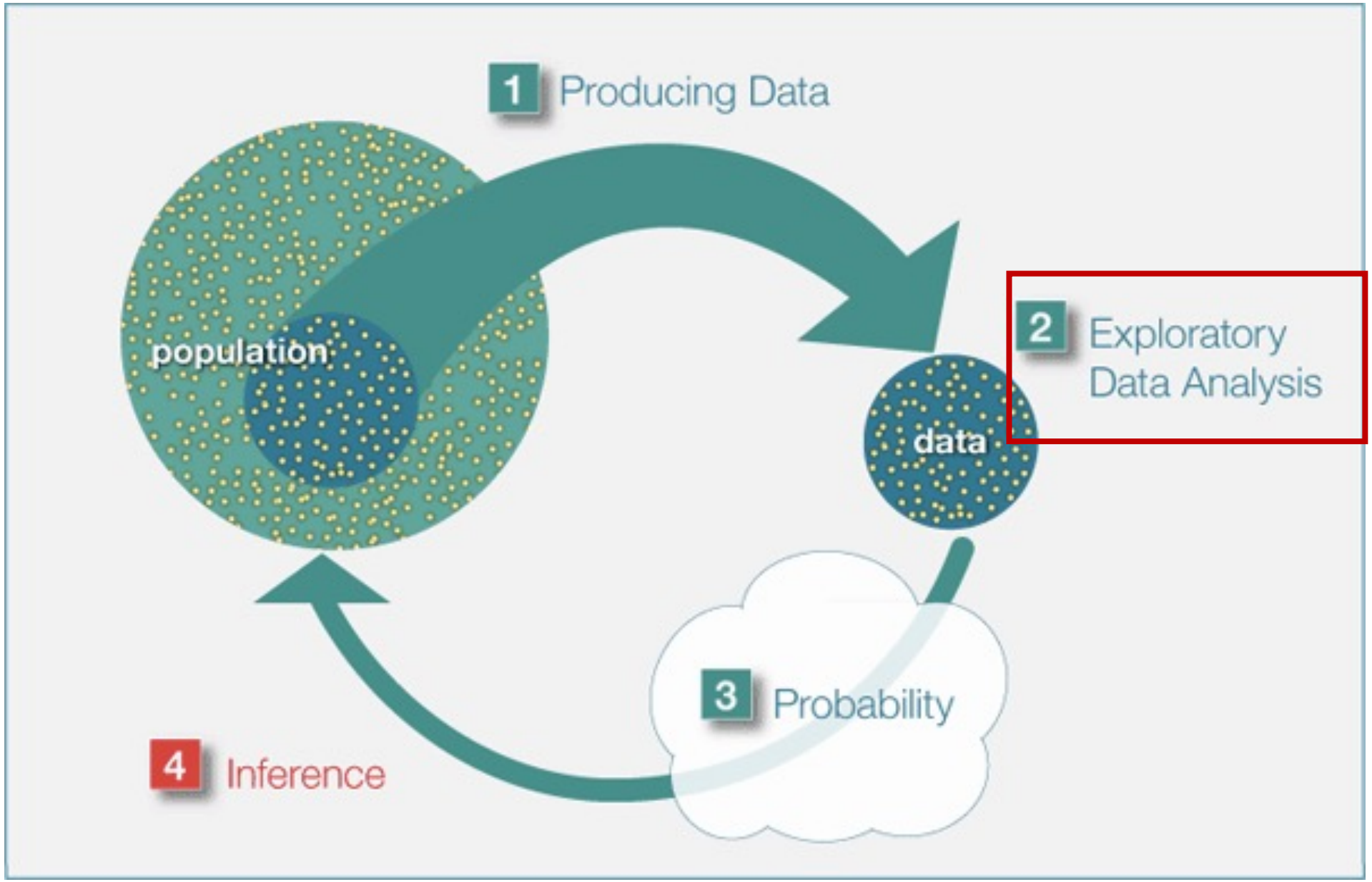
Special Topics in Biostatistics and Bioinformatics
Week II

Ege Ülgen, M.D.

10 March 2022

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ



Unit 4a: introduction to statistical inference [Internet]. [cited 2021 Sep 27]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-4/

Reasons to explore data

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To debug analyses
- To communicate results

