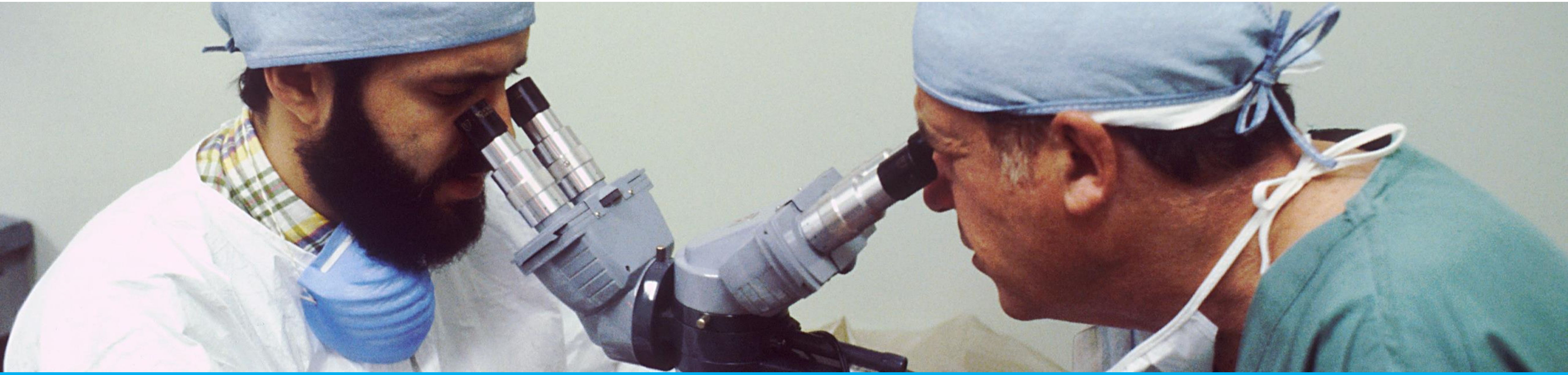


Hypothesis Testing

Image Source: https://pxhere.com/en/photo/1412604

Valentin Schwind

Learning Goals

- Getting an overview about hypothesis testing
- Learning about operationalization of concepts
- Learning more about statistical significance
- Learning about error types in hypothesis testing
- Understanding methods for increasing the statistical power

Research Question

- Must ask for new knowledge
- Formulation of the goal of a research project. It can be
  - answered in whole
  - answered in part or under certain circumstances
  - rejected as unanswerable
  - only an apparent problem
- Research questions often test one (or more) hypotheses within a paradigm or theoretical framework
  - A research question is a more general concept of a hypothesis
    - e.g., “Is there an Uncanny Valley of animals?” [1]

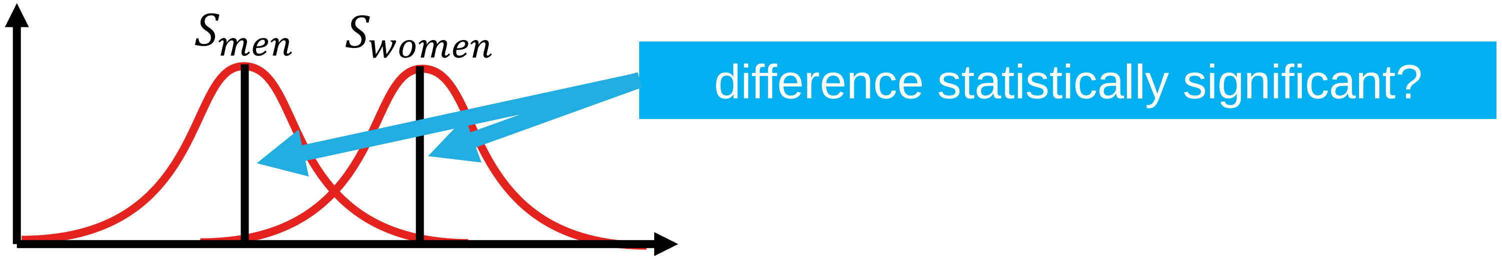
[1] V. Schwind, K. Leicht, S. Jäger, K. Wolf, N. Henze, Is there an uncanny valley of virtual animals? A quantitative and qualitative investigation, International Journal of Human-Computer Studies,

Hypothesis vs Theory?

