Jednofaktorová analýza rozptylu

David Hampel

Ústav statistiky a operačního výzkumu,
Mendelova univerzita v Brně

Kurz pokročilých statistických metod
Global Change Research Centre AS CR, 5.–7. 8. 2015

Tato akce se koná v rámci projektu: Vybudování vědeckého týmu environmentální metabolomiky a ekofyziologie a jeho zapojení do mezinárodních sítí (ENVIMET; r.č. **CZ.1.07/2.3.00/20.0246**) realizovaného v rámci Operačního programu Vzdělávání pro konkurenceschopnost

Obsah

Motivace

- V určitých případech je máme za úkol rozhodnout o rovnosti tří a více středních hodnot.
- Pro analýzu je třeba splnit určité předpoklady.
- Nakonec je žádoucí zjistit, které konkrétní střední hodnoty se od sebe liší.

Předpoklady a označení

- Jednofaktorová analýza rozptylu – též analýza rozptylu jednoduchého třídění (ONEWAY, speciální případ vícefaktorové analýzy rozptylu ANOVA) zkoumá závislost intervalové či poměrové proměnné X na nominální proměnné A, která má aspoň dvě varianty.
- Proměnná A se nazývá faktor a její varianty úrovně faktoru.
- Závislost X na A se projeví tím, že existuje statisticky významný rozdíl v průměrech proměnné X v náhodných výběrech, které vznikly tříděním podle variant proměnné A.

Předpoklady a označení

- Metodu ANOVA odvodil R. A. Fisher ve 30. letech 20. století.
- Její podstata spočívá v tom, že celkový rozptyl sledované proměnné X se rozloží na rozptyl uvnitř jednotlivých výběrů a na rozptyl mezi výběry.
- Pokud je rozptyl mezi výběry nepravděpodobně velký, svědčí to o významném vlivu faktoru A.

Předpoklady a označení

- Předpokládáme, že faktor A má $r \geq 2$ úrovní $A_1, \ldots, A_r$, přičemž i-té úrovni odpovídá n_i pozorování $X_{i1}, \ldots, X_{in_i}$, která tvoří náhodný výběr z normálního rozložení $N\left(\mu_i, \sigma^2\right)$, $i = 1, \ldots, r$.
- Celkový počet pozorování je $n = \sum_{i=1}^{r} n_i$.
- Jednotlivé náhodné výběry jsou stochasticky nezávislé.

Předpoklady a označení

Pomocí tečkové notace označujeme součet hodnot v i-tém výběru

$$X_{i.} = \sum_{j=1}^{n_i} X_{ij},$$

výběrový průměr v i-tém výběru

$$M_{i.} = \frac{1}{n_i} X_{i.},$$

součet hodnot všech výběrů

$$X_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij}$$

a celkový průměr všech r výběrů

$$M_{..} = \frac{1}{n} X_{..}.$$

Matematický model

Náhodné veličiny X_{ij} se řídí modelem

$$X_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{pro} \quad i = 1, \ldots, r,\ j = 1, \ldots, n_i,$$

- přičemž μ je společná část střední hodnoty závisle proměnné veličiny X,
- α_i je efekt faktoru A na úrovni A_i,
- ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozdělením $N\left(0, \sigma^2\right)$.

Matematický model

- Parametry μ, α_i neznáme.
- Požadujeme, aby platila tzv. reparametrizační rovnice

$$\sum_{i=1}^{r} n_i \alpha_i = 0.$$

- Pokud je třídění vyvážené, tj. pokud mají všechny výběry stejný rozsah $n_1 = n_2 = \ldots = n_r$, pak lze použít zjednodušenou podmínku

$$\sum_{i=1}^{r} \alpha_i = 0.$$

Matematický model

$$\begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1n_1} \\ \hline X_{21} \\ X_{22} \\ \vdots \\ X_{2n_2} \\ \hline \vdots \\ \hline X_{r1} \\ X_{r2} \\ \vdots \\ X_{rn_r} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & & \vdots \\ \hline 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_1} \\ \hline \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2n_2} \\ \hline \vdots \\ \hline \varepsilon_{r1} \\ \varepsilon_{r2} \\ \vdots \\ \varepsilon_{rn_r} \end{bmatrix}$$