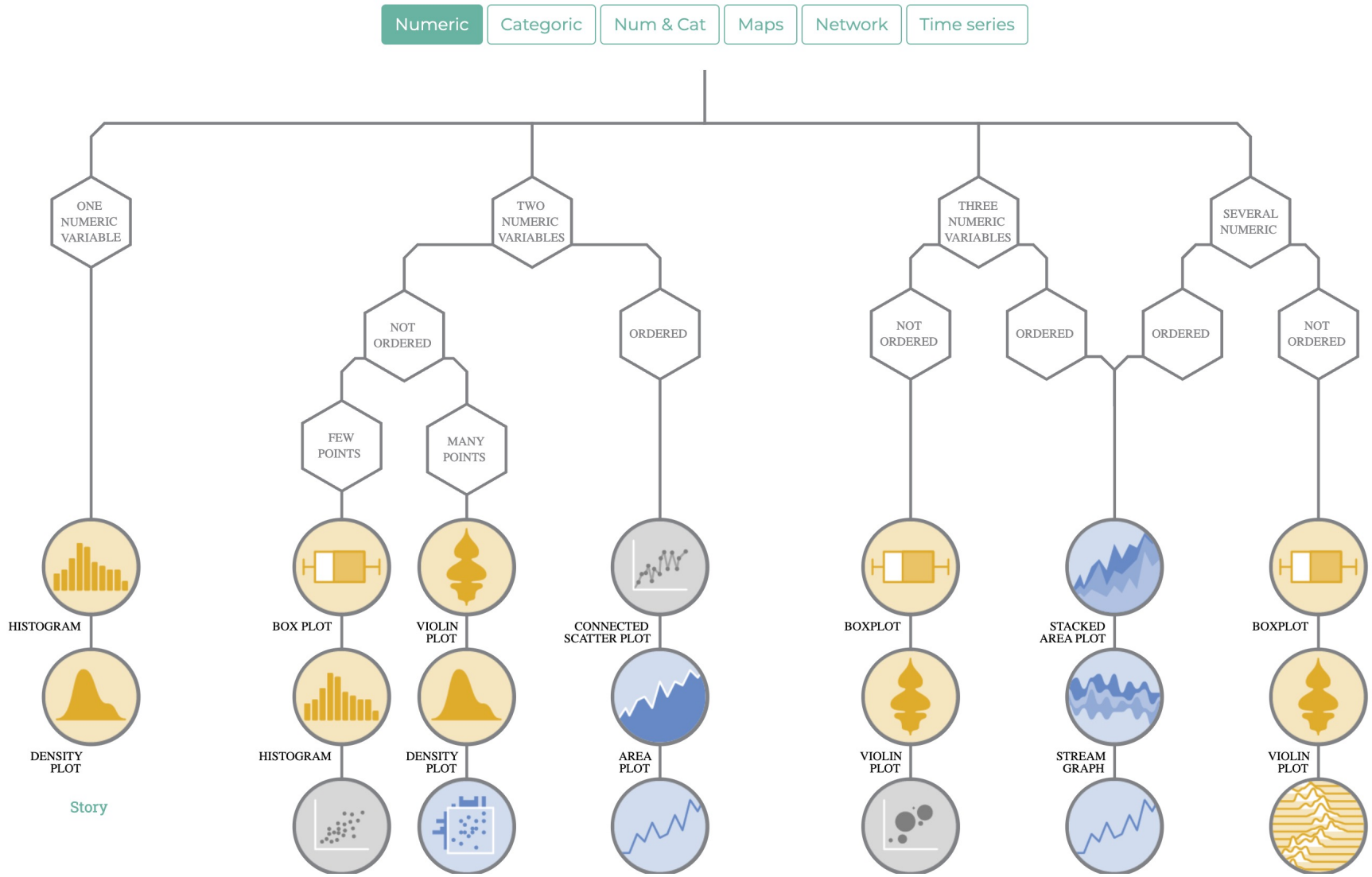
Exploratory Data Analysis

- Visualization
- Summarization
- Showing the data

**without being misled*

Examples of Visualization

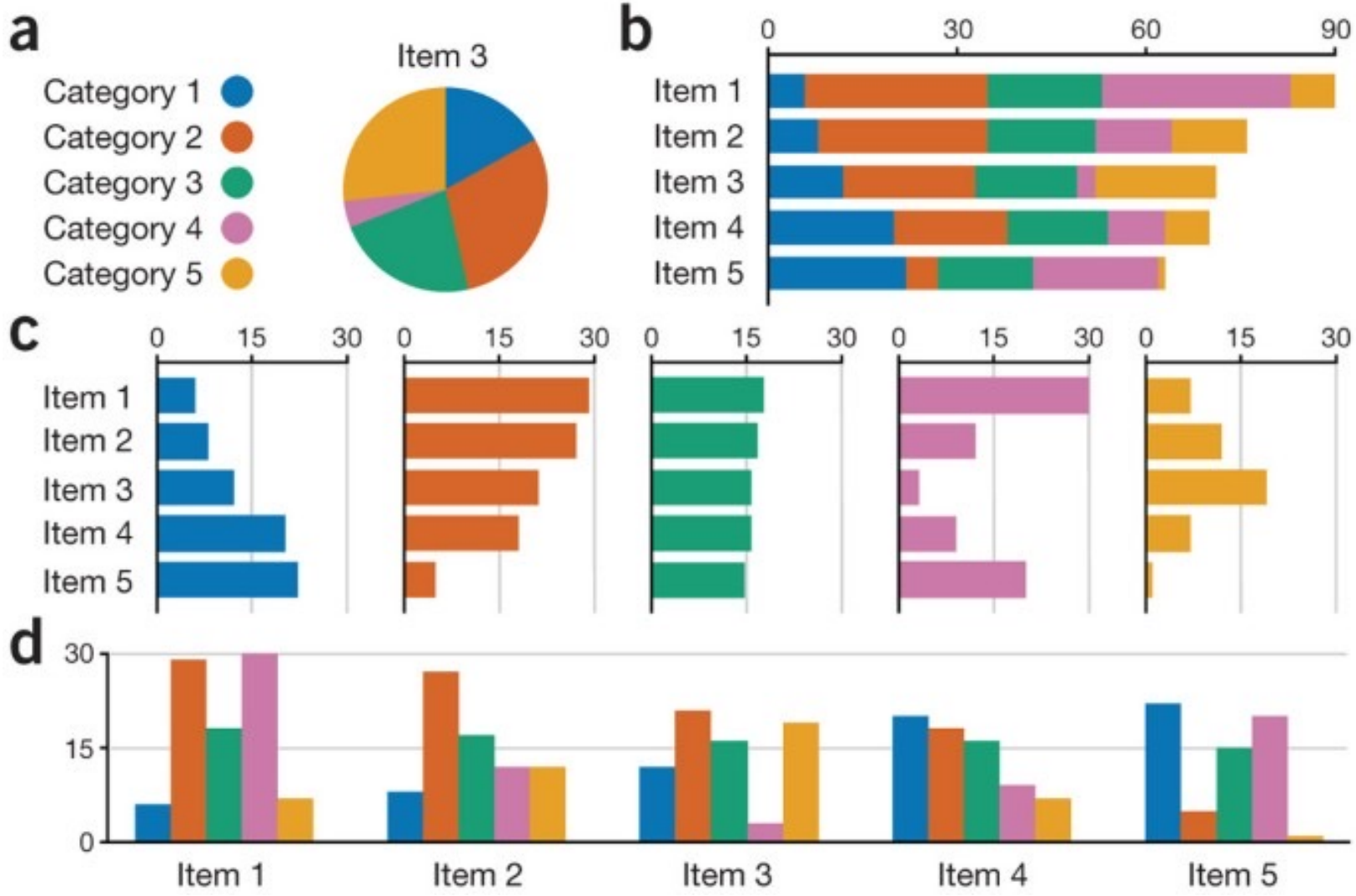
from Data to Viz



https://www.data-to-viz.com/

Streit M, Gehlenborg N. Bar charts and box plots. Nature Methods. 2014 Feb 1;11(2):117–117.

Bar Charts



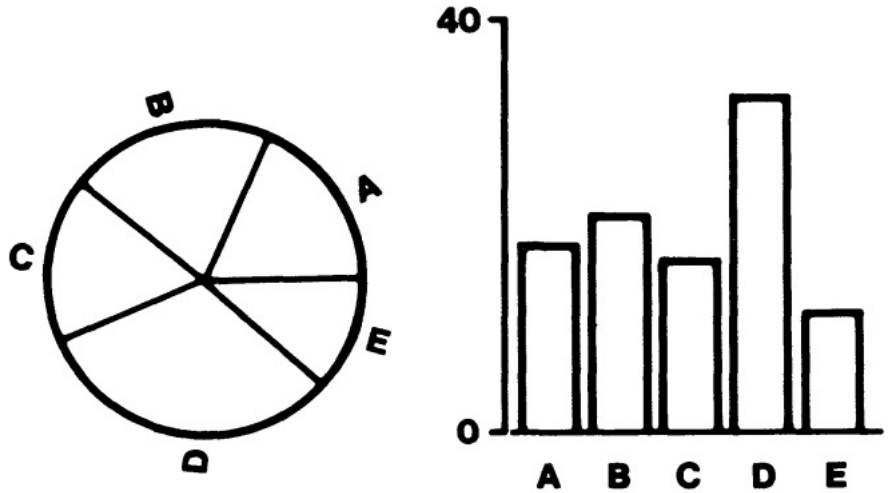


Figure 3. Graphs from position–angle experiment.

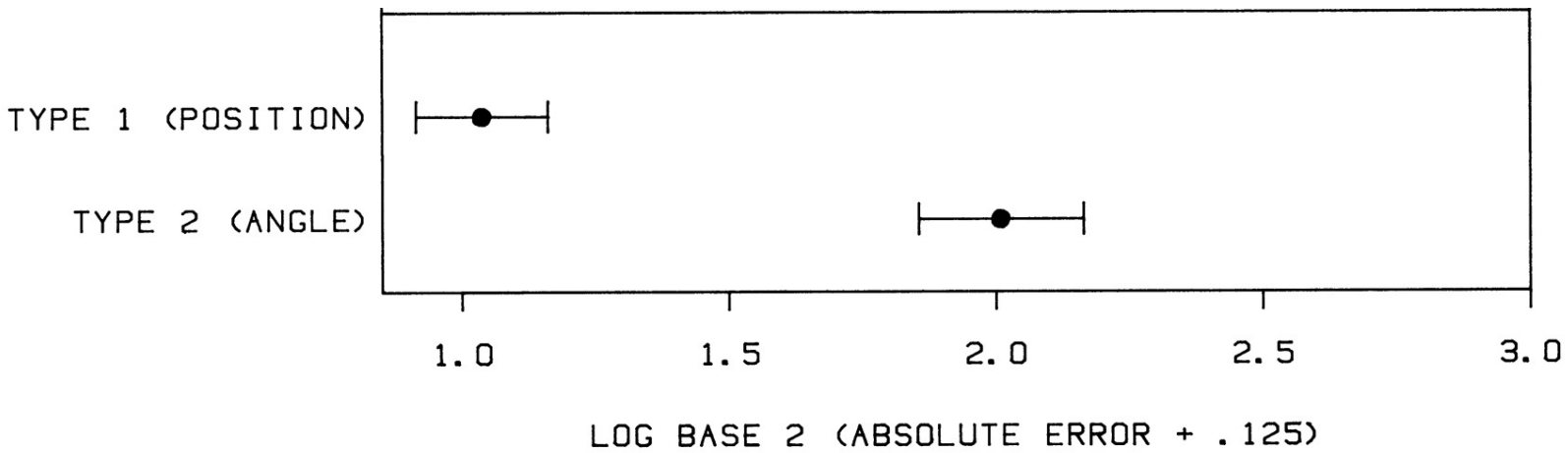
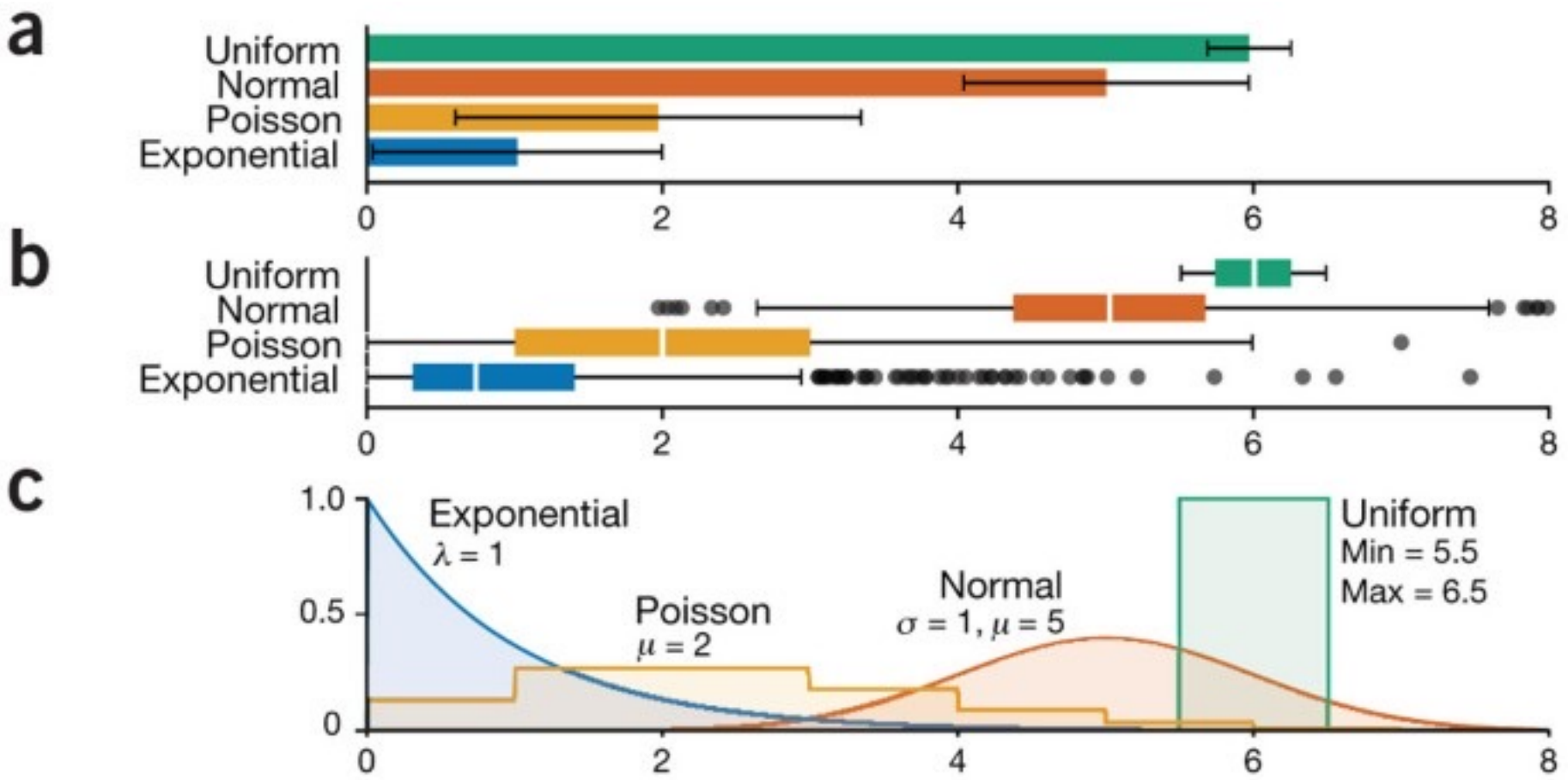


Figure 16. Log absolute error means and 95% confidence intervals for judgment types in position–length experiment (top) and position–angle experiment (bottom).

Streit M, Gehlenborg N. Bar charts and box plots. Nature Methods. 2014 Feb 1;11(2):117–117.

