Aplikovaná statistika (2)

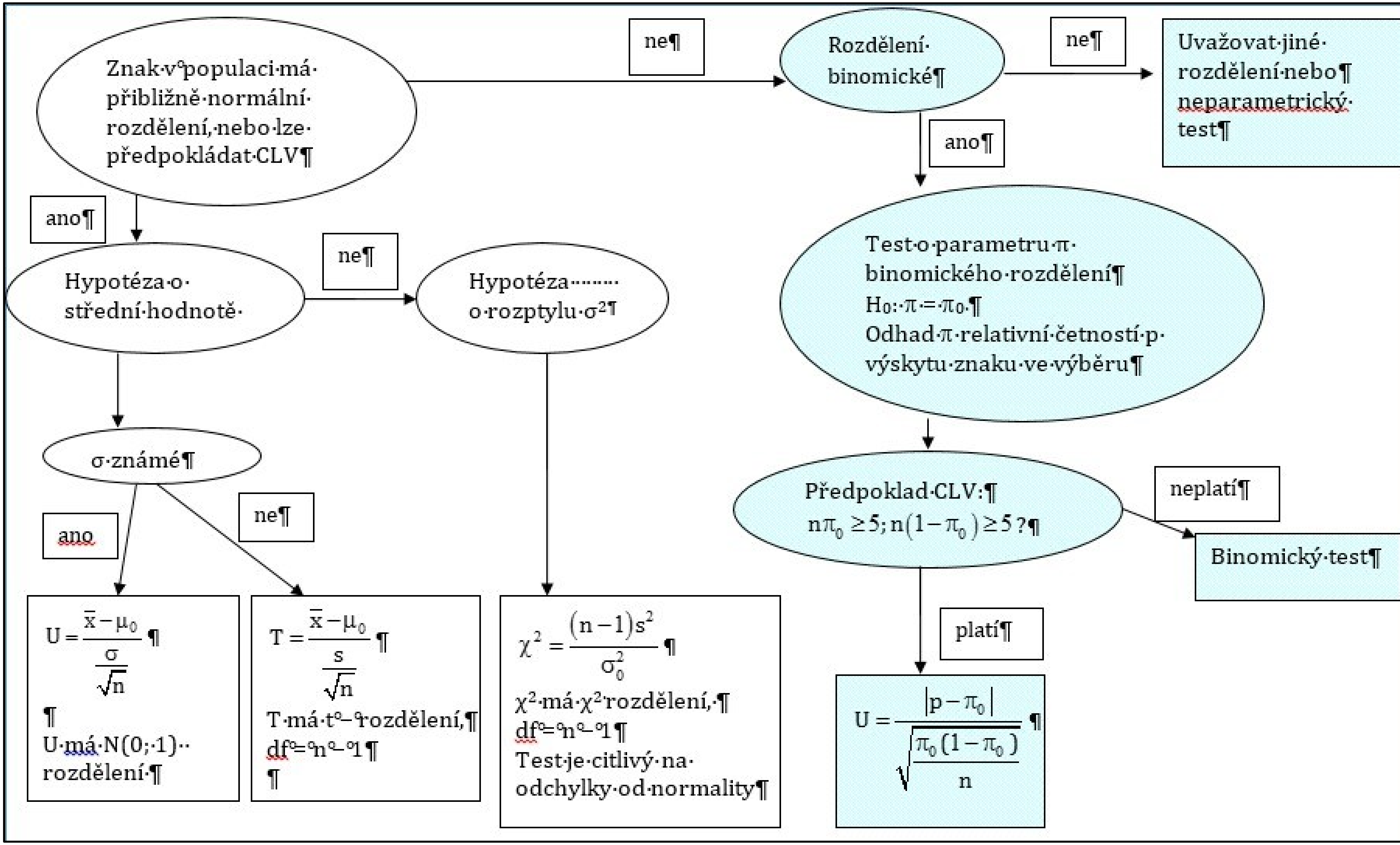
Hana Skalská

Testy hypotéz – hypotézy o jednom parametru:

