ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Master project

Gene-Disease Association Studies

Student:
Andrei Halasz

Professor:
Stephan Morgenthaler

June 3, 2011

Contents

Introduction

Gene-disease associations studies are performed in an increasing number since the creation of the HapMap Project. These studies are used to investigate regions or genes in the genome for which one has an indication or a belief that they can be associated with a particular disease. The aim is to identify the genes that cause the disease. The approach that is commonly used in such studies is the two-stage design method. Mainly, the report is divided in three parts.

In the first part we do the simulation of the two-stage design in the case of continuous measures in order to see if we can detect a surplus of mutations in the sick people. We then try to find the optimal parameters for the study (number of hypotheses to test, significance levels, sample size) for different values of the effect. A positive effect means in our case the detection of a marker (gene) that is associated with a disease. Since we are doing multiple testing we will also test which of the False Discovery Rate method or the Bonferroni correction are better from the point of view of the cost of the study, the proportion of false positives and of the power. Here the power is defined as the probability of identifying at least one true gene that causes a disease. We then do the same simulation in the case where we observe a discrete endpoint. We show that in both case we arrive at the same conclusions regarding the optimality of the parameters and the best correction method for the multiple comparison.

In the second part we simulate the evolution of mutations in the human population. The aim of this simulation is to construct the mutational spectrum for different values of the mutation rate. In order to do so, we will consider several models for the evolution of the population and also for the evolution of alleles. We start by supposing that we are in the infinite alleles model and then we add the possibility that an allele might disappear from a generation to another one.

In the third and last part we use two methods to estimate the effect with the help of the mutational spectrum. We then combine the obtained information with the results from the first part in order to see if the effect has a non-zero probability of being detected.

Model

A variety of methods for population-based gene-disease association studies have been proposed in the literature. The aim of such studies is to identify genes, or genetic regions in the human genome that favor the development of particular diseases.

A popular approach is to use the two-stage design method. This method can be described as follows (we will also give the notation that we use in our study). We assume that we have a number H of hypotheses that we want to test. An example of such a hypothesis could be: a certain gene or marker has no effect on a particular disease. A number F of these H hypotheses are the false hypotheses and usually F is much smaller than H (F<<H). In a biomedical language, H would be the total number of markers and F would be the true markers of risk. We also assume that we have a number n of subjects on which we evaluate the markers. We investigate a simple situation in which evaluation of a marker results in a normally distributed observation with variance 1. We will summarize the measurements from different subjects by computing averages.

In the first stage we evaluate all H hypotheses on a number n_1 of individuals ($n_1 < n$). We select all the rejected hypotheses R1. Thus, R1 is the number of hypotheses that were selected for the second-stage. At stage 2 we evaluate these R1 hypotheses on a number $n_2 = n - n_1$ of individuals. At the end we select the number of hypotheses that were truly rejected, number that will be denoted by R2 (rejected at stage 2).

Several studies were done using the two-stage design and it has been shown that, when the total cost of the study is the main constraint, the two-stage method is better than the one-stage method (see the reference [1]). The cost is defined by the total number of marker (gene) evaluations that are done during the study. If the total number of subjects is fixed, the two-stage design is still a better method than the one-stage design (see the reference [2]). However, if no constraints are imposed, the one-stage design has a bigger power than the two-stage design.

We are interested in evaluating the power for this design. The power is equal to the probability of detecting one of the hypotheses from F. So, in this case, the power is given by P[at least one among F is in R2]. The bigger the power, the bigger is the probability of detecting an effect, i.e., of detecting a marker that is associated with a disease. Moreover, we are also interested in calculating the cost of the test. The cost is given by H$\cdot n_1$+R1$\cdot n_2$.

We evaluate our design using simulations (for results, see next section). In the first stage we simulate H normal test statistics, where the first H$-$F are $T_i \sim \mathcal{N}(0, 1/n_1)$ and the remaining F are $T_j \sim \mathcal{N}(\Delta, 1/n_1)$, where $\Delta > 0$ is

the effect. We want to test the hypothesis that the effect is null against the alternative that the effect is positive. This is done by calculating the p-value for each hypothesis in order to determine the significance of the test. The p-values are computed as $p_i = 1 - \Phi(\sqrt{n_1}T_i)$, for $i = 1, \ldots, \mathrm{H}$. If the p-value is smaller than a certain significance level, we reject the hypothesis H_0: $\Delta = 0$. In this case we can say that there is an association between the marker and the disease. For the stage 1, the significance level of the individual tests, which is the probability of finding a false positive, is chosen as α_1. It follows that R1 contains all the hypotheses i for which

$$T_i > \frac{z_{1-\alpha_1}}{\sqrt{n_1}}.$$

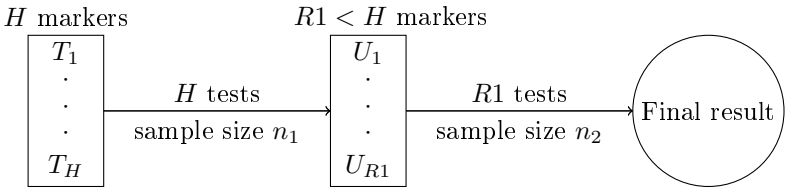
All the hypotheses that were rejected in the first stage are kept for further analysis in the second stage. Let us denote by F2 the number of truly false hypotheses that were rejected in the first stage.

In the second stage, we introduce, as before, R1 test statistics S_i, where the first R1$-$F2 are random normal variables with mean 0 and variance $1/n_2$ and the remaining F2 are random normal variables with mean Δ and variance $1/n_2$. For the markers, which make it to the second stage, we combine the first stage T_i and the second stage S_i as follows:

$$T_i \frac{n_1}{n_1 + n_2} + S_i \frac{n_2}{n_1 + n_2} = U_i.$$

We can now compute the p-values and compare them to the significance level for the second stage. An illustration of the two-stage design is presented in Figure 2.1.

Figure 2.1: Two-stage design



We remark that in the first stage we take a significance level that is bigger (less restrictive) than in the second stage because we do not want too many of the false hypotheses to escape rejection. Finally, the two-stage model can be seen as a model where in the first stage we select the hypotheses that could indicate an effect and in the second stage we validate the hypotheses that really have an effect.

2.1 Correction of p-values in the case of multiple testing

In genome-wide studies, H is very big (multiple testing) and some restriction in order to avoid excessive number of false rejections are necessary. For example, if we take a significance level α of 0.05 and we have 30 hypotheses, then we will reject at least one true hypothesis with probability

$$\begin{aligned} P(\text{rejecting at least one true hypothesis}) &= 1 - P(\text{rejecting no true hypothesis}) \\ &\geq 1 - (1 - 0.05)^{30} = 0.79. \end{aligned}$$

On average, (H-F)$\times$0.05 of the true hypotheses will be rejected. This number increases with the number of hypotheses that we want to test and it is much too large. We thus have to adjust the p-values to correct for occurrence of false positives. Two methods that are widely used to correct this problem are the False Discovery Rate (FDR) method and the Bonferroni correction.

The Bonferroni correction is used to control the Familywise Error Rate (FWER). The FWER is defined as the probability of detecting at least one false positive among all true hypotheses. This is equivalent to saying that it is the probability of making at least one type I error. The Bonferroni correction tests each individual hypothesis at a significance level of α/H. It is easy to show that if we do so, then the FWER is bounded above by α. More precisely, if we have

$$P(\text{i-th hypothesis is rejected}|\text{i-th hypothesis is true}) = \frac{\alpha}{H}, \text{ for } i = 1, \ldots, H,$$

it follows that

$$P(\text{at least one hypothesis is rejected}|\text{all hypotheses are true}) \leq \alpha,$$

because the probability on the left can by bounded by the sum over all hypotheses of the above probabilities.

The False Discovery Rate is defined as the expected proportion of erroneously rejected null hypotheses. More precisely, if we denote by R all the rejected null hypotheses and by FP the number of incorrectly rejected null hypotheses or also known as the number of false positives, then the FDR is defined as

$$\text{FDR} = E\left(\frac{\text{FP}}{R}\middle| R > 0\right) P(R > 0).$$

We now describe the controlling procedure for the False Discovery Rate. Let us consider H independent tests. We denote by $H_1, \ldots, H_H$ the null hypotheses and by $P_1, \ldots, P_H$ the corresponding p-values for each test. Let $P_{(1)}, \ldots, P_{(H)}$ be the ordered p-values and denote by $H_{(i)}$ the null hypothesis that corresponds to $P_{(i)}$. Find the largest k such that, for a given significance level α,

$$P_{(k)} \leq \frac{k}{H}\alpha$$

and reject all the null hypotheses $H_{(i)}$ for $i = 1, \ldots, k$. It then follows that FDR$\leq \alpha$. For more details on how to control the FDR in the case of a two-stage design see Yoav Benjamini et al. (1995) ([3]).

In the simulation study we will try to find which of these two methods is better with regard to the power, the proportion of false positives and the cost of the study in the case of a two-stage design.

Simulation of the two-stage design

In this section we do a simulation study for the two-stage design described in section 2. The parameters used in the simulation are the total number of hypotheses H, the number of false hypotheses F, the sample size for the first stage n1, the sample size for the second stage n2, the significance level for the first stage α_1, the significance level for the second stage α_2 and the effect Δ. When not specified, we take H= 1000, F= 50, $n_1 = 5$, $n_2 = 30$, $\alpha_1 = 0.1$ and $\alpha_2 = 0.01$. We repeat each test 200 times so the values that we will present are in fact averages over 200 replications.

As we said in the previous section, we are interested in analyzing the power, the proportion of false positives and the cost for different values of the parameters used in the study. We will also try to find out, which of the Bonferroni correction or the FDR method is better for the two-stage design.

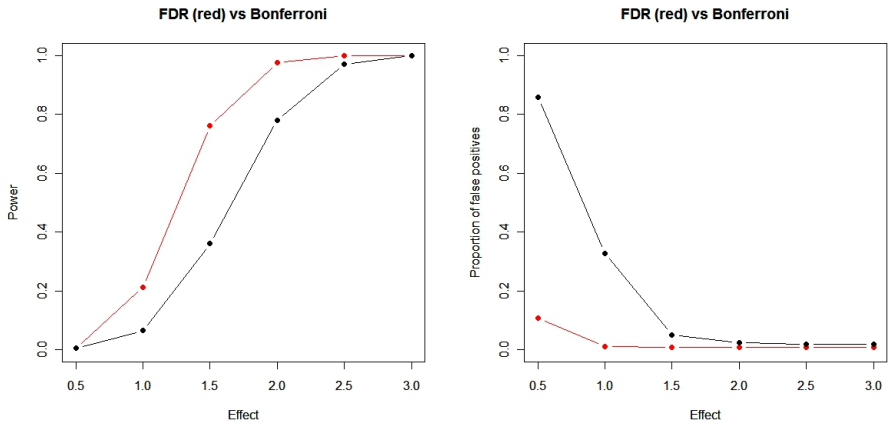
In the first part of this section we will run the simulation in the case where the test statistics are supposed to have an approximately Normal distribution. In the second part of this section we will complicate the design and suppose that the test statistics have a contaminated normal distribution.

3.1 Test statistics with a Normal distribution

We start by doing a simulation in order to see which of the Bonferroni correction or the FDR method gives a bigger power when the effect is changing. In Figure 3.1 we plotted the power and the proportion of false positives for the Bonferroni method and for the FDR method for different values of the effect. We can see that the power when using FDR is bigger than the power when using the Bonferroni correction. We also notice that the Bonferroni method leads to a bigger proportion of false positives than the FDR method.

In the previous figure the same method was used for both stages in the design. We would be interested in studying the design when we use the FDR method in the first stage and the Bonferroni or the FDR method in the second stage. This might be suggested by the fact that we are more interested in the correction method in the second stage and not in the first. Since FDR is less restrictive than Bonferroni, more hypotheses will go on to the second stage. We would like to see how the power is affected when we have more hypotheses to evaluate in stage 2 of the test.

Figure 3.1: Power and proportion of false positives using Bonferroni correction and FDR method. The significance levels are $\alpha_1 = 0.1$ and $\alpha_2 = 0.01$.



In Figure 3.2 we plotted the power and the proportion of false positives for different values of the effect when using FDR in the first stage and FDR or Bonferroni correction in the second stage. We notice that we obtain almost the same power when doing FDR in stage 1 and Bonferroni in stage 2 as when doing FDR in both stages. Moreover, for effects smaller than 1.5, using Bonferroni in the second stage gives a bigger proportion of false positives than using FDR in the second stage. For effects larger than 1.5, both methods give almost the same proportion of false positives. Since we are also interested in detecting smaller effects, in the following we will consider only the tests made with the FDR method in both stages.

We have previously shown which of the two methods is better in detecting an effect. We now try to find the best parameters in order to have an optimal power and a small number of false positives. We begin with choosing optimal significance levels for both stages.

