10. Porovnání empirického a teoretického rozložení

10.1. Motivace

Možnost použití statistických testů je podmíněna nějakými předpoklady o datech. Velmi často je to předpoklad o typu rozložení, z něhož získaná data pocházejí. Mnoho testů je založeno na předpokladu normality.

Opomíjení předpokladů o typu rozložení může v praxi vést i ke zcela zavádějícím výsledkům, proto je nutné věnovat tomuto problému patřičnou pozornost.

10.2. Kolmogorovův – Smirnovův test

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika $D_n = \sup_{-\infty<x<\infty} |F_n(x) - \Phi(x)|$. Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota. Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem $\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}$.

Upozornění: Nulová hypotéza musí specifikovat distribuční funkci zcela přesně, včetně všech jejích případných parametrů. Např. K-S test lze použít pro testování hypotézy, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení $Rs(0,1)$, což se využívá při testování generátorů náhodných čísel. Pokud však parametry distribuční funkce odhadujeme z výběru, změní se rozložení testové statistiky D_n. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

Poznámka ke K-S testu ve STATISTICE: Test normality poskytuje hodnotu testové statistiky (ozn. d) a dvě p-hodnoty. První se vztahuje k případu, kdy μ a σ^2 známe předem, druhá (ozn. Liliefors p) se vztahuje k případu, kdy μ a σ^2 neznáme. Objeví-li se ve výstupu p = n.s. (tj. non significant), pak hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Příklad: Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K- S testu zjistěte na hladině významnosti 0,05, zda tato data pocházejí z normálního rozložení.

Řešení: Odhadem střední hodnoty je výběrový průměr m = 11, odhadem rozptylu je výběrový rozptyl $s^2 = 10$. Uspořádaný náhodný výběr je (8, 9, 10, 12, 16). Vypočteme hodnoty výběrové distribuční funkce:

$$x < 8 : F_5(x) = 0$$

$$8 \leq x < 9 : F_5(x) = \frac{1}{5} = 0{,}2$$

$$9 \leq x < 10 : F_5(x) = \frac{2}{5} = 0{,}4$$

$$10 \leq x < 12 : F_5(x) = \frac{3}{5} = 0{,}6$$

$$12 \leq x < 16 : F_5(x) = \frac{4}{5} = 0{,}8$$

$$x \geq 16 : F_5(x) = 1$$

Hodnoty teoretické distribuční funkce $\Phi_T(x)$ v bodech 8, 9, 10, 12, 16:

$$\Phi_T(8) = \Phi\left(\frac{8-11}{\sqrt{10}}\right) = \Phi(-0{,}95) = 1 - \Phi(0{,}95) = 1 - 0{,}82894 = 0{,}17106$$

$$\Phi_T(9) = \Phi\left(\frac{9-11}{\sqrt{10}}\right) = \Phi(-0{,}63) = 1 - \Phi(0{,}63) = 1 - 0{,}73565 = 0{,}26435$$

$$\Phi_T(10) = \Phi\left(\frac{10-11}{\sqrt{10}}\right) = \Phi(-0{,}32) = 1 - \Phi(0{,}32) = 1 - 0{,}62552 = 0{,}37448$$

$$\Phi_T(12) = \Phi\left(\frac{12-11}{\sqrt{10}}\right) = \Phi(0{,}32) = 0{,}62552$$

$$\Phi_T(16) = \Phi\left(\frac{16-11}{\sqrt{10}}\right) = \Phi(1{,}58) = 0{,}94295$$

(Φ je distribuční funkce rozložení N(0,1).)

Rozdíly mezi výběrovou distribuční funkcí $F_5(x)$ a teoretickou distribuční funkcí $\Phi_T(x)$:
$d_1 = 0{,}2 - 0{,}17106 = 0{,}02894$; $d_2 = 0{,}4 - 0{,}26435 = 0{,}13565$; $d_3 = 0{,}6 - 0{,}37448 = 0{,}22552$;
$d_4 = 0{,}8 - 0{,}62552 = 0{,}17448$; $d_5 = 1 - 0{,}94295 = 0{,}05705$.
Testová statistika: $D_5 = 0{,}22552$, modifikovaná kritická hodnota pro $n = 5$, $\alpha = 0{,}05$ je 0,343.
Protože $0{,}22552 < 0{,}343$, hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

10.3. Shapirův – Wilkův test normality

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení $N(\mu, \sigma^2)$. Test je založen na zjištění, zda body v Q-Q grafu jsou významně odlišné od regresní přímky proložené těmito body. S-W test se používá především pro výběry menších rozsahů, $n < 50$.

