ST 437/537: Applied Multivariate and Longitudinal Data Analysis

Inference about a Mean Vector: One sample

Arnab Maity
NCSU Department of Statistics
SAS Hall 5240 919-515-1937 amaity[at]ncsu.edu

Introduction

In this lecture we will discuss inference about one population mean vector. Recall by *inference* we mean reaching conclusions concerning a population parameter using information from data.

Suppose we have a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$. Our primary goal is to carry out hypothesis testing and confidence intervals involving $\boldsymbol{\mu}$ in a variety of scenarios.

Let us consider the `LASERI` data in the `ICSNP` library. The dataset contains measurements on 32 variables on 223 healthy Finnish subjects measuring the cardiovascular responses to a passive head-up tilt. See Chapter 7.2 in [Applied Multivariate Statistics with R by Daniel Zelterman. New York: Springer] for more details on the dataset.

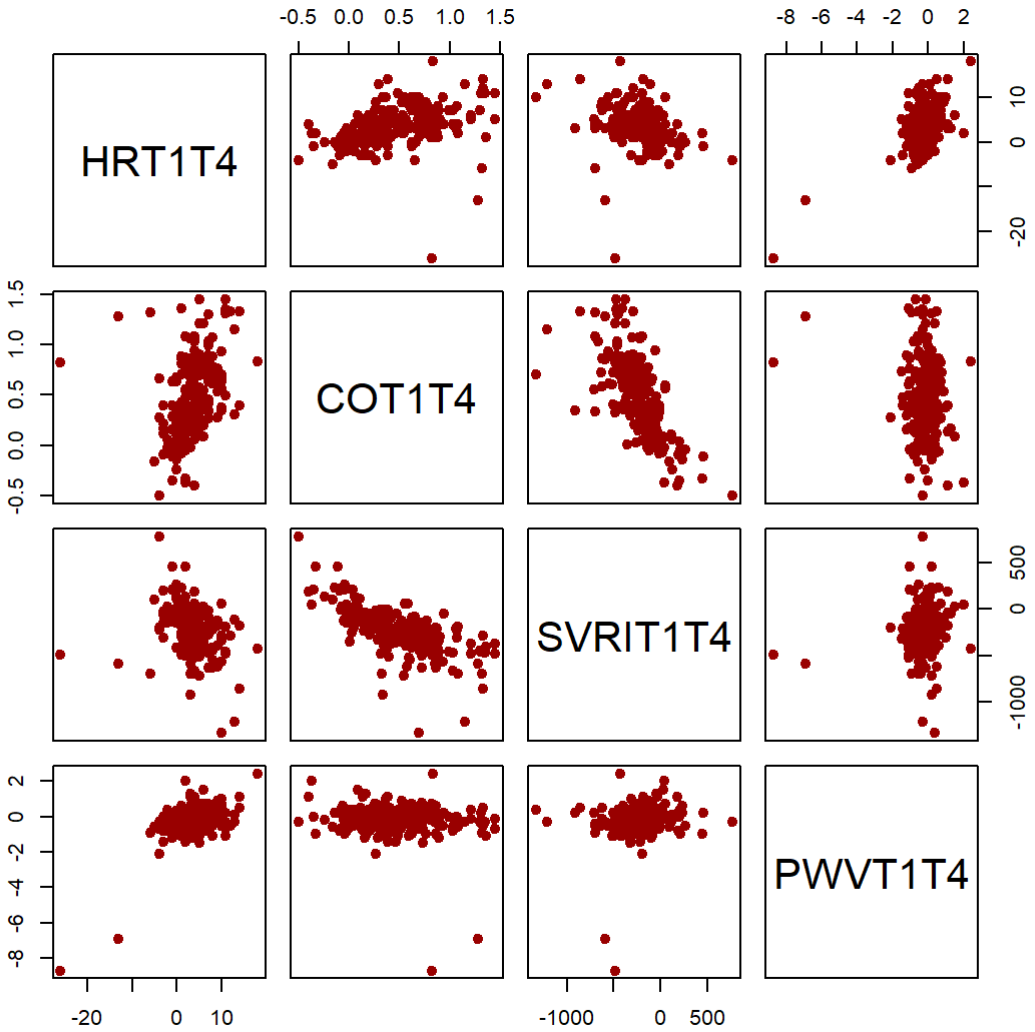
Let us consider the average differences (the pre- and post-tilt values) of `average heart rate` (HRT1T4); `average cardiac output` (COT1T4); `average systemic vascular resistance index` (SVRIT1T4); and `average pulse wave velocity` (PWVT1T4).

```
library(ICSNP)
data("LASERI")
dat <- LASERI[, 25:28]
head(dat)
```

```
##   HRT1T4 COT1T4 SVRIT1T4 PWVT1T4
## 1     11   1.35     -470    -0.3
## 2      0   0.33     -130    -0.2
## 3      1   1.36     -407    -1.1
## 4    -13   1.28     -587    -6.9
## 5      4   1.07     -305     0.0
## 6      8   0.72     -635    -0.5
```

