

Statistické minimum – příklady zkouškových otázek.

1. Definujte pojem střední hodnota náhodné veličiny a aritmetický průměr. Vysvětlete rozdíl mezi nimi.

$$\mathbb{E} X = \int_{\Omega} x f(x) dx$$

hustota, prst
náhodné veličiny

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

\mathbb{E} je blána z celé dist
 \bar{x} je jen z podskupiny
většinou odhadujeme \mathbb{E} jeho \bar{x}

2. Uvažujte náhodný vektor $\mathbf{X} = (X_1, X_2, \dots, X_n)$. Definujte kovarianční a korelační matici. Jaké mají tyto matice vlastnosti? K čemu se dají použít?

$$\Sigma \stackrel{def}{=} \text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$$

↳ štercová, symetrická
posit. semidef.

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \dots \\ \vdots & \ddots \end{pmatrix}$$

$$\text{cov}(X_i, X_j)$$

$$\text{corr}(X_i, X_j)$$

$$\rightarrow \text{corr}(X, X) = \text{var}(X)$$

$$\text{corr}(X, Y) = \text{wr}(Y, X)$$

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}X} \sqrt{\text{var}Y}} \sim (-1, 1)$$

↳ unnormalizovaná korelace

$$\rightarrow \text{corr} = \begin{pmatrix} 1 & \dots \\ \vdots & 1 \end{pmatrix}$$

$$\rightarrow \text{corr}(X, Y) \sim \text{"míra" linearity' závislosti}$$

↳ normovaný různaní mezi relacemi

↳ korelační koeficienty pro určení závislosti

3. Co je to distribuční funkce? Co je to kvantil a kvantilová funkce? Co je to funkce pravděpodobnostní hustoty? Definujte pojmy formálně a uveděte jaký je mezi nimi vztah.

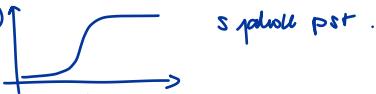
$$F_X(x) = P(X \leq x)$$

náh. vel. osa X

kvantilová funkce

$$\hookrightarrow \text{inverzní k } F_X(x) \dots x = F^{-1}(y)$$

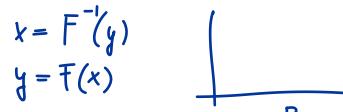
$$y = F(x)$$



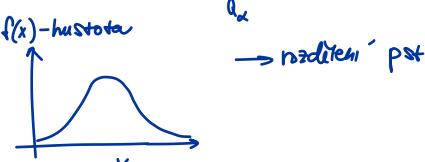
s pravou prst.

kvantil = hodnota kvantilové funkce

$$Q_\alpha = F^{-1}(P)$$



hustota prst $f(x) =$
prst ze náhodné vel.
násobek hodnoty +
dáleho inverzky



4. Vysvětlete vlastními slovy význam p -hodnoty statistického testu. Jaký je výklad hladiny významnosti α ? Předpokládejte, že p -hodnota nějakého testu je 0,045? Jaká je pravděpodobnost, že nulová hypotéza tohoto testu neplatí? Jak to souvisí s α ?

p -value = prst., že hodnota statistického testu bude taková nebo extremlí než mi vyšlo z p že H_0 platí

$\alpha = p$, že zamíříme H_0 , i když platí → chci H_0 zamířit jen 5%
→ $P(X|H_0)$

nesprávně
to co se
snažím
vyhnout

$P(H_0|X) \rightarrow$ na toto se ptají

↳ vzdáleněk určit

$$\alpha = 0,05$$

$0,045 < 0,05 \rightarrow H_0$ zamítlám

5. Vysvětlete pojem *interval spolehlivosti* (confidence interval). Co nejpřesněji popište, jak byste spočítal/a 99% interval spolehlivosti odhadu střední hodnoty normálního rozdělení z malého vzorku velikosti m .

$$[\hat{\mu} - \underbrace{t_{1-\frac{\alpha}{2}, m-2}}_{1-\frac{\alpha}{2} \text{ kvádratik}} SE(\hat{\mu}), \hat{\mu} +$$

$1-\frac{\alpha}{2}$ kvádratik

$$99\% = 100(1-\alpha)$$

6. Vysvětlete význam pojmu *chyba prvního druhu* a *chyba druhého druhu* používaných k popsání konkrétních chyb v procesu testování statistických hypotéz. Vysvětlete jak spolu tyto chyby souvisí. Popište, jak se dá výskyt těchto chyb ovlivnit.

