Multivariate Normal Distribution & Multivariate Inference

Jan Graffelman[1]

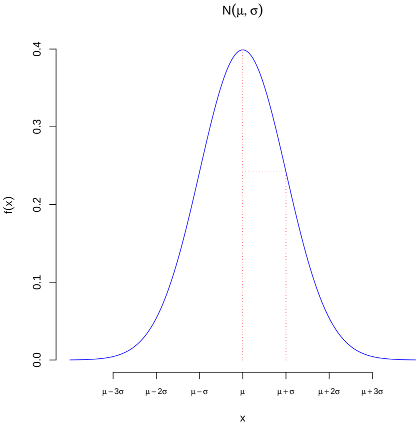
[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

jan.graffelman@upc.edu

March 19, 2020

Contents

Multivariate Normal Distribution



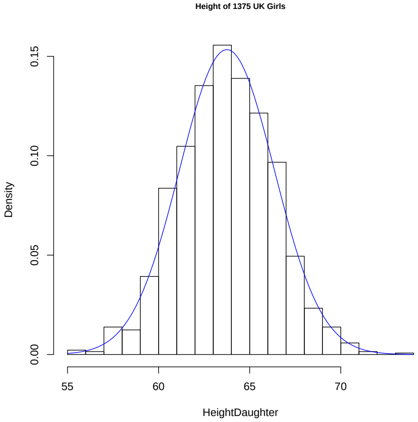
$$X \sim N(\mu, \sigma)$$

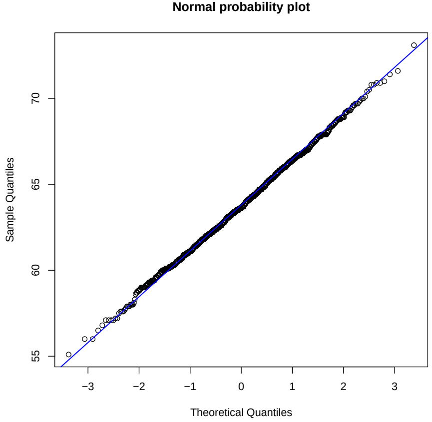
$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E(X) = \mu$$

$$V(X) = \sigma^2$$

Some normal data (Height UK girls in 1903)

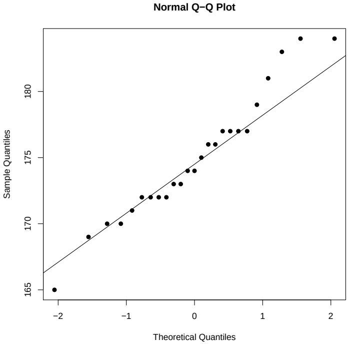




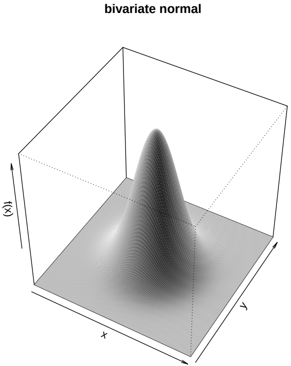
	N	N*	Mean	Stdev	Med	Q1	Q3	Min	Max
Height	1375	0	63.751	2.6	63.6	62	65.6	55.1	73.1

Normal probability plot

i	1	2	3	4	5	6	7	8	9	...	25
Height	172	174	183	175	176	184	177	169	172	...	172
Sorted	165	169	170	170	171	172	172	172	172	...	184
Rank i	1	2	3	4	5	6	7	8	9	...	25
$\frac{i-0.5}{n}$	0.02	0.06	0.10	0.14	0.18	0.22	0.26	0.30	0.34	...	0.98
$z_{(i-0.5)/n}$	-2.05	-1.55	-1.28	-1.08	-0.92	-0.77	-0.64	-0.52	-0.41	...	2.05



Some bivariate normal distributions



Density multivariate normal

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Exponent univariate normal

$$-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 = -\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)$$

Exponent multivariate normal

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Multivariate normal distribution

