Testování normality dat

Při zpracování dat se často předpokládá, že daný náhodný výběr pochází z normálního rozložení. Posuzujeme ji pomocí N-P plotu, Q-Q plotu či histogramu.
Vzhledem k důležitosti předpokladu normality se vedle grafického posouzení doporučuje též použití některého testu normality, např.
Kolmogorovova – Smirnovova testu či jeho Lilieforsovy modifikace,
Shapirova – Wilkova testu
testu dobré shody.
K závěrům těchto testů však přistupujeme s určitou opatrností. Máme-li k dispozici rozsáhlejší datový soubor (orientačně $n > 30$) a test zamítne na obvyklé hladině významnosti 0,05 nebo 0,01 hypotézu o normalitě, i když vzhled diagnostických grafů svědčí jenom o lehkém porušení normality, nedopustíme se závažné chyby, pokud použijeme statistickou metodu založenou na normalitě dat.
(V případě jednoho dvourozměrného náhodného výběru posuzujeme dvourozměrnou normalitu dat graficky pomocí dvourozměrného tečkového diagramu s proloženou $100(1-\alpha)\%$ elipsou konstantní hustoty pravděpodobnosti).

Kolmogorovův – Smirnovův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z normálního rozložení s parametry μ a σ^2. Distribuční funkci tohoto rozložení označme $\Phi_T(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce.
Testovou statistikou je statistika $D_n = \sup_{-\infty < x < \infty} |F_n(x) - \Phi_T(x)|$.

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 $\sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$.

V případě, že neznáme parametry μ a σ^2 normálního rozložení (což je nejčastější případ), změní se rozložení testové statistiky D_n. V takovém případě jde o Lilieforsovu modifikaci Kolmogorovova – Smirnovova testu. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

Tabulka kritických hodnot Lilieforsovy modifikace K- S testu

Rozsah výběru n	α				Rozsah výběru n	α			
	0,2	0,1	0,05	0,01		0,2	0,1	0,05	0,01
4	0,303	0,346	0,376	0,413	16	0,176	0,195	0,213	0,247
5	0,289	0,319	0,343	0,397	17	0,171	0,190	0,207	0,240
6	0,269	0,297	0,323	0,371	18	0,167	0,185	0,202	0,234
7	0,252	0,280	0,304	0,351	19	0,163	0,181	0,197	0,228
8	0,239	0,265	0,288	0,333	20	0,159	0,176	0,192	0,223
9	0,227	0,252	0,274	0,317	25	0,143	0,159	0,173	0,201
10	0,217	0,241	0,262	0,304	30	0,131	0,146	0,159	0,185
11	0,208	0,231	0,251	0,291	40	0,115	0,128	0,139	0,162
12	0,200	0,222	0,242	0,281	100	0,074	0,082	0,089	0,104
13	0,193	0,215	0,234	0,271	400	0,037	0,041	0,045	0,052
14	0,187	0,208	0,226	0,262	900	0,025	0,028	0,030	0,035
15	0,181	0,201	0,219	0,254					

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:

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

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

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

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

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

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

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

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

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

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

$$\Phi_T(12) = \Phi\left(\frac{12-1}{\sqrt{10}}\right) = \Phi(0{,}32) = 0{,}62552$$

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

(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.

Výpočet pomocí systému STATISTICA:

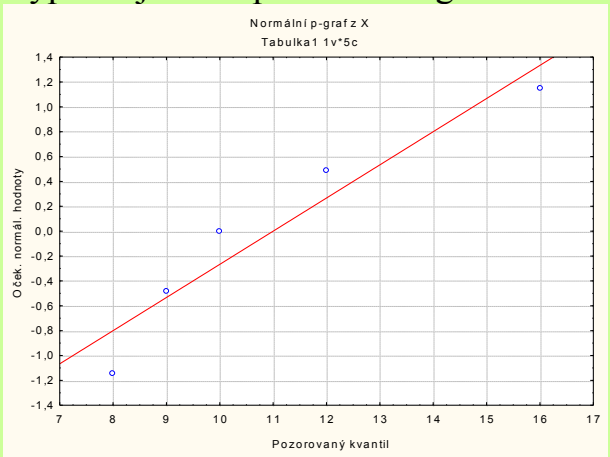
Vytvoříme nový datový soubor o jedné proměnné X a pěti případech.

Statistiky – Základní statistiky a tabulky – Tabulky četností - OK – Proměnné X, OK – Normality – zaškrtneme Lilieforsův test –Testy normality

Proměnná	N	max D	Lilliefors p
X	5	0,224085	p > .20

Ve výstupní tabulce je uvedena hodnota testové statistiky a dolní mez pro p-hodnotu. Protože p > 0,2 nezamítáme hypotézu o normalitě na hladině významnosti 0,05.

