Statistical Methods - Computer Assignment

Thijs J. Quast

November 7, 2018

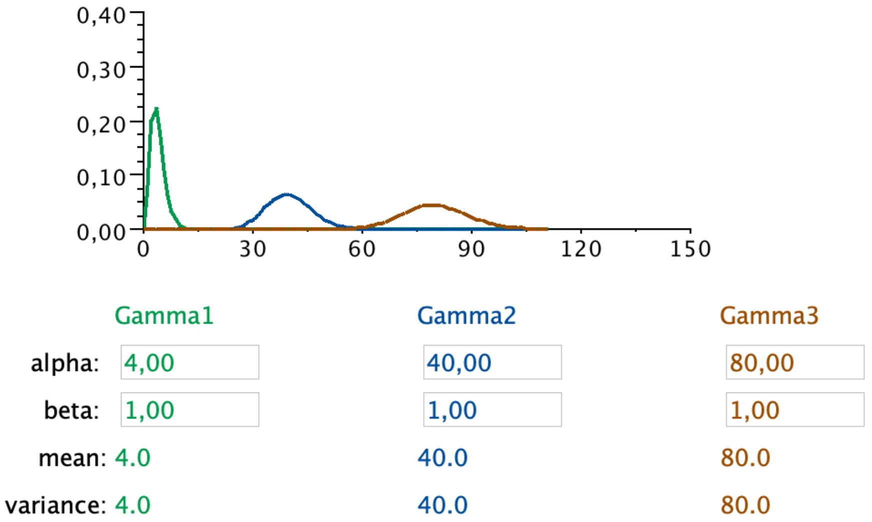
LiU ID: thiqu264
Course code: 732A93

Contents

1 Applet Exercises from Course's book

1.1 Exercise 4.84

Figure 1: Comparison of Gamma Density functions



	Gamma1	Gamma2	Gamma3
alpha:	4,00	40,00	80,00
beta:	1,00	1,00	1,00
mean:	4.0	40.0	80.0
variance:	4.0	40.0	80.0

a) The higher the alpha parameter becomes, the more skewed to the left the graph is, since the tail of the graph is more oriented towards the left of the distribution. The Gamma1 function is skewed to the right, with high concentration of density in the left side of the distribution, the distribution has a high but narrow curve. The Gamma2 density function is rather symmetrical and not very skewed. One can say it partially looks like a normal distribution. Gamma3 is skewed to the left, and has a flat shape.

b) The location of the centers differs among the three graphs. The center of the Gamma1 function is on the left side of the distribution. The center of the Gamma2 function is rather in the middle of the distribution, the center of the Gamma3 function is towards the right of the distribution.

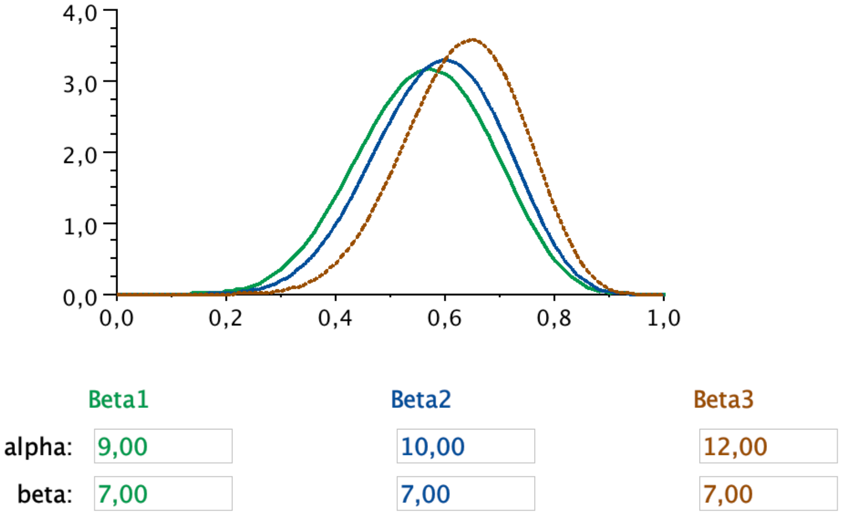
c) The differences in location of the centers of the distributions are determined by the way the alpha and beta parameters are set. The mean of a gamma distribution is computed as following:

$$Mean = \alpha\beta$$

In all three density functions the beta is set at a value of 1.0. Therefore, if the alpha increases, the mean of the distribution increases and thus the center of the density function shifts towards the right as the alpha parameter increases, conditionally that the beta parameter remains at a value of 1.0.

1.2 Exercise 4.117

Figure 2: Comparison of Beta Density Functions



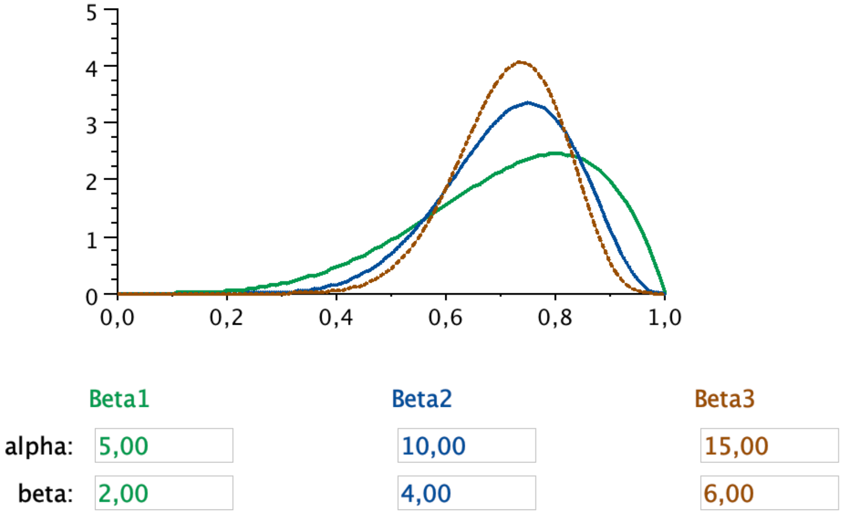
	Beta1	Beta2	Beta3
alpha:	9,00	10,00	12,00
beta:	7,00	7,00	7,00

