. Z nich záporných je $22{,}\overline{2}\,\%$, tj. 2. Hodnota testové statistiky $s_z^+ = 9 - 2 = 7$. Asymptotická testová statistika U_0 (zde označená jako Z) se realizuje hodnotou $1{,}\overline{3}$. Odpovídající asymptotická p-hodnota je 0,1824, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu, že mediány příjmů manželů a manželek jsou stejné.

Upozornění: V tomto případě není splněna podmínka pro využití asymptotické normality statistiky s_z^+, tj. $n > 20$. Je tedy vhodnější najít v tabulkách kritické hodnoty pro znaménkový test. Pro n = 9 a $\alpha = 0{,}05$ jsou kritické hodnoty $k_1 = 1$, $k_2 = 8$. Protože kritický obor $W = \langle 0,1\rangle \cup \langle 8,9\rangle$ neobsahuje hodnotu 7, nezamítáme H_0 na hladině významnosti 0,05. Dostáváme týž výsledek jako při použití asymptotického testu.

Jednovýběrový Wilcoxonův test a jeho asymptotická varianta



Frank Wilcoxon (1892 – 1965): Americký statistik a chemik

Nechť $X_1, ..., X_n$ je náhodný výběr ze spojitého rozložení s hustotou $\varphi(x)$, která je symetrická kolem mediánu $x_{0,50}$, tj. $\varphi(x_{0,50} + x) = \varphi(x_{0,50} - x)$. Nechť c je reálná konstanta.
Testujeme hypotézu H_0: $x_{0,50} = c$
proti oboustranné alternativě H_1: $x_{0,50} \neq c$ nebo
proti levostranné alternativě H_1: $x_{0,50} < c$ nebo
proti pravostranné alternativě H_1: $x_{0,50} > c$.

Postup provedení testu:

a) Utvoříme rozdíly $D_i = X_i - c$, $i = 1, ..., n$. (Jsou-li některé rozdíly nulové, pak za n bereme jen počet nenulových hodnot.)

b) Absolutní hodnoty $|D_i|$ uspořádáme vzestupně podle velikosti a spočteme pořadí R_i.

c) Zavedeme statistiky

$s_W^+ = \sum_{D_i>0} R_i^+$, což je součet pořadí přes kladné hodnoty D_i,

$s_W^- = \sum_{D_i<0} R_i^-$, což je součet pořadí přes záporné hodnoty D_i.

Přitom platí, že součet $S_W^+ + S_W^- = n(n+1)/2$.
Je-li H_0 pravdivá, pak $E(S_W^+) = n(n+1)/4$ a $D(S_W^+) = n(n+1)(2n+1)/24$.

d) Testová statistika = $\min(S_W^+, S_W^-)$ pro oboustrannou alternativu,
= S_W^+ pro levostrannou alternativu,
= S_W^- pro pravostrannou alternativu.

e) H_0 zamítáme na hladině významnosti α, když testová statistika je menší nebo rovna tabelované kritické hodnotě.

Asymptotická varianta jednovýběrového Wilcoxonova testu:

Pro $n \geq 30$ lze využít asymptotické normality statistiky S_W^+.

Platí-li H_0, pak $U_0 = \frac{S_W^+ - E(S_W^+)}{\sqrt{D(S_W^+)}} = \frac{S_W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \approx N(0,1)$.

Kritický obor:

pro oboustrannou alternativu $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$,

pro levostrannou alternativu $W = (-\infty, -u_{1-\alpha})$,

pro pravostrannou alternativu $W = (u_{1-\alpha}, \infty)$

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Předpoklady použití jednovýběrového Wilcoxonova testu:

- rozložení, z něhož daný náhodný výběr pochází, je spojité
- hustota tohoto rozložení je symetrická kolem mediánu
- sledovaná veličina X má aspoň ordinální charakter

(Není-li splněn předpoklad o symetrii hustoty kolem mediánu, lze použít např. znaménkový test.)

Příklad: U 12 náhodně vybraných zemí bylo zjištěno procento populace starší 60 let:

4,9 6,0 6,9 17,6 4,5 12,3 5,7 5,3 9,6 13,5 15,7 7,7.

Na hladině významnosti 0,05 testujte hypotézu, že medián procenta populace starší 60 let je 12 proti oboustranné alternativě.

Řešení:

Testujeme hypotézu H_0: $x_{0,50} = 12$ proti oboustranné alternativě H_1: $x_{0,50} \neq 12$.

Vypočteme rozdíly pozorovaných hodnot od čísla 12: -7,1 -6,0 -5,1 5,6 -7,5 0,3 -6,3 -6,7 -2,4 1,5 3,7 -4,3.

Absolutní hodnoty těchto rozdílů uspořádáme vzestupně podle velikosti. Kladné rozdíly přitom označíme červeně:

usp. $\lvert x_i - 12 \rvert$	0,3	1,5	2,4	3,7	4,3	5,1	5,6	6	6,3	6,7	7,1	7,5
pořadí	1	2	3	4	5	6	7	8	9	10	11	12

$S_W^+ = 1 + 2 + 4 + 7 = 14$,

$S_W^- = 3 + 5 + 6 + 8 + 9 + 10 + 11 + 12 = 64$,

$n = 12$, $\alpha = 0,05$, tabelovaná kritická hodnota pro $n = 12$ a $\alpha = 0,05$ je 13,

testová statistika = $\min(S_W^+, S_W^-) = \min(14,64) = 14$.

Protože $14 > 13$, H_0 nezamítáme na hladině významnosti 0,05. Znamená to, že na hladině významnosti 0,05 se nepodařilo prokázat, že aspoň v polovině zemí by se podíl populace nad 60 let odlišoval od 12 %.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 12 případy. Do proměnné procento napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 12.
Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných rozdil, Druhý seznam proměnných konst – OK – Wilcoxonův párový test.

	Wilcoxonův párový test (populace_nad_60) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
procento & konst	12	14,00000	1,961161	0,049861

Výstupní tabulka poskytne hodnotu testové statistiky SW^+ (zde označena T), hodnotu asymptotické testové statistiky U_0 a p-hodnotu pro U_0. V tomto případě je p-hodnota 0,049861, tedy nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. Tento výsledek je v rozporu s výsledkem, ke kterému jsme dospěli při přesném výpočtu. Je to způsobeno tím, že není splněna podmínka pro využití asymptotické normality statistiky SW^+, tj. $n \geq 30$.

Dvouvýběrové testy (Jedná se o neparametrickou obdobu dvouvýběrového t-testu)

Dvouvýběrový Wilcoxonův test a jeho asymptotická varianta

Nechť $X_1, ..., X_n$ a $Y_1, ..., Y_m$ jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit pouze posunutím. Označme $x_{0,50}$ medián prvního rozložení a $y_{0,50}$ medián druhého rozložení. Na hladině významnosti 0,05 testujeme hypotézu, že distribuční funkce těchto rozložení jsou shodné neboli mediány jsou shodné proti alternativě, že jsou rozdílné, tj.

H_0: $x_{0,50} - y_{0,50} = 0$ proti H_1: $x_{0,50} - y_{0,50} \neq 0$.

Postup provedení testu:

a) Všech $n + m$ hodnot $X_1, ..., X_n$ a $Y_1, ..., Y_m$ uspořádáme vzestupně podle velikosti.
b) Zjistíme součet pořadí hodnot $X_1, ..., X_n$ a označíme ho T_1.
   Součet pořadí hodnot $Y_1, ..., Y_m$ označíme T_2.
c) Vypočteme statistiky $U_1 = mn + n(n+1)/2 - T_1$, $U_2 = mn + m(m+1)/2 - T_2$.
   Přitom platí $U_1 + U_2 = mn$.
d) Pokud $\min(U_1,U_2) \leq$ tabelovaná kritická hodnota (pro dané rozsahy výběrů m, n a dané α), pak nulovou hypotézu o totožnosti obou distribučních funkcí zamítáme na hladině významnosti α. V tabulkách: $n = \min\{m,n\}$ a $m = \max\{m,n\}$.

Asymptotická varianta dvouvýběrového Wilcoxonova testu:

Pro velká n, m (n, m > 30) lze využít asymptotické normality statistiky U_1.

Platí-li H_0, pak $U_0 = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \approx N(0,1)$, kde $U_1 = \min(U_1, U_2)$.

Kritický obor:

pro oboustrannou alternativu $W = \langle -\infty, -u_{1-\alpha/2} \rangle \cup \langle u_{1-\alpha/2}, \infty \rangle$,

pro levostrannou alternativu $W = \langle -\infty, -u_{1-\alpha} \rangle$,

pro pravostrannou alternativu $W = \langle u_{1-\alpha}, \infty \rangle$

H_0 zamítáme na asymptotické hladině významnosti α, když $U_0 \in W$.

