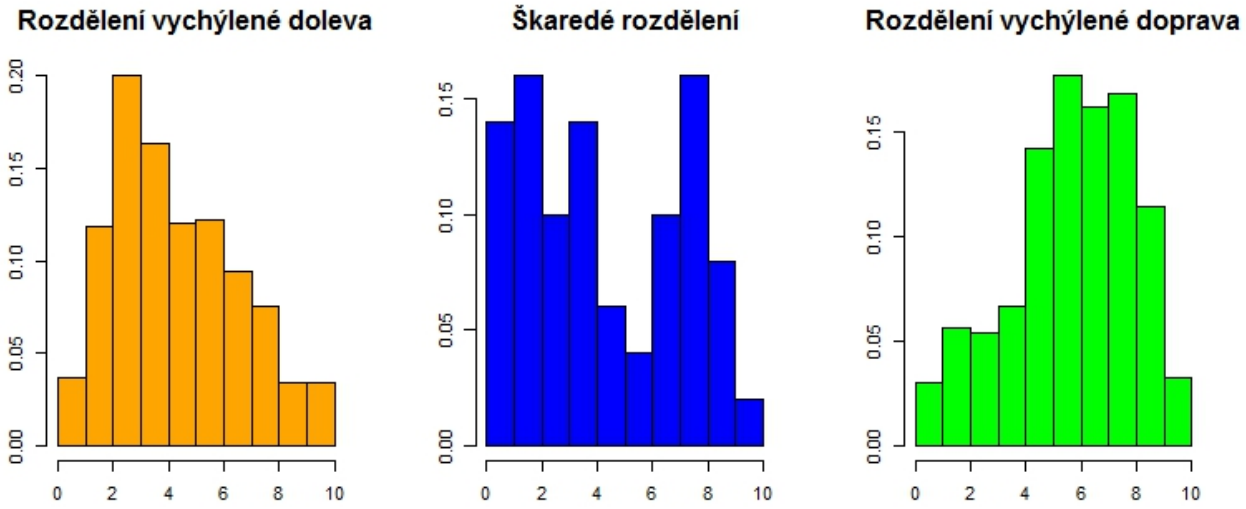
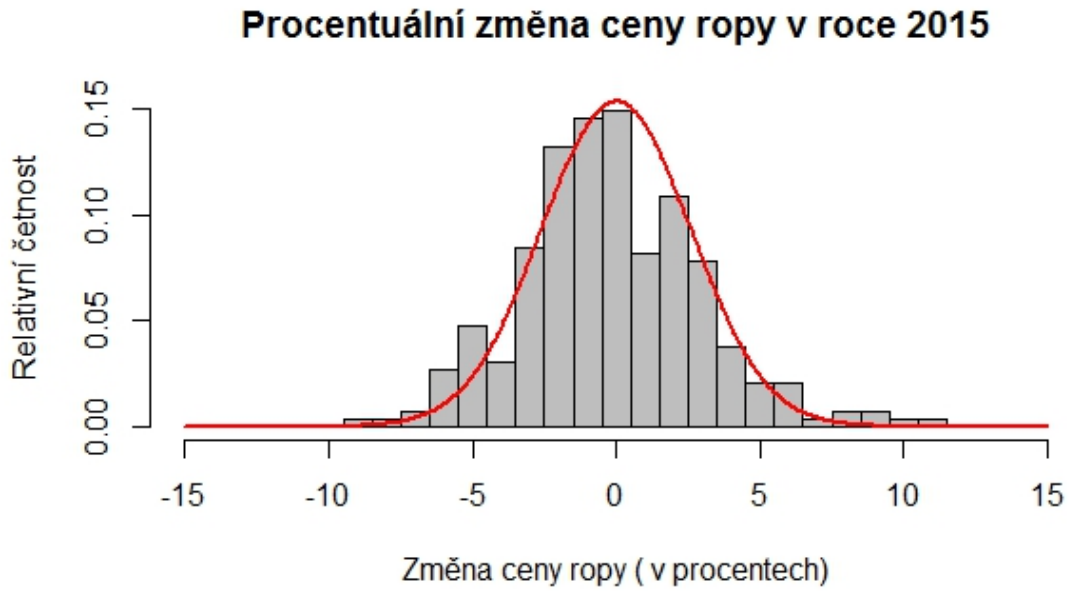
I Normální rozložení a odvozená rozložení

