Některá DISKRÉTNÍ rozdělení

Alternativní rozdělení [alternative distribution]

$$X \sim Alt(p)$$

- Příklad: Chytnu zvíře. Je to samec?
- Nejjednodušší případ, X nabývá pouze hodnot $\Omega = \{0, 1\}$.
- Data jako „nastal – nenastal“, „přítomný – nepřítomný“, „úspěch – neúspěch“.
- Popis rozdělení: $P(X = 1) = p$

  a $P(X = 0) = 1 - p = q$
- Potom $E(X) = p = 1 \cdot p + 0 \cdot q$

  $var(X) = p \cdot (1 - p) = p \cdot q = (0 - p)^2 \cdot q + (1 - p)^2 \cdot p =$

  $= p^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p \cdot (1 - p) + (p + 1 - p)$

Binomické rozdělení [binomial distribution] (dbinom)

Příklad: Mám 10 pastelek. Kolik z nich je červených?

Kolik se narodí chlapců do rodiny se třemi dětmi?

$Y \sim Bi(n, p)$ *[čti: náh. vel. Y má binomické rozdělení s parametry* n *a* p*]*

- Zjišťujeme pouze výskyt či nevýskyt jevu ***B*** v pokusu
- Parametr ***n*** udává celkový počet pokusů
- Pokusy jsou na sobě nezávislé
- Prst ***p*** výskytu jevu ***B*** je v každém pokusu stejná
- Diskrétní rozdělení

- ***Y*** nabývá jedné z hodnot ***Ω = {0, 1, 2, 3, …, n}*** s pravděpodobností

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

$$= \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!\,(n-k)!} p^k q^{n-k}$$

Binomické rozdělení

Příklad: holčičky (●) a kluci (□) v rodině se 3 dětmi:

(●,●,●); (●,●,□); (●,□,●); (□,●,●); (●,□,□); (□,□,●); (□,●,□); (□,□,□).

→ $\mathrm{P}(0\ \text{kluků}) = 1 \cdot \left(\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}\right)$ $\mathrm{P}(2\ \text{kluci}) = 3 \cdot \left(\frac{1}{2} \cdot \frac{1}{2}\right)\left(\frac{1}{2}\right)$

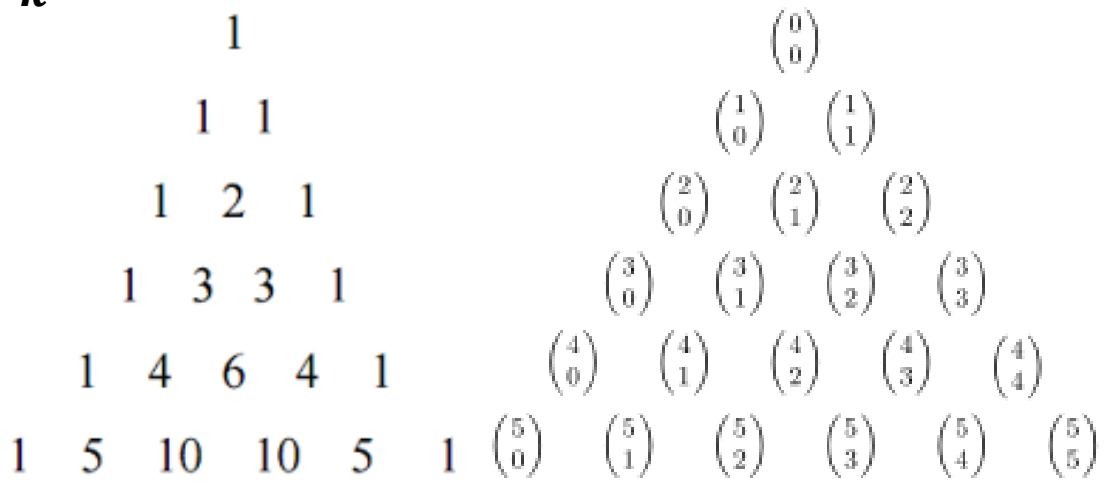
$\mathrm{P}(1\ \text{kluk}) = 3 \cdot \left(\frac{1}{2}\right)\left(\frac{1}{2} \cdot \frac{1}{2}\right)$ $\mathrm{P}(3\ \text{kluci}) = 1 \cdot \left(\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}\right)$

- ***Y*** nabývá jedné z hodnot ***0, 1, 2, 3, …, n*** s pravděpodobností

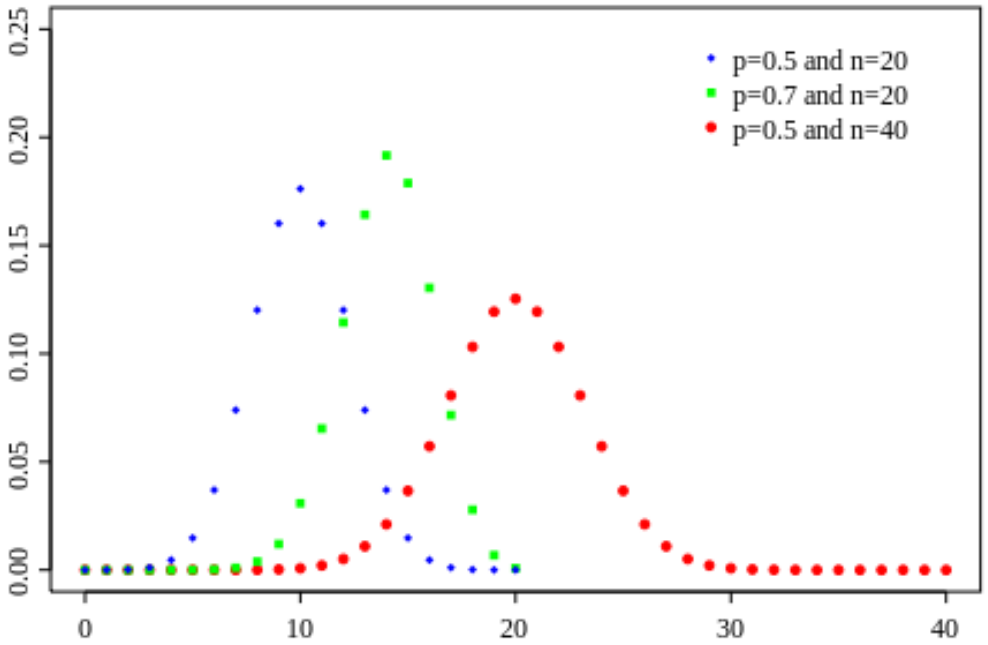
$$\boldsymbol{P(Y = k) = \binom{n}{k} p^k q^{n-k}}$$

