# Multivariate Inference - I

# Hotelling's T2

# Confidence region, simultaneous conf. intervals

STAT 32950-24620

Spring 2023 (4/6)



## Multivariate statistical distance

$\mathbb{R}^p$ vectors:

$$\mathbf{x} = (x_1, \cdots, x_p)', \qquad \mu = (\mu_1, \cdots, \mu_p)'$$

The **Euclidean distance** between vector $\mathbf{x}$ and its mean $\mu$ is

$$\sqrt{(x_1 - \mu_1)^2 + \cdots + (x_p - \mu_p)^2} = \sqrt{(\mathbf{x} - \mu)'(\mathbf{x} - \mu)}$$

The Mahalanobis distance (or **statistical distance**) between observed random vector $\mathbf{x}$ and its mean vector $\mu$ is defined as

$$\sqrt{(\mathbf{x} - \mu)' S^{-1} (\mathbf{x} - \mu)}$$

where $S$ denotes the sample covariance matrix of $\mathbf{x}$.



## Example (turtle data)

Data: Turtle shell measurements (n=48 observations, p=4 variables)

```
load(file="turtles.rda")
attach(turtles)  # from package Flury
str(turtles)    #original data n=48, var=4: Sex, L, W, H
```

```
## 'data.frame':    48 obs. of  4 variables:
##  $ Gender: Factor w/ 2 levels "Male","Female": 1 1 1 1
##  $ Length: int  93 94 96 101 102 103 104 106 107 112 ...
##  $ Width : int  74 78 80 84 85 81 83 83 82 89 ...
##  $ Height: int  37 35 35 39 38 37 39 39 38 40 ...
```



## Take a subset of the data (Male only)

```
summary(turtles)
```

```
##     Gender       Length          Width             Height
##  Male  :24    Min.   : 93    Min.   : 74.0    Min.   :35.0
##  Female:24    1st Qu.:107    1st Qu.: 86.0    1st Qu.:40.0
##               Median :122    Median : 93.0    Median :44.5
##               Mean   :125    Mean   : 95.4    Mean   :46.3
##               3rd Qu.:136    3rd Qu.:102.0    3rd Qu.:51.0
##               Max.   :177    Max.   :132.0    Max.   :67.0
```

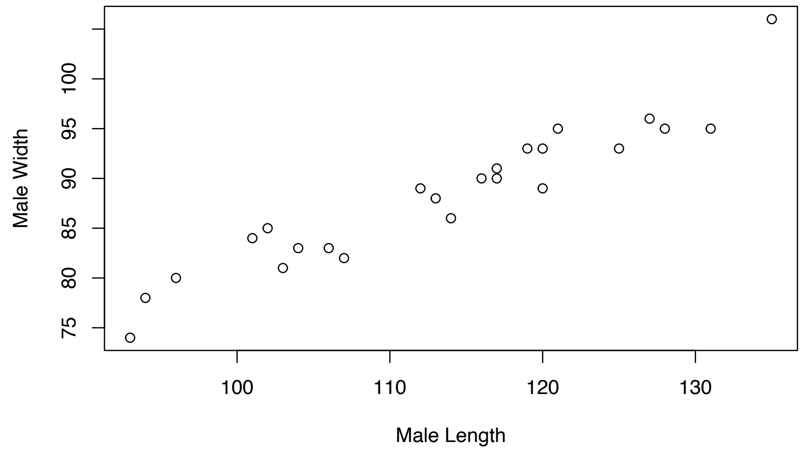
Subset observations into two groups by Gender
($g = 2$ "treatment" groups)

```
male=subset(turtles[,2:4],Gender=="Male")
female=subset(turtles[,2:4],Gender=="Female")
```

## Example scatter plot on two variables

Plot: Turtle data, male only, on 2 variables

```
x = male$Length; y = male$Width; male2=cbind(x,y)
plot(x,y, xlab="Male Length",ylab="Male Width")
```






## $T^2$ statistic

$T^2$ **statistic for multivariate data** is defined as

$$\mathbf{T}^2 = n(\bar{\mathbf{X}} - \mu)' S^{-1} (\bar{\mathbf{X}} - \mu) = (\bar{\mathbf{X}} - \mu)' \left( \widehat{Cov}(\bar{\mathbf{X}}) \right)^{-1} (\bar{\mathbf{X}} - \mu)$$