We show a pairs plot of the data below.

```
# Pairs
pairs(dat, pch=19, col="#990000")
```



We want to investigate:

- What are possible values of the true mean difference of each variable?
- How can we formally test whether the true mean diferrence vector is zero or not?

Confidence intervals

Before we study how to answe these questions in the multivariate setting, let us review the univariate case first.

Univariate Analysis: Confidence Interval

Let us first review the method for estimation of an univariate mean. Consider the following question.

Question: *Provide a point estimate of* `HRT1T4` *and provide a 95% confidence interval.*

Since the population variance in this case is unknown, and has to be estimated by the sample variance, we need to use a t-distribution based confidence interval.

Theory: *Let $X_1, X_2, ..., X_n$ be a sample from a normal distribution with mean μ and unknown variance σ^2. An estimator of μ is the sample mean $\bar{X}$. To contruct a confidence interval, we rely on the result:*

$$T := \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim t_{n-1},$$

where s^2 is the sample variance, and t_{n-1} is the Student's t distribution with $n-1$ degrees of freedom. Let $t_{n-1}(\alpha/2)$ be the upper-tail probability corresponding to the t_{n-1} distribution. Then the $100(1-\alpha)\%$ confidence interval (CI) for μ is

$$\left(\bar{X} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}},\ \bar{X} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}\right).$$

The interval described above relies upon the assumption that the random sample $X_1, \ldots, X_n$ is generated from a normal distribution. The assumption of normality can be relaxed when we have a large sample size (typically taken as ≥ 40). We can then apply the central limit theorem, and construct a **large sample interval** using the same formula above **with replacing $t_{n-1}(\alpha/2)$ by $z(\alpha/2)$**.

We first isolate the `HRT1T4` measurements and look at some exploratory plots to assess normality.

