Odhad parametrů $N(\mu, \sigma^2)$

- Mějme statistický soubor $x_1, x_2, \ldots, x_n$ modelovaný jako realizaci náhodného výběru z normálního rozdělení $N(\mu, \sigma^2)$ s neznámými parametry μ a σ.
- Jaký je maximální věrohodný odhad pro μ a σ?
- Parametr θ je vektor = (μ, σ) a věrohodnostní funkce musí být funkcí dvou proměnných:

$$L(\mu, \sigma) = f_{\mu,\sigma}(x_1) f_{\mu,\sigma}(x_2) \cdots f_{\mu,\sigma}(x_n)$$

kde každá $f_{\mu,\sigma}(x)$ je hustota pravděpodobnosti rozdělení $N(\mu, \sigma^2)$:

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

- Musí platit:

$$\ln\left(f_{\mu,\sigma}(x)\right) = -\ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

Odhad parametrů $N(\mu, \sigma^2)$

- Pak pro logaritmus věrohodnostní funkce můžeme psát:

$$\begin{aligned}\ell(\mu,\sigma) &= \ln\left(f_{\mu,\sigma}(x_1)\right) + \cdots + \ln\left(f_{\mu,\sigma}(x_n)\right) \\ &= -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left((x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right)\end{aligned}$$

- Pak parciální derivace $l(\mu, \sigma)$ jsou:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\left((x_1-\mu) + (x_2-\mu) + \cdots + (x_n-\mu)\right) = \frac{n}{\sigma^2}\left(\bar{x}_n - \mu\right) \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\left((x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2\right) \\ &= -\frac{n}{\sigma^3}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right).\end{aligned}$$

Odhad parametrů $N(\mu, \sigma^2)$

- Maximum $l(\mu, \sigma)$ bude odpovídat současné nulové hodnotě obou parciálních derivací:

$$\frac{\partial \ell}{\partial \mu} = 0 \qquad \frac{\partial \ell}{\partial \sigma} = 0$$

- Řešením těchto rovnic dostáváme, že:

  $\mu = \bar{x}_n$ a $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}.$

- Již snadno ukážeme, že věrohodnostní funkce $L(\mu, \sigma)$ nabývá maxima pro stejné hodnoty parametrů.
- Tedy vidíme, že $\bar{x}_n$ je maximální věrohodný odhad pro μ a $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$ je maximální věrohodný odhad pro σ.

Vlastnosti věrohodnostních funkcí

- Pravidlo maximální věrohodnosti poskytuje obecný návod na konstrukci odhadových funkcí.
- Věrohodnostní odhadové funkce mají několik důležitých vlastností.
- **Neměnnost principu**.
- Jestliže $D_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$

  je maximum věrohodnostní funkce parametru σ normálního rozdělení $N(\mu, \sigma^2)$, tak potom je D^2_n věrohodnostní funkce pro parametr σ^2?
- Platí to!!! Navíc je to obecná vlastnost parametru θ s věrohodnostní funkcí T a libovolnou funkci $g(\theta)$ s věrohodnostní funkci $g(T)$.

Vlastnosti věrohodnostních funkcí

- **Asymptotická nestrannost**.
- Maximum věrohodnostní funkce *T* může být stranné.
- Protože $D_n^2 = \frac{n-1}{n} S_n^2$ jak plyne z předchozí vlastnosti, lze psát:

$$\mathrm{E}\left[D_n^2\right] = \mathrm{E}\left[\frac{n-1}{n} S_n^2\right] = \frac{n-1}{n} \mathrm{E}\left[S_n^2\right] = \frac{n-1}{n}\,\sigma^2$$

- Vidíme, že D^2_n je stranný odhad parametru σ^2, ale pro *n* konvergující k nekonečnu střední hodnota D^2_n konverguje k σ^2.
- Výše uvedené platí obecně: pokud velikost statistického souboru *n* jde limitně k nekonečnu, potom maximum věrohodnostní funkce je nestranné.
- Jestliže $T_n = h(X_1, X_2, \ldots, X_n)$ je maximum věrohodnostní funkce pro parametr θ potom: $\lim_{n\to\infty} \mathrm{E}\left[T_n\right] = \theta.$

Vlastnosti věrohodnostních funkcí

- **Asymptotické minimum rozptylu**.
- Platí, že rozptyl nestranné odhadové funkce pro parametr θ je vždy $\geq$ jak nějaké kladné číslo – Cramér-Rao spodní mez.
- Maximum věrohodnostní funkce má asymptoticky nejmenší rozptyl mezi nestrannými odhadovými funkcemi.
- Tedy pro *n* konvergující k nekonečnu, rozptyl maxima věrohodnostní funkce pro parametr θ dosahuje Cramér-Rao spodní meze.

Metoda nejmenších čtverců

Odhad nejmenších čtverců

- Princip maximální věrohodnosti nám poskytuje návod jak odhadnout neznámé modelové parametry.
- Je to v podstatě obvyklá metoda v matematické statistice, ale bohužel není univerzální.
- Např. pro lineární regresní model je nutné znát distribuční funkci závislé náhodné proměnné Y, abychom našli maximální věrohodný odhad pro regresní parametry α a β.
- Odhad pomocí nejmenších čtverců nám umožní tyto parametry určit.

Odhad nejmenších čtverců

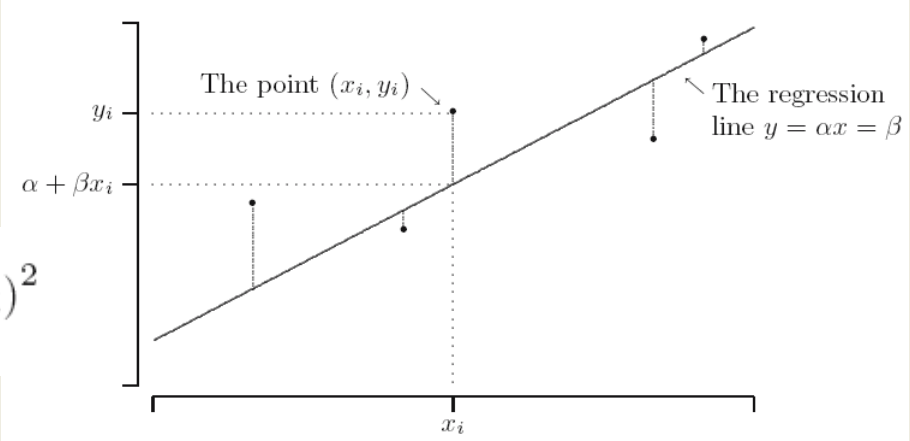
- Mějme statistický soubor tvořený dvojicí proměnných *(*x_1, y_1*)*, (x_2, y_2),…, (x_n, y_n).
- Čísla x_1, x_2,…, x_n nejsou náhodné a čísla y_1, y_2,…, y_n jsou realizace náhodné proměnné Y_1, Y_2,…, Y_n, které splňují rovnici:

$$Y_i = \alpha + \beta x_i + U_i \qquad \text{for} \quad i = 1, 2, \ldots, n,$$

  kde nezávislá náhodná proměnná U_i má nulovou střední hodnotu s rozptylem σ^2.
- Naším úkolem je najít odhady pro parametry *α, β* a σ^2 v tomto lineárním regresním modelu.
- Nevíme nic o distribuci náhodné proměnné U_i a tudíž nic ani o Y_i. Nelze tedy použít metody maximálního věrohodného odhadu.
- Chceme najít takové *α* a *β*, aby přímka nejlépe odpovídala statistickému souboru.

Odhad nejmenších čtverců

- Klasický postup spočívá v minimalizaci čtverce vzdálenosti mezi pozorovanou hodnotou y_i a hodnotou $\alpha + \beta x_i$ ležící na regresní přímce.
- Metoda nejmenších čtverců tedy předepisuje vybrat takové parametry α a β, že suma $S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ bude nabývat svého minima.



Odhad nejmenších čtverců

- K nalezení odhadu nejmenších čtverců je třeba nalézt minimum funkce $S(\alpha, \beta)$.
- Musíme tedy provést parciální derivace funkce S podle parametrů α, β a ty se musí rovnat nule:

$$\frac{\partial}{\partial\alpha}S(\alpha,\beta)=0 \quad \Leftrightarrow \quad \sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0$$

$$\frac{\partial}{\partial\beta}S(\alpha,\beta)=0 \quad \Leftrightarrow \quad \sum_{i=1}^{n}(y_i-\alpha-\beta x_i)\,x_i=0$$

- To lze přepsat na rovnice:

$$n\alpha+\beta\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i$$

$$\alpha\sum_{i=1}^{n}x_i+\beta\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i$$

Odhad nejmenších čtverců

- Dostali jsme dvě rovnice o dvou neznámých α a β.
- Lze najít obecné řešení dvojice lineárních rovnic (odhadových funkcí) pro neznámé parametry α a β.

$$\hat{\beta} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n.$$

kde sumu pro *i* od 1 do *n* jsme nahradili jen znakem suma.

- Rovnice $S(\alpha, \beta)$ je v podstatě rovnicí eliptického paraboloidu, který musí mít jen jedno maximum/minimum. Tedy existuje jen jedno řešení soustavy rovnic a tedy α a β jsou určeny jednoznačně.

Nestrannost odhadové funkce $\hat{\beta}$ a $\hat{\alpha}$

- Odhadové funkce pro parametry α a β se zapisují také pomocí náhodných proměnných:

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n, \qquad \hat{\beta} = \frac{n\sum x_i Y_i - (\sum x_i)(\sum Y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

- Dá se ukázat, že odhadové funkce $\hat{\alpha}$ a $\hat{\beta}$ jsou nestranné.
- Platí, že: $\mathrm{E}[Y_i] = \alpha + \beta x_i$ viz str. 20 přednáška 6.
- Jestliže je $\hat{\beta}$ nestranné (tedy $\mathrm{E}[\hat{\beta}] = \beta$), pak pro $\hat{\alpha}$ platí, že:

Nestrannost odhadové funkce $\hat{\beta}$ a $\hat{\alpha}$

$$\begin{aligned}\mathrm{E}[\hat{\alpha}] &= \mathrm{E}\left[\bar{Y}_n\right] - \bar{x}_n \mathrm{E}\left[\hat{\beta}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[Y_i] - \bar{x}_n\beta \\ &= \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \bar{x}_n\beta = \alpha + \beta\bar{x}_n - \bar{x}_n\beta \\ &= \alpha.\end{aligned}$$

- Pro $\hat{\beta}$ platí, že:

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{n\sum x_i \mathrm{E}[Y_i] - (\sum x_i)(\sum \mathrm{E}[Y_i])}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{n\sum x_i(\alpha + \beta x_i) - (\sum x_i)\left[n\alpha + \beta\sum x_i\right]}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{n\alpha\sum x_i + n\beta\sum x_i^2 - n\alpha\sum x_i - \beta(\sum x_i)^2}{n\sum x_i^2 - (\sum x_i)^2}$$

- Z posledního jednoduše plyne, že $\mathrm{E}[\hat{\beta}] = \beta$.

Nestranná odhadová funkce pro σ^2

- Náhodné proměnné Y_i jsou nezávislé s rozptylem σ^2.
- Bohužel nemůžeme aplikovat známou odhadovou funkci $(1/(n-1))\sum_{i=1}^{n}\left(Y_i-\bar{Y}_i\right)^2$ na odhad rozptylu náhodné proměnné Y_i, protože každé Y_i má jinou střední hodnotu.
- Nicméně dá se ukázat, že: $T=\frac{1}{n}\sum_{i=1}^{n}U_i^2$

  je nestranná odhadová funkce pro σ^2.
- Protože známe jen hodnoty x_i a Y_i a ne U_i , lze užít rovnice $U_i=Y_i-\alpha-\beta x_i$ k přepsání rovnice pro T na tvar: $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2$

Nestranná odhadová funkce pro σ^2

- Střední hodnota odhadové funkce *T* se rovná ((*n*-2)/*n*)σ^2. Potom jednoduše odhadová funkce pro parametr σ^2 je:

AN UNBIASED ESTIMATOR FOR σ^2. In the simple linear regression model the random variable

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

is an unbiased estimator for σ^2.

