Tilburg University

Assignment 1 880260, 2019/20

Due at 23:55 on Sunday, 10-11-2019

Requirements: Assignments have to be handed in, one for each group of students. Please write all group member names with corresponding student (SNR) and administrative (ANR) numbers on the top of the first sheet that you hand in. Remember that assignments have to be typeset in English and submitted in Canvas as a single PDF file.

Always justify your answers! When asked to use R, always make sure you provide your code at the end of the assignment; the results you obtained should be reproducible.

Exercise 1. [6 points in total] Statistics on trial: The case of Lucia de Berk

In this exercise, we consider one of the most famous cases of miss-carriage of justice in Dutch history in which statistics played a central role. Lucia de Berk is a pediatric nurse who worked at three hospitals, including Juliana Children's hospital JKZ. Because of the unexpected death of a baby, several other ones, and some serious incidents, JKZ decided to press charges against de Berk. After investigation of this and other cases, Lucia de Berk was sentenced to life imprisonment by the court for the murder of four patients and the attempted murder of three others. During the trial, the court had consulted a statistician who calculated that the probability of having the same number of incidents as Lucia purely by chance would be 1 in 342 million. In 2008, the case was re-opened after which Lucia de Berk was freed and exonerated (in 2010).

Let us take a closer look at some statistics used to support the innocence of Lucia de Berk. The probability of 1 in 342 million was calculated using questionable statistics. Furthermore, it relied on the assumption that nurses are interchangeable with respect to the occurrence of medical emergencies. Let us study the influence this assumption has on the probability of a high number of incidents during a nurse's shifts (and using better statistics). Please refer to the paper of Gill, Groenenboom, & de Jong (2010):

https://arxiv.org/abs/1009.0802 .

Table 1. Number of shifts and number of incidents in total and for Lucia de B.

	Nr. Of incidents	Nr. Of shifts
Total	26	1734
Lucia de B.	7	203

1. [1 point] Based on the numbers reported in Table 1, calculate the probability that 7 or more incidents occur during 203 shifts. Rely on the hypergeometric distribution. Report the probability. What do you conclude with respect to Lucia de B. from this probability?
2. [2 points] Now relax the assumption –underlying the hypergeometric test- that the total number of shifts with is fixed to the numbers reported in Table 1. To do this, make use of the rpois function in R to generate data under the model of 'an innocent nurse'. The Poisson distribution is used to model the probability of a number of events occurring in a given time

interval: $X \sim Poiss(\lambda)$ with $\lambda > 0$ expressing the event rate (e.g., if you receive on average 4 letters per month the event rate would be 4 (per month) or 12*4 (per year)).

- What value will you use for λ, the rate parameter (assuming an innocent nurse)?
- Using simulation, what is the probability of 7 or more incidents during 203 shifts under the model of an innocent nurse? What do you conclude from this probability?

3. [2 points] Repeat the exercise above but now relax the assumption that nurses are exchangeable (this is, have the same likeliness of encountering an incident during their shift). To do this, first generate a nurse specific incidence rate using the rexp function in R which generates incidence rates from the exponential distribution. The exponential distribution is the probability distribution describing time between events, $t \sim Exp\left(\frac{1}{\lambda}\right)$, thus its parameter is the inverse rate (e.g., if you receive on average 10 letters per month, you expect the time between two letters to be 30/10 = 3 days). Having generated a nurse specific incidence rate, generate the number of incidents under the Poisson model of an ‘innocent nurse with nurse-specific incidence parameter’. Repeat these two steps many times and estimate the probability of 7 or more incidents under the model of an innocent nurse.
   - From which distribution do you generate the nurse-specific incidence rate?
   - Using simulation, what is the probability of 7 or more incidents under the model of an innocent though not exchangeable nurse?
4. [1 point] Comment on your findings and discuss the issue of (not) assuming exchangeability of nurses.

Exercise 2. Should one use Welch’s t-test instead of Student’s t-test by default?

In this assignment, we focus on two statistics that have been proposed for estimating and testing the difference in means between two independent samples:

*For a first group n_1 data are sampled from $N(\mu_1, \sigma^2_1)$.

*For a second group n_2 data are sampled from $N(\mu_2, \sigma^2_2)$.

A popular tool to assess and test the difference in means between two independent populations, is Student’s t-test. An alternative but less known tool is Welch’s t-test. Here we will use Monte Carlo simulation to compare Student’s t-test with Welch’s t-test.

Useful reading: Delacre, Lakens, Leys (2017). Why Psychologists Should by Default Use Welch’s t-test Instead of Student’s t-test. *International review of social psychology, 30(1),* 92–101. DOI: http://doi.org/10.5334/irsp.82. (https://www.rips-irsp.com/article/10.5334/irsp.82/).

Part 1[1]: [9 points in total] **Properties of the estimators of the standard error of the difference in means**: Student’s t statistic compared to Welch’s t statistic.

In this first part, interest is in the standard deviation of $\bar{x}_1$-$\bar{x}_2$ based on the population variances, $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Popular estimators of this population parameter are the standard error based on the pooled standard deviation that is used in Student’s t-test and the standard error used in Welch’s t-test. Here are the formulas of these estimators:

[1] Part 2 will be part of the second graded assignment

-Standard error in Student's t-test statistic: $SE_{student} = S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$,

with S_p the pooled standard deviation: $S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.
Note that $s_1^2$and $s_2^2$are the unbiased variance estimates.

-Standard error in Welch's t-test statistic: $SE_{Welch} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

1. [6 points in total] Conduct a Monte Carlo simulation experiment to estimate the bias, variance, mean squared error (MSE), and relative efficiency (RE) of the estimated standard error according to both Student's t-test and Welch's t-test. Use the following values for the population parameters and sample sizes: Fix n_2 = 100 and vary the sample size for n_1 (n_1=10, n_1=100, or n_1=200); also fix σ^2_1=2 and vary σ^2_2 (σ^2_2=1, σ^2_2=2, σ^2_2=10); finally, fix μ_1=0 and μ_2=1.
   a. [1 point] What is the value of the population parameter that is estimated by $SE_{Student}$ and SE_{Welch} in the condition with n_1=10 and σ^2_2=1? Show how you obtained this value.
   b. [2 points] Give pseudo code for the Monte Carlo simulation experiment you will use to estimate bias, variance, MSE, and RE for the two types of standard errors in the condition with n_1=5 and σ^2_2=1.
   c. [2 points] Implement the simulation study in R and report, for each of the nine conditions, the results with respect to:
   -bias,
   -variance,
   -MSE, and
   -RE.
   Make use of tables and/or plots.
   d. [1 point] Which assumptions underlie Student's t-test? Which assumptions underlie Welch's t-test? Given these assumptions, what do you expect in terms of the relative performance of the Student's t estimator of the standard error compared to Welch's t estimator in the different conditions?
2. [3 points] Discuss the results obtained in 1c by briefly answering the following questions:
   -Which estimator is biased, and under what conditions of sample size and variance?
   -What is the influence of the sample size and variance σ^2_2 on the variance of the two estimators?
   -Which of the two methods (Student's t versus Welch) is the more efficient one, and under what circumstances?

[BONUS: 2 points but with a maximum of 15 points in total for the assignment] Give a mathematical proof for the equivalence of the two estimators under equal population variances ($\sigma^2_1 = \sigma^2_2$) or equal sample sizes of the two samples ($n_1 = n_2$).