Box plots



Hintze J, Nelson R. Violin plots : A box plot-density trace synergism. 1998

Violin Plots

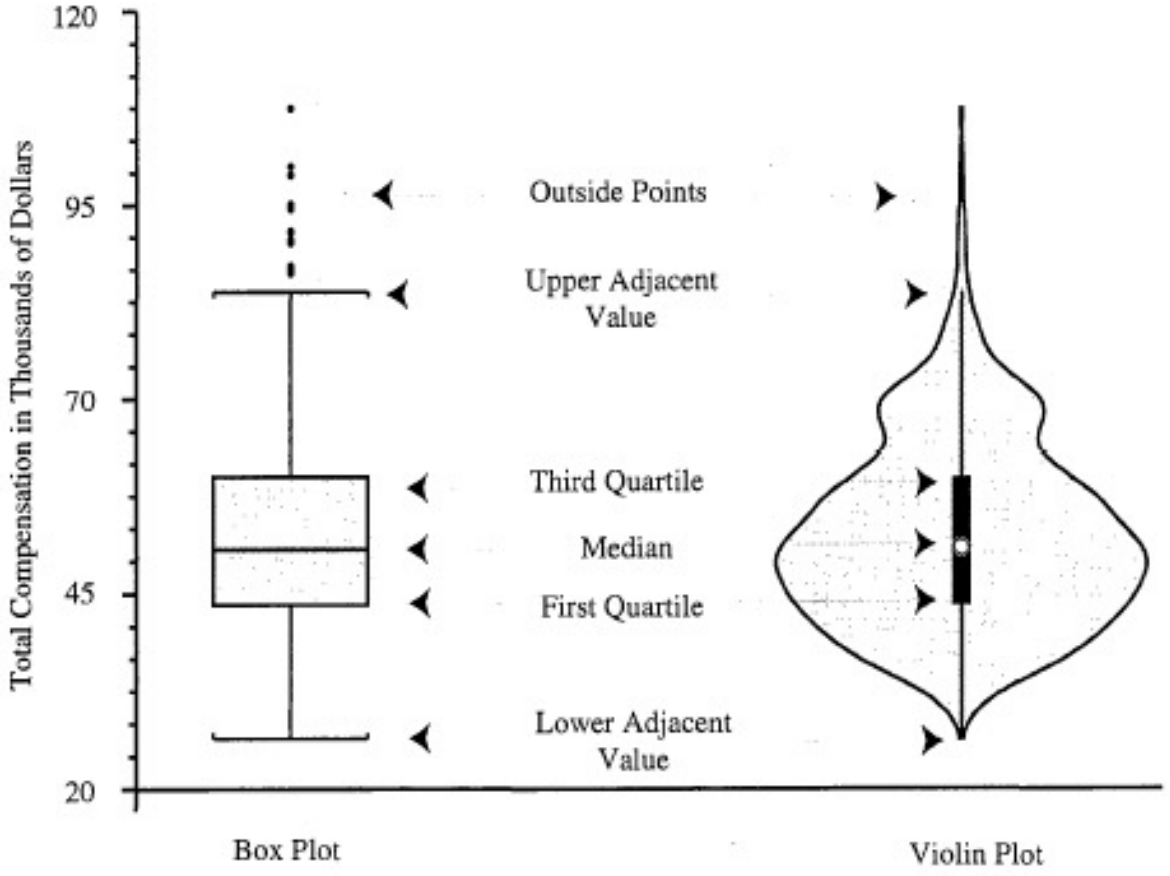


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

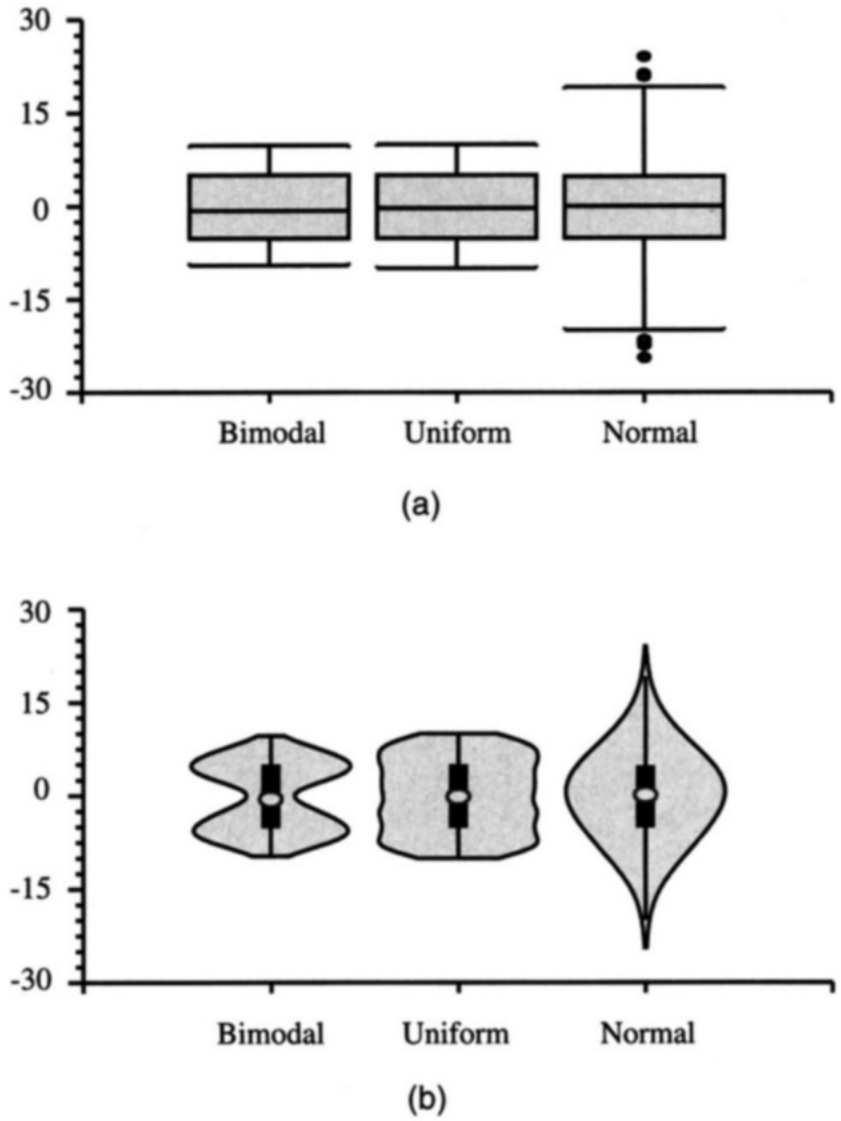
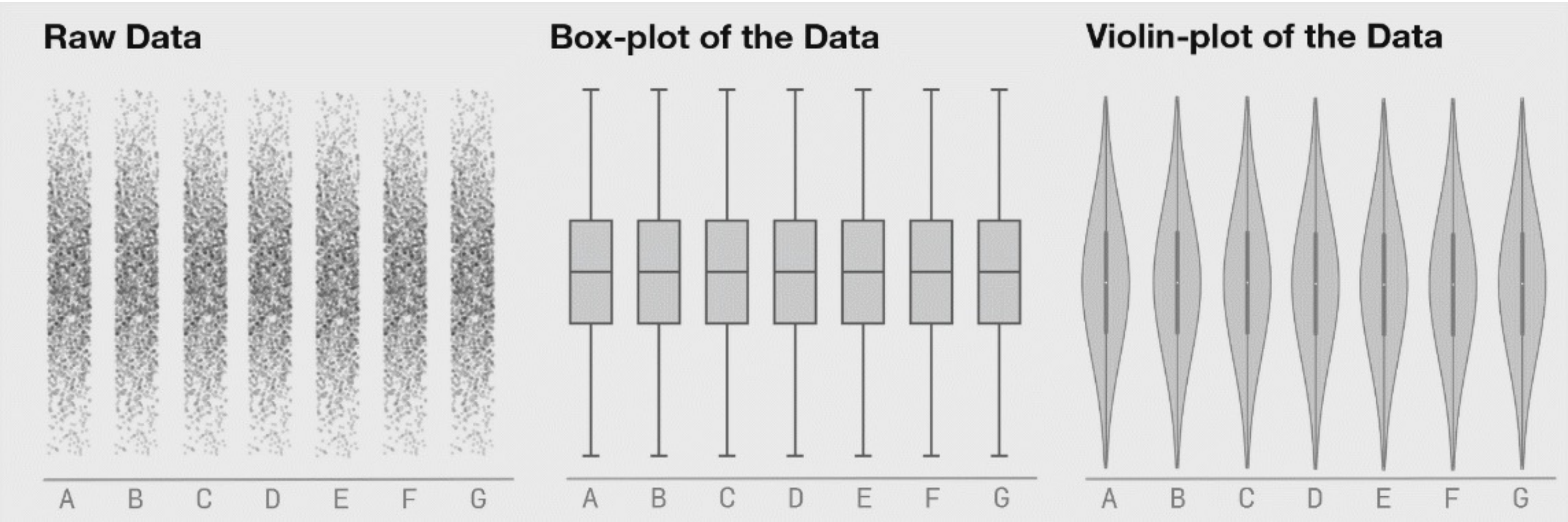