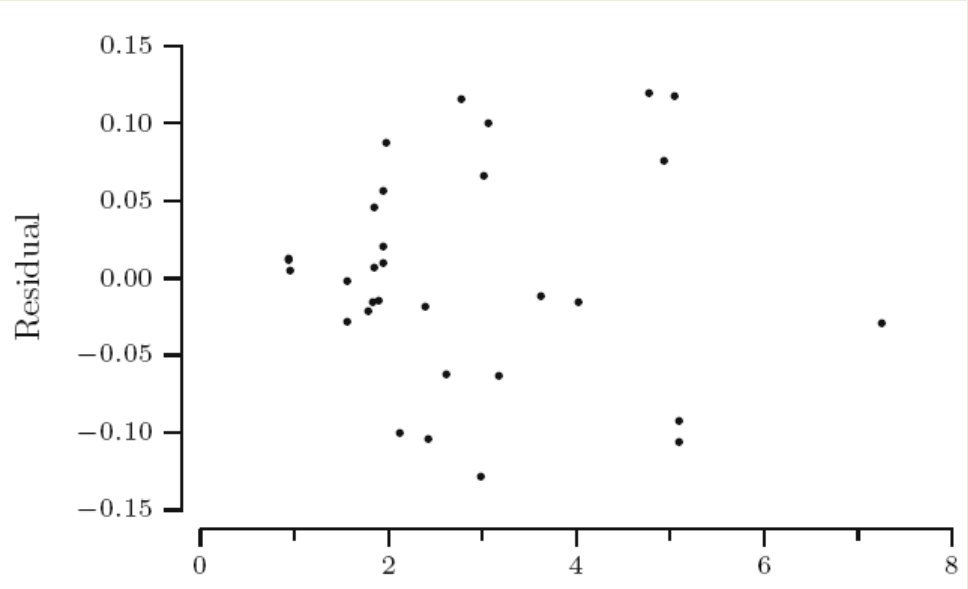
Chyba lineárního regresního modelu

- Pokud chceme studovat, jak dobře jednoduchý lineární regresní model pasuje na daný (x_i, y_i) statistický soubor, musíme zkoumat, jak se mění chyba proložené přímky od hodnot y_i v závislosti na x_i.
- Fitovací chyba r_i je definována jako vertikální vzdálenost mezi *i*-tým prvkem statistického souboru a odhadnutou regresní přímkou:

$$r_i = y_i - \hat{\alpha} - \hat{\beta} x_i, \qquad i = 1, 2, \ldots, n.$$

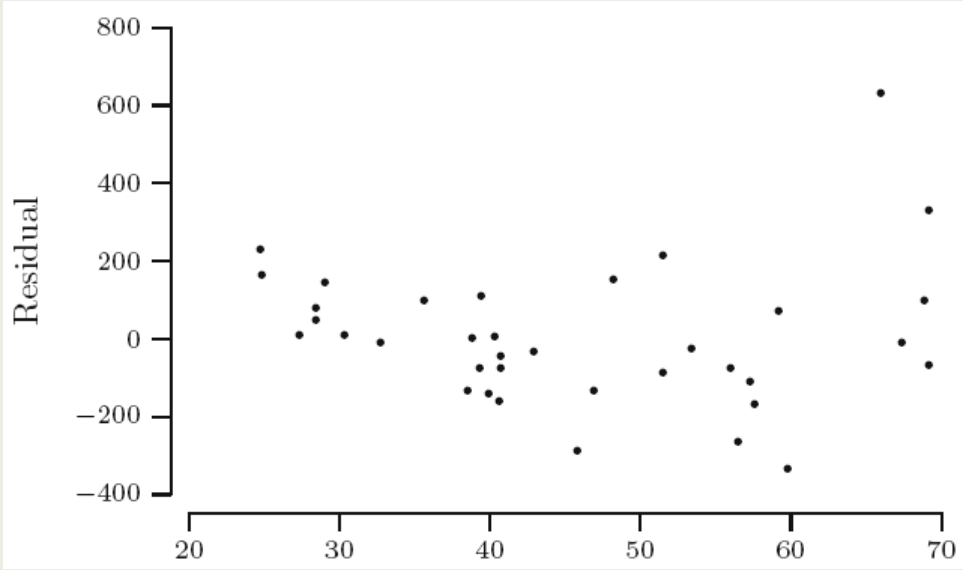
- Pokud je regresní model zvolen vhodně, potom hodnoty r_i jako funkce x_i musí náhodně fluktuovat kolem nuly a nemůžeme zde pozorovat žádný trend.

Chyba lineárního regresního modelu



- Příklad fitovací chyby správně zvoleného regresního modelu.

Chyba lineárního regresního modelu



- Fitovací chyba pro příklad ze str. 20 přednášky 6.
- Je vidět, že chyby nejsou rovnoměrně a náhodně rozděleny, ale mají „parabolický“ tvar.
- Tedy jednoduchý lineární regresní model není vhodný model pro tento statistický soubor

Chyba lineárního regresního modelu

- Lepší regresní model bude, když zvolíme:

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + U_i \qquad \text{for} \quad i = 1, 2, \ldots, n.$$

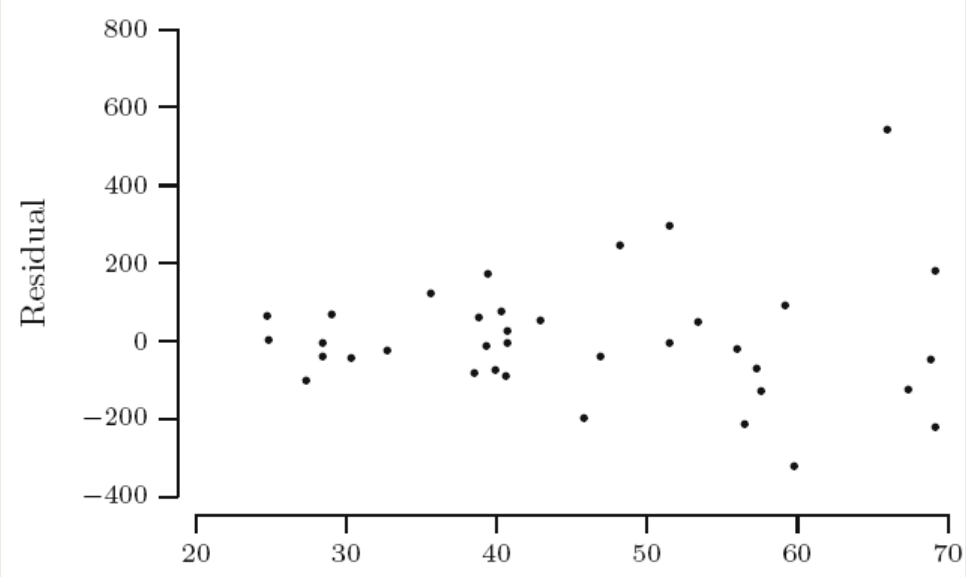
- Fitovací chyba pak bude: $r_i = y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} x_i^2$

  kde $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ jsou odhady parametrů nejmenších čtverců získaných minimalizací odhadové funkce:

$$\sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i - \gamma x_i^2 \right)^2$$

