MCP - Final Project

Eyal Bar-Natan, Nitzan Shamir, Omer Shubi

05/04/2020

Contents

Task 1

Part A.

$$P(P_i \leq \alpha) \overset{p\_value\ definition}{=} P[P_{H_0}(X > X_i) \leq \alpha] = P[P_{H_0}(\frac{X-\mu_0}{1} > \frac{X_i-\mu_0}{1}) \leq \alpha] \overset{X \underset{H_0}{\sim} N(0,1)}{=} P[P(Z > X_i) \leq \alpha]$$

$$= P[1 - \Phi(X_i) \leq \alpha] = P[\Phi(X_i) \geq 1 - \alpha] = P[\Phi(X_i) \geq \Phi(Z_{1-\alpha})] \overset{\Phi\ is\ monotone\ increasing}{=} P(X_i \geq Z_{1-\alpha})$$

For the case where P_i is a p-value of a false null hypothesis:

$$P(P_i \leq \alpha) = \cdots = P(X_i \geq Z_{1-\alpha}) = P(\frac{X_i-\mu_i}{1} \geq \frac{Z_{1-\alpha}-\mu_i}{1}) \overset{X_i \underset{H_1}{\sim} N(\mu_1,1)}{=}$$

$$P(\frac{X_i-\mu_1}{1} \geq \frac{Z_{1-\alpha}-\mu_1}{1}) \overset{X_i \underset{H_1}{\sim} N(\mu_1,1)}{=} P(Z_i \geq Z_{1-\alpha} - \mu_1) = 1 - \Phi(Z_{1-\alpha} - \mu_1)$$

For the case where P_i is a p-value of a true null hypothesis:

As we showed above, $P(P_i \leq \alpha) = \cdots = P(X_i \geq Z_{1-\alpha})$

Therefore,

$$P(P_i \leq \alpha) = \cdots = P(X_i \geq Z_{1-\alpha}) = 1 - P(X_i \leq Z_{1-\alpha}) \underset{X_i \underset{H_0}{\sim} N(0,1)}{=} 1 - \Phi(Z_{1-\alpha}) = 1 - (1-\alpha) = \alpha$$

$\Longrightarrow P_i\ 's\ distribution\ is\ uniform[0,1]$

To conclude,

When P_i is a p-value of a false null hypothesis, $P(P_i \leq \alpha) = 1 - \Phi(Z_{1-\alpha} - \mu_1)$,

and when P_i is a p-value of a true null hypothesis, $P(P_i \leq \alpha) = \alpha$.

Part B.

$\text{FWER} = P[V \geq 1] = 1 - P[V = 0] \underbrace{=}_{\text{Def. } V} 1 - P[\Sigma_{i \in M_0} I\{H_i \text{ is rejected}\} = 0]$

$= 1 - P[\forall i \in M_0 : H_i \text{ is not rejected}] = 1 - P[\forall i \in M_0 : P_i > \alpha]$

$= 1 - P[\bigcap_{i \in M_0} P_i > \alpha] = 1 - \Pi_{i \in M_0} P[P_i > \alpha] \underbrace{=}_{\text{Part A.}} 1 - \Pi_{i \in M_0} \Phi(Z_{1-\alpha} - \mu_i)$

$= 1 - \Phi(Z_{1-\alpha})^{m_0} = 1 - (1-\alpha)^{m_0}$

FWER depends on the variables m_0 and α.

As $\alpha \in (0,1), m_0 \geq 0$, it increases as each of α and m_0 increases.

This behaviour is intuitive.

For each true null hypothesis, the probability of performing a type I error is constant given α (and it is equal to α). Therefore, if there are more true null hypotheses (m_0 increases), the probability of having at least one type I error (FWER) increases as well. The intuition here says that there is more room for this kind of mistakes.

Moreover, for a given m_0, when alpha increases, our approach for rejecting null hypotheses is more liberal i.e. we allow more freedom for type I errors, thus increasing the chance of making at least one type I error.

$\text{PFER} = \mathbb{E}(v) = \mathbb{E}[\Sigma_{i \in M_0} I\{H_i \text{ is rejected}\}] \overset{\text{linearity+expectancy of indicator}}{\widetilde{=}}$

$= \Sigma_{i \in M_0} P(H_i \text{ is rejected}) = \Sigma_{i \in M_0} P(P_i \leq \alpha) \overset{\text{according to part A}}{\widetilde{=}} m_0 \alpha$

We can note that this expression could be achieved also by the expectancy of a binomial random variable.

PFER depends on the variables m_0 and α.

It increases as each of α and m_0 increases.

The intuition for such behavior, derives from the fact that the expectancy of an r.v increases as the probability that the r.v is positive, increases. In our case, this probability defined as FWER, which we saw before that is increasing while these variables are increasing.

For $m_0 \neq m$:

$\Pi^{AVG} = \mathbb{E}(S/m_1) = \frac{1}{m_1}\mathbb{E}(S) = \frac{1}{m_1}\mathbb{E}[\Sigma_{i \in M_1} I\{H_i \text{ is rejected}\}]$

$\overset{\text{linearity+expectancy of indicator}}{\widetilde{=}} \frac{1}{m_1}\Sigma_{i \in M_1} P(H_i \text{ is rejected})$

$= \frac{1}{m_1}\Sigma_{i \in M_1} P(P_i \leq \alpha) \overset{\text{according to part A}}{\widetilde{=}} \frac{1}{m_1}\Sigma_{i \in M_1}[1 - \Phi(Z_{1-\alpha} - \mu_1)]$

$= 1 - \Phi(Z_{1-\alpha} - \mu_1)$

For $m_0 = m$:

This means that $m_1 = 0 \Rightarrow S = 0$ and then $\Pi^{AVG} = 0$.

Π^{AVG} depends on μ_1 and α.

It increases when each of α and μ_1 increases (because Φ is an increasing function).

This behavior squares with the intuition that as the distance of μ_1 from 0 (μ_0) is bigger, it makes the alternative hypothesis stronger. It helps us to distinguish between the hypotheses. And it is known that the power of hypotheses testing increases when it is easier to make null hypotheses rejections (when α increases).

Part C.

According to the previous calculations (of FWER), we can note that:

$$P(V=k) = \binom{m_0}{k}[P_{H_0}(P_i \leq \alpha)]^k \cdot [1 - P_{H_0}(P_i \leq \alpha)]^{m_0-k}$$

$$\overset{according\ to\ part\ A}{=} \binom{m_0}{k}\alpha^k \cdot (1-\alpha)^{m_0-k}$$

This outcome is applied, because the p-values of the hypotheses are independent and identically distributed.

V counts the number of the true null hypotheses (m_0) that are rejected ($w.p.\ \alpha$).

Therefore,

$$V \sim Bin(m_0, \alpha)$$

According to the previous calculations (of Π^{AVG}), we can note that:

$$P(S=k) = \binom{m_1}{k}[P_{H_1}(P_i \leq \alpha)]^k \cdot [1 - P_{H_1}(P_i \leq \alpha)]^{m_1-k}$$

$$\overset{according\ to\ part\ A}{=} \binom{m_1}{k}[1-\Phi(Z_{1-\alpha} - \mu_1)]^k \cdot [\Phi(Z_{1-\alpha} - \mu_1)]^{m_1-k}$$

This outcome is applied, because the p-values of the hypotheses are independent and identically distributed.

S counts the number of the false null hypotheses (m_1) that are rejected (w.p $1 - \Phi(Z_{1-\alpha} - \mu_1)$).

Therefore,

$$S \sim Bin(m_1, 1 - \Phi(Z_{1-\alpha} - \mu_1))$$

Part D.

1.

It is given that $m = 2$.

For the case of $\underline{m_0 = 0}$:

$$\underline{FWER} = 1 - (1-\alpha)^0 = 0$$

$$\underline{PFER} = m_0\alpha = 0$$

$$m_0 = 0 \Rightarrow V = 0 \Rightarrow FDP = 0 \Rightarrow \underline{FDR} = 0$$

For the case of $\underline{m_0 = 1}$:

$$\underline{FWER} = 1 - (1-\alpha)^1 = \alpha$$

$$\underline{PFER} = m_0\alpha = \alpha$$

$m_0 = 1 \Rightarrow V \leq 1 \Rightarrow FDP \leq 1$.

$$P(FDP = \frac{V}{R} = k) = \begin{cases} p_1 & k = 0 \quad (V = 0 \ or \ R = 0) \\ p_2 & k = 1 \quad (V = R = 1) \\ p_3 & k = 0.5 \quad (V = 1, R = 2) \end{cases}$$

$\Rightarrow FDR = \mathbb{E}[FDP] = 0 \cdot P_1 + 1 \cdot P_2 + 0.5 \cdot P_3$

$= 0 \cdot P(V = 0 \vee R = 0) + 1 \cdot P(V = R = 1) + 0.5 \cdot P(V = 1, R = 2)$

$= P(V = R = 1) + 0.5 \cdot P(V = 1, R = 2) = P(V = 1, S = 0) + 0.5 \cdot P(V = 1, S = 1)$

$= P(H_{0_1} \ is \ true) \cdot P_{H_0}(P_1 \leq \alpha) \cdot P_{H_1}(P_2 > \alpha) + P(H_{0_2} \ is \ true) \cdot P_{H_1}(P_1 > \alpha) \cdot P_{H_0}(P_2 \leq \alpha) +$

$+0.5 \cdot [P(H_{0_1} \ is \ true) \cdot P_{H_0}(P_1 \leq \alpha) \cdot P_{H_1}(P_2 \leq \alpha) + P(H_{0_2} \ is \ true) \cdot P_{H_0}(P_2 \leq \alpha) \cdot P_{H_1}(P_1 \leq \alpha)]$

$\underbrace{=}_{(*,***)} 0.5 \cdot P_{H_0}(P_1 \leq \alpha) \cdot P_{H_1}(P_2 > \alpha) + 0.5 \cdot P_{H_1}(P_1 > \alpha) \cdot P_{H_0}(P_2 \leq \alpha) +$

$+0.5 \cdot [P_{H_0}(P_i \leq \alpha) \cdot P_{H_1}(P_i \leq \alpha)]$

$\underbrace{=}_{(**,***)} P_{H_0}(P_i \leq \alpha) \cdot P_{H_1}(P_i > \alpha) + 0.5 \cdot [\alpha \cdot (1 - \Phi(Z_{1-\alpha} - \mu_1))]$

$\underbrace{=}_{(**)} \alpha \cdot \Phi(Z_{1-\alpha} - \mu_1) + 0.5 \cdot [\alpha \cdot (1 - \Phi(Z_{1-\alpha} - \mu_1))]$

$$\Rightarrow \underline{FDR} = 0.5\alpha[1 + \Phi(Z_{1-\alpha} - \mu_1)]$$

$(*)$ Given that m, m_0 are both known, the true null hypotheses are distributed uniformal over all the null hypotheses. Thus given that $m = 2, m_0 = 1$ the probability that each null hypothesis is true is 0.5.

$(**)$ Part A.

$(***)$ Pvalues of true null hypotheses are i.i.d. and Pvalues of false null hypotheses are i.i.d.

For the case of $\underline{m_0 = 2}$:

$$\underline{FWER} = 1 - (1-\alpha)^2 = 2\alpha - \alpha^2$$

$$m_0 = m \Rightarrow V = R \Rightarrow FDP = I(V > 0)$$

$$\Rightarrow \underline{FDR} = P(V > 0) = FWER = 2\alpha - \alpha^2$$

$$\underline{PFER} = m_0\alpha = 2\alpha$$

For any $\underline{m_0 < m}$:

$\Pi^{AVG} = 1 - \Phi(Z_{1-\alpha} - \mu_1)$

And for $m_0 = m = 2$:

$\Pi^{AVG} = 0$, because all hypotheses are right. Therefore, no rejections are justified.

