NÁHODNÝ VZOREK (RANDOM SAMPLE)

- setříděná konečná sekvence (zde reálných čísel) : $\{\ell_i\}_{i=1}^{n}$
  a) definiční obor (zde R) tvoří pravděpodobnostní prostor
  b) **funkce pravděpodobnosti** pr definuje pravděpodobnost výskytu prvku $\tilde{\ell}_j$: $pr(\tilde{\ell}_j) = c_j/n$, kde c_j je četnost (frekvence) prvku $\tilde{\ell}_j$ v sekvenci ℓ
- poznámka : četnost výskytu všech možných hodnot v sekvenci je celková pravděpodobnost

$$pr(\ell) = \sum_{j=1}^{m} pr(\tilde{\ell}_j) = 1$$

m ... počet různých prvků v sekvenci ℓ

STATISTICKÉ PARAMETRY NÁHODNÉHO VZORKU

- střední hodnota vzorku $\ell = \{\ell_i\}_{i=1}^{n}$:

$$\boxed{m_\ell = \frac{1}{n} \sum_{i=1}^{n} \ell_i = \sum_{j=1}^{m} \tilde{\ell}_j \, pr(\tilde{\ell}_j)}$$

  - použití expektace $m_\ell = E(\ell)$

- variance náhodného vzorku :

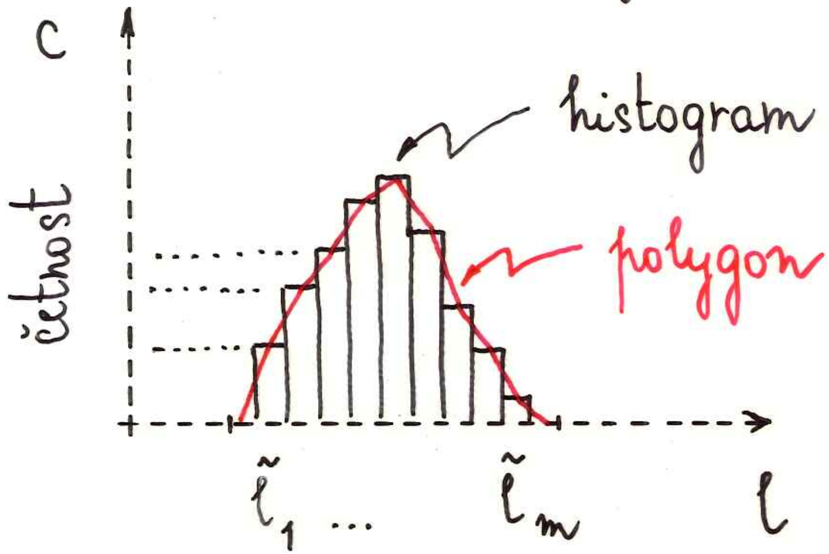
$$\boxed{s_\ell^2 = \frac{1}{n} \sum_{i=1}^{n} (\ell_i - m_\ell)^2 = \sum_{j=1}^{m} (\tilde{\ell}_j - m_\ell)^2 \, pr(\tilde{\ell}_j)}$$

  - použití expektace $s_\ell^2 = E[\ell - E(\ell)]^2$

- další parametry : $\left.\begin{array}{l}\text{- medián} \\ \text{- rozsah}\end{array}\right\}$ náhodného vzorku

HISTOGRAM A POLYGON

- setříděný náhodný vzorek $\{l_i\}_{i=1}^{n}$ lze rozdělit do tříd $\{\tilde{l}_j\}_{j=1}^{m}$



- pravděpodobnost, že $l_i = \tilde{l}_j$

$$pr(\tilde{l}_j) = \frac{c_j}{n} ,$$

kde c_j ... četnost l v $\tilde{l}_j$

- pospojované středy tvoří tzv. polygon (nutno přidat 2 třídy)
- celková oblast pod polygonem je rovna 1
- polygon lze využít pro určení pravděpodobnosti hodnot vzorku
  - integrace

NÁHODNÁ PROMĚNNÁ (RANDOM VARIABLE)

