Multivariate Statistics in Ecology and Quantitative Genetics
Hotellings T^2

Dirk Metzler & Martin Hutzenthaler

http://evol.bio.lmu.de/_statgen

Summer semester 2012

Contents

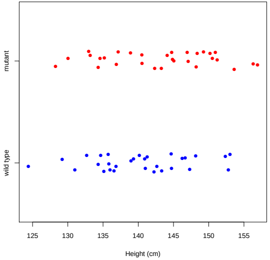
We want to compare the vegetative growth of a mutated raspberry (*Rubus idaeus*) with the wildtype raspberry.

We want to compare the vegetative growth of a mutated raspberry (*Rubus idaeus*) with the wildtype raspberry.

We use different quantities as a measure for “growth”:

- Height of bush (cm)
- Width of bush (cm)

Comparison of height:

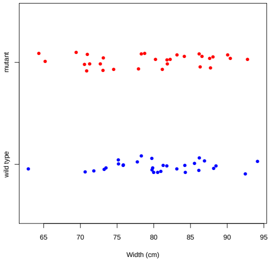



```
> raspberry <- read.table("raspberry.csv",sep=",",h=T)
> attach(raspberry)
> height.wt <- height[type=="wild type"]
> height.mu <- height[type=="mutant"]
> t.test(height.wt,height.mu)

        Welch Two Sample t-test

  data:  height.wt and height.mu
  t = -1.5441, df = 57.334, p-value = 0.1281
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   -6.7627773  0.8735087
  sample estimates:
  mean of x mean of y
   140.2768  143.2215
```

Comparison of width:




```
> width.wt <-  width [type=="wild type"]
> width.mu <-  width [type=="mutant"]
> t.test(width.wt,width.mu)

        Welch Two Sample t-test

data:  width.wt and width.mu
t = 0.5717, df = 56.645, p-value = 0.5698
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.709472  4.874489
sample estimates:
mean of x mean of y
 80.47013  79.38762
```

Both height and width are not significantly different for the the types of raspberry.

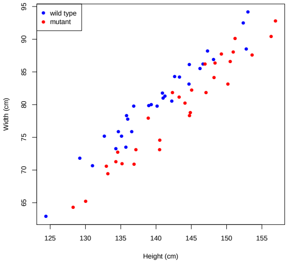
Both height and width are not significantly different for the the types of raspberry.

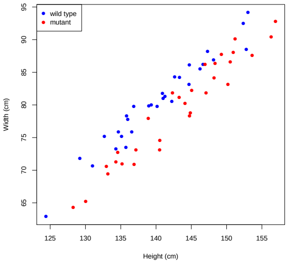
Is there no difference?

Both height and width are not significantly different for the the types of raspberry.

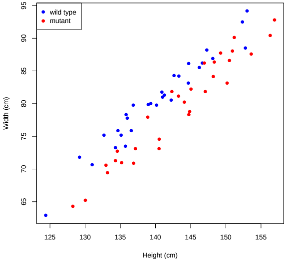
Is there no difference?

So far, we have not exploited the correlation between height and width. Let's look at the bivariate (two variables) data:





We see a difference “by eye”. Which test can detect that?



We see a difference "by eye". Which test can detect that?
Answer: We need the multivariate analogon of the t-test. This multivariate version of the t-test is called

Hotelling's T^2-test

(Before we come to this test, we need to learn some theory)

Contents

- The vector (height,width) is a two-dimensional random vector (with values in $\mathbb{R}^2$)

- The vector (height,width) is a two-dimensional random vector (with values in $\mathbb{R}^2$)
- An d-dimensional random vector is a vector of d random elements

- The vector (height,width) is a two-dimensional random vector (with values in $\mathbb{R}^2$)
- An d-dimensional random vector is a vector of d random elements
- The expectation of a random vector $X = (X_1, X_2, \ldots, X_d)^T$ is the vector of the expectations:

$$\mathbb{E}X = \mathbb{E}\begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_d \end{pmatrix}$$

- The vector (height,width) is a two-dimensional random vector (with values in $\mathbb{R}^2$)
- An d-dimensional random vector is a vector of d random elements
- The expectation of a random vector $X = (X_1, X_2, \ldots, X_d)^T$ is the vector of the expectations:

$$\mathbb{E}X = \mathbb{E}\begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_d \end{pmatrix}$$

