Statistik und Datenanalyse: Aufbau

Mittelwert, Standardabweichung, Korrelation

Benjamin Fretwurst

▶ PDF-Version der Folien

Inhalt

1 Modulintro

Lernziele heute

Modulorga

- Lernziele und Inhalte von «Statistik: Aufbau»
- Orga
- Lernmaterialien

Anschluss an Statistik: Einführung

- Mittelwert
- Streuung
- Inferenz
  - Punktschätzung (Standardfehler)
  - Intervallschätzung
  - Tests
- Kovarianz und Korrelation

1.1 Lernziele des Moduls

Grundlagenkenntnisse

- Sie erlangen Kenntnisse multivariater Statistik.
- Verständnis empirischer Forschungsbeiträge
- Umgang mit Vorlagen für die Anwendung der Verfahren in R.

Praktische Statistik für «Methoden Aufbau»

Sie können Analysemethoden anwenden, die Sie für Ihr Projekt in Methoden Aufbau brauchen.

Verständnis fortgeschrittener Statistik für die Forschungsseminare

Vorbereitung auf die Forschungsseminare, in denen Sie weiterführende Analyseverfahren verstehen müssen (Texte und Analysen) und teilweise auch anwenden.

1.1.1 Voraussetzungen

Inhaltliche Voraussetzungen

- Deskriptive und induktive Statistik aus «Statistik und Datenanalyse: Einführung»
- heute Rückblick
- Folien der Statistik-Einführungs-Vorlesung (+ Buch Ihrer Wahl)

Technische Voraussetzung

- R
- R-Studio (oder VS-Code)
- Weiteres siehe Anleitung R-Installationsanleitung

1.2 Leistungsnachweis

Was ist Prüfungsstoff?

- Folien
- Vorlesung
- Begleittext

Prüfungstermin (Major + Minor)

- Hauptklausur
  - 03.01.2024
  - 14:30–15:30 (60 Minuten)
  - BYOD ➪ [Y-15-G-60 + Y24-G-45]
- Wiederholungsklausur Major und Minor
  - 07.02.2024
  - 14:30–15:30 (60 Minuten)
  - BYOD ➪ Y-15-G-60

1.3 Inhalt und Aufwandt

Inhalte

1. Uni- und Bivariate Statistik
2. GLM – Regression
3. GLM – BLUE
4. Übung: GLM I
5. GLM – Kategoriale UV
6. GLM – Interaktionen
7. GLM – Übung II
8. Dimensionsreduktion
9. Übung: Dimensionsreduktion
10. LogReg und ML
11. Übung: Machine Learning
12. Clusteranalyse
13. Übung: R
14. Zusammenfassung

Aufwandt für 6 ECTS

Aufwand	in h	h/Woche	Punkte
Besuch der Vorlesung	21	1.5	0.7
Vor und Nachbereitung	21	1.5	0.7
Lesen der Texte	42	3.0	1.4
Übungsaufgaben in R	50	3.6	1.7
Prüfungsvorbereitung	42	3.0	1.4
Studienteilnahmepunkte	4	0.3	0.1
Summe für 6 ECTS * 30h	180	12.9	6.0

1.4 Material

E-Learning – OLAT

OLAT

- Mitteilungen
- Forum
  - Diskussionen und Fragen
  - Klausurvorbereitung
- Studienteilnahmepunkte 🔗

R-Seite und Begleittext (BETA!)

R-Seite

Auf r.ikmz.uzh.ch/Wissen_macht_R! finden Sie Anleitungen zur Installation von R und R-Studio sowie Beispielskripte und Vorlagen, die wir vor allem auch für brauchen, bzw. Ihnen helfen sollen.

Begleittext ist noch BETA!

Der Begleittext auf stat.ikmz.uzh.ch/Aufbau wird während des Semesters deutlich überarbeitet und ist daher BETA! Es gibt keine vollständige Gewähr für Fehlerfreiheit!