I.I Normální rozložení

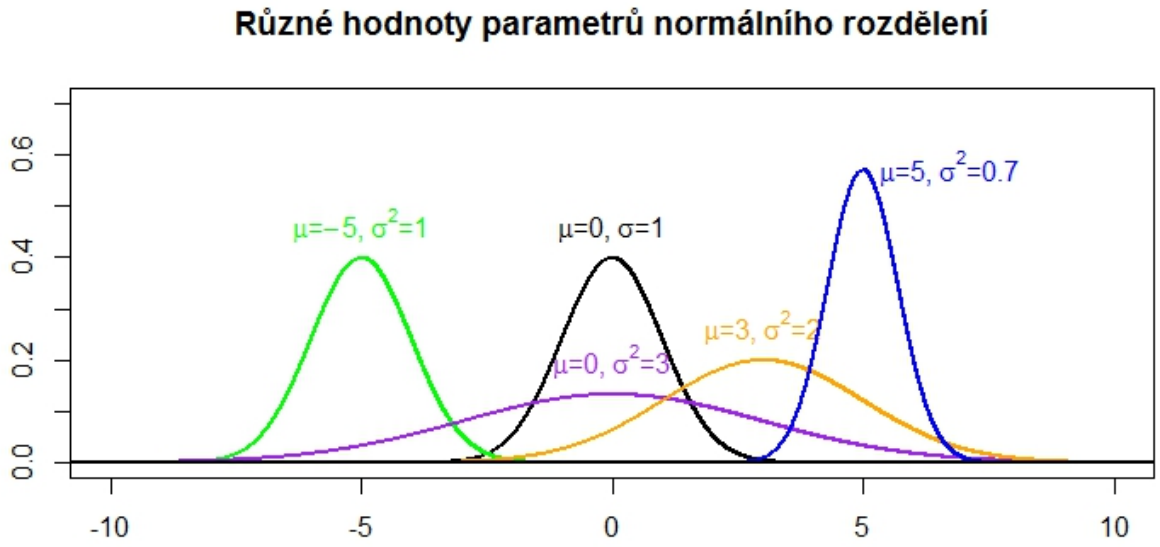
Data, se kterými pracujeme, pocházejí z různých rozložení. Mohou být vychýlena (doleva popř. doprava), nebo v nich není na první pohled vidět žádná tendence.



Často se ale setkáváme s daty, která jsou symetricky rozložená okolo jedné hodnoty. Příkladem je denní procentuální změna ceny ropy na mezinárodních trzích během roku 2015. Nejčastěji se cena změnila velmi málo, zatímco prudké změny byly velmi vzácné.



Podobnou tendenci sleduje spousta náhodných veličin, například výška (hmotnost) člověka, IQ, počet bodů získaných v testu, množství srážek... Tento jev nás vede k zavedení tzv. normálního rozdělení, které je charakterizováno dvěma parametry μ a σ. To, že náhodá veličina X pochází z normálního rozdělení, značíme $X \sim N(\mu, \sigma^2)$. Parametr μ je střední hodnota, ve které nabývá hustota normálního rozdělení maxima, parametr σ^2 je rozptyl, charakterizující variabilitu náhodné veličiny X. Čím je σ^2 menší, tím spíše se realizace náhodné veličiny X budou blížit střední hodnotě. Někdy se také normálnímu rozložení říká Gaussovo rozložení a grafu hustoty normálního rozložení gaussova (nebo zvonová) křivka (anglicky bell curve).



