Lecture 2
Intro into Quantitative Research

Statistical Methods

UNIVERSITY OF AUCKLAND

COMPSCI 705 / SOFTENG 702
Advanced Human-Computer Interaction (HCI)

Dr. Gerald Weber

Lecture Overview

- Median as robust statistic for nonparametric methods
- Confidence and Significance
- Statistical tests
- Parametric methods

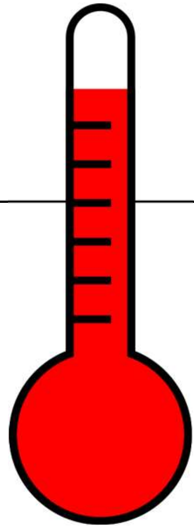
System Usability Scale (SUS)

SUS Measures subjective usability with a standard 5-point Likert scale:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Brooke, J. (1996). "SUS: a "quick and dirty" usability scale". in P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. *Usability Evaluation in Industry*. London: Taylor and Francis.

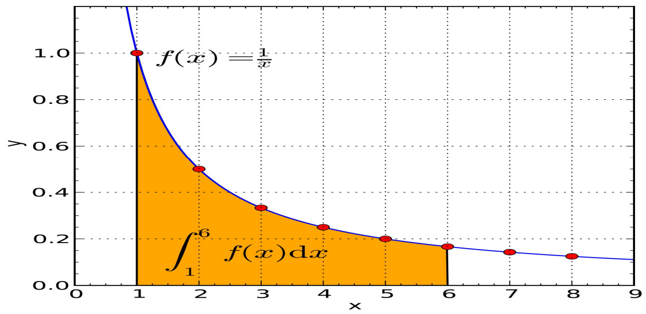
Measures of Central Tendency



Median: middle value
Mode: most frequent value
Arithmetic Mean:

- Sum over whole population, divided by population size: (a1+a2+….+an)/n.
- See also the expected value in probability theory E(X):
- E(X) is the arithmetic mean, if X is a uniformly random member of the population.

- Arithmetic mean is easy to compute incrementally.
- Requires interval variables.
- Not a robust statistic: can be heavily influenced by extreme points in skewed distributions:
- Especially by tail-heavy distributions.



Nonparametric vs parametric

- Nonparametric statistics:
  do not assume a specific distribution in the phenomenon.
- Are also working for ordinal data.

- Parametric statistics:
  do assume a distribution,
  often relying on assumption of normal distribution.

- Median fits naturally to nonparametric methods
- For arithmetic mean, often parametric methods are used.

Median: a robust statistic

- Works for ordinal variables; values have an ordering.
- Based on the concept of rank**:** Sort and number all values of the variable in the collection. ***rank*** = position in that order.
- ***Median***: the value with middle rank (position in that order).
- Often seen as a good choice of a "typical value" particularly for skewed distributions.
- Workaround for even number of values:
  - arithmetic mean of two middle values.
- But for uneven number of values:
  - Median is always an original data point.
- Also as percentile: boundary of 50^{th} percentile.
- Also as probability: 0.5 probability to be above (below) median.
- Robust statistic: Breakdown point of 50%: if less than 50 percent of data are outliers of a systematic kind (all too big), the median still gives a valid data point.