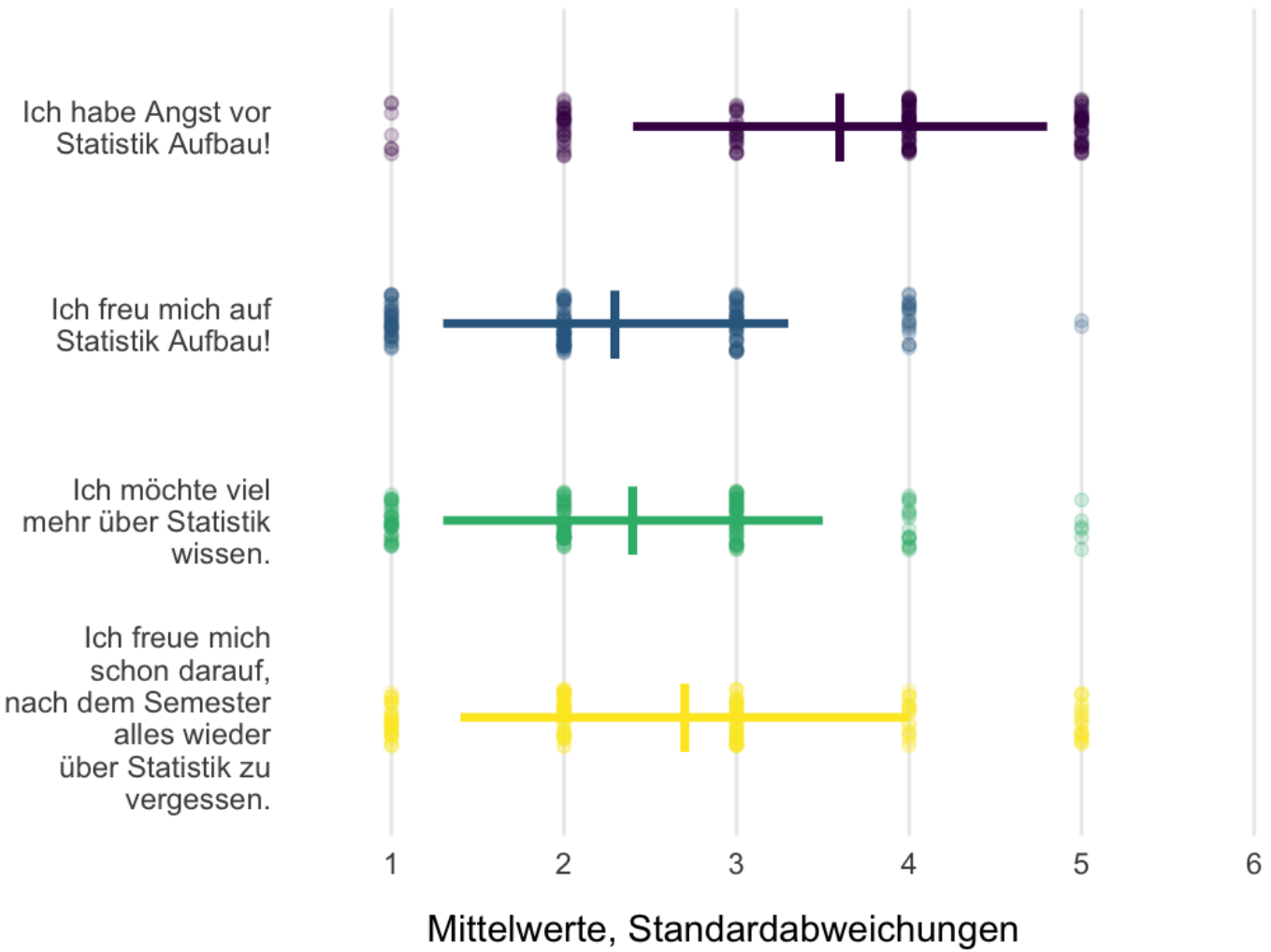
Zusatzliteratur (Wiederholung Statistik Einführung)

Für die Wiederholung von «Statistik: Einführung».