Značení parametrů, rozhodovací oblast
Hypotézy pro jeden výběr jednovýběrové):

- Hypotéza o střední hodnotě
  Test a rozhodovací oblast pro různé alternativy
  Hodnota p a rozhodnutí o hypotéze
- Hypotéza o podílu (o parametru binomického rozdělení):
  Test, předpoklady a rozhodovací oblast
  Výpočet p hodnoty

Hypotézy pro jeden výběr: Přehled



Značení

Parametr populace	Název, případně označení pro odhad	Bodový odhad (pomocí výběru)
μ	Střední hodnota Odhad aritmetickým průměrem	$\bar{x}$
σ	Směrodatná odchylka	s
σ^2	Rozptyl	s^2
π	Pravděpodobnost výskytu jevu Odhad relativní četností	p
ρ	Koeficient korelace	r
β	Koeficient v regresi	b

Univerzita Hradec Králové
Fakulta informatiky a managementu

Hana Skalská

Aplikovaná statistika

GAUDEAMUS
2013

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Recenzenti:
Doc. RNDr. Zdeněk Boháč, CSc.
Mgr. Irena Hlavičková, Ph.D.

ISBN 978-80-7435-320-8

1.8.1 Hustota, distribuční funkce a kvantily

Uvažujeme **spojitou náhodnou veličinu** X, která může nabývat hodnot $x \in (-\infty; +\infty)$. Rozdělení pravděpodobnosti veličiny X je popsáno hustotou f(x) a distribuční funkcí F(x). Budou shrnuty základní pojmy a vlastnosti, které se využívají při rozhodování o hypotéze.

Hustota f(x) **rozdělení pravděpodobnosti spojité náhodné veličiny** X je nezáporná na celém definičním oboru veličiny X, její integrál přes celý definiční obor má hodnotu jedna.

Pravděpodobnost, že hodnoty x veličiny X jsou obsaženy na intervalu $(x_1; x_2)$, lze stanovit jako plochu, vymezenou intervalem a hustotou:

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(t)\,dt = P(x_1 < X \le x_2) = P(x_1 < X < x_2).$$

Distribuční funkce $F(x) = P(X \le x) = P(-\infty < X \le x) = \int_{-\infty}^{x} f(t)\,dt.$

Pro náhodnou veličinu X definovanou na intervalu $(-\infty; +\infty)$ s hustotou f(x) se hodnota distribuční funkce F(x) v bodě x rovná pravděpodobnosti výskytu hodnot veličiny X na intervalu $(-\infty, x)$.

Pravděpodobnost, že hodnoty x veličiny X jsou z intervalu $(x_1; x_2)$, lze vyjádřit pomocí distribuční funkce a platí $P(x_1 < X < x_2) = F(x_2) - F(x_1)$.

Pro spojitou náhodnou veličinu X v jednotlivém bodě x je $P(X = x) = 0$ na celém definičním oboru X. Pravděpodobnost hodnot na intervalu se tedy nezmění zahrnutím nebo nezahrnutím krajních bodů intervalu.

Pro **spojitá** rozdělení nezáleží na tom, zda uvažujeme ostrou nebo neostrou nerovnost. Například pro pravděpodobnost jevu doplňkového lze psát $P(X > x) = 1 - F(x) = P(X \geqq x)$.

Hodnotu $F(x)$ v bodě x lze interpretovat jako plochu pod křivkou (hustotou) rozdělení $f(x)$ na intervalu $(-\infty; x)$.

Kvantil x_p spojité náhodné veličiny X **je hodnota** pro kterou platí, že hodnoty menší než kvantil x_p mají pravděpodobnost p, současně hodnoty větší než x_p mají pravděpodobnost $(1 - p)$.