Under $H_o$ that $\mu$ is the true mean,

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

$$\Longrightarrow \qquad \frac{n-p}{(n-1)p} n(\bar{\mathbf{X}} - \mu)' S^{-1} (\bar{\mathbf{X}} - \mu) \sim F_{p,n-p}$$



## F distribution and coverage probability for $T^2$

Probability statement about sample mean $\bar{\mathbf{X}}$:

$$P\left( \frac{n-p}{(n-1)p} n(\bar{\mathbf{X}} - \mu)' S^{-1} (\bar{\mathbf{X}} - \mu) \leq F_{p,n-p,\alpha} \right) = 1 - \alpha$$

where the quantile $F_{p,n-p,\alpha}$ is defined as

$$P\left( F_{p,n-p} \leq F_{p,n-p,\alpha} \right) = 1 - \alpha$$



## Region of $\bar{\mathbf{X}}$ near $\mu$ with probability $1 - \alpha$

Therefore, when $\mu$ is fixed,

random sample mean vector $\bar{\mathbf{X}}$ would be in the region

$$\sqrt{(\bar{\mathbf{X}} - \mu)' S^{-1} (\bar{\mathbf{X}} - \mu)} \leq \sqrt{\frac{(n-1)p}{(n-p)n} F_{p,n-p,\alpha}}$$

with probability $1 - \alpha$.

The inequality gives an upper bound on the statistical distance (a.k.a. Mahalanobis distance) between $\bar{\mathbf{X}}$ and $\mu$.

Given observed $\bar{\mathbf{x}}$, the inequality can be converted to a $(1 - \alpha)100\%$ confidence region for the unknown parameter vector $\mu$.

## Construct ellipsoidal confidence region for $\mu$ from $T^2$'s F

- Obtain $\bar{\mathbf{x}}$ from data.
- Consider all possible values of $\mu$ satisfying the inequality.
- Such $\mu$ forms an ellipsoidal shape centered at $\bar{\mathbf{x}}$.
- Such $\mu$ forms a $(1-\alpha)100\%$ confidence region for the true $\mu$.
- Given $\alpha$, the F-quantile can be found in F-tables or software.
  R command `qf(1-alpha,df1=p,df2=n-p)`
  gives the F-quantile $F_{p,n-p,\alpha}$.

```
qf(0.95, df1=5, df2=20)  # Example of F(df1,df2) quantile
```

```
## [1] 2.711
```



## Example: Construct $T^2$ elliptic confidence region

For male turtle data (partial):

Construct a **98% confidence region** for the mean vector of $(\mathbf{x}, \mathbf{y})$

$\mathbf{x}$ = Length (male)
$\mathbf{y}$ = Width (male)

First, find the center:

```
# Find the center of the elliptic region
xbar=mean(x)
ybar=mean(y)
c(xbar,ybar)
```

```
## [1] 113.38  88.29
```



## Example (cont.): Construct elliptic confidence region

Find the F quantile:

```
# Find the 0.98 percentile of F(p,n-p)
n=24
p=2
qf(.98,df1=p,df2=n-p)      # = 4.698
```

```
## [1] 4.698
```

Find the statistical distance "radius" or boundary of the region:

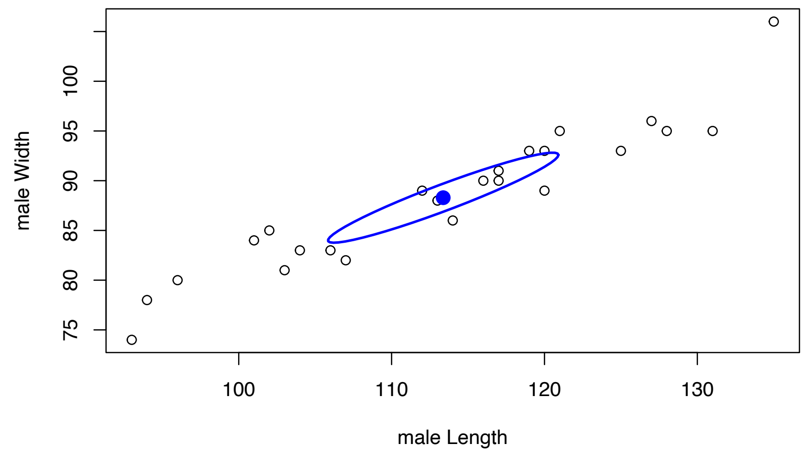
```
# Radius for sqrt((X-mu)'(Sinv)(X-mu))
sqrt(qf(.98,df1=p,df2=n-p)*(n-1)*p/(n*(n-p)))    #=.64
```

```
## [1] 0.6398
```



