

Základy štatistiky

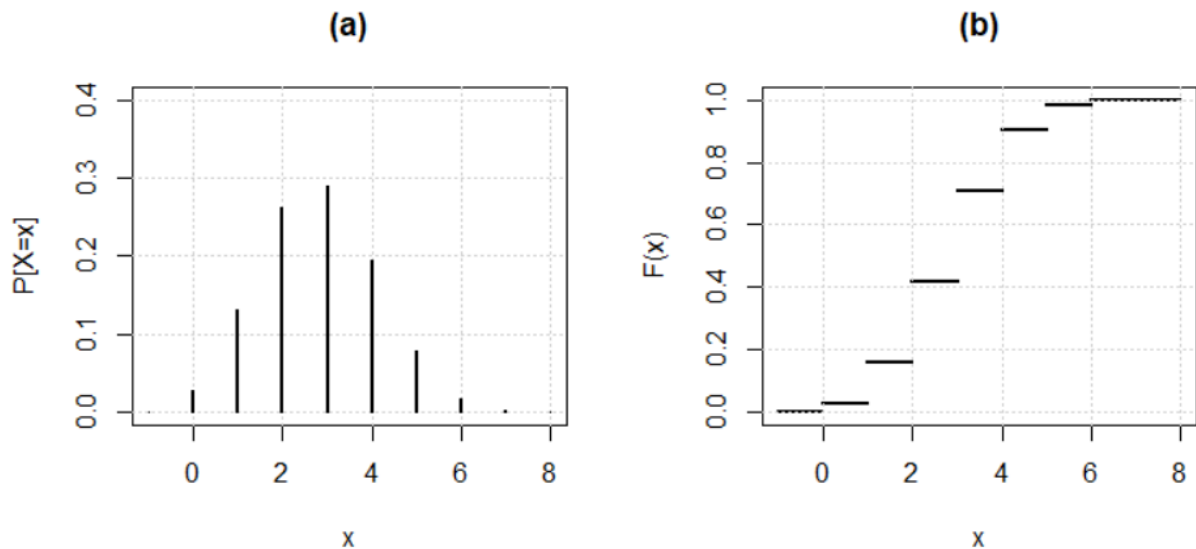
Jakub Skaloš

1 Binomické rozdelenie

Definícia 1. Hovoríme, že diskretná náhodná premenná X má binomické rozdelenie s parametrami $p \in (0, 1)$ a $n \in \mathbb{N}$, ak pre každé $k = 0, 1, \dots, n$ platí

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Túto skutočnosť značíme $X \sim \text{Bin}(n, p)$.



Obr. 1: a) sú zobrazené pravdepodobnosti ($\sim$ hustota), b) je graf kumulatívnej distribúcie

Funkcia rozdelenia pravdepodobnosti (Probability mass function, PMF). Kumulatívna distribučná funkcia (CDF) $F(x)$ funkcie má v každom bode súčet pravdepodobností po bod x vrátane, čiže:

$$F(x) = P[X \leq x]$$

Definícia 2. Pre diskretné premenné je očakávaná hodnota $E(X)$ definované ako:

$$E(X) = \sum_{\forall x \in \mathbb{Z}} P(X = x) \cdot x$$

a disperzia $D(X)$ definovaná ako:

$$D(X) = E(X^2) - E(X)^2$$

Veta 1. Očakávaná hodnota $E(X) = np$, disperzia $D(X) = \sigma^2 = np(1 - p)$:

Dôkaz. Rozpíšeme cez vzorec pre $E(X)$: ($k = 0$ zanedbáme, komb. číslo rozpíšeme)

$$E(X) = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n k \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} =$$

(k do faktoriálu, n pred z faktoriálu, p mimo mocniny) (substitúcia $i = k - 1$)

$$= np \sum_{k=1}^n \frac{(n-1)!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{n-k} = np \sum_{i=0}^{n-1} \frac{(n-1)!}{((n-1)-i)!i!} p^i (1-p)^{(n-1)-i}$$