Figure 3.2: Power and proportion of false positives using Bonferroni correction and FDR method in the second stage while using FDR in the first stage

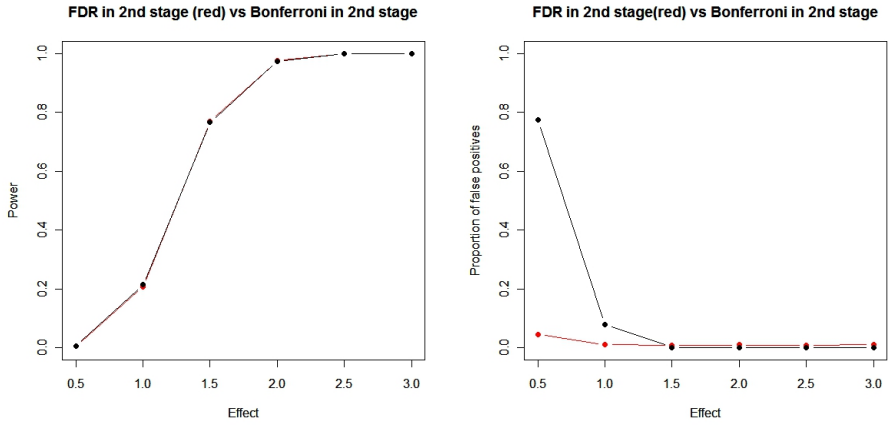


Figure 3.3: Power using different values of α_1 (left panel) and α_2 (right panel)

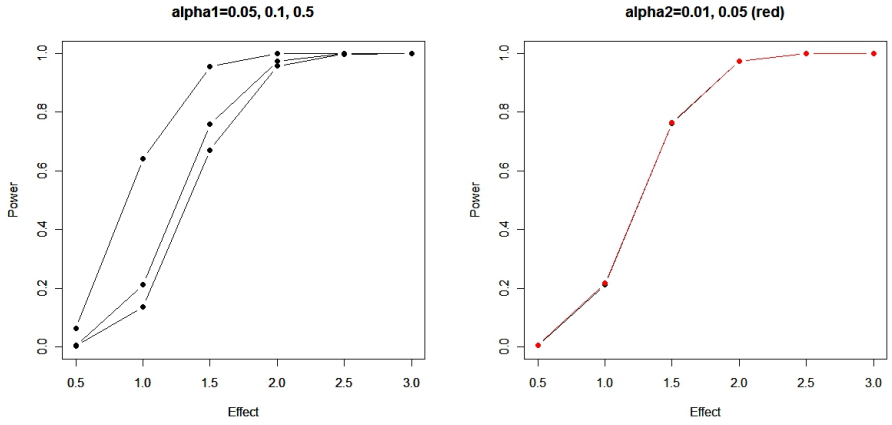
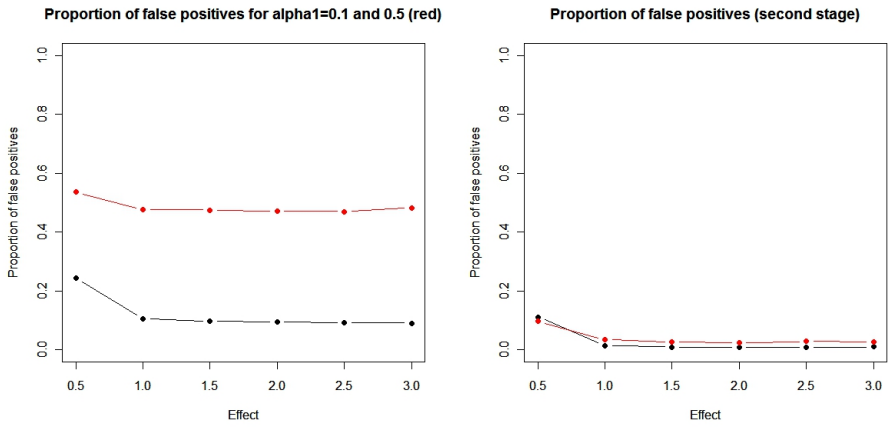


Figure 3.4: Proportion of false positives using different values for α_1 in the first stage (left panel) and in the second stage (right panel)



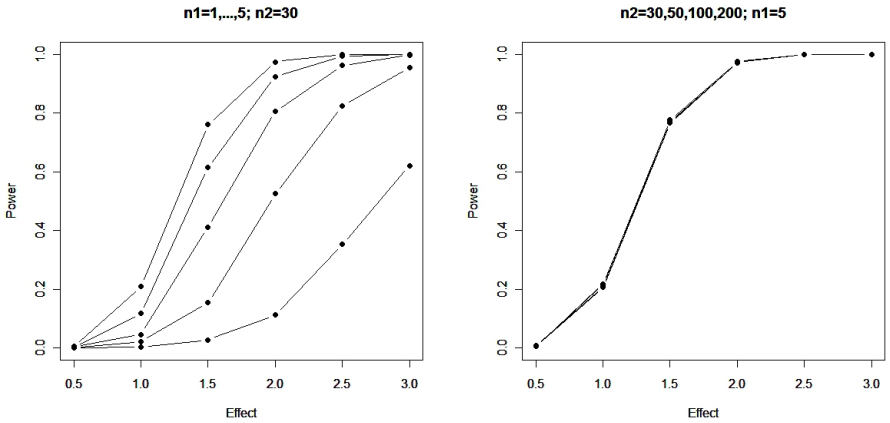
In Figure 3.3 we plotted the power for different values of α_1 and α_2. We observe that when the significance level for the first stage increases, the power also increases. For α_2 we have almost the same power for both values, so we will take the most restrictive one, which is $\alpha_2 = 0.01$. Since $\alpha_1 = 0.1$ and $\alpha_1 = 0.5$ have bigger power than $\alpha_1 = 0.05$, we will take a look at the proportion of false positives to determine which value is better for our test.

In Figure 3.4 we plotted the proportion of false positives in the first stage and in the second stage for different values of α_1. We observe that, when $\alpha_1 = 0.5$, the proportion of false positives is augmented a lot in the first stage and remains bigger in the second stage compared with $\alpha_1 = 0.1$.

Since we want to have a number of false positive that tends to zero, we prefer taking $\alpha_1 = 0.1$ even if its power is smaller than the power when $\alpha_1 = 0.5$ for small effects.

Now we will take a look at the power when the sample sizes are changing. The results are shown in Figure 3.5. We notice that when n_1 increases the power also increases, while when n_2 increases significantly, the power rests almost the same. This is linked to the cost of the process, which is equal to H$\cdot$ n_1+R1$\cdot$ n_2. Since H>R1, n_1 has more influence than n_2. Thus, taking $n_2 = 30$ will give almost the same power as taking $n_2 = 200$. Since it is not always easy to find 200 individuals for a test, $n_2 = 30$ is an optimal sample size for the second stage.

Figure 3.5: Power using different values of n_1, n_2



One can ask what happens when the number of hypotheses that we want to test is changing. In this purpose, in Figure 3.6 we plotted the power and the proportion of false positives for different number of hypotheses. We remark that when the number of hypotheses increases, the power decreases. In the right panel we see that for a small effect, the proportion of false positives is increasing when H increases. We have the same observation for a bigger effect, but this time the increase in the proportion of false positives is smaller. Thus, since we want a small number of false positives, we conclude that using H=1000 is a better value than the two others.

Figure 3.6: Power and proportion of false positives using different values of H

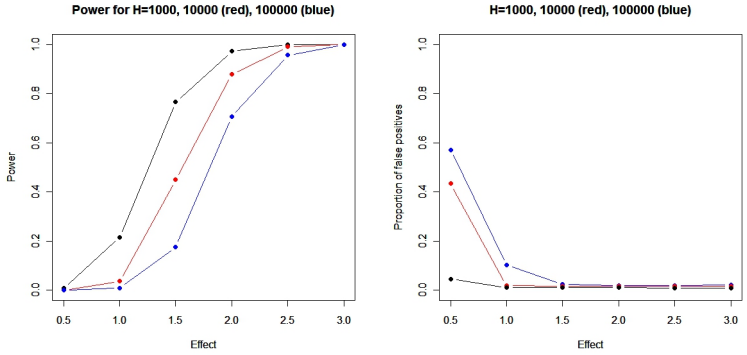
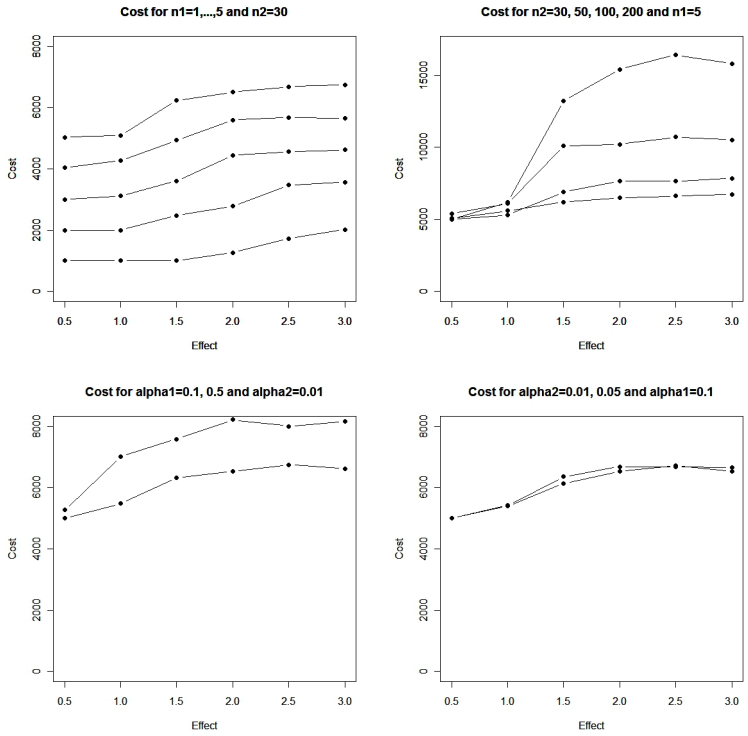


Figure 3.7: Cost for different values of n_1, n_2, α_1 and α_2

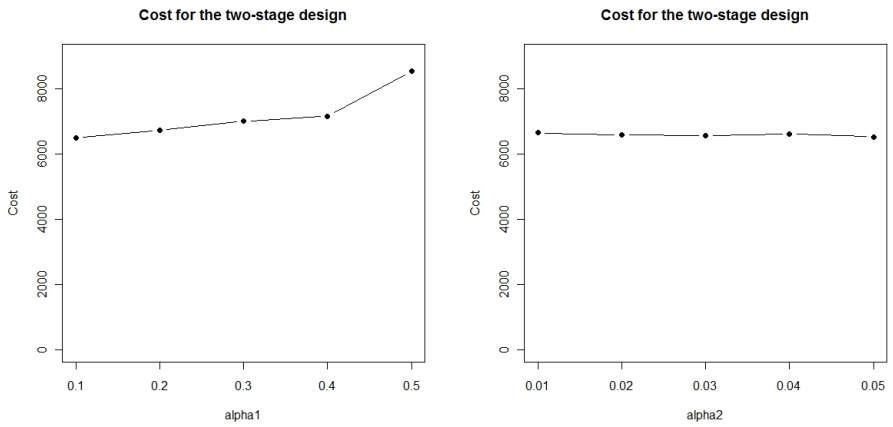


We looked at the behavior of the power when the model's parameters are changing and found that the values with which we started the simulation are in fact good values for our design. We now analyze the cost of the test when some of the parameters are changing. In Figure 3.7 we computed the cost for different values of n_1, n_2, α_1 and α_2.

We observe an increase in the cost whenever n_1, n_2 and α_1 are increasing. Thus, from the point of view of the cost, $\alpha_1 = 0.1$ is better than $\alpha_1 = 0.5$. This is the same conclusion that we made from the point of view of the power. From the last panel in Figure 3.7 we see that α_2 has almost the same cost when it is increasing. Thus we can say that the significance level for the second stage has a small impact on the value of the total cost of the test. We conclude, as we did from the perspective of the power, that $\alpha_2 = 0.01$ is the optimal value for this parameter.

In order to see if the conclusions made before are correct, we plot the cost of the two-stage design for fixed values of n_1, n_2, H and effect $= 2$ as function of α_1 and α_2. The results are shown in Figure 3.8. We notice that there is an important difference in the cost when $\alpha_1 = 0.1$ and when $\alpha_1 = 0.5$. We thus

Figure 3.8: Cost for different values of n_1, n_2, α_1 and α_2



obtain the same results as in the previous figure. Thus a smaller α_1 will result in a smaller cost and in a more selective method. For α_2, the cost is slightly diminishing when the value of the significance level is increasing. Since the variations are small, we will take the more restrictive α_2, which is 0.01. These are the same conclusions as before.

3.2 Contaminated Normal Distribution

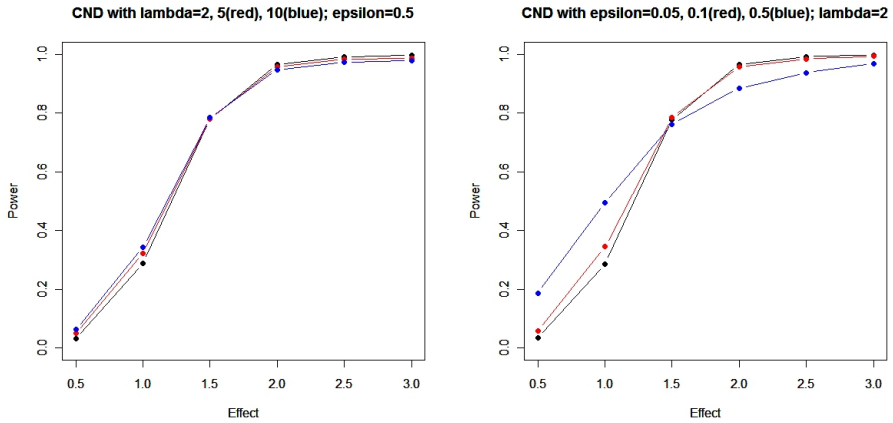
Until now we considered that the test statistic used in the simulation were approximately Normal distributed. We now complicate the study and suppose that the distribution is not approximately normal and it is of the form

$$(1-\varepsilon)\mathcal{N}(\mu,\sigma_1^2)+\varepsilon\mathcal{N}(\mu,\lambda\sigma_1^2),$$

where $\lambda > 1$ and $0 < \varepsilon < 1$. This is called the contaminated normal distribution (CND) and it has heavier tails than the Normal distribution. An example of the use of the CND is when blood pressure is calculated in a population. The males could have a normal distribution, the females could also have a normal distribution, but the two distributions don't have the same variance. The mixture of these two distributions is not a normal distribution.

In Figure 3.9, we simulated a two-stage experimental design with CND for different values of λ and ε.

Figure 3.9: Power when CND for different values of λ, ε



We notice that when λ increases, up to an effect of 1.5 the power also increases, but for effects bigger than 1.5 we have the opposite result. The same remark can be made for the probability parameter ε. In the following computations, we took $\lambda = 2$ and $\varepsilon = 0.05$.