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Parameters:

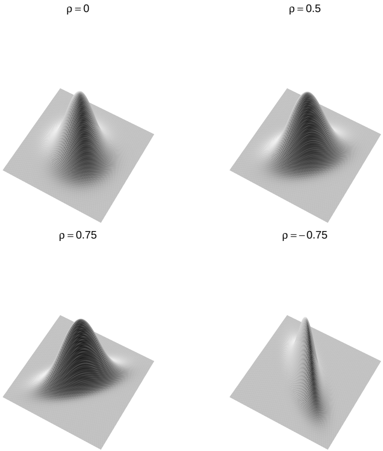
- Population mean vector:

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$$

- Population variance-covariance matrix:

$$Cov(\mathbf{X}) = \boldsymbol{\Sigma}_{p\times p} = E\left((\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\right) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Bivariate normal distribution



Parameter estimation

Maximum likelihood estimator for $\boldsymbol{\mu}$:

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$$

Maximum likelihood estimator for $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \mathbf{S}_n$$

In practice, $\mathbf{S}_{n-1}$ is often used to estimate $\boldsymbol{\Sigma}$:

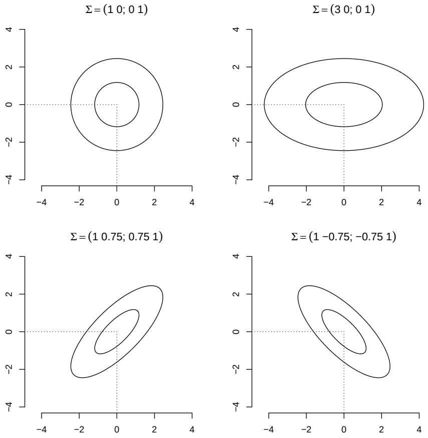
$$\mathbf{S}_{n-1} = \frac{n}{n-1} \mathbf{S}_n$$

Some Properties of MVN random variates

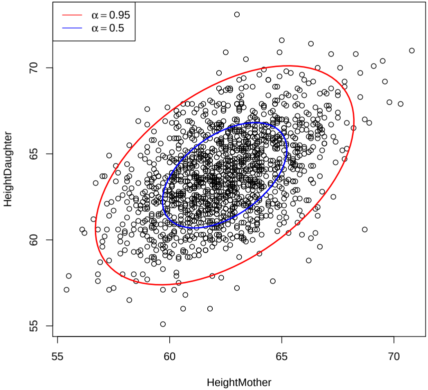
Let $\mathbf{X}$ be a $p \times 1$ random vector, and $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- Linear combinations of the components of $\mathbf{X}$ are normally distributed.
- Basic result: if $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A} q \times p$, then $\mathbf{AX} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$
- Subsets of components have a (multivariate) normal distribution.
- Components with covariance zero $\Leftrightarrow$ components are independent.
- Conditional distributions of components are (multivariate) normal.

Contours of normal densities (0.50 and 0.95)



Contours for empirical data



Making contours in R

```
X <- read.table("http://www-eio.upc.es/~jan/data/MVA/PearsonLeeheights.txt",
                 header=TRUE)

plot(X)

m <- colMeans(X)
m

S <- cov(X)
S

Z1 <- ellipse(S,level=0.95,centre=m)
points(Z1,type="l",col="red",lwd=2)

Z2 <- ellipse(S,level=0.50,centre=m)
points(Z2,type="l",col="blue",lwd=2)
```

Assessing multivariate normality

Some basic ideas:

- Individual variables (marginal distributions) should have bell-shaped (normal) histograms
- Bivariate scatterplots should have clouds of points with an elliptic shape
- Some outliers can be expected, in particular in larger samples

χ^2 plot for multivariate normality

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2_p$$

The ellipsoid traced by $\mathbf{x}$ described by

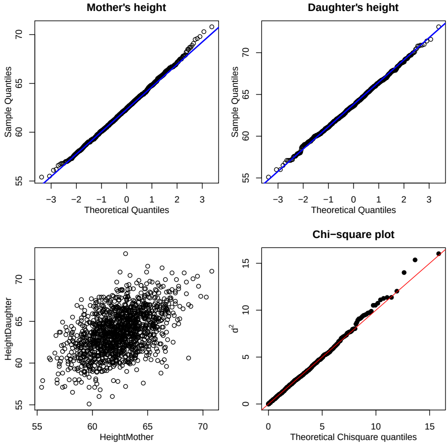
$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi^2_p(1-\alpha)$$

should contain $100 \cdot (1-\alpha)\%$ of the observations.

