Multivariate Normal Statistical Testing

Tarek El-Hajjaoui

2023-05-11

Type I Error (False Positive) - The type I error rate of a test is the probability of rejecting the null hypothesis when it is actually true. - In most research contexts, the acceptable type I error rate is set at 0.05, meaning that there is a 5% chance of falsely rejecting the null hypothesis. - In many tests, critical values are chosen to aim to control the type I error rate of a test - The true type I error rate of a test depends on whether the underlying assumption of a test is met - A simulation study can help assess whether a statistical test maintains the desired type I error rate under various conditions.

Power (probability of a True Positive) - The power of a statistical test is the probability of correctly rejecting a null hypothesis when it is false. - In other words, power is the ability of a test to detect a true effect or relationship between variables. - Power is affected by several factors, including sample size, effect size, and significance level. - For simple tests, analytical formulas for calculation power might exist. For complicated tests, simulations can be used to evaluate power

In research we often need to compare different test statistics. Consider three multivariate normal populations. Here I would like you to assess their type I error rate and power of the following four tests for testing $H_0 : \mu_A = \mu_B = \mu_C$ vs $H_1 : H_0$is not true, where $\mu_A, \mu_B, \mu_C \in \mathbf{R}^p$ are the population mean vectors of the three subpopulations, respectively.

We assume that the three underlying populations have the same variance-covariance matrix Σ. It is helpful to consider different situations. For example, here is a set of four different situations:

1. Independent and equal variance:
$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
2. Positively dependent:
$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$$
3. Negatively dependent:
$$\Sigma = \begin{pmatrix} 1 & -0.2 & -0.2 \\ -0.2 & 1 & -0.2 \\ -0.2 & -0.2 & 1 \end{pmatrix}$$
4. Independent but unequal variance:
$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Independent and equal variance, Positively dependent, Negatively dependent

- Wilk's Lambda should perform well, with good control of Type I error and reasonable power. However, its power might be somewhat lower than some of the other tests when the effect size is large.
- Lawley-Hotelling trace should have good control of Type I error and might have greater power than Wilks' Lambda for large effect sizes or many dependent variables.
- Pillai's trace should perform well, with good control of Type I error and reasonable power. Its power might be somewhat lower than Lawley-Hotelling's trace when the effect size is large or when there are many dependent variables.
- Roy's largest root should have good control of Type I error and might have the greatest power of these four tests when there is one large root. However, its power might be lower when there are several smaller roots.

Independent but unequal variance

- If the variances are unequal, Pillai's trace might perform best (both power and control of Type I error) because it is robust to violations in assumptions.

Use Trial-and-Error to choose reasonable values for $\Sigma, \mu_A, \mu_B, \mu_C$.

Empirical power. To compare the performance of the four methods, you can estimate their power by running simulations. In each simulation, you first generate a random sample from each of the three populations; you then calculate the p-values of the four tests; finally you reject a test if the p-value is less than 0.05. By running B simulations, you can estimate the power of a particular method using the number rejected tests divided by B. B should be chosen such that your estimated power is accurate enough. In my experience, B should be at least a few hundred.

Type I error rate. The method for estimating type I error rate is similar to estimate power empirically, with the difference being that type I error rate is obtained when the null hypothesis is true.

Sample Size. Assume n observations will be obtained from each population. You need to choose n carefully such that the empirical power is neither too small nor too large. This is because when n is too small (large), all methods would have low (high) power and you cannot distinguish their performance. Ideally, you should choose n such that the power of the best method is between 0.8 and 0.9.

```r
#rearrange the data such as the response matrix is
#an n-by-p matrix
Y=cbind(SepalL=c(iris3[,1,1],iris3[,1,2],iris3[,1,3]),
SepalW=c(iris3[,2,1],iris3[,2,2],iris3[,2,3]),
PetalL=c(iris3[,3,1],iris3[,3,2],iris3[,3,3]),
PetalW=c(iris3[,4,1],iris3[,4,2],iris3[,4,3]))
#for unknown reasons, data.frame won't work but cbind works
#alternatively, we can use the following way to define y
#Y=aperm(iris3,c(1,3,2));dim(y)=c(150,4)
#define the covariate variable X, which is vector of labels
iris.type=rep(c("Setosa","Versicolor","Virginica"),each=50)
obj=manova(Y~iris.type)
summary(obj, test="Wilks")
```

```
##            Df    Wilks approx F num Df den Df    Pr(>F)
## iris.type   2 0.023439   199.15      8    288 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(tidyr) #the pipe (%>%) tool is extremely useful
library(MASS)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v dplyr   1.1.2
## v tibble  3.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
# Independent and equal variance matrix
Sigma_ind_eq = matrix(c(1,0,0,
                      0,1,0,
                      0,0,1),nrow=3)
# Positively dependent
Sigma_pos_dep = matrix(c(1,0.5,0.5,
                       0.5,1,0.5,
                       0.5,0.5,1),nrow=3)
# Negatively dependent
Sigma_neg_dep = matrix(c(1,-0.2,-0.2,
                       -0.2,1,-0.2,
                       -0.2,-0.2,1),nrow=3)
# Independent but unequal variance
Sigma_ind_uneq = matrix(c(1,0,0,
                        0,2,0,
                        0,0,3),nrow=3)
```

```
# n = the number of samples
# mu = vector of means for p variables
# Sigma = covariance matrix (positive-definite and symmetric)
gen_multivariate_norm <- function(n, mu, Sigma) {
  mvrnorm(n, mu, Sigma)
}
```

```
#sample = gen_multivariate_norm(10,c())
```

1. Wilk's lambda statistic $\Lambda = \frac{|W|}{|B+W|}$

- recommended when have small sample sizes or unequal group sizes

```
# obj=manova(Y~iris.type)
# summary(obj, test="Wilks")
# recommended when you have small sample sizes or unequal group sizes
```

2. Lawley-Hotelling trace $tr[BW^{-1}]$

- The Lawley-Hotelling trace is generally more powerful than Wilks' Lambda when the effect size is large or when there are many dependent variables.

```
# obj=manova(Y~iris.type)
# summary(obj, test="Hotelling-Lawley")
```

3. Pillai trace $tr[B(B+W)^{-1})]$

- Pillai’s trace is known for being robust to violations of assumptions, particularly the assumption of equal variance-covariance matrices.

```
# obj=manova(Y~iris.type)
# summary(obj, test="Pillai")
```

4. Roy’s largest root: the largest eigenvalue of BW^{-1}

- Roy’s largest root is typically the most powerful of these four tests when there is only one significant root or when the first root accounts for most of the effect. However, it can be less powerful when there are several smaller roots.

```
# obj=manova(Y~iris.type)
# summary(obj, test="Roy")
```