Modeling Biological Data

Regression models

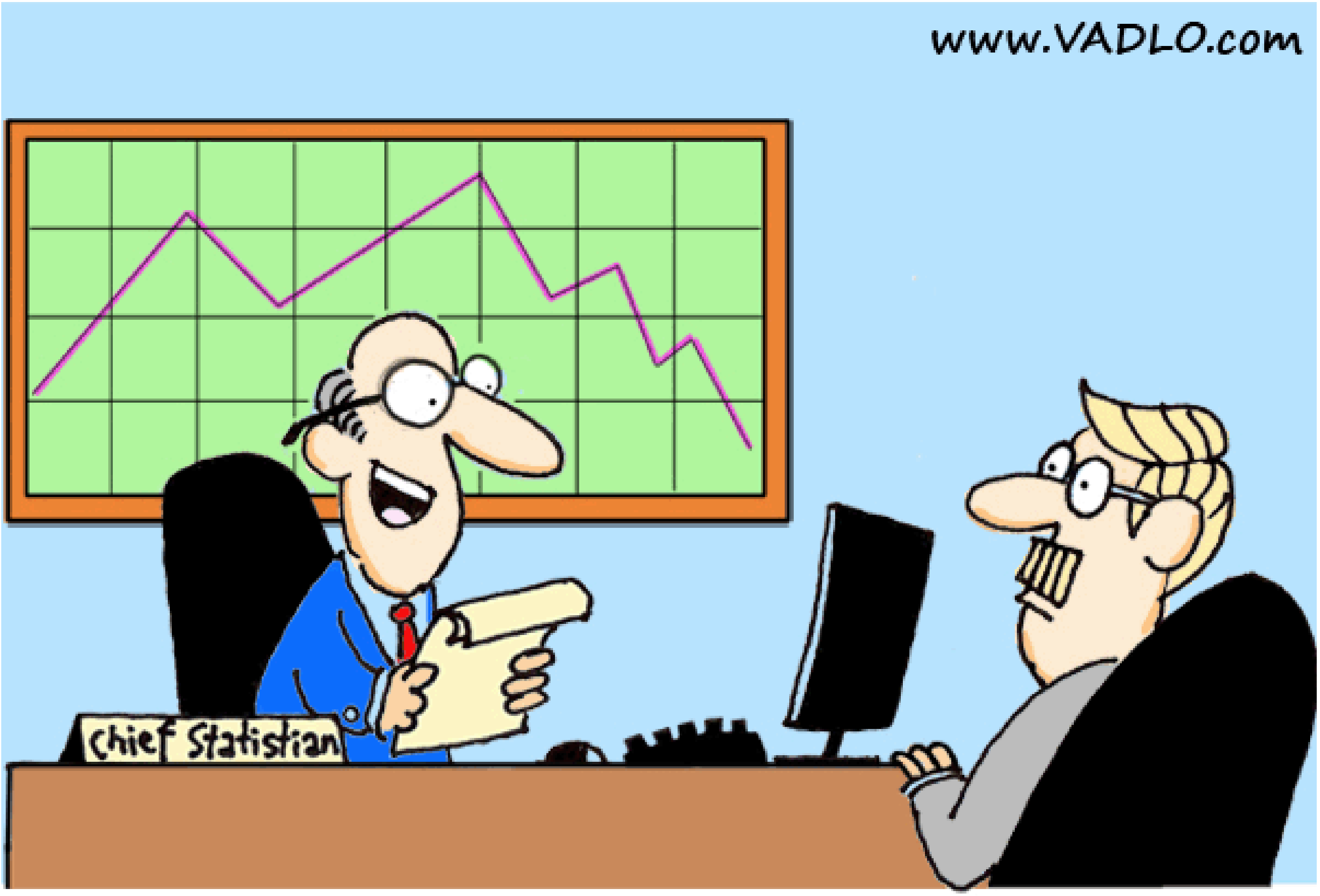
Ruy M Ribeiro

Los Alamos National Laboratory

DRAFT NOTES FOR PRESENTATION

- “Using statistical thinking to reach conclusions in clinical practice and the biomedical sciences amounts to much more than memorizing a few formulas and looking for P values.”

S. Glantz, “Primer of Biostatistics”



“I can prove it or disprove it! What do you want me to do?”

Why regression modeling?

- “Almost all of statistics is linear regression, and most of what is left over is non-linear regression.”

R. Jennrich *in* P.J. Green, J Royal Stat Soc B **46**: 149 (1984)

Introduction

Reality and Models

- Conceptual Models
- Mechanistic Models
- Statistical Models

- **Martin et al “Predictors of Limb Fat Gain in HIV Positive Patients Following a Change to Tenofovir-Emtricitabine or Abacavir-Lamivudine”**

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026885

  - Conceptual model
  - Mechanistic model
  - Statistical model

  - Statistics section: is it all linear regression?

Objectives of Statistics

- Analyze and interpret data
- Inference from individual to population
  - Confidence intervals
  - Statistical tests

Minimal approach for regression models

- Describe the data
- Understand probability distributions (infer from the data)
- Define/apply the model

The First Step is the most important:

Experimental design

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.”

Ronald Fisher (1890- 1962)

Cortesia Prof. Doutor João Carriço

Data set

- Gene expression data
  - dataGeneExp.csv
  - Read it as `genExp`

- Describe the data
  - Use summaries
  - Use plots

- **Exploratory data analyses**

- **Inference**
  - **Need to know a little more about the data**

Descriptive statistics

- Measures of location
  - Mean (several different); Mode; Median
  - Percentiles; quantiles
- Measures of spread
  - Variance (standard deviation)
  - Absolute mean deviation
  - Range, or Interquartile range
  - Coefficient of variation
- Measures of correlation:
  - Pearson correlation / coefficient of determination
  - Risk; odds-ratio
- Plotting the data (e.g. boxplots)

Plotting Data

- Clear labels
- Include consistent scales
- Label axes and include units
- Legends should tell what variables are plotted
- Make it simple, avoid clutter

- Caption should include source of data
- Focus on the data

Concepts

- Population
- Sample
- Observation

- Statistic and Parameter
  - The first is a single quantity calculated from the sample. A statistic (an estimator) can be used to estimate a population parameter.
  - Sample mean is a statistic that estimates the *population mean*, which is a parameter.

Concepts

- Dependent Variable
- Independent (Explanatory) Variable

- Variables
  – Quantitative (discrete, continuous) - scale
  – Qualitative (nominal, ordinal)
    - Factors / Levels

- Mean as a statistical linear model
  - $E[y_i]=\mu$
  - $y_i \sim 1$ (Model formula in R)

  (Details are missing ☺)

- Exploratory data analyses

- Inference
  - Need to know a little more about probabilities

- “All who drink of this remedy recover in a short time, except those whom it does not help, who all die. Therefore, it is obvious that it fails only in incurable cases.”
  - Galen (130 – 210 A.D.)

- “To be uncertain is uncomfortable, but to be certain is to be ridiculous.”
  - Chinese proverb

Fundamental concept of statistical test

Given that the null hypothesis is TRUE, what is the probability of observing the result that we obtained (or more extreme)?

p-value

Probability Distributions

- The probability that a random variable takes a certain value (discrete) or value interval (continuous)
  - How are the values of the variable “distributed”?

Some examples

- A die
- Sum of two dice

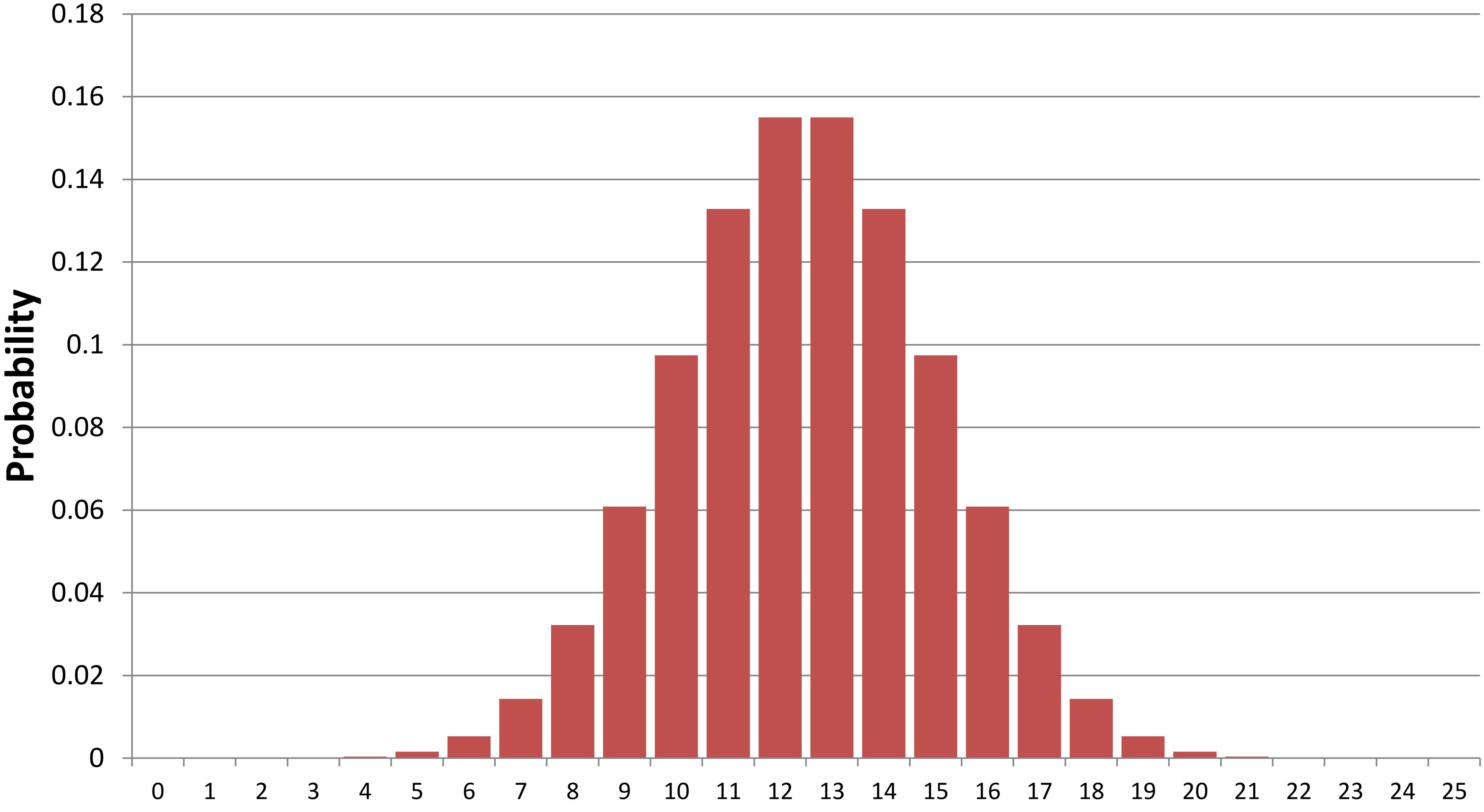
Distributions

Total on dice	Pairs of dice	Probability
2	1+1	1/36 = 3%
3	1+2, 2+1	2/36 = 6%
4	1+3, 2+2, 3+1	3/36 = 8%
5	1+4, 2+3, 3+2, 4+1	4/36 = 11%
6	1+5, 2+4, 3+3, 4+2, 5+1	5/36 = 14%
7	1+6, 2+5, 3+4, 4+3, 5+2, 6+1	6/36 = 17%
8	2+6, 3+5, 4+4, 5+3, 6+2	5/36 = 14%
9	3+6, 4+5, 5+4, 6+3	4/36 = 11%
10	4+6, 5+5, 6+4	3/36 = 8%
11	5+6, 6+5	2/36 = 6%
12	6+6	1/36 = 3%

