Practical issues in preclinical data analysis

Martin C. Michel

Dept. of Pharmacology, Johannes Gutenberg University

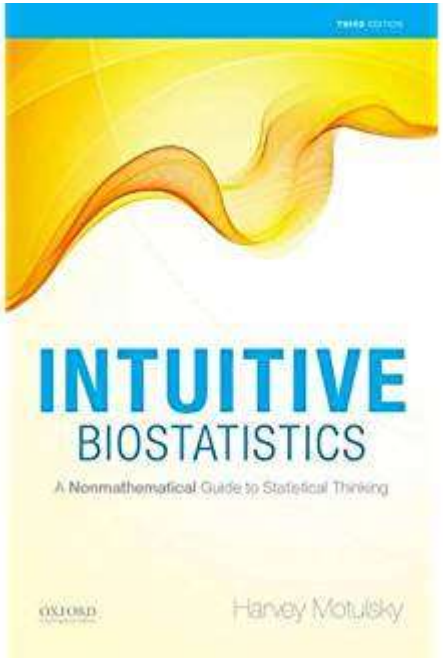
PAASP

Martin C. Michel

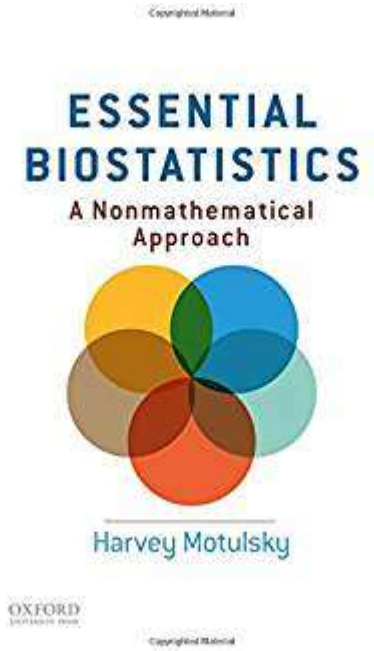
- Physician with board certification in Pharmacology and in Clinical Pharmacology
- >30 years experience in academia (e.g. Dept. Head Univ Amsterdam)
- >5 years experience in industry (Head Translational Research Boehringer Ingelheim)
- Head of IUPHAR guideline committee for increased robustness

www.dilbert.om (3.4.18)

Recommended books



Full textbook,
4th edition 2017
~ 60 €



Shorter version, focus on essentials
~ 20 €

… and the helpscreens of Prism (freely available under
http://www.graphpad.com/guides/prism/7/statistics/index.htm

Bias *versus* variability

control group

versus

experimental group

Factors causing systematic differences between groups can cause bias

control group

versus

experimental group

Factors causing differences between subjects (not groups) cause variability

Bias

- Bias can limit the validity of conclusions from scientific studies
- Bias can occur at multiple levels
  - Pre-study reading
  - Design
  - Execution
  - Data analysis
  - Data reporting
  - Publication

- Measures to reduce different forms of bias overlap

Lessons

- Analyses can be biased

Reducing bias in data analysis

- Understanding and implementing difference between exploratory and hypothesis-testing (confirmatory) study
- Randomization and blinding (can also be applied to data analysis)
  - Unblinding only after database lock (as in clinical trial)
- Pre-specification of analysis strategy

But which pre-specifications are best?

Topics

- **Exploratory vs. confirmatory study**
- Implications of Gaussian vs. non-Gaussian distribution
- Statistical analysis
- Outlier handling

Two types of study

Confirmatory (hypothesis testing)

- A scientifically plausible hypothesis exists

  e.g. KO mouse has predicted phenotype
  Typical in phase III clinical trial for regulatory purposes

Exploratory

- Absence of scientifically plausible hypothesis

  e.g. general characterization of KO mouse

Two types of study

Confirmatory (hypothesis testing)

- A scientifically plausible hypothesis exists
- Requires pre-specification of null-hypothesis, experimental methods incl. sample size and analytical methods

Exploratory

- Absence of scientifically plausible hypothesis
- Methods can be adapted (to some degree)

Two types of study

Confirmatory (hypothesis testing)

- A scientifically plausible hypothesis exists
- Requires pre-specification of null-hypothesis, experimental methods incl. sample size and analytical methods
- Leads to statement of significance regarding null-hypothesis

Exploratory

- Absence of scientifically plausible hypothesis
- Methods can be adapted (to some degree)
- P-values cannot be interpreted at face value
  - Impact of prior probability on FDR

Confirmatory vs. exploratory

- Both types of study have a place in science
  - But serve different roles
- Elements of both can be combined in a single study
  - Confirmatory for primary (and key secondary) endpoint
  - Exploratory for other endpoints
  - An (*in vitro*) study may have multiple, sequential steps; some are confirmatory, some exploratory
- Exploratory studies do not necessarily need P-values
  - Effect sizes with CI may do
- Explorative studies have a lower prior probability
  - Intrinsically higher FDR
- Readers need clarity what is what

Lessons

- Analyses can be biased
- Differentiate exploratory and confirmatory work and be transparent about it

Topics

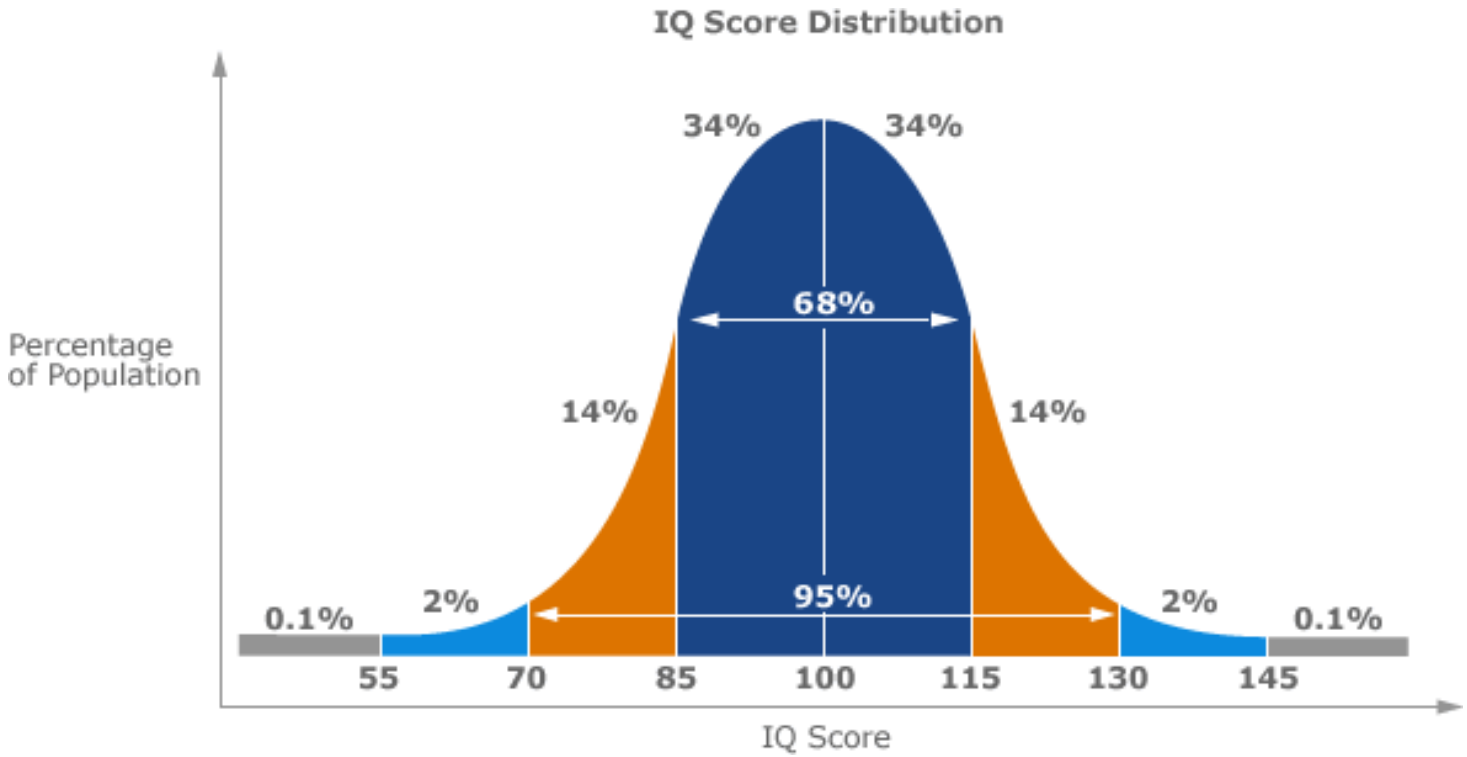
- Exploratory vs. confirmatory study
- **Implications of Gaussian vs. non-Gaussian distribution**
- Statistical analysis
- Outlier handling

What is a Gaussian distribution?

- Also called “normal” or “bell-shaped” distribution
  - Cave: not all bell-shaped distributions are Gaussian
- Arises when variability is caused by sum of many random and independent factors
  - Example: imprecise weighing of reagents, imprecise pipetting, random nature of radioactive decay, non-homogenous suspensions of cells or membranes, …
- The term Gaussian always refers to the population, not to the sample