In Figure 3.10 we plotted the power and the proportion of false positives for different number of hypotheses. We observe that the power decreases when the number of hypotheses increases. This is the same conclusion that we made when the test statistics were following apporximately a Normal distribution. However, when passing from H=10000 to H=1000000, the decrease in the power is less significant compared with a Normal distribution (see Figure 3.5). We also notice that the proportion of false positives increases when the number of hypotheses that we want to test is increasing. Thus, as in the Normal case, H=1000 is a reasonable choice.

We remark that in Figure 3.10 we used the same proportion, $\varepsilon = 0.05$, for each number of hypotheses. This means that we had 50 for H=1000, 500 for H=10000 and 5000 for H=100000 distributions that contaminated our original Normal distribution. We are interested in observing the behavior of the power and of the proportion of false positives when we have the same number, 50, of distributions that contaminates the original one. We present the results in Figure 3.11.

Figure 3.10: Power and proportion of false positives using CND for different values of H

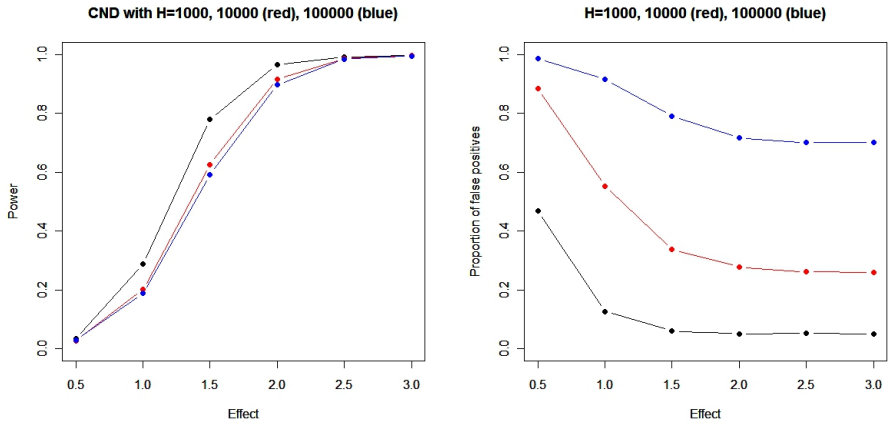
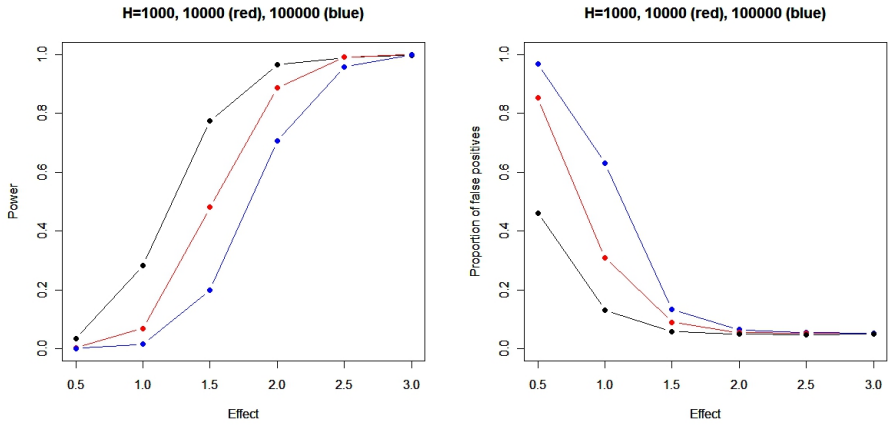


Figure 3.11: Power and proportion of false positives using CND with 50 contaminated distributions for different values of H

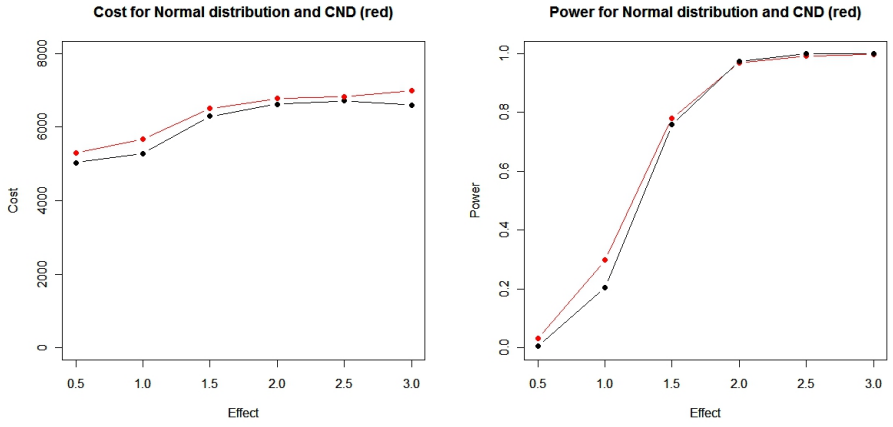


We notice that in this case, the power has a behavior similar with that of a Normal distribution. There is a bigger difference in power when the number of hypotheses is increasing. Moreover we see that for H=10000 and H=100000, we have almost the same values for the power as in the case of a Normal distribution. This might be due to the fact that in these cases the proportion of contaminating distributions is of 0.005 and 0.0005 respectively, which is quite small. From the

right panel we observe that the proportion of false positives decreased for effects larger than 0.5 comparing with the same plot from Figure 3.10.

In Figure 3.12 we do a comparison between the Normal distribution and the Contaminated Normal distribution from the point of view of the cost (left panel) and of the power (right panel). In the left panel, we see that the cost is slightly bigger when using the CND than when using the Normal distribution. This is due to the fact that when using the CND, the number of rejected hypotheses is bigger than the number of rejected hypotheses while using the Normal distribution. For the power, we notice that, for a small effect (up to 1.7 approximately), the power is bigger for the CND than for the Normal distribution. However, for a bigger effect we have the opposite result, but the difference between the two is small.

Figure 3.12: Comparison between the Normal distribution and the CND



Simulation when observing a discrete endpoint

4.1 Theoretical design

In the previous section we did a simulation study for the two-stage design using continuous measures. However, sometimes the measures that are taken, for example from a clinical trial, are discrete.

Table 4.1: Example of contingency table

Number of mutations							
H-F				F			
1	0	1	$\cdots$	1	1	1	
1	1	0	$\cdots$	1	1	1	M
0	0	1	$\cdots$	1	0	1	
.	.	.	$\cdots$	.	.	.	
.	.	.	$\cdots$	.	.	.	
.	.	.	$\cdots$	.	.	.	S
0	0	1	$\cdots$	0	1	0	
1	0	0	$\cdots$	0	0	1	
0	1	1	$\cdots$	1	0	0	

The aim of this section is to simulate a two-stage design in the discrete case. Imagine that we have a contingency table filled with 0 and 1, where 1 means that the mutation is present and 0 means that the mutation is absent. The number of lines in the table represent the number of patients. The first M lines are the sick patients and the rest of the lines, in number of S, are the healthy patients. The columns represent the mutations. We have H$-$F mutations where there is no effect present and F mutations where we have a positive effect Δ on the sick people. The effect will increase the number of mutations. An example of such a contingency table is presented in table 4.1.

We are interested in seeing if we can associate mutations with the disease. In order to do this we compute the mean for the sick patients and for the healthy patients. Mathematically we write this like

$$\hat{p}_{j1} = \frac{\sum_{i \in M} x_{ij}}{M} \quad \text{and} \quad \hat{p}_{j2} = \frac{\sum_{i \in S} x_{ij}}{S},$$

for j= $1, \ldots, H$, where x_{ij} is the element in the contingency table corresponding to the i^{th} line and j^{th} column.

We want to test the Null hypotheses $H_{j0} : \hat{p}_{j1} = \hat{p}_{j2}$ against the Alternative: $\hat{p}_{j1} = \hat{p}_{j2} + \Delta$. In order to do these tests we compute the Chi-squared test statistic

$$T_j = \frac{(\hat{p}_{j1} - \hat{p}_{j2})^2}{\frac{\hat{p}_{j1}(1-\hat{p}_{j1})}{M} + \frac{\hat{p}_{j2}(1-\hat{p}_{j2})}{S}},$$

for j= $1, \ldots, H$.

With these test statistics we can now compute p-values as

$$1 - \chi_1^2(T_j)$$

and compare them with some significance level. If the p-value is smaller than the significance level we reject the null hypothesis in favor of the alternative.

4.2 Results

As we did in the continuous mode, since we have multiple tests we correct the significance level using the Bonferroni correction of the FDR method. In this simulation we used, when not specified, H=2000, F=50, S=900, M=100, $\varepsilon = 0.05$, $\alpha_1 = 0.1$ and $\alpha_2 = 0.01$, where ε is the probability of having a mutation and α_1 and α_2 are the significance levels for the first stage and the second stage respectively. We repeated the simulation 100 times and took the mean of these results.

In Figure 4.1 we plotted the power and the proportion of false positives when using the Bonferroni correction and the FDR method in both stages.

We observe that for an effect up to 0.3, FDR has a bigger power than Bonferroni. For an effect bigger than 0.3, both methods have a power of approximately 1. Thus we see that we should be more interested in small effects than in bigger ones. For an effect between 0.1 and and approximately 0.2, FDR has a smaller proportion of false positives than the Bonferroni correction. For an effect larger than 0.2 we have that the FDR method has more false positives than the Bonferroni correction.

Figure 4.1: Power (left panel) and proportion of false positives (right panel) for the Bonferroni correction and the FDR method

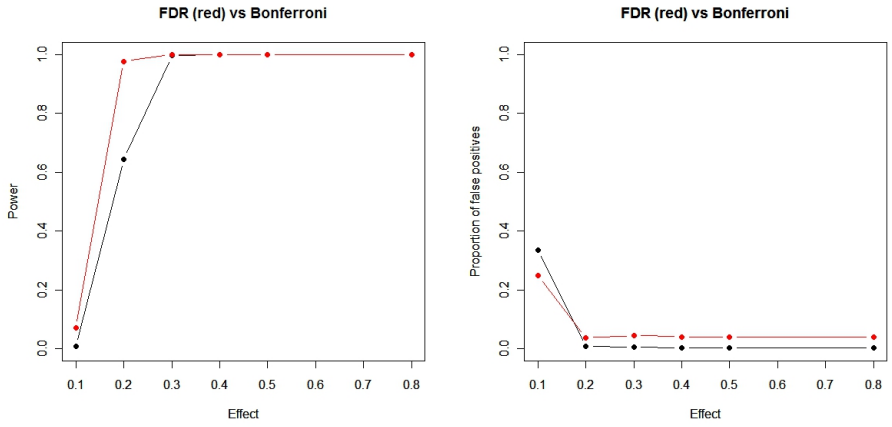


Figure 4.2: Power (left panel) and proportion of false positives (right panel) for the FDR method in stage 1 and FDR method and Bonferroni correction in stage 2

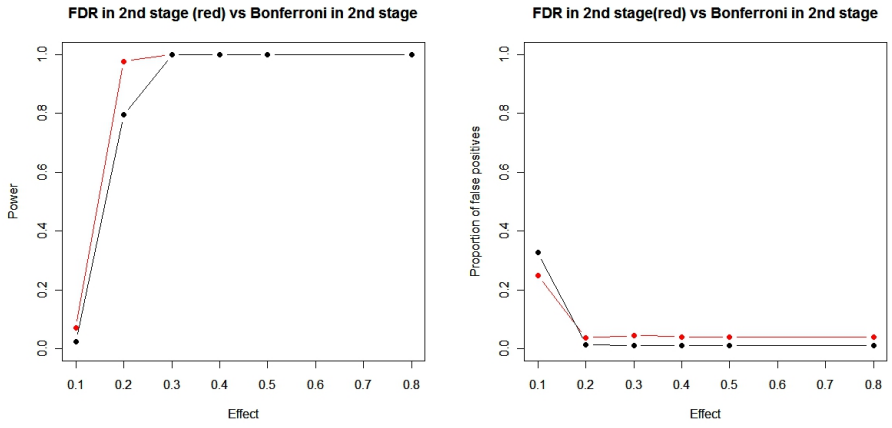
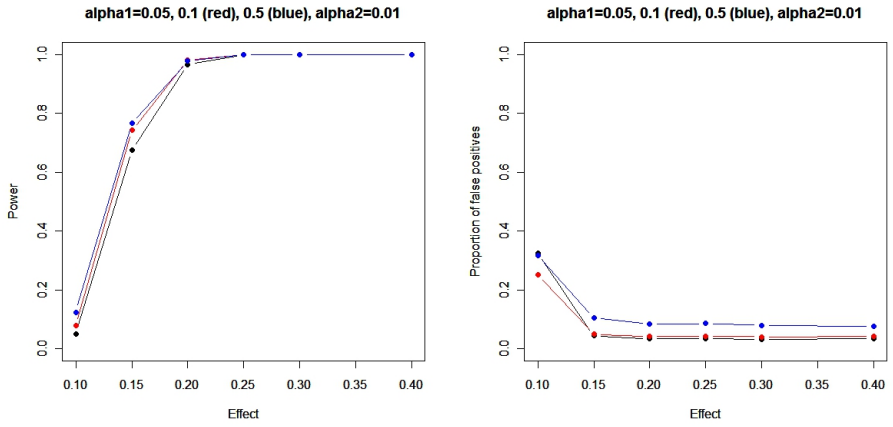


Figure 4.3: Power (left panel) and proportion of false positives (right panel) for different values of α_1



In Figure 4.2 we plotted the power and the proportion of false positives when using the FDR method in the first stage and the Bonferroni correction and the FDR method in the second stage. We notice that for the proportion of false positives we have almost the same results as before. However, we see that there is an increase in the power when using FDR-Bonferroni than when using Bonferroni in both stages. This power is still smaller than the power obtained using FDR in both stages.

In the following we will use the FDR method in both stages since for a very small effect the proportion of false positives is smaller and the power is bigger.