- zobecnění konceptu náhodného vzorku
- každé hodnotě $l \in \mathbb{R}$ je přiřazeno kladné reálné číslo $\phi(l)$
- reálná funkce $\phi \in \{ \mathbb{R} \to \langle 0, \infty) \}$ se nazývá **funkce hustoty pravděpodobnosti** proměnné l
- funkce ϕ představuje vyhlazený polygon, tj. situaci, kdy $n \to \infty$ a také $m \to \infty$ $(\tilde{l}_j \to dl)$
- stále platí, že $\int_{\mathbb{R}} \phi(l)\, dl = 1$

(porovnej s výrazem pro náhodný vzorek)

NÁHODNÝ VZOREK VS. NÁHODNÁ PROMĚNNÁ

- náhodný vzorek = vybrané (měřené) hodnoty náhodné proměnné
- střední hodnota náhodné proměnné ℓ :

$$\boxed{\mu_\ell = \int_{-\infty}^{\infty} \phi(\ell, \xi_1^\ell, \xi_2^\ell, \dots)\, \ell \, d\ell} \neq m_\ell$$

- variance náhodné proměnné :

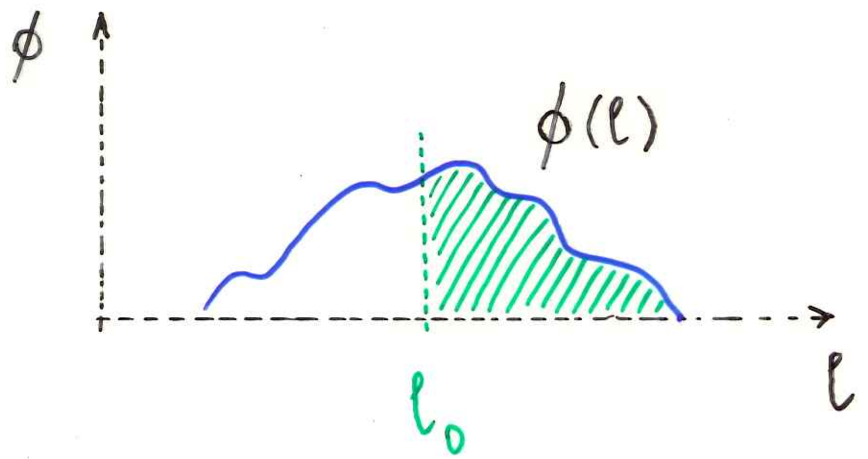
$$\boxed{\sigma_\ell^2 = \int_{-\infty}^{\infty} \phi(\ell, \xi_1^\ell, \xi_2^\ell, \dots)\,(\ell - \mu_\ell)^2 \, d\ell} \neq s_\ell^2$$

- mělo by ale stále platit : $\mu_\ell = E(m_\ell)$

$$\sigma_\ell^2 = E(s_\ell^2)$$

FUNKCE HUSTOTY PRAVDĚPODOBNOSTI

- funkce $\phi \in \{ R \rightarrow \langle 0, \infty) \}$: platí, že



$$\int_{-\infty}^{\infty} \phi(\ell)\, d\ell = pr(-\infty \leq \ell \leq \infty) = 1$$

$$\int_{\ell_0}^{\infty} \phi(\ell)\, d\ell = pr(\ell \geq \ell_0)$$

$\vdots$ atd.

- funkce ϕ má obecně libovolný počet parametrů, zde pouze 2:

$$\phi(\ell, \xi_1^\ell, \xi_2^\ell) \; : \; \xi_1^\ell = \mu_\ell \, , \; \xi_2^\ell = \sigma_\ell^2$$

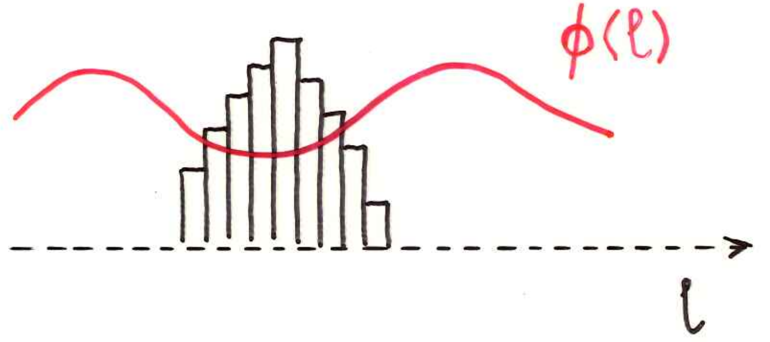
- nebo se použijí hodnoty m_ℓ a s_ℓ^2 (vzorek)