2 Befragung

Erwartungen

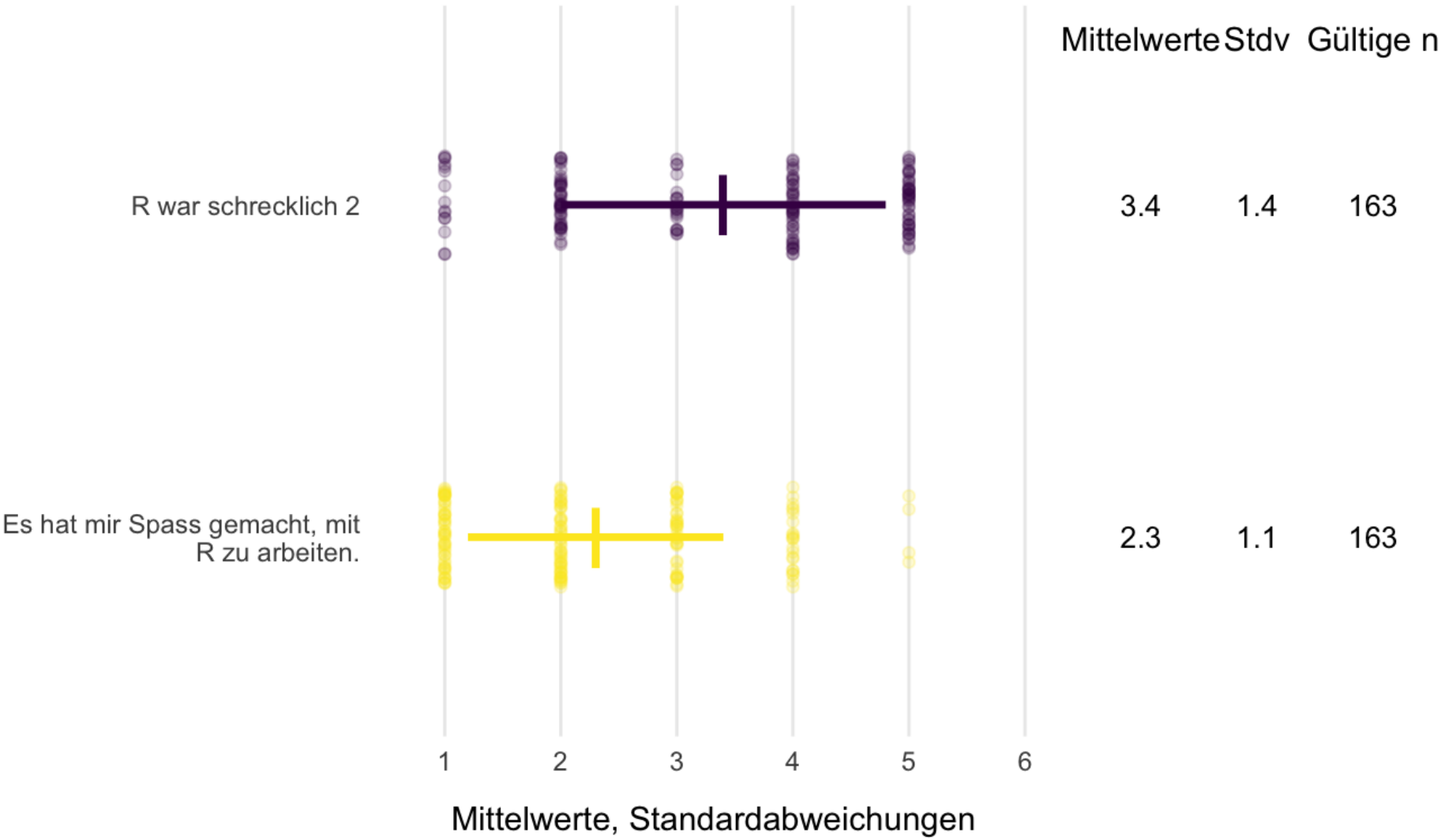
- R-Code anzeigen



	Mittelwerte	Stdv	Gültige n
Ich habe Angst vor Statistik Aufbau!	3.6	1.2	163
Ich freu mich auf Statistik Aufbau!	2.3	1.0	164
Ich möchte viel mehr über Statistik wissen.	2.4	1.1	164
Ich freue mich schon darauf, nach dem Semester alles wieder über Statistik zu vergessen.	2.7	1.3	161