Hustota $f(x)$ náhodné veličiny X z normálního rozdělení $N(\mu, \sigma^2)$ je dána vztahem

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

pro všechna $x \in \mathbb{R}$. Distribuční funkce $F(x)$ má pak tvar

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}\frac{(t-\mu)^2}{\sigma^2}} \,\mathrm{d}t. \tag{2}$$

Normální rozdělení má několik užitečných vlastností:

- Pokud $X \sim N(\mu, \sigma^2)$, pak $E(X) = \mu = \tilde{x} = \hat{x}$ (tj. střední hodnota, medián i modus jsou rovny μ, z toho také plyne, že je hustota symetrická podle přímky $x = \mu$) a rozptyl $D(X) = \sigma^2$. Směrodatná odchylka je tedy σ.
- Normální rozdělení zachovává lineární transformaci – pokud $X \sim N(\mu, \sigma^2)$, pak pro $a, b \in \mathbb{R}$ platí, že $(a + bX) \sim N(a + b\mu, b^2\sigma^2)$. Pravdivá je dokonce i silnější verze, normální rozdělení dokonce zachovává lineární kombinace...

- Pokud $X_1, X_2 \dots X_n$ jsou stochasticky nezávislé náhodné veličiny takové, že $X_i \sim N(\mu_i, \sigma_i^2)$ pro $i = 1, 2, \dots, n$, pak náhodná veličina Y vzniklá lineární kombinací $Y = \sum_{i=1}^{n}(a_i + b_i X_i)$ má zase normální rozdělení $Y \sim N(\sum_{i=1}^{n}(a_i + b_i\mu_i), \sum_{i=1}^{n} b_i^2\sigma_i^2)$.

Bohužel pro distribuční funkci náhodné veličiny X, která má normální rozdělení $N(\mu, \sigma^2)$ neexistuje explicitní formule bez integrálu, do které by stačilo dosadit hodnoty μ, σ^2 a x, abychom snadno spočítali $F(x)$. Ve skutečnosti lze hodnotu $F(x)$ počítat pouze numericky. Nicméně pro tzv. standardizované normální rozdělení jsou hodnoty distribuční funkce i kvantily tabelizované. Vhodnou transformací pak dokážeme spočítat hodnoty distribuční funkce $F(x)$ pro libovolné normální rozdělení.

Standardizované normální rozdělení se používá tak často, že si vysloužilo vlastní značení. Náhodnou veličinu, která pochází z standardizovaného normálního rozdělení bývá zvykem značit $U \sim N(0, 1)$ a její distribuční funkci $\Phi(x)$.

Důležité tvrzení

Předpokládejme, že $X \sim N(\mu, \sigma^2)$. Pak $U = \frac{X-\mu}{\sigma}$ je standardizovaná normální náhodná veličina, tj. $U \sim N(0, 1)$.

Interpretační poznámka. Pokud chceme z obecné normální náhodné veličiny X vytvořit standardizovanou normální náhodnou veličinu, tak od X odečteme střední hodnotu μ, čímž se nám vrchol hustoty posune do 0 a podělíme směrodatnou odchylkou σ, čímž graf hustoty přeškálujeme (změníme měřítko).

V tabulkách se nacházejí hodnoty distribuční funkce $\Phi(u)$ pro $u \geq 0$. Pro záporná u se používá přepočtový vztah, $\Phi(-u) = 1 - \Phi(u)$. Pro α-kvantil standardizovaného normálního rozdělení se používá značení u_α. Ty jsou tabelovány pro $\alpha \geq 0,5$. Ostatní α-kvantily spočítáme ze vztahu $u_\alpha = -u_{1-\alpha}$.

Jak se tato transformace používá v praktických výpočtech ukazuje následující příklad.

Motivační příklad. Je známo, že denní počet platících zákazníků, kteří navštíví e-shop je náhodná veličina, která se řídí normálním rozložením se střední hodnotou $\mu = 1286$ a rozptylem $\sigma^2 = 7584$. Jaká je pravděpodobnost, že:

a) v daném dni bude mít e-shop aspoň 1400 platících zákazníků?

b) v daném dni nakoupí v e-shopu více, než 1200, ale méně, než 1500 zákazníků?