čo sa rovná $E(X) = np$, pretože:

$$\sum_{i=0}^k \binom{k}{i} p^i q^{k-i} = (p+q)^k$$

Pre disperziu:

$$E(X(X-1)) = \sum_{k=2}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} =$$

$$= n(n-1)p^2 \sum_{k=2}^n \frac{(n-2)!}{(n-k)!(k-2)!} p^{k-2} (1-p)^{n-k} = n(n-1)p^2 \sum_{i=0}^{n-2} \frac{(n-2)!}{(n-2-i)!i!} p^i (1-p)^{n-2-i} = n(n-1)p^2$$

a preto:

$$\begin{aligned} D(X) &= E(X^2) - (E(X))^2 = E(X(X-1)) + E(X) - (E(X))^2 \\ &= n(n-1)p^2 + np - n^2p^2 = np(1-p) \end{aligned}$$

□

Úloha 1. Aká je šanca, že z 10 hodov férovou 6-stennou kockou mi padne buď 3 alebo 4 menej ako 3-krát?

Riešenie.

$$\begin{aligned} P[X < 3] &= P[X = 0] + P[X = 1] + P[X = 2] = \\ &= \binom{10}{0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^{10} + \binom{10}{1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^9 + \binom{10}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^8 = \frac{2^{10} + 10 \cdot 2^9 + \frac{10 \cdot 9}{2} \cdot 2^8}{3^{10}} = \frac{2^{11} + 15 \cdot 2^8}{3^9} \end{aligned}$$

2 Binomický test

Ak vieme, že $X \sim \text{Bin}(n, p)$, ale nepoznáme p , iba hodnotu n . Pýtame sa otázky (overujeme hypotézy) typu $p \neq p_0$? alebo $p < p_0$? ($> \leq \geq$)

Definícia 3. Pod p -value označujeme štatistickú pravdepodobnosť, že za predpokladu nulovej hypotézy H_0 nastane prípad extrémny rovnako, alebo viac. Nulovú hypotézu zvykneme zamietť, ak hodnota p -value je menšia ako 5%.

Úloha 2. Máme neférovú kocku, ale nevieme, ktoré čísla ako preferuje. Číslo 6 nám zo 100 pokusov padlo 4-krát. Vieme s istotou povedať, že na kocke je šanca hodenia 6-ťky menej ako 10%?

Riešenie. Naša nulová hypotéza H_0 bude $H_0 : p \geq 10\%$. Túto hypotézu optimálne chceme vyvrátiť, a to vtedy, ak $p \ll 10\%$. V tom prípade bude platiť alternatívna hypotéza $H_1 : p < 10\%$

$$p\text{-value} = \sum_{x=0}^4 P[X = x | p = 10\%] = \sum_{x=0}^4 \binom{100}{x} 0.1^x \cdot 0.9^{100-x} \approx 0.0237110827 < 0.05$$

Nulovú hypotézu H_0 zamietame a teda štatisticky potvrdzujeme alternatívnu hypotézu H_1

Dôsledok 1. Prišli sme na to, ako funguje jednostranný Binomický test.

