List 5

Kamil Zaborniak

Contents

Problem definition

{ Let $\underline{X}_i = (x_{i,1}, \ldots, x_{i,j})$, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$, be sample of i.i.d. random variables from the exponential distribution with the expected value $\mu_i = \mathbb{E}(x_{i,j})$. We know that the parameter of this distribution is $\lambda_i = \frac{1}{\mu_i}$. We will construct the most powerful test on significance level α, using Carlin-Rubin Theorem for the problem:

$$\begin{aligned} H_{0,i} :&\lambda_i = \frac{1}{3}, \\ H_{1,i} :&\lambda_i =< \frac{1}{3}. \end{aligned}$$

The likelihood ratio is:

$$\begin{aligned} L(\underline{X}_i) &= \frac{\prod_{j=1}^m f(x_{i,j}, \lambda_1)}{\prod_{j=1}^m f(x_{i,j}, \frac{1}{3})} \\ &= \frac{\prod_{j=1}^m \lambda_1 \exp(-\lambda_1 x_{i,j})}{\prod_{j=1}^m \frac{1}{3} \exp(-\frac{1}{3} x_{i,j})} \\ &= (3\lambda_1)^m \exp\left(\left(\frac{1}{3} - \lambda_1\right) \sum_{j=1}^m x_{i,j} \right) \end{aligned}$$

Let $T(X_i) = \sum_{j=1}^m$. We know that $\frac{1}{3} - \lambda_1$ is non-decreasing function, so the test function has form:

$$\gamma(\underline{X}_i) = \begin{cases} 1, & T(\underline{X}_i) > c \\ \xi, & T(\underline{X}_i) = c \\ 0, & T(X_i) < c \end{cases},$$

where

$$\mathbb{E}_{H_{0,i}}[\gamma(\underline{X}_i)] = \alpha.$$

Under $H_{0,i}$ the test statistic $T(\underline{X}_i) \sim Gamma\left(m, \frac{1}{3}\right)$, so we reject $H_{0,i}$ when

$$T(\underline{X}_i) > F^{-1}_{Gamma(m,1/3)}(1 - \alpha).$$

P-value of this test is equal to:

$$p_i = 1 - F_{Gamma(m,1/3)}(T(\underline{X}_i)).$$

}

Task 2

Lets consider mixture model with:

$$P(\mu_i = 3) - 1 - \varepsilon = 1 - P(\mu_i = 5.5),$$

for $i \in \{1, \ldots, n\}$, $n \in \{200, 1000\}$, $m \in \{20, 100\}$, $\varepsilon \in \{0.01, 0.05, 0.1, 0.2$ and significance level $q \in \{0.1, 0.1\sqrt{\frac{m}{200}}\}$. For each case, let's perform a test using the BFDR controlling procedure, Bonferroni procedure and B-H procedure.

Simulation

for each case, let's carry out a simulation consisting of repeating iterations 1000 times. A single iteration involves drawing a Bernoulli sample of size m and a probability of success equal to ε, which will show us the real alternatives. Based on the obtained vector, we generate n trials of size m according to the vector of alternatives. Based on the obtained samples, we calculate the corresponding test statistics and p-values and carry out testing using each procedure. Based on the tests performed, we calculate the False Discovery Rate, True Discovery Rate and each cost. After 1000 iterations, we obtain estimators. The BFDR estimator is the average of all FDPs, the power estimator is the average number of TDRs. We calculate each cost as: $c_0\mathbb{E}[V] + c_A\mathbb{E}[T]$, where $\mathbb{E}[V]$ is the expected number of false discoveries and $\mathbb{E}[T]$ is the expected number of non-discoveries.