- The expectation of a random matrix $M = (M_{ij})_{i=1..n, j=1..d}$ is the matrix of the expectations:

$$\mathbb{E}\begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1d} \\ \vdots & & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nd} \end{pmatrix} = \begin{pmatrix} \mathbb{E}M_{11} & \mathbb{E}M_{12} & \cdots & \mathbb{E}M_{1d} \\ \vdots & \ddots & & \vdots \\ \mathbb{E}M_{n1} & \mathbb{E}M_{n2} & \cdots & \mathbb{E}M_{nd} \end{pmatrix}$$

- reminder: The variance of a univariate random variable X is $\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2$.

- reminder: The variance of a univariate random variable X is $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2$.
- The analog in the multivariate case is the so called *covariance matrix* (or dispersion matrix or variance-covariance matrix). The covariance matrix $\mathrm{Var}(X) = \Sigma$ of $X = (X_1, \ldots, X_d)^T$ is

$$
\begin{aligned}
\Sigma &= \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots \mathrm{Cov}(X_2, X_d) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots \mathrm{Cov}(X_d, X_d) \end{pmatrix} \\
&= \mathbb{E}\left[\begin{pmatrix} X_1 - \mathbb{E}X_1 \\ \vdots \\ X_d - \mathbb{E}X_d \end{pmatrix} \cdot \left(X_1 - \mathbb{E}X_1, \cdots, X_d - \mathbb{E}X_d\right)\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}X) \cdot (X - \mathbb{E}X)^T\right] \\
&= \mathbb{E}\left[X \cdot X^T\right] - \mathbb{E}X \cdot (\mathbb{E}X)^T
\end{aligned}
$$

- Linearity of the expectation is analogos to the univarite case: Let $X = (X_1, \ldots, X_d)$ be a random vector and $C = (C_{ij})_{i=1..n, j=1..d}$ be a deterministic matrix. Then

$$\mathbb{E}(C \cdot X) = C \cdot \mathbb{E}(X)$$

- Linearity of the expectation is analogos to the univarite case: Let $X = (X_1, \ldots, X_d)$ be a random vector and $C = (C_{ij})_{i=1..n, j=1..d}$ be a deterministic matrix. Then

$$\mathbb{E}(C \cdot X) = C \cdot \mathbb{E}(X)$$

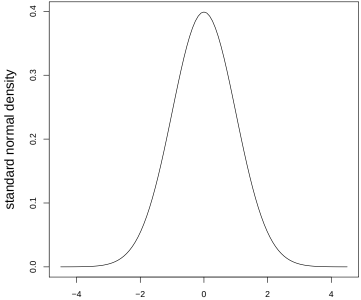
- If $Y := X - \mathbb{E}(X)$, then

$$\begin{aligned}
\operatorname{Var}(C \cdot X) &= \operatorname{Var}(C \cdot Y) \\
&= \mathbb{E}\left[C \cdot Y \cdot (C \cdot Y)^T\right] \\
&= \mathbb{E}\left[C \cdot Y \cdot Y^T \cdot C^T\right] \\
&= C \cdot \mathbb{E}\left[Y \cdot Y^T\right] \cdot C^T \\
&= C \cdot \operatorname{Var}(Y) \cdot C^T \\
&= C \cdot \operatorname{Var}(X) \cdot C^T
\end{aligned}$$

Contents

- Reminder: Univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$ has the density

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



Remember: $\mathbb{P}(\mu - \sigma < X < \mu + \sigma) = 0.68$ and
$\mathbb{P}(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95$

- The density of the *d*-dimensional normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is analogous:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right)$$

for $x \in \mathbb{R}^d$ where $\det(\Sigma)$ is the determinand of Σ, and Σ^{-1} is the inverse matrix.
We write $\mathcal{N}_d\left(\mu, \Sigma\right)$ for this distribution.

- The density of the *d*-dimensional normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is analogous:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right)$$

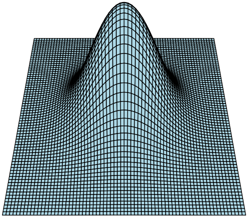
  for $x \in \mathbb{R}^d$ where $\det(\Sigma)$ is the determinand of Σ, and Σ^{-1} is the inverse matrix.
  We write $\mathcal{N}_d(\mu, \Sigma)$ for this distribution.

