# Introduction to Statistical Tests (part 2)

**Emmanuel Jeannot**

INRIA
LaBRI

January, 2024

# Outline

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)
  - Compute a quantile from a probability (e.g. `q*` functions)

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)
  - Compute a quantile from a probability (e.g. `q*` functions)
  - Draw numbers according to a given law (e.g. `r*` functions))

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)
  - Compute a quantile from a probability (e.g. `q*` functions)
  - Draw numbers according to a given law (e.g. `r*` functions))
- The CDF and the quantile are used to solve $\mathbb{P}(X \leq C) = p$ :

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)
  - Compute a quantile from a probability (e.g. `q*` functions)
  - Draw numbers according to a given law (e.g. `r*` functions))
- The CDF and the quantile are used to solve $\mathbb{P}(X \leq C) = p$ :
  - If you know $C$ and need to compute $p$, use the CDF function. Ex. $X \sim \mathcal{N}(0, 1)$ and $C = 2$

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)
  - Compute a quantile from a probability (e.g. `q*` functions)
  - Draw numbers according to a given law (e.g. `r*` functions))
- The CDF and the quantile are used to solve $\mathbb{P}(X \leq C) = p$ :
  - If you know $C$ and need to compute $p$, use the CDF function. Ex. $X \sim \mathcal{N}(0, 1)$ and $C = 2$
  - then $p =$`pnorm(2, mean = 0, sd = 1)`.

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)
  - Compute a quantile from a probability (e.g. `q*` functions)
  - Draw numbers according to a given law (e.g. `r*` functions))
- The CDF and the quantile are used to solve $\mathbb{P}(X \leq C) = p$ :
  - If you know $C$ and need to compute $p$, use the CDF function. Ex. $X \sim \mathcal{N}(0, 1)$ and $C = 2$
  - then $p =$`pnorm(2, mean = 0, sd = 1)`.
  - if you know $p$ and need to compute $C$, use the quantile function. Ex. $X \sim \mathcal{N}(0, 1)$ and $p = 0.9$

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)
  - Compute a quantile from a probability (e.g. `q*` functions)
  - Draw numbers according to a given law (e.g. `r*` functions))
- The CDF and the quantile are used to solve $\mathbb{P}(X \leq C) = p$ :
  - If you know $C$ and need to compute $p$, use the CDF function. Ex. $X \sim \mathcal{N}(0, 1)$ and $C = 2$
  - then $p =$`pnorm(2, mean = 0, sd = 1)`.
  - if you know $p$ and need to compute $C$, use the quantile function. Ex. $X \sim \mathcal{N}(0, 1)$ and $p = 0.9$
  - then $C =$`qnorm(0.9, mean = 0, sd = 1)`.

# Summary of last session

## Cumulative Distribution Function vs Quantile

- You need four functions:
  - Draw the density function (e.g. `d*` functions)
  - Compute a probability with the CDF (e.g. `p*` functions)
  - Compute a quantile from a probability (e.g. `q*` functions)
  - Draw numbers according to a given law (e.g. `r*` functions))
- The CDF and the quantile are used to solve $\mathbb{P}(X \leq C) = p$ :
  - If you know $C$ and need to compute $p$, use the CDF function. Ex. $X \sim \mathcal{N}(0, 1)$ and $C = 2$
  - then $p =$`pnorm(2, mean = 0, sd = 1)`.
  - if you know $p$ and need to compute $C$, use the quantile function. Ex. $X \sim \mathcal{N}(0, 1)$ and $p = 0.9$
  - then $C =$`qnorm(0.9, mean = 0, sd = 1)`.
  - Remark : the $\alpha^{\text{th}}$ quantile of the standard normal –$\mathcal{N}(0, 1)$– is called $z_\alpha$. E.g. if $X \sim \mathcal{N}(0, 1)$ then: $\mathbb{P}(X \leq C) = p \Leftrightarrow C = z_p$.

# Summary of last session

## Important

- The sample mean, $\bar{x}$ estimates the population mean, $\mu$.
- The sample standard deviation, $S$ estimates the population standard deviation, $\sigma$.
- S, the standard deviation, talks about how variable the population is.
- When $n$ is large enough (CLT) ($X_n = \sum_n X$):

$$\bar{X}_n = \frac{X_n}{n} \sim \mathcal{N}(\mu, \sigma/\sqrt{n}) \simeq \mathcal{N}(\bar{x}, S/\sqrt{n})$$

- The standard deviation of the sample mean is $\sigma/\sqrt{n}$
- Its logical estimate is $S/\sqrt{n}$.
- The logical estimate of the standard error is $S/\sqrt{n}$.
- $S/\sqrt{n}$, the standard error, talks about how variable averages of random samples of size n from the population are.

# Comparing systems using sample data

## [Jain 91, Chap 13]

- Determine the confidence interval of the mean
- Comparing two alternatives
- Confidence interval for proportion
- Determining sample size

# Outline

# Determine the confidence interval of the mean

## Problem

- $S = \{x_1, \ldots, x_n\}$: a set of results
- Determine the mean $\mu$ of $S$, such that:
- $\mathbb{P}(c_1 \leq \mu \leq c_2) = 1 - \alpha$
- $\alpha$: significance level (*e.g.* 0.01)
- $1 - \alpha$: confidence level (*e.g.* 0.99)

## Notations

- $n$: number of experiments
- $\bar{x} = \frac{1}{n} \sum x_i$: sample mean
- $s = \sqrt{\frac{1}{n-1} \sum (\bar{x} - x_i)^2}$: unbiased estimation of the standard deviation

# When $n$ is large ($n \geq 30$)

Central-limit theorem: $\bar{x} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
$\mu$ (resp. $\sigma$): population mean (resp. the population std. dev.) of the distribution of the $x_i$.

# When $n$ is large ($n \geq 30$)

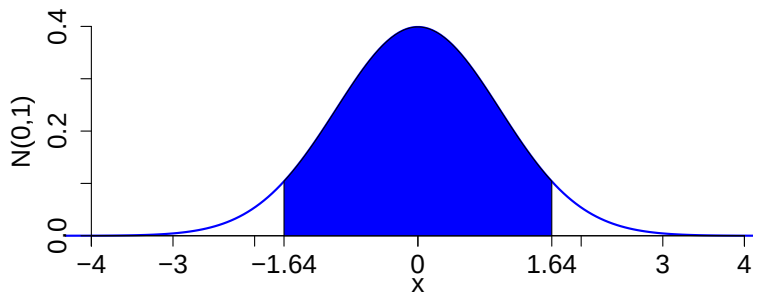
Central-limit theorem: $\bar{x} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
$\mu$ (resp. $\sigma$): population mean (resp. the population std. dev.) of the distribution of the $x_i$.

$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. $Z \sim \mathcal{N}(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}, 1) \sim \mathcal{N}(0, 1)$.
$\mathbb{P}(-c \leq Z \leq c) = 1 - \alpha \Leftrightarrow c = z_{1-\alpha/2}$
$z_i$: value of the $i^{\text{th}}$ quantile of the standard normal.
$\alpha = 0.1 : z_{1-\alpha/2} = z_{0.95} = 1.64$

# When $n$ is large ($n \geq 30$)

Central-limit theorem: $\bar{x} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
$\mu$ (resp. $\sigma$): population mean (resp. the population std. dev.) of the distribution of the $x_i$.

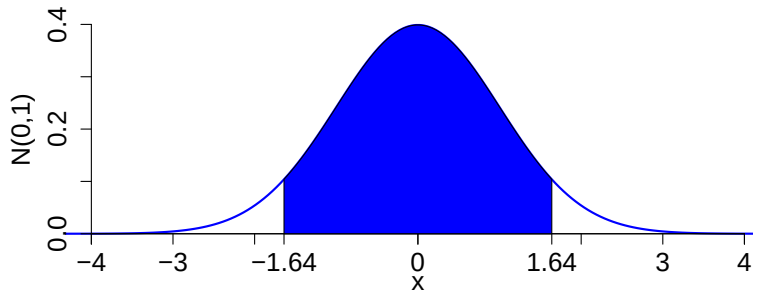
$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. $Z \sim \mathcal{N}(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}, 1) \sim \mathcal{N}(0, 1)$.
$\mathbb{P}(-c \leq Z \leq c) = 1 - \alpha \Leftrightarrow c = z_{1-\alpha/2}$
$z_i$: value of the $i^{\text{th}}$ quantile of the standard normal.
$\alpha = 0.1 : z_{1-\alpha/2} = z_{0.95} = 1.64$



# When $n$ is large ($n \geq 30$)

Central-limit theorem: $\bar{x} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
$\mu$ (resp. $\sigma$): population mean (resp. the population std. dev.) of the distribution of the $x_i$.

$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. $Z \sim \mathcal{N}(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}, 1) \sim \mathcal{N}(0,1)$.
$\mathbb{P}(-c \leq Z \leq c) = 1 - \alpha \Leftrightarrow c = z_{1-\alpha/2}$
$z_i$: value of the $i^{\text{th}}$ quantile of the standard normal.
$\alpha = 0.1 : z_{1-\alpha/2} = z_{0.95} = 1.64$

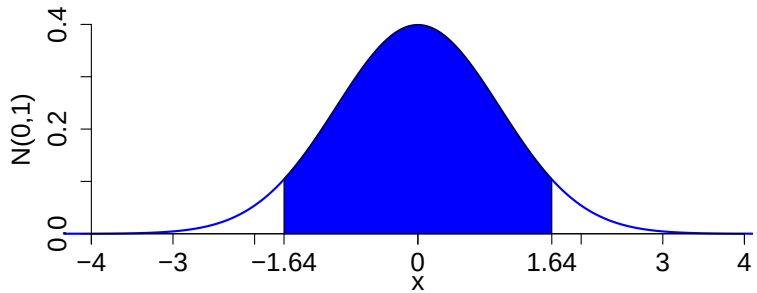



$-c \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq c \Leftrightarrow \bar{x} - c\sigma/\sqrt{n} \leq \mu \leq \bar{x} + c\sigma/\sqrt{n}$. However, $s \approx \sigma$

# When $n$ is large ($n \geq 30$)

Central-limit theorem: $\bar{x} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$
$\mu$ (resp. $\sigma$): population mean (resp. the population std. dev.) of the distribution of the $x_i$.

$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. $Z \sim \mathcal{N}(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}, 1) \sim \mathcal{N}(0,1)$.
$\mathbb{P}(-c \leq Z \leq c) = 1 - \alpha \Leftrightarrow c = z_{1-\alpha/2}$
$z_i$: value of the $i^{\text{th}}$ quantile of the standard normal.
$\alpha = 0.1 : z_{1-\alpha/2} = z_{0.95} = 1.64$




$-c \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq c \Leftrightarrow \bar{x} - c\sigma/\sqrt{n} \leq \mu \leq \bar{x} + c\sigma/\sqrt{n}$. However, $s \approx \sigma$

**With $(1-\alpha)100\%$ confidence**

$$\mu \in [\bar{x} - z_{1-\alpha/2}s/\sqrt{n}, \bar{x} + z_{1-\alpha/2}s/\sqrt{n}]$$

# Exercise

- $\bar{x} = 10$, $n = 64$, $s = 2$

# Exercise

- $\bar{x} = 10$, $n = 64$, $s = 2$
- Find the confidence interval with confidence level of 0.9 (90%) and 0.99 (99%).

# Exercise

- $\bar{x} = 10$, $n = 64$, $s = 2$
- Find the confidence interval with confidence level of 0.9 (90%) and 0.99 (99%).
- CL of 0.9. $\alpha = 0.1 \Rightarrow z_{0.95} = 1.64 \Rightarrow \mu \in [10 - 1.64 \times 2/8, 10 + 1.64 \times 2/8] \Rightarrow \mu \in [9.59, 10.41]$

# Exercise

- $\bar{x} = 10$, $n = 64$, $s = 2$
- Find the confidence interval with confidence level of 0.9 (90%) and 0.99 (99%).
- CL of 0.9. $\alpha = 0.1 \Rightarrow z_{0.95} = 1.64 \Rightarrow \mu \in [10 - 1.64 \times 2/8, 10 + 1.64 \times 2/8] \Rightarrow \mu \in [9.59, 10.41]$
- CL of 0.99. $\alpha = 0.01 \Rightarrow z_{0.995} = 2.58 \Rightarrow \mu \in [10 - 2.58 \times 2/8, 10 + 2.58 \times 2/8] \Rightarrow \mu \in [9.35, 10.65]$

# Exercise

- $\bar{x} = 10$, $n = 64$, $s = 2$
- Find the confidence interval with confidence level of 0.9 (90%) and 0.99 (99%).
- CL of 0.9. $\alpha = 0.1 \Rightarrow z_{0.95} = 1.64 \Rightarrow \mu \in [10 - 1.64 \times 2/8, 10 + 1.64 \times 2/8] \Rightarrow \mu \in [9.59, 10.41]$
- CL of 0.99. $\alpha = 0.01 \Rightarrow z_{0.995} = 2.58 \Rightarrow \mu \in [10 - 2.58 \times 2/8, 10 + 2.58 \times 2/8] \Rightarrow \mu \in [9.35, 10.65]$

**Link between confidence interval and confidence level**

When the confidence level increases, $\alpha$ decreases and the interval increases.

# Code for one vector

## R code

```
interval <-function(x,conf_level=0.9){
  n<-length(x)
  m<-mean(x)
  se<-sd(x)/sqrt(n) # standard error
  alpha<-1-conf_level
  q<-qnorm(1-alpha/2)
  return m+c(-1,1)*q*se
}
```

# Outline

# Paired or unpaired?

In the following situations, are the samples paired or unpaired?

- You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.

# Paired or unpaired?

In the following situations, are the samples paired or unpaired?

- You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.
- Solution: Paired.

# Paired or unpaired?

In the following situations, are the samples paired or unpaired?

- You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.
- Solution: Paired.
- You want to compare expected starting salaries between males and females using the class survey data.