Matematický model

$$\begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1n_1} \\ \hline X_{21} \\ X_{22} \\ \vdots \\ X_{2n_2} \\ \hline \vdots \\ \hline X_{r1} \\ X_{r2} \\ \vdots \\ X_{rn_r} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_1} \\ \hline \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2n_2} \\ \hline \vdots \\ \hline \varepsilon_{r1} \\ \varepsilon_{r2} \\ \vdots \\ \varepsilon_{rn_r} \end{bmatrix}$$

Matematický model

Zavedeme celkový součet čtverců

$$S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{..})^2 ,$$

který charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru, má počet stupňů volnosti $f_T = n - 1$, dále skupinový součet čtverců

$$S_A = \sum_{i=1}^{r} n_i (M_{i.} - M_{..})^2 ,$$

jež charakterizuje variabilitu mezi jednotlivými náhodnými výběry, má počet stupňů volnosti $f_A = r - 1$,

Matematický model

a nakonec reziduální součet čtverců

$$S_E = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{i.})^2 ,$$

který charakterizuje variabilitu uvnitř jednotlivých výběrů, má počet stupňů volnosti $f_E = n - r$.

Lze dokázat, že $S_T = S_A + S_E$.

Testování hypotézy o shodě středních hodnot

- 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.

- Tato hypotéza vlastně říká, že vliv faktoru A na proměnnou X je nevýznamný.

Testování hypotézy o shodě středních hodnot

- Testová statistika

$$F_A = \frac{S_A/f_A}{S_E/f_E}$$

se řídí rozložením $F\left(r-1, n-r\right)$, je-li H_0 pravdivá.

- Nulovou hypotézu tedy zamítneme na hladině významnosti α, když F_A se bude realizovat v kritickém oboru

$$W = \langle F_{1-\alpha}\left(r-1, n-r\right), \infty\rangle .$$

Testování hypotézy o shodě středních hodnot

Výsledky výpočtů zapisujeme do tabulky ANOVA:

Zdroj variability	součet čtverců	stupně volnosti	prům. čtverec	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	$f_T = n - 1$	–	–

Příklad – zadání

- Je dáno pět nezávislých náhodných výběrů o rozsazích 5, 7, 6, 8, 5, přičemž i-tý výběr pochází z rozložení $N(\mu_i, \sigma^2)$, $i = 1, \ldots, 5$.
- Byl vypočten celkový součet čtverců $S_T = 15$ a reziduální součet čtverců $S_E = 3$.
- Na hladině významnosti 0,05 testujte hypotézu o shodě středních hodnot.

Příklad – řešení

- Počet výběrů je $r = 5$, celkový rozsah všech pěti výběrů je $n = 5 + 7 + 6 + 8 + 5 = 31$.
- Vypočteme skupinový součet čtverců: $S_A = S_T - S_E = 15 - 3 = 12$.
- Testovou statistiku získáme jako

$$F_A = \frac{S_A/(r-1)}{S_E/(n-r)} = \frac{12/4}{3/26} = 26.$$

- Kritický obor je $W = (F_{0,95}(4, 26), \infty\rangle = (2{,}7426, \infty\rangle$.
- Protože se testová statistika realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,05.

