Method validation & correlation

Johannes Müller

With material from

Robert Haase

Anna Poetsch

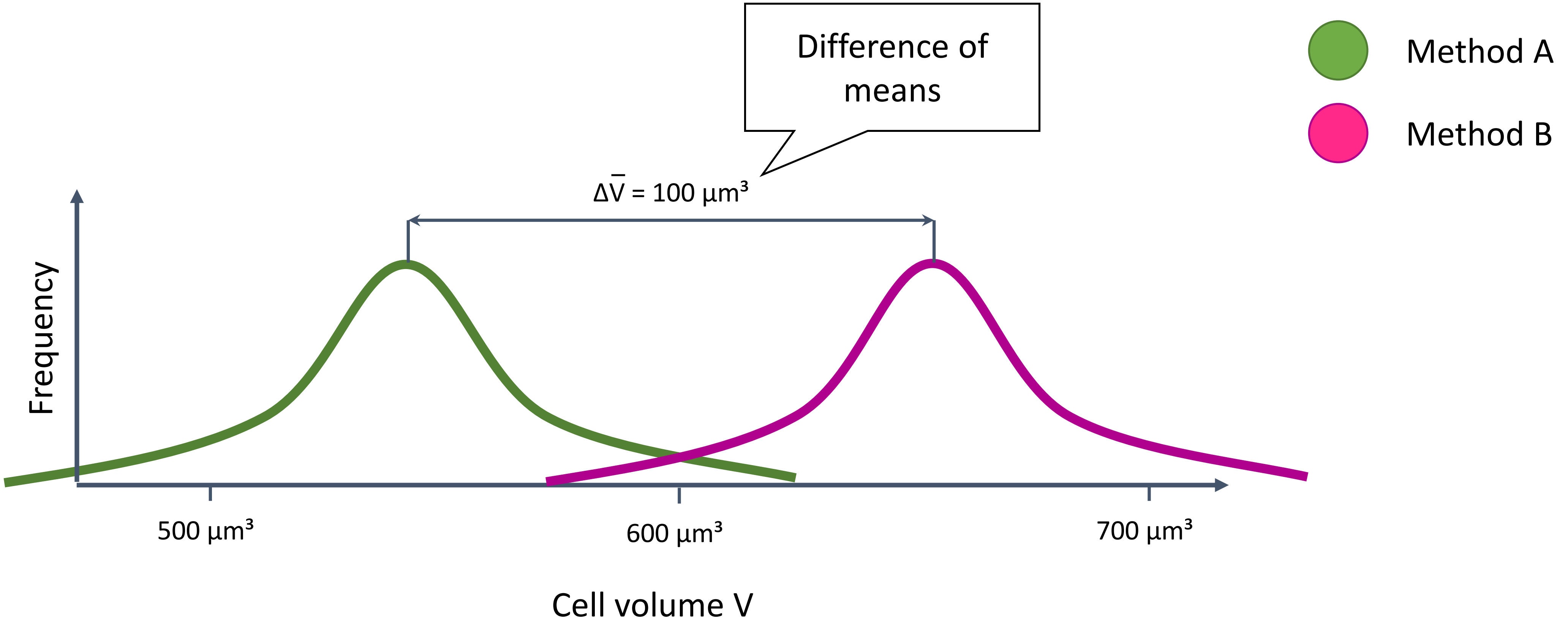
Martin J. Bland

Douglas G. Altman

@jm_mightypirate

Comparison of means

- Comparing mean measurements appears reasonable on the first view.



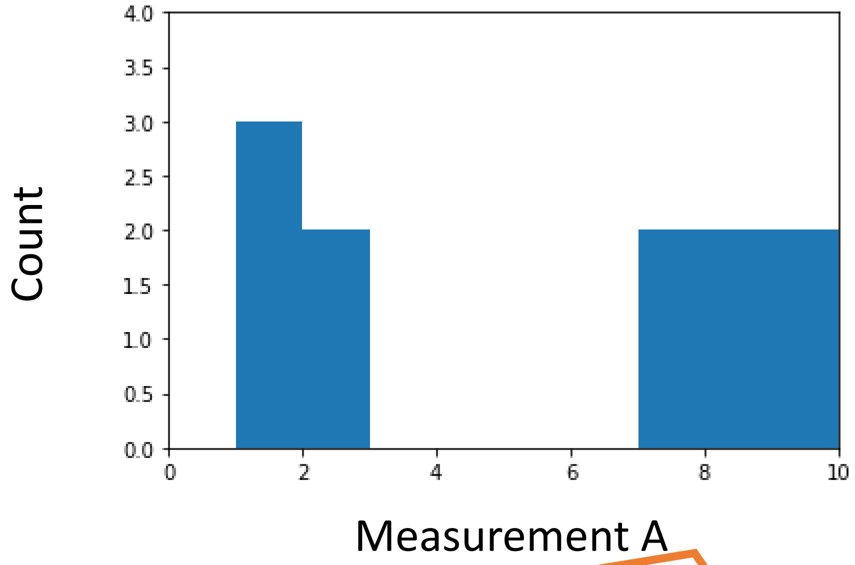
Comparison of means

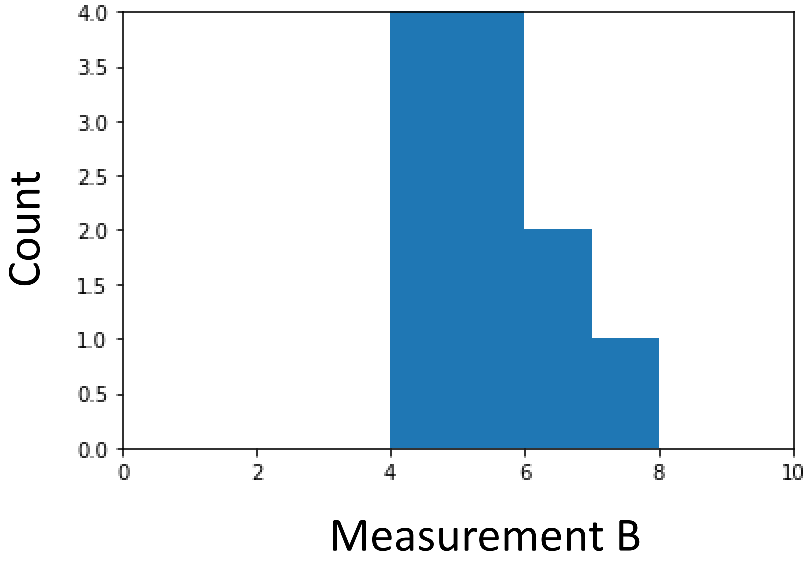
- Are two methods doing the same if their mean measurement is similar?

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
```

- Draw histograms! How can two methods do the same if histograms from their measurements are different?





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

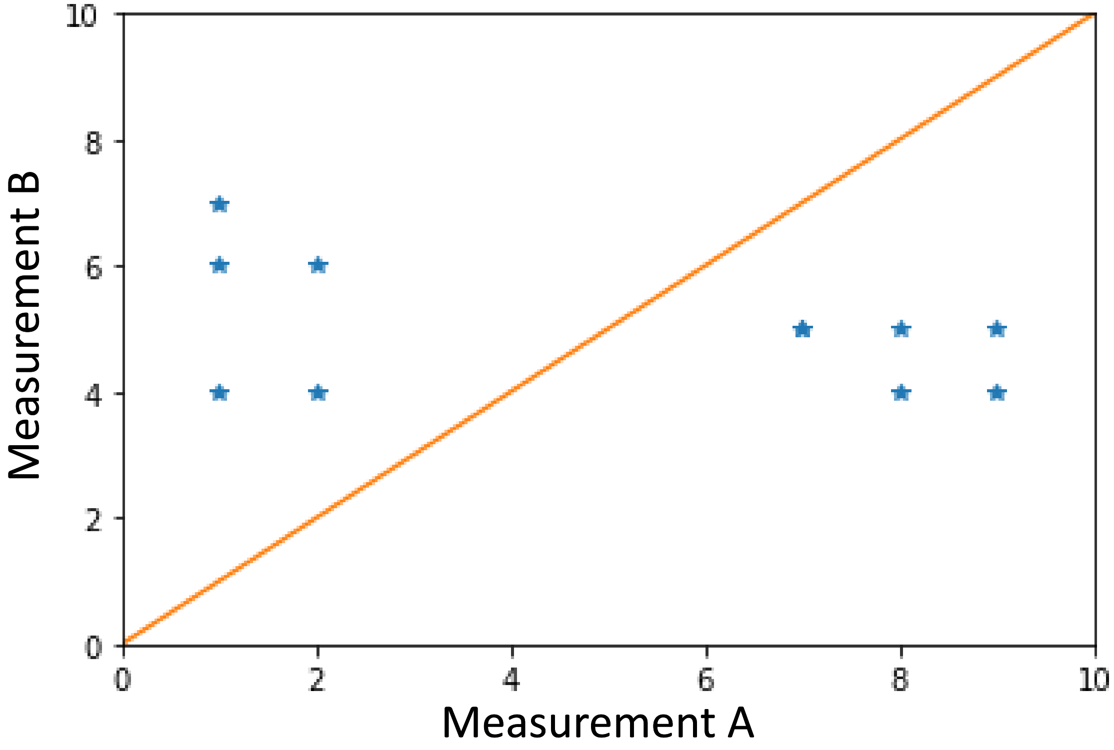
Comparison of means

- Are two methods doing the same if their mean measurement 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
```

- Plot the measurements against each other. What does it mean if they lie on a straight line? What if not?

