Research Methods

Hypothesis Testing 2

Dr. Sven Magg, Prof. Dr. Stefan Wermter

http://www.informatik.uni-hamburg.de/WTM/

Plan for today!

1. t-Tests and t-distribution
2. Test hypotheses about correlations
3. What does the p-Value mean
4. How to correctly use the p-Value
5. Can samples be too big?

Recap: Z-Test

- The Z-Test does 3 things:
  - Estimates the sampling distribution of the mean
  - Transform this distribution into a standard normal distribution
  - Express sample mean $\bar{x}$ as Z standard deviations from μ
- Z-Score: $Z = \frac{(\bar{x}-\mu)}{\sigma_{\bar{x}}}$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$
- Find critical values: $p \leq 0.05$, 2-tailed $\Rightarrow \bar{x}_{crit} = \mu \pm 1.96\sigma_{\bar{x}}$

- What if I don't know the population's standard deviation σ?
  - We can estimate it from the sample standard deviation s:

$$\hat{\sigma} = s$$

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{N}}$$

t-Test

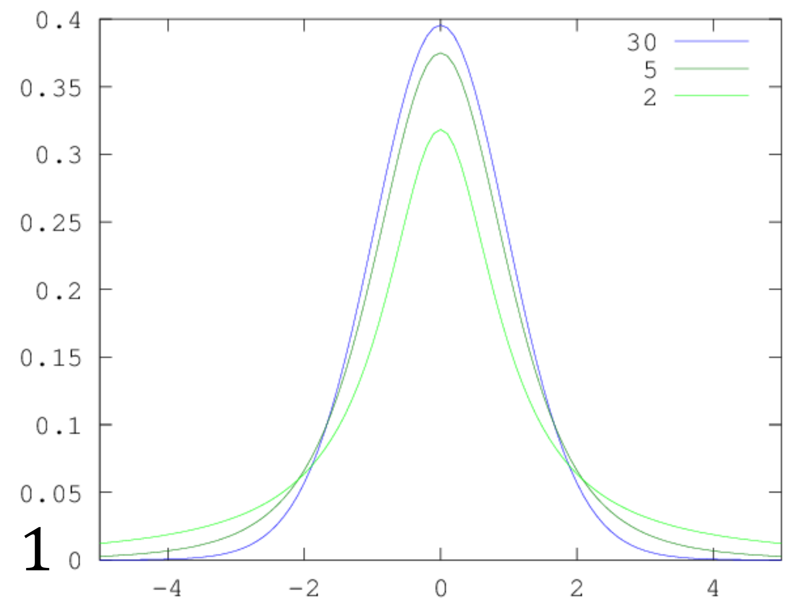
- What can we do if we have a sample with a size smaller than 30?
- Set of sampling distributions for small N: t-distributions

t-Score

$$t = \frac{(\bar{x} - \mu)}{\hat{\sigma}_{\bar{x}}} = \frac{(\bar{x} - \mu)}{s/\sqrt{N}}$$

- Look up t-score in a table for distribution with corresponding degrees of freedom
- For tests with one mean:

$$df = N - 1$$



t-Test

- Calculate t-score for 2-tailed test of size 5:

$$\mu = 1.$$
$$\bar{x} = 4.$$
$$t = \frac{3.5}{1.5} = 2.33$$

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

from Wikipedia.com

- We can only reject with $p \leq .1$ (with Z-Test: p=.02)

- Two-sample t-Test
  - What if we have two means and want to see whether they are drawn from two different populations?
  - Common test in many experiments

Two-Sample t-Test

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$(two-tailed test) or $\mu_1 > \mu_2$(one-tailed test)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}}}$$

Which $\hat{\sigma}_{\bar{x}}$ to use??

- Let's remember what the sample standard deviation was:

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{N}} = \sqrt{s^2/N}, \qquad s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} = \frac{SS}{df}$$

- Under H_0, all values were drawn from the same population, i.e. the variances have to be pooled

Pooled Variance

$$s_{pooled}{}^2 = \frac{\sum_{i=1}^{k}(N_i - 1)s_i{}^2}{\sum_{i=1}^{k}(N_i - 1)} = \frac{(N_1 - 1)s_1{}^2 + (N_2 - 1)s_2{}^2}{N_1 + N_2 - 2}$$

$$\hat{\sigma}_{pooled} = \sqrt{\frac{s_{pooled}{}^2}{N_1 + N_2}}$$

- Example:
  - $\bar{x}_1 = 87,\ s_1 = 5.7,\ N_1 = 23$
  - $\bar{x}_2 = 95,\ s_2 = 3.2,\ N_2 = 15$
  - $s_{pooled}{}^2 = \frac{(22)32.49+(14)10.24}{36} = 23.837,\ \ \hat{\sigma}_{pooled} = 0.792$
  - $t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{pooled}} = \frac{-8}{0.792} = -10.1\ (p = 4.7 * 10^{-12})$