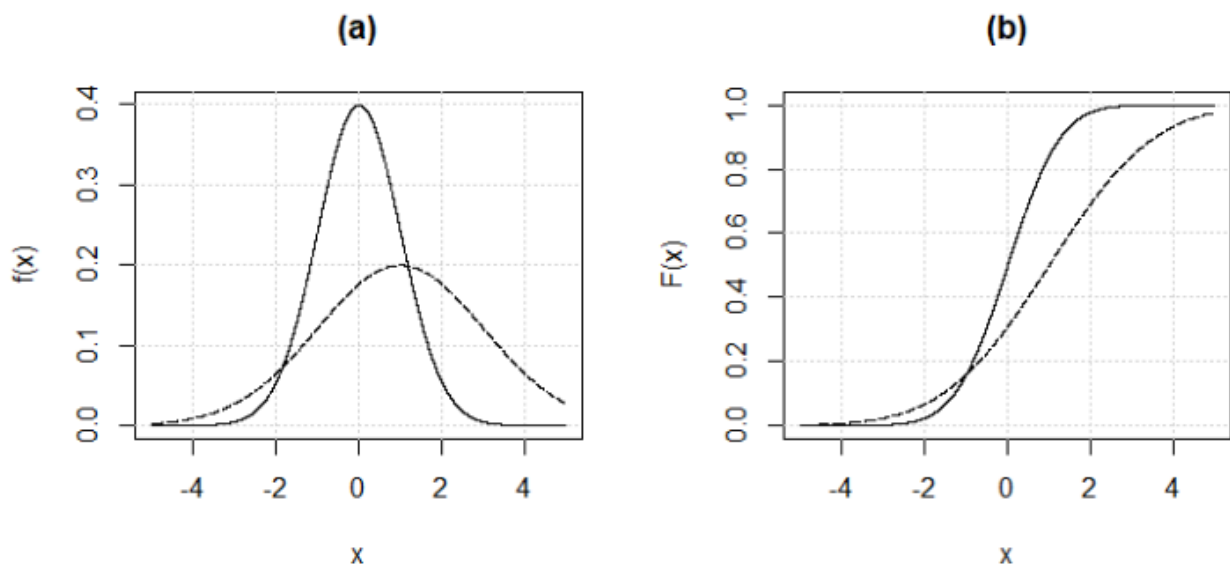
3 Normálne rozdelenie

Definícia 4. Nech $\mu \in \mathbb{R}$ a $\sigma > 0$. Hovoríme, že náhodná premenná X má *normálne rozdelenie* s parametrami μ a σ^2 , ak je X spojitá náhodná premenná s hustotou

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

pre všetky $x \in \mathbb{R}$. Túto skutočnosť značíme $X \sim N(\mu, \sigma^2)$. Ak $X \sim N(0, 1)$, tak hovoríme, že X má *štandardizované normálne rozdelenie*, nazývané aj „normalizované“ normálne rozdelenie. Distribučnú funkciu náhodnej premennej X s rozdelením $N(0, 1)$ budeme označovať symbolom Φ .

Nech $X \sim N(\mu, \sigma^2)$ a nech $a, b \in \mathbb{R}$, $a \neq 0$. Potom platí: $aX + b \sim N(a\mu + b, a^2\sigma^2)$. Špeciálne: $\frac{X-\mu}{\sigma} \sim N(0, 1)$, ak $a = \frac{1}{\sigma}$ a $b = \frac{\mu}{\sigma}$.



Obr. 2: Hustota $f(x)$ a distribučná funkcia $F(x)$ náhodnej premennej s rozdelením $N(0, 1)$ (plná čiara) a $N(1, 2)$ (prerušovaná čiara).

Čo sa správa normálne? Všetko, na čo vplýva veľa malých faktorov, ktoré podobne vplývajú a ktoré sa nasčítajú (môžeme uvažovať o binomickom rozdelení, kde p je veľmi malé a $n \rightarrow \infty$). To znamená napríklad rozdelenia populácie podľa biologických parametrov, pretože na ne vplývajú gény - výška, hmotnosť, IQ... Taktiež veci spojené s prírodou - veľkosť úrody, počet upršaných dní v roku...

Úloha 3. Predpokladajme, že výsledok štandardizovaného testu IQ pre človeka náhodne zvoleného z populácie má rozdelenie $N(\mu, \sigma^2)$, kde $\mu = 100$ a $\sigma = 15$. Predpokladajme tiež, že máme k dispozícii program, ktorý numericky počíta funkčné hodnoty funkcie Φ . Určme pravdepodobnosť, že výsledok IQ testu človeka náhodne vybraného z populácie bude nižší ako 130.