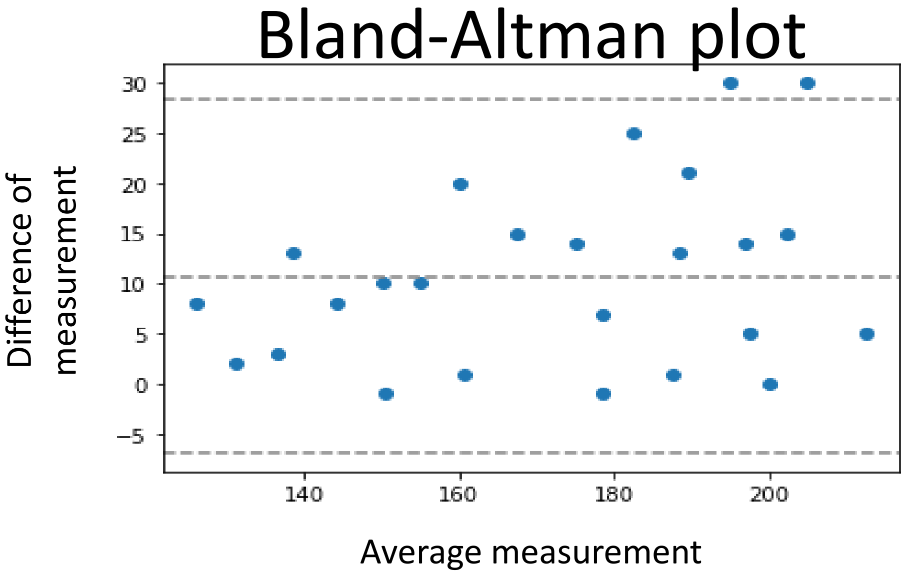


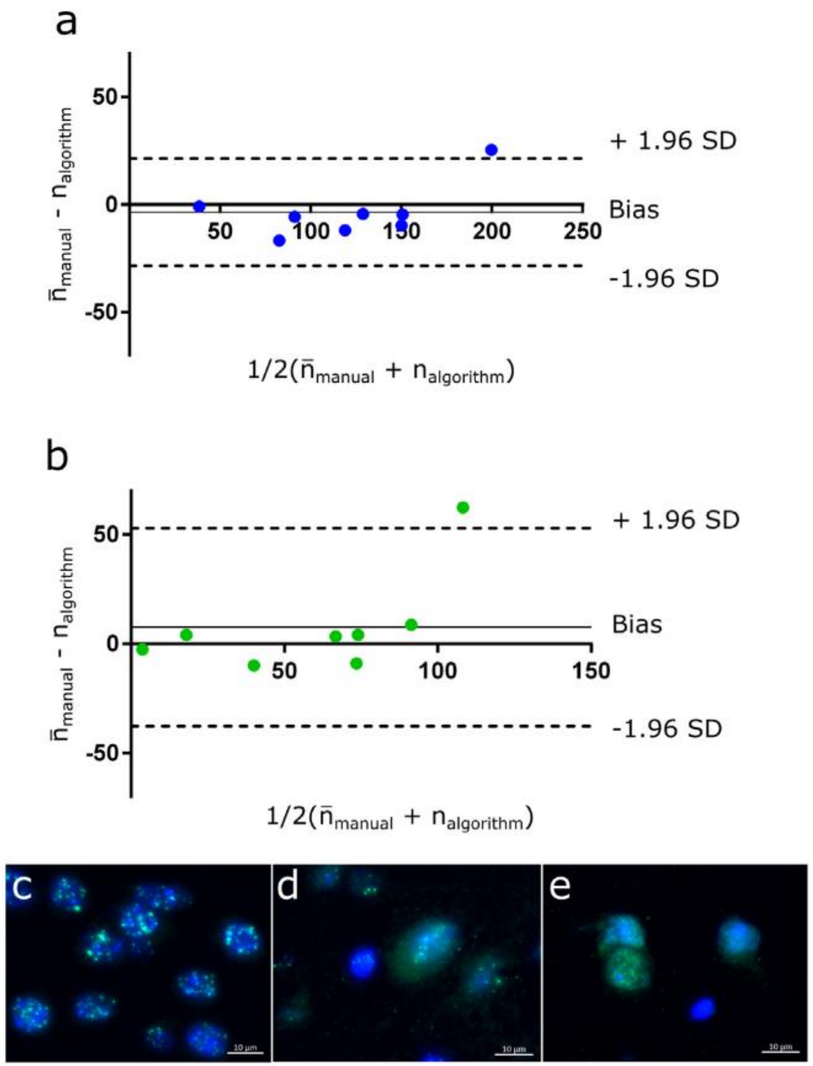
The "criterion of agreement was that the two methods gave the same mean measurement; 'the same' appears to stand for 'not significantly different'. Clearly, this approach tells us very little about the accuracy of the methods."[1]

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

Bland-Altman plots

- **Practical application:**
  - Script counts cell damage (gH2AX expression) in nucleus
  - Damage and nuclei are counted independently and require settings some parameters
  - Bland-Altman analysis used to validate parameters



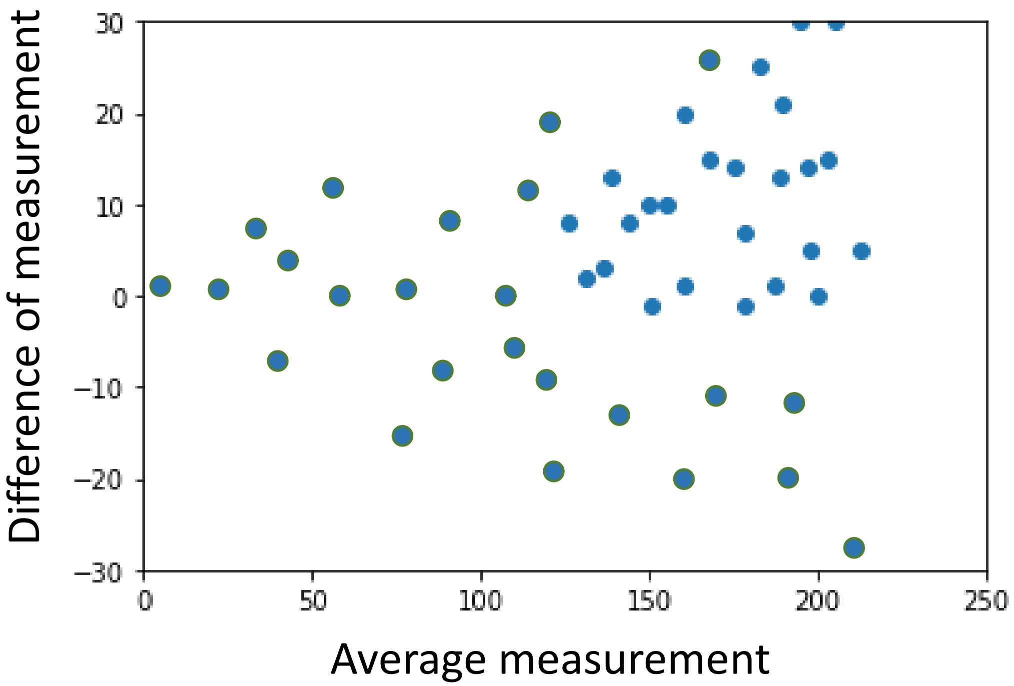


Suckert, Beyreuther, Müller et al. Rad. Onc. (2020)