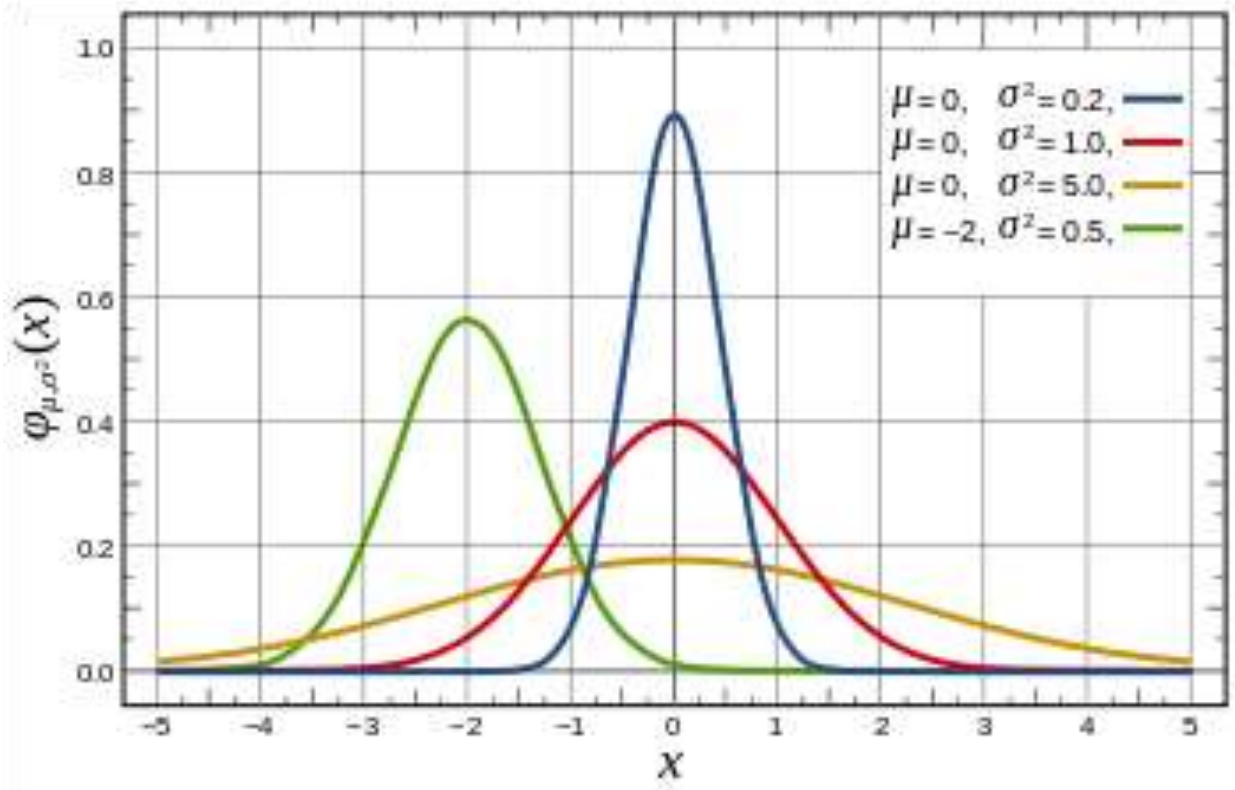
1777-1855

Distribution of IQ



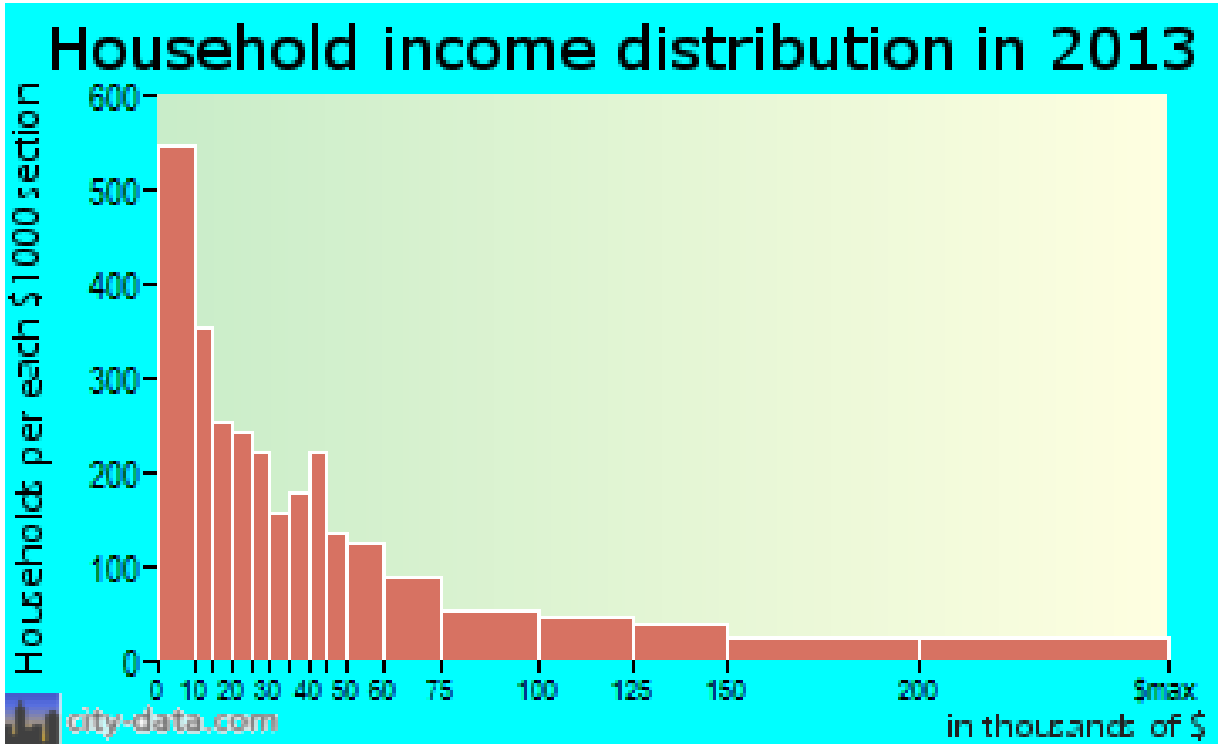
IQ is one of the few truly normally distributed variables
Why: IQ tests are designed to yield a normal distribution!

Gaussian (“normal”) distribution



Samples with Gaussian distribution can differ by their mean (position of peak on x-axis) and/or their SD (width of curve)

Income distribution in Seattle



At the very far end of the distribution: the two richest people on the planet live in Seattle, Jeff Bezos (Amazon) and Bill Gates (Microsoft)

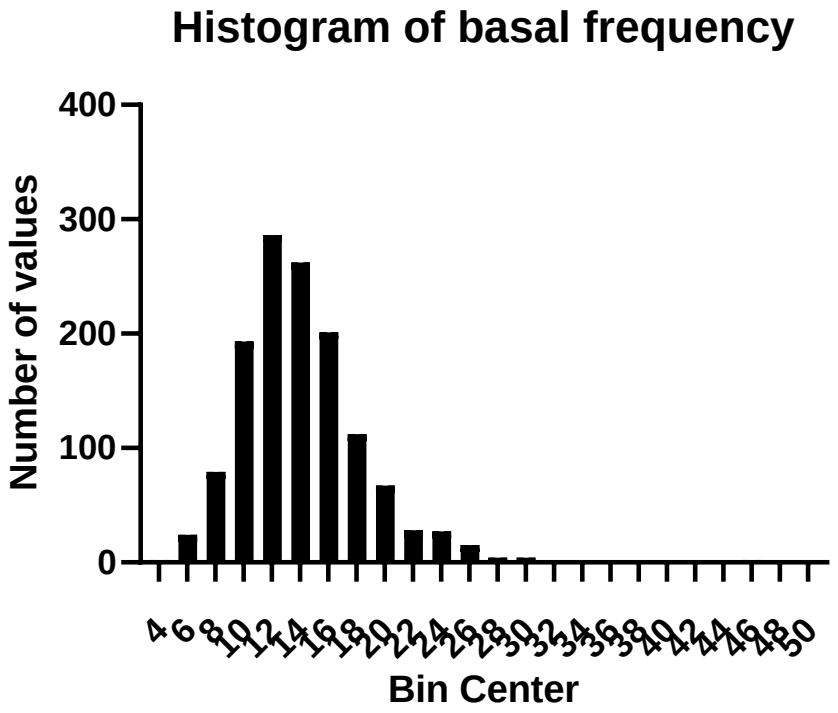
Gaussian distribution

- SD describes spread or width of distribution
- ~68% fall within ± 1 SD
- ~95% fall within ± 2 SD
- Unless the sample is huge, real distribution within a sample is less symmetrical

- Not all data sets exhibit a Gaussian distribution
  - True Gaussian distribution is rare

- Question is not “Does the underlying population exhibit a purely Gaussian distribution?” but rather “Is it close enough to Gaussian to apply tests making this assumption?”

Example: Micturition frequency



Look at the data!

Micturition frequency in 1335 patients seeking treatment for overactive bladder syndrome (1-7/24 h considered normal)

Amiri et al. (2018) Neurourol Urodyn 37 Suppl 5: S397-S398

Tests for Gaussian distribution

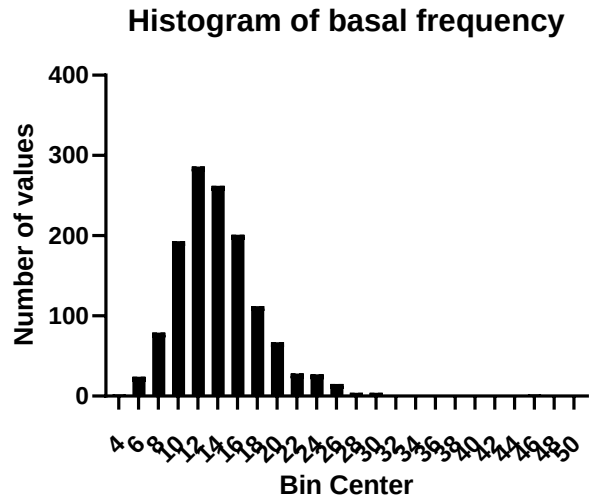
- “Gaussian” relates to populations, not samples
  - Issue for preclinical data where sample is very small relative to population
- Few populations, if any, exhibit a true Gaussian distribution
  - But may be close enough to apply parametric tests
- Several statistical tests to determine whether sample consistent with Gaussian distribution
  - I.e. whether we can assume that underlying populations is sufficiently close to Gaussian to use statistical methods assuming this
  - They test for “any” deviation from normality; with large data sets, even a trivial deviation may become “significant”

Tests for Gaussian distribution

- D’Agostino-Pearson (omnibus K2)
  - Looks how far distribution is from Gaussian for asymmetry and shape
  - Calculates P-value for presence
  - Recommended
- Anderson-Darling
  - Compares cumulative distribution of data against ideal Gaussian
- Shapiro-Wilk
  - Does not work well when a given number occurs multiple times
- Kolmogorov-Smirnov
  - Had historical value but is not very sensitive

Example: Micturition frequency

Test for normal distribution	
Anderson-Darling test	
A2*	20,18
P value	<0,0001
Passed normality test (alpha=0.05)?	No
P value summary	****
D'Agostino & Pearson test	
K2	546,4
P value	<0,0001
Passed normality test (alpha=0.05)?	No
P value summary	****
Shapiro-Wilk test	
W	0,9006
P value	<0,0001
Passed normality test (alpha=0.05)?	No
P value summary	****
Kolmogorov-Smirnov test	
KS distance	0,1149
P value	<0,0001
Passed normality test (alpha=0.05)?	No
P value summary	****



Statistical test shows: not normally distributed
Does it matter?