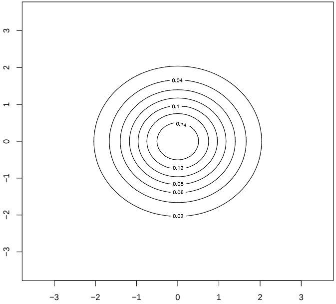
- The *standard multivariate normal distribution* has mean $\mu = 0$ and the identity matrix $\Sigma = \mathbb{1}$ as covariance matrix.

Plots for $d = 2$

Correlation 0.0:

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathrm{Var}(X_1) = 1 = \mathrm{Var}(X_2), \mathrm{Cov}(X_1, X_2) = 0.0$$

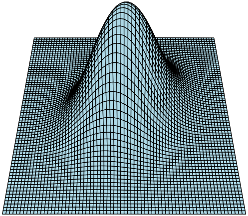


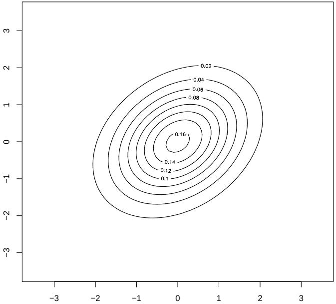


Plots for $d = 2$

Correlation 0.3:

$$\Sigma = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}, \text{Var}(X_1) = 1 = \text{Var}(X_2), \text{Cov}(X_1, X_2) = 0.3$$

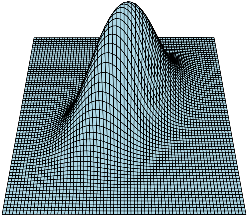


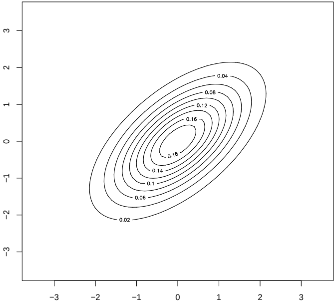


Plots for $d = 2$

Correlation 0.6:

$$\Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}, \text{Var}(X_1) = 1 = \text{Var}(X_2), \text{Cov}(X_1, X_2) = 0.6$$

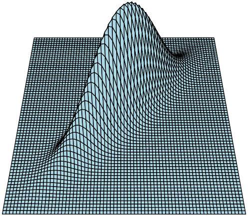


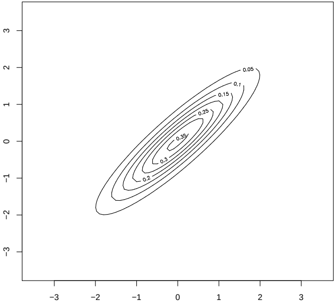


Plots for $d = 2$

Correlation 0.9:

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, \ \mathrm{Var}(X_1) = 1 = \mathrm{Var}(X_2), \ \mathrm{Cov}(X_1, X_2) = 0.9$$





Properties: (Let the distribution of X be $\mathcal{N}_d(\mu, \Sigma)$)

- Linear combinations are univariate normal distributed: $\langle c, X \rangle \sim \mathcal{N}\left(\langle c, \mu \rangle, c\Sigma c^T\right)$

Properties: (Let the distribution of X be $\mathcal{N}_d(\mu, \Sigma)$)

- Linear combinations are univariate normal distributed: $\langle c, X \rangle \sim \mathcal{N}\left(\langle c, \mu \rangle, c\Sigma c^T\right)$
- X_i and X_j are independent $\iff$ $\mathrm{Cov}(X_i, X_j) = 0$

Properties: (Let the distribution of X be $\mathcal{N}_d(\mu, \Sigma)$)

- Linear combinations are univariate normal distributed: $\langle c, X \rangle \sim \mathcal{N}\left(\langle c, \mu \rangle, c\Sigma c^T\right)$
- X_i and X_j are independent $\iff$ $\mathrm{Cov}(X_i, X_j) = 0$
- The standardized normal distribution is standard normal distributed
$$\Sigma^{-\frac{1}{2}} \cdot (X - \mu) \sim \mathcal{N}_d(0, \mathbb{1})$$
where $M = \Sigma^{-\frac{1}{2}}$ is a matrix such that $M^T \cdot M \cdot \Sigma = \mathbb{1}$.