Výpočet ještě doplníme N-P grafem:



Ze vzhledu diagramu je patrné, že datový soubor vykazuje kladné sešikmení.

Shapirův – Wilkův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení $N(\mu, \sigma^2)$.

Testová statistika má tvar:

$$W = \frac{\sum_{i=1}^{m} a_i^{(n)} \left(X_{(n-i+1)} - X_{(i)} \right)^2}{\sum_{i=1}^{m} \left(X_i - M \right)^2},$$

kde m = n/2 pro n sudé a m = (n-1)/2 pro n liché. Koeficienty $a_i^{(n)}$ jsou tabelovány.

Na testovou statistiku W lze pohlížet jako na korelační koeficient mezi uspořádanými pozorováními a jim odpovídajícími kvantily standardizovaného normálního rozložení. V případě, že data vykazují perfektní shodu s normálním rozložením, bude mít W hodnotu 1. Hypotézu o normalitě tedy zamítneme na hladině významnosti α, když se na této hladině neprokáže korelace mezi daty a jim odpovídajícími kvantily rozložení N(0,1).

Lze také říci, že S – W 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, ale v systému STATISTICA je implementováno jeho rozšíření i na výběry velkých rozsahů, kolem 2000.)

Příklad: 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.

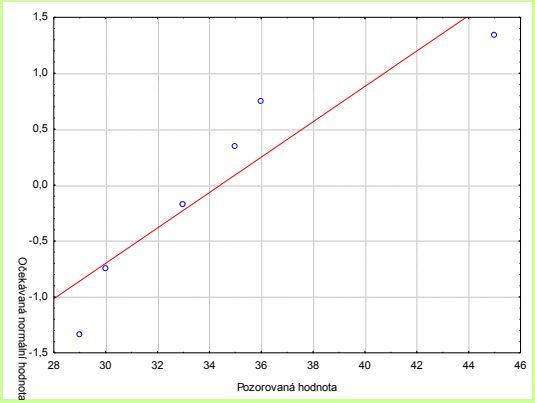
Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a 7 případech. Do proměnné X jsou zapíšeme zjištěné ceny.
Statistiky – Základní statistiky a tabulky – Tabulky četností - OK – Proměnné X, OK – Normality – zaškrtneme Lilieforsův test a Shaphiro - Wilksův W test –Testy normality

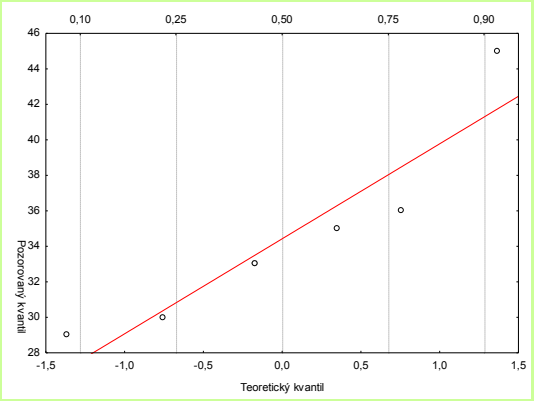
	Testy normality (Tabulka22)				
Proměnná	N	max D	Lilliefors p	W	p
x	7	0,240290	p > .20	0,868661	0,180679

V tabulce je uvedena hodnota testové statistiky pro Lilieforsův test (d = 0,24029) a pro S-W test (W = 0,86866) a odpovídající p-hodnoty. Lilieforsovo p je počítáno na základě parametrů odhadnutých z dat. V našem případě p > 0,2 a pro S-W test p = 0,18068. Ani jeden z testů nezamítá nulovou hypotézu o normalitě.
Výpočet doplníme normálním pravděpodobnostním grafem a kvantil – kvantilovým grafem:
Grafy – 2D Grafy – Normální pravděpodobnostní grafy (resp. Grafy typu Q - Q) - Proměnné X – OK.

N-P plot



Q-Q plot



Test dobré shody pro normální rozložení

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

- Data rozdělíme do r třídicích intervalů $\langle u_j, u_{j+} \rangle$, j = 1, ..., r.
- Zjistíme absolutní četnost n_j j-tého třídicího intervalu.
- 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 $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.
- Vypočteme testovou statistiku: $\kappa = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}$. Platí-li nulová hypotéza, pak $K \approx \chi^2(r-1-k)$, kde k je počet odhadovaných parametrů normálního rozložení. (Obvykle z dat z dat odhadujeme střední hodnotu i rozptyl, tedy k = 2.)
- Stanovíme kritický obor $W = \langle \chi^2_{1-\alpha}(r-1-k), \infty \rangle$
- Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $\kappa \in W$.

