Statistical learning
Assignment 2

Jakub Skalski

June 1, 2024

1. Problem 1

1.1 Define chi-squared and F distributions

Let $Z_i \sim \mathcal{N}(0,1)$. A random variable χ^2_p follows chi-squared distribution if $\chi^2_p = \sum_{i=1}^p Z_i^2$. Similarly, $F_{p,n}$ has an F distribution when $F_{p,n} = \frac{U^2/p}{V^2/n}$, where $U^2 \sim \chi^2_p$ and $V^2 \sim \chi^2_n$.

1.2 What distribution can $F_{p,n}$ be approximated by for large n values?

Let $F_{p,n} = \frac{U^2/p}{V^2/n}$, where $V^2 = \sum_i^n Y_i$ and $Y_i = X_i^2 \sim \chi^2_1$. By the strong law of large numbers $\frac{V^2}{n} = \frac{\sum_i^n Y_i}{n} \xrightarrow[n\to\inf]{D} n\frac{E[Y_1]}{n} = E[X^2] = V[X] + E^2[X] = 1$. Finally, applying Slutsky's theorem we obtain $F_{p,n} = \frac{U^2/p}{V^2/n} \xrightarrow[n\to\inf]{D} U^2/p$.

1.3 Proof of χ^2 distribution

Let $X_i \sim \mathcal{N}_p(\mu, \Sigma)$ implying $\bar{X} \sim \mathcal{N}_p(\mu, \frac{1}{n}\Sigma)$. Assuming Σ is symmetric and its decomposition exists we can rearrange the terms.

$$n(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu) = n(\bar{X}-\mu)^T\sqrt{\Sigma}^{-T}\sqrt{\Sigma}^{-1}(\bar{X}-\mu) = n[\sqrt{\Sigma}^{-1}(\bar{X}-\mu)]^T[\sqrt{\Sigma}^{-1}(\bar{X}-\mu)]$$

Let $Y := \sqrt{n}\sqrt{\Sigma}^{-1}(\bar{X}-\mu)$. Now, the equation simplifies to just a product of two normal random vectors.

$$n[\sqrt{\Sigma}^{-1}(\bar{X}-\mu)]^T[\sqrt{\Sigma}^{-1}(\bar{X}-\mu)] = Y^TY$$

But $Y \sim \mathcal{N}_p(0, \frac{\sqrt{n}^2}{n}\sqrt{\Sigma}^{-1}\Sigma\sqrt{\Sigma}^{-1}) = \mathcal{N}_p(0, I)$ and therefore $Y^TY \sim \chi^2_p$.

1.4 Properties of Hotelling T^2 statistic

Let $X_i \sim \mathcal{N}_p(\mu, \Sigma)$ and S_n is the sample covariance. We define Hotelling statistic as follows:

$$T^2 = n(\bar{X} - \mu_0)^T S_n^{-1} (\bar{X} - \mu_0)$$

We say that a random variable follows the Hotelling distribution $T^2(p, m)$ if it is of the form $md^T M^{-1} d$, where $d \sim \mathcal{N}_p(0, \Sigma)$ and $M \sim W_p(\Sigma, m)$ (Wishart distribution with m degrees of freedom). Let us show that $n(\bar{X} - \mu_0)^T S_n^{-1} (\bar{X} - \mu_0) \sim T^2(p, m)$.

Suppose $d = \sqrt{n}(\bar{X} - \mu_0)$. Clearly, $\sqrt{n}(\bar{X} - \mu_0) \sim \mathcal{N}(0, \Sigma)$.

$$T^2 = n(\bar{X} - \mu_0)^T S_n^{-1} (\bar{X} - \mu_0) = d^T S_n^{-1} d$$

We know that $(n-1)S_n \sim W(\Sigma, n-1)$ and therefore $M = (n-1)S_n$, while $m = n-1$.

$$T^2 = \frac{n-1}{n-1} d^T S_n^{-1} d = (n-1) d^T [(n-1)S_n]^{-1} d = md^T M^{-1} d$$