Paired t-Test

- Assume you have two algorithms A and B and want to compare performance
- One testing strategy:
  - Select 10 random tests for each $T_1 - T_{10}$ and $T_{11} - T_{21}$
  - Calculate means and standard deviation for both
  - Run 2-Sample t-Test
- Better strategy:
  - Run both algorithms on the same 10 tests
  - Calculate differences δ in performance for each test
  - Calculate mean $\bar{x}_\delta$ and s_δ standard deviation
  - $H_0: \mu_\delta = 0,\ H_1: \mu_\delta = k$

Paired t-Test

- Use a paired t-Test when the samples are dependent
  - Same sample has been tested twice
  - Samples have been matched into meaningful groups
- Now run 1-Sample t-Test using $\bar{x}_\delta$, s_δ and N_δ :

$$t = \frac{(\bar{x}_\delta - \mu_\delta)}{s_\delta / \sqrt{N_\delta}}$$

- N_δ is number of pairs and $df = N_\delta - 1$

- This way we can reduce the variance due to test problems

$\Rightarrow$ Boosting significance by reducing variance of tests

t-Test Summary

- Can be used for all sample sizes, for larger N it approaches the Z-distribution
- Common usage: Comparing two means
- Parametric Test!
  - Assumptions:
    - samples are drawn from normal distribution
    - samples are independent
    - sample variances are equal
  - Pretty robust against violations of normality
  - Pretty robust to variance heterogeneity if sample sizes equal
  - Not robust to violations of independency!

Correlations

Pearson's Correlation Coefficient

$$r_{XY} = \frac{cov(x,y)}{s_X s_Y} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_X s_Y}$$

- r_{XY} is a value between -1 and 1 and is a measure of linear association between two variables X and Y
- Used to test independence of two variables
- Can we test the hypothesis that the correlation is zero?
- We need a sampling distribution of the correlation coefficient! Complicated!
- We can get them by empirical sampling or transformation

Fisher's r-transform

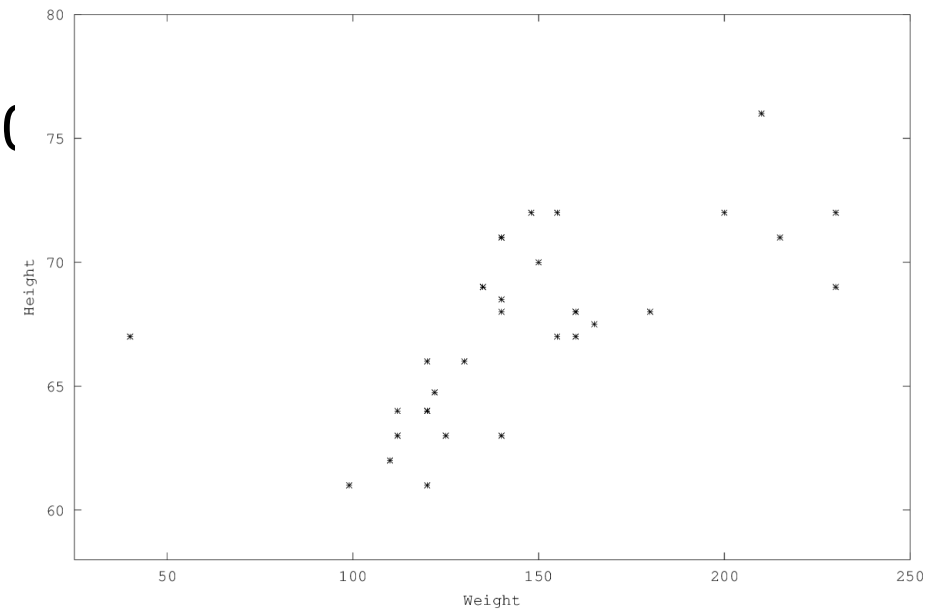
$$z(r_{XY}) = 0.5ln\frac{1 + r_{XY}}{1 - r_{XY}}$$

