ERROR PROPAGATION AND ERROR BARS

Bioinformatics Awareness Days @ TIGEM
July 10th, 2023

Eugenio Del Prete, M. Eng., Ph.D.
BIOINFORMATICS CORE
e.delprete@tigem.it

Bioinformatics Core: Tasks

- **STATISTICAL DATA ANALYSIS**
  Experimental Design, Hypothesis Testing, Power Analysis Differential Expression Analysis, Cluster Analysis, Time Series Data Analysis, Survival Analysis, Correlation Analysis
- **OMICS**
  Microarray Analysis, Gene Networks, Pathway Analysis, TFBS Identification, Gene Annotation, Integration, Protein Analysis, Drug Networks
- **NEXT GENERATION SEQUENCING**
  Whole Exome, Targeted Gene, RNA, miRNA, ChIP, Visualization, Interpretation
- **DATABASE AND SOFTWARE**
  DB Creation, DB Maintenance, Web Sites Creation, Web Service Support
- **BIOINFORMATICS AND (BIO)STATISTICS TRAINING**

Bioinformatics Core: People

DIEGO DI BERNARDO

https://www.tigem.it/research/facilities/core-facilities/bioinformatics

https://bioinformatics.tigem.it/

DIEGO CARRELLA

ROSSELLA DE CEGLI

XAVIER BUJANDA CUNDIN

EUGENIO DEL PRETE

Bioinformatics Core: Something about Me

- **TLC ENGINEER** @ UNIVERSITY OF ROME 'SAPIENZA'
  MAIN TOPICS: Signal Processing, Remote Sensing, Bioinformatics
  THESIS: miRNA Analysis, Genomic Data Mining, Consensus Analysis, PSSM Creation
- **BIOINFORMATICS RESEARCH FELLOW** @ INSTITUTE OF FOOD SCIENCES (CNR)
  Protein Prediction and Classification, Protein Analysis, Proteomic Mass Spectra Analysis, Sequence Alignment and Phylogenetic Tree, Docking
- **PHD IN APPLIED BIOLOGY** @ UNIVERSITY OF BASILICATA
  Celiac Disease and Comorbities, Microarray Data Analysis, Ontologies, Gene Set Enrichment Analysis, Semantic Similarity, Proteomic Mass Spectra Analysis
- **BIOINFORMATICS RESEARCH FELLOW** @ INSTITUTE OF APPLIED MATHEMATICS (CNR)
  Proteomic Mass Spectra Analysis, Metabolomic (Lipidomic) Data Analysis, Web Tools Developer, Hypothesis Tests, Omics Data Integration
- **BIOSTATISTICIAN AND DATA SCIENTIST** @ TIGEM

Outline

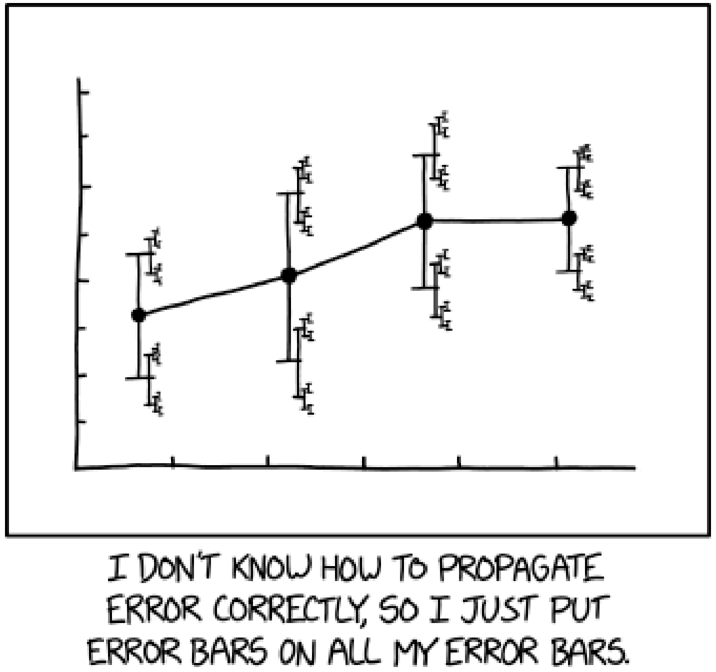
ERROR TYPES
- Playing around Error Bars
- Measurement Error
- Absolute Error and Relative Error

