

# (JC)2BIM 2018 Research School

G. Rigai

June, 2018

Introduction

Probability survival kit

Estimation

Confidence intervals

Hypothesis testing

Multiple testing (a quick introduction)

# Statistical Inference

(JC)2BIM 2018 Research School

Guillem Rigail



# Introduction

# A short introduction

1. More and more data. . .
  - ▶ Collecting, Analyzing and Interpreting data ?
2. Statistical reasoning
  - ▶ Is now a fundamental part of experimental science

# Studying a population

## 1. One often make statements like:

- ▶ this gene is downregulated in lung cancer
- ▶ in France the price of 1 kg of apple rised by 5 cents last year
- ▶ 99% of the seeds in these bags are viable

## 2. In most of these cases

- ▶ the population we are (implicitly) taking about is very large
- ▶ collecting data is time consuming, costly and possibly it destroys the object
- ▶ our measurements are inherently noisy

# Studying a population (inference)

Hence the data we collect on this population are not “perfect”.

- ▶ How can we make statements about the whole population ?
- ▶ We need assumptions about the way data point were collected
- ▶ Those assumptions should be known and explicit
- ▶ These assumptions are formulated mathematically as a model
- ▶ Draw a schematic representation of this. . .

# Understanding statistical reasoning (1)

## 1. Cooking recipe level

- ▶ if the data is such and such do this and this. . .
- ▶ apply the code instructions of a vignette/tutorial online

## 2. **Applied statistics**

- ▶ understand statistical models
- ▶ understand how to assess whether a method is valid (or not) to infer a model

this is understandable using mostly highschool mathematics and some R coding



# Understanding statistical reasoning (2)

## 3. **Apprentice statistician**

- ▶ understand mathematical and algorithmic techniques behind statistical methods

for simple models this is understandable using mostly highschool mathematics and basic algorithmics

## 4. Master statistician...

## Probability survival kit

# Outline

1. **Probability**
2. Expectation, Variance and Covariance

# Probability Space

## Informal definition

- ▶  $\Omega$  the set of all possible outcomes
- ▶  $F$  a set of subsets of  $\Omega$ , an  $\omega$  in  $F$  is called an event
- ▶  $p$  a function from  $F$  to  $[0, 1]$ 
  - ▶  $p(\Omega) = 1$
  - ▶ For two disjoint events  $\omega_1, \omega_2$ , i.e.  $\omega_1 \cap \omega_2 = \emptyset$ ,

$$p(\omega_1 \cup \omega_2) = p(\omega_1) + p(\omega_2)$$

- ▶ more generally  $p$  is countably additive

# Probability Space, a few examples and R

## 1. Throw of a coin

```
Omega <- c("Head", "Tail")  
sample(Omega, 1)
```

```
## [1] "Head"
```

```
sample(Omega, 4, replace=TRUE)
```

```
## [1] "Tail" "Head" "Head" "Head"
```

# Probability Space, a few examples and R

## 2. Throw of a dice

```
Omega <- c("As", "2", "3", "4", "5", "6")  
sample(Omega, 1)
```

```
## [1] "6"
```

```
sample(Omega, 4, replace=TRUE)
```

```
## [1] "3" "2" "6" "4"
```

# Probability Space, a few examples and R

## 3. Coordinates of an arrow

```
radius <- 2  
dist_center <- runif(0, radius)  
angle <- runif(1, min=0, max=360)
```

# Probability Space, a few examples and R

## 4. Expression of a gene in an RNAseq experiment

```
## Poisson  
rpois(n=1, lambda=100)
```

```
## [1] 93
```

```
## or Neg-Binomial  
rnbinom(n=1, prob = 1/2, size = 40)
```

```
## [1] 51
```



## Some usefull properties

- For an event  $\omega$

$$p(\bar{\omega}) = p(\Omega \setminus \omega) = 1 - p(\omega)$$

- For two events  $\omega_1, \omega_2$

$$p(\omega_1 \cup \omega_2) = p(\omega_1) + p(\omega_2) - p(\omega_1 \cap \omega_2)$$

# Independence and conditionnal probability

1.  $\omega_1$  is independent of  $\omega_2$  if

$$p(\omega_2 \cap \omega_1) = p(\omega_1)P(\omega_2)$$

2. For an event  $\omega_1$  with  $p(\omega_1) > 0$  we define the conditionnal probability  $p(\omega_2|\omega_1)$  as

$$p(\omega_2|\omega_1) = p(\omega_1 \cap \omega_2)/p(\omega_1)$$

**Note:** If  $\omega_2$  is independent of  $\omega_1$  then  $P(\omega_2|\omega_1) = P(\omega_2)$

# Random Variables

## Definition

$Y$  is a function from  $\Omega$  to a space  $Def(Y)$

- ▶ Typically  $Def(Y)$  is
  - ▶ the set of integers
  - ▶ the set of real numbers
- ▶ We have:

$$p(Y \in S) = p(\{\omega \in F \mid Y(\omega) \in S\})$$

# Random Variables

## Some examples

- ▶  $Y$  a binary variable - throw of a coin
- ▶  $Y$  an integer smaller than 6 - a throw of a six face dice
- ▶  $Y$  a real number - distance of a javelin throw
- ▶  $Y$  an integer - expression of a gene in an RNAseq experiment
- ▶ ...

# Independence and random variables

## Definition

Two random variables  $Y_1$  and  $Y_2$  are independent if for all  $y_1$  in  $Def(Y_1)$  and  $y_2$  in  $Def(Y_2)$  we have

$$p(Y_1 = y_1 \cap Y_2 = y_2) = p(Y_1 = y_1)p(Y_2 = y_2)$$

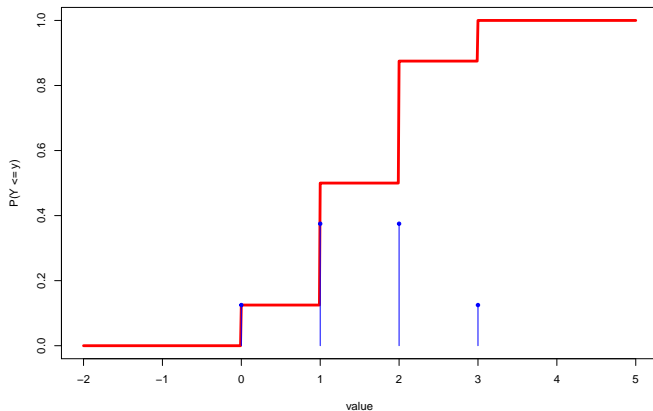
# Probability and cumulative probability for discrete variables

0. We call  $Def(Y)$  the discrete set of values taken by  $Y$  (e.g.  $\{0, 1\}$ ,  $\mathbb{N}$ )
1. For any  $y$  in  $Def(Y)$  we have access to  $p(Y = y) = p(y)$
2. We define the cumulative distribution function as  $P(Y \leq y)$ .

$$P(Y \leq y) = \sum_{\substack{y' \leq y \\ y' \in Def(Y)}} p(y')$$

# Probability and cumulative probability for discrete variables

3. A graphical example Binomial with parameter  $n = 3$  and  $\pi = 0.5$



# Density and cumulative probability for absolutely continuous random variable

For continuous variable we can proceed fairly similarly:

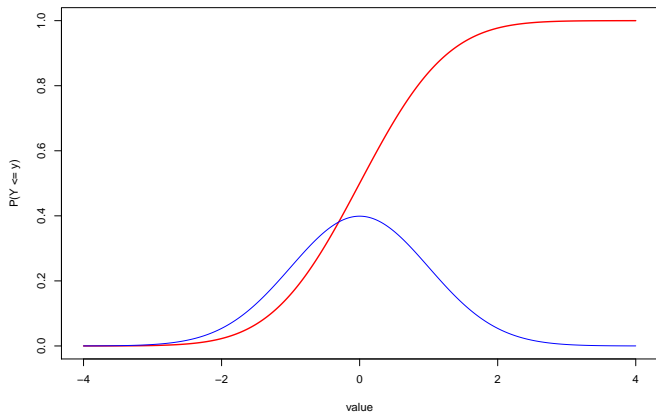
0. Take  $\text{Def}(Y) = \mathbb{R}$  the set of values taken by  $Y$
1. For any  $y$  in  $\text{Def}(Y)$  we have a continuous density function  $p(y)$  (or  $f(y)$ )
2. We define the cumulative distribution function as  $P(Y \leq y)$  as

$$P(Y \leq y) = \int_{y' \leq y} p(y') dy'$$



# Density and cumulative probability for a Gaussian

3. A graphical example with  $p(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2}$



# Simulation exercises

# Simulating simple random variables (Bernoulli)

Throw of a coin or Bernoulli variable:

- ▶  $Y = 0$  with probability  $\pi$
- ▶  $Y = 1$  with probability  $1 - \pi$

```
## One throw  
rbinom(n=1, prob=0.5, size=1)
```

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

## Simulating simple random variables (Bernoulli)

```
## 10^4 independent throws Y_1, Y_2, Y_3...  
Y <- rbinom(n=10^4, prob=0.5, size=1)  
table(Y)
```

```
## Y  
##      0      1  
## 4979 5021
```

## Simulating simple random variables (Binomial)

```
## 10^4 independent throws Y_1, Y_2, Y_3...  
Y <- rbinom(n=10^4, prob=0.5, size=5)  
table(Y)
```

```
## Y  
##      0      1      2      3      4      5  
## 277 1536 3195 3168 1537 287
```

# Simulating simple random variables (Normal)

Throw of a Normal o variable:

- ▶  $Y$  takes continuous values  $\mathcal{N}(\mu, \sigma^2)$
- ▶ the density is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

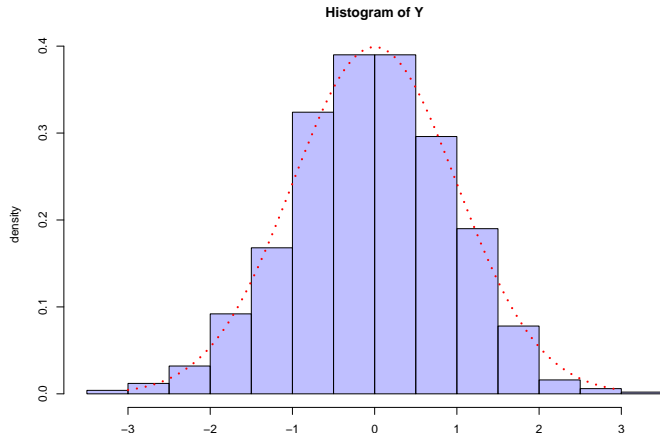
```
## One throw
```

```
rnorm(n=1, mean=0, sd=1)
```

```
## [1] -1.242426
```

