Statistical learning

Report 2

Maciej Szczutko

2024-05-07

μ testing in multivariate normal

The χ_p and F distribution.

Let $Z_1, Z_2, \ldots$ be independent copies of a $\mathcal{N}(0,1)$ variable. Use these to define a chi-squared random variable with p degrees of freedom χ^2_p. Similarly, recall the definition of an F distribution with d_1 and d_2 degrees of freedom F_{d_1,d_2}.

We write some random variable has χ^2_p distribution when is define as $\sum_{i=1}^{p} Z_i^2$.

We write some random variable has F_{d_1,d_2} distribution when is define as ratio of two, independent chi-squares distribution, scaled by their degres of freedom. Formally $X \sim F_{d_1,d_2}$ when $X = \frac{S_1/d_1}{S_2/d_2}$ where S_1 and S_2 are independent random variables with chi-square distributions with respective degrees of freedom d_1 and d_2.

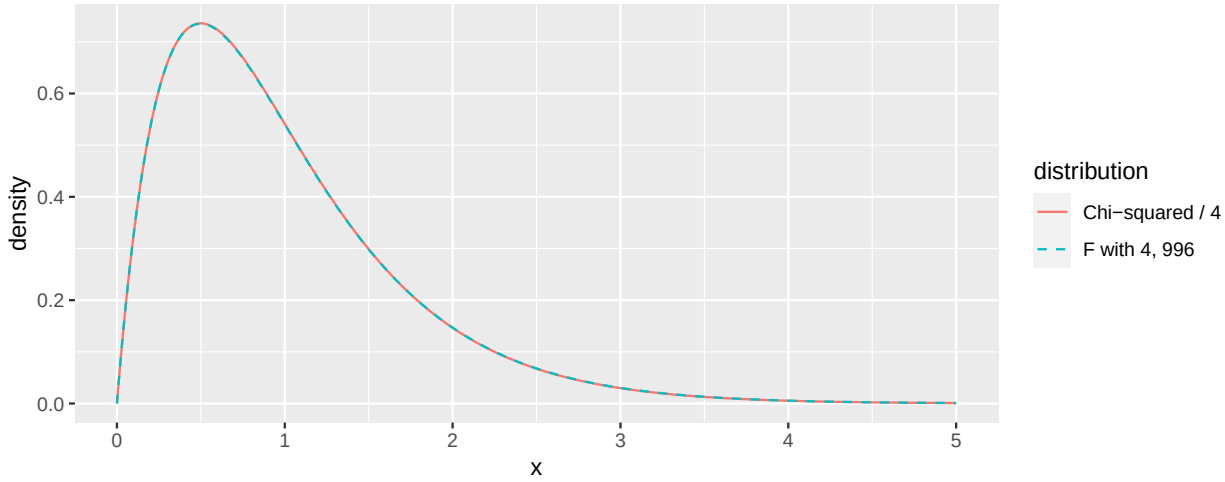
What distribution will $F_{p,n-p}$ approximately follow for $p = 4$ and $n = 1000$?

We want to investigate case we n is much larger than p. To do this we look closly on F distribution when d_2 goes to ∞.

$F_{p,n-p} \sim \frac{\chi^2_{(p)}/p}{\chi^2_{(n-p)}/(n-p)}$. By the LLW $\chi^2_{(n-p)}/(n-p) \to E(\chi^2_{(1)}) = 1$ when $n \to \infty$. Thus

$\frac{\chi^2_{(p)}/p}{\chi^2_{(n-p)}/(n-p)} \to \frac{\chi^2_{(p)}}{p}$ when $n \to \infty$.

We can use Jacobian formula to derive the density of $\frac{\chi^2_{(4)}}{4}$ and verify the denisty plots.



```
## NULL
```

As we can see the density of $\frac{\chi^2_{(4)}}{4}$ is approximately the same as F density.

Distirbution of statistical distance

Let $X_1, .., X_n \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that $n(X - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu})$ follows a χ^2_p distribution.