- A hypothesis…
  - is a proposed explanation (for a phenomenon)
  - is a logical consequence („if… then“…)
  - can be tested
- A *working* hypothesis…
  - is a hypothesis that is *provisionally* accepted as a basis for further research
- A theory…
  - is an abstract and generalized thinking about a phenomenon
  - is a group of logical explanations based on empirical data

Types

- Alternative Hypothesis (“H1”, “H2”…)
  - e.g., *“There is a difference in typing speed between males and females“*
  - Directional Hypothesis („H1a”):
    - e.g., *“Males have a lower typing speed than females“*
- Null hypothesis (“H0”)
  - e.g., *“There is no difference in typing speed between males and females“*



Types

- Deterministic
  - e.g., *“The difference in typing speed between males and females is between 12-19 WPM.“*
- Probabilistic
  - e.g., *“The difference in typing speed between males and females is between 12-19 WPM with a probability of 75%.“*
- Classifying
  - e.g., *„People that are trained on typing with keyboards have an increased typing speed to those that are not trained.“*
- Comparative
  - e.g., *„The more training people have in typing on keyboards, the higher their typing speed.“*

Five What‘s

- What is the research question?
- What is the hypothesis?
- What is the correct test for the hypothesis (falsifiability)?
- What are the independent variables?
  - Is the factor within or between subjects?
- What are the dependent variables?
  - What is the concept that should be measured?
  - Objective or subjective?
    - e.g., performance, usability, fun, immersion, fitness, health, …

  → What is the consensus about how a *concept* should be *operationalized?*

Subjective Measures need Concepts

- Ambigous mental representations
  - e.g., „health“
- Are composed of different variables
  - e.g., „mental health“ and „physical health“
- Variable definitions
  - e.g., „mental health is the absence of mental illness“ or „physical health is the capacity to carry out daily activities“
- An operational definition [1] is used to determine the existance of a phenomenon
  - e.g., „for assessing mental illness: Mental Well-being Scale (WEMWBS) and the GHQ-9 tool “

[1] P. W. Bridgman: The Logic of Modern Physics. MacMillian, New York 1927.

Operationalization of Concepts

- The process of defining the measurement for a concept that is not directly measurable
- Making a fuzzy concept (e.g., emotions, likeability, memorability, usability, health, …) clearly
  - distinguishable
  - measurable
  - understandable
- Helps infer the existence of a concept
- Should be repeatable
- Depends on theoretical definitions
- Often defined by standardized tool and consensus

Examples

- iPhone users type very fast

Example

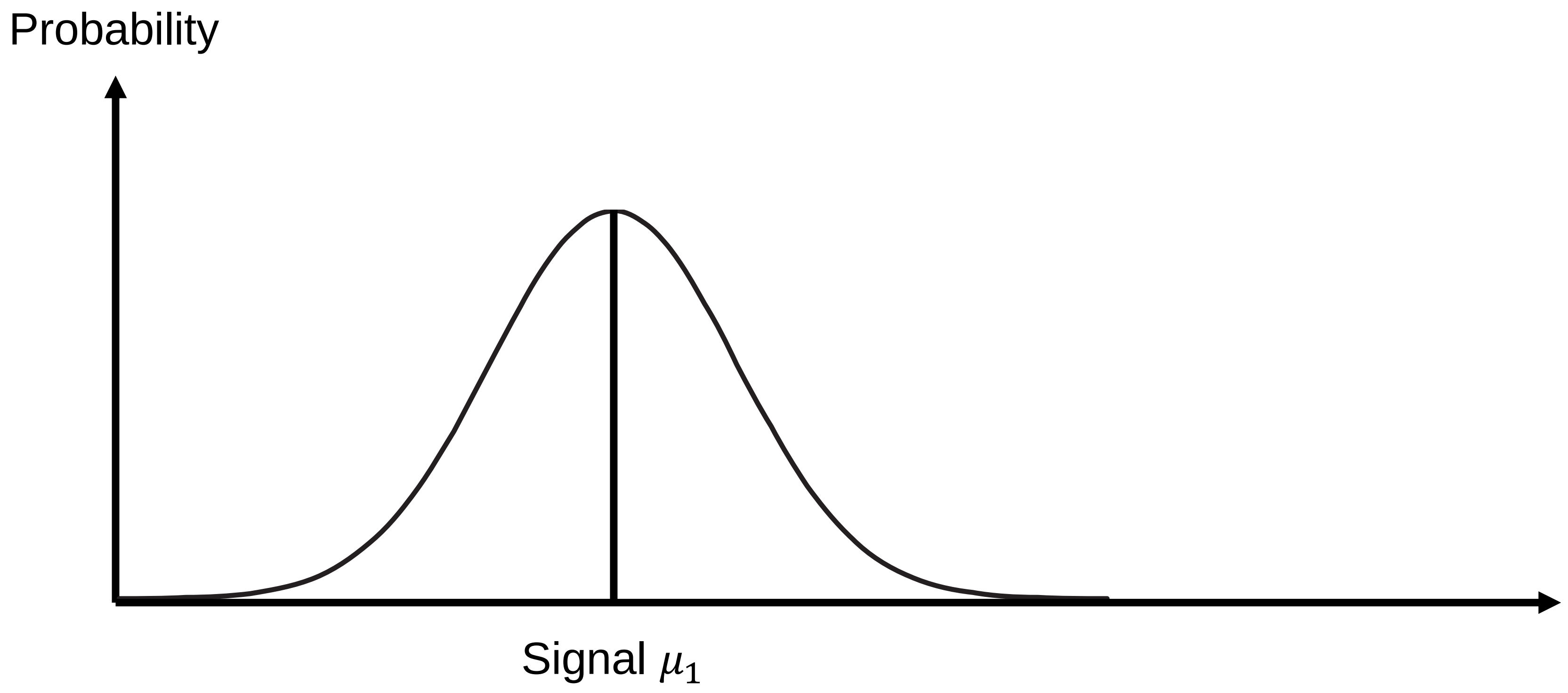
- ~~iPhone users type very fast~~
  - → Getting an iPhone increases the typing performance (H1) and decreases workload (H2) compared to getting an Android phone
  - Typing performance can operationalized by
    - words per minute (WPM)
    - characters per minute (CPM)
    - error rate
      - number of wrong / number of total words
      - number of backspace presses / number of characters
  - Workload can be operationalized by
    - NASA TLX score
  - Hypothesis tests on single or all measures?