Riešenie. Riešenie má dva kroky - prvý vyjadriť si pravdepodobnosť a druhý že upravíme tak, aby sme mohli použiť funkciu Φ :

$X \sim N(100, 15^2)$, tak platí $\frac{X-100}{15} \sim N(0, 1)$, preto

$$P[X < 130] = P\left[\frac{X - 100}{15} < \frac{130 - 100}{15}\right] = \Phi(2) \approx 0,97725.$$

4 Vlastnosti vzorky

Definícia 5. Nech $X_1, \dots, X_n$ je postupnosť nezávislých rovnako rozdelených náhodných premenných s rozdelením Q . Potom hovoríme, že $X_1, \dots, X_n$ je náhodná vzorka z rozdelenia Q . Číslo n nazývame rozsah výberu.

Definícia 6. Definujeme:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Hodnota $\bar{X}$ sa nazýva výberový priemer a hodnota S^2 výberový rozptyl. Hodnota $\sqrt{S^2}$ sa nazýva výberová smerodajná odchýlka. Hodnota S^2 je definovaná len pre $n \geq 2$. (Niektorým autorom, napr. J. Andělovi sa nepáči označovať S^2 ako výberový rozptyl, lebo nesplňa vlastnosti rozptylu a teda je to máťúce označenie)

Veta 2. Nech $X_1, \dots, X_n$ je náhodná vzorka z rozdelenia, ktoré má konečný stred μ a konečný rozptyl σ^2 . Potom platí:

$$E(\bar{X}) = \mu \quad D(\bar{X}) = \frac{\sigma^2}{n} \quad E(S^2) = \sigma^2.$$

Dôkaz. Prvé tvrdenie možno dokázať priamo úpravou, použijeme $E(X_i) = \mu$

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu$$

21 Druhé tvrdenie je zložitá na dôkaz bez potrebného aparátu na počítanie s disperziou, použijeme vzorec: $D(\sum_{i=1}^n Y_i) = \sum_{i=1}^n \sum_{j=1}^n \text{cov}(Y_i, Y_j)$ a $\text{cov}(Y_i, Y_i) = D(Y_i) = \sigma^2$

$$D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \left[\sum_{i=1}^n D(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j) \right] = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

Tretie tvrdenie, rozpíšeme si cez vzorec:

$$\begin{aligned} E(S^2) &= E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \\ &= \frac{1}{n-1} E\left(\sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right) = \frac{1}{n-1} E\left(\sum X_i^2 - 2\bar{X} \sum X_i + n\bar{X}^2\right) \end{aligned}$$

pričom $\sum X_i = n \cdot \bar{X}$, čiže:

$$= \frac{1}{n-1} E\left(\sum X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1} \cdot n \cdot E(X_i^2) - n \cdot E(\bar{X}^2)$$

použijeme vzorec $D(Y) = E(Y^2) - E(Y)^2$, čo znamená že $E(Y^2) = D(Y) + E(Y)^2$:

$$\begin{aligned} &= \frac{1}{n-1} (n \cdot D(X_i) + n \cdot E(X_i)^2 - n \cdot D(\bar{X}) - n \cdot E(\bar{X})^2) = \frac{1}{n-1} (n\sigma^2 + n\mu^2 - n\frac{\sigma^2}{n} - n\mu^2) = \\ &= \frac{1}{n-1} (n-1) \cdot \sigma^2 = \sigma^2 \end{aligned}$$