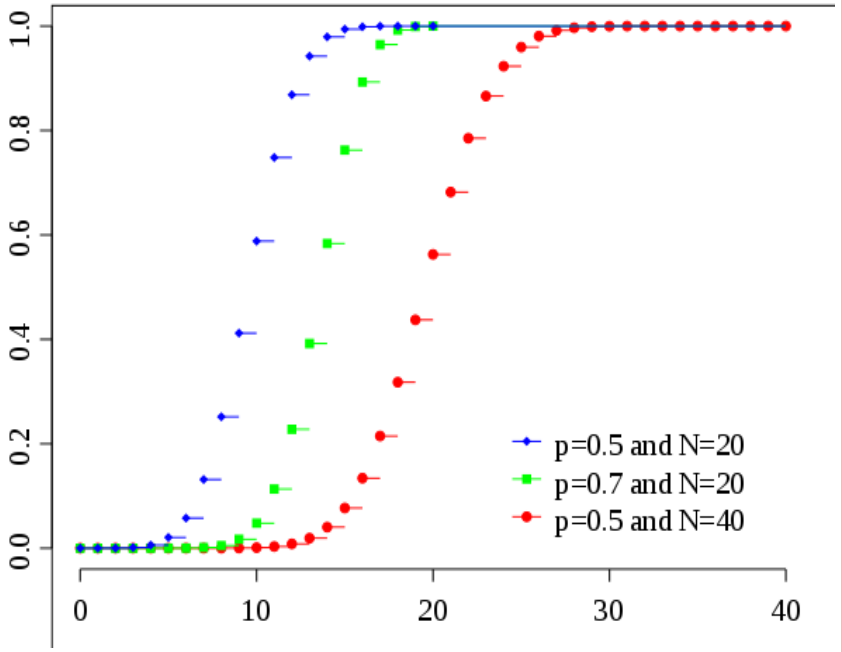
- Pomůcka:

  Pascalův trojúhelník



Binomické rozdělení graficky:





R: x <- c(1:40)
plot(x, dbinom(x, size=40, prob=0.5), pch=16, col="red")
points(x, dbinom(x, size=40, prob=0.25, pch=1, col=„blue")

Střední hodnota a rozptyl binomického rozdělení:

Y mohu vyjádřit jako součet „úspěchů“ z alternativního rozdělení:

$$\boldsymbol{Y = \sum_{i=1}^{n} X_i}, \text{ kde } \boldsymbol{X_i \sim Alt(p)} \text{ a jsou nezávislé.}$$

Potom: $\boldsymbol{\mu_Y} = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} EX_i = \sum_{i=1}^{n} p = \boldsymbol{n \cdot p}$

$$\boldsymbol{\sigma^2_Y} = var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(X_i) = \sum_{i=1}^{n} pq = \boldsymbol{npq}$$

Další příklady: Kolik je samic mezi ***n*** náhodně chycenými zvířaty?

Kolik laboratorních krys přežije naočkování virem? (!?)

Úlohy v praxi:

1) odhad parametru (prsti) ***p*** a konstrukce intervalu spolehlivosti;

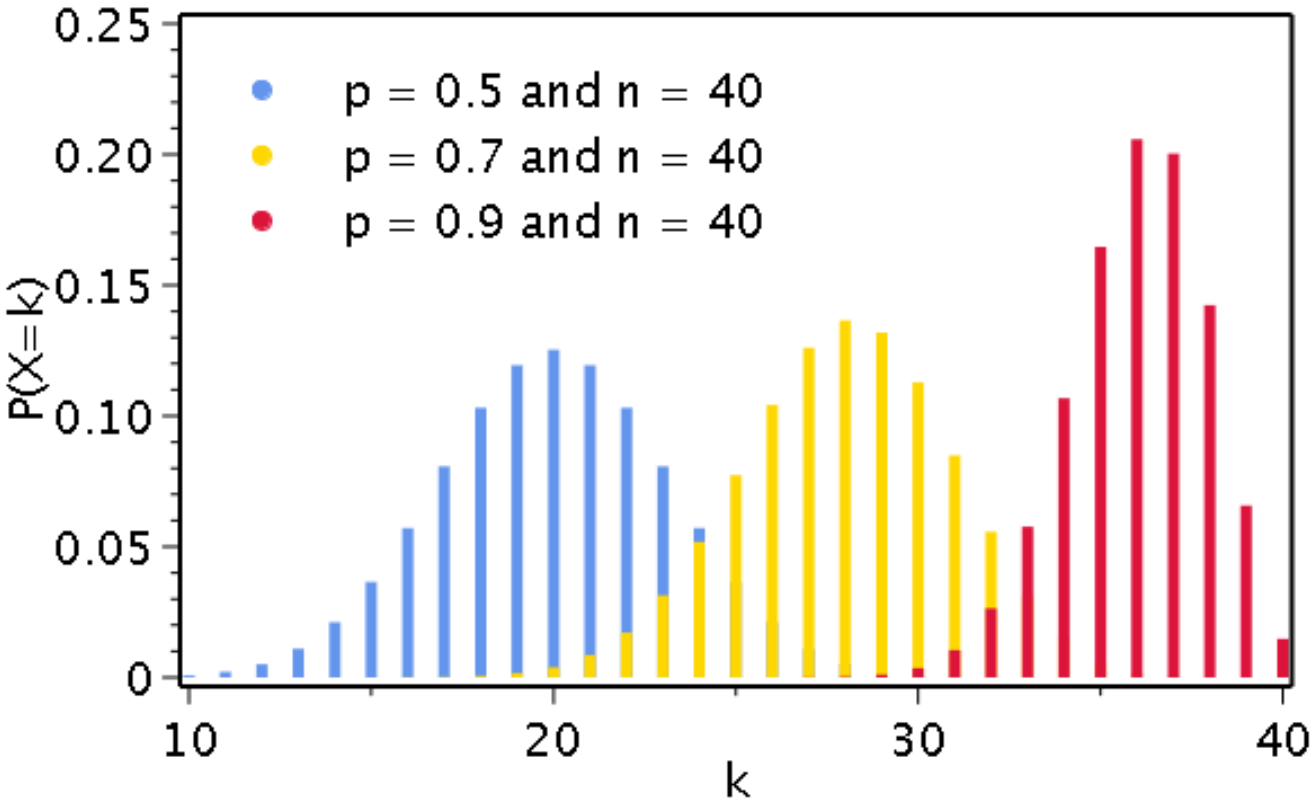
2) odhad a porovnání relativních četností jevů;

3) výpočet potřebného rozsahu výběru ***n*** tak, aby odhad parametru ***p*** měl požadovanou přesnost.

Aproximace binomického rozdělení

(aproximace = přibližný odhad, nahrazení jiným vhodným rozdělením)

Pozn.: pokud je $\boldsymbol{p = q = 0,5}$, potom je binom. rozdělení symetrické, jinak je asymetrické.



Aproximace binomického rozdělení

(aproximace = přibližný odhad, nahrazení jiným vhodným rozdělením)

Pozn.: pokud je $p = q = 0,5$, potom je binom. rozdělení symetrické, jinak je asymetrické.