Příklad – řešení

Tabulka ANOVA:

Zdroj variability	součet čtverců	stupně volnosti	průměrný čtverec	F_A
skupiny	$S_A = 12$	$f_A = r - 1 = 4$	$S_A/f_A = 3$	$\frac{S_A/f_A}{S_E/f_E} = 26$
reziduální	$S_E = 3$	$f_E = n - r = 26$	$S_E/f_E = 3/26$	–
celkový	$S_T = 15$	$f_T = n - 1 = 30$	–	–

Testování hypotézy o shodě rozptylů

Před provedením analýzy rozptylu je zapotřebí ověřit předpoklad o shodě rozptylů v daných r výběrech, tedy je nutné provést test nulové hypotézy

$$H_0 : \sigma_1^2 = \ldots = \sigma_r^2$$

proti alternativní hypotéze H_1, která tvrdí, že aspoň jedna dvojice rozptylů se liší.

Testování hypotézy o shodě rozptylů

Levenův test.

- Položme $Z_{ij} = |X_{ij} - M_{i.}|$. Označíme

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

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

Testování hypotézy o shodě rozptylů

- Platí-li hypotéza o shodě rozptylů, pak statistika

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

  se asymptoticky řídí rozložením $F(r-1, n-r)$.

- Hypotézu o shodě rozptylů tedy zamítáme na asymptotické hladině významnosti α, když testová statistika $F_{ZA} \in W$, kde $W = \langle F_{1-\alpha}(r-1, n-r), \infty)$ je kritický obor.

- Levenův test je vlastně založen na analýze rozptylu absolutních hodnot centrovaných pozorování.

Testování hypotézy o shodě rozptylů

- 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.

Testování hypotézy o shodě rozptylů

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 S_i^2, $i = 1, \ldots, r$.

- Hypotézu o shodě rozptylů zamítáme na asymptotické hladině významnosti α, když se B realizuje v kritickém oboru $W = \langle \chi^2_{1-\alpha}(r-1), \infty)$.

Post-hoc (následné) 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.
- Existuje celá řada post-hoc testů, k nejznámějším patří metoda Tukeyova, Scheffého, Duncanova, Fisherova LSD, Newmanova-Keulsova a další.
- Každá z těchto metod má svoje přednosti a nedostatky a žádná není všeobecně přijímána jako ideální.
- Zde stručně popíšeme Tukeyovu a Scheffého metodu.

Post-hoc (následné) metody mnohonásobného porovnávání

Tukeyova metoda.

- Mají-li všechny výběry týž rozsah p, pak rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$$|M_{k.} - M_{l.}| \geq q_{1-\alpha}(r, n-r) \frac{S_*}{\sqrt{p}},$$

  kde kvantily $q_{1-\alpha}(r, n-r)$ studentizovaného rozpětí najdeme ve statistických tabulkách.

- Existuje modifikace Tukeyovy metody pro nestejné rozsahy výběrů, Tukey HSD.

Post-hoc (následné) metody mnohonásobného porovnávání

Scheffého metoda.

- Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

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

Post-hoc (následné) metody mnohonásobného porovnávání

- Metody mnohonásobného porovnávání mají obecně menší sílu než ANOVA.
- Může proto 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) Je zapotřebí ověřit, že jednotlivé náhodné výběry pocházejí z normálních rozdělení.

- Můžeme použít grafickou metodu (např. N-P plot, Q-Q plot, histogram)
- nebo testy hypotéz o normálním rozložení (např. Lilieforsovu variantu Kolmogorovova-Smirnovova testu nebo Shapirův-Wilkův test).
- Doporučuje se kombinace obou způsobů.

Doporučený postup při provádění analýzy rozptylu

a) Je zapotřebí ověřit, že jednotlivé náhodné výběry pocházejí z normálních rozdělení.

- Obecně lze říci, že analýza rozptylu není příliš citlivá na porušení předpokladu normality, zvláště při větších rozsazích výběrů (nad 20).
- Mírné porušení normality tedy není na závadu, při větším porušení použijeme např. Kruskalův-Wallisův test jako neparametrickou obdobu analýzy rozptylu jednoduchého třídění.

Doporučený postup při provádění analýzy rozptylu

b) Po ověření normality musíme testovat homogenitu rozptylů.

