Bodové odhady

Náhodná proměnná ξ s množinou hodnot V = ***výběrový prostor***. Máme k dispozici n pozorování této náhodné proměnné $\{x_1, x_2, ..., x_n\}$ o kterých předpokládáme, že jsou stejně rozdělené (identically distributed) tj. se stejnou distribuční funkcí $F(x)$ a nezávislé (independent) {iid}. Pozorování $\{x_1, x_2, ..., x_n\}$ nazýváme ***náhodným výběrem rozsahu n.***

Příklad: náhodný výběr $\{x_1, x_2, ..., x_n\}$ rozsahu n náhodné proměnné ξ s normálním rozdělením $N(\mu, \sigma^2)$ má pak hustotu:

$$f(x_1, x_{21}, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\left(\sqrt{2\pi\sigma}\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Výběrovým prostorem je pak celá reálná osa R_1.

Nadále budeme předpokládat, že pro uvažovanou náhodnou proměnnou ξ existují první čtyři momenty, tj. např. $E\{\xi\} = \mu; E\{\xi^2\} = \mu_2; E\{\xi^3\} = \mu_3; E\{\xi^4\} = \mu_4$.

Výběrová střední hodnota – průměr

Výběrová střední hodnota = průměr $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Střední hodnota průměru $E\{\bar{x}\} = \frac{1}{n}\sum_{i=1}^{n} E\{x_i\} = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$. Střední hodnota průměru je rovna střední hodnotě pozorované náhodné proměnné (samozřejmě pokud tato existuje).

Rozptyl průměru:

$$\sigma^2\{\bar{x}\} = E\{(\bar{x}-\mu)^2\} = E\{(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu)^2\} = E\{(\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{n}{n}\mu)^2\} = E\{\frac{1}{n^2}(\sum_{i=1}^{n}(x_i-\mu))^2\} =$$

$$= \frac{1}{n^2}E\{(\sum_{i=1}^{n}(x_i-\mu))^2\} = \frac{1}{n^2}E\{\sum_{i=1}^{n}(x_i-\mu)^2 + \sum_{i\neq j}(x_i-\mu)(x_j-\mu)\} = \frac{\sigma^2}{n},$$

protože $\sum_{i=1}^{n} E\{(x_i-\mu)^2\} = n\sigma^2$ a $E\{(x_i-\mu)(x_j-\mu)\} = 0$ pro $i \neq j$.

Shrnuto: $\sigma^2\{\bar{x}\} = \frac{\sigma^2}{n}$ a $\sigma\{\bar{x}\} = \frac{\sigma}{\sqrt{n}}$. Tj. směrodatná odchylka průměru se snižuje oproti směrodatné odchylce původní náhodné proměnné faktorem $\frac{1}{\sqrt{n}}$.

Pokud máme dostatečný rozsah náhodného výběru ($n \ge 30$ zde pouze empirický poznatek) pak můžeme využít centrální limitní věty o konvergenci k normálnímu rozdělení a tedy: $\bar{x} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$ a proto má náhodná veličina $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\,\frac{\bar{x}-\mu}{\sigma}$ rozdělení $N(0,1)$. Potom:

$P\{d < \bar{x}-\mu < h\} = P\{d\frac{\sqrt{n}}{\sigma} < \frac{\sqrt{n}}{\sigma}(\bar{x}-\mu) < h\frac{\sqrt{n}}{\sigma}\} = \Phi(h\frac{\sqrt{n}}{\sigma}) - \Phi(d\frac{\sqrt{n}}{\sigma})$, kde $\Phi(x)$ je distribuční funkcí rozdělení $N(0,1)$. Toho lze využít pro intervalové odhady střední hodnoty, pokud známe nebo umíme „kvalitně odhadnout“ σ - viz dále.