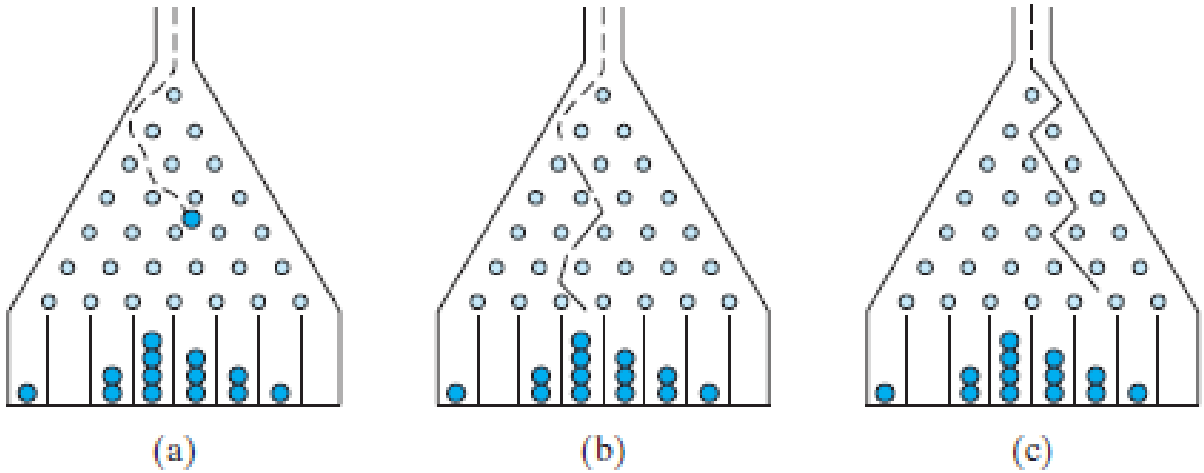
- Pro dostatečně velké ***n*** a rozumné $p \in \langle 0.1, 0.9 \rangle$ můžeme nahradit normálním rozdělením.
- Pro dostatečně velké ***n*** a malé $p < 0.1$ můžeme nahradit Poissonovým rozdělením.

Jak velké je „dostatečně velké“ ***n***?

Buď tak, aby $\boxed{n \cdot p \cdot q > 9 \quad n > \frac{9}{p \cdot q}}$ nebo podle tabulky:

p	→	n
0.5		≥ 30
0.4 a ***0.6***		≥ 50
0.3 a ***0.7***		≥ 80
0.2 a ***0.8***		≥ 200
0.1 a ***0.9***		≥ 600

Galtonova deska či opilcova procházka



https://www.youtube.com/watch?v=3m4bxse2JEQ

Poissonovo rozdělení [Poisson distribution]

Příklad: Sleduju, kolik trolejbusů projede zastávkou za jednotku času.

$X \sim Po(\lambda)$ *[čti: X má Poasonovo rozdělení s parametrem lambda]*

X nabývá hodnot **Ω = {0, 1, 2, 3, …}** (bez omezení shora) s prstí

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$\mu_X = \lambda$ … střední hodnota

$\sigma_X^2 = \lambda$ … rozptyl

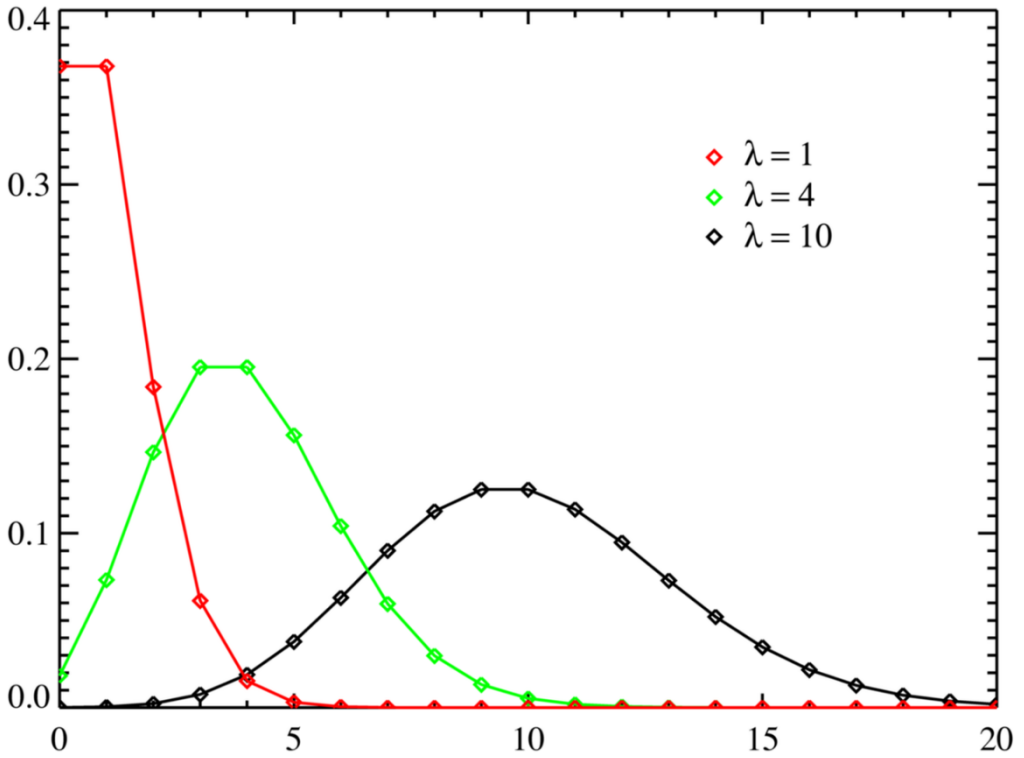
$\lambda > 0$ … kladné reálné číslo (0.25; 1.8)

- Popisuje počet náhodných, vzájemně nezávislých jevů v jednotce času nebo prostoru.
- **Jevy v čase nastávající zřídka** – „zákon vzácných jevů“.
- **Jevy v prostoru** mají být o počtech subjektů v malých objemech nebo v řídké suspenzi.
- Užití: pomocí Poissonova rozdělení testujeme otázky o náhodnosti rozmístění jedinců v ploše/času; také zda výskyt jednoho jedince ovlivňuje výskyt dalších jedinců, nebo zda žijí na sobě nezávisle.

Vlastnosti Poissonova rozdělení

Typický tvar dostává Poissonovo rozdělení pro malá λ (< 2): výrazně pozitivně šikmé.

(To vadí při regresní analýze i při analýze rozptylu. Řešíme to odmocninovou transformací nebo lépe užitím GLM /zobecněných lineárních modelů/)



- Platí, že když nezávislé $\boldsymbol{X \sim Po(\lambda_1)}$ a $\boldsymbol{Y \sim Po(\lambda_2)}$, tak $\boldsymbol{X + Y \sim Po(\lambda_3)}$.
- Pro vyšší hodnoty λ (> 10) lze data aproximovat normálním rozd.

Další příklady Poissonova rozdělení

V čase:

- Počet nezávislých kolonizací vzdáleného ostrova za jednotku času;

V prostoru:

- Počet bakterií v jednotce objemu vodní suspenze, pokud se bakterie nevyskytují ve shlucích (mikrobiologie);
- Rozmístění klíšťat v srsti myši (parazitologie);
- Počet jedinců kruštíku bahenního ve 100 pokusných čtverců (ekologie);

Rozmístění rovnoměrné – náhodné – shlukovité

Mám-li data o počtu jedinců na pokusnou plochu, dostávám pro

- rovnoměrné rozmístění: $\mu_X > \sigma^2{}_X$ (Binomické rozdělení)
- náhodné rozmístění: $\mu_X = \sigma^2{}_X$ (Poissonovo rozdělení)
- shlukovité rozmístění: $\mu_X < \sigma^2{}_X$ (negativně–binomické rozd či Neymanovo rozdělení)

Řešené příklady v Lepš & Šmilauer, str. 361

Negativně binomické rozdělení [negative binomial distribution]

Příklad: Házím korunou. Kolikrát za sebou padne lev, než mi poprvé padne korunka?