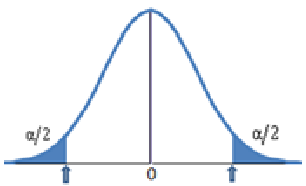
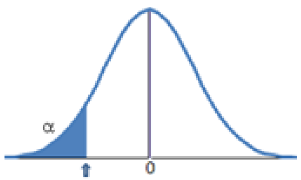
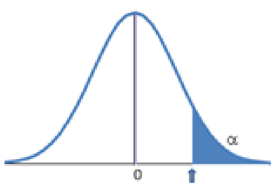
Kvantil „oddělí“ zvolený podíl p nejmenších hodnot v rozdělení pravděpodobnosti.

Při vyjádření pomocí distribuční funkce je x_p hodnota, pro kterou $p = P(X \leqq x_p) = F(x_p)$.

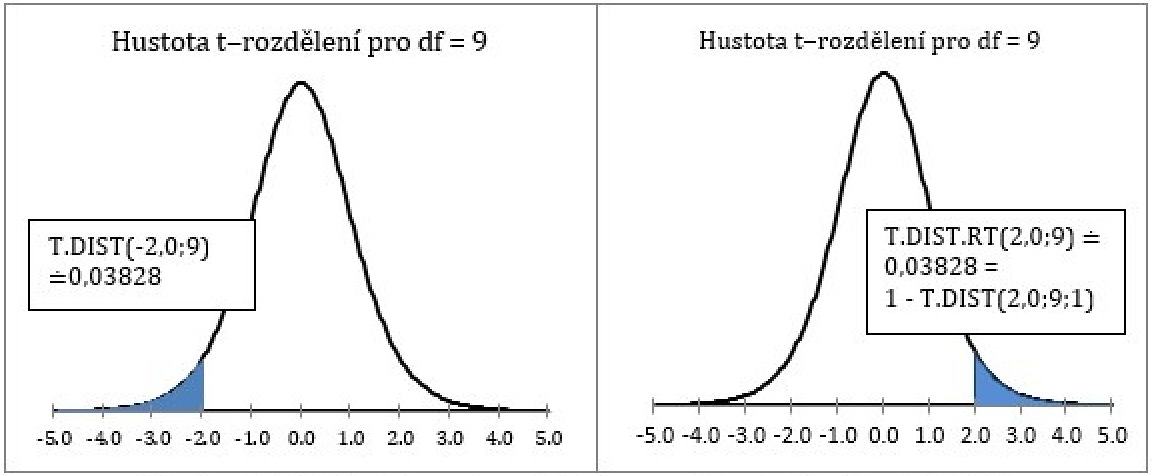
Percentil je kvantil vyjádřený v procentech. Podíl hodnot menších než x_p je p, současně $(1 - p)$ je podíl hodnot, které jsou větší než x_p.

Obecné vlastnosti kvantilu nezávisí na určitém tvaru rozdělení. Rozdělení spojitých náhodných veličin popisuje hustota $f(x)$ nebo distribuční funkce $F(x)$, jejich průběhem a volbou α jsou určené hodnoty x_α.

Oblasti zamítání pro různě formulované alternativní hypotézy

Příklad: Hypotézy o střední hodnotě	Oblast zamítání H_0 je zvýrazněna stínováním Hladina významnosti = α Kritická hodnota je označena šipkou ⬆
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ Oboustranná alternativa	
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$ Levostranná alternativa	
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$ Pravostranná alternativa	

Při stejné volbě hladiny významnosti α jsou různě vymezeny oblasti zamítání H_0 . Křivka naznačuje tvar rozdělení testového kritéria.