We can now use that fact to devise an α-level test, rejecting for large values of T^2.

$$T^2 \geq \frac{(n-1)p}{n-p} F_{p,n-p}(1-\alpha)$$

From 1.2 we conjecture that $\frac{(n-1)p}{n-p} F_{p,n-p} \xrightarrow[n \to \inf]{D} \frac{n-1}{n-p} \chi_p^2$. Thus, the probability of type I error for large n is $P(\frac{n-p}{n-1} T^2 \leq \chi_p^2(\alpha)) = P(X \leq \chi_p^2(\alpha)) = \Phi(\chi_p^2(\alpha)) = \Phi(\Phi^{-1}(\alpha)) = \alpha$, where Φ is the cumulative distribution function of the underlying distribution. When alternative hypothesis is true, X follows a noncentral chi-squared distribution and the power of the test tends to 1.0 with the increase in sample size.

2. Problem 2

Let $X = (1.7, 1.6, 3.3, 2.7, -0.04, 0.35, -0.5, 1.0, 0.7, 0.8)$ be a vector sampled from $\mathcal{N}(\mu, \Sigma)$. We consider a $H_{0,i} : \mu_i = 0$ testing problem at $\alpha = 0.05$.

2.1 Bonferroni testing procedure

First, compute the rejection boundary for $\Phi^{-1}(1 - \frac{\alpha}{2p}) = \Phi^{-1}(1 - 0.0025) = c \approx 2.8$. Then, reject all $H_{0,i}$ where $|X_i| > c$. For this case, only $H_{0,3}$ is to be rejected.

2.2 Benjamini-Hochber testing procedure

First, sort the absolute values of X and compute $c_i = \Phi^{-1}(1 - \frac{\alpha i}{2p})$ for each.

(i)	1	2	3	4	5	6	7	8	9	10
$\lvert x\rvert_{\geq}$	3.3	2.7	1.7	1.6	1.0	0.8	0.7	0.5	0.35	0.04
c_i	2.8	2.5	2.5	2.8	2.3	2.2	2.1	2.0	2.0	1.9

Largest index satisfying $|X|_{(i)} \geq \Phi^{-1}(1 - \frac{\alpha i}{2p})$ is i_{SU}=2. Reject local hypotheses for which their sorted index is at most equal to i_{SU}, which in this case would be $H_{0,3}$ and $H_{0,4}$.

2.3 False discovery proportion

Assuming that only the first three coordinates of μ are different form zero the FDP is 0 and 0.5 for Bonferroni and BH respectively.