Summary of the theoretical measures for each m_0, as a function of α and μ_1:

	FWER	PFER	FDR	AVG-Power
$m_0 = 0$	0	0	0	$1-\Phi(Z_{1-\alpha}-\mu_1)$
$m_0 = 1$	α	α	$0.5\alpha[1 + \Phi(Z_{1-\alpha} - \mu_1)]$	$1-\Phi(Z_{1-\alpha}-\mu_1)$
$m_0 = 2$	$2\alpha - \alpha^2$	2α	$2\alpha - \alpha^2$	0

2.

For $\alpha = 0.05$:

Theoretical Values for $m = 2, \mu = 1$, as a function of m_0:

m_0	FWER	PFER	FDR	avg.power
0	0.0000	0.00	0.0000	0.26
1	0.0500	0.05	0.0435	0.26
2	0.0975	0.10	0.0975	0.00

Table 2: Simulation results for m = 2, mu = 1, as a function of m0

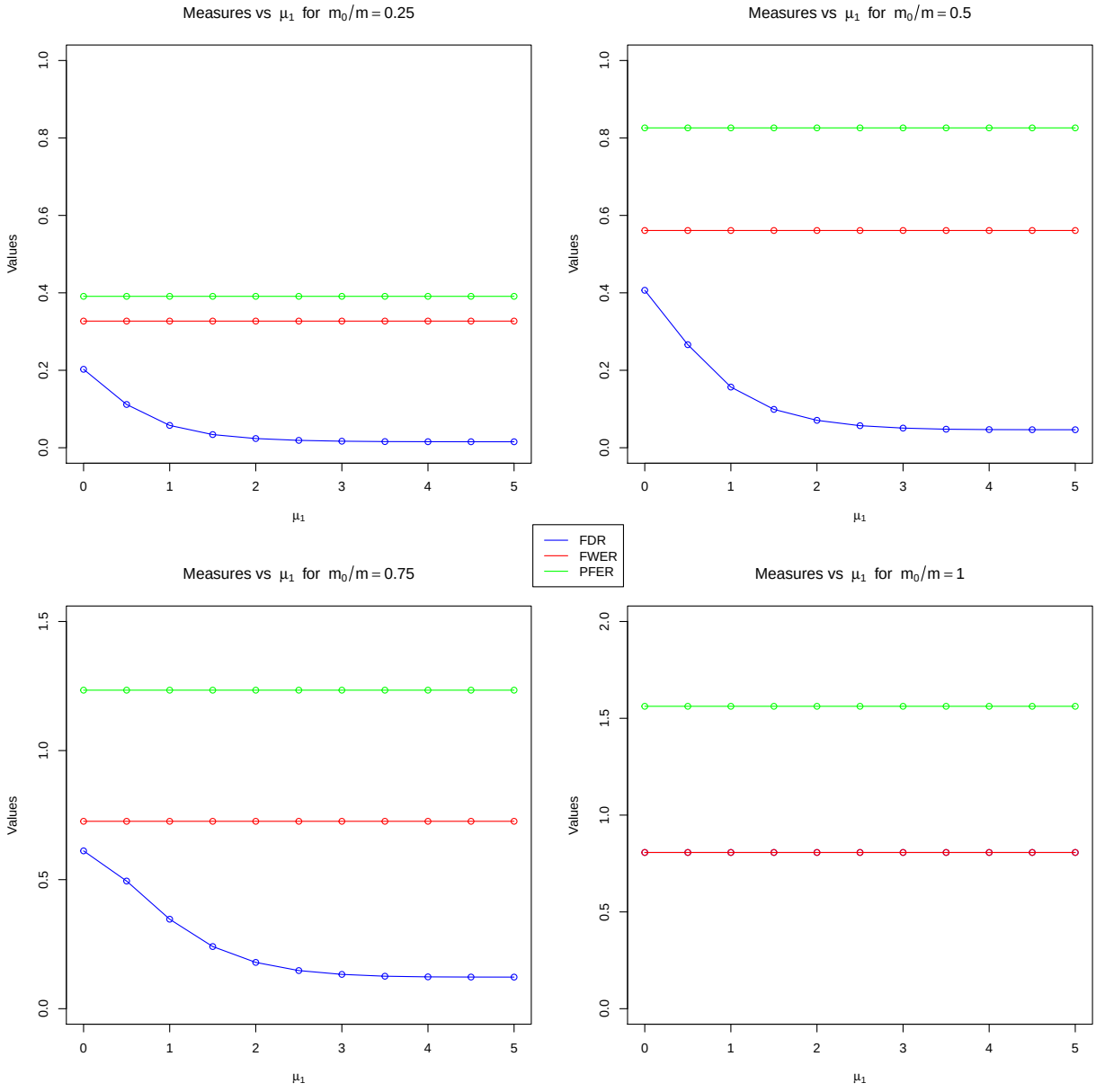
m_0	FWER	PFER	FDR
0	0.0000	0.0000	0.0000
1	0.0470	0.0470	0.0411
2	0.1013	0.1044	0.1013

As can be seen, the measures obtained from the simulation are close to the theoretical values, as expected.

Part E

1.

Error measures as a function of mu_1, for different $\frac{m_0}{m}$ proportions, given $m = 32$:



*Note that the Y axis is different in the different plots

**Note that for $\frac{m_0}{m} = 1$, it is proved that FDR = FWER. In the bottom right graph the red (FWER) is on top of the blue (FDR), therefore cannot be seen.

2.

In this section we assume $\mu_1 = 1$

Table 3: Results for m = 8

m_0/m	FWER	PFER	FDR
0.25	0.098	0.101	0.0470333
0.50	0.184	0.200	0.1106667
0.75	0.285	0.335	0.2237500
1.00	0.338	0.418	0.3380000

Table 4: Results for m = 32

m_0/m	FWER	PFER	FDR
0.25	0.340	0.403	0.0590837
0.50	0.549	0.777	0.1546414
0.75	0.720	1.210	0.3443079
1.00	0.797	1.576	0.7970000

Table 5: Results for m = 512

m_0/m	FWER	PFER	FDR
0.25	0.998	6.262	0.0591498
0.50	1.000	12.933	0.1624974
0.75	1.000	19.142	0.3650279
1.00	1.000	25.597	1.0000000

Table 6: Results for m = 1000

m_0/m	FWER	PFER	FDR
0.25	1	12.556	0.0603422
0.50	1	24.738	0.1598791
0.75	1	37.927	0.3679836
1.00	1	49.955	1.0000000

Part F.

1.

There is an order between these three error rate metrics: $PFER \geq FWER \geq FDR$. This order is consistent through all values of μ_1 of which we used in part E, as can be seen in the simulations and the plots.

First we prove that $FDR \leq FWER$:

1. If $m_0 = m$, then $FDR = FWER$.

Proof:

$m_0 = m \Rightarrow V = R \Rightarrow$

$$FDP = \begin{cases} \frac{V}{R} = 1 & R = V > 0 \\ 0 & R = V = 0 \end{cases}$$

$\Rightarrow FDP = I(V > 0)$

$\Rightarrow FDR = \mathbb{E}[I(V > 0)] = P(V > 0) = FWER$

2. If $m_0 < m$, then FDR≤FWER.

Proof:

Recall that $FWER = P(V > 0) = \mathbb{E}[I(V > 0)]$, $FDR = \mathbb{E}(FDP)$. We can notice that it is sufficient to prove that FDP$\leq I(V > 0)$.

$$FDP = \begin{cases} \frac{V}{R} & R > 0 \\ 0 & R = 0 \end{cases}$$

It holds that

$\forall R \geq 0 : V \leq R \Rightarrow FDP \leq 1 \Rightarrow$

a. If $I(V > 0) = 1$, then $FDP \leq I(V > 0)$.

b. If $I(V > 0) = 0$, this means that $V = 0$ and then $FDP = 0 \leq I(V > 0)$.

We got that for all $0 \leq m_0 \leq m$ it holds that $FDR \leq FWER$.

This result is independent of the procedure used (according to the proof above) thus, it is possible to conclude that $FDR \leq FWER$ for any procedure.

Proof that $FDR \leq FWER \leq PFER$:

$$PFER = \mathbb{E}(V) = \sum_{v=0}^{\infty} v \cdot P(V = v) = \sum_{v=1}^{\infty} v \cdot P(V = v) \geq \sum_{v=1}^{\infty} 1 \cdot P(V = v) = P(V > 0) = FWER$$

$$\Rightarrow FWER \leq PFER$$

This result is independent of the procedure used (according to the proof above) thus, it is possible to conclude that $FWER \leq PFER$ for any procedure.

And in total,

$$\Rightarrow FDR \leq FWER \leq PFER$$

For any procedure.

2.

Let us recall the following expressions applied in part B in our case (meaning test each hypothesis at level α):

$FWER = 1-(1-\alpha)^{m_0}$,$PFER = m_0\alpha$.

PFER and FWER are independent of μ_1, therefore - constant as a function of μ_1 with constant m_0, m.

This can also be seen in the plots.

When $\frac{m_0}{m} \in (0,1)$,

according to the graphs applied in part E, it seems that FDR is decreasing, as a function of μ_1 with constant m_0, m, and converges to a constant value.

This result is intuitive as V isn't affected by changes in μ_1, as as we showed before $V \sim Bin(m_0, \alpha)$ and m_0, α aren't related to μ_1. Intuitively because it measures the number of type 1 errors out of the true null hypotheses which are unaffected by the false null hypotheses because each hypothesis is tested independently at level α.

And meanwhile R potentially increases as S can increase when μ_1 increases (higher probability) as as we showed before $S \sim Bin(m_1, p = 1-\Phi(Z_{1-\alpha}-\mu_1))$ as μ_1 increases p increases. Intuitively because as the distance of μ_1 from $0(\mu_0)$ is bigger, it makes each false null hypothesis stronger (independently) thus easier to reject.

When $\frac{m_0}{m} = 0, 1$ the FDR is constant as well, as FDR equals FWER.

3.

For constant μ_1, m, all three error rate metrics are increasing while $\frac{m_0}{m}$ is increasing.

With constant m, while $\frac{m_0}{m}$ is increasing, m_0 is increasing.

As explained in part B, FWER and PFER are increasing while m_0 is increasing, thus also while $\frac{m_0}{m}$ is increasing.

This corresponds with the simulation results.

According to the tables applied in part E, it seems that while m and μ_1 are constant, in each of the four tables, through the increment of the $\frac{m_0}{m}$ values, FDR is increasing.

FDP can be written as $\frac{V}{V+S}$. Given m and μ_1, when m_0 increases m_1 decreases as $m = m_0 + m_1$.

As m_0 increases V can only increase, and S can only decrease accordingly therefore in this case $\frac{V}{V+S}$ increases as well.

For example, in the case where we change H_i that was false to true meaning $m_0 = m_0 + 1$ and $m_1 = m_1 - 1$, we separate into cases,

If beforehand H_i was not rejected and now it is also not rejected $\Rightarrow$ no change in FDP.

If beforehand H_i was not rejected and now it is rejected $\Rightarrow V = V+1, S = S \Rightarrow$ FDP increases.

If beforehand H_i was rejected and now it is also rejected $\Rightarrow V = V+1, S = S-1 \Rightarrow$ FDP increases.

If beforehand H_i was rejected and now it is not rejected $\Rightarrow$ no change in FDP.

4.

From part B:

$$FWER = 1 - (1 - \alpha)^{m_0}$$

$$PFER = m_0 \alpha$$

So, when $\underline{\frac{m_0}{m} = 1}$:

$\Rightarrow m_0 = m$

So $\underline{FWER} = 1 - (1 - \alpha)^m$ and $\underline{PFER} = m\alpha$.

