BIOTEC
Biotechnology Center
TU Dresden

PoL
Physics of Life
TU Dresden

CENTER FOR
SYSTEMS BIOLOGY
DRESDEN

[Bio-] statistics

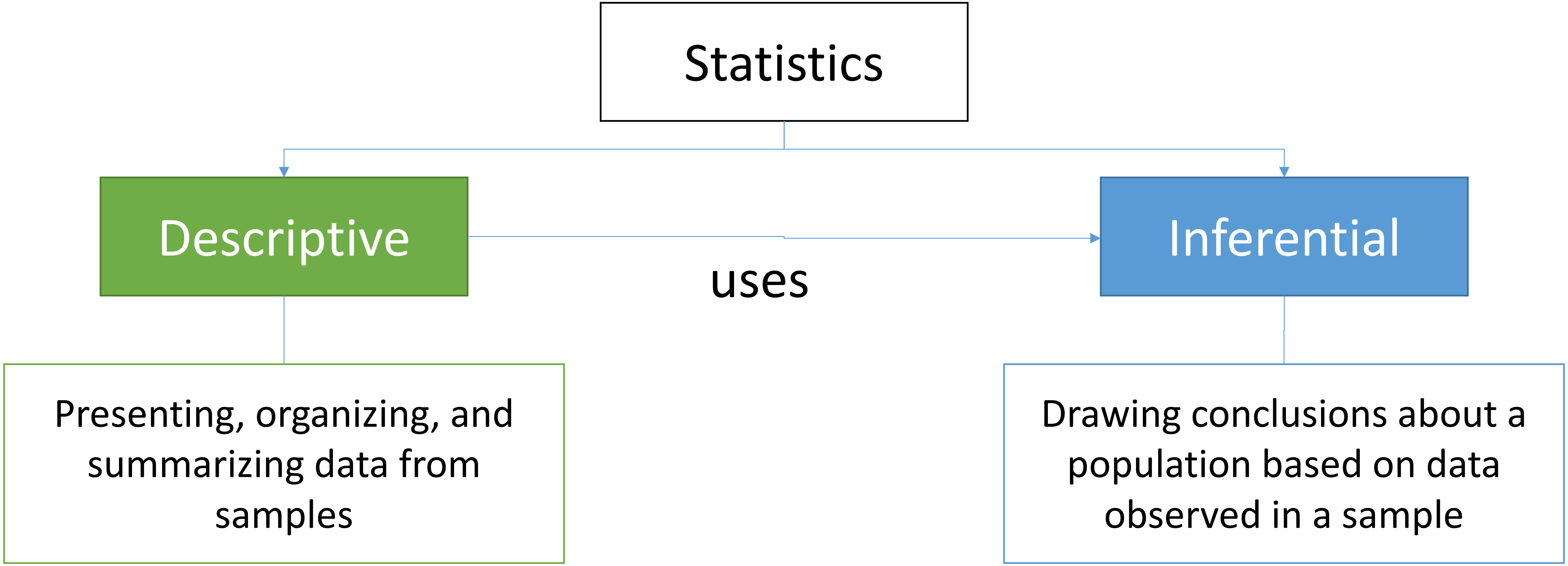
Till Korten

With material from
Justin Bois, Caltech
Marcelo L. Zoccoler, Johannes Müller, Robert Haase, PoL – TU Dresden
Aldo Acevedo Toledo, Biotec, TU Dresden
Martin J. Bland and Douglas G. Altman

December 2022

@TillKorten

Descriptive statistics



Statistics

Descriptive

Inferential

uses

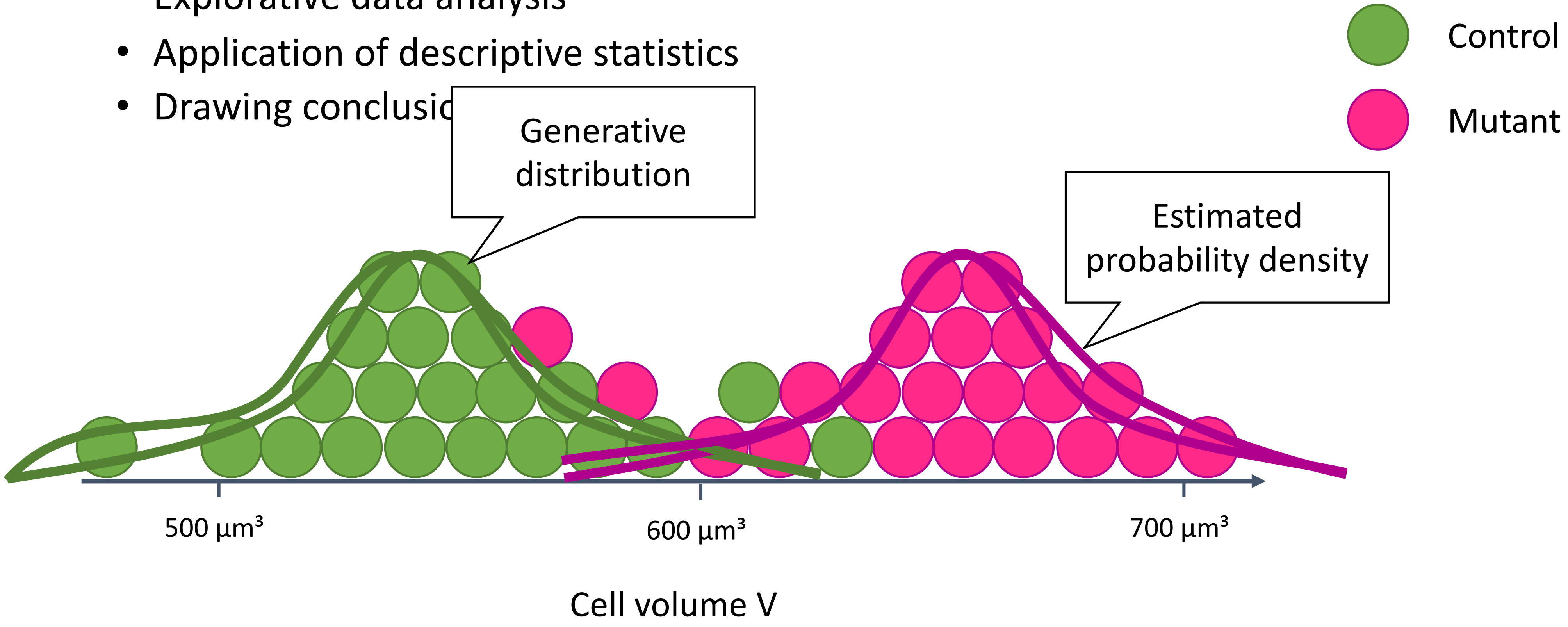
Presenting, organizing, and summarizing data from samples

Drawing conclusions about a population based on data observed in a sample

Slide adapted from
Aldo Acevedo Toledo, Biotec, TU Dresden

Measurements are Samples of a Generative Distribution

- Repeated sampling approximates the unknown underlying generative distribution, enabling
  - Explorative data analysis
  - Application of descriptive statistics
  - Drawing conclusio

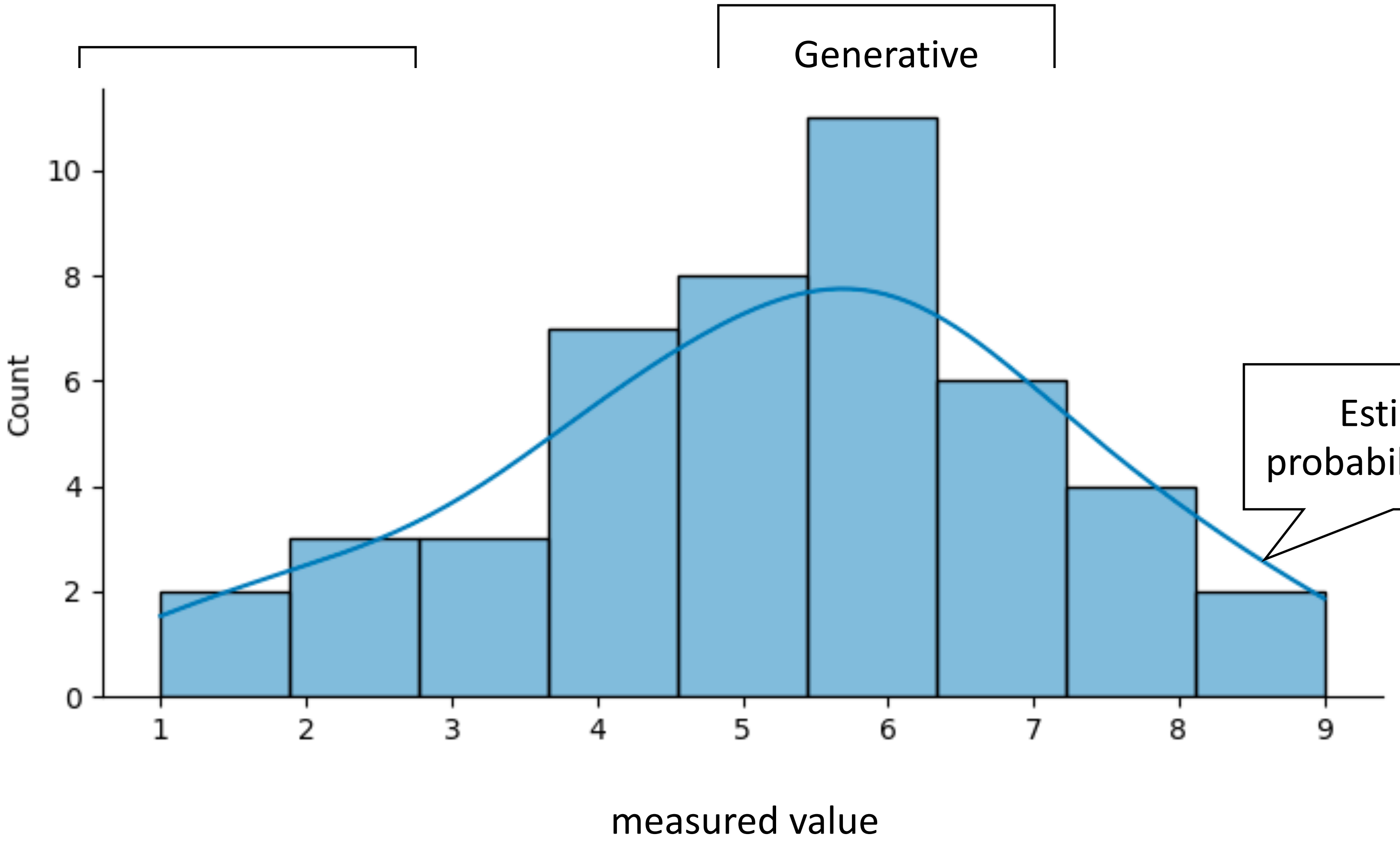


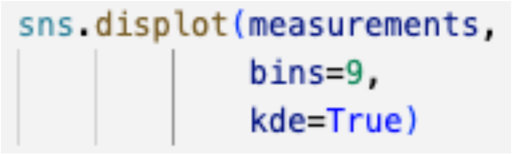
Generative distribution

Estimated probability density

How do we get from Data to Distributions?