- Opět můžeme do grafu vynést závislost fitovací chyby jako funkce x_i a pro výše uvedený regresní model.
- Už zde není vidět žádný trend ani „tvar“ rozdělení chyb, ale s rostoucím x_i se chyba vzdaluje od nuly.

Chyba lineárního regresního modelu



- Tedy rozptyl náhodné proměnné Y_i je funkcí x_i. Tuto vlastnost nazýváme jako **heteroskedasticita**.

Heteroskedasticita

- Pokud rozptyl náhodné proměnné Y_i (potažmo U_i) se neměnní nazýváme to jako **homoskedasticita**.
- Heteroskedasticita se projeví hlavně v těch případech, kdy náhodná proměnná Y_i s větší střední hodnotou má rozptyl větší než Y_i s menší střední hodnotou.
- To pak způsobí, že fitovací chyby s rostoucím x_i se „rozbíhají“ dál od nulové hodnoty.
- Tento problém lze odstranit modelem tzv. vážených nejmenších čtverců nebo použitím rozptyl stabilizujících transformací.

Obecný model lineární regrese

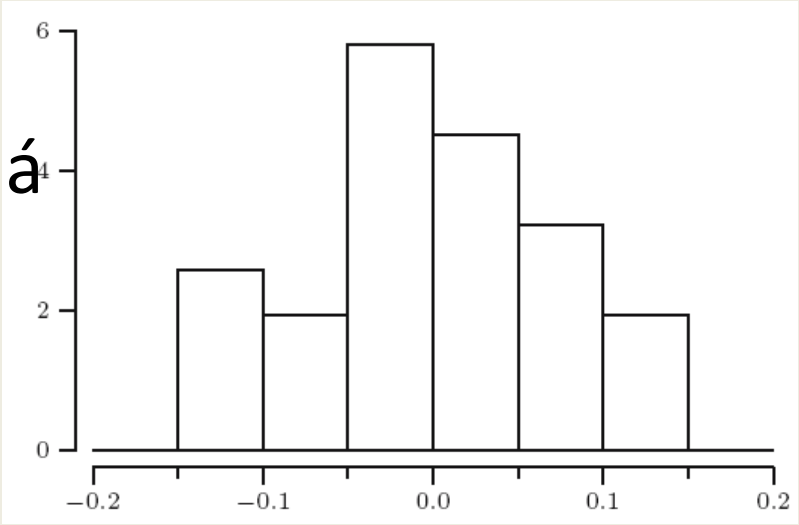
- Jak jsme viděli na příkladu z přednášky 6 str. 20, tak pod pojmem (obecná) lineární regrese si lze představit proložení libovolného polynomu skrze naměřený statistický soubor.
- Jde tedy o lineární kombinaci regresních parametrů $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$,… a prvky statistického souboru x_i se mohou vyskytovat libovolně umocněné nebo na ně může být aplikována libovolná funkce.
- Lineárnost spočívá ve skutečnosti, že odhadová funkce obecného lineárního regresního modelu je lineární pro proměnné $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$,…

Lineární regrese a princip maximální věrohodnosti

- Obecně k použití metody nejmenších čtverců nepotřebujeme znát pravděpodobnostní distribuci náhodné proměnné U_i.
- Pokud distribuci U_i známe, pak princip maximální věrohodnosti může být použit.
- Mějme, že např. U_i je popsáno pravděpodobnostní distribucí $N(0, \sigma^2)$.
- Jaký je maximální věrohodnostní odhad pro parametry α a β?
- V tomto případě Y_i jsou nezávislé a náhodná proměnná Y_i musí být popsána distribucí $N(\alpha + \beta x_i, \sigma^2)$.

Lineární regrese a princip maximální věrohodnosti

- Pokud lineární regresní model je správně zvolen pro daný statistický soubor, pak fitovací chyba r_i musí být realizací náhodného výběru R_i z normálního rozdělení.
- Histogram četnosti r_i z grafu na str. 18.
- Histogram opravdu připomíná hustotu pravděpodobnosti normálního rozdělení.



Lineární regrese a princip maximální věrohodnosti

- Pokud Y_i má $N(\alpha + \beta x_i, \sigma^2)$ distribuci, pak Y_i je popsána hustotou pravděpodobnosti:

$$f_i(y) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(y-\alpha-\beta x_i)^2/(2\sigma^2)} \qquad \text{for} \quad -\infty < y < \infty.$$

- Po zlogaritmování dostaneme:

$$\ln\left(f_i(y_i)\right) = -\ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right)^2$$

- Pak logaritmus věrohodnostní funkce musí být:

$$\begin{aligned} \ell(\alpha, \beta, \sigma) &= \ln\left(f_1(y_1)\right) + \cdots + \ln\left(f_n(y_n)\right) \\ &= -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2. \end{aligned}$$

Lineární regrese a princip maximální věrohodnosti

- Pokud je $\sigma > 0$, pak $l(\alpha, \beta, \sigma)$ dosahuje svého maxima právě tehdy, když $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ je minimální.
- Tedy pokud U_i jsou nezávislé náhodné proměnné s $N(0, \sigma^2)$ distribucí, pak princip maximální věrohodnosti a metoda nejmenších čtverců poskytují stejné odhadové funkce!!!
- Maximální věrohodnostní odhad parametru σ nalezneme derivací $l(\alpha, \beta, \sigma)$ podle σ.

$$\frac{\partial}{\partial\sigma}\ell(\alpha,\beta,\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