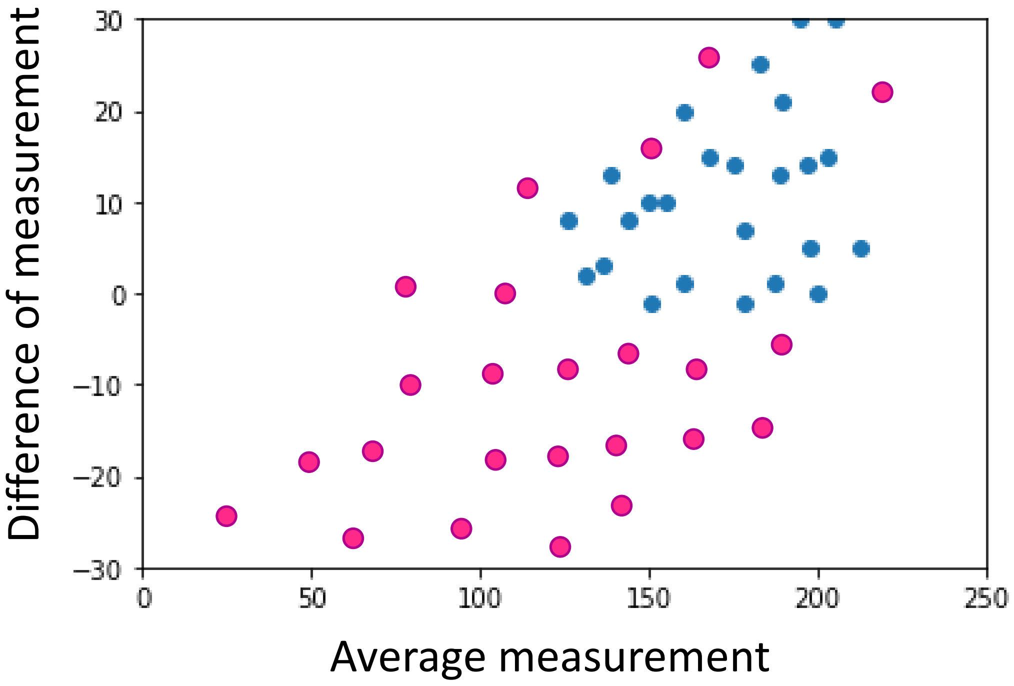
In practice:

- Bland-Altman plots allow us to differentiate various kinds of bias.

Agreement with random relative error



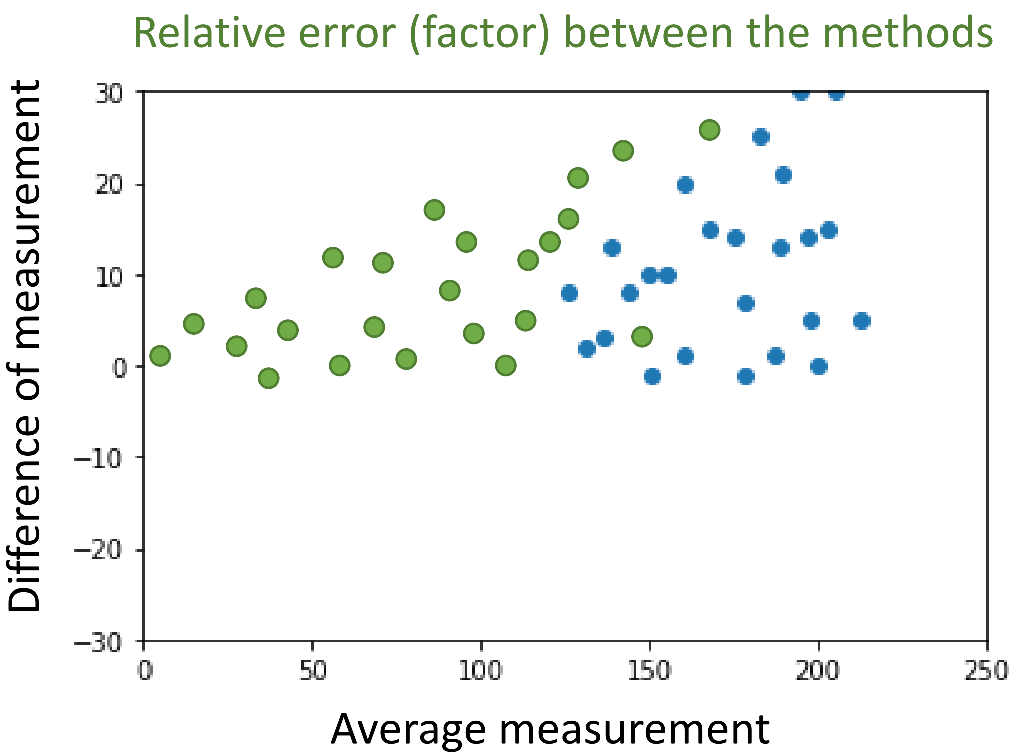
Systematic bias

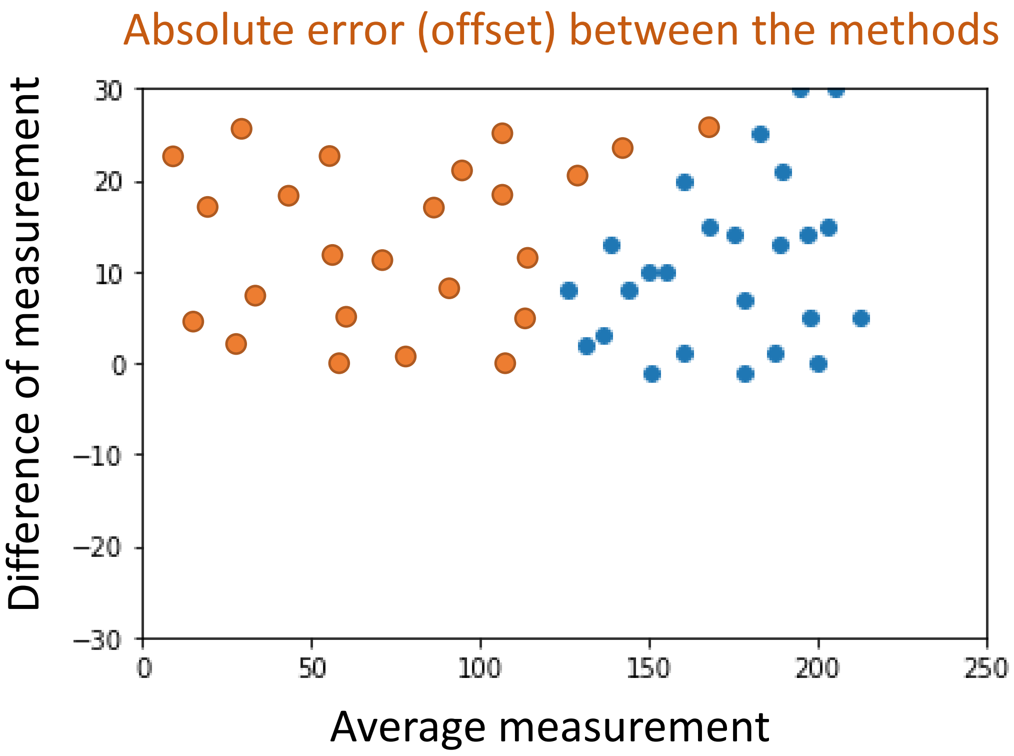


- Both distributions could have the same mean difference and confidence interval.

In practice:

- Bland-Altman plots allow us to differentiate various kinds of bias.