Simulation Result

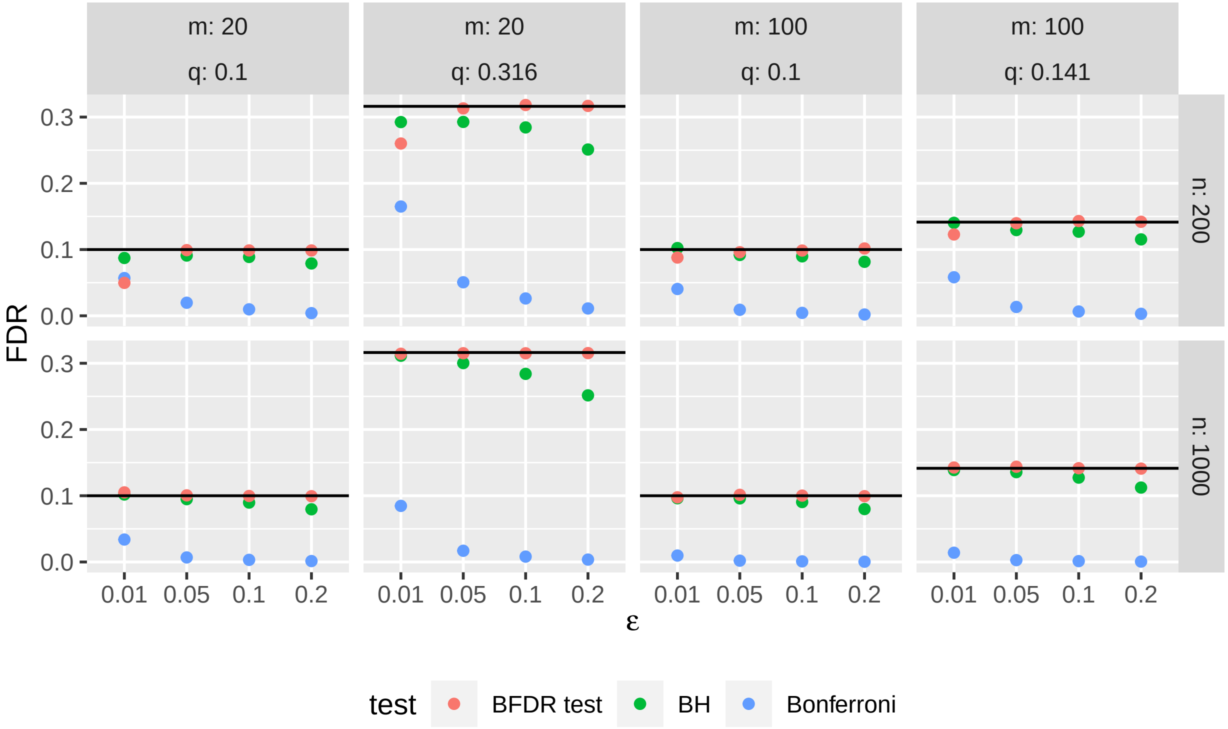


Figure 1: BFDR of each procedure depending on ϵ, number of tests(n), size sample(m) and q.

We see that each procedure controls the BFDR in each case. The smallest BFDR is achieved by the Bonferroni procedure, and the largest by the BFDR controlling procedure.

The Bonferroni procedure and the BH procedure achieve lower results as the value of ε increases. The BFDR controlling procedure tends to cause BFDR to increase as ε increases (for $m = 20$ and $n = 200$), and in other cases its BFDR values are close to q.

The graph also shows that for the Bonferroni and Benjamini-Hochberg procedures, the BFDR decreases as the sample size increases and the number of tests increases.