Řešení. Označíme X náhodnou veličinu, která udává počet platících zákazníků za jeden den v e-shopu. Podle zadání $X \sim N(1286, 7584)$. V zadání a) nás zajímá $P(X \geq 1400) = 1 - P(X \leq 1400) + P(X = 1400)$. Víme, že $P(X = 1400) = 0$, protože X je spojitá náhodná veličina.Dále upravíme tak, abychom získali výraz, obsahující distribuční funkci standardizovaného normálního rozdělení.

$$P(X \geq 1400) = 1 - P(X \leq 1400) + P(X = 1400) = 1 - P(X \leq 1400) =$$
$$= 1 - P(\underbrace{\frac{X-\mu}{\sigma}}_{U \sim N(0,1)} \leq \frac{1400-\mu}{\sigma}) = \dots$$

Dosadíme známe hodnoty a vyčíslíme.

$$\dots = 1 - P(U \leq \frac{1400 - 1286}{\sqrt{7584}}) \doteq 1 - P(U \leq 1,31) = 1 - \Phi(1,31).$$

Podle tabulek je $\Phi(1,309) = 0.905$, a tedy hledaná pravděpodobnost $1-0,905 = 0,095 = 9,5\%$.

V řešení b) hledáme $P(1200 < X < 1500)$. Postup je zcela analogický.

$$P(1200 \leq X \leq 1500) = P(X \leq 1500) - P(X \leq 1200) =$$
$$= P(\frac{X-\mu}{\sigma} \leq \frac{1500-\mu}{\sigma}) - P(\frac{X-\mu}{\sigma} \leq \frac{1200-\mu}{\sigma})$$

Dosadíme za μ a σ.

$$\dots \doteq P(U \leq 2,46) - P(U \leq -0,99) = \Phi(2,46) - \Phi(-0,99)$$

Tyto hodndoty nalezneme v tabulce, pro $\Phi(-0,99)$ využijeme přepočtový vztah $\Phi(-0,99) = 1 - \Phi(0,99)$. Dohromady tedy

$$\dots = \Phi(2,46) - (1 - \Phi(0,99)) = 0,993 - (1 - 0,839) = 0,832 = 83,2\%.$$

Pravděpodobnost, že e-shop navštíví během daného dne 1200 až 1500 zákazníků je 83,2%.

Důležité tvrzení: Pravidlo 3σ

Mějme normální náhodnou veličinu X se střední hodnotou μ a směradatnou odchylkou σ. Pak platí:

- $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68\%$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95\%$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 99,7\%$

Interpretační poznámka. Pravidlo 3σ se dá používat k jednoduchému testu normality dat – spočítáme výběrový průměr $\overline{X}$ (odhad střední hodnoty) a výběrovou směrodatnou odchylku s. Pokud data pochází z normálního rozdělení, tak musí přibližně platit procentuální zastoupení, jaké udává pravidlo 3σ s tím, že místo teoretických hodnot μ a σ použijeme jejich odhady.