In Figure 4.3 we show the behavior of the power and of the proportion of false positives when α_1 is changing. We notice that when α_1 is increasing the power also increases. Moreover, we see that there is a difference in the power only up to an effect of 0.25 and that this difference is not so big when passing from $\alpha_1 = 0.1$ to $\alpha_1 = 0.5$. From the right panel we observe that for an effect between 0.1 and 0.15, $\alpha_1 = 0.1$ has the smallest proportion of false positives. For an effect bigger than 0.15 this proportion is quasi the same for $\alpha_1 = 0.1$ and $\alpha_1 = 0.05$. Furthermore we notice that the proportion of false positives is always bigger when $\alpha_1 = 0.5$ than for the other values of this significance level. Since there is not a big difference in power for $\alpha_1 = 0.1$ and $\alpha_1 = 0.5$ but there is a bigger difference in the proportion of false positives for the same values of α_1, we could say that the best value between these two is 0.1. Finally, since the power is slightly bigger for $\alpha_1 = 0.1$ than for $\alpha_1 = 0.05$ and the proportion of false positives is almost the same, we could say that the optimal value for α_1 is 0.1.

Figure 4.4: Power (left panel) and proportion of false positives (right panel) for different values of α_2

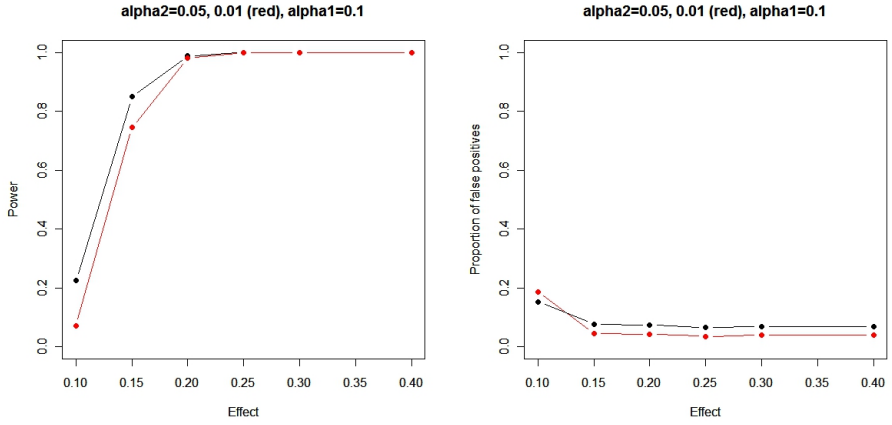
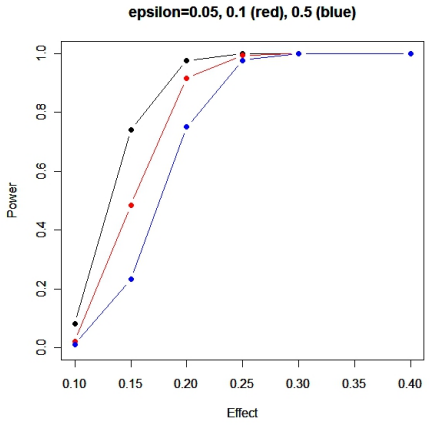


Figure 4.5: Power for different values of ε



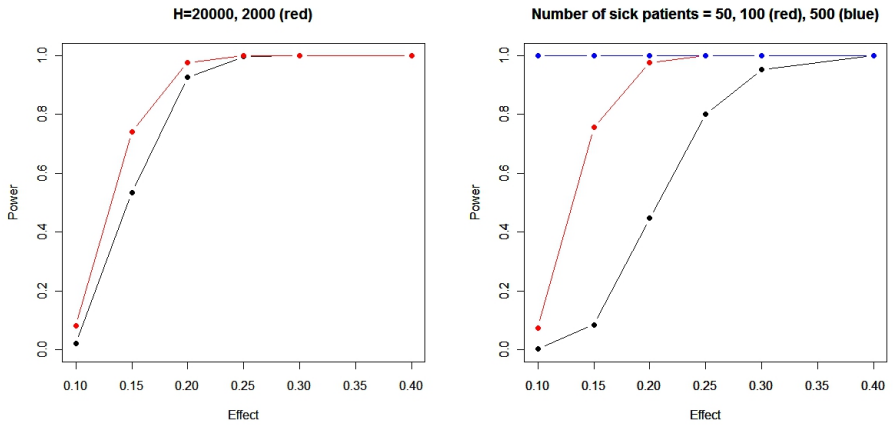
In Figure 4.4 we show the behavior of the power and of the proportion of false positives when α_2 is changing. We see that up to an effect of 0.2, the power is bigger when α_2 increases. For an effect larger than 0.2 the power is the same for both values of α_2. The proportion of false positives is bigger when $\alpha_2 = 0.05$ than when $\alpha_2 = 0.01$ except for an effect between 0.1 and 0.15. In the second stage of the design we want to be as restrictive as possible in order to have a small number (or zero) of false positives. Since the difference in power is not so

big between the two values and since $\alpha_2 = 0.01$ has a smaller proportion of false positives, we conclude that 0.01 is the optimal value for the significance level in stage 2.

We now want to see how the power fluctuates when the probability of having a mutation changes while the other parameters are fixed. The result is shown in Figure 4.5. We notice that up to an effect of 0.3, when ε increases, the power decreases. This might be explained by the difficulty of detecting a small effect when ε is big.

In Figure 4.6 we show the power as a function of the effect when the number of mutations H and the number of sick patients M is changing. We remark from the left panel that when H increases, the power decreases. From the right panel we see that when the number of sick patients is increasing (while the number of healthy patients is the same), the power also increases. When we have 500 sick patients and 900 healthy ones, the probability of incorrectly rejecting a true null hypothesis is 1. But if we want to be realistic, for example if we are testing a very rare disease, it is difficult to find 500 patients that have that disease, this is why we did all the above tests with 100 sick patients. For a very small effect, as 0.1, there is a big difference in the power between M= 100 and M= 500. When the effect increases, the difference becomes smaller and from an effect of 2.5 there is no more difference in the power. Furthermore, for an effect bigger than 0.15, the power, when using M= 100 is bigger than 0.7. In conclusion, the use of M= 100 is justified.

Figure 4.6: Power for different values of H(left panel) and M (right panel)



Simulation of the evolution of mutations

5.1 Neutral mutations

Until now we made simulations for the two-stage design in the cases where we had continuous and discrete measurements. In this section we will concentrate on a different model that we want to simulate, which is the accumulation of mutations in the DNA of the human population. The purpose is to see how the number of alleles change between generations and how many new alleles were created by mutations, where mutations are defined as a permanent change in the DNA sequence of a gene. For this, we suppose that we are in the case of the infinite alleles model. Furthermore, we suppose that we are in the theory called the Garden of Eden for the evolution of the population. This theory states that humans evolved from Africa about 200.000 years ago. If we consider a generation time of 20 years, we obtain around 10000 generation. For more information see [7]. Thus we do the simulation for a number of G= 10000 generations.

In this model we start by simulating the evolution of the population from generation 0 up to generation G. We start from an initial population number of N_0. Furthermore we suppose that the population is constant up to generation T and then it has an exponential growth. Mathematically this can be written as

$$N(t) = \begin{cases} N_0, & \text{if } t < T \\ N_0 \exp(\rho(t-T)), & \text{if } t \geq T \end{cases} \tag{5.1}$$

where ρ is the growth factor (< 1).

Once we have the population's number in each generation we can introduce the infinite alleles model. We start from $2N_0$ alleles. At the beginning all the alleles are identical. At each new generation t, we will have 2N(t) alleles. In order to obtain the alleles at generation t+1, we select randomly with replacement 2N(t) alleles from the previous generation. In the infinite alleles model, when we select an allele from the current generation to the next one, a mutation might occur. Mutations are supposed to follow a Poisson distribution with rate μ. Each time a mutation takes place, a new allele, never seen before, is created. In this case, every time we randomly choose an allele from the previous generation, the allele remains the same with probability $1 - \mu$, or a new allele is created with probability μ.

Remark 1. *Since the rate of mutation μ is very small («1), the generation of Poisson random variables is not so obvious. As the rate is small, one expects to*

have 0, 1 or 2 mutations per allele when going from one generation to another. We know that the density function of a Poisson is given by $f_\mu(k) = \frac{\mu^k}{k!}e^{-\mu}$. *By doing a Taylor expansion of order 2 we obtain, by selecting only the terms of at most power 2*

$$\begin{cases} 1-\mu+\frac{\mu^2}{2}, & \text{for } k=0, \\ \mu-\mu^2, & \text{for } k=1, \\ \frac{\mu^2}{2}, & \text{for } k=2 \end{cases}$$

It is easily verified that this defines a probability distribution. With this distribution we can generate the Poisson random variables as follows. We generate a uniform random variable U *on the interval* $[0,1]$. *Then, if* U *is smaller than* $1-\mu+\frac{\mu^2}{2}$ *we obtain* $k=0$, *if* U *is between* $1-\mu+\frac{\mu^2}{2}$ *and* $1-\frac{\mu^2}{2}$ *we obtain* $k=1$ *and else* $k=2$.

In our simulation we are interested in the number of alleles that did not mutate, the number of alleles that have one mutations, the number of alleles that have two mutations and so on. We denote by a_i the number of alleles that have mutated i times. For the simulation we took the following values for the parameters: $N_0 = 10000$, T= 8000 and $\rho = 0.001$. In Table 5.1 we present the results that we obtained for different values of μ after one simulation.

Table 5.1: Number of mutated alleles for different values of the mutation rate μ

μ	a_0	a_1	a_2	a_3	a_4	a_5	a_6	$a_0/2N_0$
10^{-9}	20000	0	0	0	0	0	0	1
10^{-8}	19998	2	0	0	0	0	0	0.999
10^{-7}	19983	17	0	0	0	0	0	0.992
10^{-6}	19794	205	1	0	0	0	0	0.989
10^{-5}	18092	1805	102	1	0	0	0	0.905
10^{-4}	7344	7429	3656	1187	303	66	14	0.367
10^{-3}	1	4	44	154	379	742	1269...	$5 \cdot 10^{-5}$

We notice that for $\mu = 10^{-9}$, which is very small mutation rate, no mutated allele appears. If $\mu = 10^{-8}$ there are two mutated alleles and this number increases for $\mu = 10^{-7}$. The first time when a double mutation appears is when $\mu = 10^{-6}$. Furthermore we observe that, as μ is getting bigger, alleles with multiple mutations exist. Finally, we observe that the proportion of the initial alleles (the non mutated ones) is decreasing as μ is increasing. For $\mu = 10^{-3}$ we have that only one allele remains non mutated. The suspension points in the case of a_6 mean that there are a lot more alleles with more than 6 mutations.

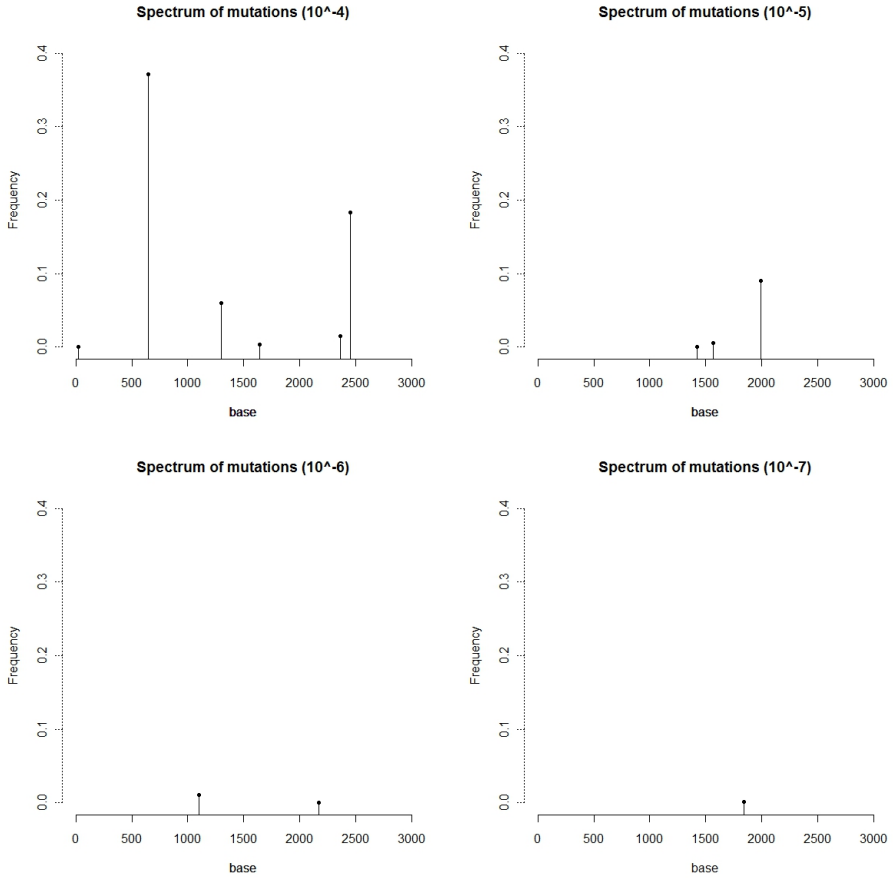
Remark 2. *One could ask if the results of our simulation are realistic. In order to verify this we compute the integral of the population size over all the generations and we multiply it by the constant rate of mutation. Since we have*

a big number of generations we can approximate the integral by a sum. Finally, we compute

$$T = \mu \cdot \sum_{t=0}^{G} N(t)$$

and we want to see if this number is close to the total number of mutated alleles that we obtain from the simulation. With the values used for the simulation we obtain that $T = 14392.25$ *for* $\mu = 10^{-4}$, $T = 1439.225$ *for* $\mu = 10^{-5}$ *and* $T = 143.9225$ *for* $\mu = 10^{-6}$. *We compare these values with those obtained in Table 5.1, which are* 12655 *for* $\mu = 10^{-4}$, 1908 *for* $\mu = 10^{-5}$ *and* 206 *for* $\mu = 10^{-6}$. *We notice that these values are not very far from the theoretical ones. We also remark that these values can change and be more or less close to the theoretical values since the results are based on a simulation study.*

Figure 5.1: Spectra of mutations for different values of μ



At this moment, one could ask about the frequency of a particular allele in this model. The frequency will certainly depend on the time. One expects to have a larger number of this allele in the population if the allele appeared in an earlier generation than if the same allele was created in a later generation. Thus, the frequency, when the allele is created, would be of $1/N(t)$, where N(t) is the number of individuals in the generation t. In our simulation, with a mutation rate of 10^{-6} and an initial number of individuals at generation 0 of 10000, we obtain that a first mutated allele is created at generation 186, which gives a frequency of 10^{-4}. At the final generation, we obtain that the frequency of this allele is of 0.0007 for a number of individuals of 73891. We finally obtain that 52 persons will carry the particular allele in the final generation. As explained before, we supposed that the population's number is constant up to generation T= 8000 from an overall of G= 10000 generations. This suggests that if an allele is created before generation T, its frequency at generation G will be the same as any other allele created before generation T. After generation T we obtain a different frequency, a smaller one, since the size of the population increases. For example, if an allele is created at generation 9426, its frequency will be of 0.0002 and we obtain that at generation G, 13 persons will have this particular allele.