Obrázek 1.2 Hustota a kvantily t–rozdělení a funkce T.DIST

Zvýrazněná část plochy pod křivkou $f_{t\text{–}dist}(x)$ na obrázku 1.2 vlevo odpovídá pravděpodobnosti 0,03828, výskytu hodnot menších (více extrémních) než kvantil $t_{0,038276;9}$ = – 2 0. Na grafu vpravo je zvýrazněn kvantil $t_{0,96172;9}$ = 2,0 a plocha velikosti 0,03828, která odpovídá pravděpodobnosti hodnot větších než +2,0. Celková plocha pod křivkou má hodnotu jedna. Funkce T.DIST.RT (right tail) – viz kapitola 1.8.3.

Test hypotézy o střední hodnotě (1 výběr)

Předpoklady: Veličina má v populaci přibližně normální rozdělení $N(\mu, \sigma^2)$.
Populační rozptyl není známý, je odhadnutý z výběru.
Vzorek o rozsahu n získaný prostým náhodným výběrem.

$H_0 : \mu = \mu_0$ **proti alternativě** $H_1 : \mu \neq \mu_0$ **na hladině významnosti α.**

Testové kritérium $T = \frac{\overline{x} - \mu_0}{s_{\overline{x}}}$, kde směrodatná chyba odhadu $s_{\overline{x}} = \frac{s}{\sqrt{n}}$

T má při platnosti H_0 Studentovo t – rozdělení pro n – 1 stupňů volnosti.

Kritický obor testu H_0 proti oboustranné alternativě H_1

$$W_\alpha = \left\{ t : |t| > t_{1-\alpha/2,n-1} \right\} = \left\{ t : t < t_{\alpha/2,n-1} \quad \text{nebo} \quad t > t_{1-\alpha/2,n-1} \right\}.$$

Rozhodovací pravidlo I: H_0 se zamítá pro $|t| > t_{1-\alpha/2,n-1}$

Přesná hladina významnosti testu při dvoustranné alternativě

$$p = 2 \cdot \left[1 - F\left(|t|\right) \right] = 2 \cdot \left(1 - \int_{-\infty}^{|t|} f(x) dx \right)$$

t je hodnota testového kritéria T, F(t) je distribuční funkce t – rozdělení

Rozhodovací pravidlo II: H_0 se zamítá při $p \leq \alpha$.

Příklad

Výrobce koření dodává malá balení jednoho druhu koření. Deklaruje hmotnost balení 16 g. Zákazníci si v internetové diskusi stěžují, že toto nemusí být pravda a hmotnost je asi menší. Zástupce ligy pro ochranu zákazníků zajistil přesné měření náhodně vybraných 50 dávek. Průměrná hmotnost balení ve vzorku byla 14,9 g, výběrová směrodatná odchylka 3,9 g.

Hypotéza nulová a alternativní, hladina významnosti:

$$H_0 : \mu = 16, \;\; H_1 : \mu < 16, \quad \alpha = 0{,}05$$

Informace z náhodného výběru:

$$n = 50 \quad \overline{x} = 14{,}9 \quad s = 3{,}9$$

Testové kritérium $t = \frac{14{,}9 - 16}{3{,}9} \cdot \sqrt{50} = -1{,}994$

Kritická hodnota (kvantil t – rozdělení, 49 stupňů volnosti): $t_{0,05}(49) = -1{,}6765$

Rozhodovací pravidlo I: Hypotéza nulová se zamítá na hladině významnosti 0,05, přijímá se H_1. Deklarovaná hmotnost je vyšší, než skutečná. Stížnost zákazníků je oprávněná.

Příklad – pokračování: Přesná hladina významnosti

.