(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 slučovat, což vede ke ztrátě informace.

Příklad:
Byl pořízen náhodný výběr rozsahu n = 100. Jeho číselné realizace byly roztříděny do 5 ekvidistantních třídicích intervalů o délce 0,04, přičemž dolní mez prvního třídicího intervalu je 3,92. Absolutní četnosti jednotlivých třídicích intervalů jsou: 11, 20, 44, 19, 6.
Výběrový průměr se realizoval hodnotou m = 4,02 a výběrová směrodatná odchylka hodnotou s = 0,04.
Na asymptotické hladině významnosti 0,05 testujte hypotézu, že náhodný výběr pochází z normálního rozložení.
Řešení:
Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.
Přitom symbolem Φ značíme distribuční funkci rozložení $N(\mu,\sigma^2)$, kde $\mu = 4{,}02$ a $\sigma = 0{,}04$.

$\langle u_j, u_{j+} \rangle$	n_j	$p_j=\Phi(u_{j+1})- \Phi(u_j)$	np_j	$(n_j - np_j)^2$	$\frac{(n_j - np_j)^2}{np_j}$
$\langle 3{,}92, 3{,}96 \rangle$	11	0,060598	6,0598	24,4060	4,0276
$\langle 3{,}96, 4{,}00 \rangle$	20	0,241730	24,1730	17,4142	0,7204
$\langle 4{,}00, 4{,}04 \rangle$	44	0,382925	38,2925	32,5756	0,8507
$\langle 4{,}04, 4{,}08 \rangle$	19	0,241730	24,1730	26,7608	1,1070
$\langle 4{,}08, 4{,}12 \rangle$	6	0,060598	6,0598	0,0036	0,0006

$K = 4{,}0276 + 0{,}7204 + 0{,}8507 + 1{,}1070 + 0{,}0006 = 6{,}7063$

Kritický obor: $W = \langle \chi^2_{1-\alpha}(k-1-r), \infty \rangle = \langle \chi^2_{0{,}95}(5-1-2), \infty \rangle = \langle 5{,}9915, \infty \rangle$

Protože testová statistika se realizuje v kritickém oboru, hypotézu o normalitě zamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Protože nemáme k dispozici původní data, ale jenom třídicí intervaly a jejich četnosti, do nového datového souboru o dvou proměnných xj a nj zadáme středy třídicích intervalů a jejich absolutní četnosti:

	1 xj	2 nj
1	3,94	11
2	3,98	20
3	4,02	44
4	4,06	19
5	4,1	6

Statistiky – Prokládání rozdělení – ponecháme implicitní nastavení pro Normální rozdělení – OK – Proměnná xj – klikneme na ikonu se závažím – Proměnná vah nj – Stav Zapnuto – OK – Parametry – Počet kategorií 5, Průměr 4,02, Rozptyl 0,0016, OK.

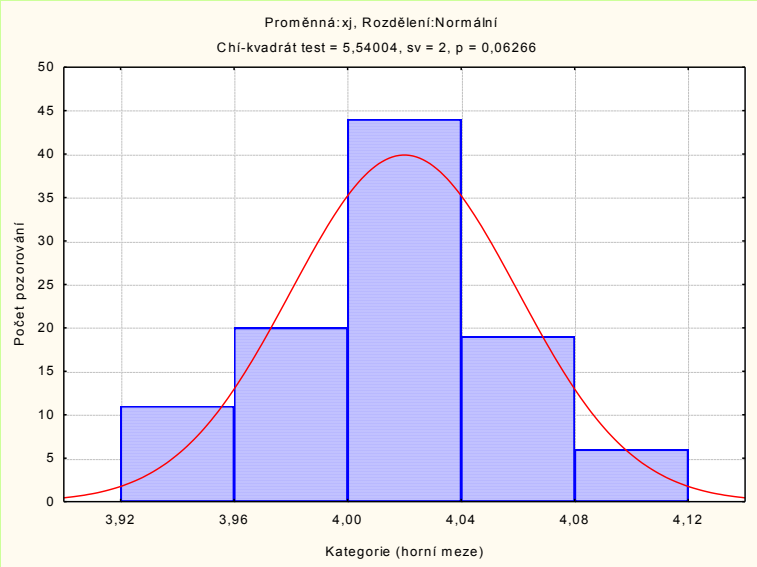
Dostaneme výstupní tabulku:

	Proměnná: xj, Rozdělení:Normální (Tabulka10) Chí-kvadrát = 5,54004, sv = 2, p = 0,06266							
Horní hranice	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 3,96000	11	11	11,00000	11,0000	6,68072	6,6807	6,68072	6,6807
4,00000	20	31	20,00000	31,0000	24,17303	30,8538	24,17303	30,8538
4,04000	44	75	44,00000	75,0000	38,29249	69,1462	38,29249	69,1462
4,08000	19	94	19,00000	94,0000	24,17303	93,3193	24,17303	93,3193
< Nekonečno	6	100	6,00000	100,0000	6,68072	100,0000	6,68072	100,0000

V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (5,54004), počet stupňů volnosti = 2 a p-hodnota (0,06266). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Rozdíl oproti ručnímu výpočtu je způsoben tím, že systém STATISTICA uvažuje první interval $\langle -\infty, 3{,}96 \rangle$ a poslední interval $\langle 4{,}08, \infty \rangle$.

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



Další testy normality

Existují testy normality založené na výběrové šikmosti a špičatosti.

Pro náhodnou veličinu s normálním rozložením platí, že její šikmost i špičatost jsou nulové. Pro výběr z normálního rozložení by tedy výběrová šikmost a špičatost měly být blízké 0.

Nechť $X_1, \ldots, X_n$ je náhodný výběr.

Výběrová šikmost: $A_3 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2}\right]^3}$

Výběrová špičatost: $A_4 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^4}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2}\right]^4} - 3$

Lze dokázat, že pro výběr z normálního rozložení platí:

$E(A_3) = 0$, $D(A_3) = \dfrac{6(n-2)}{(n+1)(n+3)}$, $E(A_4) = -\dfrac{6}{n+1}$, $D(A_4) = \dfrac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$.

Pro $n \to \infty$ se statistiky $A_3\sqrt{n}$ a $A_4\sqrt{n}$ asymptoticky řídí normálním rozložením.

Test založený na šikmosti zamítne hypotézu o normalitě na asymptotické hladině významnosti α, když

$$\frac{|A_3|}{\sqrt{D(A_3)}} \geq u_{1-\alpha/2}.$$

Test založený na špičatosti zamítne hypotézu o normalitě na asymptotické hladině významnosti α, když

$$\frac{|A_4 - E(A_4)|}{\sqrt{D(A_4)}} \geq u_{1-\alpha/2}.$$

Úlohy o parametru ϑ alternativního rozložení

S náhodným výběrem rozsahu n z alternativního rozložení se setkáváme v situaci, kdy provádíme n opakovaných nezávislých pokusů a v každém z těchto pokusů sledujeme nastoupení úspěchu. Pravděpodobnost úspěchu je pro všechny pokusy stejná. Náhodná veličina X_i nabude hodnoty 1, pokud v i-tém pokusu nastal úspěch a hodnoty 0, pokud v i-tém pokusu úspěch nenastal, i = 1, 2, …, n. Realizací náhodného výběru $X_1, \ldots, X_n$ je tedy posloupnost 0 a 1.

Opakování:

Alternativní rozložení: Náhodná veličina X udává počet úspěchů v jednom pokusu, přičemž pravděpodobnost úspěchu je ϑ. Píšeme $X \sim A(\vartheta)$.

$$\pi(x) = \begin{cases} 1-\vartheta & \text{pro } x = 0 \\ \vartheta & \text{pro } x = 1 \\ 0 & \text{jinak} \end{cases} \quad \text{neboli} \quad \pi(x) = \begin{cases} \vartheta^x (1-\vartheta)^{1-x} & \text{pro } x = 0, 1 \\ 0 & \text{jinak} \end{cases}$$

Binomické rozložení: Náhodná veličina X udává počet úspěchů v posloupnosti n nezávislých opakovaných pokusů, přičemž pravděpodobnost úspěchu je v každém pokusu ϑ. Píšeme $X \sim Bi(n, \vartheta)$.

$$\pi(x) = \begin{cases} \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x} & \text{pro } x = 0, \ldots, n \\ 0 & \text{jinak} \end{cases}$$

$E(X) = n\vartheta$, $D(X) = n\vartheta(1-\vartheta)$

(Alternativní rozložení je speciálním případem binomického rozložení pro n = 1.

Jsou-li $X_1, \ldots, X_n$ stochasticky nezávislé náhodné veličiny, $X_i \sim A(\vartheta)$, i = 1, ..., n, pak $X = \sum_{i=1}^{n} X_i \sim Bi(n, \vartheta)$.)