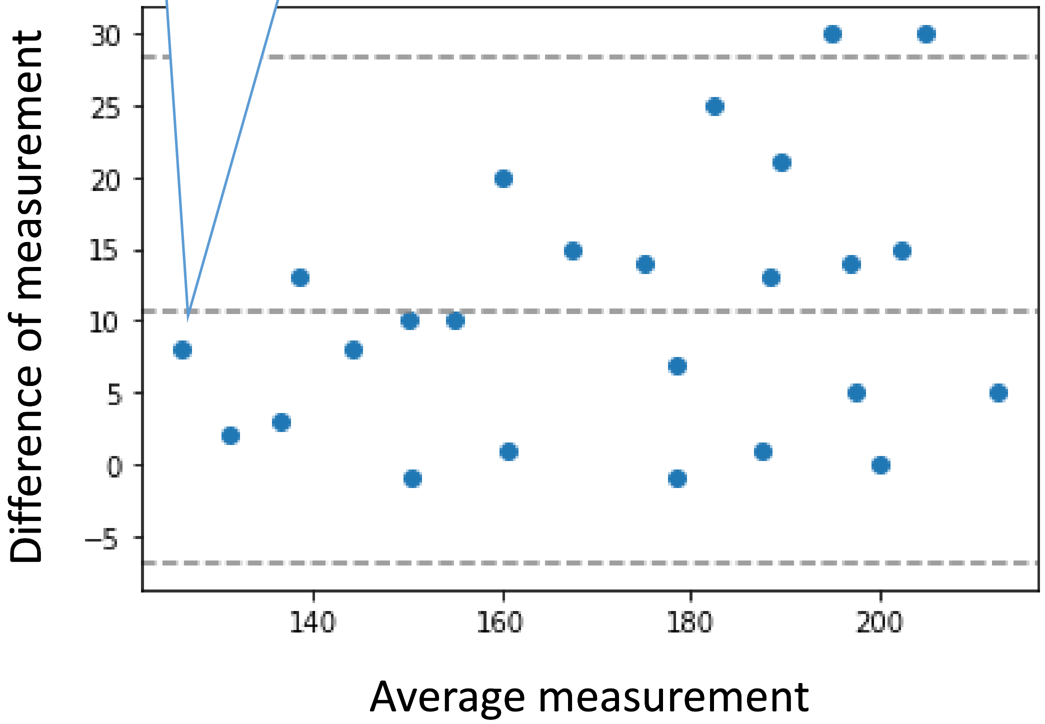
- Both effects can be corrected by calibration.

The confidence interval

- “The British Standards Institution (1979) define a coefficient of repeatability as ‘the value below which the difference between two single test results ... may be expected to lie with a specified probability; in the absence of other indications, the probability is 95 per cent’.”[1]

The mean difference between the methods

Bland-Altman plot



The confidence interval *CI* of agreement

$$CI = (\mu - 2\sigma, \mu + 2\sigma)$$

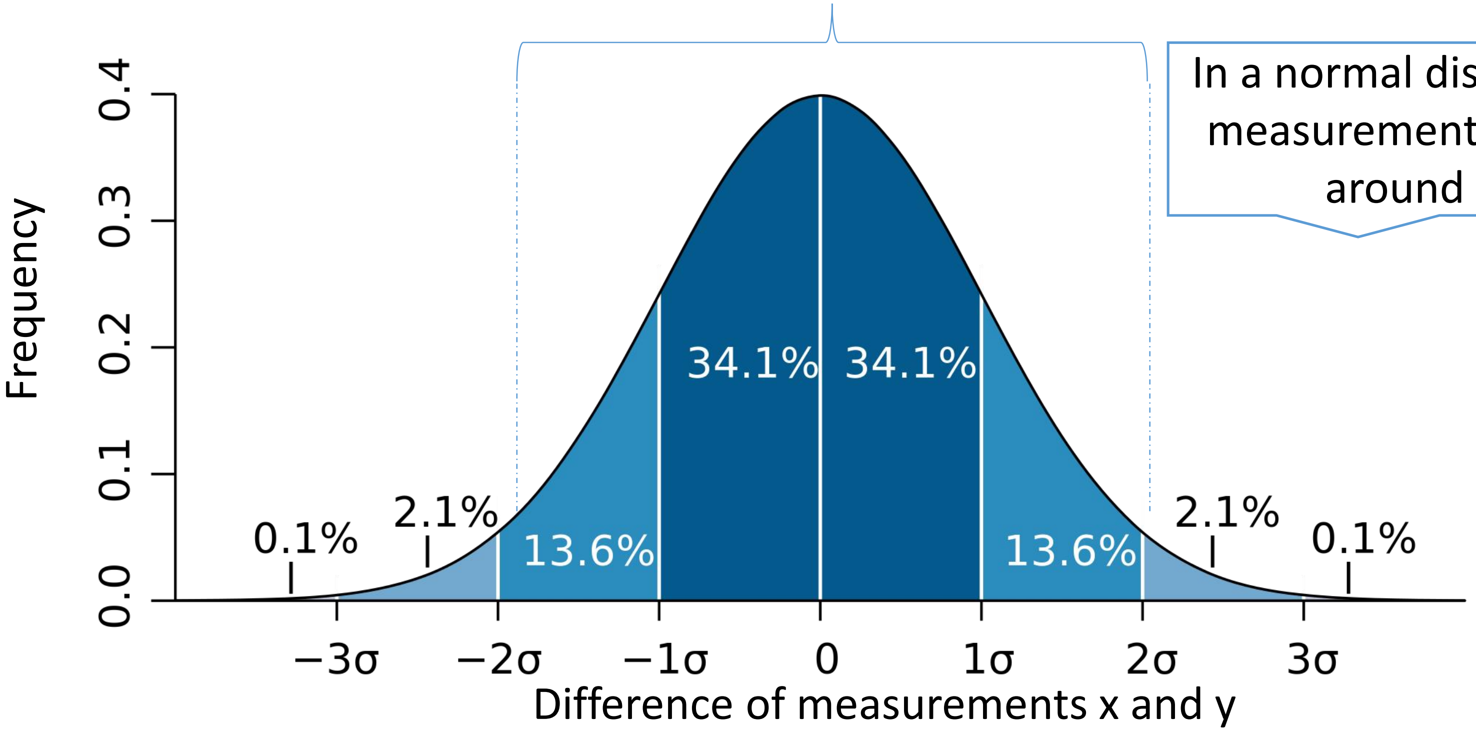
Mean difference

Standard deviation of differences

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

The confidence interval & the coefficient of repeatability.

- "The British Standards Institution (1979) define a coefficient of repeatability as 'the value below which the difference between two single test results ... may be expected to lie with a specified probability; in the absence of other indications, the probability is 95 per cent'."[1]
- If the two measurements come from the same method which just repeated twice, we can assume that the mêan difference is zero. The coefficient of repeatability *CR* can then be estimated: It's the standard deviation of differences.[2]



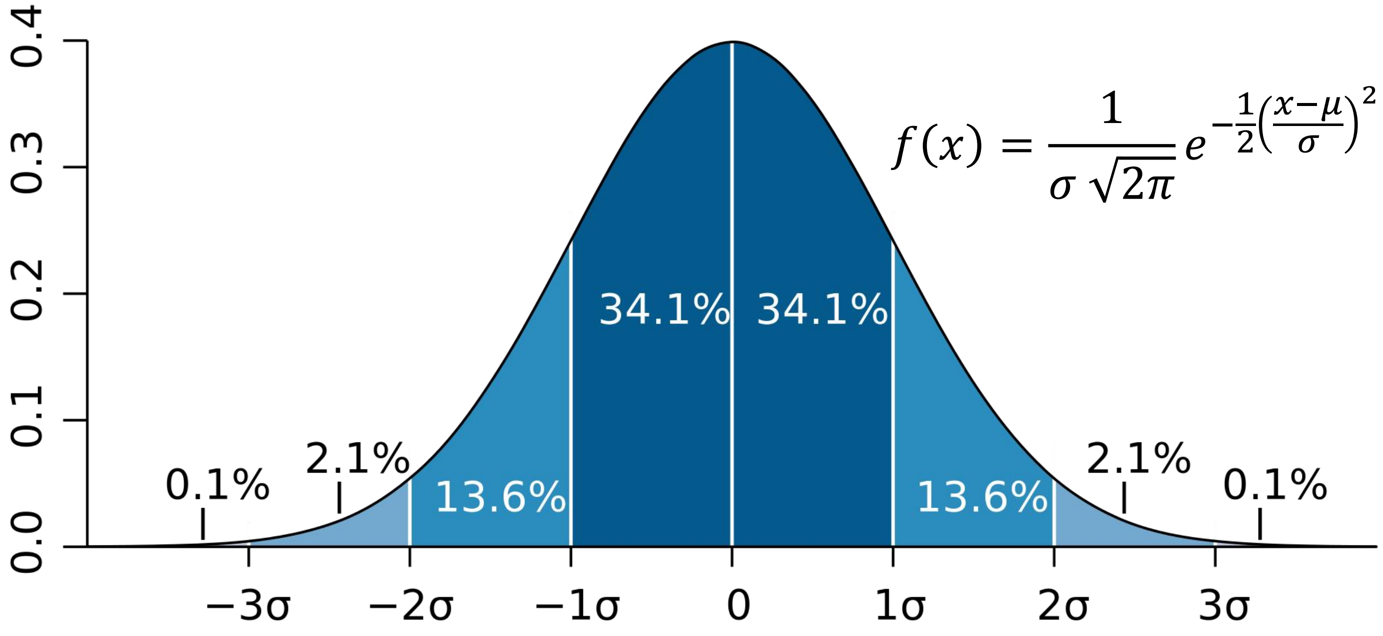
$$\mathrm{CR}(X,Y) = \sqrt{\sum_{x\in X, y\in Y} \frac{(x-y)^2}{n}}$$