10.4. Testy dobré shody

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$.

a) Je-li distribuční funkce spojitá, pak data rozdělíme do r třídicích intervalů $\langle u_j, u_{j+1} \rangle$, $j = 1, \ldots, r$. Zjistíme absolutní četnost n_j j-tého třídicího intervalu a vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.
b) Má-li distribuční funkce nejvýše spočetně mnoho bodů nespojitosti, pak místo třídicích intervalů použijeme varianty $x_{[j]}$, $j = 1, \ldots, r$. Pro variantu $x_{[j]}$ zjistíme absolutní četnost n_j a vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat variantou $x_{[j]}$. Platí-li nulová hypotéza, pak $p_j = \Phi(x_{[j]}) - \lim_{x \to x_{[j]}-} \Phi(x) = P(X = x_{[j]})$.

Testová statistika: $K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}$. Platí-li nulová hypotéza, pak $K \approx \chi^2(r\text{-}1\text{-}p)$, kde p je počet odhadovaných parametrů daného rozložení. (Např. pro normální rozložení $p = 2$, protože z dat odhadujeme střední hodnotu a rozptyl.) Nulovou hypotézu zamítáme na

asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}(r-1-p)$. Aproximace se považuje za vyhovující, když $np_j \geq 5$, j = 1, ..., r.

Upozornění: Hodnota testové statistiky K je silně závislá na volbě třídicích intervalů. Navíc při nesplnění podmínky $np_j \geq 5$, j = 1, ..., r je třeba některé intervaly resp. varianty slučovat, což vede ke ztrátě informace.

Příklad: Byl zjišťován počet poruch určitého zařízení za 100 hodin provozu ve 150 disjunktních 100 h intervalech. Výsledky měření:

Počet poruch za 100 hodin provozu	0	1	2	3	4 a víc
Absolutní četnost	52	48	36	10	4

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že náhodný výběr $X_1, ..., X_{150}$ pochází z rozložení Po(1,2).

Řešení:
Pravděpodobnost, že náhodná veličina s rozložením Po(λ), kde λ = 1,2 bude nabývat hodnot $p_0, ..., p_4$ a víc je $p_j = \frac{\lambda^j}{j!}e^{-\lambda} = \frac{1,2^j}{j!}e^{-1,2}$, $j = 0,1,2,3$, $p_4 = 1 - (p_0 + p_1 + p_2 + p_3)$.
Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
0	52	0,301	150.0,301=45,15	1,039
1	48	0,361	150.0,361=54,15	0,698
2	36	0,217	150.0,217=32,55	0,366
3	10	0,087	150.0,087=13,05	0,713
4	4	0,034	150.0,034=5,1	0,237

K = 1,039 + 0,698 + 0,713 + 0,237 = 3,053, r = 5, $\chi^2_{0,95}(4) = 9,488$. Protože 3,053 < 9,488, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Poznámka k testu dobré shody: Tento test může být použit i v těch případech, kdy rozložení, z něhož daný náhodný výběr pochází, neodpovídá nějakému známému rozložení (např. exponenciálnímu, normálnímu, Poissonovu, ...), ale je určeno intuitivně nebo na základě zkušenosti.

Příklad: Ve svých pokusech pozoroval J.G. Mendel 10 rostlin hrachu a na každé z nich počet žlutých a zelených semen. Výsledky pokusu:

č.rostliny	1	2	3	4	5	6	7	8	9	10
počet žlutých	25	32	14	70	24	20	32	44	50	44
počet zelených	11	7	5	27	13	6	13	9	14	18
celkem	36	39	19	97	37	26	45	53	64	62

Z genetických modelů vyplývá, že pravděpodobnost výskytu žlutého semene by měla být 0,75 a zeleného 0,25. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že výsledky Mendelových pokusů se shodují s modelem.