# Paired or unpaired?

In the following situations, are the samples paired or unpaired?

- You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.
- Solution: Paired.
- You want to compare expected starting salaries between males and females using the class survey data.
- Solution: Unpaired.

# Paired or unpaired?

In the following situations, are the samples paired or unpaired?

- You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.
- Solution: Paired.
- You want to compare expected starting salaries between males and females using the class survey data.
- Solution: Unpaired.
- Your company can use one of two possible advertisements. You show one ad to one group of people, and ask them to rate the likelihood of buying your product after seeing the ad. You show the second ad to a second group of people, and ask them the same question.

# Paired or unpaired?

In the following situations, are the samples paired or unpaired?

- You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.
- Solution: Paired.
- You want to compare expected starting salaries between males and females using the class survey data.
- Solution: Unpaired.
- Your company can use one of two possible advertisements. You show one ad to one group of people, and ask them to rate the likelihood of buying your product after seeing the ad. You show the second ad to a second group of people, and ask them the same question.
- Solution: Unpaired.

# Paired or unpaired?

In the following situations, are the samples paired or unpaired?

- You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.
- Solution: Paired.
- You want to compare expected starting salaries between males and females using the class survey data.
- Solution: Unpaired.
- Your company can use one of two possible advertisements. You show one ad to one group of people, and ask them to rate the likelihood of buying your product after seeing the ad. You show the second ad to a second group of people, and ask them the same question.
- Solution: Unpaired.
- Your company can use one of two possible advertisements. You show both ads to a group of people, and ask them to rate their opinions of both ads.

# Paired or unpaired?

In the following situations, are the samples paired or unpaired?

- You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.
- Solution: Paired.
- You want to compare expected starting salaries between males and females using the class survey data.
- Solution: Unpaired.
- Your company can use one of two possible advertisements. You show one ad to one group of people, and ask them to rate the likelihood of buying your product after seeing the ad. You show the second ad to a second group of people, and ask them the same question.
- Solution: Unpaired.
- Your company can use one of two possible advertisements. You show both ads to a group of people, and ask them to rate their opinions of both ads.
- Solution: Paired.

# Comparing two paired observations

## Example with two algorithms

1. You want to test two algorithms with different input data
2. $x$ vector of results of the first algorithms
3. $y$ vector of results of the second algorithms (experiments made in the same order)
4. $z = x - y$
5. Compute the confidence interval of the mean of $z$
6. If the interval contains 0 you cannot conclude on the superiority of one algorithm compared to the other.

## Exercise

Compute the 90% confidence interval for $x$ and $y$ drawn in a continuous uniform distribution (resp in $[0, 10]$ and in $[1, 11]$ with 40 samples).

# Answer

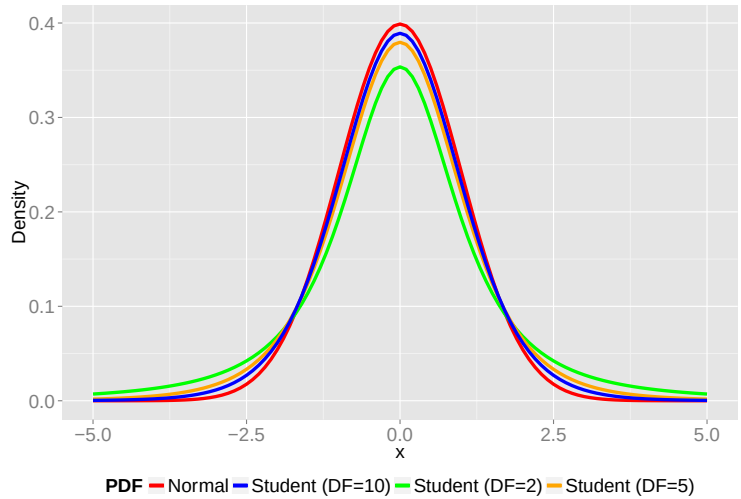
## R code

```
y<-runif(40,0,10)
x<-runif(40,1,11)
z <-x-y
shapiro.test(z)
mean(z)+c(-1,1)*qnorm(0.95)*sd(z)/sqrt(length(z))
[1] 0.1886824 1.6171149
```

# Dealing with small number of samples

## The Student distribution

- The normal approximation of the CLT works for $n$ large.
- When $n$ is small we have to use the Student distribution (t distribution) with $n - 1$ degree of freedom
- The t distribution tends to the normal one when $n$ is large.
- When $n$ is small it has a thicker tails that the normal
- This tails enable to account for the greater uncertainty when $n$ is small
- Nevertheless, it accounts for iid and normal distribution of the data.



# $n \leq 30$ and $x_i$ follow a normal distribution

$t(n)$: Student distribution with $n$ degree of freedom.
$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. $Z \sim t(n-1)$.
$\mathbb{P}(-c \leq Z \leq c) = 1-\alpha \Leftrightarrow c = t_{n-1,1-\alpha/2}$
$t_{k,i}$: value of the $i^{\text{th}}$ quantile of a Student variate with $k$ degree of freedom.
$\alpha = 0.1, n = 5 : t_{4,0.95} = 2.13$

# $n \leq 30$ and $x_i$ follow a normal distribution

$t(n)$: Student distribution with $n$ degree of freedom.
$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. $Z \sim t(n-1)$.
$\mathbb{P}(-c \leq Z \leq c) = 1-\alpha \Leftrightarrow c = t_{n-1,1-\alpha/2}$
$t_{k,i}$: value of the $i^{\text{th}}$ quantile of a Student variate with $k$ degree of freedom.
$\alpha = 0.1, n = 5 : t_{4,0.95} = 2.13$

**With $(1-\alpha)100\%$ confidence**

$$\mu \in [\bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n}, \bar{x} + t_{n-1,1-\alpha/2}s/\sqrt{n}]$$

# $n \leq 30$ and $x_i$ follow a normal distribution

$t(n)$: Student distribution with $n$ degree of freedom.
$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. $Z \sim t(n-1)$.
$\mathbb{P}(-c \leq Z \leq c) = 1-\alpha \Leftrightarrow c = t_{n-1,1-\alpha/2}$
$t_{k,i}$: value of the $i^{\text{th}}$ quantile of a Student variate with $k$ degree of freedom.
$\alpha = 0.1, n = 5 : t_{4,0.95} = 2.13$

## With $(1-\alpha)100\%$ confidence

$$\mu \in [\bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n}, \bar{x} + t_{n-1,1-\alpha/2}s/\sqrt{n}]$$

## R code

```
student_interval <-function(x,conf_level=0.9){
  n<-length(x); X<-mean(x);s<-sd(x);alpha<-1-conf_level
  q<-qt(1-alpha/2,n-1)
  return(c(X-q*s/sqrt(n),X+q*s/sqrt(n)))
}
```

# Comparing two alternatives (paired observations)

6 benchmarks were used to compare two systems.
The observations are:
{(5.4,19.1),(16.6,3.5),(0.6,3.4),(7.3,1.7),(1.4,2.5),(0.6,3.6)}.

# Comparing two alternatives (paired observations)

6 benchmarks were used to compare two systems.
The observations are:
{(5.4,19.1),(16.6,3.5),(0.6,3.4),(7.3,1.7),(1.4,2.5),(0.6,3.6)}.

Is one system better than the other?

# Comparing two alternatives (paired observations)

6 benchmarks were used to compare two systems.
The observations are:
{(5.4,19.1),(16.6,3.5),(0.6,3.4),(7.3,1.7),(1.4,2.5),(0.6,3.6)}.

Is one system better than the other?

Differences: 6 observations: {-13.7,13.1,-2.8,-1.1,-3.0,5.6}
Sample means $\bar{x} = -0.32$
Sample standard deviation $s = 9.03$
These observation are likely to follow a *normal distribution* (*P value* of Shapiro/Wilk test = 0.82>0.1): we can use the student distribution.

# Comparing two alternatives (paired observations)

6 benchmarks were used to compare two systems.
The observations are:
{(5.4,19.1),(16.6,3.5),(0.6,3.4),(7.3,1.7),(1.4,2.5),(0.6,3.6)}.

Is one system better than the other?

Differences: 6 observations: {-13.7,13.1,-2.8,-1.1,-3.0,5.6}
Sample means $\bar{x} = -0.32$
Sample standard deviation $s = 9.03$
These observation are likely to follow a *normal distribution* (*P value* of Shapiro/Wilk test = 0.82>0.1): we can use the student distribution.

$\alpha = 0.1$, $t_{5,0.95} = 2.015$. 90% confidence interval:
$\mu \in [-0.32 - 2.015 \times 9.03/\sqrt{6}, -0.32 + 2.015 \times 9.03/\sqrt{6}] = [-7.76, 7.12]$

# Comparing two alternatives (paired observations)

6 benchmarks were used to compare two systems.
The observations are:
{(5.4,19.1),(16.6,3.5),(0.6,3.4),(7.3,1.7),(1.4,2.5),(0.6,3.6)}.

Is one system better than the other?

Differences: 6 observations: {-13.7,13.1,-2.8,-1.1,-3.0,5.6}
Sample means $\bar{x} = -0.32$
Sample standard deviation $s = 9.03$
These observation are likely to follow a *normal distribution* (*P value* of Shapiro/Wilk test = 0.82>0.1): we can use the student distribution.

$\alpha = 0.1$, $t_{5,0.95} = 2.015$. 90% confidence interval:
$\mu \in [-0.32 - 2.015 \times 9.03/\sqrt{6}, -0.32 + 2.015 \times 9.03/\sqrt{6}] = [-7.76, 7.12]$
The interval contains 0: hence the two systems are not different (with a confidence of 90%)

# Outline

# Population variance vs. sample variance

**Be careful**

Distinguish between

- the population variance $\sigma^2$
- the sample variance $S^2$

Two sample variance $S_x^2$ and $S_y^2$ taken from the same population can be different.

For unpaired observation we need to distinguish between the cases where the population variance is the same or different.

# With same population variance in the two groups

## Cookbook

- Sometime elements in the groups are not comparable pairwise and not of the same size
- Ex: comparing people that received treatment vs placebo.
- $\bar{X}$ average for first group and $\bar{Y}$ average for second groups.
- The $(1-\alpha)\times 100\%$ confidence interval of the difference of the mean is : $\bar{X}-\bar{Y}\pm t_{n_x+n_y-2,1-\alpha/2}S_p\left(\frac{1}{n_x}+\frac{1}{n_y}\right)^{1/2}$
- Where $n_x$ (resp. $n_y$) is the size of the first (resp. second) group.
- $t_{n_x+n_y-2,1-\alpha/2}$ is the $1-\alpha/2$ Student quantile with $n_x+n_y-2$ degrees of freedom.
- Where $S_p^2=\left((n_x-1)S_x^2+(n_y-1)S_y^2\right)/(n_x+n_y-2)$ is the pooled variance estimator. **Assumes a constant variance across the groups** (but sample variance can be different, if they are the same then $\forall n_x,n_y,<S_p=S_x=S_y$).

# Exercise

- Draw two vectors with resp 100 and 200 samples using respectively $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$.
- Compute the confidence interval 90% of the mean of these two vectors.

# Exercise

- Draw two vectors with resp 100 and 200 samples using respectively $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$.
- Compute the confidence interval 90% of the mean of these two vectors.

## Solution

```
nx<-100 ; ny<-200
x<-rnorm(nx,0,1)
y<-rnorm(ny,1,1)
sx2<-var(x) ; sy2<-var(y)
mx<-mean(x) ; my<-mean(y)
sp=sqrt(((nx-1)*sx2+(ny-1)*sy2)/(nx+ny-2))
alpha<-1-0.9
z<-qt(1-alpha/2,nx+ny-2)
mx-my+c(-1,1)*z*sp*sqrt(1/nx+1/ny)
```

# Unequal population variance

## Cookbook

- Under unequal variance the test becomes:

$$\bar{Y} - \bar{X} \pm t_{df} \times \left( \frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right)^{1/2}$$

- Where $t_{df}$ is the student quantile calculated with degrees of freedom: $df = \frac{\left(S_x^2/n_x + S_y^2/n_y\right)^2}{\left(\frac{S_x^2}{n_x}\right)^2/(n_x-1) + \left(\frac{S_y^2}{n_y}\right)^2/(n_y-1)}$

If you do not know if the population variance is equal or unequal use the unequal case.

# Exercise

Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90.
What is the 90% confidence interval for the difference in test scores at the two schools, assuming that test scores came from normal distributions in both schools?

- Use the recipe:

# Exercise

Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90.
What is the 90% confidence interval for the difference in test scores at the two schools, assuming that test scores came from normal distributions in both schools?

- Use the recipe:
- $df = 1148469/40378.93 = 28.44$

# Exercise

Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90.
What is the 90% confidence interval for the difference in test scores at the two schools, assuming that test scores came from normal distributions in both schools?

- Use the recipe:
- $df = 1148469/40378.93 = 28.44$
- $t_{0.95,28.44} = 1.7$

# Exercise

Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90.
What is the 90% confidence interval for the difference in test scores at the two schools, assuming that test scores came from normal distributions in both schools?

- Use the recipe:
- $df = 1148469/40378.93 = 28.44$
- $t_{0.95,28.44} = 1.7$
- $\text{SE} = \sqrt{100^2/15 + 90^2/20} = 32.74$

# Exercise

Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90.
What is the 90% confidence interval for the difference in test scores at the two schools, assuming that test scores came from normal distributions in both schools?

- Use the recipe:
- $df = 1148469/40378.93 = 28.44$
- $t_{0.95,28.44} = 1.7$
- $\text{SE} = \sqrt{100^2/15 + 90^2/20} = 32.74$
- 90% confidence interval : $= 50 \pm 55.66$

# Outline

# Confidence interval for proportions

System A is better than system B for $y_1$ among $n$ experiments.
Sample proportion: $\hat{p}_1 = \frac{y_1}{n}$ $\hat{p}_2 = 1 - \hat{p}_1 = \frac{n - y_1}{n}$

# Confidence interval for proportions

System A is better than system B for $y_1$ among $n$ experiments.
Sample proportion: $\hat{p}_1 = \frac{y_1}{n}$ $\hat{p}_2 = 1 - \hat{p}_1 = \frac{n - y_1}{n}$

$y_1 \sim \mathcal{B}(n, p_1)$ ($p_1$ the true probability that A outperforms B).