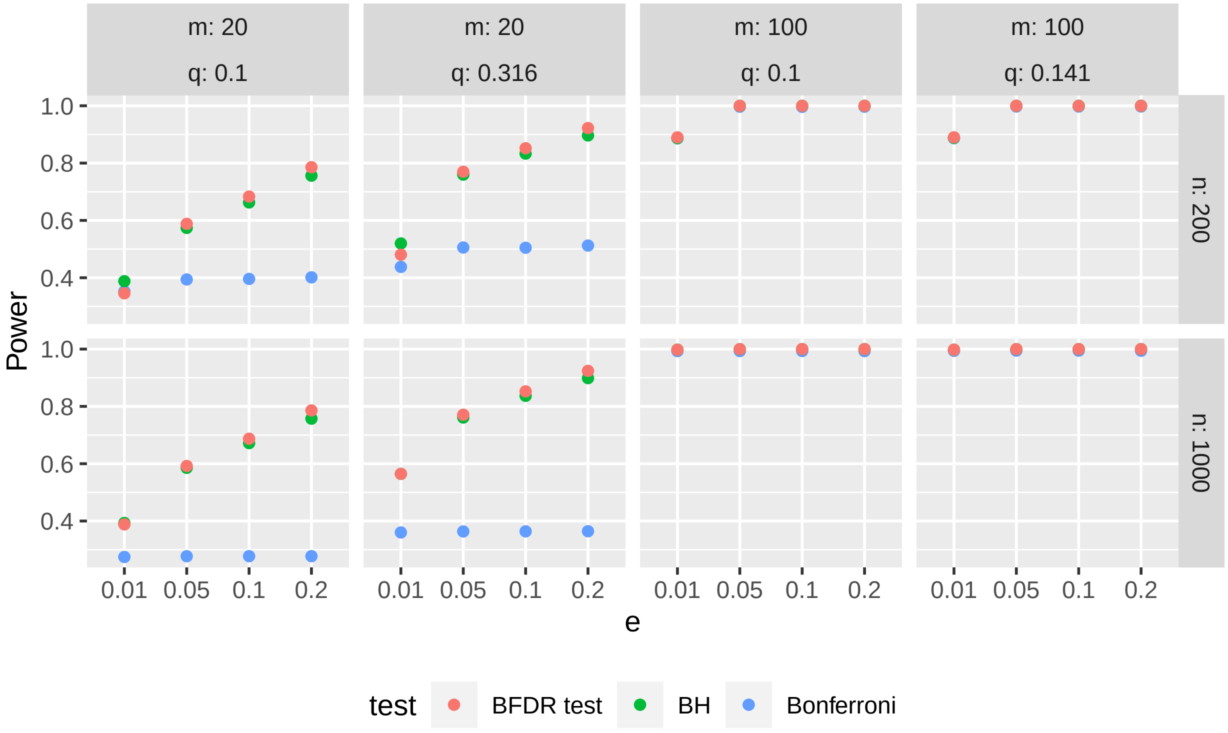


Figure 2: Power of each procedure depending on ϵ, number of tests(n), size sample(m) and q.

The BFDR controlling procedure has the best power in almost every case, the BH procedure has a result close to it, and the Bonferroni procedure has the least power. We see that procedure powers increase as ε increases. We also see that the power of procedures increases as m or n increases.

We see that for $m = 100$ and $n = 200$ the powers are equal to 1 in almost all cases, and for $n = 100$ and $m = 100$ the powers are always equal to 1.

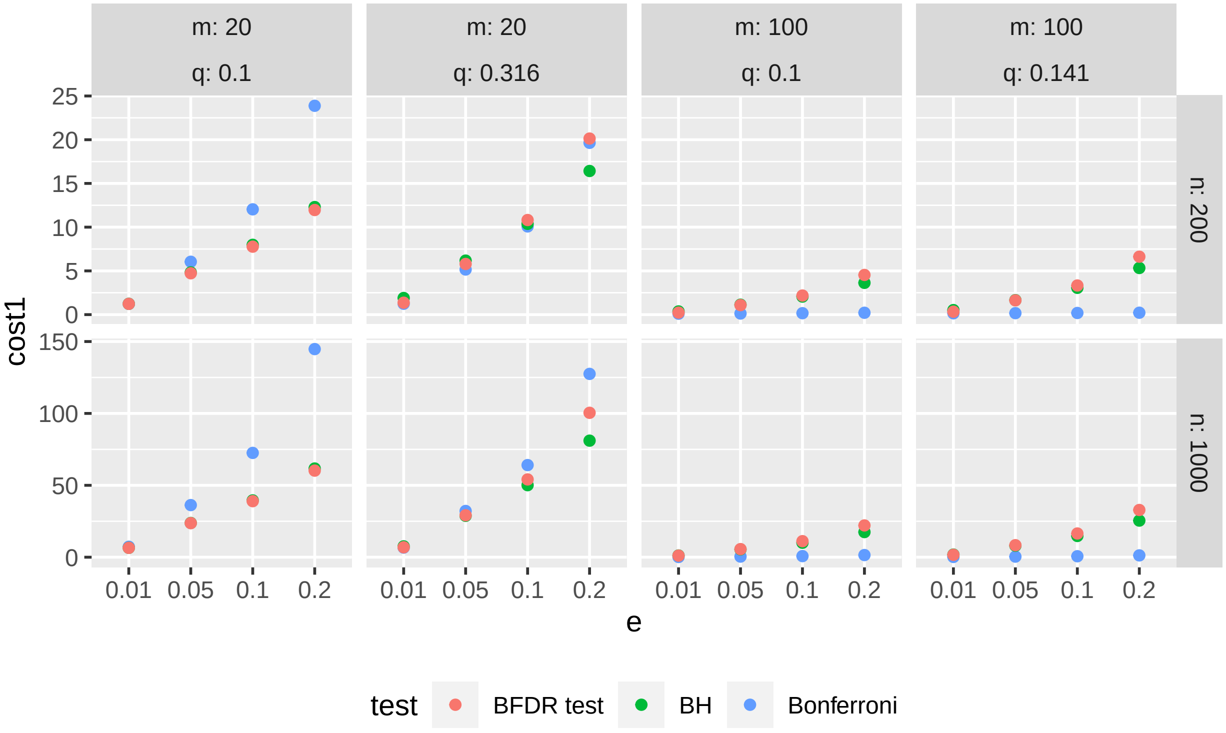


Figure 3: Cost of each procedure for $c_0 = c_A = 1$ depending on ϵ, number of tests(n), size sample(m) and q.

We know that the expected cost depends on the probability of making a type I error and the probability of making a type II error, and therefore depends on the number of false discoveries and false non-discoveries, which depend on the number of tests and ε.

We see that a given procedure takes on a higher cost the less power it has or the less it controls the BFDR. For example, for the case where $m = 20$ and $q = 0.1$. Bonferroni controls BFDR best, but its power is much much lower than that of the other procedures, so its costs are the highest. In the case where $m = 100$ procedures usually have a power equal to 1, therefore in these cases the procedure incurs a higher cost because it controls the BFDR less strongly.

