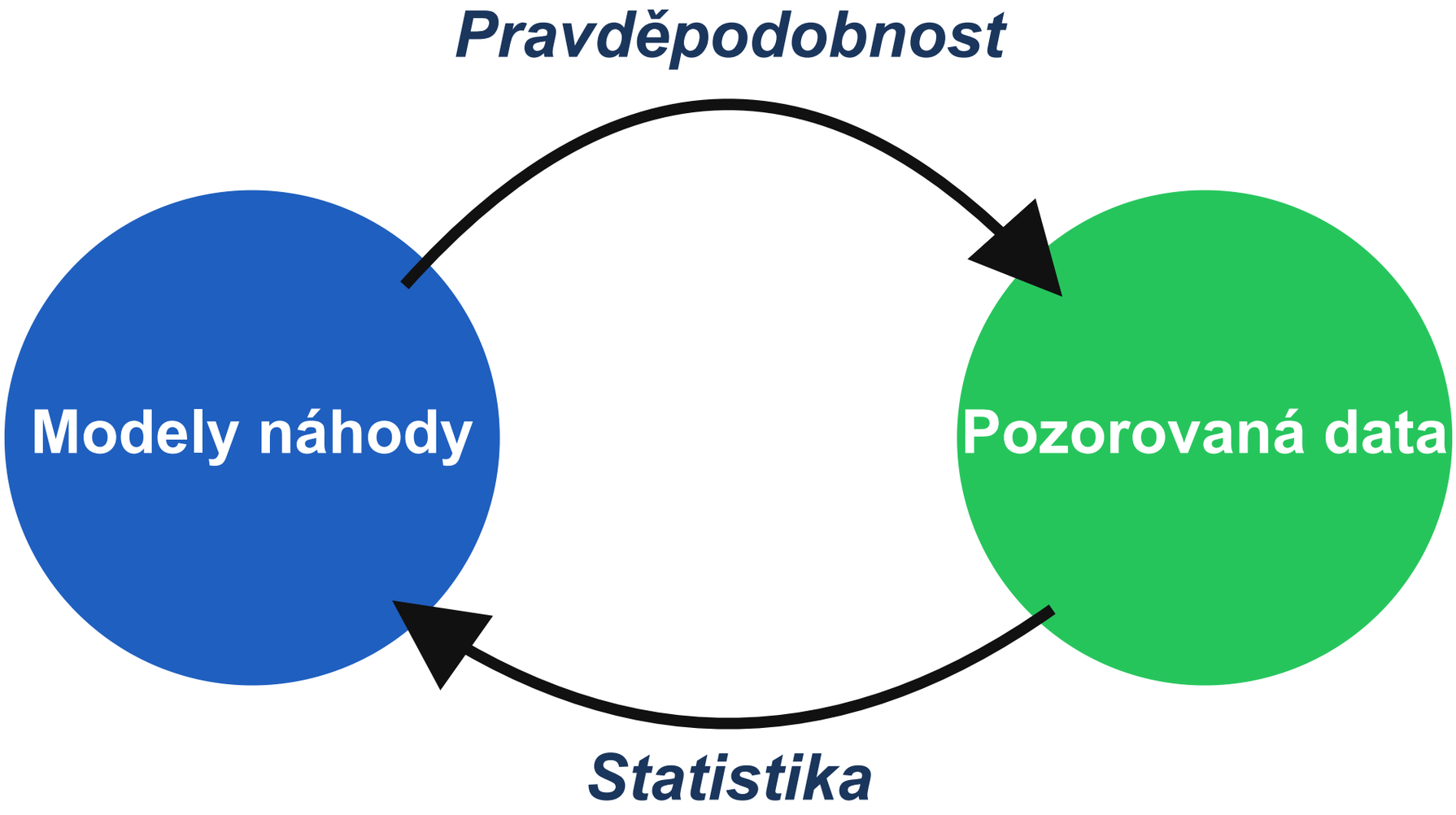
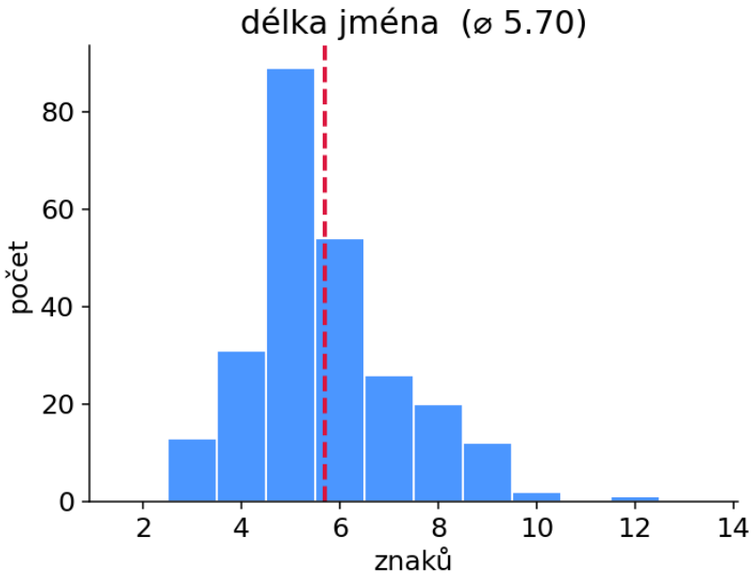
