Report 2

Kamil Zaborniak

Contents

Relationsship between the distributions

First, consider the relationship between the standard normal distribution and χ^2 with p degrees of freedom. Let us recall the definition of the F distribution with d_1 and d_2 degrees of freedom. Then we will check the relationship between the multivariate normal distribution and χ^2, and recall the theory of hypothesis testing using Hotelling statistic.

1. Let's assume that we have a sample of independent random variables: $Z_1, \ldots, Z_p$, coming from a standard normal distribution, then the sum of their squares:

$$\sum_{i=1}^{p} Z_i^2 = S^2 \sim \chi_p^2.$$

Where $\mathbb{E}[Q] = k$ and $Var[Q] = 2p$. Moreover, let us assume independent random variables: $U^2 = \sum_{j=1}^{k} Z_j^2 \sim \chi_k^2$ and $V^2 = \sum_{i=1}^{l} Z_i^2 \sim \chi_l^2$, then:

$$\frac{\frac{U^2}{k}}{\frac{V^2}{l}} = Q \sim F_{k,l}$$

.

2. Let $X \sim F_{p,n-p}$. We can write the variable X as $\frac{\chi_p^2/p}{\chi_{n-p}^2/(n-p)}$. From Weak Law of Large Numbers we know:

$$\frac{Z_1^2 + \ldots + Z_n^2}{n} \longrightarrow \frac{n}{n} = 1,$$

so

$$\frac{\chi_p^2/p}{\chi_{n-p}^2/(n-p)} \xrightarrow{n\to\infty} \frac{\chi_p^2/p}{1} = \frac{\chi_p^2}{p}.$$

For $p = 4$ and $n = 1000$, $F_{p,n-p}$ will approximately follow $\frac{\chi_4^2}{4}$.

3. Let $X_1 \ldots, N_n \sim N_p(\mu, \Sigma)$, where:

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_{1,1} & \ldots & \sigma_{1,p} \\ \vdots & \ddots & \vdots \\ \sigma_p, 1 & \ldots & \sigma_{p,p} \end{bmatrix}.$$

Then

$$\begin{aligned} n\left(\overline{X} - \mu\right)^T \Sigma^{-1} \left(\overline{X} - \mu\right) &= n\left(\overline{X} - \mu\right)^T \Sigma^{-1/2}\Sigma^{-1/2} \left(\overline{X} - \mu\right) \\ &= n\left(\Sigma^{-1/2}\left(\overline{X} - \mu\right)\right)^T \Sigma^{-1/2} \left(\overline{X} - \mu\right) \\ &= \left(\sqrt{n}\Sigma^{-1/2}\left(\overline{X} - \mu\right)\right)^2 \\ &= Z^T Z \\ &= \sum_{i=1}^{n} Z_i^2 \xrightarrow{D} \chi_n^2 \end{aligned}$$

Where $\sqrt{n}\Sigma^{-1/2}\left(\overline{X} - \mu\right) = Z \sim N_p(0, \mathbb{I})$, $\mu - \mu = 0$, $\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \mathbb{I}$, $Z = [Z_1, \ldots, Z_p]^T$.

4. Let $X_1 \ldots, N_n \sim N_p(\mu, \Sigma)$. Assume we do not know either μ or Σ. We want to test the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. The Hotelling statistic has form:

$$T^2 = n\left(\overline{X} - \mu_0\right)^T S^{-1} \left(\overline{X} - \mu_0\right),$$

where S denotes the sample covariance matrix. When null hypothesis is true, then $\frac{n-p}{(n-1)p}T^2$ is distributed as $F_{p,n-p}$, what is more, large values of T^2 lead to rejection of the null hypothesis. It follows that we reject H_0 at level α if $\frac{n-p}{(n-1)p}T^2 \geq F^{-1}_{p,n-p}(1-\alpha)$, where $F^{-1}_{p,n-p}(1-\alpha)$ is $1-\alpha$ quantile of $F_{p,n-p}$.