Three ways to define distributions

- Counting the possibilities
- Doing the experiment
- Theoretical analysis

Binomial



Probability
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Probability distributions

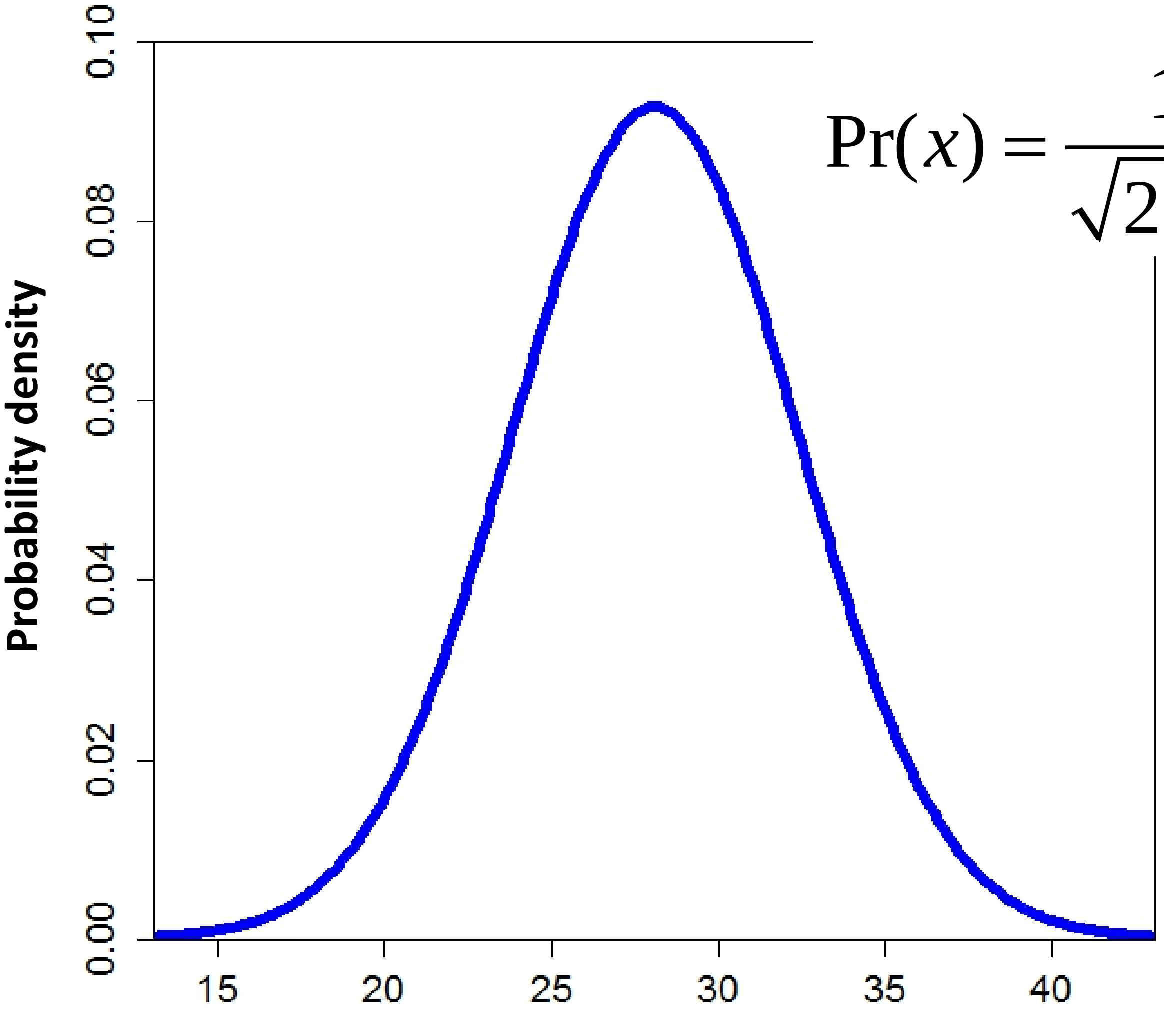
Distributions

- Negative binomial
- Poisson
- Normal
- T distribution
- Chi-squared
- Exponential
- Log normal
- Uniform
- F distribution

Properties

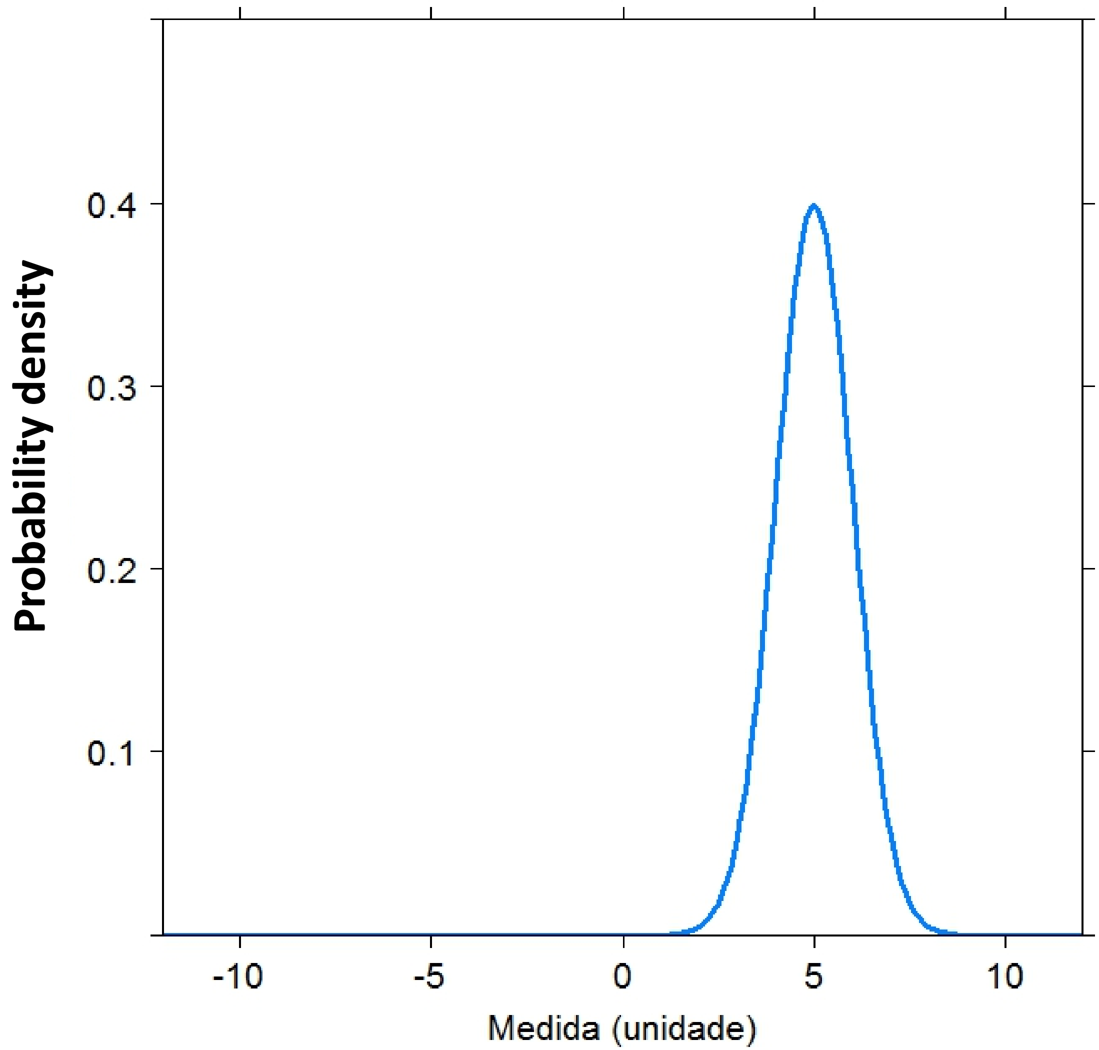
- Meaning
- Use
- Parameters
- Relation with other
- Any other interesting info