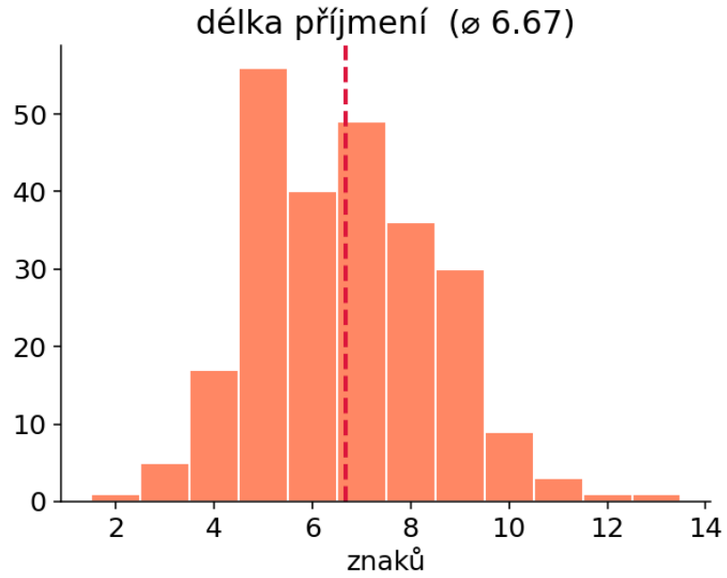
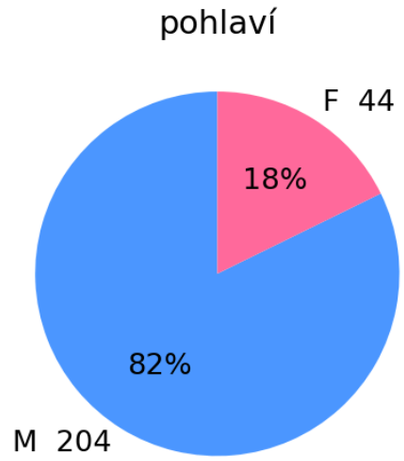
Plán přednášky