Řešení:
Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
1	25	0,75	36.0,75=27	0,148148
2	32	0,75	39.0,75=29,25	0,258547
⋮	⋮	⋮	⋮	⋮
10	44	0,75	62.0,75=46,5	0,134409

K = 0,148148 + 0,258547 + ... + 0,134409 = 1,797495, r = 10, $\chi^2_{0,95}(9) = 16,9$.
Protože 1,797495 < 16,9, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Příklady k 10. kapitole

Příklad 1: Při 60 hodech kostkou jsme dosáhli těchto výsledků: 9 x jednička, 11 x dvojka, 10 x trojka, 13 x čtyřka, 11 x pětka a 6 x šestka. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že kostka je homogenní.

Řešení: n = 60

j	n_j	p_j	np_j	$(n_j - np_j)^2$	$(n_j - np_j)^2 / np_j$
1	9	1/6	10	1	1/10
2	11	1/6	10	1	1/10
3	10	1/6	10	0	0
4	13	1/6	10	9	9/10
5	11	1/6	10	1	1/10
6	6	1/6	10	16	16/10

K = 2,8, r = 6, p = 0, $\chi^2_{0,95}(5) = 11,07$. Protože K < 11,07, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

b) Test dobré shody pro Poissonovo rozložení

Příklad 2: Ze záznamů autosalónu byl ve 100 náhodně vybraných dnech zjištěn počet prodaných aut.

Počet prodaných aut za den	0	1	2	3	4	5 a víc
Počet dnů	9	43	29	11	5	3

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že počet prodaných aut za den se řídí Poissonovým rozložením.

Řešení: Parametr λ Poissonova rozložení neznáme, odhadneme ho pomocí výběrového průměru.

$m = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]} = \frac{1}{100}(0 \cdot 9 + 1 \cdot 43 + 2 \cdot 29 + 3 \cdot 11 + 4 \cdot 5 + 5 \cdot 3) = 1,7 = \hat{\lambda}$. Pravděpodobnost, že náhodná veličina $X \sim Po(1,7)$ bude nabývat hodnot p_j, j = 0,1,2,3,4,5 a víc, je

$p_j = \frac{1,7^j}{j!} e^{-1,7}$, $j = 0,1,2,3,4$, $p_5 = 1 - (p_0 + p_1 + p_2 + p_3 + p_4)$

j	n_j	p_j	np_j	$(n_j - np_j)^2$	$(n_j - np_j)^2 / np_j$
0	9	0,1827	18,27	85,9329	4,7035
1	43	0,3106	31,06	142,5636	4,5899
2	29	0,264	26,4	6,76	0,2561
3	11	0,1496	14,96	15,6816	1,0482
4	5	0,0636	6,36	1,8496	0,2908
5	3	0,0296	2,96	0,0016	0,0005

K = 10,8891, r = 6, p = 1, $\chi^2_{0,95}(4) = 9,488$. Protože $K \geq 9,488$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Příklad 3: Při parlamentních volbách získaly 4 nejsilnější strany 30%, 20%, 15% a 10% hlasů, zbytek hlasů byl rozdělen mezi ostatní strany. Při volbách do obecního zastupitelstva v jedné obci získaly zmíněné strany (ve stejném pořadí) 1400, 900, 900 a 600 hlasů z 5000 odevzdaných hlasů. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozložení hlasů při parlamentních a místních volbách (v uvedené obci) je stejné.

Řešení: n = 60

j	n_j	p_j	np_j	$(n_j - np_j)^2$	$(n_j - np_j)^2 / np_j$
1	1400	0,3	5000.0,3=1500	10000	6,67
2	900	0,2	5000.0,2=1000	10000	10
3	900	0,15	5000.0,15=750	22500	30
4	600	0,1	5000.0,1=500	10000	20
5	1200	0,25	5000.0,25=1250	2500	2

K = 68,67, r = 5, p = 0, $\chi^2_{0,95}(4) = 9,488$. Protože $K \geq 9,488$, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Práce se systémem STATISTICA

