Poznámky k předmětu Aplikovaná statistika, 7. téma

7. Náhodný výběr

Doposud jsme se setkali s úlohami patřícími do teorie pravděpodobnosti, kde rozdělení náhodné veličiny je přesně dáno. My pak zkoumáme jeho vlastnosti. Hod spravedlivou kostkou jsme modelovali náhodnou veličinou s diskrétním rovnoměrným rozdělením na množině $\{1, \dots, 6\}$. Představme si nyní situaci, že po nás protihráč chce, abychom hráli jeho kostkou, jejíž vlastnosti neznáme. Naše rozhodování ve hře je ale ovlivněno tím, je-li ona kostka spravedlivá, či nikoli. Proto je v našem zájmu zjistit, přibližně jaké rozdělení má hod touto kostkou. Zmíněná úloha již patří do matematické statistiky. Pokusíme se ji intuitivně řešit. Půjčíme si onu kostku a budeme s ní opakovaně házet a zaznamenávat počet ok padlých v každém hodu. Rozdělení každého hodu je stejné a hody se vzájemně neovlivňují. Proto můžeme (viz statistickou definici pravděpodobnosti) např. pravděpodobnost, že na kostce padne jednička, odhadnout počtem jedniček v našich hodech ku celkovému počtu hodů. Získáme tak odhad rozdělení hodu kostkou protihráče, kterým se můžeme při hře řídit.
Obecněji, provádíme-li nějaký experiment opakovaně, zajímá nás, co lze na základě naměřených dat o experimentu říci. Modelujeme-li výsledek experimentu hodnotou náhodné veličiny X, hodila by se nám znalost rozdělení této náhodné veličiny, což ovšem v praxi není možné. Proto se snažíme z naměřených dat odhadnout rozdělení náhodné veličiny X, anebo aspoň některé jeho charakteristiky.
Pro níže uvedený postup jsou nezbytné (a vlastně i přirozené, jak jsme již viděli) požadavky na experiment:

- experiment lze provádět opakovaně při stejných podmínkách,
- jednotlivé experimenty se neovlivňují.

To vše nás povede k následující definici. Modelujme experiment náhodnou veličinou X a proveďme ho n–krát.

Definice 1. Náhodným výběrem z rozdělení náhodné veličiny X nazveme n–tici nezávislých stejně rozdělených náhodých veličin $X_1, \dots, X_n$ pocházejících z rozdělení náhodné veličiny X.

Realizace náhodného výběru z rozdělení náhodné veličiny X je n naměřených hodnot $x_1, \dots, x_n$ získaných při opakování experimentu.

ZÁKLADNÍ VÝBĚROVÉ CHARAKTERISTIKY náhodného výběru $X_1, \dots, X_n$

(i) Výběrový průměr

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(ii) Výběrový rozptyl

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad \left(= \frac{1}{n-1} \Big(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \Big) \right)$$

(iii) Výběrová směrodatná odchylka

$$S = \sqrt{S^2}$$

VLASTNOSTI VÝBĚROVÝCH CHARAKTERISTIK

Buď $X_1, \ldots, X_n$ náhodný výběr z rozdělení X, kde

$$\mathbb{E}X_i = \mathbb{E}X = \mu,\ \mathrm{var}X_i = \mathrm{var}X = \sigma^2,\ i = 1, \ldots, n.$$

Pak

$$\mathbb{E}\bar{X} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \mu.$$

Obdobně lze odvodit

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

Proto lze za odhad

- střední hodnoty μ vzít $\hat{\mu} = \bar{X}$,
- rozptylu σ^2 vzít $\hat{\sigma}^2 = S^2$.

Poznámka 1. *Mějme realizaci $x_1, \ldots, x_n$ náhodného výběru pro velké n. Pak lze očekávat, že realizace*

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *náhodné veličiny $\bar{X}$ se bude pohybovat kolem μ,*
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$ *náhodné veličiny S^2 se bude pohybovat kolem σ^2.*

Lze to vysvětlit tím, že rozptyly $\bar{X}$ a S^2 jdou k nule, pokud n jde do nekonečna.