Amiri et al. (2018) Neurourol Urodyn 37 Suppl 5: S397-S398

Consequence of lack of normality

	Study 1			Study 2		
Symptom	n	Mean ± SD	Median (IQR)	n	Mean ± SD	Median (IQR)
Urgency	1136	10.7±6.6	10 (6;14)	621	10.0±5.5	10 (6; 13)
Frequency	1309	13.7±4.5	13 (11;16)	730	13.2±4.2	13 (10; 15)
Nocturia	1270	3.5±1.9	3 (2;4)	706	3.5±1.7	3 (2; 4)
Incontinence	785	5.1±3.9	4 (2;7)	418	5.5±3.9	5 (2;7)

P<0.0001 for each parameter in each study

Over-estimation by mean more relevant for some parameters than for others

Amiri et al. (2018) Neurourol Urodyn 37 Suppl 5: S397-S398

Distribution in preclinical data

- In experimental life sciences, sample sizes often are too small to allow meaningful exclusion of non-Gaussian distribution in the underlying population
- Can only be done robustly, if you already have large or lots of data sets for that parameter (own or literature)
- If not, logical reasoning may help
- If impossible, not considering Gaussian distribution is the more conservative assumption

Always remember:

- Absence of proof is not the same as proof of absence
- Look at the data!

Log-normal distribution

- Remember calculus 101:
  - Log(A·B) = logA + logB (the slide ruler was based on this principle)
- Gaussian distribution arises from multiple additive factors causing variation
- If multiple factors are multiplicative (rather than additive), distribution of data will be skewed but becomes Gaussian when shown as logarithms
  - Log normal distributions cannot include 0 or negative values
- As a consequence, data from apparently non-Gaussian distributions may become suitable for parametric analysis after log transformation
- This approach is standard for some pharmacokinetic analyses and for analysis of concentration-response curves

Example

- Concentration-response curves are typically constructed by varying concentration on a log scale
  - Reflecting how the law of mass action works
- EC_{50} distribution within a series of experiments tends to be highly skewed
- Log EC_{50} (a.k.a. pEC_{50}) typically much closer to Gaussian distribution
  - This implies that it is ok to report mean pEC_{50} but misleading to calculate mean EC_{50}

Implications

- The presence or absence on Gaussian distribution has implications for the choice of appropriate statistical tests and reporting
- Parametric tests only appropriate if Gaussian distribution in the population can be assumed
- If Gaussian distribution cannot be assumed, means are not very informative and can be misleading

Mean vs. median

- Mean
  - Sum of all values divided by number of values
- Median
  - Half of all values are above and half below

Mean vs. median

- Mean
  - Sum of all values divided by number of values
  - Poor description of non-Gaussian distributions
- Median
  - Half of all values are above and half below
  - Better description of non-Gaussian distributions

Mean vs. median

- Mean
  - Sum of all values divided by number of values
  - Poor description of non-Gaussian distributions
  - Good description of Gaussian distributions
- Median
  - Half of all values are above and half below
  - Better description of non-Gaussian distributions
  - Good description of (large) Gaussian distributions
  - Mean and median are (almost) identical in large samples of Gaussian distribution

- Both describe position of peak on x-axis
- How to describe variability (width of distribution curve; variability)?

Quantifying variability

- Range
  - Difference between largest and smallest observed value
- Percentiles
  - x^{th} percentile is % of all values smaller than x
  - 50^{th} percentile = median
  - Used for reporting of confidence intervals
- Interquartile range
  - 75^{th} minus 25^{th} percentile
- 5 number summary
  - Minimum, 25^{th} percentile, median, 75^{th} percentile, maximum

- Applicable to all types of data, except categorical

Quantifying variability (Gaussian)

- Standard deviation (SD)
  - Variation of data around mean
  - ±1 SD covers 68%
- Calculation of SD
- Variance = SD squared (statistical definition)

Quantifying precision

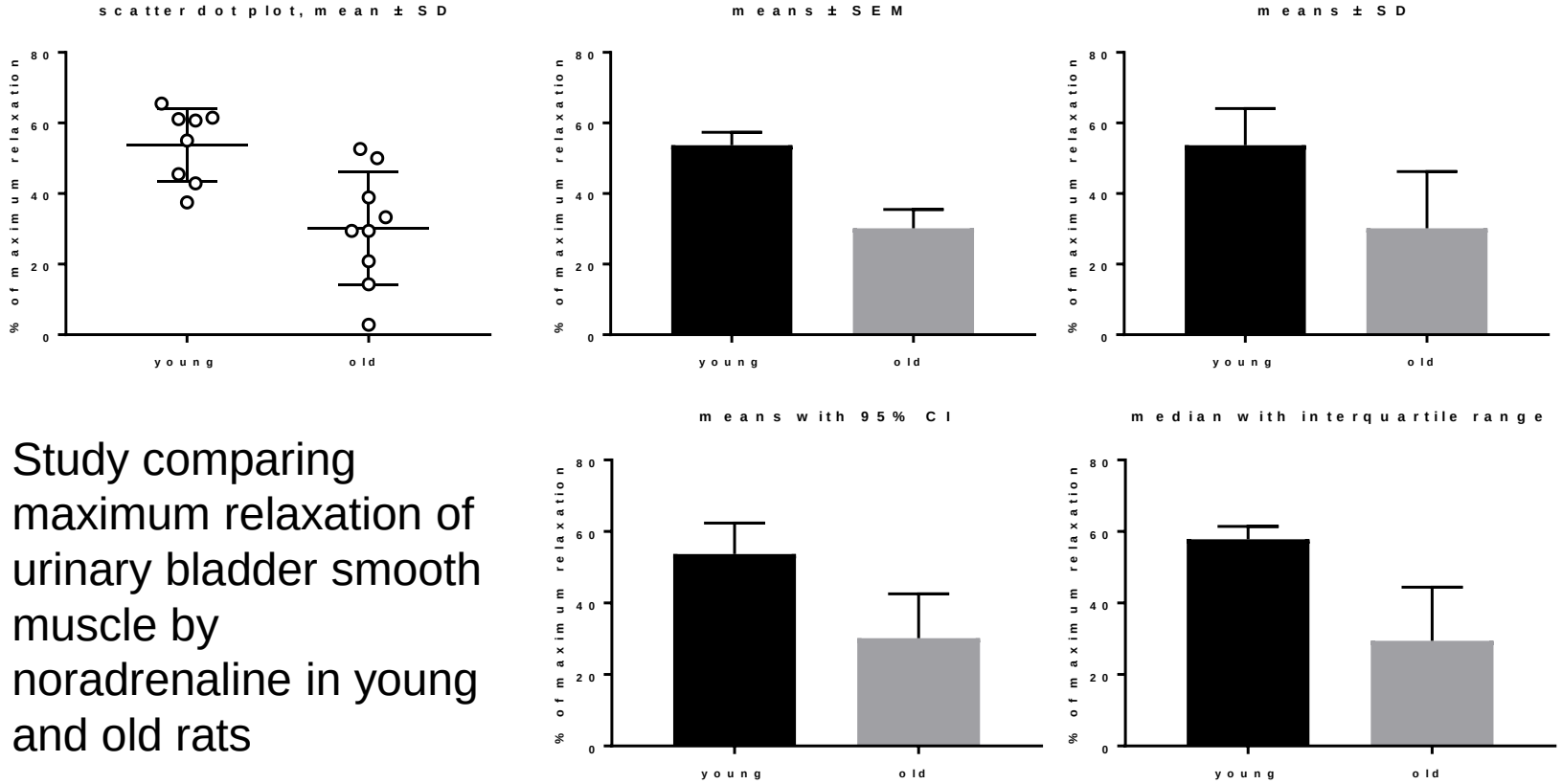
- Confidence interval (CI) is an indicator of precision, not of variability
  - Tells you about margin of error for a parameter estimate
  - Helpful to make general conclusion based on specific sample
    - E.g. where is the true mean of the population expected based on my sample
  - CI is not necessarily symmetric around mean
  - CI is sample-size dependent
- CIs can be calculated for any % range but 95% CI is most common
- CIs make several assumptions
  - Based on random sample
  - Based on independent observations

Quantifying precision

- Standard error of the mean (SEM; also sometimes referred to as SE) also is an indicator of precision, not of variability
  - Defined as SD divided by square root of n (sample size-dependent!)
  - Only applicable to data from populations with Gaussian distribution
  - SE can be calculated also for parameters other than mean

- SD tells you about the population, SE about mean within the sample
- We are typically only interested in the sample as an indicator of the population it came from → variability should be reported as SD (if Gaussian)

Indicators of variability and precision



Study comparing maximum relaxation of urinary bladder smooth muscle by noradrenaline in young and old rats

Frazier et al. (2006) Naunyn-Schmiedeberg‘s Arch Pharmacol 373: 300-309

Informative choice depends on goal

- Goal: show variability in sample (information on population sample came from)
  - Scatter plot optimal; not assuming Gaussian distribution
  - Alternatives: box & whisker, ranges or frequency distribution; violin plot
  - SD possible and perhaps necessary with large samples but assumes Gaussian distribution
- Goal: show precision of mean (how well does mean describe the sample)
  - CI optimal
  - SEM “looks nicer” but more difficult to interpret and depending on Gaussian distribution
    - May make small differences look more meaningful than they are
  - CI and SEM are highly sample size-dependent

Lessons

- Analyses can be biased
- Differentiate exploratory and confirmatory work and be transparent about it
- Assumption of Gaussian drives choice of meaningful analysis and reporting

Topics

- Exploratory vs. confirmatory study
- Implications of Gaussian vs. non-Gaussian distribution
- **Statistical analysis**
- Outlier handling

Statistical inference

- Statistics analyze data from a sample to make conclusions about the population from which it had been taken (inference)
- Standard statistical methods assume that the population is much larger than the sample
  - Consider the chance element of a small sample being seen as representative

Small samples may be misleading

Major assumption

- The sample being analyzed is a random representation of the population of interest
- Small samples run a greater risk of being non-representative, just by chance
- Even for larger samples, “random” is the key phrase

Randomness principle

Randomness principle

- A P-value reports the probability of seeing a difference as large as you observed, or larger, even if the two samples had been selected **randomly** from populations with the same mean

- Only meaningful if all factors other than primary variable **randomly** distributed among groups
  - Randomization and blinding helpful
  - P-values cannot be interpreted at face value if investigator-induced violations of randomness exist