3. Project 2

3.1 Examine the underlying data distribution

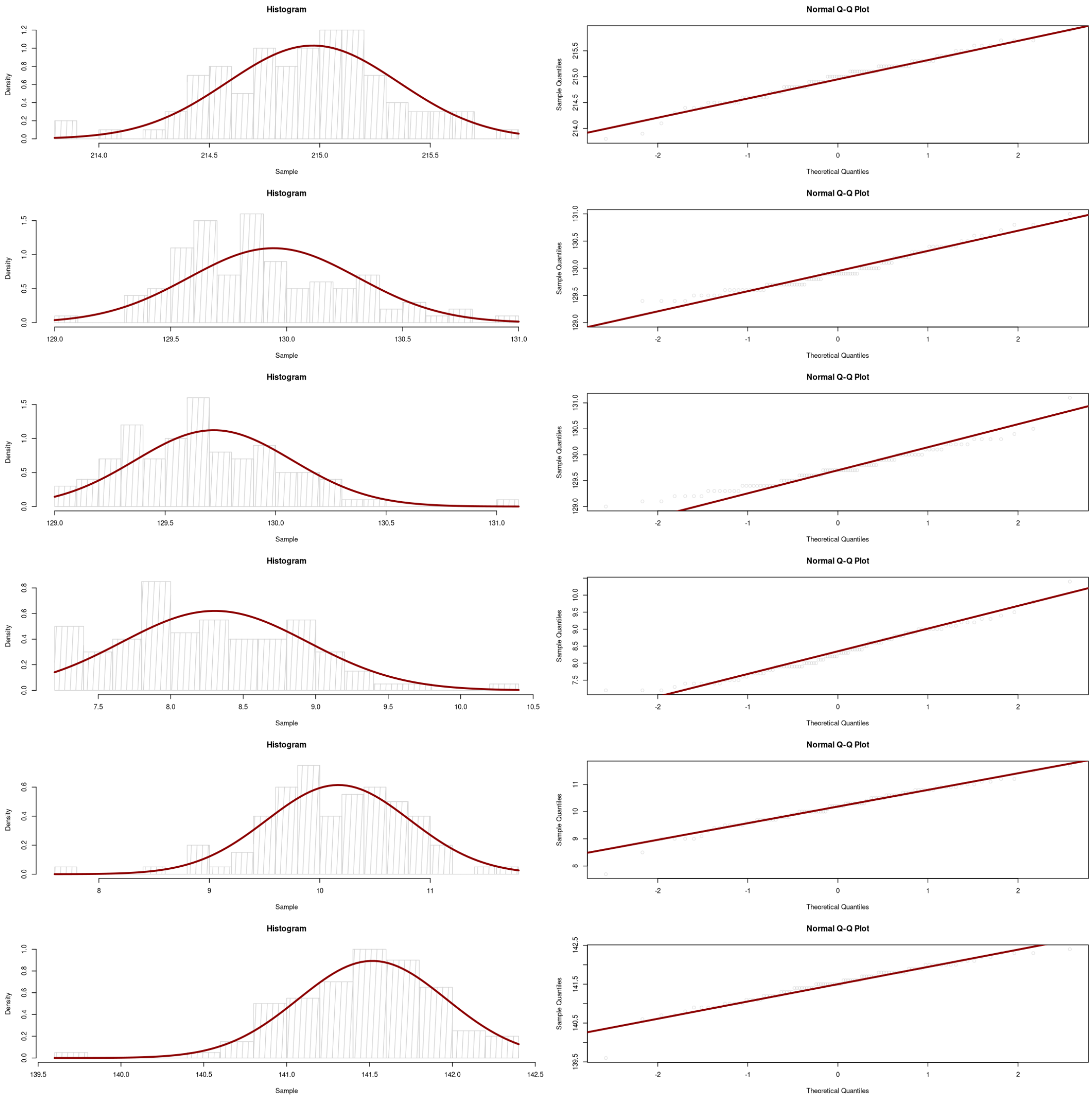

Figure 1: Histograms and qq-plots for each column of the data

Admittedly, there is a slight skew to the data, but overall there should be no problems modelling it with a normal distribution.

3.2 Estimated means and covariance

LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
1214.969	129.943	129.72	8.305	10.168	141.517

Table 1: Vector of means

LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
0.150241414	0.05801313	0.05729293	0.0571262626	0.01445253	0.0054818182
0.058013131	0.13257677	0.08589899	0.0566515152	0.04906667	-0.0430616162
0.057292929	0.08589899	0.12626263	0.0581818182	0.03064646	-0.0237777778
0.057126263	0.05665152	0.05818182	0.4132070707	-0.26347475	-0.0001868687
0.014452525	0.04906667	0.03064646	-0.2634747475	0.42118788	-0.0753090909
0.005481818	-0.04306162	-0.02377778	-0.0001868687	-0.07530909	0.1998090909

Table 2: Covariance matrix

3.3 Bank notes testing

A new production line that will be replacing the old one for printing the bank notes is tested and one of the requirements is that the average dimensions of the bank notes are comparable to these represented in the provided sample of the original bank notes. After printing a very long series of bank notes in the new production line, it was found that the mean values of the dimensions are as follows.

LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
214.97	130	129.67	8.3	10.16	141.52

Table 3: Bank notes means

The given estimate does not lie inside the Hotelling's ellipsoid, but fits inside the Bonferroni's rectangular region. We can examine this graphically by plotting the confidence intervals for some subsets of data where it can be observed that the new mean satisfies all line intervals but fails to fit inside of some of the elliptical confidence regions. This happens, because the confidence regions shrink in higher dimensions due to the correction methods used in order to maintain control over FWER and keep it constant. As it turns out the LEFT and RIGHT coordinates are somewhat off and require adjustment.

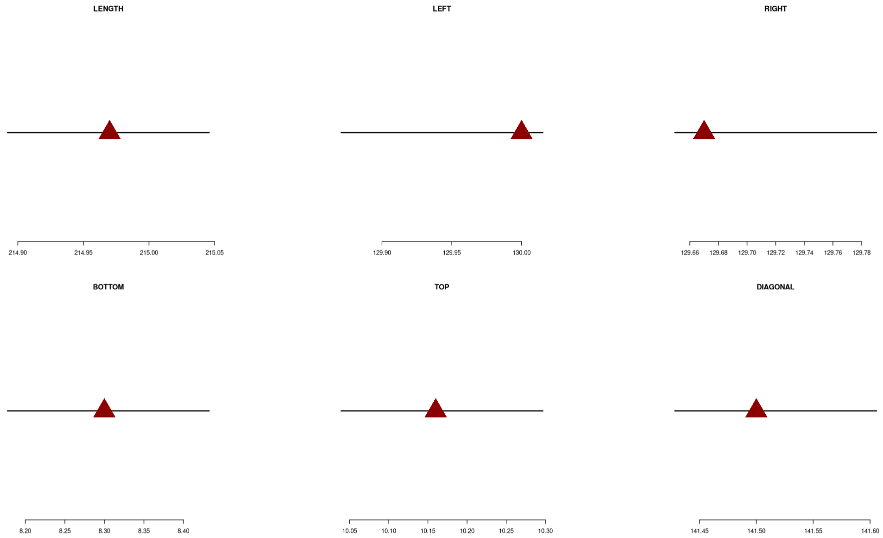

Figure 2: Line projections

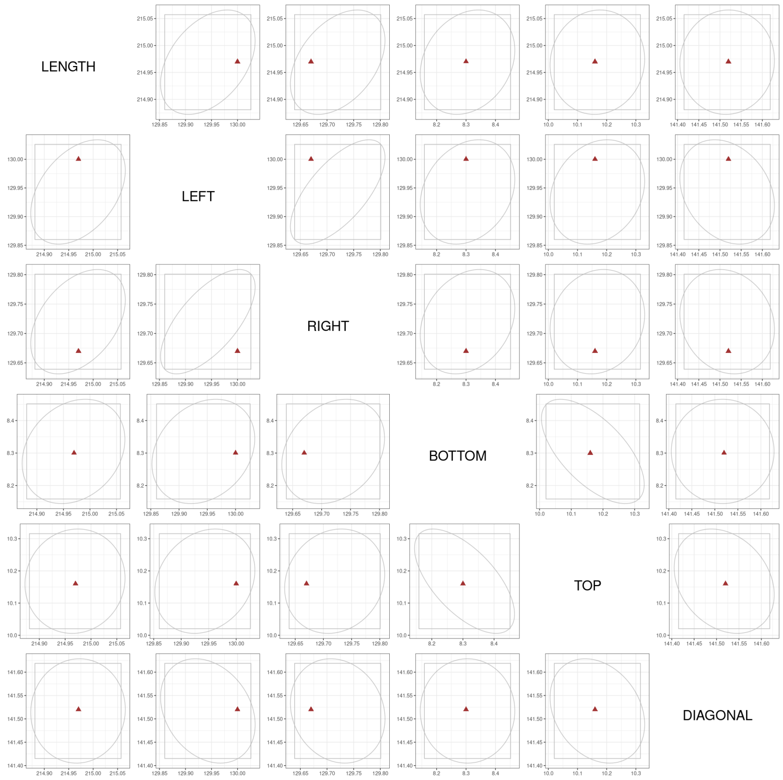

Figure 3: Pairwise plane projections

3.4 Bank note dimensions propositions

LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
214.99	129.95	129.73	8.51	9.96	141.55