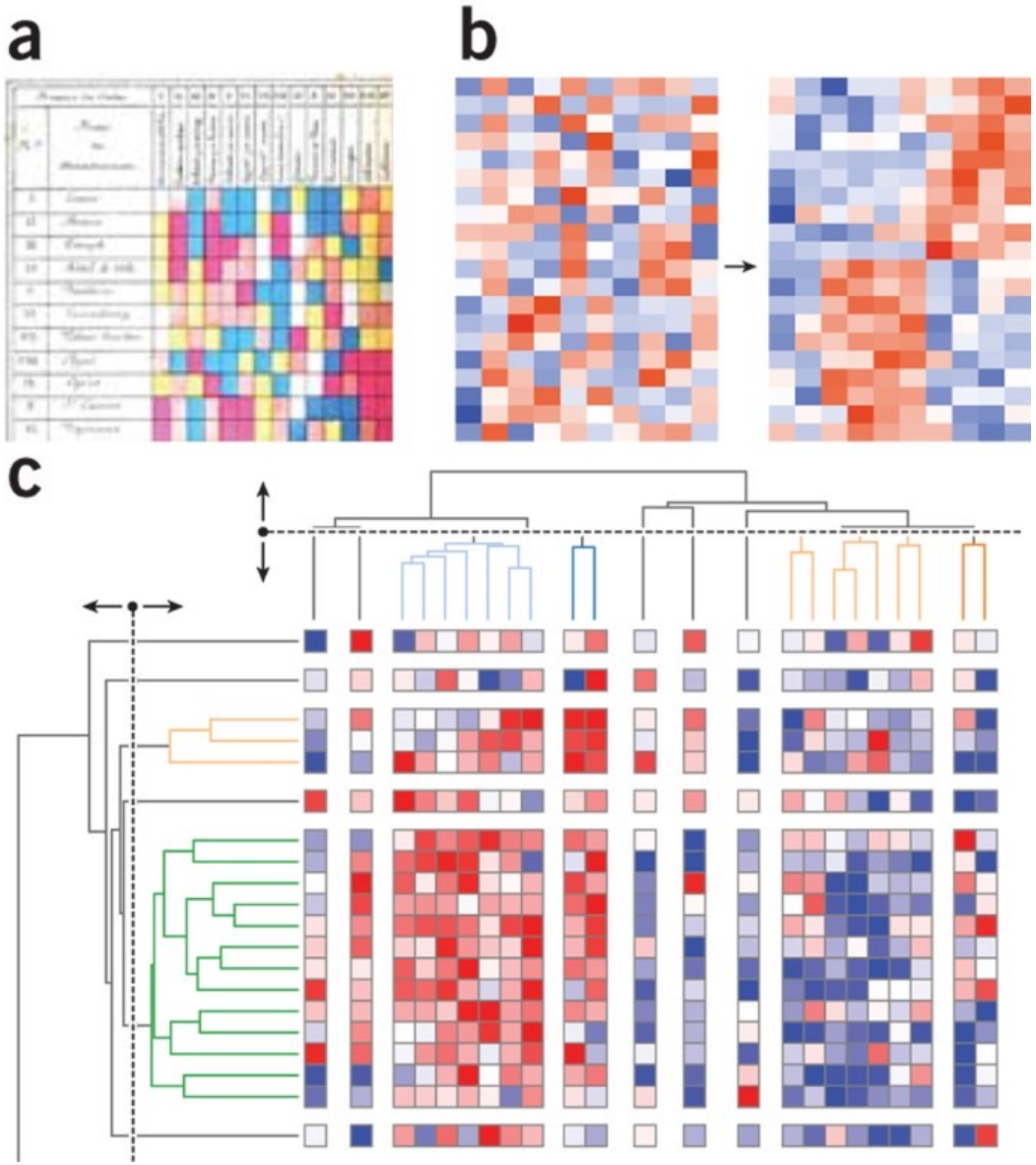
Figure 2. Comparison of Box Plots and Violin Plots fo Known Distributions. (a) Box plots; (b) violin plots.

https://www.autodeskresearch.com/publications/samestats



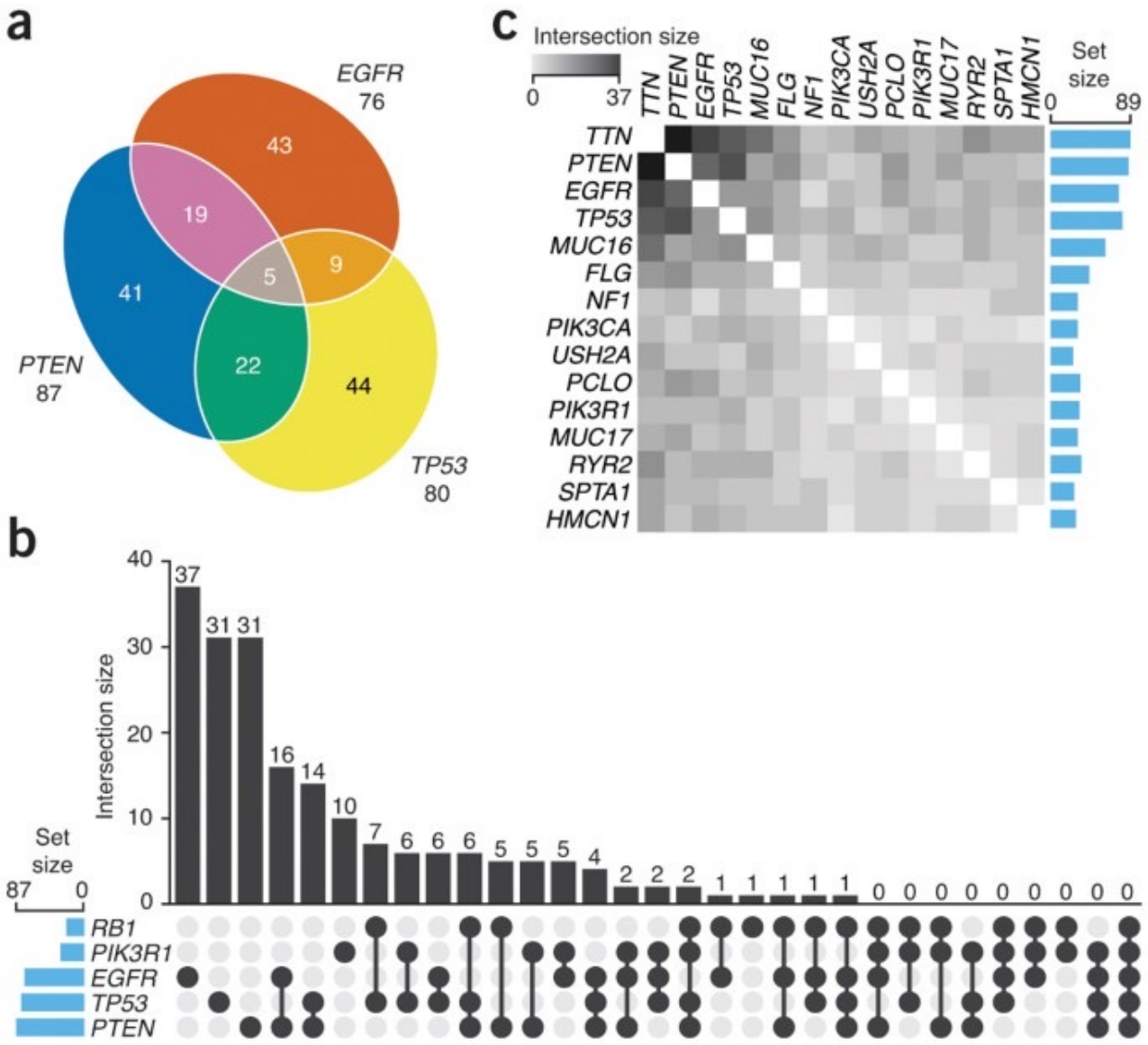
Gehlenborg N, Wong B. Heat maps. Nature Methods. 2012 Mar 1;9(3):213–213.

Heat maps



Lex A, Gehlenborg N. Sets and intersections. Nature Methods. 2014 Aug 1;11(8):779–779.

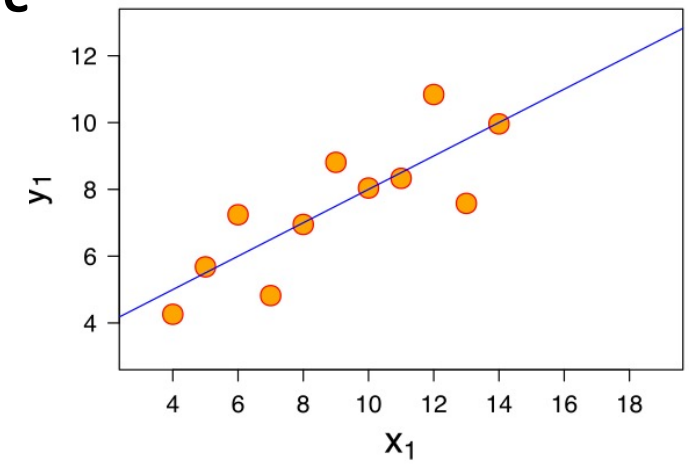
Sets and intersections

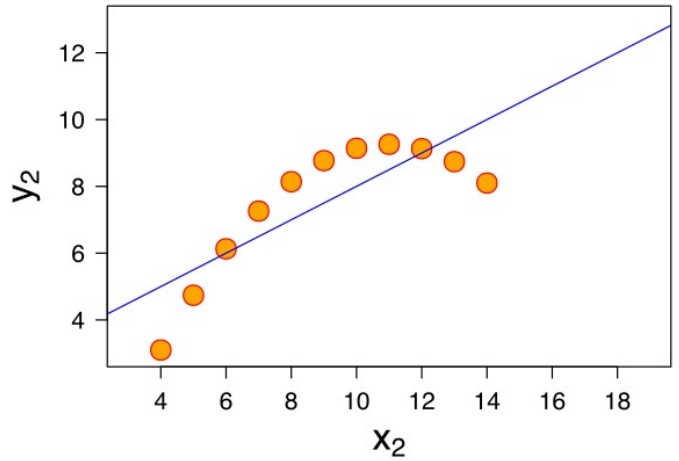


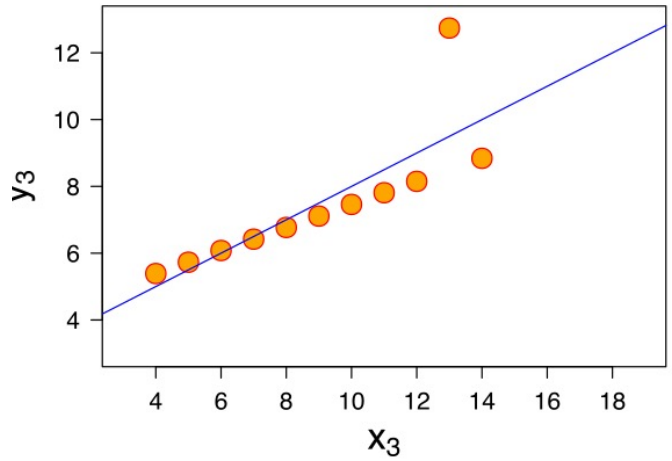
Avoid being misled

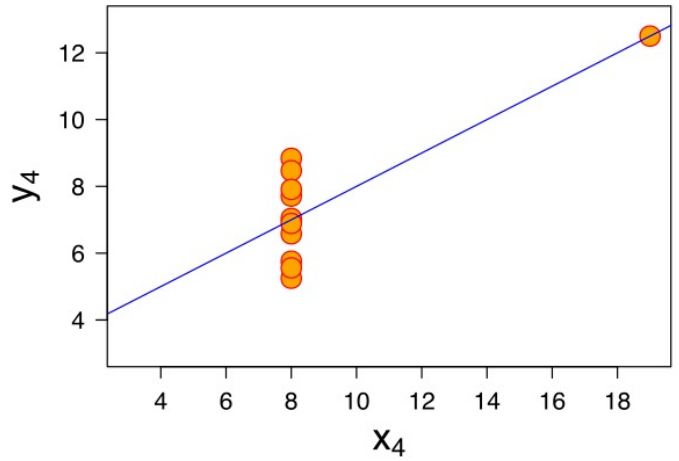
Anscombe's quartet

Property	Value
Mean of x	9
Sample variance of x : s_x^2	11
Mean of y	7.50
Sample variance of y : s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67



