Datum poslední aktualizace: 13. května 2016

NMSA202 PRAVDĚPODOBNOST A MATEMATICKÁ STATISTIKA POZNÁMKY O ZKOUŠCE

Zkouška má písemnou a ústní část. Nejdříve je písemná část, která se dále dělí na početní část (75 min) a teoretickou část (75 min). Pro připuštění k ústní zkoušce je třeba zvládnout obě části písemné zkoušky alespoň na 50 % možných bodů.

Doporučujeme **nejdříve** každou otázku alespoň **stručně zodpovědět** (tj. např. zformulovat tvrzení, uvést definici, apod.) a pak se teprve pouštět do podrobnější odpovědi (tj. např. důkazu tvrzení, odvozování, apod.).

1. Příprava ke zkoušce

1.1. **Početní část.** Jako přípravu na početní část zkoušky lze doporučit

- aktivní účast na cvičení;
- samostatné propočítání příkladů počítaných, resp. zadaných na cvičení;
- propočítání příkladů ze skript;
- propočítání příkladů z doporučené cvičebnice.

Upozorňuji, že ve srovnání se zápočtovými písemkami bude na příklady více času, příklady však budou o něco méně šablonovité.

Pro početní část budete mít k dispozici vzorce pro intervalové odhady a testy a dále tabulky s hodnotami distribuční funkce $\mathsf{N}(0,1)$ a důležitými kvantily. Nelze používat list A4 popsaný vzorci, natož jakékoliv jiné taháky.

1.2. **Teoretická část.** Doporučuji věnovat pozornost následujícím pojmům, definicím, větám a úlohám. Upozorňuji však, že se **nejedná o seznam otázek**, natož aby to byl úplný seznam všech možných otázek.

1.2.1. *Základy.*

- definice pravděpodobnostního prostoru, co je to sigma-algebra, co je to pravděpodobnost a jaké jsou její vlastnosti;
- definice podmíněné pravděpodobnosti, věta o násobení pravděpodobností, věta o úplné pravděpodobnosti, Bayesova věta;
- nezávislost systému náhodných jevů;
- definice náhodné veličiny – co zaručuje měřitelnost? Co je to sigma algebra indukovaná náhodnou veličinou?;
- distribuční funkce náhodné veličiny a její vlastnosti;
- hustota náhodné veličiny, její vztah k distribuční funkci a její využití pro výpočet $\mathsf{P}(X \in B)$;
- diskrétní vs. spojitá rozdělení;
- střední hodnota náhodné veličiny a její vlastnosti;
- momenty náhodné veličiny (absolutní, centrální);
- rozptyl náhodné veličiny a jeho vlastnosti;
- Čebyševova nerovnost;
- momentová vytvořující funkce a její využití pro výpočet momentů;
- definice náhodného vektoru a jeho distribuční funkce;
- sdružená versus marginální distribuční funkce náhodných vektorů;
- vlastnosti sdružené distribuční funkce (podrobně si to rozmyslete pro $d = 2$);
- sdružená a marginální hustota;
- kovariance, koeficient korelace, varianční matice a její vlastnosti;
- nezávislost náhodných veličin;
- konvoluce;
- rozdělení lineární kombinace nezávislých normálních rozdělení;

1.2.2. *Limitní věty.*

- Cantelliho věta, Borelova věta;
- konvergence skoro jistě implikuje konvergenci v pravděpodobnosti a jejich vztah;
- silný (slabý) zákon velkých čísel;
- Kolmogorovova nerovnost a její využití;
- SZVČ pro stejně a nestejně rozdělené náhodné veličiny, porovnání nutných předpokladů;
- Centrální limitní věta (Ljapunovova) a její důsledek pro stejně rozdělené náhodné veličiny;
- použití CLV na binomické rozdělení;
- věta o asymptotickém rozdělení $Z_n + Y_n$ (resp. $Z_n\, Y_n$), kde Z_n má asymptoticky normované normální rozdělení a Y_n konverguje k nule (resp. k jedničce) v pravděpodobnosti;