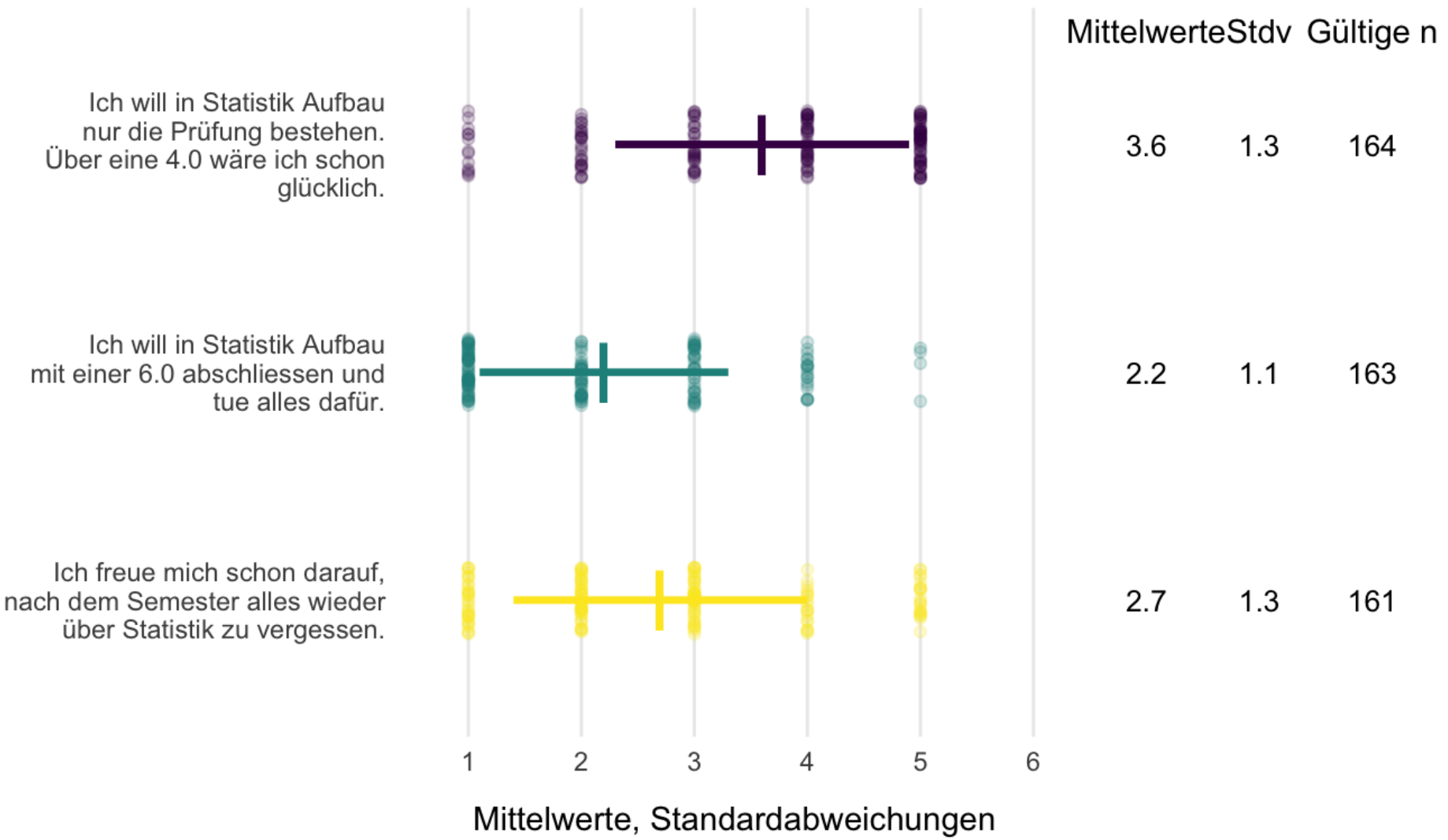
Umgang mit R

- R-Code anzeigen



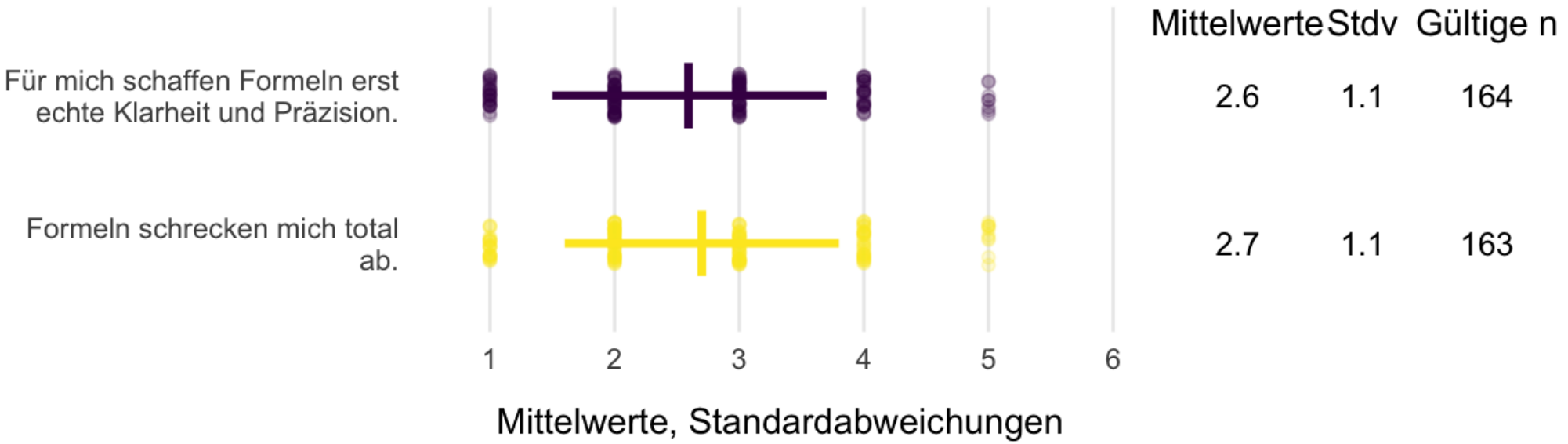