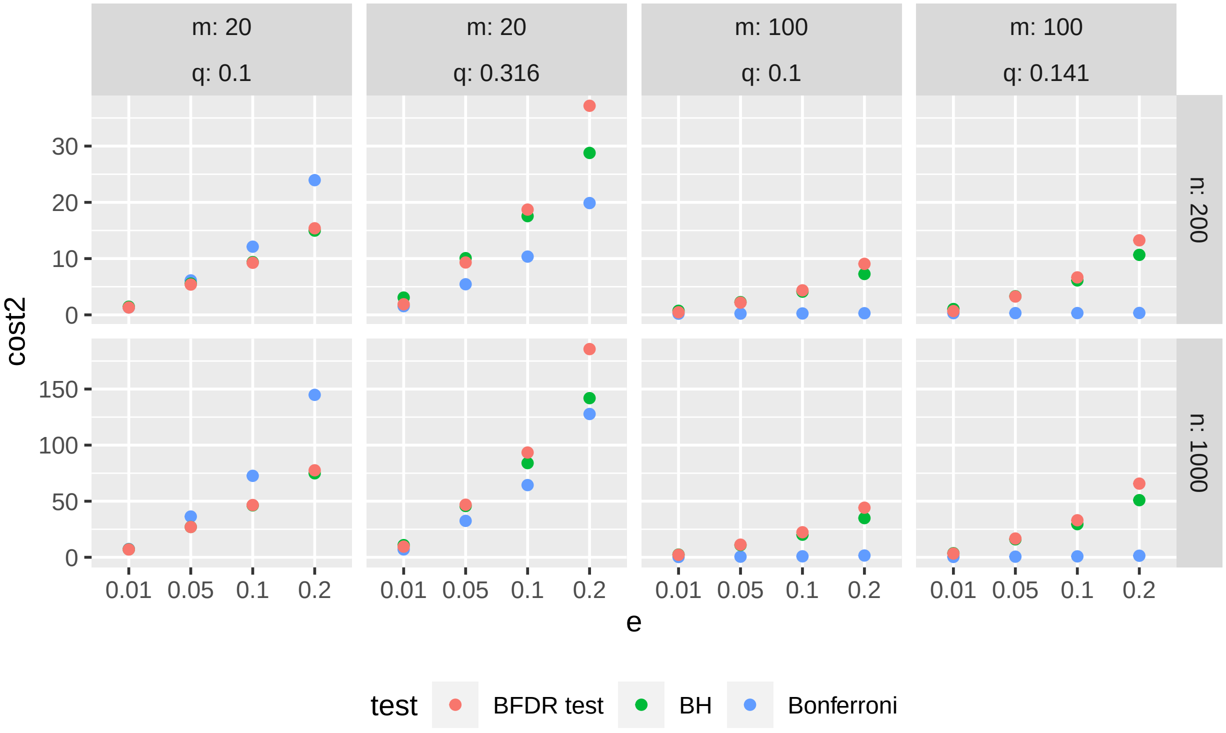


Figure 4: Cost of each procedure for $c_0 = 2$, $c_A = 1$ depending on ϵ, number of tests(n), size sample(m) and q.

For $c_0 = 2$ and $c_A = 1$ we see that the cost values for the given procedures depend more on BFDR than on power. We see that Bonferroni has the lowest costs among the procedures (except for the cases in the first column of the plot), and the BFDR controlling procedure has the highest costs.