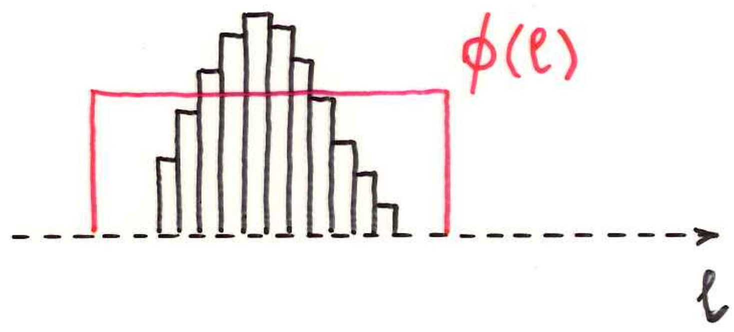
ZÁKLADNÍ POSTULÁT STATISTIKY

- každý náhodný vzorek má svoji "mateřskou" náhodnou proměnnou, které odpovídá příslušná funkce hustoty pravděpodobnosti
- správnost této hypotézy bude muset být testována:
  - správnost matematického vyjádření funkce ϕ
  - správnost volby parametrů $\xi_1^{\ell}, \xi_2^{\ell}, \dots$
- správně zvolená funkce ϕ = správně zodpovězené otázky ohledně pravděpodobnosti výskytu hodnot ve vzorku a dalších statistických vlastností

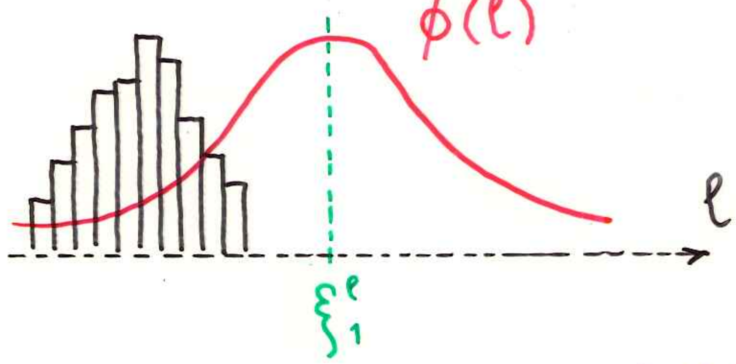
PŘÍKLADY ŠPATNÉHO ODHADU FUNKCE HUSTOTY PRAVDĚPODOBNOSTI

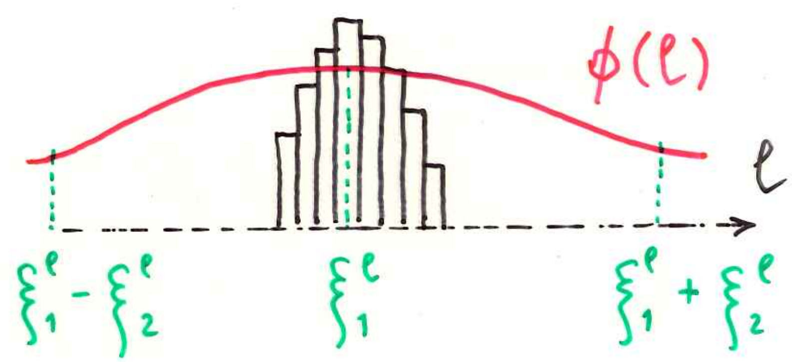
- matematický popis funkce ϕ :





- parametry funkce ϕ : $\phi(\ell, \xi_1^\ell, \xi_2^\ell)$





$\Rightarrow$ nutné testy!