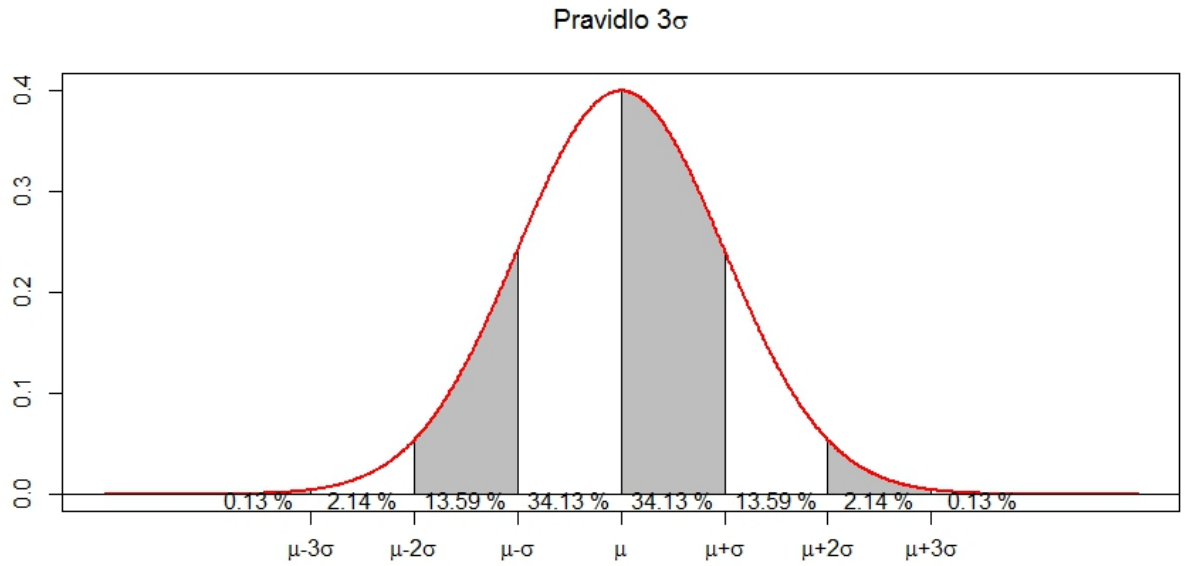
Naopak, Pravidlo 3σ nám říká, jak často bude docházet k extrémním hodnotám náhodné veličiny.

Motivační příklad. Procentuální denní změna ceny ropy je normálná náhodná veličina $X \sim N(0,2,5^2)$. Tedy $\sigma = 2,5$ a pravděpodobnost, že se cena ropy změní přes noc o více, než 5% je podle Pravidla 3σ přibližně 1-0,95=0,05=5%.

Pravděpodobnost, že změna ceny ropy bude dokonce větší, než 7,5% je dokonce menší, než 0,3%. Naopak pravděpodobnost, že se cena ropy změní jen málo, tj. max o 2,5%, je 68%.

I.II Rozložení pravděpodobnosti odvozená od normálního

Zatímco normální rozdělení používáme pro modelování rozličných náhodných veličin, následující trojice rozložení se používá pro testování statistických hypotéz – náhodné veličiny, které by se řídily těmito rozděleními se v praxi příliš nevyskytují. Všechna tato rozložení vznikla algebraickými úpravami ze standardizovaného normálního rozložení.



Pearsonovo χ^2 rozložení

Jako první uvedeme Pearsonovo rozložení, kterému se také říká χ^2 [chí kvadrát] rozložení, zejména v anglické literatuře.

Nový pojem: Pearsonovo rozložení

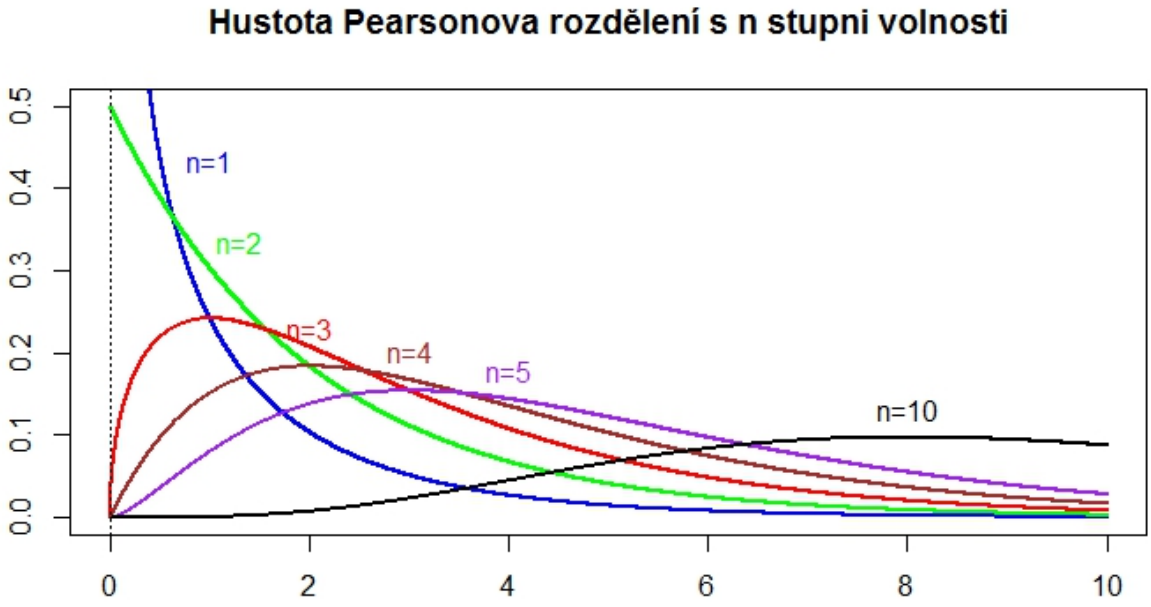
Předpokládejme, že $U_1, U_2, \ldots, U_n$ jsou stochasticky nezávislé standardizované normální veličiny, tj. $U_i \sim N(0,1)$ pro $i = 1, 2, \ldots, n$. Náhodná veličina