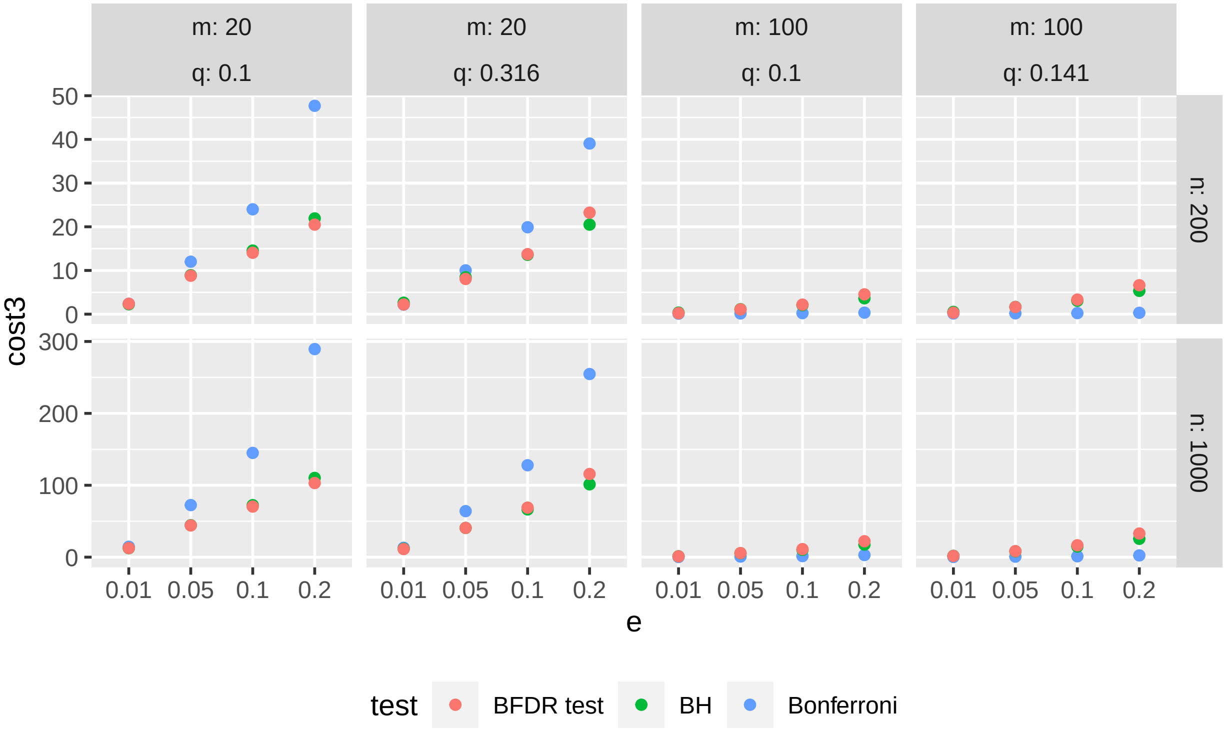


Figure 5: Cost of each procedure for $c_0 = 1$, $c_A = 2$ depending on ϵ, number of tests(n), size sample(m) and q.

The third graph shows that the case in which the cost is influenced more by power than by BFDR, so when $m = 20$, the controlling BFDR procedure achieves the lowest costs, and the Bonferroni procedure achieves the worst. In the case where $m = 100$ I see that the BFDR controlling procedure has the highest cost, because for these cases the procedures usually have a power of 1, but the BFDR controlling procedure achieves the highest BFDR.

Bayesian Classifier (task 3)

We reject $H_{0,i}$ when:

$$\frac{f_{H_{1,i}}}{f_{H_{0,i}}} \geq \frac{c_0(1-\varepsilon)}{c_A\varepsilon}$$
$$\Updownarrow$$
$$T(X_i) \geq \frac{33}{5}\log\left(\frac{11c_0(1-\varepsilon)}{6c_a\varepsilon}\right) = \tau.$$

The probability of the type I error is equal to:

$$t_1 = P_0(T(X_i) \geq \tau) = 1 - F_{Gamma(m,1/3)}(\tau).$$

The power is equal to:

$$power = P_1(T(X_i) \geq \tau) = 1 - F_{Gamma(m,1/5.5)}(\tau) = 1 - t_2,$$

Where t_2 is the probability of the type II error.

The expected cost function is:

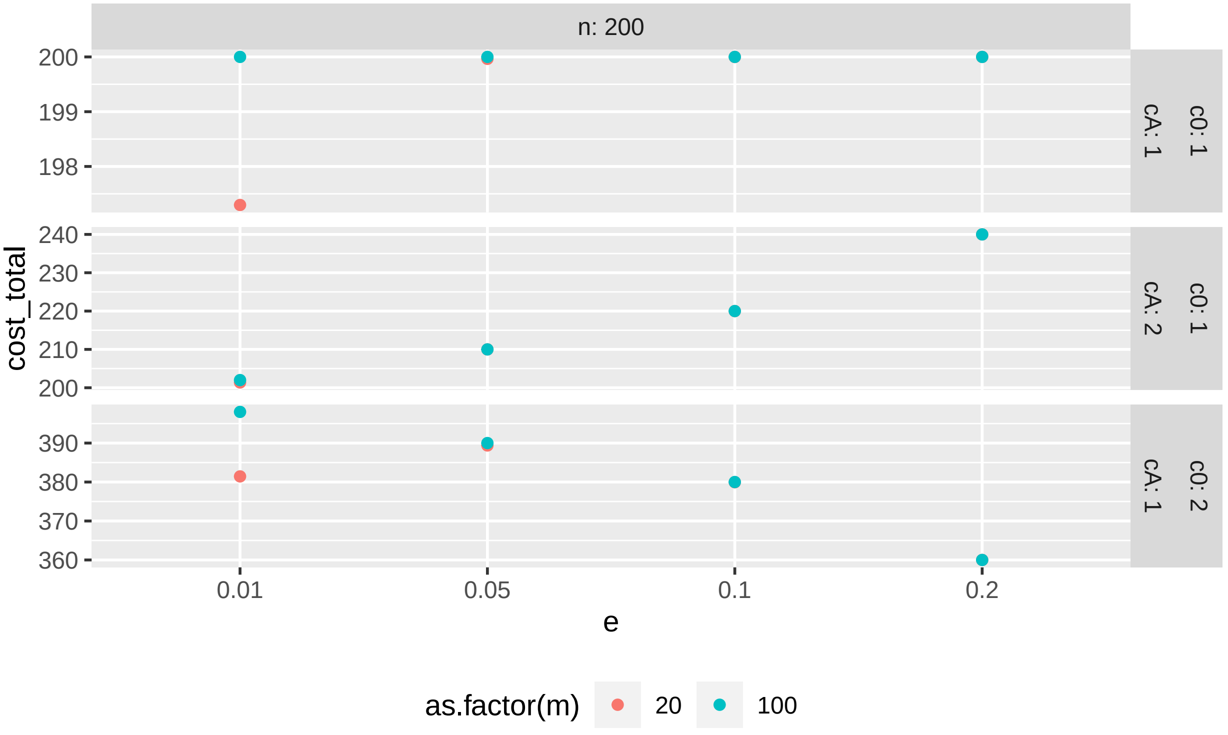
$$\mathbb{E}[C] = n((1-\varepsilon)t_1 c_0 + \varepsilon t_2 c_A).$$