Testing Using Multiple Measures

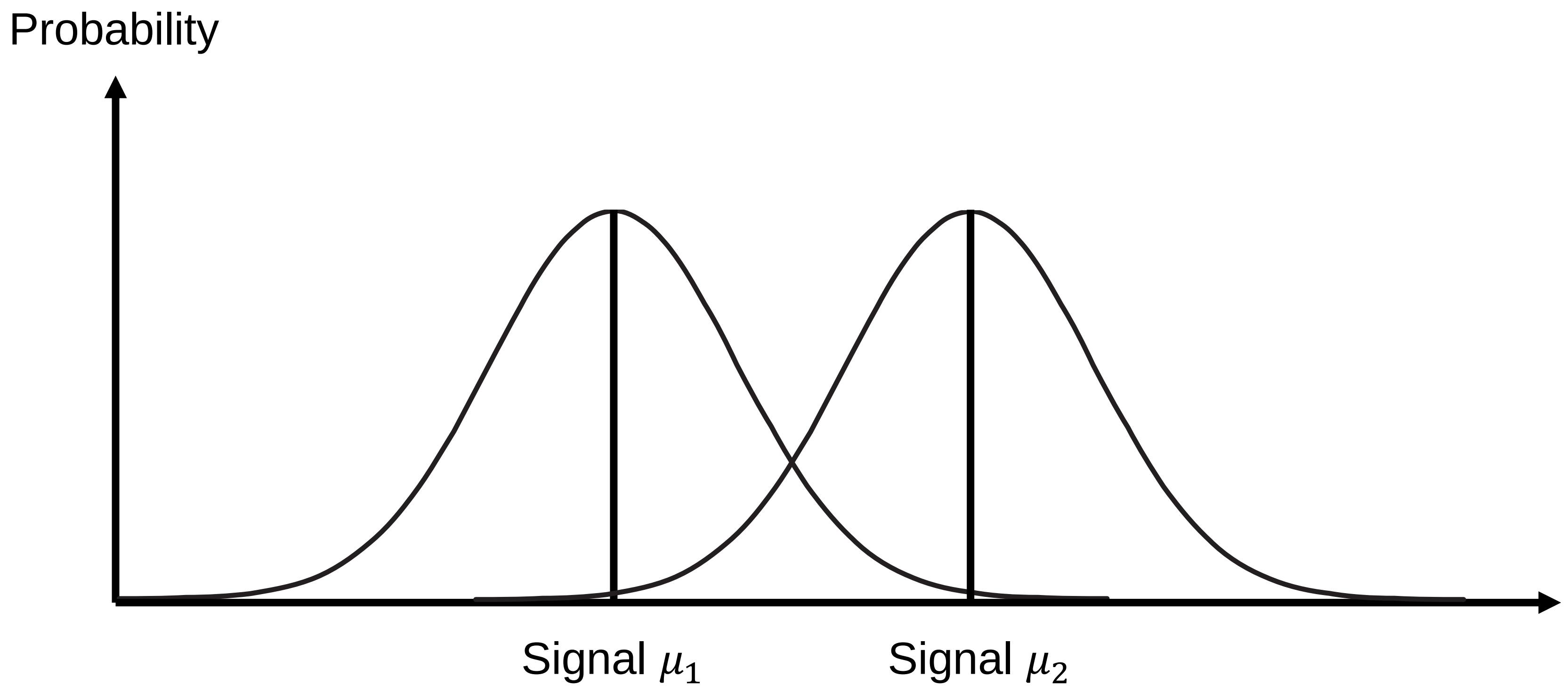
- Multiple measures
  - often reflect aspects of the same concept
  - we can expect correlations
    - e.g., more WPM = more CPM = less errors
  - increases the internal validity
- What if the hypothesis postulates that one measure has an impact on another?
  - e.g., more errors increase workload