# Confidence interval for proportions

System A is better than system B for $y_1$ among $n$ experiments.
Sample proportion: $\hat{p}_1 = \frac{y_1}{n}$ $\hat{p}_2 = 1 - \hat{p}_1 = \frac{n - y_1}{n}$

$y_1 \sim \mathcal{B}(n, p_1)$ ($p_1$ the true probability that A outperforms B).

if $np_1 \geq 10$ and $n(1 - p_1) \geq 10$
$y_1 \sim \mathcal{B}(n, p_1) \sim \mathcal{N}(np_1, \sqrt{np_1(1 - p_1)})$
$\Leftrightarrow \hat{p}_1 = \frac{y_1}{n} \sim \mathcal{N}\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n}}\right) \sim \mathcal{N}\left(\hat{p}_1, \sqrt{\frac{\hat{p}_1\hat{p}_2}{n}}\right)$

# Confidence interval for proportions

System A is better than system B for $y_1$ among $n$ experiments.
Sample proportion: $\hat{p}_1 = \frac{y_1}{n}$ $\hat{p}_2 = 1 - \hat{p}_1 = \frac{n-y_1}{n}$

$y_1 \sim \mathcal{B}(n, p_1)$ ($p_1$ the true probability that A outperforms B).

if $np_1 \geq 10$ and $n(1-p_1) \geq 10$
$y_1 \sim \mathcal{B}(n, p_1) \sim \mathcal{N}(np_1, \sqrt{np_1(1-p_1)})$
$\Leftrightarrow \hat{p}_1 = \frac{y_1}{n} \sim \mathcal{N}\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n}}\right) \sim \mathcal{N}\left(\hat{p}_1, \sqrt{\frac{\hat{p}_1\hat{p}_2}{n}}\right)$

## With $(1-\alpha)100\%$ confidence

$$p_1 \in \left[\hat{p}_1 - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{p}_2}{n}}, \hat{p}_1 + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{p}_2}{n}}\right]$$

# Confidence interval for proportions

System A is better than system B for $y_1$ among $n$ experiments.
Sample proportion: $\hat{p}_1 = \frac{y_1}{n}$ $\hat{p}_2 = 1 - \hat{p}_1 = \frac{n-y_1}{n}$

$y_1 \sim \mathcal{B}(n, p_1)$ ($p_1$ the true probability that A outperforms B).

if $np_1 \geq 10$ and $n(1-p_1) \geq 10$
$y_1 \sim \mathcal{B}(n, p_1) \sim \mathcal{N}(np_1, \sqrt{np_1(1-p_1)})$
$\Leftrightarrow \hat{p}_1 = \frac{y_1}{n} \sim \mathcal{N}\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n}}\right) \sim \mathcal{N}\left(\hat{p}_1, \sqrt{\frac{\hat{p}_1\hat{p}_2}{n}}\right)$

## With $(1-\alpha)100\%$ confidence

$$p_1 \in \left[\hat{p}_1 - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{p}_2}{n}}, \hat{p}_1 + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{p}_2}{n}}\right]$$

If the interval contains 0.5, we cannot conclude that A outperforms B.

# Example

An experiment is repeated 40 times. System A is found superior to system B 30 times, can we state with 99% confidence that system A is superior?

# Example

An experiment is repeated 40 times. System A is found superior to system B 30 times, can we state with 99% confidence that system A is superior?

$n = 40$, $y_1 = 30$
$\hat{p}_1 = 30/40 = 0.75$ ($n\hat{p}_1 = 30$, $n(1 - \hat{p}_1) = 10$)
$\sqrt{\frac{\hat{p}_1 \hat{p}_2}{n}} = \sqrt{\frac{0.75 \times 0.25}{40}} = 0.068$

# Example

An experiment is repeated 40 times. System A is found superior to system B 30 times, can we state with 99% confidence that system A is superior?

$n = 40$, $y_1 = 30$
$\hat{p}_1 = 30/40 = 0.75$ ($n\hat{p}_1 = 30$, $n(1 - \hat{p}_1) = 10$)
$\sqrt{\frac{\hat{p}_1 \hat{p}_2}{n}} = \sqrt{\frac{0.75 \times 0.25}{40}} = 0.068$

$\alpha = 0.01$, $z_{0.995} = 2.58$

# Example

An experiment is repeated 40 times. System A is found superior to system B 30 times, can we state with 99% confidence that system A is superior?

$n = 40$, $y_1 = 30$
$\hat{p}_1 = 30/40 = 0.75$ ($n\hat{p}_1 = 30$, $n(1-\hat{p}_1) = 10$)
$\sqrt{\frac{\hat{p}_1\hat{p}_2}{n}} = \sqrt{\frac{0.75\times 0.25}{40}} = 0.068$

$\alpha = 0.01$, $z_{0.995} = 2.58$

$p_1 \in [0.75 - 2.58 \times 0.068, 0.75 + 2.58 \times 0.068] = [0.57, 0.92]$
The confidence interval does not include 0.5. Hence, we can conclude with 99% confidence that system A is superior than system B.

# Code

## R code

```
proportion_test <-function(x,conf_level=0.9){
  n<-length(x)
  X<-mean(x)
  n1<-sum(findInterval(x,1))
  n2<-n-n1
  p1<-n1/n
  p2<-n2/n
  if(p1*n<10 || p2*n<10){
    stop("Cannot apply normal approximation!")
  }
  alpha<-1-conf_level
  q<-qnorm(1-alpha/2)
  s<-sqrt(p1*p2/n)
  return(c(p1-q*s,p1+q*s))
}
```

# Outline

# Confidence interval for two proportions

we want to compare $\hat{p}_1 = \frac{y_1}{n_1}$ with $\hat{p}_2 = \frac{y_2}{n_2}$.

# Confidence interval for two proportions

we want to compare $\hat{p}_1 = \frac{y_1}{n_1}$ with $\hat{p}_2 = \frac{y_2}{n_2}$.

if $n_1\hat{p}_1 \geq 10$ and $n_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2 \geq 10$ and $n_2(1-\hat{p}_2) \geq 10$
Compute

$$s = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

# Confidence interval for two proportions

we want to compare $\hat{p}_1 = \frac{y_1}{n_1}$ with $\hat{p}_2 = \frac{y_2}{n_2}$.

if $n_1\hat{p}_1 \geq 10$ and $n_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2 \geq 10$ and $n_2(1-\hat{p}_2) \geq 10$
Compute

$$S = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

**With $(1-\alpha)100\%$ confidence**

$$(p_1 - p_2) \in \left[(\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2}S, (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}S\right]$$

# Confidence interval for two proportions

we want to compare $\hat{p}_1 = \frac{y_1}{n_1}$ with $\hat{p}_2 = \frac{y_2}{n_2}$.

if $n_1\hat{p}_1 \geq 10$ and $n_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2 \geq 10$ and $n_2(1-\hat{p}_2) \geq 10$
Compute

$$S = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

**With $(1-\alpha)100\%$ confidence**

$$(p_1 - p_2) \in \left[(\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2}S, (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}S\right]$$

# Confidence interval for two proportions

we want to compare $\hat{p}_1 = \frac{y_1}{n_1}$ with $\hat{p}_2 = \frac{y_2}{n_2}$.

if $n_1\hat{p}_1 \geq 10$ and $n_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2 \geq 10$ and $n_2(1-\hat{p}_2) \geq 10$
Compute

$$S = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

**With $(1-\alpha)100\%$ confidence**

$$(p_1 - p_2) \in \left[(\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2}S, (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}S\right]$$

If the interval contains 0, we cannot conclude that $\hat{p}_1$ is larger than $\hat{p}_2$.

# Example

What is the prevalence of anemia in developing countries?

| | African Women | Women from Americas |
|---|---|---|
| Sample size | 2100 | 1900 |
| Number with anemia | 840 | 323 |
| Sample proportion | 840/2100 = 0.40 | 323/1900 = 0.17 |

Confidence interval = 95%

# Example

What is the prevalence of anemia in developing countries?

| | African Women | Women from Americas |
|---|---|---|
| Sample size | 2100 | 1900 |
| Number with anemia | 840 | 323 |
| Sample proportion | 840/2100 = 0.40 | 323/1900 = 0.17 |

Confidence interval = 95%

$2100 \times 0.4 > 10$ and ... $1900 \times (1 - 0.17) > 10$

$$S = \sqrt{\frac{0.4(1-0.4)}{2100} + \frac{0.17(1-0.17)}{1900}} = 0.01373131$$

# Example

What is the prevalence of anemia in developing countries?

| | African Women | Women from Americas |
|---|---|---|
| Sample size | 2100 | 1900 |
| Number with anemia | 840 | 323 |
| Sample proportion | 840/2100 = 0.40 | 323/1900 = 0.17 |

Confidence interval = 95%

$2100 \times 0.4 > 10$ and ... $1900 \times (1 - 0.17) > 10$

$S = \sqrt{\frac{0.4(1-0.4)}{2100} + \frac{0.17(1-0.17)}{1900}} = 0.01373131$

$\alpha = 0.05$, $z_{0.975} = 1.96$

# Example

What is the prevalence of anemia in developing countries?

| | African Women | Women from Americas |
|---|---|---|
| Sample size | 2100 | 1900 |
| Number with anemia | 840 | 323 |
| Sample proportion | 840/2100 = 0.40 | 323/1900 = 0.17 |

Confidence interval = 95%

$2100 \times 0.4 > 10$ and ... $1900 \times (1 - 0.17) > 10$

$S = \sqrt{\frac{0.4(1-0.4)}{2100} + \frac{0.17(1-0.17)}{1900}} = 0.01373131$

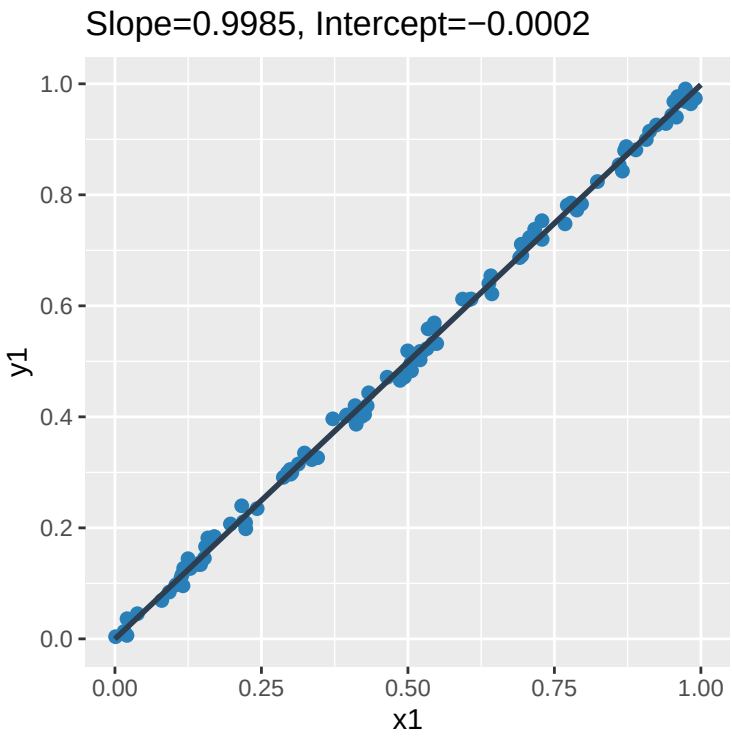
$\alpha = 0.05$, $z_{0.975} = 1.96$

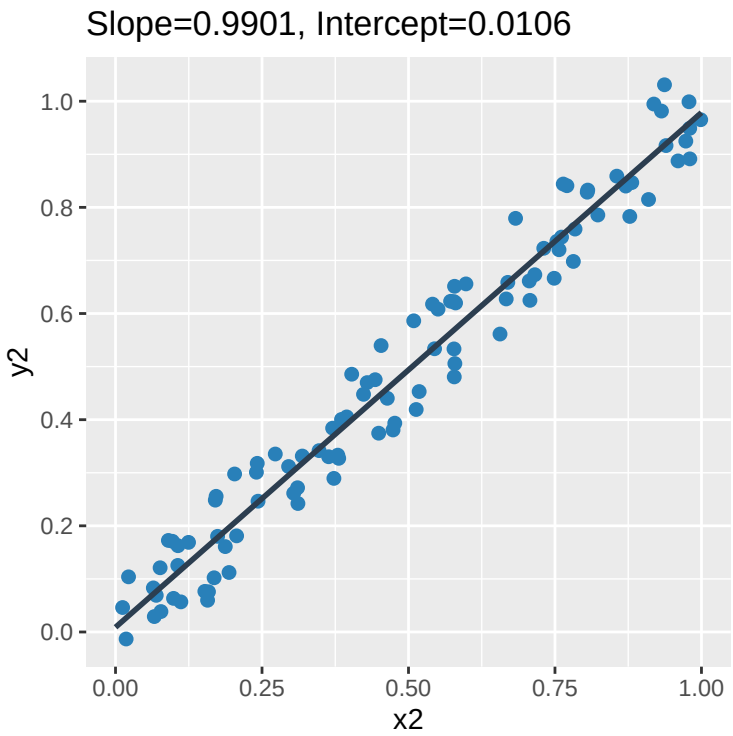
$(p_1 - p_2) \in [(0.4 - 0.17) - 1.96 \times 0.01373131, (0.4 - 0.17) + 1.96 \times 0.01373131] = [0.203, 0.257]$

With confidence 95% we can conclude that there are between 20.3% and 25.7% more African women with anemia than women from the Americas with anemia.

