Matematika III – 10. týden
Matematická statistika

Jan Slovák

Masarykova univerzita
Fakulta informatiky

18. – 22.11. 2013

Obsah přednášky

Plán přednášky

Kde je dobré číst?

- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.
- J. Slovák, M. Panák, M. Bulant, Matematika drsně a svižně, Muni Press, Brno 2013, v+773 s., elektronická edice `www.math.muni.cz/Matematika_drsne_svizne`
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, Teorie pravděpodobnosti a matematická statistika (sbírka příkladů), Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, Základní statistické metody, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.
- Riley, K.F., Hobson, M.P., Bence, S.J. Mathematical Methods for Physics and Engineering, second edition, Cambridge University Press, Cambridge 2004, ISBN 0 521 89067 5, xxiii + 1232 pp.

Plán přednášky

Statistiky statistikové zkoumají zpravidla u nějakého výběru z daného základního souboru a snaží se postihnout, do jaké míry jsou zjištěné výsledky relevantní pro celou populaci, případně se ze zjištěných dat pokouší zjistit nebo upřesnit vhodný teoretický model pro chování celého souboru (a z něj pak třeba odhadovat pravděpodobnost nějakého budoucího jevu).

Statistiky statistikové zkoumají zpravidla u nějakého výběru z daného základního souboru a snaží se postihnout, do jaké míry jsou zjištěné výsledky relevantní pro celou populaci, případně se ze zjištěných dat pokouší zjistit nebo upřesnit vhodný teoretický model pro chování celého souboru (a z něj pak třeba odhadovat pravděpodobnost nějakého budoucího jevu).

Dva základní přístupy:

- **frekvenční statistika** (nebo také klasická statistika)
- **bayesovská statistika**.

frekvenční statistika

vychází z matematické abstrakce, že skutečné pravděpodobnosti jsou dány četnostmi výskytů jevů v tak velkých vzorcích dat, že je můžeme dobře aproximovat nekonečnými modely a využít pro odhady spolehlivosti centrální limitní věty.
Statistik zde na pravděpodobnost pohlíží jako na idealizaci relativní četnosti případů, v nichž se vyskytne určitý výsledek při opakovaných pokusech.

frekvenční statistika

vychází z matematické abstrakce, že skutečné pravděpodobnosti jsou dány četnostmi výskytů jevů v tak velkých vzorcích dat, že je můžeme dobře aproximovat nekonečnými modely a využít pro odhady spolehlivosti centrální limitní věty.
Statistik zde na pravděpodobnost pohlíží jako na idealizaci relativní četnosti případů, v nichž se vyskytne určitý výsledek při opakovaných pokusech.
Tato zdánlivá výhoda/rigoróznost se může ale rychle stát nevýhodou, jakmile se začneme zabývat spolehlivostí samotných dat a vhodností zvoleného experimentu. Stejně tak je obtížné frekvenční statistiku dobře použít pro odhad pravděpodobnosti výskytu jednorázového děje.

Bayesovská statistika

Tento přístup můžeme brát jako příklad matematizace „selského rozumu“. Vstupujeme do procesu s jistým původním přesvědčením, které jsme připraveni postupně pozměňovat ve světle nových dat.

Bayesovská statistika

Tento přístup můžeme brát jako příklad matematizace „selského rozumu“. Vstupujeme do procesu s jistým původním přesvědčením, které jsme připraveni postupně pozměňovat ve světle nových dat. Je zajímavé, že historicky byl zjevně první bayesovský přístup (např. Laplace a další již v 18. století), který byl prakticky zcela vystřídán frekvenční statistikou ve 20. století. V posledních desetiletích se však ale bayesovská statistika vrátila, společně s dalšími novými přístupy, do popředí zájmu.

Plán přednášky

Máme k dispozici (velký) základní statistický soubor s N jednotkami, který nazýváme **populace**, a zároveň nějaký číselný znak pro každou z jednotek, tj. soubor hodnot $(x_1, \ldots, x_N)$. Z něj ovšem máme k dispozici pouze **výběrový soubor** s hodnotami $(X_1, \ldots, X_n)$.