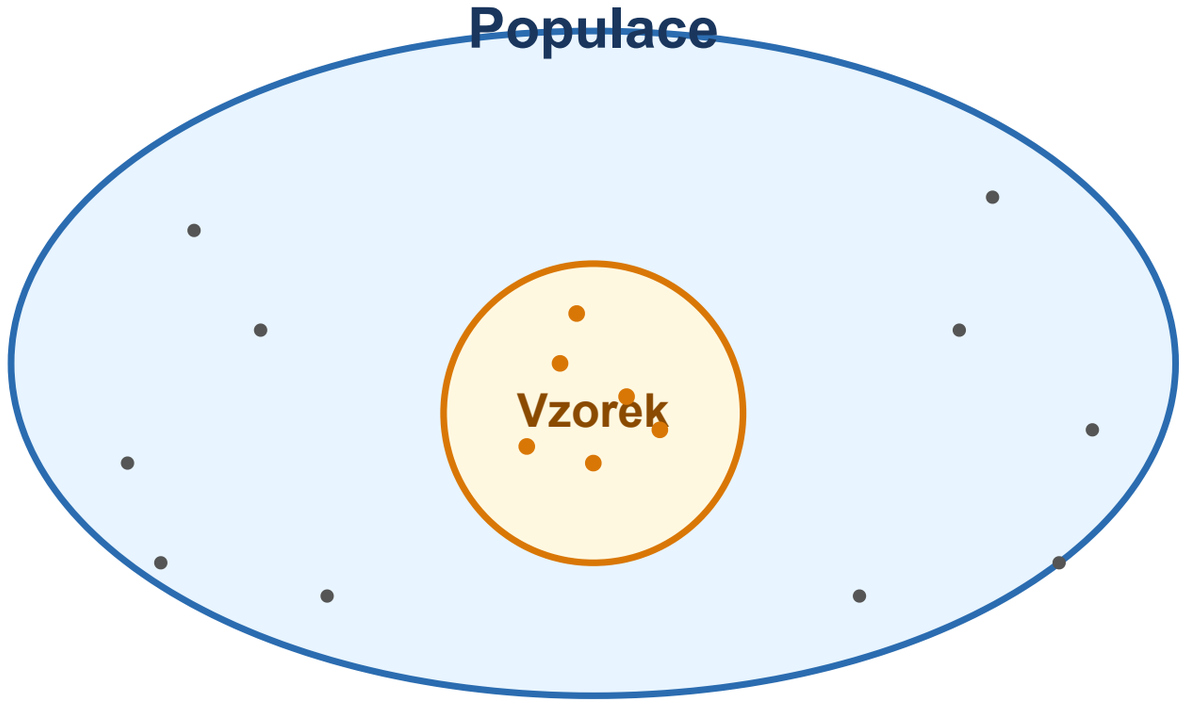
Data z tohohle předmětu





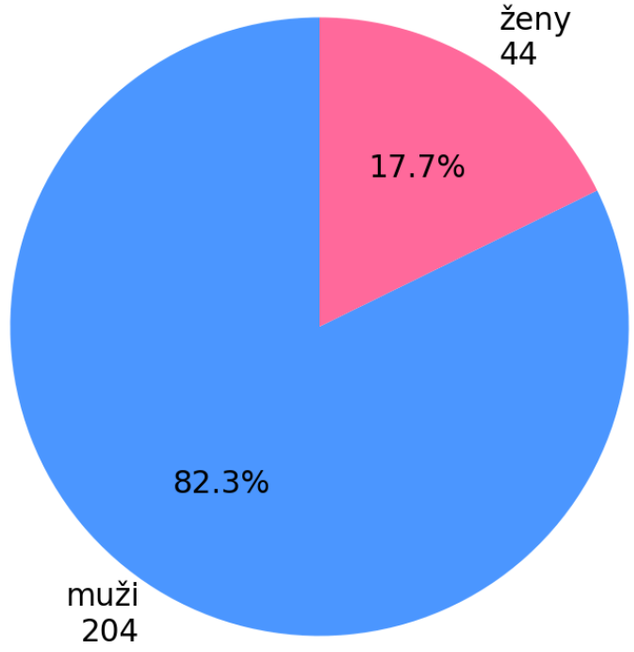


Kdo je populace?