We saw earlier that when $m_0 = m$, $FDR = FWER$, so $\underline{FDR} = 1 - (1 - \alpha)^m$.

When $\underline{\frac{m_0}{m} = 0}$:

$\Rightarrow m_0 = 0$,

So $\underline{FWER} = 1 - 1 = 0$ and $\underline{PFER} = 0$.

We saw earlier that when $m_0 < m$: $0 \leq FDR \leq FWER$.

So if $m_0 = 0$, then $\underline{FDR} = 0$, as $FWER = 0$.

This corresponds with the simulation results.

5.

With constant $\frac{m_0}{m}$ and μ_1, when m is increasing, all three error rate metrics are increasing, and FWER converges to 1.

We can see this behavior represented in the tables in part E.2, when examining each of the four $\frac{m_0}{m}$ values, through all four tables (m is increasing from one table to another).

5.1. With constant $\frac{m_0}{m}$ and μ_1, when m is increasing, we can conclude that m_0 is increasing as well.

Thus, we have the same results we had in section F.3, when m_0 was increasing, which means that FWER and PFER are increasing, as they are not dependent on m and μ_1 (but only on m_0 and α).

Regarding the convergence of FWER, When m_0 increases, V can stay the same or increase (depending on if the hypothesis was rejected or not).

Therefore,

$$FWER = P(V > 0) = 1 - P(V = 0) \underbrace{=}_{V \sim Bin(m_0, \alpha)} 1 - (1-\alpha)^{m_0} \underbrace{\longrightarrow}_{m_0 \to \infty} 1 - 0 = 1$$

5.2. What Benjamini and Hochberg called the false discovery rate (FDR) they denote by E(FDR), and receive:

$E(FDR) = \frac{\frac{m_0}{m}\alpha}{\frac{m_0}{m}\alpha + \frac{m_1}{m}F(\alpha)} + O(\frac{1}{\sqrt{m}})$, when $F(\alpha)$ depends on μ_1

Where F is the distribution of the p-value under the alternative. In our case, $F(\alpha) = P_{H_1}(P_i \leq \alpha))$.

*In the article the authors used c, which in our case is exactly α as we test each hypothesis at level α.

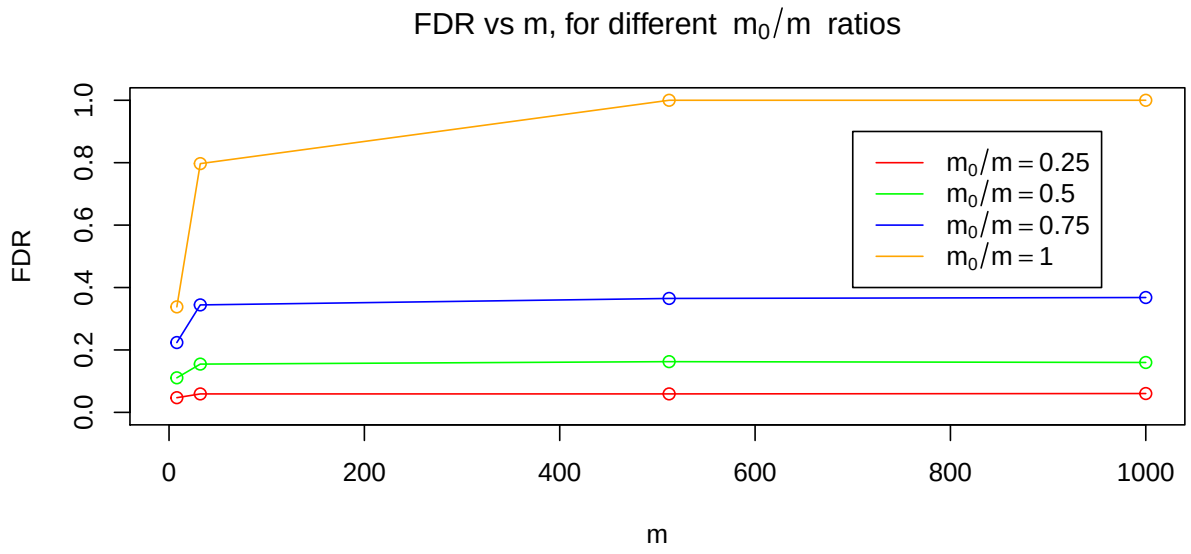
According to Genovese and Wasserman, 2002 theory, the expected FDR is controlled regardless of m_0 and regardless of the marginal distributions for the p-values corresponding to the false null hypotheses.

In their results, they used asymptotics in which the fraction of true null hypotheses $\frac{m_0}{m}$ is kept fixed.

Thus, we can see that according to their theory, $\underline{FDR}$ converges to a specfic value for large m values and constant μ_1, $\frac{m_0}{m}$.

This is as $O(\frac{1}{\sqrt{m}}) \to 0$ as $m \to \infty$, $\frac{m_1}{m} = \frac{m-m_0}{m} = 1 - \frac{m_0}{m}$ and $F(\alpha)$ (as we show below) stays constant as well.

We can see this behavior represented in the tables in part E.2, when examining each of the four $\frac{m_0}{m}$ values, through all four tables (m is increasing from one table to another). For easier visualization we plotted the results in one graph.



We saw in part A that for false null hypotheses we have:

$P(P_i \leq \alpha) = 1 - \Phi(Z_{1-\alpha} - \mu_1)$

Therefore $F_{P_i}(\alpha) = 1 - \Phi(Z_{1-\alpha} - \mu_1)$

We can notice that this result is independent of any variables except α, when μ_1 is fixed.

Therefore, we can conclude that all p-values of false null hypotheses, have the same distribution.

The equation assumes that the p-values are i.i.d. (true and false seperately) and then the authors used the binomial distribution for rejection with probability of $F(\alpha)$ to reach eq. 8 (from the article).

We showed that in our case all p-values of false null hypotheses, have the same distribution therefore, as the other conditions hold from the building method of the simulation, the equation can be used.

5.3.

$$E(FDR) = \frac{\frac{m_0}{m}\alpha}{\frac{m_0}{m}\alpha + \frac{m_1}{m}F(\alpha)} + O(\frac{1}{\sqrt{m}})$$

When $F(\alpha)\ is\ P_{H_1}(P_i \leq \alpha)$, thus according to our results, $F(\alpha) = 1 - \Phi(Z_{1-\alpha} - \mu_1)$.

$$\Rightarrow E(FDR) = \frac{\frac{1}{2}\alpha}{\frac{1}{2}\alpha + \frac{1}{2}(1 - \Phi(Z_{1-\alpha} - 1))} + 0$$

In our case, for $\alpha = 0.05$, $E(FDR) = \frac{\frac{1}{2}\cdot 0.05}{\frac{1}{2}\cdot 0.05 + \frac{1}{2}(1-\Phi(Z_{1-0.05}-1))} = \frac{\frac{1}{2}\cdot 0.05}{\frac{1}{2}\cdot 0.05 + \frac{1}{2}(1-\Phi(Z_{0.95}-1))} = \frac{0.025}{0.025+\frac{1}{2}(1-\Phi(0.645))}$
$= \frac{0.025}{0.025+\frac{1}{2}(1-0.7405364)} = \frac{0.025}{0.025+\frac{1}{2}(0.2594636)}$

$$\Rightarrow \underline{E(FDR) = 0.1615699}$$

This corresponds with the results we got in the simulation, as can be seen in the plot of section 5.2 and in the table in section E.2.

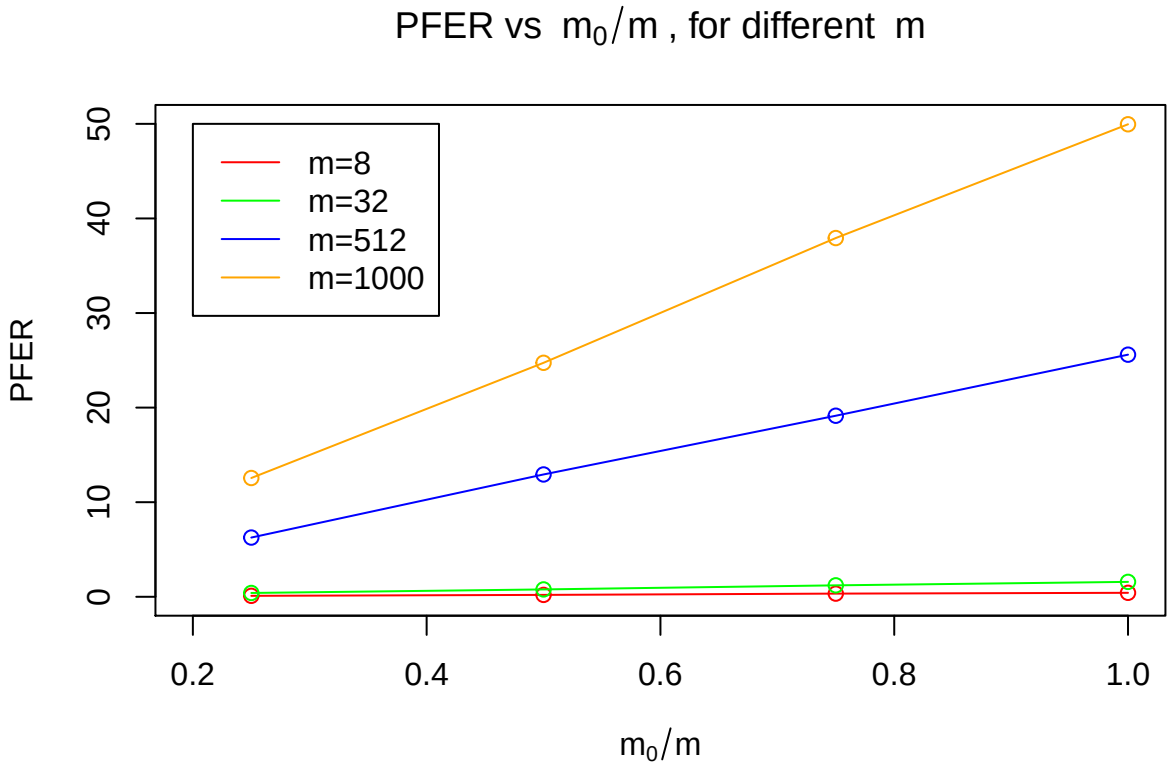
6.

6.1 Without accounting for multiple comparisons, all error measures for all tested values of m and $\frac{m_0}{m}$ are not controlled at level α for small enough μ_1.