Properties: (Let the distribution of X be $\mathcal{N}_d(\mu, \Sigma)$)

- Linear combinations are univariate normal distributed: $\langle c, X \rangle \sim \mathcal{N}\left(\langle c, \mu \rangle, c\Sigma c^T\right)$
- X_i and X_j are independent $\iff$ $\mathrm{Cov}(X_i, X_j) = 0$
- The standardized normal distribution is standard normal distributed
$$\Sigma^{-\frac{1}{2}} \cdot (X - \mu) \sim \mathcal{N}_d(0, \mathbb{1})$$
where $M = \Sigma^{-\frac{1}{2}}$ is a matrix such that $M^T \cdot M \cdot \Sigma = \mathbb{1}$.
- The square of the standardized normal distribution is chi-squared distributed with d degrees of freedom:
$$(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2_d.$$

Properties: (Let the distribution of X be $\mathcal{N}_d(\mu, \Sigma)$)

- Linear combinations are univariate normal distributed: $\langle c, X\rangle \sim \mathcal{N}\left(\langle c, \mu\rangle, c\Sigma c^T\right)$
- X_i and X_j are independent $\iff$ $\mathrm{Cov}(X_i, X_j) = 0$
- The standardized normal distribution is standard normal distributed
$$\Sigma^{-\frac{1}{2}} \cdot (X - \mu) \sim \mathcal{N}_d(0, \mathbb{1})$$
where $M = \Sigma^{-\frac{1}{2}}$ is a matrix such that $M^T \cdot M \cdot \Sigma = \mathbb{1}$.
- The square of the standardized normal distribution is chi-squared distributed with d degrees of freedom:
$$(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2_d.$$
- If $Y_1, Y_2, \dots, Y_d$ are independent and standard normal distributed, then $(Y_1, \dots, Y_d) \sim \mathcal{N}(0, \mathbb{1})$.

Properties: (Let the distribution of X be $\mathcal{N}_d(\mu, \Sigma)$)

- If $M \in \mathbb{R}^{p \times d}$ is a non-random matrix, then $M \cdot X \sim \mathcal{N}_p\left(M \cdot \mu, M \Sigma M^T\right)$

Estimating μ and Σ

Let $Y_{1\cdot}, \ldots, Y_{n\cdot}$ be a sample of independent observations from a $\mathcal{N}_d(\mu, \Sigma)$ distribution. As usual, we write the variables as columns and the different observations as rows:

$$Y = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1d} \\ Y_{21} & Y_{22} & \cdots & Y_{2d} \\ \vdots & & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nd} \end{pmatrix}$$

- The sample mean $\bar{Y}$ where $\bar{Y}_j := \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$ is an estimator for μ.

Estimating μ and Σ

Let $Y_{1\cdot}, \ldots, Y_{n\cdot}$ be a sample of independent observations from a $\mathcal{N}_d(\mu, \Sigma)$ distribution. As usual, we write the variables as columns and the different observations as rows:

$$Y = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1d} \\ Y_{21} & Y_{22} & \cdots & Y_{2d} \\ \vdots & & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nd} \end{pmatrix}$$

- The sample mean $\bar{Y}$ where $\bar{Y}_j := \frac{1}{n} \sum_{i=1}^{n} Y_{ij}$ is an estimator for μ.
- The sample covariance matrix $S := (S_{ij})_{i,j=1..d}$ where

$$S_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \left(Y_{ki} - \bar{Y}_i\right)\left(Y_{kj} - \bar{Y}_j\right)$$

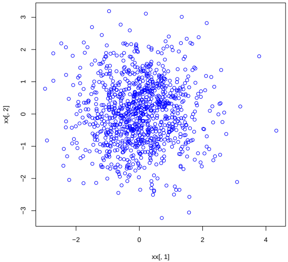
is an (unbiased) estimator for Σ.

