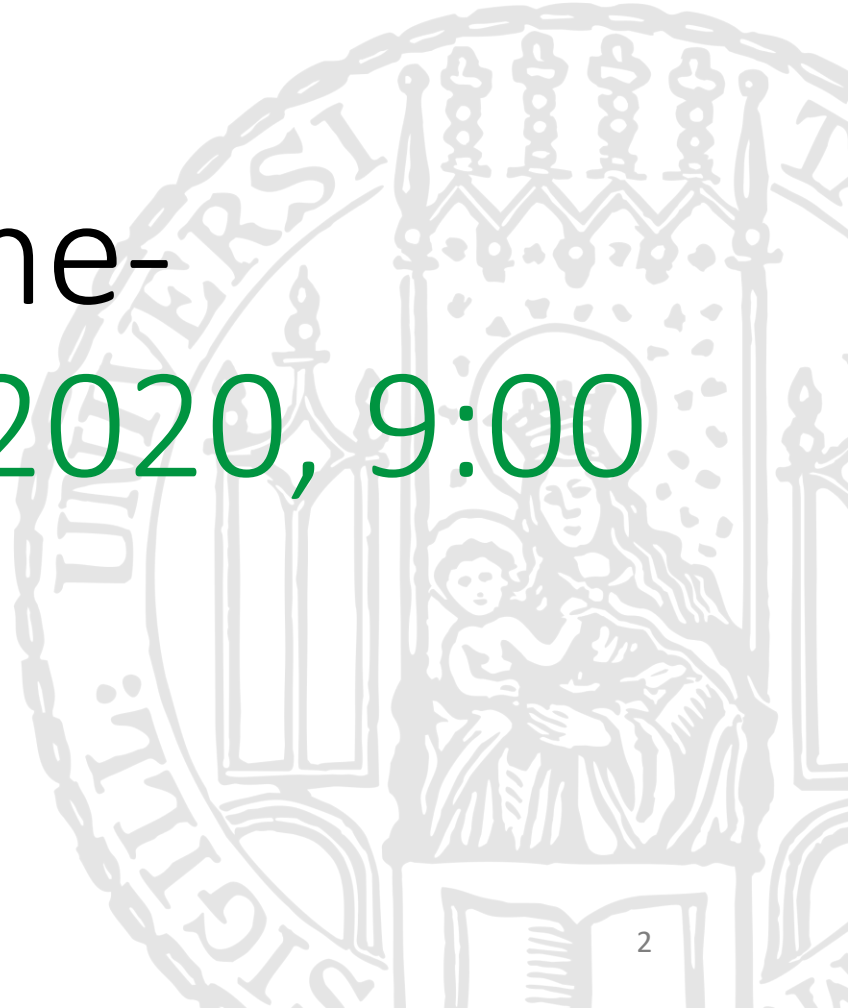
Mensch Maschine Interaktion

Übung 09
Evaluation

Nicht vergessen:

Abgabe der ersten Online-Hausarbeit ist am 26.6.2020, 9:00



Evaluation

Klassifizierung von Evaluation

- Formativ vs. Summativ
- Qualitativ vs. Quantitativ
- Analytisch vs. Empirisch

Evaluationsmethoden

	Formativ vs. summativ	Qualitativ vs. quantitativ	Analytisch vs. empirisch
Heuristische Evaluation			
GOMS KLM			
Beobachtungsstudie			
Kontrollierte Experimente			
Feldstudien			
Tagebuch-Studien			

