Pravděpodobnost a statistika

opora pro studium

Informace o předmětu

Základní informace

Název předmětu: **Pravděpodobnost a statistika**
Forma výuky: přednáška + cvičení
Ověření studijních výsledků: zápočet + zkouška

Anotace

Studenti se seznámí s možnými popisy zákonitostí u statisticky stabilních náhodných pokusů, aplikacemi nejdůležitějších pravděpodobnostních modelů i simulací náhodných pokusů a dějů (pravděpodobnostní prostor, náhodné veličiny, důležitá diskrétní a spojitá rozdělení pravděpodobnosti, centrální limitní věta). Dále se seznámí s metodami přenosu informace z výběrového souboru na základní soubor (teorie odhadu, testování hypotéz, regrese a korelace). Studenti získají praktické dovednosti ve vhodném software (Excel, Statistica, R) při práci s daty.

Sylabus

1) Shrnutí základních poznatků z teorie pravděpodobnosti.
2) Náhodná veličina, hustota, pravděpodobnostní funkce, distribuční funkce, základní charakteristiky.
3) Diskrétní rozdělení pravděpodobnosti, vlastnosti, aplikace.
4) Spojitá rozdělení pravděpodobnosti, vlastnosti, aplikace.
5) Centrální limitní věta.
6) Speciální statistická rozdělení – Pearsonovo, Studentovo, Fisherovo.
7) Náhodný výběr, základní statistiky a jejich charakteristiky.
8) Výběry z normálních rozdělení - rozdělení pravděpodobnosti statistik.
9) Bodové odhady – některé základní podmínky kladené na bodové odhady, intervaly spolehlivosti a jejich konstrukce, zpracování výsledků měření, odhad chyb měření, šíření chyb a nejistot.
10) Testy hypotéz o parametrech normálních rozdělení – jeden výběr.
11) Testy hypotéz o parametrech normálních rozdělení – dva výběry.
12) Testy hypotéz o parametrech některých dalších rozdělení. Párový t-test. Testy shody.
13) Kontingenční tabulky.

Zdroje pro studium

Materiály k probírané látce

- BUDÍKOVÁ, Marie a kol. *Statistika a pravděpodobnost*. 1. vyd. Brno: Masarykova univerzita, 2016. ISBN 978-80-210-8206-9. Dostupné na: `https://is.muni.cz/do/rect/el/estud/prif/ps15/statistika/web/index.html`
- HENDL, Jan. Přehled statistických metod: analýza a metaanalýza dat. Páté, rozšířené vydání. Praha: Portál, 2015. ISBN 978-80-262-0981-2.
- HINDLS, Richard, Markéta ARLTOVÁ, Stanislava HRONOVÁ, Ivana MALÁ, Luboš MAREK, Iva PECÁKOVÁ a Hana ŘEZANKOVÁ. Statistika v ekonomii. [Průhonice]: Professional Publishing, 2018. ISBN 978-80-88260-09-7.
- HOLČÍK, Jiří, KOMENDA, Martin (eds.) a kol. Matematická biologie: e-learningová učebnice [online]. 1. vyd. Brno: Masarykova univerzita, 2015. ISBN 978-80-210-8095-9. Dostupné na: `http://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-\matematickou-biologii`
- MAREK, Luboš. Statistika v příkladech. Druhé vydání. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-7431-153-6.
- OTIPKA, Petr a Vladislav ŠMAJSTRLA. *Pravděpodobnost a statistika*. Ostrava: Vysoká škola báňská – technická univerzita Ostava. Dostupné na: `http://homel.vsb.cz/~oti73/cdpast1/`
- PAVLÍK, Tomáš a Ladislav DUŠEK. *Biostatistika*. Brno: AKADEMICKÉ NAKLADATELSTVÍ CERM, s.r.o., 2012. ISBN 978-80-7204-782-6. Dostupné na: `https://www.matematickabiologie.cz/res/file/ucebnice/pavlik-biostatistika-v2.pdf`
- ZVÁROVÁ, Jana. Základy statistiky pro biomedicínské obory. Dotisk. Praha: Karolinum, 2002. ISBN 80-7184-786-0.

Návody k programu R

- DE VRIES, Andrie a Joris MEYS. *R for Dummies*. Hoboken: John Wiley & Sons, Inc., 2015. ISBN 978-1-119-05585-3. Dostupné na: `http://sgpwe.izt.uam.mx/files/users/uami/gma/R_for_dummies.pdf`
- NAVARRO, Danielle. *Learning statistics with R*. Sydney: University of New South Wales, 2018. Dostupné na: `https://learningstatisticswithr.com/lsr-0.6.pdf`

Online kurz programu R (DataCamp)

V rámci výuky předmětu Pravděpodobnost a statistika mají studenti přístup do databáze interaktivních online kurzů k programu R na webu `www.datacamp.com`. Tato databáze obsahuje několik desítek kurzů, přičemž některé z nich jsou vhodným doplňkem ke studiu předmětu. Jde především o kurzy základního ovládání programu R a statistických metod prováděných v programu R.

Studium kurzů je možné jak na webu `www.datacamp.com`, tak v mobilní aplikaci *DataCamp* (Android, iOS). Všechny kurzy jsou v anglickém jazyce.

O přístupu ke kurzům budou studenti informováni vyučujícím na začátku semestru.

V následujícím seznamu jsou uvedeny kurzy, které z velké části pokrývají probíranou látku. Je vhodné procházet kurzy v uvedeném pořadí, neboť je tak zaručena návaznost mezi kurzy. Kurzy je také možné procházet jednotlivě, je ale potřeba pohlídat si předpoklady ke studiu kurzu (jsou uvedeny u každého kurzu).

Základní kurzy R

Tyto kurzy je doporučeno probrat jako první, neboť studenty seznamují se základy ovládání programu R.

- Introduction to R (úvod do ovládání programu R)
- Introduction to Tidyverse (vizualizace a agregování dat)
- Intermediate R (pokročilejší funkce R)
- Introduction to Data Visualisation with ggplot2 (nastavení vzhledu grafu)

Kurzy pravděpodobnosti v R

První z kurzů seznamuje se základními pravděpodobnostními rozděleními a generováním dat z daného rozdělení. Druhý z kurzů rozšiřuje probíranou látku o zajímavé úlohy z pravděpodobnosti (hod kostkou, karetní hry).