Contents

```
> library("mvtnorm") # multivariate t- and normal distrib.
> Sigma0 <- matrix(c(1,0,0,1),ncol=2)
> dmvnorm(c(0,0),mean=c(0,0),sigma=Sigma0)
[1] 0.1591549
> dmvnorm(c(0,0))  # same, the standard normal is the default
[1] 0.1591549
```

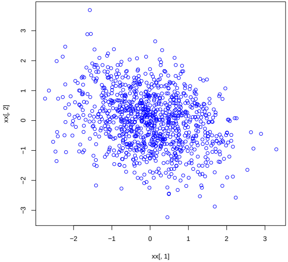
```
> library("mvtnorm") # multivariate t- and normal distrib.
> Sigma0 <- matrix(c(1,0,0,1),ncol=2)
> dmvnorm(c(0,0),mean=c(0,0),sigma=Sigma0)
[1] 0.1591549
> dmvnorm(c(0,0))  # same, the standard normal is the default
[1] 0.1591549

> xx  <- rmvnorm(1000,mean=c(0,0),sigma=Sigma0)
> plot(xx[,1],xx[,2])
```



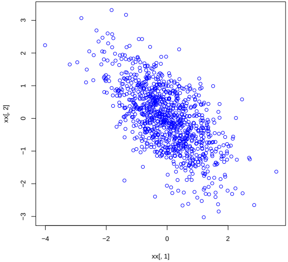
Correlation -0.3:

```
> Sigma3 <- matrix(c(1,-0.3,-0.3,1),ncol=2)
> xx  <- rmvnorm(1000,mean=c(0,0),sigma=Sigma3)
> plot(xx[,1],xx[,2])
```



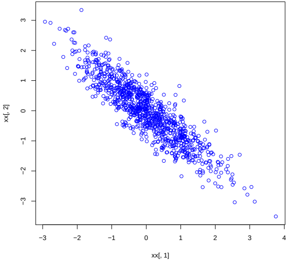
Correlation -0.6:

```
> Sigma6 <- matrix(c(1,-0.6,-0.6,1),ncol=2)
> xx  <- rmvnorm(1000,mean=c(0,0),sigma=Sigma6)
> plot(xx[,1],xx[,2])
```



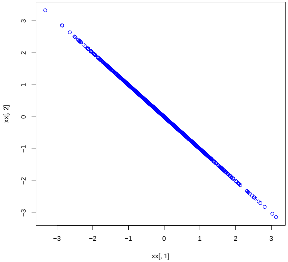
Correlation -0.9:

```
> Sigma9 <- matrix(c(1,-0.9,-0.9,1),ncol=2)
> xx  <- rmvnorm(1000,mean=c(0,0),sigma=Sigma9)
> plot(xx[,1],xx[,2])
```



Correlation -1:

```
> Sigma10 <- matrix(c(1,-1,-1,1),ncol=2)
> xx  <- rmvnorm(1000,mean=c(0,0),sigma=Sigma10)
> plot(xx[,1],xx[,2])
```



Contents

- Reminder for univarite case $d = 1$: If $X = (X_1, \ldots, X_n)$ is an independent sample from the $\mathcal{N}_1(\mu, \sigma^2)$ distribution, then the t-statistics

$$t := \frac{\bar{X} - \mu}{sd(X)/\sqrt{n}}$$

is t-distributed with $n - 1$ degrees of freedom. This fact is used for the t-test.

- Reminder for univarite case $d = 1$: If $X = (X_1, \dots, X_n)$ is an independent sample from the $\mathcal{N}_1(\mu, \sigma^2)$ distribution, then the t-statistics
$$t := \frac{\bar{X} - \mu}{sd(X)/\sqrt{n}}$$
is t-distributed with $n - 1$ degrees of freedom. This fact is used for the t-test.

- Reminder for univarite case $d = 1$: If $X = (X_1, \dots, X_n)$ is an independent sample from the $\mathcal{N}_1(\mu, \sigma^2)$ distribution, then the t-statistics

$$t := \frac{\bar{X} - \mu}{sd(X)/\sqrt{n}}$$

is t-distributed with $n - 1$ degrees of freedom. This fact is used for the t-test. In addition, $t^2 \sim \mathcal{F}_{1,n-1}$ is Fisher-distributed with 1 numerator degree of freedom and $n - 1$ denominator degrees of freedom.