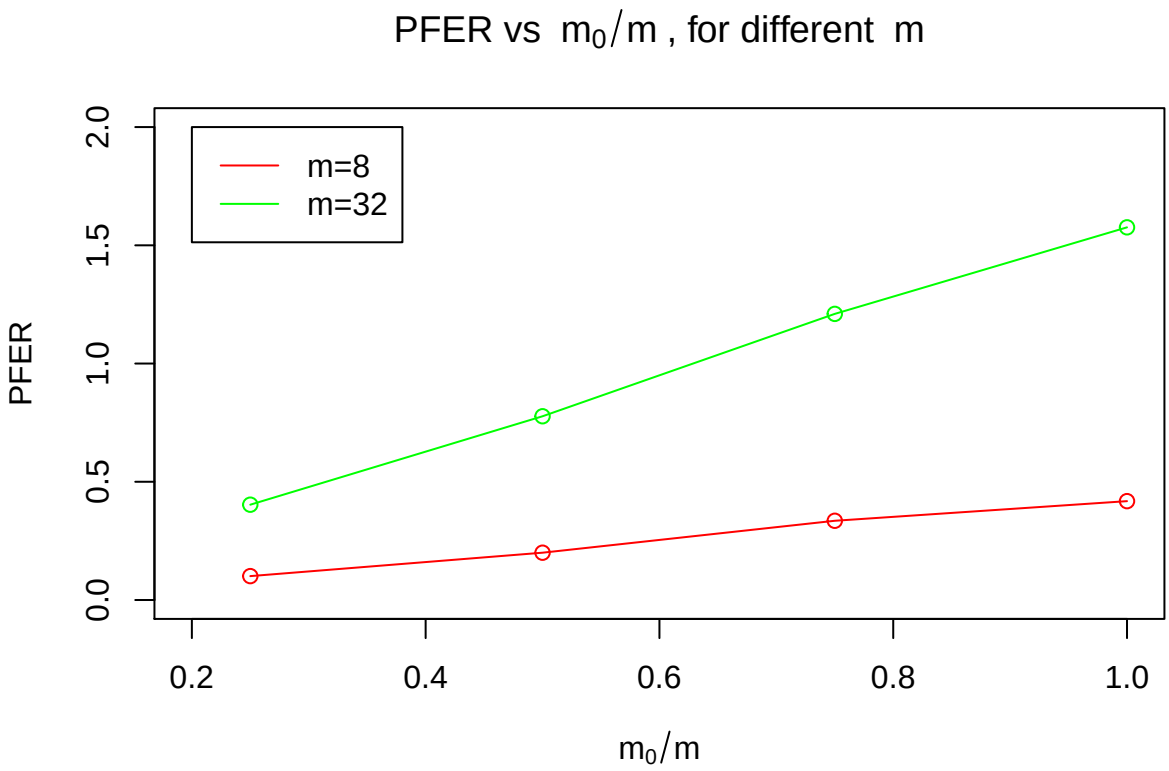
For higher μ_1 values, FWER is still not controlled.

This result corresponds with what we learned all year - multiple comparisons require adjustments if control of the error measures is desired.

6.2 PFER as a function of $\frac{m_0}{m}$, for different values of m



And zoomed in on the case where $m = 8, 32$:



As can be seen, PFER increases linearly as the proption of m_0 out of m increases.

This result correspond with the theoretical function that we've shown - $PFER = m_0 \cdot \alpha$.

6.3 Note that for $\frac{m_0}{m} = 1$, we proved in section F.1. that FDR = FWER.

From the simulation results as displayed in the tables in section E.2 we can see that the simulation corresponds with the theory.

6.4 From theory we know that $FWER \leq 1, FDR \leq 1$.

From the simulation results as displayed in the tables and graphs in section E we can see that the simulation corresponds with the theory.

6.5 Given a $\frac{m_0}{m}$ proportion ,if m grows times c, m_0 grows times c as well. Therefore PFER grows times c as well.

This can be seen in simulation table results, for example - for $\frac{m_0}{m} = 0.5$, when $m = 8 : PFER \approx 0.2$, and when $m = 32 : PFER \approx 0.8$. Meaning m increased times 4. PFER increased times 4 as well.

This result correspond with the theoretical function that we've shown - $PFER = m_0 \cdot \alpha$.

Task 2 - Unit 1

Part A.

As learnt in class, if the test statistics are independent and continuous FDR of the BH procedure for the hypotheses is $FDR_q = \frac{m_0}{m} \cdot q$.

Part B.

Define $\tilde{q} = \frac{m}{m_0} \cdot q$.

As $m \geq m_0$ we get that $\tilde{q} \geq q$ as desired.

Under the assumptions of part A,

$$FDR_{\tilde{q}} = \frac{m_0}{m} \cdot \tilde{q} \underbrace{=}_{\text{by def. of } \tilde{q}} \frac{m_0}{m}\frac{m}{m_0} \cdot q = q$$

$\tilde{q}$ depends only on q and on m_0 when m is known, as required.

And FDR is controlled at level q by definition, and furthermore, FDR equals q.

Part C.

The thresholds for the BH procedure at level q are $\forall i \in \{1, \cdots, m\} : c_i = \frac{i \cdot q}{m}$

Therefore for $\tilde{q} = \frac{m}{m_0} \cdot q$ we get

$\forall i \in \{1, \cdots, m\} : \tilde{c_i} = \frac{i \cdot \tilde{q}}{m} = \frac{i \cdot q \frac{m}{m_0}}{m} = \frac{i \cdot q}{m_0}$

As $m_0 \leq m$ we get that $\forall i : \tilde{c_i} \geq c_i$.

In BH procedure at level q we look for the k such that $k_{BH_q} = max\{i : P_{(i)} \leq \frac{qi}{m}\}$. If we compare each PV to a bigger or equal thershold we get a K only equal or higher. So it holds that

$$k_{BH_q} \leq k_{BH_{\tilde{q}}}$$

As the K_{BH} is equal or larger when using $\tilde{q}$, by def of BH we reject more hypotheses and therefore the group of rejections of the BH procedure at level q is contained in the group of rejections of the BH procedure at level $\tilde{q}$.

Formally,

$$k_{BH_q} = max\{i : P_{(i)} \leq \frac{qi}{m}\} \Rightarrow P_{(k_{BH_q})} \leq \frac{qk_{BH_q}}{m} \leq \frac{\tilde{q}k_{BH_q}}{m}$$
$$\Rightarrow P_{(k_{BH_q})} \leq \frac{\tilde{q}k_{BH_q}}{m}$$
$$\Rightarrow k_{BH_q} \leq k_{BH_{\tilde{q}}}$$

Where,

$$k_{BH_{\tilde{q}}} = max\{i : P_{(i)} \leq \frac{\tilde{q}i}{m}\}$$

This is because we found a PV, specifically $P_{(k_{BH_q})}$ which is smaller or equal to its threshold when using BH at level $\tilde{q}$, and because we are looking for the maximum k then $k_{BH_{\tilde{q}}}$ can only be equal or greater than k_{BH_q}.

The BH procedure at level q rejects $H_{(1)} \cdots H_{(k_{BH_q})}$.

The BH procedure at level $\tilde{q}$ rejects $H_{(1)} \cdots H_{(k_{BH_{\tilde{q}}})}$.

Therefore

$$\mathfrak{R}^{BH_q} \subseteq \mathfrak{R}^{BH_{\tilde{q}}}$$

Part D.

From part C,

$$\mathfrak{R}^{BH_q} \subseteq \mathfrak{R}^{BH_{\tilde{q}}}$$

Therefore,

$$\mathbb{S}^{BH_q} = \mathfrak{R}^{BH_q} \cap \mathbb{M}_1 \subseteq \mathfrak{R}^{BH_{\tilde{q}}} \cap \mathbb{M}_1 = \mathbb{S}^{BH_{\tilde{q}}}$$

Where $\mathbb{M}_1$ is the group of false hypotheses, and $|\mathbb{M}_1| = m_1$, and $\mathbb{S}$ is the group of false hypotheses that were rejected, and $|\mathbb{S}| = S$.

Therefore

$$S^{BH_q} \leq S^{BH_{\tilde{q}}}$$

And finally,

$$\Pi^{AVG}_{BH_q} \underbrace{=}_{Def.\ S} \mathbb{E}(\frac{S^{BH_q}}{m_1}) \leq \mathbb{E}(\frac{S^{BH_{\tilde{q}}}}{m_1}) = \Pi^{AVG}_{BH_{\tilde{q}}}$$

Part E.

Given independent and continuous test statistics, q and m_0, we would use the BH procedure at level $\tilde{q} = \frac{m}{m_0} \cdot q$ and would get that BH still controls FDR at level q, by part A.

As we showed before (part C), this would result in an equal or higher number of rejected hypotheses, meaning higher power compared to BH at level q(Part D).

Thus we would get a more powerful prodecure, while still controlling FDR at the given level.

Part F.

Generalised adaptive linear step-up procedure at level q:

Let P_i be the observed p-values of the test for H_i.

1. Compute $\hat{m_0}$
2. If $\hat{m_0} = 0$ reject all hypotheses.
3. Else, test the hypotheses using the linear step-up procedure at level $\hat{q} = \frac{qm}{\hat{m_0}}$.

Where the linear step-up procedure at level $\hat{q} = \frac{qm}{\hat{m_0}}$ is defined as:

Order the pvalues.

If $P_{(m)} \leq \frac{mq}{\hat{m_0}}$ reject $H_{(1)}, \cdots, H_{(m)}$ and stop.

Else if $P_{(m-1)} \leq \frac{(m-1)q}{\hat{m_0}}$ reject $H_{(1)}, \cdots, H_{(m-1)}$ and stop.

...

Else if $P_{(1)} \leq \frac{q}{\hat{m_0}}$ reject $H_{(1)}$ and stop.

Otherwise, reject no hypothesis.

Using the adaptive linear step-up procedure, assuming that $\hat{m_0} \leq m$:

$$\forall i : \hat{c_i} = \frac{i \cdot q}{\hat{m_0}} \geq \frac{i \cdot q}{m} = c_i$$

This means that the adaptive method is more permissive (less conservative) than BH, thus we would get a more powerful prodecure, as we showed in the previous sections. Note that in this case FDR control is not promised, as it depends on the method of calculation of $\hat{m_0}$.

Part G.

The estimator for FDP that the BH procedure makes use of is

$$\hat{FDP}(\alpha) = \frac{m\alpha}{R(\alpha)}$$

Where $R(\alpha) = \sum_{i=1}^{m} I(P_i \leq \alpha)$, the number of hypotheses such that their p values are less than α, meaning the number of rejected hypotheses.

Part H.

BH adaptive at level q uses the BH procedure at level $\tilde{q}$, therefore:

$$k_{BH-ADP} = max\{i : P_{(i)} \leq \frac{\tilde{q}i}{m} = \frac{qi}{\hat{m_0}}\}$$

We will prove that $P_{(k_{BH-ADP})}$ is the solution of the next maximum optimization problem:

$$\max_{\alpha} \;\; R(\alpha) \;\; s.t. \frac{\hat{m_0}\alpha}{R(\alpha)} \leq q$$

we will show that we can look only at $\alpha \in P_{(1)}, ..., P_{(m)}$ to find the solution of the problem.

Define $P_{(m+1)} = 1$.

(1) If for some i it holds that $P_{(i)} \leq \alpha \leq P_{(i+1)}$ so $R(\alpha) = i$ because we reject $H_{(1)}, \cdots, H_{(i)}$
(2) In addition assume α sustains the constraint of the problem

From (1) and (2):

$$\frac{\hat{m_0}P_{(i)}}{i} \underbrace{\leq}_{1} \frac{\hat{m_0}\alpha}{R(\alpha)} \underbrace{\leq}_{2} q$$

so we can look only on $\alpha \; s.t. \alpha \in \{P_{(1)}, ..., P_{(m)}\}$

for $\alpha = P_{(k_{BH-ADP})}$ we get:

(3) $R(\alpha) = k_{BH-ADP}$ because we reject $H_{(1)}, \cdots, H_{(k_{BH-ADP})}$

(4) $P_{(k_{BH-ADP})} \leq \frac{k_{BH-ADP} \; q}{\hat{m_0}}$ by def of procedure.

From (3) and (4) we get:

$$\frac{\hat{m_0}P_{(k_{BH-ADP})}}{k_{BH-ADP}} \leq q$$

$$\rightarrow \frac{\hat{m_0}\alpha}{R(\alpha)} = \frac{\hat{m_0}P_{(k_{BH-ADP})}}{k_{BH-ADP}} \leq q$$

so the constraint of the optimization problem holds, and from def of the procedure:

$$P_{(k_{BH-ADP}+1)} > \frac{(k_{BH-ADP}+1)q}{\hat{m_0}}$$

$$\rightarrow \frac{P_{(k_{BH-ADP}+1)}\hat{m_0}}{k_{BH-ADP}+1} > q$$

so we got that k_{BH-ADP} is the maximum α that holds the constraints.