Table 4: Bank notes means

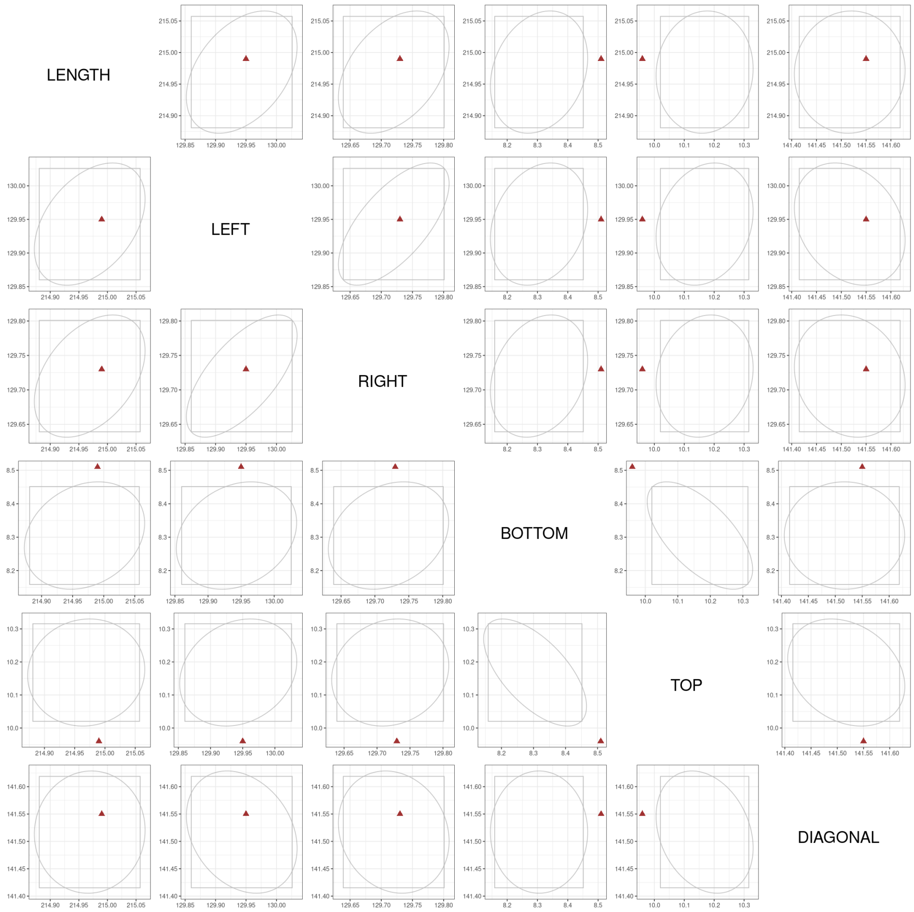

Figure 4: Pairwise plane projections

The mean fits inside the higher-dimensional ellipsoid but fails to fit inside of the Bonferroni's region. Looks like BOTTOM and TOP are in need of correction.