Normal distribution

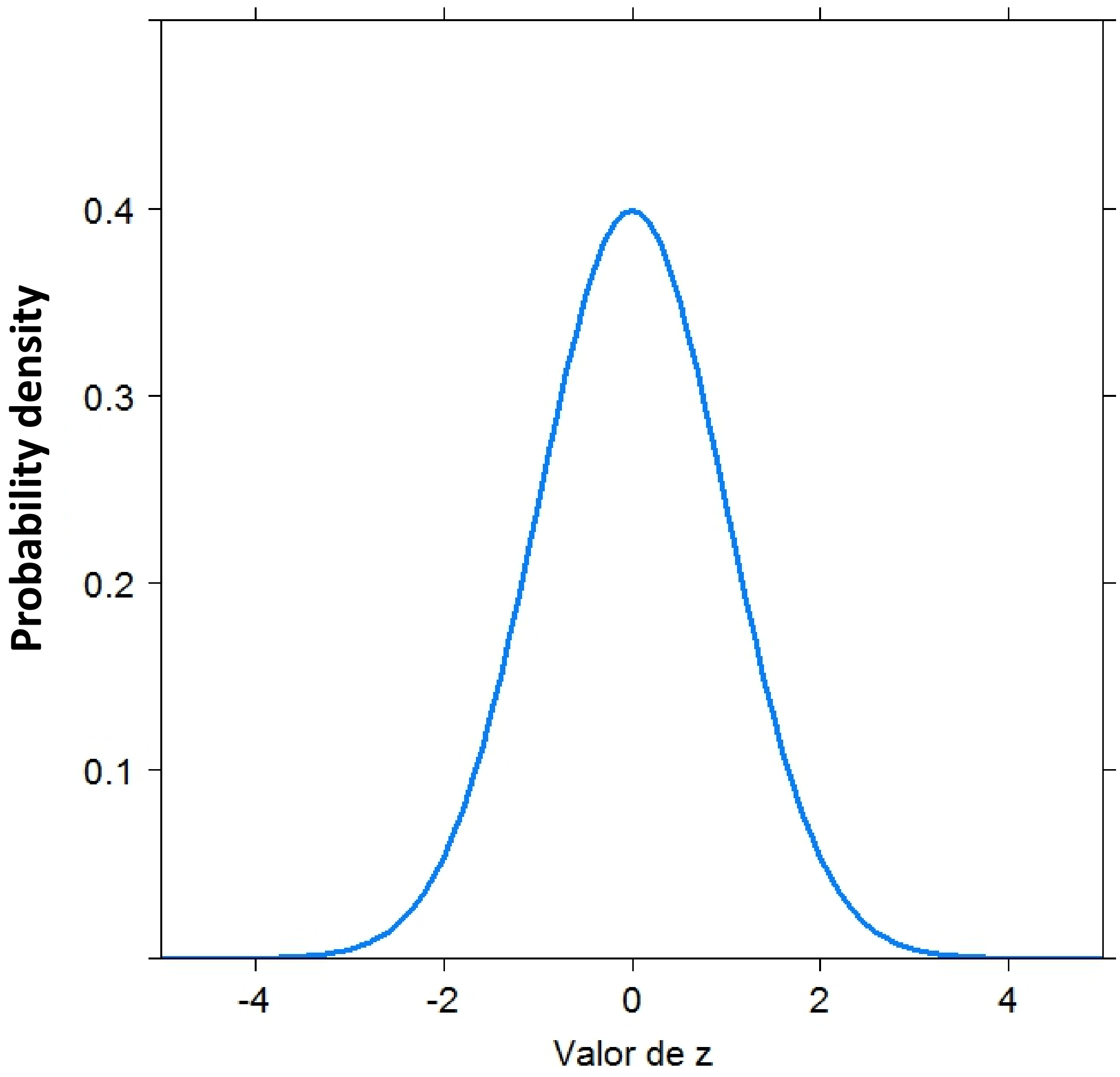


$$\Pr(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

N(μ,σ^2) a family of distributions



Standard normal distribution



$$Z = \frac{X - \mu}{\sigma}$$

Normal distribution

- Defined by mean and variance
- If μ=0 and σ^2=1, “standard normal distribution”

- Appears in many contexts because “any random variable that can be expressed as the sum of many other random variables can be well approximated by a normal distribution” (Rosner)

Central Limit Theorem

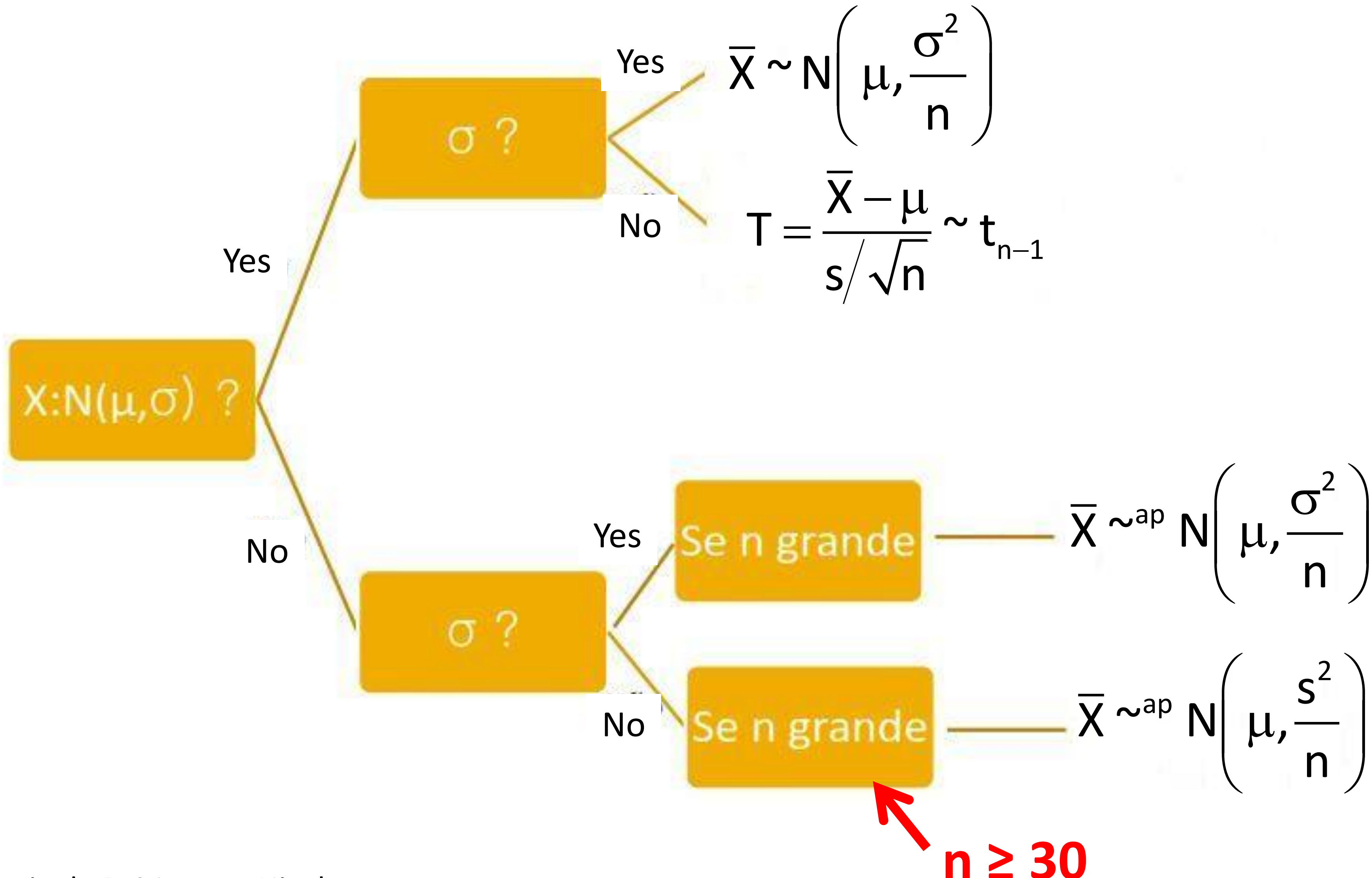
- Generate sample from a given distribution
  - runif, rpoiss, rbinom, rlnorm, …
  - Your own list of numbers(!)

- Calculate mean (save in a vector)

- Repeat 1000 times (you should have 1000 means)

