# Téma 6.: Ověřování normality dat, výběry z alternativního rozložení

## Kolmogorovův – Smirnovův test normality dat

Testujeme nulovou hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z normálního rozložení s parametry $\mu$ a $\sigma^2$. Distribuční funkci tohoto rozložení označme $\Phi_T(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika $D_n = \sup_{-\infty < x < \infty} |F_n(x) - \Phi_1(x)|$.

Nulovou hypotézu zamítáme na hladině významnosti $\alpha$, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota.

V případě, že neznáme parametry $\mu$ a $\sigma^2$ 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.

**Poznámka ke K-S testu ve STATISTICE**

Test normality poskytuje hodnotu testové statistiky (ozn. max D) a dvě p-hodnoty. (p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ podporují nulovou hypotézu, je-li pravdivá. P-hodnotu porovnáváme s námi zvolenou hladinou významnosti $\alpha$. Jestliže p-hodnota $\leq \alpha$, pak $H_0$ zamítáme na hladině významnosti $\alpha$, je-li p-hodnota $> \alpha$, pak $H_0$ nezamítáme na hladině významnosti $\alpha$.) První p-hodnota se vztahuje k případu, kdy střední hodnotu $\mu$ a rozptyl $\sigma^2$ známe předem, druhá (ozn. Lilieforsovo p) se vztahuje k případu, kdy $\mu$ a $\sigma^2$ neznáme. Objeví-li se ve výstupu p = n.s. (tj. non significant), pak hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

## Shapirův – Wilkův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ 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, $n < 50$, ale nyní již existuje modifikace pro velká n. V systému STATISTICA je implementováno rozšíření na n kolem 5000.)

## 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: $K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}$. 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 $\alpha$, když $K \in W$.

(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.

**Úkol :** U 45 studentek VŠE v Praze byla zjišťována výška a obor studia (1 – národní hospodářství, 2 – informatika). Hodnoty jsou uloženy v souboru vyska.sta. Pomocí Lilieforsovy modifikace K-S testu, pomocí S-W testu a pomocí testu dobré shody testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí N-P grafu posuďte vizuálně předpoklad normality.

**Návod:**

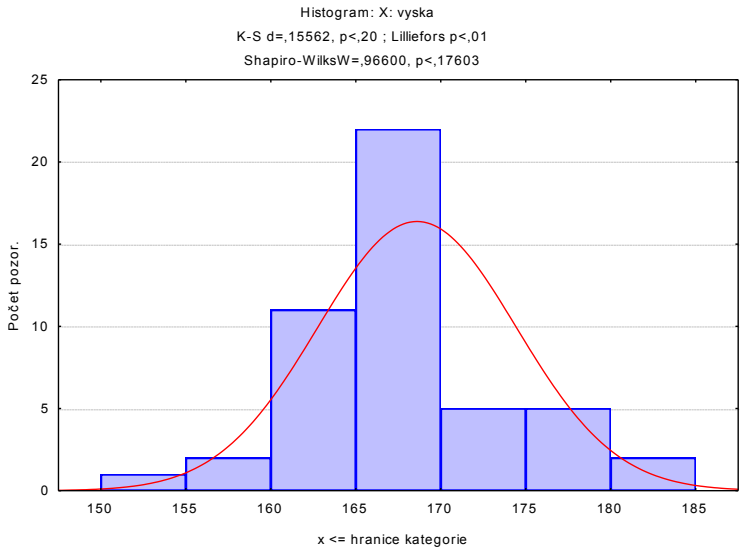
**1. způsob provedení Lilieforsova a S-W testu:** Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Proměnné X – OK – Normalita – zaškrtneme Lilieforsův test a S-W test – Testy normality.

| | Testy normality (vyska.sta) | | | | |
|---|---|---|---|---|---|
| Proměnná | N | max D | Lilliefors p | W | p |
| X: vyska | 48 | 0,155621 | p < ,01 | 0,965996 | 0,176031 |

Výstupní tabulka obsahuje počet pozorování, hodnotu testové statistiky Lilieforsovy modifikace K-S testu (max D = 0,155621), p-hodnotu (p < 0,01), testovou statistiku S-W testu (W = 0,965996) a odpovídající p-hodnotu (p = 0,176031). Vidíme, že Lilieforsův test zamítá hypotézu o normalitě na hladině významnosti 0,05, zatímco S-W test nikoli.