- Dělám pokusy s výsledkem **0 / 1** (úspěch/neúspěch), které jsou navzájem nezávislé a pocházejí ze stejného rozdělení (tj. mají stejnou prst úspěchu [iid = independent and identically distributed]
- Výsledkem je počet úspěchů ***k*** do té doby, než nastane ***r*** neúspěchů
- Počet neúspěchů ***r*** je předem stanoveno
- Počet pokusů (***n***) není omezen

$X \sim NB(r, p)$

r = předem daný počet neúspěchů (*)
p = prst úspěchu (* *čti poznámku dále*)

X nabývá hodnot ***k = 0, 1, 2, …*** (počet úspěchů) s pravděpodobností

$$P(X = k) = \binom{k + r - 1}{k} p^k (1 - p)^r = \frac{(k + r - 1)!}{k!\,(k + r - 1 - k)!}\, p^k\, q^r$$

Negativně binomické rozdělení – střední hodnota a rozptyl

$$EX = \frac{r \cdot p}{1-p} = \frac{r \cdot p}{q} \qquad var\ X = \frac{r \cdot p}{(1-p)^2} = \frac{r \cdot p}{q^2}$$

(*) POZOR! Toto rozdělení lze zavést i jinými způsoby:

a) ***X*** = počet úspěchů ***k***, dokud nenastane ***r*** neúspěchů (naše zavedení)

b) ***X*** = počet pokusů ***n***, dokud nenastane ***r*** neúspěchů (***n = k + r***)

c) ***X*** = počet neúspěchů ***r***, dokud nenastane ***k*** úspěchů

d) ***X*** = počet pokusů ***n***, dokud nenastane ***k*** úspěchů

e) ***X*** = počet úspěchů ***k***, je-li dán počet pokusů => binomické rozdělení

Podrobné vzorce na dalším listu.

Poznámka: jméno vzniklo přetvarováním binomického koeficientu

$$\binom{k+r-1}{k} = \ \ldots = (-1)^k \binom{-r}{k}$$

(*) Porovnání alternativních formulací negativně binomického rozdělení:

Stále to nejsou všechny možné tvary zápisu.

a) ***X*** = počet úspěchů ***k***, dokud nenastane ***r*** neúspěchů (naše zavedení)

$$\sim NB(r,p);\quad EX=\frac{r\cdot p}{1-p};\quad k=0,1,2,\ldots;\quad n=r,r+1,r+2,\ldots$$

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

b) ***X*** = počet pokusů ***n***, dokud nenastane ***r*** neúspěchů

$$\sim NB(r,p);\quad EX=\frac{r\cdot p}{1-p}+r;\quad n=r,r+1,r+2,\ldots$$

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

c) ***X*** = počet neúspěchů ***r***, dokud nenastane ***k*** úspěchů

$$\sim NB(k,p);\quad EX=\frac{k\cdot(1-p)}{p};\quad r=0,1,2,\ldots;\quad n=k,k+1,k+2,\ldots$$

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

d) ***X*** = počet pokusů ***n***, dokud nenastane ***k*** úspěchů

$$\sim NB(k,p);\quad EX=\frac{k\cdot(1-p)}{p}+k;\quad n=k,k+1,k+2,\ldots$$

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

SPOJITÁ ROZDĚLENÍ

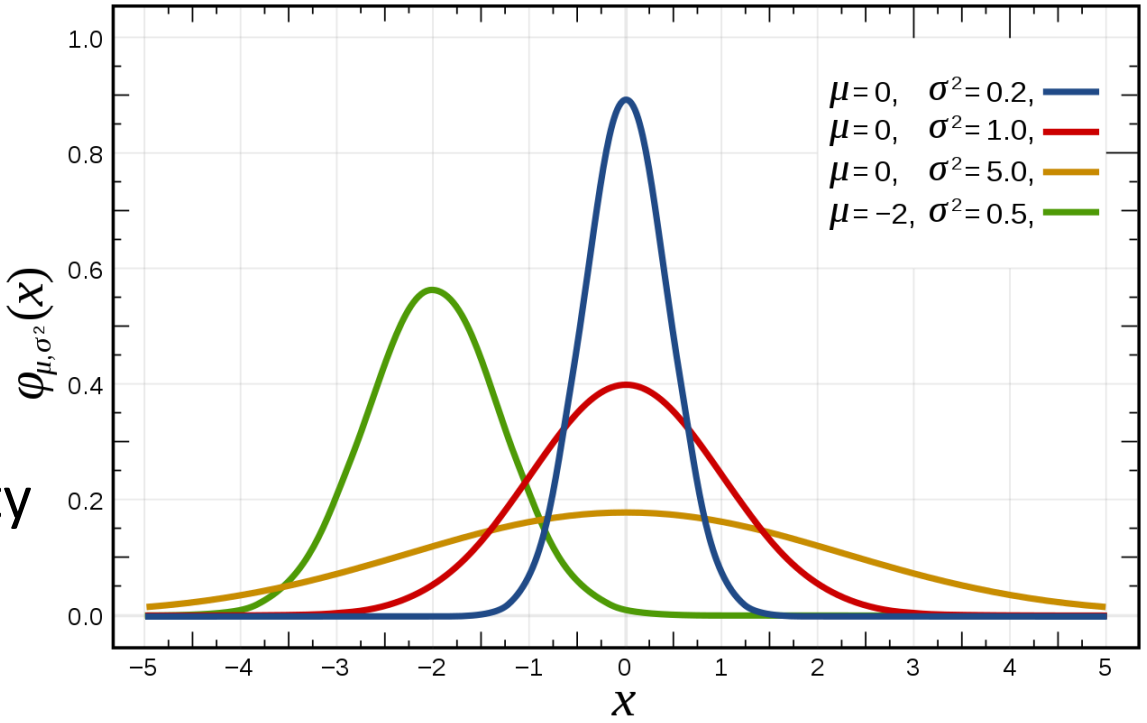
Normální rozdělení [normal distribution, Gaussian distribution] (dnorm)

Příklad: Rozložení výšek studentů ve třídě.

$X \sim N(\mu, \sigma^2)$ *[čti: X má normální rozdělení se střední hodnotou mí a rozptylem sigma na druhou]*

$$f(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} \cdot exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

- Spojitá data
- Symetrické rozdělení
- μ … poloha vrcholu
- σ^2 … šířka zvonu
- Kladné i záporné hodnoty
- Z-rozdělení (STATISTICA)