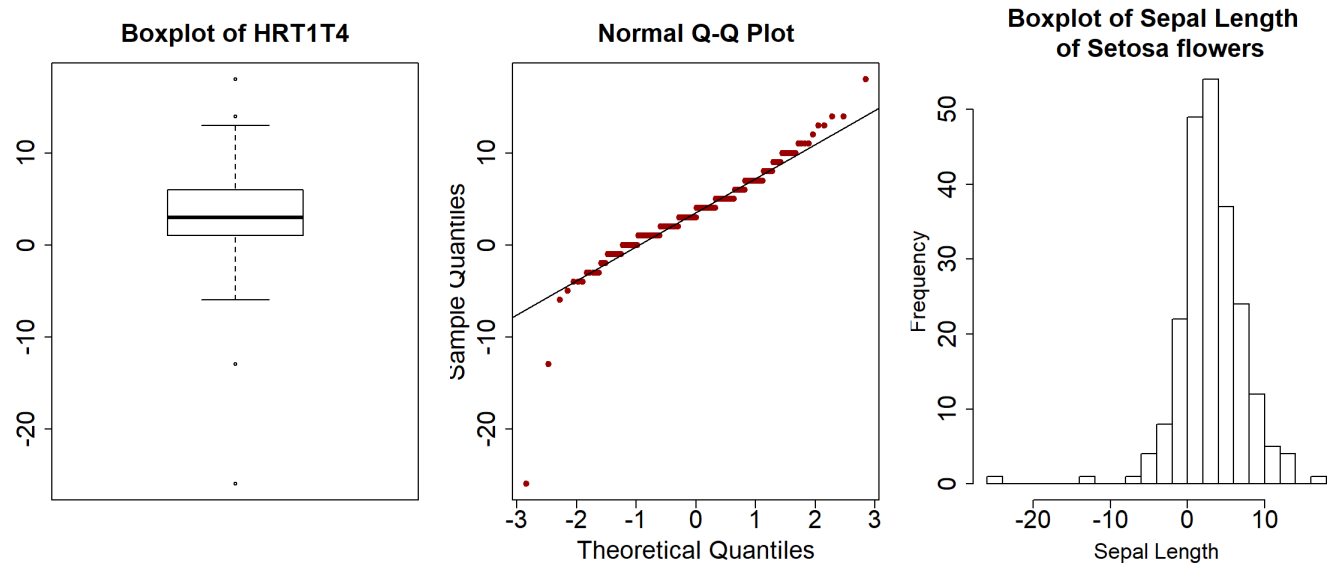
```
# sample size
n = nrow(dat)
n
```

```
## [1] 223
```

```
par(mfrow=c(1,3))
boxplot(dat$HRT1T4, main="Boxplot of HRT1T4", cex.main=2, cex.axis=2)

qqnorm(dat$HRT1T4, cex.main=2, cex.lab=2, cex.axis=2, pch=19, col="#990000")
qqline(dat$HRT1T4)

hist(dat$HRT1T4, cex.main=2, cex.lab=1.7, main = "Boxplot of Sepal Length \n of Setos
a flowers", xlab = "Sepal Length", cex.axis=2, nclass = 20)
```



The data points mostly fall on the straight line in the normal QQ-plot except a few observations close to the tails. However, since $n = 223$, we can still apply the usual t-interval here.

Estimation of the mean and 95% confidence interval is shown below.

```
xbar = mean(dat$HRT1T4)  # sample mean
xbar
```

```
## [1] 3.529148
```

```
stddev = sd(dat$HRT1T4)  # sample standard deviation
stddev
```

```
## [1] 4.335876
```

```
alpha = 0.05  # (1 - Confidence level of CI)
tupper = qt(alpha/2, (n-1), lower.tail=FALSE)  # t critical value

CI = xbar + c(-1,1)* (tupper * stddev/sqrt(n))  # CI
CI
```

```
## [1] 2.956950 4.101346
```

Compare the results with the in-built `t.test()` in R.

```
t.test(dat$HRT1T4, alternative = c("two.sided"), mu = 0,
       conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  dat$HRT1T4
## t = 12.155, df = 222, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.956950 4.101346
## sample estimates:
## mean of x
##  3.529148
```

Based on the results, we can say that `we are 95% confident` that true mean `HRT1T4` is between 2.95 and 4.10. We can also infer (with 95% confidence) that zero is not a possible value for for the mean; since the interval only contains positive values, the mean `HRT1T4` (that is, mean difference between heart rate before and after tilt) is positive.

Multivariate Inference: Simultaneous Confidence Intervals

Suppose we have random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, each of them is a $p \times 1$ vector (in our example, $p = 4$), generated from a p-variate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and unknown covariance matrix $\boldsymbol{\Sigma}$. We want to form confidence intervals for the mean parameters $\mu_1, \ldots, \mu_p$.

It is common to summarize the conclusions for the individual components of μ separately for easy computation and interpretation. One may adopt a naive approach of creating individual confidence intervals, *one variable at a time*. However, such a strategy results in lower that desired coverage. Specifically, suppose we have $p = 2$, that is, we have $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Suppose then we construct one-at-a-time 95% confidence intervals I_1 for μ_1, and I_2 for μ_2. Clearly,

$$P(\mu_1 \in I_1 \text{ and } \mu_2 \in I_2) \leq P(\mu_1 \in I_1) = 95\%.$$

This discrepancy is greater as the number of variables, p, increases. The the *overall* coverage of the intervals will be always less than 95%.

We adopt the principle that the individual intervals hold simultaneously with a specified high probability.