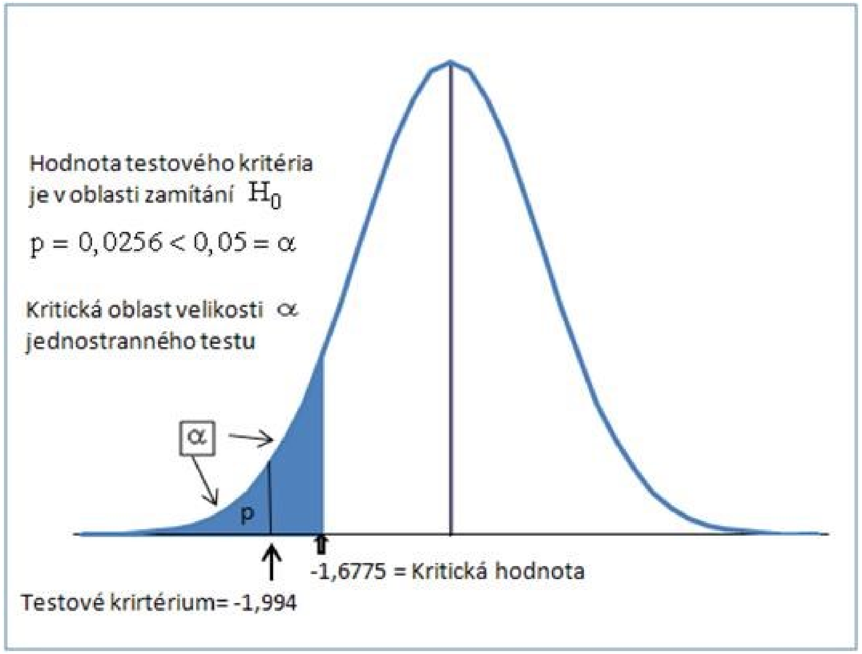
Výpočet pravděpodobnosti p (části plochy pod křivkou hustoty t-rozdělení)
f(x) je hustota, F(x) distribuční funkce t – rozdělení pro n – 1 stupňů volnosti.
TK je hodnota testového kritéria testu.

$$p = F(t) = \int_{-\infty}^{TK} f(x)\,dx$$

$$H_0 : \mu = 16 \quad H_1 : \mu < 16 \quad \alpha = 0{,}05$$

$$t = \frac{14{,}9 - 16}{3{,}9} \cdot \sqrt{50} = -1{,}994$$

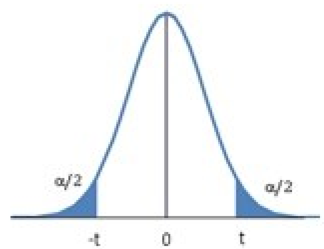
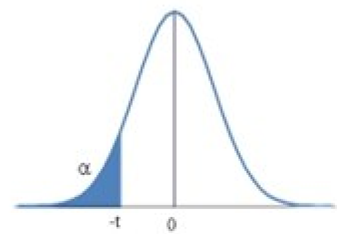
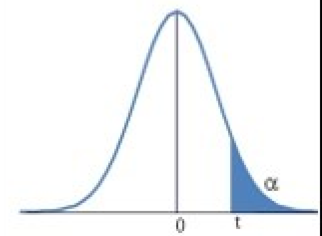
$$W_{0,05} = \{t; t < t_{0,05,50-1}\} = \{t; t < t_{0,05,50-1}\} = \\ = \begin{Bmatrix} t : t < -1{,}6765 \\ \text{MsExcel} = \text{T.INV}(0{,}05;49;1) \end{Bmatrix}$$



$$p = P(TK \le -1{,}994) = \int_{-\infty}^{-1{,}994} f(x)\,dx = F(-1{,}994)) = \text{T.DIST}(-1{,}994;49;1) = 0{,}0258$$