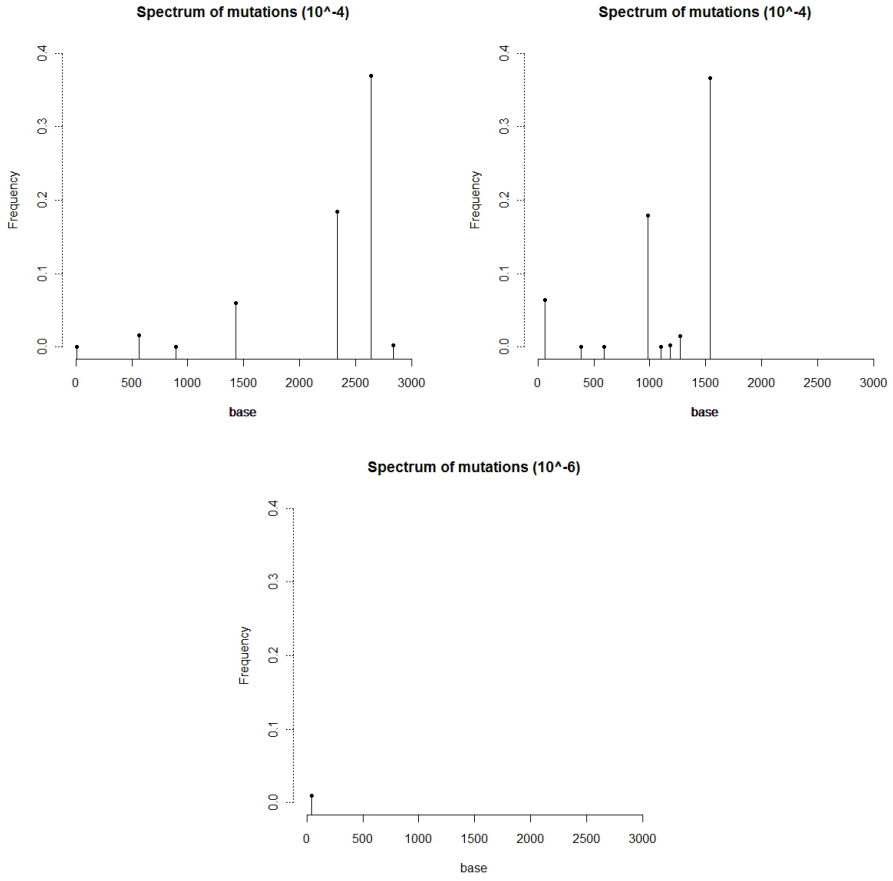
We now compute the spectrum of mutations for different values of the mutation rate. For the spectrum of mutations, we take a sample of 20000 individuals. We then select a gene and calculate the frequency of the bases that mutated. From the human genome project we know that the average number of bases in a gene is of 3000. Thus we will take 3000 bases for the spectrum of mutations. An important remark about how this spectrum is built is that we do not simulate the place on the gene where the mutation occurred, we attribute each mutation randomly without replacement to a location. The results for the values obtained from the simulation in Table 5.1 are presented in Figure 5.1.

We observe that when the rate of mutation is $\mu = 10^{-7}$ we have that approximately 0.001 of the population have a base that is different from the majority (20000). For $\mu = 10^{-6}$, we notice that two bases have mutated, one has frequency close to 0.02 and the other has frequency close to 0. When $\mu = 10^{-5}$, with low frequency, a third base is different from the majority. Finally, for $\mu = 10^{-4}$ we have that several bases have mutated. In this case the frequency of the mutated bases is bigger than the frequency of the bases that mutated with a smaller μ. One base has a frequency close to 0.4, another has a frequency of approximately 0.2 and the other bases have frequencies smaller than 0.1. This might suggest that a good value for the mutation rate would be $\mu = 10^{-4}$.

Since the simulation is based on random computations, we now present several mutation spectra for different values of μ in order to observe the variability of the process. The results are presented in Figure 5.2. We notice that, for $\mu = 10^{-4}$ we can obtain spectra that have 7 or 8 bases that have mutated and not always 6 as we obtained in Figure 5.1. We also see that the new bases that have mutated have a small frequency. For $\mu = 10^{-6}$ we observe that sometimes we can also obtain only one mutated base instead of two as we obtained before. For the other mutations rates we have almost the same spectrum after each

simulation.

Figure 5.2: Spectra of mutations for different values of μ



5.2 Hot spot mutations

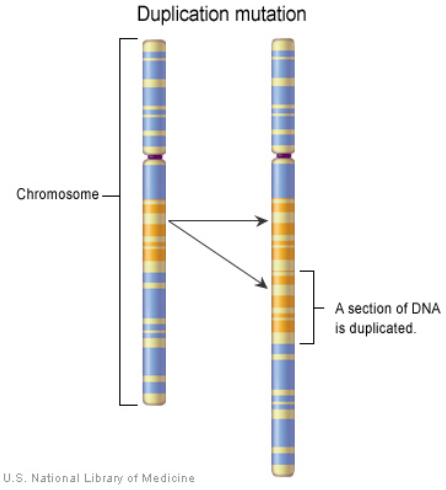
The simulation study that we did before was conducted with only one type of mutation. We now introduce multiple types of mutations, 3 types to be more specific. We denote by μ_i the rate of the type i mutation.

The type 1 would be the neutral mutation or the silent mutation, as we used in the previous subsection. An example of neutral mutation is the replacement of a base with another base in a gene that does not affect the protein encoded by the gene.

The type 2 would be mutations that are created by a bad copy of a base or

of a piece of DNA. If such a mutation occurs, it will always be the same. An example of this kind of mutation is the duplication. This takes place when a piece of DNA in a gene is abnormally copied several times, a process that might have a negative effect on the resulting protein of that gene. An illustration of a duplication is presented in Figure 5.3.

Figure 5.3: Duplication of a piece of DNA

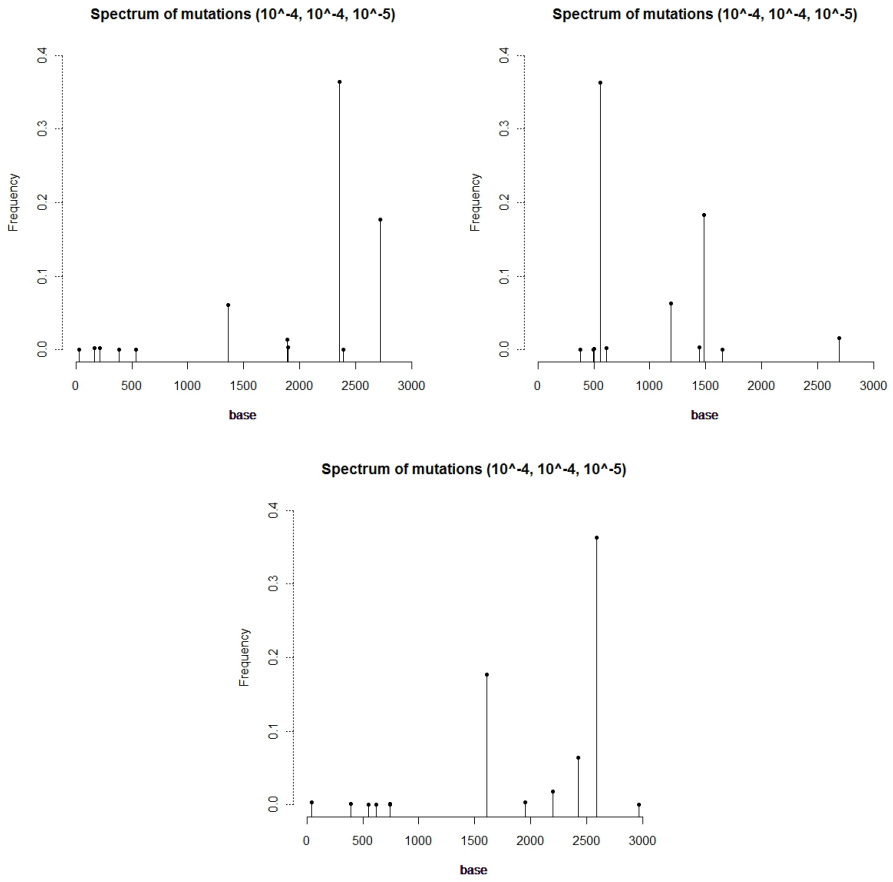


The type 3 would be all the mutations that inactivate a gene. For the type 3 mutation, there are many examples of genes that are inactivated by mutations, an interesting one is the p53 tumor-suppressor gene. As its name says it, the main role of this gene is tumor suppression by preventing genome mutation. However, if this gene is inactivated, it will no longer protect against tumors but it will create them. This is an important gene since it is mutated in over 50 % of all human cancer ([5])

With these three type of mutations, we suppose that the mutation rates satisfy the relation $\mu_3 < \mu_1 \leq \mu_2$. For our simulation we suppose that, on a DNA strand, the mutations are such that a very small proportion of the genes could have the type 3 mutation, a slightly bigger proportion could have the type 2 mutation and the rest could have the type 1 mutation. We took 20000 genes, and supposed that a very small number (between 20 and 50) could have the type 3 mutations, between 60 and 100 genes are plausible to have the type 2 mutation and the rest of the genes might be affected only by the type 1 mutation.

We want to see how the spectrum of mutations changes when the new types of mutations are introduced. In Figure 5.4 we present several spectra for the mutation rates $\mu_1 = 10^{-4}$, $\mu_2 = 10^{-4}$ and $\mu_3 = 10^{-5}$.

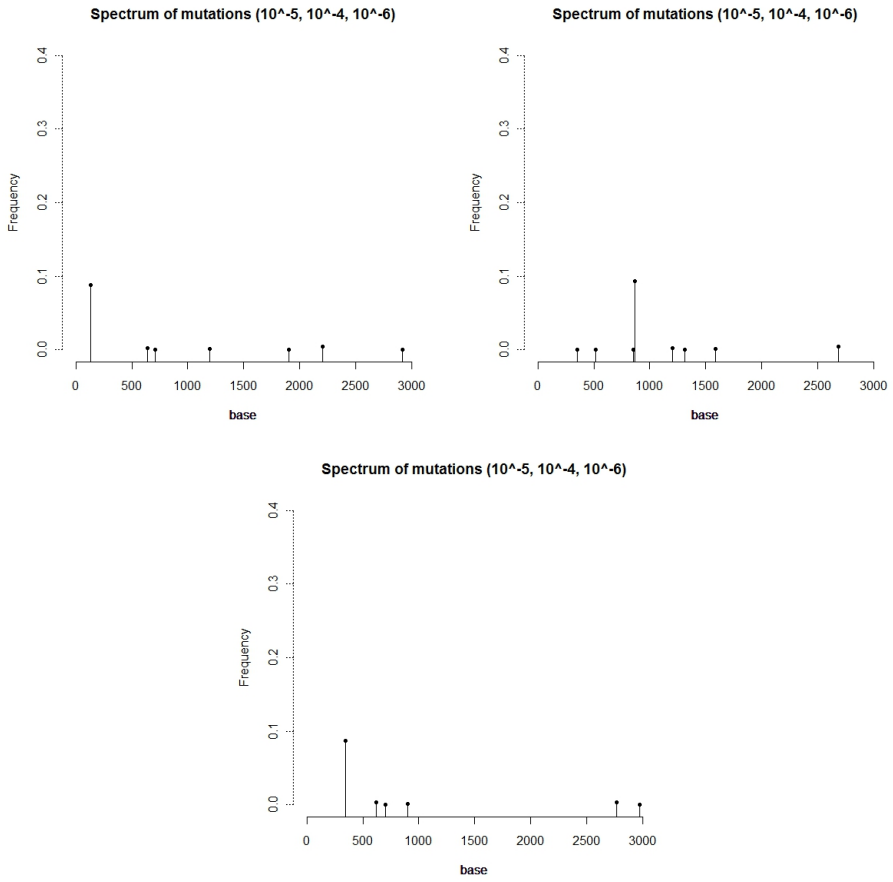
Figure 5.4: Spectra of mutations for $\mu_1 = 10^{-4}$, $\mu_2 = 10^{-4}$ and $\mu_3 = 10^{-5}$



We observe that, compared to the model with one mutation, for $\mu_1 = 10^{-4}$, we have more bases that have mutated (between 10 and 11 bases). Furthermore, we notice that the proportion of the bases that have mutated the most is almost the same in all the simulations. The new mutated bases have a frequency close to zero. Thus, the new mutations that are created in the hot spot model are presented only in a few number of individuals.