**2. způsob provedení Lilieforsova a S-W testu:** Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Normalita – zaškrtneme K-S test & Lilieforsův test a S-W test – Tabulky četností (nebo Histogram).

| | Tabulka četností: X: vyska (vyska.sta)<br>K-S d=,15562, p<,20 ; Lilliefors p<,01<br>Shapiro-Wilks W=,96600, p<,17603 | | | | | |
|---|---|---|---|---|---|---|
| Kategorie | Četnost | Kumulativní četnost | Rel.četn. (platných) | Kumul. % (platných) | Rel.četn. všech | Kumul. % všech |
| 150,0000<x<=155,0000 | 1 | 1 | 2,08333 | 2,0833 | 2,08333 | 2,0833 |
| 155,0000<x<=160,0000 | 2 | 3 | 4,16667 | 6,2500 | 4,16667 | 6,2500 |
| 160,0000<x<=165,0000 | 11 | 14 | 22,91667 | 29,1667 | 22,91667 | 29,1667 |
| 165,0000<x<=170,0000 | 22 | 36 | 45,83333 | 75,0000 | 45,83333 | 75,0000 |
| 170,0000<x<=175,0000 | 5 | 41 | 10,41667 | 85,4167 | 10,41667 | 85,4167 |
| 175,0000<x<=180,0000 | 5 | 46 | 10,41667 | 95,8333 | 10,41667 | 95,8333 |
| 180,0000<x<=185,0000 | 2 | 48 | 4,16667 | 100,0000 | 4,16667 | 100,0000 |
| ChD | 0 | 48 | 0,00000 | | 0,00000 | 100,0000 |




V tomto případě dostaneme v záhlaví tabulky či histogramu stejné informace jako pomocí předešlého způsobu.

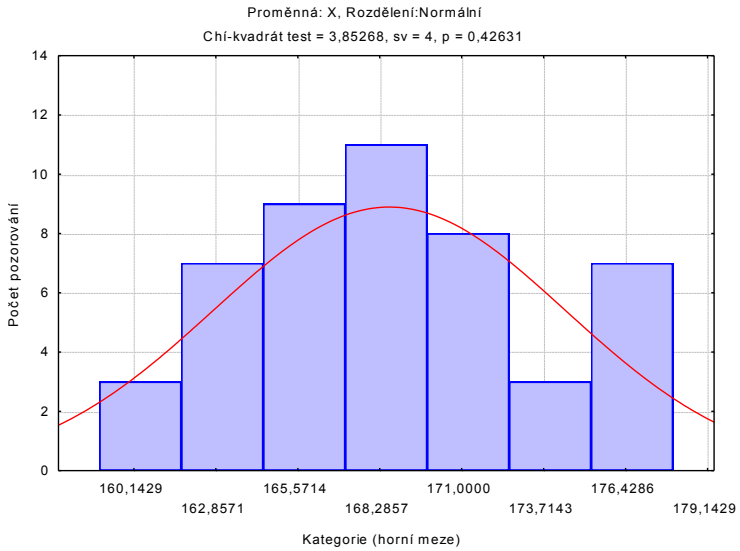
**1. způsob provedení testu dobré shody:** Statistiky - Prokládání rozdělení – ponecháme implicitní nastavení na normální rozložení – OK – Proměnná X – OK – na záložce Parametry změníme Počet kategorií na 7 (podle Sturgesova pravidla) – Výpočet.