Máme k dispozici (velký) základní statistický soubor s N jednotkami, který nazýváme **populace**, a zároveň nějaký číselný znak pro každou z jednotek, tj. soubor hodnot $(x_1, \ldots, x_N)$. Z něj ovšem máme k dispozici pouze **výběrový soubor** s hodnotami $(X_1, \ldots, X_n)$.
Abychom se vyhnuli diskusi skutečné velikosti základního statistického souboru s N jednotkami, budeme předpokládat, že vybíráme položky výběrového souboru jednu po druhé a každou vybranou jednotku poté do populace vracíme. Zároveň předpokládáme, že každá položka má stejnou pravděpodbnost výběru $1/N$. Hovoříme pak o **náhodném výběru**.

Pracujeme tedy s vektorem $(X_1, \ldots, X_n)$ nezávislých náhodných veličin a všechny tyto veličiny mají stejné rozdělení pravděpodobnosti. Zejména tedy budou sdílet distribuční funkci $F_X(x)$ a momenty

$$\mathrm{E}\, X_i = \mu, \quad \mathrm{var}\, X_i = \sigma^2.$$

Dalším naším krokem musí být odvození charakteristik výběrového průměru $\bar{X}$ a **výběrového rozptylu**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

přičemž následující věta dává hned zdůvodnění, proč volíme koeficient $\frac{1}{n-1}$ místo $\frac{1}{n}$.

Theorem

Pro výběrový průměr $\bar{X}$ spočítaný z náhodného výběru rozsahu n z rozdělení s konečnou střední hodnotou μ a konečným rozptylem σ^2 platí

$$\mathrm{E}\,\bar{X} = \mu, \quad \mathrm{var}\,\bar{X} = \frac{1}{n}\sigma^2.$$

Pro výběrový rozptyl S^2 platí

$$\mathrm{E}\,S^2 = \sigma^2.$$

Theorem

Pro výběrový průměr $\bar{X}$ spočítaný z náhodného výběru rozsahu n z rozdělení s konečnou střední hodnotou μ a konečným rozptylem σ^2 platí

$$\mathrm{E}\,\bar{X} = \mu, \quad \mathrm{var}\,\bar{X} = \frac{1}{n}\sigma^2.$$

Pro výběrový rozptyl S^2 platí

$$\mathrm{E}\,S^2 = \sigma^2.$$

Naším úkolem je odhadovat charakteristiky, jako jsou průměr μ hodnot znaku $\bar{x}$ nebo jejich rozptyl σ^2 pro celou populaci pomocí obdobných charakteristik pro náš daleko menší výběr, které budeme značit pomocí velkých písmen, např. $\bar{X}$, S^2.

Zde vstupuje do hry pravděpodobnost – budeme chtít znát pravděpodobnost přiblížení hodnot pro náš výběr těm pro celou populaci.

Zde vstupuje do hry pravděpodobnost – budeme chtít znát pravděpodobnost přiblížení hodnot pro náš výběr těm pro celou populaci.
Říkáme, že $\bar{X}$ je nestranným odhadem střední hodnoty znaku pro populaci, zatímco výběrový rozptyl je nestranným odhadem rozptylu.

Zde vstupuje do hry pravděpodobnost – budeme chtít znát pravděpodobnost přiblížení hodnot pro náš výběr těm pro celou populaci.
Říkáme, že $\bar{X}$ je nestranným odhadem střední hodnoty znaku pro populaci, zatímco výběrový rozptyl je nestranným odhadem rozptylu.
V případě, že bychom realizovali výběr z populace bez vracení, bude výběrový průměr stále nestranným odhadem střední hodnoty, výběrový rozptyl ale již ne (vyskočí tam faktor $(N-1)/N$).