Based on the previous bullet points and The Weak Law of Large Numbers, if $n \to \infty$, Statistic $T^2 \xrightarrow{D} \chi^2_p$. It follows that probability of rejecting of H_0 is:

$$P_{H_0}\left(T^2 \geq (\chi^2_p)^{-1}_{1-\alpha}\right) \to \alpha.$$

Let us assume that H_1 is true, then the power of the above test, under $n \to \infty$ is:

$$P_{H_1}\left(T^2 \geq (\chi^2_p)^{-1}_{1-\alpha}\right).$$

Under H_1, T^2 has the noncentral χ^2_p distribution with noncentrality parameter $\lambda = T^2$, so power of above test depends on difference between $\overline{X}$ and μ_0. I think the power converges to 1 with $n \to \infty$, and I'm sure the power is higher than $1-\alpha$.

Multiple testing

We have random vector $X = [1.7, 1.6, 3.3, 2.7, -0.04, 0.35, -0.5, 1.0, 0.7, 0.8]^T \sim N_{10}(0, \mathbb{I})$. For each i-th. vector coordinate, let's test a set of hypotheses:

$$H_{0,i} : \mu_i = 0 \quad vs. \quad H_{1,i} : \mu_i \neq 0$$

The Bonferroni test will reject $H_0^{(i)}$ if $|X_i| \geq \left|\Phi^{-1}\left(\frac{\alpha}{2p}\right)\right|$. So in our case Bonferroni correction reject the 3rd. null hypothesis at significance level $\alpha = 0.05$.

The Benjamini-Hochberg procedure requires sorting in descending order the absolute values of the test statistics (in our case, just vector coordinates): $|X|_{(1)} \geq |X|_{(2)} \geq \ldots \geq |X|_{(p)}$. After that, it reject all $H_{0,(i)}$ for $i \leq i_{SU}$, where i_{SU} is the largest index which it occurs: $|X|_{(i)} \geq \Phi^{-1}(1-\alpha_i)$, $\alpha_i = \alpha\frac{i}{2p}$. I our case this procedure will reject 3rd. and 4th. hypothesis.

If we assume that the first 3 null hypotheses are false, then the FDR for Bonferroni procedure is equal to 0 and 0.5 for Benjamini-Hochberg procedure.

Simulation

Let's consider the sequence of independent random variables $X_1, ..., X_p$ such that $X_i \sim N(\mu_i, 1)$ and the problem of the multiple testing of the hypotheses $H0_i : \mu_i = 0$, for $i \in \{1, ..., p\}$. We assumme $p = 5000$ and $\alpha = 0.05$. We will use the simulations (at least 1000 replicates) to estimate FWER, FDR and the power of the Bonferroni and the Benjamini-Hochberg multiple testing procedures for the following setups.

1. $\mu_1 = \ldots = \mu_{10} = \sqrt{2\log p}, \quad \mu_{11} = \ldots = \mu_p = 0$ After 1000 replicates of test using Bonferroni procedure, it's power is equall to 0.392, FWER=0.04 and FDR=0.01. For Benjamini-Hochberg: power=0.546, FWER=0.282, FDR=0.048.

2. $\mu_1 = \ldots = \mu_{500} = \sqrt{2\log p}, \quad \mu_{501} = \ldots = \mu_p = 0$ After 1000 replicates of test using Bonferroni procedure, it's power is equall to 0.385, FWER=0.034 and FDR=0.0001. For Benjamini-Hochberg: power=0.902, FWER=1, FDR=0.045.

We see that Bonferroni's method makes fewer errors as the number of false null hypotheses increases, but its power fluctuates around 0.40. As the number of false null hypotheses increases, the procedure better controls FWER and FDR.

The Benjamini-Hochberg procedure, regardless of the case, does not control the FWER, but it does control the FDR, the value of which is less than $\alpha = 0.05$. Moreover, this procedure shows that its power increases as the number of false null hypotheses increases.

Analysis of Printing Bank Notes

In the following section, we will analyze banking data. The Swiss bank data consists of 100 measurements on genuine bank notes. The measurements are:

1. X_1 — length of the bill,
2. X_2 — height of the bill (left),
3. X_3 — height of the bill (right),
4. X_4 — distance of the inner frame to the lower border,
5. X_5 — distance of the inner frame to the upper border,
6. X_6 — length of the diagonal of the central picture.

The data can be found in the file *BankGenuine.txt*.