- Maximum funkce nastane tehdy, když bude výše uvedená derivace nulová.
- Z toho dostaneme maximální věrohodnostní odhadovou funkci pro σ^2: $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$

Intervaly spolehlivosti

Interval spolehlivosti

- Odhadové funkce jsou např. výběrový průměr, výběrový rozptyl atp.
- Dostaneme odhady μ, σ^2 atp.
- Strannost a střední kvadratická chyba pak určují účinnost odhadové funkce.
- Z realizace náhodného výběru aplikovaného na odhadovou funkci *T* dostaneme odhad *t* parametru θ – **bodový odhad**.
- Typicky statistický soubor naměříme několikrát.
- Pak získáme několik odhadů hledaného parametru pravděpodobnostní distribuce. Každý bude pravděpodobně jiný, i když experiment je stejný.
- Který odhad je nejblíže zkoumanému parametru?
- Můžeme říci, že s velkou jistotou hledaný parametr leží v **intervalu od… do…** Jak velká je jistota, že θ opravdu leží v tomto intervalu?

Interval spolehlivosti

- Tento interval nazýváme jako **interval spolehlivosti**.
- Je nutné si stanovit spolehlivost hledaného parametru na základě výběrové distribuce odpovídající odhadové funkce.
- Mějme nestrannou odhadovou funkci *T* pro parametr θ – rychlost světla měřená Michelsonem – viz přednáška 6 str. 3.
- Předpokládejme, že směrodatná odchylka σ_T odhadové funkce *T* je 100 km/s.
- Z Čebyševovy nerovnosti lze odvodit (přednáška 4 str. 42), že:

$$\mathrm{P}(|Y-\mu|<k\sigma) \geq 1-\frac{\mathrm{Var}(Y)}{k^2\sigma^2} = 1-\frac{1}{k^2}$$

- Pokud náš interval zájmu bude $2\sigma_T$, pak:

$$\mathrm{P}(|T-\theta|<2\sigma_T) \geq \tfrac{3}{4}.$$

Interval spolehlivosti

- Slovy: s pravděpodobností aspoň 75% odhadová funkce *T* leží v intervalu $2\sigma_T$ = 200 km/s kolem hledané hodnoty parametru θ (rychlost světla) ->

$$T \in (\theta - 200, \theta + 200)$$

- Pokud je *T* blízko θ, tak musí být i θ blízko *T*.
- Tedy $\theta \in (T - 200, T + 200)$ s pravděpodobností 75%.
- První tvrzení: náhodná proměnná *T* je v pevném intervalu s pravděpodobností 75%.
- Druhé tvrzení: náhodný interval s pravděpodobností 75% pokrývá fixní číslo θ.
- Interval (*T* – 200, *T* + 200) se nazývá jako **interval spolehlivosti**.

Interval spolehlivosti

- Z tabulky naměřených dat (přednáška 6 str. 3) získáme odhad rychlosti světla t = 299 852,4 km/s
- Tedy interval spolehlivosti θ je:

$$\theta \in (299\,652.4,\ 300\,052.4)$$

- Máme tedy dvě tvrzení: i) rychlost světla leží buď v tomto intervalu, ii) nebo v něm neleží. Máme tedy pravdivý nebo nepravdivý výrok a my nevíme, který je správný.
- Proto můžeme jen říci, že změřená rychlost světla leží se spolehlivostí aspoň 75% ve výše uvedeném intervalu.
- Takto vytvořené intervaly spolehlivosti zahrnují jenom nestranné odhadové funkce a znalost směrodatné odchylky.

Interval spolehlivosti

- Typické intervaly spolehlivosti mají tvar: $(t - c \cdot \sigma_T,\, t + c \cdot \sigma_T)$

  kde číslo *c* je většinou mezi 2 a 3.
- Existuje tedy mnoho způsobů jak zkonstruovat intervaly spolehlivosti a obecná definice bude:

CONFIDENCE INTERVALS. Suppose a dataset $x_1, \ldots, x_n$ is given, modeled as realization of random variables $X_1, \ldots, X_n$. Let θ be the parameter of interest, and γ a number between 0 and 1. If there exist sample statistics $L_n = g(X_1, \ldots, X_n)$ and $U_n = h(X_1, \ldots, X_n)$ such that

$$\mathrm{P}(L_n < \theta < U_n) = \gamma$$

for every value of θ, then

$$(l_n, u_n),$$

where $l_n = g(x_1, \ldots, x_n)$ and $u_n = h(x_1, \ldots, x_n)$, is called a $100\gamma\%$ *confidence interval for* θ. The number γ is called the *confidence level*.