ERROR PROPAGATION
- Formula
- Operations
- Precision on Significant Figures

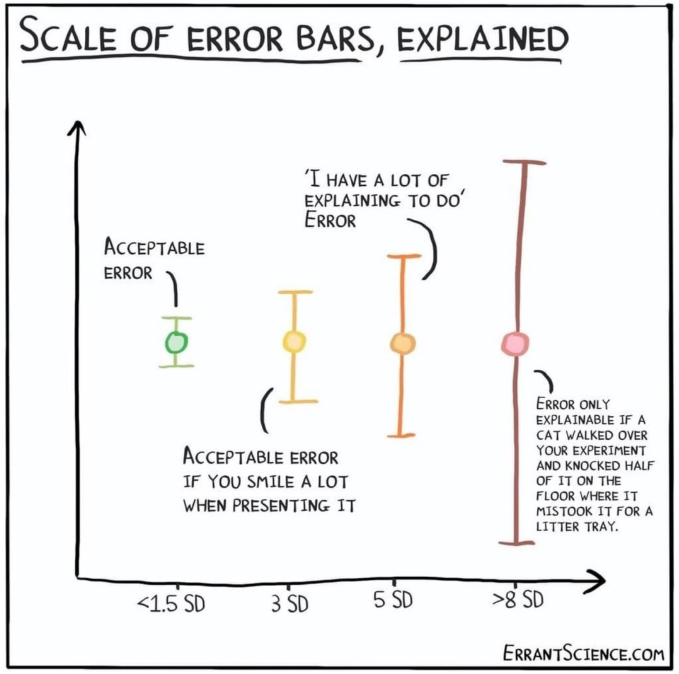
ERROR BARS
- Definition
- Practical rules
- Example: Error Bars with Prism

CONCLUSION
- Take Home Message
- Final Remarks

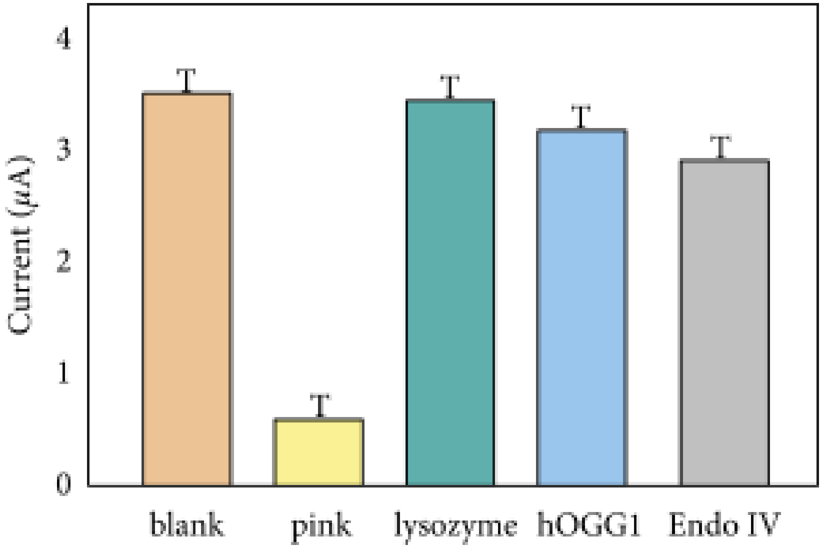
How (not) to cope with...



In case of panic...



Sell your product...



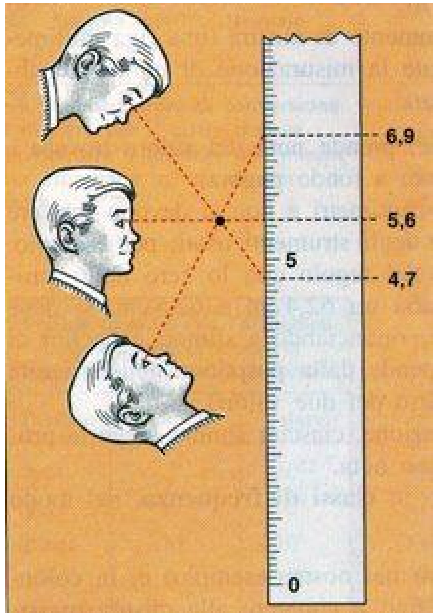
Multiple t-test...