- Measurements: [7, 5, 5, 6, 6, 7, 8, 7, 2, 4, 6, 4, 1, 6, 4, 5, 2, 7, 6, 5, 9, 5, 6, 3, 4, 8, 6, 2, 4, 6, 5, 4, 1, 5, 6, 7, 8, 3, 4, 5, 8, 6, 3, 7, 6, 9]





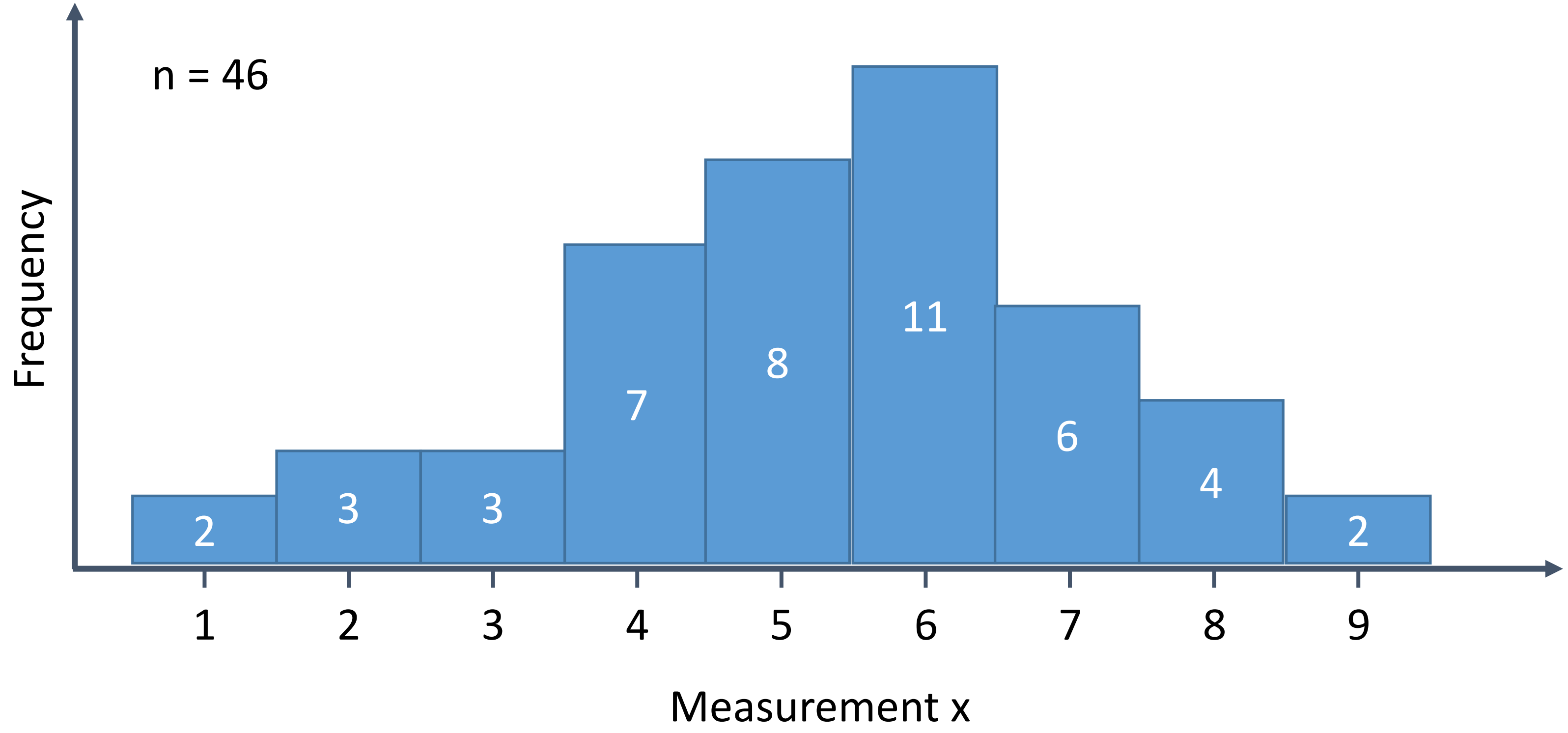
```
sns.displot(measurements,
            bins=9,
            kde=True)
```

KDE: Kernel density estimation
a method to estimate the probability density

Estimated probability density

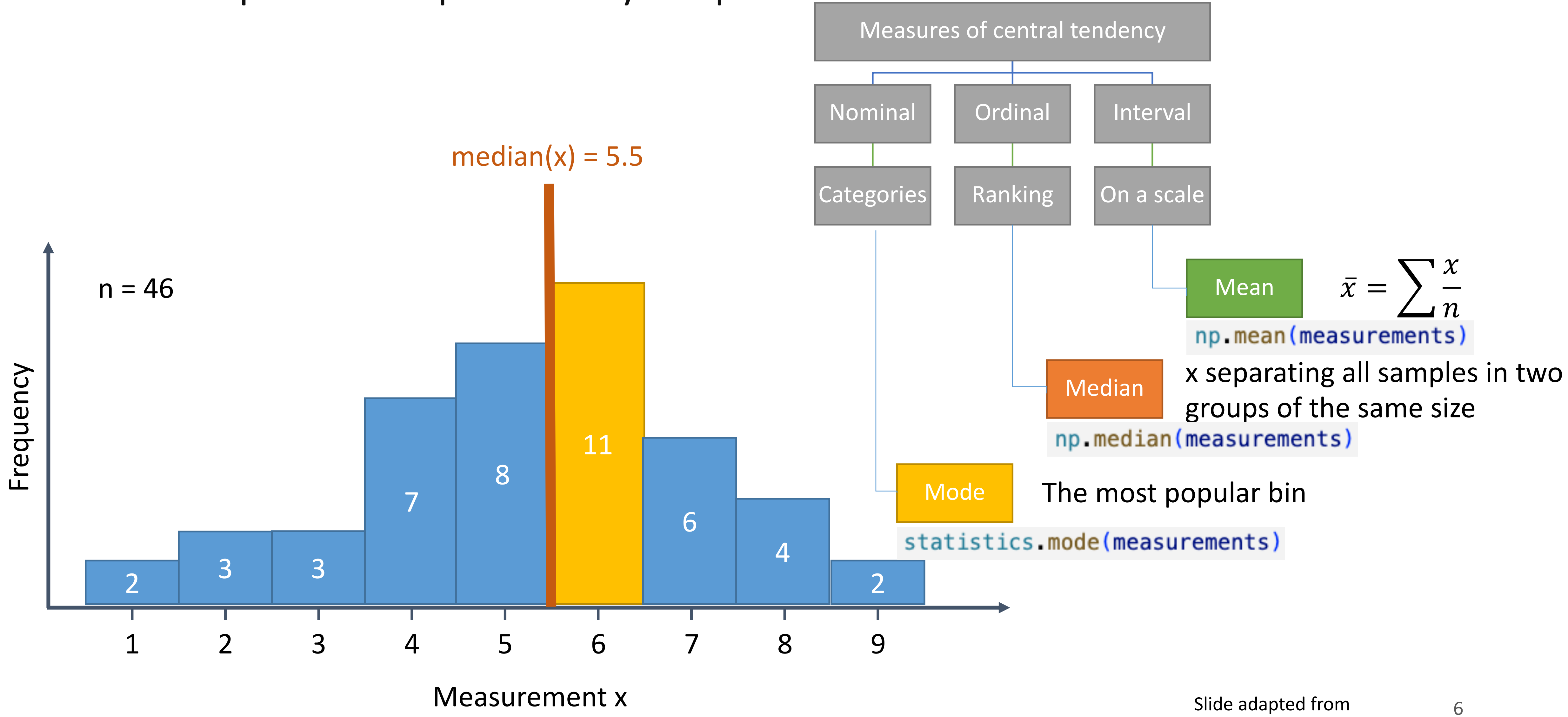
Measures of central tendency

- “Where” in parameter space are my samples located?



Measures of central tendency

- “Where” in parameter space are my samples located?



Measures of central tendency

Nominal

Ordinal

Interval

Categories

Ranking

On a scale

Mean $\bar{x} = \sum \frac{x}{n}$

```
np.mean(measurements)
```

Median x separating all samples in two groups of the same size

```
np.median(measurements)
```

Mode The most popular bin

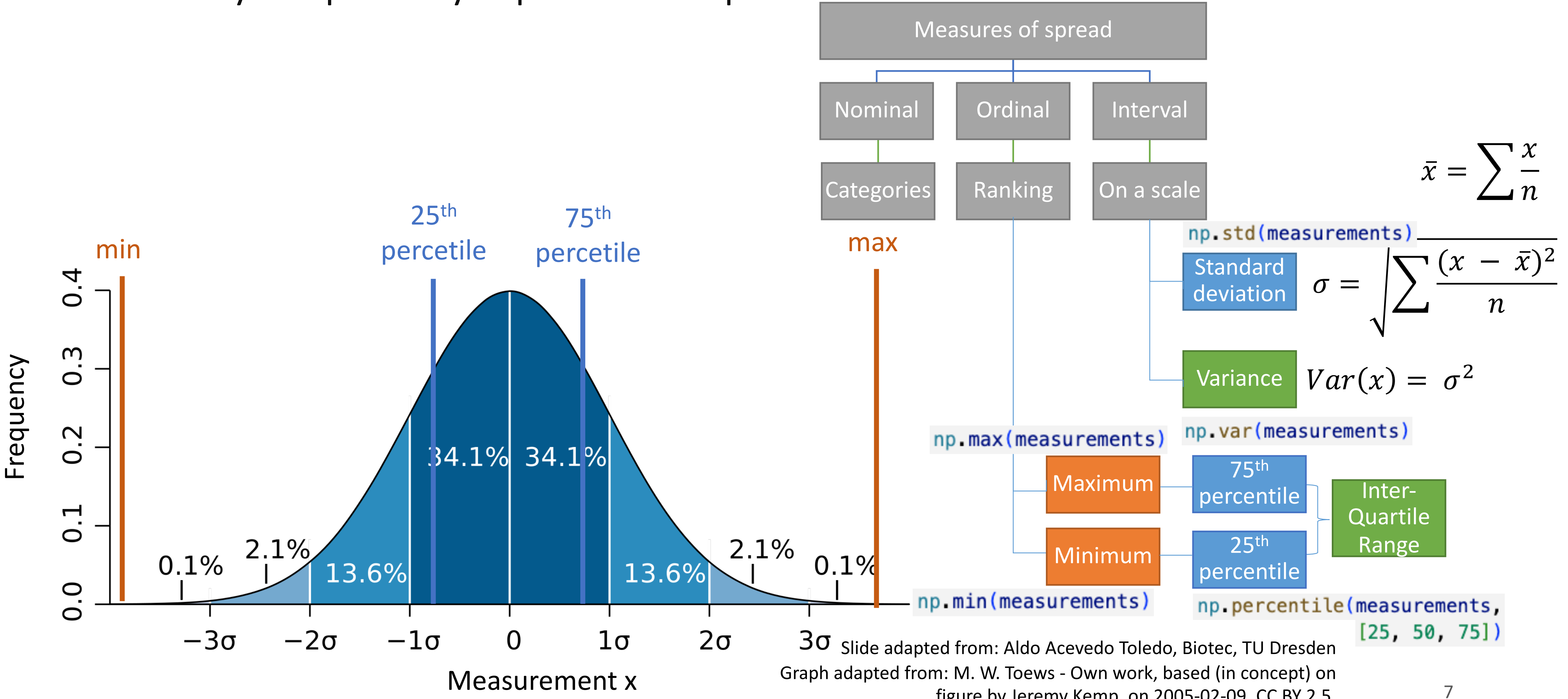
```
statistics.mode(measurements)
```

@TillKorten

Slide adapted from
Aldo Acevedo Toledo, Biotec, TU Dresden

Measures of spread

- How do my samples vary in parameters space?



$$\bar{x} = \sum \frac{x}{n}$$

$$\sigma = \sqrt{\sum \frac{(x - \bar{x})^2}{n}}$$

$$Var(x) = \sigma^2$$

```
np.std(measurements)
np.var(measurements)
np.max(measurements)
np.min(measurements)
np.percentile(measurements,
              [25, 50, 75])
```