VEKTOR NÁHODNÝCH PROMĚNNÝCH

- další zobecnění konceptu : $\underline{l} = [\, l_1 \; l_2 \dots l_n \,] \;+\; \phi(\underline{l})$

  každá složka vektoru $\underline{l}$ je náhodná proměnná

- funkce hustoty pravděpodobnosti $\phi(\underline{l}) \in \{ R^n \to \langle 0, \infty) \}$

  obecně $\phi(\underline{l}, \xi_1^l, \xi_2^l) = \phi_e(\underline{\mu}_e, \underline{\sigma}_e^2)$,

  kde $\underline{\mu}_e = [\, \mu_{e_1} \; \mu_{e_2} \dots \mu_{e_n} \,]^T$, $\underline{\sigma}_e^2 = [\, \sigma_{e_1}^2 \dots \sigma_{e_n}^2 \,]^T$

- pravděpodobnost :

$$pr(\underline{l}_1 \le \underline{l} \le \underline{l}_2) = \int_{\underline{l}_1}^{\underline{l}_2} \phi(\underline{l}, \xi_1^l, \xi_2^l)\, d\underline{l}$$

… n-dimenzionální integrace

VARIANČNĚ - KOVARIANČNÍ MATICE

- kovariance mezi dvěma prvky vektoru $\underline{l}$ - l_j a l_k :

$$\sigma_{jk} = \int_{R^n} (l_j - \mu_{l_j})(l_k - \mu_{l_k})\, \phi(\underline{l})\, d\underline{l} =$$

$$= E\left[(l_j - \mu_{l_j})(l_k - \mu_{l_k})\right] = \sigma_{kj}$$

- pro statisticky nezávislé prvky platí $\sigma_{jk} = \sigma_{kj} = 0$
- matice $\underline{C}_l = \begin{bmatrix} \sigma^2_{l_1} & \sigma_{l_1 l_2} & \dots & \sigma_{l_1 l_n} \\ \sigma_{l_2 l_1} & \sigma^2_{l_2} & \dots & \vdots \\ \vdots & & & \\ \sigma_{l_n l_1} & \dots & & \sigma^2_{l_n} \end{bmatrix} = E\left[(\underline{l} - \underline{\mu}_l)(\underline{l} - \underline{\mu}_l)\right]$

vlastnosti $\underline{C}_l$?

STATISTICKÁ ZÁVISLOST – NEZÁVISLOST

- někdy lze psát $\phi(\underline{\ell}) = \phi_1(\ell_1) \cdot \phi_2(\ell_2) \dots \phi_n(\ell_n)$
- pak také platí, že

$$\boxed{pr(\underline{\ell}_1 \leq \underline{\ell} \leq \underline{\ell}_2) = \prod_{i=1}^{n} pr(\ell_{1,i} \leq \ell_i \leq \ell_{2,i})}$$

a komponenty vektoru $\underline{\ell}$ jsou <u>statisticky nezávislé</u>

- velmi důležitý koncept, neboť variance nevyjadřují stochastické vlastnosti vektoru náhodných proměnných úplně => nutnost rozšíření vektoru variancí na

<u>varianční – kovarianční matici</u>

ZPĚT K REALITĚ : NÁHODNÉ VZORKY

- vektor měření $\underline{\ell} = [\ell_1 \; \ell_2 \; \dots \; \ell_n]^T$, ℓ_i ... náhodný vzorek

  - vektor středních hodnot $\underline{m}_\ell = [m_{\ell_1} \; m_{\ell_2} \; \dots \; m_{\ell_n}]^T$

  - vektor variancí $\underline{s}^2_\ell = [\, s^2_{\ell_1} \; s^2_{\ell_2} \; \dots \; s^2_{\ell_n} \,]^T$

+ kovariance $s_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left[(\ell_j - m_{\ell_j})(\ell_k - m_{\ell_k}) \right]$

pozor : výpočet kovariancí mezi prvky vektoru náhodných vzorků je většinou nemožný (viz. předpoklady) =>

=> varianční - kovarianční matici nelze dát dohromady

- její odhad : kritická záležitost !