- Foundations of Probability in R (pravděpodobnostní rozdělení, podmíněná pravděpodobnost)
- Probability puzzles in R (známé úlohy z pravděpodobnosti řešené v R)

Kurzy statistiky v R

Následující kurzy pokrývají nejdůležitější oblasti statistiky, které jsou vyučovány v rámci tohoto předmětu.

- Introduction to Statistics in R (popisná statistika, generování dat, vybraná pravděpodobnostní rozdělení)

- Sampling in R (náhodný výběr a jeho vlastnosti)
- Hypotesis Testing in R (p-hodnota, chyba 1. a 2. druhu, t-test, ANOVA, párový t-test, chí-kvadrát test, předpoklady testů)

Tyto kurzy navazují na předchozí a rozšiřují jednotlivé partie probírané látky.

- Introduction to Data in R (typy proměnných, vizualizace metrických a kategoriálních dat)
- Exploratory Data Analysis in R (struktura dat, hledání možných souvislostí)
- Introduction to Regression in R (jednoduchá lineární regrese, predikce, posuzování kvality regresního modelu)
- Correlation and Regression in R (bodový graf, korelace, lineární regrese, interpretace)
- Foundations of Inference (úvod do problematiky, testování hypotéz, intervaly spolehlivosti)
- Inference for Numerical Data in R (t-test, ANOVA, testy pro rozdíl dvou parametrů)
- Inference for Categorical Data in R (kontingenční tabulky, testy shody)

1 Shrnutí základních poznatků z teorie pravděpodobnosti

Základní pojmy

- ✓ kombinatorické pravidlo součinu
- ✓ faktoriál, kombinační číslo
- ✓ variace bez opakování, s opakováním
- ✓ permutace bez opakování, s opakováním
- ✓ kombinace bez opakování, s opakováním
- ✓ náhodný pokus, náhodný jev
- ✓ pravděpodobnost jevu
- ✓ nezávislé jevy
- ✓ úplná pravděpodobnost, podmíněná pravděpodobnst, Bayesova věta

Co je nutné umět

- Řešit úlohy na variace, permutace a kombinace, bez opakování i s opakováním,
- spočítat pravděpodobnost daného jevu,
- aplikovat Bayesovu větu při řešení úloh.

Poznámka: Pro studium této kapitoly stačí i středoškolské účebnice (viz Zdroje pro studium).

Zdroje pro studium

- Pravděpodobnost a statistika (VŠB) – Kombinatorika, učební text a příklady
- Pravděpodobnost a statistika (VŠB) – Pravděpodobnost jevů, učební text a příklady
- Statistika v ekonomii (Hindls a kol.), str. 53–62
- DUPAČ, Václav a Emil CALDA. *Matematika pro gymnázia: Kombinatorika, pravděpodobnost a statistika*. Dotisk 4., upr. vyd. Praha: Prometheus, 2004. Učebnice pro střední školy. ISBN 80-7196-147-7.
- HORENSKÝ, Radek, Ivana JANŮ, Martina KVĚTOŇOVÁ, Hana LUKŠOVÁ a Rita VÉMOLOVÁ. *Matematika pro střední školy. 8. díl, Kombinatorika, pravděpodobnost, statistika: učebnice*. Brno: Didaktis, 2015. ISBN 978-80-7358-238-8.

- HORENSKÝ, Radek, Ivana JANŮ, Martina KVĚTOŇOVÁ, Hana LUKŠOVÁ a Rita VÉMOLOVÁ. *Matematika pro střední školy. 8. díl, Kombinatorika, pravděpodobnost, statistika: pracovní sešit.* Brno: Didaktis, 2015. ISBN 978-80-7358-239-5.
- DataCamp kurz Foundations of Probability in R

Typové úlohy

1) Zjednodušte: $\frac{28!+29!}{30!}$.

2) Vypočítejte následující kombinační čísla: $\binom{5}{2}$, $\binom{3}{3}$, $\binom{7}{2}$, $\binom{7}{5}$.

3) U stánku nabízejí čtyři druhy zmrzliny a tři polevy. Kolik různých zmrzlin s polevou lze vytvořit, jestliže nechceme míchat více druhů zmrzliny ani více polev?

4) Na vrchol hory vedou čtyři turistické cesty a lanovka. Určete počet způsobů, kterými je možno se dostat:

   a) na vrchol a zpět;

   b) na vrchol a zpět tak, aby alespoň jednou byla použita lanovka;

   c) na vrchol a zpět tak, aby lanovka byla použita právě jednou.

5) Jméno a příjmení každého obyvatele města s 1 500 obyvateli může začínat jedním ze 32 písmen. Ukažte, že aspoň dva obyvatelé mají stejné iniciály.

6) Výbor sportovního klubu tvoří šest mužů a čtyři ženy. Určete:

   a) kolika způsoby z nich lze vybrat předsedu, místopředsedu, jednatele a hospodáře;

   b) kolika způsoby z nich lze vybrat funkcionáře podle a) tak, aby ve funkci předsedy byl muž a ve funkci místopředsedy žena nebo obráceně;

   c) kolika způsoby z nich lze vybrat funkcionáře podle a) tak, aby právě jedním z nich byla žena.

7) Určete kolika způsoby je možno ze sedmi mužů a čtyř žen vybrat šesti člennou skupinu, v níž jsou

   a) právě dvě ženy;

   b) aspoň dvě ženy.

8) Určete kolika způsoby je možno z dvaceti osob vybrat deset, požadujeme-li, aby mezi vybranými

   a) nebyl pan A;

   b) nebyli zároveň pánové A, B;

   c) byl aspoň jeden z pánů A, B.

9) V určité rodině je pravděpodobnost zdědění konkrétní nemoci u chlapce 0,2 a u děvčete 0,1. Pravděpodobnost narození chlapce je 0,485, pravděpodobnost narození děvčete je 0,515. Jaká je pravděpodobnost, že dítě narozené v uvedené rodině zdědí tuto nemoc?