$\widehat{FDP} = \widehat{\frac{V(\alpha)}{R(\alpha)}}$ when $\hat{V(\alpha)} = \hat{m_0}\alpha$ so the constraint is exactly $\widehat{FDP_{adp}} \leq q$

Explanation on the reasoning behind $\widehat{FDP}$:

$$FDP(\alpha) = \frac{V(\alpha)}{R(\alpha)}$$

We don't know $V(\alpha)$ so to estimate it we use $\mathbb{E}(V(\alpha))$

$$\mathbb{E}(V(\alpha)) = m_0 P(P_i \leq \alpha) \underbrace{\leq}_{(*)} m_0 \alpha$$

(*) Under $H_0 : P(P_i \leq \alpha) \leq \alpha$

We take the estimator for the upperbound of $\mathbb{E}(V(\alpha))$, which is $\widehat{V(\alpha)} = \hat{m_0}\alpha$

Part I

for the optimization problems :

(1)BH: $\underset{\alpha}{max} \;\; R(\alpha) \;\; s.t. \widehat{FDP} \leq q$

(2)ADP_BH: $\underset{\alpha}{max} \;\; R(\alpha) \;\; s.t. \widehat{FDP_{adp}} \leq q$

assuming $\hat{m_0} \leq m$ we get: $\widehat{FDP_{adp}} = \frac{\hat{m_0}\alpha}{R(\alpha)} \leq \frac{m\alpha}{R(\alpha)} = \widehat{FDP}$

from the assumption $\hat{m_0} \leq m$ we get: $\alpha^*_{adp} \geq \alpha^*$

this is because given α^* the solution of problem (1), in problem (2) we decrease the term that we have upperbound constraints for, therefor the α^*_{adp} can only be equal or greater than α^*.

from the conclusion above and Given q for both procedures if $H_j \in \mathfrak{R}^{BH}$ it holds that $p_{(j)} \leq \alpha^* \leq \alpha^*_{adp} \Rightarrow p_{(j)} \leq \alpha^*_{adp} \Rightarrow H_j \in \mathfrak{R}^{BH-ADP}$

Meaning $\mathfrak{R}^{BH} \subseteq \mathfrak{R}^{BH-ADP}$

Part J

The motivation to create an adaptive BH procedure is to get a more powerful procedure.

As we showed in Part D $\Pi^{AVG}_{BH_q} \leq \Pi^{AVG}_{BH_{\tilde{q}}}$.

An adaptive BH procedure at level q is basically the BH procedure at level $\tilde{q}$

Therefore an adaptive procedure at level q would result in $\Pi^{AVG}_{BH_q} \leq \Pi^{AVG}_{ADP-BH_q}$.

Note that even though it is not guaranteed that FDR is controlled at level q if $\hat{m_0} \approx m_0$ it can holds.

Task 2 - Unit 2

Part A.

```r
#' The BH (Linear Step Up) Procedure
#'
#' @param pvalues - vector of p-values of length m
#' @param q - the desired BH threshold, default is 0.05
#'
#' @return rejected_indexes, the indexes corresponding to the pvalues
#'         related to the hypotheses that the BH Procedure at level q rejects
BH_Procedure <- function(pvalues, q=0.05) {
  pv_order <- order(pvalues) # Pvalues Order (indexes) in increasing order
  m <- length(pvalues)
  i <- m
  # Step up itetarions
  while (i >= 1) {
    threshold <- i * q / m # BH threshold for the i'th ordered pvalue

    # On the first time entering the if condition -
    # stop and rejected hyptoheses 1 through i (including)
    if (pvalues[pv_order[i]] <= threshold) {
      rejected_indexes <- pv_order[1:i]
      return(rejected_indexes)
    }
    i <- i - 1
  }
}
```

```r
#' The Median Adaptive BH (Linear Step Up) Procedure
#'
#' @param pvalues - vector of p-values of length m
#' @param q - the desired BH threshold (before applying the adaptive method),
#'            default is 0.05
#'
#' @return rejected_indexes, the indexes corresponding to the pvalues
#'         corresponding to the hypotheses that the
#'         Median Adaptive BH Procedure at level q rejects
BH_ADP_Procedure <- function(pvalues, q = 0.05) {
  m <- length(pvalues)
  sorted_pv <- sort(pvalues) # Sorts the pvalues in ascending order

  # Estimate m0 according to Article definitions
  # Ceiling is used to handle the case where m is odd.
  # Note that the original instructions were to use floor
  m0_hat <- (m - ceiling(m / 2)) / (1 - sorted_pv[ceiling(m / 2)])

  # Update the threshold to be according to article defitions
  q <- q * m / m0_hat

  # Run the BH procedure with the *updated* q threshold
  rejected_indexes <- BH_Procedure(pvalues, q)
  return (rejected_indexes)
}
```

Part B.

We will show the construction of the estimator for m_0 as this shows the logic behind the estimator.

$R(\alpha) = \sum_{i=1}^{m} I(P_i \leq \alpha)$, the number of hypotheses such that their pvalues are less than α, meaning the number of rejected hypotheses. it holds that $m_0 = m - (R - V)$

but, we dont know V so we can estimate it: $\widehat{V(\alpha)} = \mathbb{E}(V(\alpha)) = \mathbb{E}[\Sigma_{i \in M_0} I\{H_i \; is \; rejected\}] \overset{linearity+expectancy \; of \; indicator}{\widehat{=}}$

$= \Sigma_{i \in M_0} P(H_i \; is \; rejected) = \Sigma_{i \in M_0} P(P_i \leq \alpha) \overset{(*)}{\widehat{=}} m_0 \alpha$

(*) the pvalues of the true null hypotheses are distributed uniform 0-1, as we proved in HW. so for $i \in M_0$: $P(p_i \leq \alpha) = \alpha$

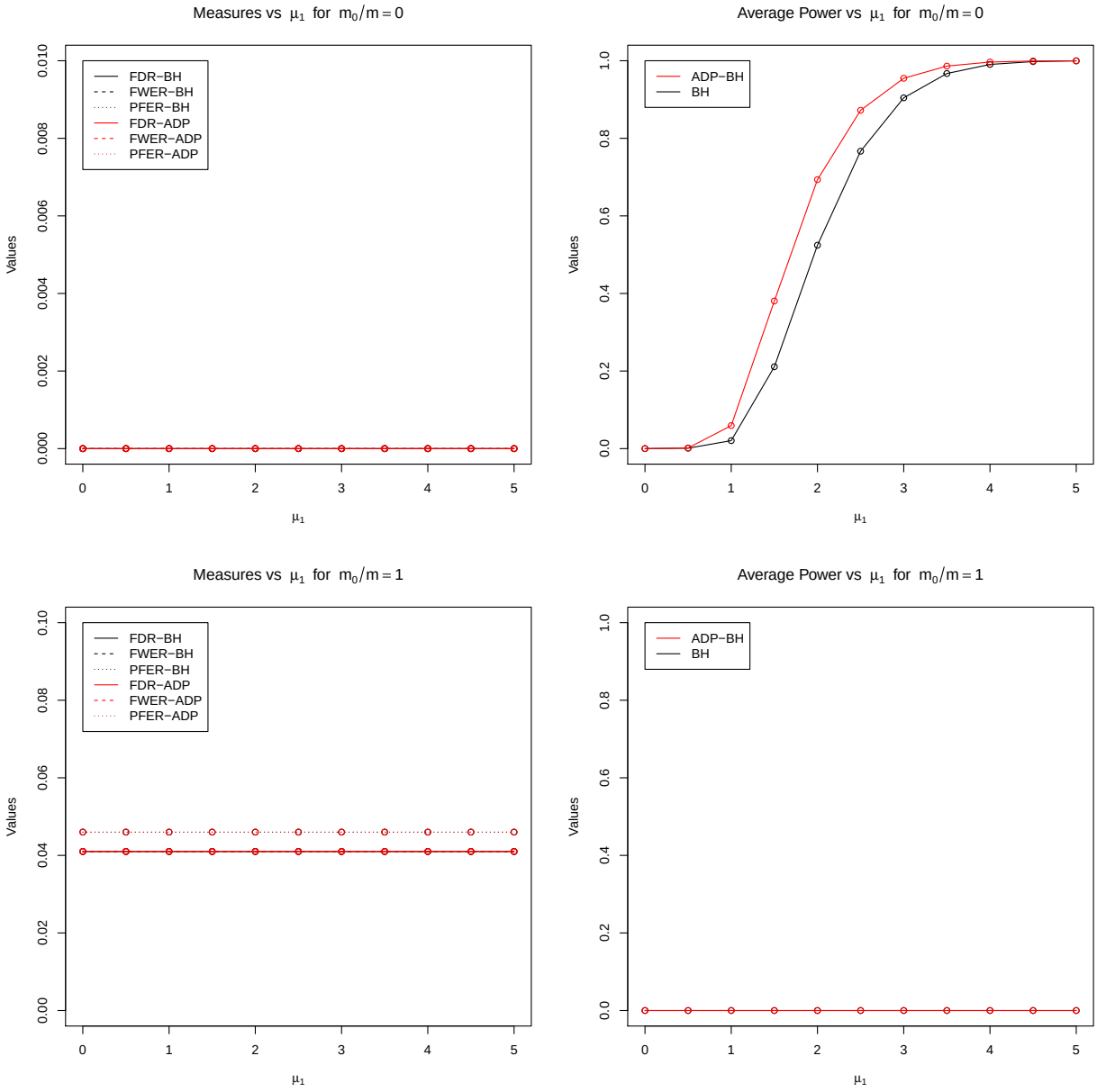
and because we don't know $m_0 \rightarrow \widehat{V(\alpha)} = \hat{m_0}\alpha$

and we get: $\hat{m_0} = m - (R(\alpha) - \hat{V(\alpha)}) = m - (R(\alpha) - \hat{m_0}\alpha) \rightarrow \hat{m_0} - \hat{m_0}\alpha = m - R(\alpha) \rightarrow \hat{m_0} = \frac{m - R(\alpha)}{1 - \alpha}$

if we take $\alpha = p_{(\lceil m/2 \rceil)}$ - the median, we reject all $p_{(1)}, .., p_{(\lceil m/2 \rceil)}$ so $R(\alpha) = \lceil m/2 \rceil$

Thus, $\hat{m_0} = \frac{m - R(\alpha)}{1 - \alpha} = \frac{m - \lceil m/2 \rceil}{1 - p_{(\lceil m/2 \rceil)}}$

Part C.



*Note that the Y axis is different in the different plots

For the case where $\frac{m_0}{m} = 0$, as $m > 0$ we get that$m_0 = 0$ and therefore by def of V, as V is the number of incorrectly rejected hypotheses, $V = 0$.

Therefore, theoretically in our case,

1. $PFER = \mathbb{E}(V|V = 0) = \mathbb{E}(0) = 0$
2. $FDR = \mathbb{E}(\frac{V}{R}|V = 0, R > 0) \cdot P(R > 0) = \mathbb{E}(\frac{0}{R}) = 0$
3. $FWER = P(V > 0|V = 0) = 0$

The simulation results lines up with what we expect in theory. This is true for both procedures, as it is indepenedent on q.

For the case where $\frac{m_0}{m} = 1$, we get that$m_0 = m$ and $m_1 = 0$ as $m_1 = m - m_0 = m - m = 0$, therefore by def of S, as S is the number of correctly rejected hypotheses, $S = 0$.

$Average\_Power = \mathbb{E}(\frac{S}{m_1}|m_1 > 0) \cdot P(m_1 > 0) + \mathbb{E}(0|m_1 = 0) \cdot P(m_1 = 0) = 0$

The simulation results lines up with what we expect in theory. This is true for both procedures, as it is indepenedent on q.