a) The curves of the densities are symmetric. However, if one is very accurate, there is a slight skewness to the left.

b) As the alpha parameter gets closer to the value of 12, and the beta parameter remains at 7, the curve of the density shape shifts towards the right. This means the higher the alpha gets, when beta remains the same, the more skewed to the left the density function becomes.

c)

Figure 3: Comparison of Beta Density Functions, $\alpha > 1$, $\beta > 1$ and $\alpha > \beta$

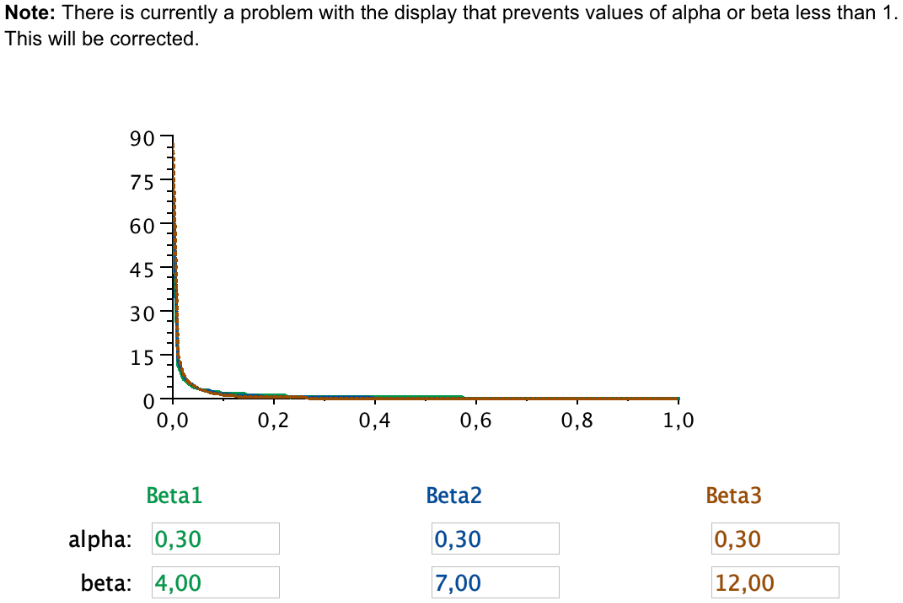


	Beta1	Beta2	Beta3
alpha:	5,00	10,00	15,00
beta:	2,00	4,00	6,00

In Beta1, with relatively low alpha and beta, the curve has more of bell shape and is skewed to the left. As I increase alpha and beta, as shown in Beta2 and Beta3, the density curve behaves more like a normal distribution, the curve increases in height, has thinner tails, and the center starts to orient more towards the left.

1.3 Exercise 4.118

Figure 4: Comparison of Beta Density Functions



Note: There is currently a problem with the display that prevents values of alpha or beta less than 1. This will be corrected.

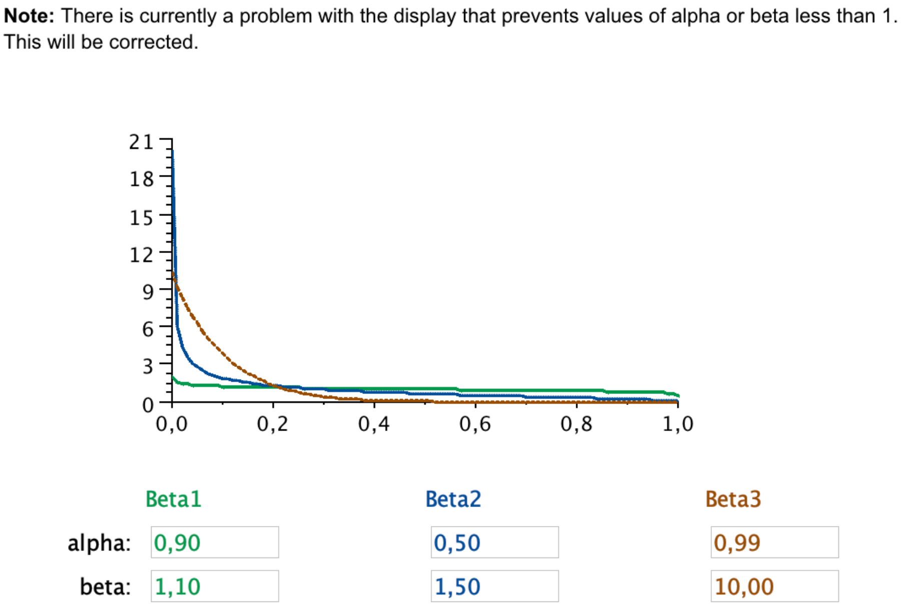
	Beta1	Beta2	Beta3
alpha:	0,30	0,30	0,30
beta:	4,00	7,00	12,00

a) No these densities are not symmetric. The tails of the densities are on the right, therefore, the density functions are skewed to the right.

b) I observe that all three density functions have a similar shape, therefore the increase of the beta is of little to no effect on the shape of the density curves.

c) Beta1, the tail of this distribution is the fattest, therefore the density is higher than the two other density functions for values greater than 0.2, and thus the probability of obtaining a value larger than 0.2 is the highest for Beta1.