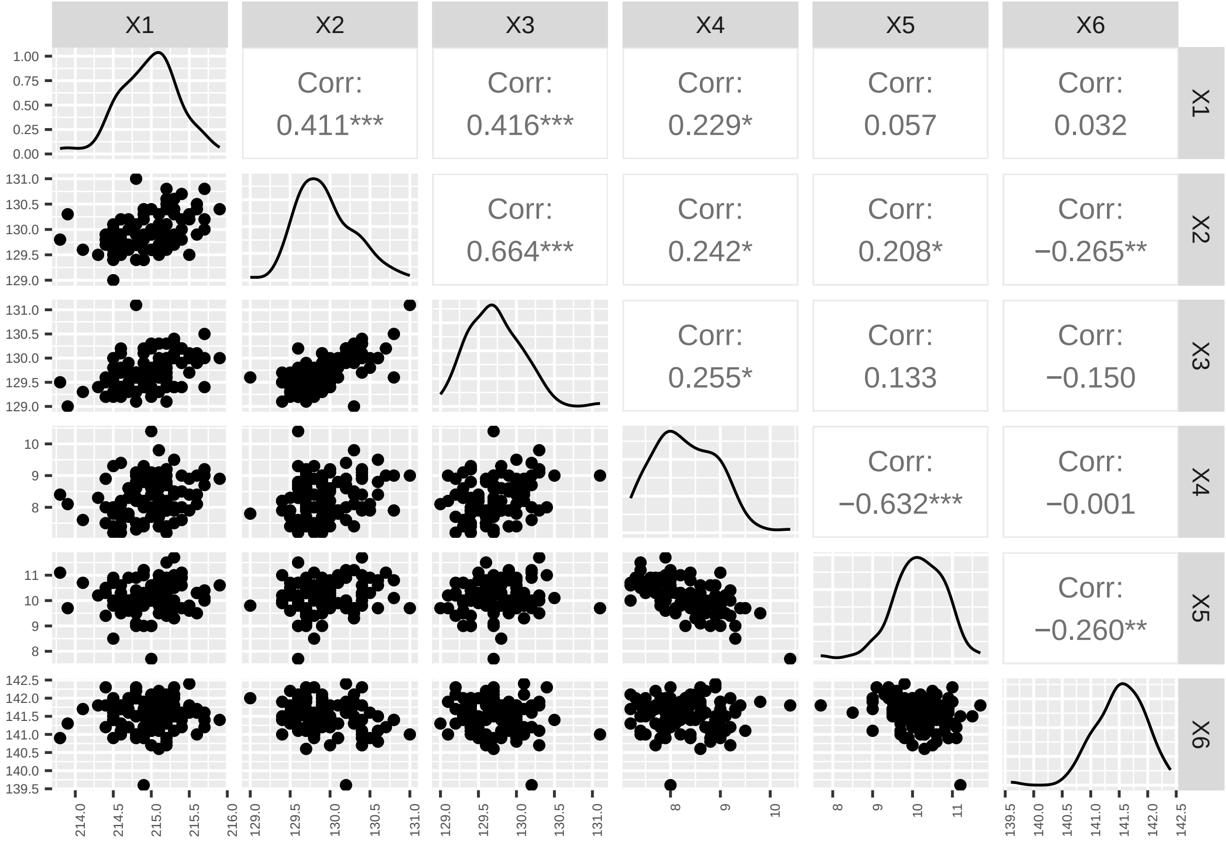

Figure 1: Scatterplots of Bank Notes data, densities and correlations between variables.