Téma: Porovnání empirického a teoretického rozložení

Příklad 1.: **Testování normality pomocí K-S testu a S-W testu**
V sedmi náhodně vybraných prodejnách byly zjištěny následující ceny určitého druhu zboží (v Kč): 35, 29, 30, 33, 45, 33, 36. Rozhodněte pomocí K-S testu a S-W testu na hladině významnosti 0,05, zda lze tyto ceny považovat za realizace náhodného výběru z normálního rozložení.
Návod: Vytvořte datový soubor o 1 proměnné (nazvěte ji X) a 7 případech. Do proměnné X uložte zjištěné ceny. Statistics – Basic Statistics / Tables – Descriptive statistics – OK – Variables X, OK – Normality – zaškrtněte Kolmogorov – Smirnov & Liliefors test for normality a Shaphiro - Wilk's W test – Frequency tables. Ve výstupu se objeví tabulka, v níž je uvedena hodnota testové statistiky pro K-S test (d = 0,24029) a S-W test (W = 0,86866) a odpovídající p-hodnoty. U K-S testu uvažujte Liliefors p, které je počítáno na základě parametrů odhadnutých z dat. V našem případě p > 0,2 a pro S-W test p = 0,18068, tedy ani jeden z testů nezamítá na hladině významnosti 0,05 hypotézu o normalitě. Výpočet je vhodné doplnit N-P plotem nebo Q-Q plotem.

Příklad 2.: **Testování shody empirického a teoretického rozložení pomocí Distribution Fitting**
Byl zjišťován počet poruch určitého zařízení za 100 hodin provozu ve 150 disjunktních 100 hodinových intervalech. Výsledky pozorování:
Počet poruch za 100 hodin provozu: 0 1 2 3 4 a víc
Absolutní četnosti: 52 48 36 10 4
Na hladině významnosti 0,05 testujte hypotézu, že uvedený náhodný výběr pochází z rozložení Po(λ), kde λ = 1,2.
Návod: Vytvořte datový soubor o dvou proměnných (POČET a ČETNOST) a pěti případech a zapište do něj hodnoty 0 1 2 3 4 a 52 48 36 10 4. Statistics – Distribution Fitting – Discrete Distribuions – Poisson – OK – Variable POČET – Weight variable ČETNOST – Status On – OK. – Parameters Lambda 1,2, OK. Ve výstupní tabulce je uvedena hodnota testového kritéria (2,10955) a p-hodnota (0,34827). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05. Pro vytvoření grafu se vraťte do Fitting Discrete Distributions – Quick – Plot of observed and expected distribution.

Příklad 3.: **Porovnání empirického a teoretického rozložení v případě, kdy teoretické rozložení není v nabídce, kterou poskytuje STATISTICA**
Firma, která vlastní několik supermarketů, se zajímá, zda zákazníci dávají přednost některému dnu v týdnu pro nákup. Náhodně bylo vybráno 300 zákazníků, kteří měli říci, který den v týdnu nejčastěji nakupují v supermarketu.
Výsledky:

Den	pondělí	úterý	středa	čtvrtek	pátek	sobota	neděle
Počet	10	20	40	40	80	60	50

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že žádný den v týdnu nemá při nakupování v supermarketu přednost před jinými dny.
Návod: Vytvořte datový soubor se dvěma proměnnými X a Y a sedmi případy. Do proměnné X zapište pozorované absolutní četnosti, do proměnné Y vypočítané teoretické četnosti (v našem případě 300/7). Statistics – Nonparametrics – Observed versus expected χ^2 – Variables Observed X, Expected Y, OK – Summary. Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Square = 78) a odpovídající p-hodnotu, kterou porovnáme se zvolenou

hladinou významnosti. V našem případě je p-hodnota velmi malá, takřka nulová, takže nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05.

Příklad k samostatnému řešení:
Na webové stránce ministerstva vnitra (www.mvcr.cz/statistiky/cro/zijici/index.html) jsou mj. uvedeny počty občanů v několika městech podle měsíce narození. Vyberte si některé město a testem dobré shody na asymptotické hladině významnosti 0,05 ověřte hypotézu, že pravděpodobnost narození je pro všechny měsíce stejná. (Pravděpodobnost narození pro libovolný měsíc získáte tak, že počet dnů v tomto měsíci podělíte počtem dnů v roce.) Počty narozených lidí v jednotlivých měsících roku rovněž znázorněte graficky. (Graphs – Stats 2D Graphs – Bar/Column Plots)