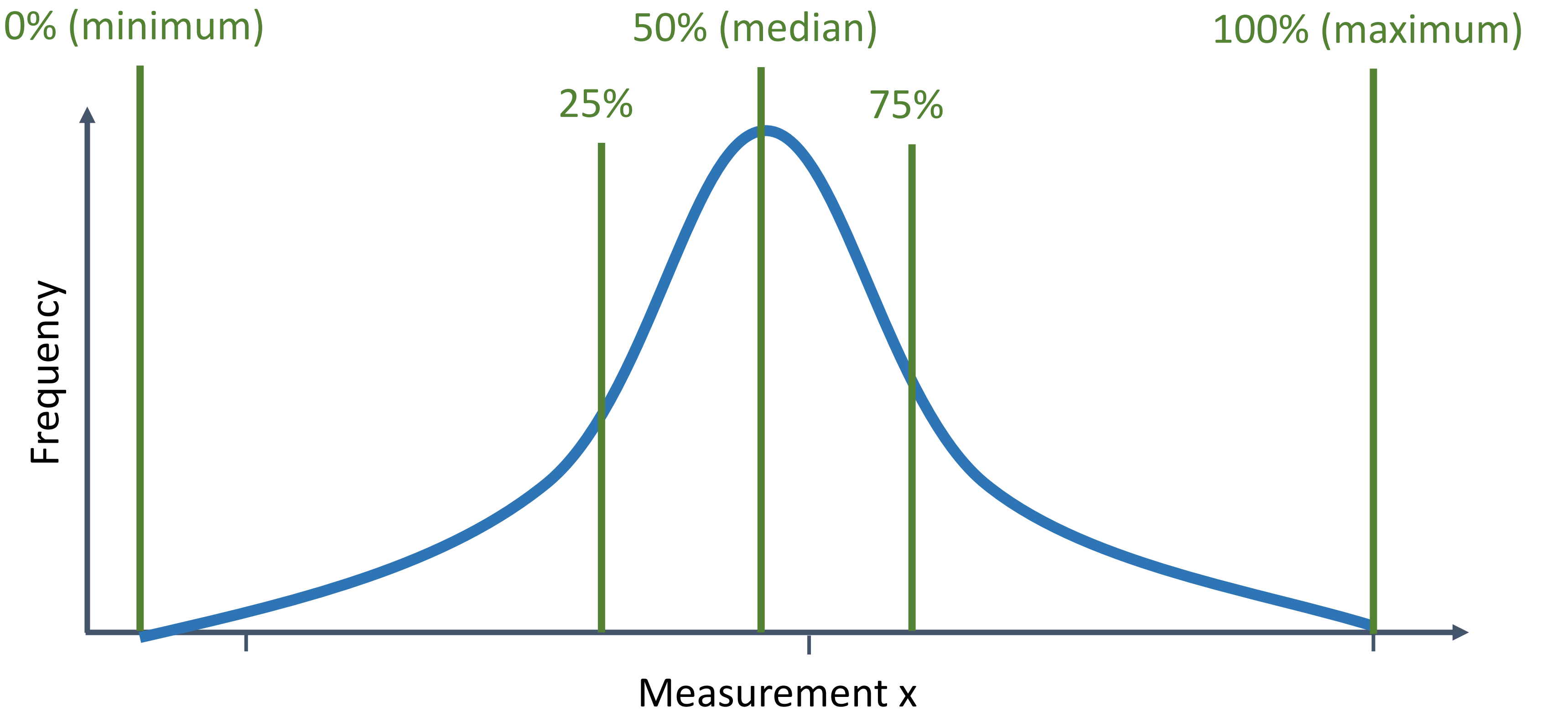
Slide adapted from: Aldo Acevedo Toledo, Biotec, TU Dresden
Graph adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

@TillKorten

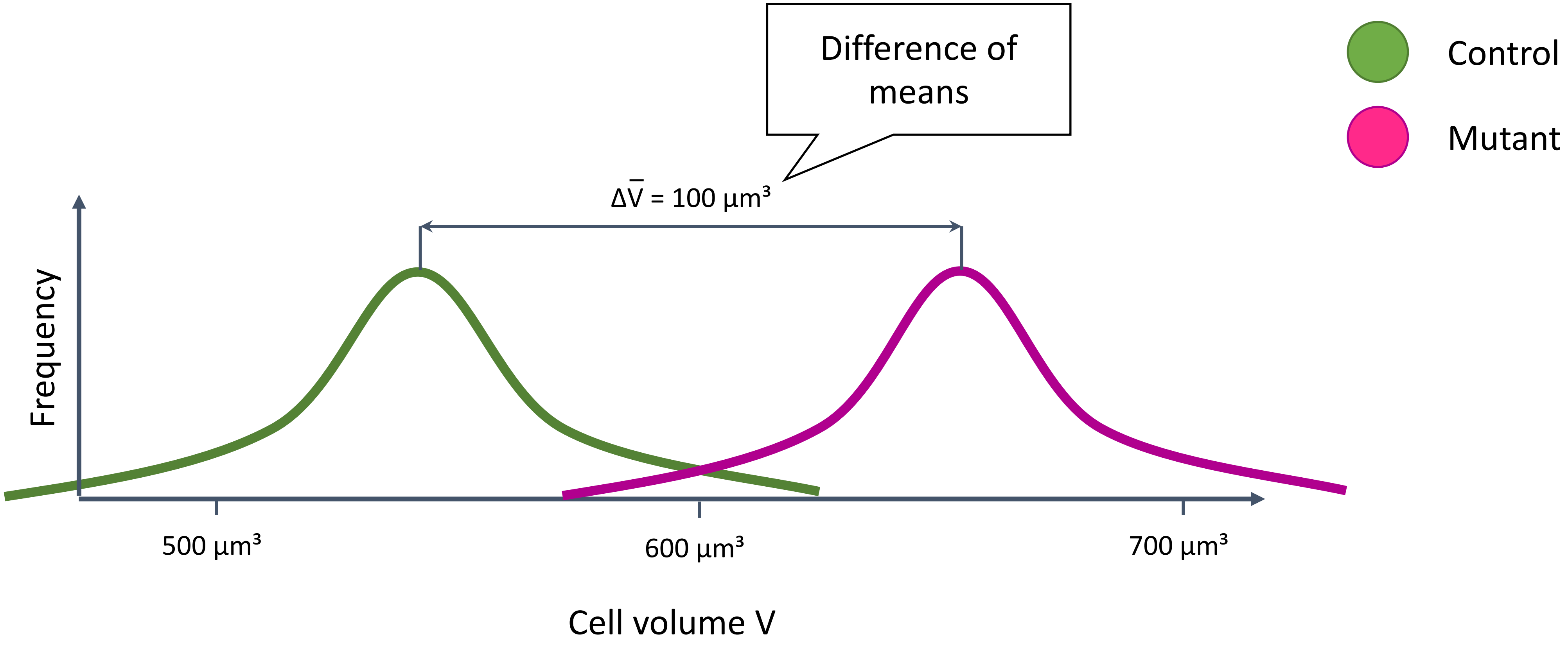
Measures of spread

- Percentiles
  - The value under which a given percentage of our samples lie
  - Independent of distribution

```
np.min(measurements)
np.quantile(measurements, [0.25, .50, .75])
np.max(measurements)
```

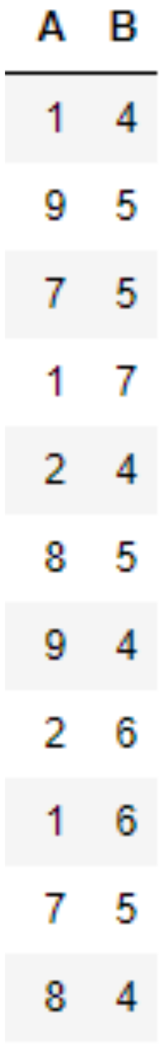


Comparing Means of Known Distributions is Reasonable

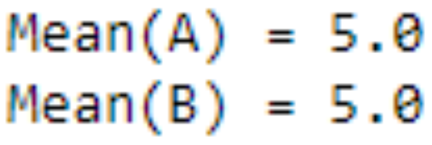


Comparing Means of Unknown Distributions can be Misleading

- Are two measurements coming from the same distribution, if their mean is similar?



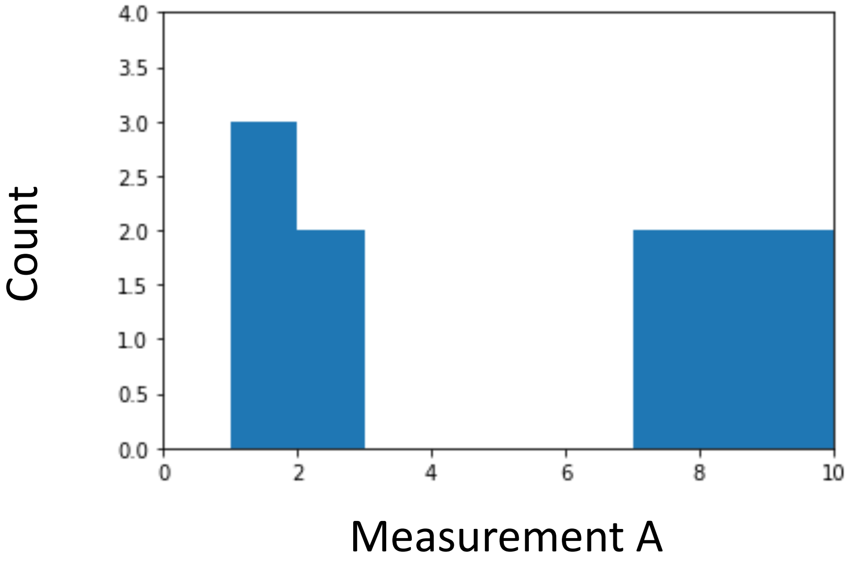
A	B
1	4
9	5
7	5
1	7
2	4
8	5
9	4
2	6
1	6
7	5
8	4

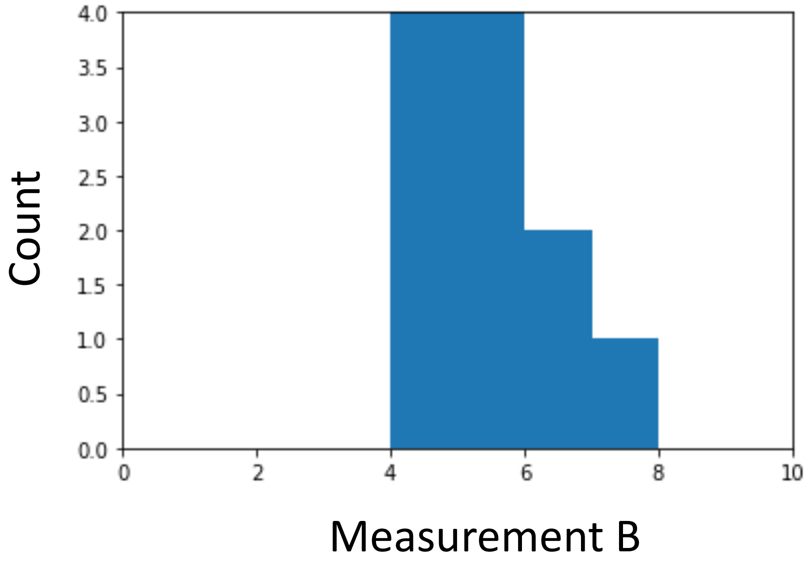


```
Mean(A) = 5.0
Mean(B) = 5.0
```

Similar means is a necessary condition, but it is NOT sufficient!