Výběrový rozptyl

Nejprve stanovíme jeho modifikaci pro případ kdy je parametr μ znám:

$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$. Označíme $y_i = (x_i-\mu)^2$, pak $E\{y_i\} = \sigma^2$ a

$$\sigma^2\{y_i\} = E\{y_i^2\} - (E\{y_i\})^2 = E\{(x_i-\mu)^4\} - \sigma^4 = m_4 - \sigma^4, \text{ kde } m_4 = E\{(x_i-\mu)^4\}$$

Proto: $E\{S_0^2\} = E\{\bar{y}\} = \sigma^2$ a $\sigma^2\{S_0^2\} = \sigma^2\{\bar{y}\} = \frac{m_4-\sigma^4}{n}$.

Pro případ kdy μ není známo, zavedeme modifikaci:

$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Pro to platí:

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu+\mu-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{n}\sum_{i=1}^{n}(\bar{x}-\mu)^2 = S_0^2 - (\bar{x}-\mu)^2$$

neboť $\frac{2}{n}\sum_{i=1}^{n}(x_i-\mu)(\mu-\bar{x}) = -2(\bar{x}-\mu)^2$.

Shrnuto: $S_1^2 = S_0^2 - (\bar{x}-\mu)^2$. Proto: $E\{S_1^2\} = E\{S_0^2\} - E\{(\bar{x}-\mu)^2\} = E\{S_0^2\} - \sigma^2\{\bar{x}\} =$

$= \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\frac{n-1}{n}$.

Protože $E\{S_1^2\} = \sigma^2\frac{n-1}{n}$, je vhodné volit novou statistiku $S^2 = \frac{n}{n-1}S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ a u té již platí: $E\{S^2\} = \sigma^2$.

Statistika $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ je nazývána výběrovým rozptylem a $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ výběrovou směrodatnou odchylkou.

Dále určíme rozptyl statistiky $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Ale

$S^2 = \frac{n}{n-1}S_1^2 = \frac{n}{n-1}\left(S_0^2 - (\bar{x}-\mu)^2\right)$, proto $(S^2)^2 = \left(\frac{n}{n-1}\right)^2\left(S_0^2 - (\bar{x}-\mu)^2\right)^2 =$

$$=\left(\frac{n}{n-1}\right)^2\left(\left(S_0^2\right)^2+(\overline{x}-\mu)^4-2S_0^2(\overline{x}-\mu)^2\right).$$

Ale $\quad E\left\{\left(S_0^2\right)^2\right\}=\sigma^2\left(\left(S_0^2\right)\right)+\left(E\left\{S_0^2\right\}\right)^2=\frac{m_4}{n}+\frac{n-1}{n}\sigma^4$,

$E\left\{S_0^2(\overline{x}-\mu)^2\right\}=\frac{1}{n^2}\left[m_4+(n-1)\sigma^4\right]$ a

$E\left\{(\overline{x}-\mu)^4\right\}=\frac{1}{n^3}\left[m_4+3(n-1)\sigma^4\right]$.

Protože: $\sigma^2\left(S^2\right)=E\left\{\left(S^2\right)^2\right\}-\left(E\left\{S^2\right\}\right)^2=E\left\{\left(S^2\right)^2\right\}-\sigma^4$, dostáváme:

$$\sigma^2\left(S^2\right)=\frac{m_4}{n}-\frac{n-3}{n(n-1)}\sigma^4.$$

Zdroj: J. Hátle – J. Likeš: Základy počtu pravděpodobnosti, SNTL/ALFA, Praha 1974 (str.151- 153) nebo novější vydání. Taktéž jakákoliv solidnější učebnice statistiky.

