Statistical Method Computer Assignment

Hoda Fakharzadehjahromy
hodfa840@student.liu.se

1 Exercises from Course's book

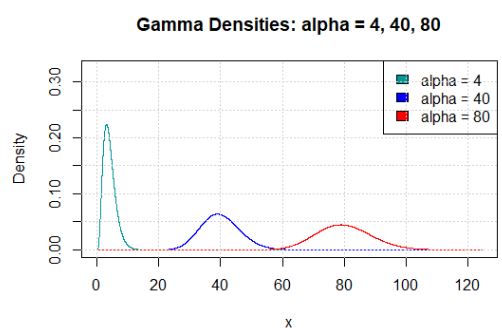
In the .R file for each exercise a function is defined

4.84

to run the code from the .R file:

```
ex.4.84()
```

Figure 1: Exercise 4.84



- (a) The density curve tends to be more symmetric as the value of α increases.
- (b) From Figure 1 it is evident that the distribution centers increases for the larger values of α.
- (c) the distribution centers increases for the larger values of α because the mean (μ)of *Density Functions* increases.

4.117

to run the code from the .R file:

```
ex.4.117()
```

- (a) The three densities are skewed left.
- (b) As the value of α gets closer to 12 the densities tends to be more symmetric.
- (c) *Beta Density Functions* are skewed right if $\alpha > \beta > 1$

4.118

to run the code from the .R file:

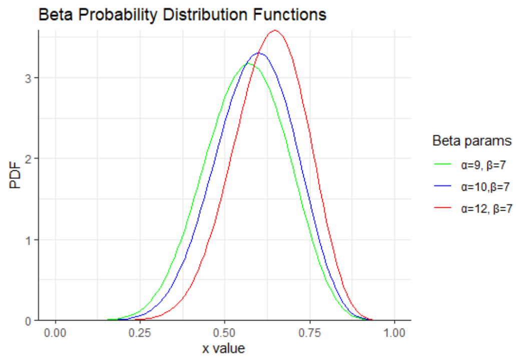


Figure 2: Exercise 4.117

```
ex.4.118()
```

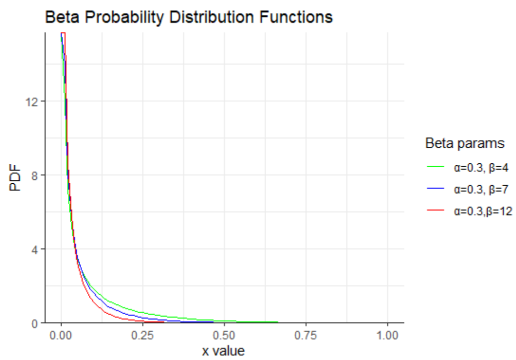


Figure 3: Exercise 4.118

- (a) The three densities are skewed right.
- (b) According Figure 3 to as the value of β gets closer to 12 the spread decreases.
- (c) At $x = 0.2$ the highest *Beta Density* value belongs to the distribution with $\alpha = 0.3$ and $\beta = 4$
- (d) When $\alpha < 1$ and $\beta > 1$ *Beta Density Function*'s shape is similar to an exponential function.

10.19

to run the code from the .R file:

```
ex.10.19()
```

$H_0 : \mu = 130$,$H_a : \mu < 130$, $n = 40$

$Z = \frac{128.6-130}{2.1/\sqrt{40}}$

so, $Z = -4.216$

test with level : $z_{0.05} = 1.645$

$H_0$is rejected because $|Z| > z_{0.05}$. There is evidence that the mean output voltage is less than 130.

10.21

to run the code from the .R file:

`ex.10.21()`

testing $H_0 : \mu_1 - \mu_2 = 0$ where μ_1 and μ_2 are the average shear strength measurements derived from unconfined compression tests for two types of soils. An estimator for μ_1 and μ_2 is $\bar{Y}_1 - \bar{Y}_2$ and the test statistic under H_0 is :

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}} = 3.65$$

Since $|Z| > z_{0.005=2.575}$ we reject $H_0 : \mu_1 \mu_2 = 0$. In other words, the soil do appear to differ with respect to average shear strength at the 0.01 confidence level.

11.31

`ex.11.31()` to test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ we fit the regression model.

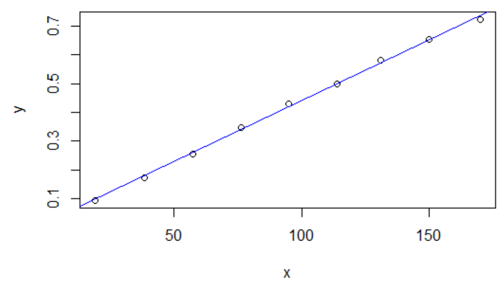


Figure 4: Exercise 11.31

The fitted model is :

$y = .01875 + .004215x$ and $p_value = 2.372e - 11$ thus H_0 is rejected and so,the peak increases as nickel concentrations increase.

11.69

The manufacturer of Lexus automobiles.

to run the code from the .R file:

`ex.11.69()`

- (a) for the Linear model: fitted model :

  $y = 32.725 + 1.812x$

- (b) Quadratic model :

  $y = 35.5625 + 1.8119x - .1351x^2$

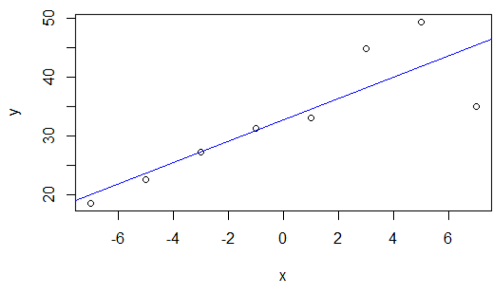


Figure 5: Exercise 11.69(a)

Imputation techniques

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

- Random regression imputations are more likely to appropriately spread across the range of the population. Matching and hot-deck imputation can be combined with regression by defining similarity as closeness in the regression predictor. Matching can be viewed as a nonparametric or local version of regression and can also be useful in some settings where setting up a regression model is challenging. In matching we replace each unit with a missing unit y, with similar values of X in the observed data . Matching imputations is more likely to avoid biased estimates.

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

- Simple random imputation: In simple random sampling imputation, samples are randomly drawn from the dataset for imputing the missing value. This approach ignores the useful information from all the other variables . This method can be a convenient starting point. A better approach is to fit a regression to the observed cases and then use that to predict the missing cases.

  Regression: A regression model is used to estimate the predict observed values of a missing data. Hence, available information for complete and incomplete cases is used to predict the value of a specific variable. The problem is that the imputed data do not have an uncertainty in them. Random regression imputation : adding the uncertainty term into the imputations by adding the prediction error into the regression. This method is less biased .

Explain shortly what Multiple Imputation is.

- Multiple Imputation: Routine multivariate imputation The direct approach to imputing missing data in several variables is to fit a multivariate model to all the variables that have missing value. The difficulty of this method is to to set up a reasonable multivariate regression model.

  Iterative regression imputation: Iterative regression imputation is used to estimate multiple values that reflect the uncertainty around the true value. As a result, the uncertainty about imputed values is carried out to our final inferences.

  Algorithm: - Add uncertainty/variation on the imputed dataset.

  - Perform analysis (ie. Mean, Simple random imputation,) on the imputed dataset.

  - Repeat this process.

  - Summarize the results to produce parameter estimate, standard error and other estimates and find a set of parameters that maximizes the probability of having seen the observed data.

  It is evident that multiple imputations requires many decisions and can be computationally intensive.