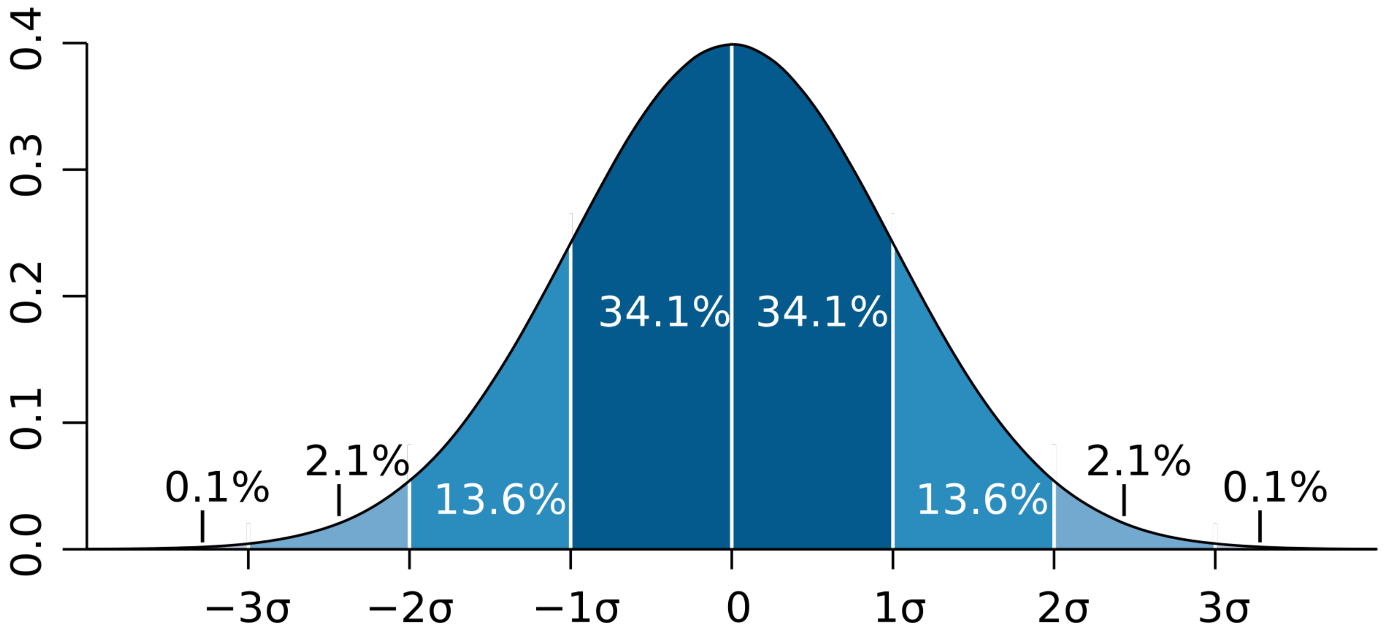
- Draw histograms. These distributions look very different!





Parametric vs. non-parametric

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



Normal distribution:

- Can be completely described by mean µ and standard deviation σ
- Allows comparing distributions (e.g., with two-sided/paired t-test)

Ranked distribution:

- Replace each value with its “rank”
- Rank = index of value in sorted list
- Robust to outliers
- Independent of underlying distribution

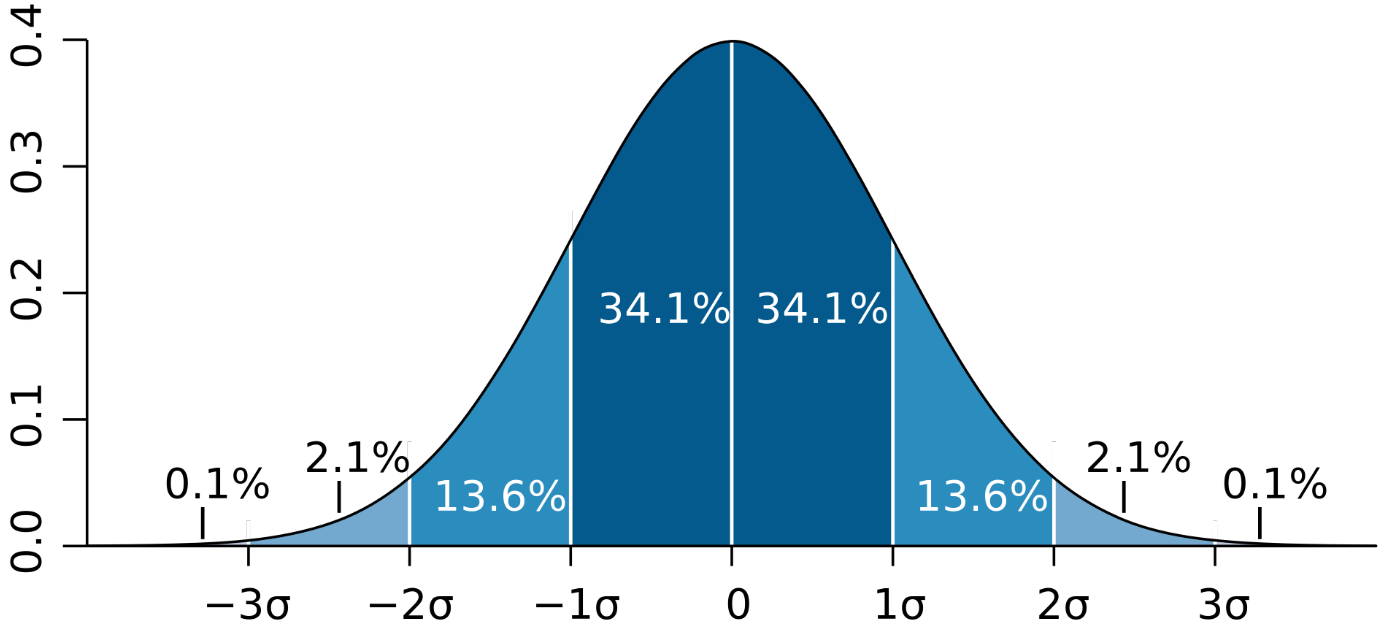
Value	Rank
10	1
15	2
3	0
97	3

Graph adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

Testing for Normality: scipy.stats.shapiro

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

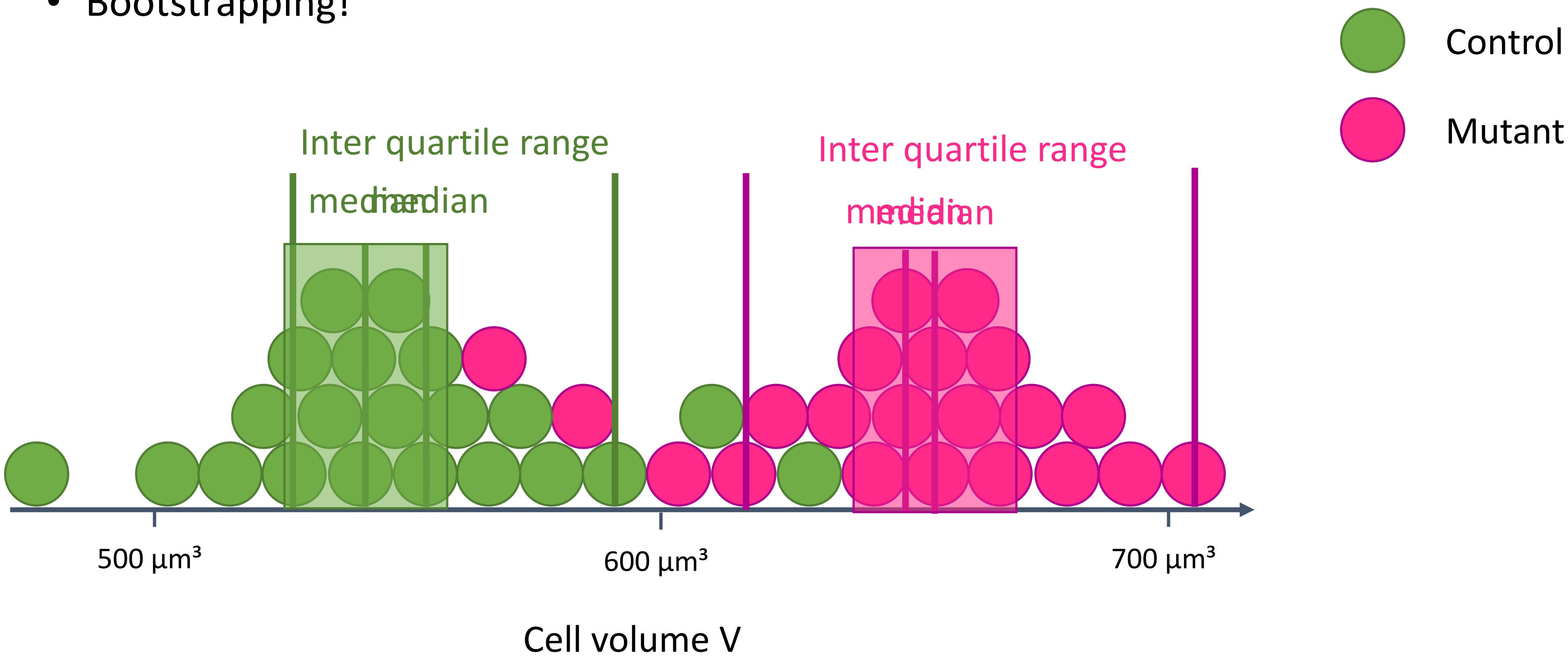
```
from scipy import stats

stats.shapiro(measurements)

ShapiroResult(statistic=0.964,
pvalue=0.161)
```



@TillKorten

Measure of Confidence

- How confident are we in our data?
  - Need 95% confidence intervals
  - Bootstrapping!



@TillKorten

The 95% Confidence Interval Explained

- “If an experiment is repeated over and over again, the estimate I compute for a parameter, $\hat{\theta}$, will lie between the bounds of the 95% confidence interval for 95% of the experiments”[1].
- “[The above definition] is correct but useless since we rarely repeat the same experiment over and over. A better interpretation is this: On day 1, you collect data and construct a 95 percent confidence interval for a parameter $\hat{\theta}_1$. On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter $\hat{\theta}_2$. […] You continue this way constructing confidence intervals for a sequence of unrelated parameters $\hat{\theta}_3, \hat{\theta}_4, \ldots \hat{\theta}_n$. Then 95 percent of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over”[2].

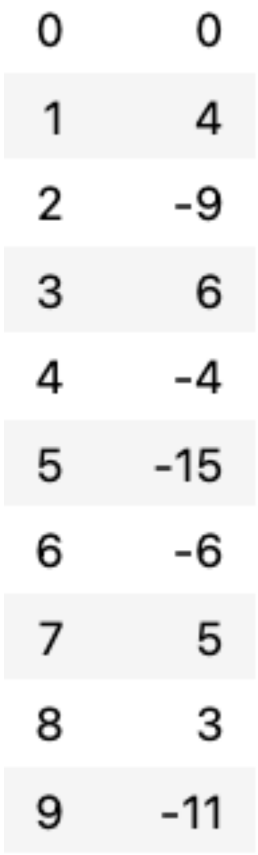
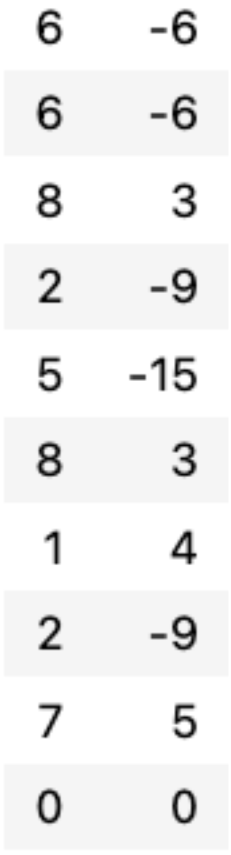
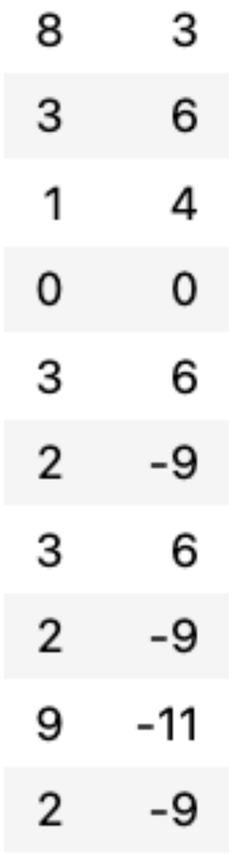
[1] Justin Bois https://justinbois.github.io/dd-pol/2022/lessons/04/confidence_intervals.html
Creative Commons Attribution License CC-BY 4.0
[2] Larry Wasserman, All of Statistics 2004