Interval spolehlivosti

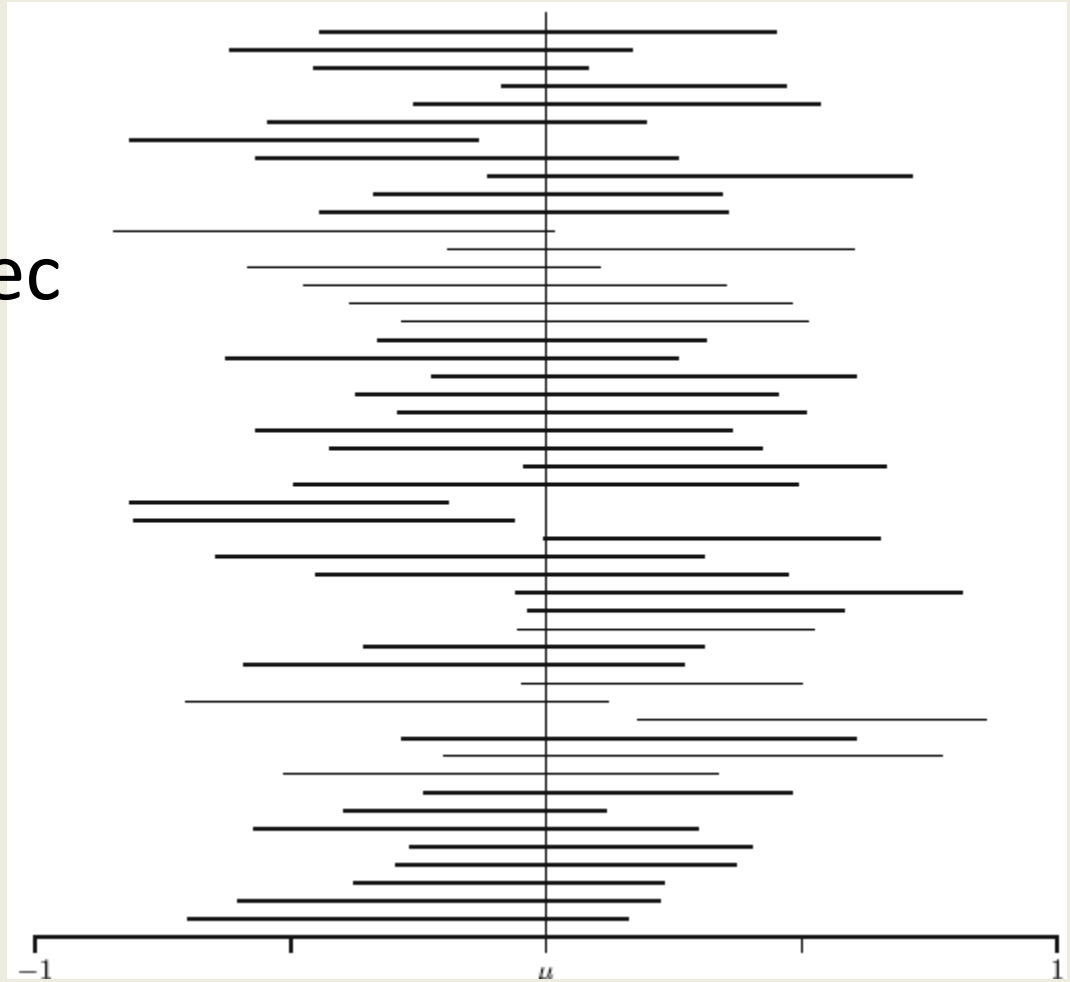
- Často se stane, že odhadové funkce parametrů distribuce L_n a U_n neexistují tak, jak jsou požadovány v definici. Ale můžeme najít takové L_n a U_n, jež splňují podmínku $\mathrm{P}(L_n < \theta < U_n) \geq \gamma$
- Takový interval spolehlivosti (l_n, u_n) se nazývá jako **konzervativní** γ interval spolehlivosti pro parametr θ. Tedy hladina spolehlivosti může být i větší.
- Žádným způsobem nemůžeme zjistit, zda-li interval spolehlivosti je správný ve smyslu, že opravdu pokrývá parametr θ.
- Metoda nám jenom garantuje, že kdykoliv vytvoříme interval spolehlivosti, tak s pravděpodobností γ pokrýváme hodnotu parametru θ.
- Tento fakt si ukážeme na příkladu:

Interval spolehlivosti

- Vygenerujeme x_1,…, x_{20} z $N(0, 1)$ distribuce. Předstírejme, že víme, že datový soubor je generován z normálního rozdělení, ale neznáme střední hodnotu a směrodatnou odchylku.
- Generování statistického souboru 50 krát zopakujeme.
- Zkonstruujeme 90% interval spolehlivosti pro každý generovaný statistický soubor. Budeme zkoumat, zda-li $\mu = 0$ leží v intervalech spolehlivosti.
- Na obr. jsou zobrazeny intervaly spolehlivosti každého generovaného statistického souboru.

Interval spolehlivosti

- Vidíme, že 4 intervaly spolehlivosti vůbec neobsahují $\mu = 0$.



Interval spolehlivosti – kritické hodnoty

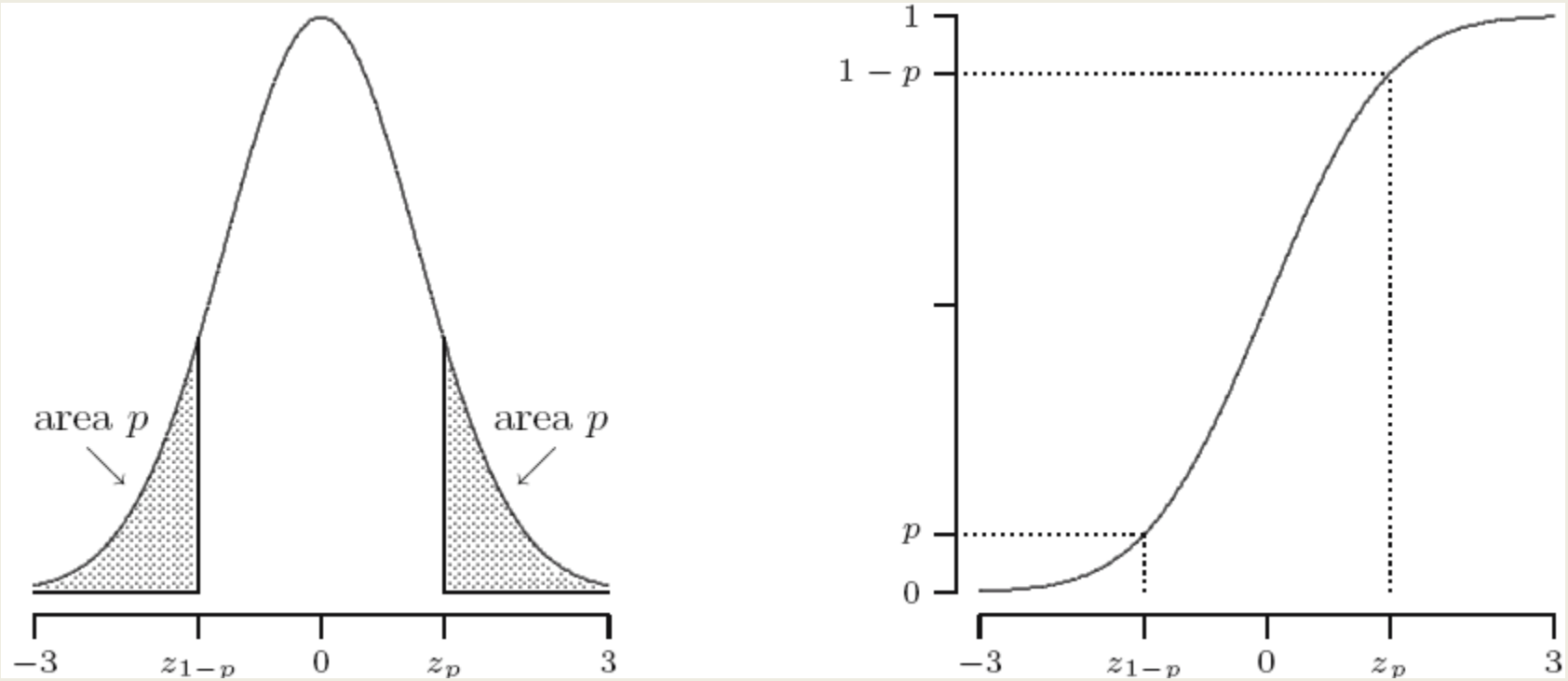
- Budeme potřebovat definovat tzv. **kritické hodnoty** pro standardní normální distribuci.
- Kritická hodnota z_p distribuce $N(0, 1)$ je takové náhodné číslo, které má pravděpodobnost p v pravé části chvostu hustoty pravděpodobnosti: $\mathrm{P}(Z \geq z_p) = p$