7. Formulujte *centrální limitní větu*. Kde se dá využít?

po hodně vzorků

průměry lib. náhodných reálných množin Norm

8. Motivujte zavedení *Studentova t-rozdělení*. Ke kterému rozdělení se t-rozdělení asymptoticky blíží s rostoucím počtem stupňů volnosti? Vysvětlete za jakých okolností a jak se od tohoto rozdělení odlišuje. K čemu se používá?

vhodné pro malý počet vzorků

bliží se k normálnímu rozdělení, když je vzorků hodně

křivka t-rozdělení je definována počtem vzorků, které mám k dispozici

odlišuje, když je málo vzorků, má těžší konce

využití pro t-test - testování hypotézy, CI, snažím se zjistit střední hodnotu malého počtu vzorků

9. Definujte *věrohodnostní funkci* (likelihood). K čemu se používá metoda maximální věrohodnosti? Pojmenujte alespoň dvě metody, které se k maximalizaci věrohodnosti používají. Vysvětlete, proč se často používá logaritmus věrohodnosti, jak se dá věrohodnosti využít při testování statistických hypotéz?

$P(\text{param} | \mathbf{x}) =$

10. Formálně definujte *multivariátní normální rozdělení*. Kolik parametrů budeme obecně potřebovat pro popis tohoto rozdělení v d dimenzích? Lze počet těchto parametrů nějak omezit? Jaké důsledky toto omezení může mít?

multidimenzionální Norm

navíc kovarianční matice - místo rozptylů

vektory průměrů

pro popis v d -dimenzích - potřeba parametrů --> průměr má d -hodnot + kovarianční matice $d \times d$, ale je symetrická ($d \times (d+1)$)

omezení --> když vezmu nekorelovanou kovarianční matici (mimo diagonálu budou nuly)
důsledky - přicházíme o informaci z korelace

11. Definujte alespoň dvě často používaná rozdělení diskrétní náhodné veličiny. Pojmenujte jejich parametry. Na příkladech naznačte, kdy se tato rozdělení dají využít.

$$\text{Binomické } (n, p) \rightarrow P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

↓
 počet
 z 8 hodin
 hostkou
 3x padne G
 ↓
 x

$$\text{Poissonovo } (\lambda) \rightarrow P(Y=k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!} \quad \longrightarrow \text{hodná žádat žádost}$$

↓
 počet výskytů

12. Vysvětlete rozdíl mezi parametrickým a neparametrickým statistickým testem. Pojmenujte základní výhody a nevýhody obou přístupů. Jmenujte alespoň jeden parametrický a jeden neparametrický test.

předpoklad, že
zkušené veličiny
pocházejí z

T-test

vyžadují
specifikovanou
Sign test
Wilcoxon

Parametrické metody (parametric survival analysis) vyžadují specifikaci konkrétního rozdělení náhodné veličiny, zatímco neparametrické metody (nonparametric survival analysis) žádné zvláštní předpoklady ohledně rozdělení pravděpodobnosti náhodné veličiny nevyžadují

13. K následující statistické úloze přiřaďte vhodný test. Bylo testováno 11 automobilů určité značky. Ověrte, zda se jejich pravé a levé přední pneumatiky ojíždějí srovnatelně. (Předpokládejte, že ojetí pneumatik [mm] má normální rozdělení. Z následujících testů vyberte ten nejlepší, svou volbu vysvětlete 2. (dvouvýběrový t-test) Friedmanův test, jednovýběrový t-test, jednovýběrový Wilcoxonův test, test o parametru alternativního rozdělení, test o rozptylu normálního rozdělení, (párový t-test). Co by se stalo pokud byste použili druhý nejlepší test z daného seznamu? A

jednovýběrový oboustraný - liší se krevní tlak (na obě strany)

jednovýběrový jednostranný - KT větší než