Bootstrapping: resample your data to estimate uncertainty

- Generate B independent bootstrap samples. Each one is generated by drawing n values out of the data array with replacement.
- Compute $\hat{\theta}^*$ for each bootstrap sample to get the bootstrap replicates.
- The central 95 percent confidence interval consists of the percentiles 2.5 and 97.5 of the bootstrap replicates.”[1]

```
np.percentile(bootstrap_samples,
[2.5, 97.5])
```

```
array([1.49875, 7.70125])
```

Experiment		Resampled Experiment		Resampled Experiment	
0	0	6	-6	8	3
1	4	6	-6	3	6
2	-9	8	3	1	4
3	6	2	-9	0	0
4	-4	5	-15	3	6
5	-15	8	3	2	-9
6	-6	1	4	3	6
7	5	2	-9	2	-9
8	3	7	5	9	-11
9	-11	0	0	2	-9

$\hat{\theta}$: `np.mean()` **-2.7** **-3.0** **-1.3**

[1] Justin Bois https://justinbois.github.io/dd-pol/2022/lessons/04/confidence_intervals.html
Creative Commons Attribution License CC-BY 4.0

[2] Larry Wasserman, All of Statistics 2004

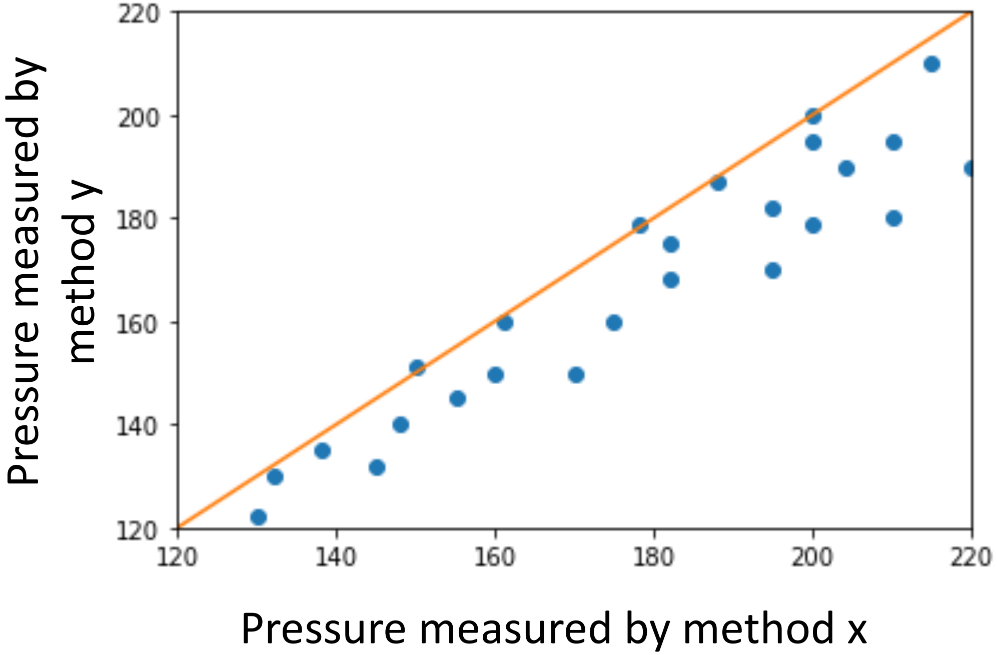
Use Bootstrapping

- Bootstrapping works for any estimate θ that you can calculate from your data
- Bootstrapping is nonparametric – it does not make assumptions about the underlying generative distribution
- Don’t let anyone tell you, that resampling is problematic – you are using it to **estimate confidence**, not to increase confidence (as you would do by adding more measurements)
- Bootstrapping is mathematically proven to work – provided you have more than 17 measurements to resample (Bradley Efron, "Bootstrap methods: another look at the jackknife" (1979))
- Be patient with long computation times – it probably took you several weeks to perform the experiment and evaluate it – show some respect to your data and calculate the confidence intervals

Pearson Correlation

- Are two methods doing the same if they correlate?
  - Correlation: Any kind of relationship.
  - Measurable; e.g. using Pearson's Correlation Coefficient *r* enumerated linear correlation.

Comparison of two methods of measuring systolic blood pressure (Data from 1)



Expectation E

Mean average μ

$$r(X,Y) = \frac{E[(X - \mu_X)(Y - \mu_y)]}{\sigma_X \sigma_Y}$$

Standard deviation σ
→ Unit independence

Disclosure: Mean and standard deviation must be obtained from the whole population or from a sample set which is sufficiently large.

In practice *E* is the weighted sum:

$$r(X,Y) = \frac{\sum_{x \in X, y \in Y} \frac{(x - \mu_X)(y - \mu_Y)}{n}}{\sigma_X \sigma_Y}$$

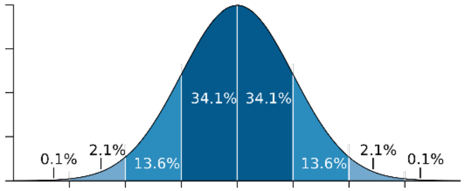
Number of measurements *n*

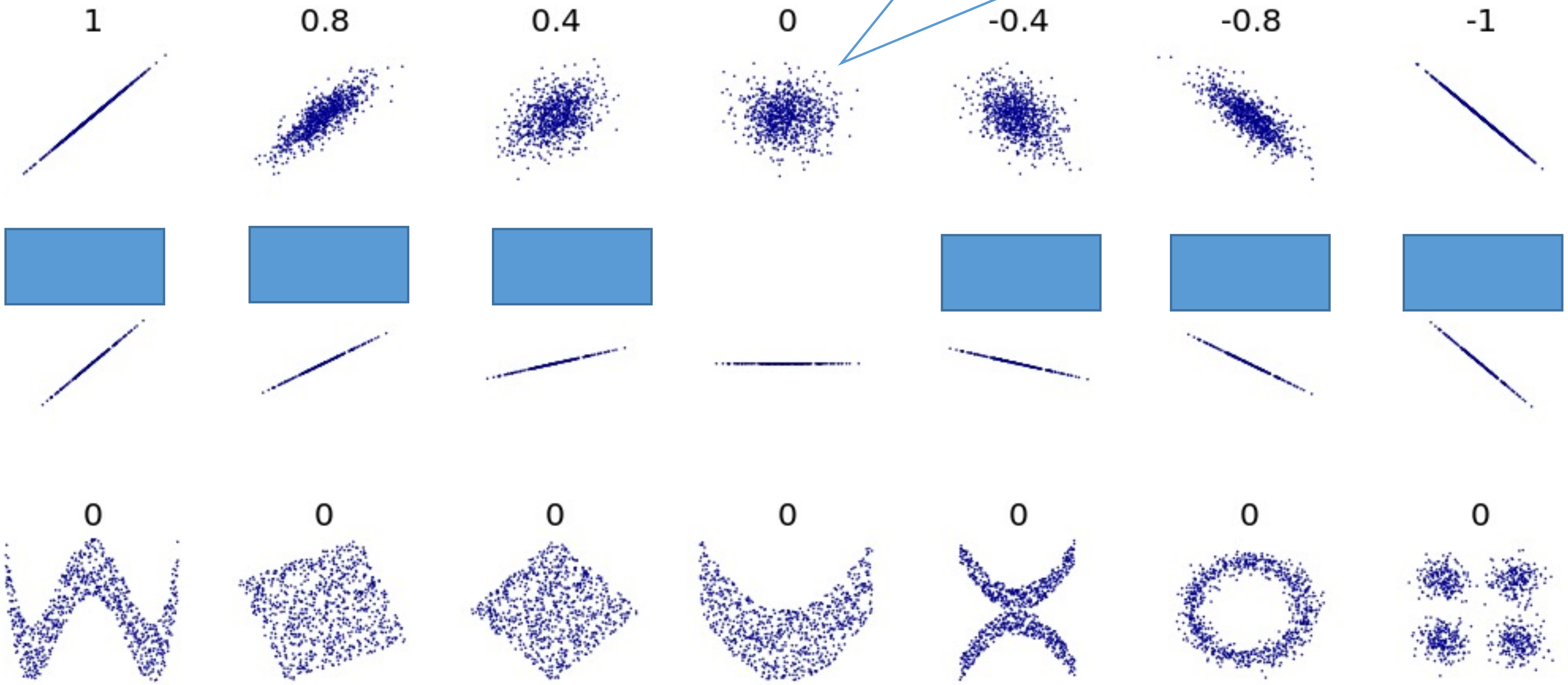
[1] Altman & Bland, The Statistician 32, 1983

Correlation: Pearson's *r*

- Pearson's *r* lies between -1 and 1
  - 1: Positive linear correlation
  - 0: No linear correlation
  - -1: Negative linear correlation

2-dimensional normal distribution





1

-1

0

0.4

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Correlation: Spearman’s r

Value x	Rank x’
10	1
15	2
3	0
97	3
…	…

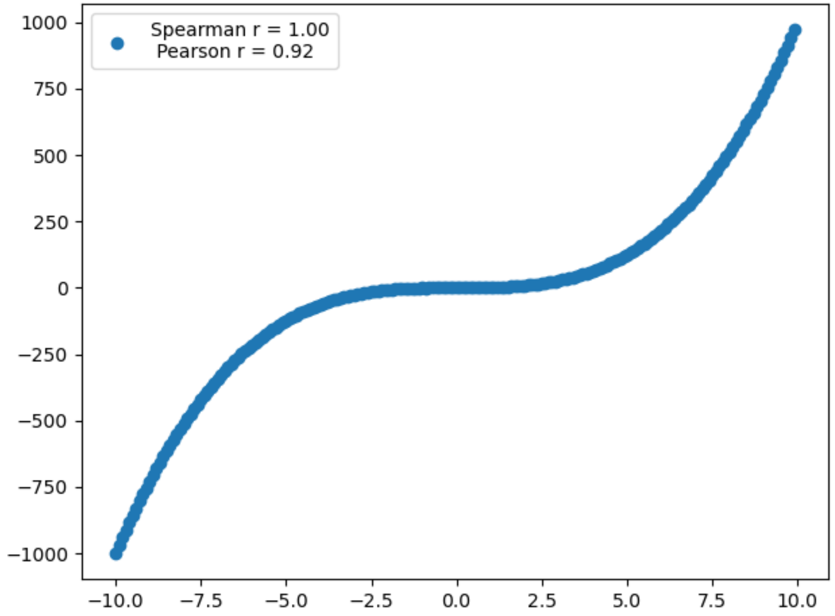
- Spearman’s *r* lies between -1 and 1
  - 1: Positive **monotonous** correlation
  - 0: No monotonous correlation
  - -1: Negative monotonous correlation

Expectation E

Mean average μ

$$r_{Spearman}(X,Y) = \frac{E[(X' - \mu_{x'})(Y' - \mu_{y'})]}{\sigma_{x'}\sigma_{y'}}$$