Ziele

- R-Code anzeigen



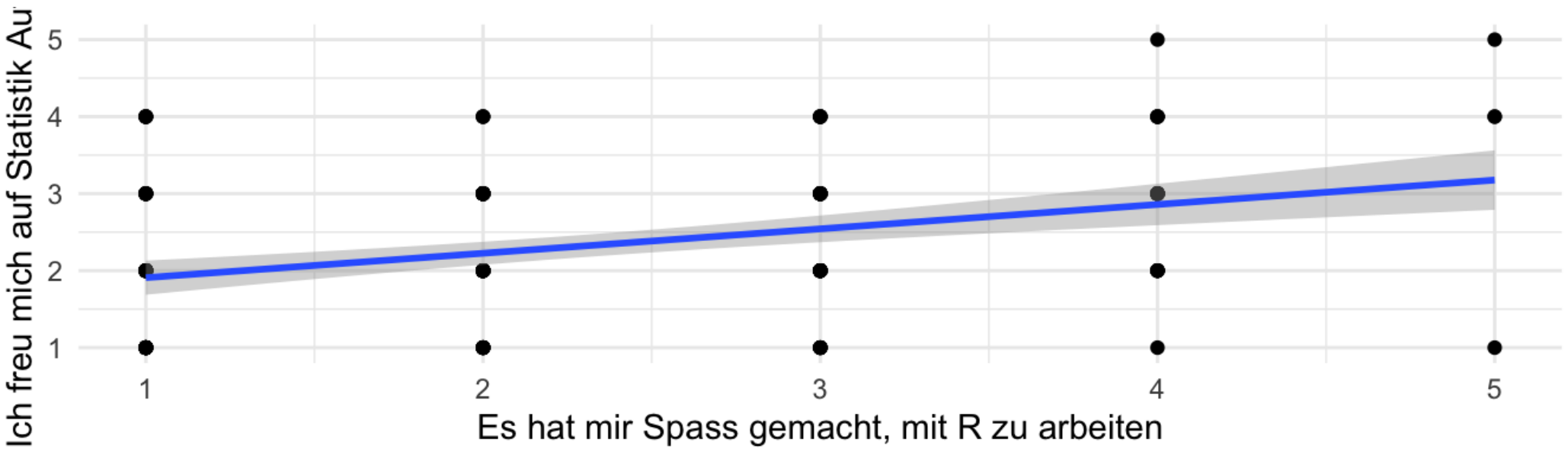
Was geht mit Formeln?