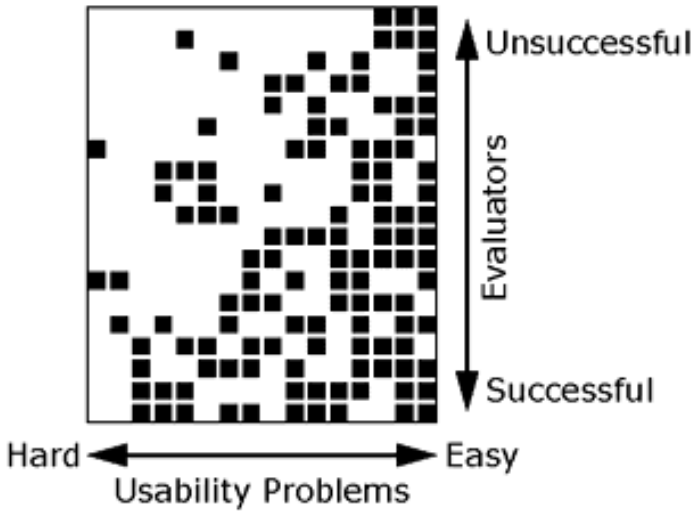
Heuristische Evaluation

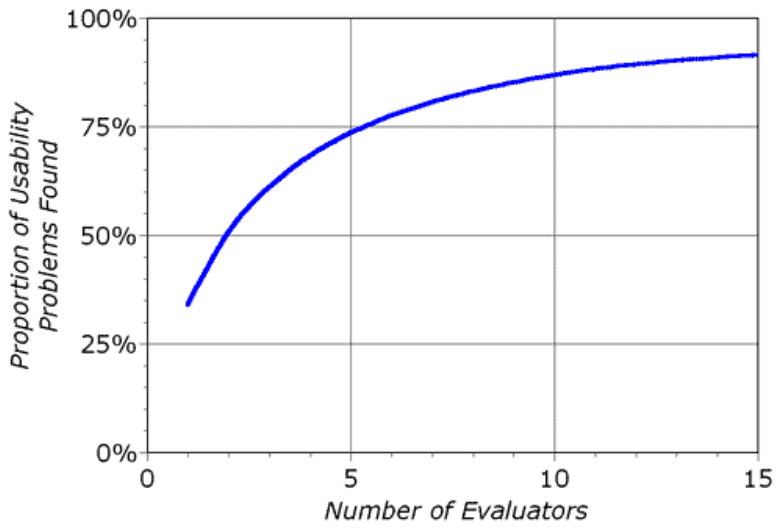
- Experten bewerten ein Interface mit einem Fragebogen entlang einer Aufgabe
- Fragebogen orientiert sich oft and den 10 Heuristiken von Nielsen

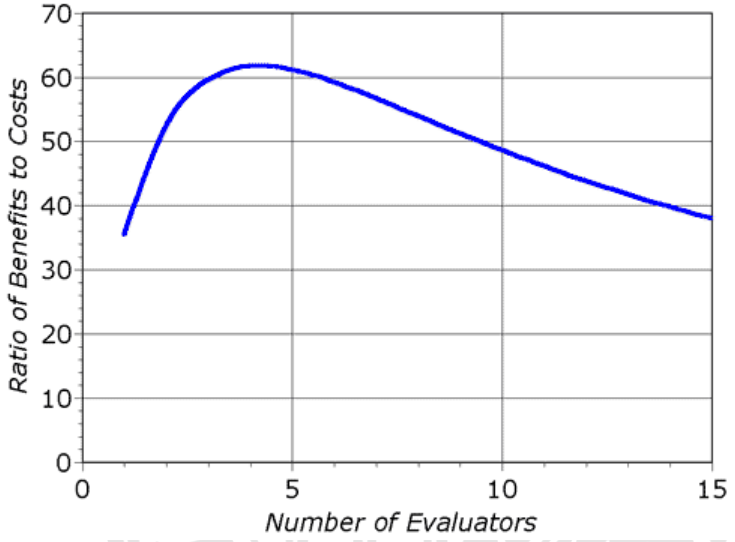
	Erfüllt	Nicht erfüllt	Kommentare
Das System hat eine Undo Funktion	○	○	

Heuristische Evaluation

- Wie viele Bewertungen/Bewerter? Idealerweise mehrere!