$$V = \sum_{i=1}^{n} U_i^2$$

má pak Pearsonovo rozložení, což značíme $V \sim \chi^2(n)$, parametr n nazýváme stupně volnosti.

Protože druhá mocnina reálného čísla je vždy nezáporná, tak i součet druhých mocnin reálných čísel je nezáporný. Z toho plyne, že náhodná veličina, řídící se Pearsonovým rozdělením je nezáporná. Všimněte si, že s rostoucím počet stupňů volnosti n se posouvá poloha maxima hustoty Pearsonova roložení doprava.

Užitečná poznámka. Kvantily Pearsonova rozdělení se počítají numericky, nejčastěji používané hodnoty jsou tabelizované. α-kvantil Pearsonova rozložení s n stupni volnosti značíme $\chi^2_\alpha(n)$. Pro velká n se používá aproximace $\chi^2_\alpha(n) \approx \frac{1}{2}(u_\alpha + \sqrt{2n-1})^2$, kde u_α je α-kvantil standardizovaného normálního rozložení.



Studentovo t-rozložení

Dalším důležitým rozložením je Studentovo t-rozložení nebo jen t-rozložení. V názvu píšeme velké S, protože jej objevil matematik William Sealy Gosset, který publikoval pod pseudonymem Student.

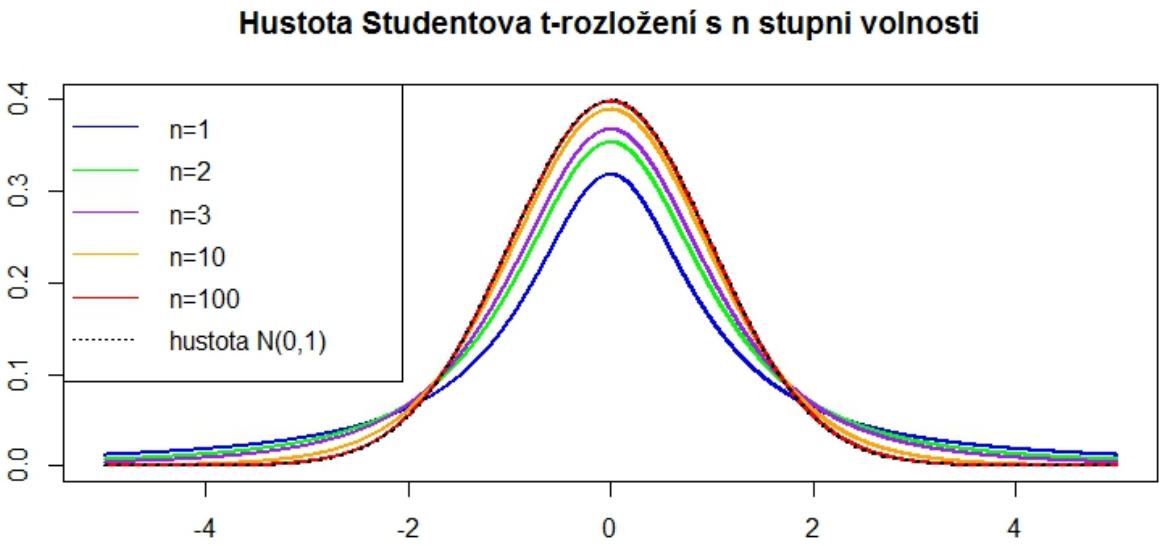
Nový pojem: Studentovo t-rozložení

Předpokládejme, že U a V jsou stochasticky nezávislé náhodné veličiny takové, že $U \sim N(0,1)$ a $V \sim \chi^2(n)$. Pak náhodná veličina