10) V první přepravce je 20 a v druhé 25 lahví bílého vína. Na lahvích nejsou etikety. V každé z přepravek je 12 lahví tramínu. Nejdříve náhodně vybereme jednu z přepravek a z ní pak vybereme

    a) jednu lahev,

    b) dvě láhve.

    Určete pravděpodobnost, že v každé vybrané lahvi bude tramín.

11) Dvanácti pacientům je podáván lék, který úspěšně léčí jejich onemocnění v 95 % případů. Jaká je pravděpodobnost, že alespoň deset z nich bude vyléčeno?

Řešení úloh:

1) $\frac{1}{29}$

2) 10, 1, 21, 21

3) 12

4) a) 25

    b) 9

    c) 8

5) Možných iniciál je pouze $32 \cdot 32 = 1024$, což je méně než lidí ve městě.

6) a) Celkem $10 \cdot 9 \cdot 8 \cdot 7 = 5040$ způsoby.

    b) Celkem $6 \cdot 4 \cdot 8 \cdot 7 + 4 \cdot 6 \cdot 8 \cdot 7 = 2688$ způsoby.

    c) Celkem $4 \cdot 6 \cdot 5 \cdot 4 + 6 \cdot 4 \cdot 5 \cdot 4 + 6 \cdot 5 \cdot 4 \cdot 4 + 6 \cdot 5 \cdot 4 \cdot 4 = 1920$ způsoby.

7) a) $\binom{4}{2} \cdot \binom{7}{4} = 210$

    b) $\binom{4}{2} \cdot \binom{7}{4} + \binom{4}{3} \cdot \binom{7}{3} + \binom{4}{4} \cdot \binom{7}{2} = 371$

8) a) $\binom{19}{10}$

    b) $\binom{20}{10} - \binom{18}{8}$

    c) $2 \cdot \binom{18}{9} + \binom{18}{8}$

9) $0{,}485 \cdot 0{,}2 + 0{,}015 \cdot 0{,}1 = 0{,}1485$

10) a) $\frac{1}{2} \cdot \frac{12}{20} + \frac{1}{2} \cdot \frac{12}{25} = 0{,}54$

    b) $\frac{1}{2} \cdot \frac{12}{20} \cdot \frac{11}{19} + \frac{1}{2} \cdot \frac{12}{25} \cdot \frac{11}{24} = 0{,}284$

11) $\binom{12}{10} \cdot 0{,}95^{10} \cdot 0{,}05^2 + \binom{12}{11} \cdot 0{,}95^{11} \cdot 0{,}05 + 0{,}95^12 \doteq 0{,}981$

2 Náhodná veličina, hustota, pravděpodobnostní funkce, distribuční funkce, základní charakteristiky

Základní pojmy

- ✓ náhodná veličina
- ✓ pravděpodobnostní funkce
- ✓ hustota pravděpodobnosti
- ✓ distribuční funkce
- ✓ střední hodnota
- ✓ rozptyl
- ✓ modus
- ✓ medián
- ✓ kvartil, percentil, kvantil
- ✓ koeficient šikmosti
- ✓ koeficient špičatosti

Co je nutné umět

- Určit pravděpodobnostní funkci/hustotu pravděpodobnosti a distribuční funkci pro danou náhodnou veličinu (včetně nakreslení grafu),
- vypočítat střední hodnotu, rozptyl, koeficient šikmosti a koeficient špičatosti pro danou náhodnou veličinu,
- pro danou náhodnou veličinu určit modus, medián a kvartily,
- pro zadaná data sestrojit histogram a empirickou distribuční funkci.

Zdroje pro studium

- Pravděpodobnost a statistika (VŠB) – Náhodná veličina, učební text a příklady
- Statistika a pravděpodobnost (MU) – Náhodná veličina
- Portál Matematická biologie – Náhodná veličina, rozdělení pravděpodobnosti a reálná data
- Statistika v ekonomii (Hindls a kol.), str. 63–84
- Náhodná veličina – příklady (VŠB)
- Biostatistika (Pavlík, Dušek), str. 26–33
- DataCamp kurz Foundations of Probability in R

Typové úlohy

1) Házeme třikrát kostkou. Nechť náhodná veličina $\mathbb{X}$ znamená počet padnutí šestky. Určete:

   a) pravděpodobnostní funkci a její graf,

   b) sestrojte graf distribuční funkce.

   Řešení:

   a) $\binom{3}{x} \cdot (\frac{1}{6})^x \cdot (\frac{5}{6})^{3-x}$

   b) $f(x) = \begin{cases} \frac{1}{3} & \leq x < 6 \\ 0 & \text{jinde} \end{cases}$

2) Je dána náhodná veličina $\mathbb{X}$, která nabývá hodnot -1, 0, 2, po řadě s pravděpodobnostmi $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{4}$. Vypočtěte střední hodnotu a rozptyl náhodné veličiny $\mathbb{X}$. Nakreslete grafy pravděpodobnostní funkce a distribuční funkce.

   Řešení: $E(\mathbb{X}) = 0$, $D(\mathbb{X}) = \frac{3}{2}$

3) Hustota náhodné veličiny $\mathbb{X}$ je dána předpisem

$$f_{\mathbb{X}}(x) = \begin{cases} \frac{x}{2} & \text{pro } x \in (0;2) \\ 0 & \text{pro } \mathbb{R} \setminus (0;2) \end{cases}$$

   Nakreslete graf distribuční funkce veličiny $\mathbb{X}$ a vypočtěte střední hodnotu a rozptyl této veličiny.

   Řešení: $E(\mathbb{X}) = \frac{4}{3}$, $D(\mathbb{X}) = \frac{2}{9}$

4) Náhodná veličina $\mathbb{X}$ má hustotu pravděpodobnosti:

$$f_{\mathbb{X}}(x) = \begin{cases} 2x & \text{pro } x \in (0;12) \\ 0 & \text{pro } \mathbb{R} \setminus (0;1) \end{cases}$$

   Určete kvartily této náhodné veličiny.