Definition: The intervals $I_1, \ldots, I_p$ are *simultaneous* $100(1-\alpha)\%$ confidence intervals for $\mu_1, \ldots, \mu_p$ if

$$P(I_k \text{ will contain true } \mu_k \text{ simultaneously for every } k = 1, \ldots, p) = 1 - \alpha.$$

Key ideas: Consider linear combinations $\boldsymbol{a}^T\boldsymbol{\mu}$; the $100(1-\alpha)\%$ CI for $\boldsymbol{a}^T\boldsymbol{\mu}$ is

$$\left| \frac{\sqrt{n}(\boldsymbol{a}^T\bar{\boldsymbol{X}} - \boldsymbol{a}^T\boldsymbol{\mu})}{\sqrt{\boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}}} \right| \leq t_{n-1}(\alpha/2).$$

By conveniently choosing $\boldsymbol{a}$ we can obtain individual confidence intervals for the components of $\boldsymbol{\mu}$. However these confidence intervals are not simultaneous; the confidence associated with all the statements taken together is smaller than $(1-\alpha)$.

In order to obtain simultaneous confidence intervals we need to evaluate

$$\max_{\boldsymbol{a}} \left| \frac{\sqrt{n}(\boldsymbol{a}^T\bar{\boldsymbol{X}} - \boldsymbol{a}^T\boldsymbol{\mu})}{\sqrt{\boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}}} \right|$$

which is an equivalent problem to evaluating

$$\max_{\boldsymbol{a}} \left| \frac{n(\boldsymbol{a}^T\bar{\boldsymbol{X}} - \boldsymbol{a}^T\boldsymbol{\mu})^2}{\boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}} \right| = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}).$$

The right hand side term is called the T^2 statistics, and its sampling distribution is $\frac{(n-1)p}{(n-p)}F_{p,n-p}$.

Thus, simultaneously *for all* p-dimensional vectors $\boldsymbol{a}$, the interval

$$\boldsymbol{a}^T\bar{\boldsymbol{X}} \pm \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\ \sqrt{\boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}/n}$$

will contain $\boldsymbol{a}^T\boldsymbol{\mu}$ with probability $1-\alpha$.

Simultaneous confidence intervals: *Let* $\bar{\boldsymbol{X}} = (\bar{X}_1, ..., \bar{X}_p)^T$ *be the sample mean vector. The simultaneous* $100(1-\alpha)\%$ *confidence intervals for* $\mu_1, ..., \mu_p$ *are*

$$For\ \mu_1: \left(\bar{X}_1 - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\frac{S_{11}}{n}},\quad \bar{X}_1 + \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\frac{S_{11}}{n}}\right),$$

$$\vdots$$

$$For\ \mu_p: \left(\bar{X}_p - \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\frac{S_{pp}}{n}},\quad \bar{X}_p + \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\frac{S_{pp}}{n}}\right),$$

where S_{kk} *is the* k*th element of the sample covariance* $\boldsymbol{S}$*.*

Let us look at the `LASERI` example, but now with all the four variables. The following function creates simultaneous intervals of the population mean given a dataset.

```
simul.CI <- function(data, level = 0.95){
  ## Input args
  ##    data: data matrix, each column is one variable, each row is one subject
  ##    level: confidence level ( default 95%)

  # Point estimate of mu; by sample mean
  xbar <- colMeans(data)

  # Sample covariance S
  S <- cov(data)

  # sample size (n) and number of variables (p)
  n <- nrow(data)
  p <- ncol(data)

  # Critical value
  crit <- (n-1)*p/((n-p)*n) * qf(1-level, p, n-p, lower.tail = FALSE)

  # half-width
  H <- sqrt( crit * diag(S) )

  # Intervals with point estimates
  out <- data.frame(Estimate = xbar, Lower = xbar - H, Upper = xbar + H)

  # Return the output
  return(out)
}

## Simulatenous CI for the LASERI data
sint <- simul.CI(dat)
sint
```

```
##              Estimate        Lower         Upper
## HRT1T4      3.5291480    2.6209599    4.43733610
## COT1T4      0.4668161    0.3865204    0.54711191
## SVRIT1T4 -224.3542601 -275.3613948 -173.34712541
## PWVT1T4    -0.2121076   -0.4086666   -0.01554866
```

Compared to the one-at-a-time confidence intervals (the t-intervals), the simulataneous intervals are wider. For example, for the variable `HRT1T4`, we have simultaneous interval is (2.6209599, 4.4373361) but the one-at-a-time t-interval is (2.9569497, 4.1013463]).