$$T = \frac{U}{\sqrt{\frac{V}{n}}}$$

má Studentovo t-rozložení s n stupni volnosti, značíme $T \sim t(n)$.

Grafem hustoty Studentova t-rozdělení je podobná zvonová křivka, jako u normálního rozdělení. S rostoucí hodnotou parametru n se Studentovo t-rozdělení blíží standardizovanému normálnímu rozdělení. Opět, α-kvantil Studentova t-rozdělení značíme $t_\alpha(n)$, tyto kvantily můžeme počítat numericky nebo vyhledat v tabulce, kde jsou uvedeny pro $\alpha \geq 0,5$. Ze symetrie hustoty Studentova t-rozdělení vyplývá, že pro $\alpha < 0,5$ můžeme využít přepočtový vztah $t_\alpha(n) = -t_{1-\alpha}(n)$.



Fisher-Snedecorovo rozdělení

Poslední rozdělení, které budeme při testování hypotéz potřebovat, je Fisher-Snedecorovo rozdělení, zkráceně F-rozdělení.

Nový pojem: Fisher-Snedecorovo rozdělení

Máme-li dvě stochasticky nezávislé náhodné veličiny V_1 a V_2, které pocházejí z Pearsonova rozdělení s n_1, resp. n_2 stupni volnosti, tj. $V_1 \sim \chi^2(n_1)$, $V_2 \sim \chi^2(n_2)$, pak náhodná veličina

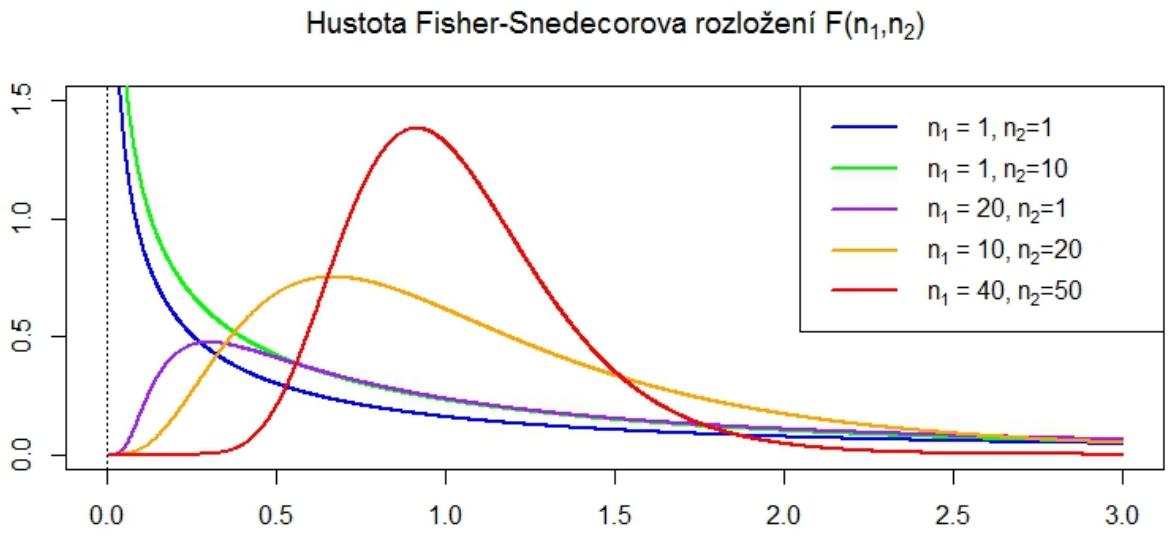
$$F = \frac{\frac{V_1}{n_1}}{\frac{V_2}{n_2}} = \frac{V_1}{V_2}\frac{n_2}{n_1}$$