| | Proměnná: X, Rozdělení: Normální (vyska.sta)<br>Chí-kvadrát = 1,09280, sv = 1 (uprav.) , p = 0,29585 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Horní hranice | Pozorované Četnosti | Kumulativ. Pozorované | Procent Pozorované | Kumul. % Pozorované | Očekáv. Četnosti | Kumulativ. Očekáv. | Procent Očekáv. | Kumul. % Očekáv. |
| <= 157,14286 | 1 | 1 | 2,08333 | 2,0833 | 1,19706 | 1,19706 | 2,49387 | 2,4939 |
| 162,28571 | 6 | 7 | 12,50000 | 14,5833 | 5,51484 | 6,71189 | 11,48924 | 13,9831 |
| 167,42857 | 12 | 19 | 25,00000 | 39,5833 | 13,46220 | 20,17409 | 28,04624 | 42,0293 |
| 172,57143 | 19 | 38 | 39,58333 | 79,1667 | 15,89146 | 36,06555 | 33,10721 | 75,1366 |
| 177,71429 | 6 | 44 | 12,50000 | 91,6667 | 9,07700 | 45,14255 | 18,91042 | 94,0470 |
| 182,85714 | 2 | 46 | 4,16667 | 95,8333 | 2,50365 | 47,64620 | 5,21594 | 99,2629 |
| < Nekonečno | 2 | 48 | 4,16667 | 100,0000 | 0,35380 | 48,00000 | 0,73708 | 100,0000 |

Při tomto roztřídění dat do 7 intervalů nejsou splněny podmínky dobré aproximace, ve třech intervalech jsou teoretické četnosti pod 5. Změníme tedy dolní mez na 159 a horní na 178.

| | Proměnná: X, Rozdělení: Normální (vyska.sta)<br>Chí-kvadrát = 3,85268, sv = 4, p = 0,42631 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Horní hranice | Pozorované Četnosti | Kumulativ. Pozorované | Procent Pozorované | Kumul. % Pozorované | Očekáv. Četnosti | Kumulativ. Očekáv. | Procent Očekáv. | Kumul. % Očekáv. |
| <= 161,71429 | 3 | 3 | 6,25000 | 6,2500 | 5,722996 | 5,72300 | 11,92291 | 11,9229 |
| 164,42857 | 7 | 10 | 14,58333 | 20,8333 | 5,675946 | 11,39894 | 11,82489 | 23,7478 |
| 167,14286 | 9 | 19 | 18,75000 | 39,5833 | 7,862633 | 19,26157 | 16,38048 | 40,1283 |
| 169,85714 | 11 | 30 | 22,91667 | 62,5000 | 8,812455 | 28,07403 | 18,35928 | 58,4876 |
| 172,57143 | 8 | 38 | 16,66667 | 79,1667 | 7,991516 | 36,06555 | 16,64899 | 75,1366 |
| 175,28571 | 3 | 41 | 6,25000 | 85,4167 | 5,863558 | 41,92910 | 12,21575 | 87,3523 |
| < Nekonečno | 7 | 48 | 14,58333 | 100,0000 | 6,070896 | 48,00000 | 12,64770 | 100,0000 |

V tomto případě jsou podmínky dobré aproximace splněny. Testová statistika se realizuje hodnotou 3,85268, p-hodnota je 0,42631, tedy na asymptotické hladině významnosti 0,05 hypotézu o normalitě nezamítáme. Podívejme se ještě na histogram s proloženou Gaussovou křivkou: Na záložce Základní výsledky zvolíme Graf pozorovaného a očekávaného rozdělení.




Je nutné upozornit,že při jiné volbě třídicích intervalů můžeme dostat zcela odlišné výsledky – vyzkoušejte sami.

**2. způsob provedení testu dobré shody:** ukážeme na jiném příkladu. 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.
**Návod:** Vytvoříme nový datový soubor o čtyřech proměnných $X_1$, $X_2$, nj, npj a pěti případech. Do proměnné $X_1$ napíšeme dolní meze třídicích intervalů (tj. 3,92 3,96 4 4,04 4,08), do proměnné $X_2$ napíšeme horní meze třídicích intervalů (tj. 3,96 4 4,04 4,08 4,12), do proměnné nj napíšeme pozorované četnosti (tj. 11 20 44 19 6) a konečně do proměnné npj uložíme teoretické četnosti tak, že do Dlouhého jména této proměnné napíšeme
=100*(INormal(X2;4,02;0,04)-INormal(X1;4,02;0,04))
Statistiky – Neparametrická statistika – Pozorované vs. očekávané $\chi 2$ – OK – Proměnné – Pozorované četnosti nj, očekávané četnosti npj – OK – Výpočet.

