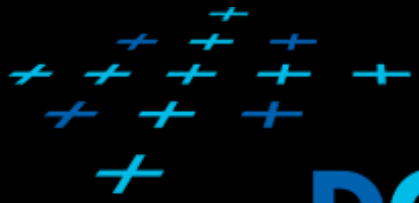


DCGI

DEPARTMENT OF COMPUTER GRAPHICS AND INTERACTION
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Experiment Evaluation and Power Analysis

SAN 2018/19

ERRORS IN EXPERIMENTS

- Type I error (False positive, α error)
  - H_0 is rejected, when in reality it is correct
- Type II error (False negative, β error)
  - H_1 is not accepted, when in reality it is correct

	H0 not rejected	**H1 accepted**
H0 is truth	Correct	Type I error
H1 is truth	Type II error	Correct

SOURCES OF ERRORS

- 1. Usability properties identification
- 2. Prototype creation
- 3. Experiment design
- 4. Participants recruitment
- 5. Test execution and data collection
- 6. Data analysis
- 7. Conclusions and recommendations statement

SOURCES OF ERRORS | CONT.

- 3. Experiment design
  - poor choice of task mix => indistinguishable results
  - wrong choice of participants => misleading results
    - unaware mixing novice and expert users can seem like design improvement or vice versa
  - accidental changes in the test conditions => insignificant or misleading results
    - large spread of measured values => insignificant results
    - shift of measured values => misleading results
- 6. Data analysis
  - analysis of influence of test conditions on the data measured
  - evaluator bias => analysis performed by more evaluators

DATA ANALYSIS | OUTLIERS

- Outliers are always there
  - but more often for "long tail" distributions

- Outliers elimination
  - selection bias => "data fishing"
  - before looking at the data measured (step 6)
  - better: before test execution (step 5)

  - perform qualitative evaluation of outliers behavior

	method A					method B				
min	26	24	22	17	15	10	9	8	7	6
max	94	98	75	82	72	41	39	31	29	27

SAN 2018 experiment

POWER ANALYSIS

- Power of a test = (1 – β)
  - probability that the test correctly rejects H_0

$$\text{power} = \mathbb{P}(\text{reject} H_0 | H_1 \text{ is true})$$

- Depends on
  - significance level α (Type I error probability)
  - sample size n
  - effect size d (min. degree of violation of H_0)
    - specify on a priori grounds

t test: $Cohen's\ d = \frac{\mu_1 - \mu_2}{\sigma}$

POWER ANALYSIS | SIZE *d*

- t tests
  - Cohen's suggestion: 0.2, 0.5, 0.8

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- ANOVA
  - Cohen's suggestion: 0.1, 0.25, 0.4

$$f = \sqrt{\frac{\sum_{i=1}^{k} p_i * (\mu_i - \mu)^2}{\sigma^2}}$$

$p_i = n_i/N$
n_i = number of observations in group i
μ = grand mean

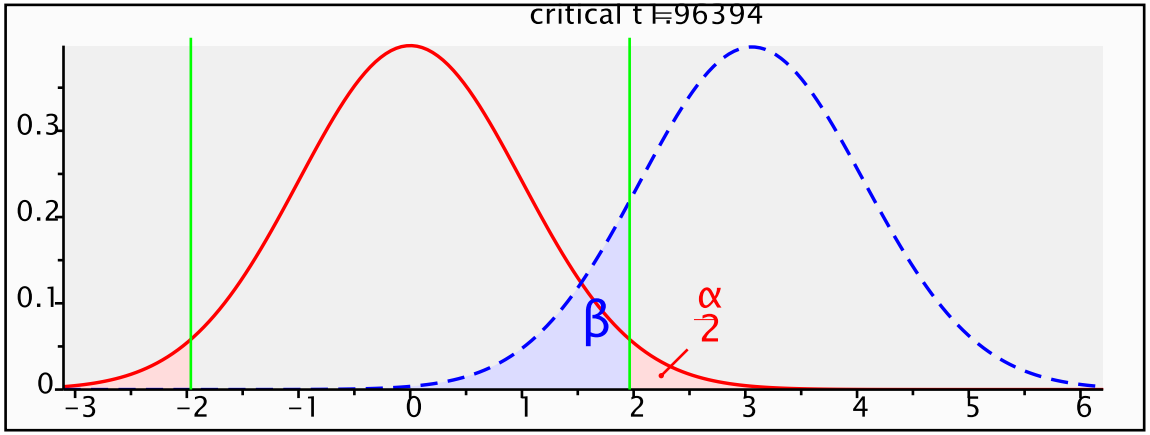
- Chi-square test
  - Cohen's suggestion: 0.1, 0.3, 0.5

$$w = \sqrt{\sum_{i=1}^{m} \frac{(p0_i - p1_i)^2}{p0_i}}$$

$p0_i$ = cell probability in i^{th} cell under H_0
$p1_i$ = cell probability in i^{th} cell under H_1

