Multivariate Analysis Lecture 7: Hotelling's T2

Zhaoxia Yu
Professor, Department of Statistics

2023-04-25

Outline of Lecture 07

- Review of Wishart and the Hotelling's T^2 distribution for one-sample problems
- Examples of one-sample Hotelling's T^2
- Two-sample Hotelling's T^2
- Examples of two-sample Hotelling's T^2
- The multivariate normality (MVN) assumption

Section 1

Review

Subsection 1

Wishart Distribution

Definition of Wishart Distribution

- A Wishart distribution can be defined in the following way
- Let $\mathbf{W}$ be a $p \times p$ random matrix. We say $\mathbf{W}$ follows $Wishart_p(k, \mathbf{\Sigma})$ if $\mathbf{W}$ can be written as $\mathbf{W} = \mathbf{X}^T\mathbf{X}$ where $\mathbf{X}$ denotes the random matrix formed by a random sample of size k from MVN $N(\mathbf{0}, \mathbf{\Sigma})$.
- The definition indicates that if we have a random sample $\mathbf{X}_1, \cdots \mathbf{X}_k$ from $N(\mathbf{0}, \mathbf{\Sigma})$, then $\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{k} \mathbf{X}_i\mathbf{X}_i^T \sim Wishart_p(k, \mathbf{\Sigma})$.
- Remark: $E[\mathbf{W}] = k\Sigma$.

Wishart vs Chi-squared

- Wishart: If $\mathbf{X}_1, \cdots \mathbf{X}_k \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Sigma})$, then

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{k} \mathbf{X}_i \mathbf{X}_i^T \sim Wishart_p(k, \mathbf{\Sigma}), \text{ where } \mathbf{X}_{k\times p} = \begin{pmatrix} X_1^T \\ \vdots \\ X_k^T \end{pmatrix}$$

- Chi-squared: If $X_1, \cdots, X_k \overset{iid}{\sim} N(0, 1)$, then

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{k} X_i^2 \sim \chi_k^2, \text{ where } \mathbf{X}_{k\times 1} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$$

Wishart vs Chi-squared (continued)

- When $p = 1$,

$$W = \sum_{i=1}^{k} X_i^2 = \sigma^2 \sum_{i=1}^{k} \left(\frac{X_i}{\sigma} \right)^2 \sim \sigma^2 \chi_k^2$$

The Sample Covariance Matrix

- Let $\mathbf{X}_1, \cdots \mathbf{X}_n$ be a random sample from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The $\mathbf{X}_{n \times p}$ follows a matrix normal distribution:

$$\mathbf{X} \sim N(\mathbf{1}_n \otimes \boldsymbol{\mu}^T, \boldsymbol{\Sigma}, \mathbf{I}_n)$$

- We have shown that

$$(n-1)\mathbf{S} \sim Wishart_p(n-1, \boldsymbol{\Sigma})$$

Subsection 2

Hotelling's T^2

Definition of Hotelling's T^2

- Hotelling generalized the student's t, which is for univarite, to Hotelling's T2, which is the multivariate version
- Definition. We say a random variable follows Hotelling's $T^2_{p,\nu}$ if the random variable can be written as $\mathbf{Z}^T \left(\frac{W}{\nu}\right)^{-1} \mathbf{Z}$ where
  1. $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{\Sigma})$
  2. $\mathbf{W} \sim W_p(\nu, \mathbf{\Sigma})$
  3. $\mathbf{Z} \perp \mathbf{W}$

Section 2

One-Sample Hotelling T^2

One-Sample Hotelling T^2

- Let $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ be a random sample from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- The sample mean vector and sample covariance matrix are denoted by $\bar{\mathbf{X}}$ and $\mathbf{S}$, respectively.
- The null hypothesis of interest $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$
- The one-sample Hotelling T^2 is defined as

$$T^2 = (\hat{\mu} - \mu_0)^T \left(Cov(\hat{\mu})\right)^{-1} (\hat{\mu} - \mu_0)$$

- We have shown that $T^2 \sim T^2_{p,n-1}$ when $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$.