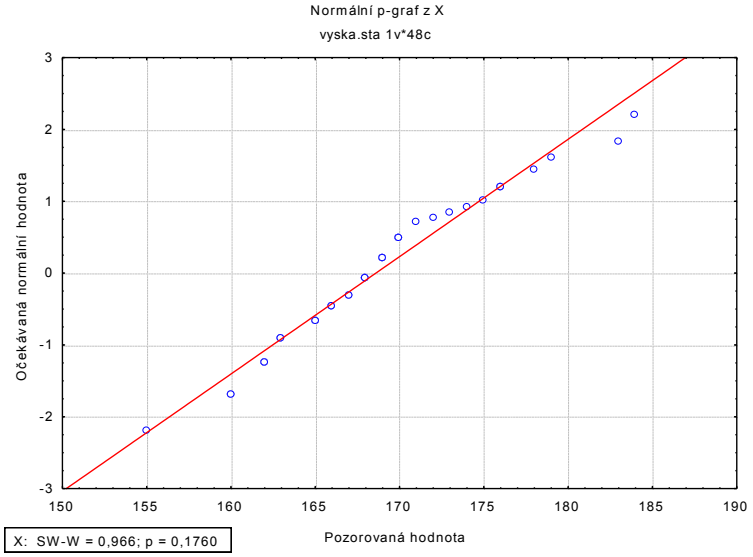
Pozorované vs. očekávané četnosti (Tabulka26)
Chi-Kvadr. = 6,706286 sv = 4 p = ,152251
POZN.: Nestejné součty pozor. a oček. četností

| Případ | pozorov. nj | očekáv. npj | P - O | (P-O)^2 /O |
|---|---|---|---|---|
| C: 1 | 11,0000 | 6,05975 | 4,94025 | 4,027562 |
| C: 2 | 20,0000 | 24,17303 | -4,17303 | 0,720398 |
| C: 3 | 44,0000 | 38,29249 | 5,70751 | 0,850706 |
| C: 4 | 19,0000 | 24,17303 | -5,17303 | 1,107030 |
| C: 5 | 6,0000 | 6,05975 | -0,05975 | 0,000589 |
| Sčt | 100,0000 | 98,75807 | 1,24193 | 6,706286 |

Testová statistika K se realizuje hodnotou 6,706286, avšak zde je uveden počet stupňů volnosti 4, což není v pořádku, neboť r-k-1 = 5 – 2 – 1 = 2. Odpovídající asymptotická p-hodnota není tedy spočtena správně. Otevřeme nový datový soubor o jedné proměnné a jednom případu a do jejího Dlouhého jména napíšeme =2*min(IChi2(6,706286;2);1-IChi2(6,706286;2))

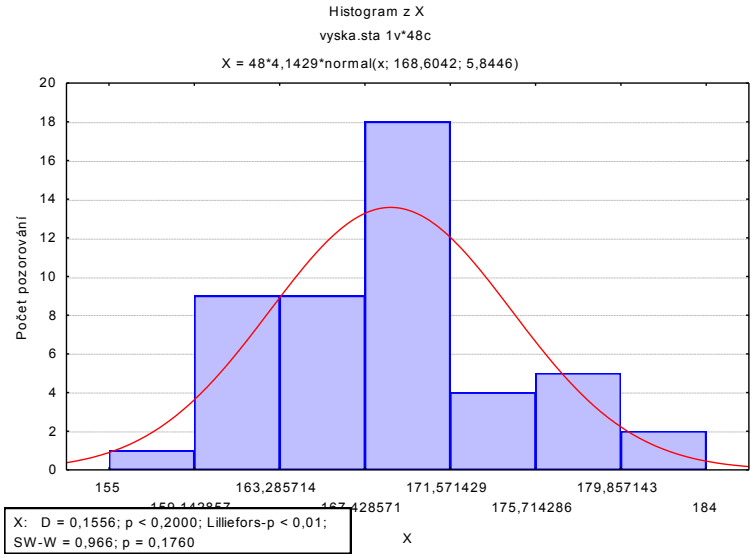
Dostaneme p-hodnotu 0,069949, tedy na asymptotické hladině významnosti 0,05 hypotézu o normalitě nezamítáme.

**Vykreslení N-P plotu pro data o výšce studentek:** Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnná X – odškrtneme Neurčovat průměrnou pozici svázaných pozorování, zaškrtneme S-W test – OK




Body se vyskytují v celkem těsné blízkosti přímky, lze je tedy považovat za realizace náhodného výběru z normálního rozložení.