Normální rozdělení $X \sim N(\mu, \sigma^2)$ také $N(\mu, \sigma)$

Střední hodnota: $EX = \int_{-\infty}^{\infty} x \cdot f(x) = \cdots = \mu$

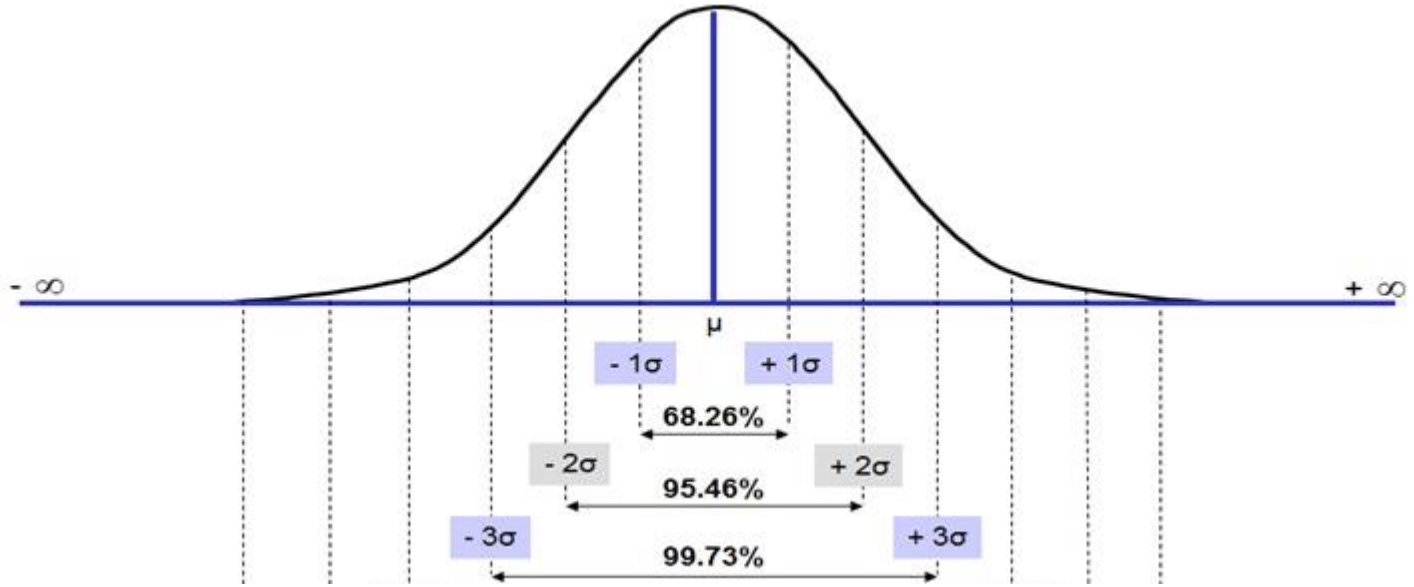
Rozptyl: $var\ X = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx = \cdots = \sigma^2$

Šikmost: $\gamma_1 = 0$

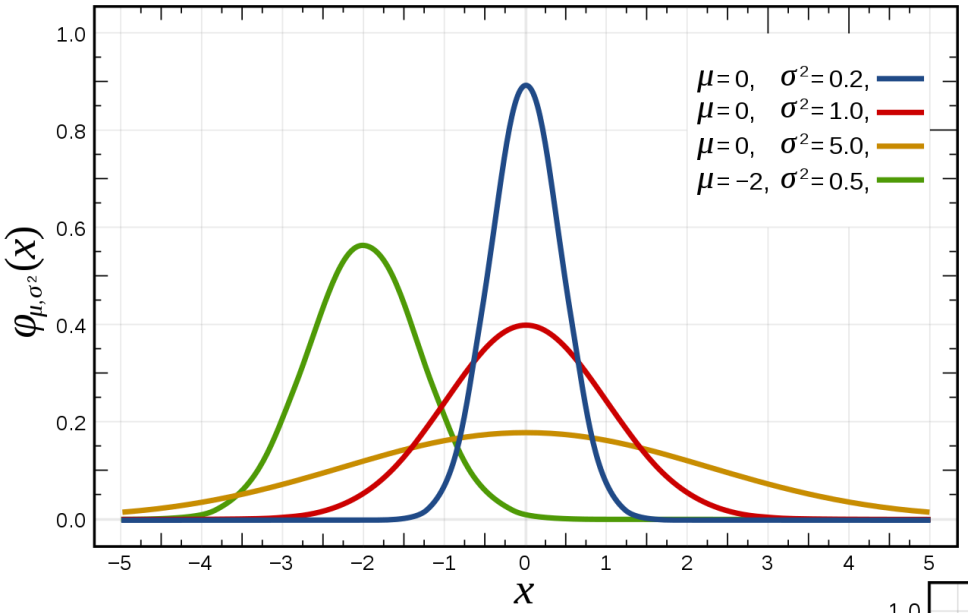
Špičatost: $\gamma_2 = 0$

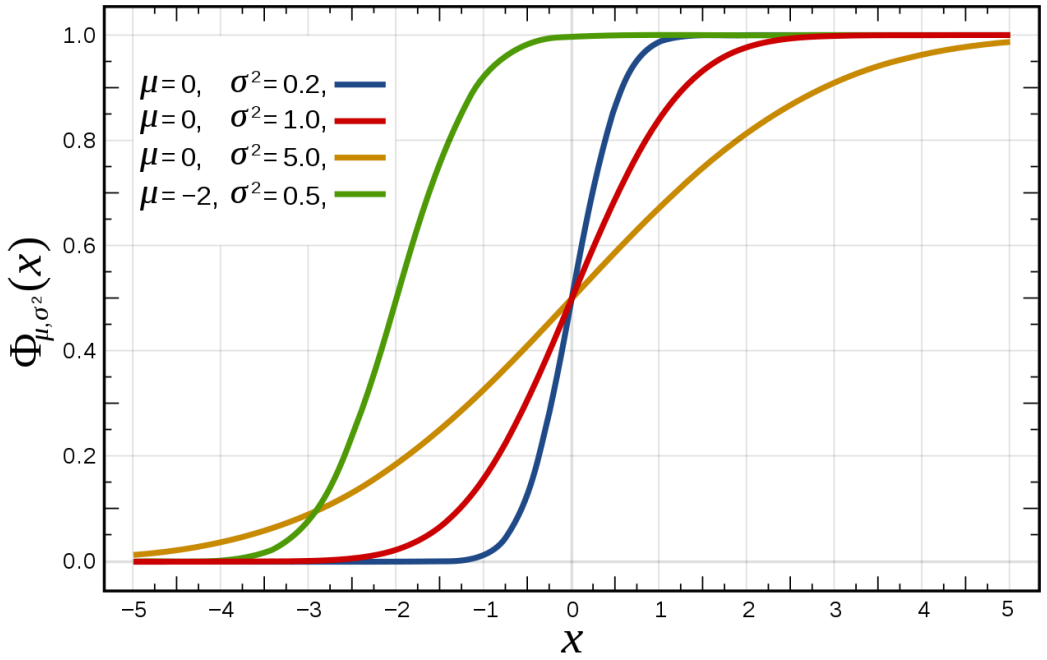
Některé prsti: $P(X \in \langle \mu - \sigma, \mu + \sigma \rangle) \cong 0,68 \sim 68\ \%$

$P(X \in \langle \mu - 2\sigma, \mu + 2\sigma \rangle) \cong 0,955 \sim 95,5\ \%$



Normální rozdělení $X \sim N(\mu, \sigma^2)$





Normální rozdělení $X \sim N(\mu, \sigma^2)$

Co modelujeme pomocí normálního rozdělení:

- spojitá data na poměrové stupnici
- spojitá data na intervalové stupnici, pokud je průměr alespoň o několik směrodatných odchylek větší než nula (arbitrární nula!)
- diskrétní data, pokud má X dostatek různých diskrétních hodnot, např. počet semenáčků v rozmezí alespoň 1 – 30