1.2.3. *Teorie odhadu a testování hypotéz.*

- formulace úlohy bodového odhadu;
- nestrannost a konzistence bodového odhadu;
- nestrannost a konzistence výběrového průměru a výběrového rozptylu;
- definice intervalového odhadu (oboustranný, horní, dolní),
- definice χ^2-rozdělení a t-rozdělení – není třeba znát hustoty, ale jejich reprezentaci pomocí náhodných veličin;
- kvantily t-rozdělení a χ^2-rozdělení pro velký počet stupňů volnosti;
- metoda maximální věrohodnosti;
- úloha testování hypotéz, chyba prvního a druhého druhu, kritický obor;
- Neymanova-Pearsonova věta a její využití při testování hypotéz;
- test podílem věrohodnosti;

1.2.4. *Regresní analýza.*

- formulace modelu lineární jednoduché regrese, metoda nejmenších čtverců;
- nestrannost a rozptyl odhadů parametrů β_1 a β_2 v modelu lineární regrese;
- speciální případy regresní analýzy – přímka procházející počátkem, dvouvýběrový t-test.

Otázky z této části učiva mohou vypadat následovně:

1. Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$. Určete rozdělení výběrového průměru. Předpokládejte, že σ^2 je známé. Využijte této znalosti ke konstrukci intervalového odhadu pro parametr μ a k testování nulové hypotézy $\mathrm{H}_0 : \mu = \mu_0$ proti oboustranné či jednostranné alternativě.

2. Nechť $X_1, \ldots, X_n$ je náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$. Označme $\overline{X}_n$ výběrový průměr a S_n^2 je výběrový rozptyl. Určete rozdělení náhodné veličiny $V = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n}$ a podrobně vysvětlete, proč tomu tak je. Využijte znalosti rozdělení náhodné veličiny V ke konstrukci intervalového odhadu pro parametr μ a k testování nulové hypotézy $\mathrm{H}_0 : \mu = \mu_0$ proti oboustranné či jednostranné alternativě.

3. Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$. Určete rozdělení náhodné veličiny $(n-1)S_n^2/\sigma^2$. Využijte této znalosti ke konstrukci intervalového odhadu pro parametr σ^2 a k testování nulové hypotézy $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$ proti oboustranné či jednostranné alternativě..

4. Nechť $X_1, \ldots, X_{n_1}$ je náhodný výběr z $N(\mu_1, \sigma^2)$ a $Y_1, \ldots, Y_{n_2}$ je náhodný výběr z $N(\mu_2, \sigma^2)$. Oba výběry jsou vzájemně nezávislé. Určete rozdělení náhodné veličiny

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{S^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad \text{kde} \quad S^{*2} = \frac{1}{n_1 + n_2 - 2} \left((n_1 - 1) S_{1,n_1}^2 + (n_2 - 1) S_{2,n_2}^2 \right).$$

Využijte této znalosti ke konstrukci intervalového odhadu pro parametrickou funkci $\mu_1 - \mu_2$ a k testování nulové hypotézy $\mathrm{H}_0 : \mu_1 - \mu_2 = \delta$ proti oboustranné či jednostranné alternativě.

5. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení s konečným nenulovým rozptylem. Odvoďte asymptotické rozdělení náhodné veličiny $\frac{\sqrt{n}(\overline{X}_n - \mathsf{E}\, X_1)}{S_n}$. Využijte této znalosti ke konstrukci intervalovému odhadu (o asymptotické spolehlivosti $1 - \alpha$) pro střední hodnotu $\mathsf{E}\, X_1$ a k testování nulové hypotézy $\mathrm{H}_0 : \mathsf{E}\, X_1 = \mu_0$ proti oboustranné či jednostranné alternativě.

6. Nechť $X_1, \ldots, X_n$ je náhodný výběr z alternativního rozdělení s parametrem p. Odvoďte asymptotické rozdělení náhodné veličiny $\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}$. Využijte této znalosti ke konstrukci intervalovému odhadu (o asymptotické spolehlivosti $1-\alpha$) pro parametr p.