   Řešení: $x_{0,25} = 0{,}5$, $x_{0,5} = \frac{\sqrt{2}}{2}$, $x_{0,75} = \frac{\sqrt{3}}{2}$

5) Náhodná veličina $\mathbb{X}$ je dána tabulkou:

x_i	1	2	3	4
p_i	0,3	0,1	0,4	?

   Určete střední hodnotu, rozptyl a koeficient šikmosti a špičatosti této náhodné veličiny.

   Řešení: $E(\mathbb{X}) = 2$, $D(\mathbb{X}) = 1$, koeficient šikmosti: $\gamma_1 = -0{,}6$, koeficient špičatosti: $\gamma_2 = -0{,}8$

3 Diskrétní rozdělení pravděpodobnosti, vlastnosti, aplikace

Základní pojmy

- ✓ alternativní rozdělení
- ✓ binomické rozdělení
- ✓ hypergeometrické rozdělení
- ✓ Poissonovo rozdělení
- ✓ rovnoměrné diskrétní rozdělení

Co je nutné umět

- Pro jednotlivá rozdělení: tvar pravděpodobnostní a distribuční funkce, střední hodnotu a rozptyl, použití v praxi,
- aktivně používat pravděpodobnostní kalkulátory, případně pravděpodobnostní tabulky.

Zdroje pro studium

- Pravděpodobnost a statistika (VŠB) – Rozdělení p-sti DNV, učební text a příklady
- Statistika a pravděpodobnost (MU) – Diskrétní rozdělení
- Biostatistika (Pavlík, Dušek), str. 34–44
- Statistika v ekonomii (Hindls a kol.), str. 85–90
- DataCamp kurz Foundations of Probability in R

Typové úlohy

1) Univerzitní student Roman někdy zaspí a nestihne přednášku, která začíná v 9 hodin. Pravděpodobnost, že zaspí, je 0,4. Určete pravděpodobnost, že v semestru přijde na přednášku Roman pozdě v důsledku zaspání v polovině nebo více případů (předpokládejme, že semestr má 12 týdnů).

   Řešení: $p = 0,3348$, jde o binomické rozdělení

2) Hráč košíkové Petr promění z každých 10 trestných hodů průměrně 7. Jaká je pravděpodobnost, že v zápase z 20 trestných hodů promění nejméně 15?

   Řešení: $p = 0,41648$, jde o binomické rozdělení

3) Student píše test, který obsahuje 15 otázek, ke každé otázce existují čtyři možné odpovědi, z nichž právě jedna je správná. Jaká je pravděpodobnost, že student odpoví správně na alespoň pět otázek (a test úspěšně splní), pokud problematiku vůbec neovládá a odpovědi volí náhodně?

   Řešení: $p = 0,3135$, jde o binomické rozdělení

4) Průměr počtu těžkých zranění v jednom ročníku hokejové ligy je 5. Zjistěte, jaká je pravděpodobnost, že počet těžkých zranění bude více než 4.

   Řešení: $p = 0,5595$, jde o Poissonovo rozdělení

5) Mezi stovkou výrobků je 20 zmetků. Náhodně vybereme deset výrobků a sledujeme počet zmetků mezi vybranými. Jaká je pravděpobnost, že mezi vybranými výrobky budou 3 zmetky?

   Řešení: $p = 0,2092$, jde o hypergeometrické rozdělení

4 Spojitá rozdělení pravděpodobnosti, vlastnosti, aplikace

Základní pojmy

- ✓ rovnoměrné spojité rozdělení
- ✓ normální rozdělení
- ✓ logaritmicko-normální rozdělení
- ✓ exponenciální rozdělení

Co je nutné umět

- Pro jednotlivá rozdělení: tvar hustoty pravděpodobnosti a distribuční funkce, střední hodnotu a rozptyl, použití v praxi,
- aktivně používat pravděpodobnostní kalkulátory, případně pravděpodobnostní tabulky.

Zdroje pro studium

- Pravděpodobnost a statistika (VŠB) – Rozdělení p-sti SNV, učební text a příklady
- Statistika a pravděpodobnost (MU) – Spojité rozdělení
- Portál Matematická biologie – Normální rozdělení
- Portál Matematická biologie – Standardizované normální rozdělení
- Biostatistika (Pavlík, Dušek), str. 34–44
- Statistika v ekonomii (Hindls a kol.), str. 91–98
- DataCamp kurz Introduction to Statistics in R

Typové úlohy

1) Pošta chodí pravidelně mezi 10:00 a 12:00 rovnoměrně. Jaká je pravděpodobnost, že obdržíme poštu mezi 11:30 a 12:30?

   Řešení: $p = 0{,}25$, jde o rovnoměrné spojité rozdělení

2) IQ je standardní škála, která má v populaci normální rozdělení No(100, 225). Jaká je pravděpodobnost, že hodnota IQ náhodně vybraného jedince bude:

   a) nižší než 95?

   b) v rozsahu 110–120?

   c) vyšší než 130?

Řešení: a) 0,4207, b) 0,1327, c) 0,1150

3) Výška má v populaci normální rozdělení No(175, 81). Jaká je pravděpodobnost, že výška náhodně vybraného jedince bude:

   a) menší než 150 cm?

   b) v rozsahu 180–200 cm?

   c) vyšší než 225 cm?

   d) menší než Vaše výška?

   e) Jakou výšku musí mít osoba, abychom mohli říci, že 95 % populace má výšku nižší než tato osoba?

   f) Jakou výšku musí mít osoba, abychom mohli říci, že 75 % populace je vyšší než tato osoba?

   Řešení: a) 0,0027, b) 0,2865, c) 0,00000001, d) ???, e) 189,8, f) 168,9

4) Stanovte hodnotu distribuční funkce $F(x)$ v bodě $x = 5$ rozdělení LN(2, 2,25).

   Řešení: $F(x) = \Phi\left(\frac{\ln 5 - 2}{2{,}25}\right) = 0{,}431$

5) Na trase mezi Ústím nad Labem a Velkým Březnem délky 10,5 km napočítali cestáři 86 děr v silnici. Jaká je pravděpodobnost, že narazíme na díru v silnici při ujetí úseku délky 100 m na této trase?