We need to show that this expression can be written as $\sum_{i=1}^p Z_i^2$. As the covariance matrix is semipositive define we can perform spectral decomposition and write.

We use following fact from previous list:

1. If e_i is eigen vector for λ_i eigen value of Σ then e_i is eigen vector for $\frac{1}{\lambda_i}$ of Σ^{-1}
2. The distribution of $\mathbf{P}^T(\mathbf{X} - \mu)$ is the same as the $\left(\sqrt{\lambda_1} Z_1, \ldots, \sqrt{\lambda_p} Z_p\right)$, where Z_i 's are independent standard normal random variables.

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p} (1/\lambda_i) (\mathbf{X} - \boldsymbol{\mu})' \mathbf{e}_i \mathbf{e}_i' (\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^{p} (1/\lambda_i) \left(\mathbf{e}_i' (\mathbf{X} - \boldsymbol{\mu})\right)^2 =$$

$$\sum_{i=1}^{p} \left[\left(1/\sqrt{\lambda_i}\right) \mathbf{e}_i' (\mathbf{X} - \boldsymbol{\mu}) \right]^2 = \sum_{i=1}^{p} Z_i^2$$

.

Testing μ in multivariate normal distirbution

Let $X_1, .., X_n \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume we do not know either $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$. We want to test the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. We introduce the Hotelling T^2 statistic:

$$T^2 := n \left(\bar{X} - \boldsymbol{\mu}_0\right)^T \boldsymbol{S}^{-1} \left(\bar{X} - \boldsymbol{\mu}_0\right),$$

where $\boldsymbol{S}$ denotes the sample covariance matrix.

T^2 distribution under H_0.

From lecture we know, if H_0 is true, then $\frac{n-p}{(n-1)p} T^2$ is distributed as $F_{p,n-p}$. The Hotteling test reject H_0 when $\frac{n-p}{(n-1)p} T^2 > F_\alpha (p, n-p)$ where $F_{1-\alpha} (p, n-p)$ denote quantile of level $1-\alpha$ from F distribution with p and $n-p$ degree of freedom.

Test behaviour under H_0 when observation number goes to ∞.

We will use the asymptotic derived before.

$$\frac{n-p}{(n-1)p} T^2 \text{ is distributed as } F_{p,n-p}.$$

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

Using the law of limits (as both exist), because $\frac{(n-1)p}{n-p} \to p$ as $n \to \infty$ and $F_{p,n-p} \to \frac{\chi^2_{(p)}}{p}$ then $T^2 \to p \cdot \frac{\chi^2_{(p)}}{p} = \chi^2_{(p)}$ as $n \to \infty$.

Using this asymptotic we can write $P\left[n(\overline{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}) \leq \chi^2_p(\alpha)\right] \doteq 1 - \alpha$

In fact, if we have large sample from population, we can leave normality assumption as is guarantee by CTL (see Richard A. Johnson, Dean W. Wichern - Applied Multivariate Statistical Analysis (6th Edition) 5.5)

Test behaviour under H_A when observation number goes to ∞.

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 $\bar{X}$ and μ_0. When n goes to ∞ then also T^2 goes ∞. Then $P\left(\frac{n-p}{(n-1)p}T^2 > F_{1-\alpha}(p, n-p)\right) \to 1$ so probability of rejecting H_0 goes to 1.

Multiple Testing

Random vector

$$X = (1.7, 1.6, 3.3, 2.7, -0.04, 0.35, -0.5, 1.0, 0.7, 0.8)$$

comes from the 10 dimensional multivariate normal distribution $N(\mu, I)$.

We want to test whether

$$H_{0(i)} : \mu_i = 0 \quad vs \quad H_{A(i)} : \mu_i \neq 0.$$

We perform 3 multiple testing procedures

1. Bonfferoni

Bonferroni reject $H_{0(i)}$ if $p_i < \frac{\alpha}{n}$.

In this case the test reject hypothesis with indexes i: 3.