Figure 5: Comparison of Beta Density Functions



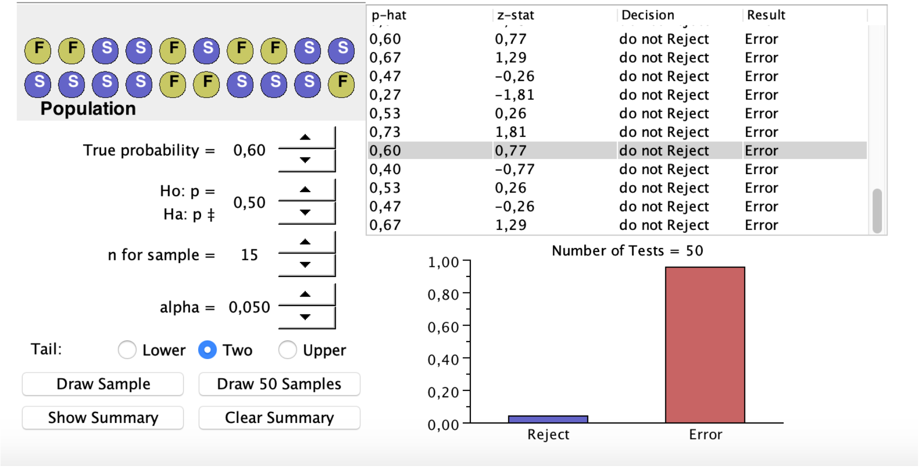
Note: There is currently a problem with the display that prevents values of alpha or beta less than 1. This will be corrected.

	Beta1	Beta2	Beta3
alpha:	0,90	0,50	0,99
beta:	1,10	1,50	10,00

d) As alpha is close to one, but smaller than one, and beta is greater than one, but close to one, the curve becomes rather flat. The larger the gap between the two parameters, with alpha <1 and Beta >1, the more the function gets the shape of an asymptote.

1.4 Exercise 10.11

Figure 6: Hypothesis testing for proportions



True probability = 0,60

Ho: p = 0,50
Ha: p ‡

n for sample = 15

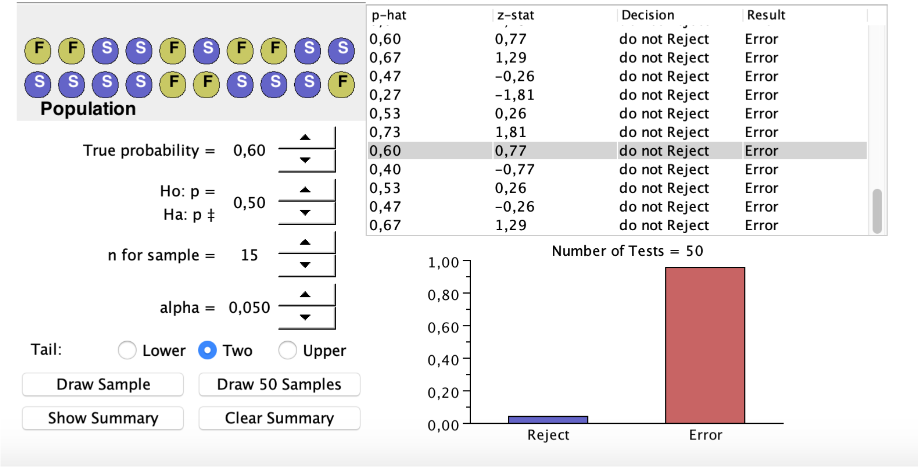
alpha = 0,050

Tail: ○ Lower ● Two ○ Upper

p-hat	z-stat	Decision	Result
0,60	0,77	do not Reject	Error
0,67	1,29	do not Reject	Error
0,47	-0,26	do not Reject	Error
0,27	-1,81	do not Reject	Error
0,53	0,26	do not Reject	Error
0,73	1,81	do not Reject	Error
0,60	0,77	do not Reject	Error
0,40	-0,77	do not Reject	Error
0,53	0,26	do not Reject	Error
0,47	-0,26	do not Reject	Error
0,67	1,29	do not Reject	Error

a) Type 2 error

Figure 7: Hypothesis testing for proportions, 200 simulations



True probability = 0,60

Ho: p = 0,50
Ha: p ‡

n for sample = 15

alpha = 0,050

Tail: ○ Lower ● Two ○ Upper

p-hat	z-stat	Decision	Result
0,60	0,77	do not Reject	Error
0,67	1,29	do not Reject	Error
0,47	-0,26	do not Reject	Error
0,27	-1,81	do not Reject	Error
0,53	0,26	do not Reject	Error
0,73	1,81	do not Reject	Error
0,60	0,77	do not Reject	Error
0,40	-0,77	do not Reject	Error
0,53	0,26	do not Reject	Error
0,47	-0,26	do not Reject	Error
0,67	1,29	do not Reject	Error

b) 92% of the tests resulted in type 2 errors. β is equal to the type2 error, 1-β is equal to the power of the test. The proportions of rejections is equal to 0,08 which is 1-β. Thus the type2 error is related to the number of rejections as together they add up to 1.