Motulsky (2014) Naunyn-Schmiedeberg‘s Arch Pharmacol 387: 1017-1023

Violation of randomness

- Violations of the randomness principle make resulting P-values difficult to interpret, perhaps even invalid
- Violations (biases) can occur unconsciously or be investigator-induced
- Unconscious biases
  - Sampling error
  - Selection bias
  - Other biases
- Investigator-induced violations are also referred to as P-hacking

Motulsky (2014) Naunyn-Schmiedeberg‘s Arch Pharmacol 387: 1017-1023

P-hacking examples

- Decision to do add 2 additional experiments when n = 6 yielded P = 0.055
  - The new n = 8 is biased by the trend in n = 6 and no longer a random sample
- Change to log-normalized data
  - Log normalization can be justified or even required when data are only normally distributed on a log scale
- Change denominator
  - From fmol/mg protein to fmol/g wet weight
- Switch to a different statistical test
  - paired vs. unpaired test
- “Outlier” removal

P-hacking

- Various design choices may be fine if pre-specified
- Post-hoc changes in design, analysis and reporting introduce bias and violate randomness principle
- This makes resulting P-values difficult to interpret, irrelevant or even misleading
  - Bias for finding a difference even if it is not there
  - Trend for exaggerated effect sizes

Motulsky (2014) Naunyn-Schmiedeberg‘s Arch Pharmacol 387: 1017-1023

P-values aren’t trophies

- The * above a data point looks nice and may be necessary to get the paper accepted
- P-values from biased experiments cannot be interpreted as face value
- P-hacking is a blatant form of biasing experiments
- Science is about answering meaningful questions
  - Papers are a tool for this but not the primary goal

Why significance tests?

The function of significance tests is to prevent you from making a fool of yourself, and not to make unpublishable results publishable

Colquhoun (1971) Lectures on Biostatistics. Clarendon Press, Oxford

What a P-value does not mean

- In contrast to a common perception, a P-value does not tell us the probability that an observed finding is true

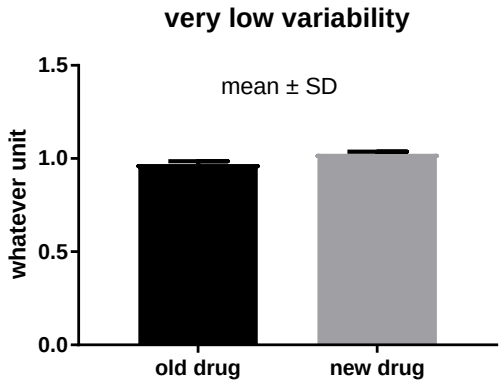
- A P-value reports the probability of seeing a difference as large as you observed, or larger, even if the data had been randomly sampled from populations with the same mean

What makes a P-value?

- A mathematic relationship exists in any data set between
  - Variability (e.g. SD) within tested sample
  - Observed/desired effect size
  - Sample size
  - P-value
- Data analysis
  - Know: sample size, variability and observed effect size
  - Calculate: P-value
- Study planning
  - Know/assume: variability, smallest effect size you care about, desired P-value
  - Calculate: required sample size

Some fake examples

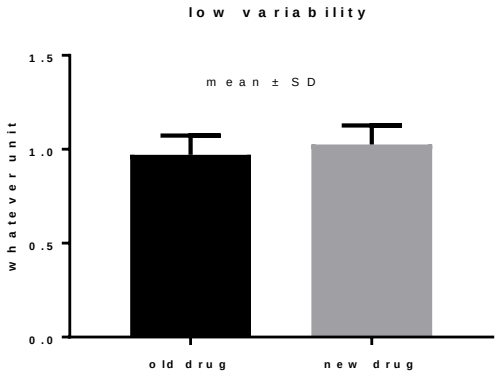
0.95	1.02
0.97	1.03
0.98	1.01
0.99	1.04
0.96	1.03



Sample size small (n = 5)
Effect size small (mean difference 0.056)
Variability very low (0.970±0.007 vs. 1.026±0.005)
P = 0.002 in t-test

Some fake examples

1.05	1.12
0.87	0.93
0.98	1.01
1.09	1.14
0.86	0.93

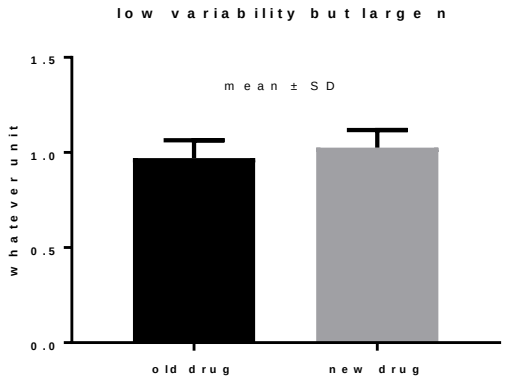


Sample size small (n = 5)
Effect size small (mean difference 0.056)
Variability low (0.970±0.046 vs. 1.026±0.045)
P = 0.4114 in t-test

Some fake examples

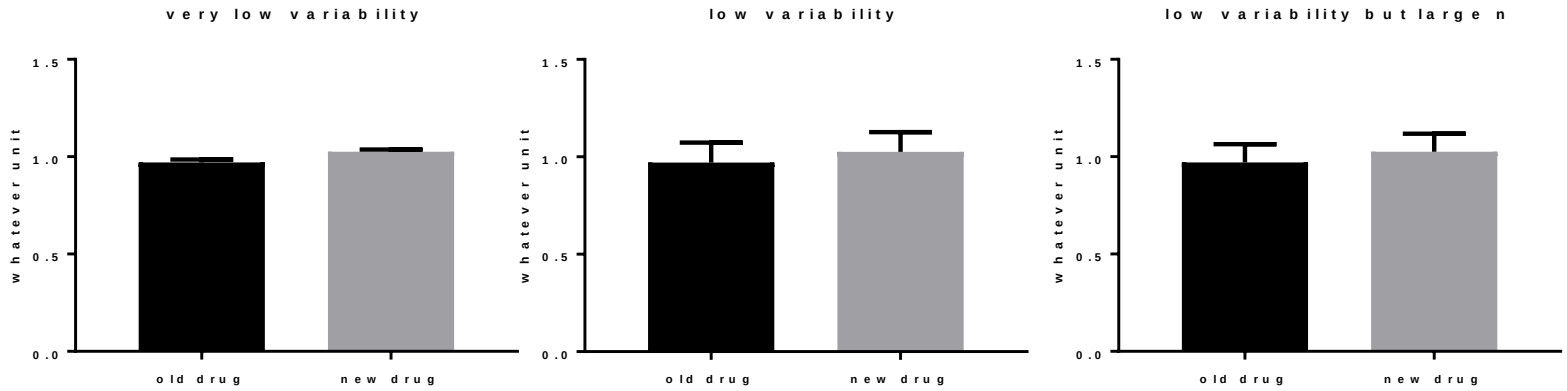
1.05	1.12
0.87	0.93
0.98	1.01
1.09	1.14
0.86	0.93

Add the very same numbers four more times to increase n without changing variance



Sample size large (n = 25)
Effect size small (mean difference 0.056)
Variability low (0.970±0.046 vs. 1.026±0.045)
P = 0.0390 in t-test

Some fake examples



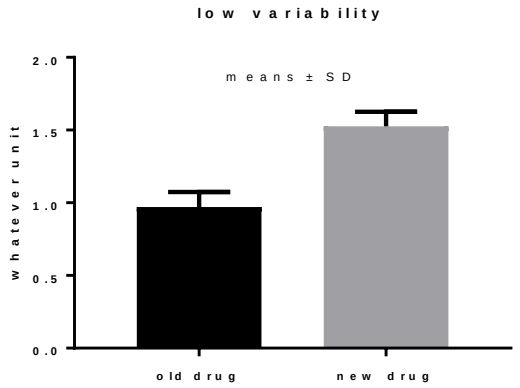
With a small effect size, minor changes in variability turn a P = 0.0020 to a P = 0.4114 but increasing sample size brings it down to P = 0.0390.

However, the effect size (difference between groups) stays the same.

Does new drug work or not?

Some fake examples

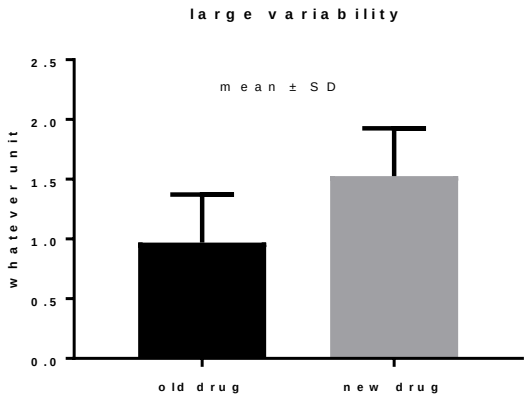
1.05	1.62
0.87	1.43
0.98	1.51
1.09	1.64
0.86	1.43



Sample small (n = 5)
Effect size large (mean difference 0.556)
Variability low (0.970±0.046 vs. 1.526±0.045)
P < 0.0001 in t-test

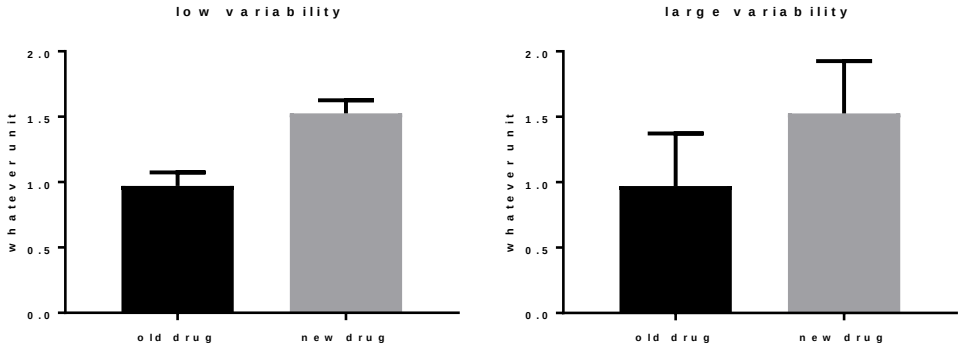
Some fake examples

1.35	1.92
0.57	1.13
0.98	1.51
1.39	1.94
0.56	1.13



Sample small (n = 5)
Effect size large (mean difference 0.556)
Variability large (0.970±0.180 vs. 1.526±0.179)
P = 0.0600 in t-test

Some fake examples



With a large effect size, a change in variability can turn a P < 0.0001 to a P = 0.0600

However, the effect (difference between group) stays just the same

Does new drug work or not?