# Outline

# Example




## Problem

Two models that are very close but the right one has data that are more spread.
Can we set a confidence interval for the model?

# Confidence Interval for linear Regression Slope

model: $\hat{y} = \hat{a}x + \hat{b}$

$\hat{a}$: slope

$\hat{b}$: intercept

$n$: number of points

$\hat{y}_i$: predicted values from the model (i.e the $y$ values for each value of $x_i$ according to the model)

# Confidence Interval for linear Regression Slope

model: $\hat{y} = \hat{a}x + \hat{b}$
$\hat{a}$: slope
$\hat{b}$: intercept
$n$: number of points
$\hat{y}_i$: predicted values from the model (i.e the $y$ values for each value of $x_i$ according to the model)
$MSE$: mean squared error. $MSE = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$

# Confidence Interval for linear Regression Slope

model: $\hat{y} = \hat{a}x + \hat{b}$
$\hat{a}$: slope
$\hat{b}$: intercept
$n$: number of points
$\hat{y}_i$: predicted values from the model (i.e the $y$ values for each value of $x_i$ according to the model)
$MSE$: mean squared error. $MSE = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$
$\bar{x}$: mean of the samples $x$ values.
Standard Error of the linear regression slope:

$$S = \frac{\sqrt{MSE}}{\sqrt{\sum(x - \bar{x})^2}}$$

# Confidence Interval for linear Regression Slope

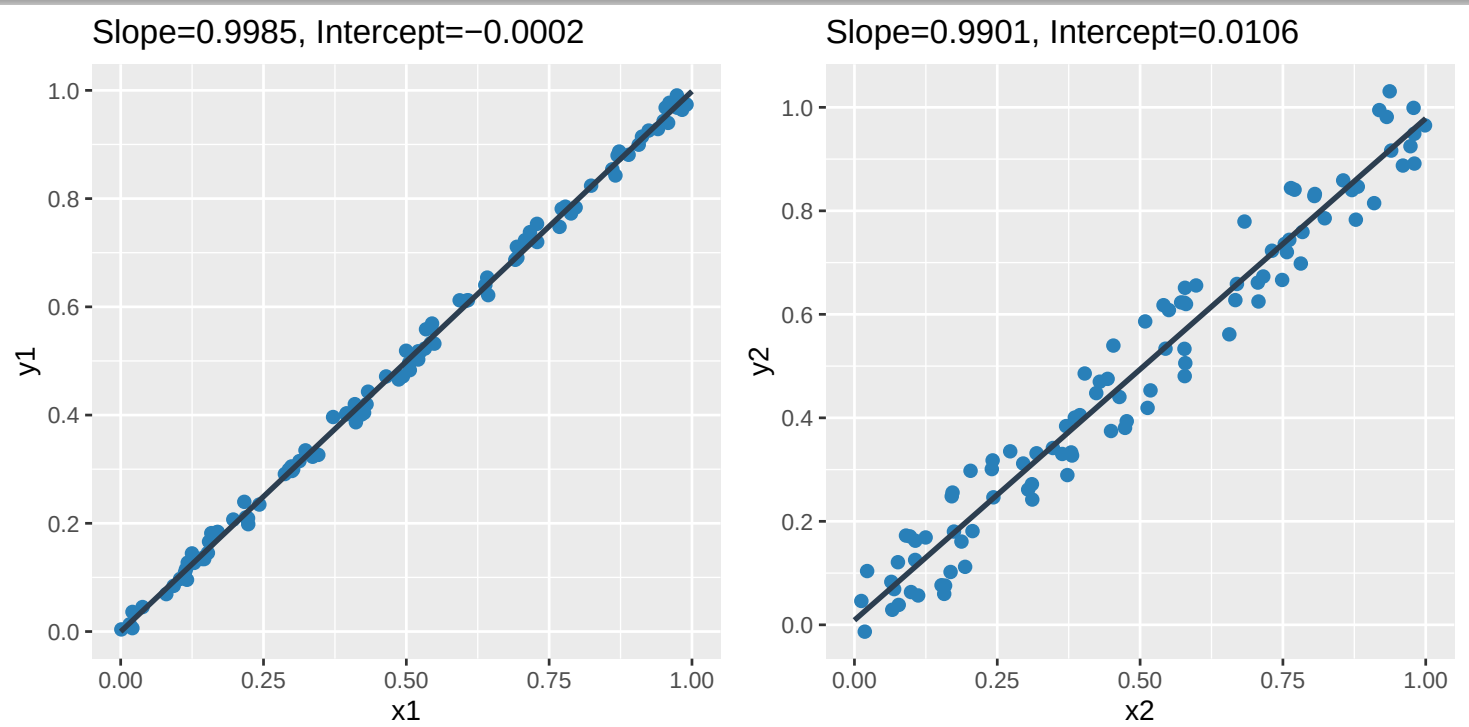
model: $\hat{y} = \hat{a}x + \hat{b}$
$\hat{a}$: slope
$\hat{b}$: intercept
$n$: number of points
$\hat{y}_i$: predicted values from the model (i.e the $y$ values for each value of $x_i$ according to the model)
$MSE$: mean squared error. $MSE = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$
$\bar{x}$: mean of the samples $x$ values.
Standard Error of the linear regression slope:

$$S = \frac{\sqrt{MSE}}{\sqrt{\sum(x - \bar{x})^2}}$$

**With $(1-\alpha)100\%$ confidence**

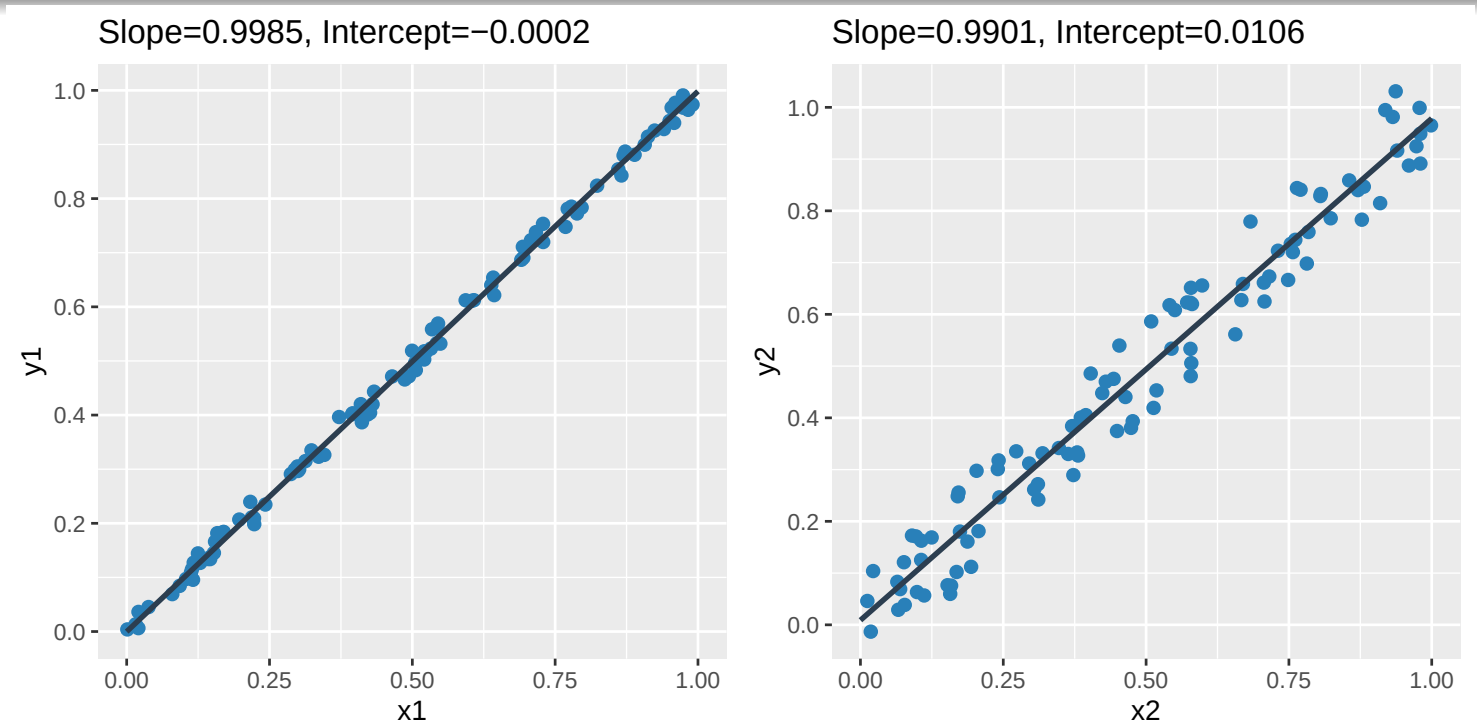
$$a \in \left[\hat{a} - t_{1-\alpha/2,n-2} \times S, \hat{a} + t_{1-\alpha/2,n-2} \times S\right]$$

# Back to the example




$\alpha = 0.05$, $n = 100$, $t_{0.975,98} = 1.9845$

# Back to the example



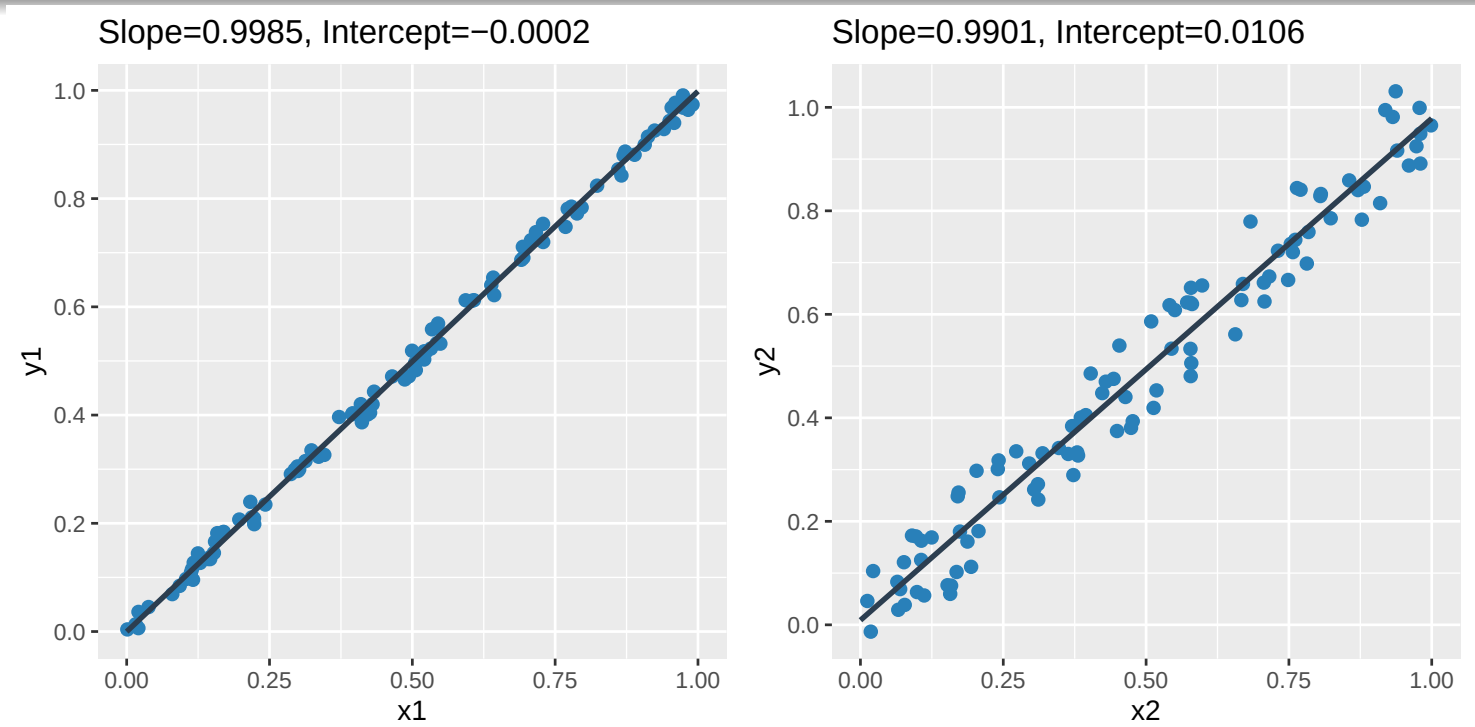

$\alpha = 0.05$, $n = 100$, $t_{0.975,98} = 1.9845$

For the left case: $S_1 = 0.0048$

$a1 \in [0.9885 - 1.9845 \times 0.0048, 0.9885 + 1.9845 \times 0.0048]$

$a1 \in [0.9889, 1.0079]$

# Back to the example




$\alpha = 0.05$, $n = 100$, $t_{0.975,98} = 1.9845$

For the left case: $S_1 = 0.0048$

$a1 \in [0.9885 - 1.9845 \times 0.0048, 0.9885 + 1.9845 \times 0.0048]$

$a1 \in [0.9889, 1.0079]$

For the right case: $S_2 = 0.0204$

$a2 \in [0.9901 - 1.9845 \times 0.0204, 0.9901 + 1.9845 \times 0.0204]$

$a2 \in [0.9496, 1.0306]$

# R code

```
lin.model <- lm(y1 ~ x1)
sry <- summary(lin.model)
sry
Call:
lm(formula = y1 ~ x1)

Residuals:
      Min         1Q     Median         3Q        Max
-0.024334 -0.010242 -0.002057  0.011115  0.025981

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0001843  0.0027465  -0.067    0.947
x1           0.9984523  0.0048120 207.493   <2e-16 ***

S<- sry$coefficients[2,2] ; b <-sry$coefficients[2,1]
q<-qt(0.975,length(x1)-2)
b+c(-1,1)*q*S
```

# Exercise

## Cars consumption

You have two cars. You drive them on different itineraries and you measure the number of kilometers and the gas consumption in litters. Assuming that only the distance impact the consumption. Discuss the consumption model of the two cars with 95% confidence.