Heuristiken nach Jacob Nielsen

1. Visibility of System Status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

https://www.nngroup.com/articles/ten-usability-heuristics/

Visibility of system status

Match between system and the real world

User control and freedom

Consistency and standards

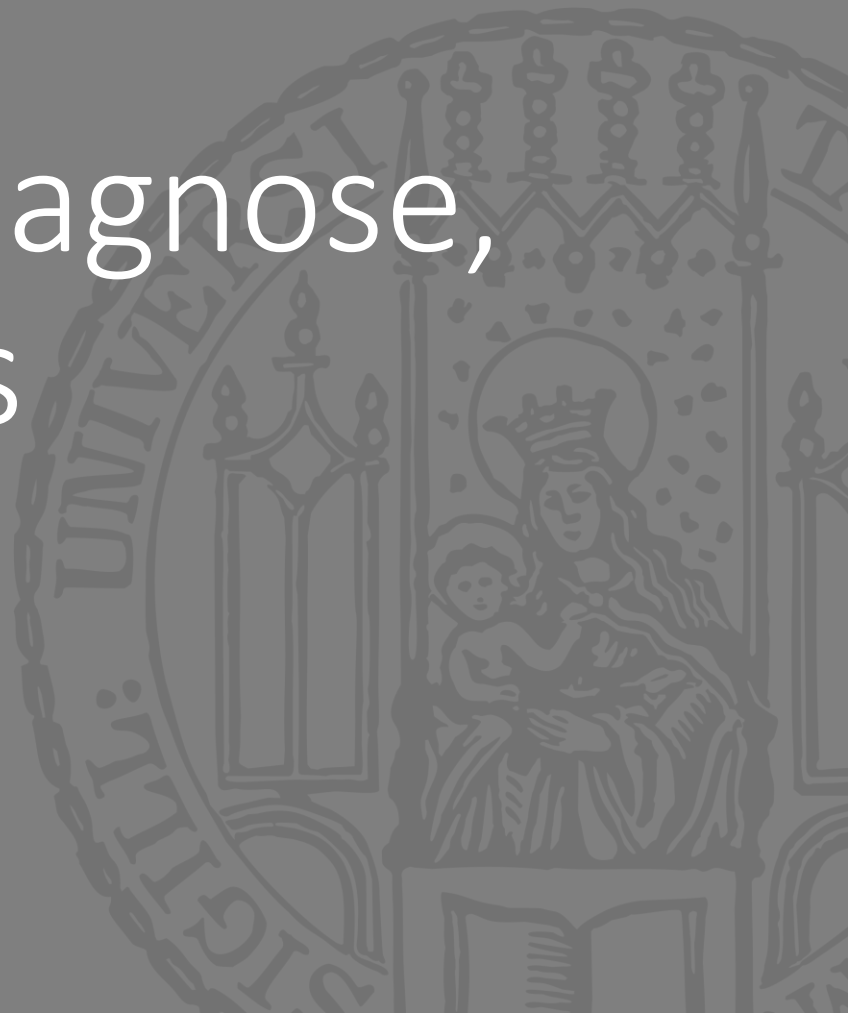
Error prevention

Recognition rather than recall

Flexibility and efficiency of use

Aesthetic and minimalist design

Help users recognize, diagnose, and recover from errors



Help and documentation

GOMS

Nach S. Card et al.

- Goals: Was will ich am Ende erreichen?
- Operators: Welche Handlungen kann ich tätigen?
- Methods: Welche Prozesse / Folgen von Operatoren führen zum Ziel?
- Selection Rules: Welche Methode/n wähle ich zum Ziel?

- "Top-down"-Methode

GOMS-KLM

Nach S. Card et al.

- "Bottom-up"-Methode

- K – Keystroke ≈ 0.28s
- P – Pointing ≈ 1.1s
- H – Homing ≈ 0.4s
- M – Mental preparation ≈ 1.35s
- R – Response of the system

GOMS KLM

- Berechnen Sie eine Abschätzung für folgendes Beispiel:

  Sie haben aktuell Word geöffnet und ihre Übungsblatt-Abgabe fertig getippt. Nun wollen Sie diese als PDF Exportieren und auf Uni2Work hochladen. Sie können davon ausgehen, dass der Browser im Hintergrund geöffnet ist und die Uni2Work upload Seite bereits geöffnet ist.