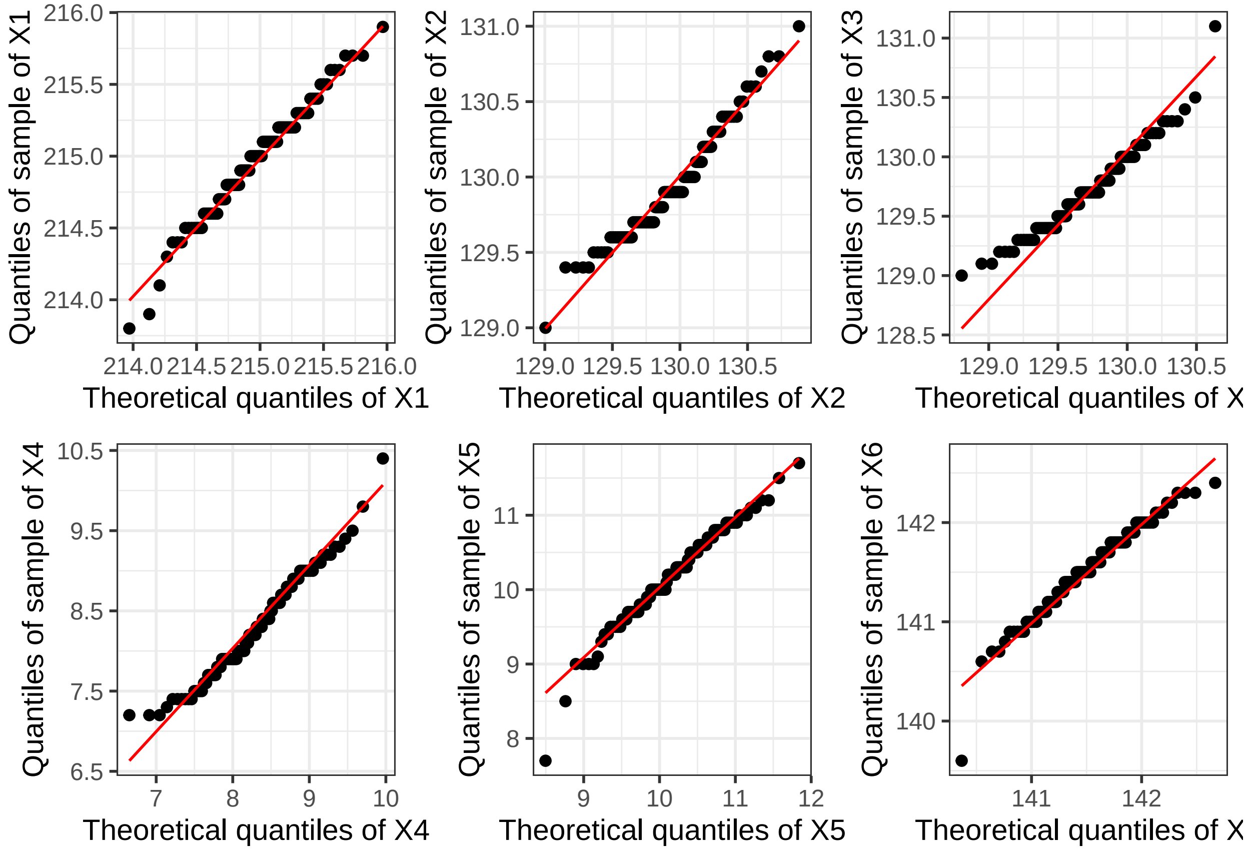

Figure 2: Q-q plots of variables.

The estimator vector of means of above variables is equal to:

$$\hat{\mu} = [214.969, 129.943, 129.720, 8.305, 10.168, 141.517]^T$$

and estimator of covariance matrix has form:

$$\hat{\Sigma} = \begin{bmatrix} 0.150 & 0.058 & 0.057 & 0.057 & 0.014 & 0.005 \\ 0.058 & 0.133 & 0.086 & 0.057 & 0.049 & -0.043 \\ 0.057 & 0.086 & 0.126 & 0.058 & 0.031 & -0.024 \\ 0.057 & 0.057 & 0.058 & 0.413 & -0.263 & 0.0001 \\ 0.014 & 0.049 & 0.031 & -0.263 & 0.421 & -0.075 \\ 0.005 & -0.043 & -0.024 & 0.0001 & -0.075 & 0.200 \end{bmatrix}.$$

Let's write, based on the Hotelling's T^2 statistics, function in R that verifying if a point lies inside of the six dimensional ellipsoid, which serve as the 95% confidence region for the value the mean value of bank notes. We treat this task as a hypothesis test: $H_0 : \mu = X$ vs. $H_1 : \mu \neq X$ in which we use Hotelling's T^2 statistic.

