Parameter estimation and hypothesis testing

September 18, 2019

In this lab, you will apply the concepts seen during the last two classes.

1. Flower measurements of three iris species

For this exercise, we will use Edgar Anderson's famous iris data set, which contains different measurements taken on 50 flowers of 3 iris species. This data frame is already loaded into R under the name `iris`.

```
data(iris)
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

a) We will first visualize some of the data. Using *ggplot2*, produce a scatter plot relating the length and width of the sepals (`Sepal.Length` and`Sepal.Width`), differentiating the points of each species by color.

How would you calculate the mean of `Sepal.Width` for species *setosa* and its 95% confidence interval?

b) What quantities do you need for this calculation?

c) Calculate the mean of `Sepal.Width` for species *setosa* and its standard error.

d) During the class on statistical distributions, we saw the functions `rnorm`,`dnorm`, `pnorm` and`qnorm` which calculate values from the normal distribution. Similar functions exist for the t distribution (`rt`,`dt`, `pt`,`qt`). Let's use the function `qt(p, df)` to determine the interval corresponding to 95% of the probability. What values of p (cumulative probability) should we use? How many degrees of freedom (df) based on sample size n?

e) Calculate the 95% confidence interval for the mean calculated in (c). How can you interpret this interval?

2. Choice of sampling methods

Imagine you want to determine the mean and standard deviation of the growth rate of yellow birch seedlings in a given region. Your sampling unit is a 1 m^2 quadrat where the annual growth of seedlings is measured. Which sampling method would you recommend for the placement of these quadrats according to the different scenarios presented, and why? You can choose stratified, cluster, systematic, or adaptive sampling.

a) The stands of this species in the area are relatively similar but far apart.

b) Growth could be strongly influenced by the north-south temperature gradient at the scale considered.

c) The species is found in several distinct stand types covering different proportions of the territory (eg 70% type A, 25% type B, 5% type C).

d) It is a rare species in the region and its distribution is little known.

3. Stratified sampling simulation

For this exercise, we will compare simple and stratified sampling from simulated samples from the iris data frame.

a) Create a boxplot of the distribution of the petal length by species.

b) Create two functions `iris_alea` and`iris_strat`. The first function chooses 30 random observations of `iris`, then returns the mean of `Petal.Length` for these observations. The second chooses 10 random observations from each of the three species, then returns the mean of `Petal.Length` (overall mean, not by species).

Notes:

- The `sample(x, size)` function in R simulates sampling a number of values (given by *size*) from the values in vector *x*.
- You can write these functions without arguments (empty parentheses after `function`), as in the example below.

```
iris_alea <- function() {
    # Insert function code here
}
```

c) Generate a vector of 1000 results of each function with `replicate`, as follows:

```
rep_alea <- replicate(1000, iris_alea())
rep_strat <- replicate(1000, iris_strat())
```

Calculate the standard error of each mean (from the standard deviation of `rep_alea` and `rep_strat`). Before running the calculation, can you predict which method will be more precise and why?

4. Ozone concentration in three gardens

For this exercise, we will use the `gardens.csv` dataset, which comes from Michael Crawley's book *Statistics: An Introduction Using R*. These data represent ozone concentrations (in parts per 100 million or pphm) measured in three gardens (A, B and C) on different days.

```
gardens <- read.csv("gardens.csv")
```

a) Observe the distribution of ozone measurements in each garden. What type of graph could you use?

b) What is the mean and standard deviation of the ozone concentration in each garden? Is the mean a good indicator of the "typical" value in each garden?

Note: The function `tapply(X, INDEX, FUN)` applies a function given by *FUN* to vector *X* separately for each factor given by *INDEX*. Therefore we can calculate the mean ozone concentration by garden with `tapply(gardens$Ozone, gardens$Garden, mean)`.

c) From these data, test the null hypothesis that gardens A and B receive the same mean ozone concentration. What is your estimate of the difference between the mean and its 99% confidence interval? Does this test give a good idea of the difference between the two gardens? Explain your answer.

d) Repeat the previous exercise for the null hypothesis that gardens A and C receive the same average ozone concentration. Comment on the difference between this result and the previous result.