Measurement Error

- Occurs when tools or instruments are used or read **incorrectly**

- Systematic Error
  - **Always** present in the measurement (but **resettable**)
  - Due to instrument **calibration** or **construction**
  - Due to the **same** misuse of the instrument
  - Causes the **bias** of the measurement

- Random Error
  - **Not always** present in the measurement (but **non-resettable**)
  - Due to the **conditions** of the measurement
  - Due to the **conditions** of the researcher
  - Causes the **bad estimate** of the measurement



PARALLAX ERROR

Random Error

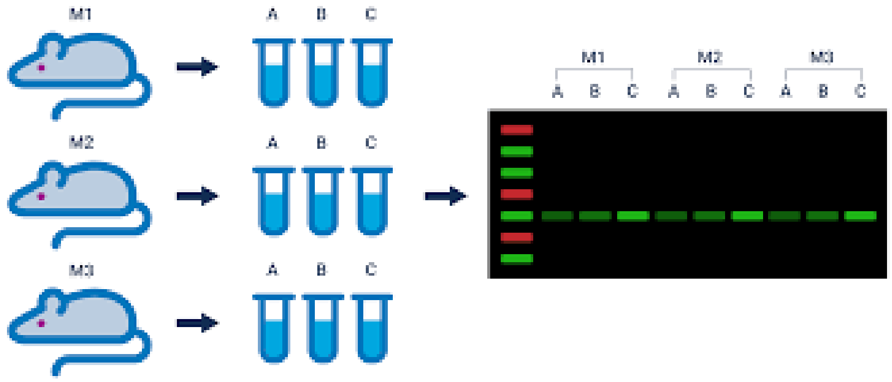
Biological sources of Random Error

- Variation in measurement readings
- Too small sample size
- Background (unpredictable) noise
- Biological intrinsic variability
- Instrument sensitivity limits
- Batch effects (time, temperature, researcher, contamination, …)

QUESTION:
Background noise in mass spectra is a random error? Is it bad?

Some solutions for Random Error

- Biological samples and technical replicates
- Keep same conditions for experiments
- Control the degrees of precision:
  - **Do not kill a fly with a sledgehammer**
  - **Comparable measurement and error**



Absolute Error and Relative Error

- A measurement can be expressed as

$$x = x_m \pm \Delta x$$

- **Absolute Error Δx**
  - **Uncertainty** of the measure
  - Acceptable **range** for the real value of the measurement (x_m)
  - Caused by all the typologies of measurement error
  - **Same unit of measurement** of the measurement

- **Relative Error $\Delta x / x_m$ (Precision)**
  - **Uncertainty** of the measure
  - Define the **quality** of the error of measurement
  - Caused by all the typologies of measurement error
  - **Adimensional**, usually reported as **percentage**

QUESTION:
An absolute error of 1 mm is always a small error?

Error Propagation: Formula

- Suppose to have several **instruments** with different **variabilities**

$$a = a_m \pm \Delta a, \qquad b = b_m \pm \Delta b, \qquad c = c_m \pm \Delta c$$

and to calculate a quantity x dependent from (a, b, c) such as

$$x = f(a, b, c)$$

- Quantity x will have its uncertainty **dependent** from the uncertainties of each measurement from the different instruments

$$\Delta x_i = f(\Delta a_i, \Delta b_i, \Delta c_i) \rightarrow dx_i = f(da_i, db_i, dc_i)$$

(considering legit a ‘movement’ **from uncertainties to derivatives**)

Error Propagation: Formula

- **Operations** (without details):
  - apply the **partial derivatives** to each instrument variability
  - verify the **independency of errors (between-within)**
  - consider the **total number of measurements**

$$\sigma_x^2 = \left(\frac{\delta x}{\delta a}\right)^2 \sigma_a^2 + \left(\frac{\delta x}{\delta b}\right)^2 \sigma_b^2 + \left(\frac{\delta x}{\delta c}\right)^2 \sigma_c^2 \quad \rightarrow \quad \sigma_x = \sqrt{\left(\frac{\delta x}{\delta a}\right)^2 \sigma_a^2 + \left(\frac{\delta x}{\delta b}\right)^2 \sigma_b^2 + \left(\frac{\delta x}{\delta c}\right)^2 \sigma_c^2}$$