In our case, from lecture notes, we know that point X lies inside of the six dimensional 95% CR if:

$$100 \cdot (\hat{\mu} - X)^T \hat{\Sigma}^{-1} (\hat{\mu} - X) \leq \frac{(100-1)\cdot 6}{100-6} F^{-1}_{6,100-6}(0.95)$$

what is equivalent to:

$$100 \cdot (\hat{\mu} - X)^T \hat{\Sigma}^{-1} (\hat{\mu} - X) \leq \frac{594}{94} F^{-1}_{6,94}(0.95)$$

The function in R of above problem has form:

```
verifyPointsInsideCR<- function(data, point, prob=0.95){
  n <- length(data[,1])
  p <- length(data)
  S_inv<- cov(data) %>% solve()
  means <- data %>% colMeans()

  T2 <- n*t(means - point) %*% S_inv %*% (means - point)
  crit_v <- (n-1)*p/(n-p) * qf(prob, p, n-p)

  return(T2<= crit_v)
}
```

Let's consider the following situation: "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 $\mu_0 = [214.97, 130, 129.67, 8.3, 10.16, 141.52]^T$. (Since the number of bank notes printed out for this purpose was very large so the error of for the obtained mean values is negligible)". Using the procedure and function *verifyPointsInsideCR* described above, let's check whether a given point belongs to 95% CR.

The T^2 is equal to 13.9149 and the critical value is equal to 13.8807, which means that μ_0 doesn't lies inside of our 95% CR.

Let's define the Bonferroni's confidence rectangular region. the below interval cover our real μ_i with probability 0.95:

$$\hat{\mu_i} \pm t_{99}\left(1-\frac{0.05}{12}\right)\sqrt{\hat{\sigma_i}^2/100}$$

Then the real vector of means is covered by the Bonferroni's confidence rectangular region with probability less or equal to $(1-0.05)^6$.

The point μ_0 lies inside the Bonferroni's confidence rectangular region.