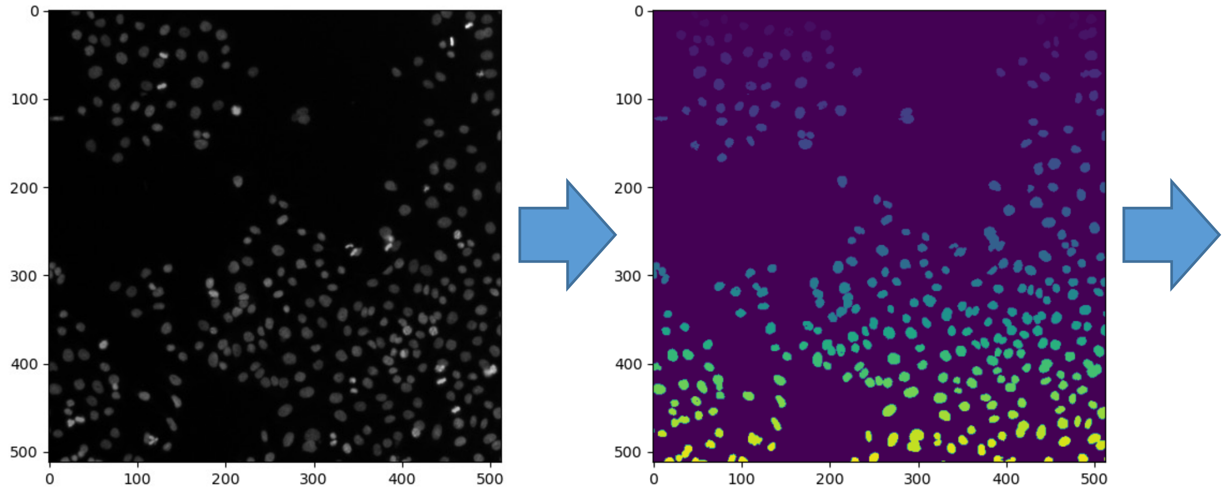
Standard deviation σ

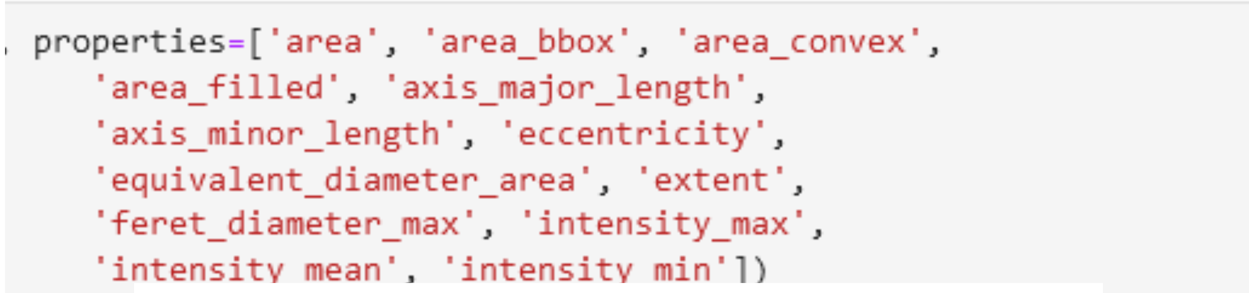


Spearman’s *r* is equivalent to using Pearson’s r on ranked data:

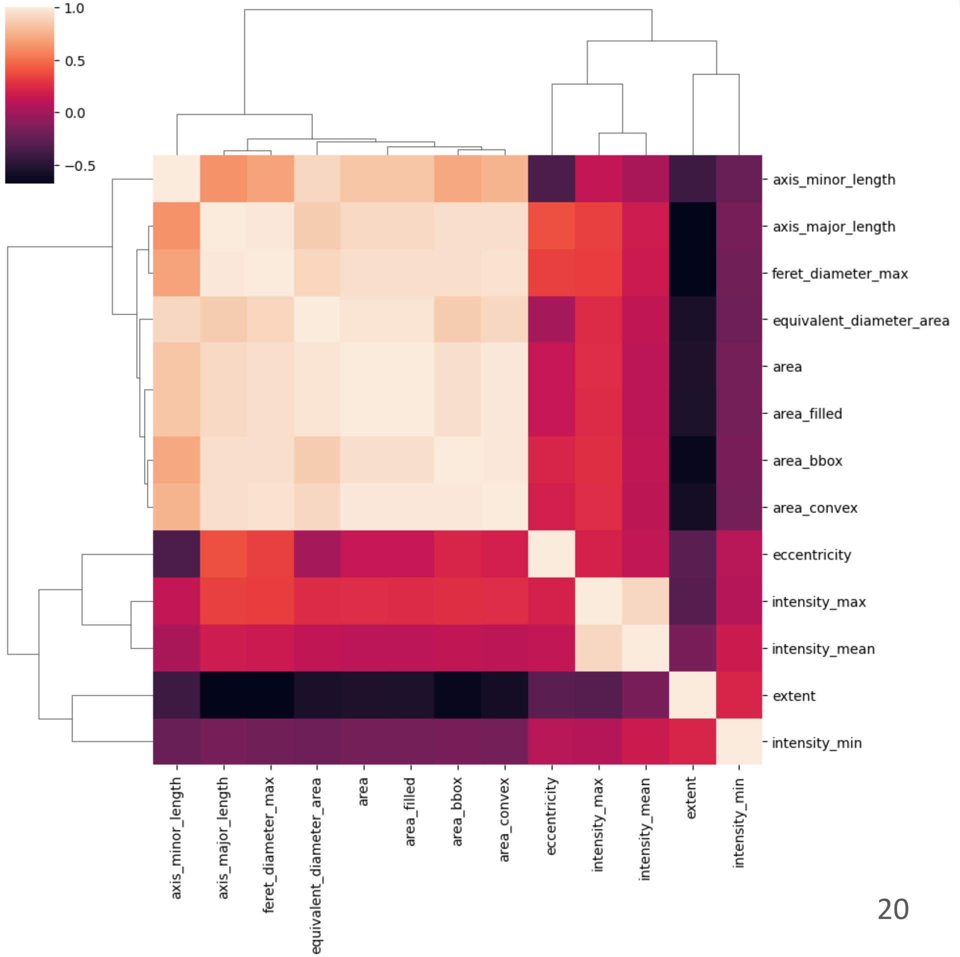
- μ_x: Mean of Samples in X
- μ_x: Mean of ranks of samples in X

Applications:





```
properties=['area', 'area_bbox', 'area_convex',
    'area_filled', 'axis_major_length',
    'axis_minor_length', 'eccentricity',
    'equivalent_diameter_area', 'extent',
    'feret_diameter_max', 'intensity_max',
    'intensity_mean', 'intensity_min'])
```



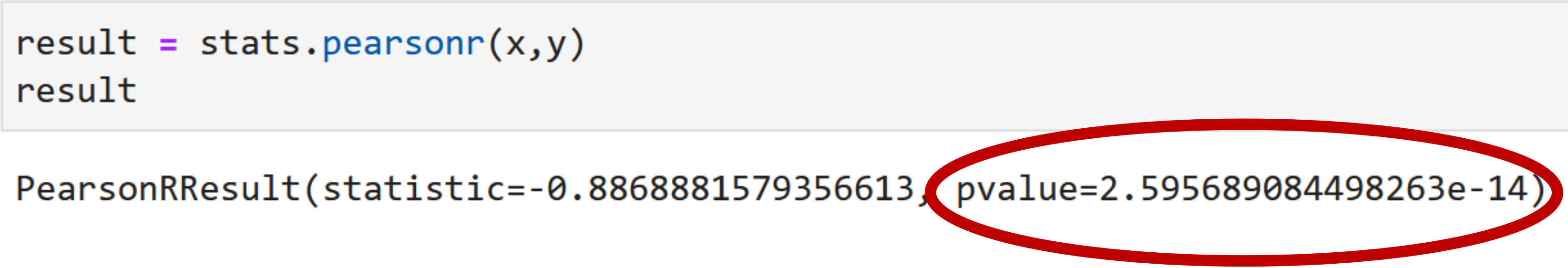
Feature selection: Measuring many features usually brings along some redundancies

→Use the correlation coefficient to remove or group such features

- →Create meta-feature (linear combination, mean, etc.) from correlating features (scaling!)
- →Pick one

→Downstream analysis works better with fewer, relevant features

p-values: Used to reject a null hypothesis H_0



```
result = stats.pearsonr(x,y)
result
```

```
PearsonRResult(statistic=-0.886881579356613, pvalue=2.595689084498263e-14)
```

P-values: Probability that the **null hypothesis H_0** is true, but rejected by chance

General: It is (usually) much easier to falsify a statement than proving it true

Example 1: $x^n + y^n = z^n$ for $n \geq 3$ and $x, y, z \in Z$

→ this took 358 years to prove – If we could have found just a single combination of x,y & z, we would immediately be done

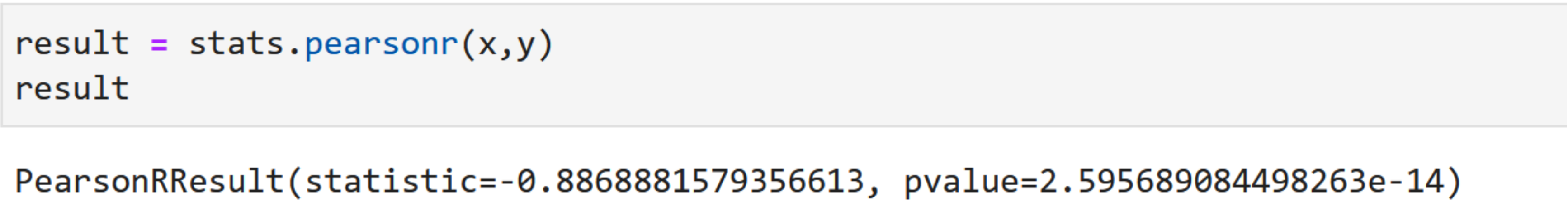
Example 2: Albert Hammond (1972): *It never rains in southern california*

→ Very hard to prove – very easy to disprove

H_0 : A treatment is ineffective/there is no difference between two groups/cell fate is not correlated to feature$_x$

https://en.wikipedia.org/wiki/Fermat%27s_Last_Theorem

In the context of correlation

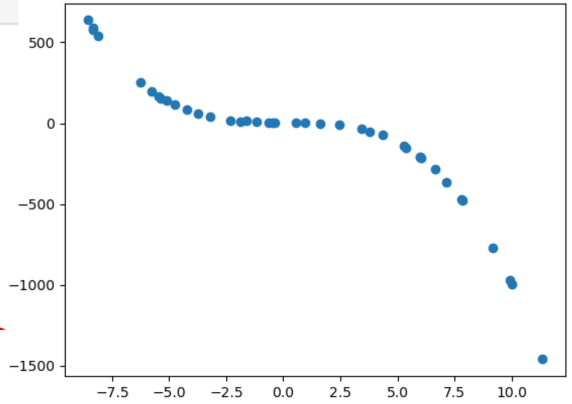


```
result = stats.pearsonr(x,y)
result
```

```
PearsonRResult(statistic=-0.886881579356613, pvalue=2.595689084498263e-14)
```

H_0 hypothesis: correlation coefficient r=0

False positive



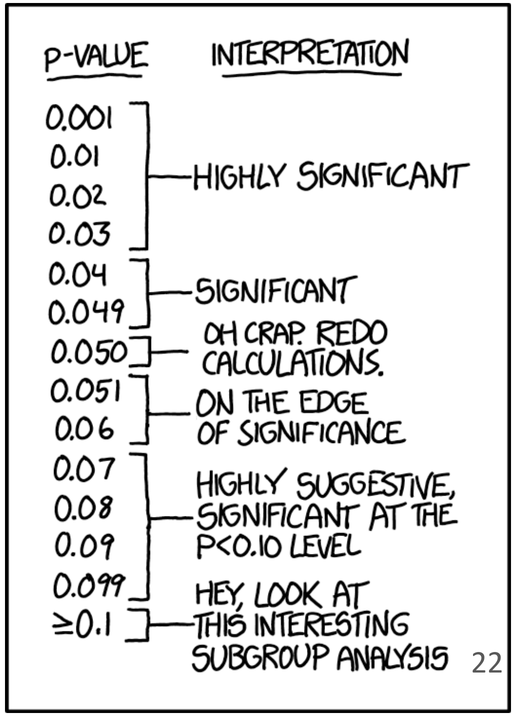
P-value: Probability that correlation coefficient r≠0 although r=0
“We just happened to draw an unfortunate selection of points from our data that looked like correlation – the odds of this happening was p”

How small should the p-value be to confidently reject H_0? → alpha-value

→ Don’t set a threshold – just report

→ Some pleasant number (0.05, 0.001, etc.

→ A common value in the field (0.05, 5σ, etc.)