- Graficky ověřujeme shodu rozptylů pomocí krabicových diagramů, kdy sledujeme, zda je šířka krabic přibližně stejná.
- Numericky testujeme homogenitu rozptylů pomocí Levenova testu, Brownova-Forsytheova testu či Bartlettova testu.
- Slabé porušení homogenity rozptylů nevadí, při větším se doporučuje mediánový test.

Doporučený postup při provádění analýzy rozptylu

c) Pokud jsou splněny předpoklady normality a homogenity rozptylů, můžeme přistoupit k testování shody všech středních hodnot.

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 – zadání

- Zjišťujeme možnou závislost výnosu na obsahu určité látky v půdě.
- K dispozici je 68 pozorování.
- Obsah sledované látky je rozdělen do 6 skupin (18–24, 25–31, . . . , 53–59) – proměnná A.
- Výnos je proměnná X (viz tabulka níže).

Příklad – zadání

A	X	A	X	A	X	A	X
18–24	27	39–45	34	18–24	37	25–31	29
18–24	31	39–45	34	18–24	34	46–52	28
18–24	39	39–45	43	53–59	25	53–59	26
18–24	38	39–45	44	18–24	32	39–45	31
18–24	39	39–45	40	25–31	28	39–45	28
25–31	37	39–45	47	25–31	30	46–52	26
25–31	35	39–45	45	39–45	34	53–59	28
25–31	40	46–52	35	46–52	28	18–24	28
25–31	40	46–52	34	46–52	31	18–24	27
25–31	31	46–52	34	18–24	30	46–52	26
32–38	34	46–52	41	25–31	29	46–52	27
32–38	36	53–59	35	32–38	32	46–52	21
32–38	34	53–59	37	53–59	34	53–59	29
32–38	41	25–31	25	53–59	28	25–31	25
32–38	30	32–38	32	18–24	26	46–52	20
32–38	44	32–38	31	18–24	27	53–59	18
32–38	44	53–59	28	25–31	24	53–59	26