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

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

2 Bland & Altman, Lancet , 1986

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 on underlying distribution

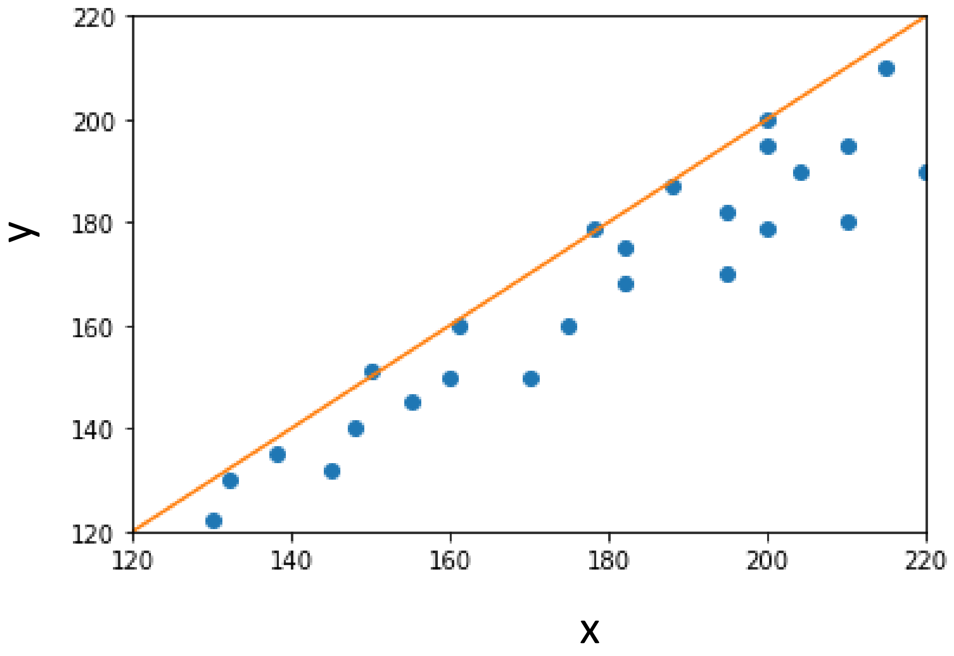
Value	Rank
10	1
15	2
3	0
97	3

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

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 correation.

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:

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

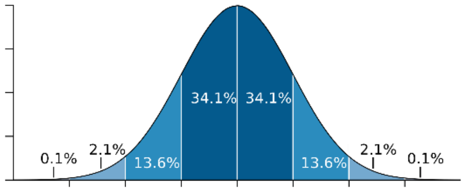
Number of measurements *n*

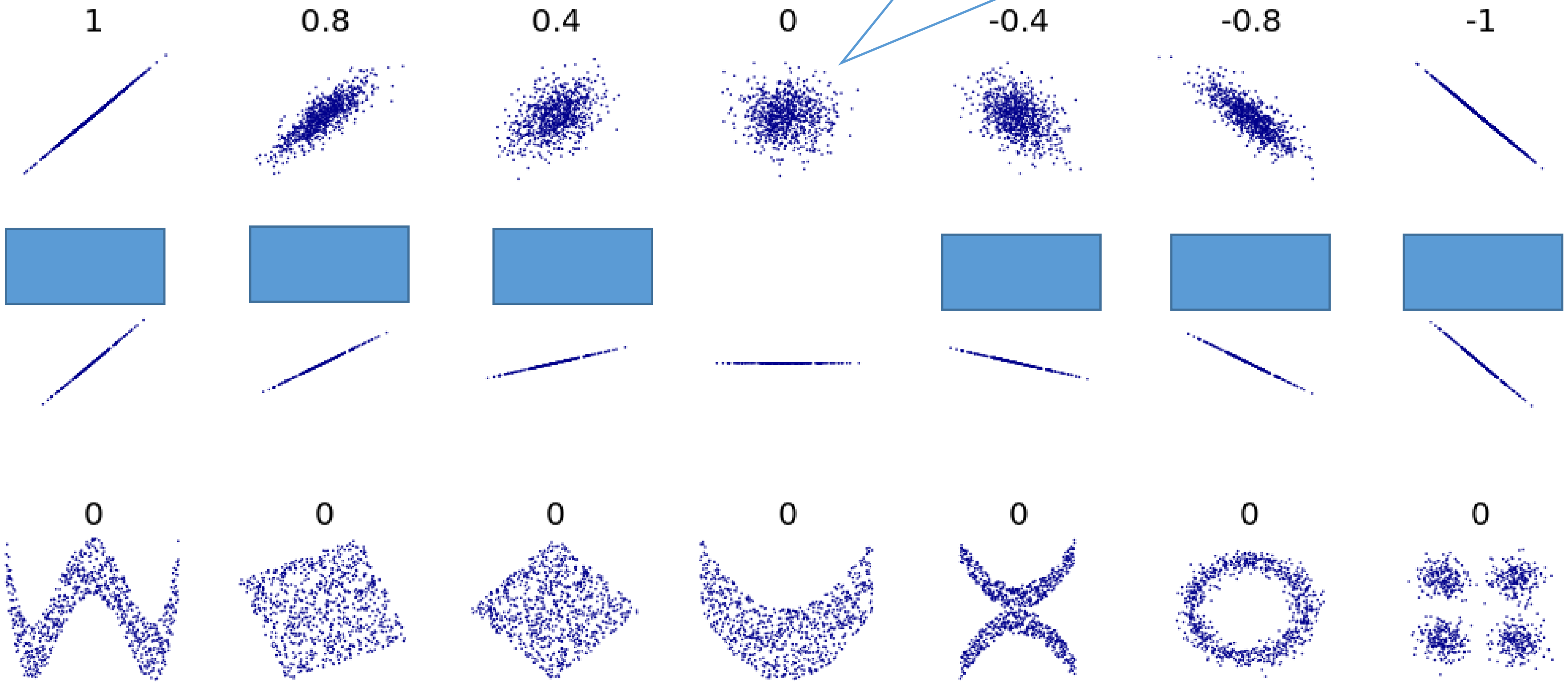
[1] Altman & Bland, The Statistician 32, 1983

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#/media/File:Correlation_examples2.svg

Correlation: Spearman's r

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

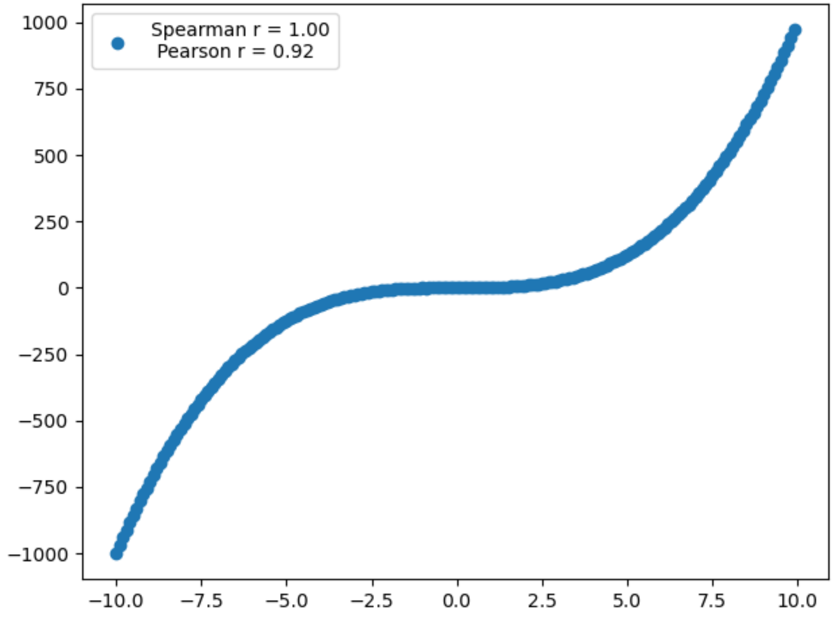
- Spearman's *r* lies between -1 and 1
  - 1: Positive **monotonous** correlation
  - 0: No monotonous correlation
  - -1: Negative monotonous correlation

