I Bodové a intervalové odhady parametrů rozdělení pravděpodobnosti

I.I Motivace

Získali jsme data $X_1, X_2, \ldots, X_n$, o kterých víme, že pocházejí z normálního rozdělení $N(\mu, \sigma^2)$, ale neznáme hodnoty parametrů μ, σ^2. Chtěli bychom na základě získaných dat tyto hodnoty odhadnout. Zajímá nás, jak tento odhad zkostruovat a, pokud jich máme několik, tak určit ten lepší.

Ve skutečnosti nemusíme pracovat zrovna s normálním rozdělením, data mohou pocházet z libovolného rozdělení, které je závislé na parametru ϑ. Toto obecné rozdělení budeme značit $L(\vartheta)$. Nabízí se dva způsoby, jak k odhadu parametru ϑ přistoupit. Buď nás bude zajímat jedna konkrétní hodnota, kterou budeme považovat za odhad – pak mluvíme o bodovém odhadu. Může nás ale zajímat interval, ve kterém parametr ϑ s pravděpodobností $1-\alpha$ leží. Pak konstruujeme tzv. intervalový odhad.

Interpretační poznámka. Parametr ϑ může být ve skutečnosti i vektor parametrů $\boldsymbol{\vartheta}$, jako například u normálního rozdělení. Pak $\boldsymbol{\vartheta} = (\mu, \sigma^2)$. Pro jednoduchost budeme pracovat se skalárním případem.

I.II Bodový odhad

Máme náhodný výběr $X_1, X_2, \ldots, X_n$, který pochází z rozdělení pravděpodobnosti $L(\vartheta)$. My hodnotu ϑ neznáme, a tak ji chceme odhadnout jednou hodnotou, tu označíme stříškou: $\hat{\vartheta}$. Protože tento odhad bude vypočítaný pomocí náhodného výběru, jedná se o *statistiku*.

Nový pojem: Bodový odhad

Bodovým odhadem $\hat{\vartheta}$ parametru ϑ rozdělení pravděpodobnosti $L(\vartheta)$ je libovolná statistika

$$\hat{\vartheta} = T(X_1, X_2, \ldots, X_n),$$

jejíž hodnoty kolísají okolo skutečné hodnoty ϑ.

Interpretační poznámka. Pro připomenutí, statistika je libovolná funkce náhodného výběru – formulka, do které dosadíme n získaných hodnot.

Tato definice je ale velmi obecná – podle ní mohou být odhady parametry μ normálního rozdělení například statistiky:

- $T_1(X_1, X_2, \ldots, X_n) = X_1 \cdot X_2 \cdot \ldots \cdot X_n$
- $T_2(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $T_3(X_1, X_2, \ldots, X_n) = X_1$

Očividně T_1 a T_3 jsou pěkně mizerné odhady μ, proto potřebujeme nějakým způsobem měřit kvalitu odhadu, abychom mohli ty špatné vyfiltrovat a vybrat ten nejlepší.

Nový pojem: Nestranný odhad

Statistika $T(X_1, X_2, \ldots, X_n)$ je nestranným (nevychýleným, nezkresleným) odhadem parametru ϑ, pokud

$$E[T(X_1, X_2, \ldots, X_n)] = \vartheta.$$

Tento požadavek znamená, že statistika $T(X_1, X_2, \ldots, X_n)$ nenadhodnocuje ani nepodhodnocuje hodnotu parametru ϑ. Je nějaká ze statistik T_1, T_2, T_3 nestranná?

$$E[T_1] = E[\underbrace{X_1 \cdot X_2 \cdot \ldots \cdot X_n}_{\text{nezávislé, } E[X_i]=\mu}] = E[X_1] \cdot E[X_2] \cdot \ldots \cdot E[X_n] = \underbrace{\mu \cdot \mu \cdot \ldots \cdot \mu}_{\text{celkem } n \text{ krát}} = \mu^n \neq \mu$$

$$E[T_2] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

$$E[T_3] = E[X_1] = \mu$$

Statistika T_1 tedy není nestranným odhadem parametru ϑ, zatímco T_2 a T_3 jsou. Tak nějak ale cítíme, že statistika T_2 by měla být lepší, než T_3, vždyť T_3 nevyužívá všechna získaná data! Proto je třeba i nestranné odhady dále porovnávat, abychom mohli nakonec vybrat ten nejlepší.

