Data Analysis: TD multivariate normal correction

Christophe Ambroise

Section 1

Multivariate normal distribution (Exercices)

IQ

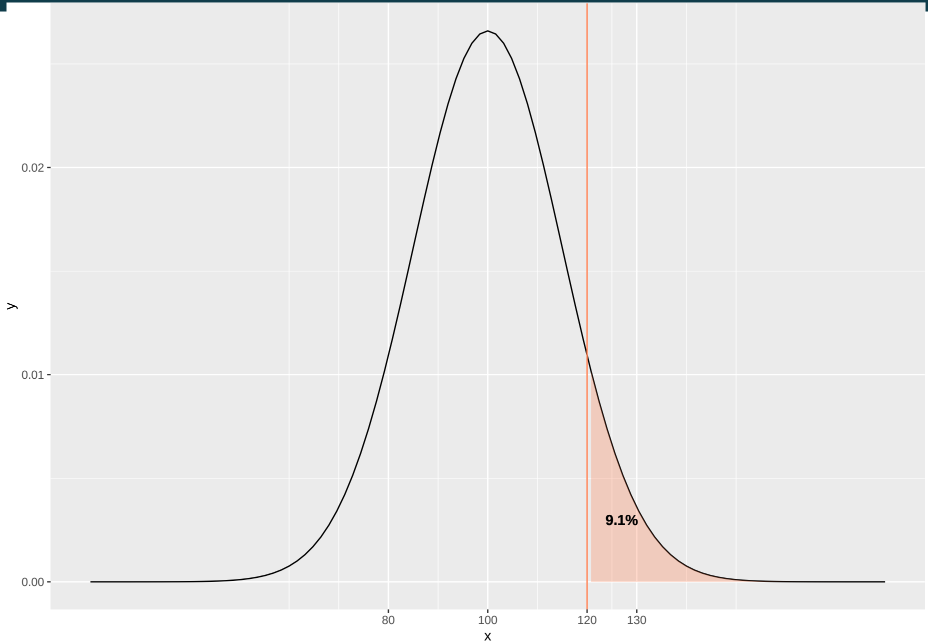
Knowing that IQ is a normal measure of mean 100 and standard deviation 15, what is the probability of having an IQ

- more than 120?
- less than 100?

IQ (Solution) I

```r
QI.sup.120<-function(x){
  ifelse(x>120,dnorm(x,mean=100,sd=15),NA)
  }
ggplot(data.frame(x=c(20, 180)),aes(x)) +
  stat_function(fun = dnorm,args = list(mean=100,sd=15)) +
  stat_function(fun =QI.sup.120 , geom = "area", fill = "coral", alpha = 0.3)
  geom_text(x = 127, y = 0.003, size = 4, fontface = "bold",
            label = paste0(round(pnorm(120,mean=100,sd=15,lower.tail = FALS
  scale_x_continuous(breaks = c(80,100,120,130)) +
  geom_vline(xintercept=120,colour="coral")
```

IQ (Solution) II



Bias of the maximum likelihood estimator of the variance

Show that the maximum likelihood estimator of the variance is biased and propose an unbiased estimator.

Solution

$$
\begin{aligned}
\mathbb{E}[\hat{\sigma}^2_{ml}] &= \mathbb{E}[\frac{1}{n}\sum_i x_i^2 - \bar{x}^2] \\
&= \sigma^2 + \mu^2 + \frac{\sigma^2}{n} - \mu^2
\end{aligned}
$$

Extreme values

Consider the Fisher irises. Find flowers whose measured widths and lengths are exceptionally large or small.

Solution {-} I

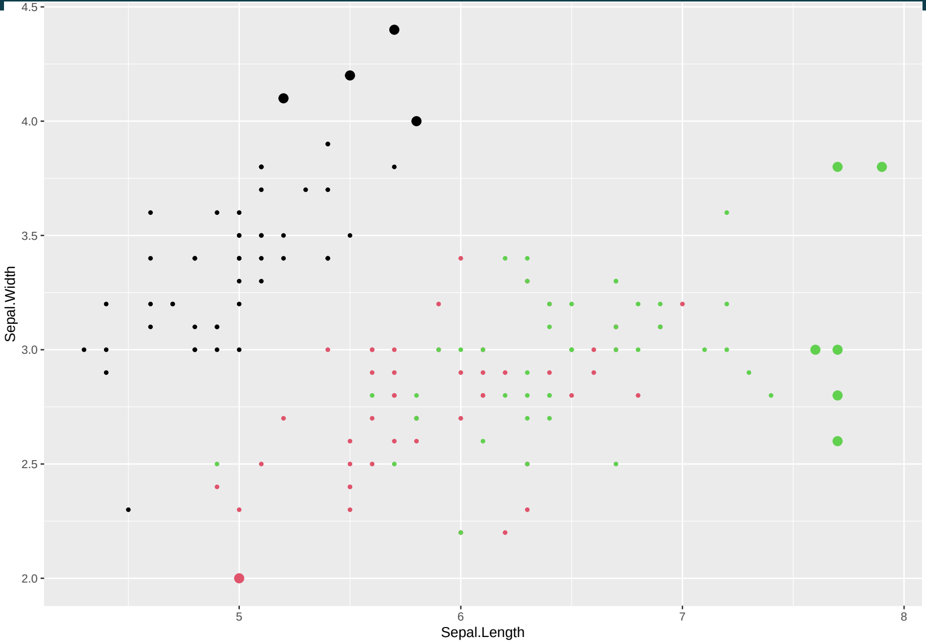
```
data(iris)
parameters <-
  as.tibble(iris) %>%
  select(-"Species") %>%
  gather(factor_key = TRUE)  %>%
  group_by(key) %>%
  summarise(mean= mean(value), sd= sd(value)) %>%
  mutate(min=mean - 2*sd,max=mean + 2*sd)
```

```
## Warning: `as.tibble()` is deprecated as of tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generate
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
flower.outliers  <-(apply(t((t(iris[,1:4]) < parameters$min) + (t(iris[,1:4]
ggplot(iris,aes(x=Sepal.Length,y=Sepal.Width))+
  geom_point(colour=as.numeric(iris$Species),size= flower.outliers*2 + 1 )
```