V praktických úlohách je třeba znát nejen číselné charakteristiky výběrového průměru a rozptylu, ale jejich úplné rozdělení pravděpodobnosti. To můžeme samozřejmě odvodit, pouze známe-li konkrétní rozdělení pravděpodobnosti X_i. Jako užitečnou ilustraci se podíváme na náhodný výběr z normálního rozdělení.
Výběrový průměr bude mít normální rozdělení a protože již známe jeho střední hodnotu a rozptyl, bude $\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n}\sigma^2)$.
O něco složitější je to s odvozením rozdělení pravděpodobnosti výběrového rozptylu. Uvažme vektor Z normovaných normálních veličin

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

Theorem

Je-li $(X_1, \ldots, X_n)$ *náhodný výběr z rozdělení* $\mathrm{N}(\mu, \sigma^2)$, *pak jsou* $\bar{X}$ *a* S^2 *nezávislé veličiny a platí*

$$\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n}\sigma^2), \quad \frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}\,.$$

Okamžitým důsledkem je, že normalizovaný výběrový průměr

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

má studentovo t-rozdělení pravděpodobnosti s $n - 1$ stupni volnosti.

Plán přednášky

Připomenutí ...

Velmi častou úlohou je pro spočtenou hodnotu $\bar{X}$ výběrového průměru určit interval, ve kterém se skutečná hodnota průměru veličiny pro celou populaci nachází s předem danou (vysokou) pravděpodobností.

Připomenutí ...

Velmi častou úlohou je pro spočtenou hodnotu $\bar{X}$ výběrového průměru určit interval, ve kterém se skutečná hodnota průměru veličiny pro celou populaci nachází s předem danou (vysokou) pravděpodobností.
Např. pro náhodnou veličinu X s normálním rozdělením máme její normovanou veličinu $Z = \frac{X - \mathrm{E}\,X}{\sqrt{\mathrm{var}\,X}}$ s výběrovým průměrem $\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$ a chceme najít takovýto interval pro pravděpodobnost $1 - \alpha$, $\alpha \in (0, 1)$.

Připomenutí ...

Velmi častou úlohou je pro spočtenou hodnotu $\bar{X}$ výběrového průměru určit interval, ve kterém se skutečná hodnota průměru veličiny pro celou populaci nachází s předem danou (vysokou) pravděpodobností.

Např. pro náhodnou veličinu X s normálním rozdělením máme její normovanou veličinu $Z = \frac{X - \mathrm{E}\,X}{\sqrt{\mathrm{var}\,X}}$ s výběrovým průměrem $\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$ a chceme najít takovýto interval pro pravděpodobnost $1 - \alpha$, $\alpha \in (0, 1)$.

Potřebujeme tedy znát hodnotu $z(\alpha)$ takovou, že $P(Z > z(\alpha)) = \alpha$.

Připomenutí ...

Velmi častou úlohou je pro spočtenou hodnotu $\bar{X}$ výběrového průměru určit interval, ve kterém se skutečná hodnota průměru veličiny pro celou populaci nachází s předem danou (vysokou) pravděpodobností.

Např. pro náhodnou veličinu X s normálním rozdělením máme její normovanou veličinu $Z = \frac{X - \mathrm{E}\,X}{\sqrt{\mathrm{var}\,X}}$ s výběrovým průměrem $\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$ a chceme najít takovýto interval pro pravděpodobnost $1 - \alpha$, $\alpha \in (0, 1)$.

Potřebujeme tedy znát hodnotu $z(\alpha)$ takovou, že $P(Z > z(\alpha)) = \alpha$.

Je-li $F(x)$ spojitá rostoucí distribuční funkce naší veličiny, pak zjevně $z(\alpha) = F^{-1}(1 - \alpha)$. Pro normální rozdělení splňuje distribuční funkce Φ tento požadavek. Takto definovaným hodnotám $z(\alpha)$ se říká **kritické hodnoty**.

