Porovnání empirického a teoretického rozložení

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.

Testy dobré shody pro diskrétní a spojité rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ 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, ..., 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, …, 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 testová statistika $K \geq \chi^2_{1-\alpha}(r\text{-}1\text{-}p)$. Aproximace se považuje za vyhovující, když teoretické četnosti $np_j \geq 5$, 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: Testování shody empirického a teoretického rozložení při úplně specifikovaném problému

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

$$p_j = \frac{\lambda^j}{j!} e^{-\lambda} = \frac{1,2^j}{j!} e^{-1,2}, \; j = 0,1,2,3, \quad p_4 = 1 - (p_0 + p_1 + p_2 + p_3).$$

Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j	$(n_j - np_j)^2 / np_j$
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

Podmínky dobré aproximace jsou splněny, všechny teoretické četnosti jsou větší než 5.
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.

Výpočet pomocí systému STATISTICA:

Načteme datový soubor poruchy.sta. Proměnná POCET obsahuje počet poruch, proměnná CETNOST pak absolutní četnosti zjištěného počtu poruch.

Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – záložka Parametry - Lambda 1,2 - Výpočet.

Proměnná: POCET , Rozdělení:Poissonovo, Lambda = 1,200 (poruchy.sta)
Chí-kvadrát = 3,03371, sv = 3, p = 0,38646

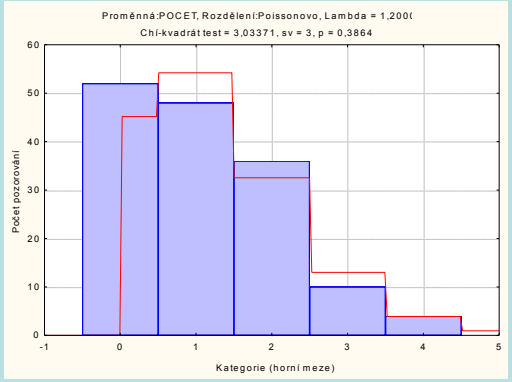
Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 0,00000	52	52	34,66667	34,6667	45,17914	45,1791	30,11943	30,1194
1,00000	48	100	32,00000	66,6667	54,21495	99,3941	36,14330	66,2627
2,00000	36	136	24,00000	90,6667	32,52897	131,9231	21,68598	87,9487
3,00000	10	146	6,66667	97,3333	13,01159	144,9347	8,67439	96,6231
< Nekonečno	4	150	2,66667	100,0000	5,06535	150,0000	3,37690	100,0000

V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (3,03371), počet stupňů volnosti = 3 a p-hodnota (0,38646). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Počet stupňů volnosti 3 však neopovídá tomu, že známe parametr λ, ve skutečnosti je počet stupňů volnosti 4. Proto pro výpočet p-hodnoty otevřeme nový datový soubor o jedné proměnné a jednom případu.

Do Dlouhého jména napíšeme =1-IChi2(3,03371;4). Dostaneme p-hodnotu 0,5522.

Pro vytvoření grafu se vrátíme do Proložení diskrétních rozložení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení



V grafu jsou patrné určité rozdíly mezi hodnotami pravděpodobnostní a četnostní funkce, ale tyto rozdíly nejsou příliš velké.

Příklad: Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému

V tabulce jsou roztříděny fotbalové zápasy určité soutěže podle počtu vstřelených branek.

Počet branek	0	1	2	3	4 a víc
Počet zápasů	19	30	17	10	8

Na hladině významnosti 0,05 testujte hypotézu, že jde o výběr z Poissonova rozložení.

Výpočet pomocí systému STATISTICA:

Načteme datový soubor branky.sta. Proměnná POCET obsahuje počet vstřelených branek, proměnná CETNOST pak počet zápasů, v nichž bylo dosaženo zjištěného počtu branek.

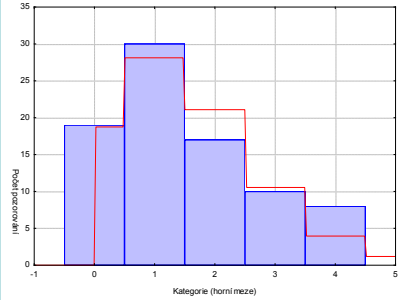
Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – Výpočet.

