R PROGRAMMING END SEMESTER

*Done by Pranav Gopalkrishna, register number 1940223, class 3CMS*

2020-11-17

# QUESTION 1

#

Create the sample numbers of 150 which are normally distributed and check whether the estimators are unbiased. (10 marks)

#

A bias is the difference between an estimate and the parameter. An unbiased estimator is an estimator with no bias. Hence, an unbiased estimator is a statistic whose expected value equals the parameter it aims to estimate. The estimator in our case will be the sample mean

#

Population parameters are

popl\_mean = 12  
popl\_sd = 6  
#

Here, I generate 1000 normal samples of size 150, with population parameters as defined above. Then, taking their mean, I substract them from the population mean to find the bias. I repeat this process 10 times, to get 10 biases.

biases = replicate(10, popl\_mean - mean(replicate(1000, mean(rnorm(150, mean = popl\_mean, sd = popl\_sd)))))

Averaging the various obtained estimates to get a more accurate estimate of population mean and standard deviation…

print(biases)

## [1] -0.0109940578 0.0107787874 -0.0050534254 -0.0148088628 -0.0146998851  
## [6] 0.0149855842 0.0007574369 0.0066657595 0.0076457971 0.0244667254

sum(biases)

## [1] 0.01974386

As can be seen, the biases are close to zero, and are not significantly mostly positive or negative. Hence, we may conclude that sample mean is an unbiased estimator of population mean.

#  
#

# QUESTION 2

#

Recently 195 North Georgia students were given a perfectionism survey. -Scores of 90 or higher indicate a person who has significant perfectionistic tendencies. The chart below compares the number of perfectionists by gender. Test at the 0.05 level of significance whether a gender difference exists for perfectionism. (10 marks)

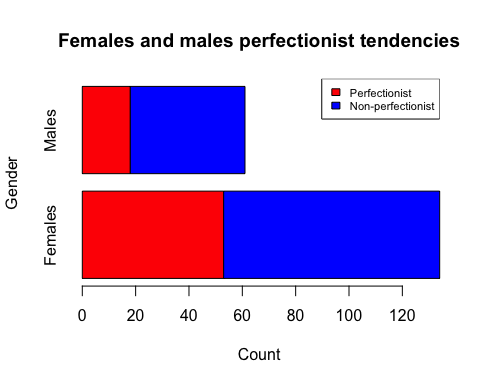
#  
data = matrix(c(53, 18, 81, 43), ncol = 2, byrow = TRUE)  
print(data)

## [,1] [,2]  
## [1,] 53 18  
## [2,] 81 43

Here, the rows denote the presence of perfectionistic tendencies {Perf, Not} and the columns denote the gender {Females, Males}

#

barplot(  
 data,  
 main = "Females and males perfectionist tendencies",  
 col = c("red", "blue"),  
 xlab = "Count",   
 ylab = "Gender",  
 names.arg = c("Females", "Males"),  
 horiz = TRUE  
 )  
legend(  
 "topright",  
 c("Perfectionist", "Non-perfectionist"),  
 fill = c("red", "blue"),  
 cex=0.70  
)



Here, since there are two attributes that we want to find an association between, we use the Chi-squared test for testing the independence of attributes.

#

Hypotheses:

# H\_0: Gender and perfectionistic tendencies are independent (no association)  
# H\_0: Gender and perfectionistic tendencies are associated  
# (Note that level of significance = 0.05)  
#

The test results are…

chisq.test(data)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: data  
## X-squared = 1.4184, df = 1, p-value = 0.2337

Here, p = 0.2337 > 0.05. The level of significance is the probability of getting significant Chi-square values (i.e. significantly low-probability Chi-squared values, since a lower probability value occurring means that the probability that the value occurred by pure chance is lesser), assuming that H\_0 is true i.e. assuming that the attributes are independent. The p-value is the probability of the current Chi-squared value occurring. Hence, since p > level of significance here, this means the probability if getting our current Chi-squared value is higher than the probability of getting a significant Chi-squared value. Hence, the current Chi-squared value has a high enough probability of occurring to not be considerd significance i.e. to be considered as probably a product of pure chance. Hence, we may accept H\_0, i.e. we may conclude that there is no association between gender and perfectionistic tendencies.