Předpoklady použití dvouvýběrového Wilcoxonova testu:

- dané dva náhodné výběry jsou nezávislé
- rozložení, z nichž dané dva náhodné výběry pocházejí, jsou spojitá
- distribuční funkce těchto rozložení se mohou lišit pouze posunutím
- sledovaná veličina má aspoň ordinální charakter

(Není-li splněn předpoklad, že distribuční funkce se mohou lišit pouze posunutím, lze použít např. dvouvýběrový Kolmogorovův – Smirnovův test.)

Příklad:
Bylo vybráno 10 polí stejné kvality. Na čtyřech z nich se zkoušel nový způsob hnojení, zbylých šest bylo ošetřeno starým způsobem. Pole byla oseta pšenicí a sledoval se její hektarový výnos. Je třeba zjistit, zda nový způsob hnojení má týž vliv na průměrné hektarové výnosy pšenice jako starý způsob hnojení.
hektarové výnosy při novém způsobu: 51 52 49 55
hektarové výnosy při starém způsobu: 45 54 48 44 53 50
Test proveďte na hladině významnosti 0,05.
Řešení:
Na hladině významnosti 0,05 testujeme H_0: $x_{0,50} - y_{0,50} = 0$ proti oboustranné alternativě H_1: $x_{0,50} - y_{0,50} \neq 0$.

usp. hodnoty	44	45	48	**49**	50	**51**	**52**	53	54	**55**
pořadí x-ových hodnot				4		6	7			10
pořadí y-ových hodnot	1	2	3		5			8	9	

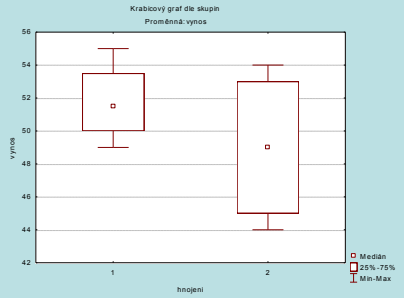
$T_1 = 4 + 6 + 7 + 10 = 27$, $T_2 = 1 + 2 + 3 + 5 + 8 + 9 = 28$
$U_1 = 4.6 + 4.5/2 - 27 = 7$, $U_2 = 4.6 + 6.7/2 - 28 = 17$
Kritická hodnota pro $\alpha = 0,05$, $\min(4,6) = 4$, $\max(4,6) = 6$ je 2. Protože $\min(7,17) = 7 > 2$, nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že nový způsob hnojení má na hektarové výnosy pšenice stejný vliv jako starý způsob.

Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 10 případy. Do proměnné vynos napíšeme zjištěné hodnoty a do proměnné hnojeni napíšeme 4x číslo 1 pro nový způsob hnojení a 6x číslo 2 pro starý způsob hnojení.
Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných vynos, Nezáv. (grupov.) proměnná hnojeni – OK – M-W U test.
Upozornění: Ve STATISTICE je dvouvýběrový Wilcoxonův test uveden pod názvem Mannův – Whitneyův test.

	Mann-Whitneyův U test (vynos) Dle proměn. hnojeni Označené testy jsou významné na hladině p <,05000									
Proměnná	Sčt poř. skup. 1	Sčt poř. skup. 2	U	Z	Úroveň p	Z upravené	Úroveň p	N platn. skup. 1	N platn. skup. 2	2*1str. přesné p
vynos	27,00000	28,00000	7,000000	1,066004	0,286423	1,066004	0,286423	4	6	0,352381

Ve výstupní tabulce jsou součty pořadí T_1, T_2, hodnota testové statistiky $\min(U_1, U_2)$ označená U, hodnota asymptotické testové statistiky U_0 (označená Z), asymptotická p-hodnota pro U_0 a přesná p-hodnota (ozn. 2*1str. přesné p – ta se používá pro rozsahy výběrů pod 30). V našem případě přesná p-hodnota = 0,352381, tedy H_0 nezamítáme na hladině významnosti 0,05.
Výpočet je vhodné doplnit krabicovým diagramem.



Je zřejmé, že výnosy při novém způsobu hnojení jsou vesměs nižší než při starém způsobu a také vykazují mnohem větší variabilitu.

Dvouvýběrový Kolmogorovův - Smirnovův test

Nechť $X_1,\dots,X_n$ a $Y_1,\dots,Y_m$ jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit nejenom posunutím, ale také tvarem. Testujeme hypotézu, že distribuční funkce těchto rozložení jsou shodné, tj., že všech $n+m$ veličin pochází z téhož rozložení proti alternativě, že distribuční funkce jsou rozdílné.

Nechť $F_1(x)$ je výběrová distribuční funkce 1. výběru a $F_2(y)$ je výběrová distribuční funkce 2. výběru. Jako testová statistika slouží $D = \max_{-\infty<x<\infty} |F_1(x) - F_2(x)|$. H_0 zamítáme na hladině významnosti α, když $D \geq D_{n,m}(\alpha)$, kde $D_{n,m}(\alpha)$ je tabelovaná kritická hodnota. Pro větší rozsahy n,m lze kritickou hodnotu aproximovat vzorcem $\sqrt{\frac{n+m}{2nm}\ln\frac{2}{\alpha}}$.

Příklad: Výrobce určitého výrobku se má rozhodnout mezi dvěma dodavateli polotovarů vyrábějících je různými technologiemi. Rozhodující je procentní obsah určité látky.
1. technologie: 1,52 1,57 1,71 1,34 1,68
2. technologie: 1,75 1,67 1,56 1,66 1,72 1,79 1,64 1,55
Na hladině významnosti 0,05 posuďte pomocí dvouvýběrového K-S testu, zda je oprávněný předpoklad, že obě technologie poskytují stejné procento účinné látky.

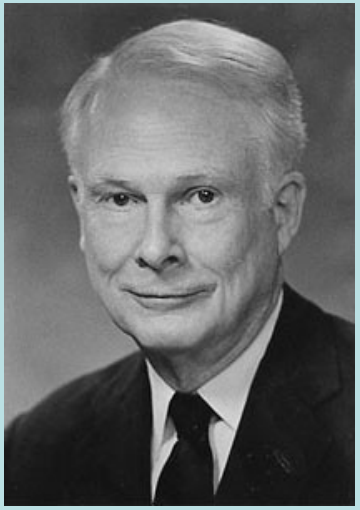
Výpočet pomocí systému STATISTICA:
Utvoříme nový datový soubor se dvěma proměnnými a 13 případy. Do proměnné X napíšeme zjištěné hodnoty a do proměnné ID napíšeme 5x číslo 1 pro první technologii a 8x číslo 2 pro starý druhou technologii.
Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných X, Nezáv. (grupov.) proměnná ID – OK – Kolmogorov-Smirnovův 2-výběrový test.

Proměnná	Max záp rozdíl	Max klad rozdíl	Úroveň p	Průměr skup. 1	Průměr skup. 2	Sm.odch. skup. 1	Sm.odch. skup. 2	N platn. skup. 1	N platn. skup. 2
obsah	-0,400000	0,025000	p > .10	1,564000	1,667500	0,147411	0,085147	5	8

Ve výstupní tabulce pro dvouvýběrový K-S test dostaneme maximální záporný a maximální kladný rozdíl mezi hodnotami obou výběrových distribučních funkcí, dolní omezení pro p-hodnotu ($p > 0,1$), průměry, směrodatné odchylky a rozsahy obou výběrů. Jelikož p-hodnota převyšuje hladinu významnosti 0,05, na této hladině nelze nulovou hypotézu zamítnout.

Kruskalův - Wallisův test

William Kruskal (1919 – 2005):
Americký matematik



Wilson Allen Wallis (1912 – 1988):
Americký matematik

Nechť je dáno $r \geq 3$ nezávislých náhodných výběrů o rozsazích $n_1, \ldots, n_r$. Předpokládáme, že tyto výběry pocházejí ze spojitých rozložení. Označme $n = n_1 + \ldots + n_r$. Na asymptotické hladině významnosti α chceme testovat hypotézu, že všechny tyto výběry pocházejí z téhož rozložení.

Postup testu:

a) Všech n hodnot seřadíme do rostoucí posloupnosti.
b) Určíme pořadí každé hodnoty v tomto sdruženém výběru.
c) Označme T_j součet pořadí těch hodnot, které patří do j-tého výběru, j = 1, ..., r (kontrola: musí platit $T_1 + ... + T_r = n(n+1)/2$).
d) Testová statistika má tvar: $Q = \frac{12}{n(n+1)} \sum_{j=1}^{r} \frac{T_j^2}{n_j} - 3(n+1)$. Platí-li H_0, má statistika Q asymptoticky rozložení $\chi^2(r-1)$.
e) Kritický obor: $w = \langle \chi^2_{1-\alpha}(r-1), \infty \rangle$.
f) H_0 zamítneme na asymptotické hladině významnosti α, když $Q \geq \chi^2_{1-\alpha}(r-1)$.