LENGTH	LEFT	RIGHT	BOTTOM	TOP	DIAGONAL
214.9473	129.9243	129.6709	8.3254	10.0389	141.4954

Table 5: Bank notes means

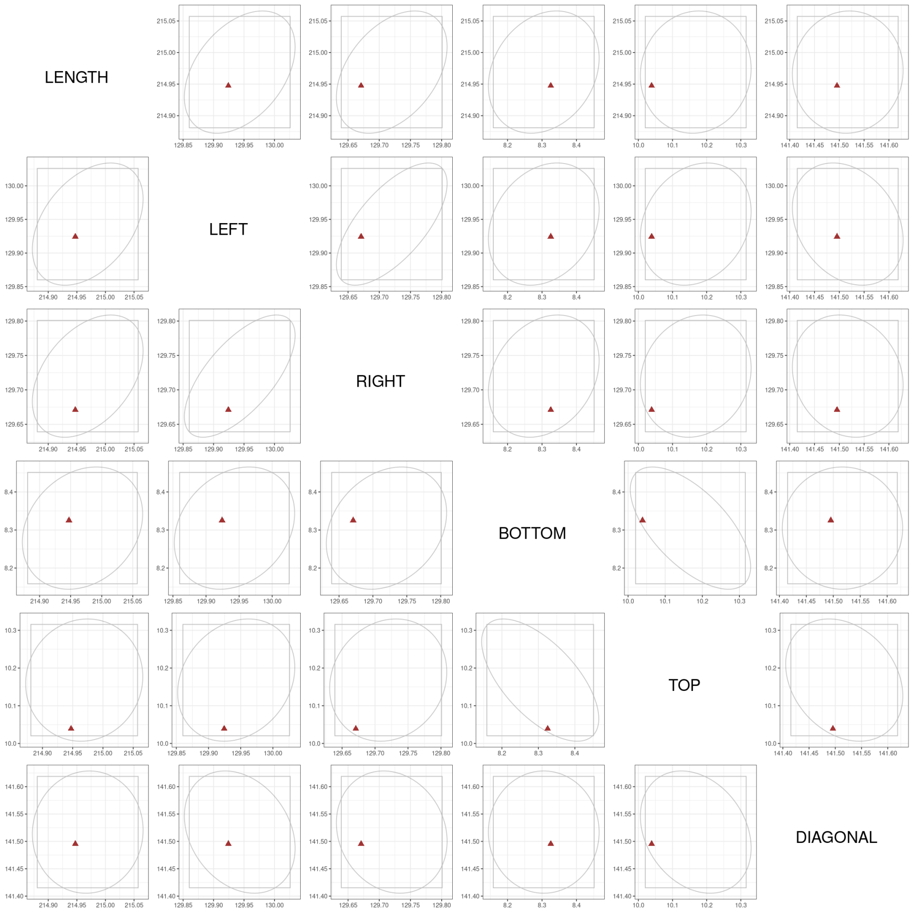

Figure 5: Pairwise plane projections

The final mean vector satisfies all high-dimensional confidence regions as well as their plane projections. It is a reasonable fit.

4. Simulation 1: Multiple testing

Consider the sequence of independent random variables $X_1, ..., X_p$ such that $X_i \sim \mathcal{N}(_i, 1)$ and the problem of the multiple testing of the hypotheses $H_{0i} : \mu_i = 0$, for $i \in 1, ..., p$ using Bonferroni and the Benjamini-Hochberg methods in the following setups:

$$\mu_1 = ... = \mu_{10} = \sqrt{2} \log p, \quad \mu_{11} = ... = \mu_p = 0$$

$$\mu_1 = ... = \mu_{500} = \sqrt{2} \log p, \ \mu_{501} = ... = \mu_p = 0$$

	bf	**bh**
rejected	0.9926667	0.9960000
fdr	0.1092331	0.0486361
power	0.3812333	0.5449000

Table 6: First 10 hypotheses are false

	bf	**bh**
rejected	1.000000000	1.00000000
fdr	0.002175374	0.04704834
power	0.386209333	0.90350533

Table 7: First 500 hypotheses are false

First few means are quite large and their respective null hypotheses are almost always rejected leading to the rejection of the global null with similar success for both methods. That being said, there is a clear power and FDR difference. Benjamini-Hochberg method is known to control FDR at or below level $\alpha \frac{p_0}{p}$ which, in this case, would be equal to 0.0499 and 0.045 for tables 6 and 7 respectively. This approach allows for more power by potentially identifying more true positives but usually results in a slightly higher proportion of false positives compared to the Bonferroni correction. It can be observed that BH method is very efficient when there are many false hypotheses present while the power of Bonferroni's method is somewhat agnostic to the change, but its FDR drops significantly.