Hotelling's T^2 Distribution vs F Distribution

Hotelling's T^2

Claim: $T^2_{p,\nu} \sim \frac{\nu p}{\nu+1-p} F_{p,\nu+1-p}$.

For the T^2 statistic, we have $T^2 \overset{H_0}{\sim} \frac{(n-1)p}{n-p} F_{p,n-p}$. We reject H_0 at significance level α when $T^2 > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$.

Corollary.

$$\frac{n-p}{p}(\bar{X}-\mu_0)^T(\hat{\Sigma})^{-1}(\bar{X}-\mu_0) \overset{H_0}{\sim} F_{p,n-p}$$

where$\hat{\Sigma} = \frac{1}{n} X^T H X = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T = \frac{(n-1)S}{n}$.

Subsection 1

A Simulation Study

Understand the Wishart Distribution

- Recall that if $W \sim Wishart_p(k, \mathbf{\Sigma})$, then $E[\mathbf{W}] = k\Sigma$.

```
library(MASS)
p=2; n=5; B=1000; rho=0.7
Sigma=diag(1+rho, p, p) - matrix(rho, p, p)
wmat.array=array(0, c(B, p, p)) #wishart-distributed
for(b in 1:B){
  X=mvrnorm(n, rep(0,p), Sigma)
  wmat.array[b,,]=(n-1)*cov(X)}
apply(wmat.array, c(2,3), mean)
```

```
##           [,1]      [,2]
## [1,]  4.135352 -2.859370
## [2,] -2.859370  4.036825
```

```
Sigma*(n-1)
```

```
##      [,1] [,2]
## [1,]  4.0 -2.8
## [2,] -2.8  4.0
```

Write an R function to conduct Hotelling's T^2

- There is no R base function for conducting Hotelling's T^2 test
- We will write an R function

```
#Hotelling's T^2 for testing H0: mu=mu0 vs mu != mu0
Hotelling.T2.1sample=function(X, mu0)
{
  n=dim(X)[1]
  p=dim(X)[2]
  X.bar=colMeans(X)
  X.S=cov(X)
  T2=n*t(X.bar-mu0)%*%solve(X.S)%*%(X.bar-mu0)
  p.value=1-pf(T2/((n-1)*p/(n-p)),p,n-p)
  return(list(X.bar=X.bar, X.cov=X.S, T2=T2, p.value=p.value))
}
```

Example of Multivariate One-Sample Problem: Protein Intake

- For the protein intake data, it might be more interesting to estimate the means than conducting hypothesis testing
- Suppose we are interested in estimating the means of the daily protein intake from different sources

```
library(MASS)#the library "MASS" is required
my.cov=4*(diag(4) + 0.3* rep(1,4)%o%rep(1,4))
n=60;p=4
my.mean=8*c(3,2,1,1)
eigen(my.cov)#to check whether the cov matrix is p.d.
```

```
## eigen() decomposition
## $values
## [1] 8.8 4.0 4.0 4.0
##
## $vectors
##      [,1]       [,2]       [,3]       [,4]
## [1,] -0.5  0.8660254  0.0000000  0.0000000
## [2,] -0.5 -0.2886751 -0.5773503 -0.5773503
## [3,] -0.5 -0.2886751 -0.2113249  0.7886751
## [4,] -0.5 -0.2886751  0.7886751 -0.2113249
```

Example of Multivariate One-Sample Problem: Protein Intake

- Estimate the mean vector using the sample mean vector
- Estimate covariance of the sample mean vector. Recall that $cov(\bar{\mathbf{X}}) = \frac{\mathbf{\Sigma}}{n}$

```
set.seed(1)
x=mvrnorm(n, mu=my.mean, Sigma=my.cov)
protein=as.matrix(data.frame(meat=x[,1],dairy=x[,2],
                             veg=x[,3], other=x[,4]))
colMeans(protein)
```

```
##      meat     dairy       veg     other
## 24.034032 15.928361  7.660490  7.738634
```

```
cov(protein)/n
```

```
##             meat       dairy         veg       other
## meat  0.07159404 0.013584596 0.018824131 0.009220700
## dairy 0.01358460 0.073421655 0.005829816 0.003895500
## veg   0.01882413 0.005829816 0.086176323 0.009828535
## other 0.00922070 0.003895500 0.009828535 0.075478822
```