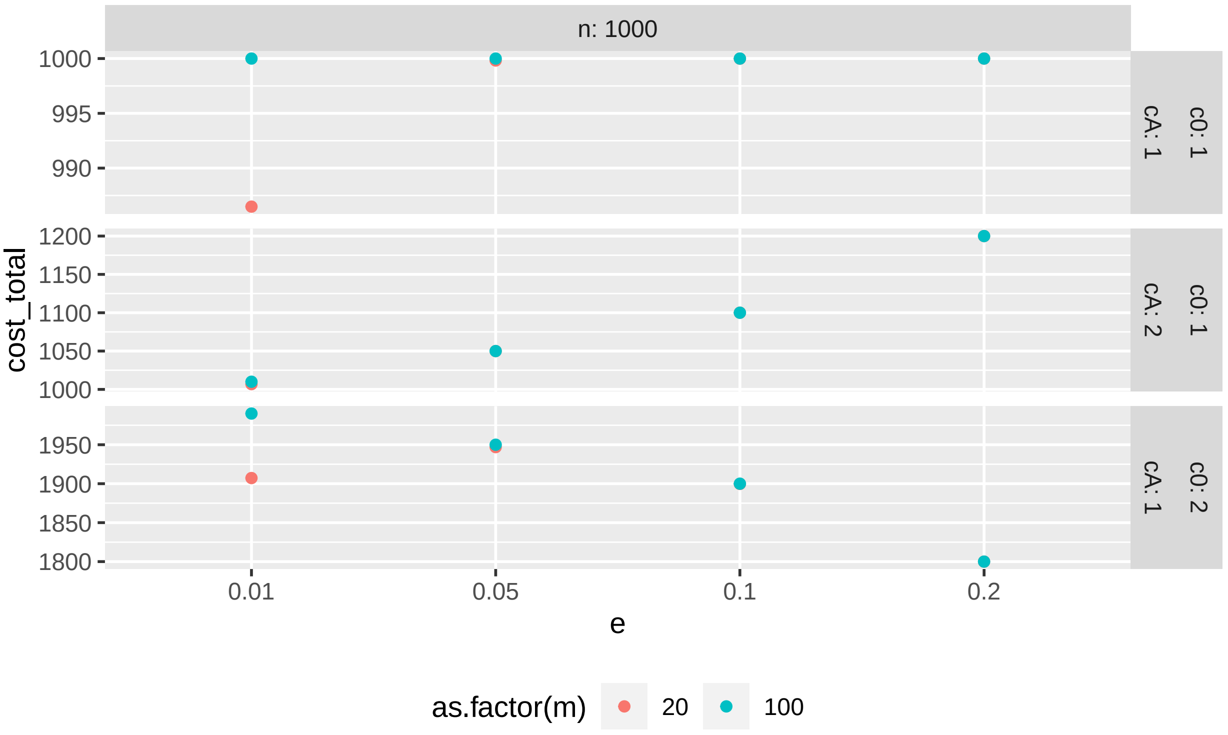
The theoretically BFDR is equal to:

$$BFDR = \frac{(1-\varepsilon)t_1}{(1-\varepsilon)t_1 + \varepsilon(1-t_2)}.$$

Table 1: Table of theoretical values.