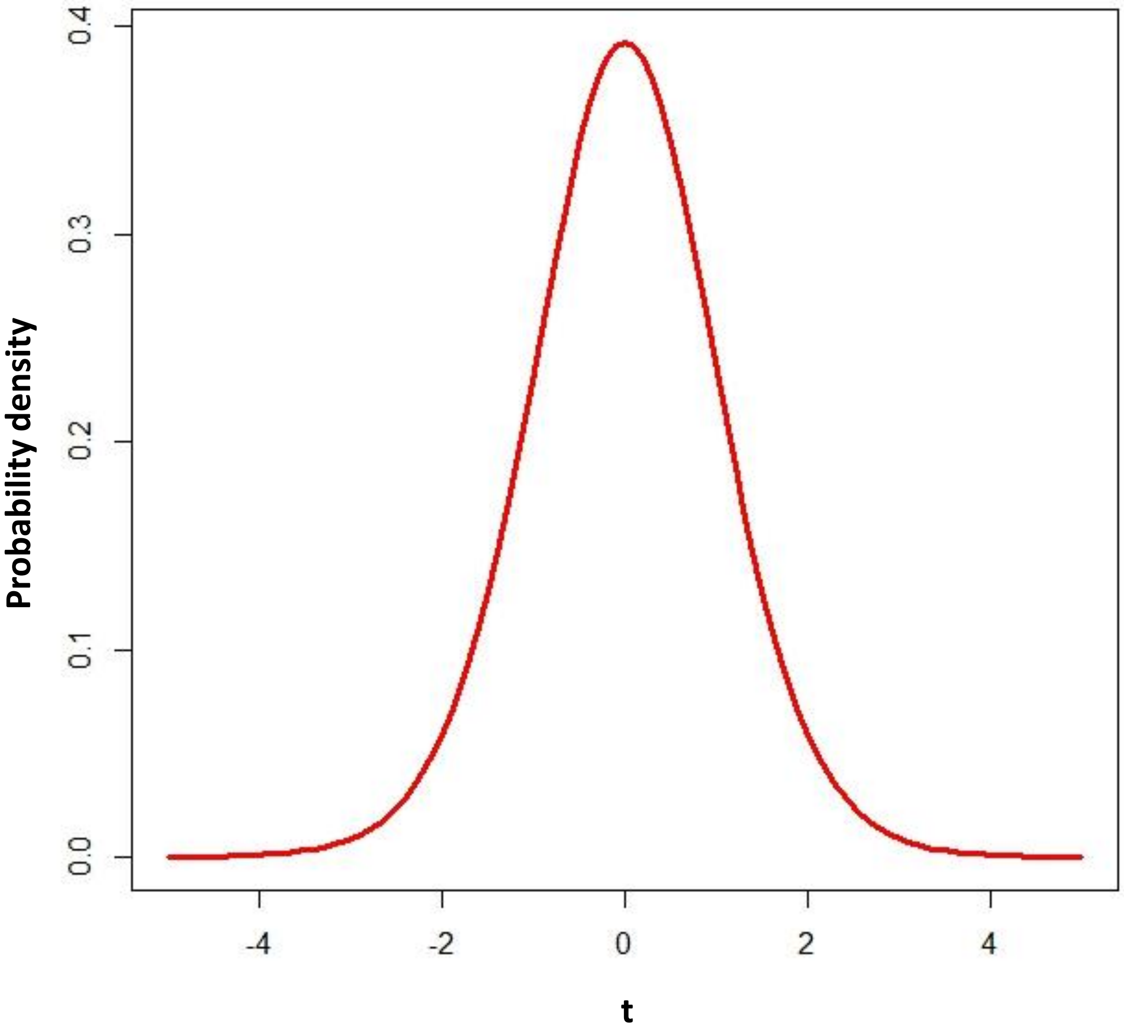
- Draw histogram of the 1000 means

The distribution of the sample mean



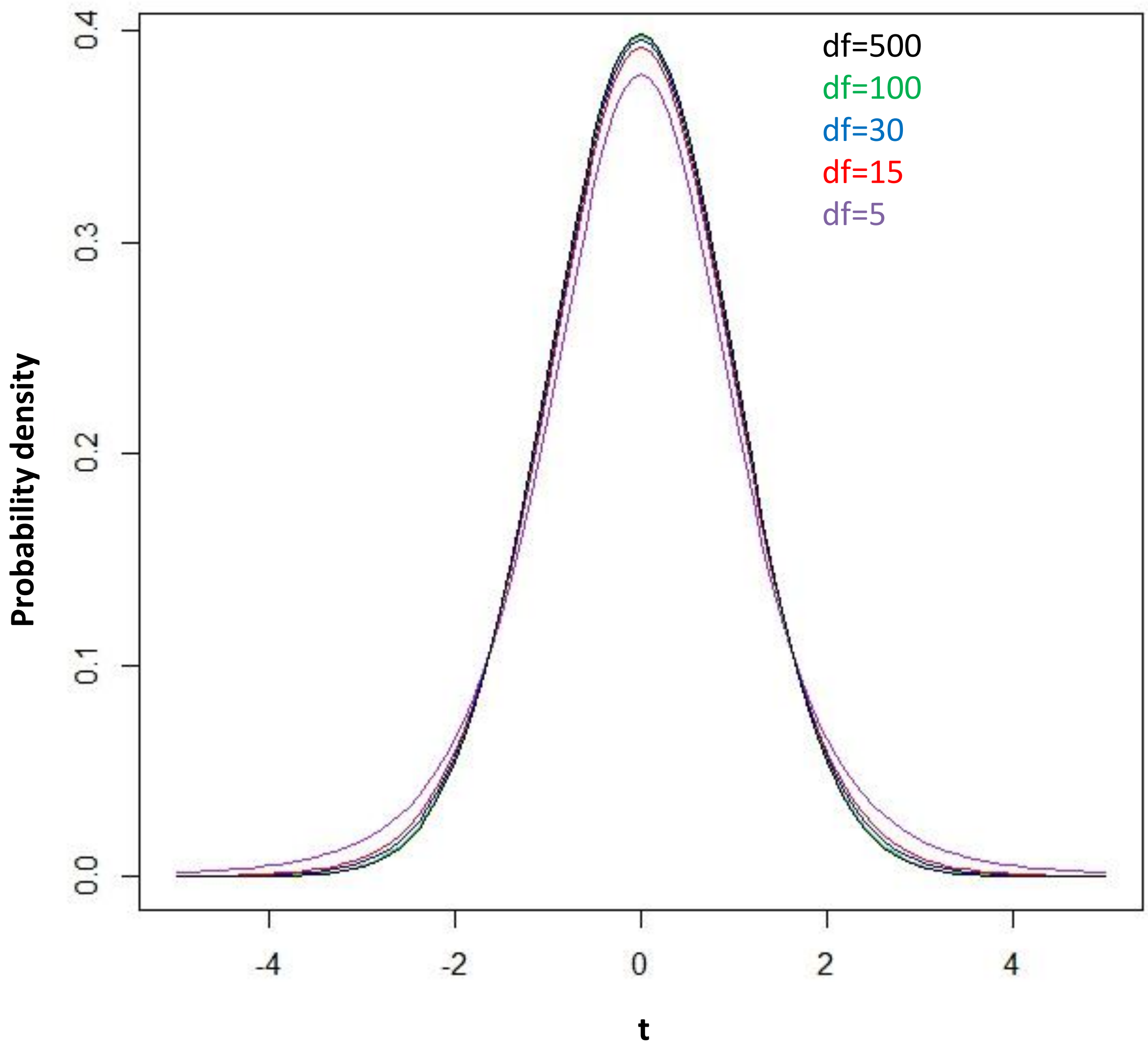
Cortesia de Drª Leonor Nicolau

The *t* distribution



$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

The *t* distributions



Confidence Intervals

Confidence Interval is an interval estimate of a population parameter and is used to indicate the reliability of an estimate.

Confidence Intervals

$$P\left(t_{df,0.025} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{df,0.975} \right) = 0.95$$

$$P\left(\bar{X} - \frac{s}{\sqrt{n}} t_{df,0.975} \leq \mu \leq \bar{X} - \frac{s}{\sqrt{n}} t_{df,0.025} \right) = 0.95$$

- If we built many such intervals (depend on $\bar{X}$ and s, here random variables), 95% of the times μ would be within

Cholesterol Treatment

- Difference (before – after)

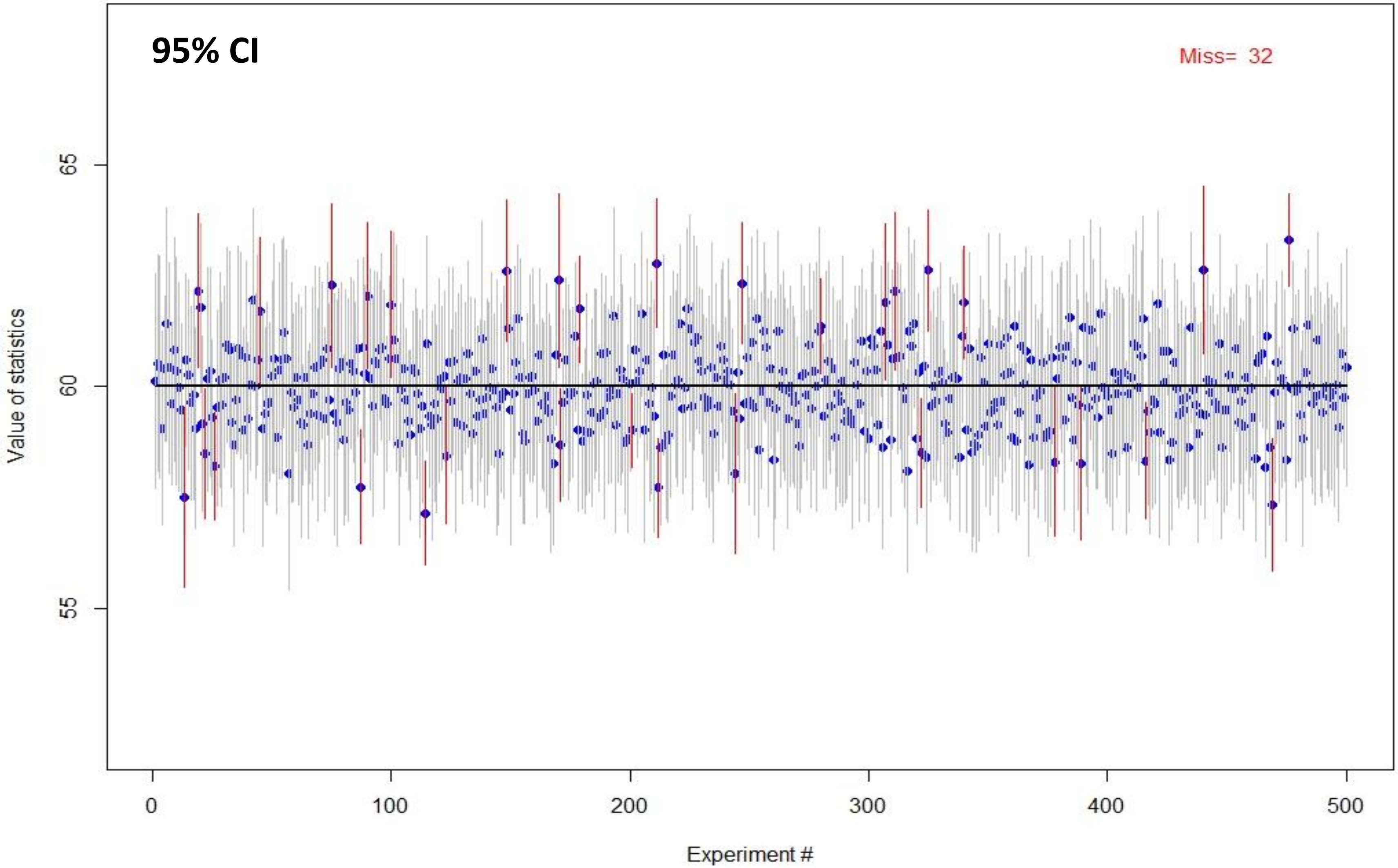
  -8, 60, 34, 39, 29, 3, -13, 32, 23, -59, 39, 20, 7, 37, 23

  Assume these differences have a normal distribution

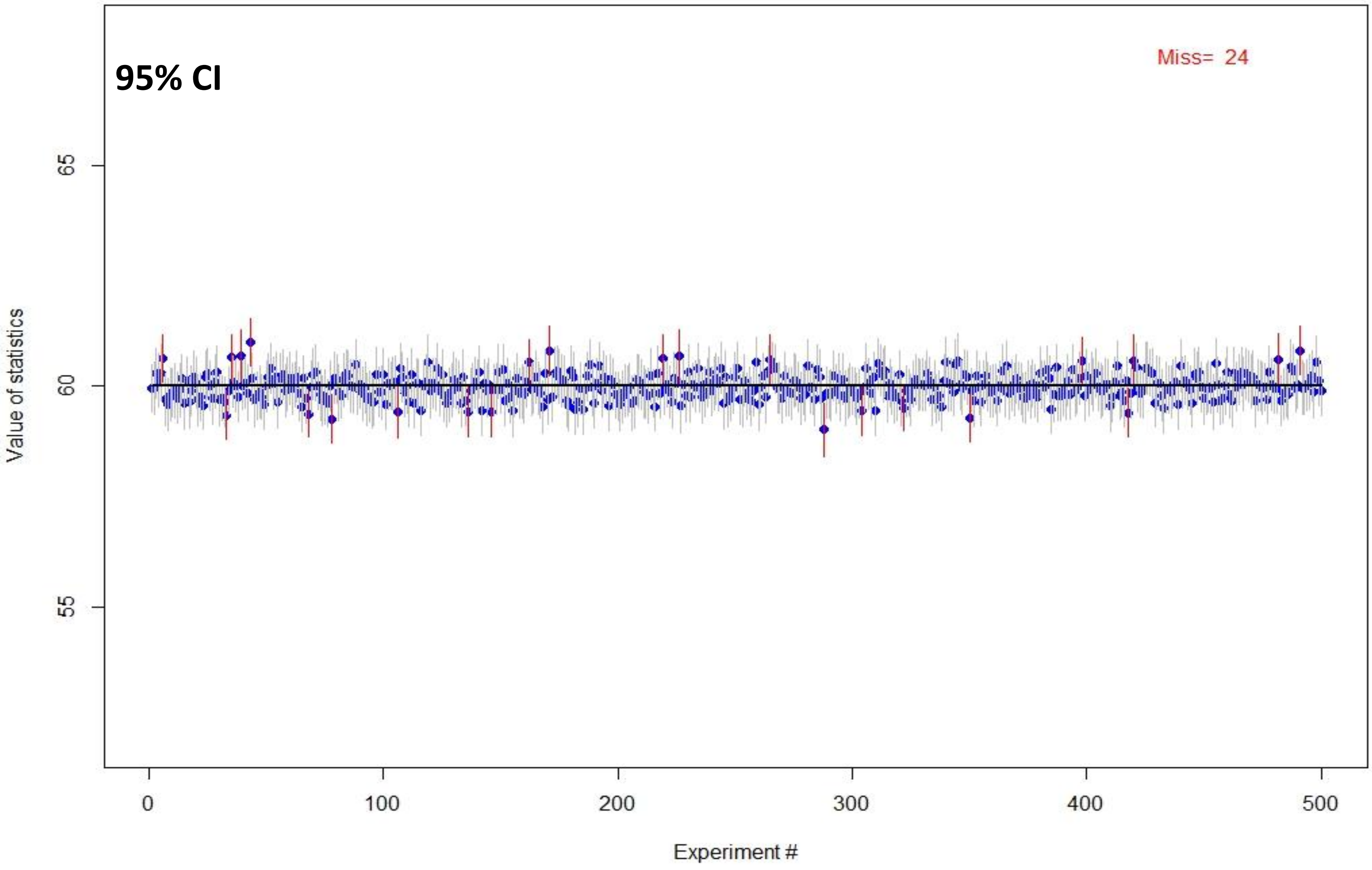
- Mean = 17.6
- Standard deviation= 28.6