- R-Code anzeigen



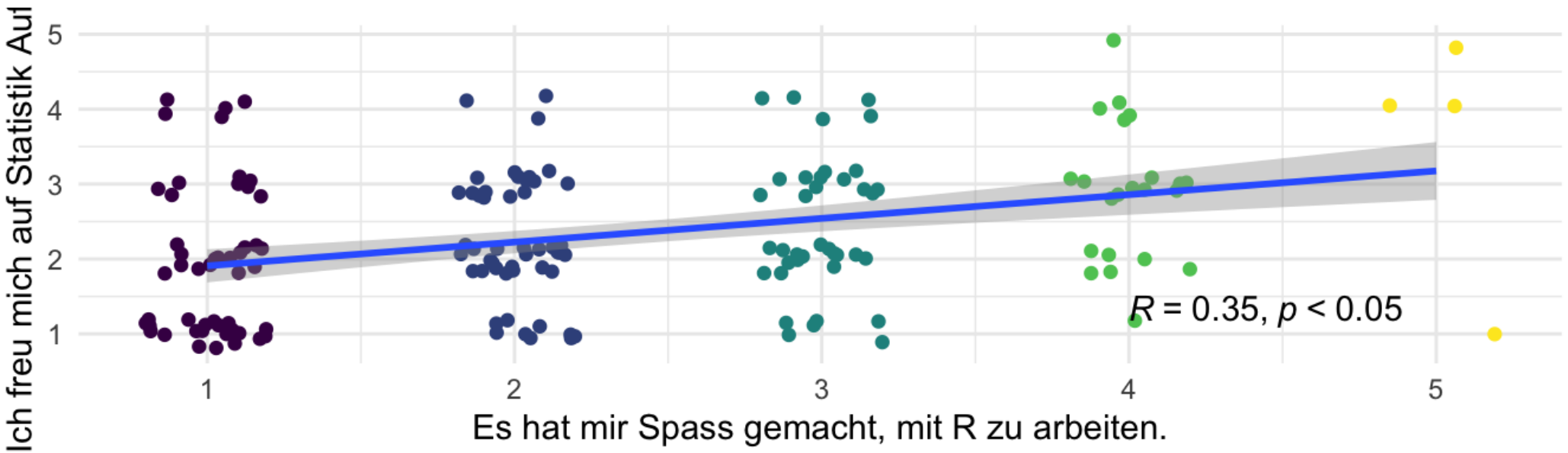
Spass und Freude

- R-Code anzeigen



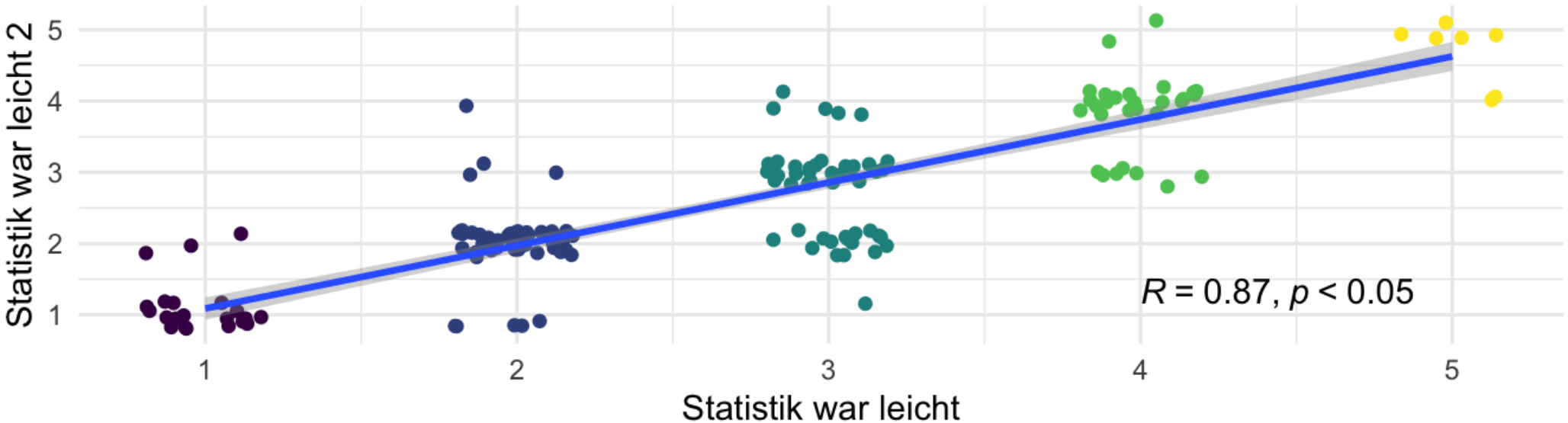