- Transforms r_{XY} to produce a sampling distribution which is approximately normal
  - Assumption: The variables are normally distributed
- Mean: $z(\rho) = 0.5ln\frac{1+\rho}{1-\rho}$, where ρ denotes the population correlation
- Estimated standard error: $\hat{\sigma}_{z(r)} = \frac{1}{\sqrt{n-3}}$
- Now we can test the hypothesis $H_0: \rho_{XY} = 0$ or $H_0: \rho_{XY} = k$

Fisher's r-transform

- Height and weight of 33 students
- Correlation Coefficient:

$$r_{XY} = ($$



$$z(r_{XY}) = 0.5ln\frac{1+0.64}{1-0.64} = 0.759$$

$$\hat{\sigma}_{z(r)} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{30}} = 0.183$$

- Now we can use a Z-Test to test our $H_0: \rho_{XY} = 0$

$$Z = \frac{(z(r_{XY}) - z(\rho))}{\hat{\sigma}_{z(r)}} = \frac{0.759-0}{0.183} = 4.16$$

- Clearly outside our critical values of +/- 1.96 (p=.00003)

Recap

- We defined a general process with 5 steps
  1. State null Hypothesis H_0
  2. Gather a sample statistic (run experiment)
  3. Find sampling distribution N_h, assuming H_0 is true
  4. Calculate p-value using the sampling distribution N_h
  5. Use p-value as evidence against H_0

- We need a hypothesis which is falsifiable in practice
- One problem is to find the sampling distribution
- Levels commonly used for rejection of H_0: p=.05 and p=.01

Recap

- We also defined another process using cut-off points
  1. State null Hypothesis H_0 and alternate hypothesis H_1
  2. Gather a sample statistic x (run experiment)
  3. Find sampling distribution, assuming H_0 is true
  4. Set maximum acceptable probability α
     - Find cut-off points c^+ and c^- such that
       $$\mathrm{P}(N_h \geq c^+) + P(N_h \leq c^-) \ \leq \ \alpha$$
  5. Decide: If $(x \geq c^+)\ or\ (x \leq c^-)$, reject H_0
     - Reject H_0 if x falls into rejection regions defined by α

- Common levels for α: α=.05 or α=.01

History excursion

- "Early" Ronald Fisher [1925]
  - Inductive inference: Use direct probability P(Data| H_0)
  - Only use a Null-Hypothesis H_0
  - Use known distribution of a test statistic T, assuming H_0
  - Set significance level (.05/.01/.001) following a convention
  - Calculate p-value to check whether there is a significant (= backed by statistics) divergence
  - Significance value is a genuine feature of the test
- "Late" Ronald Fisher [1956]
  - Calculate exact p-value from the data
  - Significance level is a feature of the data themselves
  - No need for an arbitrary convention

History excursion

- Ronald Fisher combined
  - Use known distribution of a test statistic T, assuming H_0
  - Determine density of values that exceed observed value
  - Use p value as strength of evidence against H0
  - p-Value is sample-based measure of evidence against null hypothesis
  - We report exact p-value, **NOT a decision**

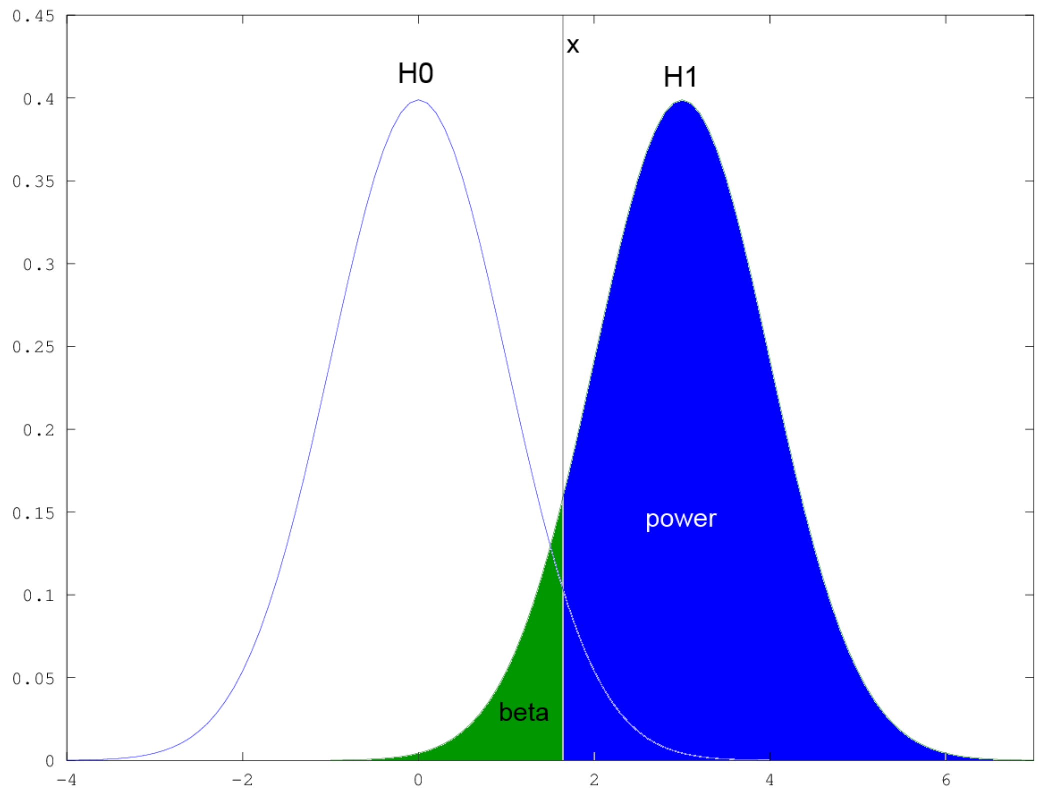
p-value	Strength of evidence
0.100	Borderline (or weak)
0.050	Moderate
0.025	Substantial
0.010	Strong
0.005	Very Strong
0.001	Overwhelming