Nový pojem: Nejlepší odhad

Statistika $T_1(X_1, X_2, \ldots, X_n)$ je lepší odhad parametru ϑ než statistika $T_2(X_1, X_2, \ldots, X_n)$, pokud pro libovolnou skutečnou hodnotu ϑ platí

$$D[T_1(X_1, X_2, \ldots, X_n)] < D[T_2(X_1, X_2, \ldots, X_n)].$$

Odhad, který má ze všech nestranných odhadů nejmenší rozptyl nazýváme nejlepší (vydatný, angl. *efficient*).

Interpretační poznámka. Malý rozptyl se nám líbí, protože nám spolu s nestranností zaručuje, že náš odhad leží blízko opravdové hodnoty parametru ϑ.

Která ze statistik T_2 a T_3 je lepší?

$$D[T_2] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$D[T_3] = D[X_1] = \sigma^2$$

Máme-li tedy náhodný výběr o velikosti n aspoň 2, pak je T_2 lepší, pro $n = 1$ jsou tyto statistiky shodné. Zajímavé je, že s rostoucí velikostí náhodného výběru n klesá rozptyl odhadu T_2, limitně až k nule. Tato vlastnost se nám líbí...

Nový pojem: Konzistentní odhad

Mějme posloupnost odhadů $T_1(X_1), T_2(X_1, X_2), \ldots, T_n(X_1, X_2, \ldots, X_n)$ parametru ϑ. Tato posloupnost je konzistentní, pokud:

$$\lim_{n\to\infty} E[T_n] = \vartheta, \tag{1}$$

$$\lim_{n\to\infty} D[T_n] = 0. \tag{2}$$

Interpretační poznámka. Podmínce (1) se říká asymptotická nestrannost a je slabší, než nestrannost (z nestrannosti plyne asymptotická nestrannost, naopak to nefunguje). Podmínka (2) říká, že s rostoucí velikostí náhodného výběru n se odhad zpřesňuje.

Důležité tvrzení: Bodové odhady základních číselných charakteristik

Pro libovolně rozdělenou náhodnou veličinu X je:

- výběrový průměr $M = \sum_{i=1}^{n} X_i$ nestranný konzistentní odhad střední hodnoty $E[X]$.
- statistika $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$ nestranný konzistentní odhad rozptylu $D[X]$. Pro normální rozdělení jsou tyto odhady nejlepší.

Pro výpočet bodových odhadů parametrů libovolného rozdělení $L(\vartheta)$ máme např.: momentovou metodu a metodu maximální věrohodnosti.

I.III Bodové odhady parametrů základních rozdělení

Následující tvrzení shrnují nejčastěji používané bodové odhady parametrů vybraných rozdělení.

Důležité tvrzení: Nestranné a konzistentní odhady parametrů základních rozdělení

Následující odhady jsou konzistentní a nestranné.

Pro odhady parametrů normálního rozdělení $N(\mu, \sigma^2)$ používáme tyto statistiky:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2.$$

Pro odhad parametru Poissonova rozložení $Po(\lambda)$ používáme výběrový průměr, tj.

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Pro odhad parametru exponenciálního rozdělení $Exp(\lambda)$ používáme převrácenou hodnotu výběrového průměru, tj.

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}.$$

Pro odhad parametru p binomického rozdělení $Bi(m, p)$ se používá statistika

$$\hat{p} = \frac{1}{mn}\sum_{i=1}^{n} X_i.$$

Speciálně, pro případ $m = 1$ (tj. alternativní rozložení), je odhadem výběrový průměr.