Příklad: V roce 1980 byly získány tři nezávislé výběry obsahující údaje o průměrných ročních příjmech (v tisících dolarů) čtyř sociálních skupin ve třech různých oblastech USA.
jižní oblast: 6 10 15 29
pacifická oblast: 11 13 17 131
severovýchodní oblast: 7 14 28 25
Na hladině významnosti 0,05 testujte hypotézu, že příjmy v těchto oblastech se neliší.
Řešení:
Výpočty uspořádáme do tabulky

Usp. hodnoty	6	7	10	11	13	14	15	17	25	28	29	131
Pořadí 1.výběru	1		3				7				11	
Pořadí 2.výběru				4	5			8				12
Pořadí 3.výběru		2				6			9	10		

$T_1 = 1 + 3 + 7 + 11 = 22,$
$T_2 = 4 + 5 + 8 + 12 = 29,$
$T_3 = 2 + 6 + 9 + 10 = 27,$

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^{r} \frac{T_j^2}{n_j} - 3(n+1) = \frac{12}{12 \cdot 13} \left(\frac{22^2}{4} + \frac{29^2}{4} + \frac{27^2}{4} \right) - 3 \cdot 13 = 0{,}5,$$

$$W = \langle \chi^2_{1-\alpha}(r-1), \infty \rangle = \langle \chi^2_{0,95}(2), \infty \rangle = \langle 5{,}991, \infty \rangle$$

Protože $Q < 5{,}991$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.
Rozdíly mezi průměrnými ročními příjmy v uvedených třech oblastech se neprokázaly.

Mediánový test

Výchozí situace je stejná jako u K-W testu

Postup testu:

a) Všech n hodnot uspořádáme do rostoucí posloupnosti.

b) Najdeme medián $x_{0,50}$ těchto n hodnot.

c) Označme P_j počet hodnot v j-tém výběru, které jsou větší nebo rovny mediánu $x_{0,50}$.

d) Testová statistika má tvar $Q_M = 4\sum_{j=1}^{r} \frac{P_j^2}{n_j} - n$. Platí-li H_0, má statistika Q_M asymptoticky rozložení $\chi^2(r-1)$.

d) Kritický obor: $w = \langle \chi^2_{1-\alpha}(r-1), \infty)$.

e) H_0 zamítneme na asymptotické hladině významnosti α, když $Q_M \geq \chi_{1-\alpha}{}^2(r-1)$.

Příklad:
Pro data o průměrných ročních příjmech proveďte mediánový test. Hladinu významnosti volte 0,05.

Řešení:
Usp. hodnoty 6 7 10 11 13 14 15 17 25 28 29 131

Medián je průměr 6. a 7. uspořádané hodnoty: $x_{0,50} = \frac{14 + 15}{2} = 14,5$.

V prvním výběru existují 2 hodnoty, které jsou větší nebo rovny 14,5, stejně tak i ve druhém a třetím výběru, tedy $P_1 = P_2 = P_3 = 2$.

Testová statistika: $Q_M = 4\sum_{j=1}^{r} \frac{P_j^2}{n_j} - n = 4\left[\frac{1}{4}(2^2 + 2^2 + 2^2)\right] - 12 = 0$

Kritický obor: $W = \langle \chi^2_{1-\alpha}(r-1), \infty \rangle = \langle \chi^2_{0,95}(2), \infty \rangle = \langle 5,991, \infty \rangle$

Protože $Q_M < 5,991$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Metody mnohonásobného porovnávání

Zamítneme-li hypotézu, že všechny náhodné výběry pocházejí z téhož rozložení, zajímá nás, které dvojice náhodných výběrů se liší na zvolené hladině významnosti. Testujeme H_0: k-tý a l-tý náhodný výběr pocházejí z téhož rozložení, k, l = 1, .., r, k ≠ l proti H_1: aspoň jedna dvojice výběrů pochází z různých rozložení.

a) Neményiho metoda (Peter Neményi 1927 – 2002: Americký matematik maďarského původu)

- Všechny výběry mají týž rozsah p (třídění je vyvážené).
- Vypočteme $|T_l - T_k|$.
- V tabulkách najdeme kritickou hodnotu (pro dané p, r, α).
- Pokud $|T_l - T_k| \geq$ tabelovaná kritická hodnota, pak na hladině významnosti α zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

b) Obecná metoda mnohonásobného porovnávání

- Vypočteme $\left|\frac{T_l}{n_l} - \frac{T_k}{n_k}\right|$.
- Ve speciálních statistických tabulkách najdeme kritickou hodnotu $h_{KW}(\alpha)$. Při větších rozsazích výběrů je možno ji nahradit kvantilem $\chi_{1-\alpha}^2(r-1)$.
- Jestliže $\left|\frac{T_l}{n_l} - \frac{T_k}{n_k}\right| \geq \sqrt{\frac{1}{12}\left(\frac{1}{n_l} + \frac{1}{n_k}\right)n(n+1)h_{KW}(\alpha)}$, pak na hladině významnosti α zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

Příklad:
Čtyři laboranti provedli analytické stanovení procenta niklu v oceli. Každý hodnotil pět vzorků.
Laborant A: 4,15 4,26 4,10 4,30 4,25
Laborant B: 4,38 4,40 4,29 4,39 4,45
Laborant C: 4,23 4,16 4,20 4,24 4,27
Laborant D: 4,41 4,31 4,42 4,37 4,43
Na asymptotické hladině významnosti 0,05 testujte hypotézu, že všechny čtyři náhodné výběry pocházejí ze stejného rozložení. Pokud nulovou hypotézu zamítnete, zjistěte, které dvojice výběrů se liší.

Výpočet pomocí systému STATISTICA:
Vytvoříme nový datový soubor o dvou proměnných a 20 případech. Do proměnné nikl napíšeme změřené hodnoty, do proměnné laborant napíšeme 5x1 pro 1. laboranta atd. až 5x4 pro 4. laboranta.
Statistiky – Neparametrická statistika – Porovnání více nezávislých vzorků - OK – Seznam závislých proměnných nikl, Nezáv. (grupovací) proměnná laborant – OK – Summary: Kruskal-Wallis ANOVA & Median test. Ve dvou výstupních tabulkách se objeví výsledky K-W testu a mediánového testu.

Kruskal-Wallisova ANOVA založ. na poř.; nikl (nikl v oceli)
Nezávislá (grupovací) proměnná : laborant
Kruskal-Wallisův test: H (3, N= 20) =13,77714 p =,0032

Závislá: nikl	Kód	Počet platných	Součet pořadí
1	1	5	29,00000
2	2	5	75,00000
3	3	5	27,00000
4	4	5	79,00000

Mediánový test, celk. medián = 4,29500; nikl (nikl v oceli)
Nezávislá (grupovací) proměnná : laborant
Chi-Kvadr. = 13,60000 sv = 3 p = ,0035

Závislá: nikl	1	2	3	4	Celkem
<= Medián: pozorov.	4,00000	1,00000	5,00000	0,00000	10,00000
očekáv.	2,50000	2,50000	2,50000	2,50000	
poz.-oč.	1,50000	-1,50000	2,50000	-2,50000	
> Medián: pozorov.	1,00000	4,00000	0,00000	5,00000	10,00000
očekáv.	2,50000	2,50000	2,50000	2,50000	
poz.-oč.	-1,50000	1,50000	-2,50000	2,50000	
Celkem: oček.	5,00000	5,00000	5,00000	5,00000	20,00000

Oba testy zamítají hypotézu o shodě mediánů v daných čtyřech skupinách na asymptotické hladině významnosti 0,05.

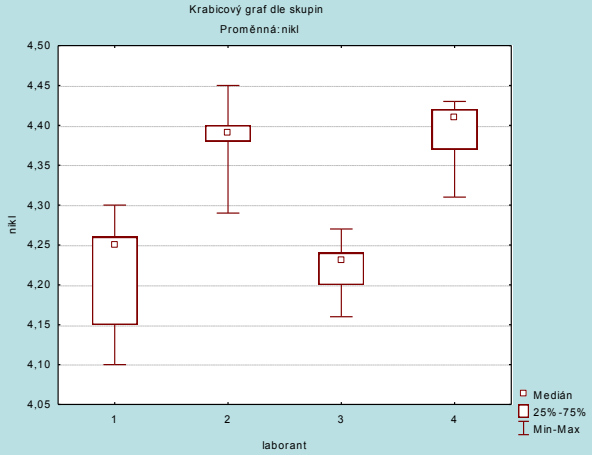
Nyní provedeme mnohonásobné porovnávání, abychom zjistili, které dvojice laborantů se liší. Zvolíme Vícenás. porovnání průměrného pořadí pro vš. skupiny.