m	$epsilon$	c_0	c_A	$ au$	BFDR	Power	Single cost
20	0.01	1	1	34.328	0.980	1	0.986
20	0.01	2	1	38.903	0.979	1	1.907
20	0.01	1	2	29.754	0.980	1	1.007
20	0.05	1	1	23.434	0.905	1	1.000
20	0.05	2	1	28.009	0.905	1	1.947
20	0.05	1	2	18.859	0.905	1	1.050
20	0.10	1	1	18.502	0.818	1	1.000
20	0.10	2	1	23.077	0.818	1	1.900
20	0.10	1	2	13.927	0.818	1	1.100
20	0.20	1	1	13.150	0.667	1	1.000
20	0.20	2	1	17.725	0.667	1	1.800
20	0.20	1	2	8.575	0.667	1	1.200
100	0.01	1	1	34.328	0.980	1	1.000
100	0.01	2	1	38.903	0.980	1	1.990
100	0.01	1	2	29.754	0.980	1	1.010
100	0.05	1	1	23.434	0.905	1	1.000
100	0.05	2	1	28.009	0.905	1	1.950
100	0.05	1	2	18.859	0.905	1	1.050
100	0.10	1	1	18.502	0.818	1	1.000
100	0.10	2	1	23.077	0.818	1	1.900
100	0.10	1	2	13.927	0.818	1	1.100
100	0.20	1	1	13.150	0.667	1	1.000
100	0.20	2	1	17.725	0.667	1	1.800
100	0.20	1	2	8.575	0.667	1	1.200





Task 4

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

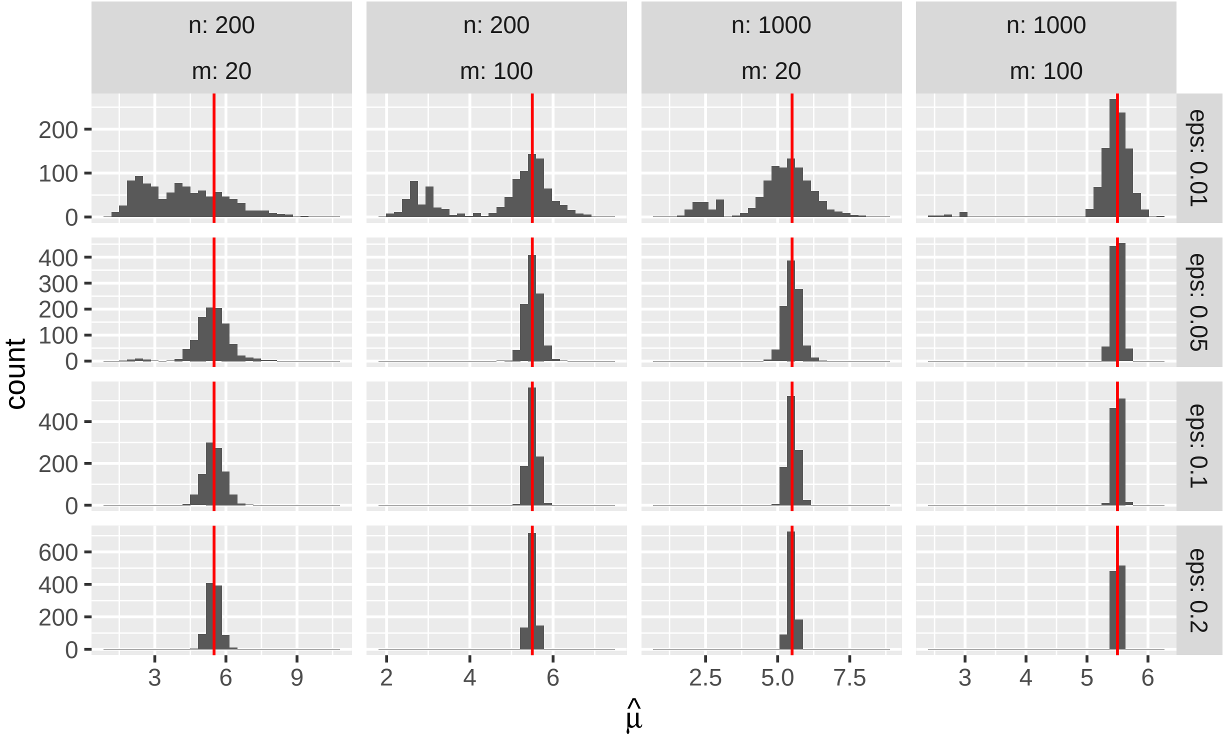


Figure 6: Histograms od estimators.