- **What would happen if we repeated the experiment?**

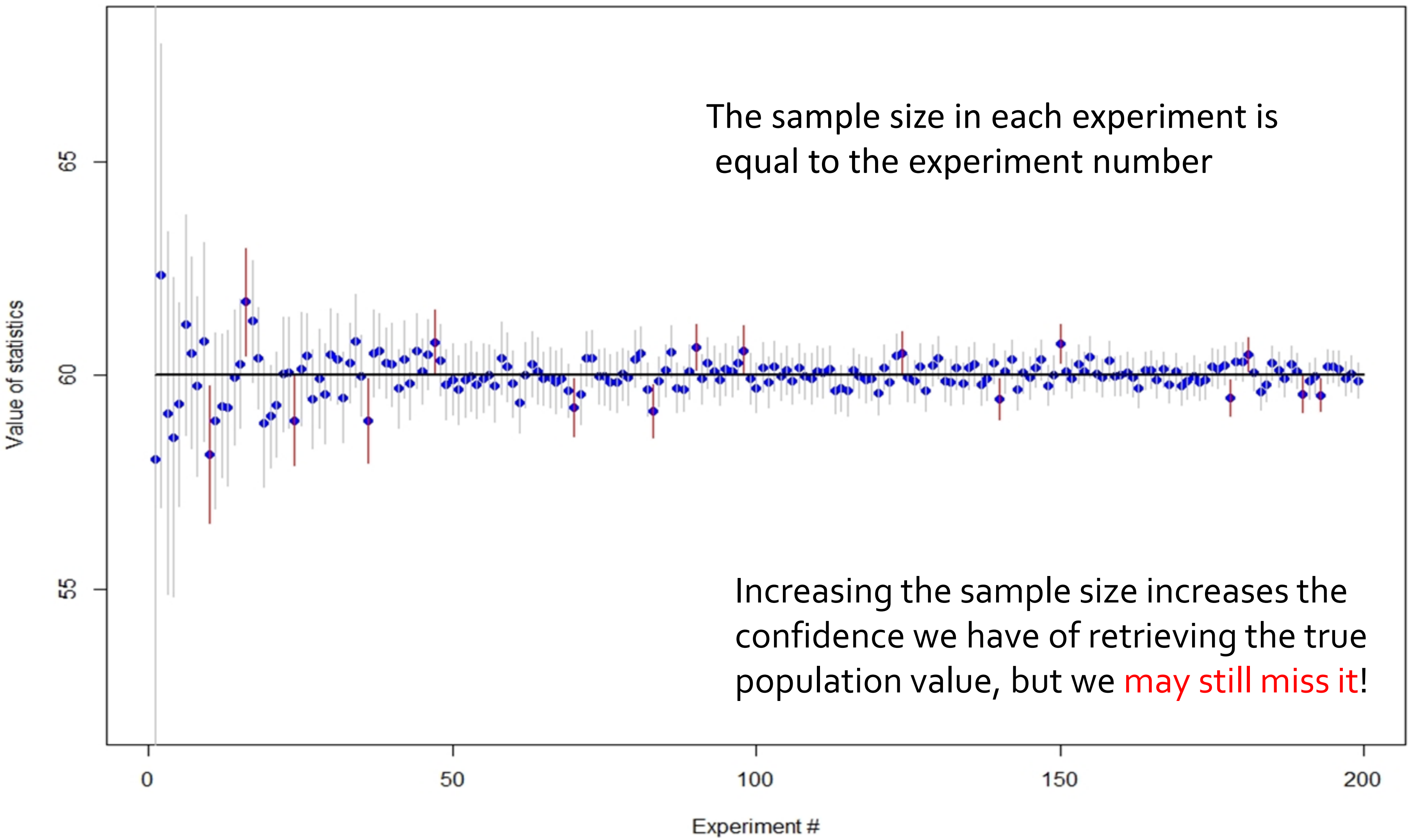
Meaning of confidence interval



Meaning of confidence interval



Meaning of confidence interval



The sample size in each experiment is equal to the experiment number

Increasing the sample size increases the confidence we have of retrieving the true population value, but we may still miss it!

Error bars

Table I. **Common error bars**

Error bar	Type	Description	Formula
Range	Descriptive	Amount of spread between the extremes of the data	Highest data point minus the lowest
Standard deviation (SD)	Descriptive	Typical or (roughly speaking) average difference between the data points and their mean	$SD = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$
Standard error (SE)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times	$SE = SD/\sqrt{n}$
Confidence interval (CI), usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean	$M \pm t_{(n-1)} \times SE$, where $t_{(n-1)}$ is a critical value of t. If n is 10 or more, the 95% CI is approximately $M \pm 2 \times SE$.

Cumming et al. Error bars in experimental biology, J Cell Biol 177: 7 (2007)

RULES FOR ERROR BARS*

- **Rule 1:** Always describe in the figure legend what the error bars are.
- **Rule 2:** The value of *n* should be indicated.
- **Rule 3:** Error bars/statistics are valid only for independent experiments, i.e. biological replicates and not technical replicates.
- **Rule 4:** It does make sense to use inferential error bars (but n should be reasonable).

*According to Cumming et al.

Practical concept of hypothesis test

Given that the null hypothesis is TRUE, what is the probability of observing the result that we obtained (or more extreme)?

p-value

Cholesterol Treatment (I)

- Before
  - 82, 163, 147, 114, 174, 128, 131, 104, 148, 147, 155, 86, 142, 130, 117
- After
  - 89, 103, 113, 75, 145, 126, 144, 72, 125, 206, 117, 66, 135, 93, 94

Assume normal distribution of cholesterol values in the population

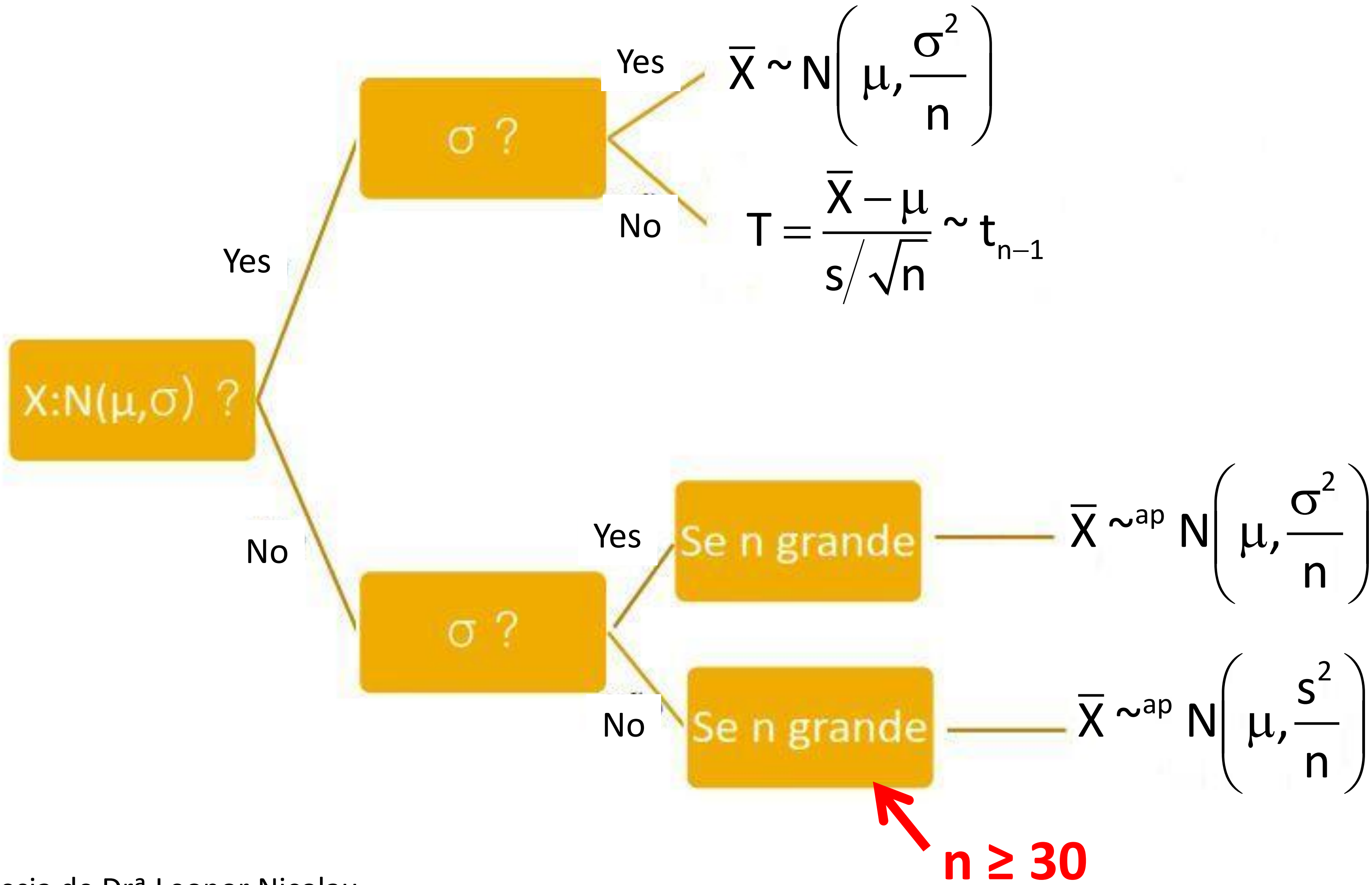
Given that the null hypothesis is TRUE, what is the probability of observing the result that we obtained (or more extreme)?

p-value

Cholesterol Treatment (II)

- Difference (before – after)
  - -8, 60, 34, 39, 29, 3, -13, 32, 23, -59, 39, 20, 7, 37, 23

