Introduction to Statistics and Data Analysis with R - Homework #2

Adi Sarid and Afek Adler

2019-11-03

This homework sheet is due on the 3rd of December. You may submit your answers in pairs, **in R** (,Rmd file). Submission will be performed electronically via Moodle.

We urge you to start solving this sheet as soon as possible and, if you have any questions, come to visit us in reception hours next week.

Question 1:

As we did in the exercise 4, show that if confidence intervals are constructed using a given confidence level from an infinite number of independent sample statistics, the proportion of those intervals that contain the true value of the parameter will be equal to the confidence level.

In the exercise we have seen it for the confidence interval for the mean of i.i.d samples from normal distribution with known variance, here we will do it for the confidence interval of sigma (as you have seen in lecture 3, slide #20).

Modify the code -

```
miu = 10
sigma = 3
n = 10
alpha = 0.1
N_tests <- 10000
counter <- 0
error = qnorm(1-alpha/2)*(sigma/sqrt(n))
for (i in 1:N_tests)
  {sample = rnorm(n,miu,sigma)
  sample_mean <- mean(sample)
  left <- sample_mean-error
  right <- sample_mean + error
  between <- (left <= miu) & (miu <= right)
  counter <- counter+between}
estimated_confidence <- (counter/N_tests)
true_confidence <- 1- alpha
print (abs(estimated_confidence - true_confidence))
```

```
## [1] 0.0012
```

Question 2:

In this question we will look at the iris classic dataset (built in in r environment):

1. Plot a box plot of the numeric features
2. Provide two sided confidence interval for the Sepal.Length & Sepal.Width. How was that confidence interval calculated? e.g what assumptions on the mean and variance fit to this particular confidence interval.
3. Provide a confidence interval for the probability to belong to the setosa species.
4. Does the above confidence interval really represent the confidence interval of this setosa species among the population of all the iris species? e.g, is this dataset biased in some meaning?
5. Plot a scatter plot such that:

```
* Sepal.Length will appear on the x axis
* Sepal.Width will appear on the y axis
* Each point will have color based on the species of that given iris
* The figure will have a legend that explains the colors of the species
```

4. Look at that splendid iris -

Suppose it has Sepal.Length of 7 and Sepal.Width of 3. Based on the previous figure, which iris type would you say it is?

Question 3:

This question is an appetizer for hypothesis test. we will create a QQ plot. A QQ plot is a visual test (not a statistical test!) to check whether a one dimensional variable is distributed normal.

Instructions are provided in the attached link:

1. What is a QQ plot? e.g, what appears in the x and y axis?
2. make a QQ plot for sample1 and sample2. are the plots the same? What's the difference?

- As always, plot the graph with `ggplot` https://www.youtube.com/watch?v=X9_ISJ0YpGw (explanation)

```
set.seed(0)
sample1 <- rnorm(100,100,36)
sample2 <- rnorm(100^2,100,36)
#######
# your code here


#######
```

Question 4:

In this question we will verify the following formulas:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\cdot n\mu = \mu$$

$$V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

1. Read about bootstrapping in Wikipedia.
2. What are the different phases of performing bootstrapping? Where does it appear in the following code?
3. What is the error metric we use in order to calculate the deviations in the result simulation, why? offer another evaluation metric.

https://en.wikipedia.org/wiki/Bootstrapping_(statistics) (bootstrapping)

```
set.seed(0)
true_mean <- 10
true_var <- true_mean
true_sd <- sqrt(true_mean)
population <- 10^3
samples <- rpois(population, lambda = true_mean)
bootstrap_samples <-1000
n_samples <- 100
means <- numeric(bootstrap_samples)
var <- numeric(bootstrap_samples)
for (i in 1:bootstrap_samples) {
   temp_sample <- sample(samples, size= n_samples, replace =T)
   means[i] <- mean(temp_sample)
   var[i] <- var(temp_sample)
}
mean_daviation <- abs(mean(means)-true_mean)/true_mean
var_deviation <- abs(var(means)-true_mean/n_samples)/true_var
print (mean_daviation)
```

```
## [1] 0.020381
```

```
print(var_deviation)
```

```
## [1] 0.0005721305
```

Bonus:

- Implement the same code with r `boot` package.

Question 5:

The breaking strength of yarn used in manufacturing drapery material is required to be at least 100 psi. Past experience has indicated that breaking strength is normally distributed and that $\sigma = 2$ psi. A random sample of nine specimens is tested, and the average breaking strength is found to be 98 psi. Find a 95% two-sided confidence interval on the true mean breaking strength.