## Student: Carles Sans Fuentes

## NIF: 920531-T252

## Contents

Applet activities ..... 2
4.84 ..... 2
4.117 ..... 2
4.118 ..... 3
10.11 ..... 4
10.12 ..... 5
11.6 ..... 5
11.50 ..... 5
16.4 ..... 6
Imputation techniques ..... 7
1. Which type of missing mechanism do you prefer to get a good imputation? ..... 7
2. Say something about simple random imputation and regression imputation of a single variable. ..... 7
3. Explain shortly what Multiple Imputation MI is. ..... 8
Table of Figures ..... 8

# Applet activities

## 4.84

This is the code for the first activity that shoud be used in R to check for results:

```
gamma4<- dgamma(0:150, 4,1)

gamma40<-dgamma(0:150, 40,1)

gamma80<-dgamma(0:150, 80,1)

mydata<-data.frame(gamma4,gamma40, gamma80)

mydata<- cbind(0:150, mydata)

colnames(mydata)[1]<-"density"

require(ggplot2)

ggplot2::ggplot(mydata, aes(x= density))+

      geom_line(aes(y = gamma4, color = "alpha = 4")) +

      geom_line(aes(y = gamma40, color = "alpha = 40"))+

      geom_line(aes(y = gamma80, color = "alpha = 80"))+

      xlab("y")+   ylab("f(y)")+ ggtitle("Gamma density distribution")
```

**a.** It is observed the larger the values of alpha, the more symmetric the density curves are.

**b.** The location of the distribution centers are increasing with alpha being higher.

**c.** The means of the distributions are increasing with alpha increasing as well.

## 4.117

This is the code I used to check the answers:

```
x <- seq(-3, 3, length=100)

Fx <-x

beta_9_7<- dbeta(Fx, 9,7)

beta_10_7<-dbeta(Fx, 10,7)

beta_12_7<-dbeta(Fx, 12,7)

mydata2<-data.frame(beta_9_7,beta_10_7, beta_12_7)

mydata2<- cbind(Fx, mydata2)

colnames(mydata2)[1]<-"density"
```

```
require(ggplot2)

ggplot2::ggplot(mydata2, aes(x= density))+

 geom_line(aes(y = beta_9_7, color = "alpha = 9, Beta = 7")) +

 geom_line(aes(y = beta_10_7, color = "alpha = 10, Beta = 7"))+

 geom_line(aes(y = beta_12_7, color = "alpha = 12, Beta = 7"))+

 xlab("y")+  ylab("f(y)")+ ggtitle("Beta density distribution")
```

**a.** All of the densities are skewed left.

**b.** The density obtains a more symmetric appearance when the value of alpha gets closer to 12.

**c.** The shape of Beta is always skewed right when alpha> Beta and  alpha > 1 and alpha > 1.

## 4.118

This is again the code used to answer the activity.

```
x <- seq(0, 1, length=1000)

Fx <- x

beta_3_3<- dbeta(Fx, 0.3,4)

beta_3_7<-dbeta(Fx, 0.3,7)

beta_3_12<-dbeta(Fx, 0.3,12)

mydata3<-data.frame(beta_3_3,beta_3_7, beta_3_12)

mydata3<- cbind(Fx, mydata3)

colnames(mydata3)[1]<-"density"

require(ggplot2)

ggplot2::ggplot(mydata3, aes(x= density))+

 geom_line(aes(y = beta_3_3, color = "alpha = 0.3, Beta = 4")) +

 geom_line(aes(y = beta_3_7, color = "alpha = 0.3, Beta = 7"))+

 geom_line(aes(y = beta_3_12, color = "alpha = 0.3, Beta = 12"))+

 xlab("y")+

 ylab("f(y)")+ ggtitle("Beta density distribution")
```

**a.** All of the densities are skewed right.

**b.** The spread decreases as the value of beta gets closer to 12.

**c.** The distribution with alpha = .3 and Beta= 4 has the highest probability.

**d.** The shapes are all really similar.

## 10.11

**a.** We can only do the type error II, which is the one about failing to reject H0.

**b.** It has been rejected 15 out of 200, so a 7.5% times of error II. Below it can be found my result on the simulations.