Centrální limitní věta:
Jsou-li náhodné veličiny $X_1, \ldots, X_n$ stochasticky nezávislé a všechny mají stejné rozložení se střední hodnotou μ a rozptylem σ^2, pak pro velká n ($n \geq 30$) lze rozložení součtu $\sum_{i=1}^{n} X_i$ aproximovat normálním rozložením $N(n\mu, n\sigma^2)$. Zkráceně píšeme $\sum_{i=} \approx N(n\mu, n\sigma)$. Pokud součet $\sum_{i=1}^{n} X_i$ standardizujeme, tj. vytvoříme náhodnou veličinu $U_n = \frac{\sum_{i=} - n\mu}{\sigma\sqrt{n}}$, pak rozložení této náhodné veličiny lze aproximovat standardizovaným normálním rozložením. Zkráceně píšeme $U_n \approx N(0,1)$

Věta: Asymptotické rozložení statistiky odvozené z výběrového průměru.
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $A(\vartheta)$ a nechť je splněna podmínka $n\vartheta(-\vartheta > 9$. Pak statistika $U = \frac{M - \vartheta}{\sqrt{\frac{\vartheta(-\vartheta}{n}}}$ konverguje v distribuci k náhodné veličině se standardizovaným normálním rozložením. (Říkáme, že U má asymptoticky rozložení N(0,1) a píšeme $U \approx N(0,1)$.)

Důkaz:

Protože $X_1, \ldots, X_n$ je náhodný výběr z rozložení $A(\vartheta)$, bude mít statistika $Y_n = \sum_{i=1}^{n} X_i$ (výběrový úhrn) rozložení $Bi(n, \vartheta)$. Y_n má střední hodnotu $E(Y_n) = n\vartheta$ a rozptyl $D(Y_n) = n\vartheta(-\vartheta$. Podle centrální limitní věty se standardizovaná statistika $U = \frac{Y_n - n\vartheta}{\sqrt{n\vartheta(-\vartheta}}$ asymptoticky řídí standardizovaným normálním rozložením N(0,1). Pokud čitatele i jmenovatele podělíme n, dostaneme vyjádření: $U = \frac{\frac{Y_n}{n} - \vartheta}{\sqrt{\frac{n\vartheta(-\vartheta}{n^2}}} = \frac{\frac{1}{n}\sum_{i=} - \vartheta}{\sqrt{\frac{\vartheta(-\vartheta}{n}}} = \frac{M - \vartheta}{\sqrt{\frac{\vartheta(-\vartheta}{n}}} \approx N(0,1)$

Věta: Vzorec pro meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ.
Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ jsou:

$$d = m - \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2},\ h = m + \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}.$$

Důkaz:

Pokud rozptyl $D(M) = \frac{\vartheta(1-\vartheta)}{n}$ nahradíme odhadem $\frac{M(1-M)}{n}$, konvergence náhodné veličiny U k veličině s rozložením N(0,1) se neporuší. Tedy

$$\forall \vartheta \in \Theta: 1-\alpha \le P\left(-u_{1-\alpha/2} < \frac{M-\vartheta}{\sqrt{\frac{M(1-M)}{n}}} < u_{1-\alpha/2}\right) =$$

$$= P\left(M - \sqrt{\frac{M(1-M)}{n}}u_{1-\alpha/2} < \vartheta < M + \sqrt{\frac{M(1-M)}{n}}u_{1-\alpha/2}\right)$$

Příklad:

Náhodně bylo vybráno 100 osob a zjištěno, že 34 z nich nakupuje v internetových obchodech. Najděte 95% asymptotický interval spolehlivosti pro pravděpodobnost, že náhodně vybraná osoba nakupuje v internetových obchodech.

Řešení:

Zavedeme náhodné veličiny $X_1, ..., X_{100}$, přičemž $X_i = 1$, když i-tá osoba nakupuje v internetových obchodech a $X_i = 0$ jinak, i = 1, ..., 100. Tyto náhodné veličiny tvoří náhodný výběr z rozložení A(ϑ).

n = 100, m = 34/100, $\alpha = 0{,}05$, $u_{1-\alpha/2} = u_{0{,}975} = 1{,}96$.

Ověření podmínky $n\vartheta(1-\vartheta) > 9$: parametr ϑ neznáme, musíme ho nahradit výběrovým průměrem. Pak $100.0{,}34.0{,}66 = 22{,}44 > 9$.

$$d = 0{,}34 - \sqrt{\frac{0{,}34(1-0{,}34)}{100}}1{,}96 = 0{,}2472\,,\ h = 0{,}34 + \sqrt{\frac{0{,}34(1-0{,}34)}{100}}1{,}96 = 0{,}4328\ .$$

S pravděpodobností přibližně 0,95 tedy $0{,}2472 < \vartheta < 0{,}4328$. Znamená to, že s pravděpodobností přibližně 95% je v uvažované populaci nejméně 24,7% a nejvíce 43,3% osob, které nakupují v internetových obchodech.