Beobachtungsstudie

- Quasi-Experiment
- Kein Eingriff in das Geschehen
- Keine Veränderung von Variablen
- Oft nachträgliches Einteilen der Teilnehmer in Gruppen
- Typische Auswertung: Korrelationen

Kontrollierte Experimente

- Unter experimentellen, kontrollierten Bedingungen
- Variieren der unabhängigen Variable
- Messung der abhängigen Variable

- Jeder Teilnehmer durchläuft jede Experimentalbedingung:
  within-subject
- Jede Teilnehmer-Gruppe durchläuft genau eine Bedingung:
  between-groups

Kontrollierte Experimente

- Experimente dienen dazu Hypothesen zu verifizieren
- Hypothese:
  "Übungsteilnehmer erzielen bessere Noten in der Klausur"
- Echt verifizieren ist schwer, falsifizieren ist oft leichter
- Ab wann sind die Noten denn "besser"?
- Deswegen Null-Hypothese:
  "Übungsteilnehmer und Nichtteilnehmer erzielen im Mittel die gleichen Noten in der Klausur"

Studien-Design

- Teilnehmer in Studien sollten aus einer möglichst diversen Gruppe kommen
- Die Studie sollte nicht zu lang sein → Ermüdungseffekte
- Die Reihenfolge des Studienablaufs sollte die Ergebnisse nicht beeinflussen → Reihenfolgeneffekte, Lerneffekte
  - Randomisierung ist gut, aber keine Sicherheit
  - Deswegen am besten alle Permutationen testen
  - Oft nicht machbar da zu viele Möglichkeiten (n!)

Latin-Square Permutation

Bed. 1	Bed. 2	Bed. 3	Bed. 4
Bed. 4	Bed. 1	Bed. 2	Bed. 3
Bed. 3	Bed. 4	Bed. 1	Bed. 2
Bed. 2	Bed. 3	Bed. 4	Bed. 1

Latin-Square Permutation

Bed. 1	Bed. 3	Bed. 4	Bed. 2
Bed. 3	Bed. 1	Bed. 2	Bed. 4
Bed. 4	Bed. 2	Bed. 1	Bed. 3
Bed. 2	Bed. 4	Bed. 3	Bed. 1

Feldstudien

- Laborstudien spiegeln oft nicht die wirklichen Situationen wieder
- Deswegen Feldstudien, beobachtung in echter Umgebung
- Geringere interne Validität
- Dafür größere externe Validität

Tagebuch-Studien

- Langzeit Studien: Studien die Effekte über lange Zeiträumen untersuchen
- Dauerhafter Aufenthalt im Labor ist nicht machbar
- Die "Messung" kann dabei eigenverantwortlich als Tagebuch durch den Probanten erfasst werden
- Auch möglich über bspw. das Smartphone via Experience Sampling (s. auch das PhoneStudy Projekt der LMU)

Evaluationsmethoden

	Formativ vs. summativ	Qualitativ vs. quantitativ	Analytisch vs. empirisch
Heuristische Evaluation			
GOMS KLM			
Beobachtungsstudie			
Kontrollierte Experimente			
Feldstudien			
Tagebuch-Studien			

Auswertung von Studienergebnissen

Daten

- Studien-Daten können in verschiedenen Formen kommen

- Nominal: bspw. Geschlecht, Herkunft, welches System genutzt wurde
- Ordinal: Reihenfolgen, ohne Aussage über Abstände, bspw. Likert-Skalen
- Interval: Reihenfolgen mit definierten Abständen, bspw. Temperatur in °C
- Ratio: Wie Interval aber mit absolutem Nullpunkt, bspw. Größe

Auswertung

- Um belastbare Aussagen über Ihre Ergebnisse zu treffen, sollten Sie diese statistisch überprüfen
- Hierfür gibt es statistische Hypothesentests
- Diese geben u.a. eine Abschätzung, wie wahrscheinlich es ist, dass Ihre Daten, unter Annahme der Null-Hypothese, so auftreten würden, wie Sie sie beobachtet haben

  $p = 0.05$ bedeutet in etwa *"Mit 5% Wahrscheinlichkeit würden wir die Daten so bekommen, wenn die Null-Hypothese gelten würde."*

Auswertung

- Bei ausreichend kleinem p-Wert können Sie die Null-Hypothese ablehnen. D.h. bei kleinen p ist ihre Hypothese wahrscheinlich richtig.