# Simulating simple random variables (Normal)

```
## 10^4 independent throws Y_1, Y_2, Y_3...  
Y <- rnorm(n=10^3, mean=0, sd=1)  
x <- seq(-3, 3, by=0.01)  
hist(Y, col=rgb(0,0,1,1/4), freq=FALSE, ylab="density")  
lines(x, dnorm(x), col="red", lty=3, lwd=3)
```



# Homework: simulating a Poisson random variables

Throw of a Poisson:

- ▶  $Y$  takes integer values  $\mathcal{P}(\lambda)$
- ▶ density  $p(y) = \frac{\lambda^y e^{-\lambda}}{y!}$

```
## One throw
```

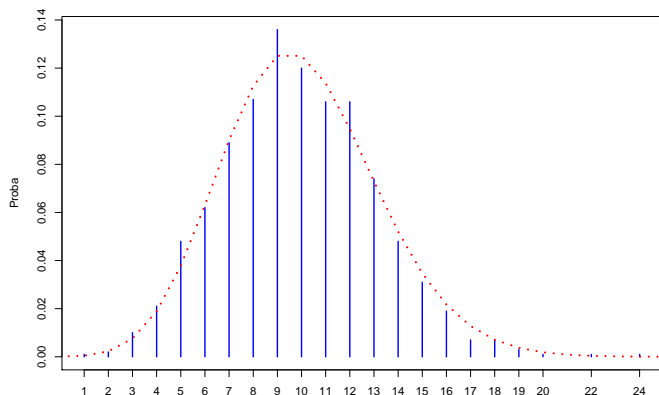
```
rpois(n=1, lambda = 10)
```

```
## [1] 14
```



## Homework: simulating a Poisson random variables

```
## 10^4 independent throws Y_1, Y_2, Y_3...  
Y <- rpois(n=10^3, lambda=10)  
x <- 0:40  
plot(table(Y)/10^3, col="blue", ylab="Proba")  
lines(x, dpois(seq(0, 40), lambda=10),  
      col="red",lty=3, lwd=3)
```





# Outline

1. Probability
2. **Expectation, Variance and Covariance**

# Expectation

## Definition

1. For discrete variables with probability  $p$

$$E(Y) = \sum_{y \in \text{Def}(Y)} yp(y)$$

2. Similarly for absolutely continuous variables with a density  $p$

$$E(Y) = \int_{y \in \text{Def}(Y)} yp(y)dy$$

## Some expectations

1. Expectation of a Bernoulli of parameter  $\pi$
2. Expectation of a Binomial distribution of parameters  $\pi$  and  $n$ .
  - ▶ number of successes in  $n$  independent experiments

$$p(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

- ▶ a bit too difficult for now...
3. Expectation of a Normal distribution of parameters  $\mu$  and  $\sigma^2$

## Expectation is linear

1. For two random variables  $Y_1, Y_2$ :

$$E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$$

2. For a constant  $c$  and a random variable  $Y_1$ :

$$E(cY_1) = cE(Y_1)$$

3. For two random variables:  $E(Y_2) = E(E(Y_2|Y_1))$

## Some expectations

1. Expectation of a Binomial distribution of parameters  $\pi$  and  $n$

$$p(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

## Some exercices on the expectation



## Ex: on the expectation (1a)

1. Expectation of the sum of  $10^4$  throws of a dice.

We have  $n$  r.v  $Y_1, \dots, Y_n$  taking value in  $\{1, 2, 3, 4, 5, 6\}$ .

- ▶ We have  $E(Y_i) = \sum_{i=1}^6 \frac{i}{6} = 3.5$
- ▶ And so we get

$$E(\sum_{i=1}^n Y_i) = \sum_{i=1}^n E(Y_i) = nE(Y_1) = 3.5 \times 10^4$$

## Ex: on the expectation (1b)

Using simulations and assuming the throws are independent

```
exper <- replicate(10^4, sum(  
  sample.int(6, 10^4, replace=TRUE)) )  
mean(exper)
```

```
## [1] 35001.99
```

## Ex: on the expectation (2)

2. Expectation of the average of  $10^4$  throws of a dice.

$$E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \frac{n}{n} E(Y_1) = 3.5$$

Using simulations and assuming the throws are independent

```
exper <- replicate(10^4, mean(  
  sample.int(6, 10^4, replace=TRUE)) )  
mean(exper)
```

```
## [1] 3.500124
```

## Homework: on the expectation (1)

3. Expectation of  $6(Y_1 - 1) + Y_2 - 1$  where  $Y_1$  and  $Y_2$  correspond to the throws of two dices.

$$E(6(Y_1 - 1) + Y_2 - 1) = 6E(Y_1) - 6 + E(Y_2) - 1 = 17.5$$

Using simulations and assuming the throws are independent

```
exper <- replicate(10^4, sum( (  
  sample.int(6, 2, replace=TRUE)-1) * c(6, 1)))  
mean(exper)
```

```
## [1] 17.3615
```

## Homework: on the expectation (2a)

- ▶ Consider  $Y_1$  a gaussian r.v. with parameters  $\mu_1 = 0$  and  $\sigma_1^2 = 1$ .
- ▶ Given  $Y_1$  the r.v.  $Y_2$  is gaussian with parameters  $\mu_2 = y_1$  and  $\sigma_2^2 = 1$

4. What is the expected value of  $Y_2$ ?

$$E(Y_2) = E(E(Y_2|Y_1)) = E(Y_1) = 0$$

## Homework: on the expectation (2b)

Using simulations and assuming independence

```
n  <- 10^4  
Y1 <- rnorm(n)  
Y2 <- rnorm(n=n, mean=Y1)  
mean(Y2)
```

```
## [1] 0.007585588
```



# Variance

## Definition

$$V(Y) = E((Y - E(Y))^2) = E(Y^2) - E(Y)^2$$

- Intuitively what does it represent ?



# Variance

## Properties

1. For two **independent** random variables  $Y_1$  and  $Y_2$ :

$$V(Y_1 + Y_2) = V(Y_1) + V(Y_2)$$

2. For a constant  $c$  and a random variable  $Y_1$ ,

$$V(cY_1) = c^2 V(Y_1)$$

3. For two random variables:

$$V(Y_2) = E(V(Y_2|Y_1)) + V(E(Y_2|Y_1))$$

## Some exercices on the variance

## Ex: on the variance (1)

1.  $X_1$  and  $-X_2$  are independant so:

$$V(X_1 - X_2) = V(X_1 + (-X_2)) = V(X_1) + V(X_2)$$

## Ex: on the variance (2)

2. Variance of the sum of  $10^4$  independent throws of a dice.

- ▶ We have  $n$  r.v  $Y_1, \dots, Y_n$  taking value in  $\{1, 2, 3, 4, 5, 6\}$ .
- ▶ We have

$$V(Y_i) = E(Y_i^2) - E(Y_i)^2 = \frac{1 + 4 + 9 + 16 + 25 + 36}{6} - 3.5^2 = \frac{35}{12}$$

As the throws are independent we have

$$V\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n V(Y_i) = nV(Y_1) = \frac{35}{12} \times 10^4$$

## Ex: on the variance (3)

- ▶ Using simulations and assuming the throws are independent.

```
exper <- replicate(10^4, sum(  
  sample.int(6, 10^4, replace=TRUE)) )  
var(exper)
```

```
## [1] 29566.77
```

## Homework: on the variance (4)

3. Variance of the average of  $10^4$  independent throws of a dice.

► Using independence:

$$V\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(Y_i) = \frac{n}{n^2} V(Y_1) = \frac{35}{12n}$$

► Using simulations and assuming the throws are independent

```
exper <- replicate(10^4, mean(  
  sample.int(6, 10^4, replace=TRUE)) )  
var(exper)
```

```
## [1] 0.000293675
```

## Homework: on the variance (5)

4. Expectation of  $6(Y_1 - 1) + Y_2 - 1$  where  $Y_1$  and  $Y_2$  correspond to the throws of two dices.

► Using independence

$$V(6(Y_1-1)+Y_2-1) = 6^2 V(Y_1) + V(Y_2) = 3 \times 35 + \frac{35}{12} = 35 \times (3 + \frac{1}{12})$$

Using simulations and assuming the throws are independent

```
exper <- replicate(10^4, sum( (
  sample.int(6, 2, replace=TRUE)-1)*c(6, 1) ))
var(exper)
```

```
## [1] 108.4646
```

## Homework: on the variance (6)

- ▶ Consider  $Y_1$  a gaussian r.v. with parameters  $\mu_1 = 0$  and  $\sigma_1^2 = 1$ .
- ▶ Given  $Y_1$ , the r.v.  $Y_2$  is gaussian with parameters  $\mu_2 = y_1$  and  $\sigma_2^2 = 1$

5. What is the variance of  $Y_2$  ?

$$V(Y_2) = E(V(Y_2|Y_1)) + V(E(Y_2|Y_1)) = E(1) + V(Y_1) = 1 + 1 = 2$$



## Homework: on the variance (7)

Using simulations and assuming independence

```
n  <- 10^4  
Y1 <- rnorm(n)  
Y2 <- rnorm(n=n, mean=Y1)  
var(Y2)
```

```
## [1] 2.019936
```



# Covariance

## Definition

$$\text{Cov}(Y_1, Y_2) = E((Y_1 - E(Y_1))(Y_2 - E(Y_2)))$$

$$\text{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2)$$

- ▶ What is  $\text{Cov}(Y_1, Y_1)$
- ▶ If  $Y_1$  and  $Y_2$  are independent ?
- ▶ Intuitively what does the covariance represent ?

# Covariance

## Properties

### 1. Covariance is bilinear:

- ▶ For two random variables  $Y_1, Y_2$ :  $Cov(Y_1, Y_2) = Cov(Y_2, Y_1)$
- ▶ For three random variables  $Y_1, Y_2, Y_3$ :  
 $Cov(Y_1 + Y_2, Y_3) = Cov(Y_1, Y_3) + Cov(Y_2, Y_3)$
- ▶ For a constant  $c$  and two random variable  $Y_1$ :  
 $Cov(cY_1, Y_2) = cCov(Y_1, Y_2)$

### 2. For three random variables

$$Cov(Y_1, Y_2) = E(cov(Y_1, Y_2|Y_3)) + cov(E(Y_1|Y_3)E(Y_2|Y_3))$$

## An exercise on the covariance

## Ex: on the covariance (1)

1. Consider  $Y_1$  a gaussian r.v. with parameter  $\mu_1 = 0$  and  $\sigma_1^2 = 1$ . Given  $Y_1$ , the r.v.  $Y_2$  is gaussian with parameters  $\mu_2 = y_1$  and  $\sigma_2^2 = 1$

What is the covariance of  $Y_2$  and  $Y_1$  ?

$$\text{Cov}(Y_1, Y_2) = E(\text{Cov}(Y_1, Y_2|Y_1)) + \text{Cov}(E(Y_1|Y_1), E(Y_2|Y_1))$$

$$\text{Cov}(Y_1, Y_2) = E(0) + \text{Cov}(Y_1, Y_1) = V(Y_1) = 1$$

## Ex: on the covariance (2)

Using simulations and assuming independence

```
n  <- 10^4  
Y1 <- rnorm(n)  
Y2 <- rnorm(n=n, mean=Y1)  
cov(Y1, Y2)
```

```
## [1] 1.011137
```





## Estimation

# Statistical inference

- ▶ A population (possibly infinite)
- ▶ Cannot do a census
- ▶ What can we say about the whole population given a sample
- ▶ We need assumptions = a model
- ▶ Small schema (population, sample, model, inference)

# Data

Given a sample of size  $n$

- ▶  $y_1, y_2 \dots y_n$
- ▶ Assume that they are realisations of  $n$  random variables

$$Y_1, Y_2, \dots, Y_n$$

# Modeling

## Model of the experiment

- ▶ Define the law of the r.v  $Y_1, \dots, Y_n$
- ▶ Sometimes it is difficult
- ▶ In simple cases one assumes that  $Y_i$  are i.i.d:

$$Y_i \sim \mathcal{P}(\theta)$$

with distribution  $p_\theta$

- ▶ Often  $\theta$  is the parameter we want to estimate.