Example of Multivariate One-Sample Problem: Protein Intake

- Use Hotelling's T^2 to quantify uncertainties. Recall that

$$T^2 = (\bar{\mathbf{X}} - \boldsymbol{\mu})^T \left(Cov(\bar{\mathbf{X}})\right)^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

where $Cov(\bar{\mathbf{X}}) = \frac{\mathbf{S}}{n}$.

- The result indicates that

$$Pr[(\bar{\mathbf{X}}-\boldsymbol{\mu})^T \left(Cov(\bar{\mathbf{X}})\right)^{-1} (\bar{\mathbf{X}}-\boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}] = 1-\alpha$$

- Thus, a $(1-\alpha)100\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\mu : (\bar{\mathbf{X}} - \boldsymbol{\mu})^T \left(Cov(\bar{\mathbf{X}})\right)^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}\}$$

Example of Multivariate One-Sample Problem: Protein Intake

- Confidence intervals, which are in the form of

$$\text{estimate} \pm \text{critical value} \times \text{standard error}$$

  are often preferred
- If there is only one parameter of interest, we can construct a C.I. using t-distribution, just as in univariate analysis
- Example. What is the mean protein intake? We have discussed the problem in Lecture 04.
  - Lecture 04: we constructed a large-sample C.I. by using 1.96 as the critical value. (See the protein intake example)
  - This lecture: we construct a C.I. by using $t_{n-1,1-\frac{\alpha}{2}}$ as the critical value

$$\bar{X}_{(j)} \pm t_{n-1,1-\frac{\alpha}{2}} \sqrt{\frac{s^2_{X_{(j)}}}{n}} \text{ for } j = 1, \cdots, p.$$

Example of Multivariate One-Sample Problem: Protein Intake

- The confidence region has exactly $(1-\alpha)100\%$ confidence; however
- In many situations, we would like to construct confidence intervals for multiple parameters
- This is known as simultaneous intervals
- If we use the conventional method to construct individual C.I.s, we will have lower than $(1-\alpha)100\%$ confidence to cover all the parameters simultaneously

Example of Multivariate One-Sample Problem: Protein Intake

- Some linear algebra result ensures that the following method gives $(1-\alpha)100\%$ confidence to cover all linear combinations of the parameters (in the form of $a^T\boldsymbol{\mu}$) simultaneously

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

- Example, consider the individual means $\mu_j, j = 1, \cdots, 4$ in the protein intake example, the following intervals are 95% simultaneous C.I. for the mean protein intake from the four sources

Example of Multivariate One-Sample Problem: Protein Intake

```
#sample means
colMeans(protein)
```

```
##      meat     dairy       veg     other
## 24.034032 15.928361  7.660490  7.738634
```

```
#standard errors
sqrt(diag(cov(protein)/n))
```

```
##      meat     dairy       veg     other
## 0.2675706 0.2709643 0.2935580 0.2747341
```

```
#critical value based on T2
sqrt((n-1)*p/(n-p)*qf(0.95, p, n-p))
```

```
## [1] 3.269537
```

```
#lower bounds
low.bound=colMeans(protein) - sqrt((n-1)*p/(n-p)*qf(0.95, p, n-p)) *sqrt(diag(cov(protein)/n))
#upper bounds
up.bound=colMeans(protein) + sqrt((n-1)*p/(n-p)*qf(0.95, p, n-p)) *sqrt(diag(cov(protein)/n))
```

Example of Multivariate One-Sample Problem: Protein Intake

```
#put everything into a table
data.frame(lower=low.bound, mean=colMeans(protein),
           upper=up.bound)
```

```
##           lower      mean     upper
## meat  23.159200 24.034032 24.908864
## dairy 15.042433 15.928361 16.814289
## veg    6.700691  7.660490  8.620288
## other  6.840381  7.738634  8.636887
```