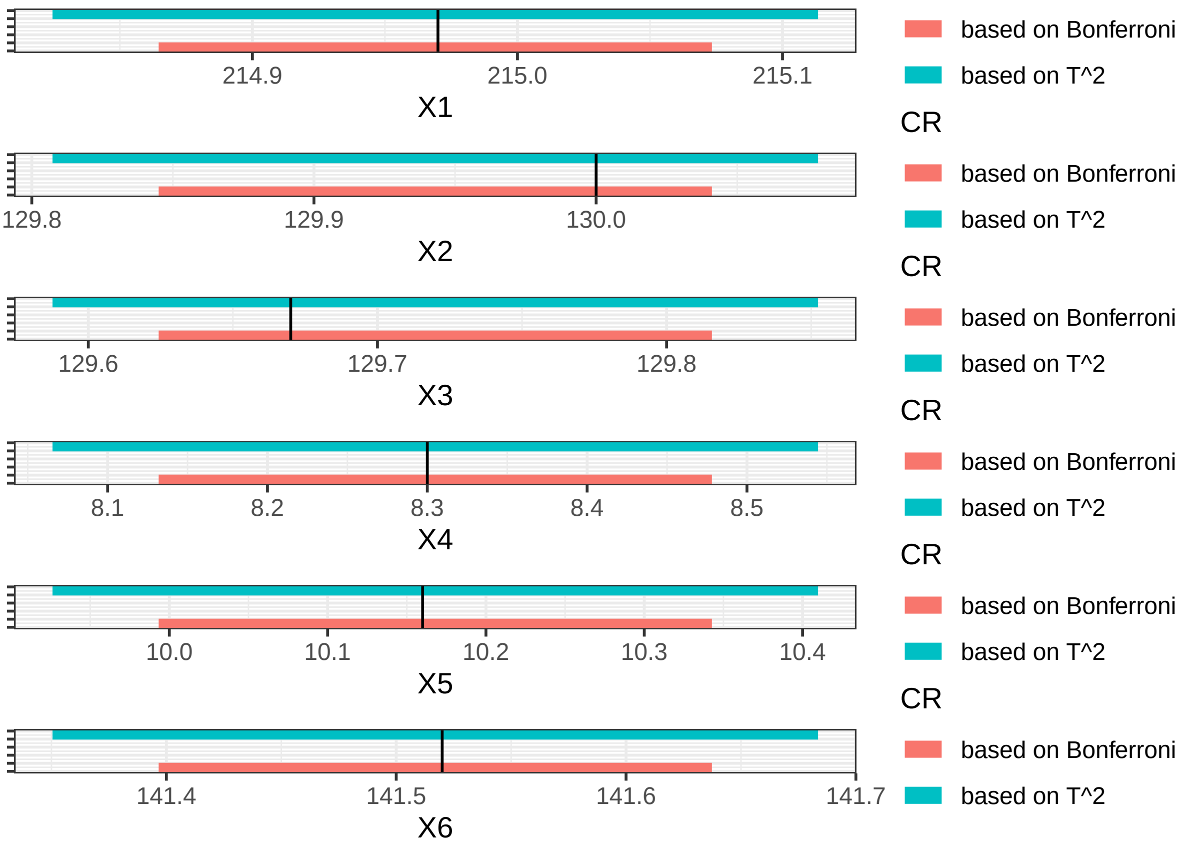

Figure 3: Projections of the both of confodence intervals to the one-dimensional spaces.

As we can see, the projection of point μ_0 belongs to every CR projection into a one-dimensional subspace, regardless of the type of CR. Moreover, depending on the type of CR, the projections differ.

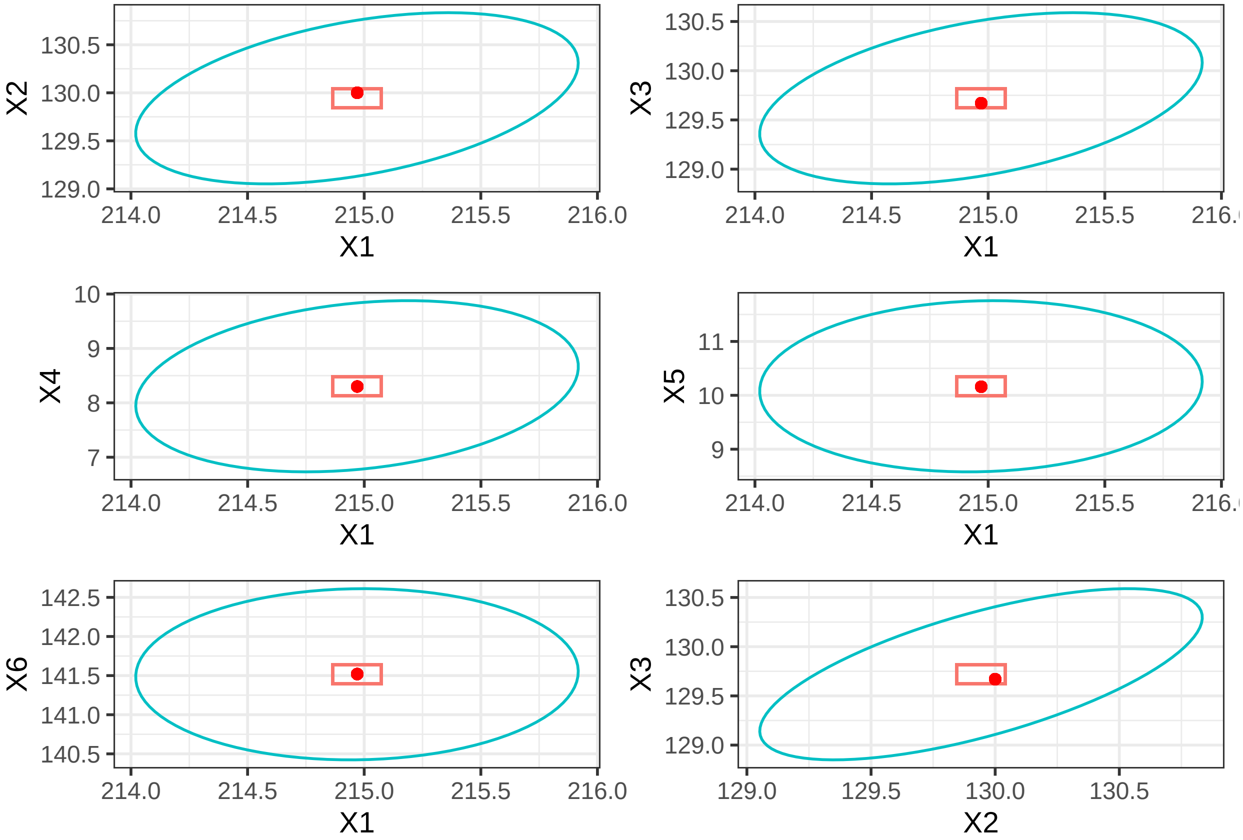

Figure 4: Projections of the both of CR types to the two-dimensional spaces.