Random sampling and interval results

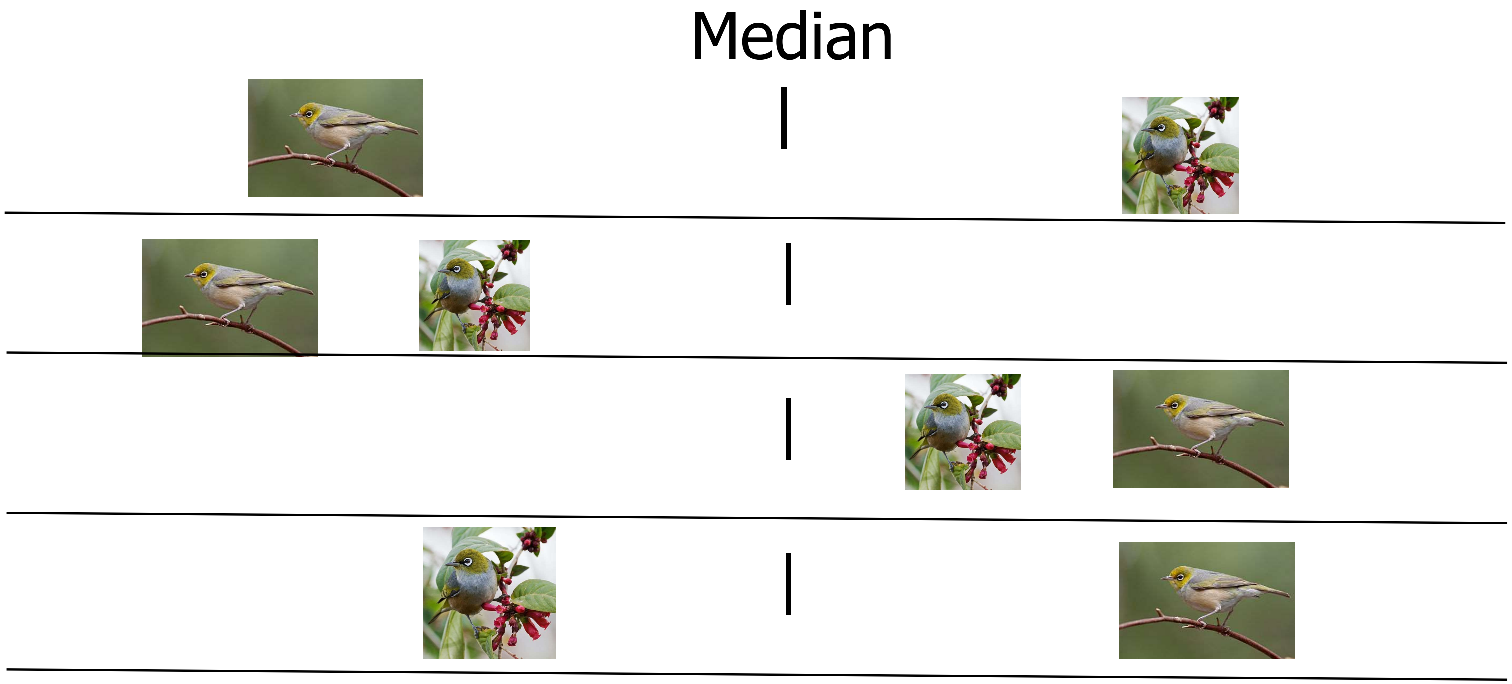
- To obtain a representative score for a system (e.g. the SUS score) we recruit a random sample of potential users.
- Different users will give different scores:
- The score we obtain, and any measure (especially median and arithmetic mean) will randomly fluctuate, depending on which participants we choose randomly.
- Therefore we know that a single mean or median is not giving enough information, we have to present an interval and a qualification how informative this interval is.

Estimating the median

- We want to measure = estimate the median weight of adult silvereyes.
- Assume we have weighed two randomly chosen adult silvereyes: 9 gram and 10 gram.
- Claim: with probability 0.5 (colloquially 50% chance) the population median weight of adult silvereyes lies between 9 and 10 gram.
- Is this true?


First estimate for the median.

- The following 4 possibilities are equally probable:



- Question whether a randomly chosen sample point is above or below median is like a random coin flip.
- With 50 percent probability the two first data points straddle the median.

Confidence

- If we take random samples, we know that the median/mean of the sample fluctuates.
- Therefore it is not enough to report only one value:
- Assuming a probability C, eg. 95% (meaning 0.95) is requested: A *C-confidence Interval* is an interval for which we know that the population average lies with probability C within the reported interval, assuming we have done good **random sampling.**
- E.g. Measuring the median weight of adult silvereye: we measure a random sample and report a weight of 9 to 11 gram with 95 percent confidence.
- Note: The confidence interval that we report will fluctuate as well! Would we have a different random sample, we would have reported a different interval;
- Confidence: With which probability is the median within the reported interval?

Symmetric Confidence Interval for Median

- The first 95% confidence interval (CI) can be obtained with 6 measurements: 1-2/64 = 1 -1/32 ~= 97%
- BUT: It is the interval between the smallest and largest value.
- No leeway for outliers.
- Notation: Interval (n, m) states the rank of the datapoints that we choose from our measurements sample.
- The first 95% confidence interval that excludes the two extreme measurements can be obtained with 9 measurements: Interval (2,8) has confidence 1- (2*10)/512 ~= 1- 1/25 ~= 96%
- These CIs are always between original data points.
- 12 measurements: (3,10) 96% confidence
- 15 (4,12) 96% confidence
- 17 (5, 13) 95,1%
- 20 (6, 15) 95,8 %

Significance

- Often we are not primarily interested in a median, but:
- For two measures m1, m2 on the same population, is one alternative better for most members of the population? E.g. mouse vs trackpad.