Table 1: Hypothesis testing summary table

True value	Null value	Sample size	N of Tests	Test Type	alpha	Reject rate	Error rate
0.60	0.50	100	200	Both	0.050	0.510	0.490
0.60	0,.50	50	200	Both	0.050	0.365	0.635
0.60	0,.50	30	200	Both	0.050	0.175	0.825
0.60	0.50	15	200	Both	0.050	0.080	0.920

c) As the sample size increases, percentage of type2 errors increases.

d) Continue with exercise 10.12

1.5 Exercise 10.12

Table 2: Hypothesis testing summary table

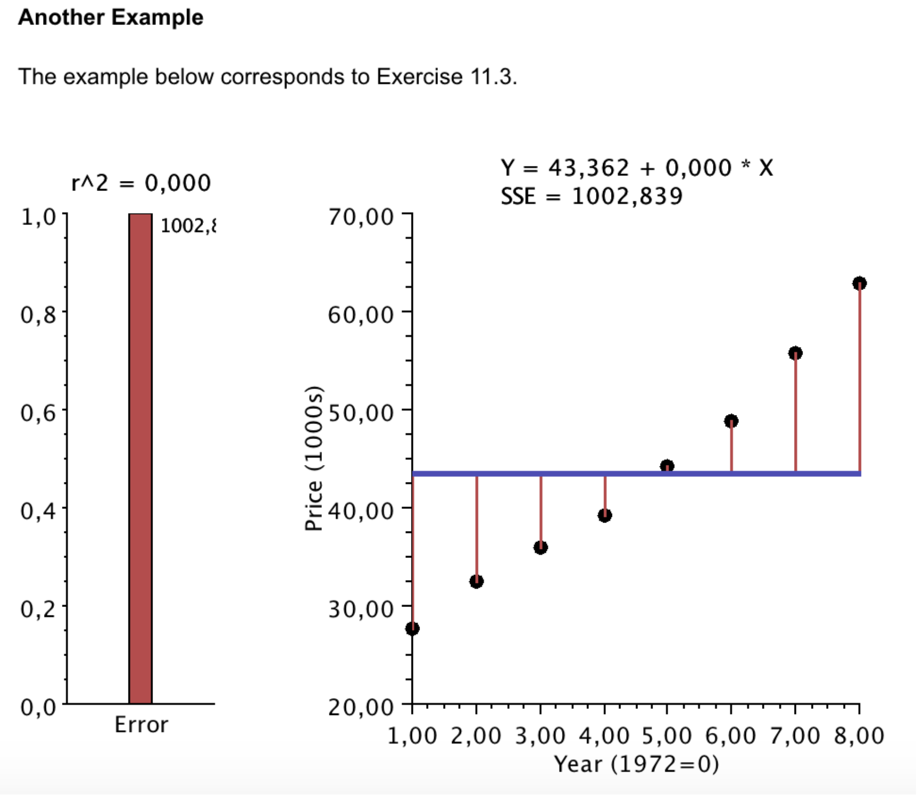
True value	Null value	Sample size	N of Tests	Test Type	alpha	Reject rate	Error rate
0.60	0.50	100	200	Both	0.050	0.510	0.490
0.60	0.50	50	200	Both	0.050	0.365	0.635
0.60	0.50	30	200	Both	0.050	0.175	0.825
0.60	0.50	15	200	Both	0.050	0.080	0.920
0.60	0.50	100	200	Both	0.100	0.625	0.375
0.60	0.50	50	200	Both	0.100	0.500	0.500
0.60	0.50	30	200	Both	0.100	0.295	0.705
0.60	0.50	15	200	Both	0.100	0.275	0.725

a) $\alpha = 0.05$

b) For n=100, smallest simulated value for β is when $\alpha = 0.05$. For n=50, smallest simulated value for β is when $\alpha = 0.05$. For n=30, smallest simulated value for β is when $\alpha = 0.05$.

1.6 Exercise 11.6

Figure 8: Linear model



Another Example

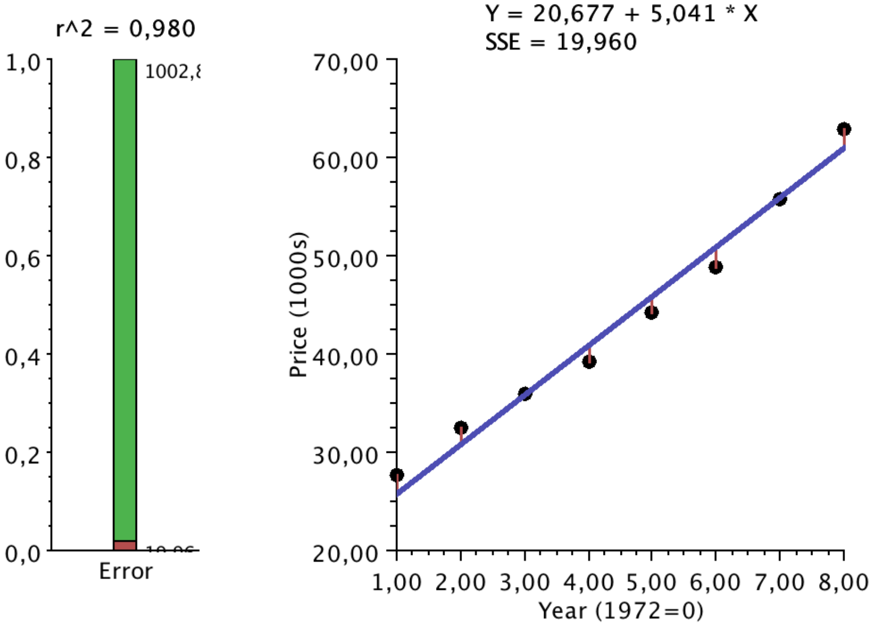
The example below corresponds to Exercise 11.3.

a) Intercept: 43,362

SSE: 1002,839

b) No, a line with negative slope would not fit the data well. This would increase the SSE.