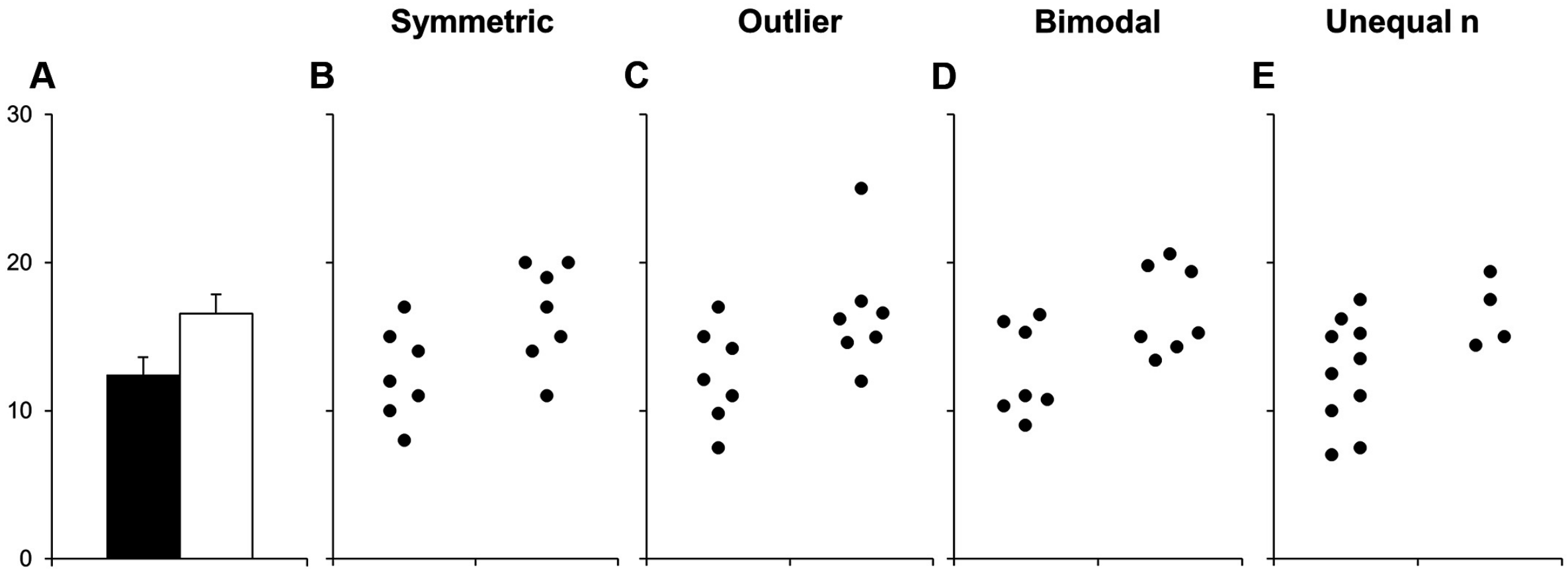






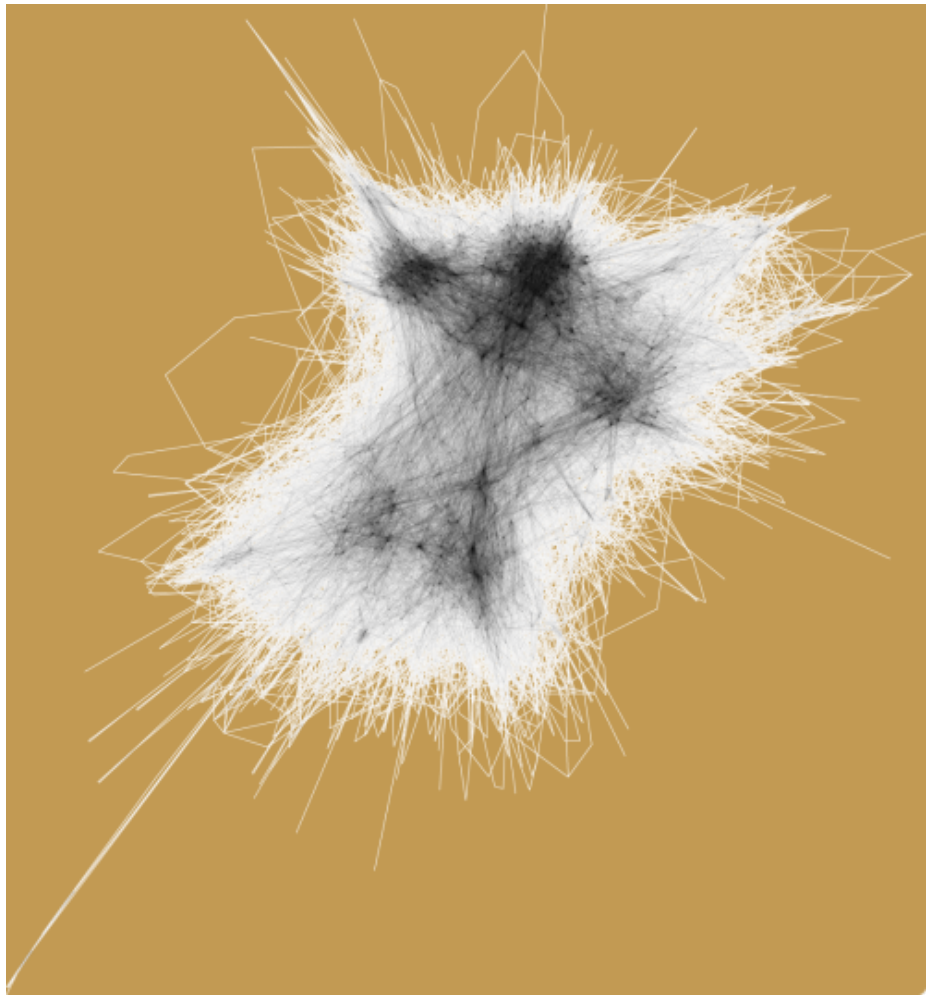
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Weissgerber TL, Milic NM, Winham SJ, Garovic VD. Beyond bar and line graphs: time for a new data presentation paradigm. PLoS Biol. 2015 Apr;13(4):e1002128



Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Avoid “Ridiculograms”