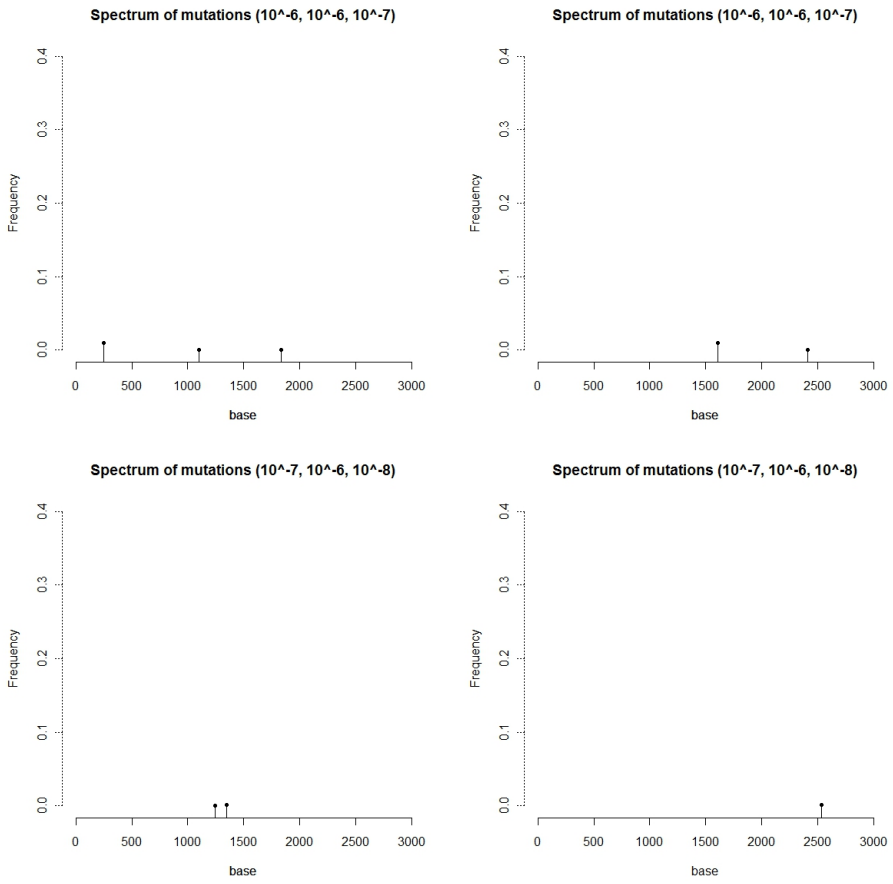
We now present several spectra for the mutations rates $\mu_1 = 10^{-5}$, $\mu_2 = 10^{-4}$ and $\mu_3 = 10^{-6}$. The results are shown in Figure 5.5. We notice that, for the same neutral mutation rate, we have more bases that have mutated in the hot spot model than in the one mutation model. We also notice that the number of mutated bases varies from 6 to 8, but the highest frequency is almost the same in each case.

Figure 5.5: Spectra of mutations for $\mu_1 = 10^{-5}$, $\mu_2 = 10^{-4}$ and $\mu_3 = 10^{-6}$



Finally, in Figure 5.6 we show several mutational spectra for the mutations rates $\mu_1 = 10^{-6}$, $\mu_2 = 10^{-6}$, $\mu_3 = 10^{-7}$ and $\mu_1 = 10^{-7}$, $\mu_2 = 10^{-6}$, $\mu_3 = 10^{-8}$. We remark that, for these mutations rates, we obtain almost the same number of mutated bases as in the one mutation model. Moreover, we notice that the frequencies of the mutated bases are very small.

Figure 5.6: Spectra of mutations



5.3 Other models for the population's evolution

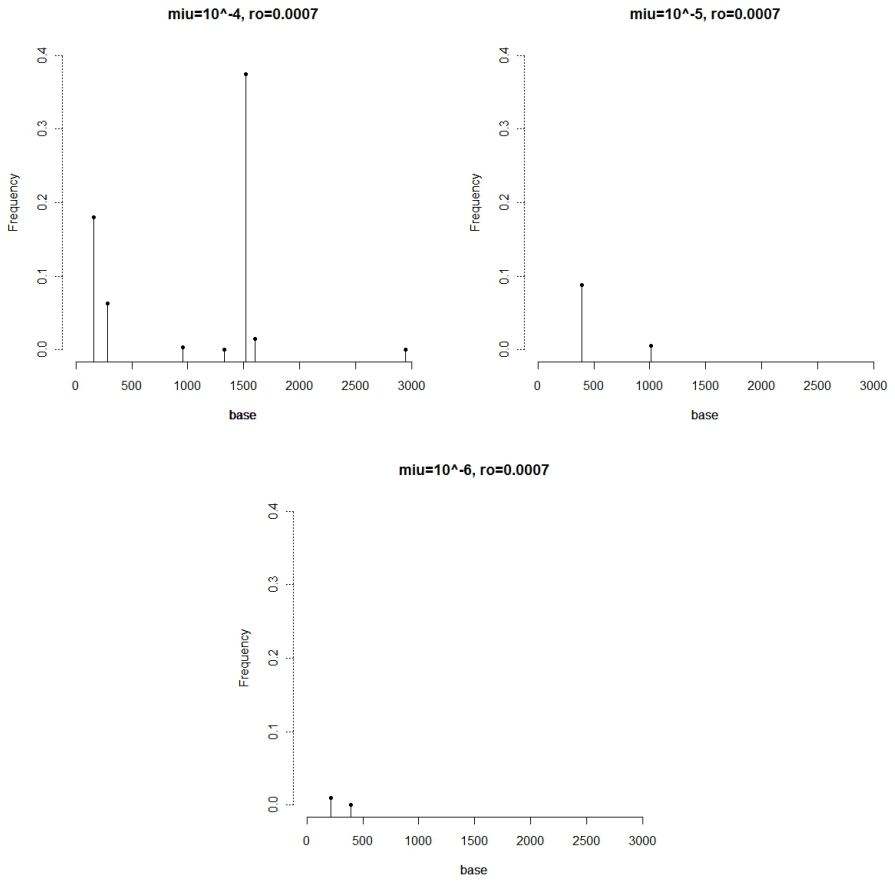
We are now interested in observing the behavior of alleles if we use different models for the growth of the population. We only consider the case of neutral mutations and want to see how the spectrum of mutations changes. In this purpose we introduce two other evolution models. The first model would be an exponential model. The second one would be linear up to a point and then it will grow exponentially. The exponential model is mathematically expressed like

$$N(t) = N_0 \exp(\rho t)$$

for $t = 0, \ldots, G$. In our simulation we used as before $N_0 = 10000$, but this time we took $\rho = 0.0007$. With this parameters we obtain that the population at

generation G is 10966332. With this model, we obtain a bigger population at the end, which is more realistic, but the computational time is heavily increased. Some spectra of mutation are shown in Figure 5.7. We observe that we obtain spectra similar to those of the model with a constant population up to generation T and an exponential growth afterwards.

Figure 5.7: Spectra of mutations for the exponential model for different values of μ



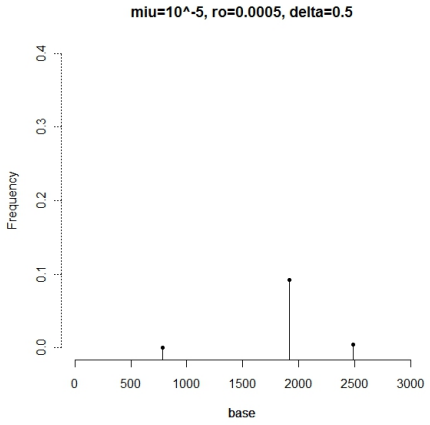
The second model resembles with the model that we used in the beginning of this section. The only difference is that the original model had a constant population up to generation T, while this model has a linear growth up to

generation T. Mathematically, this can be expressed as

$$N(t) = \begin{cases} N_0 + \delta t, & \text{if } t < T \\ (N_0 + \delta T)\exp(\rho(t-T)), & \text{if } t \geq T \end{cases}$$

where δ is the slope of the linear growth. For the simulation we took $\delta = 0.5$, $\rho = 0.0005$ and $N_0 = 10000$. With these values for the parameters, we obtain a final population of 65298566, which is bigger than the population that we had for the exponential model. However, the spectrum of mutations is the same as for the other models. For example, in Figure 5.8 we plotted the spectrum for $\mu = 10^{-5}$. We remark that, as for the exponential model, the computational time is significantly increased.

Figure 5.8: Spectrum of mutations for the linear + exponential model for $\mu = 10^{-5}$



We conclude that the model that we initially used in the beginning of this section is the best model to use for the evolution of the population in the case of the infinite alleles model.

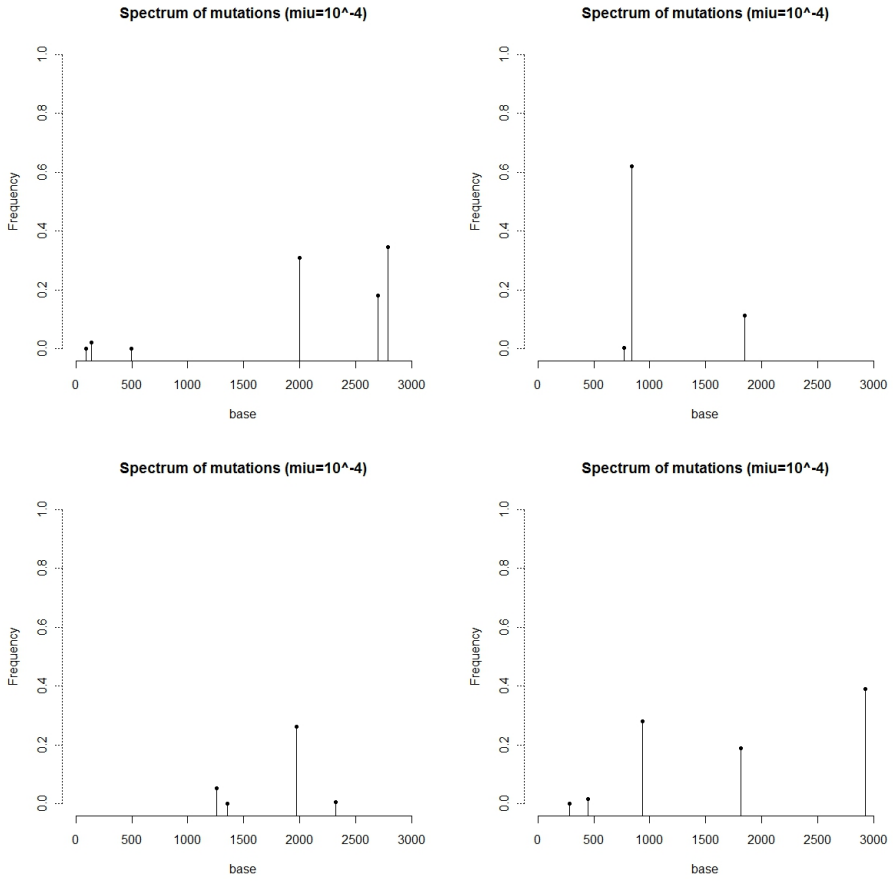
5.4 A more realistic model

In the infinite alleles model presented in the beginning of this section we have supposed that the number of alleles is constant between generations and that only the types of alleles change. We now present a more realistic model in which we suppose that the number of alleles changes with the number of the population from a generation to another. In this model, the new generation t+1 of alleles is created as follows: before entering the alleles in the mutational process, we choose $2N(t)$ alleles from generation t by selecting with probability

$a_i/2N(t)$ the alleles of type i. We expect to obtain, in proportion, a smaller number of mutated alleles since the probability of selecting a mutated alleles is lower than the probability of selecting a non-mutated one.

As before, we are interested in the spectrum of mutations that results from this new model. For the simulations, we use the model with constant and then exponential growth for the evolution of the population. Furthermore, the values of the parameters are N0= 10000, T= 8000 and $\rho = 0.001$. In Figure 5.9 we show spectra that we obtain when the mutations rate is $\mu = 10^{-4}$.

Figure 5.9: Spectra of mutations for $\mu = 10^{-4}$



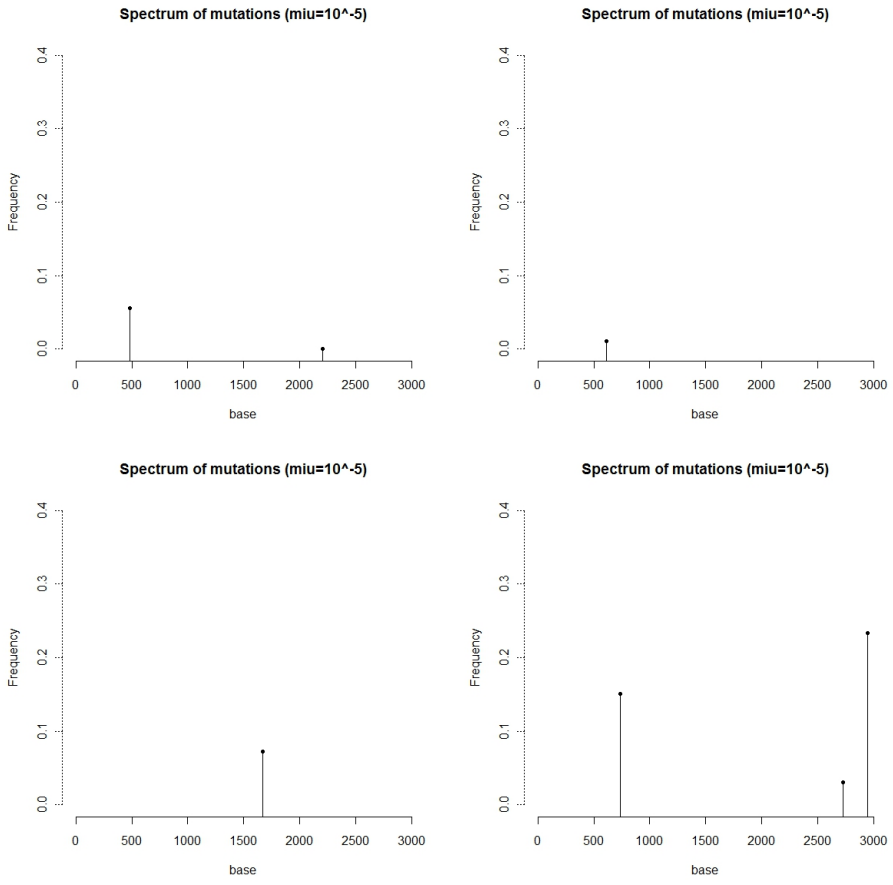
We observe that the number of mutated bases varies from 3 to 6. In our simulations, the most frequent results had 4 or 5 frequent bases and in some cases we obtained a spectrum with 3 or 6 bases that differ from the majority. We also notice that the highest frequency observed is when we only have 3 SNPs.