# Estimator

An estimator is a function of  $Y_1, \dots, Y_n$ .

- ▶ It is a random variable
- ▶ A simple example:

$$\bar{Y} = \frac{\sum_i Y_i}{n}$$

- ▶ Propose an estimator for the variance of  $Y$  ?

# Estimation

Realisation of an estimator

- ▶ This is not a random variable
- ▶ For example

$$\bar{y} = \frac{\sum_i y_i}{n}$$

## An exercise

## Exercise: viscosity of a polymer

We have 4 viscosity measurements of a polymer used by a company to make microprocessors: 78, 85, 91, 76. For the polymer to be used we need that the viscosity is between 75 and 95

### **Exercise**

- ▶ Data?
- ▶ Model?
- ▶ Estimator?
- ▶ Estimation?



## Ex: viscosity of a polymer

► Data :  $y_1 = 78, y_2 = 85, y_3 = 91, y_4 = 76$

► Model

$$Y_i \sim \mathcal{N}(\mu, \sigma^2) \quad i.i.d$$

►  $\mu$  and  $\sigma^2$  are respectively the expected value and the variance

► Estimators

1. For the mean

$$\hat{\mu} = \bar{Y} = \frac{\sum Y_i}{n}$$

2. For the variance

$$\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{\mu})^2}{(n - 1)}$$

## Ex: viscosity of a polymer

- Estimation

```
y <- c(78, 85, 91, 76)
mean(y); var(y); sd(y)
```

```
## [1] 82.5
```

```
## [1] 47
```

```
## [1] 6.855655
```

- The mean is indeed in [75, 95] but the variance seems quite large...



## Quality of an estimator (1a)

1. The mean Squared Error (MSE)

$$E((\hat{\theta}_n - \theta)^2) = MSE(\hat{\theta}_n)$$

## Quality of an estimator (1b)

### 1. The mean Squared Error (MSE)

```
## 1 simu
Y <- mean(rt(10, df=4))
theta_hat_obs <- mean(Y)
mse_obs <- (theta_hat_obs - 0)^2

## if n=10
mean(replicate(10^3, mean(rt(10, df=4))^2))

## [1] 0.2150205

# if n=100
mean(replicate(10^3, mean(rt(100, df=4))^2))

## [1] 0.01985605
```

## Quality of an estimator (2a)

Lets try to decompose the error:

$$\hat{\theta}_n - \theta = \hat{\theta}_n - E(\hat{\theta}_n) + E(\hat{\theta}_n) - \theta$$

2. The expectation of the second part is called the bias

$$E(\hat{\theta}_n) - \theta = \textit{Bias}(\hat{\theta}_n)$$

## Quality of an estimator (2b)

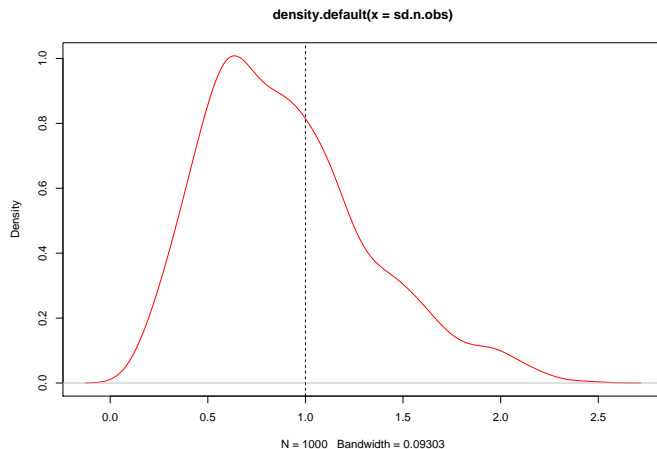
```
## a estimator with a bias
sd.n <- function(n){
  Y <- rnorm(n)
  sum((Y-mean(Y))^2)/n
}

sd.n.obs <- (replicate(1000, sd.n(10)))
mean(sd.n.obs)

## [1] 0.9043383
```

## Quality of an estimator (2c)

```
plot(density(sd.n.obs), col="red", lwd=1)  
abline(v=1, lty=2)
```





## Quality of an estimator (3)

3. The expectation of the first part is called the variance

$$E((\hat{\theta}_n - E(\hat{\theta}))^2) = V(\hat{\theta}_n)$$

- An estimator is a r.v

## Quality of an estimator (3)

4. It can be shown that

$$MSE(\hat{\theta}_n) = E((\hat{\theta}_n - \theta)^2) = Bias(\hat{\theta}_n)^2 + V(\hat{\theta}_n)$$

- ▶ Inferring a very complex model (without a little bias) is not necessarily better than inferring a simpler model (with larger bias)
- ▶ Variance counts.

# Quality of the empirical mean estimator

We consider a sample of size  $n$  :  $y_1, \dots, y_n$ . We assume

- ▶  $Y_i$  are i.i.d
- ▶  $E(Y_i) = \theta$
- ▶  $V(Y_i) = \sigma^2$

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n Y_i$$

# Quality of the empirical mean estimator

## 1. Bias

Using the linearity of the expectation

$$E(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \theta$$

- ▶ On average we do not make any mistake.

# Quality of the empirical mean estimator

## 2. Variance

Using the independence:

$$V(\hat{\theta}_n) = \frac{1}{n^2} \sum_{i=1}^n V(Y_i) = \frac{\sigma^2}{n}$$

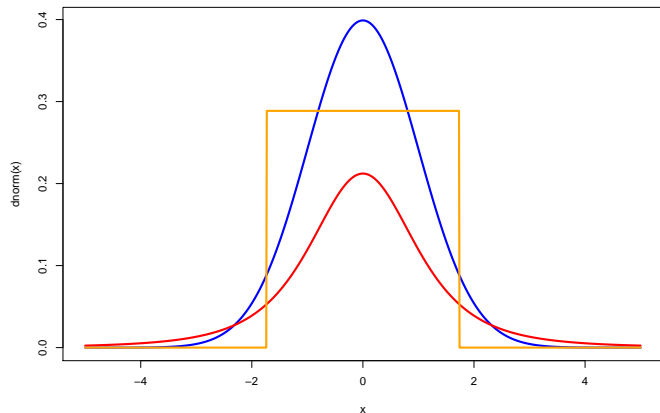
- ▶ On average we are not too far from  $\theta$
- ▶ On average we are closer if  $\sigma$  is smaller
- ▶ On average we are closer if we have more data.

## Quality of the empirical mean estimator

- ▶ Knowing the mean and variance of a distribution is usefull but not particularly precise.
- ▶ We would like to know the distribution of  $\hat{\theta}_n$

## Distribution with same mean and variances

Consider the density of a Gaussian, Student and Uniform distribution with the same mean and variance.



## An exercise on the mean and median



## Exercise: mean or median ?

Consider a sample of size  $n$ . Assume with  $Y_i$  i.i.d.

Compare (using simulations) the Bias, Variance and MSE of the empirical mean and empirical median estimators

1. if the data are drawn from a Gaussian distribution

```
rnorm(3)
```

```
## [1]  1.2773410 -0.5756087  1.2022450
```

2. if the data are drawn from a Student distribution with a degree of freedom  $k = 3$  (heavy tail)

```
rt(3,df=3)
```

```
## [1] -1.2072950 -0.7701449  0.5161675
```

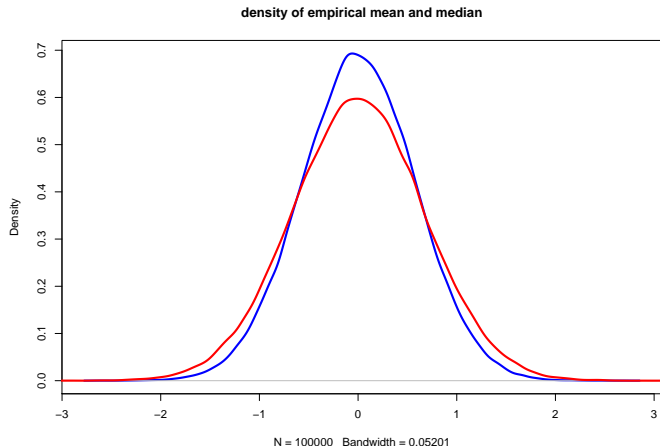
## Ex: Mean or median with a Gaussian

```
## simulation function
one.simu <- function(n){
  y <- rnorm(n)
  c(mean(y), median(y))
}

## replication
es <- t(replicate(10^5, one.simu(3)))
colnames(es) <- c("mean", "median")
```

## Ex: Mean or median with a Gaussian: distribution

```
plot(density(es[, 1]), col="blue", lwd=3,  
     main="density of empirical mean and median")  
lines(density(es[, 2]), col="red", lwd=3)
```



## Ex: Mean or median with a Gaussian: bias and variance

```
colMeans(es) ## Bias (compare to 0)
```

```
##          mean      median  
## 0.003787214 0.002776014
```

```
apply(es, 2, var) ## Variance
```

```
##          mean      median  
## 0.3340075 0.4505219
```

## Ex: Mean or median with a Gaussian: MSE

```
colMeans(es^2) ## MSE
```

```
##      mean      median  
## 0.3340185 0.4505251
```