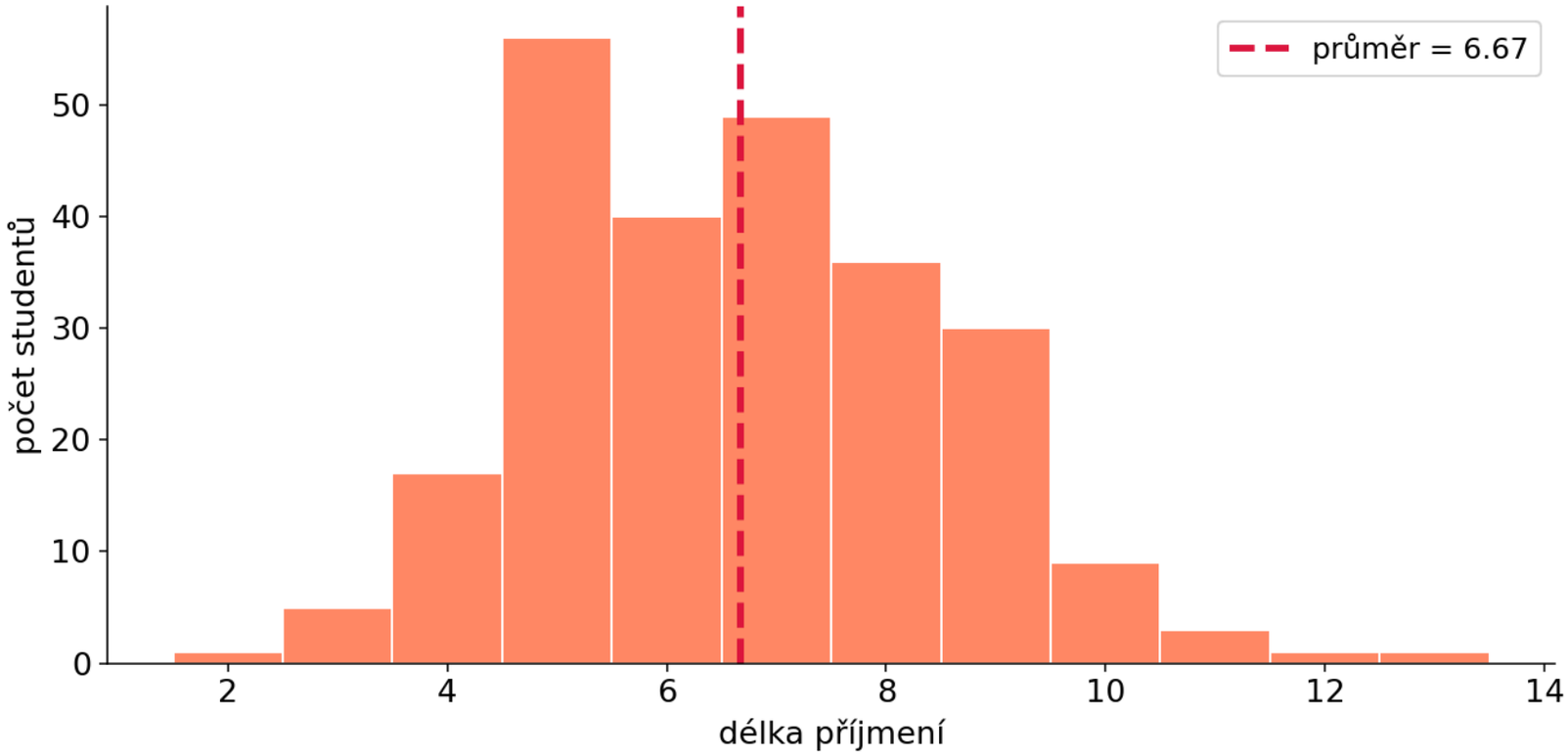
předpoklad: **vzorek je uniformní výběr z populace**