Figure 9: Linear model



c) The line has the equation:

$$y = 20,677 + 5,401X$$

$SSE = 19,960$. Most likely, if the intercept and the slope of the line change, the line will fit the data less well. And therefore SSE will increase.

d) No, the line I manually find is not the least-squares line. When pressing the button “Find Best Model” the following equation is derived, which is the least-squares line:

$$y = 21,575 + 4,842X$$

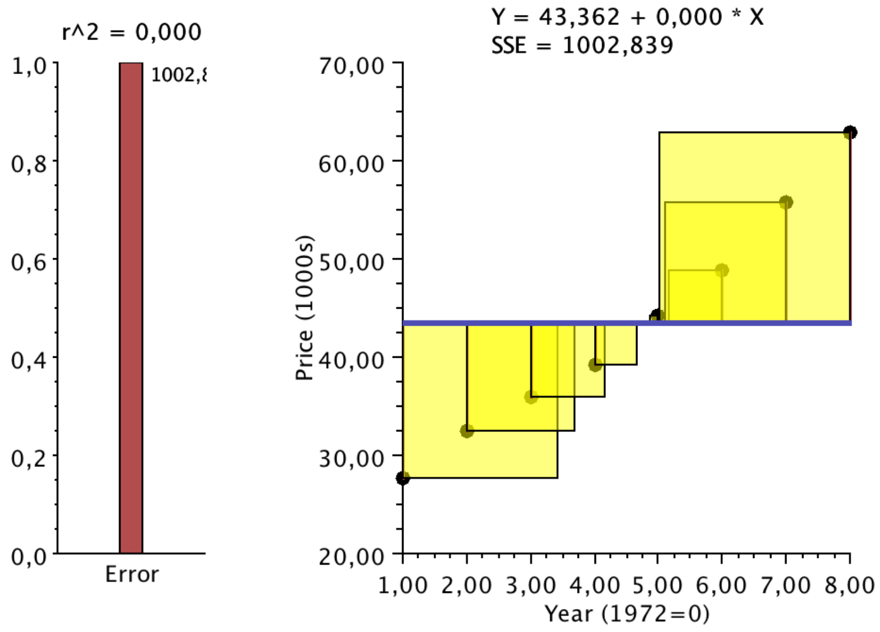
With $SSE = 18,286$. Compared to the line I manually fitted, the least-squares model has a higher intercept, however a lower/smaller slope. The SSE of the least-squares line is smaller than the SSE of the line I fitted manually.

e) Considering the least-squares line, the $SSE = 18,286$. To arrive at MSE, I apply the following formula:

$$MSE = s^2 = \frac{SSE}{n-2}$$

Resulting from this, the average error is $\sqrt{(18,286/(8-2))} = 55.21$. Therefore to estimate the y-coordinate of I take the x-coordinate of this point as input of the line. Resulting from this the fitted y-value is: $21,575+4,842(5) = 45,785$. From the graph it can be seen that this fitted line is slightly above the observed value. Therefore the mean error should be subtracted from the fitted value. Resulting from this, the y-coordinate of the observation around which the blue line pivots is $45,785 - 55.21 = 45,729.79$.

Figure 10: Sum squared errors, linear model



f) The further the vertical distance of an observation is to the fitted line with slope 0, the larger the yellow boxes become. The yellow boxes present the squared error terms for each observation. Therefore, the sum of all these boxes is equal to SSE, which is 1002,839

1.7 Exercise 11.50

a) As the fit of the line improves, the r^2 of the line also improves.

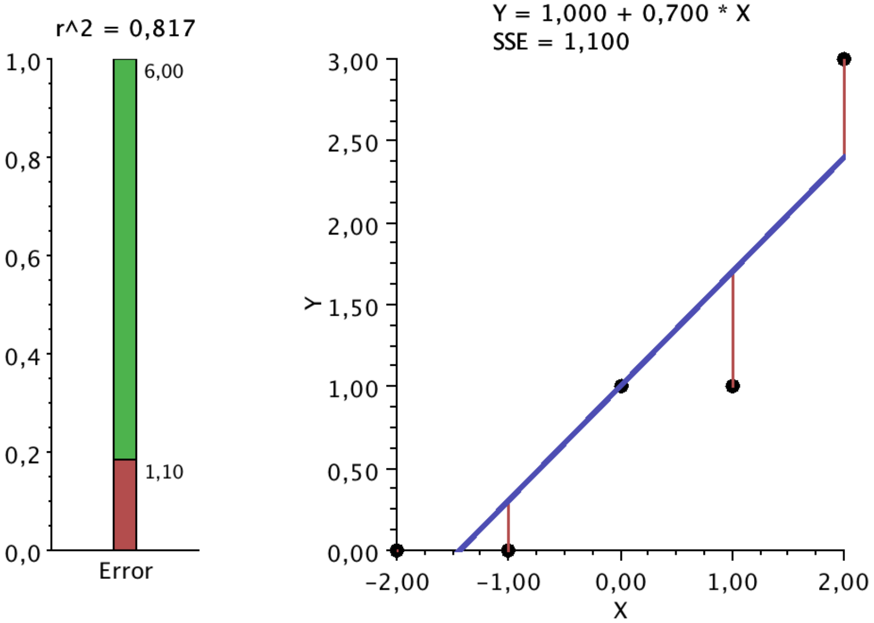
b) For the least-squares estimate: $r^2 = 0.982$.