- Relation to confidence interval:
- We make a within-subjects experiment:
- Measure for m1, m2 for each participant.
- Take difference m1-m2 for each participant.
- If the confidence interval for the mean/median of the difference m2-m1 excludes the zero, then the result is significant.
- Note, this does not require that the confidence intervals for m1 and m2 are disjoint.
- Significance expressed as P value: 0.05 or 0.01 in contrast to confidence.

One-sided vs two-sided

- We can either be naïve concerning the direction of the outcome,
- Or we can expect a certain outcome (This expectation is a so-called prior)

- Naïve: two sided test:
- E.g mini keyboard vs handwriting.
- **H1:** For two measures m1, m2 on the same population, **one of them** is better for most members of the population.
- **H0:** both are equally good. (Null Hypothesis)

- With prior: one-sided test:
- E.g. mouse vs trackpad.
- **H1:** For two measures m1, m2 on the same population, **m1** is better for most members of the population.
- **H0:** m1 is not better. (Null Hypothesis)

Two-sided sign test

- A test for paired samples: For two measures m1, m2 on the same population, is one measure better for most members of the population? E.g. mini keyboard vs handwriting.

- Random sample consists of n measurement pairs.
- Sign test essentially computes the confidence interval for the median of the difference.
- Result is significant if the confidence interval excludes the zero.
- For significant result we need 6 measurement pairs, and they have to have all the same sign.
- For 9 measurements we can have one pair with opposite sign.
- 12 measurements: 2 pairs with opposite sign allowed
- 15 measurements: 3
- 17 4
- 20 5

One-sided sign test

- If we have fixed the hypothesis we need only a one-sided confidence interval:
- Only have to deal with outliers on one side.
- The interval is open on the other side, e.g. (1, …)
- P-value for given number of opposing signs is doubled.
- Hence p-value 0.05 is reached easier:
- Interval (1, …) has confidence 1-1/2n.
- For significant (p=0.05) result we need 5 measurement pairs, and they have to have all the same sign.
- For 8 measurements we can have one pair with opposite sign.
- 11 measurements: 2 pairs with opposite sign allowed
- 13 measurements: 3
- 16 4
- 18 5

Wilcoxon Signed-Rank test

- Sign test makes minimal assumptions about distribution of data.
- But needs relatively clear results: Statistical power is low.

- **Wilcoxon signed-rank test**: takes differences into account, has higher statistical power:
- E.g. if we have 9 pairs with two pairs with opposite sign: if the difference for these two pairs is small, then the Wilcoxon Signed-rank test might still give a significant result.
- Order differences according to their absolute size:
- -0.3, 0.6, -2.1, 2.2, 3.4, 5.1, 5.7, 6.1, 6.8
- So we get an ordered list of signs:
  \- + - + + + + + +
- 1 3 W-value is 3+1=4
- Look up in table for critical W values for Wilcoxon test
- For n=9, maximally allowed W-value is 5, for p=0.05,
- so result is significant!

Type I errors and Type II errors

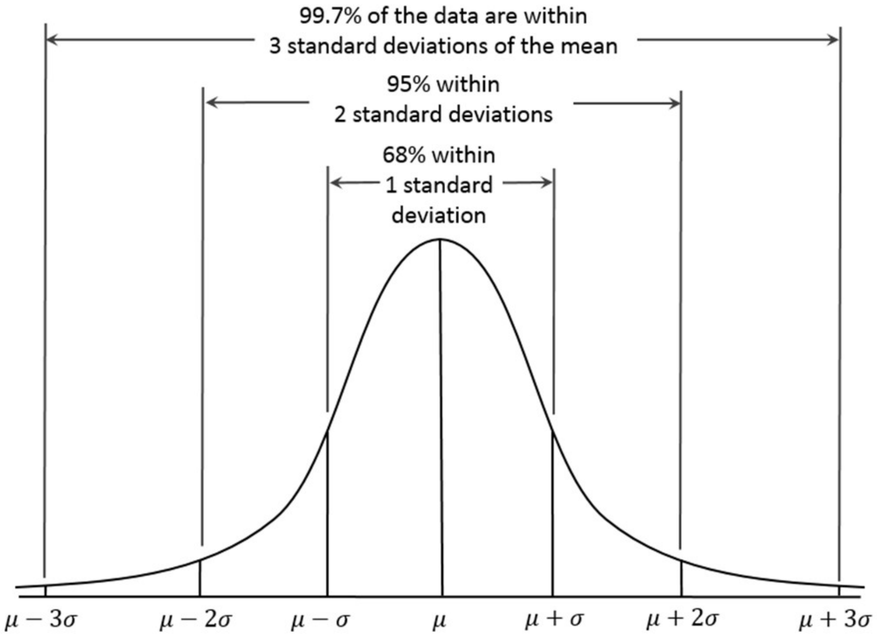
- **Type I error**:
  - Null hypothesis is rejected in experiment,
  - but should not be rejected, is true.
  - a.k.a. **false positive.**