Example of Multivariate One-Sample Problem: Protein Intake

- Three choices of criticl values
  - unadjusted: $t_{n-1,1-\alpha/2}$. Should NOT be used if multiple linear functions need to be estimated
  - T^2: $\sqrt{\frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}}$
  - Bonferroni's correction: simply replace α with α/k where k is the total number of linear functions of the mean parameters: $t_{n-1,1-\alpha/(2k)}$
- Example: the critical values for the individual means from four protein sources

Example of Multivariate One-Sample Problem: Protein Intake

```
#unadjusted, shouldn't be used when constructing simultaneous C.I.s
qt(1-0.05/2, n-1)
```

```
## [1] 2.000995
```

```
#T^2
sqrt((n-1)*p/(n-p)*qf(0.95, p, n-p))
```

```
## [1] 3.269537
```

```
#Bonferroni correction
qt(1-0.05/p/2, n-1)
```

```
## [1] 2.576588
```

Section 3

Two-Sample Hotellings T^2

One-Sample vs Two-Sample

- In the one-sample problem, the goal is to make inference of
  - univariate: a population mean (one-sample t-test problem) or
  - multivariate: a population mean vector (one-sample Hotelling T^2 problem)
- In the two-sample problem
  - univariate: compare two population means
  - multivariate: compare two population mean vectors

Univariate Two-Sample Problems

- Two independent samples
  - Sample 1 is from population 1:

$$X_{11}, \cdots, X_{1,n_1} \overset{iid}{\sim} N(\mu_1, \sigma^2)$$

  - Sample 2 is from population 2:

$$X_{21}, \cdots, X_{2,n_1} \overset{iid}{\sim} N(\mu_2, \sigma^2)$$

- Null hypothesis: $H_0 : \mu_1 = \mu_2$

Univariate Two-Sample Problems

- Pooled sample variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where

$$s_i^2 = \frac{\sum_{j=1}^{n_i} X_{ij}^2 - (\sum_{j=1}^{n_i} X_{ij})^2 / n_i}{n_i - 1}$$

- Two-sample t-statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

- Null distribution: $t \overset{H_0}{\sim} t_{n_1+n_2-2}$.

Multivarite Two-Sample Problems

- Two independent samples
  - Sample 1 is from population 1:

$$\mathbf{X}_{11}, \cdots, \mathbf{X}_{1,n_1} \overset{iid}{\sim} N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

  - Sample 2 is from population 2:

$$\mathbf{X}_{21}, \cdots, \mathbf{X}_{2,n_2} \overset{iid}{\sim} N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

- Null and alternative hypotheses: $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

Multivarite Two-Sample Problems

- SPooled sample variance

$$\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

where

$$\mathbf{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)'$$

- Two-sample Hotelling's T^2

$$T^2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T\{\mathbf{S}_p(\frac{1}{n_1} + \frac{1}{n_2})\}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$

- Null distribution:

$$T^2 \overset{H_0}{\sim} \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1}F_{p, n_1+n_2-p-1}$$

Multivariate Two-Sample Problems: Write an R Function

- No existing base function in R.

```
Hotelling.T2.2sample=function(X, Y){
  n=dim(X)[1]; m=dim(Y)[1]; p=dim(X)[2]
  if(p!= dim(Y)[2]) return("Error: the dimensions of X and Y are not the same")
  X.bar=colMeans(X); Y.bar=colMeans(Y)
  X.S=cov(X); Y.S=cov(Y)
  pooled.S=((n-1)*X.S+(m-1)*Y.S)/(m+n-2)
  T2=t(X.bar-Y.bar)%*%solve((1/n+1/m)*pooled.S)%*%(X.bar-Y.bar)
  p.value=1-pf(T2/((n+m-2)*p/(n+m-1-p)),p,n+m-1-p)
  return(list(X.bar=X.bar, Y.bar=Y.bar, T2=T2, p.value=p.value))}
```