Podíl pohlaví



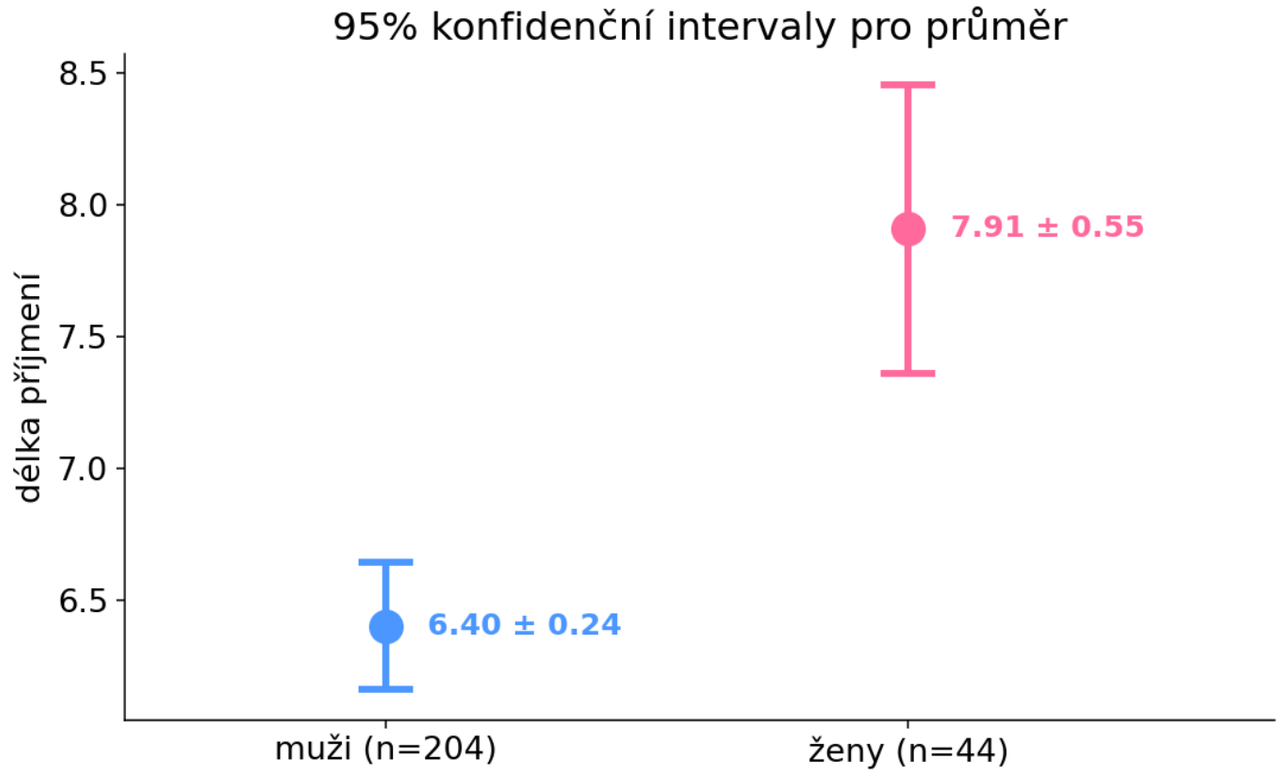
> "All models are wrong, but some are useful. "