Pro výpočet průměru známe některá výpočetní schémata, např.:

$\overline{z_{n+1}}=\frac{1}{n+1}\sum_{i=1}^{n+1}x_i=\frac{1}{n+1}\left(\frac{n}{n}\sum_{i=1}^{n}x_i+x_{n+1}\right)=\frac{1}{n+1}\left(n\overline{z_n}+x_{n+1}\right)=\overline{z_n}+\frac{1}{n+1}\left(x_{n+1}-\overline{z_n}\right)$ a

$E\left\{\overline{z_{n+1}}\right\}=E\left\{\overline{z_n}\right\}+\frac{1}{n+1}E\left\{\left(x_{n+1}-\overline{z_n}\right)\right\}$, protože však $E\left\{\overline{z_{n+1}}\right\}=E\left\{\overline{z_n}\right\}$, je $E\left\{\left(x_{n+1}-\overline{z_n}\right)\right\}=0$.

Námět: Odečtením rovnic: $\overline{z_{n+1}}=\overline{z_n}+\frac{1}{n+1}\left(x_{n+1}-\overline{z_n}\right)$ a $E\left\{\overline{z_{n+1}}\right\}=E\left\{\overline{z_n}\right\}$ lze získat vztah pro vývoj rozptylu průměru v závislosti na n.

Obecnější pohled na bodové odhady (jednorozměrný případ)

Pojem statistika:

Náhodná proměnná ξ má množinu hodnot V = ***výběrový prostor*** a hustotu $f(x,g)$, distribuční funkci $F(x,g)$, kde $g\in R_1$ je neznámý, odhadovaný, parametr. Máme k dispozici náhodný výběr z tohoto rozdělení o rozsahu n $\{x_1,x_2,...,x_n\}$. Odhadem parametru g nebo rozdělení $F(x,g)$, $f(x,g)$ budeme rozumět statistiku $\hat{g}(x_1,x_2,...,x_n)\cong\hat{g}$. Slovem **statistika** bude označováno jakékoliv „pravidlo“ přiřazující náhodnému výběru nějakou hodnotu pocházející z „definičního oboru“ odhadovaného parametru. V našem případě: $\hat{g}:V^n\to R_1$. Statistika nemusí být vázána přímo k nějakému parametru, může být „navázána“ i na nějakou funkci (parametrická funkce) takového parametru(ů).

O (parametrickém) odhadu řekneme, že je nestranný (ve smyslu střední hodnoty) $\Leftrightarrow E\left\{\hat{g}\right\}=g$. Případný rozdíl $E\left\{\hat{g}\right\}-g$ bude nazýván vychýlením (stranností).

Poznámka: nověji se objevují i definice nestrannosti vázané na medián.

Příklad 1.: Výše bylo odvozeno, že $E\left\{\frac{1}{n}\sum_{i=1}^{n}x_i\right\}=E(\xi)$. Tedy průměr je nestranným odhadem střední hodnoty, pokud tato existuje.

Příklad 2.: Pro posunuté exponenciální rozdělení $E(A,\delta)\equiv f(x)=\frac{1}{\delta}e^{-\frac{1}{\delta}(x-A)} \Leftrightarrow x>A; f(x)=0 \Leftrightarrow x\le A$ je nestranným odhadem parametrické funkce $A+\delta$ statistika $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$. Tedy v tomto případě průměr je nestranným odhadem „parametrické funkce“ $A+\delta$.

Příklad 3.: Mějme výběr z modifikovaného geometrického rozdělení $p_i=(1-q)^{i-1}q; i\ge 1$, to může modelovat situaci, kdy opakujeme nějaký experiment a pozorujeme při kolikátém opakování se daný (sledovaný) jev ***A*** realizoval (pravděpodobnost toho, že se daný jev realizuje při jakémkoliv experimentu je pak q). Ke každému pozorování označíme posloupnost experimentů po řadě $E_1,E_2,...,E_i; E_j=\bar{A} \Leftrightarrow j<i; E_i=A$. Zvolíme statistiku $t(E_1,E_2,...,E_i)=1 \Leftrightarrow E_1=A$ a $t(E_1,E_2,...,E_i)=0 \Leftrightarrow E_1=\bar{A}$. Je evidentní, že $E\{t(E_1,E_2,...,E_i)\}=1q+0(1-q)=q$, tedy taková statistika je nestranným odhadem parametru q. Jak lze nahlédnout, takový odhad je velice špatný a neužitečný pro další použití. Proto je důležitá další charakterizace odhadů:

O odhadu $\hat{g}(x_1,x_{21},...,x_n)\cong\hat{g}$ řekneme, že je vydatným odhadem g pokud platí: $\hat{g}(x_1,x_{21},...,x_n)\cong\hat{g}$ je nestranným odhadem a navíc platí: $\sigma^2\left(\hat{g}\right)\le\sigma^2\left(\hat{g}_a\right)$ pro všechny možné nestranné odhady $\hat{g}_a$ parametru g. Vydatný odhad tedy minimalizuje chybu měřenou pomocí jeho rozptylu.

Rozdělení výběrového průměru

1. Je k dispozici náhodný výběr z alternativního rozdělení o rozsahu n $\{x_1,x_2,...,x_n\}$. Potom $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i; n\bar{x}=\sum_{i=1}^{n}x_i$ má binomické rozdělení $p_i=\binom{n}{i}p^i(1-p)^{n-i}$ a distribuční funkce průměru $F_{\bar{x}}(x)=P\left(\bar{x}<x\right)=P\left(n\bar{x}<nx\right)=\sum_{i<nx}\binom{n}{i}p^i(1-p)^{n-i}$.
2. Je k dispozici náhodný výběr z binomického rozdělení $p_i=\binom{m}{i}p^i(1-p)^{m-i}$ o rozsahu n $\{x_1,x_2,...,x_n\}$. Potom $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i; n\bar{x}=\sum_{i=1}^{n}x_i$ má binomické rozdělení

$p_i = \binom{nm}{i} p^i (1-p)^{nm-i}$ a distribuční funkce průměru $F_{\bar{x}}(x) = P(\bar{x} < x) = P(nm\bar{x} < nmx) = \sum_{i<nmx} \binom{nm}{i} p^i (1-p)^{nm-i}$.

3. Je k dispozici náhodný výběr z Poissonova rozdělení $p_i = e^{-\lambda} \frac{\lambda^i}{i!}$ o rozsahu n $\{x_1, x_2, ..., x_n\}$. Potom $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i ; n\bar{x} = \sum_{i=1}^{n} x_i$ má Poissonovo rozdělení $p_i = e^{-n\lambda} \frac{(n\lambda)^i}{i!}$ a distribuční funkce průměru $F_{\bar{x}}(x) = P(\bar{x} < x) = P(n\bar{x} < nx) = \sum_{i<nx} e^{-n\lambda} \frac{(n\lambda)^i}{i!}$.
4. Je k dispozici náhodný výběr z normálního rozdělení $F_{\xi}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s-\mu)^2}{2\sigma^2}} ds$ o rozsahu n $\{x_1, x_2, ..., x_n\}$. Potom $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ má normální rozdělení $F_{\xi}(x) = \int_{-\infty}^{x} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n(s-\mu)^2}{2\sigma^2}} ds$ a je to tedy distribuční funkce průměru (normální rozdělení se střední hodnotou μ a rozptylem $\frac{\sigma^2}{n}$).
5. …….

Námět: Odvoďte distribuční funkci průměru výběru z exponenciálního rozdělení $F_{\xi}(x) = 1 - e^{-\frac{x}{\tau}}; x \geq 0$. Pomůcka součet n nezávislých náhodných exponenciálně rozdělených proměnných má Gama rozdělení $f_{\xi}(x) = \frac{x^{n-1} e^{-\frac{x}{\tau}}}{\tau^n (n-1)!}$.

Doporučená a zdrojová literatura:

Jiří Reif	Metody matematické statistiky, ZČU v Plzni 2004
Jaroslav Hátle, Jiří Likeš	Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974.
Alfréd Rényi	Teorie pravděpodobnosti, ACADEMIA, Praha 1972
C. Radhakrishna Rao	Lineární metody statistické indukce a jejich aplikace, ACADEMIA, Praha 1978