Datum poslední aktualizace: 19. prosince 2021

NMSA331 MATEMATICKÁ STATISTIKA 1
POZNÁMKY O ZKOUŠCE

Zkouška má písemnou (100 min) a ústní část. V případě distančního zkoušení (které je popsáno níže) může z povahy věci ústní část zkoušky probíhat trochu odlišným způsobem než v případě standardního prezenčního zkoušení.

Během písemné či ústní části **není dovoleno** používat jakékoliv poznámky či elektronické materiály. Také je zakázána jakokoliv komunikace s dalšími osobami.

1. Příprava ke zkoušce

1.1. Co je dobré znát?

- Všechny používané **definice**. Kromě definici vyslovených na přednášce si nezapomeňte připomenout definici *měřitelnosti, χ^2-rozdělení* a *t-rozdělení*.
- Všechna vyslovená **tvrzení** a **věty**. Pokud se na přednášce dělal důkaz, tak také **důkaz** daného tvrzení. Pokud se na přednášce nějaký důkaz (či jeho část) nedělal s tím, že je to analogické, tak si to rozmyslete, protože to může být součástí zkoušky.
- Není třeba umět dokazovat tvrzení z Appendixu *Poznámek k přednášce*. Tvrzení a definice z tohoto Appendixu je však třeba znát a umět používat v rozsahu, v jakém byla používána na přednášce.
- Otázka může také vycházet z **praktického příkladu**, viz vzorové příklady v *Poznámkách k přednášce*.
- Součástí zkoušky mohou být také modifikace příkladů, které se probíraly na **cvičení** nebo byly za domácí úlohu. Mimo jiné byste měli umět vyšetřit *konzistenci, nestrannost* a *asymptotické rozdělení odhadů*.
- I když se to u konkrétních testů zpravidla explicitně nedělalo, tak je třeba vědět, jak se z obecné definice **p-hodnoty** dospěje ke konkrétnímu vzorci pro p-hodnotu daného testu.
- Není třeba znát různé korekce testových statistik v případě, že některý předpoklad není splněn. Tyto korekce jsou uváděny pouze pro úplnost. Na druhou stranu se sluší vědět, zda v případě nesplnění předpokladů dodržuje test hladinu alespoň asymptoticky.

1.2. **Obecné doporučení.** Při přípravě na zkoušku doporučuji **psát si** vše (na papír) a pak si to kontrolovat, zda to dává smysl. Umění zkontrolovat si svůj zápis je klíčové zejména během písemné části zkoušky, při které nemáte možnost opravy a zkoušející může hodnotit pouze to, co je napsáno na papíře. Jediný hloupý „překlep“ může naprosto znehodnotit výsledek Vašeho snažení.

V matematice platí, že to co napíšete (řeknete) by mělo dávat smysl. Proto je důležité **rozumět definicím a zněním vět**. Když se učíte důkazy (resp. jiná odvození), tak je dobré si **rozmyslet jednotlivé kroky důkazu a jejich návaznost**. Je daleko přijatelnější, pokud nevíte, jak se některý krok technicky provede, než když sice znáte podrobně některé kroky, ale nějaký krok Vám chybí a v myšlence důkazu je nevysvětlený skok.

2. Některé časté chyby

1. Často se zapomíná na **předpoklad nezávislosti**. Např. ve dvouvýběrovém problému musí být oba náhodné výběry nezávislé. Podobně se na nezávislost zapomíná v definici χ^2-rozdělení, t-rozdělení, resp. F-rozdělení.

Obecně se dá říct, že kdykoliv máte více než jednu náhodnou veličinu, tak musíte vyjasnit, zda zmiňované náhodné veličiny jsou nezávislé.

2. Někdy se jako předpoklad párového testu uvádí, že $X_1, \dots, X_n$ je náhodný výběr a $Y_1, \dots, Y_n$ je náhodný výběr. To však není správně, protože není jasné, v jakém vztahu jsou náhodné veličiny $X_1, \dots, X_n$ s náhodnými veličinami $Y_1, \dots, Y_n$.

3. Pro nezávislé náhodné veličiny X_1 a X_2 **neplatí**: $\mathsf{var}\,(X_1 - X_2) = \mathsf{var}\,(X_1) - \mathsf{var}\,(X_2)$.

4. Chybí vyznačení, o jaký *typ konvergence* se jedná (obyčejnou, v pravděpodobnosti, v distribuci, skoro jistě).

5. Testová statistika (a vlastně ani jakákoliv jiná statistika) nesmí obsahovat neznámé parametry, protože ji musíme být schopni spočítat na základě dat. Tj. pro $X_1, \dots, X_n$ náhodný výběr z rozdělení s neznámou střední hodnotou μ není následující náhodná veličina

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \mu)}{S_n}$$

testovou statistiku, protože její předpis obsahuje neznámý parametr μ. Pokud však testujeme např. $H_0 : \mu = \mu_0$, kde μ_0 je nějaká předepsaná (a tudíž známá) hodnota, pak

$$T_n = \frac{\sqrt{n}\,(\overline{X}_n - \mu_0)}{S_n}$$

již testovou statistikou je.

6. V důkazu tvrzení $T_n = \frac{\sqrt{n}\,(\overline{X}_n - \mu)}{S_n} \sim t_{n-1}$ za předpokladu, že $X_1, \dots, X_n$ je náhodný výběr z normálního rozdělení $\mathsf{N}(\mu, \sigma^2)$ studenti občas argumentují tím, že $\frac{\sqrt{n}\,(\overline{X}_n - \mu)}{\sigma} \xrightarrow[n\to\infty]{\mathsf{d}} \mathsf{N}(0,1)$. To je sice pravda, ale nedostali bychom pak přesné t-rozdělení. Díky předpokladu normality, však s využitím definice mnohorozměrného normálního rozdělení platí, že $\frac{\sqrt{n}\,(\overline{X}_n - \mu)}{\sigma}$ má **přesné** rozdělení $\mathsf{N}(0,1)$.