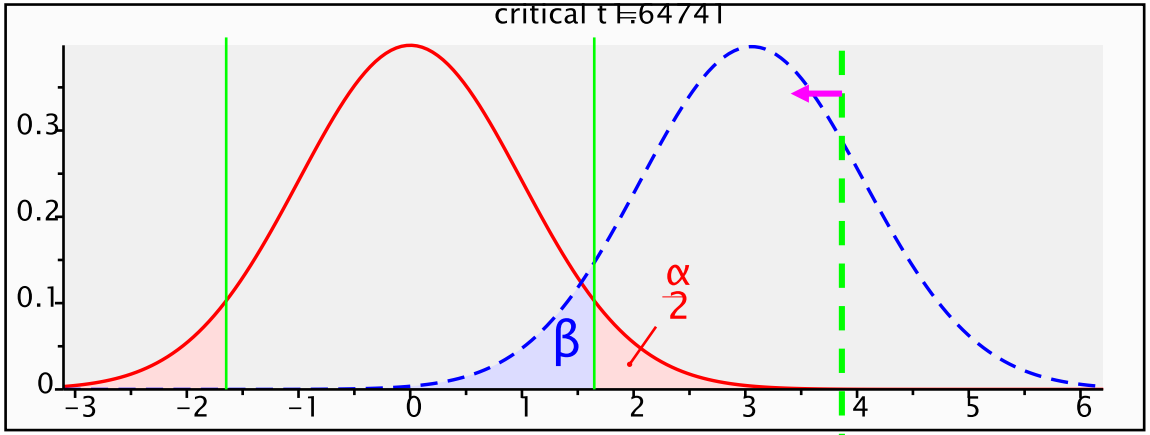
POWER ANALYSIS | DEPENDENCE

t test (difference between two independent means)



$\alpha = 0.05$

$\beta = 0.14$



$\alpha = 0.1$

$\beta = 0.08$

POWER ANALYSIS | TYPES

- A priori
  - controlling power level before conducting test
  - computing sample size *n*
  - function of required power level, specified α, *d*
- Post hoc
  - after a test was conducted
    - Does the test had fair chance to reject incorrect H_0?
  - computing the power level
- Compromise
  - fixed ratio between α and β
- Sensitivity
  - estimating/checking the size of an effect *d*

POWER ANALYSIS | DISCOVERY

- How many users do we need for discovering 95% of (**ALL**) problems?
- Golden rule of usability testing: Five users is enough to observe **all relevant** problems with **very high** probability.

- To detect X % of problems that affects Y % of users.
- To have a X % chance of detecting …

$$n = \frac{\ln(1 - X)}{\ln(1 - Y)}$$

n = 5
very high = 95 %
all relevant = 50 %

POWER ANALYSIS | COMPARING

- Determining *n* for comparing two means
  - within-subject

$$n = \frac{(t_\alpha + t_\beta)^2 s^2}{d^2}$$

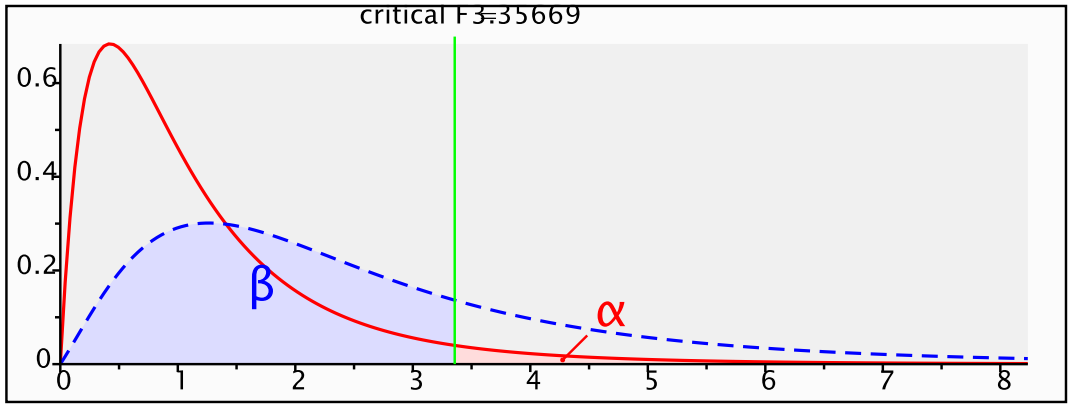
$t_\alpha = critical\ value\ for\ Confidence\ level$
$t_\beta = critical\ value\ for\ Power$
$s^2 = the\ variance\ (estimate\ of\ SD^2)$
$d^2 = the\ square\ of\ critical\ diference$