Výpočet pomocí systému STATISTICA:

a) Přesný způsob

Otevřeme nový datový soubor se dvěma proměnnými a jednom případu.
První proměnnou nazveme d a do jejího Dlouhého jména napíšeme
=0,34-sqrt(0,34*0,66/100)*VNormal(0,975;0;1)
Druhou proměnnou nazveme h a do jejího Dlouhého jména napíšeme
=0,34+sqrt(0,34*0,66/100)*VNormal(0,975;0;1)
Dostaneme výsledek:

	1 d	2 h
1	0,247155	0,432845

Vidíme, že s pravděpodobností aspoň 0,95 se pravděpodobnost nákupu v inetrnetových obchodech bude pohybovat v mezích 0,2471 až 0,4328.

b) Přibližný způsob, použitelný pro dostatečně velký rozsah výběru

Do nového datového souboru o jedné proměnné X a 100 případech uložíme 34 jedniček (nakupování v internetových obchodech) a 66 nul.
Statistika – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. – ponecháme implicitní hodnotu pro Interval 95,00 – Výpočet.
Dostaneme tabulku:

	Popisné statistiky (Tabulka3)			
Proměnná	N platných	Průměr	Int. spolehl. -95,000%	Int. spolehl. 95,000
X	100	0,340000	0,245532	0,434468

Dospěli jsme k výsledku, že s pravděpodobností aspoň 0,95 se pravděpodobnost nákupu v inetrnetových obchodech bude pohybovat v mezích 0,2455 až 0,4345.

Příklad: Kolik osob musíme vybrat, abychom podíl modrookých osob v populaci odhadli se spolehlivostí 90% a šířka intervalu spolehlivosti byla nanejvýš a) 0,06, b) 0,01?

Řešení:

Šířka 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ:

$$h - d = m + \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2} - \left(m - \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}\right) = 2\sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2}$$

Požadujeme, aby $h - d \leq \Delta$, tedy $2\sqrt{\frac{m(1-m)}{n}}u_{1-\alpha/2} \leq \Delta$. Odtud vyjádříme $n \geq \frac{4m(1-m)u_{1-\alpha/2}^2}{\Delta}$.

Předpokládejme, že nemáme žádné předběžné informace o podílu modrookých osob v populaci. Musíme tedy zvolit takové m, aby šířka intervalu spolehlivosti byla maximální. Maximalizujeme výraz $m(1-m) = m - m^2$. Derivujeme podle m a položíme rovno 0: $1 - 2m = 0 \Rightarrow m = \frac{1}{2}$. V tomto případě volíme relativní četnost m = 0,5.

ad a) $$n \geq \frac{4m(1-m)u_{1-\alpha/2}^2}{\Delta} = \frac{4 \cdot 0{,}5 \cdot 0{,}5 \cdot u_{0{,}95}^2}{0{,}06^2} = \frac{4 \cdot 0{,}5 \cdot 0{,}5 \cdot 1{,}645^2}{0{,}06^2} = 751{,}67$$

Uvedenou podmínku tedy splníme, když vybereme aspoň 752 osob.

ad b) $$n \geq \frac{4m(1-m)u_{1-\alpha/2}^2}{\Delta} = \frac{4 \cdot 0{,}5 \cdot 0{,}5 \cdot u_{0{,}95}^2}{0{,}01^2} = \frac{4 \cdot 0{,}5 \cdot 0{,}5 \cdot 1{,}645^2}{0{,}01^2} = 27060{,}25$$

Chceme-li dosáhnout podstatně užšího intervalu spolehlivosti, musíme vybrat aspoň 27 061 osob.

Modifikace: Předpokládejme, že v populaci je nanejvýš 30% modrookých osob. Pak relativní četnost m = 0,3.

ad a) $n \geq \frac{4m(1-m)u_{1-\alpha/2}^2}{\Delta} = \frac{4 \cdot 0{,}3 \cdot 0{,}7 \cdot u_{0{,}95}^2}{0{,}06^2} = \frac{4 \cdot 0{,}3 \cdot 0{,}7 \cdot 1{,}645^2}{0{,}06^2} = 631{,}41$

V tomto případě stačí vybrat 632 osob.
Ve srovnání s předešlým případem vidíme, že rozsah výběru skutečně klesl.

ad b)

$n \geq \frac{4m(1-m)u_{1-\alpha/2}^2}{\Delta} = \frac{4 \cdot 0{,}3 \cdot 0{,}7 \cdot u_{0{,}95}^2}{0{,}01^2} = \frac{4 \cdot 0{,}3 \cdot 0{,}7 \cdot 1{,}645^2}{0{,}01^2} = 22730{,}61$

V tomto případě musíme vybrat aspoň 22 731 osob.