Consider “absence of proof” vs. “proof of absence”

False Discovery Rate (FDR)

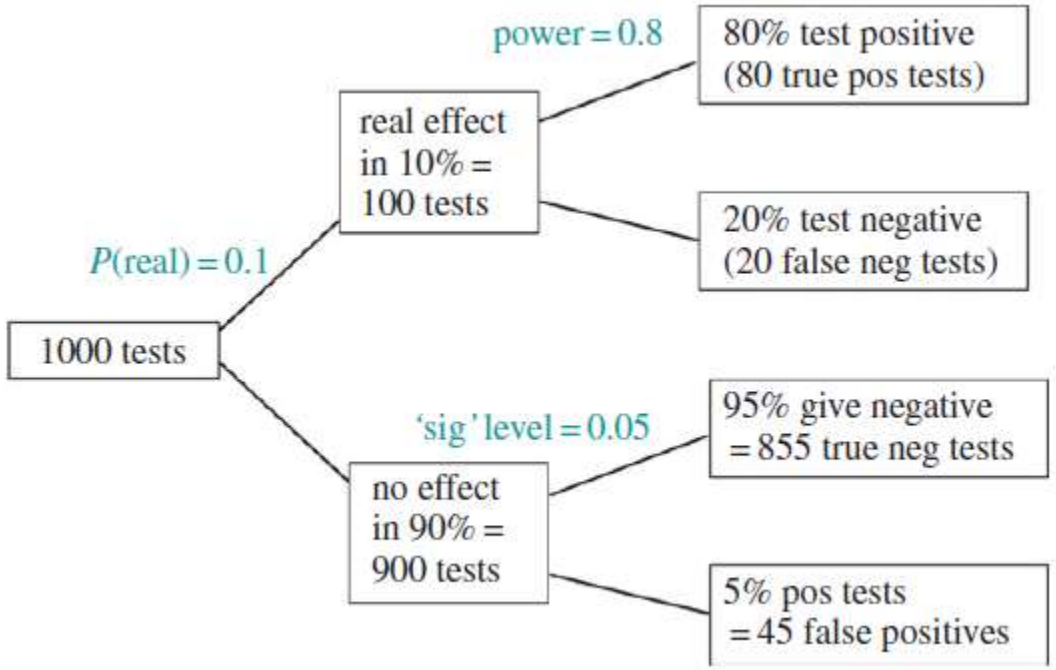
- Even if everything has been done “by the book” (randomization and blinding, detailed method description, no P-hacking), a “significant” finding may not be true
- Poor robustness even of appropriately designed studies predicted in 2005
- Underlying concept later termed False Discovery Rate
  - More recently termed False Positive Rate

Ioannidis (2005) PLoS Medicine 2: e124
Colquhoun (2015) R Soc Open Sci 1: 140216

FDR vs. PPV

- Positive predictive value (PPV)
  - Probability that a real effect exists if a “significant” result has been obtained
- FDR
  - Probability that a real effect does NOT exist if a “significant” result has been obtained
- PPV and FDR are flipsides of the same coin
  - FDR = 100 – PPV
  - PPV = 100 - FDR

FDR in significance testing



45/125 falsely detected as „significant “, i.e. FDR is 36%

Colquhoun (2015) R Soc Open Sci 1: 140216

What drives a high FDR?

- Low prior probability
- Low α
- Low power
  - Based on small effect size and small sample size and α

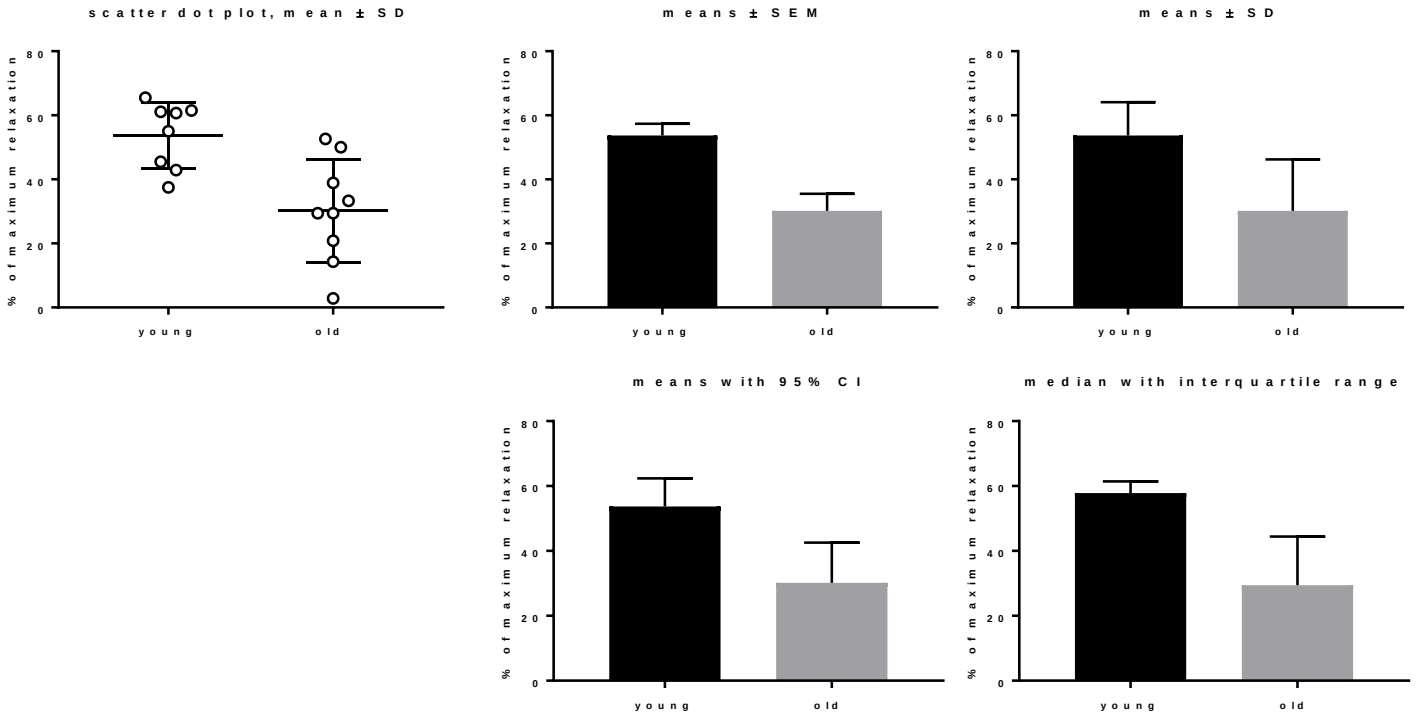
a $p \sim 0.05$ means nothing more than 'worth another look'

If you want to avoid making a fool of yourself very often, do not regard anything greater than $p < 0.001$ as a demonstration that you have discovered something

Colquhoun (2015) R Soc Open Sci 1: 140216

Comparing groups

Does relaxation of urinary bladder by noradrenaline differ between young and old rats?



Frazier et al. (2006) Naunyn-Schmiedeberg‘s Arch Pharmacol 373: 300-309

Comparing groups

Does relaxation of urinary bladder by noradrenaline differ between young and old rats?

- Analysis options
  - t-test comparing both groups
    - P = 0.0030
  - Calculation of mean difference ± CI

Mean ± SEM of column A	53,71 ± 3,664, n=8
Mean ± SEM of column B	30,17 ± 5,365, n=9
Difference between means	-23,55 ± 6,667
95% confidence interval	-37,76 to -9,335

- What do you consider more informative?

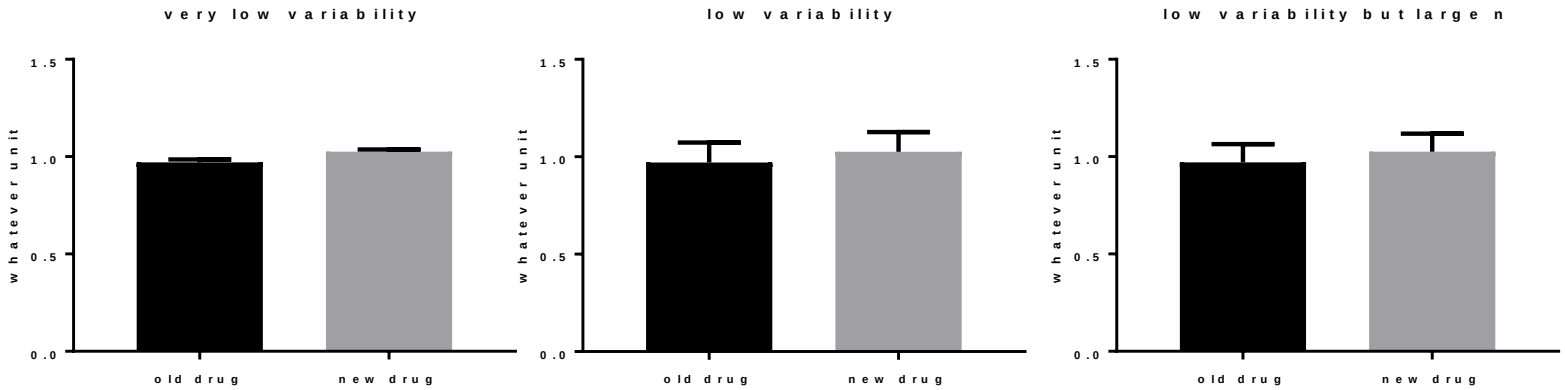
Frazier et al. (2006) Naunyn-Schmiedeberg‘s Arch Pharmacol 373: 300-309

Link between CI and P-value

- If 95% CI does **not include null-hypothesis**, P must be < 0.05
- If 95% CI does **include null hypothesis**, P must be ≥ 0.05

Four things you need to know

- P-value for same effect size and variability depends on sample size



With a small effect size, minor changes in variability turn a P = 0.0020 to a P = 0.4114 but increasing sample size brings it down to P = 0.0390.

However, the effect size (difference between groups) stays the same.

