Exercise 13

Applied Statistics, IT University of Copenhagen

Preparation

- Read pages 283–284, 305–307, 321–325 from Verzani (2014).

Exercises

1. One sample t-test (T)

We perform a t-test for the null hypothesis $H_0 : \mu = 10$ by means of a dataset consisting of $n = 16$ elements with sample mean 11 and sample variance 4. We use significance level 0.05.

a. Should we reject the null hypothesis in favor of $H_1 : \mu \neq 10$?

b. What if we test against $H_1 : \mu > 10$?

2. Major League Baseball (R)

The data set OBP (Using R) contains on-base percentages for the 2002 Major League Baseball season. Do a significance test to see wheter the mean on-base percentage is 0.330 against a two-sided alternative.

3. Easter Eggs (T & R)

Assume that you got six similar Easter eggs with 20g of chocolate reported in each. After taking one more lecture in Applied Statistics, you want to further investigate whether it is plausible that the eggs really contain 20g chocolate or if the egg producer is cheating. You weight the eggs and obtain the following six observations for the chocolate weight:

Chocolate contents (g)
20.1, 19.1, 18.2, 20.2, 19.6, 19.1

You may assume that you measurement is a realization of a random sample from a normal distribution $N(\mu, \sigma^2)$, where μ represents the true average contents.

(a) Formulate the appropriate null hypothesis and alternative hypothesis.

(b) Which test is appropriate for testing the hypothesis? Explain why.

(c) Compute the value of the test statistic and report your conclusion at significance level $\alpha = 0.05$.

(d) Compute the corresponding left tail p-value. Is it likely to observe these measurements under the null hypothesis?

4. Two-sample t-test (T)

The data in Table 28.3 (pp. 425, Dekking et al. (2010)) represents salaries (in pounds Sterling) in 72 randomly selected advertisements in The Guardian (April 6, 1992). When the range was given in the advertisement, the midpoint of the range is reproduced in the table. The data are salaries corresponding to two kinds of occupations ($n = m = 72$): (1) Creative, media, and marketing and (2) education. The sample mean and sample variance of the two datasets are, respectively:

$$
\begin{aligned}
(1)\ \bar{x}_{72} &= 17410 \text{ and } s_x^2 = 41258741, \\
(2)\ \bar{y}_{72} &= 19818 \text{ and } s_y^2 = 50744521,
\end{aligned}
$$

Suppose that the datasets are modeled as realizations of normal distributions with expectations μ_1 and μ_2, which represent the salaries for occupations (1) and (2).

a. Test the null hypothesis that the salary for both occupations is the same at level $\alpha = 0.05$ under the assumption of equal variances. Formulate the proper null and alternative hypotheses, compute the value of the test statistic, and report your conclusion.

b. Do the same without the assumption of equal variances.

c. As a comparison, one carries out an empirical bootstrap simulation for the nonpooled studentized mean difference. The bootstrap approximations for the critical values are $c_l^* = -2.004$ and $c_u^* = 2.133$. Report your conclusions about the salaries on the basis of the bootstrap results.

5. Bootstrapping in two-sample tests (R)

For the `babies (UsingR)` data set, the variable `age` contains the recorded mom's age and `dage` contains the dad´s age for several different cases in the sample. Do a significance test of the null hypothesis of equal age against a one-sided alternative that dads are older in the sample population. Use a non-normal model with bootstrapping.

Dekking, F. M., C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester. 2010. *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer-Verlag.

Verzani, John. 2014. *Using R for Introductory Statistics*. CRC Press.