Testování hypotézy o parametru ϑ

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $A(\vartheta)$ a nechť je splněna podmínka $n\vartheta(1-\vartheta) > 9$.

Na asymptotické hladině významnosti α testujeme hypotézu

H_0: $\vartheta = c$ proti alternativě H_1: $\vartheta \neq c$ (resp. H_1: $\vartheta < c$ resp. H_1: $\vartheta > c$).

Testovým kritériem je statistika $T_0 = \frac{M - c}{\sqrt{\frac{c(1-c)}{n}}}$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1).

Kritický obor má tvar $w = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$ (resp. $w = (-\infty, -u_{1-\alpha})$ resp. $w = (u_{1-\alpha}, \infty)$).

(Testování hypotézy o parametru ϑ lze samozřejmě provést i pomocí 100(1-α)% asymptotického intervalu spolehlivosti nebo pomocí p-hodnoty.)

Příklad: Podíl zmetků při výrobě určité součástky činí $\vartheta = 0{,}01$. Bylo náhodně vybráno 1000 výrobků a zjistilo se, že mezi nimi je 16 zmetků. Na asymptotické hladině významnosti 0,05 testujte hypotézu H_0: $\vartheta = 0{,}01$ proti oboustranné alternativě H_1: $\vartheta \neq 0{,}01$.

Řešení:
Zavedeme náhodné veličiny $X_1, ..., X_{1000}$, přičemž $X_i = 1$, když i-tý výrobek byl zmetek a $X_i = 0$ jinak, i = 1, ..., 1000. Tyto náhodné veličiny tvoří náhodný výběr z rozložení A(ϑ).
Testujeme hypotézu H_0: $\vartheta = 0{,}01$ proti alternativě H_1: $\vartheta \neq 0{,}01$.

Známe: n = 1000, $m = \frac{16}{1000} = 0{,}016$, c = 0,01, $\alpha = 0{,}05$, $u_{1-\alpha/2} = u_{0{,}975} = 1{,}96$

Ověření podmínky $n\vartheta(1-\vartheta) > 9$: $1000.0{,}01.0{,}99 = 9{,}9 > 9$.

a) Testování pomocí kritického oboru:

Realizace testového kritéria: $t_0 = \frac{m - c}{\sqrt{\frac{c(1-c)}{n}}} = \frac{0{,}016 - 0{,}01}{\sqrt{\frac{0{,}01 \cdot 0{,}99}{1000}}} = 1{,}907$.

Kritický obor: $W = (-\infty, -u_{0{,}975}) \cup (u_{0{,}975}, \infty) = (-\infty, -1{,}96) \cup (1{,}96, \infty)$. Protože $1{,}907 \notin W$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

b) Testování pomocí intervalu spolehlivosti

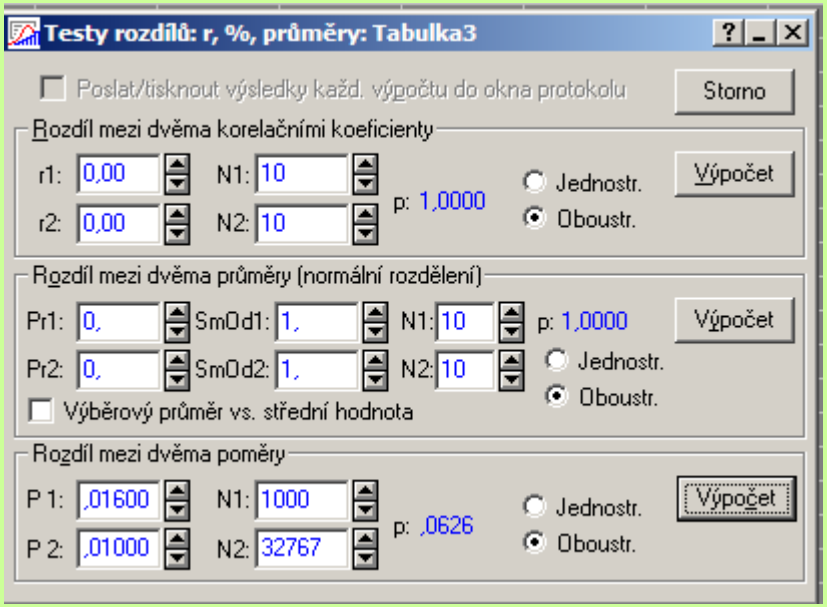
$$d = m - \sqrt{\frac{m(1-m)}{n}} u_{1-\alpha/2} = 0{,}016 - \sqrt{\frac{0{,}016 \cdot 0{,}984}{1000}} 1{,}96 = 0{,}0082$$

$$h = m + \sqrt{\frac{m(1-m)}{n}} u_{1-\alpha/2} = 0{,}016 + \sqrt{\frac{0{,}016 \cdot 0{,}984}{1000}} 1{,}96 = 0{,}0238$$