7. Nechť $X_1, \ldots, X_n$ je náhodný výběr z alternativního rozdělení s parametrem p. Odvoďte asymptotické rozdělení náhodné veličiny $\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}}$. Využijte této znalosti k testování nulové hypotézy $\mathrm{H}_0 : p = p_0$ proti oboustranné či jednostranné alternativě.

8. Nechť $X_1, \ldots, X_n$ je náhodný výběr z Poissonova rozdělení s parametrem λ. Odvoďte asymptotické rozdělení náhodné veličiny $\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\overline{X}_n}}$. Využijte této znalosti ke konstrukci intervalovému odhadu (o asymptotické spolehlivosti $1-\alpha$) pro parametr λ.

2. Časté chyby – početní část

1. V příkladech na klasický pravděpodobnostní prostor se množina elementárních jevů Ω zvolí tak nešťastně, že jednotlivé elementární jevy nemají stejnou pravděpodobnost.

2. Hustotu spojité náhodné veličiny vypočteme jako derivaci distribuční funkce. Jelikož je distribuční funkce neklesající, hustota musí být (skoro všude) nezáporná.

3. Při dokazování závislosti náhodných veličin X, Y nestačí ukázat, že $f_{X,Y}(x,y) \neq f_X(x)\, f_Y(y)$ v nějakém jednom vybraném bodě, ale alespoň na nějaké množině nenulové Lebesgueovy míry. Naopak, při dokazování nezávislosti je třeba ukázat, že $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$ pro skoro všechna $(x,y) \in \mathbb{R}^2$.

4. Při výpočtu marginální hustoty (distribuční) funkce musí tato být definována na celém $\mathbb{R}$.

5. Jelikož je rozptyl definovaný jako střední kvadratická odchylka, tj. $\mathsf{var}(X) = \mathsf{E}(X - \mathsf{E}\,X)^2$, nebude jistě správně, pokud vyjde záporný. I pokud vyjde nulový, tak je to také krajně podezřelé, protože by to znamenalo, že náhodná veličina X je rovná konstantě skoro jistě. Pokud Vám vyjde záporný rozptyl, je dobré přiznat, že někde bude asi chyba. V opačném případě nabude zkoušející nevalné mínění o Vašich znalostech.

6. Obecně jistě **neplatí**: $\mathsf{var}(aX) = a\ \mathsf{var}(X)$.

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

8. Pro výpočet středních hodnot ve tvaru $\mathsf{E}\,h(X)$, kde h je měřitelná funkce, není zapotřebí odvozovat rozdělení náhodné veličiny $h(X)$. V případě diskrétní, resp. spojité náhodné veličiny X lze zpravidla s úspěchem využít vzorců

$$\mathsf{E}\,h(X) = \sum_k h(x_k)\,\mathsf{P}(X = x_k), \qquad \text{resp.} \qquad \mathsf{E}\,h(X) = \int h(x) f_X(x)\,dx.$$

9. Pozor vyjma speciálních případů obecně **neplatí** $\mathsf{E}\,\frac{1}{X} = \frac{1}{\mathsf{E}\,X}$. Naopak, dá se ukázat, že pro g ryze konvexní na borelovské množině D a náhodnou veličinu X, takovou že X není konstantní skoro jistě a $\mathsf{P}(X \in D) = 1$ platí $\mathsf{E}\,g(X) > g(\mathsf{E}\,X)$ (tzv. Jenssenova nerovnost).

10. S výjimkou velmi speciálních degenerovaných případů jistě **neplatí**

$$\mathsf{E} \max_{1 \le i \le n} X_i = \max_{1 \le i \le n} \mathsf{E}\,X_i.$$

Chceme-li tedy vypočítat $\mathsf{E} \max_{1 \le i \le n} X_i$ je třeba zpravidla najít distribuční funkci $\max_{1 \le i \le n} X_i$, derivací dostat hustotu, ...