  - between subject

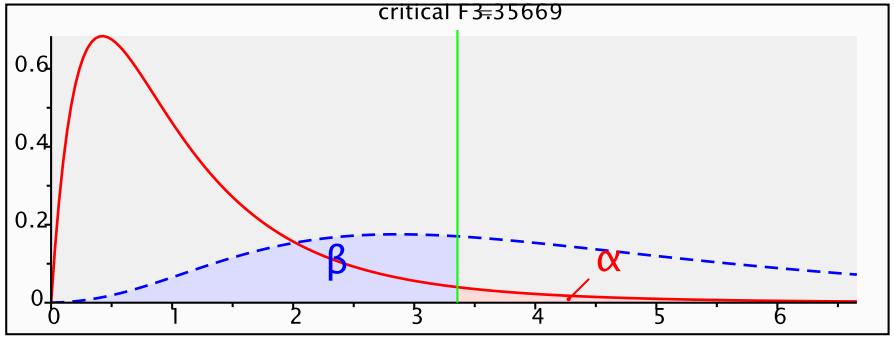
$$n = \frac{2(t_\alpha + t_\beta)^2 s^2}{d^2}$$

POWER ANALYSIS | COMPARING

F test (MANOVA: Repeated measures, within factors)



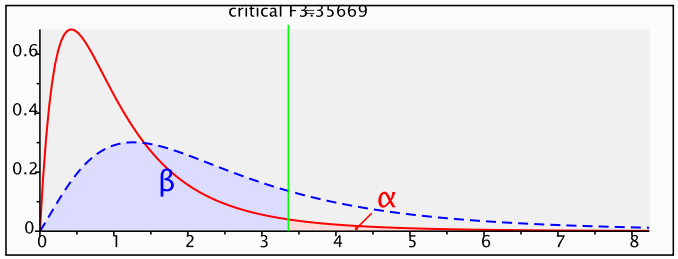
$\alpha = 0.05$
$\beta = 0.73$
$f = 0.25$ (medium)
$n = 16$



$\alpha = 0.05$
$\beta = 0.37$
$f = 0.4$ (large)
$n = 16$

POWER ANALYSIS | COMPARING

F test (MANOVA: Repeated measures, within factors)

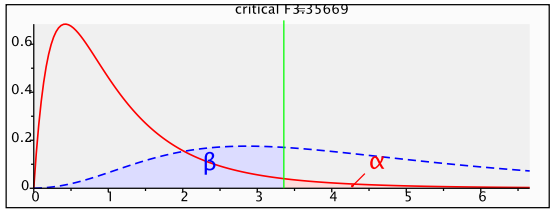


$\alpha = 0.05$

$\beta = 0.73$ for $\beta = 0.2, n = 44$

$f = 0.25$ (medium)

$n = 16$

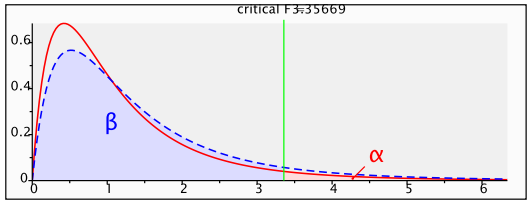


$\alpha = 0.05$

$\beta = 0.37$ for $\beta = 0.2, n = 22$

$f = 0.4$ (large)

$n = 16$



$\alpha = 0.05$

$\beta = 0.92$ for $\beta = 0.2, n = 244$

$f = 0.1$ (small)

$n = 16$

EXPERIMENT RESULTS

F test (MANOVA: Repeated measures, within factors)

```
Keyboard type means:
   A=41.86400
   B=14.40800
Group means:
   AB=29.92800
   BA=26.34400
```

Effect	df	SS	MS	F	p
Group	1	1605.632	1605.632	3.020	0.08865
Participant(Group)	48	25519.320	531.653		
Keyboard type	1	94228.992	94228.992	341.435	0.00000
Keyboard type x Group	1	1083.392	1083.392	3.926	0.05330
Keyboard type_x_P(Grou	48	13247.016	275.979		
Trails	4	8265.372	2066.343	107.509	0.00000
Trails_x_Group	4	38.148	9.537	0.496	0.73855
Trails_x_P(Group)	192	3690.280	19.220		

SAN 2018 experiment

THANK YOU FOR ATTENTION

DEPARTMENT OF COMPUTER GRAPHICS AND INTERACTION
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Zdeněk Míkovec
xmikovec@fel.cvut.cz

REFERENCES

- Power Analysis in R: http://www.statmethods.net/stats/power.html
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, ISBN 978-0805802832, Routledge.
- Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Elsevier.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). *Statistical power analyses using G* Power 3.1. 9: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences*. Behavior Research Methods, 41(4), 1149-1160.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). *Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses*. Behavior research methods, 41(4), 1149-1160.
- Kuniavsky, M. (2012). *Observing the user experience: a practitioner's guide to user research*. Morgan kaufmann.
- MacKenzie, I. S. (2012). *Human-computer interaction: An empirical research perspective*. Newnes.