  → Depends on the correct experimental design and the statistical test!

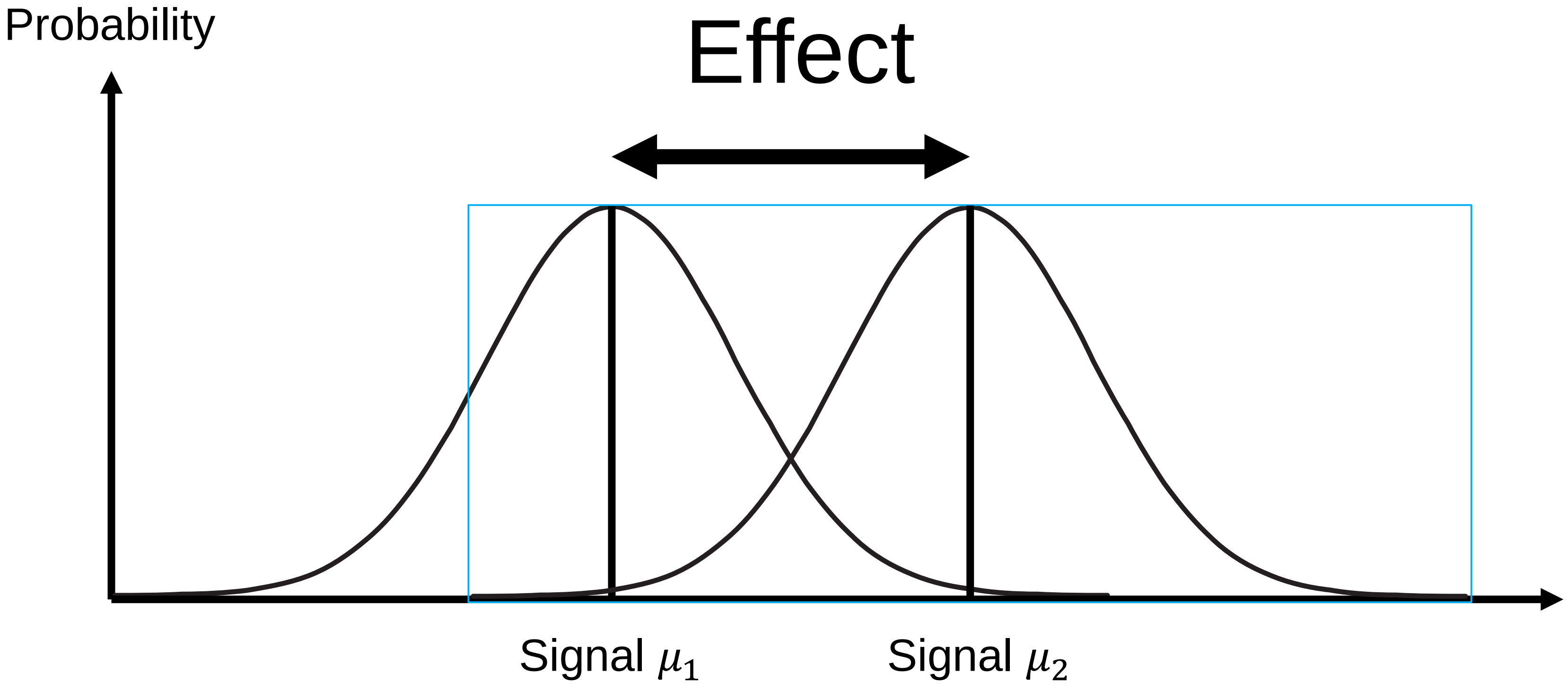
Hypothesis Testing



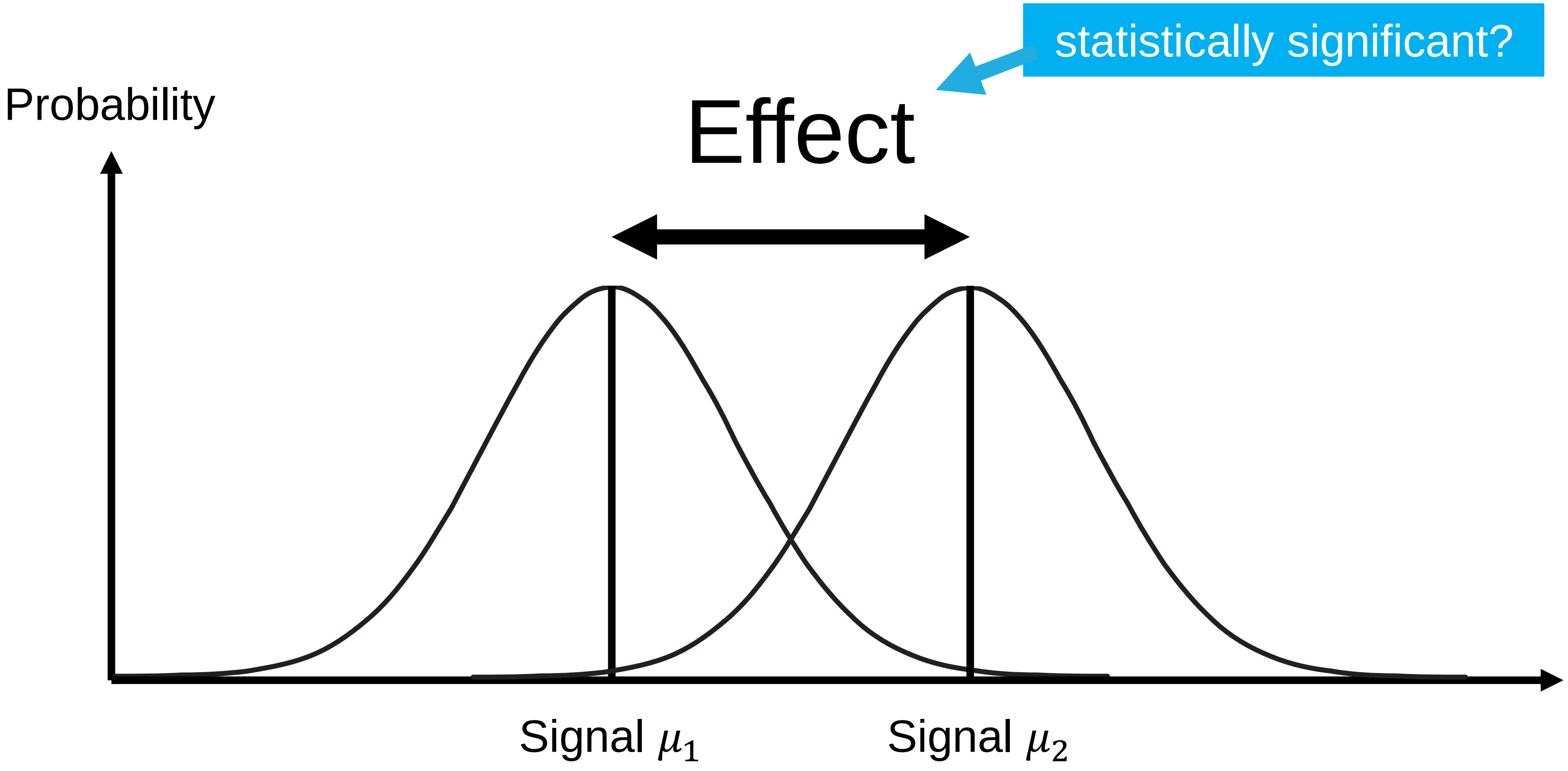
Hypothesis Testing



Hypothesis Testing



Hypothesis Testing



Statistical Significance

- A statistical significant effect exists if the probability that the difference occurred is below a certain significance level
- Significance level (α)
  - Lower significance level means higher evidence
  - Arbitrary, but typical significance level: $\alpha = 0.05$
- Significant results ($p < \alpha$)
  - Null hypothesis can be rejected
  - There is a statistical significant difference
- Non-Significant results ($p >= \alpha$)
  - Null hypothesis cannot be rejected
  - We cannot conclude anything!