má Fisher-Snedecorovo rozdělení. Parametr n_1 nazýváme počet stupňů volnosti čitatele, parametr n_2 počet stupňů volnosti jmenovatele. Značíme $F \sim F(n_1, n_2)$.

Náhodná veličina F, která se řídí Fisher-Snedecorovým rozdělením je nezáporná, protože vznikla jako podíl dvou nezáporných náhodných veličin – průměrů čtverců ze standardizovaných normálních rozdělení. Pro rostoucí hodnoty n_2 se hustota pravděpodobnosti Fisher-Snedecorova rozdělení blíží hustotě Pearsonova χ^2 rozdělení s n_1 stupni volnosti.

Pro α-kvantil Fisher-Snedecorova rozdělení s n_1 stupni volnosti čitatele a n_2 stupni volnosti jmenovatele se používá značení $F_\alpha(n_1, n_2)$. Tyto hodnoty jsou pro $\alpha \geq 0,5$ tabelované. Pro výpočet α-kvantilů při $\alpha < 0,5$ se používá vztah

$$F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)}.$$



I.III Dvourozměrné normální rozdělení

Normální rozložení náhodné veličiny X můžeme zobecnit pro (obecně n-rozměrný) náhodný vektor $\boldsymbol{X}$. Prakticky budeme ale pracovat jen s případem $n = 2$, a tedy $\boldsymbol{X} = \binom{X_1}{X_2}$. Skutečnost, že náhodný vektor $\boldsymbol{X}$ pochází z dvourozměrného normálního rozdělení s vektorem středních hodnot $\boldsymbol{\mu} = \binom{\mu_1}{\mu_2}$ a variační maticí $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, zapisujeme $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. V případě, že $\boldsymbol{\mu}$ je nulový vektor a $\boldsymbol{\Sigma}$ diagonální matice, mluvíme o standardizovaném n-rozměrném normálním rozdělením.

Vztah, mezi vícerozměrným normálním rozdělením náhodného vektoru $\boldsymbol{X}$ a marginálními rozděleními jeho složek X_i shrnuje následující tvrzení. Pro jiná rozložení podobné tvrzení neplatí.

Důležité tvrzení: Vztah mezi dvourozměrným normálním rozdělením náhodného vektoru X a marginálními rozděleními jeho složek

Mějme náhodný vektor $\boldsymbol{X} = \binom{X_1}{X_2}$ z dvourozměrného normálního rozdělení s vektorem středních hodnot $\boldsymbol{\mu} = \binom{\mu_1}{\mu_2}$ a variační maticí $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, tj. $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Potom

- Marginální náhoná veličina X_1 má normální rozdělení se střední hodnotou μ_1 a rozptylem σ_1^2.
- Marginální náhoná veličina X_2 má normální rozdělení se střední hodnotou μ_2 a rozptylem σ_2^2.
- Korelace náhodný veličin X_1 a X_2 je ρ.

Rovněž vícerozměrné normální rozložení zachovává linearitu, nicméně si musíme pohlídat rozměry vektorů a matic abychom je mohli násobit.

Důležité tvrzení: Lineární transformace dvourozměné normální náhodné veličiny

Pokud $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a máme vektor reálných čísel $\boldsymbol{a} = \binom{a_1}{a_2}$ a čtvercovou matici reálných čísel $\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$, pak transformovaný náhodný vektor $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$ má také dvourozměrné normální rozložení, a sice

$$\boldsymbol{Y} \sim N_2(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T)$$