Vícenásobné porovnání p hodnot (oboustr.); nikl (nikl v oceli)
Nezávislá (grupovací) proměnná : laborant
Kruskal-Wallisův test: H (3, N= 20) =13,77714 p =,0032

Závislá: nikl	1 R:5,8000	2 R:15,000	3 R:5,4000	4 R:15,800
1		0,083641	1,000000	0,045158
2	0,083641		0,061779	1,000000
3	1,000000	0,061779		0,032664
4	0,045158	1,000000	0,032664	

Tabulka obsahuje p-hodnoty pro porovnání dvojic skupin. Vidíme, že na hladině významnosti 0,05 se liší laboranti A, D a laboranti C, D.

Grafické znázornění výsledků



Hodnocení kontingenčních tabulek

Motivace

Při zpracování dat se velmi často setkáme s úkolem zjistit, zda dvě náhodné veličiny nominálního typu jsou stochasticky nezávislé. Např. nás může zajímat, zda ve sledované populaci je barva očí a barva vlasů nezávislá.
Zpravidla chceme také zjistit intenzitu případné závislosti sledovaných dvou veličin. K tomuto účelu byly zkonstruovány různé koeficienty, které nabývají hodnot od 0 do 1. Čím je takový koeficient bližší 1, tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.

Kontingenční tabulky

Nechť X,Y jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť X nabývá variant $x_{[1]}, ..., x_{[r]}$ a Y nabývá variant $y_{[1]}, ..., y_{[s]}$.
Označme:

$\pi_{jk} = P(X = x_{[j]} \wedge Y = y_{[k]})$ … simultánní pravděpodobnost dvojice variant $(x_{[j]}, y_{[k]})$

$\pi_{j.} = P(X = x_{[j]})$ … marginální pravděpodobnost varianty $x_{[j]}$

$\pi_{.k} = P(Y = y_{[k]})$ … marginální pravděpodobnost varianty $y_{[k]}$

Simultánní a marginální pravděpodobnosti zapíšeme do kontingenční tabulky:

	y	$y_{[1]}$	…	$y_{[s]}$	$\pi_{j.}$
x	π_{jk}				
$x_{[1]}$		π_{11}	…	π_{1s}	$\pi_{1.}$
…		…	…	…	…
$x_{[r]}$		π_{r1}	…	π_{rs}	$\pi_{r.}$
$\pi_{.k}$		$\pi_{.1}$	…	$\pi_{.s}$	1

Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), ..., (X_n, Y_n)$ rozsahu n z rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor (X, Y). Zjištěné absolutní simultánní četnosti n_{jk} dvojice variant $(x_{[j]}, y_{[k]})$ uspořádáme do kontingenční tabulky:

	y	$y_{[1]}$	...	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		n_{11}	...	n_{1s}	$n_{1.}$
...		...	...	...	...
$x_{[r]}$		n_{r1}	...	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	...	$n_{.s}$	n

$n_{j.} = n_{j1} + ... + n_{js}$ je marginální absolutní četnost varianty $x_{[j]}$
$n_{.k} = n_{1k} + ... + n_{rk}$ je marginální absolutní četnost varianty $y_{[k]}$

Simultánní pravděpodobnost π_{jk} odhadneme pomocí simultánní relativní četnosti $p_{jk} = \frac{n_{jk}}{n}$, marginální pravděpodobnosti $\pi_{j.}$ a $\pi_{.k}$ odhadneme pomocí marginálních relativních četností $p_{j.} = \frac{n_{j.}}{n}$ a $p_{.k} = \frac{n_{.k}}{n}$.

Testování hypotézy o nezávislosti

Testujeme nulovou hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny.

Kdyby náhodné veličiny X, Y byly stochasticky nezávislé, pak by platil multiplikativní vztah

$\forall j = 1,\dots,r$, $\forall k = 1,\dots,s : \pi_{jk} = \pi_{j.}\,\pi_{.k}$ neboli $\frac{n_{jk}}{n} = \frac{n_{j.}}{n} \cdot \frac{n_{.k}}{n}$, tj. $n_{jk} = \frac{n_{j.} n_{.k}}{n}$. Číslo $m_{jk} = \frac{n_{j.} n_{.k}}{n}$ se nazývá teoretická četnost dvojice variant $(x_{[j]}, y_{[k]})$.

Testová statistika: $K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.} n_{.k}}{n} \right)^2}{\frac{n_{j.} n_{.k}}{n}}$.

Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$.

Kritický obor: $w = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty \rangle$.

Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

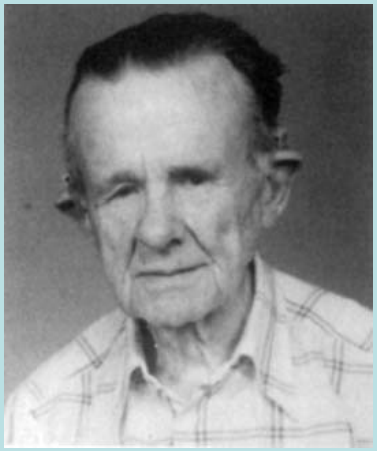
Podmínky dobré aproximace

Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti $\frac{n_{j.} n_{.k}}{n}$ aspoň v 80% případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20% neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

Měření síly závislosti

Cramérův koeficient: $v = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r,s\}$. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je k 1, tím je závislost mezi X a Y těsnější, čím blíže je k 0, tím je tato závislost volnější.

Význam hodnot Cramérova koeficientu:
mezi 0 až 0,1 … zanedbatelná závislost,
mezi 0,1 až 0,3 … slabá závislost,
mezi 0,3 až 0,7 … střední závislost,
mezi 0,7 až 1 … silná závislost.



Carl Harald Cramér (1893 – 1985): Švédský matematik

Příklad

V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází (veličina X) a typ školy, na kterou se hlásí (veličina Y). Výsledky jsou zaznamenány v kontingenční tabulce:

Sociální skupina	Typ školy			$n_{j.}$
	univerzitní	technický	ekonomický	
I	50	30	10	90
II	30	50	20	100
III	10	20	30	60
IV	50	10	50	110
$n_{.k}$	140	110	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérův koeficient.

Řešení:

Nejprve vypočteme všech 12 teoretických četností:

Sociální skupina	Typ školy			$n_{j.}$
	univerzitní	technický	ekonomický	
I	50	30	10	90
II	30	50	20	100
III	10	20	30	60
IV	50	10	50	110
$n_{.k}$	140	110	110	360

$$\frac{n_{1.}n_{.1}}{n}=\frac{90\cdot 140}{360}=35,\quad \frac{n_{1.}n_{.2}}{n}=\frac{90\cdot 110}{360}=27{,}5,\quad \frac{n_{1.}n_{.3}}{n}=\frac{90\cdot 110}{360}=27{,}5,$$

$$\frac{n_{2.}n_{.1}}{n}=\frac{100\cdot 140}{360}=38{,}9,\quad \frac{n_{2.}n_{.2}}{n}=\frac{100\cdot 110}{360}=30{,}6,\quad \frac{n_{2.}n_{.3}}{n}=\frac{100\cdot 110}{360}=30{,}6,$$

$$\frac{n_{3.}n_{.1}}{n}=\frac{60\cdot 140}{360}=23{,}3,\quad \frac{n_{3.}n_{.2}}{n}=\frac{60\cdot 110}{360}=18{,}3,\quad \frac{n_{3.}n_{.3}}{n}=\frac{60\cdot 110}{360}=18{,}3,$$

$$\frac{n_{4.}n_{.1}}{n}=\frac{110\cdot 140}{360}=42{,}8,\quad \frac{n_{4.}n_{.2}}{n}=\frac{110\cdot 110}{360}=33{,}6,\quad \frac{n_{4.}n_{.3}}{n}=\frac{110\cdot 110}{360}=33{,}6$$

Vidíme, že podmínky dobré aproximace jsou splněny, všechny teoretické četnosti převyšují číslo 5.

Dosadíme do vzorce pro testovou statistiku K:

$$K=\frac{(50-35)^2}{35}+\frac{(30-27{,}5)^2}{27{,}5}+\ldots+\frac{(50-33{,}6)^2}{33{,}6}=76{,}84\ .$$

Dále stanovíme kritický obor:

$$W=\langle \chi^2_{1-\alpha}((4-1)(3-1)),\infty)=\langle \chi^2_{0,95}(6),\infty)=\langle 12{,}6,\infty)$$

Protože $K \in W$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05.

Vypočteme Cramérův koeficient: $V=\sqrt{\frac{76{,}4}{360\cdot 2}}=0{,}3267$.