2. Benjamini-Hohberg

Algorithm: Benjamini-Hochberg Procedure
$j = n$.
while $p_{(j)} > \dfrac{a}{n-j+1}$: do
$j = j - 1$
end while
Reject $H_{(1)}, \ldots, H_{(j)}$

In this case the test reject hypothesis with indexes i: 3, 4.

The Bonferroni it's pretty naive in construction and demand strong signals to reject. BH is more suffistacted in construction and less conservative.

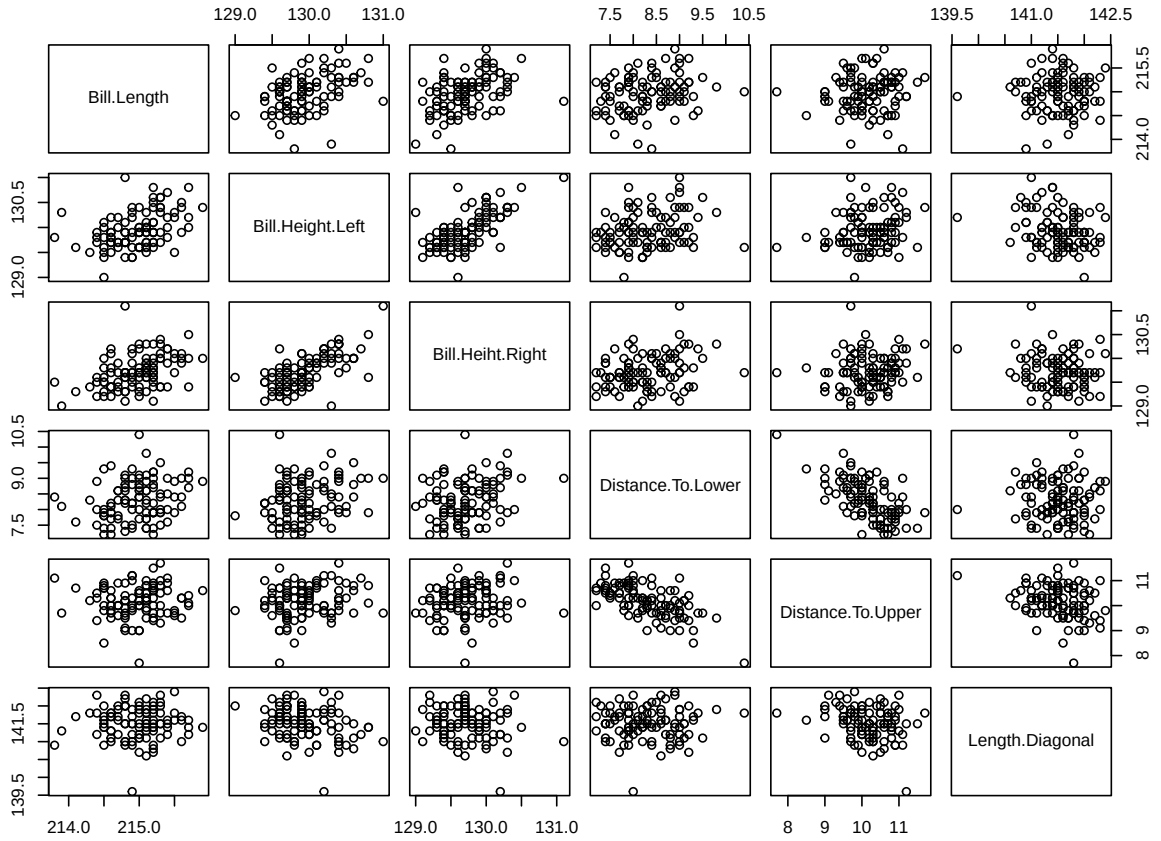
	FDP
Bonferroni	0
BH	0.5

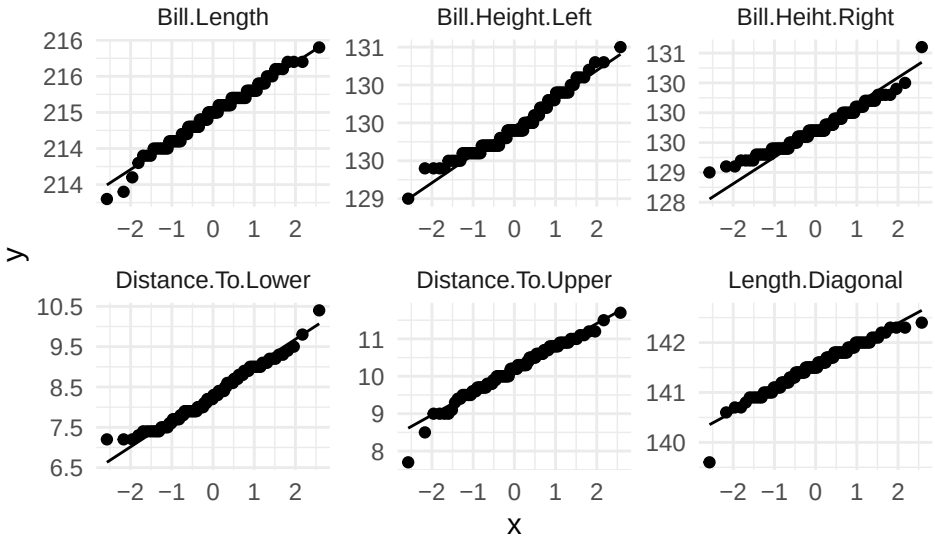
BankGenuine dataset

Now we will inspect BankGenuine dataset. We want to perform some analysis about new production line setting based on historical data. First we will check whether normality assumptions holds.

Table 2: BankGenuine dataset

Bill.Length	Bill.Height.Left	Bill.Heiht.Right	Distance.To.Lower	Distance.To.Upper	Length.Diagonal
215	131	131	9.0	9.7	141
215	130	130	8.1	9.5	142
215	130	130	8.7	9.6	142
215	130	130	7.5	10.4	142





From the graphs above, we can assume that the data are from a multivariate normal distribution.

μ and Σ estimators BankGenuine data