Four things you need to know

- P-value for same effect size and variability depends on sample size
- P-values have a strange logic
  - Based on assumption about population (null-hypothesis) but asks about data in sample

Four things you need to know

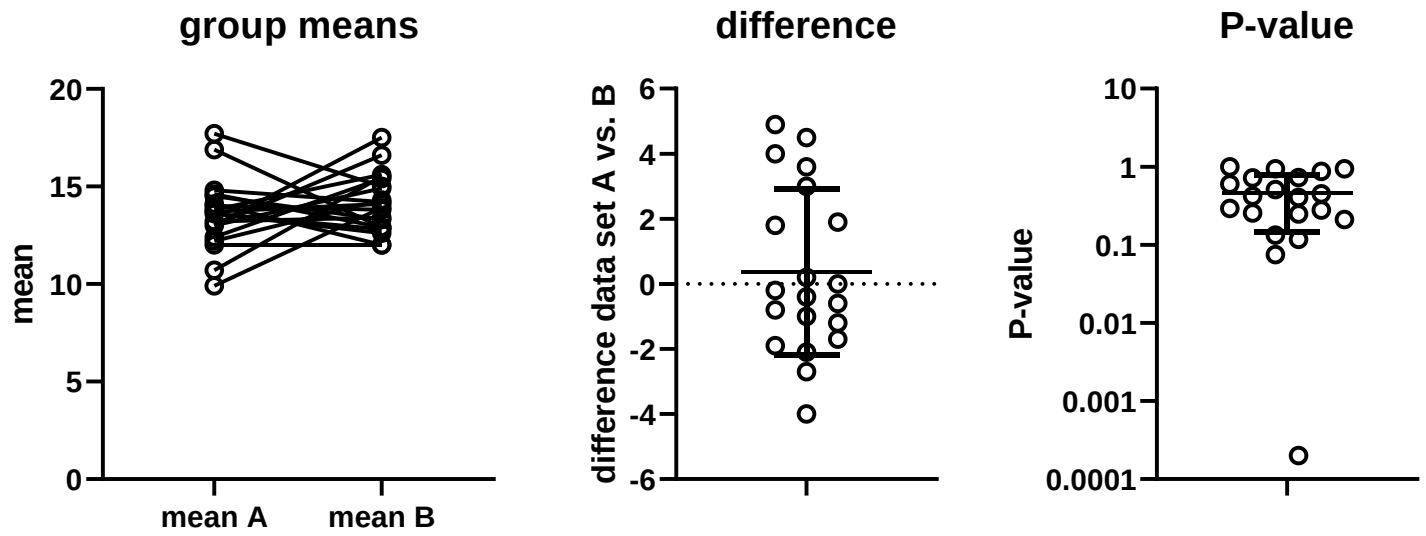
- P-value for same effect size and variability depends on sample size
- P-values have a strange logic
- The null-hypothesis is rarely true
  - Null-hypothesis typically opposite of scientific hypothesis
  - But the deviation from it may be of trivial magnitude
  - Very large samples of detect a trivial difference as “statistically significant”

Four things you need to know

- P-value for some effect size and variability depends on sample size
- P-values have a strange logic
- The null-hypothesis is rarely true
- P-values are poorly reproducible (fickle)

The fickle P-value

- 20 students randomly picked pairs of 4, 6 or 10 from micturition database



- Mathematical modeling confirms the fickleness of the P-value

Halsey et al. (2015) Nature Meth 12: 179-185

Why statistical hypothesis-testing?

- Statistical tests turn a spectrum of probabilities into a binary decision
  - Reject or accept null-hypothesis
  - But reality is more complex most of the time
- Statistical tests can enable decision making
  - $P < 0.05$ in phase III study as basis for regulatory approval
  - Decision to move a new molecule into clinical development
  - Hypothesis-testing statistical tests only helpful in hypothesis-testing (confirmatory) studies

A legal analogy

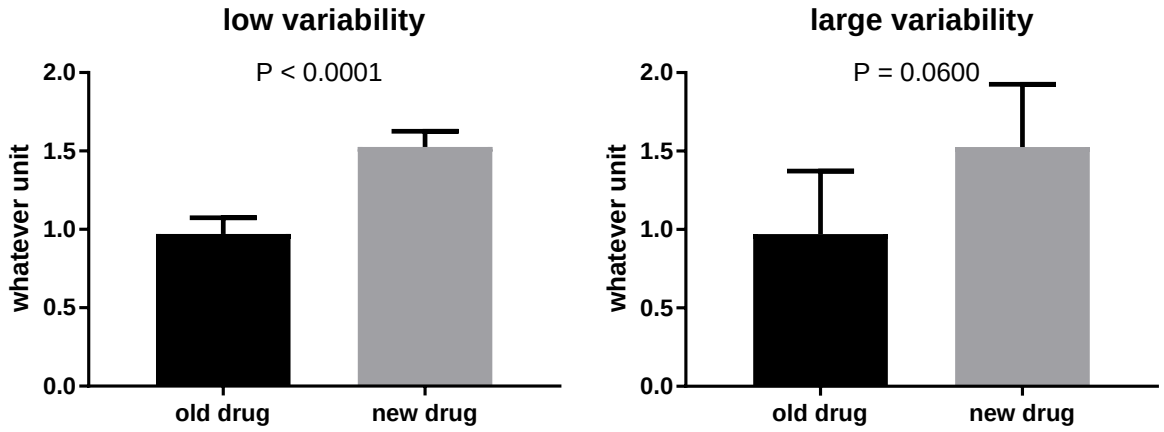
- The null-hypothesis of a scientist is comparable to the “presumed innocent” of a juror in a court trial
  - Neither can conclude that the defendant is innocent (null hypothesis true)
  - Only guilty vs. not guilty (null hypothesis rejected or not)
  - Journalists covering a trial are not forced to make a binary decision but can describe the grey zone; would be helpful for scientists as well

*, ** or ***?

- For rejection of null-hypothesis, P = 0.04 and P = 0.0004 is the same
  - The null-hypothesis is rejected in both cases
  - Nonetheless, a difference in probability exists
  - Better reflected in distance of end of CI from null-hypothesis
- Fancy wording around P-values slightly above 0.05 (“almost significant” etc.)
  - reflects the reality that probability is a grey scale
  - misses the point that “significance” is about making a crisp binary decision

What is “not significant”?

- No statistical test can prove the null-hypothesis
- If the null-hypothesis could not be rejected, it means just that. Possible explanations include:
  - There is no difference
  - Effect size too small
  - Sample size too small
  - Variability too large
  - Type II error



Absence of proof does not equal proof of absence

Commonly used statistical tests

Some assumptions are shared by **all** statistical tests

- Random (representative) sample
- Independent observations
- Measured variable is indeed the one you care about

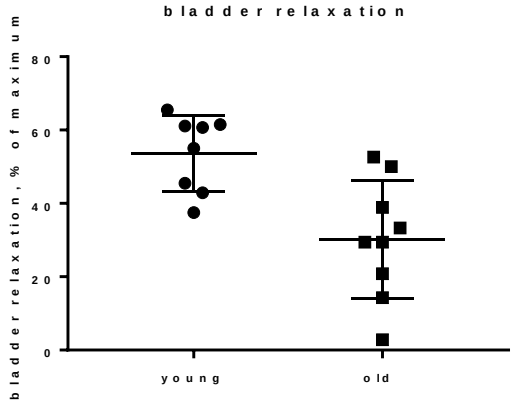
Commonly used statistical tests

• Unpaired, parametric	unpaired t-test
• Paired, parametric	paired t-test
• Unpaired, non-parametric	Mann-Whitney test
• Paired, non-parametric	Wilcoxon matched pairs

Commonly used statistical tests

Continuous variable measured in two groups

- Some tests are based on the assumption of Gaussian distribution in the population (parametric tests)
  - Standard t-test also assumes similar SD in both groups
- Non-parametric tests do not assume Gaussian distribution
  - Less sensitive to “outliers”
  - May yield different P-values



P-values

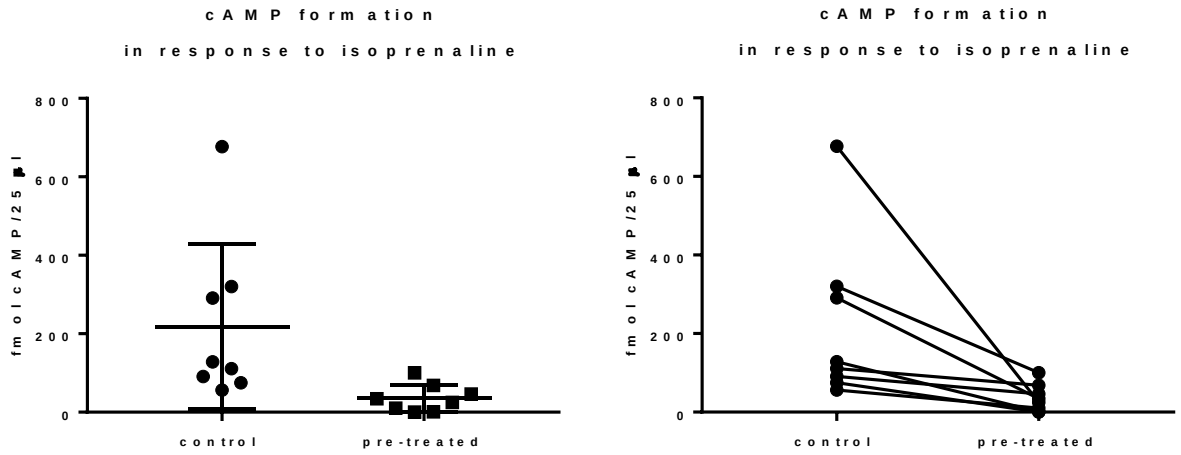
- Unpaired t-test 0.0300
- Mann-Whitney test 0.0035

Frazier et al. (2006) Naunyn-Schmiedeberg‘s Arch Pharmacol 373: 300-309

Commonly used statistical tests

Continuous variable measured in two groups

- Some tests are based on the assumption of Gaussian distribution in the population (parametric tests)
- Non-parametric tests do not assume Gaussian distribution
- Either type can be paired or unpaired



P-values

• Unpaired t-test	0.0290
• Paired t-test	0.0406
• Mann-Whitney test	0.0019
• Wilcoxon matched pairs	0.0078