$$Correlation = \sqrt{r^2} = r$$

Resulting from the formula above, the coefficient of correlation is equal to $\sqrt{0.982} = 0.991$.

c)

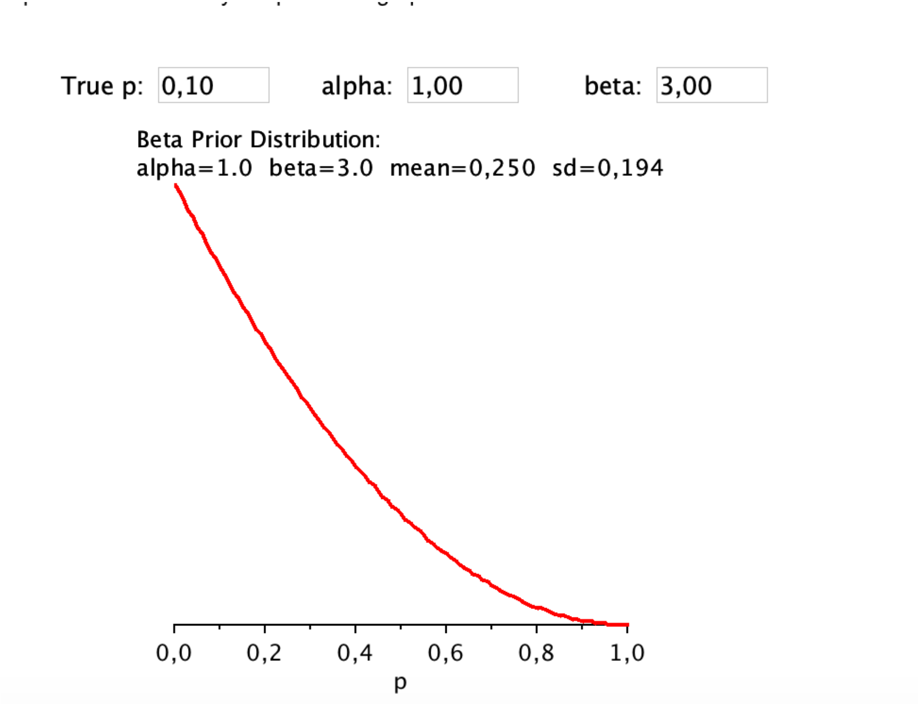
Figure 11: Two fitted linear models



The observations from the graph in exercise 11.49 are further apart from each other. Therefore the correlation between the two variables is weaker than in the graph in exercise 11.50. Resulting from this, the value for r^2 is lower in the graph from exercise 11.49.