The distribution of the sample mean



Cortesia de Drª Leonor Nicolau

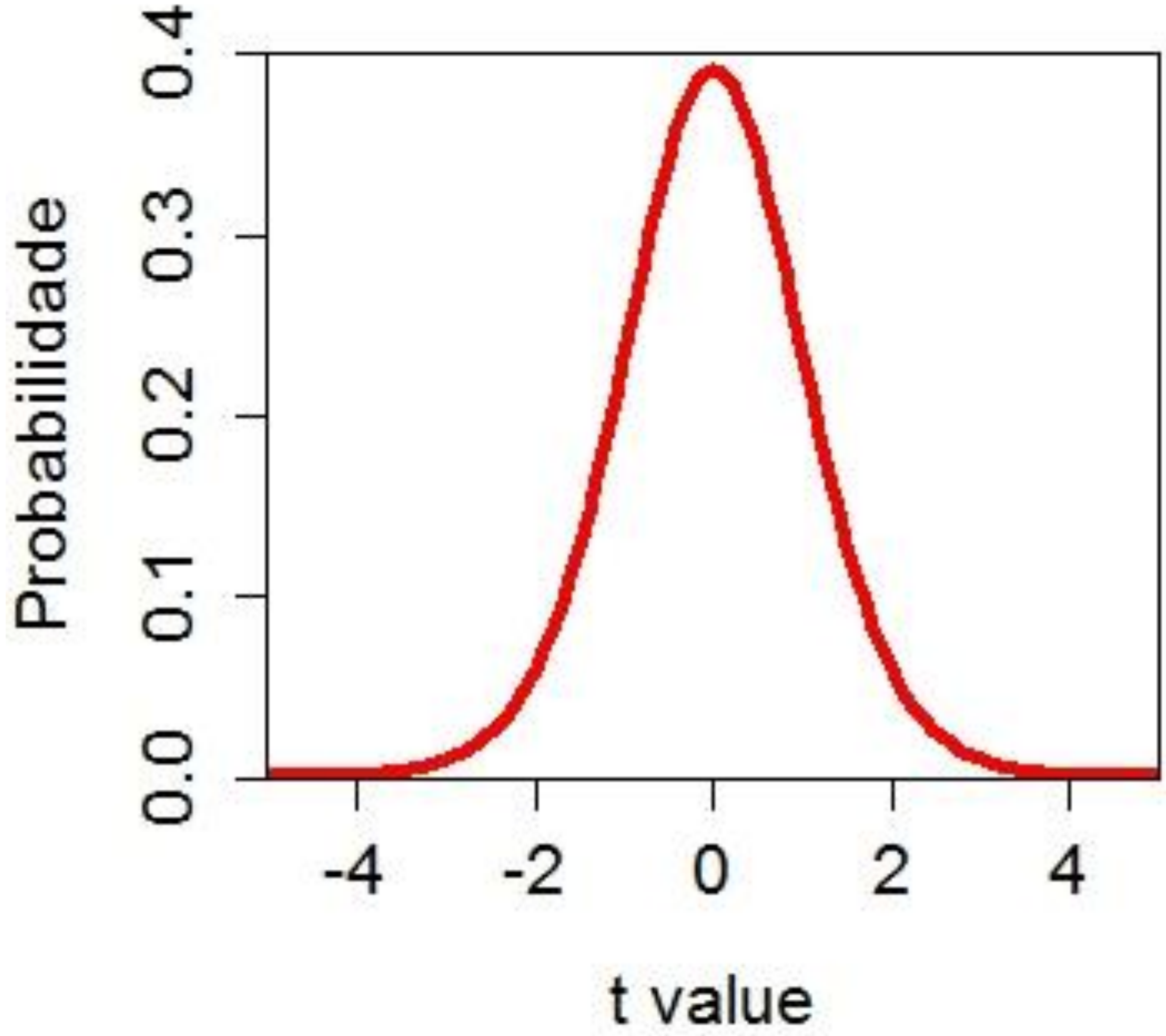
Cholesterol Treatment (II)

- Difference (before – after)

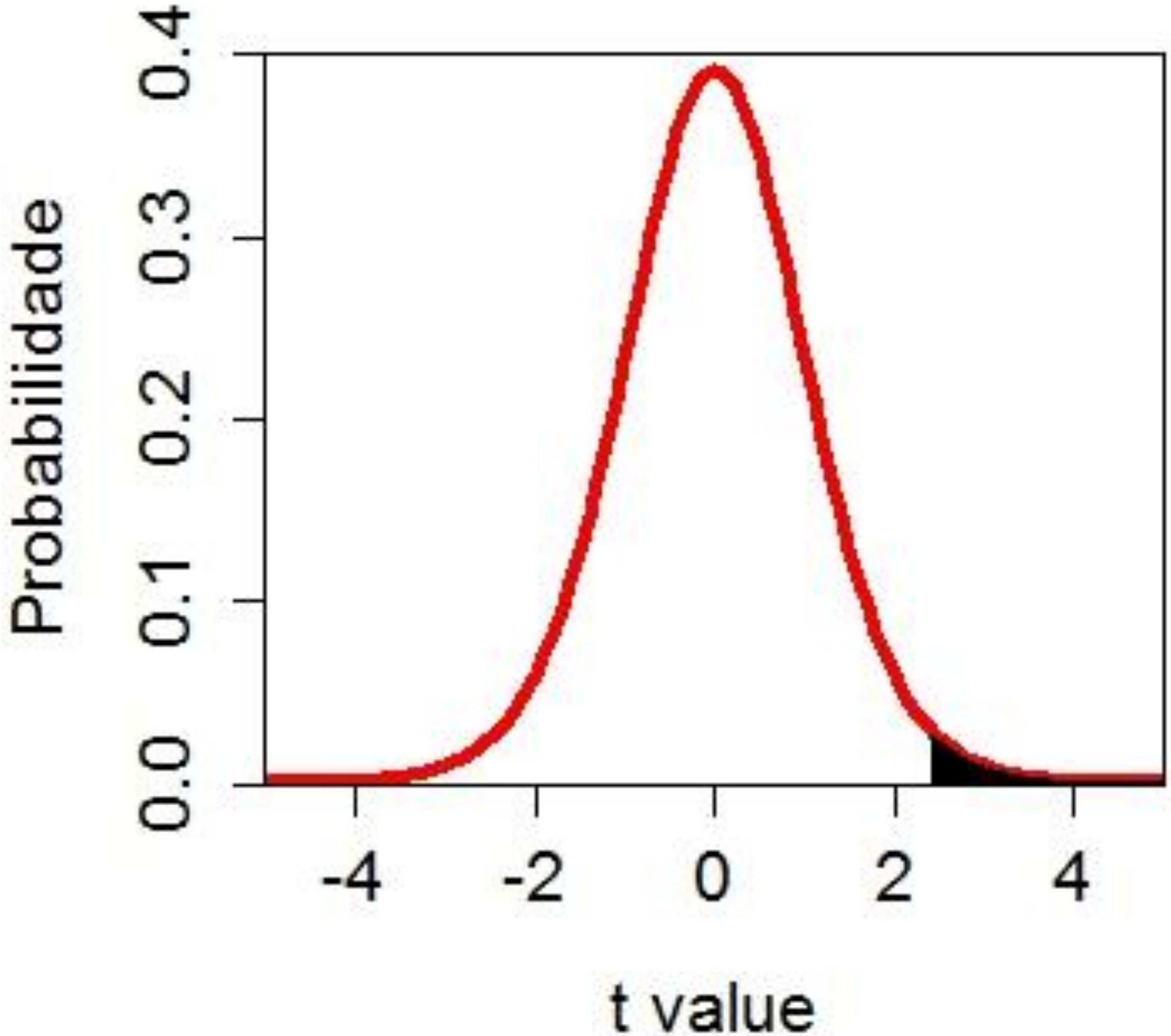
  -8, 60, 34, 39, 29, 3, -13, 32, 23, -59, 39, 20, 7, 37, 23

- Mean = 17.6
- Standard deviation= 28.6
- Statistic, $t_{14} = \dfrac{17.6 - 0}{28.6 / \sqrt{15}} = 2.38$

