Náhodný výběr a statistika

I Náhodný výběr

Začněme příkladem. Chceme získat informace o platech v ČR, např. průměrný plat nebo nás zajímají rozdíly mezi platy mužů a žen. Za tímto účelem je potřeba sesbírat data - ptát se jednotlivých občanů na jejich plat. Samozřejmě, že když se zeptáme všech, budeme naprosto přesní a budeme mít kompletní informace. Jenže tento postup je dost nákladný a dost obtížně proveditelný. Proto z celku všech lidí zvolíme náhodně pouze některé – čímž získáme náhodný výběr.

Nový pojem: Náhodný výběr

Náhodný výběr je uspořádaná n-tice náhodných veličin $X_1, X_2, \ldots, X_n$, které jsou stochasticky nezávislé a mají stejné rozdělení.

Upořádaná n-tice proto, abychom ji mohli zapsat do sloupcového vektoru (pokud se jedná o vícerozměrné rozdělení, tak do matice). Realizací náhodného výběru jsou pak konkrétní hodnoty, které značíme malými písmeny $x_1, x_2, \ldots, x_n$. Rozsah náhodného výběru je n.

Interpretační poznámka. Rozdělení, které mají náhodné veličiny $X_1, X_2, \ldots, X_n$ musí být stejné, ale nepotřebujeme jej konkrétně znát. Důležitý předpoklad je dostatečná nahodilost – zkoumáme-li výšku příjmů vysokoškoláků, nestačí poptat se svých spolužáků. Je třeba zahrnout všechny vysoké školy nebo změnit studii na „Výška příjmů mých spolužáků“.

Podle toho, na jaké otázky hledáme odpověď, můžeme rozlišit dva směry statistiky:

- Statistická indukce (inference) – podle informací z náhodného vyběru vyslovíme závěry o celém základním souboru, odhadneme parametry rozdělení, ze kterého data pochází apod. V našem příkladě přejdeme od průměrného platu v získaném výběru k střední hodnotě platu v celé populaci.
- Testování hypotéz – hledáme odpovědi na otázky, které se váží k populaci, například „Je průměrný plat v ČR vyšší, než 25 000 Kč?“ nebo „Liší se průměrný plat mužů a žen v ČR?“.

Protože ale pracujeme s náhodnými výběry a ne celým základním souborem informací, dopouštíme se chyby. Pro statistiku je důležité umět tuto chybu kvantifikovat – takto umíme určit, která metoda je lepší. Obvykle chyba, se kterou pracujeme klesá s rostoucí velikostí náhodného výběru n. V praxi je důležitý výpočet velikosti náhodného výběru (sample size estimation) při návrhu studie. Náhodný výběr musí být dostatečně velký, aby riziko chyby bylo malé (obvykle 5%), na druhou stranu získat pozorování může být dosti nákladné.

II Statistika?

Slovo statistika má dva významy. První, známější, je věda, která získává informace z dat. Druhý význam, se kterým budeme často pracovat, je statistika jakožto funkce.

Nový pojem: Statistika

Statistika je libovolná funkce náhodného výběru.

Interpretační poznámka. To znamená, že je to nějaká formulka, do které dosadíme hodnoty, které jsou obsaženy v náhodném výběru. Takže se tam vyskytují náhodné veličiny X_i (při samotném výpočtu dosazujeme ale realizace x_i – naměřené hodnoty) a možná taky n – velikost náhodného výběru.

Dále je uveden seznam některých často používaných statistik (jejich podrobnému vysvětlení jsou věnovány samostatné materiály).
Rozmyslete si, že každá z níže uvedených náhodných veličin je statistika, tedy funkce náhodného výběru.
V následující tabulce vždy uvažujeme náhodný výběr $X_1, X_2, \ldots, X_n$, rozsah je tedy n.

název	značení	výpočet
výběrový průměr	M	$M = \sum_{i=1}^{n} \frac{X_i}{n}$
medián	$X_{0,5}$	Prostřední hodnota uspořádaného náhodného výběru pro n liché, jinak aritmetický průměr dvou prostředních hodnot.
modus	$\hat{X}$	Nejčetnější hodnota náhodného výběru.
výběrový rozptyl	S^2	$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M)^2$
výběrová směrodatná odchylka	S	$S = \sqrt{S^2}$
výběrová kovariance	S_{12}	$S_{12} = \frac{1}{n-1} \sum_{i=1}^{n} ((X_i - M_1)(Y_i - M_2))$
výběrový korelační koeficient	R_{12}	$R_{12} = \frac{S_{12}}{S_1 S_2}$
empirická distribuční funkce v bodě x	$F(x)$	$F(x) = \frac{\text{card}\{i, X_i \leq x\}}{n}$