Multivariate Two-Sample Problems: Wite an R Function

- The built-in function “t.test” serves a dual function for univariate analyis
- We will write a dual function Hotelling.T2

```
Hotelling.T2=function(X, Y=NULL, mu0=NULL)
{
 if(is.null(Y) && is.null(mu0) )
   return("Error: mu0 is not specified")
 if(!is.null(X) && !is.null(mu0))
   obj=Hotelling.T2.1sample(X, mu0)
 if(!is.null(X) && !is.null(Y))
   obj=Hotelling.T2.2sample(X,Y)
 return(obj)
}
```

Multivariate Two-Sample Problems: Iris setosa vs versicolor

```
Hotelling.T2.2sample(iris[1:50,1:4], iris[51:100,1:4])
```

```
## $X.bar
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.006        3.428        1.462        0.246
##
## $Y.bar
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.936        2.770        4.260        1.326
##
## $T2
##          [,1]
## [1,] 2580.839
##
## $p.value
##      [,1]
## [1,]    0
```

Multivariate TWo-Sample Problems: Example

```
Hotelling.T2(iris[1:50,1:4], iris[51:100,1:4])
```

```
## $X.bar
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.006        3.428        1.462        0.246
##
## $Y.bar
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.936        2.770        4.260        1.326
##
## $T2
##          [,1]
## [1,] 2580.839
##
## $p.value
##      [,1]
## [1,]    0
```

Linear Functions of Differences: Iris Setosa vs Versicolor

- We might be interested in the difference between iris setosa and versicolor in the four features
- Because we are interested all the four features, we do need to construct simultaneous C.I.s for the four features. Two methods to find critical values with adjustment for multiple C.I.s:
  - Method 1 - T^2:

$$\sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1,1-\alpha}}$$

  - Method 2- Bonferroni's correction by replacing α with α/k, i.e., use the following critical value

$$t_{n_1+n_2-2,1-\alpha/(2k)}$$

Linear Functions of Differences: Iris Setosa vs Versicolor

```
n1=n2=50; p=4
mean1=matrix(colMeans(iris[1:50,1:p]), p, 1)
mean2=matrix(colMeans(iris[51:100,1:p]), p, 1)
mean.diff = mean1-mean2
S1=cov(iris[1:50,1:p]); S2=cov(iris[51:100,1:p]);
Sp=( (n1-1)*S1+(n2-1)*S2 )/ (n1+n2-2)
```

Linear Functions of Differences: Iris Setosa vs Versicolor

- Method 1: T^2

```
cv=sqrt((n1+n2-2)*p/(n1+n2-p-1)*qf(1-0.05, p, n1+n2-p-1 ))
round(data.frame(diff=mean.diff, se=sqrt(diag((1/n1+1/n2)*Sp) ),
CI.lower=mean1-mean2-qt(1-0.05/(2*p), n1+n2-2)*sqrt(diag((1/n1+1/n2)*Sp) ),
CI.upper=mean1-mean2+qt(1-0.05/(2*p), n1+n2-2)*sqrt(diag((1/n1+1/n2)*Sp) ) ), 3)
```

```
##                diff    se CI.lower CI.upper
## Sepal.Length -0.930 0.088   -1.155   -0.705
## Sepal.Width   0.658 0.070    0.481    0.835
## Petal.Length -2.798 0.071   -2.978   -2.618
## Petal.Width  -1.080 0.032   -1.161   -0.999
```

Linear Functions of Differences: Iris Setosa vs Versicolor

- Method 1 Bonferroni

```
cv=sqrt((n1+n2-2)*p/(n1+n2-p-1)*qf(1-0.05, p, n1+n2-p-1 ))
round(data.frame(diff=mean.diff, se=sqrt(diag((1/n1+1/n2)*Sp) ),
CI.lower=mean1-mean2-qt(1-0.05/(2*p), n1+n2-2)*sqrt(diag((1/n1+1/n2)*Sp) ),
CI.upper=mean1-mean2+qt(1-0.05/(2*p), n1+n2-2)*sqrt(diag((1/n1+1/n2)*Sp) ) ), 3)
```

```
##                diff    se CI.lower CI.upper
## Sepal.Length -0.930 0.088   -1.155   -0.705
## Sepal.Width   0.658 0.070    0.481    0.835
## Petal.Length -2.798 0.071   -2.978   -2.618
## Petal.Width  -1.080 0.032   -1.161   -0.999
```