   Řešení: $p = 0,5591$, jde o exponenciální rozdělení

5 Centrální limitní věta

Základní pojmy

- ✓ centrální limitní věta
- ✓ slabý zákon velkých čísel
- ✓ Čebyševova nerovnost
- ✓ Bernoulliova věta

Co je nutné umět

- Formulovat a používat uvedené limitní věty.

Zdroje pro studium

- Úvod do teorie odhadu (Marie Litschmannová), str. 5–16
- Limitní věty (Radim Briš)
- Statistika a pravděpodobnost (MU) – Centrální limitní věta
- Statistika a pravděpodobnost (MU) – Zákon velkých čísel
- Statistika v ekonomii (Hindls a kol.), str. 104–108
- Statistika a pravděpodobnost (MU) – Řešené úlohy

Typové úlohy

1) Chyba měření má rovnoměrné rozdělení $\mathrm{Ro}(-0{,}5; 0{,}5)$. Užitím centrální limitní věty určete, jaká je pravděpodobnost, že průměrná chyba z 300 měření bude menší než $-0{,}02$.

   Řešení: 11,5 %

2) Pomocí Čebyševovy nerovnosti odhadněte s 90% jistotou počet padlých líců při hodu 100 mincemi.

   Řešení: Počet padlých líců bude s 90% jistotou v intervalu (35; 65).

3) Pravděpodobnost, že zaměstnanec dodržuje bezpečnostní pokyny na na pracovišti, je 0,9. Jestliže náhodně vybereme 40 zaměstnanců, určete pravděpodobnost, že relativní četnost zaměstnanců dodržujících pokyny je z intervalu (0,85; 0,95).

   Řešení: $p = 0,91$, použijeme Čebyševovu nerovnost

6 Speciální statistická rozdělení – Pearsonovo, Studentovo, Fisherovo

Základní pojmy

- ✓ Pearsonovo rozdělení
- ✓ Studentovo rozdělení
- ✓ Fisherovo-Snedecorovo rozdělení

Co je nutné umět

- Pro jednotlivá rozdělení: tvar hustoty pravděpodobnosti a distribuční funkce, střední hodnotu a rozptyl,
- aktivně používat pravděpodobnostní kalkulátory, případně pravděpodobnostní tabulky.

Zdroje pro studium

- Úvod do teorie odhadu (Marie Litschmannová), str. 18–38
- Spojitá rozdělení pravděpodobnosti (Radim Briš), str. 31–38
- Statistika a pravděpodobnost (MU) – Spojité rozdělení (uvedený typy)
- Portál Matematická biologie – Další rozdělení pravděpodobnosti (uvedené typy)
- Statistika v ekonomii (Hindls a kol.), str. 99–103

Typové úlohy

1) Nechť náhodná veličina $\mathbb{X} \sim \chi^2(10)$. Určete kvantil $\chi^2_{0,05}(10)$.
   Řešení: $\chi^2_{0,05}(10) = 3{,}94$
2) Nechť náhodná veličina $\mathbb{X} \sim t(7)$. Určete kvantil $t_{0,025}(7)$.
   Řešení: $t_{0,025}(7) = -2,36$
3) Nechť náhodná veličina $\mathbb{X} \sim F(2, 9)$. Určete kvantil $F_{0,975}(2, 9)$.
   Řešení: $F_{0,975}(2, 9) = 5{,}71$

7 Náhodný výběr, základní statistiky a jejich charakteristiky

Základní pojmy

- ✓ náhodný výběr
- ✓ absolutní četnost, relativní četnost
- ✓ Sturgesovo pravidlo
- ✓ průměr, modus, medián
- ✓ rozptyl, směrodatná odchylku, rozpětí, kvartilové rozpětí, variační koeficient
- ✓ minimum, maximum, kvartily (kvantily, percentily)
- ✓ vylučování odlehlých hodnot, metoda vnitřních hradeb
- ✓ histogram
- ✓ empirická distribuční funkce
- ✓ krabicový graf

Co je nutné umět

- Určit relativní a absolutní četnosti pro zadaná data,
- určit základní statistické charakteristiky pro zadaná data,
- pro zadaná data zkonstruovat histogram, empirickou distribuční funkci a krabicový graf pro zadaná data,
- vyloučit odlehlé hodnoty ze zadaných dat (metodou vnitřních hradeb),
- generovat data z daného pravděpodobnostního rozdělení.

Zdroje pro studium

- Statistika a pravěpodobnost (MU) – Základní a výběrový soubor
- Statistika a pravěpodobnost (MU) – Číselné charakteristiky znaků
- Pravděpodobnost a statistika (VŠB) – Statistický soubor I, učební text a příklady
- Statistika v ekonomii (Hindls a kol.), str. 21–52
- Portál Matematická biologie – Data a jejich vizualizace
- Portál Matematická biologie – Identifikace ohlehlých hodnot
- DataCamp kurz Sampling in R
- DataCamp kurz Introduction to Statistics in R

Typové úlohy

1) Následující tabulka udává pravděpodobnost výskytu rostliny na daném stanovišti. Modelujte strukturu rostlin ve sběru o velikosti 10 ks, 100 ks, 1000 ks. Pro každý sběr vytvořte histogram a tabulku relativních četností. Porovejte se skutečnými pravděpodobnostmi výskytu.

Rostlina	Pravděpodobnost výskytu
1	0,4
2	0,1
3	0,2
4	0,2
5	0,1

2) Modelujte 10, 1000 a 1000 hodů hrací kostkou. Pro každý soubor vytvořte histogram a tabulku relativních četností. Porovejte se skutečnými pravděpodobnostmi.

3) Vygenerujte soubor z No(175, 81) o rozsahu 5000.

   a) Sestrojte histogram pro vygenerované výšky osob.

   b) Sestrojte empirickou distribuční funkci pro vygenerované výšky osob. Řešte úlohu 3 v kapitole 4 pomocí vámi vytvořené empirické distribuční funkce a výsledky porovnejte.