Průměrná délka příjmení



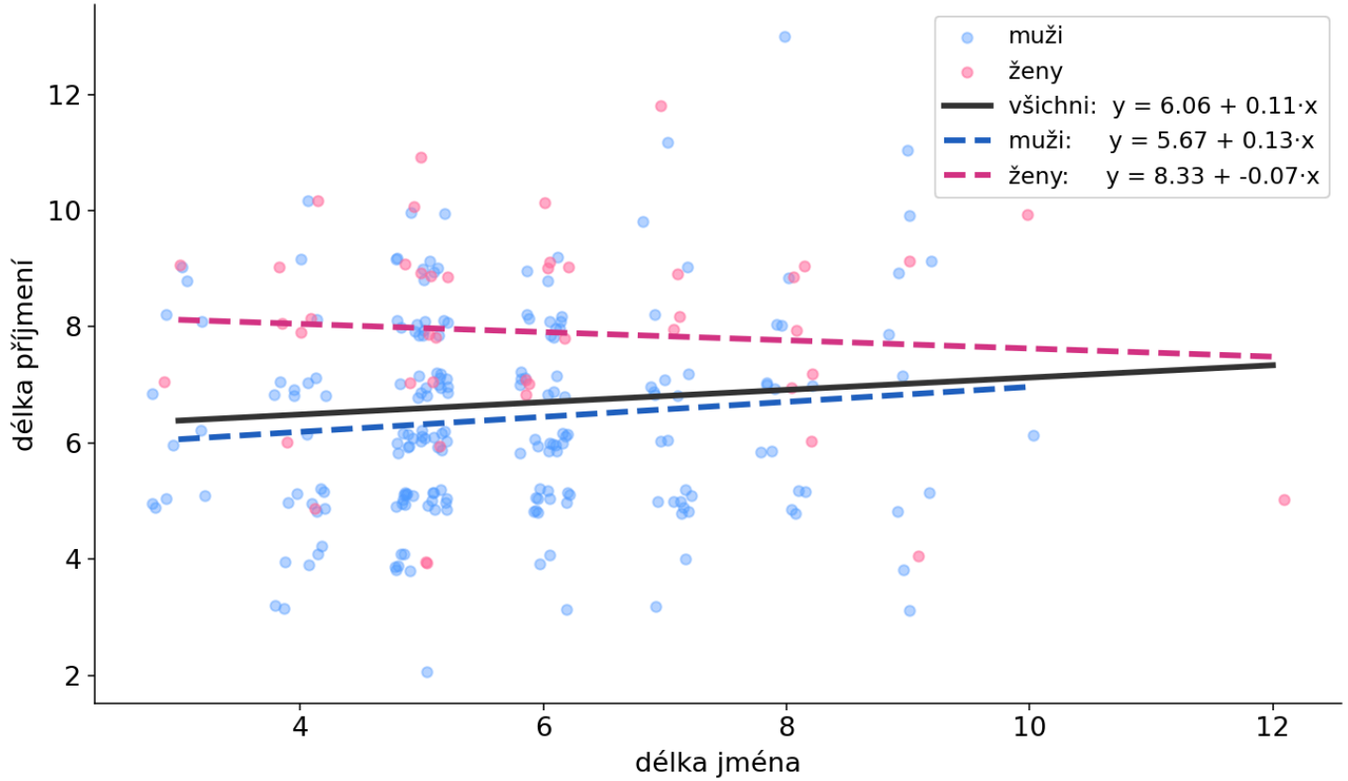
Jak přesný je průměr 6.67?

Muži vs. ženy



Je rozdíl 1.5 znaků reálný, nebo náhoda?

Souvisí délka jména a příjmení?