$$\mu = (214.969, 129.943, 129.72, 8.305, 10.168, 141.517)^T \text{ S=} \begin{bmatrix} 0.1502 & 0.058 & 0.0573 & 0.0571 & 0.0145 & 0.0055 \\ 0.058 & 0.1326 & 0.0859 & 0.0567 & 0.0491 & -0.0431 \\ 0.0573 & 0.0859 & 0.1263 & 0.0582 & 0.0306 & -0.0238 \\ 0.0571 & 0.0567 & 0.0582 & 0.4132 & -0.2635 & -0.0002 \\ 0.0145 & 0.0491 & 0.0306 & -0.2635 & 0.4212 & -0.0753 \\ 0.0055 & -0.0431 & -0.0238 & -0.0002 & -0.0753 & 0.1998 \end{bmatrix}$$

Verify if point lies in multidimensional ellipse

To check whether new production line is aligned with previous one we can perform Hoteling test for mean measured in new data. Following code will check the production lines are "statistically identical".

```
is_point_in_ellipse <- function(X, mu, sample_covariance, n, confidence_level = .05)
{
  statistical_distance <- n * (t(X-mu) %*% solve(sample_covariance) %*% (X-mu))
  df_of_asymptotic <- length(mu) #dimension of distribution
  critical_value <- qchisq(1-confidence_level, df_of_asymptotic)
  is_in_ellipse <- statistical_distance < critical_value
  returnValue(list("statistical_distance" = statistical_distance, "critical_value" = critical_value, "is
}
```

The new measured mean is (214.97, 130, 129.67, 8.3, 10.16, 141.52). The statistical distance is: $[13.9149]$ when critical value is 12.5916. We conclude that new mean lie outside ellipsoid.