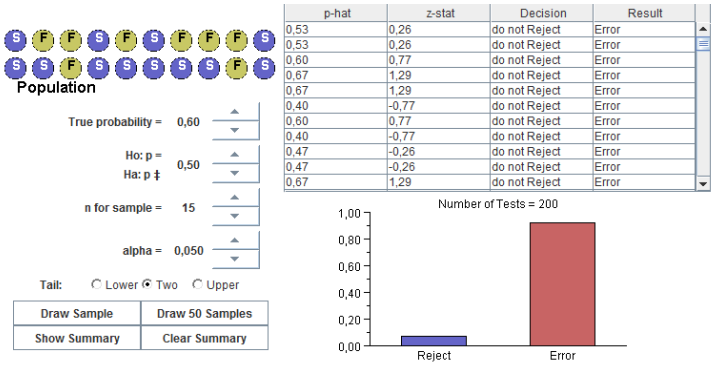


Figure 1 number of tests = 200 for n sample of 15

**c & d.** The reject rate (type error II) has increased by almost 10% when n increases (from 7.5% to 16.5%). Here below it can be seen the simulations and final summary of it.

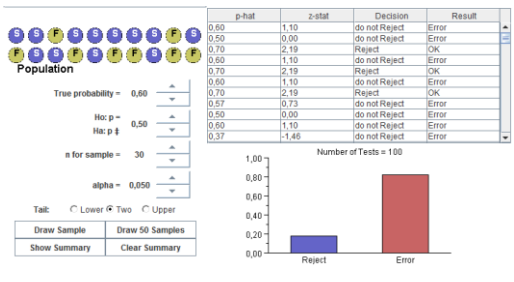


Figure 4 number of tests = 100

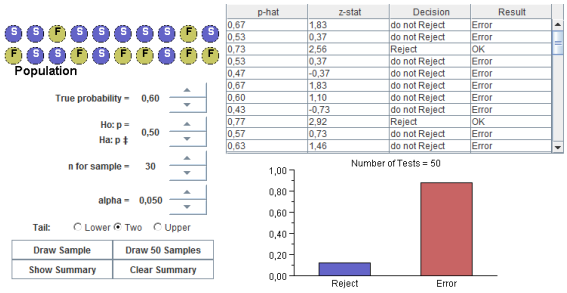


Figure 2 number of tests = 50

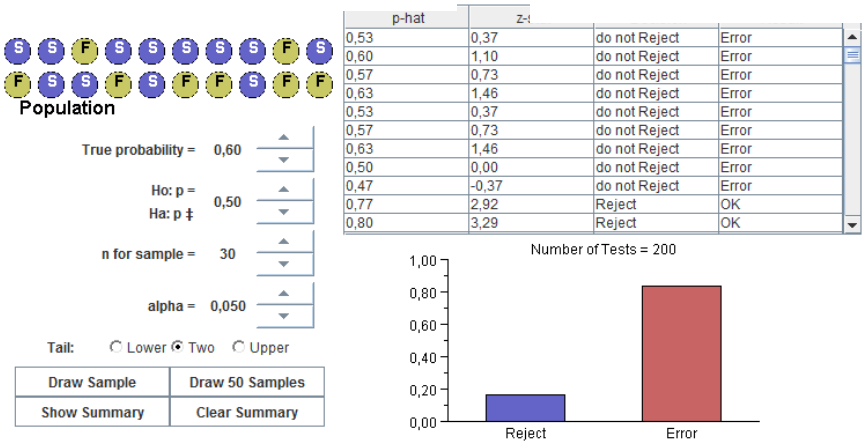


Figure 3 number of tests = 200

| True Value | Null Value | Sample Size | N of Tests | Test Type | alpha | Reject Rate | Error Rate |
|---|---|---|---|---|---|---|---|
| 0,60 | 0,50 | 30 | 200 | Both | 0,050 | 0,165 | 0,835 |
| 0,60 | 0,50 | 15 | 200 | Both | 0,050 | 0,075 | 0,925 |

Figure 5 First Summary

## 10.12

| True Val... | Null Value | Sample ... | N of Tests | Test Type | alpha | Reject R... | Error Rate |
|---|---|---|---|---|---|---|---|
| 0,60 | 0,50 | 15 | 200 | Both | 0,050 | 0,135 | 0,865 |
| 0,60 | 0,50 | 100 | 200 | Both | 0,100 | 0,640 | 0,360 |
| 0,60 | 0,50 | 50 | 200 | Both | 0,100 | 0,435 | 0,565 |
| 0,60 | 0,50 | 30 | 200 | Both | 0,100 | 0,260 | 0,740 |
| 0,60 | 0,50 | 15 | 200 | Both | 0,100 | 0,275 | 0,725 |

Figure 6 Second Summary

**a.** Since Beta and alpha behave inversely to each other, the simulated value for Beta should be smaller for alpha = .10 than for alpha = .05, and in our simulation this can be seen for n = 15 that it is like this.