- Typische “klein-genug”-Werte:

  $p < 0.05$ → signifikant *

  $p < 0.01$ → hoch signifikant **

  $p < 0.001$ → höchst signifikant ***

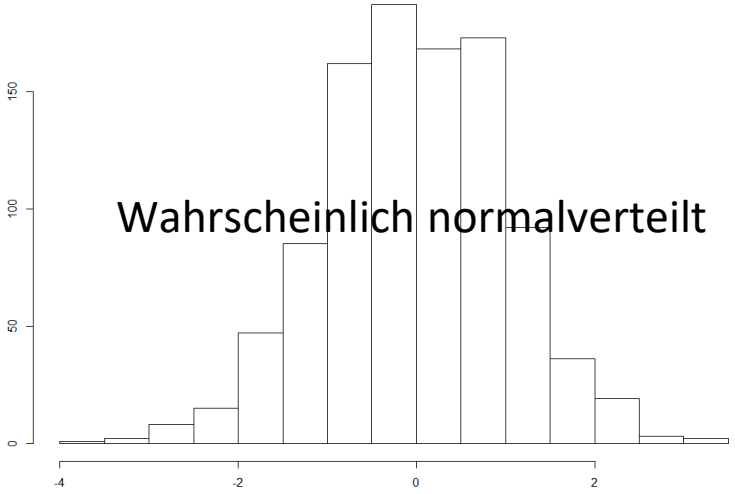
Hypothesen-Tests

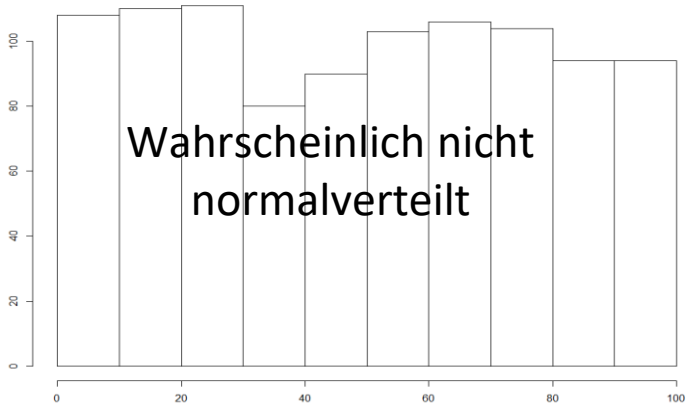
- Es gibt für verschiedene Situationen verschiedene Tests
- Nicht jeder Test ist für jede Situation geeignet oder hat Aussagekraft
- Wichtige Kriterien für die Test-Auswahl:
  - Form der Daten
  - Anzahl der Samples
  - Ziel der Analyse

https://www.youtube.com/watch?v=rulIUAN0U3w

Normalverteilung

- Ein wichtiges Kriterium für die Test-Wahl: Sind die Daten normalverteilt?
- Auch hierfür gibt es statistische Tests, bspw. den Shapiro-Wilk-Test
- Hier ist die *Null-Hypothese*, dass die Daten normalverteilt sind





t-Test als Beispiel für Hypothesentests

- Für normalverteilte Daten
- Mehrere Varianten:
  - Für 1 sample: Weicht der Mittelwert von einer Annahme ab
  - Für 2 samples unpaired/independent (bspw. zwei Gruppen, between groups): Sind die Mittelwerte signifikant unterschiedlich
  - Für 2 samples paired (bspw. Wenn Probanten zwei Systeme testen, within subject) :
    Sind die Mittelwerte signifikant unterschiedlich