Proměnná: POCET , Rozdělení:Poissonovo, Lambda = 1,500 (branky .sta)
Chí-kvadrát = 2,07051, sv = 3, p = 0,55790

Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 0,00000	19	19	22,61905	22,6190	18,74294	18,74294	22,31302	22,3130
1,00000	30	49	35,71429	58,3333	28,11440	46,85733	33,46952	55,7825
2,00000	17	66	20,23810	78,5714	21,08580	67,94313	25,10214	80,8847
3,00000	10	76	11,90476	90,4762	10,54290	78,48603	12,55107	93,4358
< Nekonečno	8	84	9,52381	100,0000	5,51397	84,00000	6,56424	100,0000

V tomto případě je parametr λ Poissonova rozložení neznámý, je odhadnut pomocí výběrového průměru a odhad činí 1,5. Dále je v záhlaví výstupní tabulky uvedena hodnota testového kritéria (Chí kvadrát = 2,07051), počet stupňů volnosti r – p – 1 = 5 – 1 – 1 = 3 a p-hodnota (0,5578). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Pro vytvoření grafu se vrátíme do Proložení diskrétních rozložení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení.



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:

číslo rostliny	1	2	3	4	5	6	7	8	9	10
počet žlutých semen	25	32	14	70	24	20	32	44	50	44
počet zelených semen	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	$(n_j - np_j)^2 / np_j$
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

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

Výpočet pomocí systému STATISTICA:

Načteme datový soubor Mendel hrach.sta. Proměnná celkem obsahuje celkový počet semen, X obsahuje pozorovaný počet žlutých semen a Y vypočítané teoretické četnosti žlutých semen (v našem případě X*0,75).

Statistiky – Neparametrická statistika – Pozorované versus očekávané χ^2 – OK - Pozorované četnosti X, Očekávané četnosti Y - OK – Výpočet. Dostaneme tabulku:

Pozorované vs. očekávané četnosti (Mendel hrach.sta)
Chi-Kvadr. = 1,797495 sv = 9 p = ,994280
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. X	očekáv. Y	P - O	(P-O)^2 /O
C: 1	25,0000	27,0000	-2,00000	0,148148
C: 2	32,0000	29,2500	2,75000	0,258547
C: 3	14,0000	14,2500	-0,25000	0,004386
C: 4	70,0000	72,7500	-2,75000	0,103952
C: 5	24,0000	27,7500	-3,75000	0,506757
C: 6	20,0000	19,5000	0,50000	0,012821
C: 7	32,0000	33,7500	-1,75000	0,090741
C: 8	44,0000	39,7500	4,25000	0,454403
C: 9	50,0000	48,0000	2,00000	0,083333
C: 10	44,0000	46,5000	-2,50000	0,134409
Sčt	355,0000	358,5000	-3,50000	1,797495

Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Kvadr = 1,797495), počet stupňů volnosti (sv = 9) a odpovídající p-hodnotu, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota 0,99428, takže nulová hypotéza se nezamítá na asymptotické hladině významnosti 0,05.

Příklad: 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.

Výpočet pomocí systému STATISTICA:

Načteme datový soubor kostka.sta. Proměnná celkem obsahuje X obsahuje pozorované četnosti jednotlivých čísel 1, …, 6 a Y vypočítané teoretické četnosti (v našem případě 10).

Statistiky – Neparametrická statistika – Pozorované versus očekávané χ^2 – OK - Pozorované četnosti X, Očekávané četnosti Y - OK – Výpočet. Dostaneme tabulku:

Pozorované v s. očekávané četnosti (kostka.sta)
Chi-Kvadr. = 2,800000 sv = 5 p = ,730786

Případ	pozorov. X	očekáv. Y	P - O	(P-O)^2 /O
C: 1	9,00000	10,00000	-1,00000	0,100000
C: 2	11,00000	10,00000	1,00000	0,100000
C: 3	10,00000	10,00000	0,00000	0,000000
C: 4	13,00000	10,00000	3,00000	0,900000
C: 5	11,00000	10,00000	1,00000	0,100000
C: 6	6,00000	10,00000	-4,00000	1,600000
Sčt	60,00000	60,00000	0,00000	2,800000

Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Kvadr = 2,8), počet stupňů volnosti (sv = 5) a odpovídající p-hodnotu, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota 0,730786, takže nulová hypotéza se nezamítá na asymptotické hladině významnosti 0,05.

Příklad: 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.

Výpočet pomocí systému STATISTICA:

Načteme datový soubor autosalon.sta. Proměnná POCET obsahuje počet prodaných aut, proměnná CETNOST pak počet dnů, v nichž byl prodán zjištěný zjištěného počet aut.

Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – Výpočet.