- car 1: distance= $\{10, 30, 80, 15, 120\}$, consumption = $\{0.5, 1.6, 3.8, 0.8, 6.2\}$
- car 2: distance= $\{5, 20, 50, 45, 90, 110\}$, consumption = $\{0.2, 0.7, 2.1, 1.9, 3.5, 4.7\}$

# Exercise

## Cars consumption

You have two cars. You drive them on different itineraries and you measure the number of kilometers and the gas consumption in litters. Assuming that only the distance impact the consumption. Discuss the consumption model of the two cars with 95% confidence.

- car 1: distance= $\{10, 30, 80, 15, 120\}$, consumption = $\{0.5, 1.6, 3.8, 0.8, 6.2\}$
- car 2: distance= $\{5, 20, 50, 45, 90, 110\}$, consumption = $\{0.2, 0.7, 2.1, 1.9, 3.5, 4.7\}$

## Hints

You have to assume that when distance = 0, consumption = 0.
To built a linear model that passes through the origin do it this way: `fit <- lm(y ~ 0+x)`
Beware that the coefficients section of `summary(fit)` have now only one line

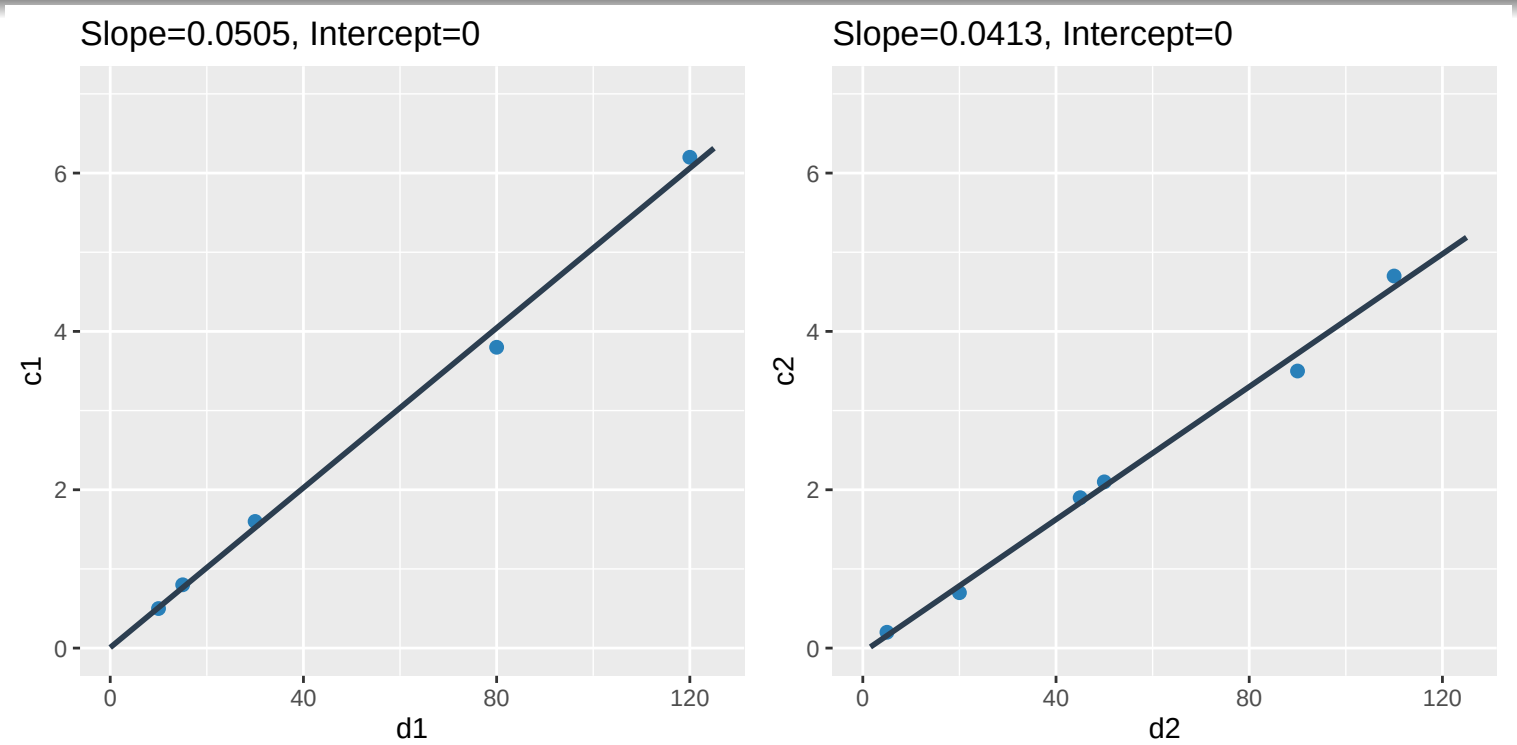
# Solution

```
d1 <- c(10,30,80,15,120); c1 <- c(0.5,1.6,3.8,0.8,6.2)
d2 <- c(5,20,50,45,90,110);c2 <- c(0.2,0.7,2.1,1.9,3.5,4.7)
alpha <- 0.05 ; q <- qt(1 - alpha/2, length(d1)-2)

m1 <- lm(c1~0+d1); s1<- summary(m1)
S1 <- s1$coefficients[1,2]
a1 <- s1$coefficients[1,1]
 0.05053348
a1+c(-1,1)*q*S1
0.04738497 0.05368200

m2 <- lm(c2~0+d2); s2<-summary(m2)
S2 <- s2$coefficients[1,2]
a2 <- s2$coefficients[1,1]
a2
0.04125249
a2+c(-1,1)*q*S2
0.03855199 0.04395298
```

# Solution (ctn.)




## Interpretation (95% interval)

- $a1 \in [0.04738497, 0.05368200]$
- $a2 \in [0.03855199, 0.04395298]$
- the unit is l/km. We can conclude with 95% confidence that car 2 has lower consumption than car 1.

# Confidence Interval around a Linear Regression Line

$n$: number of points

$\hat{y}_i$: predicted values from the model.

$MSE$: mean squared error. $MSE = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$

# Confidence Interval around a Linear Regression Line

$n$: number of points

$\hat{y}_i$: predicted values from the model.

$MSE$: mean squared error. $MSE = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$

The standard error S is, for any value $x$ and its predicted value $\hat{y}$:

$$S = \sqrt{MSE} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

# Confidence Interval around a Linear Regression Line

$n$: number of points

$\hat{y}_i$: predicted values from the model.

$MSE$: mean squared error. $MSE = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$

The standard error S is, for any value $x$ and its predicted value $\hat{y}$:

$$S = \sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

## With $(1-\alpha)100\%$ confidence

$$y \in \left[\hat{y} - t_{1-\alpha/2,n-2} \times S, \hat{y} + t_{1-\alpha/2,n-2} \times S\right]$$

# Confidence Interval around a Linear Regression Line

$n$: number of points

$\hat{y}_i$: predicted values from the model.

$MSE$: mean squared error. $MSE = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$

The standard error S is, for any value $x$ and its predicted value $\hat{y}$:

$$S = \sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

## With $(1-\alpha)100\%$ confidence

$$y \in \left[\hat{y} - t_{1-\alpha/2,n-2} \times S, \hat{y} + t_{1-\alpha/2,n-2} \times S\right]$$

## Remark

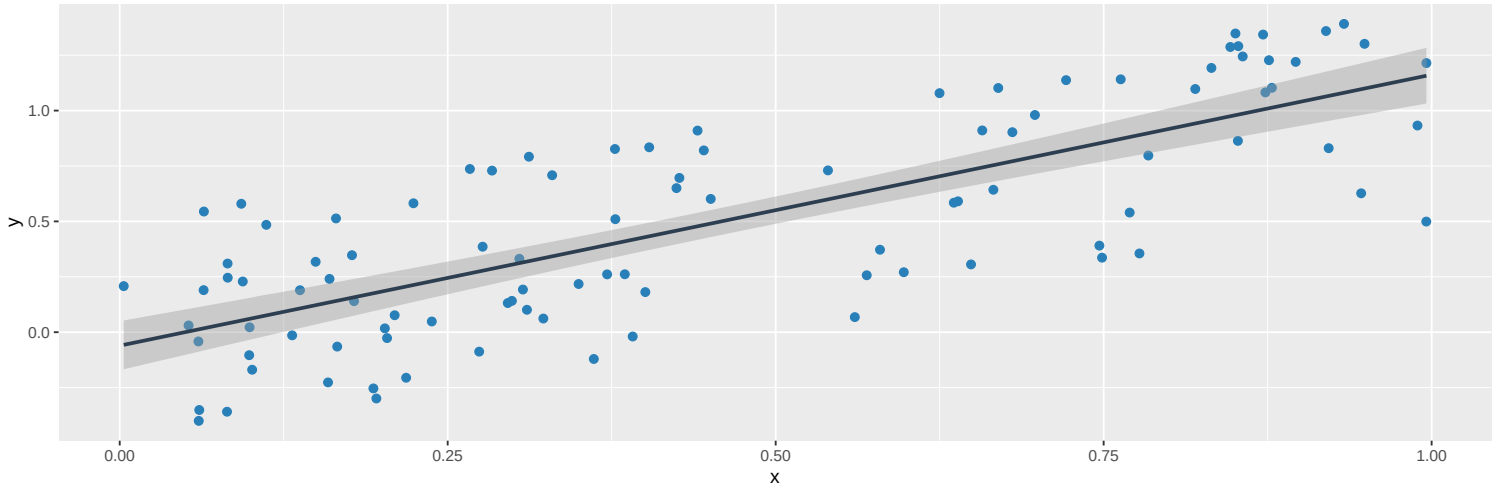
$S$ increases when $x$ is away from $\bar{x}$. This accounts from the fact that the uncertainty is larger at both ends of the range as we have less values close to $x$.

# Draw Conf. int around the Reg. Line with R

## R Code

```
library(ggplot2)
x<-runif(100)
y<-x+(runif(100)-0.5)
df <- data.frame(x,y)

reg<- ggplot(df, aes(x=x, y=y)) +
  geom_point(color='#2980B9', size = 2) +
  geom_smooth(method=lm, se=TRUE, level=0.95, fullrange=TRUE, color='#2C3E50')
reg
ggsave("./linear_reg_conf_int.pdf",plot=reg, width=12, height=4)
```



# Takeaway message

## The confidence interval

**Mean $\pm$ Quantile $\times$ Standard Error**

# R is nice!

## R provides all the above function

Confidence interval of the mean :

```
r<-t.test(x,conf.level=0.9)
```

Comparing paired experiment :

```
r<-t.test(x,y,paired=TRUE,conf.level=0.9)
```

Comparing unpaired experiment :

```
r<-t.test(x,y,conf.level=0.9)
```

Comparing unpaired experiment with equal variance :

```
r<-t.test(x,y,conf.level=0.9,var.equal=TRUE)
```

CI for proportion : `r=binom.test(n1,n,conf.level=0.9)`

CI for two proportions : `r=prop.test(Y,N,conf.level=0.9)`

```
inf<-r$conf.int[1]
sup<-r$conf.int[2]
```

# Outline

# The Null Hypothesis

- We need to make decision between two hypothesis
- The status-quo, or default hypothesis, $H_0$ is called the null hypothesis
- The alternative hypothesis is call $H_a$

# The Null Hypothesis

- We need to make decision between two hypothesis
- The status-quo, or default hypothesis, $H_0$ is called the null hypothesis
- The alternative hypothesis is call $H_a$

## Example

- We compare two algorithms ($A_1$ vs $A_2$) with 100 paired experiments.
- Performance metric $P$: the higher the better.
- Sample mean: $\tilde{\mu} = \overline{P_1 - P_2} = 2$ and $s = 10$.
- We want to test the hypothesis $H_0 : \mu = 0$ ($A_1$ is not better than $A_2$). $\mu$ population mean.
- Versus $H_a : \mu > 0$ ($A_1$ is better than $A_2$).

# 4 cases

| Truth | Decision | Result |
|---|---|---|
| $H_0$ | $H_0$ | Correctly accepting the null |
| $H_0$ | $H_a$ | Type I error (wrongly rejecting the null) |
| $H_a$ | $H_a$ | Correctly rejecting the null |
| $H_a$ | $H_0$ | Type II error (wrongly accepting the null) |

# 4 cases

| Truth | Decision | Result |
|---|---|---|
| $H_0$ | $H_0$ | Correctly accepting the null |
| $H_0$ | $H_a$ | Type I error (wrongly rejecting the null) |
| $H_a$ | $H_a$ | Correctly rejecting the null |
| $H_a$ | $H_0$ | Type II error (wrongly accepting the null) |

## Example, court case

- The null hypothesis $H_0$ is that the defendant is innocent
- $H_a$ is that the defendant is guilty
- Setting low standard $\Rightarrow$ more innocent people are convicted (Type I error)
- Setting higher standard $\Rightarrow$ more guilty people are left free (type II error)
- The way set standard impact the type of error we make. The more we make one type of error the less we make the other type.

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)
- $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha$

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)
- $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha$
- Using the CLT, under $H_0$ we have: $\tilde{\mu} \sim \mathcal{N}(0, s/\sqrt{n}) = \mathcal{N}(0, 10/\sqrt{100}) = \mathcal{N}(0, 1)$.

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)
- $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha$
- Using the CLT, under $H_0$ we have: $\tilde{\mu} \sim \mathcal{N}(0, s/\sqrt{n}) = \mathcal{N}(0, 10/\sqrt{100}) = \mathcal{N}(0, 1)$.
- So, $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha \Leftrightarrow \mathbb{P}(\tilde{\mu} \leq C; H_0) = 1 - \alpha \Leftrightarrow C = z_{1-\alpha}$

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)
- $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha$
- Using the CLT, under $H_0$ we have: $\tilde{\mu} \sim \mathcal{N}(0, s/\sqrt{n}) = \mathcal{N}(0, 10/\sqrt{100}) = \mathcal{N}(0, 1)$.
- So, $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha \Leftrightarrow \mathbb{P}(\tilde{\mu} \leq C; H_0) = 1 - \alpha \Leftrightarrow C = z_{1-\alpha}$
- "*Reject* $H_0$ *when* $\tilde{\mu} > z_{1-\alpha}$" has the property that probability of rejection is $\alpha$ when $H_0$ is true.