Standardizované normální rozdělení [standard score, normal deviate]

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

… v praxi počítám: $Z_i = \frac{X_i - \bar{X}}{sd(X)}$

$$E\,Z = 0$$

$$var\,Z = 1$$

Můžete se setkat s tímto značením:

Hustota ***Z*** se značí $\varphi(x) = \frac{1}{\sqrt{2\pi}}\,exp\left\{-\frac{x^2}{2}\right\}$ *[čti: fí x rovná se …]*

Distribuční funkce $\phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} exp\left\{-\frac{t^2}{2}\right\} dt$ *[čti: fí …;* velké fí*]*

Některé kvantily ***N(0,1)***:

$$\left.\begin{array}{l}\phi(-1,96) = 0,025 \sim 2,5\,\% \\ \phi(+1,96) = 0,975 \sim 97,5\,\%\end{array}\right\} P(X < -1,96 \cup X > 1,96) = 0,05$$

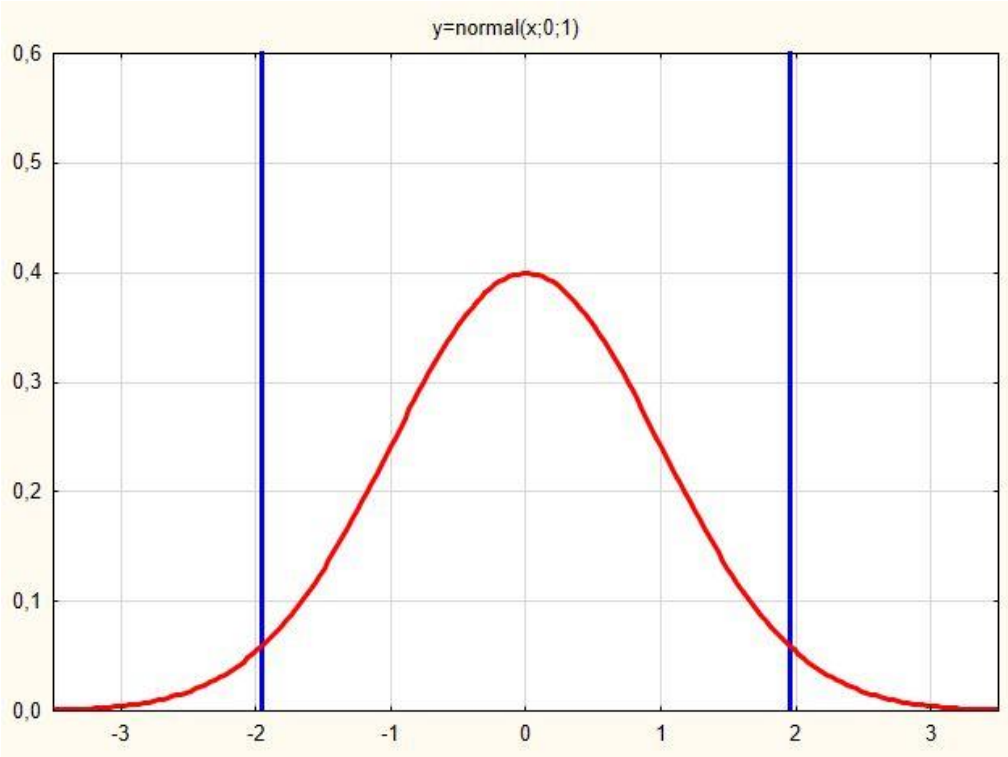
Dříve hodnoty kvantilů v tabulkách, dnes počítají počítače.

Standardizované normální rozdělení [standard score, normal deviate]

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$ … v praxi počítám: $$Z_i = \frac{X_i-\bar{X}}{sd(X)}$$

Některé kvantily ***N(0,1)***:

$$\left.\begin{array}{l}\phi(-1,96) = 0,025 \sim 2,5\ \% \\ \phi(+1,96) = 0,975 \sim 97,5\ \%\end{array}\right\} P(X < -1,96 \cup X > 1,96) = 0,05$$



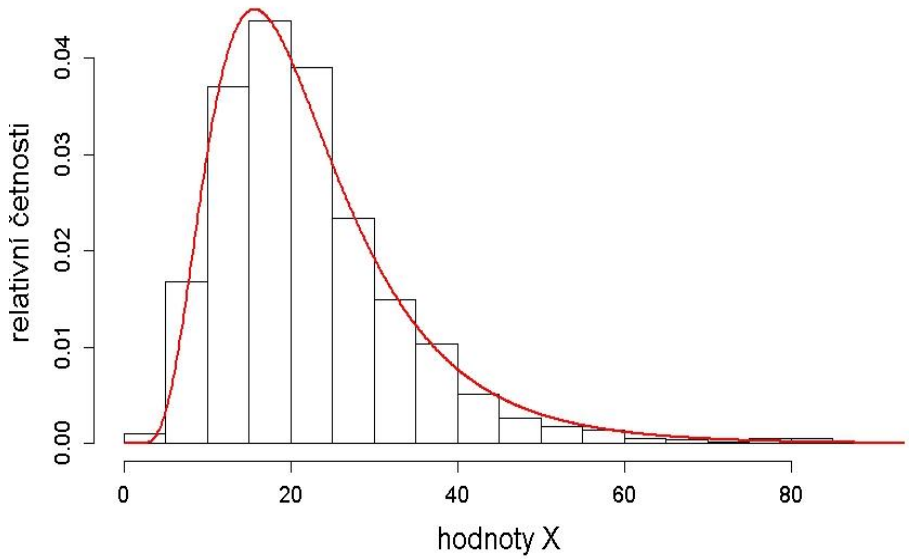
Rozdělení odvozená od normálního rozdělení:

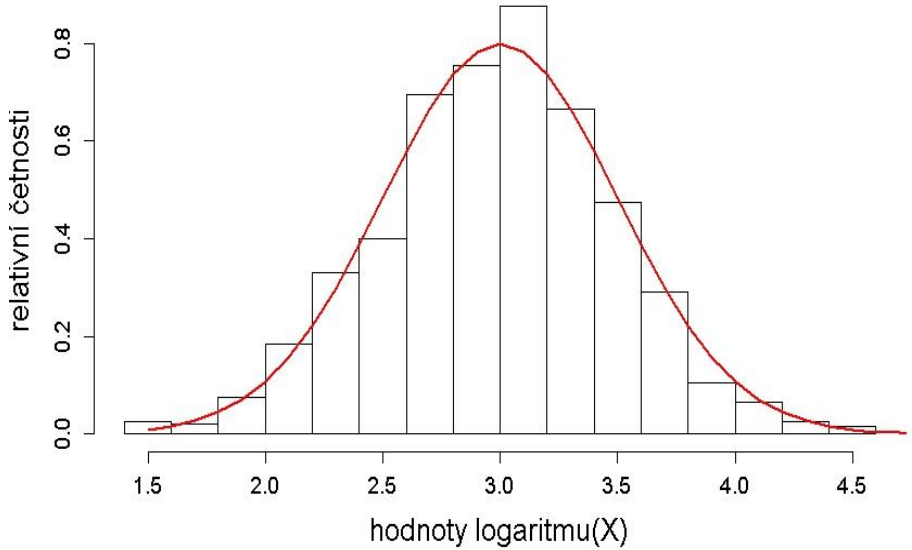
Logaritmicko–normální rozdělení [log–normal distribution]

když $\boldsymbol{X}$ nabývá jen kladných hodnot a její logaritmus má normální rozdělení

$$X \sim LN(\mu, \sigma^2) \rightarrow ln(X) \sim N(\mu, \sigma^2), \; X > 0$$

- příklady: tělesná hmotnost, abundance druhů, koncentrace látek, …
- s rostoucí mírou polohy roste také míra variability
- když $Y = ln(X)$, tak $X = e^Y$