The relationship between $FWER$, and FDR, for the case where $\frac{m_0}{m} = 1$:

For the case where $\frac{m_0}{m} = 1$, we get that$m_0 = m$ and $m_1 = 0$ as $m_1 = m - m_0 = m - m = 0$, therefore by def of V, as V is the number of incorrectly rejected hypotheses, $V = R$.

Therefore,

$FDR = \mathbb{E}(\frac{V}{R}|V = R, R > 0) \cdot P(R > 0) = \mathbb{E}(1|V = R, R > 0) \cdot P(R > 0) = P(R > 0) = P(V > 0) = FWER$

In the simulation:

For the BH Procedure:

- FWER:

```
##  [1] 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041
```

- FDR:

```
##  [1] 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041
```

For the adaptive BH Procedure:

- FWER:

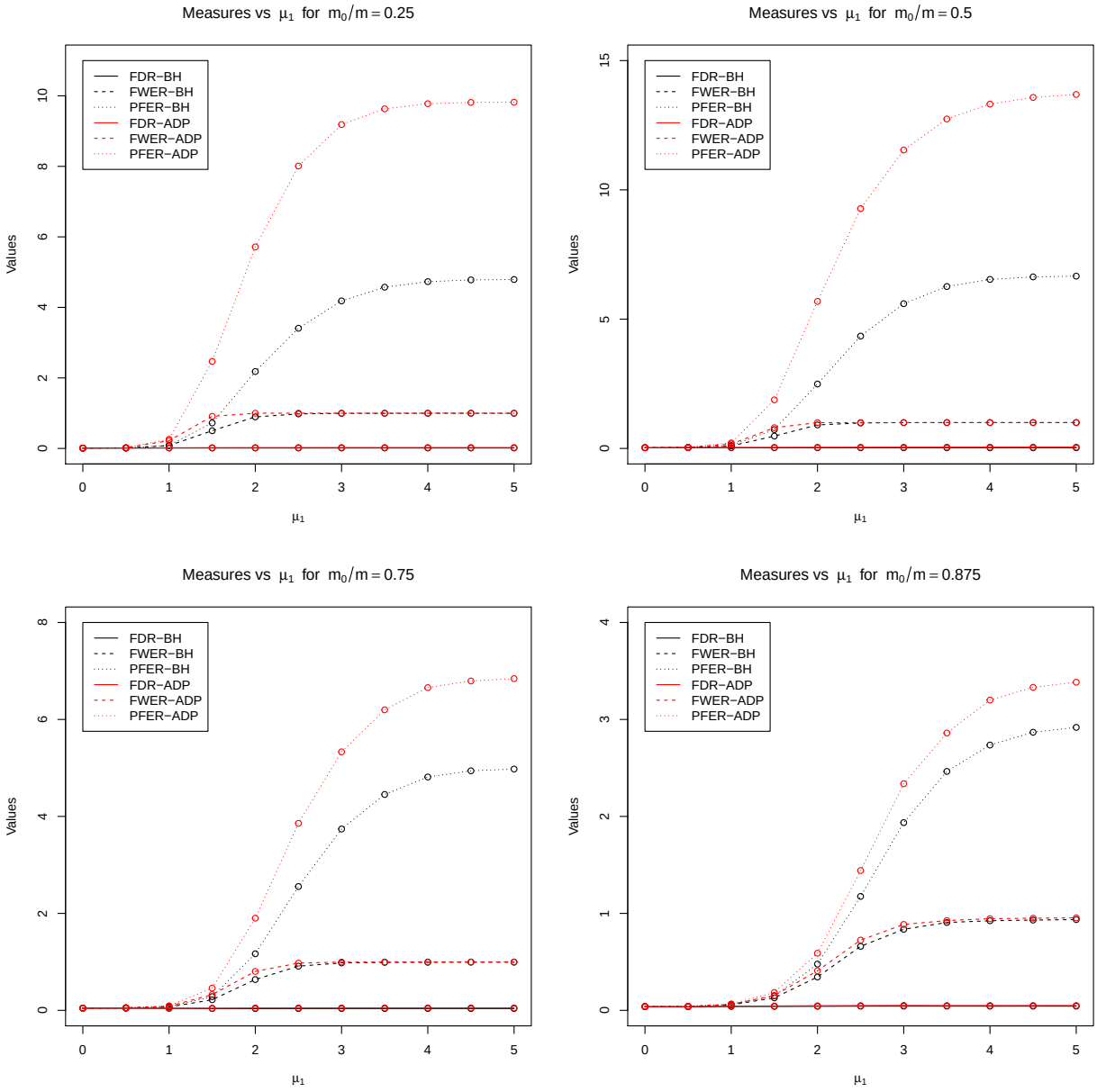
```
##  [1] 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041
```

- FDR:

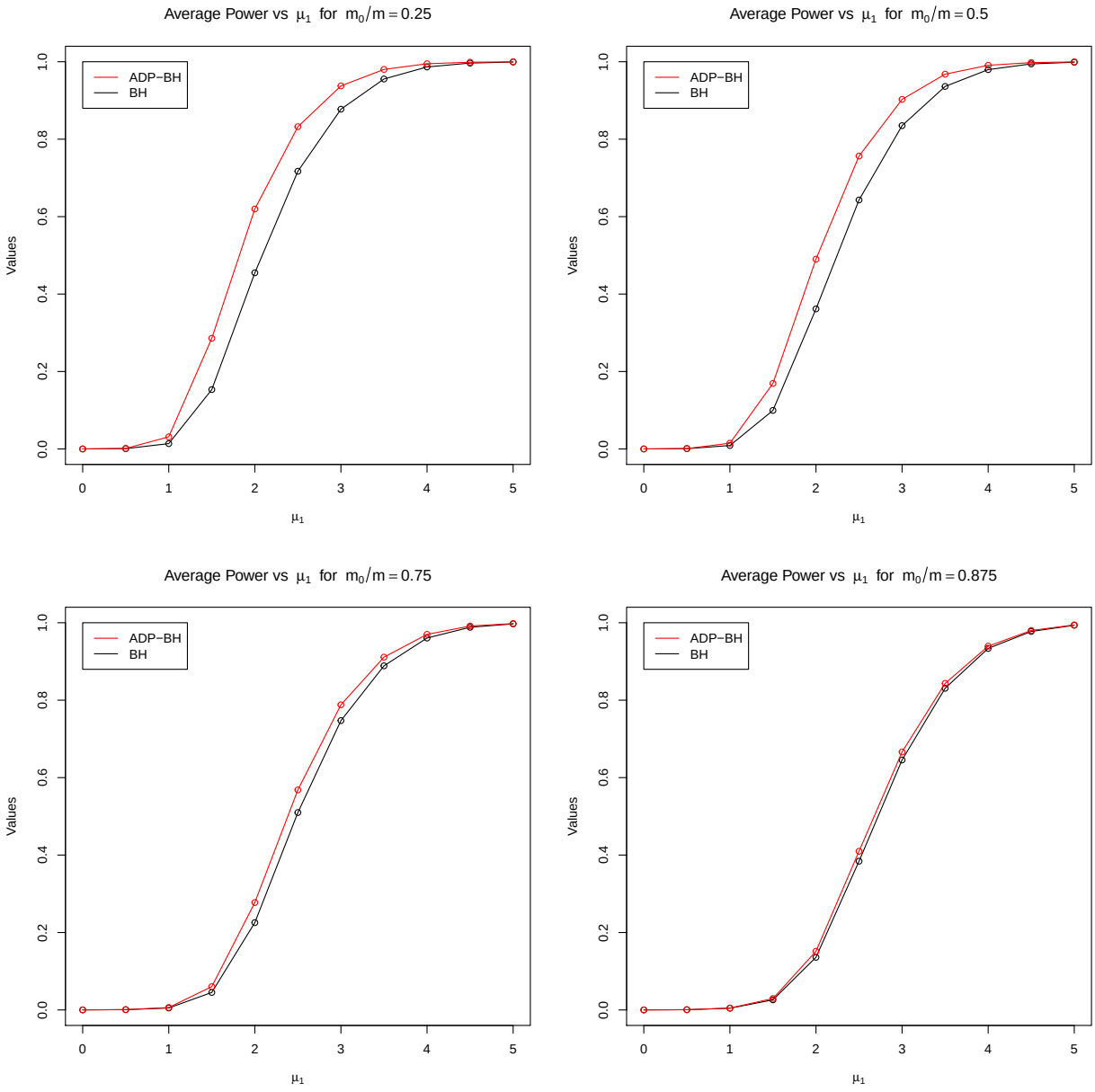
```
##  [1] 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041 0.041
```

We got that $FDR = FWER$ and the simulation results lines up with what we expect in theory. This is true for both procedures, as it is indepenedent on q.

Part D



*Note that the Y axis is different in the different plots



1.

As we proved in Unit 1 Part D, in theory $\Pi^{AVG}_{BH_q} \leq \Pi^{AVG}_{BH_{\tilde{q}}}$.

Meaning that the average power of the BH-adaptive procedure is higher or equal to that of the BH procedure.

As can be seen above, the simulation results lines up with what we expect in theory.

2.

As the proportion increases the difference in the average powers decreases.

This behaviour is expected.

Explanation:

The adaptive procedure uses an estimator for m_0.

Assuming $\hat{m_0} \leq m$ (*), and the trivial assumption that $\hat{m_0}$ has a positive correlation with m_0,

It holds that $\frac{m}{\hat{m_0}} \geq 1$.

As $\frac{m_0}{m}$ increases it means that m_0 increases (as m is const.), therefore $\hat{m_0}$ increases. Therefore $\frac{m}{\hat{m_0}}$ decreases.

Therefore, $\tilde{q} = \frac{m}{\hat{m_0}} q$ gets closer to q.

So the adaptive BH applies the BH procedure at a level that gets closer to q as the proportion increases, so the group of rejects ($\mathfrak{R}$) becomes more similar, and therefore the average power as well.