- **Considerations**
  - **standard deviation as error** ($\Delta x_i \rightarrow \sigma_{x_i}$)
  - dependency from errors, **not from cross-errors**
  - application to **all the math operations**

Error Propagation: Operations

- Addition and Subtraction

$$x = a + b - c \quad \rightarrow \quad \sigma_x = \sqrt{\sigma_a^2 + \sigma_b^2 + \sigma_c^2}$$

- Multiplication and Division

$$x = \frac{ab}{c} \quad \rightarrow \quad \frac{\sigma_x}{x} = \sqrt{\left(\frac{\sigma_a}{a}\right)^2 + \left(\frac{\sigma_b}{b}\right)^2 + \left(\frac{\sigma_c}{c}\right)^2}$$

- Power

$$x = a^k \quad \rightarrow \quad \frac{\sigma_x}{x} = |k|\frac{\sigma_a}{a}$$

- Constant

$$x = ka \quad \rightarrow \quad \sigma_x = |k|\sigma_a$$

QUESTION:
Error propagation with reciprocal quantity?

QUESTION:
Error propagation with addition and constants?

Precision on significant figures

- Significant figures in a number are accurate **except for the final digit**
- Numbers are often the **result of averages** obtained from multiple experiments
- **False precision** (**overprecision**) occurs when numerical data are presented in a manner that implies better precision than is justified

	Average	SD
Experimental numbers	7.31732	0.382521
Significant figures with 1-digit uncertainty	7.3	0.4
False significant figures with 2-digit uncertainty	7.32	0.38
False significant figures with 3-digit uncertainty	7.317	0.383

RESOLUTION: 0.1

- Significant figures **change the error propagation**

Precision on significant figures

Measurement uncertainty

$$x = x_m \pm \sigma_x$$

Measurements as MEAN VALUE
Uncertainty as STANDARD DEVIATION

Best Practice

- uncertainty cannot be more precise than the best estimate of the measured value
- uncertainty determines the number of significant figures in the real measurements
- rounding should always be performed at the end of a series of calculations

Examples

a) 87.25 u.m. + 3.0201 u.m.
b) 26.843 u.m. + 12.23 u.m.
c) (15.9994 × 9) u.m. + 2.0158 u.m.

Results

a) 90.27 u.m.
b) 39.07 u.m.
c) 143.9946 u.m. + 2.0158 u.m. = 146.0104 u.m.

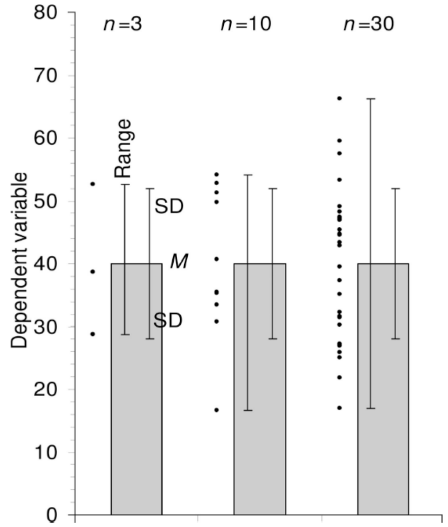
Error Bars: Definition

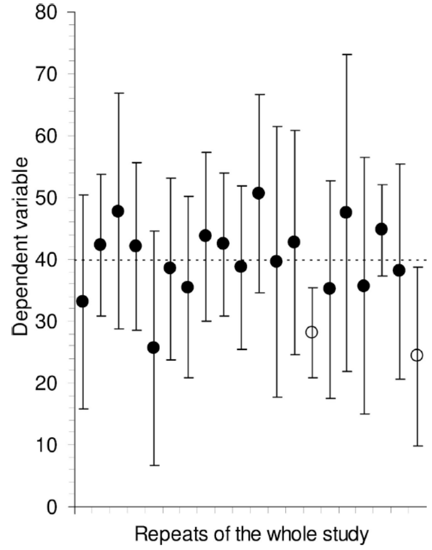
- Provide information to **describe** data or to **infer** conclusions

- ## Descriptive error bars
  - show how data are spread
  - see whether a single results fits within the normal range

- ## Inferential error bars
  - show a range where you can expect to find the mean
  - compare samples between groups

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 $M \pm 2 \times SE$.