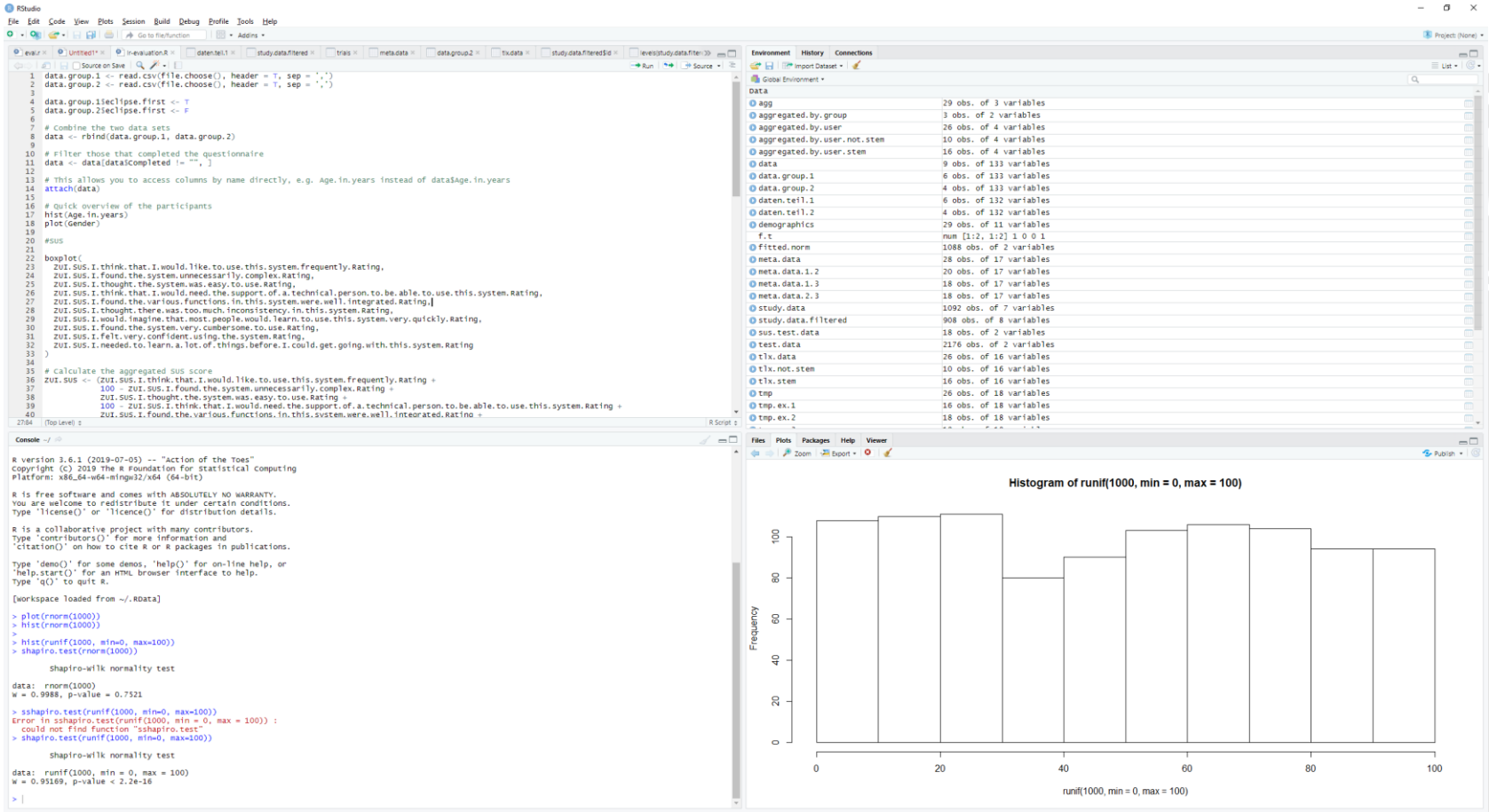
    Bei 2 samples sollten die Varianzen idealerweise ähnlich sein