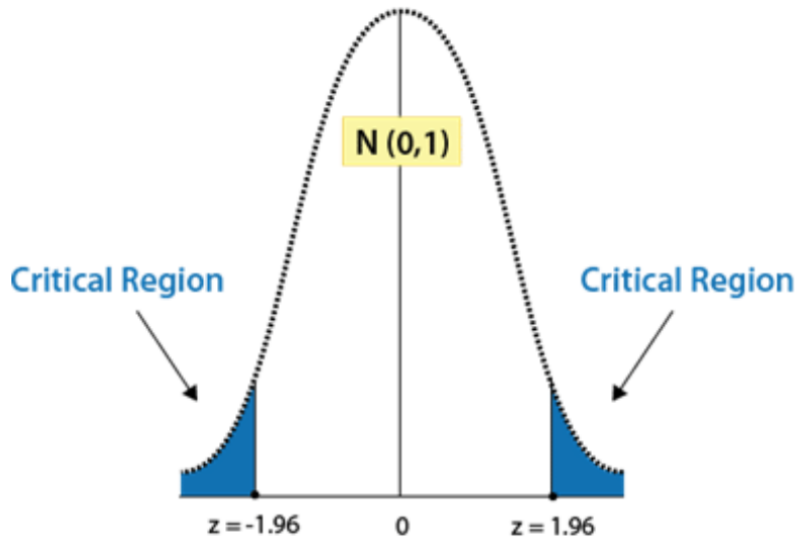
□

Dôsledok 2. $\bar{X}$ a S^2 sú dobré odhady pre μ a σ^2 respektívne.

Veta 3. (bez dôkazu) Nech $X_1, \dots, X_n$ sú iid a pochádzajú z $N(\mu, \sigma^2)$. Potom $\bar{X}$ a S^2 sú od seba nezávislé a $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, čiže $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ (Možné znenie Centrálnej limitnej vety)

5 Z-score test

Pýtame sa otázky na populačný priemer, typu či populačný priemer μ je $< \mu_0$, $> \mu_0$ alebo $\neq \mu_0$. To robíme na základe priemeru vzorky $\bar{X}$. **Dôležité: musíme poznať populačný rozptyl σ^2**



Obr. 3: Kritické hodnoty pre obojstranný 5% Z-test

Úloha 4. Chceme overiť hypotézu, že priemer výšky mladého človeka nie je 160cm, zo vzorky ľudí prítomných na prednáške. Z múdrej knihy vieme, že výška mladého človeka sa správa normálne a rozdelenie má rozptyl 7cm.

Riešenie. Vypočítame si výberový priemer. Za nulovú hypotézu si určíme $H_0 : \mu = 160cm$, tú chceme vyvrátiť, čiže dokázať platnosť $H_1 : \mu \neq 160cm$. To sa deje v prípade, ak $\bar{X} < 160cm$, alebo $\bar{X} > 160cm$. Dáta najprv normalizujeme, potom sa pozrieme na z-skóre:.

$$z\text{-skóre} = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} = \frac{\bar{X} - 160cm}{7cm} \sqrt{n},$$

za predpokladu, že $\sigma = 5cm$ a z nulovej hypotézy $\mu = 160cm$

Zamietame, ak nám z-skóre padne do kritickej oblasti (critical region, čiže hodnoty z-skóre, kde p-value je menšia ako 5%), čiže z-skóre < -1.96 alebo > 1.96 . Hranica (v našom prípade 1.96) sa nazýva kritická hodnota pre 2,5%, označovaná aj $u_{2,5\%}$. Je určená tak, aby $1 - \Phi(u_a) = a$.

Definícia 7. Chyba 1. druhu (type 1 error) nastáva vtedy, keď testom zamietneme pravdivé tvrdenie (odsúdime nevinného). Chyba 2. druhu (type 2 error) nastáva vtedy, keď testom nezamietneme nepravdivé tvrdenie (kriminálnika neodsúdime).

Veta 4. Nech $X_1, \dots, X_n$ sú iid z $N(\mu, \sigma^2)$, pričom σ^2 poznáme. Ak testujeme uvedené H_0 vs H_1 pomocou uvedeného testu, tak

$$P(\text{chyba 1. druhu}) = 5\%.$$

Dôkaz.

$$\begin{aligned} P(\text{chyba 1. druhu}) &= P(\text{zamietneme } H_0 | H_0 \text{ platí}) = \\ &= P(z < -u_{2,5\%}) + P(u_{2,5\%} < z) = 2 \cdot P(u_{2,5\%} < z) = 2 \cdot 2,5\% = 5\% \end{aligned}$$

□

Definícia 8. Ak $P(\text{chyba 1. druhu}) = \alpha$ (alebo ak $P(\text{chyba 1. druhu}) \leq \alpha$ a nijakou menšou konštantou, sa ohraničiť nedá), tak test sa volá *testom na hladine významnosti α* (level of significance α).