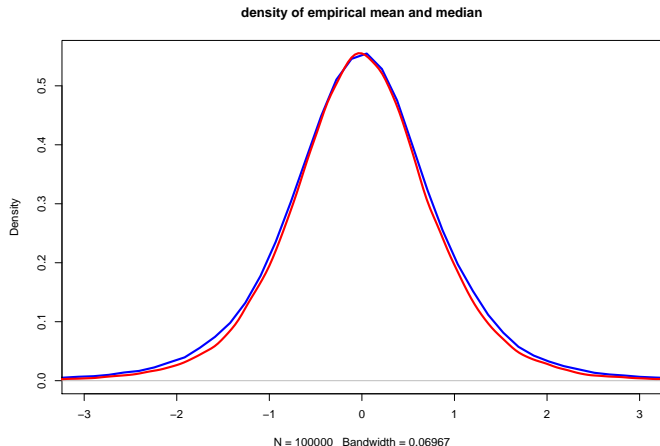
## Ex: Mean or median with a Student

```
k <- 3
one.simu <- function(n){
  y <- rt(n, df=k)
  c(mean(y), median(y))
}

es <- t(replicate(10^5, one.simu(3)))
colnames(es) <- c("mean", "median")
```

## Ex: Mean or median with a Student: distribution

```
plot(density(es[, 1]), col="blue", lwd=3,  
     main="density of empirical mean and median", xlim=c(-3, 3),  
     lines(density(es[, 2]), col="red", lwd=3))
```



## Ex: Mean or median with a Student : bias and variance

```
colMeans(es) ## Bias (compare to 0)
```

```
##          mean          median  
## -0.006067683 -0.003258026
```

```
apply(es, 2, var) ## Variance
```

```
##          mean          median  
## 1.0181870 0.7196179
```



## Ex: Mean or median with a Student : MSE

```
colMeans(es^2) ## MSE
```

```
##      mean      median  
## 1.0182136 0.7196213
```

## Homework: mean or median ?

Consider a sample of size  $n$ . Assume with  $Y_i$  i.i.d.

Compare (using simulations) the Bias, Variance and MSE of the empirical mean and empirical median estimators

3. if the data are drawn from a  $\chi^2$  distribution with a degree of freedom  $k = 5$

```
rchisq(3, df=5)
```

```
## [1] 0.4439805 10.4333834 6.1365275
```

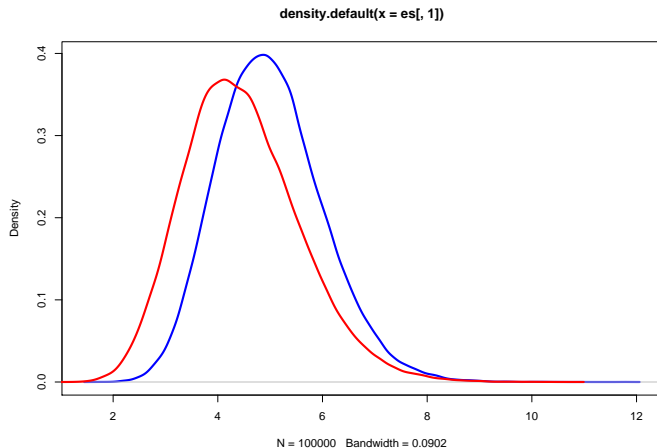
## Homework: Mean or median with a $\chi^2$

```
k <- 5
one.simu <- function(n){
  y <- rchisq(10, df=k)
  c(mean(y), median(y))
}

es <- t(replicate(10^5, one.simu(3)))
colnames(es) <- c("mean", "median")
```

## Homework: Mean or median with a $\chi^2$ : distribution

```
plot(density(es[, 1]), col="blue", lwd=3)  
lines(density(es[, 2]), col="red", lwd=3)
```



## Homework: Mean or median with a $\chi^2$ : distribution

- ▶ Looking at wikipedia we found that the mean and median of a  $\chi^2$  are not equal...
  1. The expectation is equal to the degree of freedom  $k$
  2. The median is close to  $k * (1 - 2/(9 * k))^3$
- ▶ So in fact we are not even trying to estimate the same thing...

## Homework: Mean or median with a $\chi^2$ : bias and variance

```
mean(es[, 1])-k ## mean bias
```

```
## [1] -0.0004150696
```

```
mean(es[, 2])-k*(1-2/(9*k))^3 ## median bias
```

```
## [1] 0.09734625
```

```
apply(es, 2, var) ## Variance
```

```
##      mean  median
```

```
## 1.004363 1.209067
```

## Homework: Mean or median with a $\chi^2$ : MSE

```
mean((es[, 1] - k)^2) ## mean
```

```
## [1] 1.004353
```

```
mean((es[, 1] - k*(1-2/(9*k))^3 )^2) ## median
```

```
## [1] 1.4102
```





## Homework exercise: Sampling and estimation

- ▶ Consider  $Y_1$  a random variable with a Poisson distribution of parameter  $\lambda_1 = 10$
  - ▶ Knowing  $Y_1 = y_1$   $Y_2$  is a Poisson random variable of parameter  $\lambda_2 = y_1$
1. What is the expectation and variance of  $Y_2$  ?
  2. What is the covariance of  $Y_1$  and  $Y_2$  ?

## H-Ex: Sampling and estimation an exercise

3. Estimate the expectation, variance and covariance using sampling.

```
n <- 10^3  
Y1 <- rpois(n, lambda=10)  
Y2 <- rpois(n, lambda = Y1)  
mean(Y2)      # using math we know that  $E(Y2) = 10$ 
```

```
## [1] 10.154
```

```
var(Y2)        # using math we know that  $V(Y2) = 20$ 
```

```
## [1] 19.99428
```

```
cov(Y1, Y2)    # using math we know that  $Cov(Y1, Y2) = 10$ 
```

```
## [1] 9.947644
```

## H-Ex: Estimation of $E(Y_2) = \lambda_2$

We try to estimate  $E(Y_2)$  using  $\sum_{i=1}^n Y_{2,i}/n = \hat{\lambda}_2$ .

**Quality of the estimator  $\hat{\lambda}_2$  ?**

4. Bias ?
5. Variance ?
6. Distribution ?

## H-Ex: Estimator $\hat{\lambda}_2$

### 4. Bias ?

We already checked that  $E(Y_2) = E(\hat{\lambda}_2)$

- ▶ No bias = on average we do not make any mistake
- ▶ This doesn't tell us anything about the magnitude of our mistakes

### 5. What is the variance $V(\hat{\lambda}_2)$

after some calculations we get  $V(\hat{\lambda}_2) = V(Y_2)/n$

- ▶ On average we are not too far
- ▶ Still we would like to quantify the error more precisely (probability)

## H-Ex: Distribution of $\hat{\lambda}_2$

6. What is the distribution of  $\hat{\lambda}_2$  ?

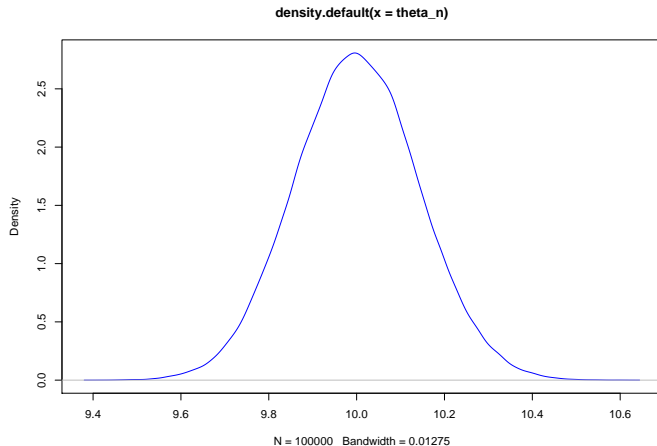
- ▶ It looks a bit difficult mathematically (harder than for the expectation or variance at least).
- ▶ But using simulations ...

```
one.rep <- function(n=10^3){  
  Y1 <- rpois(n, lambda=10)  
  Y2 <- rpois(n, lambda = Y1)  
  return(mean(Y2))  
}  
  
theta_n = replicate(10^5, one.rep())
```

## H-Ex: Distribution of $\hat{\lambda}_2$

3. What is the distribution of  $\hat{\lambda}_2$

```
plot(density(theta_n), col="blue")
```



## H-Ex: To continue at home

7. Compare the density you get for larger and smaller  $n$
8. Consider an estimator of  $V(Y_2)$ . Use simulations to get an idea of the bias, variance and distribution of this estimator.
9. Consider an estimator of  $Cov(Y_2, Y_1)$ . Use simulations to get an idea of the bias, variance and distribution of this estimator.





# How do we get formula for estimators ?

- ▶ For the mean it is fairly natural to take the empirical mean.
- ▶ For the variance it is fairly natural to take the empirical variance.
- ▶ For the covariance it is fairly natural to take the empirical covariance.
- ▶ How do you get estimators for more “complex” parameters ?
- ▶ Call a statistician
- ▶ Many more or less generic approaches
  1. Method of Moments
  2. Minimum Mean square error
  3. Maximum likelihood
  4. Bayesian inference

# A brief introduction to the maximum likelihood approach

The likelihood of a sample  $y_1, \dots, y_n$  and of parameters  $\theta$  is defined as

$$\mathcal{V}(y_1, \dots, y_n, \theta) = p_{\theta}(Y_1 = y_1, \dots, Y_n = y_n)$$

Assuming all the  $Y_i$  are i.i.d

$$\mathcal{V}(y_1, \dots, y_n, \theta) = \prod_{i=1}^n p_{\theta}(Y_i = y_i)$$

# The log-Likelihood

The likelihood of a sample  $y_1, \dots, y_n$  and of parameters  $\theta$  is defined as

$$\mathcal{V}(y_1, \dots, y_n, \theta) = p_{\theta}(Y_1 = y_1, \dots, Y_n = y_n)$$

Assuming all the  $Y_i$  are i.i.d and taking the log

$$\mathcal{L}(y_1, \dots, y_n, \theta) = \sum_{i=1}^n \log(p_{\theta}(Y_i = y_i))$$

## An example with Bernoulli variables

Assume  $Y_i$  are i.i.d Bernoulli variables of parameter  $\pi$

- ▶  $p_\theta(Y_i = 0) = 1 - \pi$
- ▶  $p_\theta(Y_i = 1) = \pi$
- ▶  $n_1$  the number of  $y_i$  equal to 1

$$\mathcal{V}(y_1, \dots, y_n, \theta) = \prod_{i=1}^n p_\theta(Y_i = y_i) = \pi^{n_1} (1 - \pi)^{n - n_1}$$

- ▶ taking the log

$$\mathcal{L}(y_1, \dots, y_n, \theta) = (n_1) \log(\pi) + (n - n_1) \log(1 - \pi)$$

## An example with Gaussian variables

Assume  $Y_i$  are i.i.d Gaussian variables of parameter  $\mu$  and  $\sigma$

$$\blacktriangleright p_{\mu,\sigma}(Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i - \mu)^2}{2\sigma^2}}$$

$$\mathcal{V}(y_1, \dots, y_n, \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i - \mu)^2}{2\sigma^2}}$$