**b.** The one that gives the smaller values of Beta is n = to 100. Beta error gets smaller as n increases.

## 11.6

**a.** Y = 43.362 + 4.842*X; , SSE = 1002.839.

**b.** No, because the data points show an increasing trend.

**c.** Intercept = 21.575, SSE = 18.286. It happens that the SSE is much lower, so that it is reduced a lot.

**d.** It is exactly the same, but if it had been another one which was not the best one, the SSE would have been bigger on c and it would have been reduced.

**e.** It is 4,5 with 43,3625 of SSE.

**f.** It is observed that the farer on the years, the bigger the variation. The sum of the areas is the SSE.

## 11.50

**a.** $r^2$ increases as the fit improves. Logically, $r^2$ is how much of the variance is explained by the variable we regress on it.

**b.** For the best model, $r^2$ = .982 and so r = .99096.

**c.** The scatterplot in this example has a lower error variance about the line.

## 16.4

**a.** I observe a success. The posterior distribution looks with a positive slope and the prior with a negative slope.

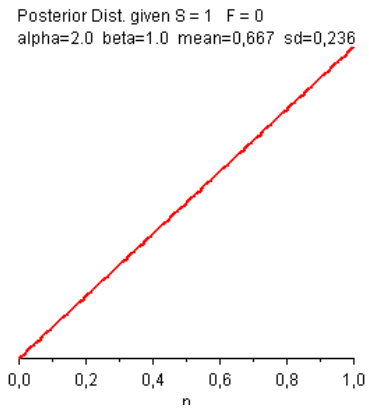


Figure 8 Posterior distributions 1 Success

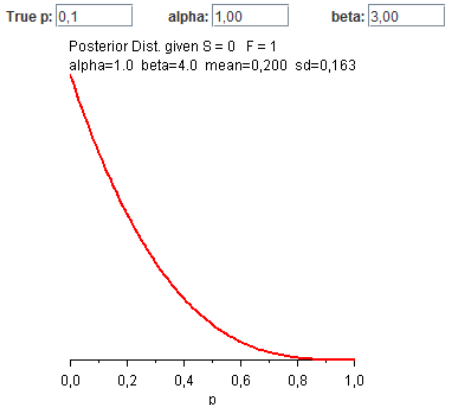


Figure 7 Posterior distributions 1 Failure

**b.** Two successes on the posterior. Yes, they look different but still it is similar to before.

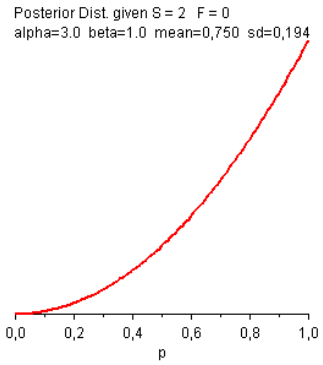


Figure 9 Posterior distributions 2 Successes

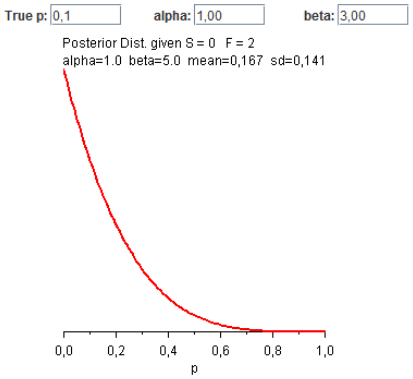


Figure 10 Posterior distributions 2 Failures

**c.** It just gets a quadratic shape.

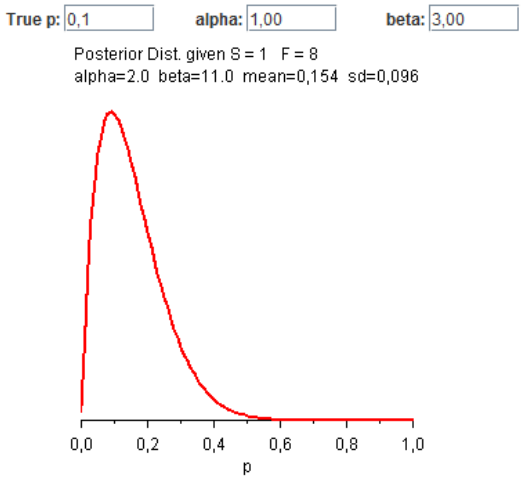


Figure 11 Posterior distributions 1 Success and 8 Failures

**d.** I just did the maximum of them, but the shape was pretty good with 150.

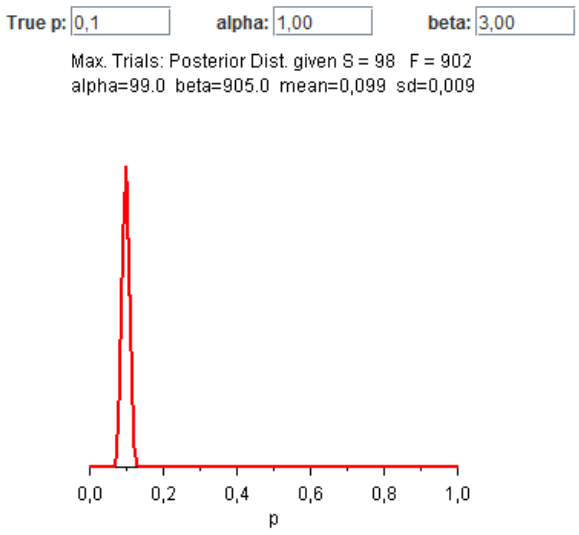


Figure 12 Posterior distributions with maximum trials

# Imputation techniques

## 1. Which type of missing mechanism do you prefer to get a good imputation?

I prefer to get a missingness that depends on unobserved predictors. If you are not able to get information for one variable but it is known that people with similar characteristics are less likely to give it (e.g. Revealing your earnings or your political beliefs if you have a degree), it can be modelled taking into account the affecting characteristics to give an accurate estimation. Of course, this model will never be totally exact, but ideally I believe it is better to deal with this problem that depends on other variables rather than dealing with a problem that depends on the variable itself.