Příklad – zadání

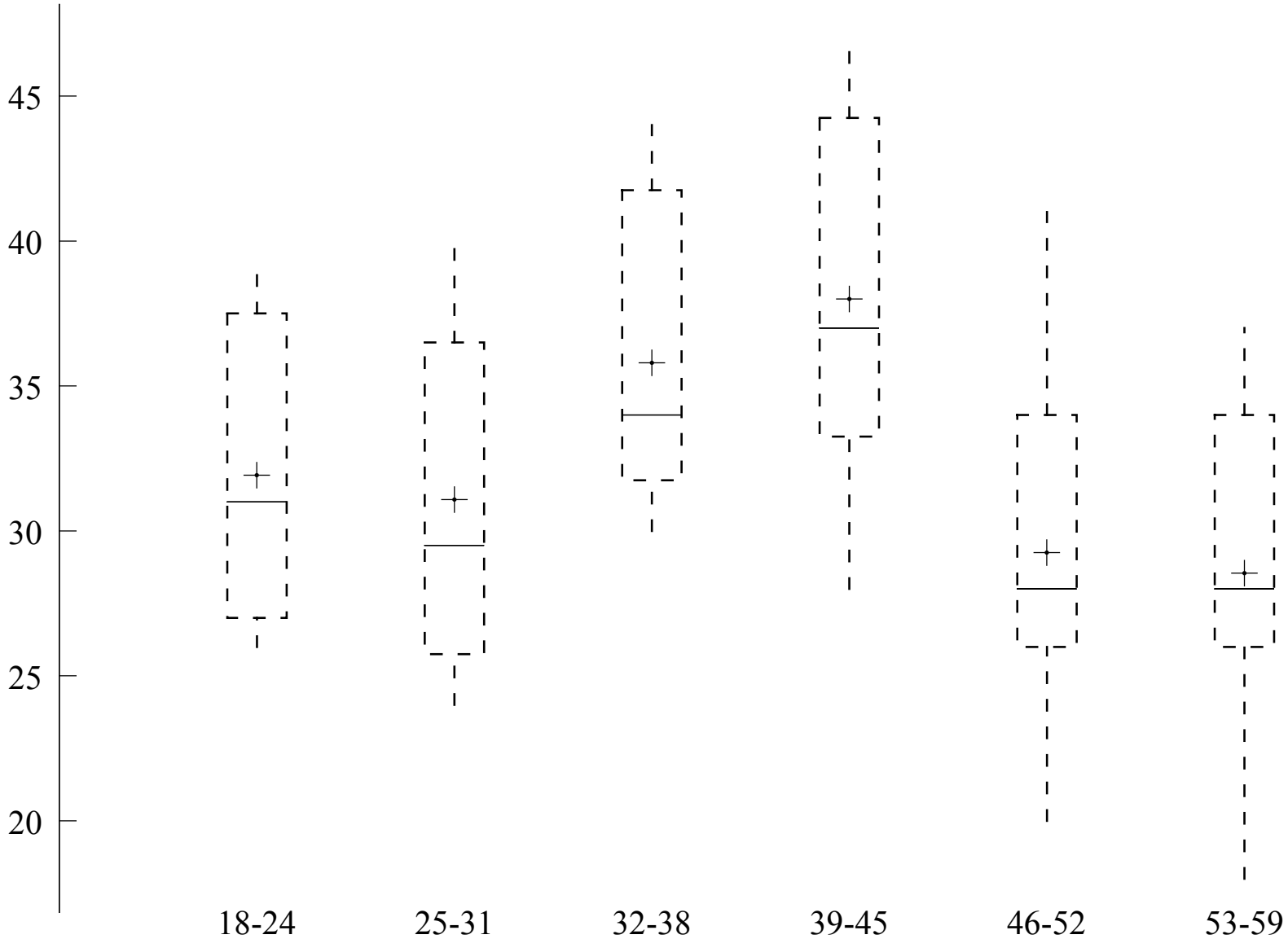
- Na hladině významnosti 0,05 máme testovat hypotézu, že rozdíly ve výnosu jsou způsobeny pouze náhodnými vlivy.
- V případě zamítnutí nulové hypotézy je třeba identifikovat, které dvojice skupin se liší na hladině významnosti 0,05.

Příklad – charakteristiky dat

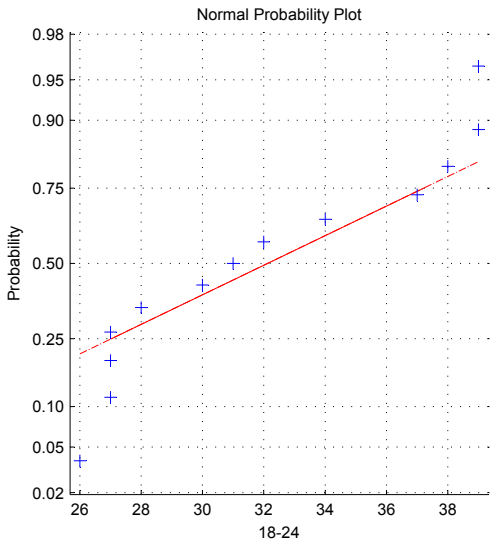
Skupina	Průměr	Medián	Minimum	Maximum
18–24	31,9231	31,0000	26,0000	39,0000
25–31	31,0833	29,5000	24,0000	40,0000
32–38	35,8000	34,0000	30,0000	44,0000
39–45	38,0000	37,0000	28,0000	47,0000
46–52	29,2500	28,0000	20,0000	41,0000
53–59	28,5455	28,0000	18,0000	37,0000