► taking the log

$$\mathcal{L}(y_1, \dots, y_n, \theta) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2$$

# Maximum likelihood ?

1. Given some sample, what value of  $\theta$  should we take ?
2. Idea: take  $\theta$  that maximise the log-likelihood
3. The log-likelihood is used as a measure of fit to the data

## **Why should we do this:**

1. Fairly generic (as soon as you have a model)
2. In a number of cases ML has good statistical properties (asymptotically unbiased and Gaussian. . . )

## Maximum likelihood for $n$ i.i.d Bernoulli r.v.

Maximization of the likelihood for  $n$  i.i.d Bernoulli variables

- ▶ Idea: derivative of  $\mathcal{L}$  as a function of  $\pi$

## Visually for $n$ i.i.d Bernoulli r.v. (1)

```
Y <- rbinom(n=20, size=1, prob=0.5)
mean(Y)
```

```
## [1] 0.7
```

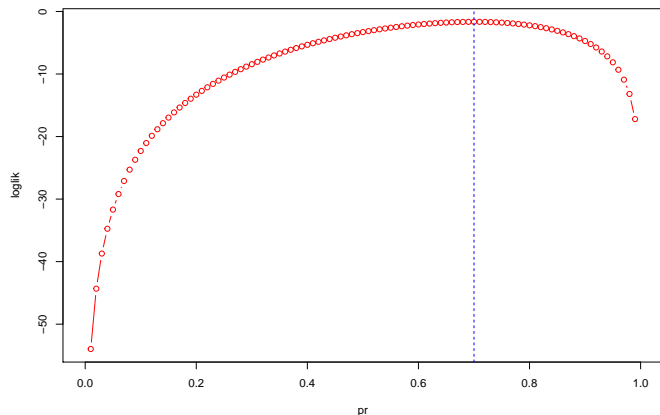
```
pr <- seq(0, 1, by=0.01)
loglik <- dbinom(sum(Y), size=20, prob=pr, log=TRUE)
pr[which.max(loglik)]
```

```
## [1] 0.7
```



## Visually for $n$ i.i.d Bernoulli r.v. (2)

```
plot(pr, loglik, col="red", type="b")  
abline(v=mean(Y), lty=2, col="blue")
```



An exercise maximum likelihood of a Gaussian r.v

## Ex: Maximum likelihood for a Gaussian r.v. (1)

Maximization of the likelihood for  $n$  i.i.d Gaussian variables

- ▶ Idea: derivative of  $\mathcal{L}$  as a function of  $\mu$  and  $\sigma^2$

## Ex: Visually for $n$ i.i.d Gaussian r.v. (2)

```
Y <- rnorm(n=10, mean=1)
mean(Y)
```

```
## [1] 1.229894
```

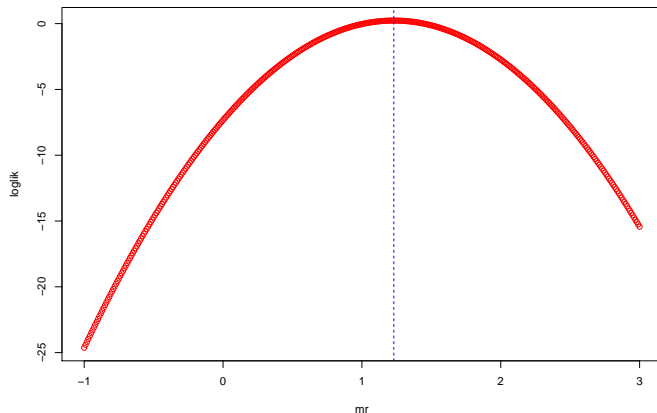
```
mr <- seq(-1, 3, by=0.01)
loglik <- dnorm(mean(Y), mean=mr, sd=1/sqrt(10), log=TRUE)

mr[which.max(loglik)]
```

```
## [1] 1.23
```

## Ex: Visually for $n$ i.i.d Gaussian r.v. (3)

```
plot(mr, loglik, col="red", type="b")  
abline(v=mean(Y), lty=2, col="blue")
```



## Confidence intervals

## Idea / Definition

**Idea:** Rather than giving one value for a parameter, we aim to give two bounds  $B_1$  and  $B_2$  and we hope that the true value is between the two

### 1. Random interval

**Definition:** Let  $B_1 = m(Y_1, \dots, Y_n)$  et  $B_2 = M(Y_1, \dots, Y_n)$  two r.v. We define a random interval for  $\theta$  with the couple  $(B_1, B_2)$ . We call  $P(B_1 < \theta < B_2)$  the level of confidence.

### 2. Confidence interval

**Definition:** A confidence interval at level  $1 - \alpha$  for  $\theta$  is a realisation  $[b_1, b_2]$  of a random interval with confidence  $1 - \alpha$

# Confidence interval for the mean knowing the variance (1)

- ▶ Data  $y_1, \dots, y_n$
- ▶ Estimator

$$\bar{Y} = \sum Y_i / n$$

- ▶ We suppose that  $V(Y_i) = \sigma^2$  is known
- ▶ Model for the estimator [using TCL]

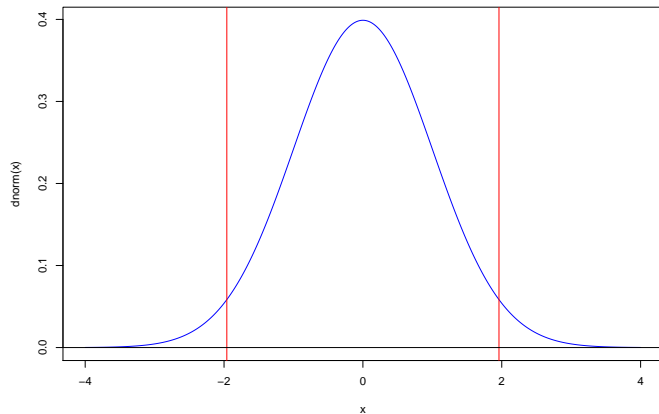
$$\bar{Y} \sim \mathcal{N}(\mu, \sigma^2 / \sqrt{n})$$

$$\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$$



## Confidence interval for the mean knowing the variance (2)

► Visually



## Confidence interval for the mean knowing the variance (3)

- So we have

$$P(u_{\frac{\alpha}{2}} \leq \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

- If we study the two inequalities

$$u_{\frac{\alpha}{2}} \leq \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{and} \quad \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}$$

- We get

$$\mu \leq \bar{Y} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{Y} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu$$

## Confidence interval for the mean knowing the variance (4)

- We get

$$P(\bar{Y} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$B_1 = \bar{Y} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad B_2 = \bar{Y} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- The interval is larger for
  - larger  $\sigma$
  - smaller  $\alpha$
  - smaller  $n$

Exercise: Check that a c.i returned by a given approach works reasonably well

## Exercise: First implement in R the previous c.i

Our two bounds are:

$$B_1 = \bar{Y} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad B_2 = \bar{Y} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

```
y <- rnorm(10) ## 0 mean and sd=1
n <- length(y)
theta.h <- mean(y)
alpha <- 0.05
b1 <- theta.h + qnorm(p=alpha/2)/sqrt(n)
b2 <- theta.h + qnorm(p=1-alpha/2)/sqrt(n)
```

## Exercise: Check using simulations that our c.i works reasonably well (1)

- ▶ What should we check exactly ?
- ▶ The claim is : the probability that the random interval with confidence level  $1 - \alpha$  contains the true mean is  $1 - \alpha$
- ▶ We got our c.i assuming

$$\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/\sqrt{n})$$

- ▶ Is it a reasonable assumption?

## Ex: Check that our c.i works reasonably well (2)

A function to do that simulating uniform  $Y_i$ :

For now we simulate uniform  $Y_i$

```
one.simu <- function(n=2, alpha=0.05){  
  y <- runif(n, -0.5, 0.5)*sqrt(12) ## 0 mean and sd=1  
  theta <- mean(y)  
  b1 <- theta + qnorm(p=alpha/2)/sqrt(n)  
  b2 <- theta + qnorm(p=1-alpha/2)/sqrt(n)  
  return((b1 < 0) & (0 < b2))  
}
```

## Ex: Check that our c.i works reasonably well (3)

A few test

```
## n=2, clearly not perfect  
res1 <- replicate(10^5, one.simu(2))  
mean(res1)
```

```
## [1] 0.96051
```

```
## n=10, fairly close  
res2 <- replicate(10^5, one.simu(10))  
mean(res2)
```

```
## [1] 0.95103
```



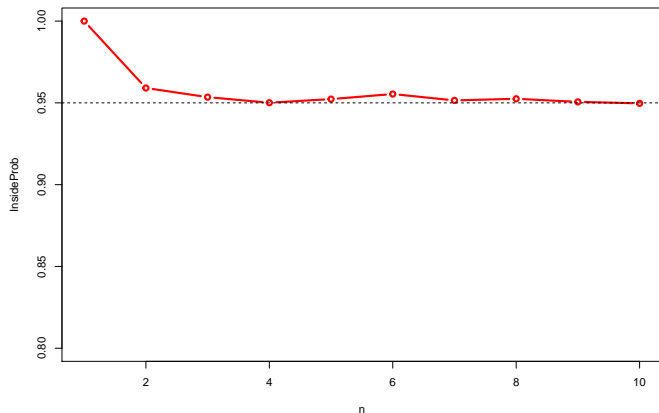
Ex: Check that our c.i works reasonably well (4)

If we now test for many values of  $n$

```
InsideProb <- numeric(10)
for(n in c(1:10)){
  InsideProb[n] <- mean( replicate(10^4, one.simu(n)) )
}
```

## Ex: Check that our c.i works reasonably well (5)

```
plot(InsideProb, type="b", lwd=3, col="red", xlab="n",  
     ylim=c(0.8, 1))  
abline(h=0.95, lty=2)
```



For large enough (in fact not so large) it works. With probability 95% the interval

## H-Ex: Check that our c.i works reasonably well (6)

- ▶ We used a uniform distribution for the  $Y_i$
- ▶ Test for two times a Bernoulli r.v. of parameter 0.5

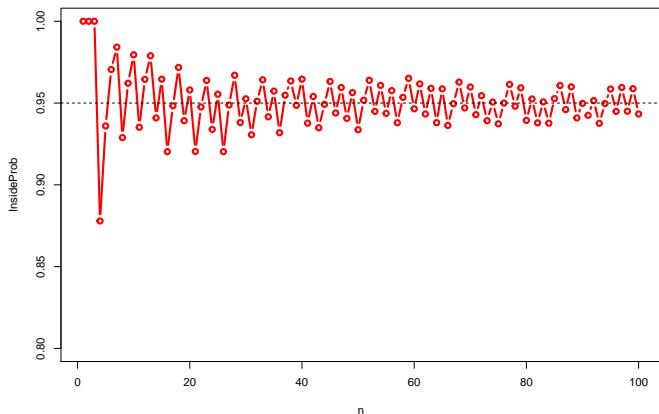
```
one.simu <- function(n=2, alpha=0.05){  
  y <- 2*rbinom(n=n, size=1, prob=0.5) ## 0 mean and sd=1  
  theta <- mean(y)  
  b1 <- theta + qnorm(p=alpha/2)/sqrt(n)  
  b2 <- theta + qnorm(p=1-alpha/2)/sqrt(n)  
  return((b1 < 1) & (1 < b2))  
}
```