For sample data:

1. Calculate $d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$
2. Order the distances from small to large
3. Calculate the rank $(i - \frac{1}{2})/n$
4. Calculate corresponding quantiles q_i according to a χ^2_p distribution.
5. Plot (d_i^2, q_i)
6. Compare with a reference line with intercept 0 and slope 1

Example χ^2 plot for multivariate normality



Inference on a mean vector

Univariate test on a population mean

- $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$
- Statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$
- $100 \cdot (1-\alpha)$ confidence interval: $CI_{1-\alpha}(\mu) = \overline{x} \pm t_{n-1,\alpha/2} s/\sqrt{n}$

Note that

$$t^2 = \frac{(\bar{x}-\mu_0)^2}{s^2/n} = n(\overline{x}-\mu_0)(s^2)^{-1}(\overline{x}-\mu_0)$$

By analogy, for the multivariate case we obtain Hotelling's T^2

$$T^2 = n(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu_0)$$

Multivariate test on a population mean vector

- $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$
- Statistic: $\frac{(n-p)}{p(n-1)} T^2 \sim F_{p,n-p}$
- $100 \cdot (1-\alpha)$ confidence region is the ellipse traced for $\boldsymbol{\mu}$:

$$n(\overline{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \leq c^2 = \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

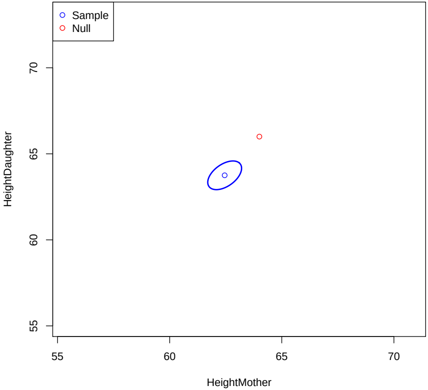
Example

Height of mothers and daughters of the Pearson-Lee data (1903)

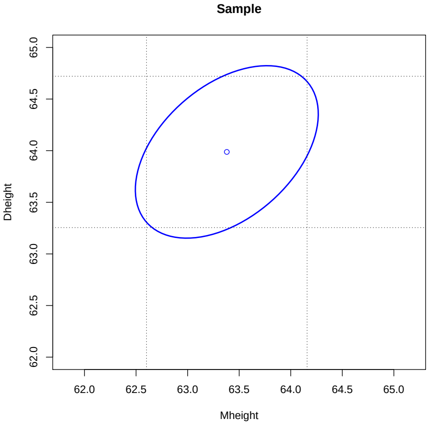
$$H_0 : (\mu_M, \mu_D) = (64, 66) \text{ vs } H_0 : (\mu_M, \mu_D) \neq (64, 66)$$

```
> # H0 values 66; 64;
> modern <- c(64,66)
>
> install.packages("ICSNP")
> library(ICSNP)
>
> HotellingsT2(X,mu=modern,test="f")

Hotelling's one sample T2-test

data:  X
T.2 = 562.9, df1 = 2, df2 = 1373, p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(64,66)
```

Confidence region



Confidence region versus confidence intervals



Univariate Student t-test for two independent samples (common σ^2)