Protože číslo c = 0,01 leží v intervalu 0,0082 až 0,0238, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

c) Testování pomocí p-hodnoty
Protože testujeme nulovou hypotézu proti oboustranné alternativě, vypočteme p-hodnotu podle vzorce:
$p = 2 \min\{ \Phi(1{,}907), 1 - \Phi(1{,}907) \} = 2 \min \{ 0{,}97104, 1 - 0{,}97104 \} = 0{,}05792$.
Protože vypočtená p-hodnota je větší než hladina významnosti 0,05, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA (pouze přibližný):
Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,016, do políčka N1 napíšeme 1000, do políčka P 2 napíšeme 0,01, do políčka N2 napíšeme 32767 (větší hodnotu systém neumožní) - Výpočet. Dostaneme p-hodnotu 0,0626, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.



Příklad: Nový léčebný postup považujeme za úspěšný, pokud po jeho ukončení bude dosaženo zlepšení zdravotního stavu u alespoň 50% zúčastněných pacientů. Nová terapie byla vyzkoušena u 40 pacientů a ke zlepšení došlo u 24 osob, tj. u 60%. Je možné na asymptotické hladině významnosti 0,05 zamítnout hypotézu, že tato terapie nedosahuje úspěšnosti aspoň 50%?

Řešení:

Zavedeme náhodné veličiny $X_1, ..., X_{40}$, přičemž
$X_i = 1$, když terapie u i-tého pacienta byl úspěšná a
$X_i = 0$ jinak,
$i = 1, ..., 40$.
Tyto náhodné veličiny tvoří náhodný výběr z rozložení $A(\vartheta)$.
Testujeme hypotézu H_0: $\vartheta \leq 0{,}5$ proti pravostranné alternativě H_1: $\vartheta > 0{,}5$.

Známe: $n = 40$, $m = \frac{24}{40} = 0{,}6$, $c = 0{,}5$, $\alpha = 0{,}05$, $u_{1-\alpha} = u_{0{,}95} = 1{,}645$

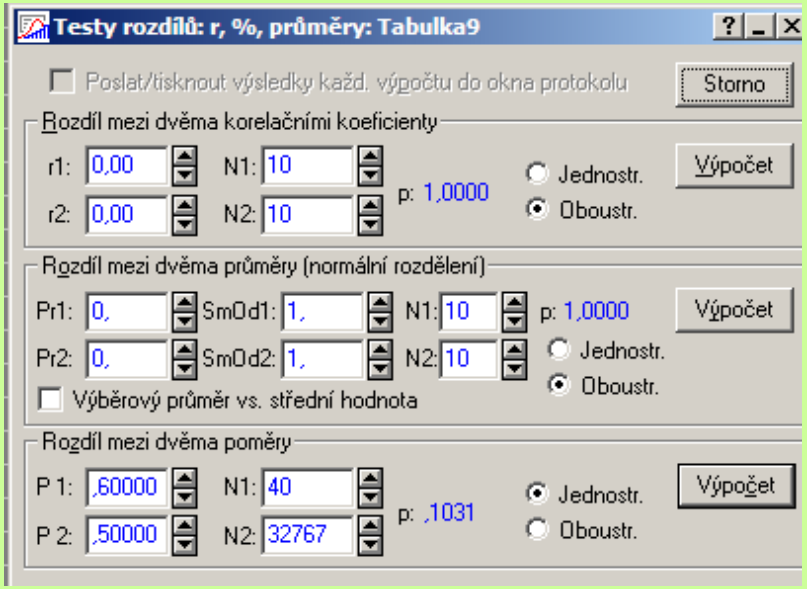
Ověření podmínky $n\vartheta(1-\vartheta) > 9$: $40.0{,}6.0{,}4 = 9{,}6 > 9$.

Realizace testového kritéria: $t_0 = \frac{m - c}{\sqrt{\frac{c \cdot (1-c)}{n}}} = \frac{0{,}6 - 0{,}5}{\sqrt{\frac{0{,}5 \cdot 0{,}5}{40}}} = 1{,}2649$.

Kritický obor: $W = \langle u_{1-\alpha}, \infty) = \langle u_{0{,}95}, \infty) = \langle 1{,}645, \infty)$.

Protože $1{,}2649 \notin W$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:



Vypočtená p-hodnota jednostranného testu je 0,1031, tedy větší než asymptotická hladina významnosti 0,05. H_0 nezamítáme na asymptotické hladině významnosti 0,05.