4) Určete základní statistické charakteristiky pro veličinu BMI ze souboru DATADETI. Tyto charakteristiky vypočtete pro celý soubor, pouze pro dívky a pouze pro chlapce. Výsledky porovnejte a interpretujte. Pro dívky a pro chlapce sestrojte krabicové grafy, porovnejte je mezi sebou a interpretujte rozdíly.

5) Metodou vnitřních hradeb určete odlehlé hodnoty u veličiny SKOKD ze souboru DATADETI. Data znázorněte pomocí krabicového grafu.

   Řešení: Odlehlými hodnotami jsou 239, 248 a 250.

6) Určete základní statistické charakteristiky pro veličinu VYSKA ze souboru DATADETI. Hodnoty této veličiny rozdělte do 5 intervalů a určete absolutní a relativní četnosti, kumulované absolutní a kumulované relativní četnosti. Sestrojte histogram.

7) Proveďte totéž, co v bodě 6, počet intervalů ale určete podle Sturgesova pravidla.

   Řešení: Dle Sturgesova pravidla pracujeme s 9 intervaly.

8 Výběry z normálních rozdělení – rozdělení pravděpodobnosti statistik

Základní pojmy

- ✓ rozdělení výběrového průměru $\overline{x}$ pro veličinu z normálního rozdělení
- ✓ pravděpodobnostní rozdělení další speciálních náhodných veličin (viz zdroje pro studium)

Co je nutné umět

- Jaké pravděpodobnostní rozdělení má výběrový průměr v případě jednoho výběru,
- jaké pravděpodobnostní rozdělení mají další speciální náhodné veličiny (rozdíl průměrů v případě dvou výběru, podíl rozptylů atd.)

Zdroje pro studium

- Portál Matematická biologie – vlastnosti výběrového průměru
- Úvod do teorie odhadu (Marie Litschmannová), str. 1–38
- Statistika v ekonomii (Hindls a kol.), str. 115–120
- DataCamp kurz Foundations of Inference

Typové úlohy

Vygenerujte základní soubor (16 veličin z No(30, 4) o rozsahu 1000).

1) Pro každý řádek (výběrový soubor) určete průměr. Množina všech průměrů bude tvořit nový statistický soubor. Spočítejte pro něj základní charakteristiky a výběrový průměr a rozptyl porovnejte s parametry základního souboru. Sestrojte histogram a porovnejte ho s histogramem dat ze základního souboru.
2) Pro každý řádek (výběrový soubor) určete rozptyl a proveďte totéž jako v předchozí úloze.

9 Bodové odhady – některé základní podmínky kladené na bodové odhady, intervaly spolehlivosti a jejich konstrukce, zpracování výsledků měření, odhad chyb měření, šíření chyb a nejistot

Základní pojmy

- ✓ bodový odhad, intervalový odhad
- ✓ vychýlenost, konzistence a vydatnost bodového odhadu
- ✓ interval spolehlivosti oboustranný, jednostranný
- ✓ interval spolehlivosti pro střední hodnotu μ a rozptyl σ^2 normálního rozdělení
- ✓ interval spolehlivosti pro podíl v populaci (relativní četnost)

Co je nutné umět

- Určit jednostranné i oboustranné intervaly spolehlivosti pro střední hodnotu a rozptyl normálního rozdělení,
- určit jednostranné i oboustranné intervaly spolehlivosti pro relativní četnost.

Zdroje pro studium

- Bodové a intervalové odhady (Radim Briš)
- Pravděpodobnost a statistika (VŠB) – Induktivní statistika, učební text a příklady (bodové a intervalové odhady)
- Základy statistiky pro biomedicínské obory, str. 86–92 (intervalové odhady)
- Portál Matematická biologie – Intervaly spolehlivosti
- Biostatistika (Pavlík, Dušek), str. 45–62
- Statistika v ekonomii (Hindls a kol.), str. 120–135
- Populárně naučný text o intervalových odhadech
- DataCamp kurz Foundations of Inference

Typové úlohy

1) Vygenerujte základní soubor (25 veličin z normálního rozdělení s $\mu = 175$, $\sigma^2 = 100$, rozsah souboru 1000). Pro každý řádek (výběrový soubor) určete průměr, rozptyl a směrodatnou odchylku. Pro každý řádek určete interval spolehlivosti pro průměr a rozptyl a zjistěte, v kolika procentech případů pokrývají jednotlivé intervaly spolehlivosti hodnoty $\mu = 175$, $\sigma^2 = 100$. Experimentujte s různými intervaly spolehlivosti (90%, 95%, 99%).

2) Byla měřena délka trvání určitého procesu. Z 12 měření byla zjištěna střední doba trvání procesu 44 s a směrodatná odchylka 4 s. Sestrojte 90% a 95% interval spolehlivosti pro očekávanou délku procesu za předpokladu normálního rozdělení.

   Řešení: 90% interval spolehlivosti je (41,83; 46,17), 95% interval spolehlivosti je (41,35; 46,65).

3) Při výlovu rybníka bylo náhodně vybráno a zváženo 15 kaprů. Naměřené hmotnosti v gramech jsou u jednotlivých kaprů následující:

   $$3100, 3000, 2500, 2500, 4200, 2100, 3250, 2500,$$
   $$4800, 2300, 4100, 3600, 4000, 3000, 3600$$

   a) S 99% spolehlivostí odhadněte průměrnou hmotnost kapra.

   b) Jakou minimální garantovanou nosnost musí mít taška, aby unesla průměrného kapra s 95% spolehlivostí?

   Řešení: a) (2624; 3849), b) Taška musí mít nosnost alespoň 3678 g.

4) Předpokládejme, že mezi 200 dotazovanými (náhodně vybranými z dané populace) je 26 leváků. Sestrojte 95% interval spolehlivosti pro podíl leváků v celé dané populaci.

   Řešení: Využijeme interval spolehlivosti pro relativní četnost, kde výběrová relativní četnost je $p = 0{,}13$. Interval spolehlivosti je (0,08; 0,18).