https://www.youtube.com/watch?v=rullUAN0U3w

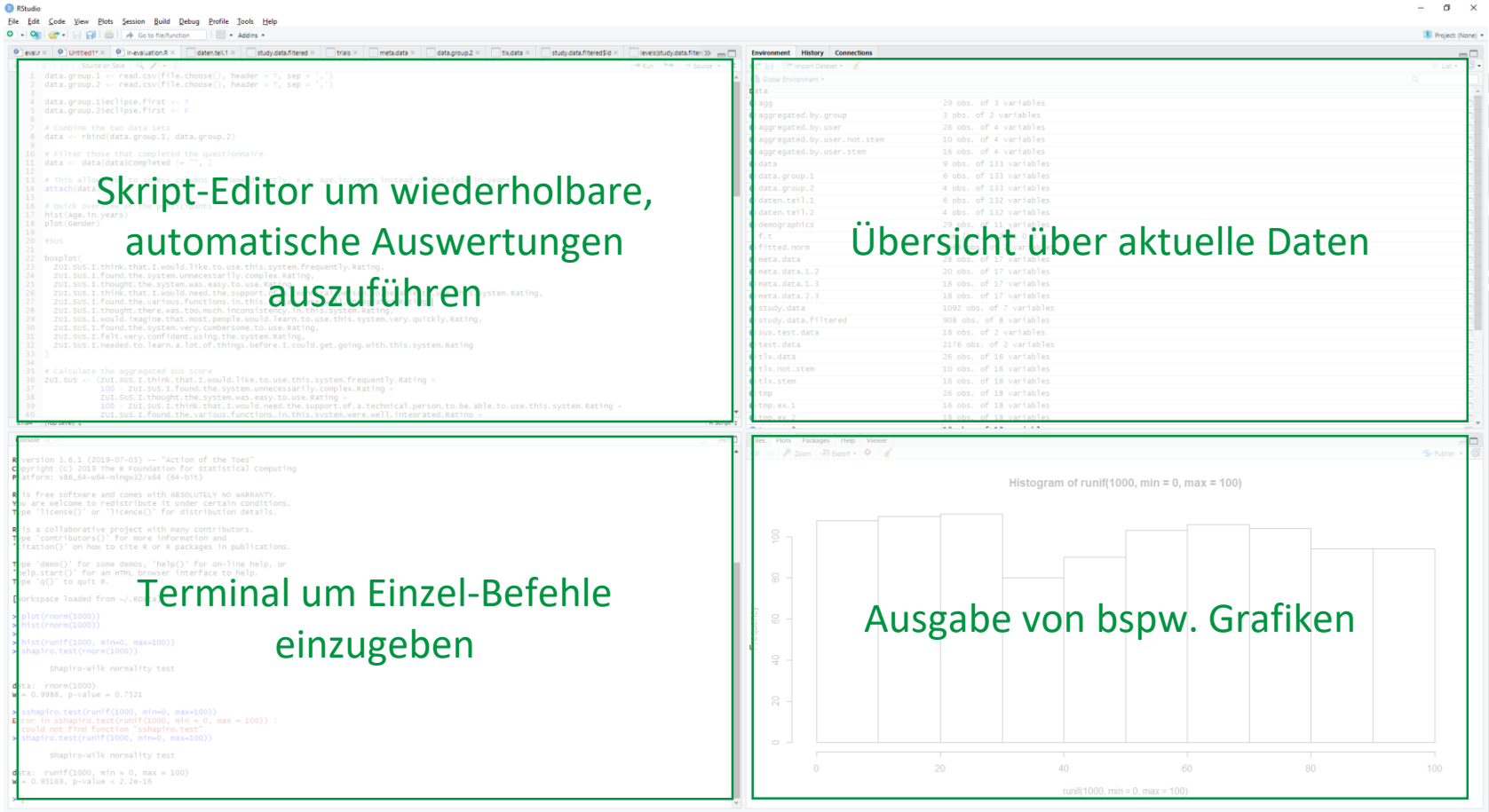
Und wie macht man den Test jetzt?

- Hypothesentests sind in div. Software vor-implementiert, bspw.:
  - Excel (eingeschränkt, Microsoft, kostenpflichtig)
  - SPSS (IBM, kostenpflichtig)
  - R (open source, kostenlos)