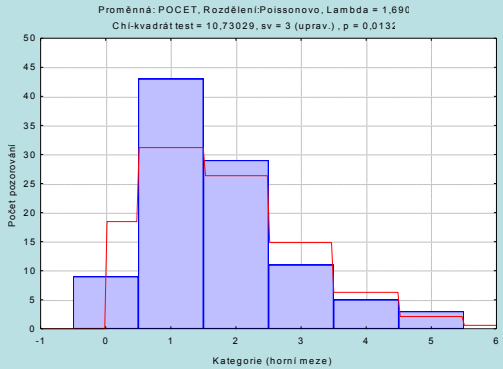
Proměnná: POCET, Rozdělení:Poissonovo, Lambda = 1,69000 (autosalon.sta)
Chí-kvadrát = 10,73029, sv = 3 (uprav.) , p = 0,01328

Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 0,00000	9	9	9,00000	9,0000	18,45196	18,4520	18,45196	18,4520
1,00000	43	52	43,00000	52,0000	31,18380	49,6358	31,18380	49,6358
2,00000	29	81	29,00000	81,0000	26,35031	75,9861	26,35031	75,9861
3,00000	11	92	11,00000	92,0000	14,84401	90,8301	14,84401	90,8301
4,00000	5	97	5,00000	97,0000	6,27159	97,1017	6,27159	97,1017
< Nekonečno	3	100	3,00000	100,0000	2,89834	100,0000	2,89834	100,0000

V záhlaví výstupní tabulky uvedena hodnota testového kritéria (10,73029), počet stupňů volnosti 3 a p-hodnota (0,01328). Nulová hypotéza se tedy zamítá na asymptotické hladině významnosti 0,05.

Vidíme, že nesouhlasí počet stupňů volnosti, měl by být 4. Proto p-hodnotu vypočteme zvlášť. Otevřeme nový datový soubor o jedné proměnné a jednom případu. Do Dlouhého jména napíšeme =1-IChi2(10,73029;4). Dostaneme p-hodnotu 0,0298.

Pro vytvoření grafu se vrátíme do Proložení diskrétních rozložení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení.



V tomto případě jsou patrné značné rozdíly mezi pozorovanými a teoretickými četnostmi.

Jednoduchý test exponenciálního rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z exponenciálního rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim Ex(\lambda)$ je $E(X) = 1/\lambda$ a rozptyl je $D(X) = 1/\lambda^2$.

Test založíme na statistice $K = \frac{(n-1)S^2}{M^2}$, která se v případě platnosti H_0 asymptoticky řídí rozložením $\chi^2(n-1)$.

Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty)$.

Jestliže $K \in W$, H_0 zamítáme na asymptotické hladině významnosti α.

Příklad: Byla zkoumána doba životnosti 45 součástek (v hodinách). Zjistili jsme, že průměrná doba životnosti činila m = 99,93 h a rozptyl $s^2 = 7328{,}91\ h^2$. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z exponenciálního rozložení.

Řešení:

Testová statistika: $K = \frac{(n-1)S^2}{M^2} = \frac{44 \cdot 7328{,}91}{99{,}93^2} = 32{,}2924$

Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty) = \langle 0, \chi^2_{0{,}025}(44)\rangle \cup \langle \chi^2_{0{,}975}(44), \infty) = \langle 0; 27{,}575\rangle \cup \langle 64{,}202, \infty)$

Protože se testová statistika nerealizuje v kritickém oboru, hypotézu o exponenciálním rozložení nezamítáme na asymptotické hladině významnosti 0,05.

Jednoduchý test Poissonova rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z Poissonova rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim Po(\lambda)$ je $E(X) = \lambda$ a rozptyl je $D(X) = \lambda$.

Test založíme na statistice $K = \frac{(n-1)S^2}{M}$, která se v případě platnosti H_0 asymptoticky řídí rozložením $\chi^2(n-1)$.

Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty\rangle$.

Příklad: Studujeme rozložení počtu pacientů, kteří během 75 dnů přijdou na pohotovost. Osmihodinovou pracovní dobu rozdělíme do půlhodinových intervalů a v každém intervalu zjistíme počet příchozích pacientů:

Počet pacientů	0	1	2	3	4	5	6	7	8	9	10
Pozorovaná četnost	79	188	282	275	196	114	45	10	7	3	1

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z Poissonova rozložení.

Řešení:

Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$$m = \frac{1}{1200}(0 \cdot 79 + 1 \cdot 188 + \ldots + 10 \cdot 1) = 2,80\overline{3}$$

$$s^2 = \frac{1}{1199}\left(79 \cdot (0 - 2,80\overline{3})^2 + 188 \cdot (1 - 2,80\overline{3})^2 + \ldots + 1 \cdot (10 - 2,80\overline{3})^2\right) = 2,708579$$

$$K = \frac{(n-1)S^2}{M} = \frac{1199 \cdot 2,708579}{2,80\overline{3}} = 1158,579,$$

Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty\rangle = \langle 0; 1104,93\rangle \cup \langle 1296,86; \infty\rangle$

H_0 nezamítáme na asymptotické hladině významnosti 0,05.