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)
- $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha$
- Using the CLT, under $H_0$ we have: $\tilde{\mu} \sim \mathcal{N}(0, s/\sqrt{n}) = \mathcal{N}(0, 10/\sqrt{100}) = \mathcal{N}(0, 1)$.
- So, $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha \Leftrightarrow \mathbb{P}(\tilde{\mu} \leq C; H_0) = 1 - \alpha \Leftrightarrow C = z_{1-\alpha}$
- "*Reject* $H_0$ *when* $\tilde{\mu} > z_{1-\alpha}$" has the property that probability of rejection is $\alpha$ when $H_0$ is true.
- In this case either :

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)
- $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha$
- Using the CLT, under $H_0$ we have: $\tilde{\mu} \sim \mathcal{N}(0, s/\sqrt{n}) = \mathcal{N}(0, 10/\sqrt{100}) = \mathcal{N}(0, 1)$.
- So, $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha \Leftrightarrow \mathbb{P}(\tilde{\mu} \leq C; H_0) = 1 - \alpha \Leftrightarrow C = z_{1-\alpha}$
- "*Reject* $H_0$ *when* $\tilde{\mu} > z_{1-\alpha}$" has the property that probability of rejection is $\alpha$ when $H_0$ is true.
- In this case either :
  - null hypothesis is false

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)
- $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha$
- Using the CLT, under $H_0$ we have: $\tilde{\mu} \sim \mathcal{N}(0, s/\sqrt{n}) = \mathcal{N}(0, 10/\sqrt{100}) = \mathcal{N}(0, 1)$.
- So, $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha \Leftrightarrow \mathbb{P}(\tilde{\mu} \leq C; H_0) = 1 - \alpha \Leftrightarrow C = z_{1-\alpha}$
- "*Reject* $H_0$ *when* $\tilde{\mu} > z_{1-\alpha}$" has the property that probability of rejection is $\alpha$ when $H_0$ is true.
- In this case either :
  - null hypothesis is false
  - we have seen a rare event in support of $H_a$ while $H_0$ is true

# Setting the correct standard

- let's go back to our example $H_0 : \mu = 0$ and $H_a : \mu > 0$.
- A reasonable strategy is to reject the null if $\tilde{\mu}$ is larger than $C$.
- $C$ chosen such that the probability of Type I error is low (e.g $\alpha = 0.05$)
- $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha$
- Using the CLT, under $H_0$ we have: $\tilde{\mu} \sim \mathcal{N}(0, s/\sqrt{n}) = \mathcal{N}(0, 10/\sqrt{100}) = \mathcal{N}(0, 1)$.
- So, $\mathbb{P}(\tilde{\mu} > C; H_0) = \alpha \Leftrightarrow \mathbb{P}(\tilde{\mu} \leq C; H_0) = 1 - \alpha \Leftrightarrow C = z_{1-\alpha}$
- "*Reject* $H_0$ *when* $\tilde{\mu} > z_{1-\alpha}$" has the property that probability of rejection is $\alpha$ when $H_0$ is true.
- In this case either :
  - null hypothesis is false
  - we have seen a rare event in support of $H_a$ while $H_0$ is true
  - our modeling is false (non iid variables, etc.)

# Go back to the example

## Example

- $\alpha = 0.05$
- $C = z_{1-\alpha} = z_{0.95} = 1.645$.
- So, as $\tilde{\mu} = 2 > C$ we can reject the null hypothesis or we have faced an unlikely event that support the alternative hypothesis ($A_1$ is better than $A_2$) while this is not true.

# Test statistic

## An other way of seeing the problem

- Let TS=$\frac{\tilde{\mu}-\mu_o}{s/\sqrt{n}}$ (TS is the number of standard deviation that $\mu$ is away from $\mu_0$).
- Under our assumptions, $TS \sim \mathcal{N}(0,1)$; Hence,
- If $TS > z_{1-\alpha}$ we would reject the null hypothesis.
- In our example TS$=2$ while $Z_{0.95} = 1.645$, so we reject the null.

# $n$ is small

## Use the student distribution

Use the $t$ quantile with $n-1$ degrees of freedom. R:

```
qt(1-alpha,df)
```

## Example

- $\alpha = 0.05$, $n = 16$
- $t_{15,0.95} = 1.7531$
- TS = $\frac{2-0}{10/\sqrt{16}} = 0.8 < t_{15,0.95}$. So, we fail to reject the null hypothesis.

# Two-sided tests

- Sometime we want to test $H_a : \mu \neq 0$ instead of $H_a : \mu > 0$.
- We reject if TS greater $z_{1-\alpha/2}$ or smaller than $z_{\alpha/2}$
- So we reject if the absolute value of TS is greater than $z_{1-\alpha/2}$.

# Two-sided tests

- Sometime we want to test $H_a : \mu \neq 0$ instead of $H_a : \mu > 0$.
- We reject if TS greater $z_{1-\alpha/2}$ or smaller than $z_{\alpha/2}$
- So we reject if the absolute value of TS is greater than $z_{1-\alpha/2}$.

## Example

- $\alpha = 0.05$, $n = 100$, $z_{0.975} = 1.96 < 2 =$TS. So, we reject.
- $\alpha = 0.05$, $n = 16$, $t_{15,0.975} = 2.13 > 0.8 =$TS. So, we fail to reject.

# Two-sided tests

- Sometime we want to test $H_a : \mu \neq 0$ instead of $H_a : \mu > 0$.
- We reject if TS greater $z_{1-\alpha/2}$ or smaller than $z_{\alpha/2}$
- So we reject if the absolute value of TS is greater than $z_{1-\alpha/2}$.

## Example

- $\alpha = 0.05$, $n = 100$, $z_{0.975} = 1.96 < 2 =$TS. So, we reject.
- $\alpha = 0.05$, $n = 16$, $t_{15,0.975} = 2.13 > 0.8 =$TS. So, we fail to reject.

## Remark

If we fail to reject a one-sided test we will fail to reject the two sided test as well (The quantile is an increasing function)

# Connection with confidence interval

## Null Hypothesis and CI

- Testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$.
- The set where we fail to reject $H_0$ is the $(1 - \alpha)100\%$ confidence interval of $\mu$
- It works the other way: if a $(1 - \alpha)100\%$ interval contains $\mu_0$, then we fail to reject $H_0$ .

# Exercice

## Questions

- A factory has a machine that dispenses 80 mL in bottles. An employee believes that average is lower. Using 40 samples, she measures that that the average amount dispensed is 78 mL with a standard deviation of 2.5. (a) state the null and alternative hypothesis. (b) with 95% confidence is there enough evidence to support the idea that the machine is not working properly?
- Same questions but when the employee believes that the average is different and the amount dispensed 80.8 mL with 99% confidence.

# P-Value: Coin flip example

## Example

- $H_0$: the coin is fair, $H_a$ the coin is biased toward heads.

# P-Value: Coin flip example

## Example

- $H_0$: the coin is fair, $H_a$ the coin is biased toward heads.
- 4 tails out 10?

# P-Value: Coin flip example

## Example

- $H_0$: the coin is fair, $H_a$ the coin is biased toward heads.
- 4 tails out 10?
- 40 tails out 100??

# P-Value: Coin flip example

## Example

- $H_0$: the coin is fair, $H_a$ the coin is biased toward heads.
- 4 tails out 10?
- 40 tails out 100??
- 400 tails out 1000???

# P-Value: Coin flip example

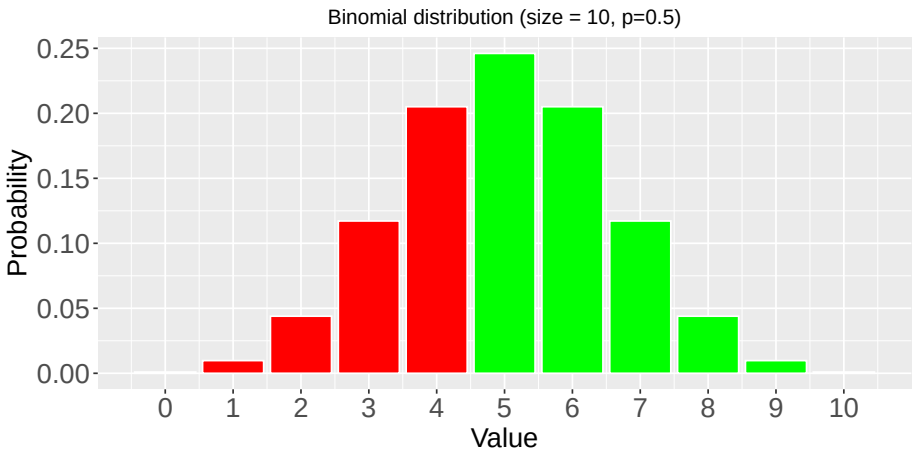
## Example

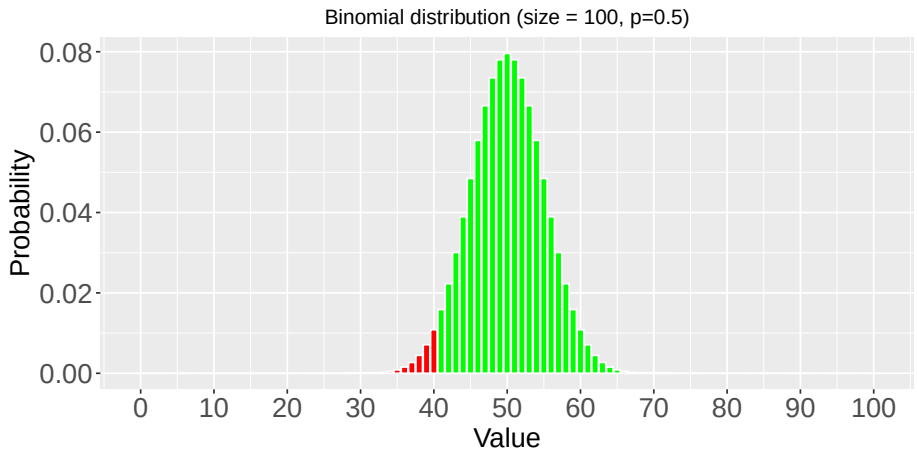
- $H_0$: the coin is fair, $H_a$ the coin is biased toward heads.
- 4 tails out 10?
- 40 tails out 100??
- 400 tails out 1000???

## P-value

Assuming the coin is fair, what is the chance (i.e. the probability) to see $n$ or less tails out of $s$ trials ($n < s/2$)