When sample size is large, we can relax the assumption of multivariate normality of the data. When n (and $n-p$) is large, one can obtain approximate simulataneous confidence intervals.

Large sample simultaneous intervals: *When n is large, the approximate simultaneous $100(1-\alpha)\%$ confidence intervals for $\mu_1, ..., \mu_p$ are*

$$For\ \mu_1\text{: } \left(\bar{X}_1 - \sqrt{\chi^2_p(\alpha)\frac{S_{11}}{n}},\ \ \bar{X}_1 + \sqrt{\chi^2_p(\alpha)\frac{S_{11}}{n}}\right),$$

$$\vdots$$

$$For\ \mu_p\text{: } \left(\bar{X}_p - \sqrt{\chi^2_p(\alpha)\frac{S_{pp}}{n}},\ \ \bar{X}_p + \sqrt{\chi^2_p(\alpha)\frac{S_{pp}}{n}}\right),$$

where S_{kk} is the kth element of the sample covariance $\boldsymbol{S}$.

The Bonferroni method for multiple correction

As we noticed in the previous section, the simultaneous confidence intervals are (much) wider than the one-at-a-time t-intervals. When p is small, it may be possible to find shorter intervals while still preserving the correct level of confidence. The method is commonly referred to as **Bonferroni correction**.

Key idea: One can preserve an overall level of confidence of $1-\alpha$ by choosing the individual level of confidence $1-\alpha^*$ such that $\alpha^* = \alpha/p$.

The Bonferroni method for multiple correction: *The Bonferroni $100(1-\alpha)\%$ confidence intervals for μ_k, $k = 1, ..., p$ are*

$$\bar{X}_k \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{S_{kk}/n}, \quad k = 1, ..., p$$

where S_{kk} is the kth element of the diagonal of the sample covariance $\boldsymbol{S}$.

Below, we see the comparison of the three types of intervals, one-at-a-time, simulataneous and Bonferroni, for `HRT1T4`.

```
##              Lower Upper
## One-at-a-time 2.957 4.101
## Bonferroni    2.798 4.260
## Simultaneous  2.621 4.437
```

Clearly, the one-at-a-time interval is narrower (and also has a smaller overall confidence level); the simultaneous confidence interval is widest, and the Bonferroni interval is in between.

Hypothesis testing

Let us take up the problem of hypothesis testing in a multivariate one-sample situation. First we review hypothesis testing in the univariate case.

Univariate Inference: Hypothesis testing

Consider the following question.

Question: Test wheather true mean of `HRT1T4` is zero or not. Specifically, suppose μ is the true mean of `HRT1T4`.Then we want to test $H_0: \mu = 0$ vs. $H_a: \mu \neq 0$.

Since the population variance in this case is unknown, and has to be estimated by the sample variance, we need to use a t-distribution based test. In general we want to test $H_0: \mu = \mu_0$ vs. $H_a: \mu \neq \mu_0$, where μ_0 is a known constant.

Theory: *Let $X_1, X_2, \ldots, X_n$ be a sample from a normal distribution with mean μ and unknown variance σ^2. An estimator of μ is the sample mean $\bar{X}$. To contruct a test, we rely on the result:*

$$T := \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim t_{n-1},$$

where s^2 is the sample variance, and t_{n-1} is the Student's t distribution with $n-1$ degrees of freedom. Let $t_{n-1}(\alpha/2)$ be the upper-tail probability corresponding to the t_{n-1} distribution.

We reject $H_0: \mu = \mu_0$ in favor of $H_a: \mu \neq \mu_0$, if

$$\left| \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right| > t_{n-1}(\alpha/2);$$

we fail to reject H_0 otherwise.

In case we have a large sample size (≥ 40), we can relax the normality assumption, and perform a z-test by applying the same rule above but replacing $t_{n-1}(\alpha/2)$ by $z(\alpha/2)$.