Error Bars: Practical Rules

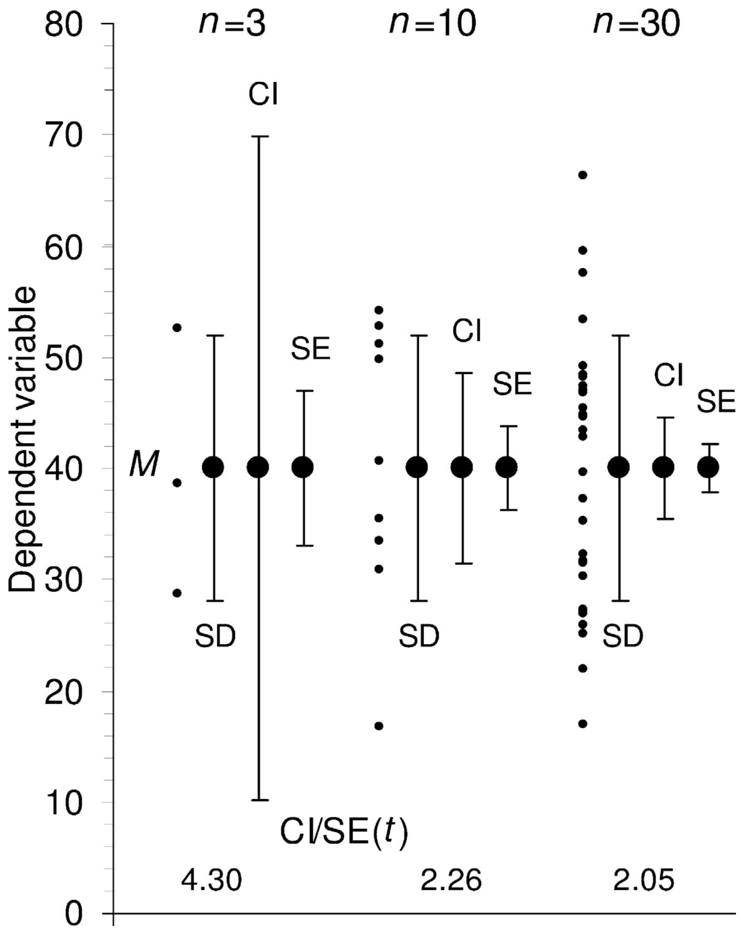
- Rule 1: When showing error bars, always describe in the figure what they are
- Rule 2: The sample size must be stated in the figure
- Number of independent results **is different** from number of (technical) replicates
- Rule 3: Error bars and statistics should only be shown for independently repeated experiments, and never for replicates
- Rule 4: It is appropriate to show inferential error rather than descriptive error

QUESTION:
Suppose to have 20 measurements from one KO mouse and one WT mouse, to determine if a gene affects the tail length. Can I answer the question? Why?