# Coin flip example

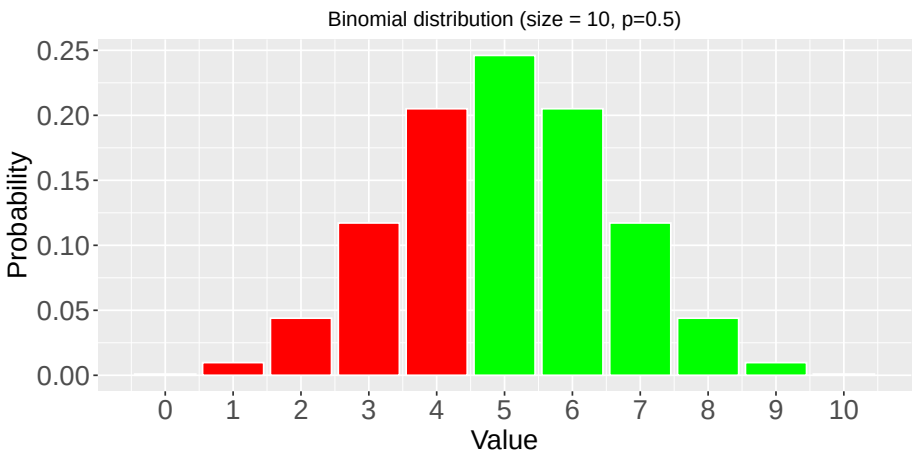


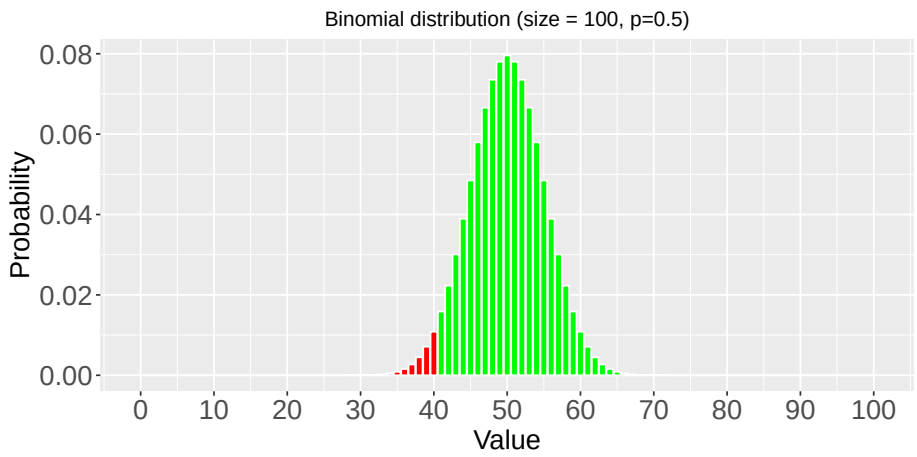




## Result

- Use `pbinom`, to compute the red area

# Coin flip example

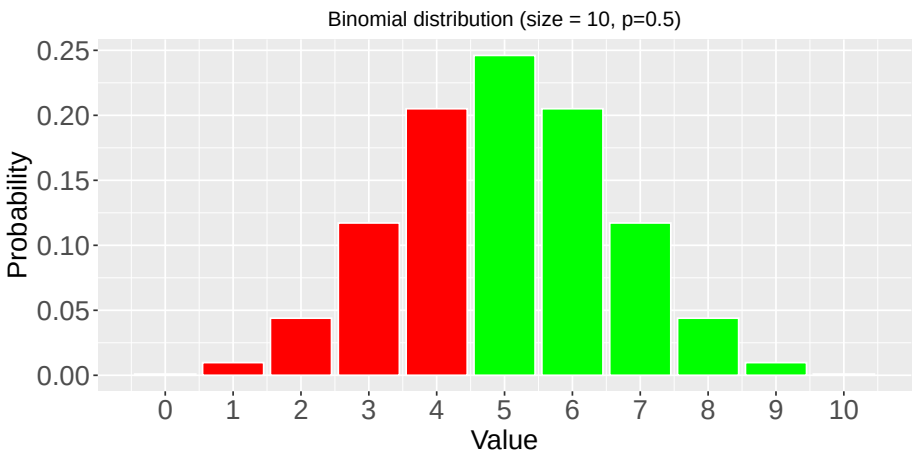


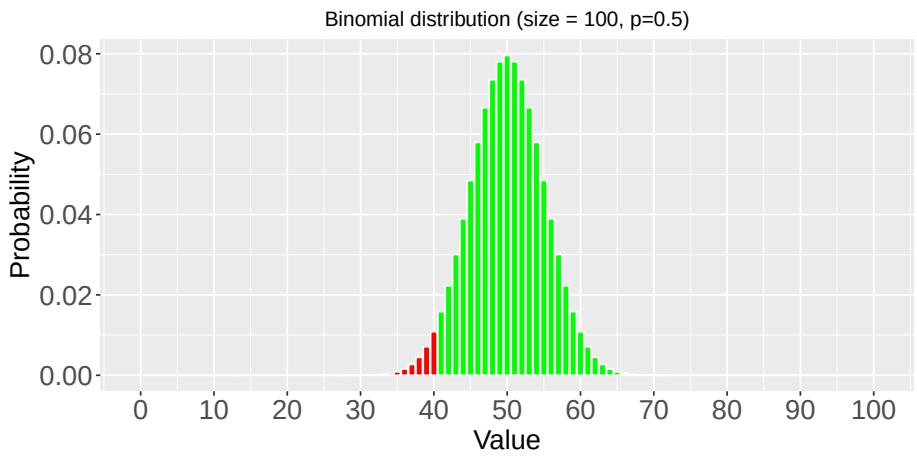




## Result

- Use `pbinom`, to compute the red area
- 4 tails out 10: 0.3769531

# Coin flip example

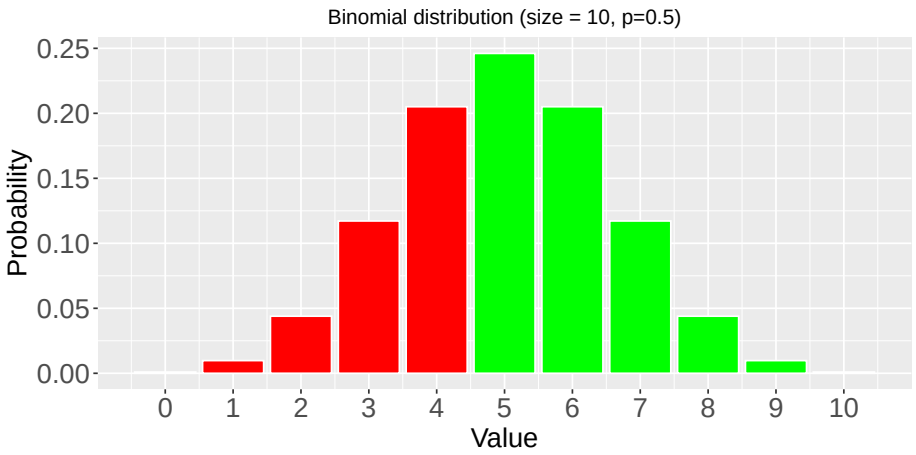


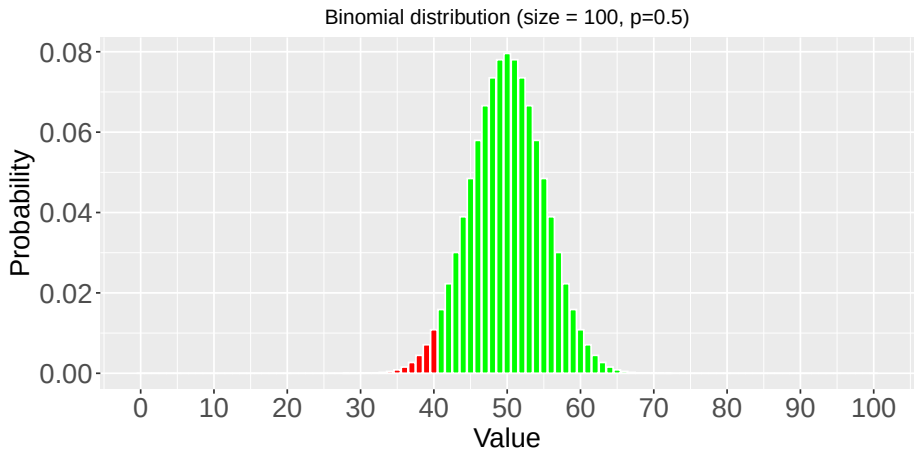




## Result

- Use `pbinom`, to compute the red area
- 4 tails out 10: 0.3769531
- 40 tails out 100: 0.02844397

# Coin flip example




## Result

- Use `pbinom`, to compute the red area
- 4 tails out 10: 0.3769531
- 40 tails out 100: 0.02844397
- 400 tails out 1000: $1.364232\ 10^{-10}$

# P-value

- Assume the null hypothesis is true.
- What is the **probability** to see data in favor of the alternative hypothesis?
- Decide :
  1. The statistic to evaluate the support of the null hypothesis
  2. The distribution of that statistic under the null hypothesis (null distribution)
  3. Compute the probability of obtaining the statistic as or more extreme in favor of the alternative hypothesis under the distribution given in 2 while $H_0$ is true

# P-value with our example

## Example

- 16 experiences, TS = 2.5.

# P-value with our example

## Example

- 16 experiences, TS = 2.5.
- The null distribution is a student distribution with 15 degrees of freedom

# P-value with our example

## Example

- 16 experiences, TS = 2.5.
- The null distribution is a student distribution with 15 degrees of freedom
- What is the probability to get a statistic as large as 2.5

# P-value with our example

## Example

- 16 experiences, TS = 2.5.
- The null distribution is a student distribution with 15 degrees of freedom
- What is the probability to get a statistic as large as 2.5
- `1-pt(2.5,df=15) = pt(2.5,df=15,lower.tail=False)` = 0.01225.

# P-value with our example

## Example

- 16 experiences, TS = 2.5.
- The null distribution is a student distribution with 15 degrees of freedom
- What is the probability to get a statistic as large as 2.5
- `1-pt(2.5,df=15) = pt(2.5,df=15,lower.tail=False)` = 0.01225.
- $t_{15,1-0.01225} = t_{15,0.98775} = 2.5$.

# P-value with our example

## Example

- 16 experiences, TS = 2.5.
- The null distribution is a student distribution with 15 degrees of freedom
- What is the probability to get a statistic as large as 2.5
- `1-pt(2.5,df=15) = pt(2.5,df=15,lower.tail=False)` = 0.01225.
- $t_{15,1-0.01225} = t_{15,0.98775} = 2.5$.

## p-value

In this case, the probability that we have seen an event as or more extreme in favor of the alternative hypothesis while the null hypothesis is true is: 0.01225.
So, (assuming our model is correct) either we observed data that was pretty unlikely under the null, or the null hypothesis is false.

# p-value

## Definitions

The p-value or probability value is the **probability** of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.

How unusual is the result we got if the null hypothesis is true.

The p-value tells you how unusual your data are assuming the null-hypothesis.

## Be careful

The p-value is not the probability of making a type-1 error: $\alpha$ is this probability.

The p-value is not the probability that null hypothesis is true given the data. $\mathbb{P}(D|H_0) \neq \mathbb{P}(H_0|D)$.

# Interpretation

## Small p-value

Either:

- $H_0$ is true and we have observed a rare event
- $H_0$ is false
- or possibly the null model is incorrect (non iid samples, non normality distribution).

# Interpretation

## Cutoff

- One sided test: if the p-value is smaller than $\alpha$, reject the null hypothesis
- Two-sided test: if the smallest p-value of the two one-sided hypothesis test, is smaller than $\alpha/2$, reject the null hypothesis

## Rule of thumbs

- p-value $< 0.01$ : very strong evidence against $H_0$
- $0.01 <$ p-value $< 0.05$ : strong evidence against $H_0$
- $0.05 <$ p-value $< 0.1$ : some weak evidence against $H_0$
- p-value $> 0.1$ : little or no evidence against $H_0$

# Attained significance level

## Example

- 100 experiences, $\tilde{\mu} = 2$, $s = 10$, $H_0 : \mu = 0$. TS = 2.

# Attained significance level

## Example

- 100 experiences, $\tilde{\mu} = 2$, $s = 10$, $H_0 : \mu = 0$. TS = 2.
- the null distribution is $\mathcal{N}(0, 1)$

# Attained significance level

## Example

- 100 experiences, $\tilde{\mu} = 2$, $s = 10$, $H_0 : \mu = 0$. TS = 2.
- the null distribution is $\mathcal{N}(0, 1)$
- We have seen that with $\alpha = 0.05$ we reject the null hypothesis ($z_{0.95} = 1.645$).

# Attained significance level

## Example

- 100 experiences, $\tilde{\mu} = 2$, $s = 10$, $H_0 : \mu = 0$. TS = 2.
- the null distribution is $\mathcal{N}(0, 1)$
- We have seen that with $\alpha = 0.05$ we reject the null hypothesis ($z_{0.95} = 1.645$).
- What is the smallest $\alpha$ that still reject the null hypothesis?

# Attained significance level

## Example

- 100 experiences, $\tilde{\mu} = 2$, $s = 10$, $H_0 : \mu = 0$. TS = 2.
- the null distribution is $\mathcal{N}(0, 1)$
- We have seen that with $\alpha = 0.05$ we reject the null hypothesis ($z_{0.95} = 1.645$).
- What is the smallest $\alpha$ that still reject the null hypothesis?
- `1-pnorm(2)=pnorm(2,lower.tail=False)`=0.02275.

# Attained significance level

## Example

- 100 experiences, $\tilde{\mu} = 2$, $s = 10$, $H_0 : \mu = 0$. TS = 2.
- the null distribution is $\mathcal{N}(0, 1)$
- We have seen that with $\alpha = 0.05$ we reject the null hypothesis ($z_{0.95} = 1.645$).
- What is the smallest $\alpha$ that still reject the null hypothesis?
- `1-pnorm(2)=pnorm(2,lower.tail=False)`=0.02275.
- $z_{1-0.02275} = z_{0.97725} = 2$.

## Attained significance level

The smallest value of $\alpha$ for which we still reject the null hypothesis is called the *attained significance level*.
Here, it is: 0.02275

# Exercice

## Compute the P-value and the attained significance in this case

- A factory has a machine that dispenses 80 mL in bottles. An employee believes that average is lower. Using 40 samples, she measures that that the average amount dispensed is 78 mL with a standard deviation of 2.5. (a) state the null and alternative hypothesis. (b) What is the P-Value associated with these data (using distribution and test statistic) (c) Do you accept or reject the null hypothesis with 99.9% confidence? d() What is the attained significance level?

# Exercice

## Compute the P-value and the attained significance in this case

- A factory has a machine that dispenses 80 mL in bottles. An employee believes that average is lower. Using 40 samples, she measures that that the average amount dispensed is 78 mL with a standard deviation of 2.5. (a) state the null and alternative hypothesis. (b) What is the P-Value associated with these data (using distribution and test statistic) (c) Do you accept or reject the null hypothesis with 99.9% confidence? d() What is the attained significance level?

## Answer

- p = 2.1 10 – 7
- no
- 2.1 10 – 7

# p-value examples

## Some tests

- T test (`t.test`). $H_0$ : mean of data is zero. P-value is small we can reject the null hypothesis and assume that the data has not a zero mean or the two vector have different means.
- Shapiro-Wilk normality test (`shapiro.test`). $H_0$ : the data is normally distributed. Hence, large p-value: it is likely that the data are from a normally distributed population.

# Exercise (from `Statistical inference for data science` by B. Caffo.)

## Question

Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. Assuming that the groups are independent and that they days are iid. Compute the 95% confidence interval. Perform a Z test of equivalence. Give a p-value for the test.

## Hints

- Paired or unpaired experience? Standard Error?

# Exercise (from `Statistical inference for data science` by B. Caffo.)

## Question

Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. Assuming that the groups are independent and that they days are iid. Compute the 95% confidence interval. Perform a Z test of equivalence. Give a p-value for the test.

## Hints

- Paired or unpaired experience? Standard Error?
- Compute the confidence interval of $\bar{X} - \bar{Y}$.

# Exercise (from `Statistical inference for data science` by B. Caffo.)

## Question

Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. Assuming that the groups are independent and that they days are iid. Compute the 95% confidence interval. Perform a Z test of equivalence. Give a p-value for the test.

## Hints

- Paired or unpaired experience? Standard Error?
- Compute the confidence interval of $\bar{X} - \bar{Y}$.
- Compute TS

# Exercise (from `Statistical inference for data science` by B. Caffo.)

## Question

Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. Assuming that the groups are independent and that they days are iid. Compute the 95% confidence interval. Perform a Z test of equivalence. Give a p-value for the test.