Type I & Type II Errors

- $p = 0.028$

Type I error
(False Positive)

non-existing effect found
2.8%

true

Effect found

false

Effect exists

Type I & Type II Errors

- p = 0.028

Type I error (False Positive) non-existing effect found 2.8%	**Correct** (True Positive) existing effect was found 97.2%	**true**
false	**true**	

Effect found

Effect exists

Type I & Type II Errors

- p = 0.28 ←

Type I error (False Positive) non-existing effect found	**Correct** (True Positive) existing effect was found	**true**	**Effect found**
Correct (True Negative) no effect exists, no effect found 72%		**false**	
false	**true**		

Effect exists

Type I & Type II Errors

- p = 0.28 ←

Type I error (False Positive) non-existing effect found	**Correct** (True Positive) existing effect was found	**true**	**Effect found**
Correct (True Negative) no effect exists, no effect found 72%	**Type II error** (False Negative) effect exists, but is not found 28%	**false**	
false	**true**		
Effect exists			

Type III and Type IV Errors

- Type III error: “Wrong hypothesis, right answer”
  - Incorrect operationalization of variables
  - Poor theory (e.g., ad hoc explanations of findings)
  - Mis-identifying causal architecture
    - e.g., focusing on inter-individual factors (gender- or age-related differences) rather than structural factors
  - Researcher is either focusing on theory or on evaluation but not on the reasoning chain
- Type IV error: “Right hypothesis, wrong answer”
  - Collinearity among predictors
  - Wrong test
  - Aggregation bias

Example

- Let's assume we performed a paired t-test
- $p = 0.67 > \alpha = 0.05$
  - Reject H1. No significant difference between the conditions
  - We cannot conclude anything

	Nokia N95	**iPhone**
1	1.89	2.39
2	1.82	1.86
3	7.12	1.82
4	2.30	2.34
5	1.66	1.94
6	1.84	2.01
7	1.80	2.28
8	1.45	2.06
9	1.54	1.91
10	1.72	2.07

Example

- Let's assume we draw a better sample
- p = 0.028 < 0.05
  - Reject H0. Significant difference between the conditions
  - Typing on the iPhone results in a higher CPS than typing on the N95
- One outlier between rejecting and accepting H0 indicates a weak statistical power!

	Nokia N95	iPhone
1	1.89	2.39
2	1.82	1.86
3	**2.30**	**1.82**
4	2.30	2.34
5	1.66	1.94
6	1.84	2.01
7	1.80	2.28
8	1.45	2.06
9	1.54	1.91
10	1.72	2.07

Statistical Power

- Statistical power is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true
- Aspects that increase the statistical power
  - control all factors
  - increasing the sample size
  - increasing the effect size
  - increasing the number of conditions
  - increasing the number of measures
  - increasing the statistical significance criterion ($\alpha = 0.05$)

Statistical Power

- Statistical power is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true
- Aspects that increase the statistical power:
  - control all factors
  - increasing the sample size
  - increasing the effect size
  - increasing the number of conditions
  - increasing the number of measures
  - increasing the statistical significance criterion ($\alpha = 0.05$)

Do it

You can do that but be careful…

Don't do it.
In practice you will not find a justification

Increasing Statistical Power

- Increasing the sample size
  - More subjects
  - More trials
- Increasing the effect size
  - Reduce noise as much as possible
    - Task repetition (e.g., ask participants to enter 100 phrases instead of 1 and take the average)
    - Similar tasks (e.g., use phrases with the same difficulty instead of random phrases)
  - Remove outliers (e.g., remove samples that are 3x away from the standard deviation → only works under certain criteria)
  - Build something really good

Increasing Statistical Power

- Take multiple *similar* measures, e.g., for task performance:
  - Task completion time (TCT)
  - Error rate
  - Perceived task load (e.g., NASA TLX)
  - Subjective impression (e.g., SUS)
- Measurements or conditions that cannot be justified should not be taken
- Measure covariates (co-factors) you cannot or do not want to control (e.g., gender, hand size, height of the participants, glass wearer, etc.)