H_0 ze zamítá na hladině p = 0,0259

Rozhodovací oblast testu – oblast zamítání označená modře

A1. Hypotéza o střední hodnotě – jeden výběr			
H_0 a testové kritérium (TK)	$H_1: \mu \neq \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$
$H_0: \mu = \mu_0$ $T = \frac{\overline{x} - \mu_0}{s_{\overline{x}}},\ s_{\overline{x}} = \frac{s}{\sqrt{n}}$ T má rozdělení t pro dané stupně volnosti (df). Pro n > 100 má rozdělení N(0,1).			
Kritické hodnoty (KH) t–rozdělení, df = n – 1	$-t = t_{\alpha/2;df}$ $t = t_{1-\alpha/2;df}$	$-t = t_{\alpha;df}$	$t = t_{1-\alpha;df}$
Příklad pro α = 0,05, stupně volnosti = df = 9, různé alternativy			
t - rozdělení	KH = ± 2,2621	KH = – 1,8331	KH = 1,8331
MsExcel od v.2010	–t = T.INV(0,025;9) t = T.INV(0,975;9)	T.INV(0,05;9)	T.INV(0,95;9)
Kritické hodnoty pro N(0; 1) MsExcel od v.2010	KH = 1,9600 NORMSINV(0,975) NORM.S.INV(0,975)	KH = – 1,6449 NORMSINV(0,05) NORM.S.INV(0,05)	KH = 1,6449 NORMSINV(0,95) NORM.S.INV(0,95)

Hypotéza o parametru binomického rozdělení – jeden výběr

Alternativní znak (jev), odhadujeme pravděpodobnost výskytu.

Příklady: Výrobek má vadu (je bezvadný), domácnost má (nemá) připojení k internetu, zákazník uvažuje (neuvažuje) o zakoupení určitého výrobku.

Otázka: Zda neznámá pravděpodobnost π výskytu znaku v populaci, ze které byl vzorek vybraný, se rovná hypotetické hodnotě π_0.

Test je označovaný jako test o podílu (parametru binomického rozdělení).

Předpoklad: Prostý náhodný výběr n navzájem nezávislých prvků z populace, ve které neznáme zastoupení znaku.

Počet m prvků ve výběru, které vykazují sledovaný znak, má binomické rozdělení.

Relativní četnost $p = \frac{m}{n}$

odhaduje neznámý parametr π (pravděpodobnost výskytu jevu).

Hypotéza o podílu – jeden výběr

Testuje se

H_0: $\pi = \pi_0$ proti H_1: $\pi \neq \pi_0$ na hladině významnosti α.

Testové kriterium $U = \dfrac{|p - \pi_0|}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$

Testové kriterium s korekcí na nespojitost binomického rozdělení (sign = jen znaménko výrazu) – pro případy, když tento člen přispěje výrazněji k hodnotě testového kritéria.

$$U = \frac{p - \pi_0 + c}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \qquad c = \frac{\mathrm{sign}(\pi_0 - p)}{2n} \quad \text{pro} \quad |p - \pi_0| \geq \frac{1}{2n}$$

$$c = 0 \quad \text{pro} \quad |p - \pi_0| < \frac{1}{2n}$$

Testové kritérium má při platnosti H_0 normované normální rozdělení N(0,1).

Kritický obor testu H_0 proti oboustranné alternativě na hladině významnosti α

$$W_\alpha = \left\{u : |u| > u_{1-\alpha/2}\right\} = \left\{u : u < u_{\alpha/2} \quad \text{nebo} \quad u > u_{1-\alpha/2}\right\}$$

Vlastnosti testového kritéria plynou z konvergence binomického rozdělení k normálnímu rozdělení pro velká n. Test lze proto použít pouze když přibližně: $n \cdot \pi_0 \geq 5,\ \ n \cdot (1 - \pi_0) \geq 5.$

Test hypotézy o parametru binomického rozdělení

„Přesná“ hladina významnosti při dvoustranné alternativě:

$$p = 2 \cdot \left[1 - \Phi(u)\right]$$

u je hodnota testového kritéria U,
$\Phi(x)$ je **distribuční funkce** normovaného normálního rozdělení N(0,1).
V Excelu funkce $_{2010}$=NORM.S.DIST(x;1) $_{2007}$=NORMSDIST(x;1).

Předpokladům konvergence vyhovují výběry velkých rozsahů a případy, když π není blízké 0 nebo 1.

Pokud předpoklady nejsou splněné, doporučují se **přesné testy** (vycházejí z binomického rozdělení).