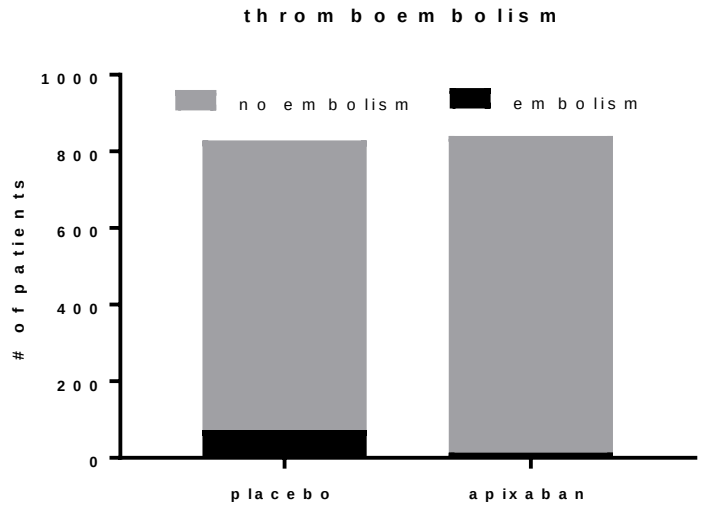
Okeke et al. : in preparation

Commonly used statistical tests

Binary variable measured in two groups

Occurrence of thromboembolism with factor X inhibitor apixaban

- Fisher's exact test
  - Chi-square ok with large samples



Agnelli et al. (2013) New Engl J Med 369: 799-808

The more you look, the more you find

- If the null-hypothesis is true (no group difference), the chance for **not** finding a “statistically significant” difference is 0.95
- If you make two such comparisons (null hypothesis true in both cases), the chance for not finding a statistically significant difference is $0.95 \cdot 0.95$, i.e. 0.9025 (about 90%)
- Thus, looking for $P < 0.05$ in two comparisons leaves an ~10% chance of finding one statistically significant difference (if no differences exist between populations)

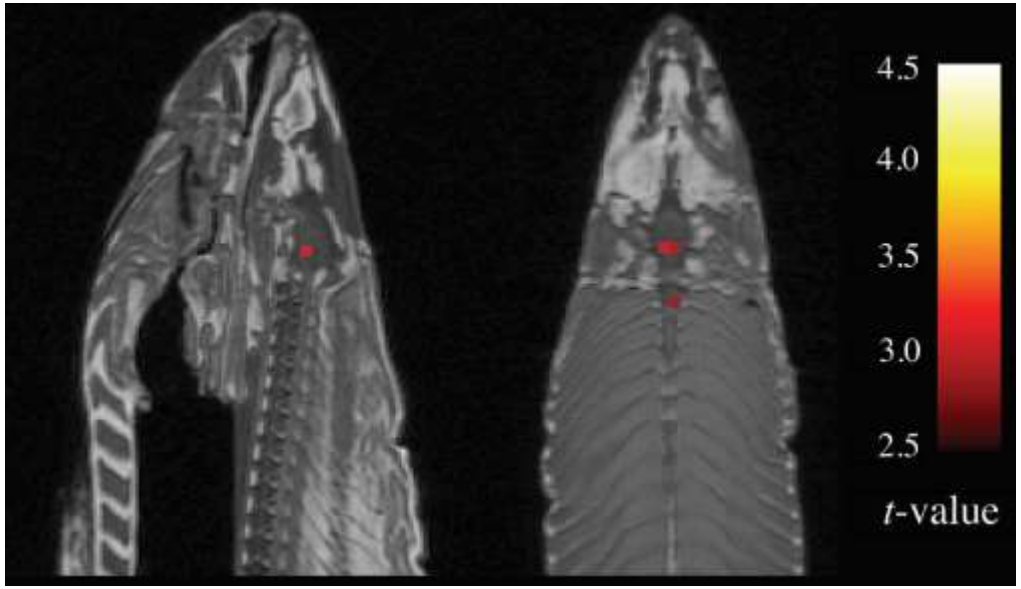
The more you look, the more you find

- The more comparisons you make, the more likely it is to find a “statistically significant” difference (if none is there)
  - With 13 comparisons, there already is a 50% chance of finding at least one to be “statistically significant”
  - A small P-value in any of those does not protect you from this
    - Rejection of null-hypothesis is a binary decision!

An extreme example

- Experimental animal shown two different pictures
- MRI assessment of cerebral blood flow in 1000 brain regions
- At pre-set α of 0.001, two „significant“ spots emerged

- Real finding or coincidence?



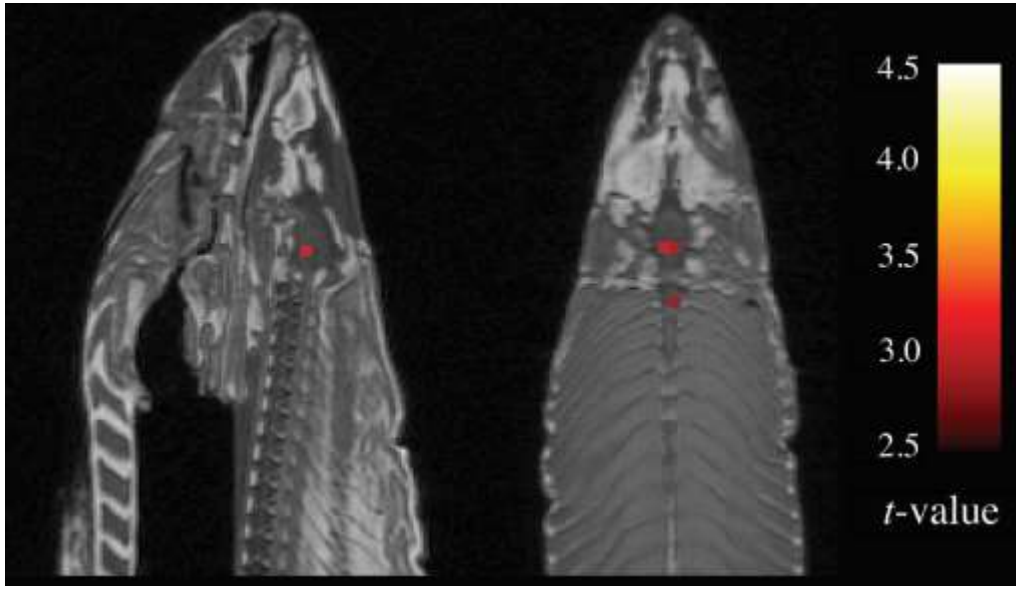
Bennett et al. (2011) J Serendip Unexpec Res 1: 1-5

An extreme example

- Experimental animal shown two different pictures
- MRI assessment of cerebral blood flow in 1000 brain regions
- At pre-set α of 0.001, two „significant“ spots emerged

- Real finding or coincidence?

The experimental animal was a dead salmon!



Bennett et al. (2011) J Serendip Unexpec Res 1: 1-5

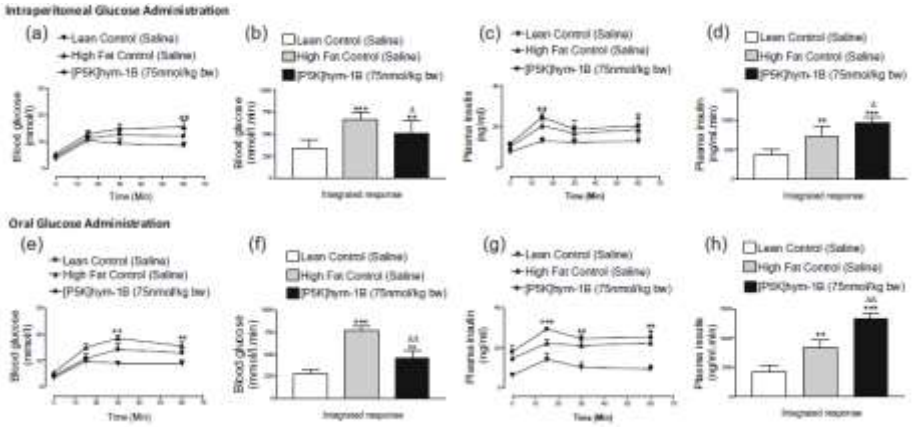
Types of multiple comparisons

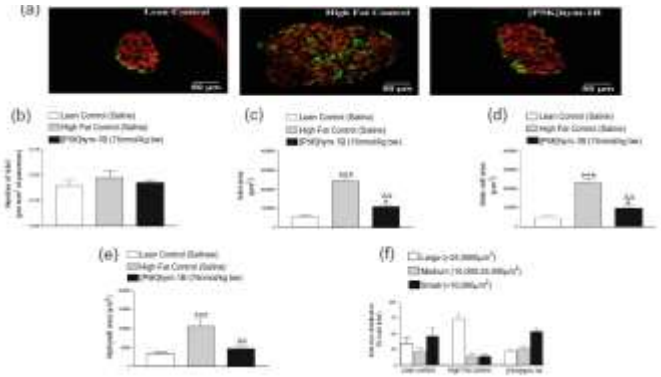
- Many subgroups
- Multiple endpoints
- One control, multiple treatments
- Multiple sample sizes (tested sequentially)

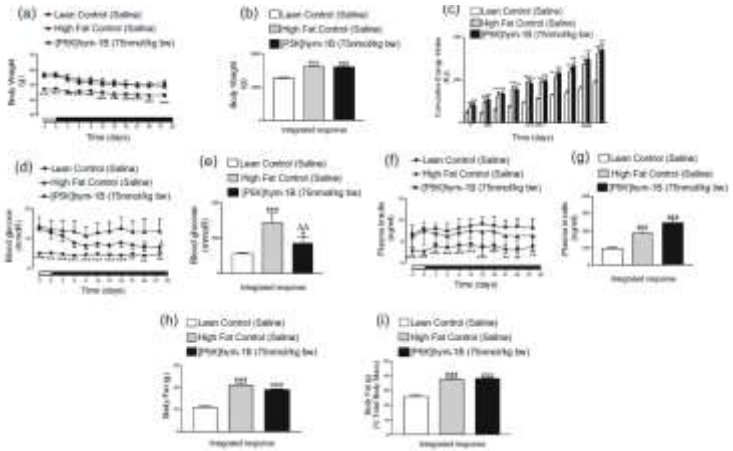
- Ask yourself what your scientific question is
- Consider 1-2 primary endpoints (hypothesis-testing), multiple secondary endpoints (exploratory)
  - Typically applied to clinical trial
  - It is unrealistic that a study tests >10 pre-specified hypotheses

Multiple endpoints

- One chronic treatment, partly combined with multiple acute interventions
- Many outcomes
  - mRNA expression of 9 genes
  - >20 functional parameters
  - Some at multiple time points
  - All based on 6 animals per group
  - >100 P-values reported







Owolabi et al. (2016) Naunyn-Schmiedeberg‘s Arch Pharmacol 389: 769-781

How to deal with multiple comparisons?