For the `HRT1T4` variable, we can perform the t-test for $H_0: \mu = 0$ vs $H_a: \mu \neq 0$ at level $\alpha = 0.05$ as follows.

```
mu0 <- 0 ## value of mu under H0
n <- nrow(LASERI) # sample size
alpha <- 0.05 # significance level of the test
t.crit <- qt(p = alpha/2, df = n-1, lower.tail = F) # t critical value

# test statistic
xbar = mean(dat$HRT1T4)  # sample mean
stddev = sd(dat$HRT1T4) # sample sd
test.stat <- (xbar - mu0)/(stddev/sqrt(n))

# perform the test
data.frame(test.stat, t.crit)
```

```
##   test.stat   t.crit
## 1  12.15473 1.970707
```

We can see that the absolute value of the test statiatic is larger than the t critical value. Thus we reject $H_0: \mu = 0$ and conlude that μ is not zero.

We can also use the R fucntion `t.test()`.

```
t.test(LASERI$HRT1T4)
```

```
##
##  One Sample t-test
##
## data:  LASERI$HRT1T4
## t = 12.155, df = 222, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.956950 4.101346
## sample estimates:
## mean of x
##  3.529148
```

The p-value is less than $\alpha = 0.05$, and thus we reach the same conclusion as above.

Multivariate Inference: Testing for a Mean Vector

Suppose we have random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, each of them is a $p \times 1$ vector (in our example, $p = 4$), generated from a p-variate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and unknown covariance matrix $\boldsymbol{\Sigma}$. We want to test

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_a: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

where $\boldsymbol{\mu}_0$ is a known fixed vector. We can perform the hypothesis testing $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ using the Hotelling's T^2 statistic.

Theory: *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be the observed data. We will use the test statistic $T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$. Recall that when the null hypothesis is true ($\boldsymbol{\mu} = \boldsymbol{\mu}_0$) then*

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

So we would reject H_0 at level α if

$$\frac{n(n-p)}{(n-1)p} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{s}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) > F_{p, n-p}(\alpha).$$

In our example, we have four variables. The **Hotelling's T^2** test can be done in R as follows.

```
# dataset with four variables
dat <- LASERI[, 25:28]

# Hotelling's T2 test
library(ICSNP)
HotellingsT2(X = dat)
```

```
##
##  Hotelling's one sample T2-test
##
## data:  dat
## T.2 = 101.67, df1 = 4, df2 = 219, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to c(0,0,0,0)
```

Based on the p-value, we strongly reject H_0.

We can also manually perform the test.

```
p = ncol(dat) # number of variables
n = nrow(dat) # sample size
xbar = colMeans(dat) # sample mean vector
S = cov(dat) # sample covariance matrix
mu0 = c(0,0,0,0) # value of mean under null
alpha = 0.05 # significance level
## Test statistic
T2 <- n*(n-p)/(p*(n-1)) * t(xbar-mu0) %*% solve(S) %*% (xbar-mu0)
T2
```

```
##          [,1]
## [1,] 101.6741
```

```
## Critical value
qf(alpha, p, (n-p), lower.tail=FALSE)
```

```
## [1] 2.41287
```

Since T^2 value is larger than the critical value, we reject H_0.

When we have a large sample size, n, we can again relax the normality assumption and conduct an approximate test:

Reject H_0 at level α if

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{s}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \chi^2_p(\alpha).$$

When sample size is small, in the lack of information about the parent distribution one needs to check the multivariate normality assumption using the methods mentioned in the previous lectures. In the absence of normality, one can apply nonparametric rank-based procedures. We ddemonstrate one such procedure, named Tyler's angle rank test, is demonstrated below.

```
library(ICSNP)
HP.loc.test(dat)
```

```
##
##  TYLER ANGLES RANK TEST
##
## data:  dat
## Q.W = 108.81, df = 4, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to c(0,0,0,0)
```

This test is described in **[Ian, M. and Paindaveine, D. (2002b), Randles' interdirections or Tyler's angles?, In Y. Dodge, Ed. Statistical data analysis based on the L1-norm and related methods, 271–282.] (https://link.springer.com/chapter/10.1007/978-3-0348-8201-9_23)**.

Main page: **ST 437/537: Applied Multivariate and Longitudinal Data Analysis (https://maityst537.wordpress.ncsu.edu/)**