Při sběru dat z experimetu (zvláště diskrétní povahy) se často používá pojem

SETŘÍDĚNÁ DATA

Jde o uložení dat ve formě tzv. tabulky četností. Buď $x_1, \ldots, x_n$ realizace náhodného výběru z **diskrétního** rozdělení, kde hodnoty y_i se v realizaci opakují n_i–krát, $i = 1, \ldots, k$, a $\sum_{i=1}^{k} n_i = n$. Číslo n_i se nazývá četnost hodnoty y_i a k počet tříd.

Příklad 1. *Stokrát opakovaným hodem spravedlivou šestistěnnou kostkou jsme získali data, která jsme zapsali do tabulky četností*

y_i	1	2	3	4	5	6
n_i	10	17	20	18	22	13

Data lze znázornit též graficky pomocí tzv. diagramu četností.

[1] Poněkud obtížnější je spočítat rozptyl S^2,

$$\mathrm{var}S^2 = \frac{\sigma^4}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right),$$

kde $\mu_4 = \mathbb{E}(X - \mathbb{E}X)^4$ je tzv. čtvrtý centrální moment.

V tomto případě je výhodnější pro výpočet $\bar{x}$ a s^2 použít vzorce

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i y_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (y_i - \bar{x})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{k} n_i y_i^2 - n\bar{x}^2\Big).$$

I data pocházející ze **spojitého** rozdělení se někdy setřiďují např. do k tříd. Pak se jako zástupce y_i i–té třídy často volí střed i–té třídy.

Příklad 2. *Byly změřeny koncentrace vzorků jisté kyseliny. Naměřené údaje jsou uvedeny v následující tabulce četností*

x	$5,5$–$6,5$	$6,5$–$7,5$	$7,5$–$8,5$	$8,5$–$9,5$	$9,5$–$10,5$	$37,5$–$38,5$
y_i	6	7	8	9	10	38
n_i	3	7	20	10	9	1

Obdobou diagramu četností je ve spojitém případě histogram, což je odhad hustoty rozdělení náhodné veličiny, ze kterého data pocházejí.

Nejdůležitější pro nás bude
NÁHODNÝ VÝBĚR POCHÁZEJÍCÍ Z NORMÁLNÍHO ROZDĚLENÍ
$X \sim \mathcal{N}(\mu, \sigma^2)$, kde μ, σ^2 jsou neznámé parametry. Pak pro

(i) výběrový průměr $\bar{X}$ platí

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

a jako důsledek platí

$$\frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1),$$

(ii) pro výběrový rozptyl platí

$$\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2(n-1), \ n > 1,$$

(iii) $\bar{X}, S^2$ jsou nezávislé,

(iv) $\bar{X}$ a S^2 platí

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \left(= \frac{\frac{\bar{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}}} \right) \sim t(n-1), \ n > 1.$$

Cvičení k tomuto tématu

(i) Spočtěte hodnoty realizací náhodných veličin $\bar{X}$ a S^2 v Příkladě 1.

(ii) Spočtěte $\bar{x}$ a s^2 v Příkladě 2. Jak se úloha změní, pokud bychom neuvažovali naměřenou odlehlou koncentraci příslušnou třídě $37,5$–$38,5$?

(iii) Nechť $X_1, \ldots, X_9$ jsou nezávislé stejně rozdělené náhodné veličiny s rozdělením $\mathcal{N}(1, 16)$. Určete

(a) $\mathbb{P}(\bar{X} > 5)$,

(b) $\mathbb{P}\big(|\bar{X} - 2| < 2\big)$.

(iv) Určete hodnotu m, kterou náhodná veličina $X \sim \mathcal{N}(1, 4)$ překročí s pravděpodobností

(a) $0,4$;

(b) $0,8$.

(v) Buď $X \sim \mathcal{N}(\mu, \sigma^2)$. Spočítejte pro $k = 1, 2, 3$

$$\mathbb{P}(\mu - k\sigma < X < \mu + k\sigma).$$

(vi) Proveďme náhodný výběr z rozdělení $\mathcal{N}(2, 9)$ o rozsahu $n = 16$. Stanovte hodnotu h tak, aby pravděpodobnost, že výběrový rozptyl S^2

(a) nepřekročí h, byla $0,05$,

(b) překročí h, byla $0,05$.

(dcv) Byl proveden průzkum počtu dětí v 99 náhodně zvolených rodinách. Spočtěte hodnoty výběrového průměru a výběrové směrodatné odchylky pro tento průzkum, víte-li

počet dětí	0	1	2	3	4	5	6	7	8
četnost	20	29	30	12	3	4	0	0	1