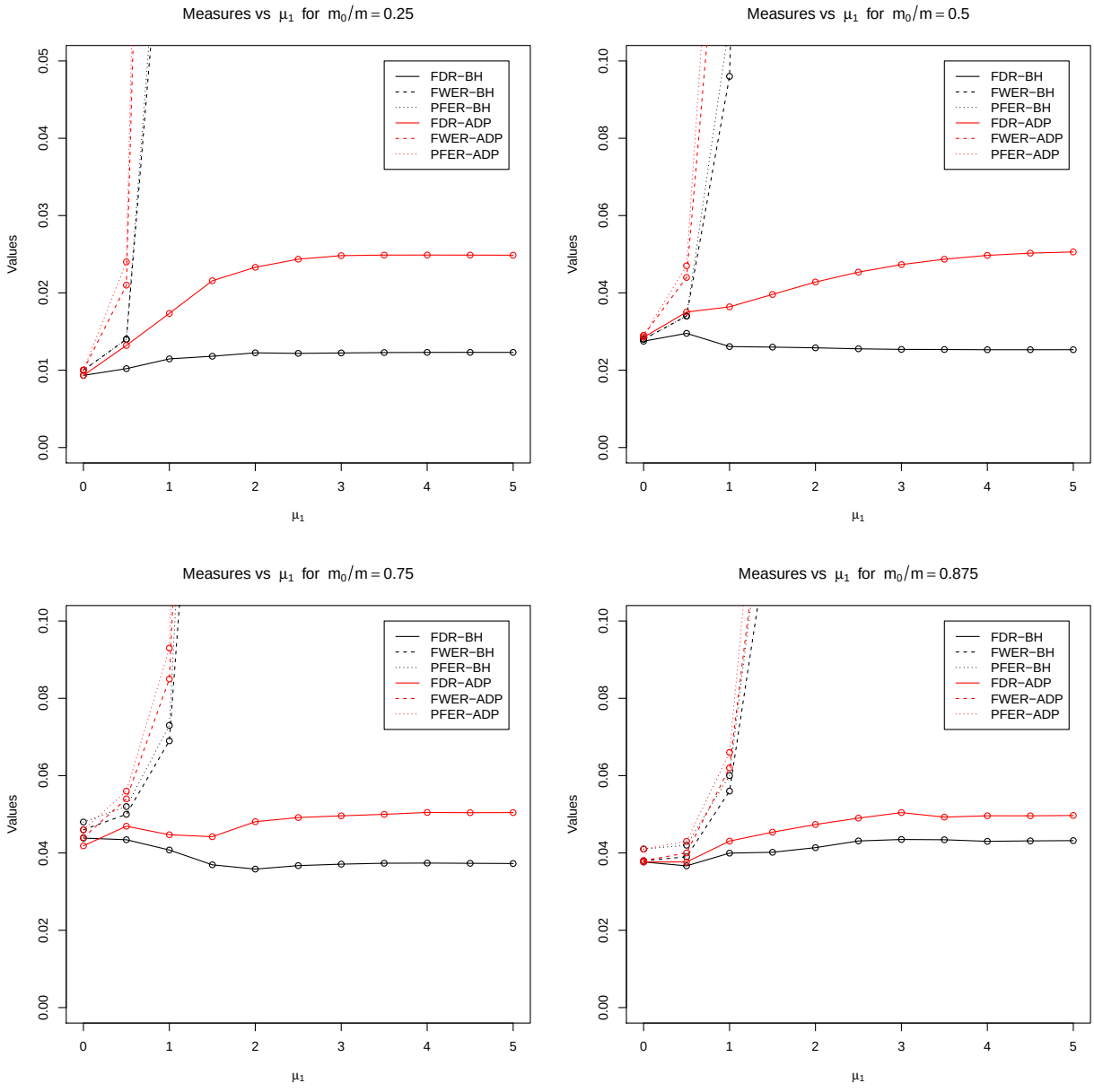
(*) $\hat{m_0} \geq m$ can happen by def of $\hat{m_0}$, but this is an extreme case, thus we ignore it.

3.

As learnt in class, if the test statistics are independent and continuous FDR of the BH procedure for the hypotheses is $FDR = \frac{m_0}{m} \cdot q$.

As a function of μ_1 the FDR is constant. It depends on the q BH-threshold, m and m_0.

Let's zoom in on the graphs:



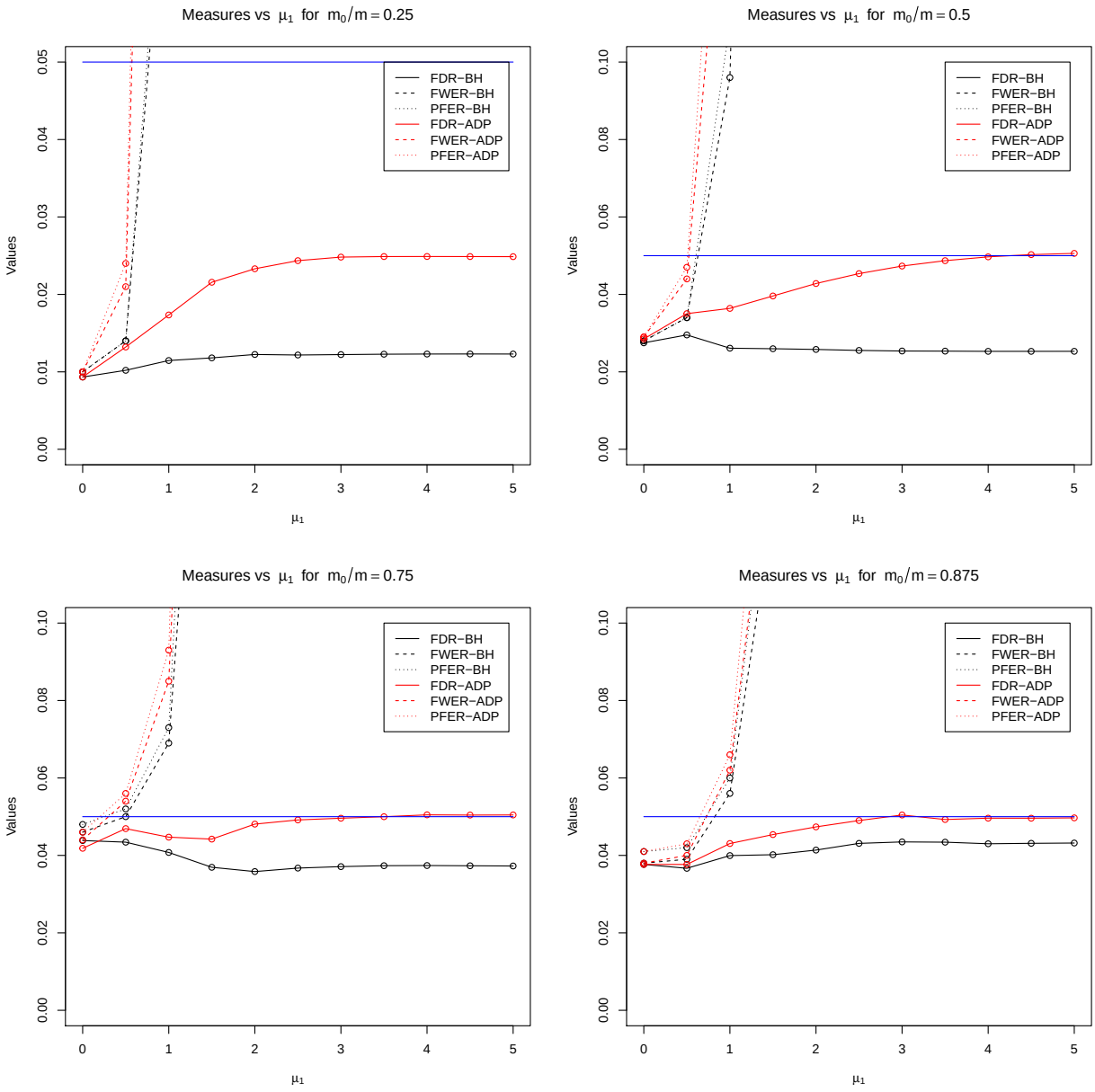
*Note that the Y axis is different in the different plots

As can be seen above, the simulation results lines up with what we expect in theory.

Note that there are slight changes in the FDR for each μ_1, as even though we drawed the random variable only once for all μ_1s, the actual pvalues still change for each μ_1 and therefore the procedures reject a few hypotheses more or less due to randomness.

4.

Let's plot the upper bound, $FDR = 0.05$:



The FDR of the adaptive BH is closer to the upper bound than the BH procedure.

According to Unit 1, Part A & B:

$FDR_{BH_q} = \frac{m_0}{m} \cdot q$

$FDR_{\tilde{q}} = \frac{m_0}{m} \cdot \tilde{q} \underbrace{=}_{\text{by def. of } \tilde{q}} \frac{m_0}{m} \frac{m}{\hat{m}_0} \cdot q = \frac{m_0}{\hat{m}_0} \cdot q$

And if $\hat{m}_0$ is close enough to m_0 then $FDR_{\tilde{q}} \approx q$

$FDR_{BH_q} \leq FDR_{ADP-BH\tilde{q}} \approx q =$ Desired FDR.

As can be seen above, the simulation results lines up with what we expect in theory.

5

5.1 FWER:

For both procedures, we can see that $FWER$ can get to 1 (it's upper bound), meaning that the procedures do not control FWER, at any reasonable level.

Specifically $FWER > 0.05$.

PFER:

For both procedures, $PFER$ can increase unboundedly, as as m and m_0 increase V can increase as well.

Specifically $PFER > 0.05$.

Meaning that the procedures do not control PFER, at any reasonable level.

5.2 For both FWER and PFER, the adaptive BH has higher values than the BH procedure.

This is intuitive.

As we proved in Unit 1 part C, the adaptive BH rejects equal or more hypotheses than BH.

Therefore it is intuitive that the adaptive BH will make more false rejections than BH, thus resulting in a higher or equal PFER and FWER, by their defitions.

Explanation:

As proven in Unit 1 part C:

$\mathfrak{R}^{BH_q} \subseteq \mathfrak{R}^{ADP-BH_{\tilde{q}}}$

Given the same M_0 for both procedures, we get $\mathbb{V}^{BH_q} \subseteq \mathbb{V}^{ADP-BH_{\tilde{q}}}$ as $\mathbb{V} = \mathfrak{R} \cap M_0$.

Therefore $V^{BH_q} \leq V^{ADP-BH_{\tilde{q}}}$, so from probability:

Thus, $FWER_{BH} = P(V_{BH} > 0) \leq P(V_{ADP-BH} > 0) = FWER_{ADP-BH}$

And, $PFER_{BH} = \mathbb{E}(V_{BH}) \leq \mathbb{E}(V_{ADP-BH}) = PFER_{ADP-BH}$

5.3 For both FWER and PFER, and for both the adaptive BH procedure and the BH procedure as μ_1 increases, the measures increase as well (till reaching an upper bound).

Explanatation:

As μ_1 increases, thus the difference between μ_1 and $\mu_0 = 0$ increases. Therefore the p-values of the false null hypotheses (M_1) gets smaller and approach 0.

In the procedures, the p-values are sorted in ascending order. Thus we get a situation in which the p-values of the false null hypotheses are (probably) smaller than the p-values of the true null hypotheses.

The p-values of the false null hypotheses increase the threshold for the p-values of the true null hypotheses, thus it is easier to reject the true null hypotheses and make a type 1 error (V increases).

Upper bound of FWER is 1 as it is a probability, and upper bound of PFER is m_0.

6.

6.1 As μ_1 increases, the average power increases, for both procedures.

This is intuitive as as μ_1 increases, p-values of the false null hypotheses (M_1) gets smaller and approach 0. Thus the procedures reject more hypotheses that increase S.

6.2 The average power increases as $\frac{m_0}{m}$ decreases (m_0 decreases), for a given μ_1.

When there are more false null hypotheses we get more justified rejections (S).

6.3 The adaptive BH procedure doesn't control FDR.

This can be seen in part D section 3, for example for $\frac{m_0}{m} = 0.875$ and for $\mu_1 = 2$ we get $FDR > 0.05$.

6.4 As learnt in class, for any procedure it is supposed to hold that

$$FDR \leq FWER \leq PFER$$

In this simulation the results we got correspond with the theory.

Task 3

Part A.

Number of bacteria populations of phylum type 'Proteobacteria', found in at least one sample:

```
## [1] 6335
```

Part B.

B.1

The ANOVA model (normal based, linear model) for a given bacteria population K:

$$Y_{ij}^K = \mu_i^K + \varepsilon_{i,j}^K$$
$$Y_{ij}^K \sim N(\mu_i^K, \sigma_K^2),\ i = 1, \cdots, 9\ , j = 1, \cdots, n_i$$

Where each i represents a sample type (environment), and n_i is the number of samples taken from sample type i.

B.2

$$H_0^K : \mu_1^K = \cdots = \mu_9^K$$
$$H_1^K : Otherwise$$

The null hypothesis claims that the expectation of the prevalence of bacteria population K is the same in each environment.

If H_0^K is rejected this means that their prevalence distribution differs in different environments, meaning K is an 'interesting' population.

B.3

The assumptions in ANOVA are that

1. Each ε_{ij}^K comes from a normal distribution
2. The variance (σ_K^2) of prevalence of bacteria population K is the same in each environment (Homogeneity of variances)
3. The data points are independent, so each sample is randomly selected and independent

By the above assumptions we get that each ε_{ij}^K is independent and is distributed $N(0, \sigma^2)$.