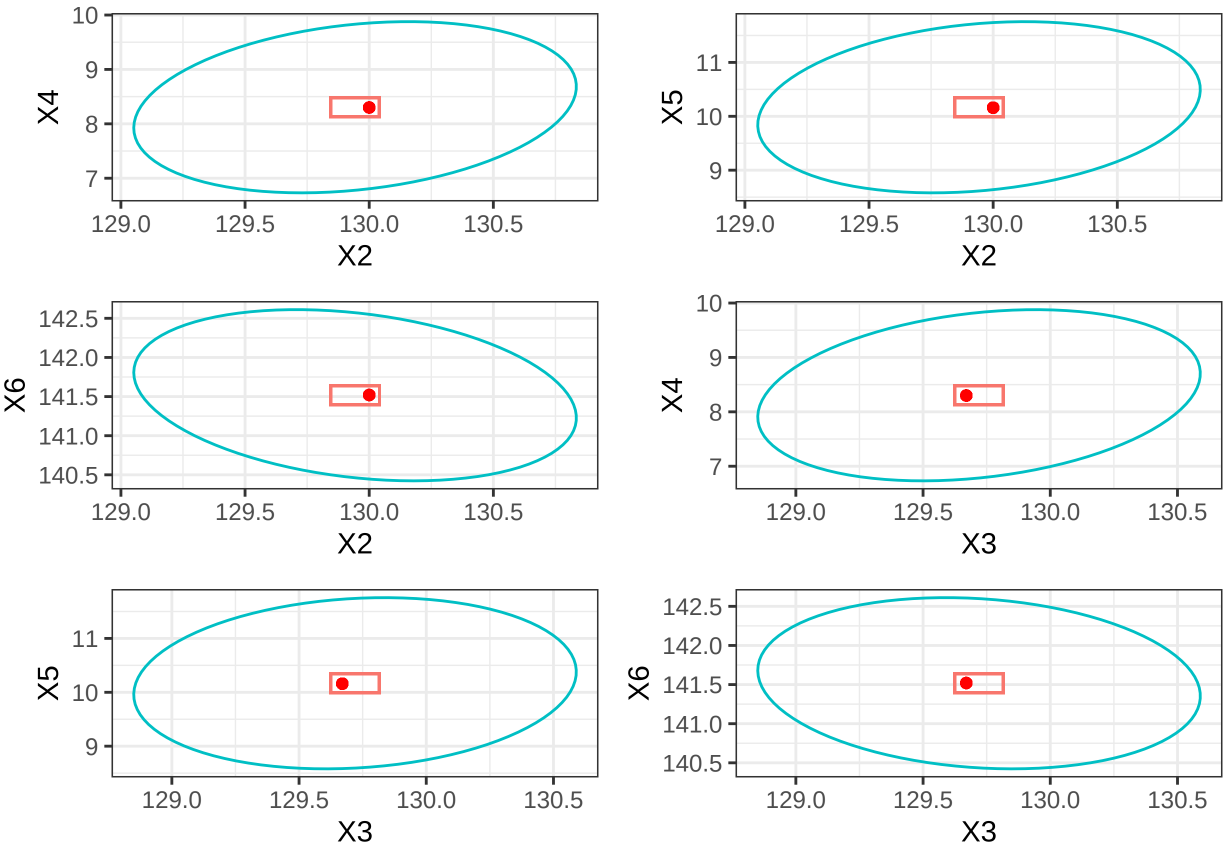

Figure 5: Projections of the both of CR types to the two-dimensional spaces.

```
FALSE adding dummy grobs
```