⇦ For small sample size (n = 3), depicting error bars is misleading

Error Bars: Practical Rules

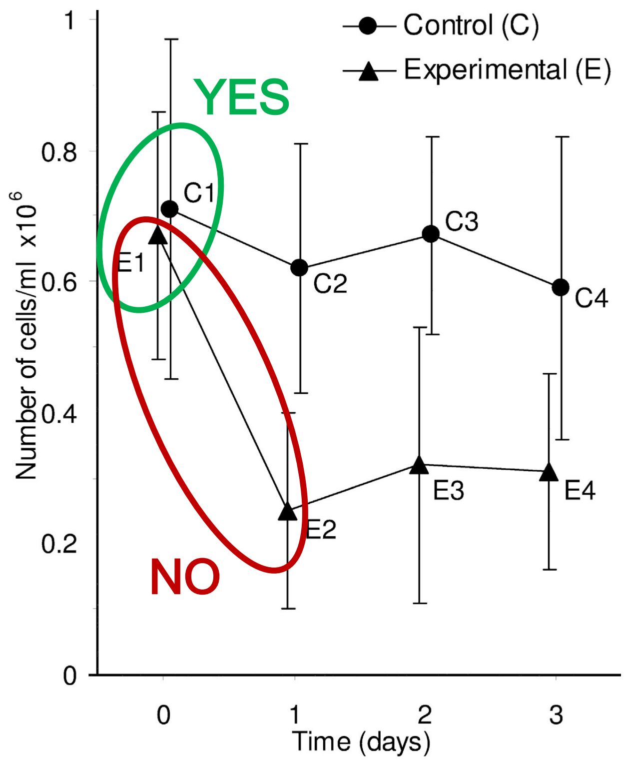
- **Standard Error (SE)**
  - an increment of the sample size **reduce** the SE
  - a reduction of the SE **improves** the estimate of true mean
- **Confidence Interval (CI)**
  - more complicate to calculate (but not nowadays)
  - interpretation **independent** from the sample size (but not the formula)
- Rule 5: **95% CI capture the true mean on 95% of occasions. In order to ‘mimic’ the 95%CI, SE bars can be 2 times for n ≥ 10**



Error Bars: Practical Rules

- Other visual considerations are available in order to compare 95% CIs between different conditions and ‘predict’ the statistical significance (i.e., **p-value**)
- Suppose to have **repeated measurements**, i.e., number of cells in three independent clonal experimental cell cultures (E) and three independent clonal control cell cultures (C) was measured over time
- Rule 6: **In the case of repeated measurements on the same group, CI or SE bars are irrelevant to comparisons within the same group**

These solutions are fast but dangerous (I do not suggest them!)



Example: Error bars with Prism 9.4.0 (GraphPad)

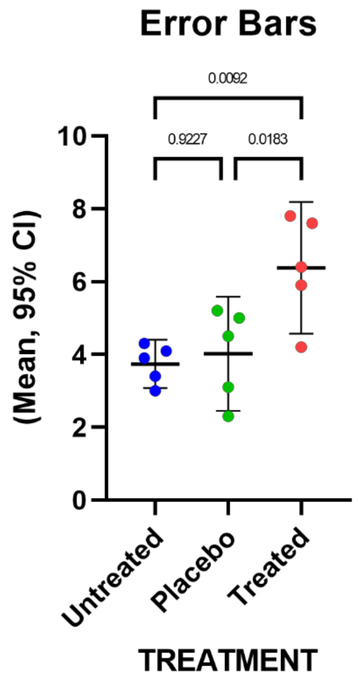
- **Upload dataset (already in Prism)**
  - control, placebo and treatment (3 conditions)
  - 5 biological replicates per condition (15 samples)
- **General statistics**
  - Add column sum (descriptive)
  - Add CI for mean (inferential)
- **Statistical significance**
  - Check the normality of the samples
  - Perform One Way ANOVA (parametric)
  - Correct for multiple comparisons
  - Control the homoscedasticity

1. Column → Start with sample data to follow tutorial → Column → Error bars in column tables → Entering replicate data → Create
2. Rename Data Tables
3. Analysis → Analyze → Column Analysis → Descriptive statistics → Basics & Confidence Interval
4. Analysis → Analyze → Column Analysis → Normality and Lognormality test → Which distribution to test? → Normal (Gaussian) distribution → Method to test distributions → Shapiro-Wilk normality test
5. Analysis → Analyze → Column Analysis → One-way ANOVA → Multiple Comparisons → Followup test → …every other column → Residuals → Homoscedasticity plot → Diagnostic for residuals → Are residuals clustered or heteroscedastic?