**Upozornění:** K-S test a S-W test lze provést i při kreslení histogramu. Při vytváření histogramu zaškrtneme na záložce Detaily K-S test a S-W test.




**Samostatný úkol:** Testy normality a grafické ověření normality proveďte jak pro výšky studentek oboru národní hospodářství, tak pro výška studentek oboru informatiky.

**Pro kontrolu:**

Výsledky pro obor národní hospodářství:

| Proměnná | Testy normality (vyska.sta) Zhrnout podmínku: z=1 N | max D | Lilliefors p | W | p |
|---|---|---|---|---|---|
| X: vyska | 28 | 0,167473 | p < ,05 | 0,970969 | 0,606793 |

Vidíme, že Lilieforsova varianta K-S testu zamítá hypotézu o normalitě na hladině významnosti 0,05 (p-hodnota je menší než 0,05), zatímco S-W test hypotézu o normalitě nezamítá (p-hodnota je větší než 0,05).

Výsledky pro obor informatika:

| Proměnná | Testy normality (vyska.sta) Zhrnout podmínku: z=2 N | max D | Lilliefors p | W | p |
|---|---|---|---|---|---|
| X: vyska | 20 | 0,172301 | p < ,15 | 0,922747 | 0,111924 |

V tomto případě ani jeden z testů hypotézu o normalitě nezamítá na hladině významnosti 0,05.

**Upozornění:** V archivu závěrečných prací https://is.muni.cz/auth/th/77721/prif_m/ je uložena diplomová práce Dominika Grůzy „Ověřování normality“.

## Úlohy o výběrech z alternativního rozložení

### Úkol 1.: Vlastnosti výběrového průměru z alternativního rozložení

Mezi americkými voliči 60% osob volí republikány a 40% demokraty. Jaká je pravděpodobnost, že v náhodném výběru 100 amerických voličů budou voliči republikánů v menšině? Výpočet proveďte jak přesně, tak pomocí aproximace normálním rozložením.

**Návod:**

$X_1, ..., X_{100}$ je náhodný výběr z A(0,6), $X_i = 1$, když i-tá osoba volí republikány, $X_i = 0$ jinak, i = 1, ..., 100. Zavedeme statistiku $Y_{100} = X_1 + ... + X_{100}$, $Y_{100} \sim Bi(100; 0,6)$ (viz skripta Teorie pravděpodobnosti a matematická statistika, sbírka příkladů, příklad 8.10.),

$E(Y_{100}) = n\vartheta = 100.0,6 = 60$, $D(Y_{100}) = n\vartheta(1-\vartheta) = 100.0,6.0,4 = 24$ Označme $\Phi_{100}(y)$ distribuční funkci náhodné veličiny $Y_{100}$, $\Phi_{100}(y) = \sum_{t=} \binom{100}{t} ),6^{t}0,4^{100-}$ .

Přesný výpočet: $P(Y_{100} < 50) = P(Y_{100} \leq 49) = \Phi_{100}(49) = 0,016761686$.

Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =IBinom(49;0,6;100). Funkce IBinom(x;p;n) počítá hodnotu distribuční funkce rozložení Bi(n,p) v bodě x.

Přibližný výpočet: užijeme důsledek Moivreovy - Laplaceovy integrální věty (viz skripta Základní statistické metody, věta 6.3.1.1.). Nejdříve ověříme splnění podmínky dobré aproximace $n\vartheta(1-\vartheta) = 100.0,6.0,4 = 24 > 9$. Podmínka je splněna.

$P(Y_{100} < 50) = P(Y_{100} \leq 49) \approx \Phi(49)$, kde $\Phi(49)$ je hodnota distribuční funkce rozložení N(60; 24) v bodě 80.

Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =INormal(49;60;sqrt(24)).

Zjistíme, že $\Phi(49) = 0,012372$.