- Multiple comparisons within a study are not intrinsically wrong and may even be required for some scientific questions
- Adjust α to number of comparisons
  - Divide planned α by number of planned comparisons (Bonferroni correction)
  - Alternative: apply hierarchical testing

There is no free lunch

Multiple comparison corrections have a price

- They reduce risk of type I error
  - Fewer false positives
- They increase risk of type II error
  - More false negatives

Recommended tests

- Statistical comparison between 3 or more groups is done by analysis of variance (ANOVA)
- Compares means of 3 or more groups
- Compares overall variability with variability within groups
- If variability within groups is smaller than overall variability, means must differ substantially
- ANOVA does not tell you which means differ from one another!

ANOVA post-tests

- Multiple tests exist to compare specific group pairs
- Tukey test
  - Compares all group means with each other
- Dunnett test
  - Compares several groups to one shared control group
  - Other options available (e.g. Bonferroni-corrected t-tests)
- Post-tests yield P-values
  - Options for paired and/or non-parametric testing
- Post-tests also yield mean inter-group difference with CI
- Either accounts for number of comparisons being made
- Significance levels and CI apply to the entire family of comparisons

Multiple comparison planning

- Plan every comparison
  - Additional comparison that had not been pre-planned should be marked as such
- Do every planned comparison
- Report every comparison that has been done

- State whether reported α for specific pair-wise comparisons includes correction for (0.05/K for Bonferroni) or not ($P<0.05$)

Lessons

- Analyses can be biased
- Differentiate exploratory and confirmatory work and be transparent about it
- Assumption of Gaussian drives choice of meaningful analysis and reporting
- P-values are fickle
- Choice of statistical test affects P-value
- Effect size with confidence interval may be more informative than P-value
  - Particularly for explorative experiments

Topics

- Exploratory vs. confirmatory study
- Implications of Gaussian vs. non-Gaussian distribution
- Statistical analysis
- **Outlier handling**

Outliers

Outliers

- An outlier is a value so far away from the others that it appears to have come from a different population
- Possible reasons include
  - Invalid data entry (typo, missing decimal etc.)
  - Experimental mistake (e.g. double pipetting)
  - Biological diversity
  - Random chance
  - Skewed distribution (remember income graphs from lecture 2)

Outlier problem

- Outliers can spoil analysis
  - Creating the impression of a difference/association
  - Blocking discovery of real difference/association
  - Non-parametric tests less sensitive to outliers

- Remember: experimenters are highly biased
  - Unreliably witness to identify outliers

5 questions before outlier removal

1. Was there a mistake in data entry?
   - Fix it!
2. Is extreme value code for a missing value?
3. Evidence of experimental error?
   - Best to make a note pro-actively during experiment if you suspect a mistake (e.g. double pipetting)
4. Could it represent biological variability?
   - May be an exciting finding
5. Is the sample possible from a population with non-Gaussian distribution?
   - Log transformation?

Formal outlier tests

Based on the question:

- What is the chance that one value will be so far from the others, if all were sampled from a Gaussian distribution?
- E.g. any value >2x or even >3x SD away from mean
  - <5% and 1% of values in Gaussian distribution should be that far away

- Small P-value may indicate that value came from a different population
  - Example for different population: double pipetting
  - Removal of value can be justified if the “5 questions” were answered with “no”

Is outlier removal legitimate?

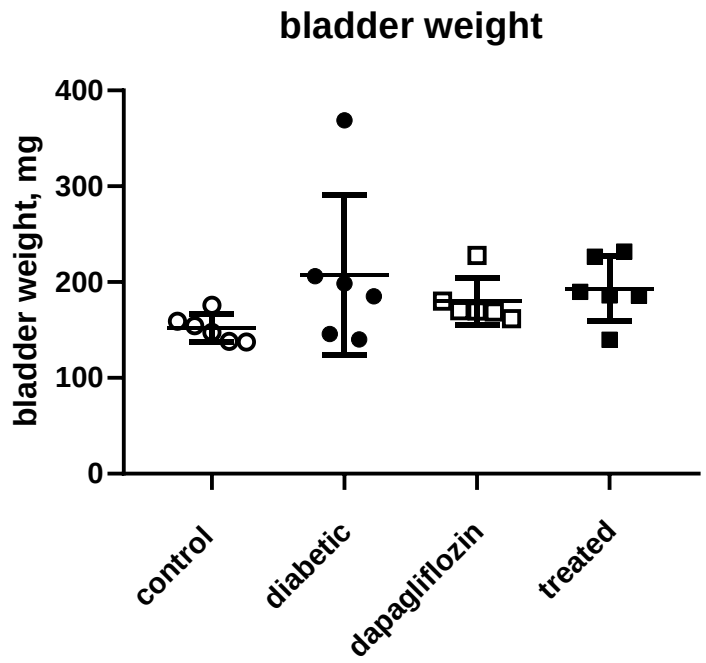
- Ad hoc outlier removal to obtain results you like is cheating
- It is not cheating if criteria for outlier removal had been pre-specified
  - In my experience, thinking of all possible types of outliers (and defining rules how to handle them) is easier said than done

- If there is more than one outlier, do not remove them sequentially
  - Each removal shifts the estimate of the underlying population

Is outlier removal legitimate?

- Pilot study to test whether experimental diabetes affects bladder weight and whether treatment (dapagliflozin) ameliorates this

- Is the highest point in the diabetic group an outlier?



Erdogan et al.; in preparation

Consequences

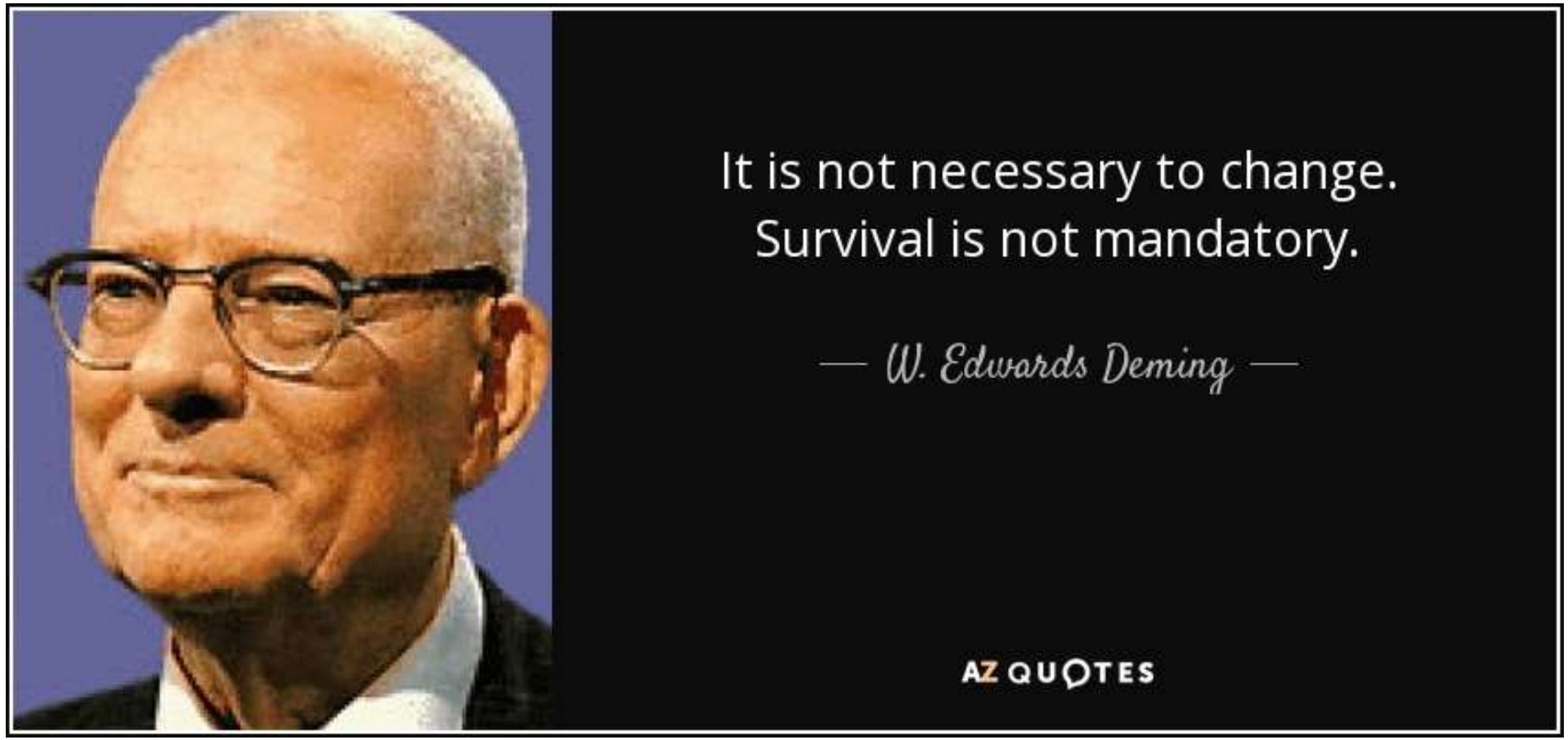
- A well designed experiment is less sensitive to outliers
  - Large sample size
  - Multiple replicates
- Be as conservative a possible in removing “outliers”
- “Informal” outlier removal should be done by someone not familiar with the experiment (blinded assessor)
  - Less bias
- Outliers can be removed from the analysis, but should never be removed from the record
  - See “exclude” function in Prism

Outlier tests

- An apparent outlier can have two reasons
  - Value comes from the same population as the others (just coincidentally from an extreme end)
    - Should not be removed
  - Can come from a different population
    - Biology e.g. gene polymorphism → could be informative
    - Mistake e.g. bad pipetting → should be removed
- But how to be certain which type applies?
  - You never really can

Lessons

- Analyses can be biased
- Differentiate exploratory and confirmatory work and be transparent about it
- Type of variable under investigation dictates which analyses are meaningful
- Assumption of Gaussian drives choice of meaningful analysis and reporting
- P-values are fickle
- Choice of statistical test affects P-value
- Effect size with confidence interval may be more informative than P-value
- Outlier removal can be legitimate, but measures should be taken to avoid bias
  - Preset rules
  - Blinding


It is not necessary to change.
Survival is not mandatory.
— W. Edwards Deming —
AZ QUOTES