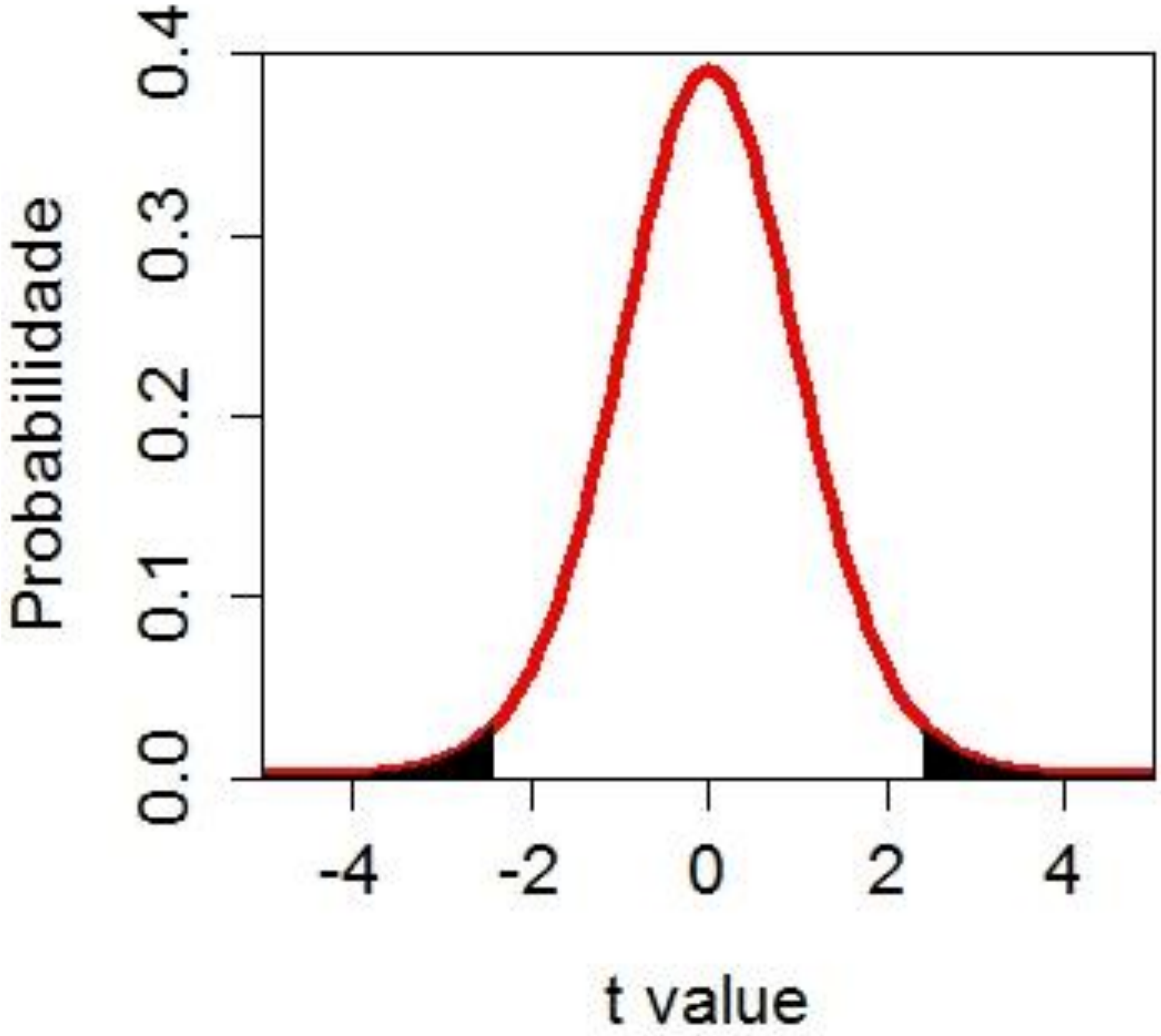
T distribution



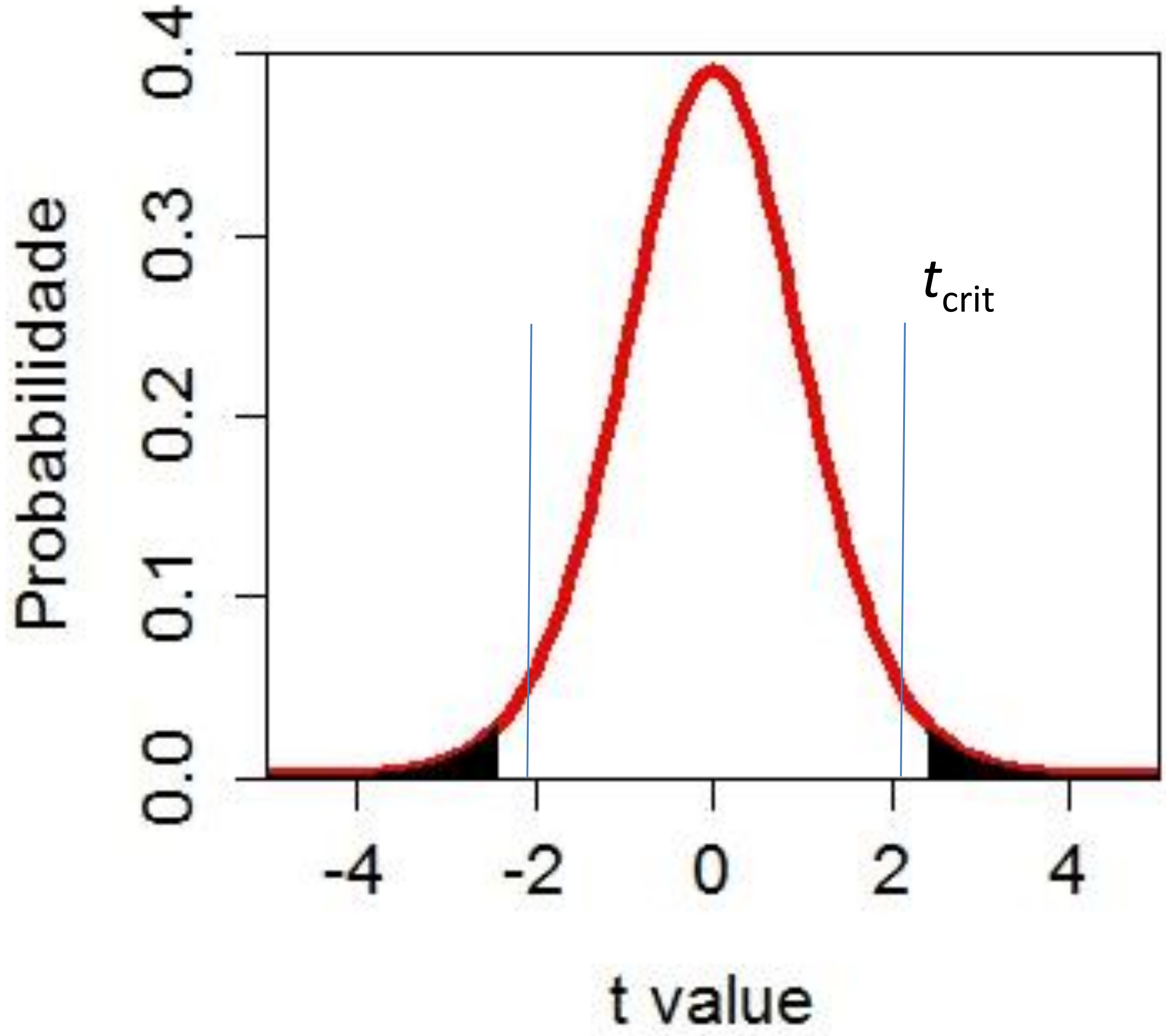
T distribution



T distribution



T distribution



Fundamental concept of statistical test

Given that the null hypothesis is TRUE, what is the probability of observing the result that we obtained (or more extreme)?

p-value

Is the probability large or small?

- The significance level, α
- Decided a priori
- Depending on study objectives

The P value is not…

- … the probability that the null hypothesis is true.
- … the probability that a finding is "merely by chance“.
- … the probability of falsely rejecting the null hypothesis
- … the probability that a replicate experiment would not yield the same conclusion.

Cortesia Prof. João Carriço

More on the P value

- 1−p is not the probability of the alternative hypothesis being true
- The significance level of the test (α) is not determined by the p-value.
- The p-value does not indicate the size or importance of the observed effect

Cortesia Prof. João Carriço

Modeling Biological Data

Regression models

Ruy M Ribeiro

Los Alamos National Laboratory

DRAFT NOTES FOR PRESENTATION

Given that the null hypothesis is TRUE, what is the probability of observing the result that we obtained (or more extreme)?

p-value

Possible decisions

		NULL HYPOTHESIS	
		NOT REJECTED	REJECTED
GIVEN NULL HYPOTHESIS IS	TRUE	√	Type I Error
	FALSE	Type II Error	√

		NULL HYPOTHESIS	
		NOT REJECTED	REJECTED
GIVEN NULL HYPOTHESIS IS	TRUE	√	Significance level
	FALSE	β	√

POWER OF THE TEST = 1-β

TRADE-OFF

Controlling type I error

- P(rejecting H_0 when H_0 is true)= *p*
- Typically we want $p<\alpha$ (the significance level)

- If $\alpha=0.05$, we are willing to make a mistake 1 in 20 times. What happens if we make repeated comparisons (tests)?

Correcting multiple comparisons

- Bonferroni correction
  - k comparisons
  - $\alpha = \alpha_T / k$ (where, e.g., $\alpha_T = 0.05$)

- Holm correction
  - Order p values (smaller to larger)
  - $\alpha = \alpha_T / (k-j+1)$, where j is order of comparison

  - Sidak's version uses (solution of) $\alpha_T = 1-(1-\alpha)^{(k-j+1)}$

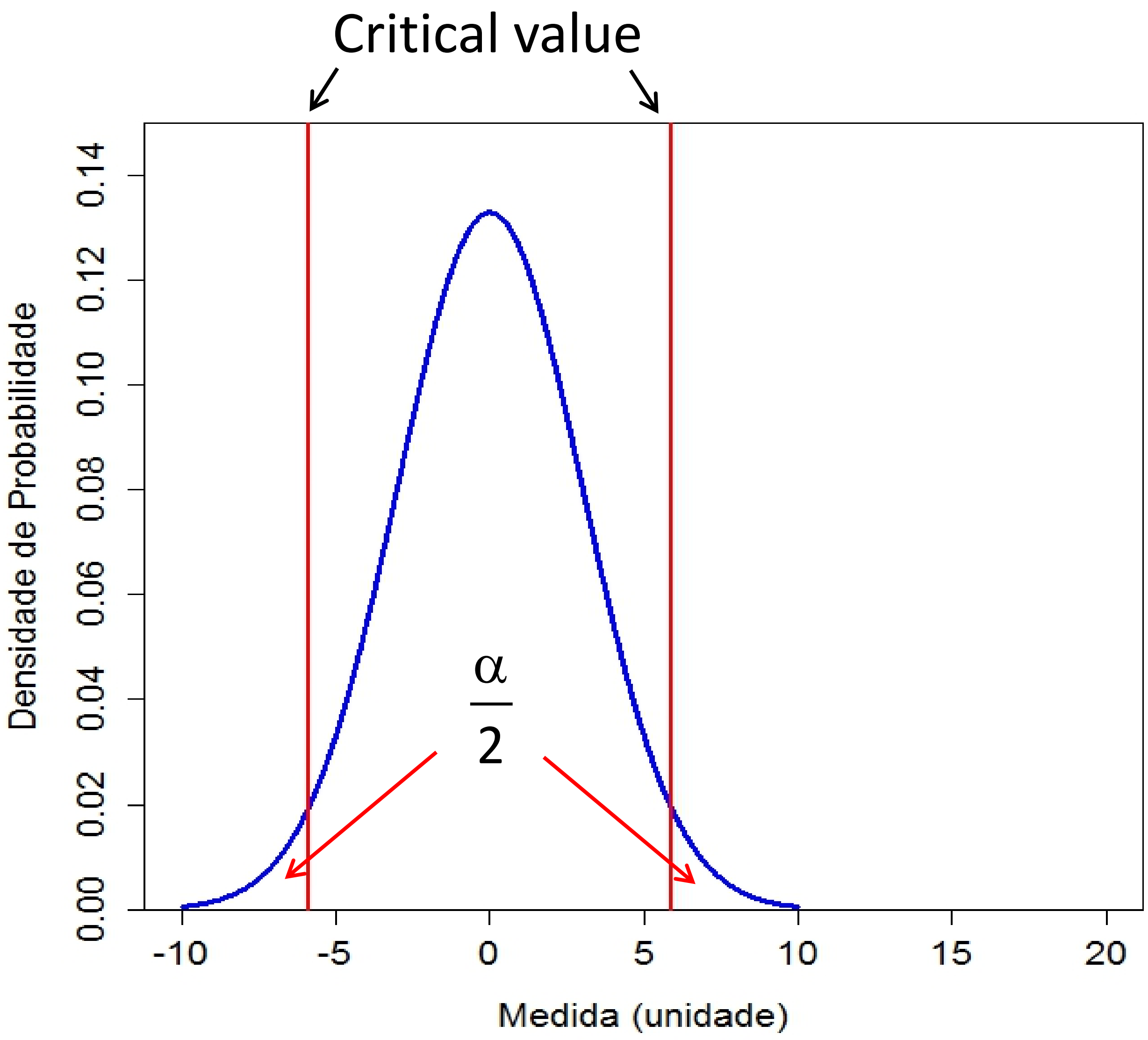
Controlling type II error (Power)

- Probability of rejecting the null hypothesis when it is not true (1 – p(type II error); or 1-β).

- If power is low, one is less likely to find a difference, even if it exists.