Example: Error bars with Prism 9.4.0 (GraphPad)

Error bars

- Select the suitable plot
- Add the error bars
- Add the statistical significance
- Report all the p-values and methods

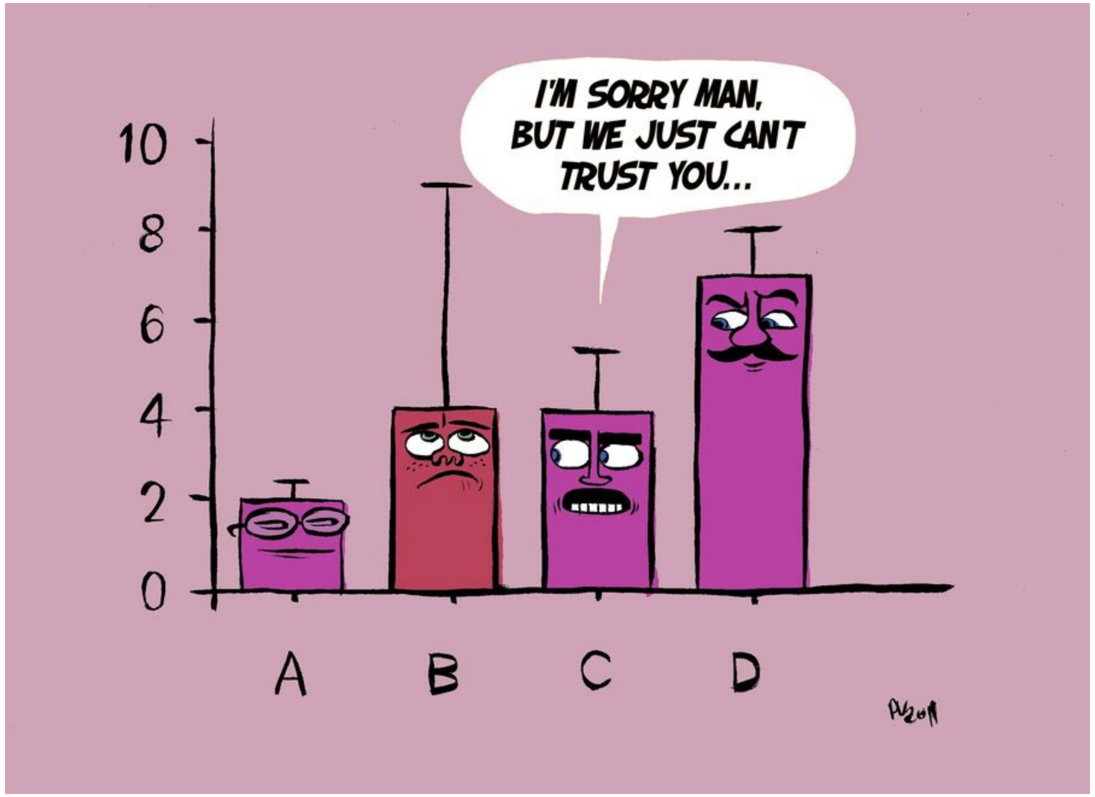


11. Graphs → Individual values → Scatter plot → Mean with 95% CI
12. Define title and labels
13. Double click on point and error bar to change
14. Draw → Asterix → Format Pairwise Comparisons → Display options → P value → Line/bracket… → Second Plot
15. Info → Project Info

Take Home Message

- Errors are (nearly) always present, but this **does not mean** they are always an issue
- Error propagation has **a differential formula** from which it is possible to extract each case
- Sample size is the most important value for an experiment, for the **strength** and the **reproducibility** of the results
- Select with accuracy the **type of error bar** and describe the selected type of error bar in the figure caption (or elsewhere)
- **Not all the types of graph** (scatter plot, barplot, boxplot, …) are suitable to depict the same experiment

Final Remarks



Eugenio Del Prete, M.Eng., Ph.D.
Biostatistician and Data Analyst
Telethon Institute of Genetics and Medicine (TIGEM)
Pozzuoli (NA), Italy
e-mail: e.delprete@tigem.it

References

[1] Cumming, G. **Error bars in experimental biology.** JCB (2007)
[2] Habibzadeb, F. **How much precision in reporting statistics is enough?** Croat Med J (2015)
[3] Morral, J. **Significant Figures and False Precision.** J. Phase Equilib Diffus (2018).
[4] Brown, A. W. **Issues with data and analyses: Errors, underlying themes, and potential solutions.** Proc Natl Acad Sci (2018)

[h1]*https://www.epfl.ch/labs/lben/wp-content/uploads/2018/07/Error-Propagation_2013.pdf*
[h2]*https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Quantifying_Nature/Significant_Digits/Propagation_of_Error*