1.8 Exercise 16.4

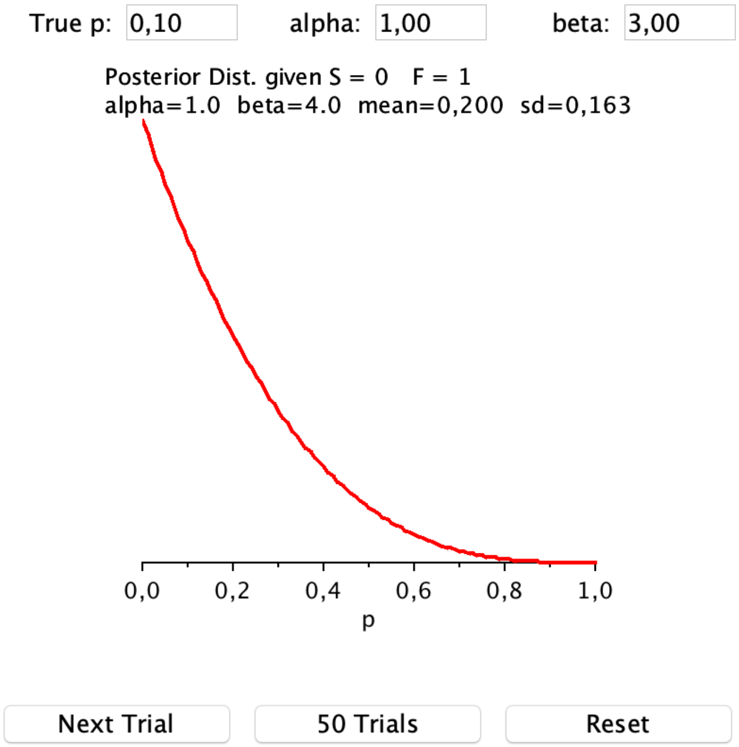
Figure 12: Beta distribution



True p: 0,10 alpha: 1,00 beta: 3,00

Beta Prior Distribution:
alpha=1.0 beta=3.0 mean=0,250 sd=0,194

Figure 13: Posterior Beta distribution

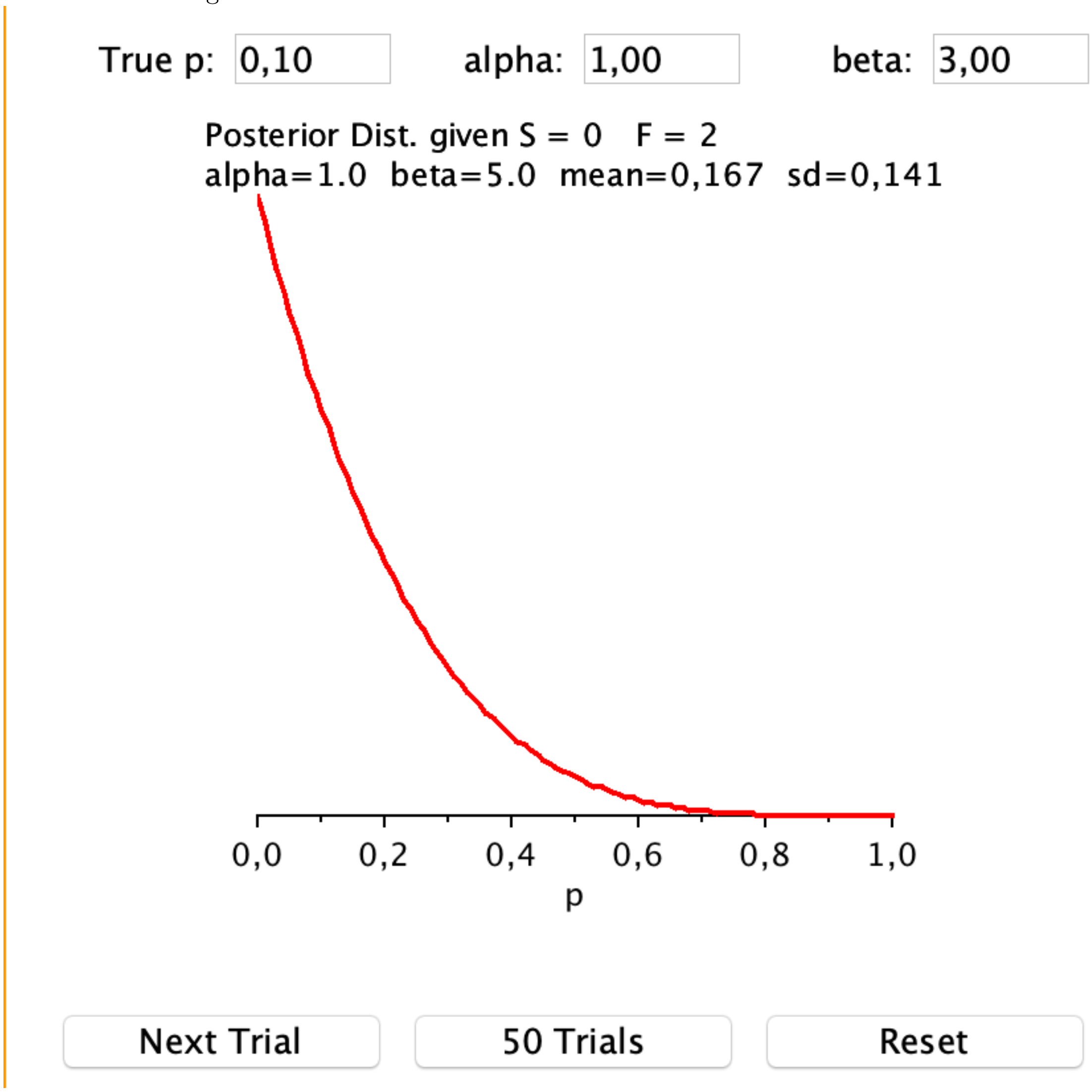


True p: 0,10 alpha: 1,00 beta: 3,00

Posterior Dist. given S = 0 F = 1
alpha=1.0 beta=4.0 mean=0,200 sd=0,163

a) A failure is observed. The graph looks different. I would say the curve in the posterior distribution is stronger.

