Universität Bielefeld
Tanja Pasurek
Ehsan Abedi

WiSe 2024/25

Foundations of Statistics

Homework 11

Topic I: Confidence intervals for parameters of Bernoulli distribution

Exercise 1. Given $n \in \mathbb{N}$, let $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Ber}(p)$ be a random sample with unknown parameter $p \in [0, 1]$. We know that $\sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$. Recall that both ML and MM estimators for p is the sample mean $\hat{p} = \overline{X}_n$. When n is large, an application of the CLT gives:

$$\begin{aligned} 1 - \alpha &\approx \mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2}\right) \\ &= \mathbb{P}\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right), \end{aligned} \tag{1}$$

where $z_{1-\alpha/2}$ is $(1-\alpha/2)$-quantile of the standard normal distribution $\mathcal{N}(0, 1)$.

(a) As discussed in Ch. 4.2, the formula above does not yet provide a confidence interval (CI) because the unknown parameter p is still present in the bounds (1). An *approximate* confidence interval for p can be obtained by *plugging-in* $\hat{p}$ for p, which then gives the bounds

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right). \tag{2}$$

However, it is known that for p close to 0 and 1, this approximation may not be sufficiently accurate. The goal is to check this using a simulation in `R`.

Let $n = 60$ and assume we know the unknown parameter, say $p = 0.02$, chosen here to be close to 0. Run a simulation in `R` with $N = 10000$ experiments to estimate the coverage probability above for $\alpha = 0.05$.

(b) Let's visualize the first 30 confidence intervals obtained in (a). To this end, use `plotCI()` in the package `plotrix`. To your plot, add the horizontal line representing the true parameter $p = 0.02$.

(c) The *Wilson confidence interval*, which was proposed in 1927, aims at circumvent this issue.

To derive the formula for $(1-\alpha)\,100\%$ CI for p, we do not plug in $\hat{p}$ for p; insted we solve both inequalities in (1) for p. Show that the resulting confidence interval is

$$\left(\frac{\hat{p} + \frac{z_{1-\alpha/2}^2}{2n} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{z_{1-\alpha/2}^2}{n}}, \frac{\hat{p} + \frac{z_{1-\alpha/2}^2}{2n} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{z_{1-\alpha/2}^2}{n}} \right). \tag{3}$$

(d) Run a simulation in `R` to find the coverage probability for this improved CI for $n = 60$, $\alpha = 0.05$, and parameter $p = 0.02$. Compare the results with (a).

Exercise 2. Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Ber}(p)$ represent the outcome of n flipping a Dutch 1 Euro coin ($X_i = 1$ corresponds to getting "heads" and $X_i = 0$ corresponds to getting "tails" of the i-th experiment.) We assume the coin is fair, that is, the unknown parameter is about $p \approx 0.5$. Suppose we want to make a 95% confidence interval for p, and it should be at most $w = 0.01$ wide. The goal of this exercise is to determine the required sample size n.

(a) Under the assumption above, derive the width of the 95% CI for p of both methods (2) and (3) in Exercise 1. Plot these widths in `R` as a function n.

(b) Determine how large n should be so that $w = 0.01$ for both methods. Mark these points in your plot.

(c) The coin is thrown the number of times computed in (b) for the classical method (2), resulting in 19 477 times heads. Construct the 95% CI. Compare the results with `R` function `prop.test`.

Topic II: Confidence intervals for parameters of normal distribution

Exercise 3. Consider the following sampling of a normal distribution

34.40, 37.70, 55.59, 40.71, 41.29, 57.15, 44.61, 27.35, 33.13, 35.54,
52.24, 43.60, 44.01, 41.11, 34.44, 57.87, 44.98, 20.33, 47.01, 35.27

(a) Calculate a 99% confidence interval for μ with given $\sigma^2 = 100$.

(b) Calculate a 99% confidence interval for μ with unknown σ.

(c) Use the `R` function `t.test()` to check your results in (b).

(d) Now, we require that the lengths of the confidence intervals in (a) and (b) be less than or equal to 5. What sample size is necessary in each case?

Exercise 4. Let $X_1, ..., X_n$ be a random sample from a normal distribution with mean μ_1 and variance σ_1^2, and let $Y_1, ..., Y_m$ be a random sample from a normal distribution with mean μ_2 and variance σ_2^2. Moreover, the two samples are independent. Assuming that both σ_1^2 and σ_2^2 are known, construct the exact $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$.

Hint: Check that

$$\overline{X}_n - \overline{Y}_m \overset{\mathrm{d}}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right).$$