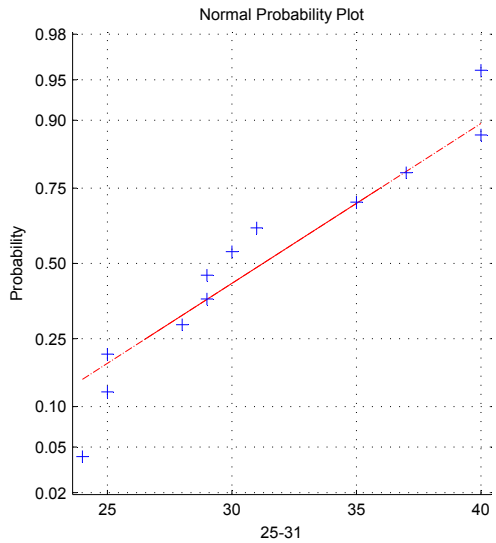
Skupina	Odchylka	C.V.	Šikmost	Četnost
18–24	4,95751	0,155296	0,312563	13,0
25–31	5,66422	0,182227	0,434070	12,0
32–38	5,30827	0,148276	0,639137	10,0
39–45	6,59966	0,173675	−0,0464489	10,0
46–52	6,04716	0,206741	0,255538	12,0
53–59	5,29837	0,185612	−0,151714	11,0

