Applied Statistics

Sofia Triantafillou

sof.triantafillou@gmail.com

University of Crete
Department of Mathematics and Applied Mathematics

Lecture Summary

- Confidence Intervals
- Bootstrap Confidence Intervals (material on the website).

Confidence Intervals for the mean of Normal

- Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.
- I know that

$$U = \frac{n^{1/2}(\overline{X}_n - \mu)}{\sigma'} \sim t_{n-1}, \text{ where}$$

$$\sigma' = \left(\frac{\sum_{n=1}^{n}(X_i - \overline{X}_n)^2}{n}\right)^{1/2}$$

- I can compute $P(-c < U < c)$.
- I can compute

$$P(\overline{X}_n - \frac{c\sigma'}{n^{1/2}} < \mu < \overline{X}_n + \frac{c\sigma'}{n^{1/2}})$$

- $(\overline{X}_n - T_{n-1}^{-1}(\frac{\gamma+1}{2})\frac{\sigma'}{n^{1/2}}, \overline{X}_n + T_{n-1}^{-1}(\frac{\gamma+1}{2})\frac{\sigma'}{n^{1/2}})$ is the $\gamma 100\%$ confidence interval for μ.

Confidence Intervals for the mean of Normal

- Sometimes we use the notation $1-\alpha$ confidence interval, so $a/2$ is the probability mass on each side of the interval.
- $(\overline{X}_n - T_{n-1}^{-1}(1-a/2)\frac{\sigma'}{n^{1/2}}, \overline{X}_n + T_{n-1}^{-1}(1-a/2)\frac{\sigma'}{n^{1/2}})$ is the $(1-a)100\%$ confidence interval for μ.

Confidence Intervals: Interpretation

- After observing our sample, we find that (a, b) is our 95%-CI for μ.
- This does not mean that $P(a < \mu < b) = 0.95$. In fact, we can not make such statements if we consider μ to be a number (frequentist view).
- We can think of our interpretation as repeated samples.
  - Take a random sample of size n from $\mathcal{N}(\mu, \sigma^2)$.
  - Compute (a, b).
  - Repeat many times.
  - There is a 95% chance for the random intervals to include the value of μ.

Computing Confidence Intervals

In python:

- Sample $n = 10$ data points from a $\mathcal{N}(3, 2^2)$ distribution.
- Compute the 90% confidence interval for the mean μ.
- Repeat 100 times.
- How many times does the interval include the true mean?
- Repeat with $\sigma^2 = 4^2$. What happens to the CIs?
- Repeat with $n = 50$. What happens to the CIs?

The Bootstrap

What if we do not know the distribution of X_i?

- Data $x_1, x_2, \ldots x_n, \sim F$ with true mean μ.
- F^*: empirical distribution (resampling distribution).
- $x_1^*, x_2^*, \ldots, x_n^*$ resample same size data.

Example: Die rolling: $3, 1, 2, 4, 3, 2, 1, 6, 1, 6$

F	1/6	1/6	1/6	1/6	1/6	1/6
F^*	0.3	0.2	0.2	0.1	0	0.2

Bootstrap Confidence Intervals

- $\delta^* = \overline{x}_n^* - \overline{x}_n$.
- $\delta = \overline{x}_n - \mu$.

The bootstrap principle: The distribution of δ is the approximately the same as the distribution of δ^*.

- You want to find an $1 - a$ confidence interval for μ.
- Let q_β^* be the β sample quantile of the empirical distribution of δ^*
- $q_{a/2}^* \leq \overline{x}_n - \mu \leq q_{1-a/2}^*$
- $\overline{x}_n - q_{1-a/2}^* \leq \mu \leq \overline{x}_n - q_{a/2}^*$

Empirical Bootstrap Confidence Intervals

- Data $x_1, x_2, \ldots x_n$ draw from a distribution F.
- Use the data to estimate the variation of the estimates (based on the data).
  - Generate many bootstrap samples $x_1^*, x_2^*, \ldots, x_n^*$.
  - Compute the statistic θ^* for each sample.
  - Compute the bootstrap difference $\delta^* = \theta^* - \hat{\theta}$.
  - Use the quantiles q^* of δ^* to approximate the quantiles of $\delta = \hat{\theta} - \theta$.
  - Compute a confidence interval $(\hat{\theta} - q^*_{1-a/2}, \hat{\theta} - q^*_{a/2})$

Computing bootstrap Confidence Intervals

In python:

- Sample $n = 10$ data points from a $\mathcal{N}(3, 2^2)$ distribution.
- Compute bootstrap confidence intervals for 100 bootstrap samples.
- Repeat 100 times.
- How many times does the interval include the true mean?
- Repeat with $\sigma^2 = 4^2$. What happens to the CIs?
- Repeat with $n = 50$. What happens to the CIs?

Example

Data on calorie content in 20 different beef hot dogs from Consumer Reports (June 1986 issue):

$$186, 181, 176, 149, 184, 190, 158, 139, 175, 148,$$

$$152, 111, 141, 153, 190, 157, 131, 149, 135, 132$$

- Compute bootstrap intervals for the mean.
- (a) using the normality assumption.
- (b) usign bootstrap.