Neyman-Pearson

- Neyman-Pearson: [1928]
- Define H_0 and alternative Hypothesis H_1
- There are two errors you can make:
  - Type I: False rejection (probability α)
  - Type II: False acceptance (probability β)

	H_1 is true	H_0 is true
Reject H_0	Correct Outcome Power (1-β) True Positive (TP)	Wrong Outcome Type I (α-)Error False Positive (FP) Significance Level
Accept H_0	Wrong Outcome Type II (β-)Error False Negative (FN)	Correct Outcome Specificity (1-α) True Negative (TN)

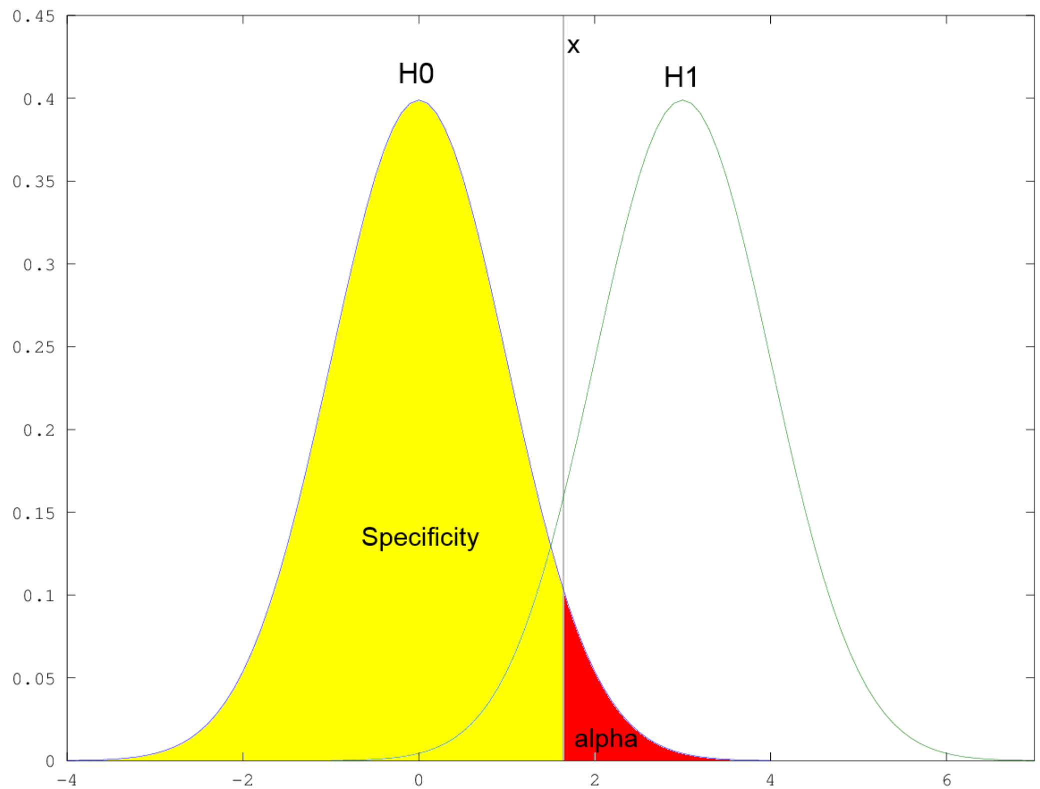
Group Task!