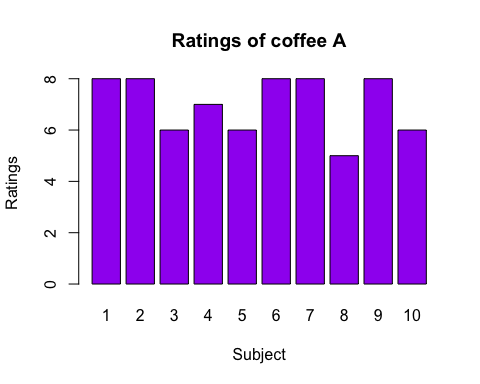
#  
#

# QUESTION 3

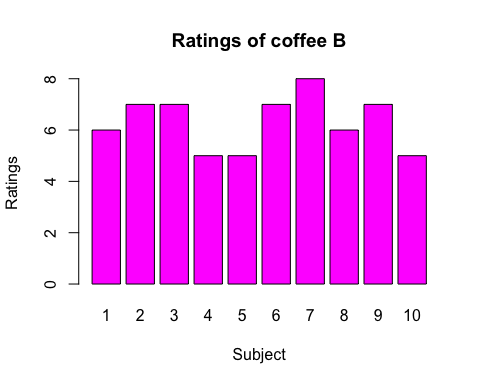
#

The same group of people rated both brands of coffee. Half of the subjects tasted brand A first. Half of the subjects tasted brand B first. Is there a difference in the mean ratings between the two brands of coffee? (10 marks)

#  
library(ggplot2)  
subjects = c(1:10)  
coffee\_A = c(8, 8, 6, 7, 6, 8, 8, 5, 8, 6)  
coffee\_B = c(6, 7, 7, 5, 5, 7, 8, 6, 7, 5)  
data = data.frame(subjects, coffee\_A, coffee\_B)  
barplot(main = "Ratings of coffee A", data$coffee\_A, xlab = "Subject", ylab = "Ratings", names.arg = c(1:10), col = "purple")



barplot(main = "Ratings of coffee B", data$coffee\_B, xlab = "Subject", ylab = "Ratings", names.arg = c(1:10), col = "magenta")



#

Here, we have two samples, and we need to infer the means of the populations of the two samples, and compare them to see if they are significantly different. Hence, we use t-test. The samples are not paired, since half the people tasted A first, and the other half tasted the B first, so we cannot determine for any one observation whether it was the A’s rating that affected B’s rating or vice versa.

#

By performing a t-test, we make the following assumptions…

# 1. Samples are independent (which they are, in our case)  
# 2. Samples follow normal distribution  
#

Also, we need to check if the samples’ inferred populations have the same or very similar variances, since there are different variations of the t-test, depending on this answer.

#

To first test whether the samples’ inferred populations can be said to have the same variances, we use the F-test. Hypotheses:

# H\_0: Population variances are equal  
# H\_1: Population variances are different  
# (Note that level of significance = 0.05)  
var.test(data$coffee\_A, data$coffee\_B)

##   
## F test to compare two variances  
##   
## data: data$coffee\_A and data$coffee\_B  
## F = 1.1881, num df = 9, denom df = 9, p-value = 0.8016  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.2951119 4.7833594  
## sample estimates:  
## ratio of variances   
## 1.188119

Here we have p = 0.8016 > 0.05. Hence, for reasons similar to the ones discussed in question 2’s conclusion, we may accept H\_0 i.e. we may conclude that the variances of the inferred populations of the samples of coffee ratings are equal. In other words, in the population, we may expect with 95% confidence that the varinances in the ratings of coffee A and B in the population will be close enough to be considered equal.

#

Now, to test whether the samples’ inferred populations each can be said to follow a normal distribution, we use the Shapiro-Wilk test. Hypotheses:

# H\_0: The population follows normal distribution  
# H\_1: The population follows some non-normal distribution  
# (Note that level of significance = 0.05)  
shapiro.test(data$coffee\_A)

##   
## Shapiro-Wilk normality test  
##   
## data: data$coffee\_A  
## W = 0.79745, p-value = 0.0135

shapiro.test(data$coffee\_B)

##   
## Shapiro-Wilk normality test  
##   
## data: data$coffee\_B  
## W = 0.87375, p-value = 0.1105

For coffee A’s ratings, p = 0.0135 < 0.05 (H\_0 rejected), while for coffee B’s ratings, p = 0.1105 > 0.05 (H\_0 accepted). Hence, while the coffee ratings of A from the entire population cannot be said to follow a normal distribution, coffee ratings of B from the entire population can be said follow a normal distribution.

#

However, since the t-test is robust, assumptions not being met can still allow for an accurate test. So we continue with the t-test. Hypotheses:

# H\_0: Population means are the same  
# H\_1: Population means are significantly different  
# (Note that level of significance in 0.05)  
t.test(data$coffee\_A, data$coffee\_B, var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: data$coffee\_A and data$coffee\_B  
## t = 1.4126, df = 18, p-value = 0.1748  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.3410817 1.7410817  
## sample estimates:  
## mean of x mean of y   
## 7.0 6.3

Here, we have p = 0.1748. Hence, we may accept H\_0. Hence, we may conclude that for the entire population, the ratings of coffee A and coffee B will not be significantly different. This suggests that there will be strong competition between the two brands.