## Plot of 98% elliptic confidence region of $\mu$

```
library(car)
plot(x,y,xlab="male Length",ylab="male Width")
ellipse(c(xbar,ybar),shape=cov(male2),radius=.64)
```



## Marginal simultaneous Confidence Intervals

Marginal simultaneous confidence intervals are formed by univariate $t$ confidence interval. For each component $\bar{X}_k$,

$$\frac{\bar{X}_k - \mu_k}{\sqrt{s_{kk}/n}} \sim t_{n-1}, \quad k = 1, \cdots, p.$$

Without considering dependency among components,

$$\Longrightarrow \quad P\left(\left|\frac{\bar{X}_k - \mu_k}{\sqrt{s_{kk}/n}}\right| \leq t_{n-1,\alpha/2}\right) = 1-\alpha$$

$$\Longrightarrow \quad \bar{X}_k - t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}} \leq \mu_k \leq \bar{X}_k + t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}}$$

Each interval contains its mean parameter $\mu_k$ at $(1-\alpha)100\%$ confidence level, regardless or **ignoring dependence** among component variables.



## Special case: independent components

If the component variables were **independent**, then

$$Pr\left\{\bar{X}_k - t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}} \leq \mu_k \leq \bar{X}_k + t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}},\ k = 1, \cdots, p\right\}$$

$$= \prod_{k=1}^{p} Pr\left\{\bar{X}_k - t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}} \leq \mu_k \leq \bar{X}_k + t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}}\right\}$$

That is, the $p$ intervals, forming a hyper-rectangle in $\mathbb{R}^k$, contain all $\mu_k$ simultaneously with confidence level $(1-\alpha)^p 100\%$ – not $(1-\alpha)100\%$ – when independence holds, which is not true in general.



## Example marginal t C.I. for individual component mean $\mu_k$

For male turtle length-width data: Construct marginal 99% C.I. for the mean length and mean width individually.

```
n=24; p=2; alpha=.01
se=sqrt(diag(cov(male2)))/sqrt(n)
q = 1-(alpha/(2))
cr=qt(q,n-1)
# t-C.I. limits;  marginal
x1 = xbar - cr*se[1]; x2 = xbar + cr*se[1]
y1 = ybar - cr*se[2]; y2 = ybar + cr*se[2]
c(x1,x2,y1,y2)
```

```
##      x      x      y      y
## 106.62 120.13  84.24  92.35
```



## Example: Comparison of multivariate vs. univariate

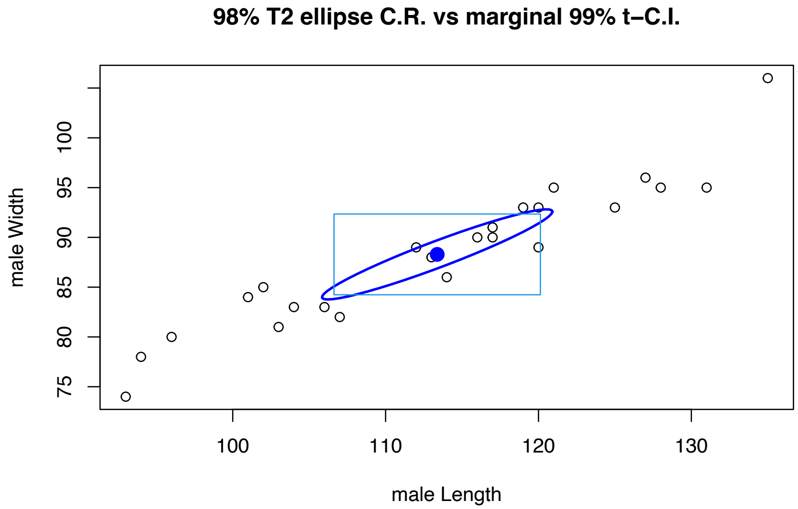
Comparison:

Ellipsoical onfidence Regions vs marginal C.I.

multivariate vs. univariate approaches

```
plot(x,y,xlab="male Length",ylab="male Width")
ellipse(c(xbar,ybar),shape=cov(male2),radius=.64);
rect(x1,y2,x2,y1,border=4) #blue, .99 marg
title("98% T2 ellipse C.R. vs marginal 99% t-C.I.")
```

## Plot: $T^2$ ellitic C.R. vs marginal t C.I.'s






## Bonferroni simultaneous C.I.'s

Bonferroni simultaneous C.I.'s are also formed by univariate confidence intervals. For each component random variable $\bar{X}_k$,

$$\frac{\bar{X}_k - \mu_k}{\sqrt{s_{kk}/n}} \sim t_{n-1} \qquad \Longrightarrow \quad P\left(\left|\frac{\bar{X}_k - \mu_k}{\sqrt{s_{kk}/n}}\right| \leq t_{n-1,\alpha_k/2}\right) = 1 - \alpha_k$$

$$\Longrightarrow \quad P\left(\left|\frac{\bar{X}_k - \mu_k}{\sqrt{s_{kk}/n}}\right| > t_{n-1,\alpha_k/2}\right) = \alpha_k, \qquad k = 1, \cdots, p.$$

Then

$$P\left(\left|\frac{\bar{X}_k - \mu_k}{\sqrt{s_{kk}/n}}\right| > t_{n-1,\alpha_k/2} \text{ for some } k = 1, \cdots, p\right) \leq \alpha_1 + \cdots + \alpha_p$$

even when the components are dependent.



## Choose confidence level for Bonferroni simultaneous C.I.'s

For Bonferroni simultaneous confidence intervals, choose

$$\alpha_k = \alpha/p, \qquad \textit{then} \quad \alpha_1 + \cdots + \alpha_p = \alpha$$

$$\Longrightarrow \quad P\left(\left|\frac{\bar{X}_k - \mu_k}{\sqrt{s_{kk}/n}}\right| > t_{n-1,\alpha/2p}, \quad \text{for some } k = 1, \cdots, p\right) \leq \alpha$$

$$\Longrightarrow \quad P\left(\left|\frac{\bar{X}_k - \mu_k}{\sqrt{s_{kk}/n}}\right| \leq t_{n-1,\alpha/2p}, \quad \text{for all } k = 1, \cdots, p\right) \geq 1 - \alpha$$

Then for all $k = 1, \cdots, p$,

$$\bar{X}_k - t_{n-1,\alpha/2p}\sqrt{\frac{s_{kk}}{n}} \leq \mu \leq \bar{X}_k + t_{n-1,\alpha/2p}\sqrt{\frac{s_{kk}}{n}}$$

It's a hyper-rectangular region with confidence level $\geq (1-\alpha)100\%$.



## Example: Bonferroni t-C.I.'s

Example: Bonferroni t C.I.'s for individual component mean $\mu_k$

For male turtle length-width data:

Construct Bonferroni 99% C.I. for the mean length and mean width.

```
n=24; p=2; alpha=.01
se=sqrt(diag(cov(male2)))/sqrt(n)
q = 1-(alpha/(2*p))
cr=qt(q,n-1)
c(xbar - cr*se[1], xbar + cr*se[1]) # Bonf. CI, x
```

```
##     x     x
## 105.9 120.8
```

```
c(ybar - cr*se[2], ybar + cr*se[2])# Bonf. CI, y
```