Protože je hustota pro normální rozdělení symetrická kolem jeho střední hodnoty, dostáváme $1 - \alpha = P(|Z| < z(\alpha/2))$.

$$1-\alpha = P\left(\left|\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right| < z(\alpha/2)\right)$$
$$= P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z(\alpha/2)\right)$$

což je interval s náhodnými konci, který s námi určenou pravděpodobností pokrývá neznámý parametr μ. V kontextu takových úloh hovoříme o **intervalu spolehlivosti** s **koeficientem spolehlivosti** $1-\alpha$.

$$1-\alpha = P\left(\left|\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right| < z(\alpha/2)\right)$$
$$= P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z(\alpha/2)\right)$$

což je interval s náhodnými konci, který s námi určenou pravděpodobností pokrývá neznámý parametr μ. V kontextu takových úloh hovoříme o **intervalu spolehlivosti** s **koeficientem spolehlivosti** $1-\alpha$.

Pro normální rozdělení je velice populární kritická hodnota $z(0,025) = 1,96$, která odpovídá naší úloze se zvolenou pravděpodobností 95%.

$$1-\alpha = P\left(\left|\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right| < z(\alpha/2)\right)$$
$$= P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z(\alpha/2)\right)$$

což je interval s náhodnými konci, který s námi určenou pravděpodobností pokrývá neznámý parametr μ. V kontextu takových úloh hovoříme o **intervalu spolehlivosti** s **koeficientem spolehlivosti** $1-\alpha$.

Pro normální rozdělení je velice populární kritická hodnota $z(0,025) = 1,96$, která odpovídá naší úloze se zvolenou pravděpodobností 95%.

Kritické hodnoty jsou dány pomocí tzv. **kvantilové funkce**

$$F^{-1}(u) = \inf\{x \in \mathbb{R}; F(x) \geq u\}, \quad 0 < u < 1.$$

Kvantilová funkce skutečně dává přímo příslušné kvantily, např. $F^{-1}(0,5)$ je medián, atd.

Example

Před deseti lety byl uskutečněn rozsáhlý výzkum výšky desetiletých chlapců a zjistilo se, že střední výška byla $\mu_0 = 136,1$cm se směrodatnou odchylkou $\sigma = 6,4$cm. Nyní byly na náhodném výběru 15 desetiletých chlapců zjištěny následující výšky: 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147. Je známo, že variabilita výšek v populaci se mění velice pomalu, zatímco výšky se mohou měnit rychle. **Otázka: došlo ke změně střední výšky populace desetiletých chlapců?**

Example

Před deseti lety byl uskutečněn rozsáhlý výzkum výšky desetiletých chlapců a zjistilo se, že střední výška byla $\mu_0 = 136,1$cm se směrodatnou odchylkou $\sigma = 6,4$cm. Nyní byly na náhodném výběru 15 desetiletých chlapců zjištěny následující výšky: 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147. Je známo, že variabilita výšek v populaci se mění velice pomalu, zatímco výšky se mohou měnit rychle. **Otázka: došlo ke změně střední výšky populace desetiletých chlapců?**

Ze zadání předpokládáme, že výběr 15 hodnot je z normálního rozdělení se známým rozptylem σ^2 a otázku si upřesníme tak, že hledáme v jakém intervalu je nyní střední hodnota výšky populace se spolehlivostí 95% :

Example

Před deseti lety byl uskutečněn rozsáhlý výzkum výšky desetiletých chlapců a zjistilo se, že střední výška byla $\mu_0 = 136,1$cm se směrodatnou odchylkou $\sigma = 6,4$cm. Nyní byly na náhodném výběru 15 desetiletých chlapců zjištěny následující výšky: 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147. Je známo, že variabilita výšek v populaci se mění velice pomalu, zatímco výšky se mohou měnit rychle. **Otázka: došlo ke změně střední výšky populace desetiletých chlapců?**

Ze zadání předpokládáme, že výběr 15 hodnot je z normálního rozdělení se známým rozptylem σ^2 a otázku si upřesníme tak, že hledáme v jakém intervalu je nyní střední hodnota výšky populace se spolehlivostí 95% : $\bar{x} = 139,133$ a tedy interval spolehlivosti je $(139,133 - (6,4/\sqrt{15})1,96, 139,133 + (6,4/\sqrt{15})1,96) = (135,9, 142,4)$.