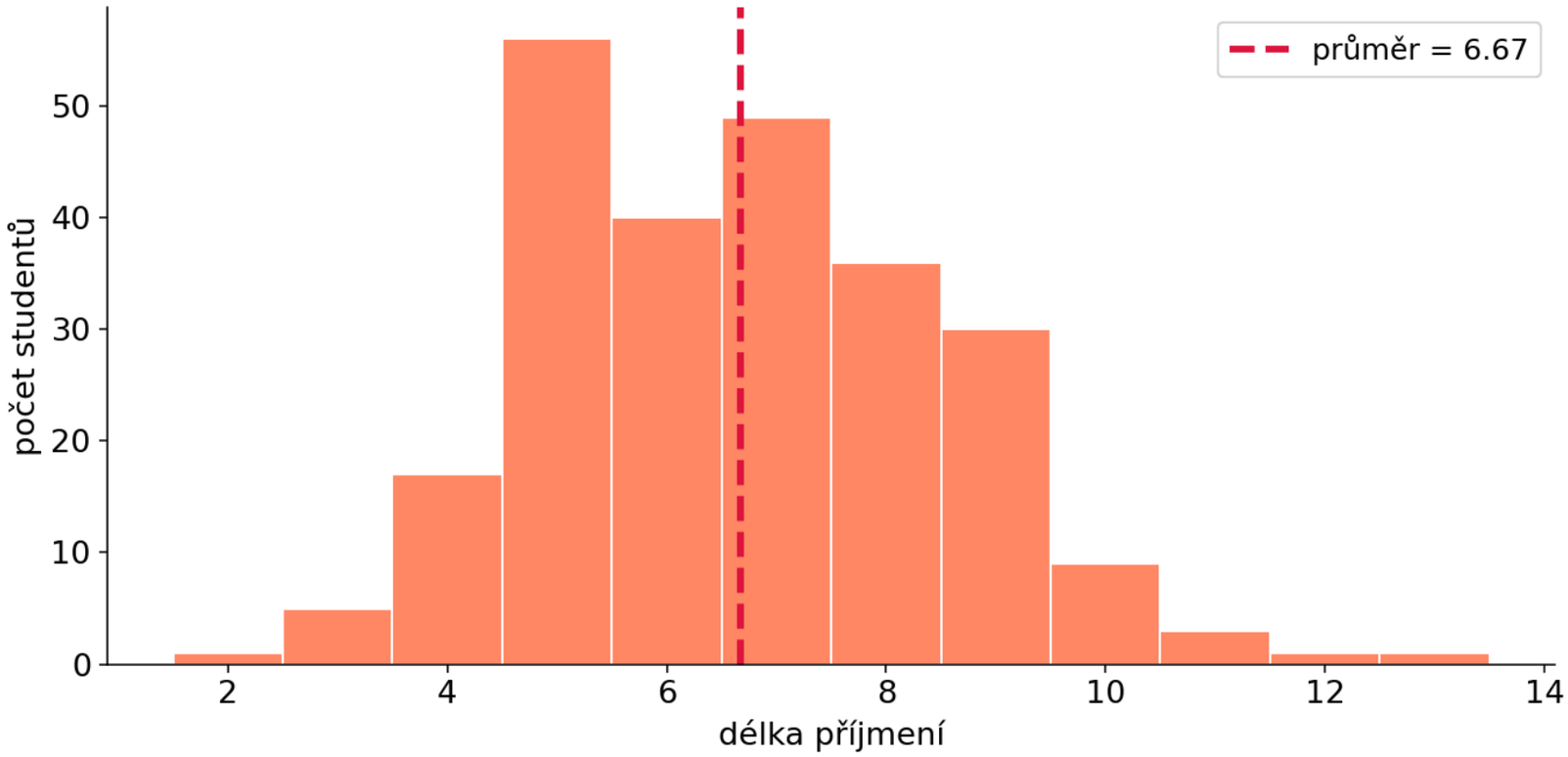
Overfitting / p-hacking / data dredging

Začátek: bodové odhady

Technika #1: Maximum likelihood estimation (maximální věrohodnost)

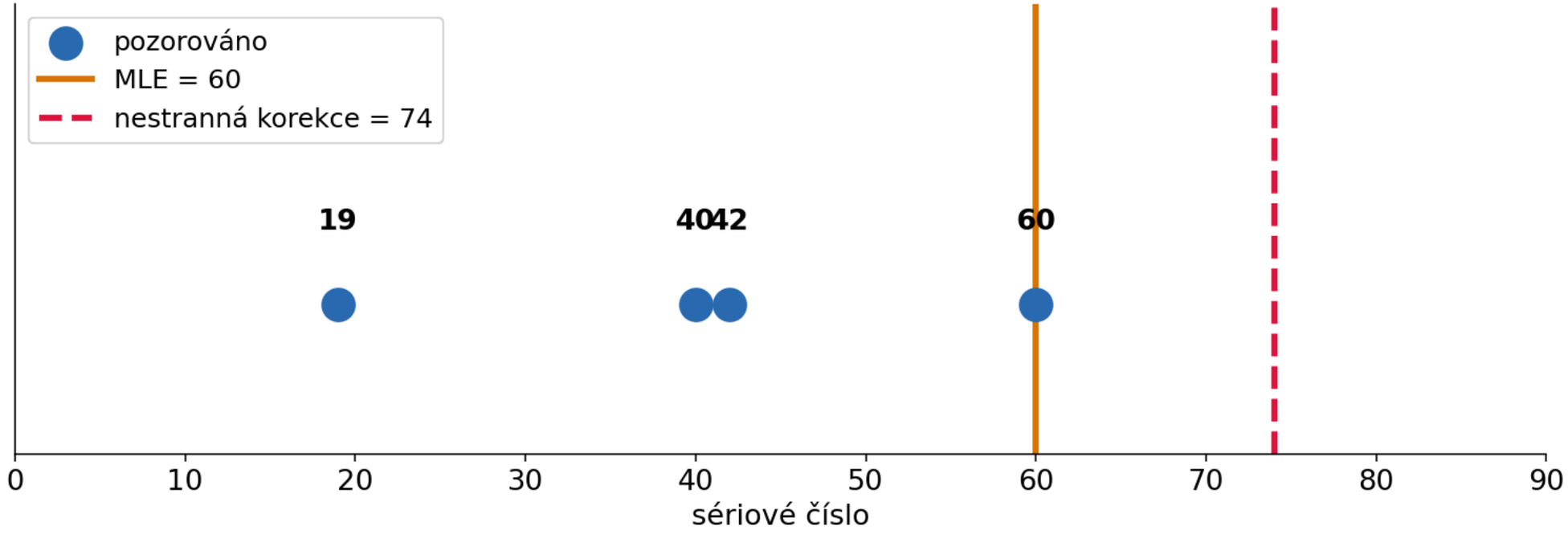
Technika #2: Unbiased estimators (nestranný estimátor)