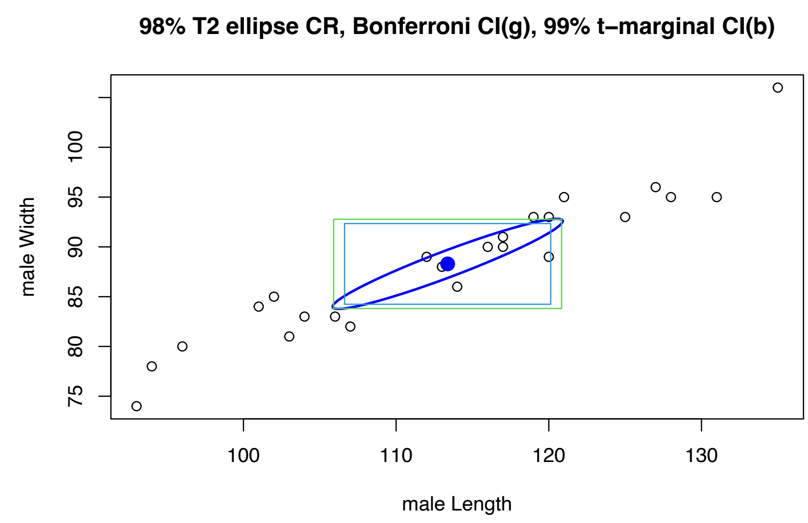
```
##     y     y
## 83.81 92.77
```

## Bonferroni vs Marginal vs $T^2$-ellipse

```
plot(x,y,xlab="male Length",ylab="male Width")
ellipse(c(xbar,ybar),shape=cov(male2),radius=.64);
rect(x1,y2,x2,y1,border=4)   #t-marginal,blue
title("98% T2 ellipse CR, Bonferroni CI(g), 99% t-marginal
rect(105.91,83.81,120.84,92.77,border=3) #green,Bonferroni
```



## PLot of $T^2$ ellipse vs t-marginal vs Bonferroni CI's






## Mazimization Lemmma results on $T^2$

By the Mazimization Lemmma (generalized Cauchy-Schwarz inequality)

$$n\frac{[a'(\bar{X}-\mu)]^2}{a'Sa} \le n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu)$$

Under $H_o$,

$$n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) = T^2 \sim (n-1)\frac{p}{n-p}F_{p,n-p}$$

$$P\left(T^2 \le (n-1)\frac{p}{n-p}F_{p,n-p}\right) = 1-\alpha$$

Thus

$$P\left(n\frac{[a'(\bar{X}-\mu)]^2}{a'Sa} \le (n-1)\frac{p}{n-p}F_{p,n-p}\right) \ge 1-\alpha$$



## Hyper-rectangle Simultaneous C.I. by $T^2$

Choose $a = [\,0 \cdots 0\;1\;0 \cdots 0\,]'$

with $k$th component $= 1$ and 0 for other components.

Then

$$a\bar{x} = \bar{x}_k, \qquad a\bar{\mu} = \mu_k, \qquad a'Sa = s_{kk} = s_k^2$$

We obtain **simultaneous confidence intervals by** $T^2$ for the component means:

$$\bar{x}_k - \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p,\alpha}}\sqrt{\frac{s_{kk}}{n}} \le \mu_k \le \bar{x}_k + \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p,\alpha}}\sqrt{\frac{s_{kk}}{n}}$$

for all $k = 1, \cdots, p$.

## Example: Construct $T^2$ simultaneous C.I.

Example: Construct $T^2$ 98% simultaneous C.I. for the mean length and mean width.

```
n=24; p=2; alpha=.02
se=sqrt(diag(cov(male2)))/sqrt(n)
cr=sqrt(qf(1-alpha,p,n-p)*(n-1)*p/(n-p))
c(xbar - cr*se[1], xbar + cr*se[1])    # 105.8 120.9
```

```
##     x     x
## 105.8 120.9
```

```
c(ybar - cr*se[2], ybar + cr*se[2])    # 83.77 92.82
```

```
##     y     y
## 83.77 92.82
```