7. Žádný z t-testů nepotřebuje spojitost rozdělení. Na druhou stranu znaménkový test (resp. testy o kvantilech) sice vyžadují spojitost rozdělení, ale nevyžadují existenci konečného rozptylu (ani střední hodnoty).

8. Je třeba si dát pozor na to, co tvrdíme o rozdělení testové statistiky v případě **jednostranných hypotéz**. Např. uvažujme $X_1, \dots, X_n$ náhodný výběr z alternativního rozdělení s parametrem p_X a testujeme jednostrannou nulovou hypotézu $H_0 : p_X \leq p_0$ proti alternativě $H_1 : p_X > p_0$. Potom **nelze říct**, že testová statistika

$$Z_n = \frac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}, \quad \text{kde } \widehat{p}_n = \overline{X}_n,$$

má za nulové hypotézy asymptoticky $\mathsf{N}(0,1)$ rozdělení. Toto tvrzení by platilo pouze pro $p_X = p_0$ (tj. skutečná hodnota parameteru je na hranici nulové hypotézy a alternativy).

9. Při výpočtu *hladiny testu* je zapotřebí si dávat, aby byly vysvětleny všechny symboly. Tedy v obecné definici hladiny testu

$$\sup_{F\in\mathcal{F}_0} \mathsf{P}(S_n(\boldsymbol{X})\in\mathcal{C}), \quad \text{resp.} \quad \sup_{F\in\mathcal{F}_0} \lim_{n\to\infty} \mathsf{P}(S_n(\boldsymbol{X})\in\mathcal{C})$$

je třeba u konkrétního testu vysvětlit, co je $S_n(\boldsymbol{X})$, $\mathcal{F}_0$ a $\mathcal{C}$.

10. U každého testu je třeba se zamyslet nad tím, **jaké hodnoty testové statistiky svědčí proti hypotéze** (a ve prospěch alternativy). Od toho se totiž odvíjí tvar kritického oboru a vzorec pro p-hodnotu.

11. Při porovnávání pravděpodobností v multinomickém rozdělení nelze považovat

$$\frac{\sqrt{n}\left(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{p}}_n-\gamma_0\right)}{\sqrt{\boldsymbol{c}^{\mathsf{T}}[\mathsf{diag}\,(\boldsymbol{p})-\boldsymbol{p}^{\otimes 2}]\boldsymbol{c}}}$$

za testovou statistiku. Podobně nelze považovat

$$\left(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{p}}_n-u_{1-\alpha/2}\sqrt{\tfrac{\boldsymbol{c}^{\mathsf{T}}[\mathsf{diag}\,(\boldsymbol{p})-\boldsymbol{p}^{\otimes 2}]\boldsymbol{c}}{n}},\ \boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{p}}_n+u_{1-\alpha/2}\sqrt{\tfrac{\boldsymbol{c}^{\mathsf{T}}[\mathsf{diag}\,(\boldsymbol{p})-\boldsymbol{p}^{\otimes 2}]\boldsymbol{c}}{n}}\right).$$

za interval spolehlivosti pro parametr $\theta=\boldsymbol{c}^{\mathsf{T}}\boldsymbol{p}$.

12. Pokud mluvíme o rozdělení testové statistiky, tak je třeba říct, zda je to rozdělení za nulové hypotézy nebo za alternativy. Na předpoklad platnosti nulové hypotézy se často zapomíná zejména při uvádění vlastností Wilcoxonovy a Kolmogorovovy-Smirnovovy statistiky.

13. Součástí „praktických příkladů“, kdy je třeba navrhnout a popsat vhodný test, by měly být i předpoklady o zúčastněných náhodných veličinách. Aby bylo jasné, co je s čím **nezávislé**, případně **stejně rozdělené**.

14. Veškeré symboly (obzvláště však ty zavedené až v této přednášce) by měly být vysvětleny, resp. definovány. Např. $\widehat{F}_X$, S_X,...

15. Je zapotřebí psát, zda je test (p-hodnota) **přesný** nebo **asymptotický**.

16. Vzhledem k nesymetrickému postavení nulové hypotézy a alternativy umíme ve statistice prokázat pouze alternativu. Proto to, co chceme dokázat, dáváme do alternativy.

17. Ačkoliv máme dualitu intervalů spolehlivosti a testování hypotéz, tak nelze zaměňovat pojmy *spolehlivost, resp. pravděpodobnost pokrytí* (používá se pro intervalové odhady) a *hladina* (používá se při testování).

18. Neznalost jak dokázat konzistenci empirické distribuční funkce pro pevné x.

19. Při výpočtu nestrannosti se někdy může hodit využít toho, že pro každou náhodnou veličinu Y s konečným druhým momentem platí

$$\mathsf{E}\,Y^2=\mathsf{var}\,(Y)+\left(\mathsf{E}\,Y\right)^2.$$

Tedy například střední hodnotu $\mathsf{E}\left(\overline{X}_n\right)^2$ lze snadno spočítat pomocí

$$\mathsf{E}\left(\overline{X}_n\right)^2=\mathsf{var}\left(\overline{X}_n\right)+\left(\mathsf{E}\,\overline{X}_n\right)^2.$$