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

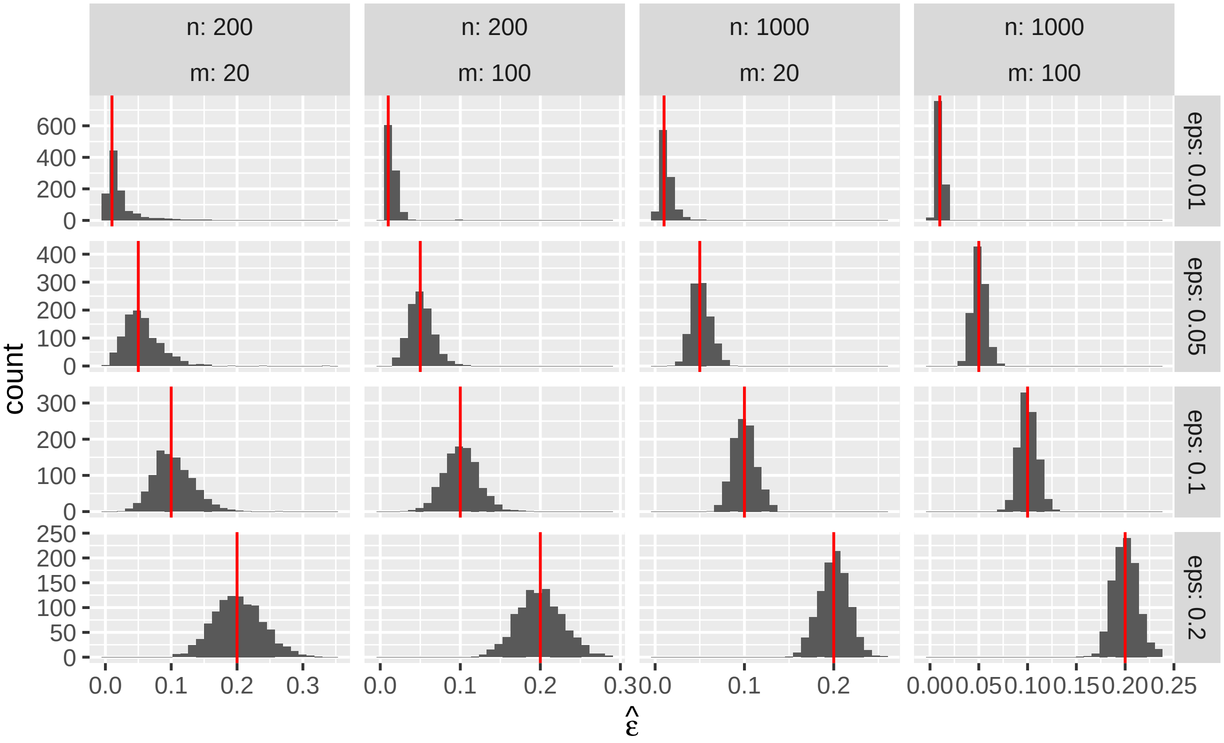


Figure 7: Histograms od estimators.

n	m	eps	Var_mu_est	Bias_mu_est	MSE_mu_est	Var_eps_est	Bias_eps_est	MSE_eps_est
200	20	0.01	2.9063673	-1.3410009	4.7046506	0.0011073	0.0151071	0.0013355
200	20	0.05	0.6325010	-0.0986437	0.6422316	0.0008393	0.0063049	0.0008790
200	20	0.10	0.1837321	0.0053830	0.1837611	0.0008969	0.0038201	0.0009115
200	20	0.20	0.0694738	-0.0096380	0.0695667	0.0013665	0.0016880	0.0013693
200	100	0.01	1.6443816	-0.7682805	2.2346365	0.0003395	0.0043706	0.0003586
200	100	0.05	0.0324771	0.0081715	0.0325438	0.0002326	-0.0007016	0.0002330
200	100	0.10	0.0150140	0.0058388	0.0150481	0.0004860	0.0002027	0.0004861
200	100	0.20	0.0082803	-0.0012896	0.0082820	0.0008278	0.0002231	0.0008278
1000	20	0.01	1.6221944	-0.4714898	1.8444969	0.0000605	0.0029200	0.0000690
1000	20	0.05	0.0760681	0.0029500	0.0760768	0.0001161	0.0013306	0.0001178
1000	20	0.10	0.0335260	-0.0003738	0.0335262	0.0001651	-0.0001402	0.0001651
1000	20	0.20	0.0143820	-0.0025357	0.0143884	0.0002712	0.0001022	0.0002712
1000	100	0.01	0.1917672	-0.0678604	0.1963722	0.0000099	0.0002185	0.0000100
1000	100	0.05	0.0062370	-0.0018456	0.0062405	0.0000487	0.0004139	0.0000488
1000	100	0.10	0.0030047	0.0012865	0.0030064	0.0000884	0.0004312	0.0000886
1000	100	0.20	0.0016654	-0.0000539	0.0016654	0.0001585	0.0003816	0.0001587

We see that the EM algorithm estimates the values of ε and μ better as the parameters n, m, ε increase.