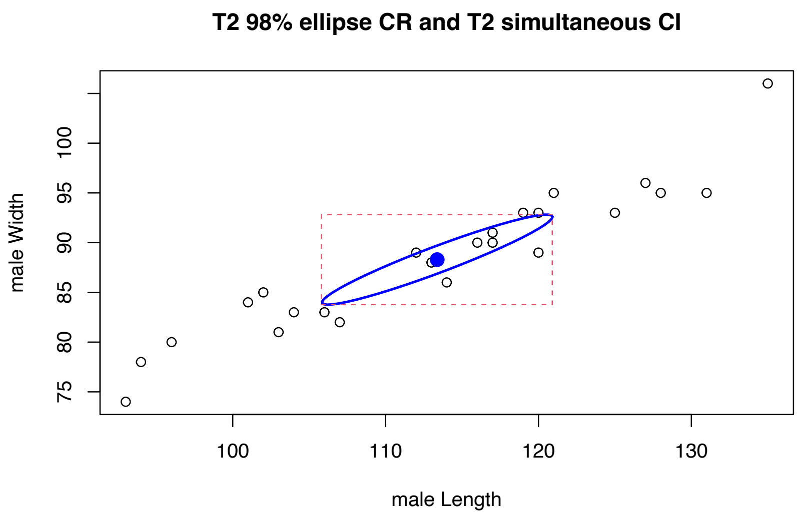


## Plot code (T2 C.R. and Simultaneous C.I.)

```
plot(x,y,xlab="male Length",ylab="male Width")
ellipse(c(xbar,ybar),shape=cov(male2),radius=.64);
#rect(x1,y2,x2,y1,border=4)
#blue, t-marginal 99% each
# rect(105.91, 83.81, 120.84, 92.77, border=3,lwd=2)
#green, Bonf, 99%
title("T2 98% ellipse CR and T2 simultaneous CI")
rect(105.8, 83.77, 120.9, 92.82,border=2, lty=2) #red, T2
```



## Plot T2 C.R. and Simultaneous C.I. 98%






## Asymptotic confidence intervals

When the sample size $n$ is large, Hotelling's $T^2$

$$n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \sim \chi^2_p \quad \textit{approximately}$$

The $\chi^2$ approximation gives the **asymptotic simultaneous confidence intervals** for the component means:

$$\bar{x}_k - \sqrt{\chi^2_{p,\alpha}}\sqrt{\frac{s_{kk}}{n}} \le \mu_k \le \bar{x}_k + \sqrt{\chi^2_{p,\alpha}}\sqrt{\frac{s_{kk}}{n}}$$

## Example Construct asymptotic 99% C.I.'s

Construct asymptotic 99% C.I. for the mean length and mean width.

```
n=24; p=2; alpha=.01
se=sqrt(diag(cov(male2)))/sqrt(n)
cr=sqrt(qchisq(1-alpha,df=2))
c(xbar - cr*se[1], xbar + cr*se[1])     # 106.1 120.7
```

```
##     x     x
## 106.1 120.7
```

```
c(ybar - cr*se[2], ybar + cr*se[2])     # 83.91 92.67
```

```
##     y     y
## 83.91 92.67
```

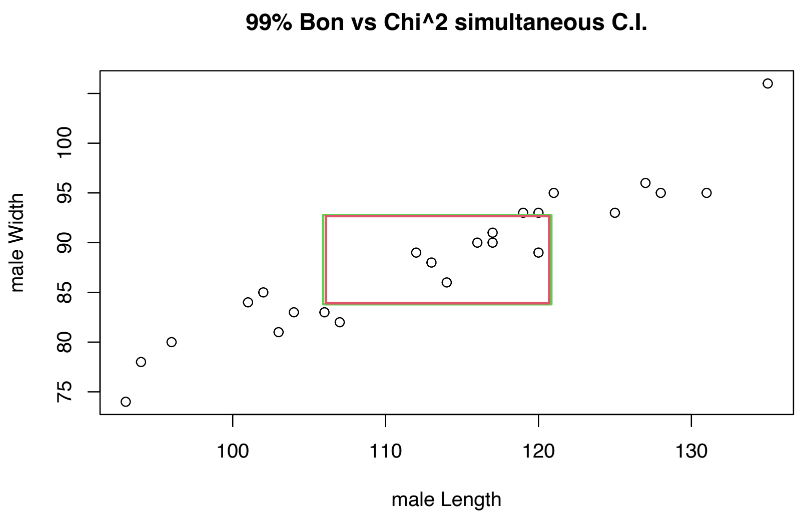


## Comparison: Bonferroni vs asymptotic

```
plot(x,y,xlab="male Length",ylab="male Width");
title("99% Bon vs Chi^2 simultaneous C.I.")
rect(105.91, 83.81, 120.84, 92.77, border=3,lwd=2)
#green Bonf
rect(106.1,83.91, 120.7, 92.67,border=2,lwd=2)
#red, asymptotic
```