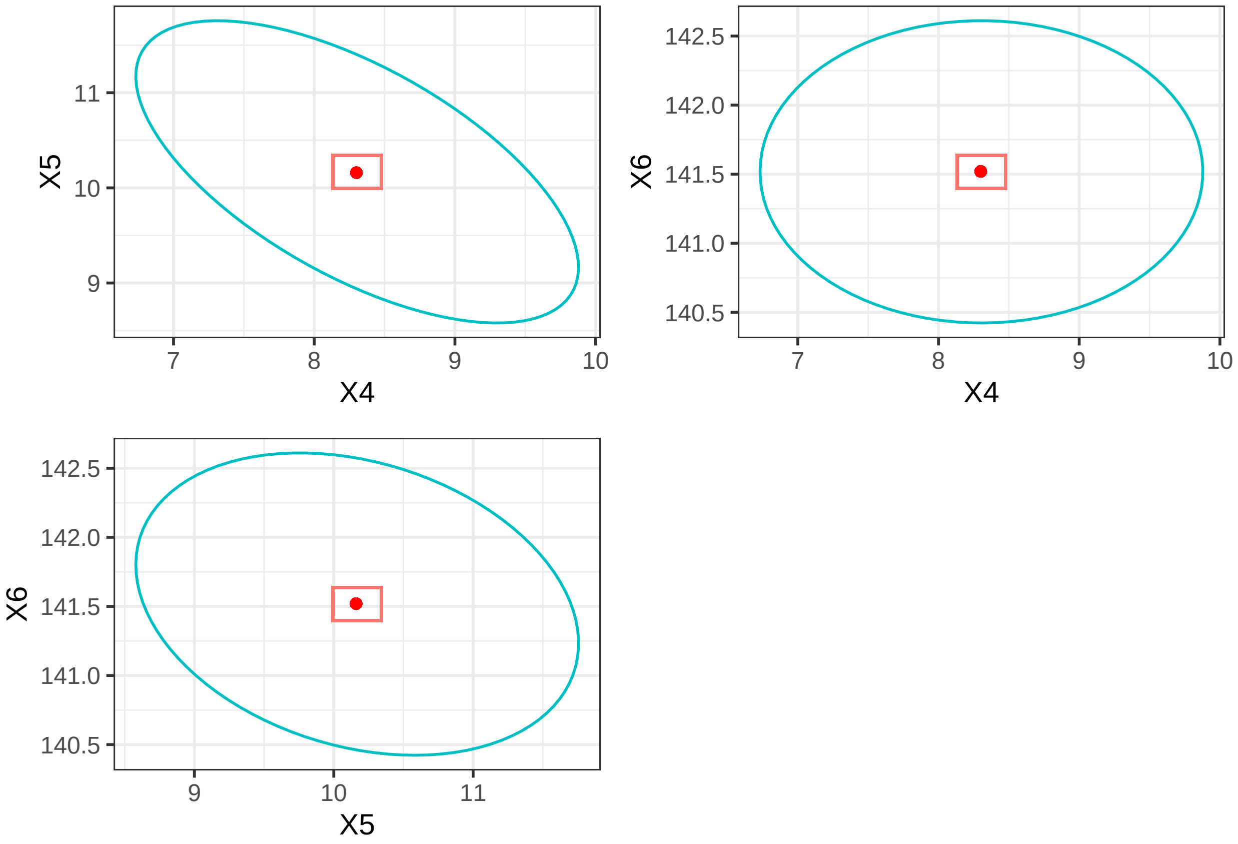

Figure 6: Projections of the both of CR types to the two-dimensional spaces.

As we can see in the graphs above, each projection of our point onto two-dimensional subspaces belongs to the CR projection, regardless of the CR type. It may seem contradictory that the point μ_0 does not belong to the CR determined by the T^2 statistic. However, if the projection of this CR covers the projection of point μ_0, it does not mean that point μ_0 belongs to this CR. This is also confirmed by the fact that this CR has the form of a 6-dimensional ellipse, unlike the CR determined using the Bonferroni method.

It has been decided that the settings of the production line needs to be tuned better to match original dimensions of banknotes. After such tuning, another test has been carried out and the resulting means were $\mu_1 = [214.99, 129.95, 129.73, 8.51, 9.96, 141.55]^T$. This point is covered by the Hotelling's 95% CR, but is not covered by Bonferroni 95% CR.

After yet another tuning, the vector of means was $\mu_2 = [214.9473, 129.9243, 129.6709, 8.3254, 10.0389, 141.4954]^T$. This point is covered by both Confidence Regions. It's means that the probability of the event "μ_2 is the nearest the real vector of means", is the highest.