Hodnota Cramérova koeficientu svědčí o tom, že mezi veličinami X a Y existuje středně silná závislost.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o třech proměnných (X - sociální skupina, Y – typ školy, četnost) a 12 případech:

	1 X	2 Y	3 četnost
1	I	univerzitní	50
2	I	technický	30
3	I	ekonomický	10
4	II	univerzitní	30
5	II	technický	50
6	II	ekonomický	20
7	III	univerzitní	10
8	III	technický	20
9	III	ekonomický	30
10	IV	univerzitní	50
11	IV	technický	10
12	IV	ekonomický	50

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Očekávané četnosti. Dostaneme kontingenční tabulku teoretických četností:

Souhmná tab.: Očekáv ané četnosti (ty p skoly)
Četnost označených buněk > 10
Pearsonův chí-kv. : 76,8359, sv=6, p=,000000

X	Y univerzitní	Y technický	Y ekonomický	Řádk. součty
I	35,0000	27,5000	27,5000	90,0000
II	38,8889	30,5556	30,5556	100,0000
III	23,3333	18,3333	18,3333	60,0000
IV	42,7778	33,6111	33,6111	110,0000
Vš.skup.	140,0000	110,0000	110,0000	360,0000

Všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny. V záhlaví tabulky je uvedena hodnota testové statistiky K = 76,8359, počet stupňů volnosti 6 a odpovídající p-hodnota. Je velmi blízká 0, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o nezávislosti typu školy a sociální skupiny.
Hodnotu testové statistiky a Cramérův koeficient dostaneme také tak, že na na záložce Možnosti zaškrtneme Pearsonův & M-V chí kvadrát a Cramérovo V, na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

Statist.	Chí-kv adr.	sv	p
Pearsonův chí-kv.	76,83589	df=6	p=,00000
M-V chí-kv adr.	84,53528	df=6	p=,00000
Fí	,4619881		
Kontingenční koef icient	,4193947		
Cramér. V	,3266749		

Čtyřpolní tabulky

Nechť r = s = 2. Pak hovoříme o čtyřpolní kontingenční tabulce a používáme označení: $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$.

X	Y		$n_{j.}$
	$y_{[1]}$	$y_{[2]}$	
$x_{[1]}$	a	b	a+b
$x_{[2]}$	c	d	c+d
$n_{.k}$	a+c	b+d	n

Test nezávislosti ve čtyřpolní tabulce

Testovou statistiku pro čtyřpolní kontingenční tabulku lze zjednodušit do tvaru:

$$K = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

Platí-li hypotéza o nezávislosti veličin X, Y, pak K se asymptoticky řídí rozložením $\chi^2(1)$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}(1), \infty)$

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \in W$.

Povšimněte si, že za platnosti hypotézy o nezávislosti ad = bc.

Pro čtyřpolní tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako Fisherův faktoriálový test.

Sir Ronald Aylmer Fisher (1890 – 1962): Britský statistik a genetik.
(Fisherův přesný test je popsán např. v knize K. Zvára: Biostatistika, Karolinum, Praha 1998. Princip spočívá v tom, že pomocí kombinatorických úvah se vypočítají pravděpodobnosti toho, že při daných marginálních četnostech dostaneme tabulky, které se od nulové hypotézy odchylují aspoň tak, jako daná tabulka.)

Upozornění: STATISTICA poskytuje p-hodnotu pro Fisherův přesný test. Jestliže vyjde $p \leq \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.

Příklad: V náhodném výběru 50 obézních dětí ve věku 6 – 14 let byla zjišťována obezita rodičů. Veličina X – obezita matky, veličina Y – obezita otce. Výsledky průzkumu jsou uvedeny v kontingenční tabulce:

X	Y		$n_{j.}$
	ano	ne	
ano	15	9	24
ne	7	19	26
$n_{.k}$	22	28	50

Pomocí Fisherova exaktního testu ověřte, zda lze na hladině významnosti 0,05 zamítnout hypotézu o nezávislosti náhodných veličin X a Y.

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor o třech proměnných X, Y (varianty 0 – neobézní, 1 – obézní) a četnost a čtyřech případech:

	1 X	2 Y	3 četnost
1	obézní	obézní	15
2	obézní	neobézní	9
3	neobézní	obézní	7
4	neobézní	neobézní	19

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Fisher exakt., Yates, McNemar (2x2). Dostaneme výstupní tabulku:

	Statist. : X(2) x Y(2) (obezita rodicu)		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	6,410777	df=1	p=,01134
M-V chí-kvadr.	6,548348	df=1	p=,01050
Yatesův chí-kv.	5,048207	df=1	p=,02465
Fisherův přesný, 1-str.			p=,01188
2-stranný			p=,02163
McNemarův chí-kv. (A/D)	,2647059	df=1	p=,60691
(B/C)	,0625000	df=1	p=,80259

Vidíme, že p-hodnota pro Fisherův exaktní oboustranný test je 0,02163, tedy na hladině významnosti 0,05 zamítáme hypotézu, že obezita matky a otce spolu nesouvisí.

Podíl šancí ve čtyřpolní kontingenční tabulce

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá výběrový podíl šancí (odds ratio). Považujeme ho za odhad neznámého teoretického podílu šancí $\omega = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$. Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		$n_{j.}$
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
$n_{.k}$	a+c	b+d	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za 1. okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$. Podíl šancí je tedy $OR = \frac{ad}{bc}$.

Jsou-li veličiny X, Y nezávislé, pak $\pi_{jk} = \pi_{j.}\pi_{.k}$, tudíž teoretický podíl šancí $\omega = 1$. Závislost veličin X, Y bude tím silnější, čím více se ω bude lišit od 1. Avšak $\omega \in \langle 0, \infty \rangle$, tedy hodnoty ω jsou kolem 1 rozmístěny nesymetricky. Z tohoto důvodu raději používáme logaritmus teoretického či výběrového podílu šancí.

Testování nezávislosti ve čtyřpolních tabulkách pomocí podílu šancí

Na asymptotické hladině významnosti α testujeme hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny (tj. $\ln \psi = 0$) proti alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $\ln \psi \neq 0$).

Testová statistika $T_0 = \dfrac{\ln OR}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$ se asymptoticky řídí rozložením $N(0,1)$, když nulová hypotéza platí.

Kritický obor: $W = \langle -\infty, -u_{1-\alpha/2} \rangle \cup \langle u_{1-\alpha/2}, \infty \rangle$.

Nulovou hypotézu tedy zamítáme na asymtotické hladině významnosti α, když se testová statistika realizuje v kritickém oboru W.

Testování nezávislosti lze provést též pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro logaritmus podílu šancí ψ, který je dán vzorcem:

$$\langle d, h \rangle = \left(\ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2}, \ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} \right)$$

Jestliže interval spolehlivosti neobsahuje 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

Příklad (testování nezávislosti pomocí podílu šancí a pomocí statistiky K):
U 135 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

přijetí	dojem		$n_{j.}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{.k}$	56	69	125

Řešení:
a) Testování pomocí podílu šancí:
$OR = \frac{ad}{bc} = \frac{17 \cdot 58}{11 \cdot 39} = 2{,}298$. Podíl šancí nám říká, že uchazeč, který zapůsobil na komisi dobrým dojmem, má asi 2,3 x větší šanci na přijetí než uchazeč, který zapůsobil špatným dojmem.

Provedeme další pomocné výpočty:
$\ln OR = 0{,}832,$

$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}} = 0{,}439$, $u_{0{,}975} = 1{,}96$

Dosadíme do vzorců pro meze asymptotického intervalu spolehlivosti pro podíl šancí:
$\ln d = \ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} = 0{,}832 - 0{,}439 \cdot 1{,}96 = -0{,}028$, $\ln h = \ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} = 0{,}832 + 0{,}439 \cdot 1{,}96 = 1{,}692$

Protože interval (-0,028; 1,692) obsahuje číslo 0, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

b) Testování pomocí statistiky K:

přijetí	dojem		$n_{j.}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{.k}$	56	69	125

Ověříme splnění podmínek dobré aproximace:

$$\frac{n_{1.}n_{.1}}{n} = \frac{28 \cdot 56}{125} = 12{,}544\ , \quad \frac{n_{1.}n_{.2}}{n} = \frac{28 \cdot 69}{125} = 15{,}456\ ,$$

$$\frac{n_{2.}n_{.1}}{n} = \frac{97 \cdot 56}{125} = 43{,}456\ , \quad \frac{n_{2.}n_{.2}}{n} = \frac{97 \cdot 69}{125} = 53{,}544$$

Podmínky dobré aproximace jsou splněny.

Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{125 \cdot (17 \cdot 58 - 11 \cdot 39)^2}{28 \cdot 97 \cdot 56 \cdot 69} = 3{,}6953$$

Kritický obor: $W = (\chi^2_{0,95}(1), \infty) = (3{,}841, \infty)$.

Protože testová statistika se nerealizuje k kritickém oboru, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Vypočteme ještě Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}} = \sqrt{\frac{3{,}6953}{125\,(2-1)}} = 0{,}1719$

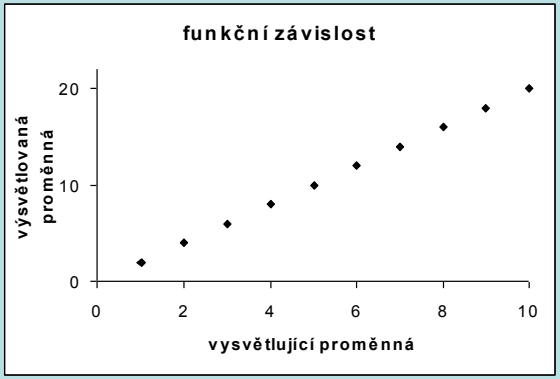
Vidíme, že mezi dojmem u přijímací zkoušky a přijetím na fakultu je pouze slabá závislost.

Jednoduchá korelační analýza

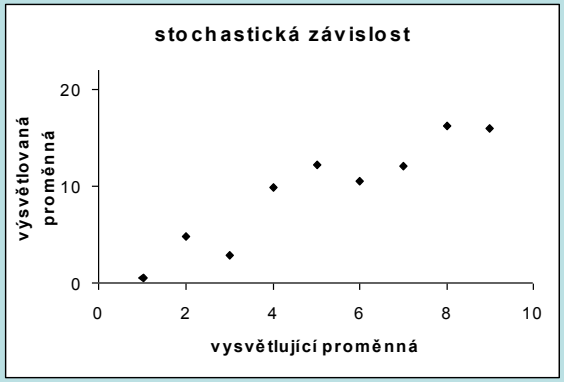
Motivace

Uvažme náhodné veličiny X, Y, které jsou aspoň ordinálního typu. Tyto náhodné veličiny mohou mít různý vztah:

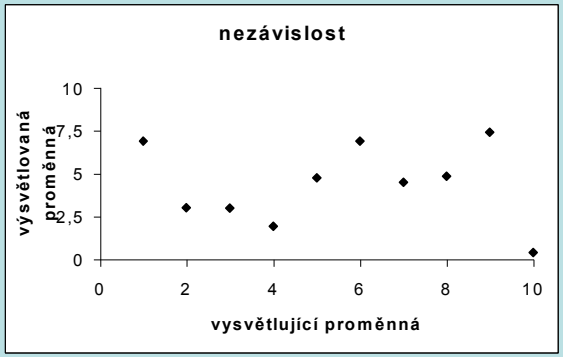
- Deterministická (funkční) závislost: jedna náhodná veličina je spjata s druhou náhodnou veličinou funkční závislostí vyjádřenou předpisem Y = g(X), např. X – poloměr náhodně vybrané sériově vyráběné kuličky do kuličkových ložisek, Y = $\frac{4}{3}\pi X^3$ - objem této kuličky. Každé realizaci náhodné veličiny X (vysvětlující proměnná) je přiřazena právě jedna realizace náhodné veličiny Y (vysvětlovaná proměnná).



- Stochastická závislost: jedna náhodná veličina ovlivňuje v různé míře druhou náhodnou veličinu, např. X – věk pracovníka v letech, Y – počet dnů absence za rok. Každé realizaci náhodné veličiny X může být přiřazeno více realizací náhodné veličiny Y. Závislost může být jednostranná i oboustranná.



- Stochastická nezávislost: náhodné veličiny se navzájem neovlivňují, např. házíme-li naráz dvěma kostkami a označíme X – počet ok padlých na jedné kostce, Y – počet ok padlých na druhé kostce, pak náhodné veličiny X, Y jsou stochasticky nezávislé.



X a Y jsou stochasticky nezávislé, když platí: $\forall (x, y) \in \mathbb{R}^2 : \Phi(x, y) = \Phi_1(x)\,\Phi_2(y)$

X a Y jsou nekorelované, když platí C(X, Y) = 0 (tj. mezi X a Y není žádný lineární vztah).

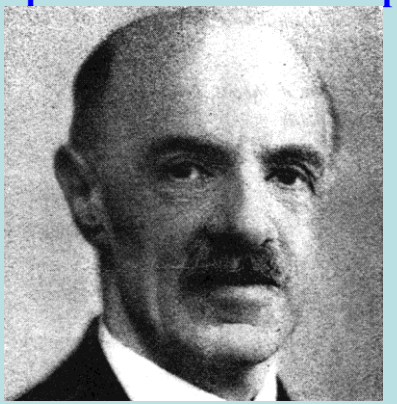
Ze stochastické nezávislosti vyplývá nekorelovanost, avšak z nekorelovanosti nevyplývá stochastická nezávislost.

Korelační analýza:

- zkoumá, zda existuje závislost mezi dvěma náhodnými veličinami X, Y, které jsou buď ordinálního nebo intervalového či poměrového typu. Důležité – nelze se spokojit s formálním matematickým popisem závislosti, závislost musí být logicky zdůvodnitelná!
- pomocí Pearsonova či Spearmanova koeficientu korelace měří těsnost této závislosti
- pro náhodné veličiny intervalového a poměrového typu je založena na předpokladu, že dvourozměrný náhodný vektor $\begin{pmatrix} X \\ Y \end{pmatrix}$ se řídí dvourozměrným normálním rozložením $N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$, kde

  $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = D(X)$, $\sigma_2^2 = D(Y)$, $\rho = R(X,Y)$
- při výraznějším porušení předpokladu dvourozměrné normality doporučuje použití metod, které jsou určeny pro náhodné veličiny ordinálního typu

Spearmanův koeficient pořadové korelace



Charles Edward Spearman (1863 – 1945): Britský psycholog a statistik, zakladatel faktorové analýzy

Nechť X,Y jsou náhodné veličiny ordinálního typu (tj. obsahová interpretace je možná jenom u relace rovnosti a relace uspořádání).
Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), ..., (X_n, Y_n)$ z rozložení, jímž se řídí náhodný vektor (X, Y). Označíme R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny Y_i, i = 1, ..., n.

Spearmanův koeficient pořadové korelace: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$.

Tento koeficient nabývá hodnot mezi –1 a 1. Čím je bližší 1, tím je silnější přímá pořadová závislost mezi veličinami X a Y, čím je bližší –1, tím je silnější nepřímá pořadová závislost mezi veličinami X a Y. Teoretická hodnota Spearmanova koeficientu se značí ρ_S.

Vlastnosti Spearmanova koeficientu pořadové korelace

Pro Spearmanův koeficient pořadové korelace platí $-1 \leq r_S \leq 1$. Čím je bližší 1, tím je silnější přímá pořadová závislost mezi veličinami X a Y, čím je bližší –1, tím je silnější nepřímá pořadová závislost mezi veličinami X a Y.

Je-li $r_S = 1$ resp. $r_S = -1$, pak realizace $[x_i, y_i]$, $i = 1, \ldots, n$ daného náhodného výběru leží na nějaké rostoucí resp. klesající funkci.

Hodnoty r_S se nezmění, když provedeme vzestupnou transformaci původních dat.

Hodnoty r_S se vynásobí -1, když provedeme sestupnou transformaci původních dat.

Koeficient je symetrický.

Koeficient je rezistentní vůči odlehlým hodnotám.

Význam absolutní hodnoty Spearmanova koeficientu:
mezi 0 až 0,1 … zanedbatelná pořadová závislost,
mezi 0,1 až 0,3 … slabá pořadová závislost,
mezi 0,3 až 0,7 … střední pořadová závislost,
mezi 0,7 až 1 … silná pořadová závislost.

Spearmanův koeficient pořadové korelace se používá v situacích, kdy
- zkoumaná data mají ordinální charakter
- nelze předpokládat, že vztah mezi veličinami X, Y je lineární
- náhodný výběr nepochází z dvourozměrného normálního rozložení

Testování nezávislosti ordinálních veličin

Na hladině významnosti α testujeme hypotézu H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti

- oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny
- levostranné alternativě H_1: mezi X a Y existuje nepřímá pořadová závislost
- pravostranné alternativě H_1: mezi X a Y existuje přímá pořadová závislost).

Jako testová statistika slouží Spearmanův koeficient pořadové korelace r_S.