- **Type II error**:
  - Null hypothesis is not rejected in experiment,
  - but should be rejected, is false.
  - a.k.a. **false negative**.

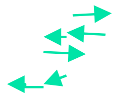
Normal Distribution

Normal Distribution:

- aka “Gaussian Bell curve”
- Symmetric, infinite distribution, never going to zero.
- Standard deviation: the x-value of the inflection points (where the curvature changes).
- Dying out rather quickly:
- Empirical rule 68-95-99.7: within 3 standard deviations lie more than 99 percent of values.
- Is result of a random zigzac walk: equal-sized legs of the walk added up.
- Appears as sampling distribution, e.g. distribution of mean with random sampling



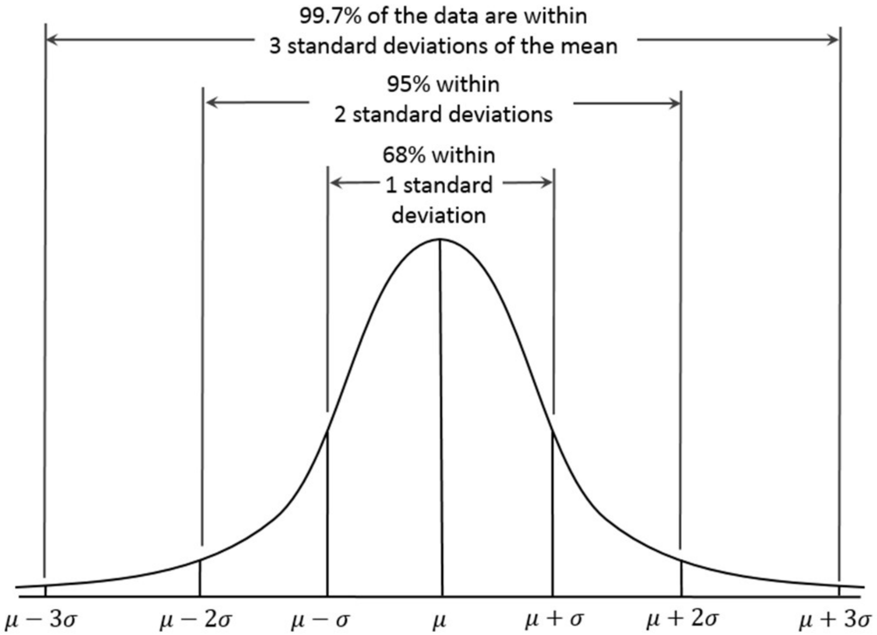
Dan Kernler, commons.wikimedia.org/wiki/File:Empirical_Rule.PNG