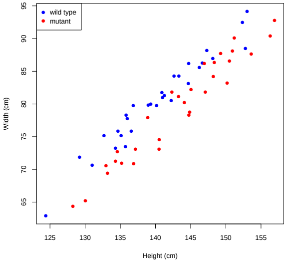
- Multivariate case: If $X := (X_{1,\cdot}, \dots, X_{n,\cdot}$ be a sample from the $\mathcal{N}_d(\mu, \Sigma)$ distribution with $n > d$, then the statistics

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

is called *Hotellings* T^2. The matrix S^{-1} is the inverse of the sample covariance matrix. It has been shown that

$$T^2 \sim \frac{(n-1)d}{n-d} \mathcal{F}_{d,n-d}$$

Back to our raspberry example:



We see a difference "by eye". Does Hotellings T^2-test detect this difference?

```
> raspberry <- read.table("raspberry.csv",sep=",",header=T)
> raspberry.wt<-subset(raspberry,type=="wild type",select=-3)
> raspberry.mu<-subset(raspberry,type=="mutant",select=-3)
> library(rrcov) #one out of many libraries with Hotelling T2
> T2.test(raspberry.wt,raspberry.mu)

        Hotelling's two sample T2-test

data:  raspberry.wt and raspberry.mu
T.2 = 36.7558, df1 = 2, df2 = 57, p-value = 5.577e-11
alternative hypothesis: true location difference is not
equal to c(0,0)
sample estimates:
               height    width
mean x-vector 143.2215 79.38762
mean y-vector 140.2768 80.47013
```

Hotellings T^2-test detects a significant difference

```
# alternative syntax with a formula
> T2.test(cbind(height,width)~type,data=raspberry)
```

```
# alternative syntax with a formula
> T2.test(cbind(height,width)~type,data=raspberry)
```

Syntax:

```
## Default S3 method:
T2.test(x, y = NULL, mu = 0, conf.level = 0.95, ...)

## S3 method for class 'formula':
T2.test(formula, data, subset, na.action, ...)
```

If `y = NULL`, then perform a one sample T^2-test with null hypothesis $H_0:$ true mean is `mu`

Directions for use:

- The T^2-test (like the t-test) is prone to outliers. A single outlier can considerably decrease the power of the test. So check for outliers (possible mismeasurements)

Directions for use:

- The T^2-test (like the t-test) is prone to outliers. A single outlier can considerably decrease the power of the test. So check for outliers (possible mismeasurements)
- The T^2-test (like the t-test) is prone to dependence in the data. The test assumes that different samples (the rows of X) are independent. So be careful when sampling your data.

Directions for use:

- The T^2-test (like the t-test) is prone to outliers. A single outlier can considerably decrease the power of the test. So check for outliers (possible mismeasurements)
- The T^2-test (like the t-test) is prone to dependence in the data. The test assumes that different samples (the rows of X) are independent. So be careful when sampling your data.
- The T^2-test assumes $n > d$, so make sure that you have more samples than variables.

Directions for use:

- The T^2-test (like the t-test) is prone to outliers. A single outlier can considerably decrease the power of the test. So check for outliers (possible mismeasurements)
- The T^2-test (like the t-test) is prone to dependence in the data. The test assumes that different samples (the rows of X) are independent. So be careful when sampling your data.
- The T^2-test assumes $n > d$, so make sure that you have more samples than variables.
- The T^2-test (more than the t-test) looses test power if the true distribution (of the population) deViates from the normal distribution (e.g., the Shapiro-Wilk test tests for normality)

Directions for use:

- The T^2-test (like the t-test) is prone to outliers. A single outlier can considerably decrease the power of the test. So check for outliers (possible mismeasurements)
- The T^2-test (like the t-test) is prone to dependence in the data. The test assumes that different samples (the rows of X) are independent. So be careful when sampling your data.
- The T^2-test assumes $n > d$, so make sure that you have more samples than variables.
- The T^2-test (more than the t-test) looses test power if the true distribution (of the population) deViates from the normal distribution (e.g., the Shapiro-Wilk test tests for normality)
- If one of the variables shows already a significant difference between the group, then also the T^2-test yields a significant difference.

Contents

A multivariate Shapiro-Wilk test for normality is `mshapiro.test()` in the library `mvnormtest`:

```
# install.packages("mvnormtest")  # install library if i
> library(mvnormtest)
> library(mvtnorm)
> x <- rmvnorm(100,mean=c(0,0))
> mshapiro.test(t(x))

      Shapiro-Wilk normality test

data:  Z
W = 0.9915, p-value = 0.7839
```