Průměrná délka příjmení

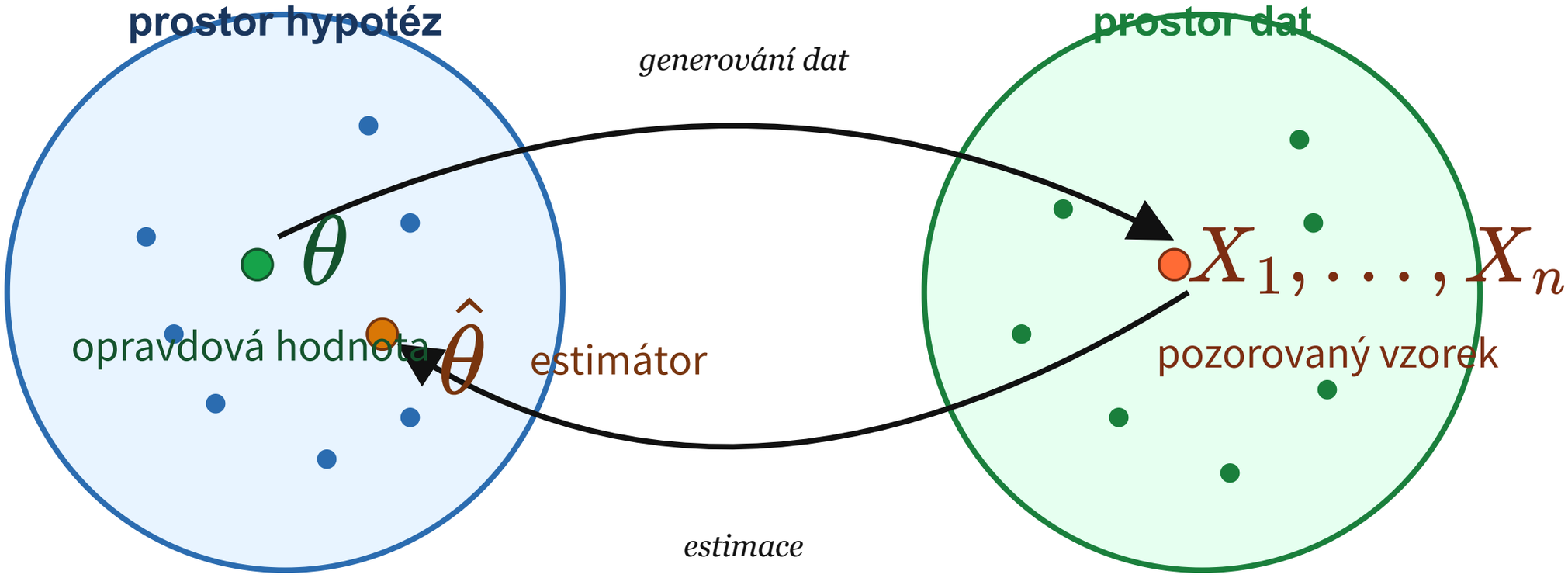


Kolik tanků vyrobili Němci?

19 40 42 60



Začátek: bodové odhady



z neznámého θ vzniká vzorek; z něj uhodneme $\hat{\theta} \in \Theta$

Dva vzorečky pro rozptyl

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{vs.} \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Proč dva? Co každý optimalizuje?

Jak trénovat ML model?

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log p_\theta(y_i \mid x_i)$$

Odkud se bere vzoreček?