Section 4

Assess MVN

The assumption of MVN

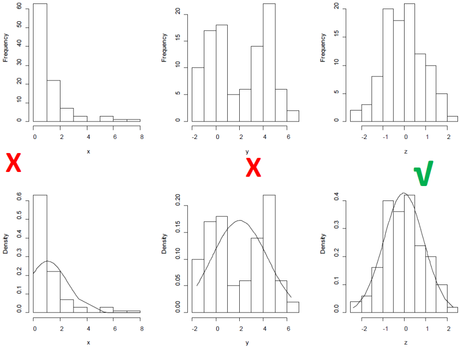
- We assume each observation $\mathbf{X}_i$ follows a MVN
- Assessing the assumption of multivariate normality is more difficult than assessing the assumption of normality (univariate)
- This is because univariate normality does not guarantee multivariate normality. Typically, we look at the following two items:
- It is difficult to examine joint normality in more than 2d. In practice, we do 1d and 2d
    - Marginal normality
    - Are pairs of variables show elliptical contours?
- Are there outliers in the data?

Assess Marignal Normality

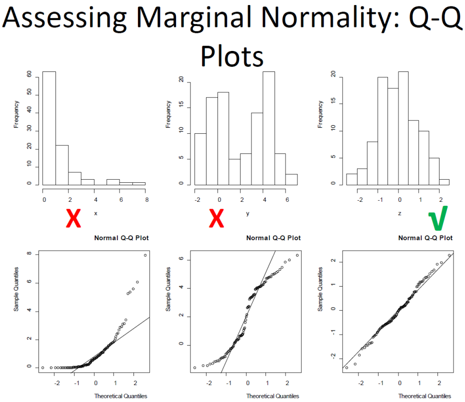
- Useful visual tools:
  - histogram
  - QQ plot
  - scatter plot
- Less useful tools (formal tests)
  - Kolmogorov-Smironov test
  - Shapiro-Wilk test (correlation coefficient between data and normal scores)

Histograms

Assessing Marginal Normality: Histogram

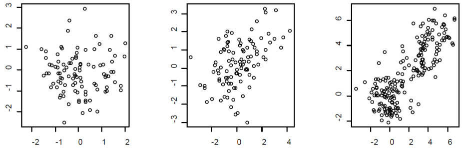


QQ plots


Assessing Marginal Normality: Q-Q Plots
Frequency
x
y
z
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

Bivariate Scatter Plots

Assessing Bivariate Normality



√ √ X

Large-Sample Results

- Multivariate CLT

$$
\begin{aligned}
& \sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{\mathbf{D}} N(\mathbf{0}, \boldsymbol{\Sigma}) \\
\Rightarrow \quad & n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \to \chi^2_p
\end{aligned}
$$

- When $n - p$ is large, we replace $\frac{(n-1)p}{n-p} F_{p,n-p}$ with χ^2_p
- When $n_1 - p$ and $n_2 - p$ are large, we replace $\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}$ with χ^2_p

Assignment 2: Due on Monday, May 1st

- Problem 1: Choose a $3 - by - 3$ covariance matrices with non-zero covariances. For each covariance matrix, simulate 1,000 data sets from a trivariate normal distribution.
  0. Hints:
     - Hint 1: the R library MASS provides a function to generate a random sample from a multivariate normal distribution.
     - Hint 2: Make sure that the covariance matrix you choose is positive definite. You can compute the eigenvalues by the "eigen" function in R and and check whether all the eigenvalues are positive.
  1. Try to make sense of the covariance matrix by examining the pairwise scatter plots using the data you simulate.
  2. During the simulation, you will generate 1,000 Wishart distributed random matrices. Calculate the trace for each of them. Explain what distribution the traces should follow and examine their histogram.
- Problem 2: Find a good data example to conduct a