Figure 14: Posterior distribution with two trials

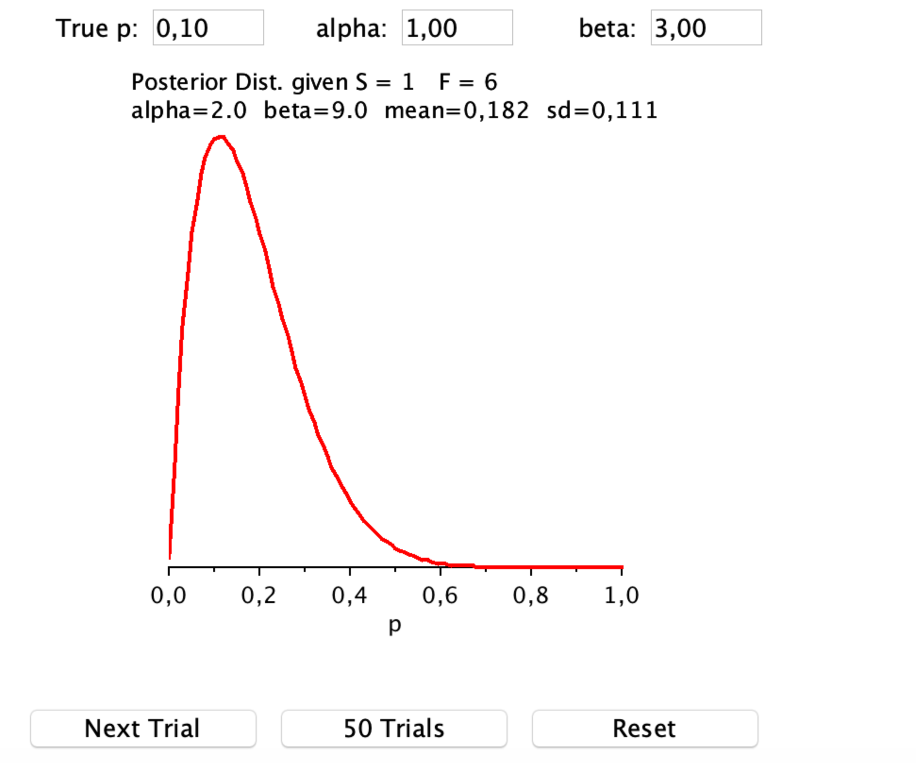


True p: 0,10 alpha: 1,00 beta: 3,00

Posterior Dist. given S = 0 F = 2
alpha=1.0 beta=5.0 mean=0,167 sd=0,141

b) Two failures out of two trials. The shape of the posterior is different than the shape of the previous posterior estimated. The curve of the second posterior graph is stronger again.

Figure 15: Beta distribution until first succes

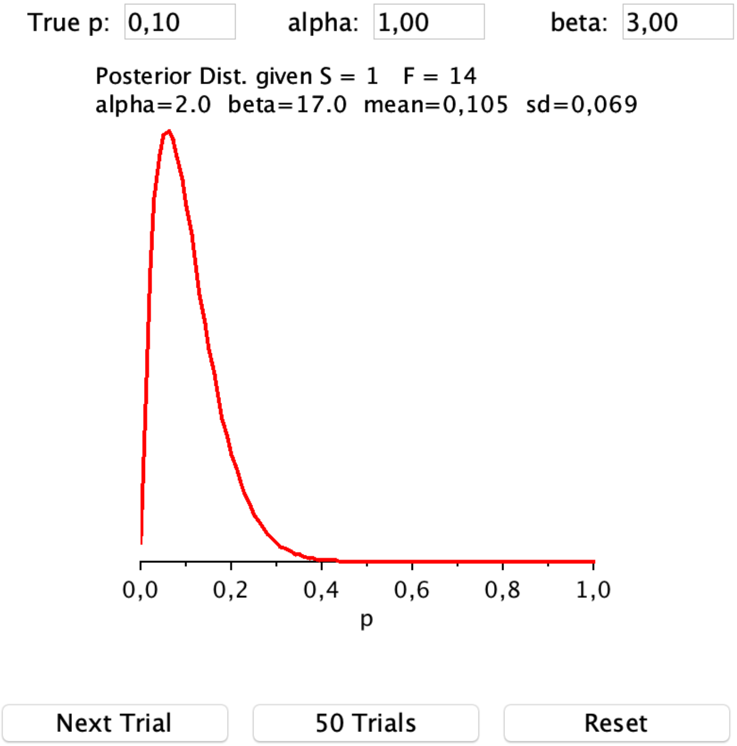


True p: 0,10 alpha: 1,00 beta: 3,00

Posterior Dist. given S = 1 F = 6
alpha=2.0 beta=9.0 mean=0,182 sd=0,111

c) Once a success is achieved, the shape of the graph changes drastically. Previously the shape was like an exponential distribution, whereas after achieving a succes the graph looks like a beta function.

Figure 16: Beta distribution, mean approximately 0.1



True p: 0,10 alpha: 1,00 beta: 3,00

Posterior Dist. given S = 1 F = 14
alpha=2.0 beta=17.0 mean=0,105 sd=0,069

d) A total of 15 trials are necessary in order to obtain a graph with mean equal to 0.105.

2 Imputation techniques

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

I would prefer *Missingness completely at random,* this is based on two reasons. Firstly, I believe Missingness that depends on unobserved predictors and missingness that depends on the missing value itself, are suitable when the data was obtained by means of survey. This is not always the case however, in many cases data is obtained from databases that are not based on surveys, in financial data for example. In such a case if an observation has missing data in my opinion this is just completely random. Also, my choice for this mechanism is based on what I believe will be relevant for the master in Statistics and Machine Learning at LiU. Given my prior experience in statistics and Machine Learning I know that several computation techniques rely on huge amounts of data (big data). I argue this data is mostly not obtained through surveys, as this will be difficiult to implement survey results from huge audiences. When some observations are missing at complete randomness, I argue it is valid to remove these observations, as most likely sufficient observations will remain to draw statistical inferences from. And larger standard errors because of a reduced sample size will not be a problem because of the amount of data available.

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

When using simple random imputation, one replaces the missing value of a variable by drawing a random value from the non-missing observations of that variable. This is not a very strong way of dealing with missing data, as it imputes a random number into an observation, this random number is irrevevant to other variables of the observation. In real life this is not realistic, as most likely the value of a variable of an observation is influenced by other variables of the same observation.

A better way of dealing with missing data for observations is to fit a linear regression model. The literature mentions a way of data transformation which included the predictive power of the model, which I think is very strong. In the example income and working ours are topcoded, as these outliers will strongly influence results. After preparing the data, the software estimates a linear regression model through the available data. Missing observations are then replaced by predictions of the model.

3.Explain shortly what Multiple Imputation is.

Multiple Imputation does not replace a missing value with one imputed value, however it replaces it with multiple values. Afterwards, based on these different datasets (based on the different imputed values for missing observations). Then on each of these datasets models are ran, afterwards e.g. coefficients on the model are determined by taking an average of all computed coefficients from the different datasets.