Bonferroni's confidence interval for orginal banknote.

$$\bar{X}_i \pm t_{n-1}\left(\alpha_i/2\right)\sqrt{s_i^2/n}$$

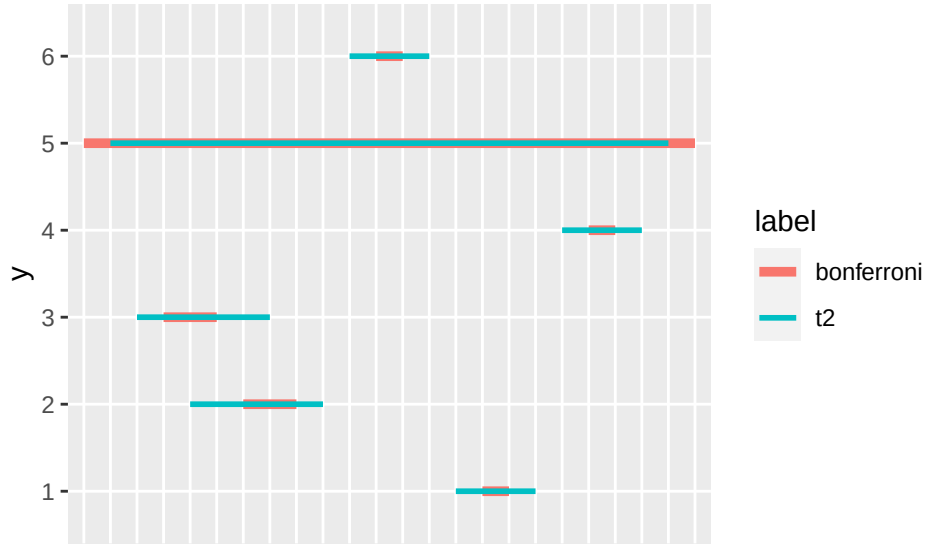
Here we set $\alpha_i = \frac{\alpha}{p}$ for each coordinate.

Following function does check whether point falls into rectangular condifence region.

This time we conclude that point lie in confidence intervals.

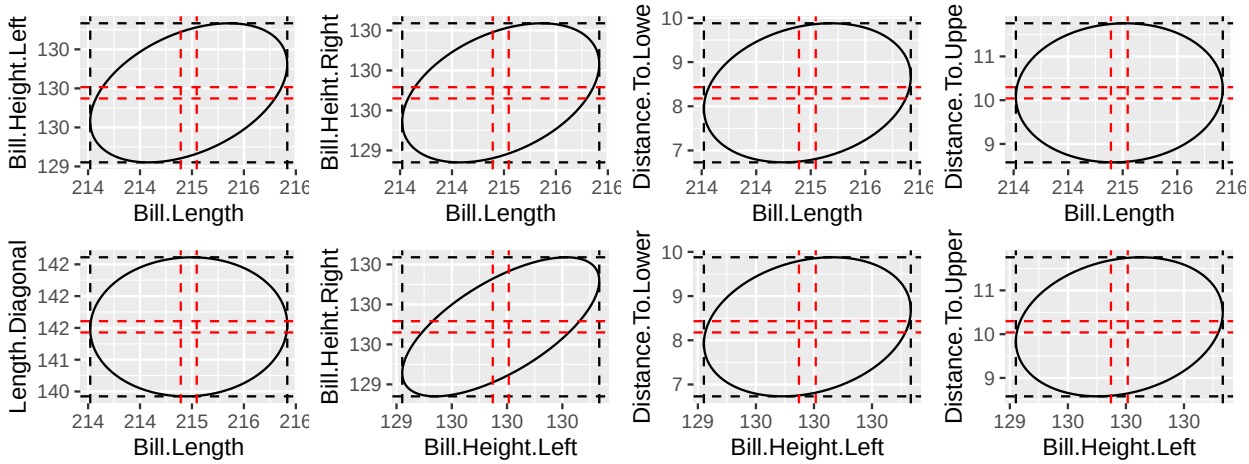
1d projections of confidence intervals

Now we will visual inspect difference between both types of confidence intervals.