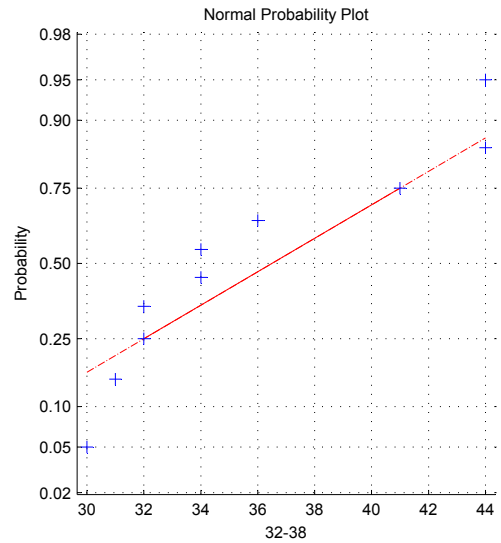
Příklad – charakteristiky dat

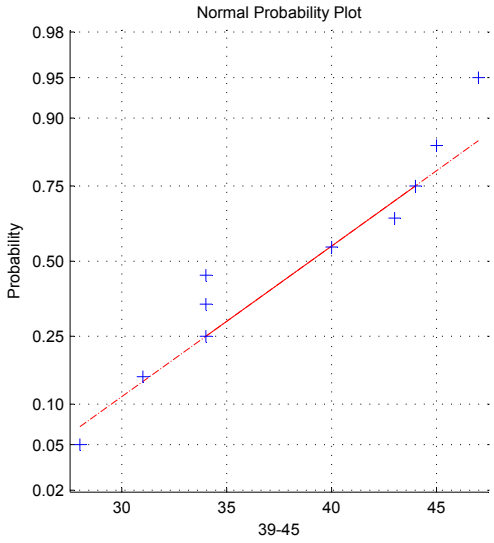


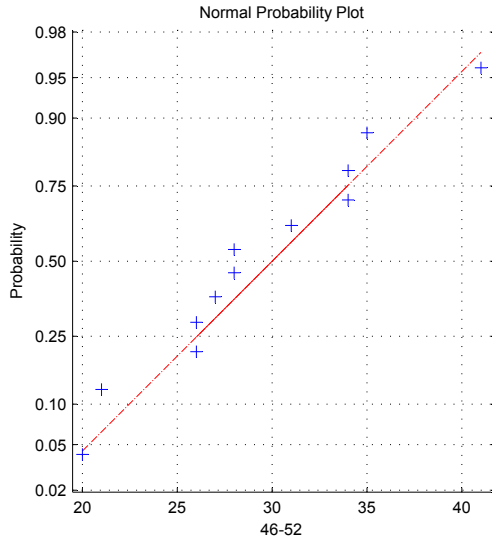
Příklad – normalita dat

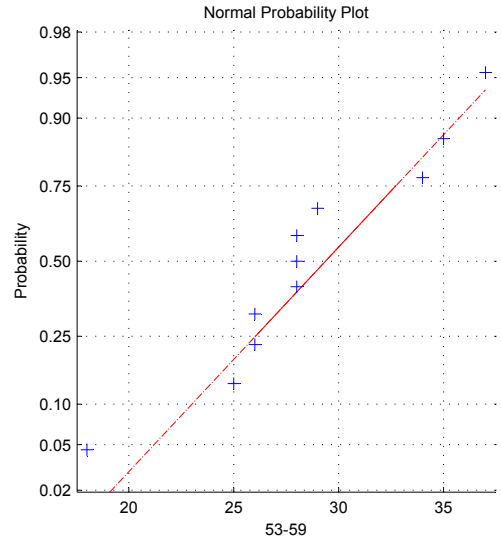












Příklad – normalita dat

- Pro posouzení normality dále použijeme Shapirův-Wilkův test.

Skupina	18–24	25–31	32–38	39–45	46–52	53–59
p-hodnota	0,0653	0,1941	0,0674	0,3608	0,8088	0,4538

Příklad – shoda rozptylů

- Nyní se zaměříme na ověření předpokladu o homogenitě rozptylů, tj. na hladině významnosti 0,05 testujeme hypotézu

$$H_0 : \sigma_1^2 = \ldots = \sigma_r^2$$

proti alternativní hypotéze H_1, která tvrdí, že aspoň jedna dvojice rozptylů se liší.

- Použijeme Levenův test.

Příklad – shoda rozptylů

- Skupinový součet čtverců S_{ZA} nabývá hodnoty 28,79949,
- skupinový počet stupňů volnosti f_{ZA} je 5,
- reziduální součet čtverců S_{ZE} je 729,6711,
- reziduální počet stupňů volnosti f_{ZE} je 62,
- testová statistika F_{ZA} se realizuje hodnotou 0,489417,
- odpovídající p-hodnota je 0,78290,
- tedy na hladině významnosti 0,05 nelze zamítnout hypotézu o homogenitě rozptylů.

Příklad – shoda středních hodnot

- Dále se budeme věnovat testování hypotézy o shodě středních hodnot normálních rozložení, z nichž pocházejí sledované náhodné výběry, tj. testujeme hypotézu

$$H_0 : \mu_1 = \ldots = \mu_r$$

proti alternativní hypotéze H_1, která tvrdí, že aspoň jedna dvojice středních hodnot se liší.

Příklad – shoda středních hodnot

- Vidíme, že skupinový součet čtverců S_A je 733,2742,
- skupinový počet stupňů volnosti f_A je 5,
- reziduální součet čtverců S_E je 1976,417,
- reziduální počet stupňů volnosti f_E je 62,
- testová statistika F_A se realizuje hodnotou 4,600547,
- odpovídající p-hodnota je 0,001239.

Příklad – shoda středních hodnot

- Na hladině významnosti 0,05 (a dokonce i na hladině významnosti 0,01) zamítáme hypotézu o rovnosti středních hodnot.
- Znamená to, že s rizikem omylu nejvýše 0,05 jsme prokázali, že střední hodnoty výnosu se pro různé skupiny obsahu sledované látky liší.

Příklad – párová porovnání

- Vzhledem k tomu, že na hladině významnosti 0,05 jsme zamítli hypotézu o shodě středních hodnot, provedeme mnohonásobné porovnávání, abychom identifikovali, které dvojice náhodných výběrů přispěly k zamítnutí nulové hypotézy.
- Výsledek Scheffého metody ukazuje, že na hladině významnosti 0,05 se liší skupiny 39–45 a 46–52 a dále 39–45 a 53–59.
- Rozdíly mezi ostatními dvojicemi skupin nejsou prokazatelné na hladině významnosti 0,05.

ANOVA – příklad

- `data_ANOVA_nekomplet.sta`