11. Hustota maxima dvou náhodných veličin není maximem jejich hustot.

12. Při použití Borelovy věty se často nezdůvodní, proč jsou uvedené jevy nezávislé.

13. Při použití CLV či SZVČ nezapomeňte uvést všechny předpoklady - tj. nezávislost, případnou stejnou rozdělenost, podmínky na momenty (střední hodnota, rozptyl, ...).

14. Zápis

$$\lim_{n\to\infty} \left(\overline{X}_n - \mathsf{E}\,\overline{X}_n\right) = 0,$$

není v pořádku, protože není řečeno, o jaký druh konvergence se zde jedná (bodová, skoro jistě, v pravděpodobnosti, ...).

15. Konzistencí se v tomto předmětu rozumíme *silná konzistence*, která vyžaduje konvergenci s.j. Konzistenci tedy **nelze** dokazovat pomocí Věty 4.3 ze skript, která za platnosti

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathsf{var}(X_i) = 0$$

dává *slabý zákon velkých čísel*.

16. V příkladech na maximálně věrohodný odhad je dobré psát si hustoty včetně indikátorů. Pokud indikátor závisí na parametru, který chceme odhadovat, nelze hledat maximum pomocí derivace věrohodnosti.

17. Řešení příkladu na testování hypotéz by mělo obsahovat: formulaci modelu včetně předpokladů, nulovou hypotézu, alternativní hypotézu, obor zamítnutí (kritický obor) nebo interval spolehlivosti, závěr (zda jsme hypotézu zamítli nebo nezamítli a na jaké hladině).

18. V předpokladech dvouvýběrového t-testu chybí shodnost rozptylů.

19. Když při testování nezamítneme nulovou hypotézu, tak z toho nelze vyvodit, že nulová hypotéza platí.

Koneckonců i lékaři říkají, že neexistuje zdravý pacient (tj. nulová hypotéza je, že pacient je zdravý), ale pouze špatně vyšetřený pacient.

20. Ačkoliv spolu intervaly spolehlivosti a testování hypotéz úzce souvisí, nelze zaměňovat pojmy *spolehlivost* (používá se pro intervalové odhady) a *hladina* (používá se při testování).

3. Časté chyby – teoretická část

21. V definici nezávislosti náhodných jevů se zapomene, že příslušná rovnost platí pro každou **konečnou** k-tici jevů.

22. V definici nezávislosti náhodných veličin pomocí distribučních funkcí (hustot) není napsáno, že příslušná rovnost platí pro (skoro) všechna $x_1, x_2, \ldots$

23. Aby se předešlo nedorozumění, je zapotřebí **poctivě vypisovat indexy**.

24. Neplést náhodné jevy a náhodné veličiny. Náhodný jev je (měřitelná) množina a náhodná veličina je (měřitelná) funkce. Nejčastěji tuto chybu vídáme při formulaci Borelovy, resp. Cantelliho věty.

25. V důkazu Borelovy věty se nezávislost využívá na průnik spočetně mnoha jevů. Z definice nezávislosti však pouze víme, jak počítat pravděpodobnost průniku konečně mnoha jevů.

26. Občas lze vidět následující kritéria pro SZVČ pro nestejně rozdělené náhodné veličiny, která jistě nejsou správně.

$$\sum_{i=1}^{\infty} \frac{\mathsf{var}(X_i)}{n^2} < \infty, \qquad \text{nebo} \qquad \sum_{i=1}^{n} \frac{\mathsf{var}(X_i)}{i^2} < \infty$$

27. Předpoklad $\mathsf{E}\,X < \infty$ nezaručuje konečnost $\mathsf{E}\,X$, neboť nevylučuje možnost $\mathsf{E}\,X = -\infty$.

28. Matematicky značně podezřelý je zápis

$$\lim_{n\to\infty} \overline{X}_n = \mathsf{E}\,\overline{X}_n, \quad \text{[P]-s.j.}$$

Pokud $\mathsf{E}\,\overline{X}_n$ se nemění s rostoucím n, tak proč tuto závislost na n zdůrazňovat? Pokud se však $\mathsf{E}\,\overline{X}_n$ s rostoucím n mění, pak je výše uvedený zápis vyloženě špatně.