Solution {-} II



Equiprobability Ellipses I

- Generate 1000 observation of a two-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with
  - $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 0.75 \end{pmatrix}$
  - $\boldsymbol{\mu}^t = (0, 0)$
- Draw the ellipses of equiprobability of the multiples of 5%.

Solution {-} I

- Let $x^1, \ldots, x^p$ i.i.d. variables following $\mathcal{N}(0,1)$, then $= (x^1, \ldots, x^p)) \sim \mathcal{N}_p(0, I_p)$
- Find a matrix A of size (p,p) such that $A\boldsymbol{x}$ has variance Σ, i.e. $AA' = \Sigma$. Sevral solutions are possible - Cholesky : $\Sigma = T'T$ where T is triangular $(A = T')$ - SVD : $\Sigma = UDU'$ where D is a diagonal matrix of eigenvalues and U an orthogonal matrix of eigenvectors $(A = UD^{\frac{1}{2}})$
- then $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\mu} \sim \mathcal{N}_p(0, \Sigma)$

If $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ alors $\boldsymbol{y} = \Sigma^{-1/2}(\boldsymbol{x} - \mu) \sim \mathcal{N}_p(0, I_p)$ and

$$Q = \boldsymbol{y}^t \boldsymbol{y} \sim \chi^2_p$$

.

The equation

$$P(Q \leq q) = \alpha$$

with $q = \chi^2_{p,\alpha}$ defines an α level equiprobability ellipsoid .

Solution {-} II

```
par(mfrow=c(1,3)) # partage l'affichage en 2
Q<-qchisq(p=seq(0.05,0.95,by=0.1),df=2)
sigma<-matrix(c(2,1,1,0.75),2,2)
Y<-matrix(rnorm(2000),1000,2)%*%chol(sigma)
plot(Y,xlab="x",ylab="y",pch='.')
x<-seq(-4,4,length=100)
y<-seq(-4,4,length=100)
sigmainv<-solve(sigma)
a<-sigmainv[1,1]
b<-sigmainv[2,2]
c<-sigmainv[1,2]
z<-outer(x,y,function(x,y) (a*x^2+b*y^2+2*c*x*y))
image(x,y,z)
contour(x,y,z,col="blue4",levels=Q,labels=seq(from=0.05,to=0.95,by=0.1),add=
persp(x,y,1/(2*pi)*det(sigmainv)^(-1/2)*exp(-0.5*z),col="cornflowerblue",the
```

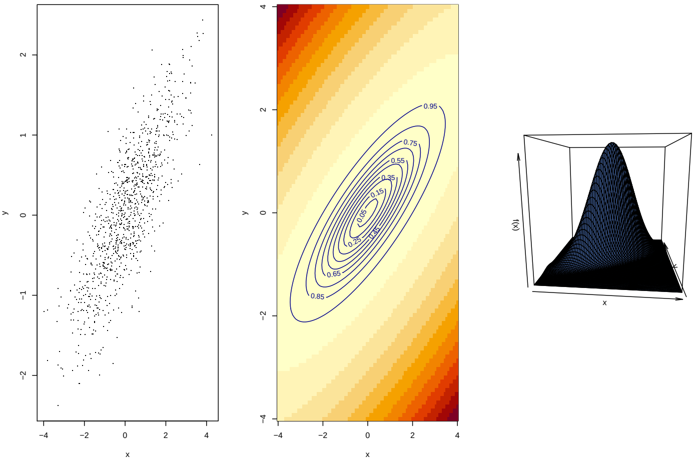


Figure 2: Ellipso<U+00EF>de d'<U+00E9>quiprobabilit<U+00E9> dans le plan

Limit between two bidimensional Gaussian

Simulate to Gaussian multivariate densities in 2d with respective mean vectors $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$

a. With the same covariance matrix $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 0.75 \end{pmatrix}$
b. With different covariance matrices $\Sigma_1 = \begin{pmatrix} 2 & 1 \\ 1 & 0.75 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Consider a mixture of the two densities in proportion π, $1 - \pi$ and draw the limit between the two posterior densities (where probabilities of being drawn from each component is equal) for diffent values of π.

Correction

The distribution if a mixture

$$f(\boldsymbol{x}) = \pi f_1(\boldsymbol{x}) + (1 - \pi) f_2(\boldsymbol{x}).$$

The posterior of the first class is

$$p(\boldsymbol{x}|k = 1) = \frac{\pi f_1(\boldsymbol{x})}{f(\boldsymbol{x})}$$

The equation to use for the contour line is

$$\log p(\boldsymbol{x}|k = 1) = \log p(\boldsymbol{x}|k = 2)$$