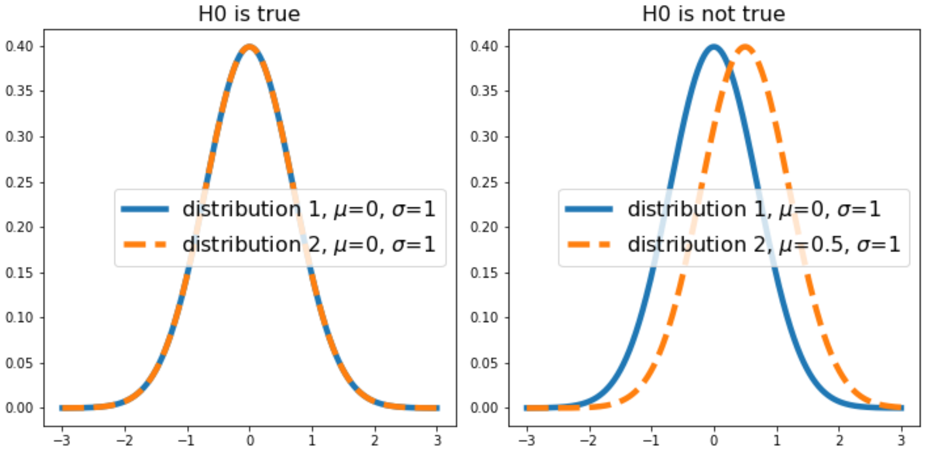


P-VALUE	INTERPRETATION
0.001 0.01 0.02 0.03	HIGHLY SIGNIFICANT
0.04 0.049	SIGNIFICANT
0.050	OH CRAP. REDO CALCULATIONS.
0.051 0.06	ON THE EDGE OF SIGNIFICANCE
0.07 0.08 0.09 0.099	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

https://xkcd.com/1478/, shared under CC BY-NC 2.5

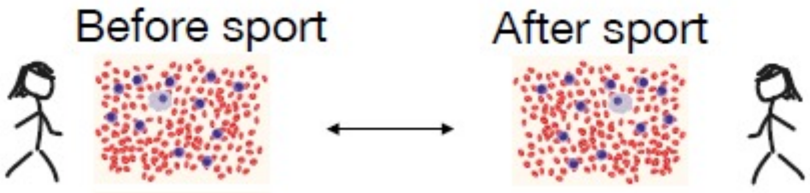
Other tests



Comparing two (normal) distributions

→ Unpaired t-test (H0: The means are different)

→ Paired t-test (H0: $X_{after} - X_{before} = 0$)



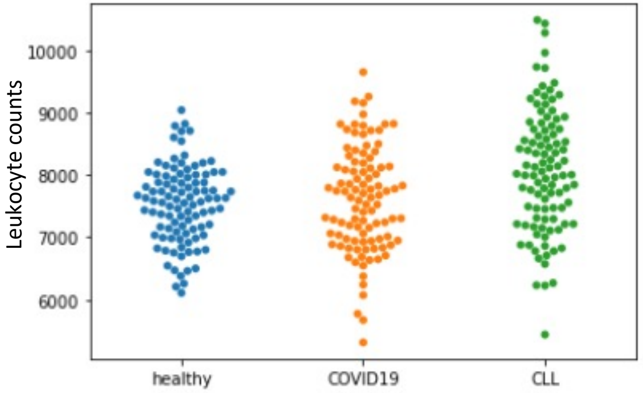
Alternative: Wilcoxon-Mann-Whitney-Test if assumptions are violated

Many observations don't follow normal distributions

→ (Cell) count data: Poisson distribution

→ Binary outcomes (e.g., coin flip): Binomial distribution

→ Each provides appropriate tests



Comparing multiple groups:

→ ANOVA (analysis of variances), H0: No differences between distributions

→ Requires "post-hoc" tests to find out which groups are different

Data skewed by outliers:

→ Consider comparing ranks rather than raw data

Pitfalls

Warning! p-values cannot be used to estimate the probability of your hypothesis!
A p-value of 0.01 does NOT mean that your original hypothesis (e.g. treatment is effective, there is a difference between groups, cell fate is correlated to feature x) is true with 99%

Explanation: the null hypothesis could also be false because of some other reason

More Pitfalls

Multiple testing: More tests → more type I errors (false positives)

Strategies:

1. Control family-wise error rate $\mathrm{FWER} = P\left(n_{false\ positives} \geq 1\right) = 1 - (1-\alpha)^N$

   Bonferroni correction: $\alpha_{adj} = \frac{\alpha}{N}$

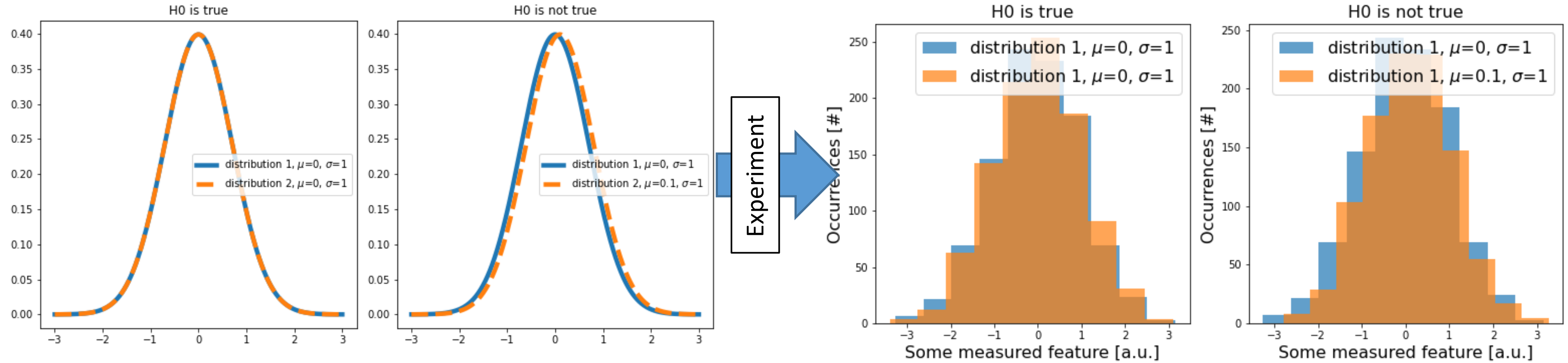
   → Prevents false positives (type I error)

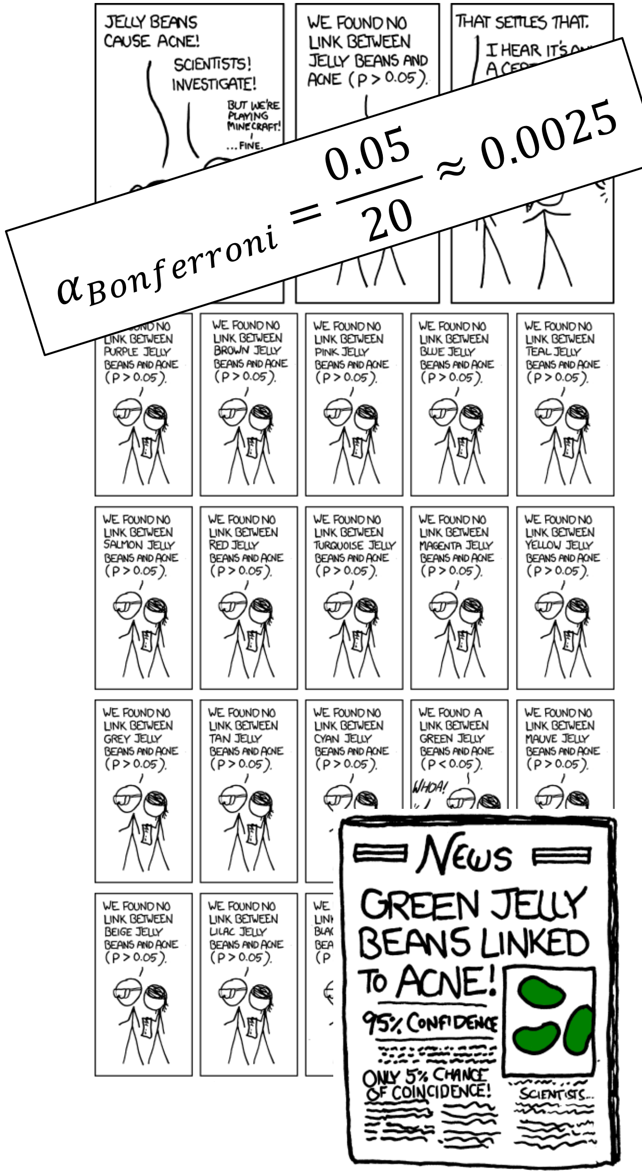
   → Introduces false negatives (type II error)

2. *Benjamini-Hochberg adjustment*: Control false discovery rate $FDR = \frac{FP}{FP+TP}$

   → Find largest k so that $p_k \leq \frac{k}{m}\alpha$ (p_k: p-value of rank k)

3. Tukey range test: Typically done after ANOVA, controls type I errors




JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

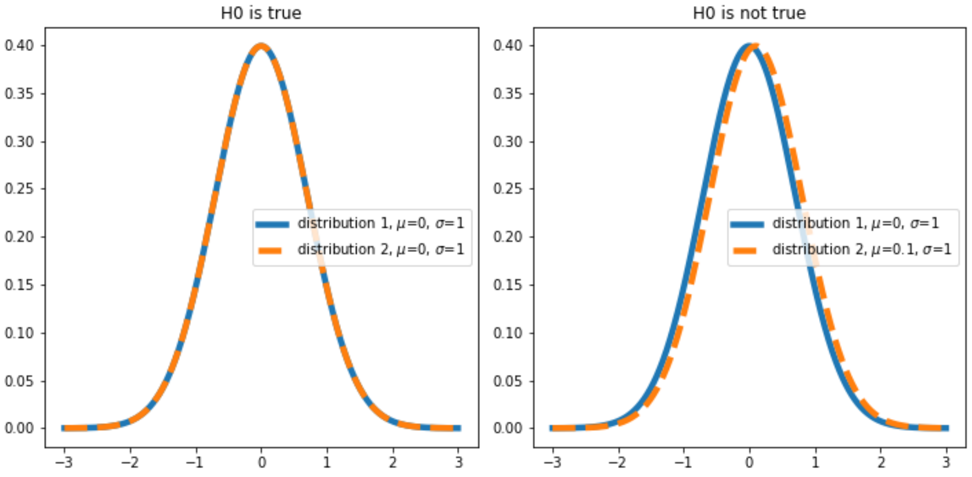
$$\alpha_{Bonferroni} = \frac{0.05}{20} \approx 0.0025$$