Mírné porušení předpokladů se řeší **přidáním korekčního členu** v čitateli testového kritéria (eliminuje nepřesnost, že nespojité binomické rozdělení je aproximováno spojitým normálním).

Příklad A - Hypotéza o parametru binomického rozdělení

Firma má podíl na trhu 56 %. Po reklamní kampani bylo ve vzorku 500 náhodně vybraných nákupů zjištěno, že výrobek dané firmy zakoupilo 298 zákazníků, ostatní zakoupili výrobek konkurenčních firem. Lze odvodit, že se firmě změnil podíl na trhu (dvoustranná alternativa)?

$H_0: \pi = 0{,}56$ $H_1: \pi \neq 0{,}56$ $\alpha = 0{,}05$? Použití limitní věty: 500*0,56 i 500*0,54 >5 √

$$p = \frac{298}{500} = 0{,}596 \qquad u = \frac{\left|0{,}596 - 0{,}560 + \frac{\operatorname{sign}(0{,}560 - 0{,}596)}{2 \times 500}\right|}{\sqrt{\frac{0{,}560(1-0{,}560)}{500}}} = \frac{\left|0{,}036 - \frac{1}{1000}\right|}{0{,}022199} = 1{,}5766$$

$$\text{NORM.S.INV}(0{,}975) = 1{,}96 \qquad \text{NORM.S.INV}(0{,}025) = -1{,}96$$

$$\text{Kritická hodnota} = 1{,}96 > 1{,}5766 = \text{Testové kritérium}$$

$$W_{0,05} = \left\{u : |u| > u_{1-0,05/2}\right\} = \left\{u : u < u_{0,025} = -1{,}96 \quad \text{nebo} \quad u > u_{0,975} = 1{,}96\right\}$$

H_0 nezamítáme na hladině významnosti 0,05.
Nelze zamítnout hypotézu, vzorek mohl být vybrán z populace s podílem 0,56.
Nebyla prokázána statisticky významná změna podílu na trhu po provedené reklamní kampani (na hladině významnosti 0,05).

Příklad A – pokračování: Rozhodovací pravidlo II

Protože jsou splněné předpoklady limitní věty, tak testové kritérium má přibližně normované normální rozdělení N(0,1).

Hodnota p (při jednostranné alternativě by se uvažovala pouze jedna strana příslušného typu rozdělení) je:

$$p = 2\times\left[1-\Phi\left(1,5766\right)\right] = 2\times\left[1-\mathrm{NORM.S.DIST}\left(1,5766\right)\right] = 2\,\mathrm{x}\,0,0574 = 0,1148 > 0,05$$

nebo: $p = 2\times\Phi\left(-1,5766\right)] = 2\times\mathrm{NORM.S.DIST}\left(-1,5766\right) = 2\times 0,0574 = 0,1148 > 0,05$

Protože hodnota p je větší, než zvolená hladina významnosti, nulová hypotéza se nezamítá.

Nepodařilo se prokázat vliv marketingové kampaně na změnu původního 56% podílu na trhu.

Na hladině významnosti α = 0,05 je odchylka od hypotetické hodnoty náhodná.

Příklad B – Zadání a řešení pro jednostrannou alternativu

Firma má podíl na trhu 56 %. **Od reklamní kampaně** **očekává jeho zvýšení**.
Po reklamní kampani bylo ve vzorku 500 náhodně vybraných nákupů zjištěno, že výrobek dané firmy zakoupilo 298 zákazníků, ostatní zakoupili výrobek konkurenčních firem. Lze odvodit, že se firmě zvýšil podíl na trhu?
Příklady A a B se liší ve formulaci alternativní hypotézy, rozdílné je stanovení W_α a p.