Nulovou hypotézu zamítáme na hladině významnosti α ve prospěch

- oboustranné alternativy, když $|r_S| \geq r_{S,1-\alpha/2}(n)$
- levostranné alternativy, když $r_S \leq -r_{S,1-\alpha}(n)$
- pravostranné alternativy, když $r_S \geq r_{S,1-\alpha}(n)$,

kde $r_{S,1-\alpha}(n)$ je kritická hodnota, kterou pro $\alpha = 0{,}05$ nebo $0{,}01$ a $n \leq 30$ najdeme v tabulkách.

Asymptotické varianty testu

Pro n > 20 lze použít testovou statistiku $T_0 = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}}$, která se v případě platnosti nulové hypotézy asymptoticky řídí rozložením t(n-2).

Kritický obor pro oboustrannou alternativu: $W = (-\infty, -t_{1-\alpha/2}(n-2)) \cup (t_{1-\alpha/2}(n-2), \infty)$

Kritický obor pro levostrannou alternativu:
$W = (-\infty, -t_{1-\alpha}(n-2))$

Kritický obor pro pravostrannou alternativu:
$W = (t_{1-\alpha}(n-2), \infty)$.

Hypotézu o pořadové nezávislosti náhodných veličin X, Y zamítáme na asymptotické hladině významnosti α, když $t_0 \in W$.

Upozornění: Systém STATISTICA používá tuto variantu testu pořadové nezávislosti bez ohledu na rozsah náhodného výběru.

Pro n > 30 lze použít testovou statistiku $r_s\sqrt{n-1}$. Platí-li H_0, pak $r_s\sqrt{n-1} \approx N(0, 1)$. Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α ve prospěch

oboustranné alternativy, když $r_S\sqrt{n-1} \in (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$,

levostranné alternativy, když $r_S\sqrt{n-1} \in (-\infty, -u_{1-\alpha})$,

pravostranné alternativy, když $r_S\sqrt{n-1} \in (u_{1-\alpha}, \infty)$

Příklad na testování pořadové nezávislosti (jsou známa pořadí):
Dva lékaři hodnotili stav sedmi pacientů po témž chirurgickém zákroku. Postupovali tak, že nejvyšší pořadí dostal nejtěžší případ.

Číslo pacienta	1	2	3	4	5	6	7
Hodnocení 1. lékaře	4	1	6	5	3	2	7
Hodnocení 2. lékaře	4	2	5	6	1	3	7

Vypočtěte Spearmanův koeficient a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.

Řešení:
Na hladině významnosti 0,05 testujeme H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě H_1: X, Y jsou pořadově závislé náhodné veličiny. V tomto příkladě přímo známe pořadí R_i (tj. hodnocení 1. lékaře) a pořadí Q_i (tj. hodnocení 2. lékaře). Vypočteme

$$r_S = 1 - \frac{6}{7(7^2-1)}\left[(4-4)^2 + (1-2)^2 + (6-5)^2 + (5-6)^2 + (3-1)^2 + (2-3)^2 + (7-7)^2\right] = 0{,}857 .$$

Kritická hodnota: $r_{S,0,95}(7) = 0{,}745$. Protože $0{,}857 \geq 0{,}745$, nulovou hypotézu zamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor o dvou proměnných X (hodnocení 1. lékaře), Y (hodnocení 2. lékaře) a sedmi případech. Do proměnných X a Y zapíšeme zjištěná hodnocení.

	1 X	2 Y
1	4	4
2	1	2
3	6	5
4	5	6
5	3	1
6	2	3
7	7	7

Statistiky – Neparametrické statistiky – Korelace – OK – vybereme Vytvořit detailní report - Proměnné X, Y – OK – Spearmanův koef. R. Dostaneme tabulku

	Spearmanovy korelace (dva lekari.sta) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	7	0,857143	3,721042	0,013697

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,857, testová statistika se realizuje hodnotou 3,721, odpovídající p-hodnota je 0,0137, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení dvou lékařů ve prospěch oboustranné alternativy.

Příklad na testování pořadové nezávislosti (pořadí musíme stanovit):
Jsou dány realizace náhodného výběru z dvourozměrného rozložení, kterým se řídí náhodný vektor (X,Y): (2,5 13,4), (3,4 15,2), (1,3 11,8), (5,8 13,1), (3,6 14,5). Na hladině významnosti 0,05 testujte hypotézu, že náhodné veličiny jsou pořadově nezávislé proti oboustranné alternativě.

Řešení:

x_i	2,5	3,4	1,3	5,8	3,6
y_i	13,4	15,2	11,8	13,1	14,5
R_i	2	3	1	5	4
Q_i	3	5	1	2	4
$(R_i-Q_i)^2$	1	4	0	9	0

Testová statistika: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1} (R_i - Q_i)^2 = 1 - \frac{6}{5 \cdot 24} 14 = 0{,}3$

Kritická hodnota: pro n = 5 a α = 0,05 je kritická hodnota 0,9. Protože testová statistika se realizuje hodnotou 0,3, hypotézu o pořadové nezávislosti veličin X a Y nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA
Postupujeme úplně stejně jako v předešlém případě. Výstupní tabulka má tvar:

	Spearmanovy korelace (poradova korelace.sta) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	Úroveň p
X & Y	5	0,300000	0,544705	0,623838

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,3, testová statistika se realizuje hodnotou 0,5447, odpovídající p-hodnota je 0,6238, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o pořadové nezávislosti veličin X, Y.

Pearsonův koeficient korelace

Karl Pearson (1857 – 1936): Britský statistik

Číslo

$$R(X,Y) = \begin{cases} E\left\{\left(\frac{X - E(X)}{\sqrt{D(X)}} \cdot \frac{Y - E(Y)}{\sqrt{D(Y)}}\right)\right\} = \frac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} & \text{pro } \sqrt{D(X)}\sqrt{D(Y)} > 0 \\ 0 \text{ jinak} \end{cases}$$

se nazývá Pearsonův koeficient korelace.
(Pro výpočet Pearsonova koeficentu korelace musíme znát simultánní distribuční funkci $\Phi(x,y)$ v obecném případě resp. simultánní hustotu pravděpodobnosti $\varphi(x,y)$ ve spojitém případě resp. simultánní pravděpodobnostní funkci $\pi(x,y)$ v diskrétním případě.)

Vlastnosti Pearsonova koeficientu korelace

a) $R(a_1, Y) = R(X, a_2) = R(a_1, a_2) = 0$

b) $R(a_1 + b_1X, a_2 + b_2Y) = \operatorname{sgn}(b_1b_2)\, R(X, Y) = \begin{cases} R(X, Y) & \text{pro } b_1b_2 > 0 \\ -R(X, Y) & \text{pro } b_1b_2 < 0 \end{cases}$

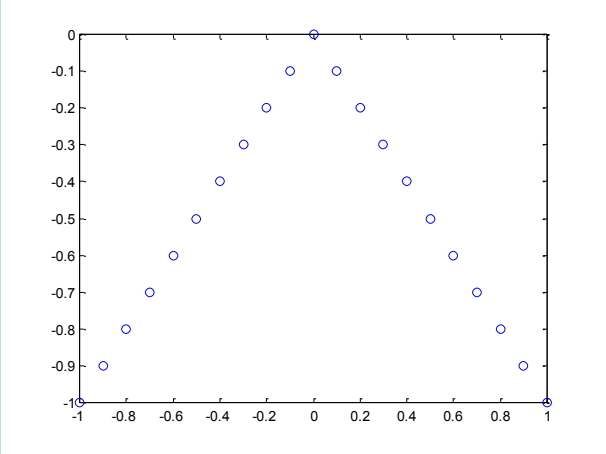
c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

d) $R(X, Y) = R(Y, X)$

e) $|R(X, Y)| \leq 1$ a rovnost nastane tehdy a jen tehdy, když mezi veličinami X, Y existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a, b tak, že pravděpodobnost $P(Y = a + bX) = 1$. Přitom $R(X, Y) = 1$, když $b > 0$ a $R(X, Y) = -1$, když $b < 0$. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)

Z vlastností Pearsonova koeficientu korelace vyplývá, že se hodí pouze k měření těsnosti lineárního vztahu veličin X a Y. Při složitějších závislostech může dojít k paradoxní situaci, že Pearsonův koeficient korelace je nulový.

Ilustrace:



Definice nekorelovanosti

Je-li $R(X, Y) = 0$, pak řekneme, že náhodné veličiny jsou nekorelované. (Znamená to, že mezi X a Y neexistuje žádná lineární závislost. Jsou-li náhodné veličiny X,Y stochasticky nezávislé, pak jsou samozřejmě i nekorelované.)

Je-li $R(X, Y) > 0$, pak řekneme, že náhodné veličiny jsou kladně korelované. (Znamená to, že s růstem hodnot veličiny X rostou hodnoty veličiny Y a s poklesem hodnot veličiny X klesají hodnoty veličiny Y.)