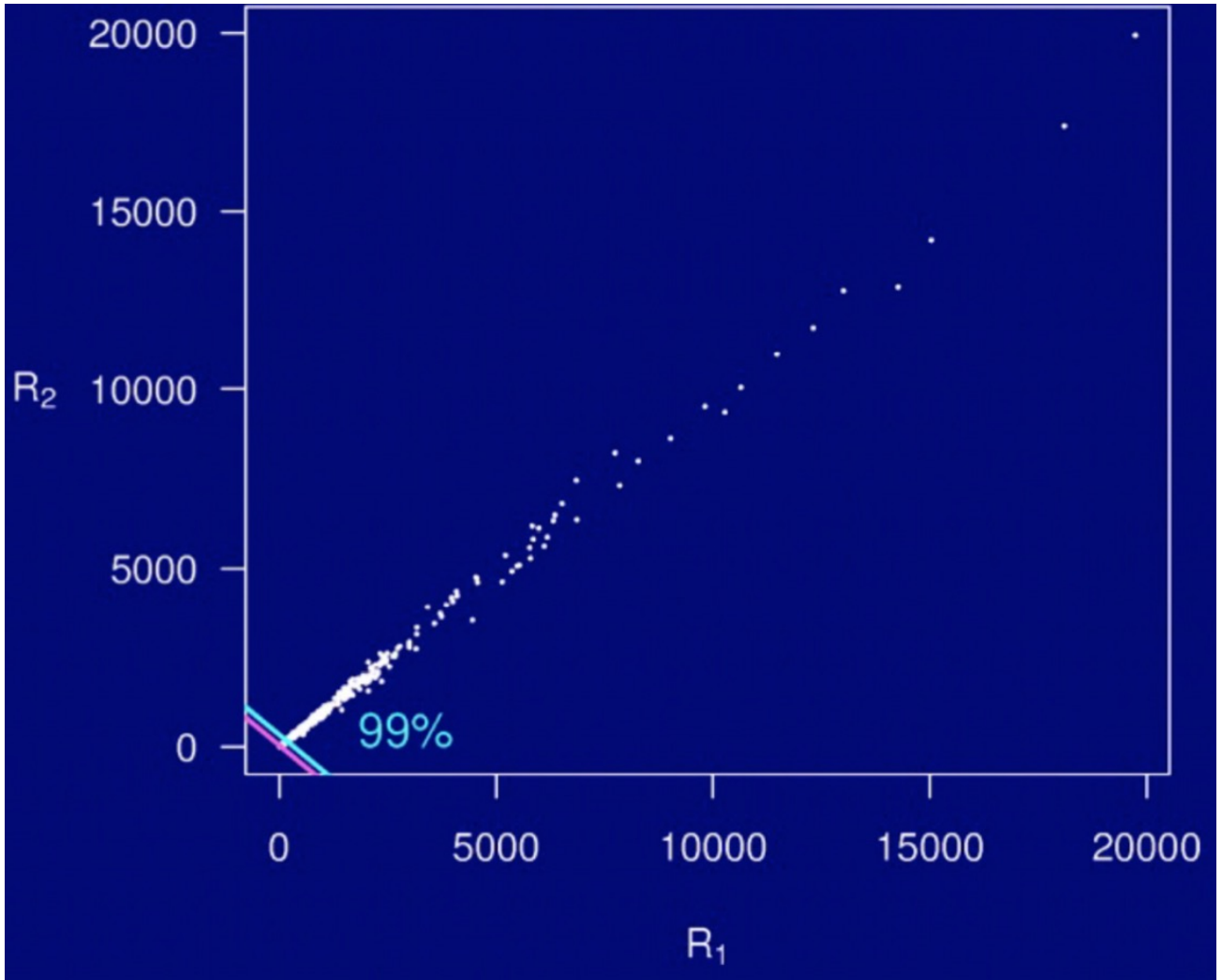


https://petterhol.me/2018/05/28/ridiculograms-a-ridiculous-dialogue/

https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf

Scale - Basic Principles

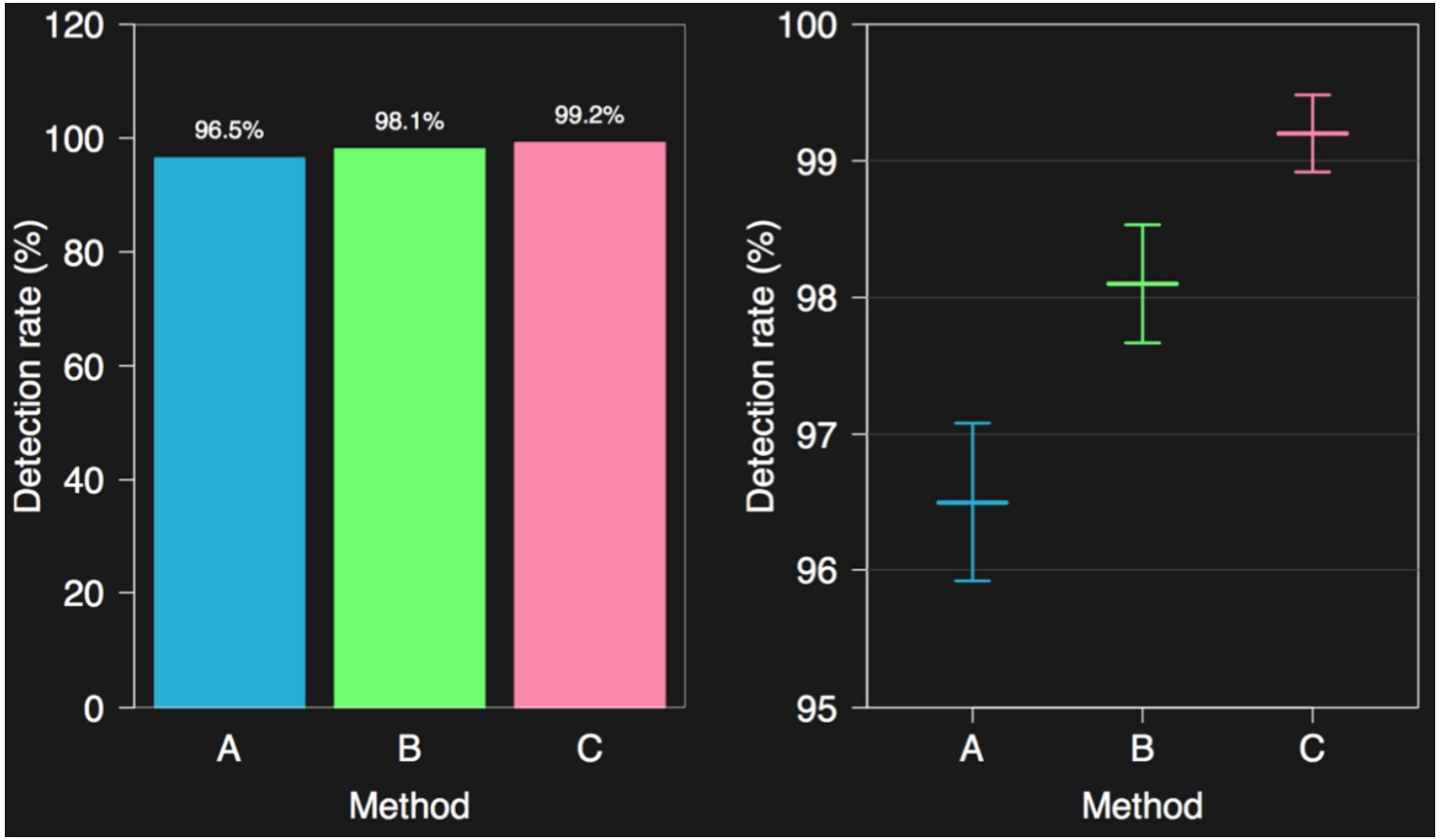
- Be careful with scale



https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf

Scale - Basic Principles

- Use common scales and start at 0



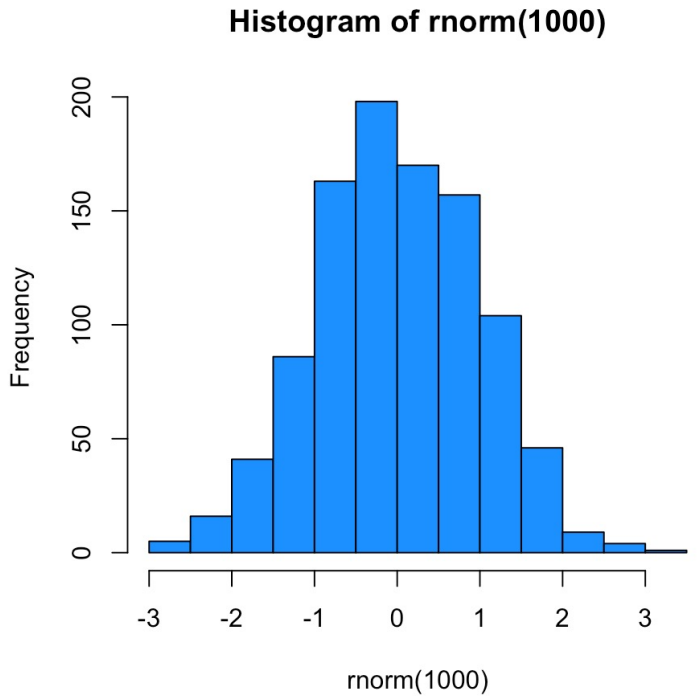
A nice Resource

- http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html

Data Transformations

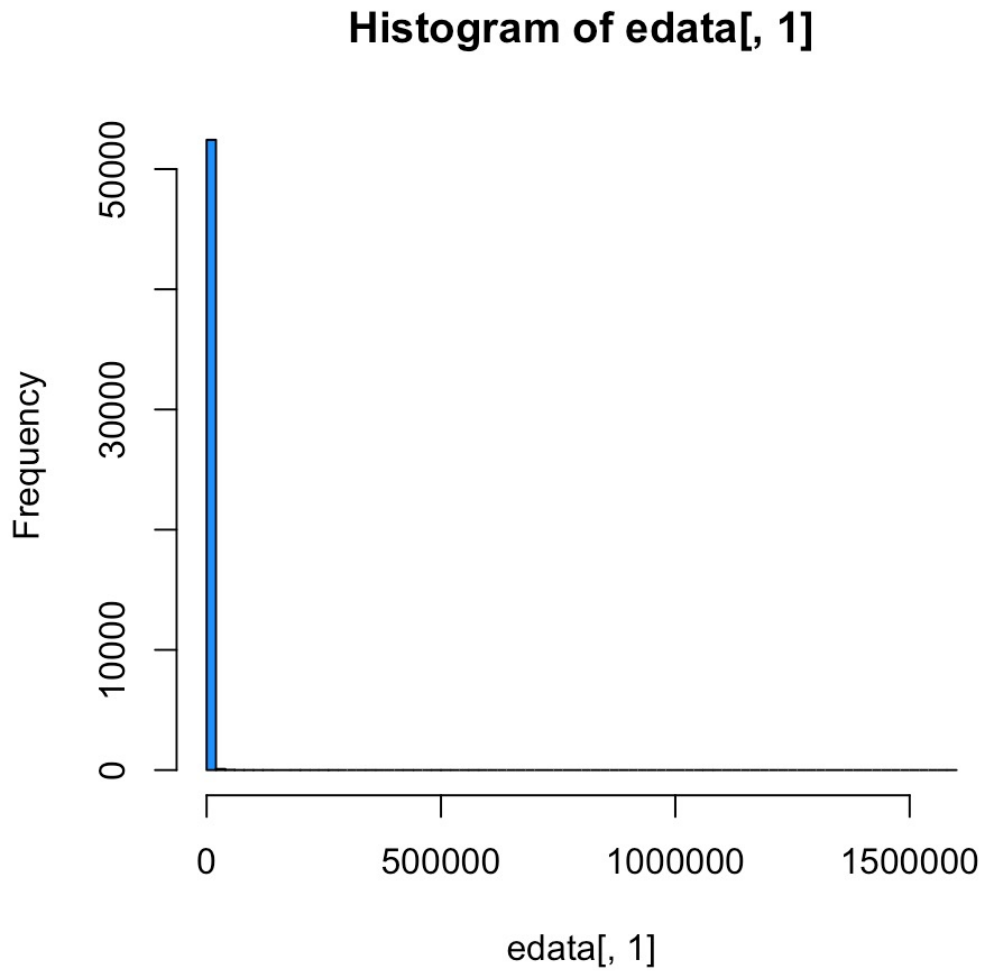
Reasons for need for symmetric data

- plots are easier to see this way
- most statistical methods are designed to work better for non-skewed data