H-Ex: Check that our c.i works reasonably well (7)

```
InsideProb <- numeric(100)
for(n in c(1:100)){
  InsideProb[n] <- mean( replicate(104, one.simu(n)) )
}
```

## H-Ex: Check that our c.i works reasonably well (8)

```
plot(InsideProb, type="b", lwd=3, col="red", xlab="n",  
     ylim=c(0.8, 1))  
abline(h=0.95, lty=2)
```





## Confidence interval for the mean not knowing the variance

- ▶ If  $\sigma$  is not known, similar calculations using the T distribution leads to a confidence interval:
- ▶ Namely we start from

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim \mathcal{T}_{n-1}$$

# Confidence interval for the mean not knowing the variance

## In R

```
x <- runif(10)
t.test(x)$conf.int
```

```
## [1] 0.1622967 0.6318180
## attr(,"conf.level")
## [1] 0.95
```



## Exercise and the student confidence interval

## Exercise: Student confidence interval

1. Check that the Student c.i “works” when simulating  $Y_i$  as

- ▶ independent Student r.v of degree 2.1 (in R `rt`).
- ▶ (HW) independent  $\chi^2$  r.v of degree 3 (in R `rchisq`)

2. Study the effect of  $n$ .

```
one.simu <- function(n=2, alpha=0.05){  
  y <- rt(n=n, df=2.1) ## 0 mean and sd=1  
  CI <- t.test(y)$conf.int  
  return((CI[1] < 0) & (0 < CI[2]))  
}
```

## Ex: Student c.i

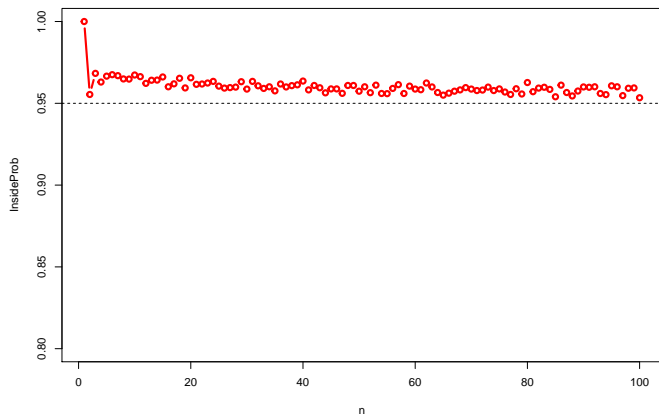
We test for various  $n$

```
InsideProb <- numeric(100)
InsideProb[1] <- 1
for(n in c(2:100)){
  InsideProb[n] <- mean( replicate(10^4, one.simu(n)) )
}
```

## Ex: Student c.i

We plot

```
plot(InsideProb, type="b", lwd=3, col="red", xlab="n",  
     ylim=c(0.8, 1))  
abline(h=0.95, lty=2)
```



## H-Ex: $\chi^2$ distribution

```
one.simu <- function(n=2, alpha=0.05){  
  y <- rchisq(n=n, df=5) ## 0 mean and sd=1  
  CI <- t.test(y)$conf.int  
  return((CI[1] < 5) & (5 < CI[2]))  
}
```

## H-Ex: $\chi^2$ c.i

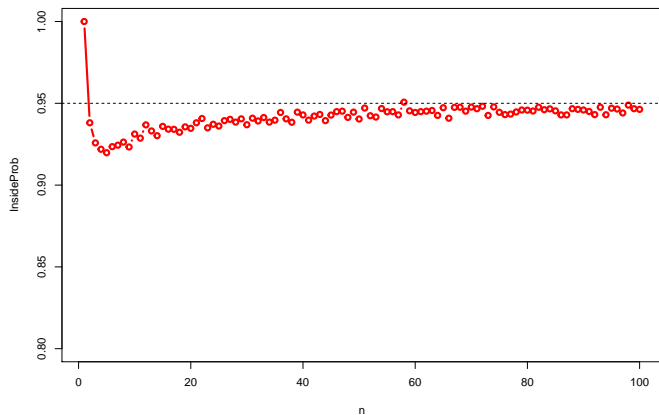
We test for various  $n$

```
InsideProb <- numeric(80)
InsideProb[1] <- 1
for(n in c(2:100)){
  InsideProb[n] <- mean( replicate(10^4, one.simu(n)) )
}
```

## H-Ex: $\chi^2$ c.i

We plot

```
plot(InsideProb, type="b", lwd=3, col="red", xlab="n",  
     ylim=c(0.8, 1))  
abline(h=0.95, lty=2)
```







# Confidence intervals

- ▶ Many statistical methods provide confidence intervals
- ▶ Computational or mathematical derivation of those c.i can be complex
- ▶ From an application point of view always the same principle

**Principle** A statistical model with some assumptions on the signal

1. Check that those assumptions are reasonable for your application
2. In doubt check using simulations that this is working

## Exercise on polymer

## Student interval

- Polymer viscosity

Is the viscosity of the polymer in the interval  $[75, 95]$  ?

```
## the data was  
y <- c(78, 85, 91, 76)  
t.test(y)$conf.int[1:2]
```

```
## [1] 71.59112 93.40888
```

What do we conclude ?



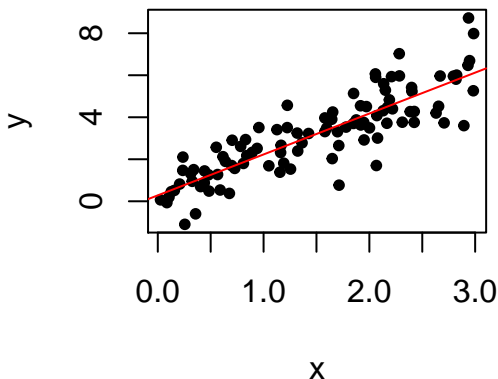
## An other example simple linear regression

- ▶ Linear regression

$$Y_i = \alpha x_i + \beta + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad i.i.d$$

```
model <- lm(y ~ x) ## Regression  
plot(x, y, pch=20); abline(model, col="red"); ## graphe
```



## Simple linear regression (2)

```
confint(model) ## IC à 95%
```

```
##                2.5 %    97.5 %  
## (Intercept) -0.1200145 0.6829409  
## x           1.7156597 2.1797113
```

- ▶ Can we conclude that the slope is different from 0?
- ▶ What about the origin ?

## A schematic view of what a 95% confidence interval does

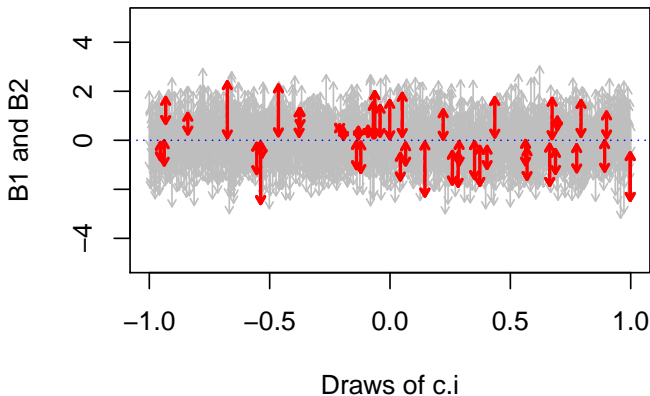


Figure 2: 200 confidence intervals

## Hypothesis testing



## A few examples of test-like questions

- ▶ Is the expression of gene HER2 large in breast cancer ?
- ▶ Is a new variety of tomato more resistant to mildew than the previous one ?
- ▶ Is a drug better than a placebo ?
- ▶ Should the +1 increased popularity of a candidate be commented ?

## When can you use a test ?

- ▶ A Yes/No question
- ▶ You have data
- ▶ These data can be considered as the result of some r.v. (known through a model)
- ▶ The question should be about a parameter of the distribution

# Outcome of a test

Only two possibilities:

1. Either you accept the hypothesis  $H_0$  (data are not in disagreement with your assumption)
2. Or you reject it (data are in disagreement with your assumption)

# Four elements of a test

1. Data  $y_1, \dots, y_n$  realisation of r.v.  $Y_1, \dots, Y_n$
2. A statistical model:
  - ▶ distribution of  $Y_1, \dots, Y_n$  depending on some parameters  $\theta$
3. An assumption:
  - ▶ A statement about  $\theta$ .
  - ▶ This is the so called  $H_0$  hypothesis ( $H_1$  is the alternative)
4. A decision rule
  - ▶ If  $T = f(X_1, \dots, X_n)$  is a test statistic
  - ▶  $R$  is subset of values for  $T$  that is improbable if  $H_0$  is true

## Four elements of a test

1. Data  $y_1, \dots, y_n$  realisation of r.v.  $Y_1, \dots, Y_n$
  2. A statistical model:
  3. An assumption:
  4. A decision rule
- 
- ▶ A test can be viewed as a probabilistic extension of “argument to absurdity”

## Efficiency of a test (intuition)

- ▶ Two types of error
  1. Reject  $H_0$  when it is true
  2. Keep  $H_0$  when it is false
- ▶ Typically it is not possible to control both of these errors at the same time.

## Efficiency of a test (intuition)

- ▶ Type I risk:  $\alpha$  the probability under  $H_0$  to reject  $H_0$
- ▶ Type II risk:  $\beta$  the probability under  $H_1$  to keep  $H_0$ .
- ▶ We call power  $\pi = 1 - \beta$
- ▶ Make a table. . .

# Construction of a test

- Fabrication of a 100 cl bottle

1. **Data**  $y_1, \dots, y_n$  some measurements realisation of r.v  $Y_1, \dots Y_n$ .