This frequency is approximatively 0.4 for the other cases.

We remark that, in this case, the highest peak does not necessary represent the alleles that have mutated only one time. It could represent the alleles that have mutated two or more times since the probability of selection can increase near the end of the evolution.

For a smaller mutation rate like $\mu = 10^{-5}$ the results of different spectra are presented in Figure 5.10. We see that we obtain a number of mutated bases that oscillates from 1 to 3. When we obtain 3 SNPs, the highest frequencies are greater than 0.1, one being even greater than 0.2. However, as expected, these frequencies are smaller than those when $\mu = 10^{-4}$.

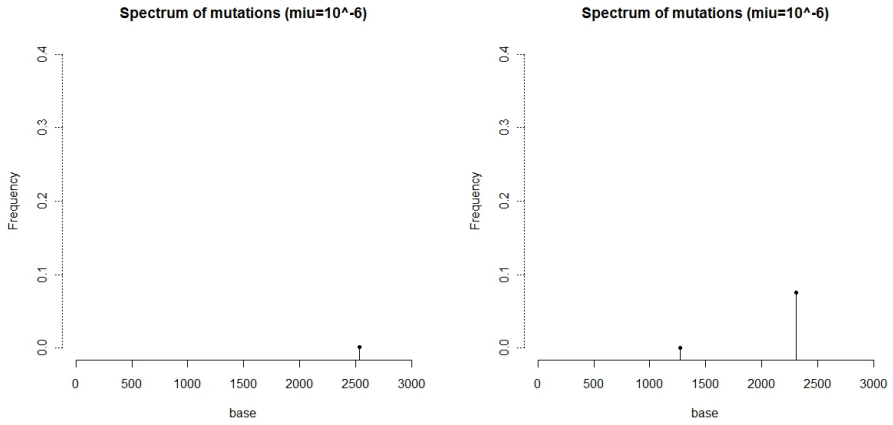
Figure 5.10: Spectra of mutations for $\mu = 10^{-5}$



Finally, in Figure 5.11 we present spectra of mutation when $\mu = 10^{-6}$. In most cases we obtain spectra that resemble the one in the left panel, i.e., with

only one mutated base. We can also obtain spectra that have a quite important frequency (of almost 0.1) for this mutation rate as we can observe in the right panel. Compared to all the other spectra that we presented for $\mu = 10^{-6}$, this is the highest frequency that we have obtained. We remark that in this case we can also obtain a spectrum with no SNPs.

Figure 5.11: Spectra of mutations for $\mu = 10^{-6}$

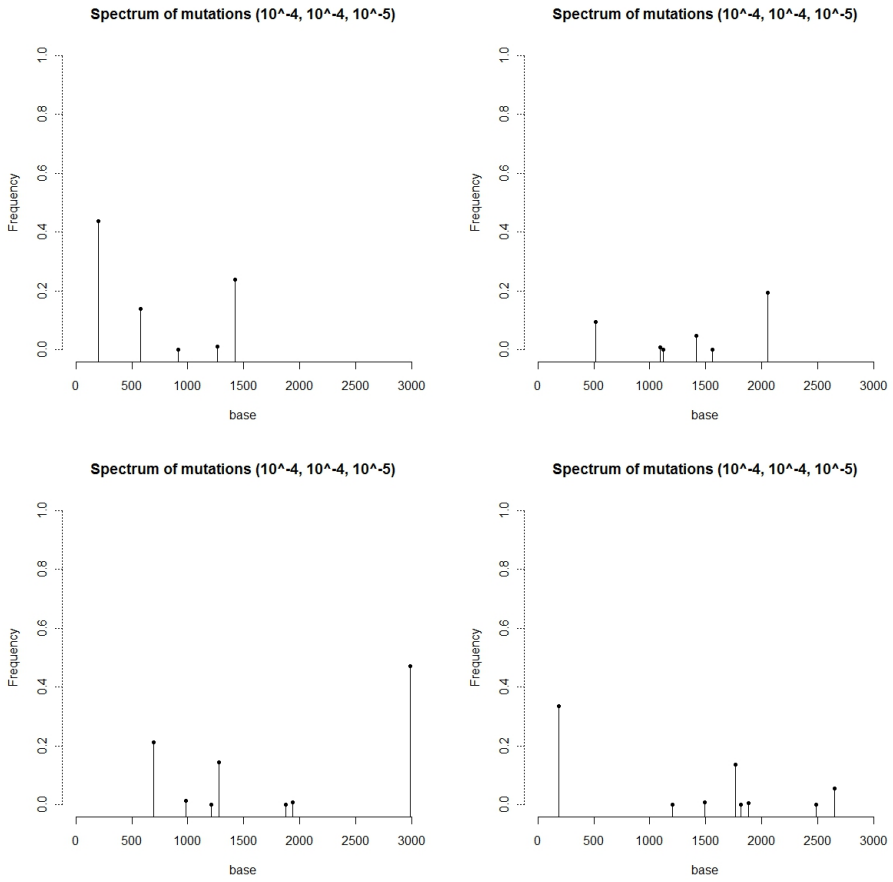


5.4.1 Adding hot spot mutations

In this section we want to see how this model behaves in the case where hot spot mutations are present. We proceed as we did in the subsection 5.2 and suppose that there are three type of mutations. We obtained different spectra for different values of the mutation rates. In Figure 5.12 we present some spectra for $\mu_1 = 10^{-4}$, $\mu_2 = 10^{-4}$ and $\mu_3 = 10^{-5}$.

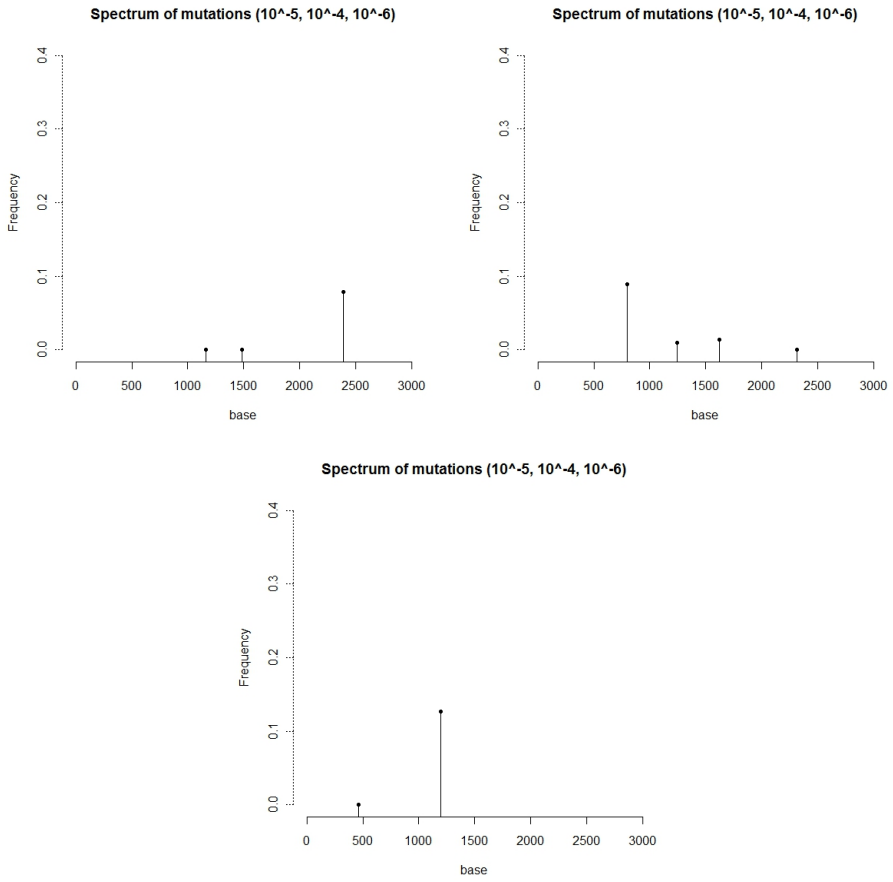
We remark that, compared to the spectra obtained in Figure 5.4, we have fewer bases that have mutated. This number is between 5 and 8. We also see that the highest frequency varies between 0.2 and about 0.5. Furthermore, for the same rate of mutation for the neutral mutation type (10^{-4}), we now obtain a slightly bigger number of mutated bases (see Figure 5.9). These new mutated bases are created by the type 2 or type 3 mutations that we introduced.

Figure 5.12: Spectra of mutations for $\mu_1 = 10^{-4}$, $\mu_2 = 10^{-4}$ and $\mu_3 = 10^{-5}$



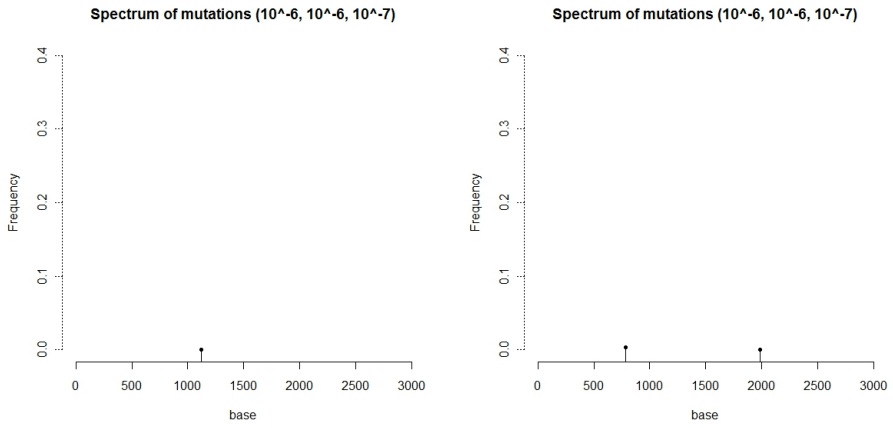
We now take smaller mutations rates for the type 1 and 3 mutations like $\mu_1 = 10^{-5}$ and $\mu_3 = 10^{-6}$. The results are presented in Figure 5.13. If we compare these spectra with those obtained in Figure 5.5 we notice that the number of mutated bases has significantly decreased. However, we now obtain highest frequencies which are not always close to 0.1 as we obtained before. In the bottom panel the highest frequency is of almost 0.15. Finally, we remark that for the same neutral mutation rate (10^{-5}), we obtain spectra that are similar with those from the one mutation scenario (see Figure 5.10).

Figure 5.13: Spectra of mutations for $\mu_1 = 10^{-5}$, $\mu_2 = 10^{-4}$ and $\mu_3 = 10^{-6}$



Finally, we consider the case when the mutation rates are $\mu_1 = 10^{-6}$, $\mu_2 = 10^{-6}$ and $\mu_3 = 10^{-7}$. The possible spectra that we can obtain with this rates resembles those presented in Figure 5.14. As before, if we compare the results with those obtained in Figure 5.6 (the two above panels) we notice that we obtain at most 2 mutated bases instead of 3. This might be explained by the fact that in this new model, we introduced the possibility that an allele might disappear over generations. We remark that in this case, as in the case where we considered only one mutation with rate 10^{-6}, we can also obtain spectra with no mutated bases.