## Plot Bonferroni vs asymptotic simultaneous C.I.






## Comparison of C.R. and simultaneous C.I.

Comparison at 99% confidence level:

- Marginal t confidence interval for each component
- Bonferroni simultaneous confidence intervals
- Asymptotic $\chi_p^2$ simultaneous confidence intervals
- Simultaneous confidence intervals from $T^2$
- Ellipsoidal $T^2$ confidence region

```
n=24; p=2; alpha=.01
se=sqrt(diag(cov(male2)))/sqrt(n)
cr=sqrt(qf(1-alpha,p,n-p)*(n-1)*p/(n-p))
c(xbar - cr*se[1], xbar + cr*se[1])     # 105.1 121.7
```

```
##     x     x
## 105.1 121.7
```

```
c(ybar - cr*se[2], ybar + cr*se[2])     # 83.30 93.28
```

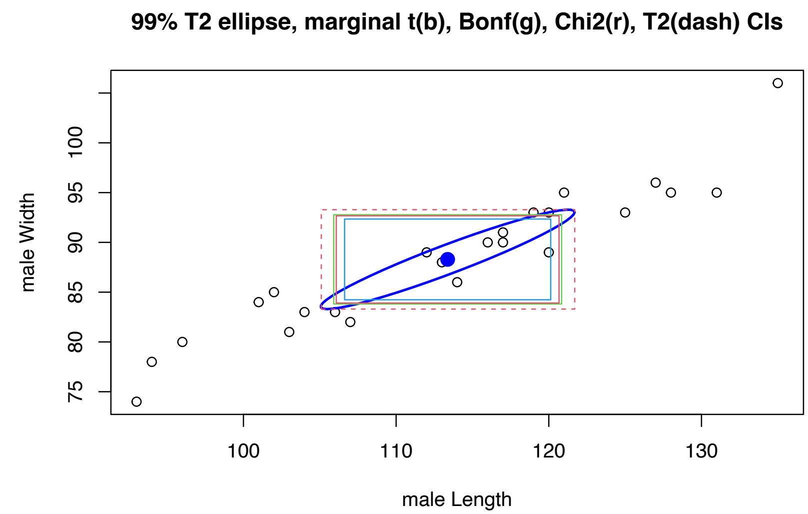
```
##     y     y
## 83.30 93.28
```

## Example: Comparison of C.R. and simult. C.I.'s at 99%

```
plot(x,y,xlab="male Length",ylab="male Width")
ellipse(c(xbar,ybar),shape=cov(male2),radius=.7059);
rect(x1,y2,x2,y1,border=4)  #blue, marginal t
rect(105.9112, 83.80957, 120.8388, 92.77377, border=3)
#green, Bonferroni
rect(106.08,83.91, 120.67, 92.67,border=2) #red,Chi
rect(105.1, 83.3, 121.7, 93.282,border=2, lty=2) #dash,T2
title("99% T2 ellipse, marginal t(b), Bonf(g),
      Chi2(r), T2(dash) CIs")
```



## Plot comparison of C.R. and simultaneous C.I.'s at 99%






## Sample R code

Find the 99% C.I.'s for component means using R function cregion:

(courtesy of Prof. R. Tsay)

```
# Find the length of 99% C.I.
source("cregion.R")
cregion(male2,alpha=.01)
```



```
## [1] "C.R. based on T^2"
##       [,1]   [,2]
## [1,] 105.1 121.69
## [2,]  83.3  93.28
## [1] "CR based on individual t"
##        [,1]   [,2]
## [1,] 106.62 120.13
## [2,]  84.24  92.35
## [1] "CR based on Bonferroni"
##        [,1]   [,2]
## [1,] 105.91 120.84
## [2,]  83.81  92.77
## [1] "Asymp. simu. CR"
##        [,1]   [,2]
## [1,] 106.08 120.67
## [2,]  83.91  92.67
```

‘‘