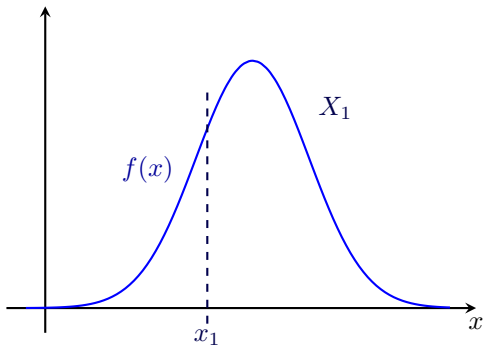
Náhodný výběr

Aplikovaná statistika
16. dubna 2024

Jaroslav Schmidt

Sadě n náhodných veličin $X_1, \ldots, X_n$, které mají všechny shodné rozdělení, říkáme náhodný výběr. Velkým písmenem X značíme, že se stáe jedná o náhodnou veličinu. Občas je totiž zaměňována se sadou $x_1, \ldots, x_n$ kterou označujeme malými písmeny a míníme tím konkrétní realizaci náhodného výběru.

Představme si to například následovně. Zkoumáme průměrnou výšku populace. Vydáme se do terénu a zeptáme se první oběti na jeho výšku. Než nám odpoví, dá se to vnímat tak, že jeho výška je neznámá veličina X_1. Jeho výška je tedy prozatím dána jakousi hustotou pravděpodobnosti $f(x)$. To je v obrázku 1 označeno modrou křivkou. Jeho výška je tedy prozatím jen funkcí nabývající několikonečně mnoha hodnot.



Obrázek 1: Náhodná veličina X_1 a její konkrétní realizace x_1.

Poté co nám prozradí svoji skutečnou výšku, dostáváme to, čemu říkáme realizace náhodné veličiny. Získáváme číslo x_1. Číslo si zaznamenáme a pokračujeme k další oběti. Opět neznáme její výšku, tedy jedná se opět o náhodnou veličinu X_2 reprezentovanou stejnou funkcí hustoty náhodné veličiny $f(x)$ a poté co nám prozradí svoji výšku, bude se jednat o další realizaci náhodné veličiny x_2. Předpokládejme na chvilku, že známe funkci $f(x)$, potom pro každou náhodnou veličinu X_i můžeme spočítat střední hodnotu a rozptyl, označme jej μ a σ^2. Co když chceme vyšetřovat průměrnou hodnotu? Tedy hodnotu

$$\bar{X} = \frac{\sum_i X_i}{n}. \tag{1}$$

Je dobré podotknout, že průměrujeme X_i nikoliv x_i. Jedná se tedy o náhodné veličiny, jejichž hodnotu zatím neznáme, může nabývat různých hodnot s různou pravděpodobností. Díky tomu je i veličina $\bar{X}$ náhodnou veličinou se svojí hustotou pravděpodobnosti. Nejedná se tedy o jediné číslo. Naproti tomu $\bar{x} = \sum_i x_i/n$ je prosté číslo a jedná se o pouhý aritmetický průměr naměřených hodnot. Už by nám měly být známé tyto dvě pravidla pro výpočet středních hodnot a rozptylů:

$$\mathbb{E}\left(aX + bY\right) = a\mathbb{E}\left(X\right) + b\mathbb{E}\left(Y\right), \tag{2}$$

$$\operatorname{var}\left(aX + bY\right) = a^2 \operatorname{var}\left(X\right) + b^2 \operatorname{var}\left(Y\right). \tag{3}$$

Ty můžeme ihned použít pro výpočet střední hodnoty a rozptylu naší nové veličiny $\bar{X}$ jako

$$\mathbb{E}\left(\frac{X_1+\cdots+X_n}{n}\right)=\frac{\mathbb{E}\left(X_1\right)+\cdots+\mathbb{E}\left(X_n\right)}{n}=\frac{n\mu}{n}=\mu \tag{4}$$

$$\operatorname{var}\left(\frac{X_1+\cdots+X_n}{n}\right)=\frac{\operatorname{var}\left(X_1\right)+\cdots+\operatorname{var}\left(X_n\right)}{n^2}=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n} \tag{5}$$

Hodí se i vědět, jaké má tato náhodná veličina $\bar{X}$ rozdělení. To ovšem není triviální spočítat a navíc k tomu potřebujeme znám funkci $f(x)$, kterou často dopředu neznáme. Naštěstí nás zachraňuje centrální limitní věta. Tu zde nebudeme odvozovat, jen ji uvedeme.

Ta říká, že bez ohledu na to, z jakého rozdělení pochází náhodné veličiny X_i, jejich průměrná hodnota $\bar{X}$ je asymptoticky normálně rozdělená. Asymptoticky v tom smyslu, že plná platnost nastává jen při $n \to \infty$. Pro menší n je tato platnost jen přibližná. A jelikož už známe střední hodnotu i rozptyl náhodné veličiny $\bar{X}$, známe její rozdělení

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right). \tag{6}$$

Připomeňme, že toto platí přesně jen asymptoticky. Tento výsledek zaslouží komentář. Povšimněme si výsledného rozptylu. Pokud budeme n zvětšovat nade všechny meze, potom se rozptyl bude blížit nule, tedy

$$n \to \infty \quad \Rightarrow \quad \bar{X} \sim \mathcal{N}\left(\mu, 0\right). \tag{7}$$

Tento výsledek znamená, že pokud budeme mít výběr o velikosti $n = \infty$, potom $\bar{X}$ již nebude náhodná veličina a bude přímo platit $\bar{X} = \mu$