MÍRA STATISTICKÉ ZÁVISLOSTI – KORELACE

- korelační koeficient $\varrho_{jk} = \dfrac{S_{jk}}{S_j \cdot S_k}$, $S_j \neq 0$, $S_k \neq 0$

- pro korelační koeficient platí :

  a) $\varrho \in \langle -1, +1 \rangle$

  b) $\varrho_{jk} = 0 \Rightarrow \xi_j$ a ξ_k nekorelované

  c) $|\varrho_{jk}| < 1 \Rightarrow \xi_j$ a ξ_k korelované

  d) $\varrho_{jk} = \pm 1 \; \ldots \; \xi_j$ a ξ_k úplně korelované

  vztah ke statistické závislosti ?

- vektor náhodných proměnných : $S \rightarrow \varrho$

MATEMATICKÁ STATISTIKA

- systematická metoda pro studium stochastických vlastností parametrů
- upozornění: statistika musí být používána opatrně ve spojení s praktickými zkušenostmi, dalšími důkazy a "selským rozumem"
- statistika může určit stupeň důvěryhodnosti v řešení vyrovnání
- možné podělení statistiky:
  - parametrická: normální rozdělení
  - robustní: i pro nenormální rozdělení
  - neparametrická: nezávislé na rozdělení

  $\Big\}$ bez apriori informací
  - Bayesovská: apriori informace o $\underline{x}$ a $\underline{C}_x$

STATISTICKÉ TESTOVÁNÍ

- může poskytnout odpovědi na otázky typu:
  - odpovídá zvolená funkce hustoty pravděpodobnosti danému vzorku?
  - jsou parametry funkce hustoty pravděpodobnosti zvoleny správně?
  - jsou tyto parametry konsistentní s jejich jinými odhady?
- hypotéza: kvantitavní výrok o funkci hustoty pravděpodobnosti a hodnotách jejích parametrů
- nulová hypotéza vs. alternativní hypotézy:

  testování na základě statistické proměnné, která je funkcí parametrů funkce hustoty pravděpodobnosti

STATISTICKÁ PROMĚNNÁ

- nástroj k ověření či vyloučení platnosti dané hypotézy
- obecně: statistická proměnná y má funkci hustoty pravděpodobnosti s obdobné funkce vzorku ℓ:

$$\phi_y(\xi_1^y, \xi_2^y) = \det \underline{J}\, \phi_\ell(\xi_1^\ell, \xi_2^\ell)$$

- důležité: každá hypotéza má svoji funkci ϕ_y a proměnnou y
- problém: většina testů hypotéz platí pro náhodné proměnné a jejich platnost pro náhodné vzorky je limitována

$\Rightarrow$ <u>nutné testovat vztah vzorku a proměnné!</u>

TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

- testování platnosti hypotézy $\mathcal{H}_0$: přijmout - vyloučit
- stejný osud se týká testu alternativní hypotézy $\mathcal{H}_1$:
  testování $\mathcal{H}_0$ a $\mathcal{H}_1$ by mělo ukázat, které z hypotéz se dá více věřit => žádná absolutně platná odpověď !
- **pozor** : neplatnost hypotézy $\mathcal{H}_0$ (dle výsledku testu) neznamená platnost hypotézy $\mathcal{H}_1$ a obráceně !
- každý výsledek testu má pouze omezenou platnost resp. je spojen s jistou mírou pravděpodobnosti => chyby jsou možné (např. test přijme $\mathcal{H}_0$, ale $\mathcal{H}_0$ neplatí)

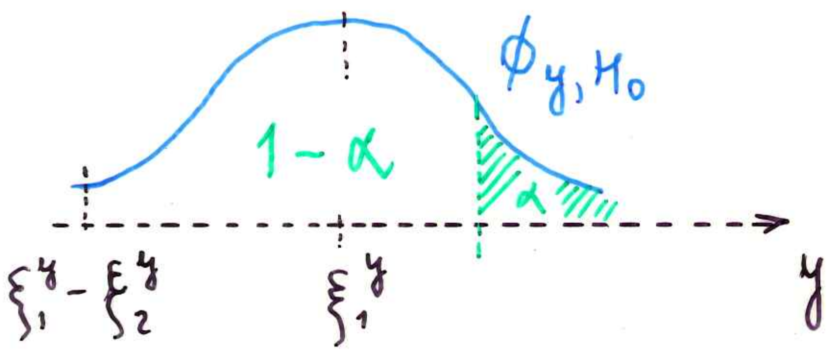
TYPY CHYB PŘI STATISTICKÉM TESTOVÁNÍ HYPOTÉZ