Figure 5.14: Spectra of mutations for $\mu_1 = 10^{-6}$, $\mu_2 = 10^{-6}$ and $\mu_3 = 10^{-7}$



5.4.2 From Adam and Eve

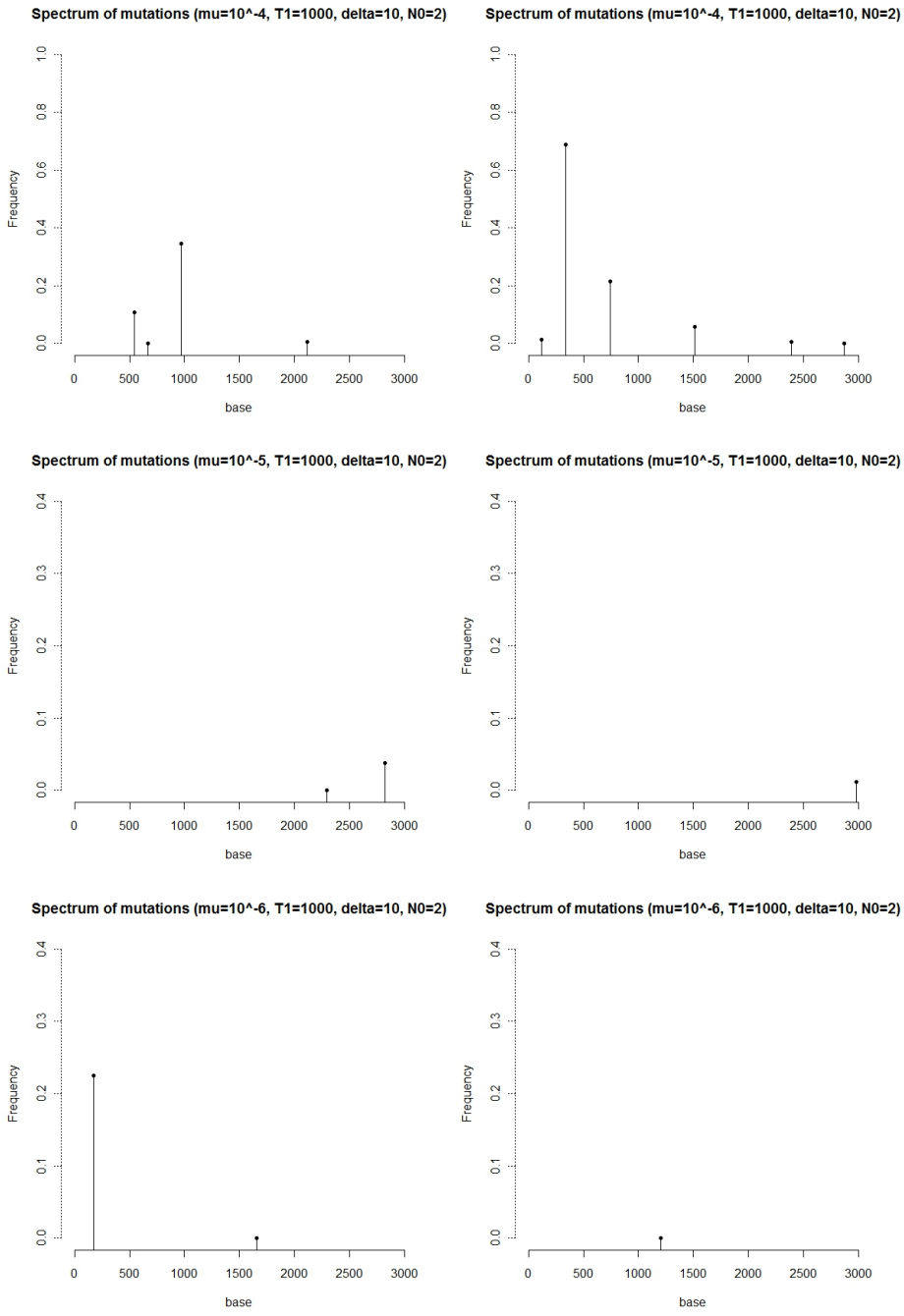
Up to this point, in all the simulation that we used we supposed that the initial population was of 10000 individuals. However, as the Bible says, initially there were only two people that started populating the Earth. We thus do a simulation in which we suppose that the initial population size was of two individuals (a male and a female) and up to a certain generation the population grew linearly. Afterwards the population was constant up to another time point and finally it grew exponentially. Mathematically this can be expressed as

$$N(t) = \begin{cases} N_0 + \delta t, & \text{if } t < T_1 \\ N_0 + \delta T_1, & \text{if } T_1 \leq t < T_2 \\ (N_0 + \delta T_1)\exp(\rho(t - T_2)), & \text{if } t \geq T_2 \end{cases}$$

In the simulation we took $T_1 = 1000$, $T_2 = 8000$, $\rho = 0.001$ and $\delta = 10$. With these parameters we obtain a final population of 73905, which is almost the same as in the previous model. We run the simulation for the mutation rates $\mu = 10^{-4}$, $\mu = 10^{-5}$ and $\mu = 10^{-6}$. The spectra that we obtained are presented in Figure 5.15. We notice that we obtain spectra that are similar with those obtained in Figures 5.9, 5.10 and 5.11. We also observe that for the mutation rate $\mu = 10^{-6}$ we might obtain a mutated base with a frequency higher than 0.2, which was not the case in the constant + exponential model used in the beginning of this section.

In conclusion, the initial population size has not an important effect on the mutational spectrum if the final population size is almost the same. One could think at an even more realistic model in which he could add the effect of bottleneck that has occurred in the human history, but goes beyond the purpose of this report.

Figure 5.15: Spectra of mutations for $\mu = 10^{-4}$, $\mu = 10^{5}$ and $\mu = 10^{-6}$



Estimating the effect

In this chapter we want to estimate the effect in the model with discrete measures. We will use the model in which the number of alleles changes with the number of the population from a generation to another (model presented in section 5.4)

We present a couple of scenarios for the causality of the disease:

1) In the first scenario, we suppose that, in a particular gene, there are several SNPs that are at risk. More precisely, we suppose that an individual that has a disease has one of the SNPs of the gene but the healthy individuals don't have any SNPs from that gene. We thus have that the effect is equal to

$$\Delta = P(A_i|M) = \frac{P(A_i \cap M)}{P(M)} = \frac{P(A_i \cap M)}{\sum_k P(A_k \cap M)} \tag{6.1}$$

where the event A_i represents the presence of the i^{th} SNP and the event M stands for the sick people.

We compute the effect from the spectra obtained in the previous section. For a mutation rate of 10^{-4}, we obtain for the spectrum presented in the upper left panel of Figure 5.9, effects for each SNP equal to 0.403, 0.361, 0.209, 0.026, 0.002 and $0.008 \cdot 10^{-3}$. We now want to see if these effects have a high probability of being detected. For this we take a look at the power of the test in the two-stage design with discrete measures. We have seen that for this design, the optimal significance parameters are $\alpha_1 = 0.1$ and $\alpha_2 = 0.05$. From Figure 4.4 we observe that the first three effects obtained have a power of 1, while the other effects, which are very small, have a power close to 0 of being detected. We notice that these last three effects correspond to SNPs that have a very small frequency (close to 0).

For $\mu = 10^{-5}$ we obtain, from the upper left panel of Figure 5.10, that the effect is of 0.999 for the SNP with the highest frequency and of 0.001 for the other SNP. From Figure 4.4 we remark that the SNP with the highest frequency has a power 1 of being detected while the other SNP has a power close to 0 of being detected.

We now consider the case where hot spot mutations are present. For the spectrum in the lower left panel of Figure 5.12 we obtain effects of 0.249, 0.553, 0.169, 0.018, 0.001, 0.009 and 0.001. The first two SNPs (those with the highest frequencies) have power 1 of being detected. The effect of the third SNP has power close to 0.9. We notice that in the upper left panel of Figure 5.9 we also had a mutated base with a frequency close to 0.3 (as our third SNP) and in that case we had a power of 1 of detecting its effect.

By adding hot spot mutations we increase the number of mutated bases and in consequence the probability of detecting an effect is slightly diminished. Finally, for the other SNPs we have power close to 0.

For mutation rates of $\mu_1 = 10^{-5}$, $\mu_2 = 10^{-4}$ and $\mu_3 = 10^{-6}$, we obtain from the spectrum in the upper right panel of Figure 5.13 effect equal to 0.788, 0.124, 0.086 and 0.002. We notice that the SNP with the highest frequency has power 1 of detecting its effect. For the effect of 0.124 we have a power of approximately 0.4 of observing it. For the other effects we obtain a power close to 0.

2) The second scenario implies that a particular SNP in a particular gene is causing the disease. This means that only one mutation from the spectrum of mutations is causing the disease. The difficulty comes here from the fact that we have to know which SNP is responsible for the disease.

   We suppose that the mutation that causes the disease is a type 2 mutation (a recurrent mutation). Moreover, we suppose that we have the spectrum of a normal mutational process. Let us call this process the background noise. The process of the type 2 mutation will be added on the spectrum of the background noise. The SNP that will differ from the background noise will be the one that it is associated with the disease. The effect can now be computed as we did in 6.1. We simulated a background noise process with $\mu = 10^{-4}$ and a process for the type 2 mutation with $\mu = 10^{-5}$. The result is presented in Figure 6.1.

Figure 6.1: Spectrum of mutations for the second scenario with $\mu = 10^{-4}$ for the background noise process and $\mu = 10^{-5}$ for the type 2 mutational process (in red)

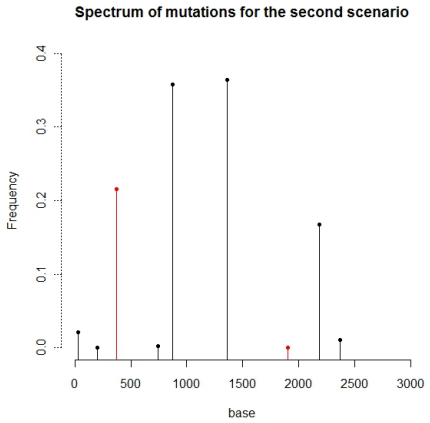
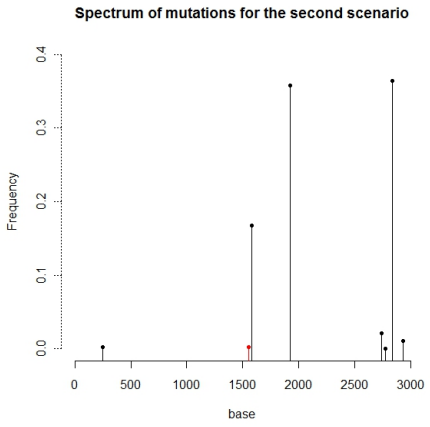


Figure 6.2: Spectrum of mutations for the second scenario with $\mu = 10^{-4}$ for the background noise process and $\mu = 10^{-6}$ for the type 2 mutational process (in red)



If we assume that the SNP that causes the disease is the one corresponding to the mutated base (in red) with the highest frequency, we find an effect of 0.189. This effect has a power of approximately 0.9 of being detected. If we consider that the second mutation of type 2 is the one that causes the disease, we find that the effect is $0.005 \cdot 10^{-1}$. This time, the effect has a power close to 0 of being observed. We notice that in this scenario, the effect has smaller values than in the first scenario.

If we suppose that the mutation rate for the type 2 mutation is $\mu = 10^{-6}$ we obtain the spectrum presented in Figure 6.2. In this case we obtain an effect of 0.003, which has power close to 0 of being detected.

Conclusion

We begun this report by doing a simulation of the two-stage design for the situation of continuous measures. Since we were in the case of multiple testing we had to adjust the p-values to correct for occurrence of false positives. To do so, we used the False Discovery Rate (FDR) method and the Bonferroni correction. We wanted to see which one of these two methods was the better to use from the point of view of the power, the cost, and the proportion of false positives. We found that the better method was the FDR method. Furthermore, we tried to find the optimal values for the parameters used in the study, i.e., for the number of hypotheses H, the sample sizes n_1 and n_2 for the first and second stage and the significance levels α_1 and α_2 for stage 1 and stage 2. We found that the optimal values are H= 1000, $n_1 = 5$, $n_2 = 30$, $\alpha_1 = 0.1$ and $\alpha_2 = 0.01$. For this simulation we used test statistics that followed a Normal distribution. We then complicated the study and supposed that the test statistics were following a Contaminated Normal distribution (CND). We compared the results with those obtained with the Normal distribution and observed that the cost of the study is increasing when using CND. For the power we noticed that for smaller effects (up to approximately 1.7) we have a bigger power for the CND than for the Normal distribution. For bigger effects we had the opposite result.

Since the evaluation of biomarkers does not always result in continuous observations, we also did a simulation of the two-stage design when observing a discrete endpoint. As we did in the continuous case, we wanted to see which of the FDR or the Bonferroni method was better. We have seen that for smaller effects the FDR method was better from the point of view of the power and the proportion of false positives. We then tried to found the optimal parameters for the study and found the same values as before for the significance levels. We thus arrived at the same conclusions as in the continuous case.

Next, we concentrated on something different, which was the simulation of the evolution of mutations in the human population. First, we supposed that we were in the infinite alleles model and that the number of alleles is constant over time. We used a model for the evolution of the population in which we supposed that the population is constant up to generation 8000 and then it has an exponential growth with parameter $\rho = 0.001$. We considered that we had a number of 10000 generation and an initial population of 10000, which gave a final population of 73891. We then constructed the spectrum of mutations for mutation rates of 10^{-4}, 10^{-5}, 10^{-6} and 10^{-7} and we saw that we obtain a plausible spectrum for 10^{-4}. Up to here we only considered one type of mutation, the neutral one. We thus introduced other two type of mutations. These types were mutations that are created by a bad copy of a base or of a

piece of DNA and mutations that inactivate a gene. As before we computed the spectrum of mutations and observed that in this case we had more SNPs than in the one-mutation case. We then tried different models for the evolution of the population, models that would give a final population that is closer to the actual population. We saw that we obtained the same spectrum as in the initial model and that the computational time was heavily increased. We concluded that the model that we initially used was the best model for the evolution of the population in the case of infinite alleles model.

Afterwards, we simulated a more realistic model, in which alleles can disappear from a generation to another one. We obtained, when adding hot spot mutations, spectra with fewer SNPs than in the initial model with hot spot mutations. We then considered a model in which the initial population was of 2 individuals. With this model, that has almost the same final population as before, we obtained results that did not differ from those with model 5.1.

In the final chapter we presented two methods of estimating the effect from the spectrum of mutations. We observed that the SNPs with high frequency had also a big power of detecting their effects. We have also seen that when we introduce hot spot mutations we diminish the probability of detecting an effect.

In conclusion we have seen that we can estimate the effect of detecting a marker that is associated with a disease. We saw from the second method of estimating the effect that, when the mutation rate is of 10^{-4} or 10^{-5}, the effect has non-zero power of being detected. We have also seen that the spectra obtained with these two values were closer to what we expected than with the other values. Finally, we can conclude that these two values might be the plausible values for the mutation rate.

Bibliography

[1] *Two-Stage Designs for Gene-Disease Association Studies*, Jaya M. Satagopan, David. A Verbel, E.S. Venkatraman, Kenneth E. Offit, and Colin B. Begg, Biometrics 58, 163-170, March 2002.

[2] *Two-Stage Designs for Gene-Disease Association Studies with Sample Size Constraints*, Jaya M. Satagopan, E.S. Venkatraman, and Colin B. Begg, Biometrics 60, 589-597, September 2004

[3] *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing*,Yoav Benjamini and Yosef Hochberg, J.R. Statist. Soc. B (1995) 57, No.1, pp.289-300

[4] *Génétique statistique*, Stephan Morgenthaler, Springer 2008

[5] *Pathologic basis of disease*, Robbins and Cotran and Kumar and Collins, sixth edition

[6] http://ghr.nlm.nih.gov/handbook/genomicresearch

[7] http://www.ramsdale.org/dna11.htm