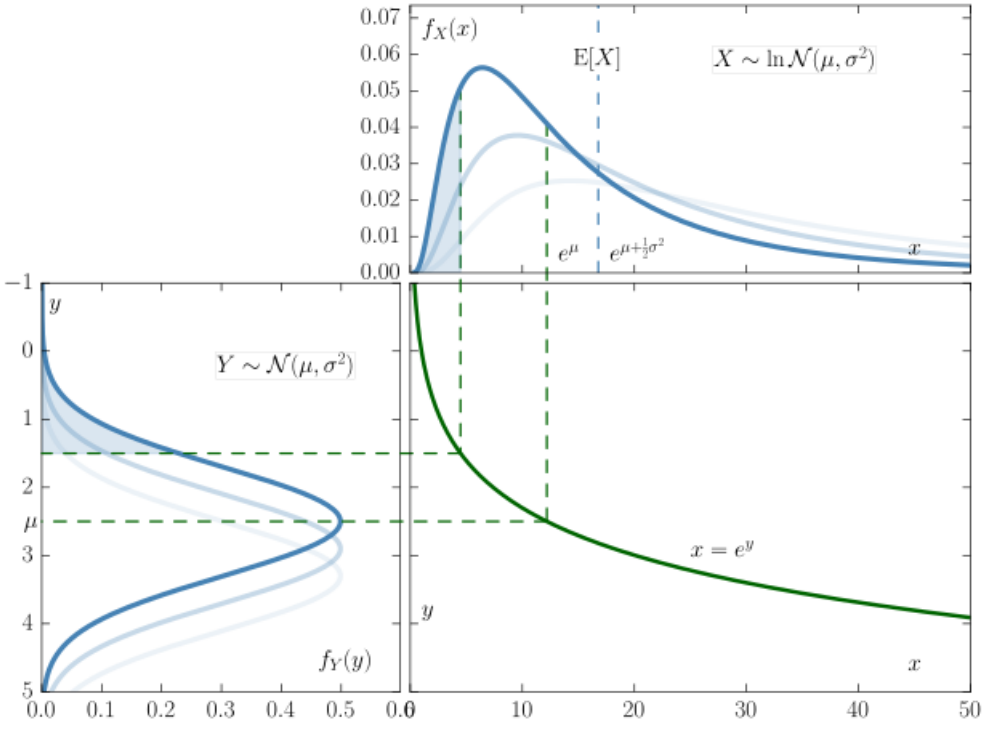
Rozdělení odvozená od normálního rozdělení:

Logaritmicko–normální rozdělení

když ***X*** nabývá jen kladných hodnot a její logaritmus má normální rozdělení

$$X \sim LN(\mu, \sigma^2) \rightarrow ln(X) \sim N(\mu, \sigma^2)\ ,\ X > 0$$

$$EX = e^{\mu + \frac{\sigma^2}{2}} \qquad median\ X = e^{\mu} \qquad modus\ X = e^{\mu - \sigma^2}$$



Rozdělení odvozená od normálního rozdělení:

χ^2, Chí-kvadrát rozdělení [chi-square distrib. [čti: kaj-skvér]] (dchisq)

jsou-li $Z_1, \dots, Z_k \sim N(0,1)$ iid. (nezávislé, stejně rozdělené),

pak platí: $W = \sum_{i=1}^{k} {Z_i}^2 \sim {\chi_k}^2$

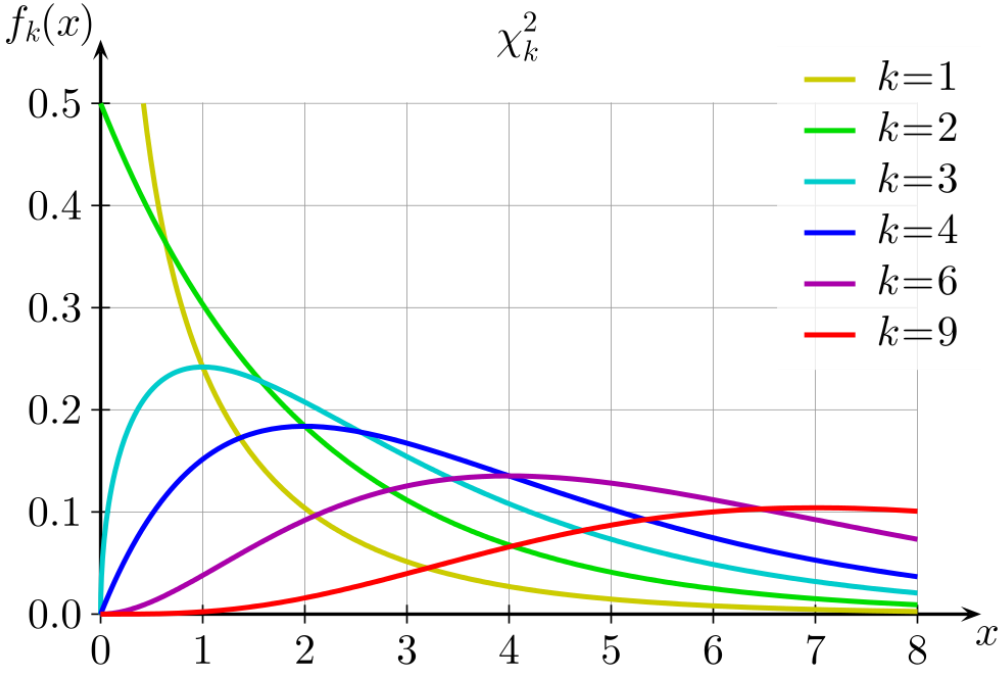
[čti: W *má chí kvadrát rozdělení o k stupních volnosti]*

[stupně volnosti = degrees of freedom]

- hodnoty $W \geq 0$
- tvar hustoty i distribuční funkce

závisí na počtu stupňů volnosti

- středí hodnota $EW = k$
- rozptyl $varW = 2k$



Chí-kvadrát rozdělení $W \sim \chi^2(k)$

Používáme například v těchto situacích:

- testy o rozdělení výběrového rozptylu
- testování nezávislosti faktorů v kontingenčních tabulkách
- testy dobré shody četností pozorovaných dat v kategoriích vůči předpokládanému rozdělení (goodness of fit test)
- testy poměrem věrohodností pro „nested“ modely (likelyhood-ratio test)
- analýza přežívání – „log-rank“ testy

Rozdělení odvozená od normálního rozdělení:

t-rozdělení, Studentovo rozdělení [Student's t-distribution] (dt)