Expectation E

Mean average μ

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

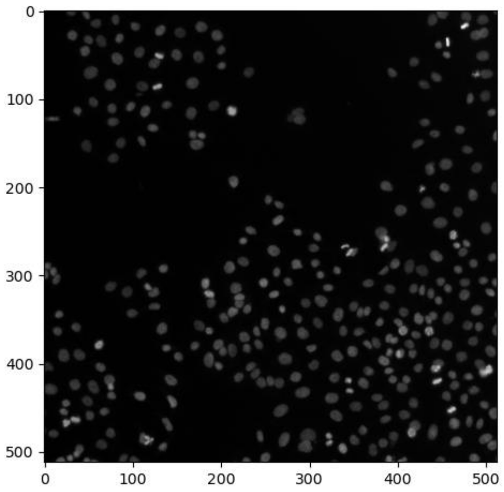
Standard deviation σ

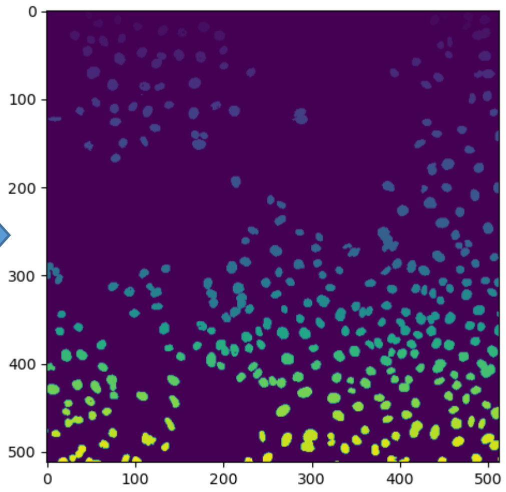


Spearman's *r* is equivalent to using Pearson's r on ranked data:

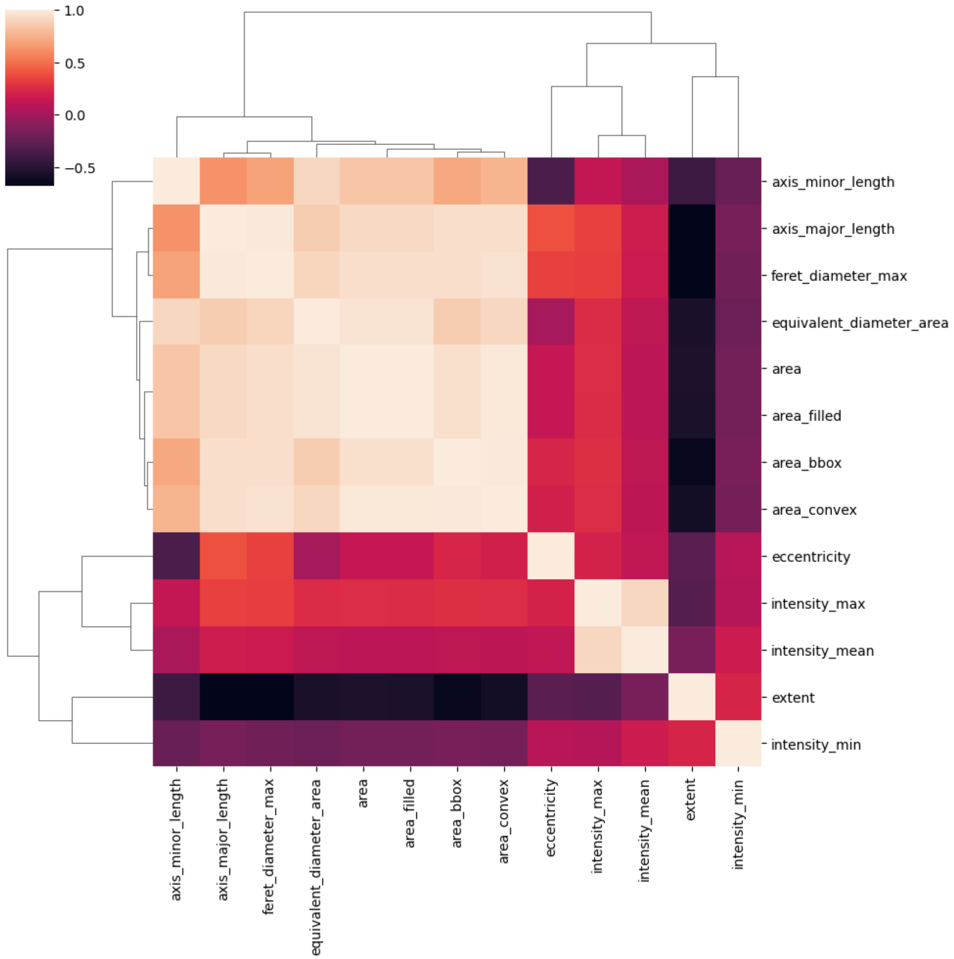
- μ_x: Mean of Samples in X
- μ_x: Mean of ranks of samples in X

Applications:






```python
, 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

Hypothesis testing

Johannes Müller

With material from
Anna Poetsch

@jm_mightypirate

We need to talk

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

```
PearsonRResult(statistic=-0.8868881579356613, 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 californa*