- chyba 1. druhu : H_0 je zamítnuta, ale ve skutečnosti je správná
- chyba 2. druhu : H_0 je přijata, ale ve skutečnosti není správná
- pravděpodobnost chyby 1. druhu je dána úrovní významnosti α, $\alpha \in \langle 0,1 \rangle$ - volí se pokud možno co nejmenší; bohužel pro $\alpha = 0$ neexistuje žádná přijatelná hypotéza H_0
- pravděpodobnost chyby 2. druhu je dána sílou testu $1-\beta$, $\beta \in \langle 0,1 \rangle$, hodnota β by měla být také malá a $\beta > 0$, hodnoty α a β jsou na sobě závislé !

HLADINA VÝZNAMNOSTI α

- v geodézii se nejčastěji používá $\alpha \in \langle 0{,}01 ; 0{,}05 \rangle$
- rozdíl $1-\alpha$ se nazývá hladina pravděpodobnosti

$$pr(y > y_{1-\alpha}) = \int_{y_{1-\alpha}}^{\infty} \phi(y, \xi_1^y, \xi_2^y)\, dy = \alpha$$

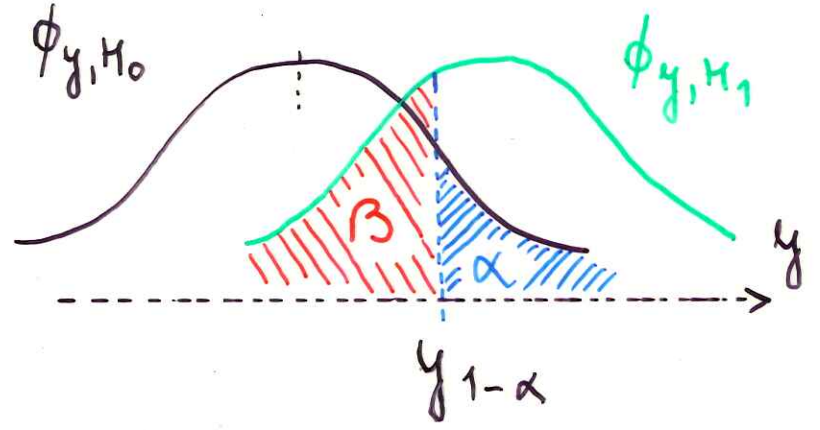


- alternativní hypotéza H_1: funkce ϕ_{y,H_1} má většinou stejnou formu jako ϕ_{y,H_0}, liší se hodnoty parametrů tj.

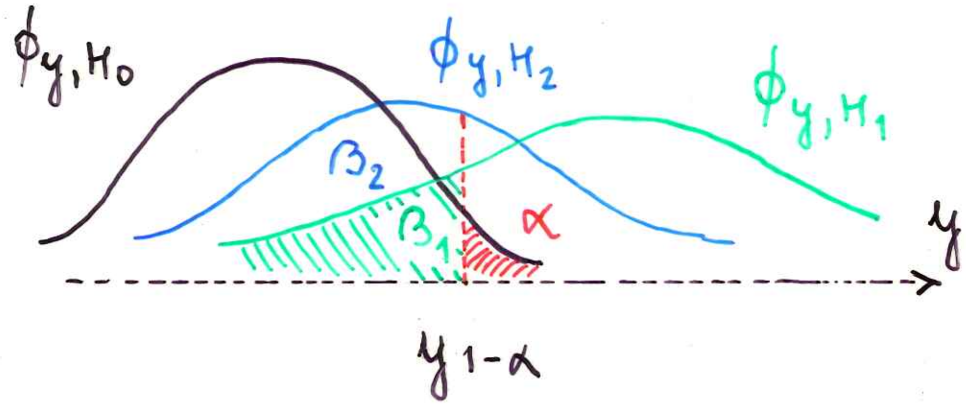
$$\phi_{y,H_0}(\xi_y^1, \xi_2^y) \text{ vs. } \phi_{y,H_1}(\dot{\xi}_1^y, \dot{\xi}_2^y)$$