Thus each $Y_{ij}^K \sim N(\mu_i^K, \sigma^2)$

B.4

For the 5 bacteria populations that we display, the prevalence in almost each population and for each sample in each environment is zero, meaning the prevalence is many times exactly the same. This is unlikely if we assume the random variables to come from a normal distribution, because the r.v. is continous and the probability that it will get the same values is zero, thus it seems that the normality assumption does not hold.

Table 7: Sample type of each sample

	Type
CL3	Soil
CC1	Soil
SV1	Soil
M31Fcsw	Feces
M11Fcsw	Feces
M31Plmr	Skin
M11Plmr	Skin
F21Plmr	Skin
M31Tong	Tongue
M11Tong	Tongue
LMEpi24M	Freshwater
SLEpi20M	Freshwater
AQC1cm	Freshwater (creek)
AQC4cm	Freshwater (creek)
AQC7cm	Freshwater (creek)
NP2	Ocean
NP3	Ocean
NP5	Ocean
TRRsed1	Sediment (estuary)
TRRsed2	Sediment (estuary)
TRRsed3	Sediment (estuary)
TS28	Feces
TS29	Feces
Even1	Mock
Even2	Mock
Even3	Mock

Table 8: Otu Table for the 5 selected Bacteria Populations

	CL3	CC1	SV1	M31Fcsw	M11Fcsw	M31Plmr	M11Plmr	F21Plmr
249529	0	0	0	0	0	0	0	0
129416	0	0	0	0	0	0	0	0
103231	0	0	0	0	0	0	0	0
327333	0	0	0	0	0	0	0	0
255272	0	0	0	0	0	0	0	0

Table 9: Otu Table for the 5 selected Bacteria Populations

	M31Tong	M11Tong	LMEpi24M	SLEpi20M	AQC1cm	AQC4cm	AQC7cm	NP2	NP3
249529	0	0	0	0	0	0	0	0	0
129416	0	0	0	0	0	1	0	0	0
103231	0	0	0	0	0	1	0	0	0
327333	0	0	0	0	0	0	0	0	0
255272	0	0	0	0	0	0	0	3	2

Table 10: Otu Table for the 5 selected Bacteria Populations

	NP5	TRRsed1	TRRsed2	TRRsed3	TS28	TS29	Even1	Even2	Even3
249529	0	0	1	0	0	0	0	0	0
129416	0	0	1	1	0	0	0	0	0
103231	0	0	0	0	0	0	0	0	0
327333	0	0	0	1	0	0	0	0	0
255272	2	0	0	0	0	0	0	0	0

Part C.

```
pvalues <- vector()
g <- as.vector(GlobalPatterns@sam_data$SampleType) # Sample Types in order as in data
for (row_index in 1:nrow(abundances)) {
  # For each bacteria population perform the kruskal test
  x <- as.vector(abundances[row_index])
  pvalues[row_index] <- kruskal.test(x, g)[["p.value"]]
}
```

Part D.

Number of interesting bacteria populations, meaning number of pvalues that are smaller than 0.05:

```
## [1] 3507
```

The problem is that we did not account for the multiple comparisons!

If we do not account for the multiple comparisons, then the error measures such as FWER and FDR are not controlled.

Usually researchers need to control these measures in order to show that their results are meaningful thus if they are not controlled their results are deemed not reliable.

Not accounting for the multiple comparisons can result in us rejecting more hypotheses just because we performed more comparisons.

So possibly more hypotheses are rejected just due to chance and not because they are false.

In particular this can result in rejecting more true null hypotheses, thus making more type I mistakes.

Part E.

```
# BH from Mission 2
reject.indexes.bh <- BH_Procedure(pvalues, q=0.05)

# Adaptive BH from Mission 2
reject.indexes.adp.bh <- BH_ADP_Procedure(pvalues, q=0.05)

# Hochberg
pvalues.hochberg <- p.adjust(pvalues,method = 'hochberg')
```

We define 'Discovery' by procedure X - If the procedure X determines a population to be interesting. So we think (aka the procedure rejects the null hypothesis of this population) that their prevalence distribution differs in different environments, and 'False Discovery' by procedure X - If the procedure X determines a population as a discovery, but it is not. Meaning that the population's actual prevalence distribution is equal in all different environments, but we thought otherwise.

All three procedures deal with the problem by accounting for the multiple comparisons in someway or another.

1. BH

The BH procedure controls the FDR rate at level 0.05 (if the test statistics are independent/independent and continuous/positive dependent).

Note that if we have no knowledge about the test statistics then the BH procedure controls the FDR rate at level $0.05 * (1 + \frac{1}{2} + \cdots + \frac{1}{m})$

However, the FWER is not controlled at level 0.05 (unless all null hypotheses are true and then FWER=FDR)

In our case it means that the expectancy of the number of false discoveries by BH out of the total number of discoveries by BH is at most 5 percent (FDR).

2. Adaptive BH

The adaptive BH does not control the FDR rate. However, if the estimator for m_0 as defined in the procedure is equal (or close enough) to the actual m_0 then it does control the FDR.

In our case it means that the expectancy of the number of false discoveries by the Adaptive BH out of the total number of discoveries by the Adaptive BH is at most 5 percent (FDR, if the estimator is good enough).

Unfortunately FWER is not controlled at level 0.05.

3. Hochberg

Hochberg's procedure promises us control of the FDR and of the FWER, under the assumption that the test statistics are continuous and independent/positive dependent.

In our case it means that the expectancy of the number of false discoveries by Hochberg out of the total number of discoveries by Hochberg is at most 5 percent (FDR).

In addition, the probability that the number of false discoveries by Hochberg is greater than zero is at most 5 percent (FWER).

Part F.

- Number of discoveries (interesting bacteria populations) by BH procedure:

```
## [1] 1512
```

This is smaller than the number of discoveries we got in part D (3507).

- Number of discoveries (interesting bacteria populations) by the Adaptive BH procedure:

```
## [1] 3552
```

This is greater than the number of discoveries we got in part D (3507).

- Number of discoveries (interesting bacteria populations) by the Hochberg procedure:

```
## [1] 0
```

This is smaller than the number of discoveries we got in part D (3507).

Part G.

It can be predicted that the number of rejected hypotheses by the BH procedure will be less or equal to the number of rejected hypotheses made by the researcher in Part D.

This corresponds with the result we got before.

Proof:

For a given $\alpha = 0.05$, and given $H_i \in \mathfrak{R}^{BH}$ it holds that $P_{(i)}^{ADJ-BH} \leq \alpha$

It holds that

$$(1) \quad \forall i \in \{1, \cdots, m\} : P_{(i)} \leq p_{(i)}^{ADJ-BH}$$

From (1) (proved below the main proof)

$$P_{(i)} \leq P_{(i)}^{ADJ-BH} \leq \alpha \implies P_{(i)} \leq \alpha$$

Therefore, given that α is the same α used by the researcher, it holds that

$$H_i \in \mathfrak{R}^{Researcher}$$

Meaning

$$\mathfrak{R}^{BH} \subseteq \mathfrak{R}^{Researcher}$$

Thus concluding the proof.

Proof of (1):

Define

$$j^* = arg \min_{i \leq j \leq m} \frac{m}{j} p_{(j)}$$

It holds that $\forall i \leq j \leq m : P_{(i)} \leq P_{(j)}$ by def of order statistics.

Therefore as $\frac{m}{j} \geq 0$ it holds that $\frac{m}{j} P_{(i)} \leq \frac{m}{j} P_{(j)}$.

In particular,

$$(2) \quad \frac{m}{j^*} P_{(i)} \leq \frac{m}{j^*} P_{(j^*)}$$

So the adjusted pvalue according to the BH procedure is:

$$p_{(i)}^{ADJ-BH} = \min_{i \leq j \leq m} \frac{m}{j} p_{(j)} = \frac{m}{j^*} \cdot P_{(j^*)} \underbrace{\geq}_{(2)} \frac{m}{j^*} P_{(i)} \underbrace{\geq}_{\frac{m}{j^*} \geq 1} P_{(i)}$$

Note, if $min\{1, \min_{i \leq j \leq m} \frac{m}{j} p_{(j)}\} = 1$ it is trivial that $p_{(i)}^{ADJ-BH} = 1 \geq P_{(i)}$ and the inequality holds as well.

We got that

$$(1) \quad \forall i \in \{1, \cdots, m\} : p_{(i)}^{ADJ-BH} \geq P_{(i)}$$

Part H.

From the previous sections,

$$0 \leq 1512 \leq 3551$$

Meaning that,

$$|\mathfrak{R}^{Hochberg}| \leq |\mathfrak{R}^{BH}| \leq |\mathfrak{R}^{ADP-BH}|$$

This is expected as even without seeing the actual data:

From Task 2, Unit 1 part c and the fact that an adaptive BH procedure at level q is basically the BH procedure at level $\tilde{q}$:

$$\mathfrak{R}^{BH_q} \subseteq \mathfrak{R}^{BH_{\tilde{q}}} = \mathfrak{R}^{ADP-BH_q}$$

And from HW 5 Q3 part D we know that,

$$\mathfrak{R}^{Hochberg} \subseteq \mathfrak{R}^{BH}$$

So in total,

$$\mathfrak{R}^{Hochberg} \subseteq \mathfrak{R}^{BH} \subseteq \mathfrak{R}^{ADP-BH}$$

$$\Rightarrow |\mathfrak{R}^{Hochberg}| \leq |\mathfrak{R}^{BH}| \leq |\mathfrak{R}^{ADP-BH}|$$

Part I.

I.1

Case 1.

The researcher is interested in constructing Confidence Intervals (CIs) for the difference in expectation for interesting populations from the soil and fresh water (creek) environments, and that the False coverage rate (FCR) be controlled at level $q = 0.05$.

Where FCR is the expectation of proportion of V out of R, using the same notation as before.

In order to control the FCR, according to the theorem as learnt in class, the CIs should be constructed at level $1 - \frac{|S(T)|}{m}q$.

Where $|S(T)|$ is the number of interesting populations and m is the number of hypotheses (bacteria populations).

Note that the theorem assumptions are that the decision rule $S(T)$ is simple, each marginal CI is monotone in the confidence level and the components of T are independent, where T is the vector of estimators.

In our case, each CI should be constructed at level:

```
m <- nrow(abundances)
q <- 0.05
s.T <- num.rejects.bh

1- q*s.T/m
```

```
## [1] 0.9880663
```

Case 2.

The researcher is interested in constructing CIs, and that $P(\exists i \in S : \Theta_i \notin CI_i) = P(V_{CI}(S) > 0) \leq q = 0.05$

Where S is any subgroup of $\{1, \cdots, m\}$, and $V_{CI}(S)$ is the number of CIs that the parameter is not inside them, out of the constructed CIs (S).

To get the desired results, the CIs should be simultaneous at level $1 - q$, meaning each CI should be constructed at level $1 - \frac{q}{m}$ (by bonferroni, as we proved in HM 2 question 2).

As learnt in class this would provide control of $P(V_{CI}(S) > 0)$ at level q.

In our case, each CI should be constructed at level:

```
1-q/m
```

```
## [1] 0.9999921
```

I.2

As the confidence level 1- α increases the CI is getting wider as learnt in class.

In the previous section we got that the CIs of case 2 should be constructed at a level greater than that of the CIs of case 1.

Therefore, in case 2 we would get wider CIs than in Case 1, for the same populations.

This result is intuitive because we saw in class that $P(V_{CI}(S) > 0)$ (which is the required assurance in case 2) is a more conservative measure than FCR (which is the required assurance in case 1). Therefore, the CIs in case 2 will be wider (more conservative) than the CIs in case 1, as they control a more conservative measure.