$H_0: \pi = 0{,}56$ $H_1: \pi > 0{,}56$ $\alpha = 0{,}05$? Použití limitní věty je možné: 500*0,56 i 500*0,54 >5 √

$$p = \frac{298}{500} = 0{,}596 \qquad u = \frac{0{,}596 - 0{,}560 + \dfrac{\operatorname{sign}(0{,}560 - 0{,}596)}{2 \times 500}}{\sqrt{\dfrac{0{,}560(1-0{,}560)}{500}}} = \frac{0{,}036 - \dfrac{1}{1000}}{0{,}022199} = 1{,}5766$$

$$\text{NORM.S.INV}(0{,}95) = 1{,}64 \qquad W_{0,05} = \{u : u > u_{0.95}\} = \{u : u > 1{,}64\}$$

$$\text{Kritická hodnota} = 1{,}64 > 1{,}5766 = \text{Hodnota testového kritéria}$$

H_0 nezamítáme na hladině významnosti 0,05.
Nelze zamítnout nulovou hypotézu, vzorek mohl být vybraný z populace s podílem 0,56.
Nebylo prokázané statisticky významné zvýšení podílu na trhu po reklamní kampani (na hladině významnosti 0,05).

Příklad B – pokračování: Rozhodovací pravidlo II

Protože jsou splněné předpoklady limitní věty, tak testové kritérium má přibližně normované normální rozdělení N(0,1).

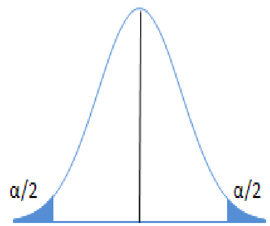
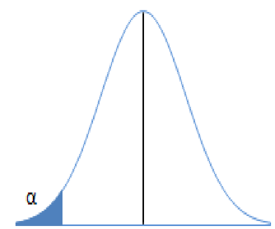
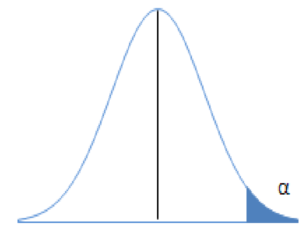
Přesná hodnota p při jednostranné alternativě (uvažuje se pouze jedna strana teoretického rozdělení) je:

$$p = 1 - \Phi\left(1{,}5766\right) = 1 - \mathrm{NORM.S.DIST}\left(1{,}5766\right) = 1 - 0{,}9426 = 0{,}0574 > 0{,}05$$

Protože hodnota p je větší, než zvolená hladina významnosti, nulová hypotéza se nezamítá.

Nepodařilo se na hladině významnosti 0,05 prokázat, že vlivem marketingové kampaně se statisticky významně zvýšil původní 56% podíl firmy na trhu (p = 0,0574). Pozorované zvýšení ve vzorku lze vysvětlit náhodnou chybou.

Rozhodovací oblast testu – oblast zamítání označená modře

Hypotéza o parametru binomického rozdělení (jeden výběr) – limitní test			
$H_0 : \pi = \pi_0$	$H_1 : \pi \neq \pi_0$	$H_1 : \pi < \pi_0$	$H_1 : \pi > \pi_0$
Testové kritérium má při platnosti H_0 rozdělení N(0,1) $U = \dfrac{p - \pi_0}{\sqrt{\dfrac{\pi_0 (1 - \pi_0)}{n}}}$ kde $p = \dfrac{m}{n}$ Ověření platnosti CLV: $n\pi_0 \geq 5$, $n(1 - \pi_0) \geq 5$			
Příklad:	α = 0,05	α = 0,05	α = 0,05
	KH = ± 1,9600 NORM.S.INV(0,025) NORM.S.INV(0,975)	KH = – 1,6449 NORM.S.INV(0,05)	KH = 1,6449 NORM.S.INV(0.95)
Hodnota p	$2[1 - \Phi(\|u\|)]$	$\Phi(u)$	$1 - \Phi(u)$

$\Phi(u)$ distribuční funkce normovaného normálního rozdělení N(0; 1).