$$k\, e^{-x^2}$$

Confidence Interval of the mean

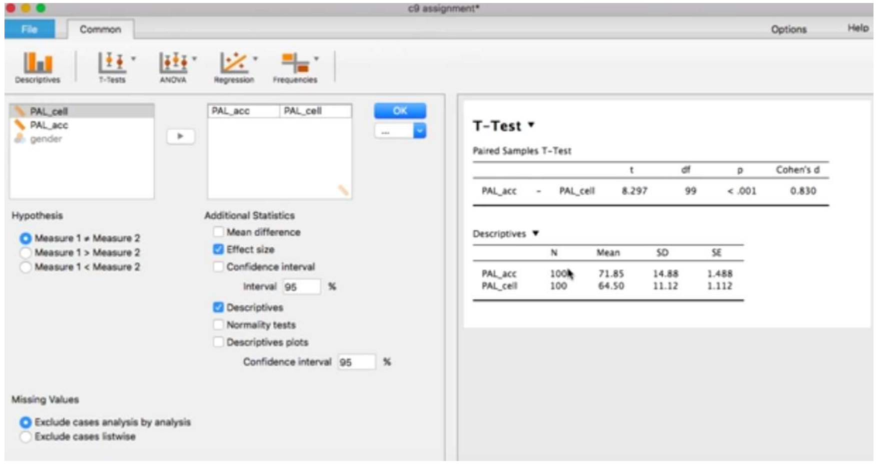
- Is typically given as a parametric confidence interval:
- Makes assumptions about the distribution (arguably always necessary for mean, because it is not robust)
- Sample mean m
- Standard deviation: σ
- Square root of sample size: √n
- Correction factor for desired level α of confidence: $t_{\alpha,n}$
  Can be looked up in table.
- Confidence interval:
- (m + σ $t_{\alpha,n}$ 1/√n, m - σ $t_{\alpha,n}$ 1/√n)



t-test

- A frequently used parametric test
- Used for hypothesis testing
- Exists as unpaired and paired test.
- Paired test: used for within-subject design

- Tool support for tests:
- E.g. Jasp, an open source tool designed for users familiar with SPSS.

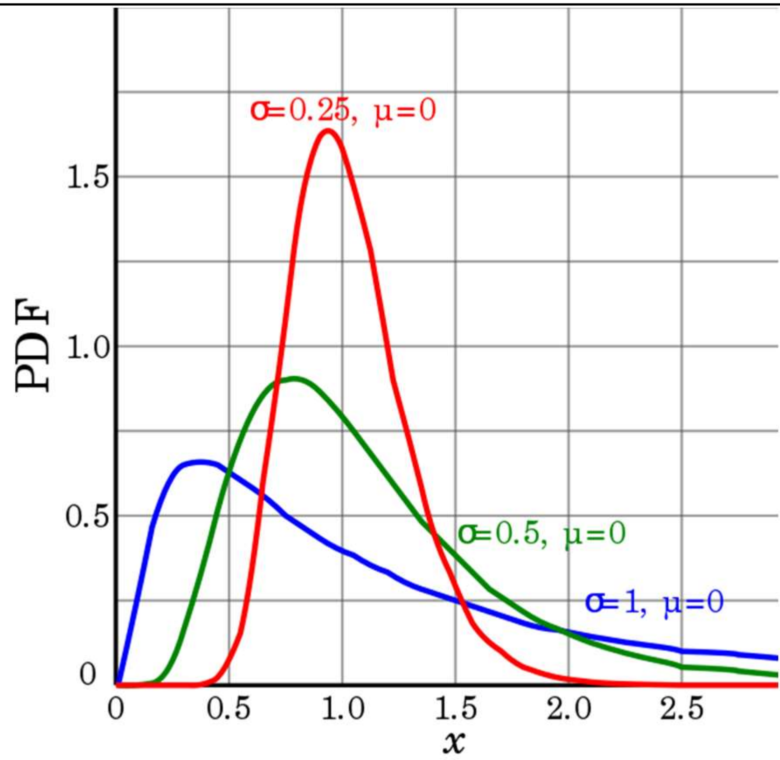


https://jasp-stats.org/

Log-normal Distributions

Log-normal Distribution:

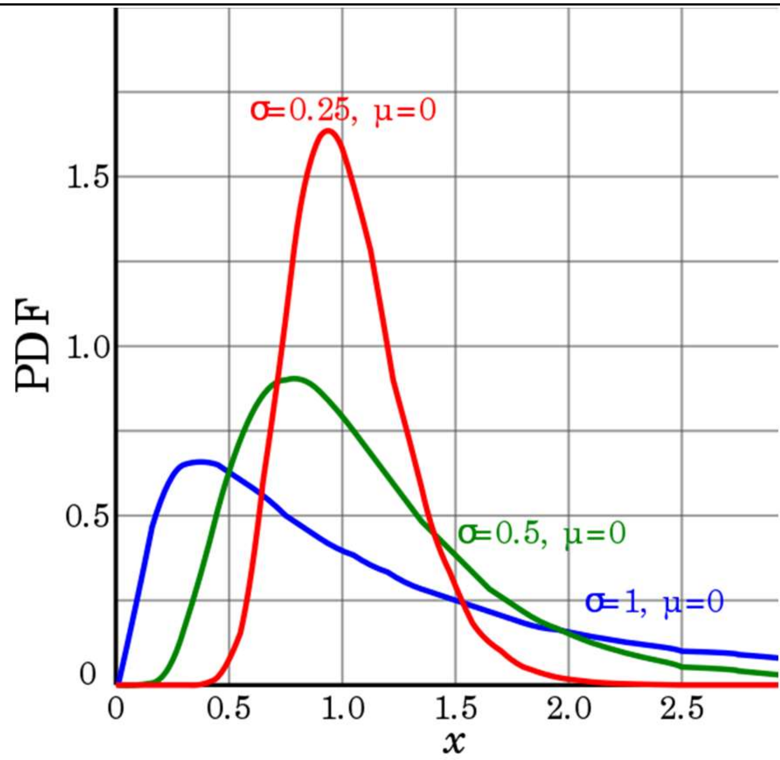
- Result of a random repeated multiplication with small factors around 1.
- Appears in many natural processes, e.g. growth processes
- Also many HCI phenomena:
  - Dwelltime on online content
  - Length of online content
- Positive/negative influences work as factor, not additive.
- By using logarithmic scale, factors turn into additive steps.
- Typical for parameters that are only positive.

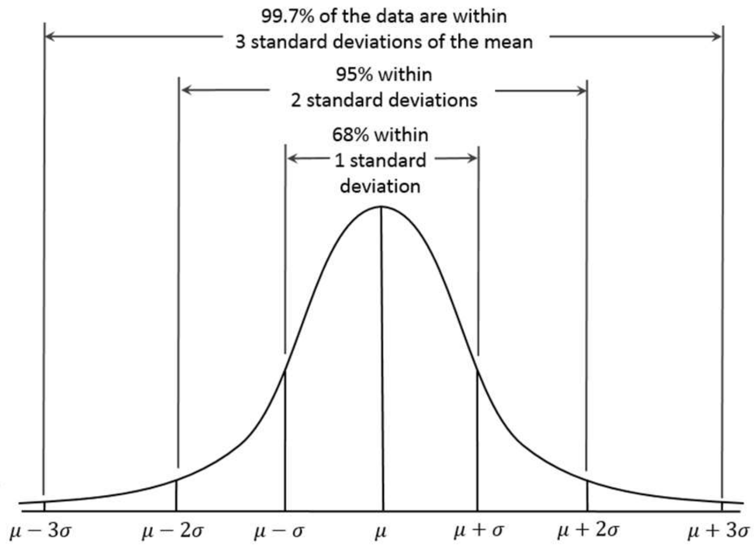


Different log-normal distributions with same arithmetic mean

Normal and Log Normal Distributions: Median vs. Mean

- Median gives for log-normal distribution the mean after taking logarithm.
- Example of the robustness of the median.
- Shows that median can be used without making assumptions on the underlying distribution.
- But: If it can be shown from the data that the distribution is e.g. log-normal, then transforming it and making use of the properties of the normal distribution can give stronger results.





Issues with low confidence levels

- The accepted confidence levels 95% and 99% are just a convention (stemming from parametric methods).
- For descriptive studies reporting e.g. a median: 1 in 20 of the 95% confidence intervals that we find are spurious.
- Type 1 error (false positive): The null hypothesis is true but is rejected.

- Since inconclusive studies are rarely reported: The reported results are selected for being affirmative of H1.
- It may be that false positive are overrepresented
- More than 1/20 of the reported positive 95% confidence intervals or p=0.05 significance tests will be false positives.

Summary 1

- Different measures of central tendency have different uses: mode, median, arithmetic mean.

- The confidence interval for the median is given by actual datapoints and is independent of the distribution.
- Confidence and significance are measures of whether our findings are just coincidence or true relationships: conclusion validity.
- Significance can be explained with confidence.
- Established mininmum requirement: 95% confidence resp significance of 0.05.
- We need enough data to exclude outliers.
- The Wilcoxon signed-rank test is considered stronger than the sign test.

Summary 2

- Different measures of central tendency have different uses: mode, median, arithmetic mean.

- Parametric methods are particularly suitable for data with roughly normal distribution.
- The parametric confidence interval for the mean is symmetric around the sample mean.
- The t-test is a popular parametric test.
- Parametric methods are not robust against outliers

- The Wilcoxon Signed-Rank test is a nonparametric alternative to the t-test and more robust