Je-li $R(X, Y) < 0$, pak řekneme, že náhodné veličiny jsou záporně korelované. (Znamená to, že s růstem hodnot veličiny X klesají hodnoty veličiny Y a s poklesem hodnot veličiny X rostou hodnoty veličiny Y.)

Výběrový koeficient korelace

Nechť $(X_1, Y_1), ..., (X_n, Y_n)$ náhodný výběr rozsahu n z dvourozměrného rozložení daného distribuční funkcí $\Phi(x,y)$.
Z tohoto dvourozměrného náhodného výběru můžeme stanovit:

výběrové průměry $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$, $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$,

výběrové rozptyly $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - M_1)^2$, $S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - M_2)^2$,

výběrovou kovarianci $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - M_1)(Y_i - M_2)$ a s jejich pomocí zavedeme

výběrový koeficient korelace $R_{12} = \begin{cases} \frac{1}{n-1}\sum_{i=1}^{n} \frac{X - M_1}{S_1} \cdot \frac{Y - M_2}{S_2} = \frac{S_{12}}{S_1 S_2} & \text{pro } S_1 S_2 > 0 \\ 0 & \text{jinak} \end{cases}$. Vlastnosti Pearsonova koeficientu korelace se přenášejí i na výběrový koeficient korelace.

(Spearmanův koeficient pořadové korelace odpovídá Pearsonovu koeficientu korelace aplikovanému na pořadí.)

Pearsonův koeficient korelace dvourozměrného normálního rozložení

Jak bylo uvedeno v motivaci, korelační analýza předpokládá, že daný náhodný výběr pochází z dvourozměrného normálního rozložení. Proč je tento předpoklad tak důležitý? Odpověď poskytne následující věta.

Nechť náhodný vektor (X, Y) má dvourozměrné normální rozložení s hustotou

$$\varphi(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]}$$

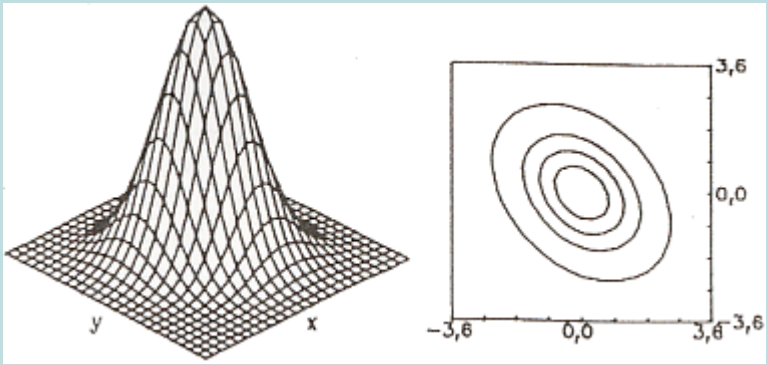
, přičemž $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = D(X)$, $\sigma_2^2 = D(Y)$, $\rho = R(X,Y)$.

Marginální hustoty jsou: $\varphi_1(x) = \int_{-\infty}^{\infty} \varphi(x,y)\,dy = \ldots = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$, $\varphi_2(y) = \int_{-\infty}^{\infty} \varphi(x,y)\,dx = \ldots = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$.

Je-li $\rho = 0$, pak pro $\forall [x,y] \in \mathbb{R}^2 : \varphi(x,y) = \varphi_1(x)\varphi_2(y)$, tedy náhodné veličiny X, Y jsou stochasticky nezávislé. Jinými slovy: stochastická nezávislost složek X, Y normálně rozloženého náhodného vektoru je ekvivalentní jejich nekorelovanosti. Pro jiná dvourozměrná rozložení to neplatí!

Upozornění: nadále budeme předpokládat, že $(X_1, Y_1), ..., (X_n, Y_n)$ je náhodný výběr rozsahu n z dvourozměrného normálního rozložení $N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$.

Předpoklad dvourozměrné normality lze orientačně ověřit pomocí dvourozměrného tečkového diagramu: tečky by měly zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy:
Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0{,}75$:



Do dvourozměrného tečkového diagramu můžeme ještě zakreslit $100(1-\alpha)\%$ elipsu konstantní hustoty pravděpodobnosti. Bude-li více než $100\alpha\%$ teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi veličinami X a Y existuje určitý stupeň přímé resp. nepřímé lineární závislosti.

Testování hypotézy o nezávislosti

Na hladině významnosti α testujeme H_0: X, Y jsou stochasticky nezávislé náhodné veličiny (tj. $\rho = 0$) proti
- oboustranné alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $\rho \neq 0$)
- levostranné alternativě H_1: X, Y jsou záporně korelované náhodné veličiny (tj. $\rho < 0$)
- pravostranné alternativě H_1: X, Y jsou kladně korelované náhodné veličiny (tj. $\rho > 0$).

Testová statistika má tvar: $T_0 = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$.

Platí-li nulová hypotéza, pak $T_0 \sim t(n-2)$.

Kritický obor pro test H_0 proti
- oboustranné alternativě: $W = (-\infty, -t_{1-\alpha/2}(n-2)) \cup (t_{1-\alpha/2}(n-2), \infty)$,
- levostranné alternativě: $W = (-\infty, -t_{1-\alpha}(n-2))$,
- pravostranné alternativě: $W = (t_{1-\alpha}(n-2), \infty)$.

H_0 zamítáme na hladině významnosti α, když $t_0 \in W$.

Příklad: Testování hypotézy o nezávislosti proti oboustranné alternativě

V dílně pracuje 15 dělníků. Byl u nich zjištěn počet směn odpracovaných za měsíc (náhodná veličina X) a počet zhotovených výrobků (náhodná veličina Y):

X 20 21 18 17 20 18 19 21 20 14 16 19 21 15 15
Y 92 93 83 80 91 85 82 98 90 60 73 86 96 64 81.

Předpokládejte, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y a na hladině 0,01 testujte hypotézu o nezávislosti X a Y proti oboustranné alternativě.

Řešení:

Vypočteme realizace

výběrových průměrů: $m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = 18{,}267$, $m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = 83{,}6$,

výběrových rozptylů: $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m_1)^2 = 5{,}6381$, $s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - m_2)^2 = 121{,}4$,

výběrové kovariance: $s_{12} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m_1)(y_i - m_2) = 24{,}2571$,

výběrového koeficientu korelace: $r_{12} = \frac{s_{12}}{s_1 s_2} = 0{,}927$.

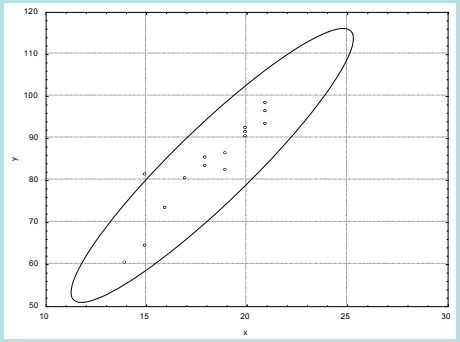
Realizace testové statistiky: $t_0 = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} = 8{,}912$,

kritický obor $W = (-\infty, -t_{0,995}(13)) \cup (t_{0,995}(13), \infty) = (-\infty, -3{,}012) \cup (3{,}012, \infty)$.

Protože $t_0 \in W$, hypotézu o nezávislosti veličin X a Y zamítáme na hladině významnosti 0,01. S rizikem omylu nejvýše 1% jsme tedy prokázali, že mezi počtem směn odpracovaných za měsíc a počtem zhotovených výrobků existuje závislost.

Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor o dvou proměnných X, Y a 15 případech. Dvourozměrnou normalitu dat ověříme pomocí dvourozměrného tečkového diagramu: Grafy – Bodové grafy – Proměnné X, Y – OK – odškrtneme Typ proložení Lineární – na záložce Detaily zaškrtneme Elipsa Normální - OK.



Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Prom. X & prom. Y	Korelace (smeny a vyrobky.sta) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD) Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	18,26667	2,37447									
X	18,26667	2,37447	1,000000	1,000000			15	0,000000	1,000000	0,000000	1,000000
X	18,26667	2,37447									
Y	83,60000	11,01817	0,927180	0,859663	8,923795	0,000001	15	5,010135	4,302365	1,562407	0,199812
Y	83,60000	11,01817									
X	18,26667	2,37447	0,927180	0,859663	8,923795	0,000001	15	1,562407	0,199812	5,010135	4,302365
Y	83,60000	11,01817									
Y	83,60000	11,01817	1,000000	1,000000			15	0,000000	1,000000	0,000000	1,000000

Výběrový koeficient korelace se realizoval hodnotou 0,92718, testová statistika nabyla hodnoty 8,924, odpovídající p-hodnota je 0,000001, tedy na hladině významnosti 0,01 zamítáme hypotézu o nezávislosti veličin X, Y.