Colour the regions for α, β,power and specificity

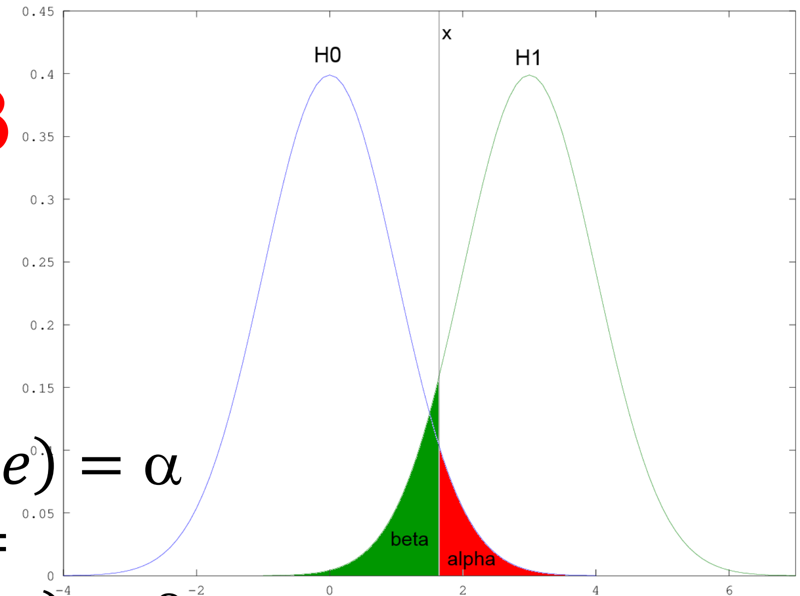
Group Task!

2 5



Colour the regions for α, β,power and specificity

α and β



- Error probabilities:

$$P(Type\ I\ Error) = P(Reject\ H_0|H_0\ true) = \alpha$$

$$P(Type\ II\ Error) = P(Accept\ H_0|H_1\ true) = \beta$$

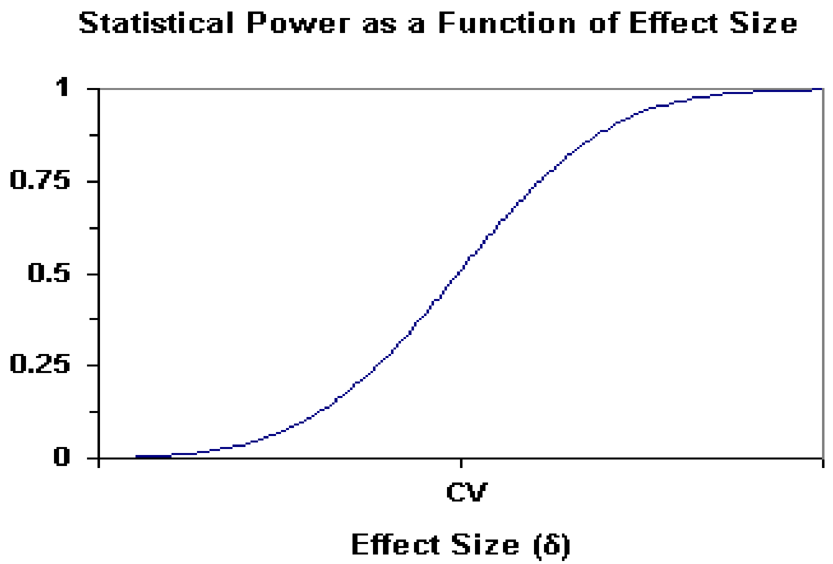
- The power of the test ($= P(Accept\ H_1|H_1\ true)$) is dependent on:
  - α, which is set a priori
  - the degree of separation between the two distributions given by $\delta = \mu_1 - \mu_0$
  - the variance(s) of the population(s)
  - the sample size N

Power of a test

- Analogy
  - You are searching for an item in your room
  - Power: “What are your chances that you would find the item”
  - Depends on:
    - How long you are searching **Sample Size**
    - The size of the tool **Size of effect, i.e. degree of separation**
    - The messiness of the room **Standard deviation**
  - There is a high chance to find a large item in a clean room if you spend a long time searching!
  - If you can’t find it, you can be confident it wasn’t there
  - Power of an experiment: If there really is an effect, how high are the chances that the experiment would find it?