Přesný výpočet

| | 1 Prom1 |
|---|---|
| 1 | 0,016762 |

Aproximativní výpočet

| | 1 Prom1 |
|---|---|
| 1 | 0,012372 |

**Úkol 2.: Asymptotický interval spolehlivosti pro parametr $\vartheta$ alternativního rozložení**
Může politická strana, pro niž se v předvolebním průzkumu vyslovilo 60 z 1000 dotázaných osob, očekávat se spolehlivostí aspoň 0,95, že by v této době ve volbách překročila 5% hranici pro vstup do parlamentu?

**Návod:**
Zavedeme náhodné veličiny $X_1, ..., X_{1000}$, přičemž $X_i = 1$, když i-tá osoba se vysloví pro danou politickou stranu a $X_i = 0$ jinak, i = 1, ..., 1000. Tyto náhodné veličiny tvoří náhodný výběr z rozložení $A(\vartheta)$. V tomto případě n = 1000, m = 60/1000 = 0,06, $\alpha = 0,05$, $u_{1-\alpha} = u_{0,95} = 1,645$.
Ověření podmínky $n\vartheta(1-\vartheta) > 9$: parametr $\vartheta$ neznáme, musíme ho nahradit výběrovým průměrem. Pak 1000.0,06.0,94 = 56,4 > 9.
95% levostranný interval spolehlivosti pro $\vartheta$ je
$\left(m - \sqrt{\frac{m(1-m)}{n}}u_{1-\alpha}\ ;\ \infty\right) = \left(0,06 - \sqrt{\frac{0,06(1-0,06)}{1000}}u_{0,95}\ ;\ \infty\right)$ (viz skripta Základní statistické metody, důsledek 6.3.2.2.)
Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =0,06-sqrt(0,06*0,94/1000)*VNormal(0,95;0;1). Vyjde 0,047647.
S pravděpodobností přibližně 0,95 tedy $\vartheta > 0,03$. Protože tento interval zahrnuje i hodnoty nižší než 0,05, nelze vyloučit, že strana získá méně než 5% hlasů.

**Úkol 3: Testování hypotézy o parametru $\vartheta$ alternativního rozložení**
Určitá cestovní kancelář organizuje zahraniční zájezdy podle individuálních přání zákazníků. Z několika minulých let ví, že 30% všech takto organizovaných zájezdů má za cíl zemi X. Po zhoršení politických podmínek v této zemi se cestovní kancelář obává, že se zájem o tuto zemi mezi zákazníky sníží. Ze 150 náhodně vybraných zákazníků v tomto roce má 38 za cíl právě zemi X. Potvrzují nejnovější data pokles zájmu o tuto zemi? Volte hladinu významnosti 0,05.

**Návod:**
Máme náhodný výběr $X_1, ..., X_{150}$ z rozložení A(0,3). Testujeme $H_0$: $\vartheta = 0,3$ proti levostranné alternativě $H_1$: $\vartheta < 0,3$. V tomto případě je testovým kritériem statistika
$T_0 = \frac{M - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}}$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1) (viz skripta Základní statistické metody, věta 6.3.3.1.). Musíme ověřit splnění podmínky $n\vartheta(1-\vartheta) > 9$: 150.0,3.0,7 = 31,5 > 9. Vypočteme realizaci testového kritéria:
$\frac{m - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} = \frac{\frac{38}{150} - 0,3}{\sqrt{\frac{0,3(1-0,3)}{150}}} = -2,4722$. Kritický obor: $W = \langle -\infty, -u_{1-\alpha}\rangle = \langle -\infty, -1,645\rangle$.
Protože testové kritérium nepatří do kritického oboru, $H_0$ nezamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% tedy naše data neprokázala pokles zájmu zákazníků cestovní kanceláře o zemi X.
Vytvoříme datový soubor o dvou proměnných a jednom případu. Vypočteme realizaci testového kritéria tak, že do Dlouhého jména první proměnné zapíšeme odpovídající vzorec, tj. =(38/150-0,3)/sqrt(0,3*0,7/150). Do Dlouhého jména druhé proměnné napíšeme =VNormal(0,95;0;1), čímž získáme kvantil $u_{0,95}$ a testové kritérium porovnáme s opačnou hodnotou tohoto kvantilu.

| | 1<br>Prom1 | 2<br>Prom2 |
|---|---|---|
| 1 | -1,24722 | 1,644854 |

Protože testové kritérium není menší než opačná hodnota příslušného kvantilu, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.