SÍLA STATISTICKÉHO TESTU β

- míra odlišitelnosti dvou statistických hypotéz



- alternativní hypotéza $\mathcal{H}_1$ se má dostatečně lišit od $\mathcal{H}_0$ ale ne zbytečně mnoho



- hypotéza $\mathcal{H}_1$ je lepší než $\mathcal{H}_2$, neboť $\beta_1 < \beta_2$

- platí: $pr(y < y_{1-\alpha}) = \int_{-\infty}^{y_{1-\alpha}} \phi_{H_1}(y, \xi_1^y, \xi_2^y)\, dy = \beta$

TESTOVÁNÍ HYPOTÉZY H_0 A JEJÍ ALTERNATIVY

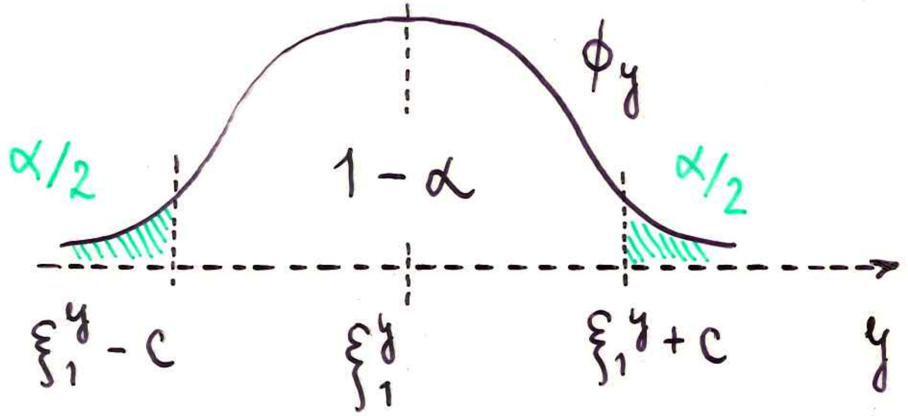
výstup / situace	test říká přijmi H_0	test říká zamítni H_0
H_0 je správné	správné rozhodnutí $pr = 1 - \alpha$	chyba 1. druhu $pr = \alpha$
H_0 je špatné	chyba 2. druhu $pr = \beta$	správné rozhodnutí $pr = 1 - \beta$

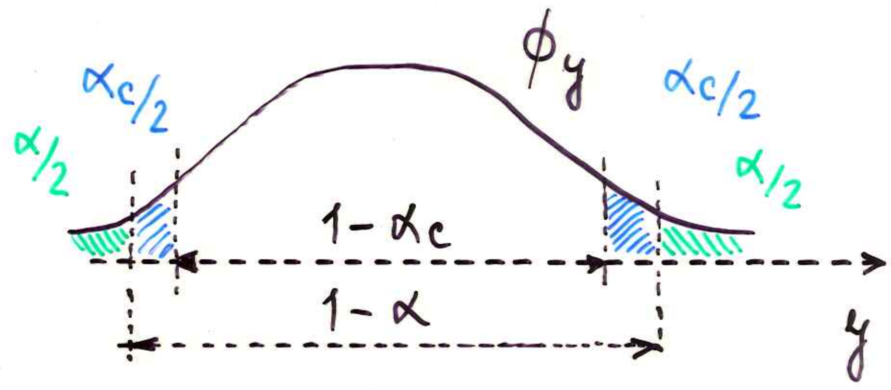
- zmenšení hodnoty α = zvětšení hodnoty β

$\Rightarrow$ nutný kompromis !!

OBLASTI PRAVDĚPODOBNOSTI

- alternativa k testování hypotéz (= nutná formulace hypotéz)
- pravděpodobnost $pr (y_{\frac{\alpha}{2}} < y < y_{1-\frac{\alpha}{2}}) = 1 - \alpha$





- kritická úroveň významnosti $\alpha_c = pr (y > y_0)$

  nastane-li $\alpha > \alpha_c \Rightarrow \mathcal{H}_0$ není přijatelná na hladině α

- testy mohou být jednostranný či dvoustranný (viz. obrázky)