Hypothesis:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Test statistic:

$$T = \frac{\overline{x}_m - \overline{x}_n - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

$$s_p^2 = \frac{(m-1)\, s_X^2 + (n-1)\, s_Y^2}{n + m - 2}$$

Under the null:

$$T \sim t_{n+m-2}$$

Example

	N	N*	Mean	Stdev	Med	Q1	Q3	Min	Max
Boys	77	0	179.506	6.5	178	175	183	165	198
Girls	14	0	167.5	4.363	168.5	165	170	160	174

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

$$s_p^2 = \frac{(m-1)\, S_X^2 + (n-1)\, S_Y^2}{n+m-2} = \frac{(77-1)(6.5)^2 + (14-1)(4.363)^2}{77+14-2} = 38.86232$$

$$T = \frac{\overline{X}_m - \overline{Y}_n - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{179.506 - 167.5}{\sqrt{38.86232}\sqrt{\frac{1}{77} + \frac{1}{14}}} = 6.628885$$

Critical value: $t_{89,0.975} = 1.986979$ p-value: $2 \cdot P\left(t_{89} > 6.628885\right) = 2.52e-09$

$$CI_{0.95}(\mu_1 - \mu_2) = \left(\left(\overline{X}_m - \overline{Y}_n \right) \pm t_{n+m-2,\alpha/2}\, s_p \sqrt{\frac{1}{m} + \frac{1}{n}} \right) = (8.408, 15.605)$$

Multivariate comparison of two groups (common $\boldsymbol{\Sigma}$)

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

Assumptions:

- Both populations are multivariate normal
- $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$

Results:

- $T^2 = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \left((1/n_1 + 1/n_2)\mathbf{S}_p\right)^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$
- $T^2 \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}$
- $\mathbf{S}_p = \frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2}{n_1+n_2-2}$ is the pooled covariance matrix

Example

- Hemophilia A data. Two groups: carriers and non-carriers of a gene for Hemophilia A
- Two variables: Anti Hemophilic Factor activity (AHF-A) and AHF antigen
- Do carriers and non-carriers the same mean vector for these variables?

```
> X <- read.table("hemophilia.dat")
> head(X)
  Group  AHFact AHFanti
1     1 -0.0056 -0.1657
2     1 -0.1698 -0.1585
3     1 -0.3469 -0.1879
4     1 -0.0894  0.0064
5     1 -0.1679  0.0713
6     1 -0.0836  0.0106
> G1 <- X[X$Group==1,2:3]
> G2 <- X[X$Group==2,2:3]
> dim(G1)
[1] 30  2
> dim(G2)
[1] 45  2
> HotellingsT2(G1,G2,test="f")

Hotelling's two sample T2-test

data:  G1 and G2
T.2 = 40.605, df1 = 2, df2 = 72, p-value = 1.562e-12
alternative hypothesis: true location difference is not equal to c(0,0)
```

Multivariate comparison of two groups (no common $\boldsymbol{\Sigma}$)

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

Assumptions:

- Both populations are multivariate normal
- $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Results:

- $T^2 = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 \right)^{-1} [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$
- $T^2 \sim \chi^2_p$

Example: salmon data

```
> head(X)
   Origin Gender Fresh Marine
1 Alaskan Female   108    368
2 Alaskan   Male   131    355
3 Alaskan   Male   105    469
4 Alaskan Female    86    506
5 Alaskan   Male    99    402
6 Alaskan Female    87    423
>
> colMeans(Y[X$Origin=="Alaskan",])
 Fresh Marine
 98.38 429.66
> colMeans(Y[X$Origin=="Canadian",])
 Fresh Marine
137.46 366.62
> cov(Y[X$Origin=="Alaskan",])
           Fresh    Marine
Fresh   260.6078 -188.0927
Marine -188.0927 1399.0861
> cov(Y[X$Origin=="Canadian",])
          Fresh   Marine
Fresh  326.0902 133.5049
Marine 133.5049 893.2608
>

> T2 <- (m1-m2)%*%solve(S1/50+S2/50)%*%(m1-m2)
> T2
         [,1]
[1,] 207.2967
> qchisq(0.95,2)
[1] 5.991465
```