Example

Před deseti lety byl uskutečněn rozsáhlý výzkum výšky desetiletých chlapců a zjistilo se, že střední výška byla $\mu_0 = 136,1$cm se směrodatnou odchylkou $\sigma = 6,4$cm. Nyní byly na náhodném výběru 15 desetiletých chlapců zjištěny následující výšky: 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147. Je známo, že variabilita výšek v populaci se mění velice pomalu, zatímco výšky se mohou měnit rychle. **Otázka: došlo ke změně střední výšky populace desetiletých chlapců?**

Ze zadání předpokládáme, že výběr 15 hodnot je z normálního rozdělení se známým rozptylem σ^2 a otázku si upřesníme tak, že hledáme v jakém intervalu je nyní střední hodnota výšky populace se spolehlivostí 95% : $\bar{x} = 139,133$ a tedy interval spolehlivosti je $(139,133 - (6,4/\sqrt{15})1,96, 139,133 + (6,4/\sqrt{15})1,96) = (135,9, 142,4)$.

Protože tento interval pokrývá i populační průměr před deseti lety, nemůžeme na této hladině spolehlivosti tvrdit, že se populační výška změnila.

Plán přednášky

Odhadování parametrů může být bodové nebo intervalové. V předchozím příkladu takovými byly výběrový průměr $\bar{x} = 139,133$ a interval spolehlivosti (135,9;142,4).
Obecně postupujeme takto: Pro náhodný výběr rozsahu n $X_1, \ldots, X_n$ z rozdělení, které závisí na (vektorovém) parametru θ hledáme funkci náhodných veličin (říkáme též statistiku nebo výběrovou statistiku) $T(X_1, \ldots, X_n)$, která bude mít v „rozumném smyslu“ blízko ke skutečné hodnotě θ.

Odhadování parametrů může být bodové nebo intervalové. V předchozím příkladu takovými byly výběrový průměr $\bar{x} = 139,133$ a interval spolehlivosti (135,9;142,4).
Obecně postupujeme takto: Pro náhodný výběr rozsahu n $X_1, \ldots, X_n$ z rozdělení, které závisí na (vektorovém) parametru θ hledáme funkci náhodných veličin (říkáme též statistiku nebo výběrovou statistiku) $T(X_1, \ldots, X_n)$, která bude mít v „rozumném smyslu“ blízko ke skutečné hodnotě θ.
Jakožto funkce náhodných veličin je T opět náhodnou veličinou (resp. náhodným vektorem). Konstanta (resp. konstantní vektor)

$$b = \mathsf{E}\, T - \theta$$

se nazývá **vychýlení** odhadu T. **Nestranný** (nevychýlený) je takový odhad, kdy $b = 0$.

Nejlepší odhad

Máme-li k dispozici jistou třídu odhadů $\mathcal{T}$, říkáme že T je **nejlepším odhadem**, má-li mezi všemi nejmenší rozptyl.

Nejlepší odhad

Máme-li k dispozici jistou třídu odhadů $\mathcal{T}$, říkáme že T je **nejlepším odhadem**, má-li mezi všemi nejmenší rozptyl.
$T = T_n$ je **konzistentním odhadem**, je-li pro každé $\epsilon > 0$

$$\lim_{n\to\infty} P(|T_n - \theta| < \epsilon) = 1.$$

Theorem

Je-li $\lim_{n\to\infty} \mathrm{E}\, T_n = \theta$, $\lim_{n\to\infty} \mathrm{var}\, T_n = 0$, *pak je* T_n *konzistentním odhadem* θ.