Cvičení ze statistiky - 7

Filip Děchtěrenko

Minule bylo..

- Probrali jsme spojité modely
- Tyhle termíny by měly být známé:
  - Rovnoměrné rozdělení
  - Střední hodnota
  - Mccalova transformace
  - Normální rozdělení

Přehled fint

- $P(Z>z)=1-P(Z\leq z)=1-\phi(z)$
  $P(Z\geq z)=1-P(Z<z)=1-\phi(z)$
  Aspoň na jedné straně musí být rovnost
- $P(X<a$ nebo $X>b)=1-P(a\leq X\leq b)$
  dá se to představit na číselné ose
- $\phi(-a)=1-\phi(a)$
  toto použijeme vždy, když budeme chtít najít zápornou hodnotu v tabulce
- $P(a<Z\leq b)=P(Z\leq b)-P(Z\leq a)=\phi(b)-\phi(a)$
  obvykle bývá a záporné, potom
  $P(-a<Z\leq b)=\phi(b)-\phi(-a)=\phi(b)-(1-\phi(a))=\phi(b)+\phi(a)-1$
- $P(-a<Z\leq a)=2\phi(a)-1$

Součet a průměr jako náhodné veličiny

- Mějme dva stánky ze zmrzlinou. Náhodné veličiny X_1 a X_2, značí počet lidí, kteří si koupili zmrzlinu u prvního a druhého stánku za jeden den.
- Může nás zajímat, kolik zmrzliny se prodalo za den celkem, případně kolik zmrzliny se prodalo průměrně. Tedy průměr a součet jsou pro nás náhodné veličiny
- Jak se ale mají vlastnosti?

Centrální limitní věta

- Máme-li n nezávislých náhodných výběrů $X_1, X_2, X_3, \dots, X_n$ ze stejných výběrů (značí se *i.i.d*), potom lze jejich součet a průměr aproximovat náhodným rozdělením
- Protože jsou ze stejného rozdělení, mají všechny stejné EX a var X (EX=EX_1=EX_2=…=EX_n)

1. Součet $X = \sum X_i$ má normální rozdělení
$$X \sim N(n \cdot EX, n \cdot VarX)$$
2. Průměr $\bar{X} = \frac{\sum X_i}{n}$ má normální rozdělení
$$\bar{X} \sim N(EX, \frac{Var\ X}{n})$$
3. Jsou-li X_i z binomického rozdělení, potom Z se dá aproximovat normálním rozdělením (Laplaceho věta)

Rozdělení součtu a průměru

- Mějme n i.i.d proměnných X_i, potom pro součet X a průměr $\bar{X}$ platí, že jsou z normálního rozdělení
- Můžeme je tedy transformovat na Z rozdělení a najít v tabulkách
- Protože $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, platí $Z = \frac{\bar{X} - E\bar{X}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$

  kde
  - $\bar{X}$ je průměr jako náhodná veličina
  - $E\bar{X}$ je střední hodnota náhodné veličiny průměru
  - $\sigma_{\bar{X}}$je směrodatná odchylka náhodné veličiny průměru
  - μ je střední hodnota X_i
  - σ je směrodatná odchylka X_i
- Obdobně součet

$$Z = \frac{X - EX}{\sigma_X} = \frac{X - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$

Příklady

- Výtah s kapacitou pro 6 míst má kapacitu 550 kg. Jaká je pravděpodobnost, že při plném obsazení bude tato hodnota překročena, má-li hmotnost cestujícího střední hodnotu 90 kg a směrodatnou odchylku 10 kg?
  - P(X>550)=1-P($\frac{X-90\cdot 6}{10\sqrt{6}}<\frac{550-90\cdot 6}{10\sqrt{6}}$)=1-P(Z<1)=1-ϕ(1)=1-0.66=0.34
- Počet bodů v testu je náhodná veličina z rozdělení N(20,25). Určete, jaká je pravděpodobnost , že průměr bodů čtyř lidí je menší než 22.
  - P($\bar{X}$<22)=P($\frac{\bar{X}-20}{5}\sqrt{4}<\frac{22-20}{5}\sqrt{4}$)=P(Z<0.8)=ϕ(0.8)=0.788
- Kolik lidí potřebuji, aby byla pravděpodobnost, že průměr bodů bude menší než 22 s pstí 99%?
  - 0.99=P($Z<\frac{22-20}{5}\sqrt{n}$) ->inverzní hledání v tabulkách pro p=0.99
    2.32<0.4$\sqrt{n}$
    33.64<n
    tedy potřebujeme aspoň 34 lidí