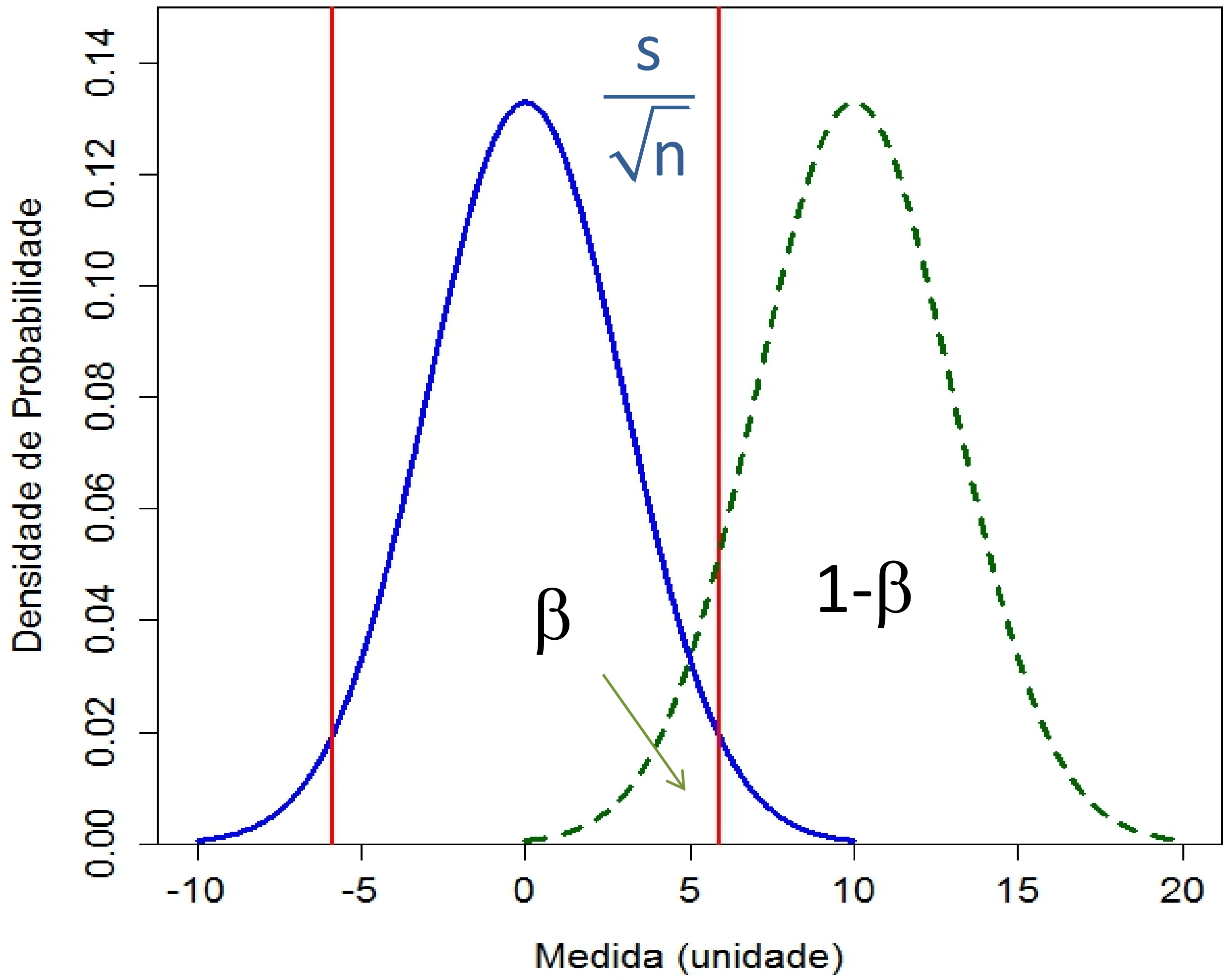
- Affected by:
  - Significance level: smaller => less power
  - Difference in means: smaller => less power
  - Standard deviation: smaller => more power
  - Sample size (n): smaller => less power

Type I Error



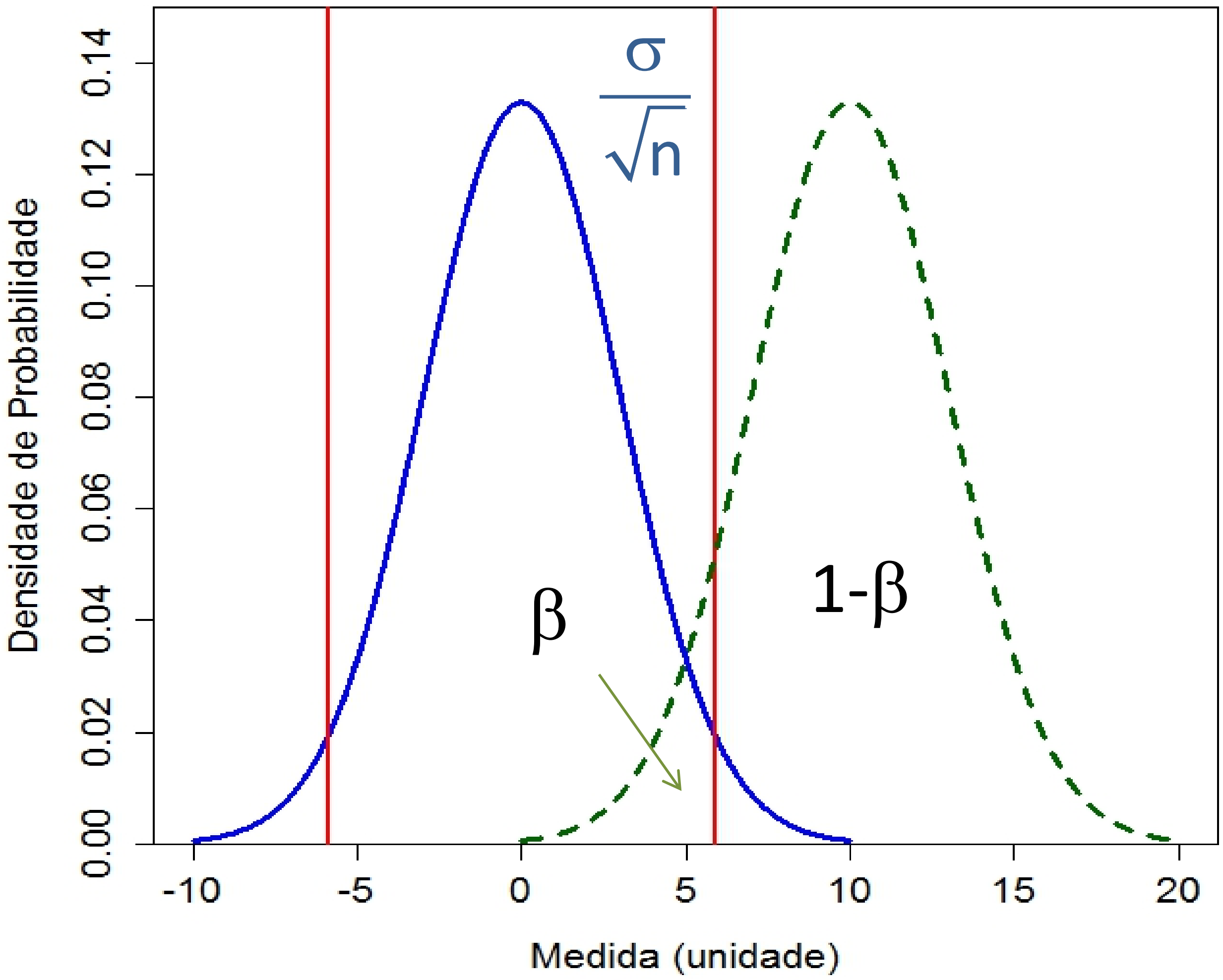
Critical value
$\frac{\alpha}{2}$
Densidade de Probabilidade
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
-10
-5
0
5
10
15
20
Medida (unidade)

Type II Error and Power of a Test



Type II Error and Power of a z-Test

Sample size: $n \approx \dfrac{\sigma^2 (z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$



Sample Size

Question:

What is the Sample Size needed for my study?

Answer:

What is the goal of the study?

What was the study design?

What is the variable that you want to study?

What variation is expected to find in that variable?

What type of analyses (statistical test)?

What other constraints: logistical, ethical, financial?

General Linear Models

- Basic: linear regression
  - But much more than traditional LR
- Foundation of generalized linear models
- Generalized linear models generalize(!) general linear models

General Linear Models

- Prediction
  - For new data
  - Accuracy of the model
- Understand / interpret
  - Analyze the relation between variables
  - Parsimony (as simple as possible)

(Recall) Concepts

- Dependent Variable (y)
- Independent (Explanatory) Variable (x)

- Variables
  - Quantitative (discrete, continuous) - scale
  - Qualitative (nominal, ordinal)
    - Factors / Levels

Linear model

$$\begin{cases} E(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \\ \mathrm{var}(y_i) = \sigma^2 \end{cases}$$

$$\begin{cases} y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2) \end{cases}$$

Normality is not strictly necessary to define the model, but for inference

Linear model

Model

$$\begin{cases} y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2) \end{cases}$$

Assumptions

- Independence
- Linearity
- Constant variance
- Normality

Normality is not strictly necessary to define the model, but for inference

Linear regression in R

- `lm(Gexp ~ Biom, data=genExp)`

- `res <- lm(Gexp ~ Biom, data=genExp)`

- The result is an object

```
> summary(res1)
Call:
lm(formula = Gexp ~ Biom, data = genExp)

Residuals:
    Min          1Q      Median       3Q      Max
-1.75167 -0.26619 -0.00401  0.24474  2.11936

Coefficients:
                     Est          Std. Err     t value   Pr(>|t|)
(Intercept) -5.432679   0.181460  -29.94   <2e-16 ***
Biom           0.131976   0.002955   44.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4307 on 652 degrees of freedom
Multiple R-squared:  0.7537,  Adjusted R-squared:  0.7533
F-statistic:  1995 on 1 and 652 DF,  p-value: < 2.2e-16
```

How are these calculated?

- Residuals
- Sum of squares

- Least squares

Inspecting the object

- coef
- resid
- confint
- predict
- anova

```
> anova(res1)
Analysis of Variance Table

Response: Gexp
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Biom	1	369.99	369.99	1994.7	< 2.2e-16 ***
Residuals	652	120.93	0.1855		
Total	653	490.92	0.7518		

From the summary

Multiple R-squared: 0.7537, Adjusted R-squared: 0.7533

F-statistic: 1995 on 1 and 652 DF, p-value: < 2.2e-16

Model evaluation and assumptions

- Independence
  - Experimental design
  - Blocks
  - Repeated measures
  - Time courses

Model evaluation and assumptions

- Normality
  - For inference: statistical tests and confidence intervals
  - Less important for larger sample sizes (CLT)
  - E.g., variable with only positive values, or proportions

Model evaluation and assumptions

- Linearity
  - Non linear relations and transformations
- Constant variance
  - Variability proportional to the mean
  - Variables with positive values, or proportions

How do we assess assumptions?

- Analyses of residuals
  - resid
  - rstandard
  - Plot rstandard against explanatory variables
  - Plot rstandard against fitted values
  - qqnorm plot
- Outliers and influential observations
  - Mistake?

Transforming the response variable

- Constraints on the possible values
  - No negative values for y: log y (or log(y+0.5))
  - Counts out of total: logistic transformation

- Normal distribution
  - log y
  - sqrt(y)

Transforming the response variable

- Stabilizing the variance
  - Remove mean-variance relationship

$$y^3 \quad y^2 \quad y \quad \sqrt{y} \quad \log(y) \quad y^{1/2} \quad \frac{1}{y} \quad (\ldots)$$

  - Proportions: $\arcsin(\sqrt{y})$

Transforming the covariates

- Examples:
  - Power law: $y = ax^b$

    $$\log(y) = \log(a) + b\log(x)$$

  - Polynomial transformation: x^2 or x^3

More complex models

- Interactions
- Marginality
- Comparing models

Generalizing the general linear model