MS-C1620 Statistical Inference

Exercise 5

Homework exercise

To be solved at home before the exercise session.

---

1. a. A simple sample size calculation can be performed for binary proportion confidence intervals as follows. We bound the standard deviation estimate from above as $\sqrt{\hat{p}(1-\hat{p})} \leq 0.5$ to obtain the *conservative* confidence interval,

$$\left(\hat{p} - z_{\alpha/2}\frac{0.5}{\sqrt{n}}, \hat{p} + z_{\alpha/2}\frac{0.5}{\sqrt{n}}\right).$$

   The half-width of a confidence interval is known as its *margin of error* and for the conservative confidence interval the margin of error does not depend on the proportion of "successes". Thus we can compute a universal sample size for which a certain desired margin of error is reached.

      i. Compute the required sample sizes to obtain the margins of error of 0.01, 0.02 and 0.03 for a 95% conservative confidence interval.
      ii. Study how much the calculations in part *i* over-estimate the required sample sizes when the proportion of successes is small $\hat{p} = 0.05$. That is, redo part *i* using the regular binary confidence interval in slide 4.6.

   b. A manufacturer claims that only 6% of their products are faulty. To investigate this, a customer picks a random sample of size n of products and observes the proportion of faulty ones to be $\hat{p} = 0.09$. He tests the manufacturer's claim using the asymptotic one-sample proportion test in slide 4.9. Is the p-value of the test smaller for sample size $n = 100$ or $n = 200$?

Class exercise

To be solved at the exercise session.

*Note: all the needed data sets are either given below or available in base **R**.*

---

1. The data set `precip` describes the average annual amounts of precipitation (rainfall) in inches for 70 United States (and Puerto Rico) cities. A city is said to be *dry* if its average annual rainfall is less than 20 inches. Treat the data as a random sample amongst all US cities and estimate a confidence interval for the proportion of dry cities in the US.
   a. Visualize the data.
   b. Create a new variable which takes the value `1` if the city is dry and `0` otherwise.
   c. Compute an approximate 95% confidence interval for the proportion of dry cities.
   d. What is the interpretation of the confidence interval in part c?

---

2. In 2018, a proportion $p_0 = 0.098$ of people living in Finland had their last name beginning with a vowel. Treat the previous fact as a hypothesis and test it using the participants of the exercise session as a sample.
   a. Observe the sample size n and the observed proportion $\hat{p}$ of participants having last names beginning with a vowel.
   b. Write down the assumptions and hypotheses of the one-sample proportion test.
   c. Conduct the test, using the exact version of the test if the requirements of the approximative test on slide 4.9 are not fulfilled.
   d. What is the conclusion of the test? Can this conclusion be taken as evidence against/for the "hypothesis"?

---

3. In the beginning of the year a total of $n_1 = 963$ people were polled and $x_1 = 537$ out of them expressed their support for a certain presidential candidate. In a poll organized one month later $x_2 = 438$ people out of $n_2 = 901$ people claimed to support the candidate. Based on the data, has the support for the candidate decreased?
   a. Visualize the data.
   b. Write down the hypotheses for a two-sample proportion test and conduct it on a significance level 5%.
   c. What are the conclusions of the test?
   d. What assumptions were required by the test in part b? How can a poll-organizer ensure that they are satisfied?

---

4. **(Optional)** Find out how the *Wilson score confidence interval* for a binary proportion is computed and locate an R package which computes it (there are several). The Wilson interval gives a better coverage probability than the standard CI given in slide 4.6 for small sample sizes. Find out how large this improvement is by conducting a simulation study. For example, simulate `m = 10000` samples from the binomial distribution with `n = 15` and `p = 0.3` and compute for both intervals the proportion of the samples in which the interval contains the true parameter value.
