Problem 1

First, we generate the data:

- matrix X of size 100×950 which elements are iid random variables from a distribution $N(0, 10^{-3/2})$,
- vector β with 950 elements such that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 3$ and the rest are zeros,
- response vector $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, 1)$.

We then proceed to analysing linear models using $10, 100, 500$ and 950 columns of X.

We generate least squares estimators of coefficients vector β and analyze the results:

Number of columns	Mean of significant coefficients	Variance of signifcant coefficients
10	2.67	1.23
100	2.86	1.29
500	3.16	1.82
950	2.67	5.47

Number of columns	Mean of insignificant coefficients	Variance of insignifcant coefficients
10	0.12	1.30
100	0.12	1.00
500	0.06	1.40
950	-0.01	4.63

In the case of signifcant coefficients, we can clearly see that the more columns we apply to the model, the bigger is the variance of the coefficients. In the case of insignifcant coefficients, variance gets smaller at first (between 10 and 100 columns) but starts to grow afterwards reaching a peak at 950 columns.

Means of both significant and insignifcant coefficients seem to oscilate around true values (respecitvely 3 and 0).

We perform t-tests for significance of individual regression coefficients at $\alpha = 0.1$.

Number of columns	Number of discoveries
10	4
100	11
500	45
950	105

When the number of used columns grow, the more discoveries we make. Since only five coefficients are significant, the extra discoveries made as the number of columns grows are mostly false discoveries.

We proceed to calculating average standard deviation of the estimators of regression coefficients and average length of the respective 90% confidence intervals.

Number of columns	Average standard deviation	Average confidence interal length
10	0.90	2.99
100	1.03	3.42
500	1.39	4.60
950	4.36	14.63

Similarly, the more columns we apply, the bigger are the values.

Last part of the problem 1 is comparing different methods of calculating true and false discoveries:

Number of columns	Method	True discoveries	False discoveries
10	Without adjusting	4	0
	Bonferroni	1	0
	Benjamini-Hochberg	3	0
100	Without adjusting	5	6
	Bonferroni	1	0
	Benjamini-Hochberg	1	0
500	Without adjusting	2	43
	Bonferroni	0	0
	Benjamini-Hochberg	0	0
950	Without adjusting	1	104
	Bonferroni	0	0
	Benjamini-Hochberg	0	0

We can observe that adjusting for multiple testing decreases the number of false discoveries but in the case of high number of columns, it fails to recognize significant coefficients.

Problem 2

We repeat the procedure of generating data and creating models 500 times.

In the table below, we compare the average variance of the estimaros of individual regression coefficients with the theoretical value derived from inverse Wishart distribution.

Number of columns	Average variance	Theoretical value
10	0.91	1.01
100	1.09	1.11
500	1.99	2.00
950	20.38	20.40

As the number of columns grows, the bigger the average variance gets. We can also see that average variance calculated from the experiments gets closer to the theoretical value.

We will now compare the average length of the 90% confidence interval with its theoretical estimate.

Number of columns	Average CI length	Theoretical estimate
10	3.01	3.31
100	3.43	3.47
500	4.65	4.66
950	14.97	15.14

Values obtained from the experiments are close to the theoretical estimates. The grow as the number of columns get larger.

In the next step, we will compare number of false discoveries, true discoveries, FWER and FDR for the procedures from the Problem 1.

Number of columns	Method	Mean number of TD	Mean number of FD	FWER	FDR
10	Without adjusting	4.622	0.582	0.456	0.097
	Bonferroni	0.988	0.004	0.004	0.004
	Benjamini-Hochberg	4.148	0.306	0.254	0.054
100	Without adjusting	4.506	9.75	1	0.668
	Bonferroni	0.856	0.006	0.006	0.006
	Benjamini-Hochberg	2.206	0.298	0.214	0.065
500	Without adjusting	3.458	49.054	1	0.932
	Bonferroni	0.256	0.066	0.066	0.066
	Benjamini-Hochberg	0.336	0.142	0.116	0.078
950	Without adjusting	0.892	93.694	1	0.989
	Bonferroni	0	0.05	0.05	0.05
	Benjamini-Hochberg	0.004	0.362	0.05	0.049

We can see that Bonferroni procedure controls FWER at the level $\alpha = 0.1$.

Lastly, we calculate the theoretical values of FWER.

Number of columns	Theoretical bound of FWER for no adjusting	Theoretical bound of FWER for Bonferroni
10	0.409	0.05
100	0.999	0.095
500	1	0.099
950	1	0.099

We can see that the theoretical bounds hold when analyzing the data from the experiments.