How to get the power?

- Power often shown as power curves for one of the four parameters α, δ, variances (σ_1 and σ_2) , and N
- Calculating the power of a test is usually difficult and tedious
- Using computers:
  1. Fix all parameters but one
  2. Draw samples from both populations
  3. Run the test and see whether you would reject H_0
  4. Repeat 2.& 3. k times and calculate #rejections/k
  5. Repeat 2.- 4. for all needed values of the free parameter



Neyman-Pearson

- First define α and β before running the test
- α and β are probabilities if making errors of type I / II in the long run and therefore features of the test
- Set α and β not by a convention but after a detailed cost-benefit analysis of the consequences

- You have to think prior to the experiment about meaningful values for α and β
  - What are the consequences of making a type I or II error?
  - Values often used: α=.05 and β=.2 (= power of 80%)
  - Sometimes you want a power of almost 100% and can accept a high probability of errors of type I

Neyman-Pearson

- Rules of inductive behaviour
- Rule gives you decision (reject/accept H_0) without final statement whether we believe H_0 is true/false
- A optimal statistical test minimises β while keeping α at a set bound.
- Nowadays often mixed forms of Fisher/Neyman-Pearson is used
  - Report significance level (usually 1-3 stars or “ns”) and exact p-value
  - Define α and use it to reject hypotheses after calculating p
  - Use a conventional α and β of 0.05 and .2
  - etc….

What should you do?

- Decide for one side of the debate!
- Either:
  1. Think about and set α and β **before** the test and report findings as significant or not, stating the significance level "There was a significant effect (α=.05)" or ".. ($p \leq .01$)", etc.
  2. Calculate p-Value for the sample **after** the test and report exact p-Value without reporting a decision about H_0

- In the first case, think about the consequences of your errors.
- If you can't think of any: Use 2.

Sample Size

- You can increase the power of the test by increasing N
- But should you?

- We do **not** increase the confidence by increasing N!
  - If we want to estimate a parameter, sample size should be as large as you can afford
  - If we want to test a hypothesis, samples should be no larger than required to show the effect
- If we test a hypothesis and are confident to reject H_0 with sample size N, we don't gain anything by increasing N
- On the contrary…..

Sample Size

- Can samples be too big?
- By increasing N you can
  - boost any real effect to significance
  - boost any meaningless effect to significance
- Do not fish for significance!

- What we want to know: How much predictive power does our result have?
- In other words: Does knowing which population the sample came from give us the power to predict it’s value?

Sample size

Sample	$\bar{x}$	s	N
A	147.95	11.10	1000
B	146.77	10.16	1000
A&B	147.36		2000

- Example from Cohen
- $t = 2.468$ with 1998 degrees of freedom
- Significant difference ($p \leq .05$)
- If I hand you one sample from A and let you guess whether it is above the combined mean A&B, how well would you do?
- 517 values from A exceed the combined mean
- 464 from B do as well
- You would guess correctly for 51.7% of samples!
- If you don't know the origin of the sample: 50%

Sample Size

- We can roughly estimate predictive power
- If we know (or can estimate):
  - the population variances $\sigma_A{}^2$ and $\sigma_B{}^2$ of populations A & B
  - The variance $\sigma_P{}^2$ of the combined population
- then predictive power means reduction in variance from knowing the population
- Relative reduction by knowing sample is from A:

$$\omega^2 = \frac{\sigma_P{}^2 - \sigma_A{}^2}{\sigma_P{}^2}$$

- If $\omega^2 = 0$, then $\sigma_P{}^2 = \sigma_A{}^2$ and no prediction is possible
- $\omega^2 = 1$ means no variance in the population, i.e. perfect prediction

Sample Size

- We can only estimate ω^2 since we try to find the population parameters!
- If both variances are equal for both populations, a rough estimate is:

$$\widehat{\omega}^2 = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1}$$

- In the example: $\widehat{\omega}^2$=0.0025
- If we would have drawn only 100 samples each: $\widehat{\omega}^2$=0.027

- Increasing N decreases the predictive power and therefore the meaningfulness of significant findings

What have we learned?

1. How to use a t-Test for low numbers of samples
2. How to test hypothesis about two means
3. Test hypothesis about correlations with Fisher's r-transform
4. Fisher's testing strategy using sample-based p-values
5. Neyman-Pearson's strategy to set α and β
6. We know now how to report significance (or when not)
7. Increasing sample size will increase the chance to find effects where there are none!