→ Very hard to prove – very easy to disprove

H_0 hypothesis: A treatment agent 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=-0.8868881579356613, pvalue=2.595689084498263e-14)
```

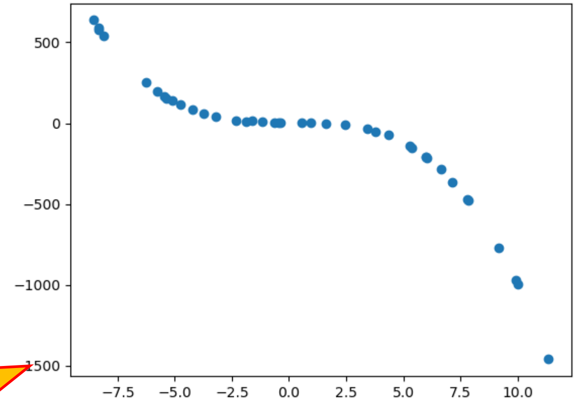
H_0 hypothesis: correlation coefficient r=0

In scipy: **alternative** : *{'two-sided', 'less', 'greater'}, optional*

Defines the alternative hypothesis. Default is 'two-sided'. The following options are available:

- 'two-sided': the correlation is nonzero H_0?
- 'less': the correlation is negative (less than zero) H_0?
- 'greater': the correlation is positive (greater than zero) H_0?

False positive



P-value: Probability that correlation coefficient r≠0 although r=0

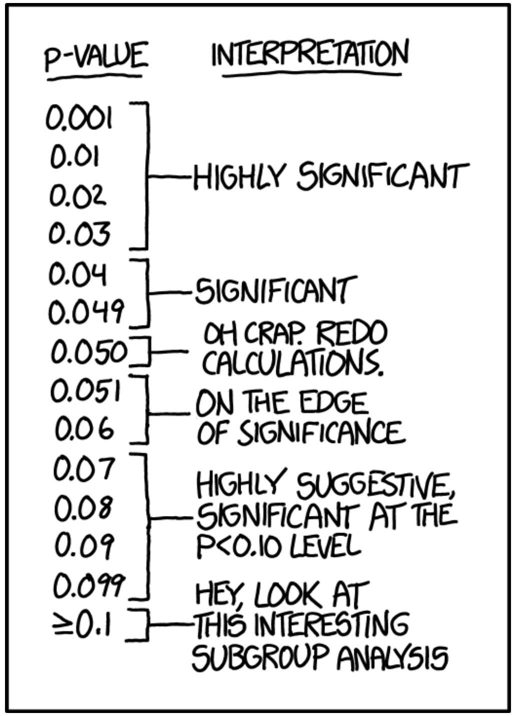
"We just happened to draw an unfortunate selection of points from our data that looked like correlation – the odds of this happening was p"

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.

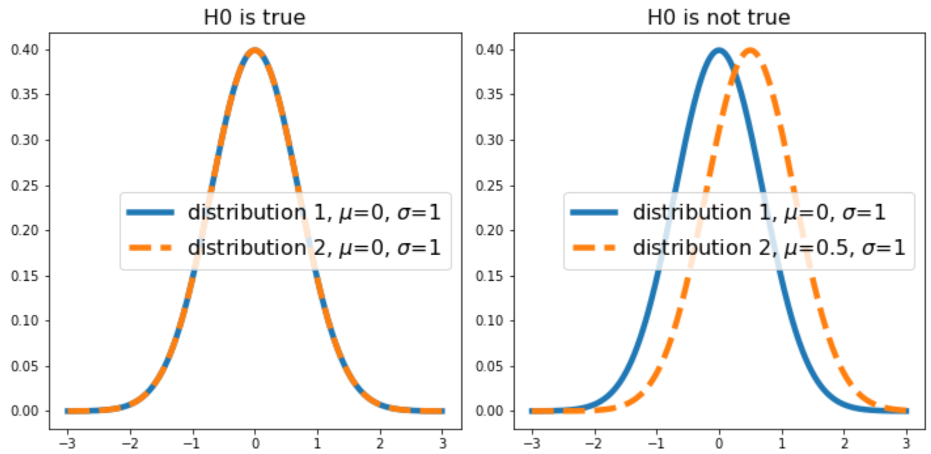
→ A common value in the field (0.05, 5σ, etc.)



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

https://xkcd.com/1478/

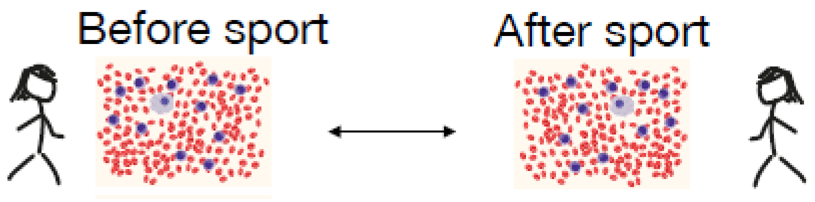
Other tests



Comparing two (normal) distributions

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

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



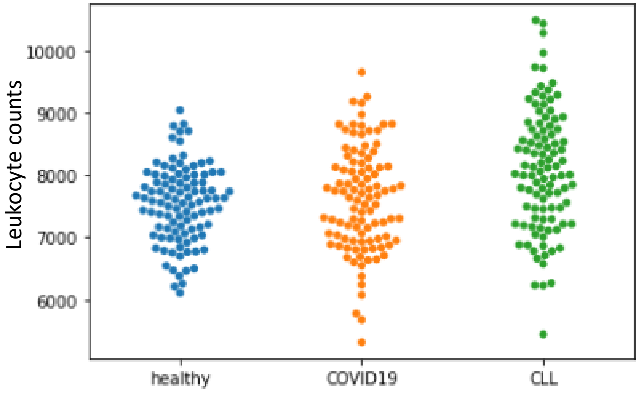