R-Studio


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
data.group.1 <- read.csv(file.choose(), header = T, sep = ',')
data.group.2 <- read.csv(file.choose(), header = T, sep = ',')
data.group.1$eclipse.first <- T
data.group.2$eclipse.first <- F
# Combine the two data sets
data <- rbind(data.group.1, data.group.2)
# Filter those that completed the questionnaire
data <- data[data$Completed != "",]
# This allows you to access columns by name directly, e.g. Age.in.years instead of data$Age.in.years
attach(data)
# Quick overview of the participants
hist(Age.in.years)
plot(Gender)
#SUS
boxplot(
ZUI.SUS.I.think.that.I.would.like.to.use.this.system.frequently.Rating,
ZUI.SUS.I.found.the.system.unnecessarily.complex.Rating,
ZUI.SUS.I.thought.the.system.was.easy.to.use.Rating,
ZUI.SUS.I.think.that.I.would.need.the.support.of.a.technical.person.to.be.able.to.use.this.system.Rating,
ZUI.SUS.I.found.the.various.functions.in.this.system.were.well.integrated.Rating,
ZUI.SUS.I.thought.there.was.too.much.inconsistency.in.this.system.Rating,
ZUI.SUS.I.would.imagine.that.most.people.would.learn.to.use.this.system.very.quickly.Rating,
ZUI.SUS.I.found.the.system.very.cumbersome.to.use.Rating,
ZUI.SUS.I.felt.very.confident.using.the.system.Rating,
ZUI.SUS.I.needed.to.learn.a.lot.of.things.before.I.could.get.going.with.this.system.Rating
)
# Calculate the aggregated SUS score
ZUI.SUS <- (ZUI.SUS.I.think.that.I.would.like.to.use.this.system.frequently.Rating +
100 - ZUI.SUS.I.found.the.system.unnecessarily.complex.Rating +
ZUI.SUS.I.thought.the.system.was.easy.to.use.Rating +
100 - ZUI.SUS.I.think.that.I.would.need.the.support.of.a.technical.person.to.be.able.to.use.this.system.Rating +
ZUI.SUS.I.found.the.various.functions.in.this.system.were.well.integrated.Rating +
Console
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
> plot(rnorm(1000))
> hist(rnorm(1000))
> hist(runif(1000, min=0, max=100))
> shapiro.test(rnorm(1000))
Shapiro-Wilk normality test
data: rnorm(1000)
W = 0.9988, p-value = 0.7521
> sshapiro.test(runif(1000, min=0, max=100))
Error in sshapiro.test(runif(1000, min = 0, max = 100)) :
could not find function "sshapiro.test"
> shapiro.test(runif(1000, min=0, max=100))
Shapiro-Wilk normality test
data: runif(1000, min = 0, max = 100)
W = 0.95169, p-value < 2.2e-16
Environment History Connections
Data
agg 29 obs. of 3 variables
aggregated.by.group 3 obs. of 2 variables
aggregated.by.user 26 obs. of 4 variables
aggregated.by.user.not.stem 10 obs. of 4 variables
aggregated.by.user.stem 16 obs. of 4 variables
data 9 obs. of 133 variables
data.group.1 6 obs. of 133 variables
data.group.2 4 obs. of 133 variables
daten.teil.1 6 obs. of 132 variables
daten.teil.2 4 obs. of 132 variables
demographics 29 obs. of 11 variables
f.t num [1:2, 1:2] 1 0 0 1
fitted.norm 1088 obs. of 2 variables
meta.data 28 obs. of 17 variables
meta.data.1.2 20 obs. of 17 variables
meta.data.1.3 18 obs. of 17 variables
meta.data.2.3 18 obs. of 17 variables
study.data 1092 obs. of 7 variables
study.data.filtered 908 obs. of 8 variables
sus.test.data 18 obs. of 2 variables
test.data 2176 obs. of 2 variables
tlx.data 26 obs. of 16 variables
tlx.not.stem 10 obs. of 16 variables
tlx.stem 16 obs. of 16 variables
tmp 26 obs. of 18 variables
tmp.ex.1 16 obs. of 18 variables
tmp.ex.2 18 obs. of 18 variables
Files Plots Packages Help Viewer
Histogram of runif(1000, min = 0, max = 100)
Frequency
0 20 40 60 80 100
runif(1000, min = 0, max = 100)

R-Studio


Skript-Editor um wiederholbare, automatische Auswertungen auszuführen
Übersicht über aktuelle Daten
Terminal um Einzel-Befehle einzugeben
Ausgabe von bspw. Grafiken
Histogram of runif(1000, min = 0, max = 100)

R-Studio

- Kommende Woche wollen wir mit R-Studio arbeiten
- Deswegen bitte bis dahin R-Studio installieren

Nächste Übung: Evaluation II & IxD

22.06.20 - 26.06.20