Laplaceova věta

- Mějme proměnnou $X \sim Bi(n, \pi)$, potom platí

$$\lim_{n \to \infty} \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0,1)$$

- Tedy pro dostatečně velká n se nám blíží binomické rozdělení k normálnímu (a můžeme ho aproximovat N(0,1)
- Podmínka pro korektní použití Laplaceovy věty:

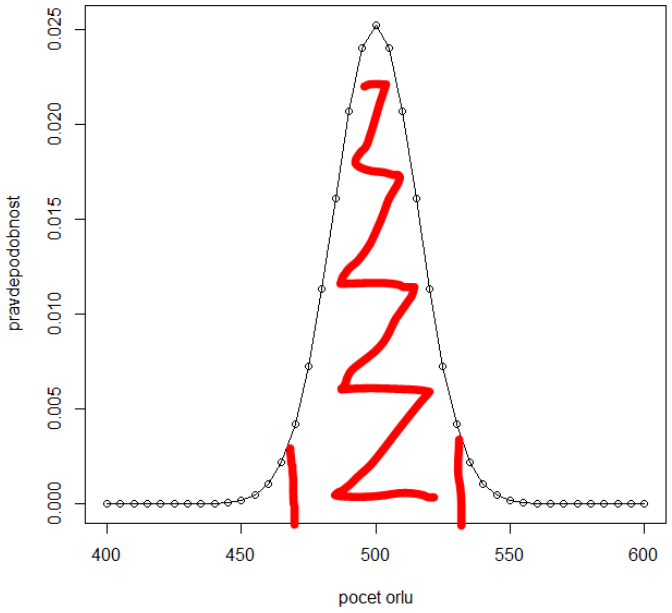
$$n\pi(1-\pi) > 9$$

- Je to klasická Z transformace s parametry

$$\mu = n\pi, \sigma^2 = n\pi(1-\pi)$$

Příklady

- Jaká je pravděpodobnost, že z tisíce hodů mincí bude 470 až 530 orlů?
  - X∼Bi(1000,0.5)
    platí 1000·0.5·(1-0.5)=250>9,můžeme aproximovat normálním rozdělením
    EX=500, var(X)=250
    P(470<X<530)=
    $P(\frac{470-500}{\sqrt{250}}<\frac{X-500}{\sqrt{250}}<\frac{530-500}{\sqrt{250}})$=P(-1.897<Z<1.897)=2·ϕ(1.897)-1=0.94



Korekce na spojitost

- Máme-li binomické rozdělení, jsou následující hodnoty stejné
  $P(X>29)=P(X\geq 30)$
- Normální rozdělení ale má tyto hodnoty odlišné!
  - Použijeme průměr, tj. $P(X\geq 29.5)$
- Obdobně $P(29<X)=P(30\leq X)$
  - Použijeme $P(29.5\leq X)$

Laplaceova věta pro relativní četnost

- Laplaceova věta patří i pro relativní četnost (máme X/n namísto X)
- Tedy E(X/n)=π a var(X/n)= π(1- π)/n (protože var(bX)=b^2var(X))
- Opět musí platit, že nπ(1-π)>9, abychom mohli aproximovat normálním rozdělením

Příklad

- V populaci má 4% lidí žloutenku. Jaká je pravděpodobnost, že ve výběru 25 osob se bude počet osob se žloutenkou lišit o více než 3%
  - Tedy nás zajímá, jaká je pst, že se výběru vyskytuje více než 7% lidí se žloutenkou nebo méně než 1% lidé se žloutenkou
  - $\mu=0.04$; $\sigma=0.002$
  - Chceme tedy $P(p<0.01 \text{ nebo } p>0.07)$
  - Nebo neumíme! Platí:
    $P(p<a \text{ nebo } p>b)=1-P(a<p<b)$
    tedy (schematicky) P(nebo)=1-P(a současně)
  - $P(p<0.01 \text{ nebo } p>0.07)=1-P(0.01<p<0.07)=1-P(\frac{0.01-0.04}{\sqrt{0.002}}<Z<\frac{0.07-0.04}{\sqrt{0.002}})=1-P(-0.67<Z<0.67)=$
    $=1-(2\phi(0.67)-1)=2-2\cdot 0.75=0.50$

Inferenční statistika

- Popsali jsme si deskriptivní statistiku a několik základních pravděpodobnostních modelů
- Inferenční statistika nám říká jak na základě vzorku (ten popíšeme pomocí deskriptivní statistiky) můžeme odvozovat parametry pravděpodobnostního modelu, ze kterého pocházejí data
- Můžeme na základě výběrových charakteristiky určit parametry modelu přesně?
- Ne -> Mluvíme o *odhadech* parametrů modelu