Alternative: Wilcoxon-Mann-Whitney-Test if assumptions are violated

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

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

Strategies:

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

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

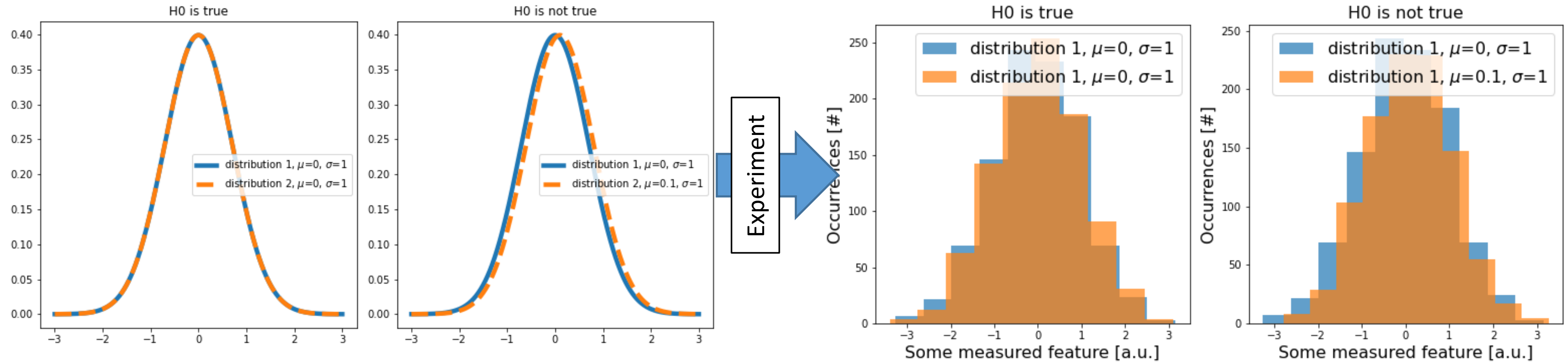
   → Prevents false positives (type I error)

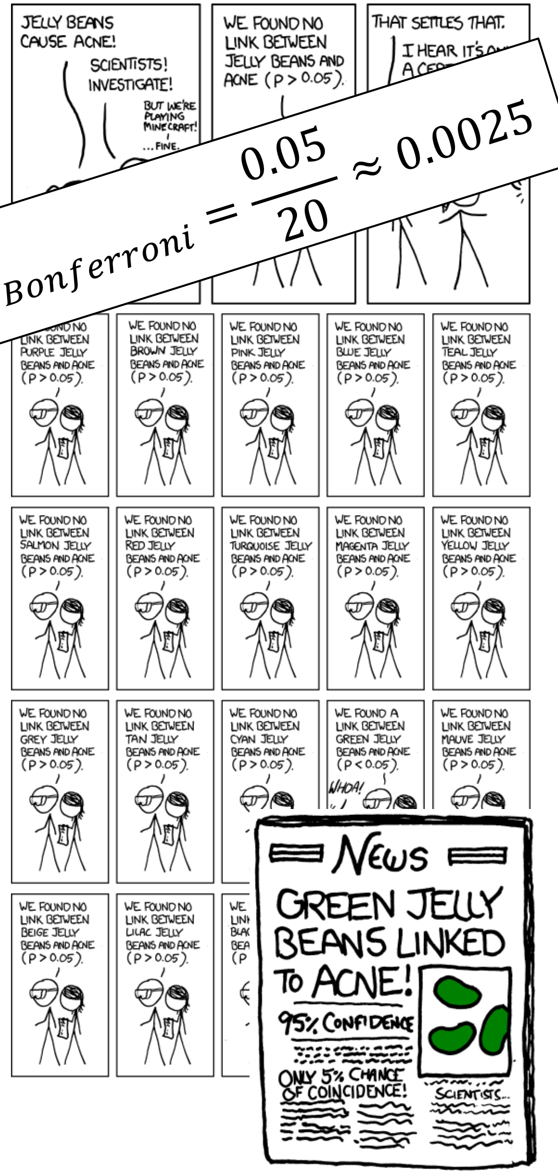
   → Introduces false negatives (type II error)

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



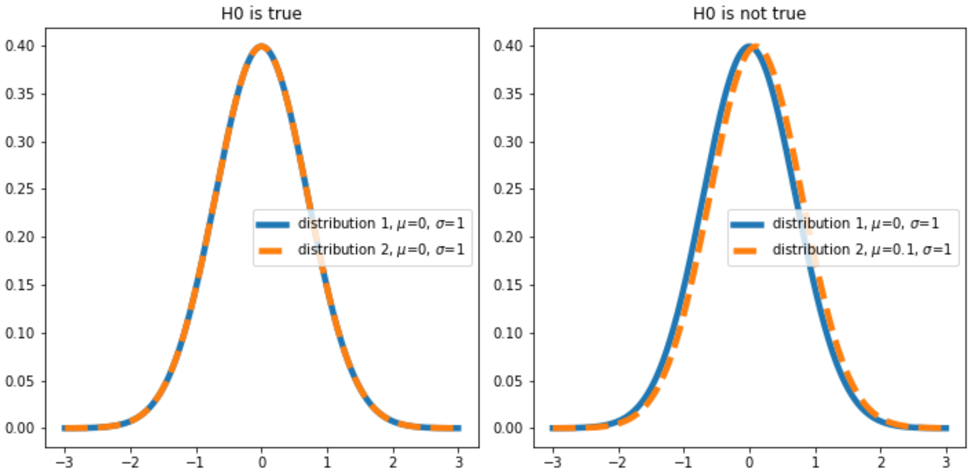


$\alpha_{Bonferroni} = \frac{0.05}{20} \approx 0.0025$

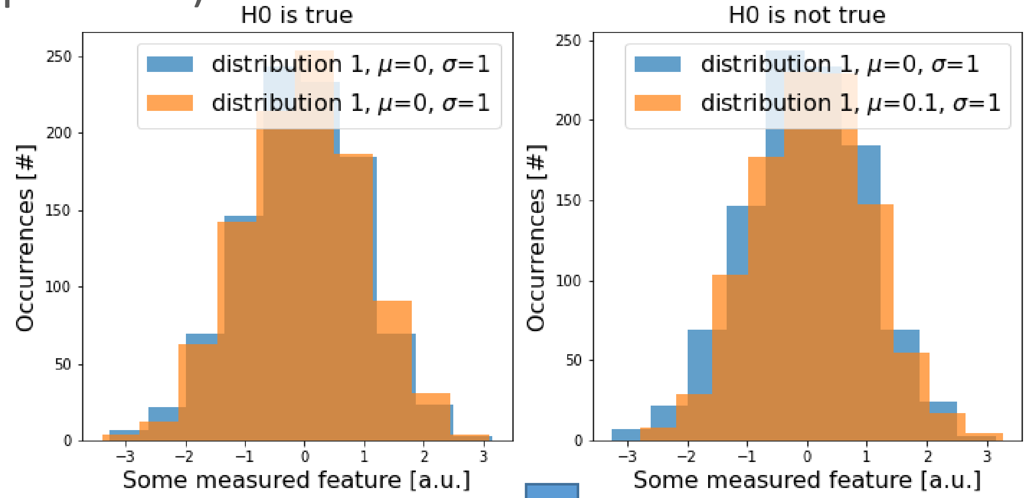
https://xkcd.com/882/

Pitfalls

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



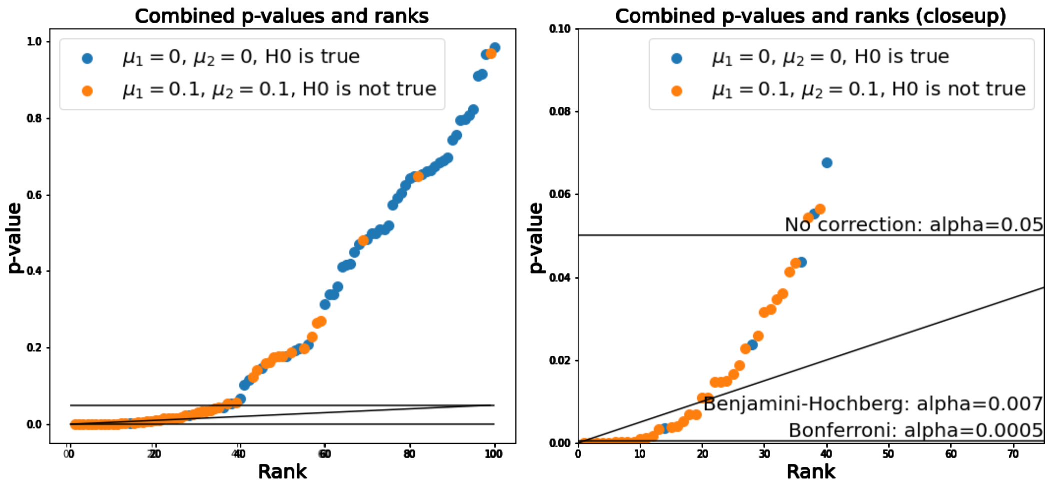
Experiment



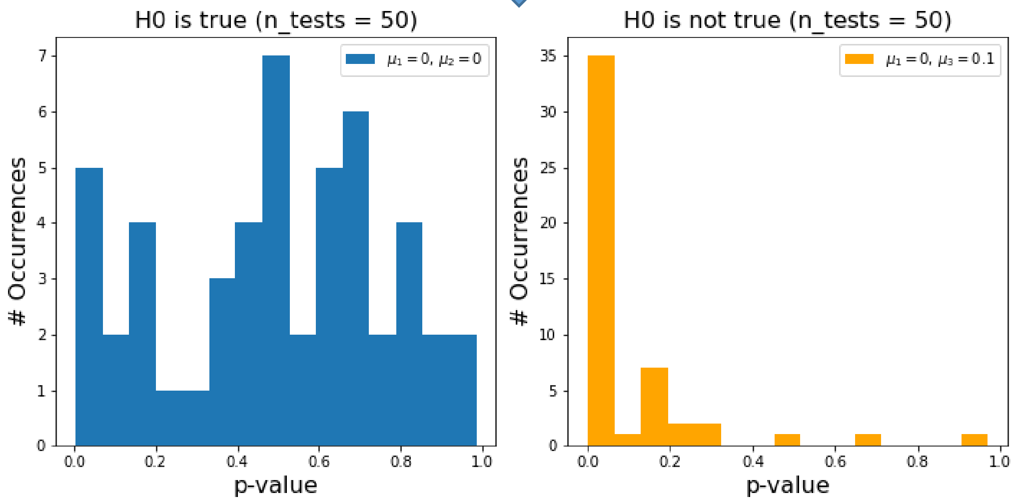
100 Experiments

Multiple testing correction:

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



Rank p-values



@jm_mightypirate

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

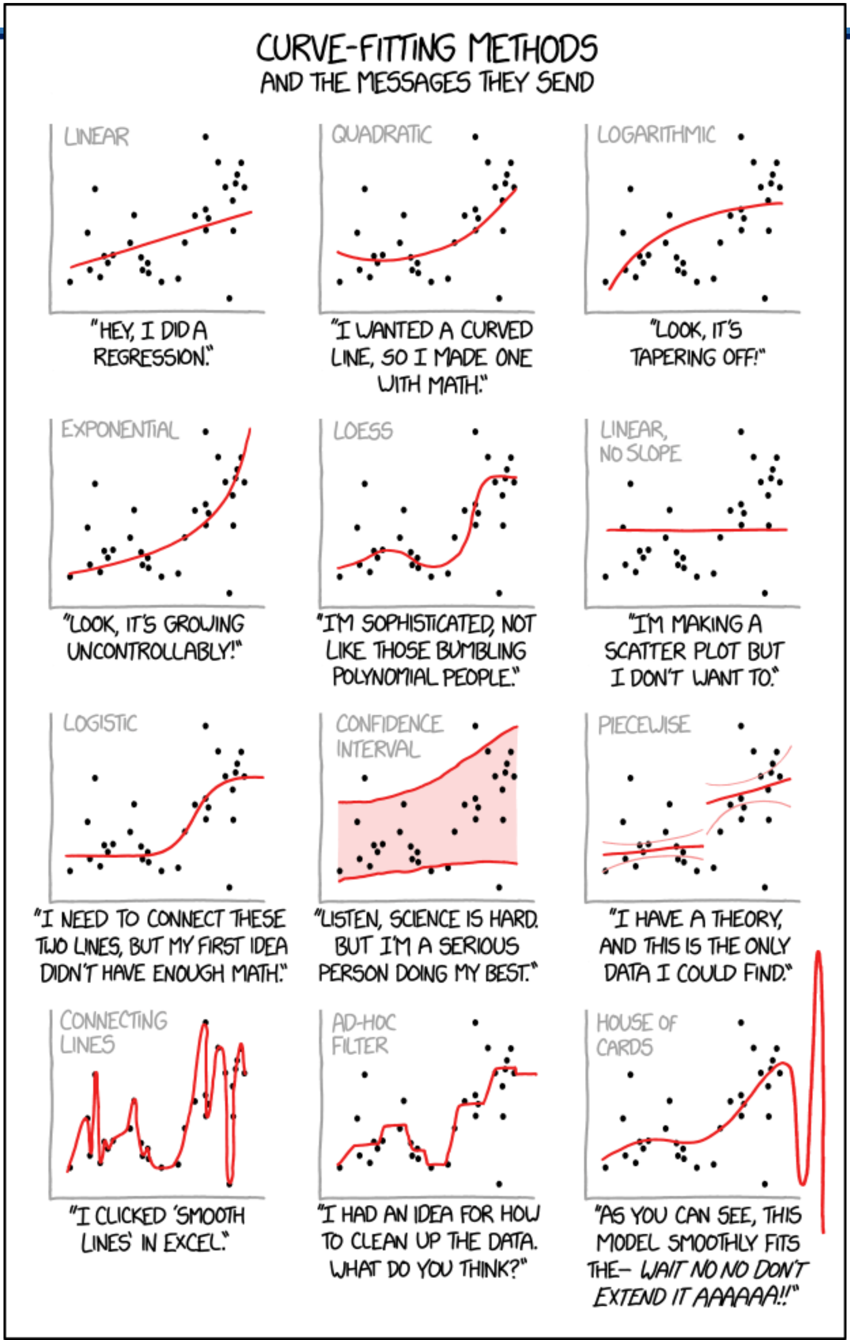
https://en.wikipedia.org/wiki/Independence_(probability_theory)

A small p-value indicates….

A big difference between datasets

Small probability of false positives

Small standard deviations of the compared groups



PoL Physics of Life TU Dresden

@jm_mightypirate

https://xkcd.com/2048/