Familywise Error Rate (FWER)

- Too many conditions increase the probability that Type I errors occur. An estimation of FWER is:

$$F \leq 1 - (1 - \alpha)^c$$

  - α = alpha level for an individual test (e.g., 0.05)
  - c = number of tests
- For example, with an alpha level of 5% and a series of 10 tests, the FWER is:

$$F = 1 - (1 - 0.05)^{10} = .401 = 40\%$$

- This means that the probability of a Type I error is just over 40%, which is very high considering only ten tests were performed.

P-Value Adjustment

- Too many tests increase the probability of an inflation of Type I errors
- Solution: Bonferroni-correction: „Divide the alpha level by the number of tests you're running and apply that alpha level to each individual test."
  - e.g., if your overall alpha level is 0.05 and you are running 10 tests, then each test will have an alpha level of 0.05/10 = 0.005
  - Apply the new alpha level to each test for finding p-values. In this example, the p-value would have to be 0.005 or less for statistical significance

Multiple Measures

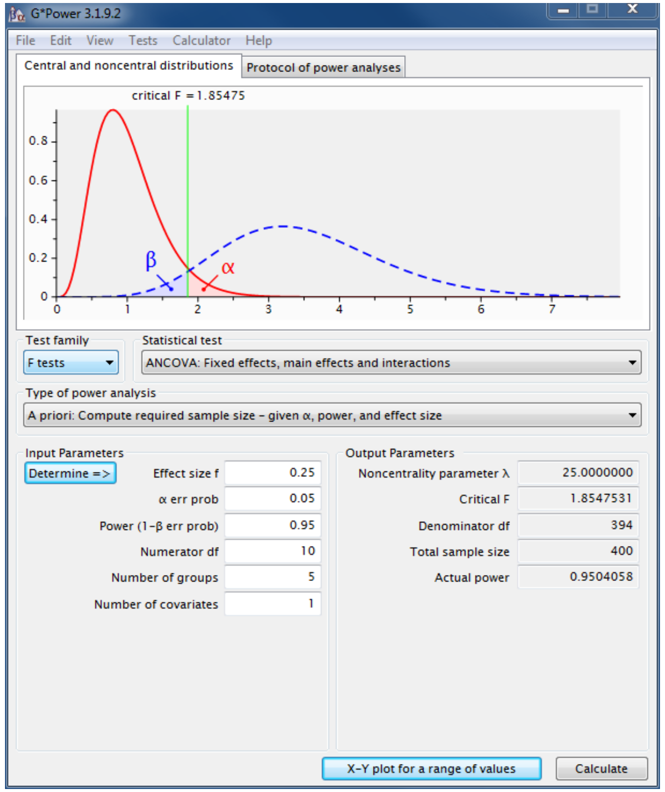
- Multiple measures allow to answer more research questions with minimal additional effort
- Multiple-item measure can be tested for internal consistency (there are consistency tests such as Cronbach’s alpha)
- When an independent variable is a construct that is manipulated indirectly, use a *manipulation check*

  → Usually a measure of the independent variable given at the end of the procedure

  → You can use statistical tests to check for manipulation

Determine the Statistical Power

- There are tools to perform a power analysis
  - can be used to determine the number of participants
  - require an effect size
  - see G*Power [1]



[1] G*Power: Statistical Power Analyses: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

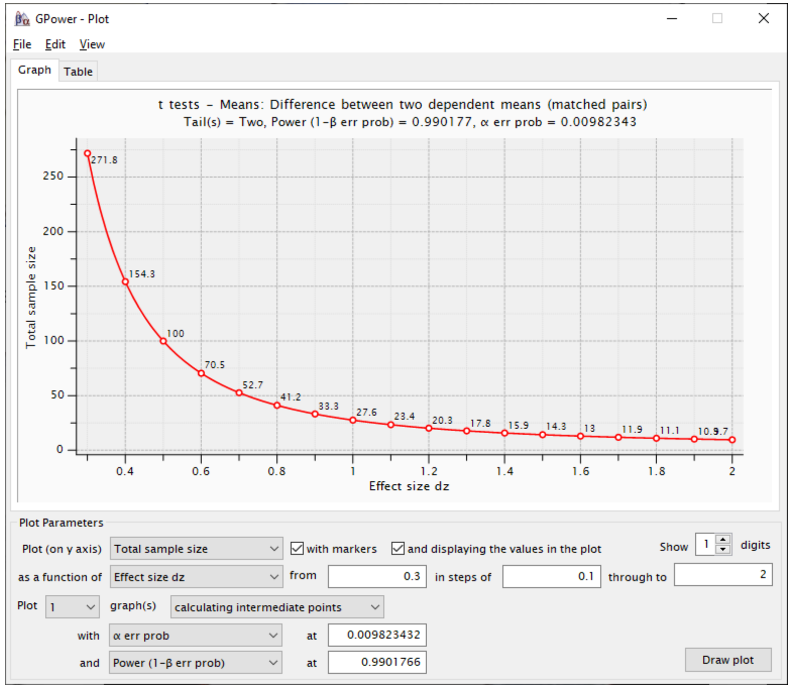
Determine the Statistical Power

- Example: The difference between two means in a paired t-Test
  - For an estimated effect size of 0.5 (medium)