Testing equality of covariance matrices (Box M test)

$$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_g \text{ vs } H_1 : \mathbf{\Sigma}_i \neq \mathbf{\Sigma}_j \text{ for some } i \neq j$$

Box M test statistic

$$M = (N - g) \ln \left(|\mathbf{S}_p| \right) - \sum_{i=1}^{g} (n_i - 1) \ln \left(|\mathbf{S}_i| \right)$$

with:

- $\mathbf{S}_p$ the pooled covariance matrix
- $\mathbf{S}_i$ covariance matrix group S_i
- N total sample size, g number of groups, n_i sample size group i

Asymptotically, the distribution of the statistic under the null:

$$X^2 = -2(1 - c) \ln (M) \approx \chi^2_{(g-1)p(p+1)/2}$$

where c is a constant for bias correction. This test is known to

- be sensitive to deviations from multivariate normality.
- have little power for small samples.
- being too liberal with large samples (rejects too often).

Example: salmon data

```
install.packages("biotools")
library(biotools)
boxM(Y,grouping=X$Origin)
> boxM(Y,grouping=X$Origin)

Box's M-test for Homogeneity of Covariance Matrices

data:  Y
Chi-Sq (approx.) = 10.696, df = 3, p-value = 0.01349
```

Multivariate ANalysis Of Variance (MANOVA)

MANOVA is the extension of Hotelling's T^2 when there are more than two groups.

Statistical model:

$$\mathbf{x}_{\ell j} = \boldsymbol{\mu} + \boldsymbol{\tau}_\ell + \mathbf{e}_{\ell j} = \boldsymbol{\mu}_\ell + \mathbf{e}_{\ell j} \quad j = 1, 2, \ldots, n_\ell \quad \ell = 1, 2, \ldots, g \quad \mathbf{e}_{\ell j} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$$

Hypothesis:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g \text{ vs } H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ for some } i \neq j$$

Equivalently,

$$H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0} \text{ vs } H_1 : \boldsymbol{\tau}_i \neq \mathbf{0} \text{ for some } i$$

- $\boldsymbol{\mu}$ can be estimated by the overall sample mean vector $\bar{\mathbf{x}}$
- $\boldsymbol{\tau}$ can be estimated by the difference vectors $(\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}})$
- $\mathbf{e}$ can be estimated by the difference vectors $(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_\ell)$

MANOVA

- In classical univariate analysis of variance (ANOVA) the analysis consists of a decomposition of the total sum-of-squares in a between part and a within part.
- In MANOVA we have the same decomposition, but in a multivariate way.
- Matrices with sums-of-squares:

$$\mathbf{T} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}})(\mathbf{x}_{\ell j} - \bar{\mathbf{x}})'$$

$$\mathbf{B} = \sum_{\ell=1}^{g} (\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\ell} - \bar{\mathbf{x}})'$$

$$\mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_{\ell})(\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_{\ell})'$$

- and it holds that

$$\mathbf{T}_{p \times p} = \mathbf{B}_{p \times p} + \mathbf{W}_{p \times p}$$

MANOVA table

Source	Sums-of-Squares	DF
Treatment	$\mathbf{B}$	$g-1$
Residual	$\mathbf{W}$	$\sum_{\ell=1}^{g} n_\ell - g$
Total	$\mathbf{T}$	$\sum_{\ell=1}^{g} n_\ell - 1$

To test the null, we use Wilks' lambda

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$$

For large samples

$$-\left(n-1-\frac{p+g}{2}\right)\ln(\Lambda) \sim \chi^2_{p(g-1)}$$

Alternative statistics, such as Pillai's trace or Roy's largest root are often used, and equivalent to Wilks' Λ for large samples.

MANOVA Example: Fisher's iris data

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
> results <- manova(cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width ) ~ Species, data = iris)
> summary(results)
           Df Pillai approx F num Df den Df    Pr(>F)
Species     2 1.1919   53.466      8    290 < 2.2e-16 ***
Residuals 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Bibliography

- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, Chapters 4 and 5, 5th edition, Prentice Hall.