https://xkcd.com/882/, Shared under CC BY-NC 2.5

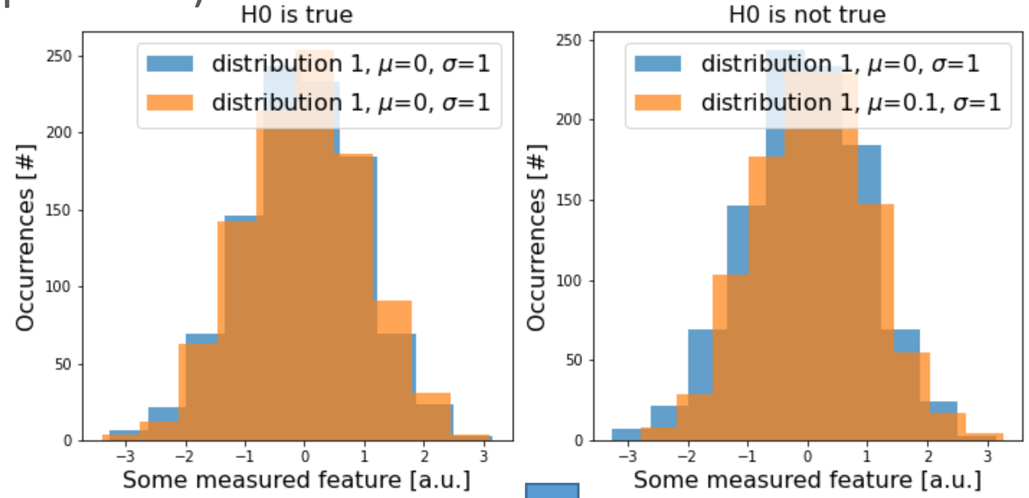
@TillKorten

More Pitfalls

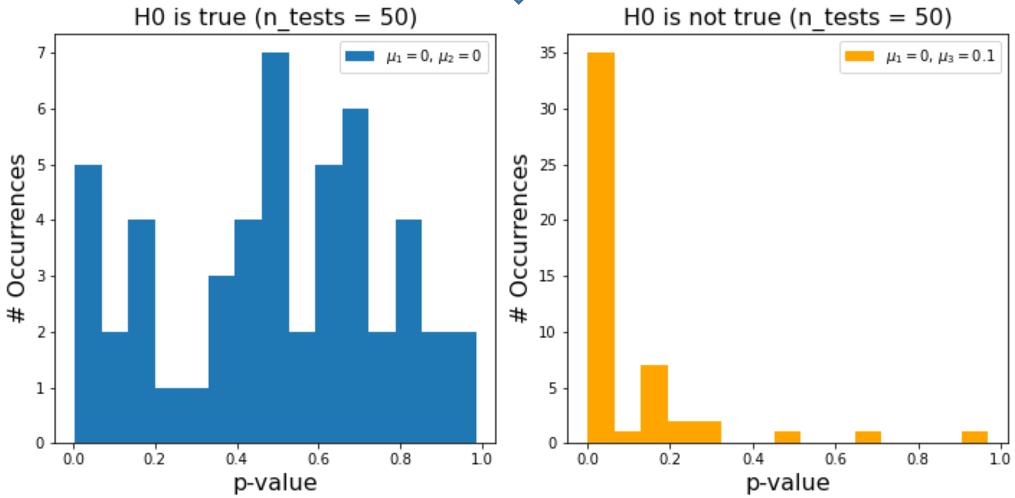
Multiple testing: More tests → more type I errors (false positives)



Experiment



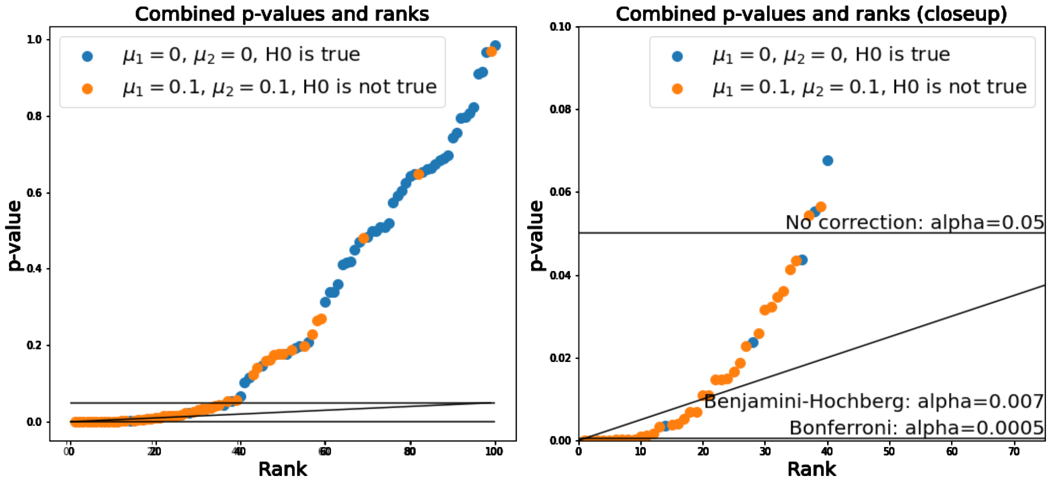
100 Experiments



Rank p-values

Multiple testing correction:

Separate cases (H0 true/false) in this plot:



Even More Pitfalls

Multiple testing: More tests → more type I errors (false positives)

Distribution type: T-test assumes normal distribution of data
→ Some data may follow different distributions (Poisson, binomial, etc.)
→ The equivalent for a t-test exists for all other distributions, too!
→ Less strict test types exist – ask your statistician!

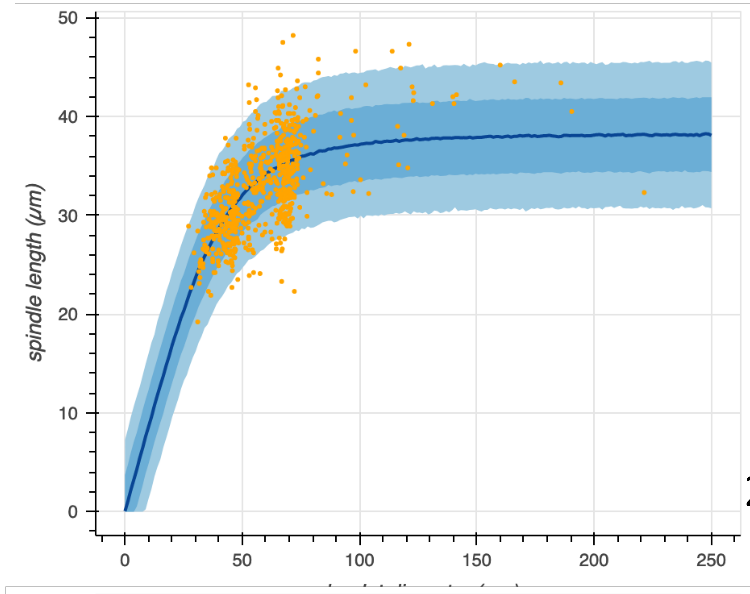
Sample size: Do not perform statistical test with small (n < 10) sample sizes.
→ If you work in this region (experiments expensive, animals, etc): Consult your local statistician!

Sample independence:
→ T-tests are only valid if samples are independent: "*Two events are independent […] if […] the occurrence of one does not affect the probability of occurrence of the other*"
Examples:
Histological slices from same animal: Not independent
Same blood test derived from two patients: Independent

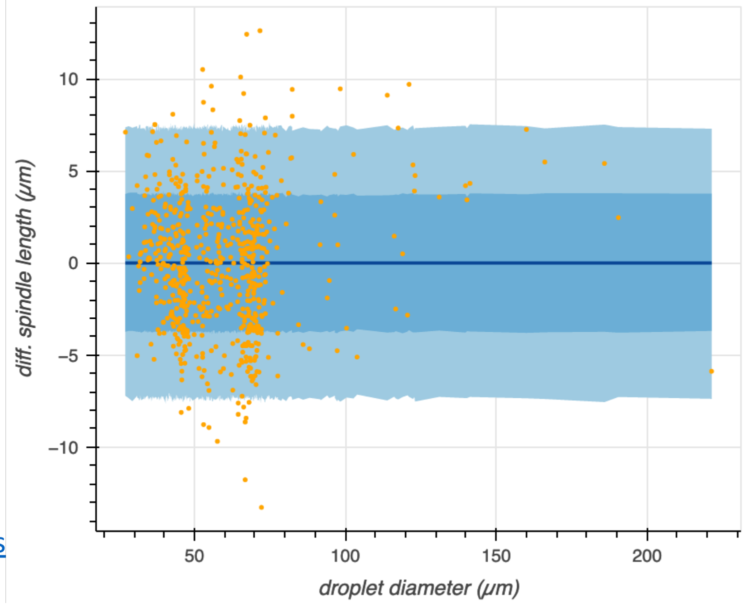
https://en.wikipedia.org/wiki/Independence_(probability_theory)

Trust Confidence Intervals Over p-Values

- p-values are less reliable than suggested by their numeric value – particularly when used for detecting significant differences (sometimes termed “The Statistical Crisis in Science”[1])
- p-values do not have good graphical representations – confidence intervals do
- Avoid basing scientific conclusions or decisions about experiments solely on p-values
- In many cases, it is better to calculate confidence intervals by bootstrapping and define non-overlapping confidence intervals as significant
- What is more trustworthy? The graphical representation of a model fit to the data to the right or a p-value of 10^{-14}?



[2]



[2]

1) Andrew Gelman, Eric Loken https://www.americanscientist.org/article/the-statistical-crisis-in-science
2) Justin Bois https://justinbois.github.io/dd-pol/2022/lessons/04/confidence_intervals.html
Creative Commons Attribution License CC-BY 4.0

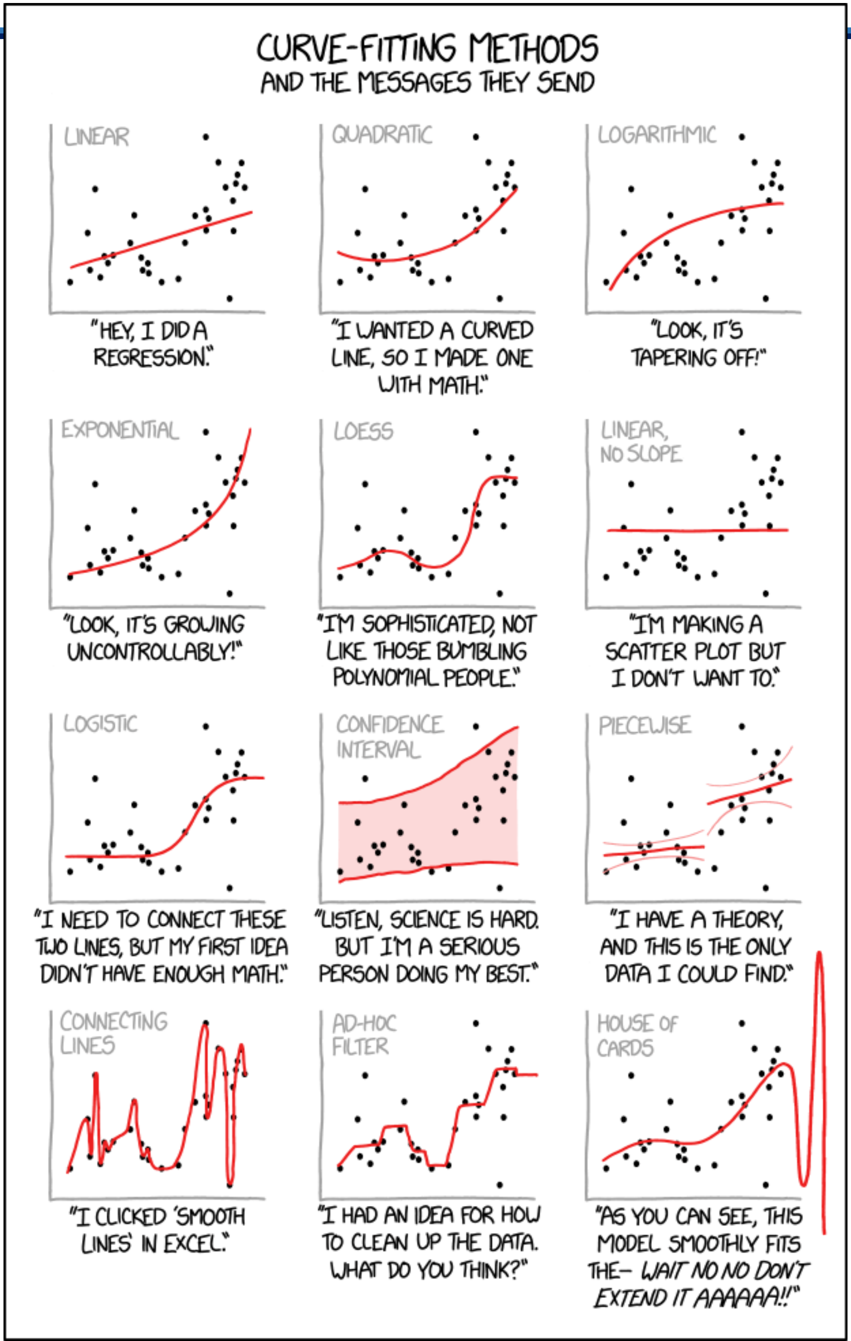
A small p-value indicates….

A big difference between datasets

Small probability of false positives

Small standard deviations of the compared groups

Not much, it's just a number that some reviewers like to see


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

https://xkcd.com/2048/