Odhady parametrů

- Můžeme odhadovat jednotlivé charakteristiky modelů, pak mluvíme o *bodových odhadech*
- Pro každý parametr máme ještě intervalový odhad (určuje, kde hledaný parametr nachází)
  - Vyberu jednoho člověka (věk 24 let). Výběrový průměr je 24 let. Co můžu říct o průměru celé populace?
    - Průměr celé populace je 24 let s pstí <0.001

Bodové odhady

- Jednotlivým výběrovým charakteristikám odpovídají parametry pravděpodobnostního modelu
- Nás budou zajímat tři
  - $\bar{x}$ je bodovým odhadem EX
  - s je bodovým odhadem σ
    s^2 je bodovým odhadem $varX$
  - p je bodovým odhadem π

Intervalové odhady

- Chceme odhadnout, kde se vyskytuje odhadovaný parametr s danou pstí
- My budeme hledat intervaly spolehlivosti (konfidenční intervaly, CI) pouze pro π a μ
- Postup:
  - Stanovme požadovanou pst
    - Typicky se používá 99%,95% a 90%
  - Pro tyto hodnoty najdeme z hodnoty ze vzorečku P(-a<Z≤a)=2ϕ(a)-1
    - Dostaneme z hodnoty 2.576, 1.96 a 1.64
  - Upravíme (ukázka např. pro μ)

$$P(-z < Z < z) = P(-z < \frac{\overline{X} - \mu}{\sigma}\sqrt{n} < z) =$$
$$P(-z\sigma < (\overline{X} - \mu)\sqrt{n} < z\sigma) =$$
$$P(\overline{X} - z\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z\frac{\sigma}{\sqrt{n}})$$

  - A dosadíme
- Interval zapisujeme jako (D;H) kde D a H jsou dolní a horní mez

Další intervalové odhady

- Neznáme-li σ, použijeme t-rozdělení (později), případně do vzorečku dosadíme *s* namísto σ (nepřesné)
- Pro relativní četnost:

$$P(p - z\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < p + z\sqrt{\frac{\pi(1-\pi)}{n}})$$

- Ve vzorečku nahradíme π pomocí *p*, tedy

$$P(p - z\sqrt{\frac{p(1-p)}{n}} < \pi < p + z\sqrt{\frac{p(1-p)}{n}})$$

Šíře intervalu

- Někdy nás zajímá, kolik potřebujeme lidí, abychom měli interval s daným rozsahem
  - Šíře klesá s rostoucí velikostí vzorku
  - Šíře roste s rostoucím koeficientem spolehlivosti
- Šíři CI určíme jako $\frac{H-D}{2}$, kde H a D jsou horní a dolní mez

Příklady

1. Opakovanými měřeními byla zjištěna tloušťka vlákna: 210, 217,209, 216, 216, 215, 220, 214, 213 (10^{-6} m). Je známo, že měření mají rozdělení N(μ, 25). Nalezněte 95% interval spolehlivosti pro μ.
   - Pro 95% CI je z hodnota 1.96, výběrový průměr je 214.4 a velikost vzorku je 9
   - CI tedy je (214.4-1.96·5/3; 214.4+1.96·5/3)=(211.1;217.7)
2. Jak by vypadal 95% CI, pokud bychom naměřili ještě 7 vláken 212,215,210,219,218,213 a 214?
3. Jak by vypadal 90% CI a 99% CI?
4. Kolik vláken bychom museli změřit, abychom měli 95% interval o šíři 2?
   - Dosadíme do vzorce $\frac{H-D}{2}$ a dostaneme

     $$\left(\bar{X}+1.96\frac{5}{\sqrt{n}}\right)-\left(\bar{X}-1.96\frac{5}{\sqrt{n}}\right)<2$$

     $1.96\frac{5}{\sqrt{n}}<1$

     n>96.04

     Potřebujeme tedy alespoň 97 vláken
5. Na dotazník ohledně preference zvířete odpovědělo 25 lidí. 7 z nich preferovalo kočku jako domácího mazlíčka. Určete 95% CI pro preferenci kočky v celém základním souboru
   - $p=\frac{7}{25}$

     95% CI tedy je $(0.28-1.96\sqrt{\frac{0.28(1-0.28)}{25}};\ 0.28-1.96\sqrt{\frac{0.28(1-0.28)}{25}})=(0.104;0.456)$
   - Co je tady za metodologický problém?