2. **Model**

$$Y_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{i.i.d}$$

3. **Hypothesis**

$$H_0 = \{\mu = 100\text{cl}\}$$

4. **Rule:**

- $T = \frac{1}{n} \sum Y_i$
- Reject if  $(T - 100)$  is too large:

$$\ell = u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$



## Construction of test (2)

- ▶ If all  $Y_i$  are normal and independent

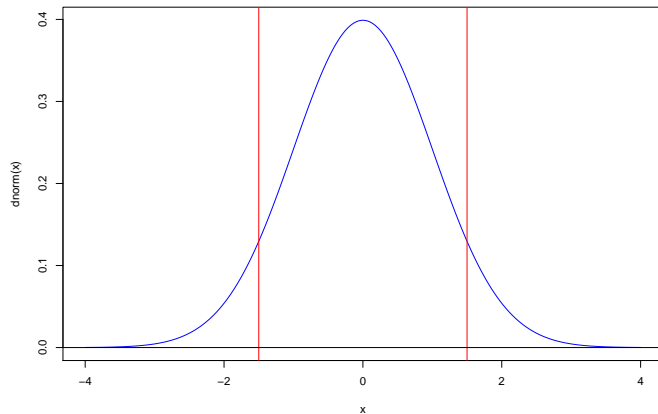
$$\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

- ▶ So under  $H_0$  (taking  $\mu_0 = 100$ )

$$P(u_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

## Construction of test (3)

- Visually, we observed  $\bar{y} = 1.5$



## Construction of a test (4)

- So we get

$$P\left(u_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

- If we study the two inequalities

$$u_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad \text{et} \quad \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}$$

- we get

$$\mu_0 + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{X} \quad \text{et} \quad \bar{X} \leq \mu_0 + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

# Construction of a test (5)

- Fabrication of a 100 cl bottle

1. **Data**  $y_1, \dots, y_n$  some measurements realisation of r.v  $Y_1, \dots Y_n$ .

2. **Model**

$$Y_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{i.i.d}$$

3. **Hypothesis**

$$H_0 = \{\mu = 100cl\}$$

4. **Rule:**

- $T = \frac{1}{n} \sum Y_i$
- Reject if  $(T - 100)$  is too large:

$$\ell = u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

# Implementation in R

- We know the variance and scale the data using the variance

```
## simulate some bottles
Y <- rnorm(10, mean=101, sd=2)
Y.scaled <- Y/2
mu0 <- 100/2

## probability of a value so far from 0 under H0
p.val <- 2*pnorm(abs(mean(Y)-mu0),
                  sd= 1/sqrt(10), lower.tail=FALSE)
```

## Exercise: Check that this is working ?

What should we check ?

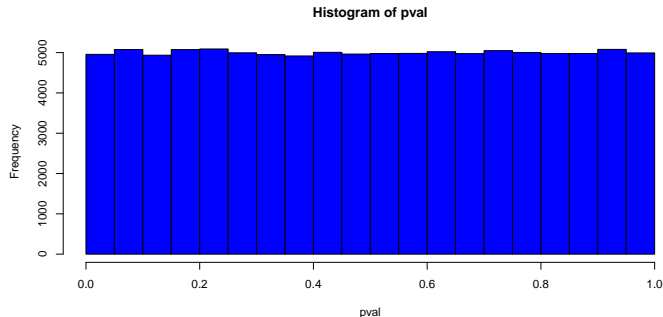
- ▶ control that we indeed control at level  $\alpha$
- ▶ control that under  $H_1$  we have some power

```
n.test <- function(Y.sc, mean.H0=0){  
  min(1, 2*pnorm(abs(mean(Y.sc-mean.H0)),  
                sd= 1/sqrt(length(Y.sc)),  
                lower.tail=FALSE) )  
}  
  
one.simu <- function(n, mean){  
  ## no-need to scale here (sd=1)  
  Y <- rnorm(n, mean=mean, sd=1)  
  return( n.test(Y, 0) )  
}
```

## Ex: Type I error control

- ▶ If we make  $n$  simulations under  $H_0$  a proportion  $\alpha$  of those experiments should have a p-value under  $\alpha$ .
- ▶ That is the p-values should be uniform

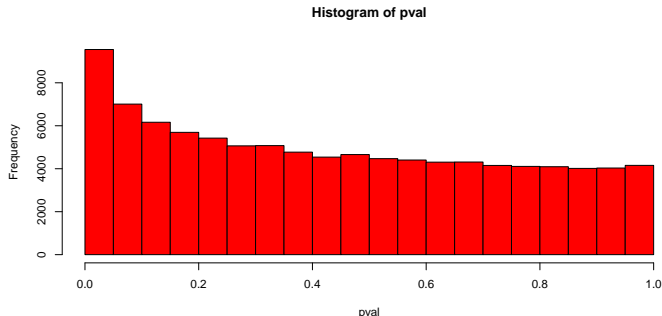
```
pval <- replicate(10^5, one.simu(10, 0))  
hist(pval, col="blue")
```



## Ex: Power

- If we make  $n$  simulations under  $H_1$  a proportion higher than  $\alpha$  of those experiments should have a p-values under  $\alpha$ .

```
pval <- replicate(10^5, one.simu(10, 0.2))  
hist(pval, col="red")
```





## Ex: Power

- ▶ At level  $\alpha = 0.05$  and  $\alpha = 0.01$  the power is

```
mean(pval <= 0.05)
```

```
## [1] 0.09549
```

```
mean(pval <= 0.01)
```

```
## [1] 0.02664
```

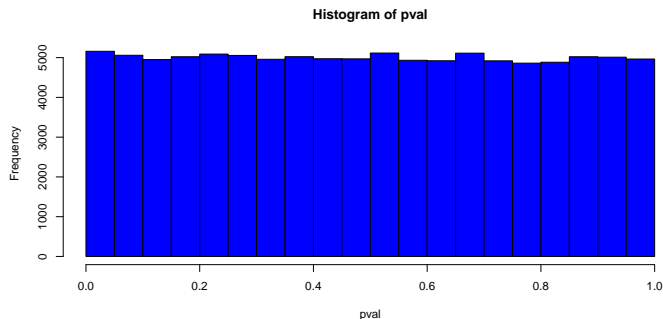
Ex:

- ▶ What happens when  $n$  increases ?

## Ex: Type I error control

- ▶ Larger  $n$
- ▶ The p-values should still be uniform

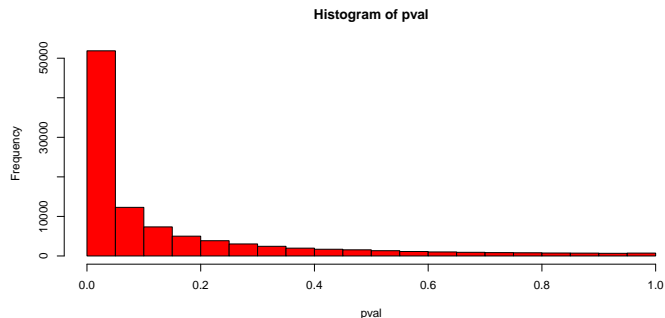
```
pval <- replicate(10^5, one.simu(100, 0))  
hist(pval, col="blue")
```



## Ex: Power

- We should have more power

```
pval <- replicate(10^5, one.simu(100, 0.2))  
hist(pval, col="red")
```



## Ex: Power

- At level  $\alpha = 0.05$  and  $\alpha = 0.01$  the power is

```
mean(pval <= 0.05)
```

```
## [1] 0.51867
```

```
mean(pval <= 0.01)
```

```
## [1] 0.28616
```

## Homework:

- ▶ What happens if you change the distribution of the  $Y_i$



## Test if $\sigma^2$ is not know

- ▶ If  $\sigma$  is not known, similar calculations using the student distribution with  $n - 1$  degrees of freedom lead to the famous T-test
- ▶ Namely we start from

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim \mathcal{T}_{n-1}$$

**In R**

```
Y <- runif(10, min=0, max=1)
t.test(Y)$p.value
```

```
## [1] 0.0001358614
```



## Many other test...

1. In general many statistical methods provide hypothesis testing
  2. Computational or mathematical derivation of those is often complex
  3. But same principle
- ▶ A statistical model: check that it is reasonable for your application
  - ▶ An  $H_0$  hypothesis: check that it adress your question
  - ▶ In doubt check using simulations that the test is working

## Two populations t-test with same variance

- ▶  $Y_{11}, \dots, Y_{1n}$  i.i.d with mean  $\theta_1$  and variance  $\sigma^2$
- ▶  $Y_{21}, \dots, Y_{2n'}$  i.i.d with mean  $\theta_2$  and variance  $\sigma^2$
- ▶  $H_0 \theta_1 = \theta_2$
- ▶  $H_1 \theta_1 \neq \theta_2$

Similar calculations lead to another student statistics. In R:

```
Y1 <- rnorm(10, sd=2)
Y2 <- rnorm(8, sd=2)

t.test(Y1, Y2, var.equal = TRUE)$p.value

## [1] 0.8927863
```

## Exercise: Power of the two sample t-test

## Ex: Power of two sample t-test

Two populations

- ▶  $Y_{11}, \dots, Y_{1n}$  i.i.d with mean  $\theta_1$  and variance  $\sigma^2$
- ▶  $Y_{21}, \dots, Y_{2n'}$  i.i.d with mean  $\theta_2$  and variance  $\sigma^2$

We fix  $N = n + n'$ .

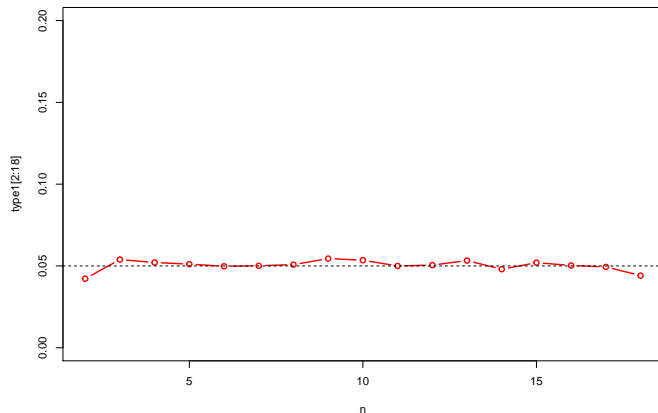
1. Check that the t-test is indeed controlling  $\alpha$
2. Assess the power of the t-test to detect a difference of 0.5 for an  $\alpha$  level of 5%

```
one.simu <- function(n=10, N=20, diff=0){  
  Y1 <- runif(n, min=0, max=1)  
  Y2 <- runif(N-n, min=0, max=1)+diff  
  t.test(Y1, Y2, var.equal = TRUE)$p.value  
}
```



## Ex: Type I error control

```
plot(2:18, type1[2:18], ylim=c(0, 0.2), type="b", xlab="n",  
     abline(h=0.05, lty=2))
```

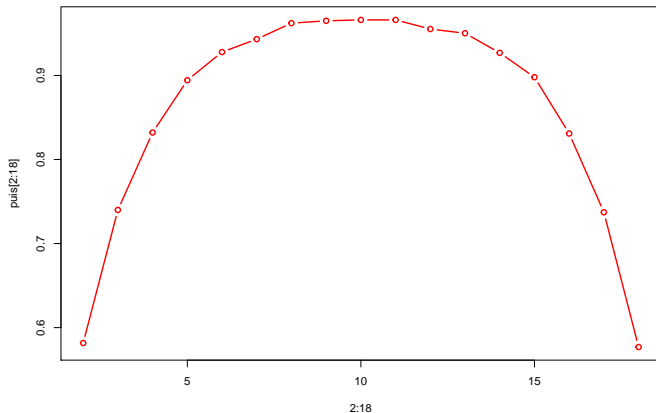


## Ex: Power

[illegible]