10 Testy hypotéz o parametrech normálních rozdělení – jeden výběr

Základní pojmy

- ✓ nulová a alternativní hypotéza
- ✓ hladina významnosti
- ✓ testové kritérium a kritický obor
- ✓ p-hodnota testu (p-value)
- ✓ chyba 1. a 2. druhu
- ✓ jednovýběrový test o střední hodnotě normálního rozdělení při neznámém rozptylu (z-test)
- ✓ jednovýběrový test o střední hodnotě normálního rozdělení při známém rozptylu (t-test)
- ✓ jednovýběrový test o rozptylu normálního rozdělení

Co je nutné umět

- Aplikovat oba jednovýběrové testy,
- určit p-hodnotu pro daný test,
- testovat pomocí intervalů spolehlivosti.

Zdroje pro studium

- Testování hypotéz (Radim Briš), str. 1–19
- Portál Matematická biologie – Úvod do testování hypotéz
- Portál Matematická biologie – Testy o parametrech jednoho rozdělení (pouze pro normální rozdělení)
- Biostatistika (Pavlík, Dušek), str. 63–74
- Statistika v ekonomii (Hindls a kol.), str. 136–146
- Populárně naučný text o testování hypotéz
- Populárně naučný text o t-testu
- DataCamp kurz Hypotesis Testing in R
- DataCamp kurz Foundations of Inference

Typové úlohy

1) Testujte hypotézu $\mu = 25$ a $\sigma^2 = 3$ na základě dat z tabulky. Pracujte s hladinou významnosti $\alpha = 0{,}05$.

22,4	20,4	23,5	25,6	25,4	26,5
18,6	22,5	25,2	20,8	21,6	19,6
19,3	21,0	21,5	18,8	21,9	22,2
23,3	22,3	22,3	22,3	25,7	22,8
22,6	22,0	26,9	24,7	27,8	21,7

Řešení: Hypotézu $\mu = 25$ zamítáme (interval spolehlivosti (21,51; 24,04)), hypotézu $\sigma^2 = 3$ zamítáme (interval spolehlivosti (3,90; 11,10)).

2) Výrobní proces produkuje miliony žárovek se střední životností 14000 hodin a směrodatnou odchylkou 2000 hodin. Novou technologií byl vyroben vzorek 25 žárovek s průměrnou životností 14740 hodin, přičemž předpkládáme, že, že směrodatná odchylka se nezměnila. Za předpokladu, že životnost žárovek má normální rozdělení, testujte na hladině významnosti i $\alpha = 0{,}05$, zda po zavedení nové technologie došlo ke zlepšení životnosti žárovek.

Řešení: Použijeme z-test, zavedení nové technologie se životnost žárovek zvýšila, $p = 0{,}032$.

3) Spotřeba automobilu byla testována 11 řidiči s následujícími výsledky. Lze výrobcem udávanou spotřebu 8,8 l/100 km považovat za pravdivou? Pracujte s hladinou významnosti $\alpha = 0{,}05$.

8,8	8,9	9,0	8,7	9,3	9,0
8,7	8,8	9,4	8,6	8,9	

Řešení: Použijeme jednovýběrový t-test, výrobce říká pravdu, $p = 0{,}145$, interval spolehlivosti: $(8{,}78; 9{,}05)$.

4) Obsah naftolu AS byl v dodávce stanoven spektrofotometricky. Podle normy je vyhovující vzorek takový, který obsahuje minimálně 94 % stanovované látky. Pomocí intervalu spolehlivosti ověřte, zda tato dodávka vyhovuje normě. Ověření proveďte na hladině významnosti $\alpha = 0{,}05$.

Obsah naftolu AS [%]: 94,1, 93,4, 94,1, 94,9, 92,6, 94,3, 93,9, 93,3, 94,0, 94,4, 93,6, 93,3, 94,6, 95,0, 94,7, 93,5.

Řešení: Dodávka vyhovuje normě, interval spolehlivosti je (55,54; 57,28).

11 Testy hypotéz o parametrech normálních rozdělení – dva výběry

Základní pojmy

- ✓ F-test pro rozptyl
- ✓ dvouvýběrový t-test s rovností rozptylů
- ✓ dvouvýběrový t-test s nerovností rozptylů (Welchova korekce)

Zdroje pro studium

- Portál Matematická biologie – Testy o parametrech dvou rozdělení
- Testování hypotéz (Radim Briš), str. 25–28 (dvouvýběrové testy o střední hodnotě a rozptylu)
- Biostatistika (Pavlík, Dušek), str. 63–74
- Statistika v ekonomii (Hindls a kol.), str. 147–155
- Přehled statistických metod (Hendl), str. 217–229
- Základy statistiky pro biomedicínské obory, str. 104–120 (testy pro dva výběry, včetně párového t-testu)
- DataCamp kurz Inference for Numerical Data in R

Co je nutné umět

- Aplikovat všechny testy,
- určit p-hodnotu pro daný test.

Typové úlohy

1) Byl měřen obsah vápníku v krevním séru skupiny zdravých lidí a skupiny nemocných. Naměřené hodnoty jsou v tabulce. Porovnejte na hladině významnosti $\alpha = 0{,}05$ obsahy vápníku obou skupin, tj. určete, zda se obě skupiny od sebe statisticky významně liší.

   Zdraví lidé (mmol/l):

2,15	2,13	2,27	2,52	2,11	2,26	2,34	2,68	2,24

   Nemocní lidé (mmol/l):

2,09	1,8	1,97	2,35	2,08	1,9	2,06	2,3	2,36

   Řešení: Používáme t-test s rovností rozptylů; obsah vápníku je u zdravých lidí vyšší než u nemocných, $p = 0,045$.

2) Ve dvou porostech byly ve výšce 1 m měřeny tloušťky kmene v cm. Na hladině významnosti $\alpha = 0{,}05$ posuďte, zda jsou oba porosty stejně tloušťkově vyspělé.