  kde Z je náhodná proměnná s $N(0, 1)$.
- Z tabelovaných hodnot Φ (0, 1) plyne: $\mathrm{P}(Z \geq 1{,}96) = 0{,}025$.
- Tedy $z_{0,025} = 1{,}96$. Jinými slovy z_p je (1-p) kvantil standardního normálního rozdělení:

$$\Phi(z_p) = \mathrm{P}(Z \leq z_p) = 1 - p$$

- Protože hustota pravděpodobnosti $N(0, 1)$ je symetrická, musí platit: $\mathrm{P}(Z \leq -z_p) = \mathrm{P}(Z \geq z_p) = p$.
- Pak $\mathrm{P}(Z \geq -z_p) = 1 - p$ a proto $z_{1-p} = -z_p$.
- Například: $z_{0,975} = -z_{0,025}$.

Interval spolehlivosti – kritické hodnoty



Interval spolehlivosti – normální rozdělení

- Mějme náhodný výběr $X_1,\ldots, X_n$ generovaný z rozdělení $N(\mu, \sigma^2)$ a hledáme interval spolehlivosti pro konkrétní statistický soubor jako realizaci náhodného výběru a známe rozptyl.
- Výběrový průměr $\overline{X_n}$ má rozdělení $N(\mu, \sigma^2/n)$.
- Pokud provedeme transformaci proměnné $\overline{X_n}$: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ tak nová proměnná Z bude mít distribuci $N(0, 1)$.
- Vybereme dvě čísla c_l a c_u tak, aby $\mathrm{P}(c_l < Z < c_u) = \gamma$
- Potom musí platit:

$$\begin{aligned}\gamma &= \mathrm{P}\left(c_l < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < c_u\right) \\ &= \mathrm{P}\left(c_l \frac{\sigma}{\sqrt{n}} < \bar{X}_n - \mu < c_u \frac{\sigma}{\sqrt{n}}\right) \\ &= \mathrm{P}\left(\bar{X}_n - c_u \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n - c_l \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

Interval spolehlivosti – normální rozdělení

- Potom lze pro odhadové funkce výběrových parametrů L_n a U_n nalézt:

$$L_n = \bar{X}_n - c_u \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad U_n = \bar{X}_n - c_l \frac{\sigma}{\sqrt{n}}$$

- Tyto parametry splňují podmínku intervalu spolehlivosti: interval (L_n, U_n) pokrývá μ s pravděpodobností γ.
- Tedy interval spolehlivosti se spolehlivostí γ% pro μ je:

$$\left(\bar{x}_n - c_u \frac{\sigma}{\sqrt{n}}, \bar{x}_n - c_l \frac{\sigma}{\sqrt{n}} \right)$$

- V praxi γ zvolíme tak, aby interval spolehlivosti se rozdělil rovnoměrně mezi oba chvosty $N(0, 1)$ distribuce. Tedy $\alpha = 1 - \gamma$.

Interval spolehlivosti – normální rozdělení

- Tedy pro parametry c_l a c_u musí platit:

$$\mathrm{P}(Z \geq c_u) = \alpha/2 \quad \text{and} \quad \mathrm{P}(Z \leq c_l) = \alpha/2$$

- Tedy $c_u = z_{\alpha/2}$ a $c_l = z_{1-\alpha/2} = -z_{\alpha/2}$.
- Tedy $(1 - \alpha)$ interval spolehlivosti pro μ je dán:

$$\left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

- Př.: jestliže $\alpha = 0{,}05$, pak kritické hodnoty intervalu spolehlivosti budou $z_{0{,}025} = 1{,}96$ a 95% interval spolehlivosti bude:

$$\left(\bar{x}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x}_n + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Interval spolehlivosti – *t*-rozdělení

- Předpokládejme nyní, že směrodatnou odchylku neznáme.
- Potom transformace: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ nemá pro nás význam, protože vedle μ neznáme také σ, které se vyskytuje ve výpočtu mezí intervalu spolehlivosti.
- Nicméně můžeme nahradit σ odhadovou funkcí S_n a transformace pak bude: $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$
- Tato nová náhodná proměnná závisí jen na n.
- Její hustotu pravděpodobnosti navíc můžeme analyticky vyjádřit.

Interval spolehlivosti – *t*-rozdělení

DEFINITION. A continuous random variable has a *t-distribution with parameter* m, where $m \geq 1$ is an integer, if its probability density is given by

$$f(x) = k_m \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} \qquad \text{for } -\infty < x < \infty,$$

where $k_m = \Gamma\left(\frac{m+1}{2}\right) / \left(\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}\right)$. This distribution is denoted by $t(m)$ and is referred to as the t-distribution with m *degrees of freedom.*

- Funkce gama je definována předpisem:

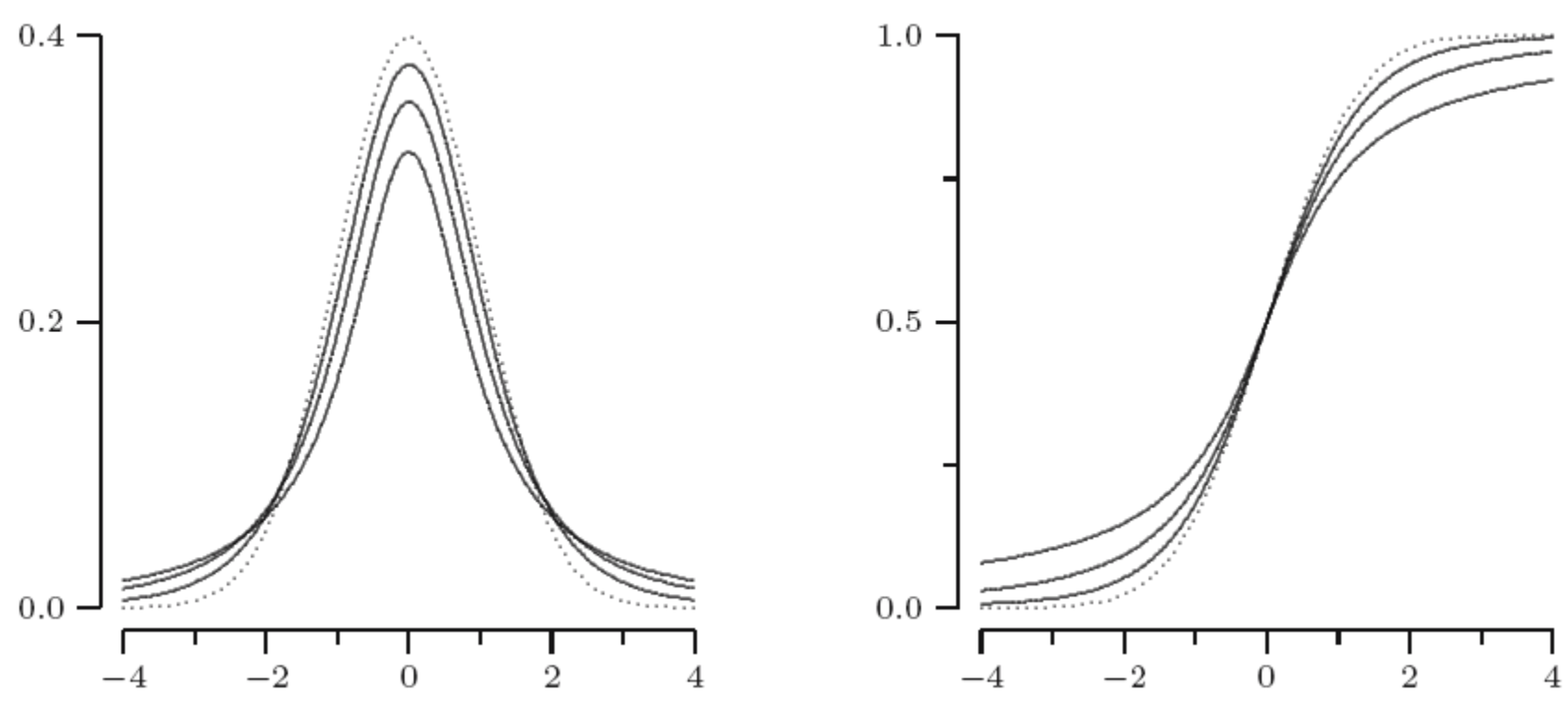
$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \mathrm{e}^{-t}\, \mathrm{d}t$$

- A platí pro ni: $\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \quad \text{and} \quad \Gamma(n) = (n-1)!$

Interval spolehlivosti – *t*-rozdělení

- k_m je normalizační konstanta.
- Pro $m = 1$ je $k_1 = 1/\pi$ a výsledná $f(x)$ rovna standardní Cauchy distribuci.
- Střední hodnota náhodné proměnné s distribucí $t(m)$ je $E[X] = 0$ pro $m \geq 2$ a $Var[X] = m/(m - 2)$ pro $m \geq 2$.
- Hustota pravděpodobnosti *t*-distribuce se podobá standardnímu normálnímu rozdělení.
- Pro $m \rightarrow \infty$ hustota pravděpodobnosti $t(m)$ konverguje k hustotě pravděpodobnosti standardního normálního rozdělení.
- Pro rostoucí x $t(m)$ klesá k nule pomaleji než $N(0, 1)$.

Interval spolehlivosti – *t*-rozdělení



- Tečkovaná křivka je $N(0, 1)$ a plná čára je $t(1)$, $t(2)$ a $t(5)$.

Interval spolehlivosti – *t*-rozdělení

- Potřebujeme ještě stanovit kritické hodnoty pro $t(m)$.
- Kritická hodnota je číslo $t_{m,p}$ splňující podmínku:

$$\mathrm{P}(T \geq t_{m,p}) = p,$$

- Díky symetrii $t(m)$ kolem nuly, ze stejných důvodů jako pro standardní normální rozdělení, dostaneme: $t_{m,1-p} = -t_{m,p}$
- Např.: $t_{10,\,0{,}01} = 2{,}764$ a tedy $t_{10,\,0{,}99} = -2{,}764$.
- Nyní můžeme zkonstruovat interval spolehlivosti pro μ náhodné proměnné $\dfrac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$

Interval spolehlivosti – *t*-rozdělení

THE STUDENTIZED MEAN OF A NORMAL RANDOM SAMPLE. For a random sample $X_1, \ldots, X_n$ from an $N(\mu, \sigma^2)$ distribution, the *studentized mean*

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

has a $t(n-1)$ distribution, regardless of the values of μ and σ.

- Ze znalosti kritických hodnot *t*-rozdělení můžeme odvodit:

$$\mathrm{P}\left(-t_{n-1,\alpha/2} < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < t_{n-1,\alpha/2}\right) = 1 - \alpha$$

- Stejně jako v případě znalosti směrodatné odchylky σ normálního rozdělení, můžeme nyní pro 1 – α interval spolehlivosti parametru μ odvodit:

$$\left(\bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}},\ \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}\right)$$

Interval spolehlivosti – normální rozdělení

- V praktických aplikacích interval spolehlivosti pro střední hodnotu náhodné proměnné X s $N(\mu, \sigma^2)$ v případě neznalosti směrodatné odchylky vyjde širší než v případě znalosti σ.
- Je to díky tomu, že $t(m)$ rozdělení nabývá vyšších hodnot pro větší hodnoty X.
- Dále pak obzvláště pro náhodné výběry s malým počtem prvků je směrodatná odchylka určená z odhadové funkce většinou větší.