  → you need 100 participants
  - For an estimated effect size of 1.0 (large)

  → you need 28 participants
  - For an estimated effect size of 2.0 (very large)

  → you need 10 participants



[1] G*Power: Statistical Power Analyses: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

Statistical Tests (Examples)

Analysis	Application	Examples
Factor/ Component	Searches for joint variations of observed variables in response to unobserved latent variables (factors)	• PCA • EFA

Statistical Tests (Examples)

Analysis	Application	Examples
Factor/ Component	Searches for joint variations of observed variables in response to unobserved latent variables (factors)	• PCA • EFA
Correlation	Determines the degree to which a pair of variables are linearly related	• Person • Spearman • Kendall

Statistical Tests (Examples)

Analysis	Application	Examples
Factor/ Component	Searches for joint variations of observed variables in response to unobserved latent variables (factors)	• PCA • EFA
Correlation	Determines the degree to which a pair of variables are linearly related	• Person • Spearman • Kendall
Regression	Models the functional relationship between a dependent variable and one or more independent variables	• Linear • Logistic • Nonlinear • Nonparametric

Statistical Tests (Examples)

Analysis	Application	Examples
Factor/ Component	Searches for joint variations of observed variables in response to unobserved latent variables (factors)	• PCA • EFA
Correlation	Determines the degree to which a pair of variables are linearly related	• Person • Spearman • Kendall
Regression	Models the functional relationship between a dependent variable and one or more independent variables	• Linear • Logistic • Nonlinear • Nonparametric
Cluster	Statistically grouping a set of objects in such a way that they are in the same group and more similar to each other than to those in other groups	• K-Means • MDS • HC

Statistical Tests (Examples)

Analysis	Application	Examples
Factor/ Component	Searches for joint variations of observed variables in response to unobserved latent variables (factors)	• PCA • EFA
Correlation	Determines the degree to which a pair of variables are linearly related	• Person • Spearman • Kendall
Regression	Models the functional relationship between a dependent variable and one or more independent variables	• Linear • Logistic • Nonlinear • Nonparametric
Cluster	Statistically grouping a set of objects in such a way that they are in the same group and more similar to each other than to those in other groups	• K-Means • MDS • HC
Variance	Analyzes the differences among group means in a sample.	• t-Test • ANOVA • ART-ANOVA

Statistical Tests (Examples)

Analysis	Application	Examples
Factor/ Component	Searches for joint variations of observed variables in response to unobserved latent variables (factors)	• PCA • EFA
Correlation	Determines the degree to which a pair of variables are linearly related	• Person • Spearman • Kendall
Regression	Models the functional relationship between a dependent variable and one or more independent variables	• Linear • Logistic • Nonlinear • Nonparametric
Cluster	Statistically grouping a set of objects in such a way that they are in the same group and more similar to each other than to those in other groups	• K-Means • MDS • HC
Variance	Analyzes the differences among group means in a sample.	• t-Test • ANOVA • ART-ANOVA
Equivalence	The null hypothesis is defined as an effect large enough to be deemed interesting, specified by an equivalence bound.	• TOST • Bayes

Summary

- Experiments and statistical analysis can isolate cause and effect and are used for testing hypotheses
- Make an hypothesis testable and falsifiable (null-hypothesis)
- Calculate an appropriate sample to increase the statistical power and to avoid Type I and Type II errors
- Decrease the variance by multiple and repeated measures
- Increase the effect size
- Below a level of significance level of 0.05, the p-value indicates if the the null hypothesis can be rejected in favor of the alternative hypothesis
- Results are never true in a sense of being 100% correct!

Literature

- Field, Andy & Hole, Graham. (2003). *How to Design and Report Experiments.*
- Field, Andy (2013). *Discovering Statistics Using IBM SPSS Statistics.*
- Lehmann, Erich Leo (1959). *Testing Statistical Hypotheses.*