Tloušťky kmene v porostu I (cm):

33,5	28,6	36,2	41,4	41,1	43,7	24,1	33,8
40,5	29,6	31,5	26,5	25,8	30,1	31,1	24,4
32,2	33	35,7	33,2	33,4	33,1	41,7	34,6
34,1							

Tloušťky kmene v porostu II (cm):

29,5	30,4	21,6	24,4	28,5	26,8	25,5	22,6
26,1	19,2	24,6	28,7	20,2	21	27,5	25,8
23,3	26,5	31,2	28,3				

Řešení: Použijeme t-test s nerovností rozptylů (F-test je hraniční), Porost I je vyspělejší než Porost II, $p = 0{,}000000364$.

3) V biometrické studii o vlcích Severní Ameriky byla položena otázka, zda je významný rozdíl mezi celkovou délkou lebky u druhů C. l. nubilus a C. l. ligoni. Máme rozhodnout na základě následujících měření (na hladině významnosti $\alpha = 0{,}05$):

Délky lebky u druhu C. l. nubilus (mm):

261,5	254,4	264,5	258,2	254,4	258,4	253,5	253,2	243,7	244,2
253	250,9	259,1	255,6	256,4	254,8	264,4	250,7	249,3	

Délky lebky u druhu C. l. ligoni (mm):

251,5	254	262,1	270,8	262,5	262	261,5	259,5	255,8

Řešení: Použije t-test srovností rozptylů, délka lebky u druhu C. l. nubilus je menší než u druhu C. l. ligoni, $p = 0,016$.

12 Testy hypotéz o parametrech některých dalších rozdělení, párový t-test, testy shody

Základní pojmy

- ✓ test pro podíl u jednoho výběru
- ✓ párový t-test
- ✓ test shody

Co je nutné umět

- Aplikovat všechny testy,
- určit p-hodnotu pro daný test.

Zdroje pro studium

- Testování hypotéz (Radim Briš), str. 19–21 (test pro podíl u jednoho výběru)
- Portál Matematická biologie – Test pro podíl u jednoho výběru
- Statistika v ekonomii (Hindls a kol.), str. 161–168
- Portál Matematická biologie – Test o rozdílu párových pozorování
- Biostatistika (Pavlík, Dušek), str. 75–88
- Přehled statistických metod (Hendl), str. 314–315 (test dobré shody)
- Základy statistiky pro biomedicínské obory, str. 104–120 (testy pro dva výběry, včetně párového t-testu)
- DataCamp kurz Inference for Numerical Data in R
- DataCamp kurz Inference for Categorical Data in R

Typové úlohy

1) Testujte hypotézu „Pravděpodobnost narození dívky a chlapce je stejná.“ na základě dat z roku 2010 (narozeno 13 929 dívek a 14 718 chlapců).

   Řešení: Hypotézu zamítáme, interval spolehlivosti pro podíl dívek je (0,48; 0,49).

2) Na skupině dobrovolníků byl testován prostředek na snížení hmotnosti. Hmotnosti 12 testovaných lidí před a po dietní kúře jsou v tabulce. Rozhodněte na hladině významnosti $\alpha = 0{,}05$ zda je prostředek účinný.

   hmotnost před dietou (kg):

85	75	90	65	150	80	110	56	88	73	67	134

hmotnost po dietě (kg):

76	75	81	64	155	72	99	45	89	66	56	110

Řešení: Používáme párový t-test; prostředek je účinný, $p = 0{,}004$.

3) Byla změřena výška 10 stromů vždy dvěma výškoměry. Na hladině významnosti $\alpha = 0{,}05$ ověřte předpoklad, že oba přístroje měří stejně, tj. rozdíly v naměřených hodnotách jsou náhodné.

Výška při 1. měření (m):

21,8	18,3	25,6	20,1	19,8	16,1	12,7	18,2	12,3	18,3

Výška při 2. měření (m):

22,4	18,7	25,3	20,3	19,6	16,2	12,5	18,4	12,5	19

Řešení: Používáme párový t-test; přístroje měří stejně, $p = 0{,}145$.

4) Na skupině osob byl testován prostředek na snížení tlaku. Tlak deseti testovaných lidí před a po podání léku jsou v tabulce. Působí lék skutečně snížení tlaku? Ověření proveďte na hladině významnosti $\alpha = 0{,}05$.

Tlak před podáním léku (mmHg):

115	125	135	150	120	105	120	108	99	95

Tlak po podání léku (mmHg):

110	121	125	152	105	99	110	102	90	90

Řešení: Používáme párový t-test; prostředek je účinný, $p = 0{,}001$.

13 Kontingenční tabulky

Základní pojmy

- ✓ kontingenční tabulka
- ✓ hypotéza homogenity a hypotéza nezávislosti
- ✓ test nezávislosti na kontingenční tabulce

Zdroje pro studium

- Základy statistiky pro biomedicínské obory, str. 130–139 (test nezávislosti)
- Portál Matematická biologie – Analýza kontingenčních tabulek
- Biostatistika (Pavlík, Dušek), str. 97–116
- Statistika v ekonomii (Hindls a kol.), str. 169–177
- Přehled statistických metod (Hendl), str. 315–330
- DataCamp kurz Hypotesis Testing in R
- DataCamp kurz Inference for Categorical Data in R

Typové úlohy

1) Bylo zkoušeno použití nového druhu léku. Na základě následující tabulky rozhodněte, zda je průběh nemoci ovlivněn aplikací tohoto léku. Testujte na hladině významnosti $\alpha = 0{,}05$.

		Průběh nemoci	
		lehký	těžký
Lék byl aplikován	ne	14	34
	ano	29	25

Řešení: Používáme chí kvadrát-test na kontingenční tabulce; průběh nemoci je pozitivně ovlivněn použitím léku, testové kritérium je rovno 6,27, kritickým oborem je interval $(3{,}84; \infty)$.

2) V populaci byly zjištěny počty praváků a leváků. Na základě tabulky rozhodněte, zda je praváctví/leváctví závislé na pohlaví. Testujte na hladině významnosti $\alpha = 0{,}05$.

	Praváci	Leváci	Celkem
Muži	43	9	52
Ženy	44	4	48
Celkem	87	13	100

Řešení: Používáme chí kvadrát-test na kontingenční tabulce; leváctví/praváctví nezávisí na pohlaví, testové kritérium je rovno 1,77, kritickým oborem je interval $(3{,}84; \infty)$.