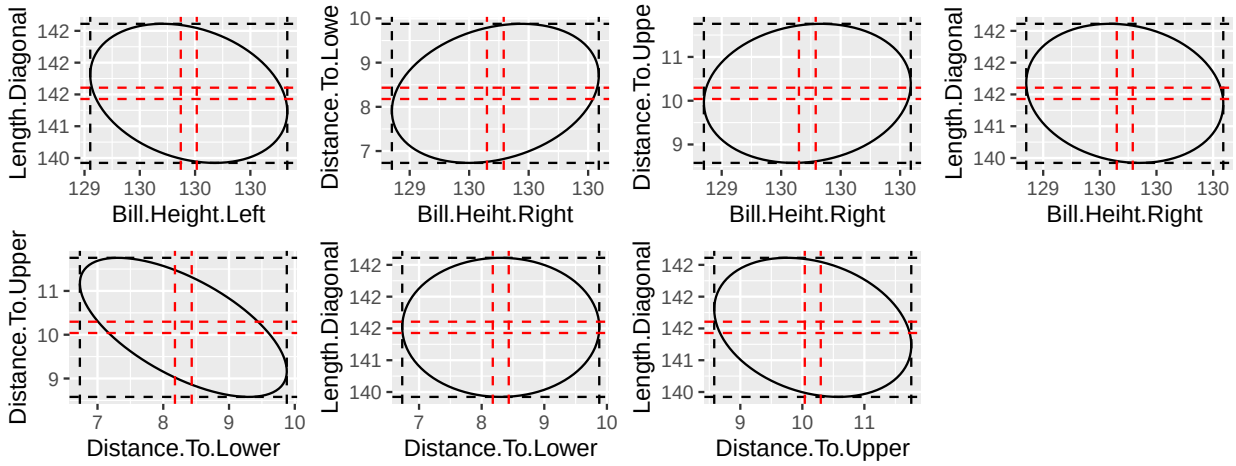


In general, the projection of confidence interval obtained with T^2 are wider because they consider relations between other variables when Bonferroni looks only on variance for given coordinates.

Better intuition coming from 2d projection



On 2d projections plots the difference is huge. The difference is so drastic because Bonferroni naive construction use $\alpha_i = \frac{\alpha}{6}$. For higher dimensional distributions it would be even worse. The Bonferroni looks only on components variance and scale quantile, when T^2 is designed to take into account the relationship between variables. I would say that Bonferroni is a correct construction, but impractical and worse for interpretation.



Interference about new settings based on measurments.

The new setting (called **m1**) does not fall into T^2 ellipsoid region and neither Bonferroni rectangular region. Most likely the setting is even worse than previous one.

The settings called **m2** is very close to original one. It falls into both test region criteria. I wasn't sure the confidence level = 0.95 is accurate for such import industry branch. But even when I was a bit more restrictive and set level to 0.99 the conclusion is the same. I would say this is correct set of parameters.

Simulation (multiple testing)

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. $\mu_1 = \ldots = \mu_{10} = \sqrt{2 \log p}, \quad \mu_{11} = \ldots = \mu_p = 0$
2. $\mu_1 = \ldots = \mu_{500} = \sqrt{2 \log p}, \quad \mu_{501} = \ldots = \mu_p = 0$

Result presented in table.

Table 3: Estimated power, FWER, FDR for Benjamini-Hochberg and Bonferroni.

	power_bonf	FWER_bonf	FDR_bonf	power_bh	FWER_bh	FDR_bh
a)	0.39	0.04	0.01	0.55	0.29	0.05
b)	0.39	0.05	0.00	0.90	1.00	0.05

We see that Bonferroni is powerless in both setup. In such distributed signals it's expected. But it control FWR and FWER as it was designed.

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, for this procedure shows, power increases as the number of false null hypotheses increases.