## Hints

- Paired or unpaired experience? Standard Error?
- Compute the confidence interval of $\bar{X} - \bar{Y}$.
- Compute TS
- $H_0$ ? $H_a$?

# Exercise (from `Statistical inference for data science` by B. Caffo.)

## Question

Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. Assuming that the groups are independent and that they days are iid. Compute the 95% confidence interval. Perform a Z test of equivalence. Give a p-value for the test.

## Hints

- Paired or unpaired experience? Standard Error?
- Compute the confidence interval of $\bar{X} - \bar{Y}$.
- Compute TS
- $H_0$ ? $H_a$?
- One sided or two sided?

# Exercise (from `Statistical inference for data science` by B. Caffo.)

## Question

Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. Assuming that the groups are independent and that they days are iid. Compute the 95% confidence interval. Perform a Z test of equivalence. Give a p-value for the test.

## Hints

- Paired or unpaired experience? Standard Error?
- Compute the confidence interval of $\bar{X} - \bar{Y}$.
- Compute TS
- $H_0$ ? $H_a$?
- One sided or two sided?
- Compute the probability to get a a value as large as $\mid TS \mid$ and as small as $(- \mid TS \mid)$?

# Answer

## R code

```
n1<-100 ; n2<-100
m1<-10 ; m2<-11 ; alpha <- 1-0.95
mu0<-0 # Null hypothesis: m1-m2 = 0
s<-4 # Pooled variance
se <- s*sqrt(1/n1+1/n2) #Stand. err. for unpaired XP
ts <- ((m1-m2)-mu0)/se ; ts
q <- qnorm(1-alpha/2)
m1-m2+c(-1,1)*q*se #95% confidence inteval
p_value<-pnorm(abs(ts),lower.tail = FALSE) +
         pnorm(-abs(ts)) #Two-sided test
p_value # .07709987
2*pnorm(-abs(ts)) # Same thing
```

p-value > $\alpha = 0.05$. Hence, we do not reject the null hypothesis and we do not conclude that one scheme is better than the other.

# Discussion about the value

## Be carefull

- A p-value is only a probability.
- If you reject the null hypothesis based on p-value you might end up to make a mistake
- Ex: you rejected the null n=20 times with the same p-value $p = \alpha = 0.05$. The probability that you make at least one type I error is $1 - (1 - p)^n = 1 - 0.95^{20} = 64.1\%$

# p-value pro

- simply statistics
  http://simplystatistics.org/2012/01/06/p-values-and-hypothesis-testing-get-a-bad-rap-but-
- normal deviate]
  http://normaldeviate.wordpress.com/2013/03/14/double-misunderstandings-about-p-values/
- Error statistics
  http://errorstatistics.com/2013/06/14/p-values-cant-be-trusted-except-when-used-to-argue

# p-value cons

- Statistical Evidence: A Likelihood Paradigm* by Richard Royall
  http://www.crcpress.com/product/isbn/9780412044113
- Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy* by Steve Goodman
  https://scholar.google.com/scholar?q=towards+evidence+based+medical+statistics+the+p-value+fallacy&hl=en&as_sdt=0&as_vis=1&oi=scholart&sa=X&ei=uOTjVNHdG4anggSMlYOwBQ&ved=0CBsQgQMwAA
- The Earth is Round (p < .05)* by Cohen
  http://www.iro.umontreal.ca/~dift3913/cours/papers/cohen1994_The_earth_is_round.pdf

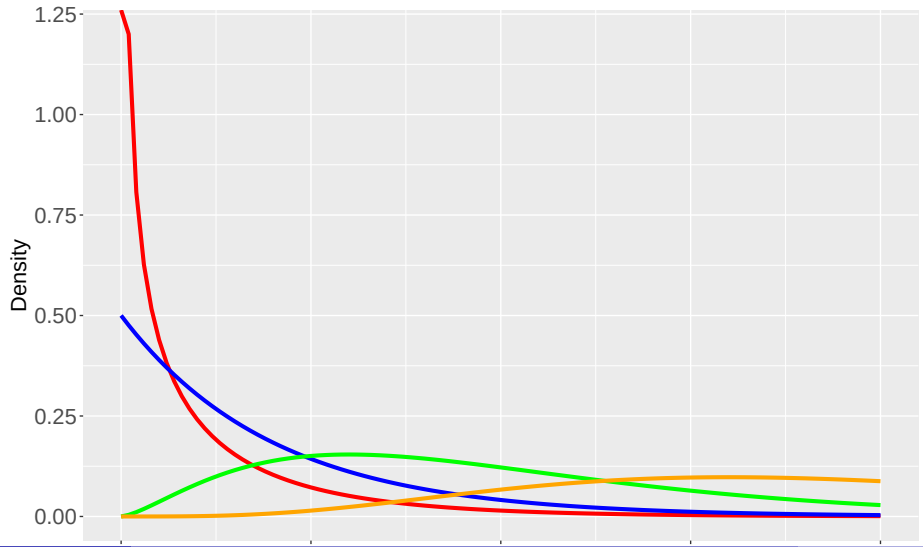
# Outline

# $\chi^2$ distribution

## Problem

- Testing if a sample follows a given distribution
- Testing if two random variables are independent

- Use the $\chi^2$ distribution.
- $\chi^2$ distribution: Let $X_1, \ldots, X_k$ $k$ random variables following the standard normal distribution $\mathcal{N}(0, 1)$.
  Then $X = \sum_n X_i^2$ follows the $\chi^2$ distribution with $k$ degrees of freedom.



# Testing if a sample follows a given distribution

## Problem

- $y_1, \ldots, y_N$ $N$ samples from a random variable $Y$
- $Y$ can take $J$ different distinct values $v_1, \ldots, v_J$
- $H_0$: $Y$ takes value $v_j$ with probability $p_j$ and $\sum p_j = 1$

## $\chi^2$ test for adequation

- $\alpha$: significance level
- $n_j$ number of sample with value $j$. $\hat{p}_j = n_j/N$
- $T = \sum_{j=1}^{J} \frac{(n_j - Np_j)^2}{Np_j}$ follow a $\chi^2$ distribution with $J-1$ degrees of freedom (Pearson 1900).
- Hence, if $T < Q_{\chi^2(J-1)}(1-\alpha)$ we fail to reject $H_0$.
- where $Q_{\chi^2(J-1)}(1-\alpha)$ is the quantile of the $\chi^2$ law with $(J-1)$ degrees of freedom with $p = 1-\alpha$ (R: use `qchisq(p, df)`).

# Example: testing the balance of a 6-faces dice

## Problem

- $\alpha = 0.05$
- 600 draws: 88, 109, 107, 94, 105, 97
- Can we reject $H_0$: "the dice is balance" with confidence $1 - \alpha$?
- Same thing for 600 draws: 89, 131,93, 92, 104, 91

# Example: testing the balance of a 6-faces dice

## Problem

- $\alpha = 0.05$
- 600 draws: 88, 109, 107, 94, 105, 97
- Can we reject $H_0$: "the dice is balance" with confidence $1 - \alpha$?
- Same thing for 600 draws: 89, 131,93, 92, 104, 91

## Answer

- Compute T
- Compare to $Q_{\chi^2(5)}(0.95)$

# Example: testing the balance of a 6-faces dice

## Problem

- $\alpha = 0.05$
- 600 draws: 88, 109, 107, 94, 105, 97
- Can we reject $H_0$: "the dice is balance" with confidence $1 - \alpha$?
- Same thing for 600 draws: 89, 131,93, 92, 104, 91

## Answer

- Compute T
- Compare to $Q_{\chi^2(5)}(0.95)$

## R code

```
Q<-qchisq(0.95,5)
T1 <- ((88-100)^2+(109-100)^2+(107-100)^2+(94-100)^2+(105-100)^2+(97-100)^2)/100
T1<Q
T2 <- ((89-100)^2+(131-100)^2+(93-100)^2+(92-100)^2+(104-100)^2+(91-100)^2)/100
T2<Q
```

# Testing if two random variables are independent

## Problem

- two random variables $X$ and $Y$.
- $I$ different possible values for $X$ and $J$ for $Y$.
- $H_0$: $X$ and $Y$ are independent.
- $O_{i,j}$ number of samples for which $X = i$ and $Y = j$.
- $E_{i,j} = \frac{O_{i+} \times O_{+j}}{N}$ : empirical expected value of having $X = i$ and $Y = j$, where:
  - $O_{i+} = \sum_{j=1}^{J} O_{i,j}$
  - $O_{+j} = \sum_{i=1}^{I} O_{i,j}$
- $T = \sum \frac{(O_{i,j} - E_{i,j})^2}{N}$ follows a $\chi^2$ distribution of (I-1)(J-1) degrees of freedom.
- If $T < Q_{\chi^2((I-1)(J-1))}(1-\alpha)$ we fail to reject $H_0$ with confidence $1-\alpha$.

# Example

## Problem

- two random variables $X$ and $Y$.
- $X$ can takes two values ('A' and 'B') and $Y$ four values (1, 2, 3 , 4).
- $H_0$: $X$ and $Y$ are independent.
- $\alpha = 0.05$

| | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| A | 50 | 70 | 110 | 60 | 290 |
| B | 60 | 75 | 100 | 50 | 285 |
| Total | 110 | 145 | 210 | 110 | 575 |

## Question

We see that number of B's are greater than number of A's for small values of $Y$ and the opposite for large values of $Y$. Is this statistically significant?

# Example

## Solution

| | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| A | 50 | 70 | 110 | 60 | 290 |
| B | 60 | 75 | 100 | 50 | 285 |
| Total | 110 | 145 | 210 | 110 | 575 |

- Degrees of freedom?
- Quantile: Q=7.814728
- T=2.423491
- $T < Q$ Failed to reject $H_0$

# Remarks

## Validity of the test

- $T$ converge towards a $\chi^2$ law
- Hence, test valid only if sample in each category is large enough
- How large? No consensus, but at least 5 (or 10 or 20, depending on the authors).

## R code

- `chisq.test(x,p)`
- `x`: sample size (vector or matrix)
- `p`: probability vector (omit if equi-probability)
- `chisq.test(c(88,109,107,94,105,97))` for 6-faces dice problem
- `chisq.test(x=matrix(data=c(50, 70, 110, 60, 60, 75,100, 50),nrow=2,ncol=4, byrow=TRUE))`

# Outline

# Computing the number of experiments

## Problem

You have a confidence interval, how many more experiments (n) you need to reduce your confidence interval to a given level ($\epsilon$)?

CI of the mean: $\mu \in [\bar{x} - z_{1-\alpha/2} s/\sqrt{n}, \bar{x} + z_{1-\alpha/2} s/\sqrt{n}]$

If you want: $\mu \in [\bar{x}(1-\epsilon), \bar{x}(1+\epsilon)]$

$$n \geq \left( \frac{z_{1-\alpha/2} s}{\bar{x}\epsilon} \right)^2$$

CI of a proportion: $p_1 \in \left[ \hat{p}_1 - z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{p}_2}{n}}, \hat{p}_1 + z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{p}_2}{n}} \right]$

If you want $p1 \in [\hat{p}_1 - \epsilon, \hat{p}_1 + \epsilon]$

$$n \geq \frac{z^2_{1-\alpha/2} \hat{p}_1 \hat{p}_2}{\epsilon^2}$$

# Example

## For the mean

$\bar{x} = 10$, $s = 2$

# Example

## For the mean

$\bar{x} = 10$, $s = 2$

$\alpha = 0.1 \Rightarrow z_{0.95} = 1.64$

# Example

## For the mean

$\bar{x} = 10$, $s = 2$

$\alpha = 0.1 \Rightarrow z_{0.95} = 1.64$

$\epsilon = 0.05 \Rightarrow n \geq \left(\frac{1.64 \times 2}{10 * 0.05}\right)^2 = 43$

# Example

## For the mean

$\bar{x} = 10$, $s = 2$

$\alpha = 0.1 \Rightarrow z_{0.95} = 1.64$

$\epsilon = 0.05 \Rightarrow n \geq \left(\frac{1.64 \times 2}{10 * 0.05}\right)^2 = 43$

## For proportion

$n = 40$, $n_1 = 30 \Rightarrow \hat{p}_1 = 30/40 = 0.75$, $\hat{p}_2 = 0.25$

# Example

## For the mean

$\bar{x} = 10$, $s = 2$

$\alpha = 0.1 \Rightarrow z_{0.95} = 1.64$

$\epsilon = 0.05 \Rightarrow n \geq \left(\frac{1.64 \times 2}{10 * 0.05}\right)^2 = 43$

## For proportion

$n = 40$, $n_1 = 30 \Rightarrow \hat{p}_1 = 30/40 = 0.75$, $\hat{p}_2 = 0.25$

$\alpha = 0.01 \Rightarrow z_{0.995} = 2.58$

# Example

## For the mean

$\bar{x} = 10$, $s = 2$

$\alpha = 0.1 \Rightarrow z_{0.95} = 1.64$

$\epsilon = 0.05 \Rightarrow n \geq \left(\frac{1.64 \times 2}{10 * 0.05}\right)^2 = 43$

## For proportion

$n = 40$, $n_1 = 30 \Rightarrow \hat{p}_1 = 30/40 = 0.75$, $\hat{p}_2 = 0.25$

$\alpha = 0.01 \Rightarrow z_{0.995} = 2.58$

$\epsilon = 0.005 \Rightarrow n \geq \frac{2.58^2 \times 0.75 \times 0.25}{0.005} = 250$

# Outline

# Conclusion

- Many sciences involve experiments (computer-science is among them).
- There are a lot of tools and methods to perform insightful experiments:
  - General methodology
  - Performance analysis
  - Statistics and probability
  - Data analysis and representation
- This course is an attempt to tackle this issue.

# Acknowledgments and further reading

- Bad Stats: Not what it Seems
  `http://www.aviz.fr/badstats`
- MOOC *Statistical Inference* on Coursera.
- `https://leanpub.com/LittleInferenceBook` that comes with the above course

# Further reading