6 Ďalšie testy

Už sme mali príklad jednostranného Binomického testu a obojstranného Z-score testu. Jednostranné, aj obojstranné vieme spraviť pre akúkoľvek testovanú štatistiku, záleží či sa pozeráme na kritickú oblasť len na jednej strane, alebo na oboch. Máme však neskutočne veľa druhov testov, každý je na iný účel. Vymenujem niekoľko druhov, ktoré sa bežne používajú:

6.1 Dvojvýberový Z-test

Používa sa na otázky typu $\bar{X} < \bar{Y}$ (jednostranný), alebo $\bar{X} \neq \bar{Y}$ (obojstranný), ak máme 2 sady iid dát: $X_1, \dots, X_n$ a $Y_1, \dots, Y_m$, ktoré pochádzajú z normálnych rozdelení so známymi rozptylami - napríklad porovnáваме výšky Čechov a Slovákov. Vtedy môžeme povedať, že náhodná premenná $\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$. Ak znormalizujeme, naša testovacia štatistika bude z-skóre $= \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1)$.

6.2 Párový Z-test

V párovom Z-teste máme 2 sady dát $X_1, \dots, X_n$ a $Y_1, \dots, Y_n$, ktoré sú však po pároch závislé (napríklad stav pred užitím lieku a po užití, výška ráno a výška večer...). Vyrobitme si náhodnú premennú D , kde $D_i = X_i - Y_i$, potrebujeme vopred poznať aj σ_D^2 . Následne pokračujeme ako pri štandardnom jednovýberovom Z-teste.

6.3 T-testy

T-testy využívajú Studentovo rozdelenie s n stupňami voľnosti, ktoré aproximuje Normálne rozdelenie (limitne sa k nemu blíži ak $n \rightarrow \infty$). Má dve výhody oproti Z-testu: funguje spoľahlivo aj pri malých n , a netreba vedieť rozptyl celého rozdelenia, ten sa odhadne pomocou výberového rozptylu. Vzorce pre t-skóre pre rôzne druhy T-testov:

- jednovýberový:

$$\text{t-skóre} = \frac{\bar{X} - \mu_0}{S} \sqrt{n} \sim t_{n-1}$$

- dvojvýberový, pričom predpokladáme rovnaký rozptyl:

$$\text{t-skóre} = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n \cdot m}{n + m}} \sim t_{n+m-2}$$

$$\text{spoločný výberový rozptyl: } S^2 = \frac{S_X^2(n-1) + S_Y^2(m-1)}{n+m-2}$$