29. Centrální limitní věta (CLV) pro nestejně rozdělené náhodné veličiny je ve skriptech formulována pro tzv. trojúhelníkové schéma. Jelikož využití této věty je až v regresi (Věta 7.3 ze skript), tak studenti často správně nepochopí to trojúhelníkové schéma náhodných veličin, špatně píšou indexy a celá formulace CLV „jde do háje".

Pokud tedy tomu dvojitému indexy nerozumíte, tak se doporučuji raději naučit jednodušší verzi Ljapunovovy CLV, jak je uvedena např. v materiálech k 9. cvičení.

30. V centrální limitní větě (CLV) pro nestejně rozdělené náhodné veličiny je podmínka na třetí **absolutní** centrální momenty, tj. $\mathsf{E}\,|X_k - \mathsf{E}\,X_k|^3$, nikoliv na obyčejné třetí centrální momenty $\mathsf{E}(X_k - \mathsf{E}\,X_k)^3$.

31. Nenapíše se, že konvergence v centrální limitní větě $\mathsf{P}(Z_n \leq x) \to \Phi(x)$ platí pro všechna $x \in \mathbb{R}$. Také je zde nesmysl psát, že výše uvedená konvergence platí *skoro jistě*.

32. Při formulaci silného zákona velkých čísel či centrální limitní věty se špatně zformulují kritéria.

33. Často se zejména při formulaci silného zákona velkých čísel či centrální limitní věty opomíjí předpoklad **nezávislosti** zúčastněných náhodných veličin. Také je nevyhovující psát, že např. $\overline{X}_n - \mathsf{E}\,\overline{X}_n \to 0$ [P]-s.j., aniž se řekne, že $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ a co jsou to X_i.

34. Při definování úlohy bodového odhadu by mělo být řečeno, co je to ta parametrická funkce g (zobrazení odkud kam). Podobně, co je to za zobrazení ta funkce $\phi(\mathbf{X})$.

35. Občas se plete, že zatímco funkční předpis $\phi(\mathbf{X})$ bodového odhadu nesmí záviset na odhadovaném parametru, tak rozdělení $\phi(\mathbf{X})$ by naopak na odhadovaném parametru záviset mělo. Např. konzistence odhadu požaduje (vulgárně řečeno), aby se rozdělení odhadu s rostoucím rozsahem výběru koncentrovalo ve „stále větší blízkosti" odhadovaného parametru (resp. odhadované parametrické funkce).

36. V definici intervalového odhadu není napsáno, že požadovaná vlastnost musí platit pro všechna θ z příslušného parametrického prostoru. Totéž by mělo platit při definici nestrannosti a konzistence bodového odhadu.

37. V definici t-rozdělení se zapomene, že čitatel i jmenovatel mají být **nezávislé** náhodné veličiny.

38. V definici χ^2-rozdělení se zapomene, že jednotlivé sčítance jsou **nezávislé** náhodné veličiny.

39. Studenti se často naučí Neymanovu-Pearsonovu větu i její důkaz, nicméně neumí ji použít v konkrétních případech a její význam pro testování hypotéz zůstává utajen.

40. Častým zdrojem zmatků je špatné porozumění pravděpodobnosti chyby prvního druhu a druhého druhu.

41. Věnujte pozornost terminologii pro intervaly spolehlivosti a testování hypotéz. **Nelze** například říkat, že:

Pravděpodobnost, že nulová hypotéza platí ...

Intervalový odhad na hladině α (*Správně*: Intervalový odhad o spolehlivosti $1-\alpha$) ...

Parametr μ náleží do intervalu spolehlivosti s pravděpodobností $1-\alpha$ (*Správně:* Intervalový odhad pokryje skutečnou hodnotu parametru μ s pravděpodobností $1-\alpha$).