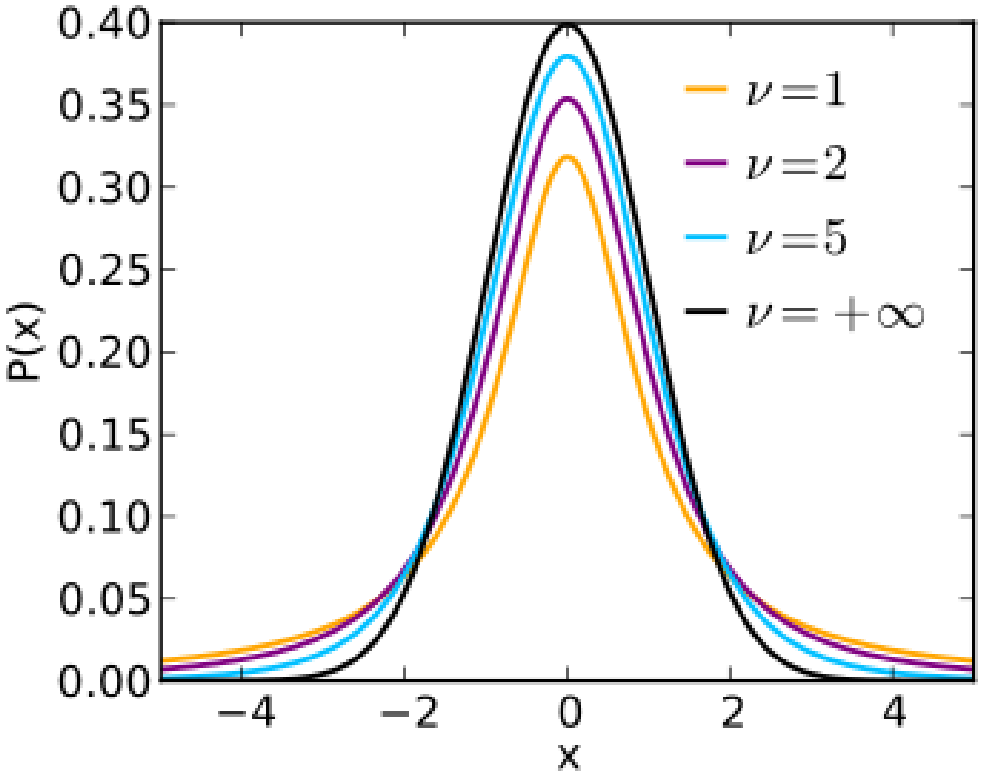
jsou-li $\boldsymbol{W = \sum_{i=1}^{k} Ž_i^{\,2} \sim \chi^2(k)}$ a $\boldsymbol{Z \sim N(0,1)}$ nezávislé náh. veličiny, pak platí:

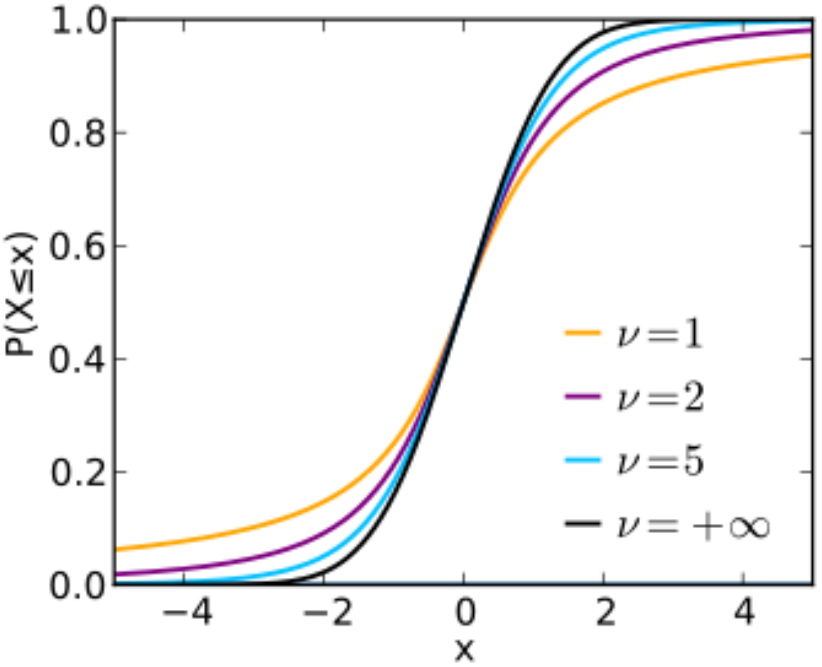
$$\boldsymbol{T = \frac{Z}{\sqrt{\frac{W}{k}}} \sim t_k}$$

[čti: *T má té rozdělení s ká stupni volnosti*]

- hodnoty $T \in \langle -\infty, \infty \rangle$
- rozdělení symetrické kolem nuly, velmi podobné normálnímu rozdělení
- střední hodnota $\boldsymbol{ET = 0}$ pro $k > 1$, jinak neexistuje
- rozptyl $\boldsymbol{var\ T = \frac{k}{k-2}}$ pro $k > 2$, jinak neexistuje; $varT \underset{k\to\infty}{\longrightarrow} 1$
- objevuje se v testech, kde neznámý populační rozptyl σ^2 nahrazujeme výběrovým rozptylem S^2

t-rozdělení $T \sim t_k$





t-rozdělení $T \sim t_k$

Odvození tvaru náhodné veličiny užitečné pro testování:

(Tentýž snímek také v další přednášce.)

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}, \qquad \text{protože } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{... Normování } \bar{X}$$

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} \quad \longrightarrow \quad (n-1) \cdot S^2 = \sum (X_i - \bar{X})^2$$

$$\frac{(n-1) \cdot S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^{\,2} = W$$

Normovaná X_i: $\frac{X_i - \mu}{\sigma}$

Chci tam $\bar{X}$ místo μ, ale ztrácím tím jeden stupeň volnosti.

Přidat σ je snadné, ale musím ji přidat na obě strany rovnice!

$$T = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1) \cdot S^2}{\sigma^2}}{n-1}}} = \frac{(\bar{X} - \mu)\frac{\sqrt{n}}{\sigma}}{\sqrt{\frac{S^2 \cdot (n-1)}{\sigma^2 \cdot (n-1)}}} = \frac{(\bar{X} - \mu)\frac{\sqrt{n}}{\sigma}}{\frac{S}{\sigma}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

F – rozdělení, Fisherovo rozdělení [Fisher – Snedecor distribution] (df)

jsou-li $V \sim \chi^2(k)$ a $W \sim \chi^2(m)$ nezávislé náhodné veličiny, potom platí

$$F = \frac{\frac{V}{k}}{\frac{W}{m}} \sim F(k, m)$$

[čti: *F má ef rozdělení s ká a em stupni volnosti]*

- hodnoty $F \geq 0$
- střední hodnota $EF = \frac{m}{m-2}$, pro $m > 2$, jinak neexistuje
- rozptyl $varF = \frac{2m^2(k+m-2)}{k(m-2)^2(m-4)}$, pro $m > 4$, jinak neexistuje
- tam, kde porovnáváme dva nezávislé odhady stejného rozptylu (ANOVA); ověřování shody populačních rozptylů před dvouvýběrovým t-testem

F-rozdělení $F \sim F(k, m)$