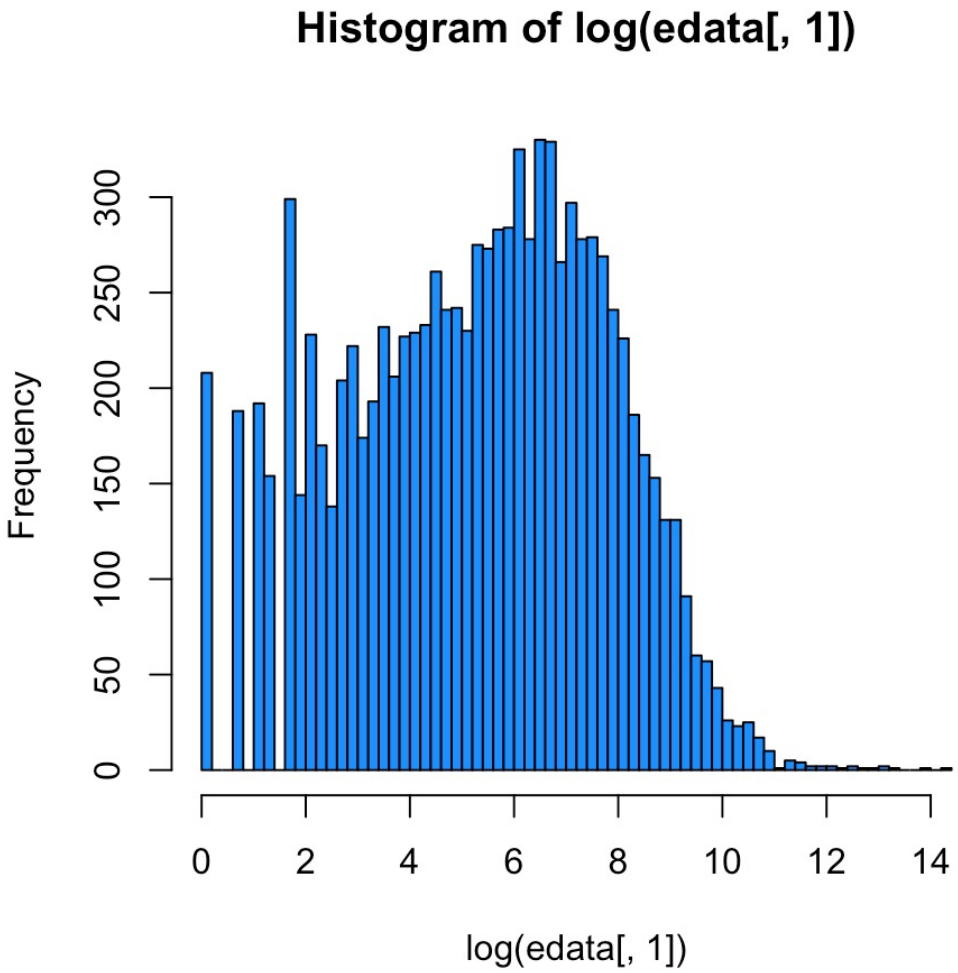
http://jtleek.com/genstats/inst/doc/01_11_transforms.html

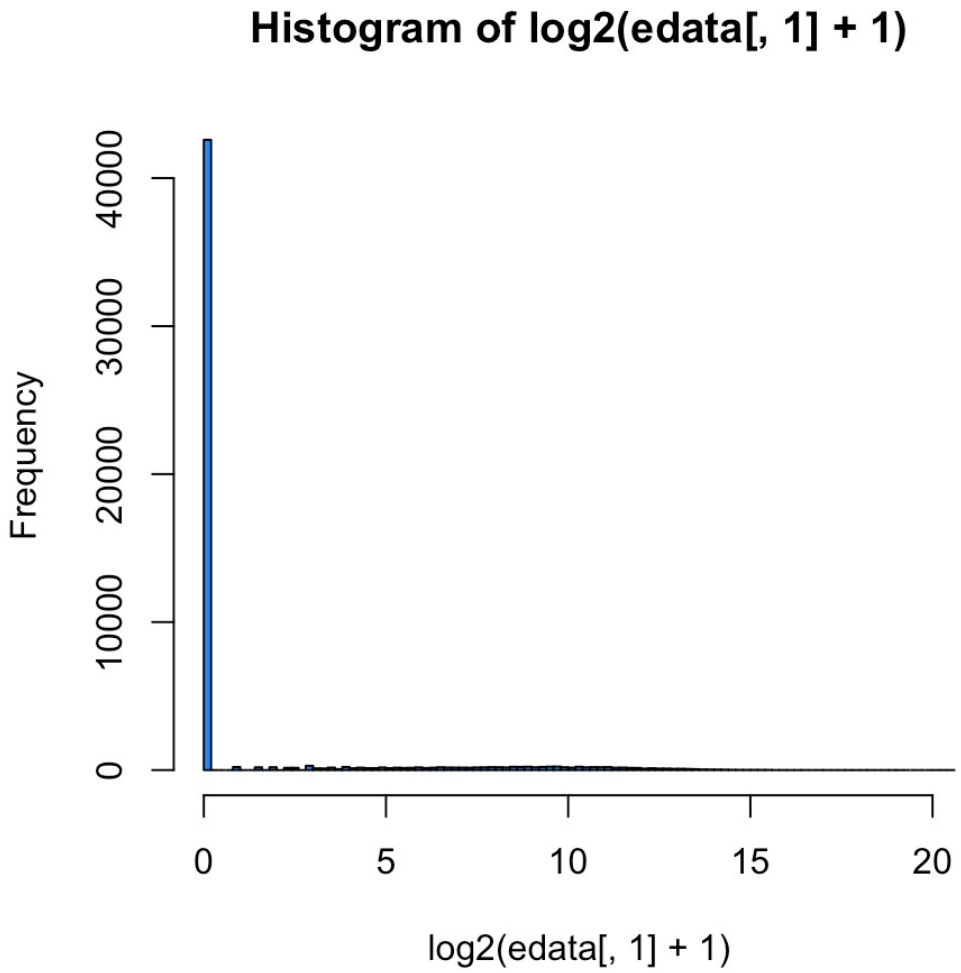
Realistically, most omics data is skewed



http://jtleek.com/genstats/inst/doc/01_11_transforms.html

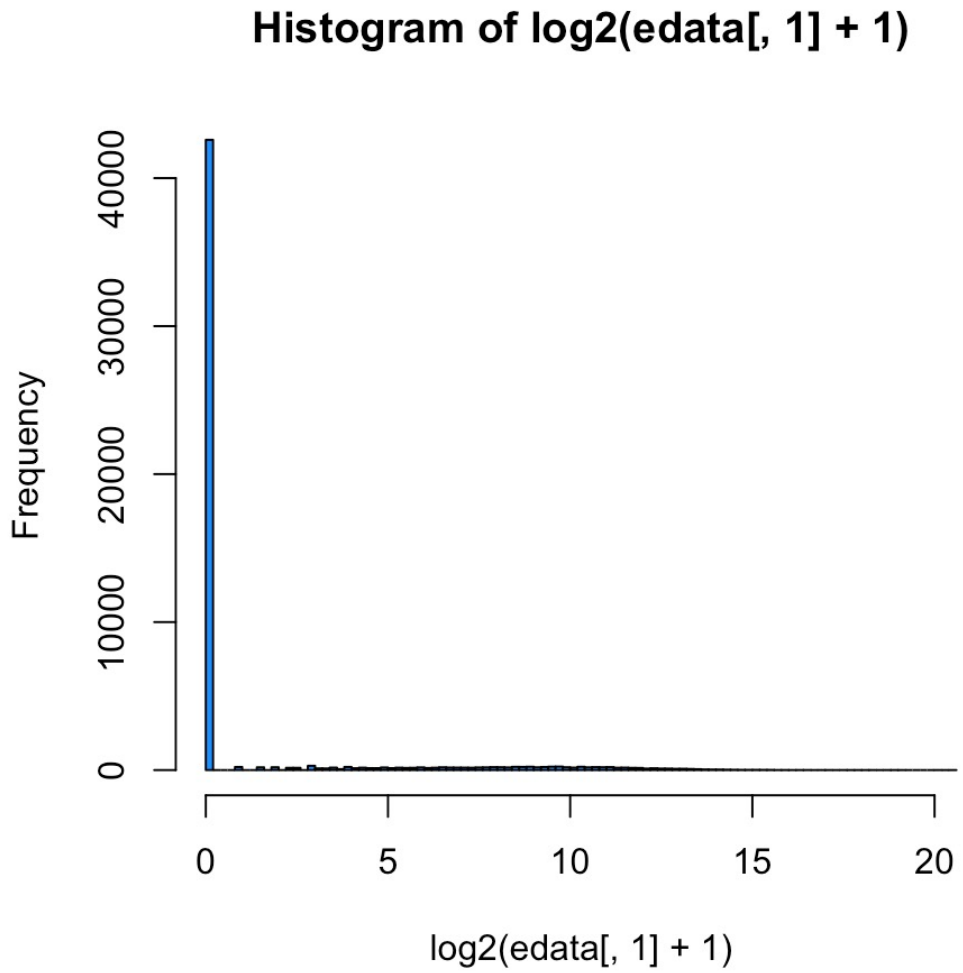
Realistically, most omics data is skewed