## 2. Say something about simple random imputation and regression imputation of a single variable.

Random imputation is probably the easiest approach when dealing with missing values but it may bias your database if you want to regress the randomized variable on others since you only take into account the same variable across all the observations, but not other each variable to predict a more accurate estimation for that observation. By doing so, (if you are for instance, modelling your dataset,) it surely will bias your model doing it less exact and with enlarging the variance.

Regression imputation is a better system since you take into account other variables to predict your outcome and also, in the non-deterministic case, the variance of your regression

estimator (the error on the estimator) to generate random estimations on the range of the estimator. This will surely give a more accurate response than the deterministic model (because it takes into account the error on the predictor) the simple random imputation way.

## 3. Explain shortly what Multiple Imputation MI is.

Multivariate imputation is a system that is used to fill missing data by applying to each missing point a similar model that estimates several coefficient estimators with its error and then it weights these different coefficients and errors to obtain a better estimation for each point.

## Table of Figures

Figure 1 number of tests = 200 for n sample of 15 ..... 4
Figure 2 number of tests = 50 ..... 4
Figure 3 number of tests = 200 ..... 4
Figure 4 number of tests = 100 ..... 4
Figure 5 First Summary ..... 5
Figure 6 Second Summary ..... 5
Figure 7 Posterior distributions 1 Failure ..... **Error! Bookmark not defined.**
Figure 8 Posterior distributions 1 Success ..... **Error! Bookmark not defined.**
Figure 9 Posterior distributions 2 Successes ..... **Error! Bookmark not defined.**
Figure 10 Posterior distributions 2 Failures ..... **Error! Bookmark not defined.**
Figure 11 Posterior distributions 1 Success and 8 Failures ..... **Error! Bookmark not defined.**
Figure 12 Posterior distributions with maximum trials ..... **Error! Bookmark not defined.**