## Ex: Power

```
plot(2:18, puis[2:18], type="b", col="red", lwd=2)
```





## Homework: Compare the t-test and wilcoxon-test

Compare the following three following testin R for various normal distributed data and  $\chi^2$  distributed data. You should assess for various sample sizes:

1. The type I error control
2. The power to detect a mean difference of 0.5 at level  $\alpha = 0.01$

```
Y1 <- runif(10)
```

```
Y2 <- runif(12)
```

```
wilcox.test(Y1, Y2)$p.value
```

```
## [1] 0.5824145
```

```
t.test(Y1, Y2, var.equal=FALSE)$p.value
```

```
## [1] 0.6209135
```

```
t.test(Y1, Y2, var.equal=TRUE)$p.value
```

# H-Ex

## Simulation function

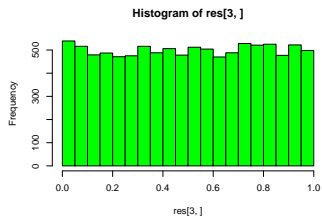
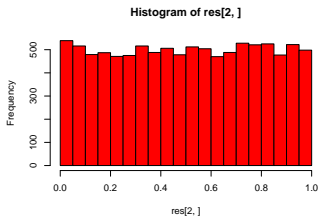
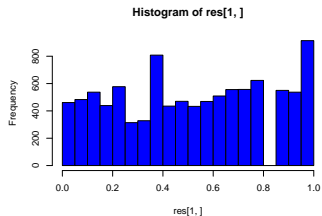
```
one.simu <- function(n1, n2, diff){  
  Y1 <- runif(n1, min=0, max=1)  
  Y2 <- runif(n2, min=0, max=1)+diff  
  p1 <- wilcox.test(Y1, Y2)$p.value  
  p2 <- t.test(Y1, Y2, var.equal = TRUE)$p.value  
  p3 <- t.test(Y1, Y2, var.equal = TRUE)$p.value  
  return(c(p1, p2, p3))  
}
```

## H-Ex: Type I error control (a)

```
res <- replicate(104, one.simu(10, 10, 0))
```

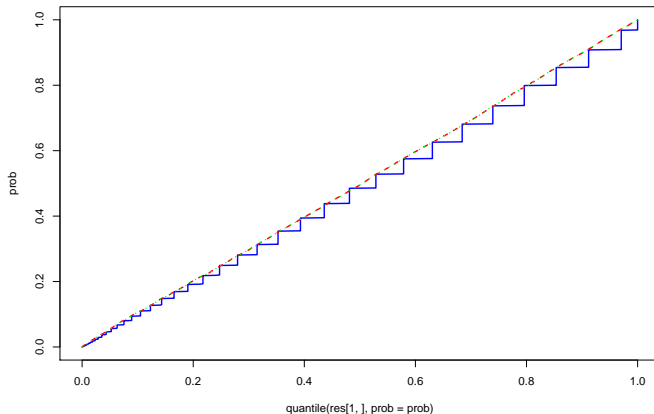
## H-Ex: Type I error control (b)

```
par(mfrow=c(2, 2))  
hist(res[1, ], col="blue")  
hist(res[2, ], col="red")  
hist(res[3, ], col="green")
```



## H-Ex: Type I error control (c)

```
prob <- seq(0, 1, by=0.001)
plot(quantile(res[1, ], prob=prob), prob, type="l", col="b")
lines(quantile(res[2, ], prob=prob), prob, type="l", col="r")
lines(quantile(res[2, ], prob=prob), prob, type="l", col="g")
```



## H-Ex: Power

```
res <- replicate(10^4, one.simu(10, 10, 0.5))  
rownames(res) <- c("Wil", "T.E", "T.I")  
## for alpha at 0.05  
rowMeans(res <=0.05)
```

```
##      Wil      T.E      T.I  
## 0.9175 0.9683 0.9683
```

```
## for alpha at 0.01  
rowMeans(res <=0.01)
```

```
##      Wil      T.E      T.I  
## 0.7370 0.8324 0.8324
```

## Multiple testing (a quick introduction)

# Multiple testing in genomics

- ▶ Differential analysis : one test per gene
- ▶ ChipSeq : one test per window
- ▶ GWAS : one test per SNP



# Why performing many tests is a problem?

Suppose you are performing  $G$  tests at level  $\alpha$ .

$$P(\text{at least one FP if } H_0 \text{ is always true}) = 1 - (1 - \alpha)^G$$

- ▶ Ex: for  $\alpha = 5\%$  and  $G = 20$ ,

$$P(\text{at least one FP if } H_0 \text{ is always true}) \simeq 64$$

- ▶ This probability increases with the number of test  $G$
- ▶ For more than 75 tests
- ▶ if  $H_0$  is always true the probability to have at least one false positive is very close to 100%!

## Error Rate for $G$ tests

Instead of the risk  $\alpha$ , control:

- ▶ the Family-Wise Error Rate:  $\text{FWER} = \mathbb{P}(U > 0)$ 
  - ▶ probability to have at least one false positive decision
- ▶ the False Discovery Rate:  $\text{FDR} = \mathbb{E}(Q)$  with

$$Q = \begin{cases} U/R & \text{if } R > 0 \\ 0 & \text{otherwise} \end{cases}$$

## Adjusted p-values

Settings: p-values  $p_1, \dots, p_G$  (*{e.g.}, corresponding to  $G$  tests*)

**Adjusted p-values** *adjusted p-values are  $\tilde{p}_1, \dots, \tilde{p}_G$  such that*

*Rejecting tests such that  $\tilde{p}_g < \alpha$*

*is equivalent to*

$$P(U > 0) \leq \alpha \quad \text{or} \quad \mathbb{E}(Q) \leq \alpha$$

# Calculating adjusted p-values

1. order the p-values  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(G)}$
2. calculate  $\tilde{p}_{(g)} = a_g p_{(g)}$ 
  - ▶ with Bonferroni method:  $a_g = G$  (FWER)
  - ▶ with Benjamini and Hochberg method:  $a_g = G/g$  (FDR)
3. if  $\tilde{p}_{(g)}$  is larger than 1 replace it by 1

# Implementation in R

We simulate 1000 test under  $H_0$ .

```
pval <- replicate(1000, t.test(rnorm(10))$p.value)

## adjustment
fdr <- p.adjust(pval, method="BH")
bfr <- p.adjust(pval, method="bonferroni")
```

## Exercise: Check that the “BH” approach is working reasonable well

- ▶ What should we do ?
- ▶ For a given threshold  $\alpha$  check that the average proportion of false positive is indeed less than  $\alpha$ .

```
one.simu <- function(n0=90, n1=10, p=20,  
                    alpha=0.05, meanH1=0.5){  
  pval0 <- replicate(n0, t.test(rnorm(p))$p.value)  
  pval1 <- replicate(n1, t.test(rnorm(p)+meanH1)$p.value)  
  pval <- c(pval0, pval1)  
  padj <- p.adjust(pval, method="BH")  
  sum(padj[1:n0] <= alpha) / max(1, sum(padj <= alpha))  
}
```

Ex: Check that “BH” is working reasonable well (1)

One simulation:

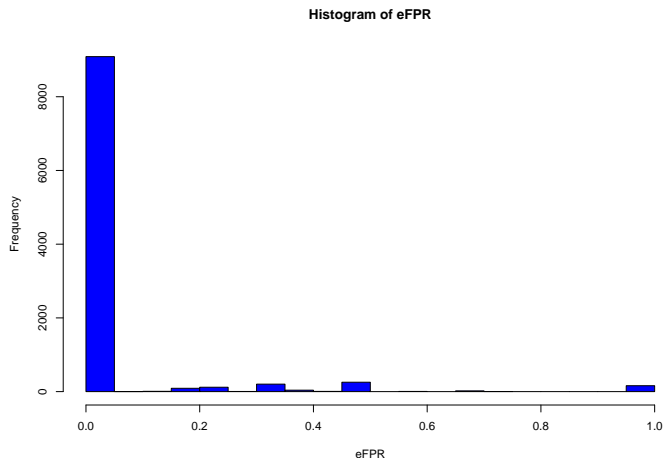
```
eFPR <- replicate(10^4, one.simu())  
mean(eFPR)
```

```
## [1] 0.04402909
```

**We do control the FPR on average.**

Ex: Check that “BH” is working reasonable well (2)

```
hist(eFPR, col="blue")
```



Sometimes we are a bit unlucky...



Ex: Check that “BH” is working reasonable well (3)

For various proportion of H1 and H0

```
res <- lapply(10*1:9, FUN=function(i)
  replicate(10^3, one.simu(n0=i, n1=100-i)))

mat <- do.call(cbind, res)
colnames(mat) <- paste0("n0=", 10*1:9)
```

## Ex: Check that “BH” is working reasonable well (4)

On average we get

```
signif(colMeans(mat), 1)
```

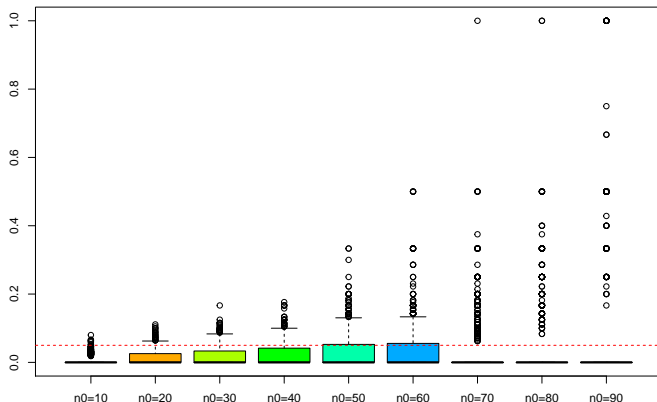
```
## n0=10 n0=20 n0=30 n0=40 n0=50 n0=60 n0=70 n0=80 n0=90  
## 0.005 0.010 0.020 0.020 0.030 0.030 0.030 0.030 0.050
```

- ▶ **We indeed control the FDR**
- ▶ Our control is not always tight.

## Ex: Check that “BH” is working reasonable well (5)

In details things are even more complex:

```
boxplot(mat, col=rainbow(9));  
abline(h=0.05, lty=2, col="red")
```



Homework: Check that “bonferroni” is working reasonable well

## Conclusion FDR, BH and beyond. . .

- ▶ There are other approaches
- ▶ possibly more complex mathematically
- ▶ see for example:  
<https://mathforgenomics.github.io/neuvial.pdf>