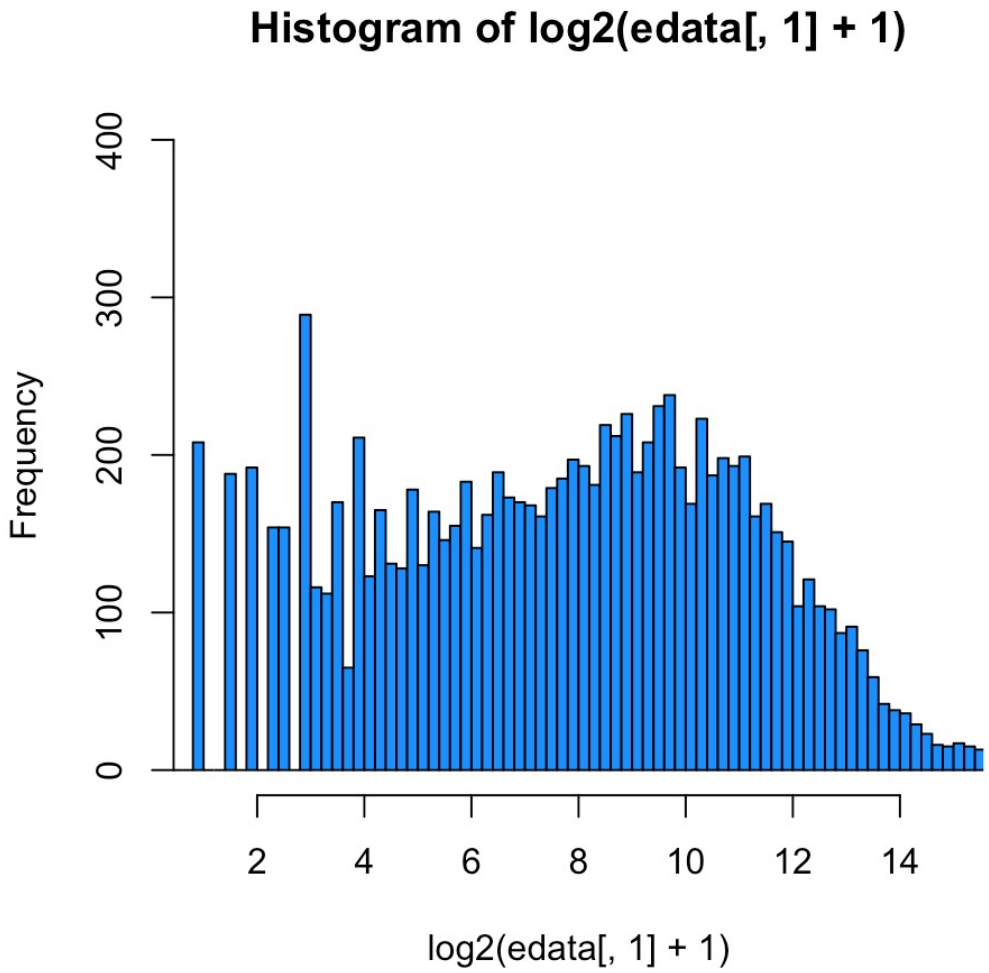




http://jtleek.com/genstats/inst/doc/01_11_transforms.html

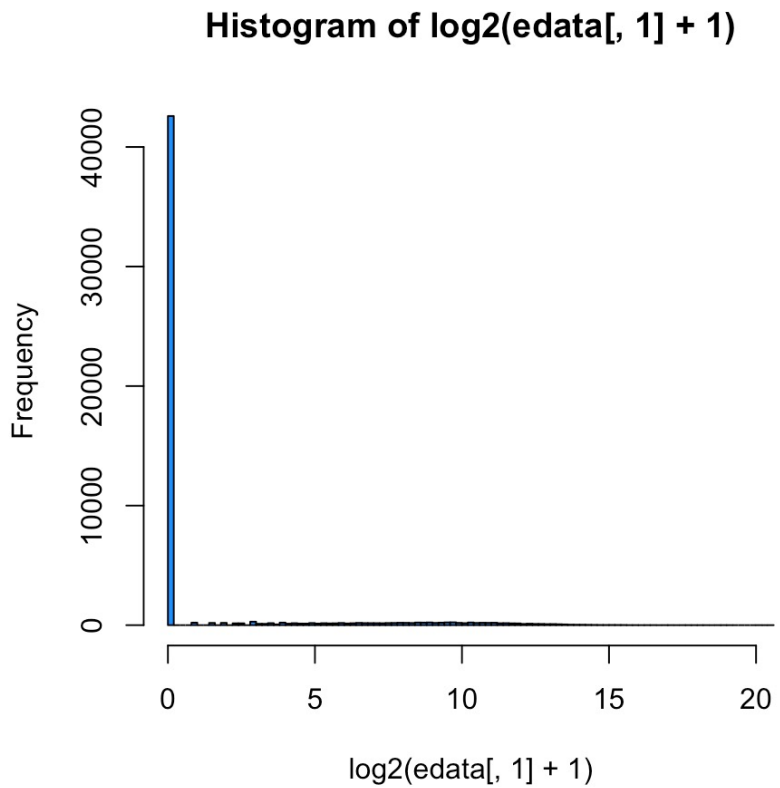
Realistically, most omics data is skewed

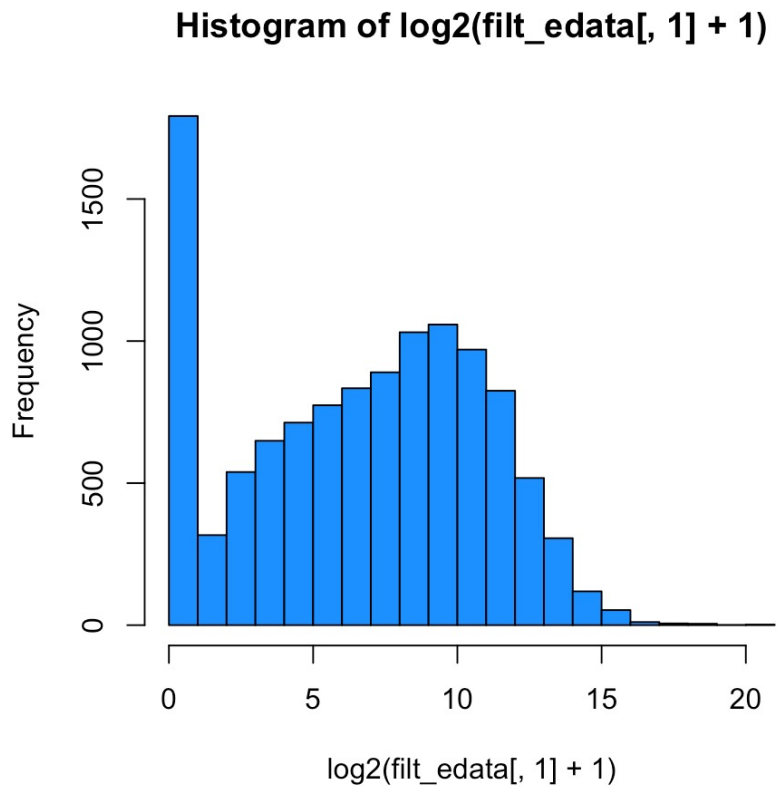




A common pre-processing technique is to remove features that don't have much data

```
low_genes = rowMeans(edata) < 5
filt_edata = filter(edata,!low_genes)
```





Other transformations

- **Variance stabilizing transforms**

  which seek to remove a mean variance relationship among the data

- **Box-Cox transforms**

  which seek to make the data approximately Normally distributed

- **rlog transform**

  unique to genomics count data, this is a regularized version of the log transform that seeks to minimize differences at low count levels

Preprocessing

- Convert raw data to “processed”, trying to remove technical artifacts
  - GC content bias
  - PCR duplicates
  - Probe sequence and fragment length
  - …
- highly platform/problem dependent

Normalization

- Remove technological biases
- Make samples comparable

- highly platform/problem dependent

Scaling data - Standardization

$$X_i' = \frac{X_i - center(X)}{scale(X)}$$

Center: mean/median
Scale: sd/IQR/MAD

Scaling data – Min-Max Scaling

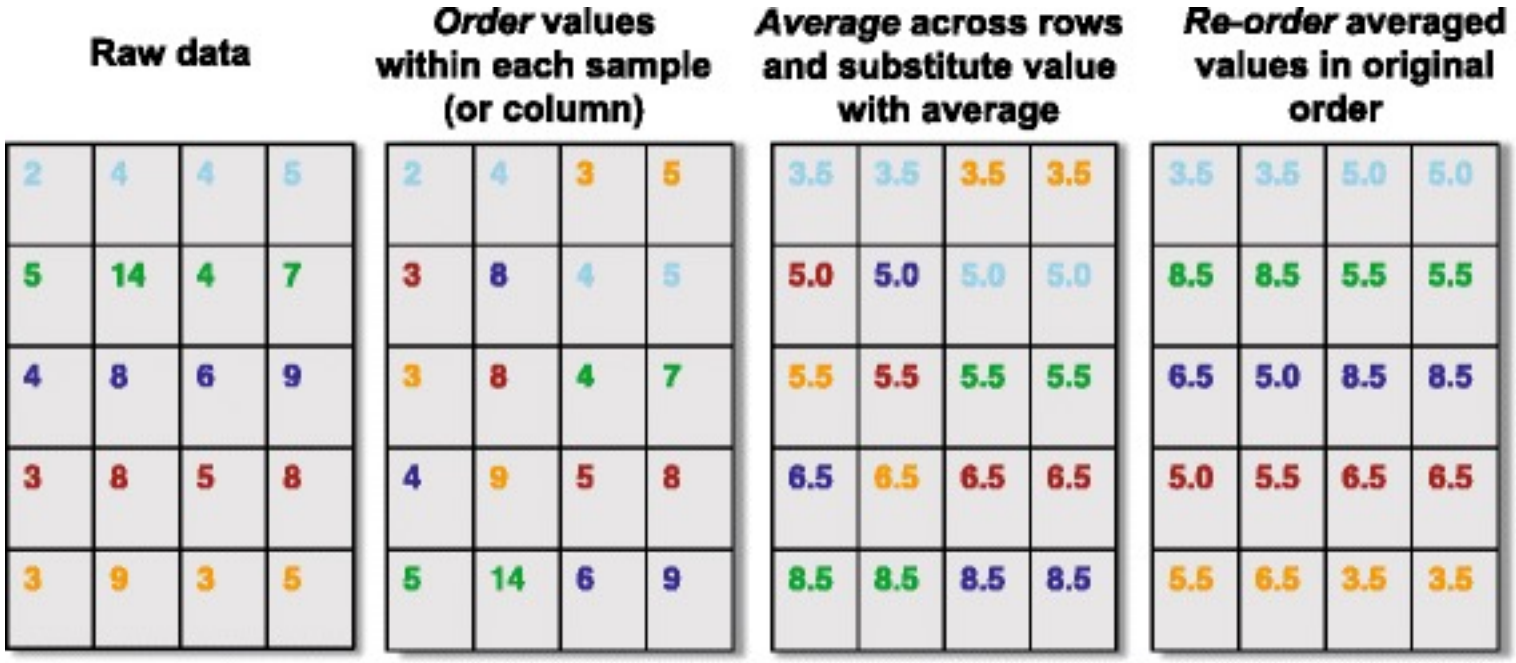
$$X_i' = \frac{X_i - \min(X)}{\max(X) - \min(X)}$$

Scaling in any interval [a,b]

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0679-0

Quantile Normalization



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0679-0

Quantile Normalization