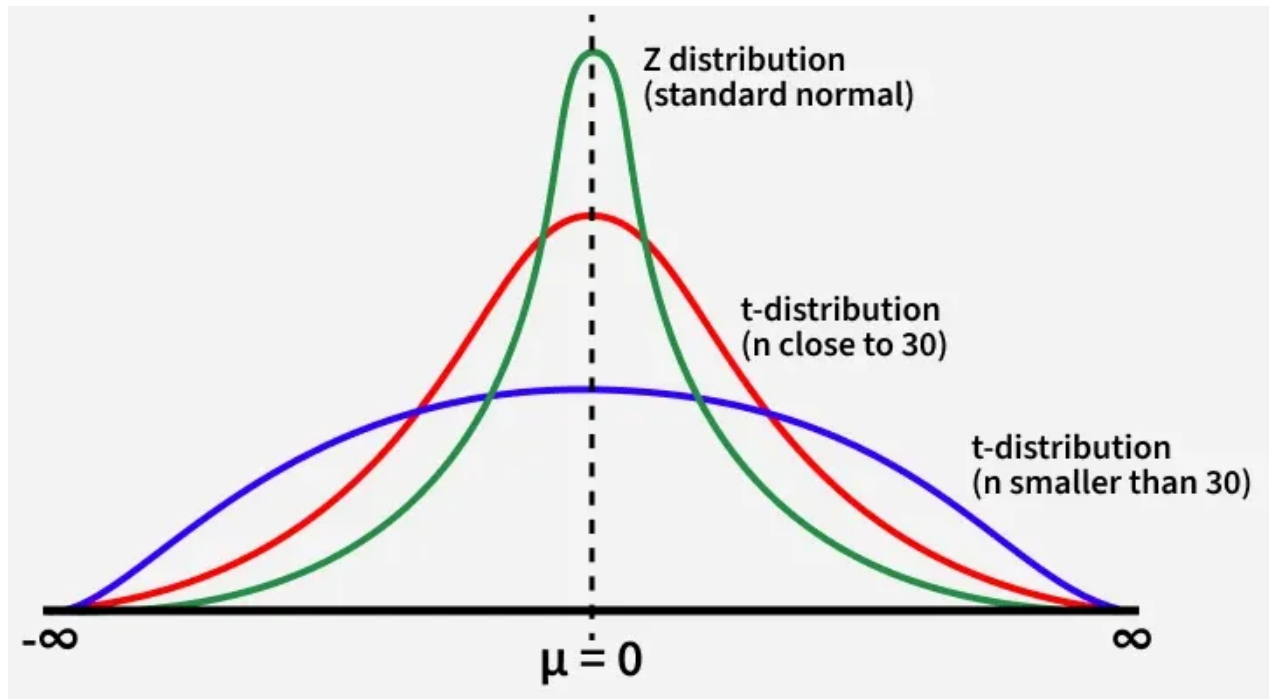
- dvojvýberový, rôzne rozptyly (Welchov T-test):

$$\text{t-skóre} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim t_{df}$$

$$\text{degrees of freedom: } df = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{\left(\frac{s_X^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_Y^2}{m}\right)^2}{m-1}}$$

- párový:

$$t\text{-skóre} = \frac{\bar{D} - \mu_0}{S_D} \sqrt{n} \sim t_{n-1}$$



Obr. 4: Studentovo rozdelenie

6.4 Likelihood ratio test

Likelihood ratio test (LRT) sa používa na otázky typu: *aké rozdelenie pravdepodobnejšie vygenerovalo túto sadu dát?* Používajú sa na to vierohodnosti $L_\theta(X)$, definované ako:

$$L_\theta(X) = L(\theta|X) = P(X|\theta),$$

Pričom θ sú parametre, ktoré určujú rozdelenie. Ak testujeme dve hypotézy: $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, H_0 zamietame v prípade, keď $L_{\theta_0}(X) \ll L_{\theta_1}(X)$

6.5 Testy normality

Testy normality sa používajú na overenie, či dáta pochádzajú z normálneho rozdelenia. Sú dôležité najmä pri rozhodovaní, či môžeme použiť parametrické testy (napr. T-testy), ktoré predpokladajú normalitu dát. Medzi najpoužívanéjšie testy patria:

- **Shapiro–Wilkov test** — veľmi citlivý test pre malé aj stredné výbery. Nulová hypotéza tvrdí, že dáta sú z normálneho rozdelenia. Testovacia štatistika porovnáva zoradené dáta s teoretickými kvantilmi normálneho rozdelenia.

- **Kolmogorov–Smirnov test** — porovnáva empirickú distribučnú funkciu s teoretickou distribučnou funkciou normálneho rozdelenia. V praxi sa často používa jeho modifikácia Lillieforsov test, keď parametre normálneho rozdelenia nepoznáme.

Testy normality sú často doplnené grafickými metódami, ako sú napríklad histogramy, ktoré poskytujú vizuálny pohľad na odchýlky od normálneho tvaru.

6.6 Neparametrické testy

Neparametrické testy nepredpokladajú konkrétne rozdelenie dát (napr. normálne) a sú vhodné pri malých výberoch alebo pri porušení predpokladov parametrických testov. Príklady zahŕňajú:

- Wilcoxonov test (pre porovnanie mediánov)
- Mann–Whitneyov U-test (alternatíva k dvojvýberovému t-testu)