Statistik war leicht

- R-Code anzeigen



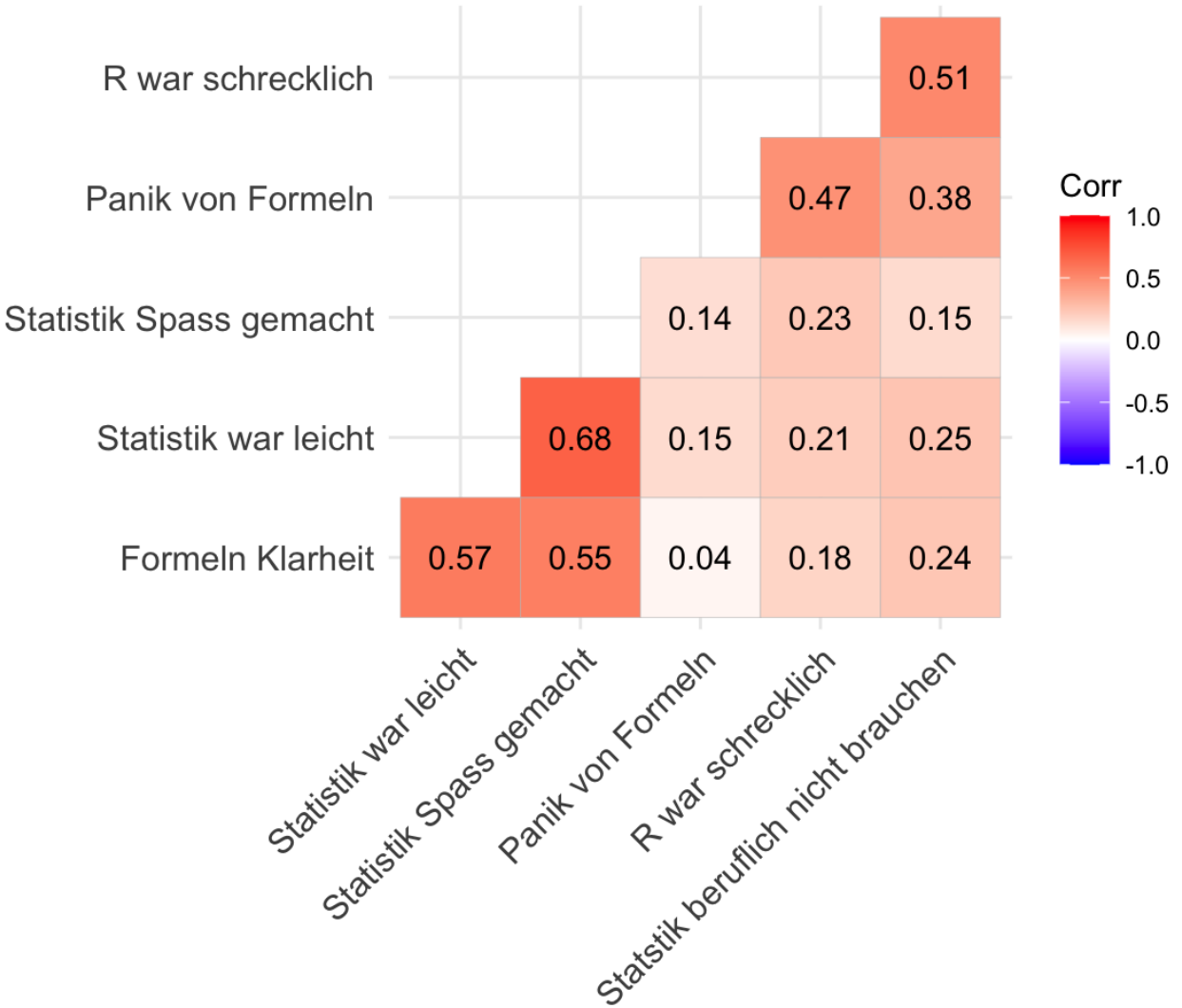
Statistik war leicht (doppelt)

- R-Code anzeigen



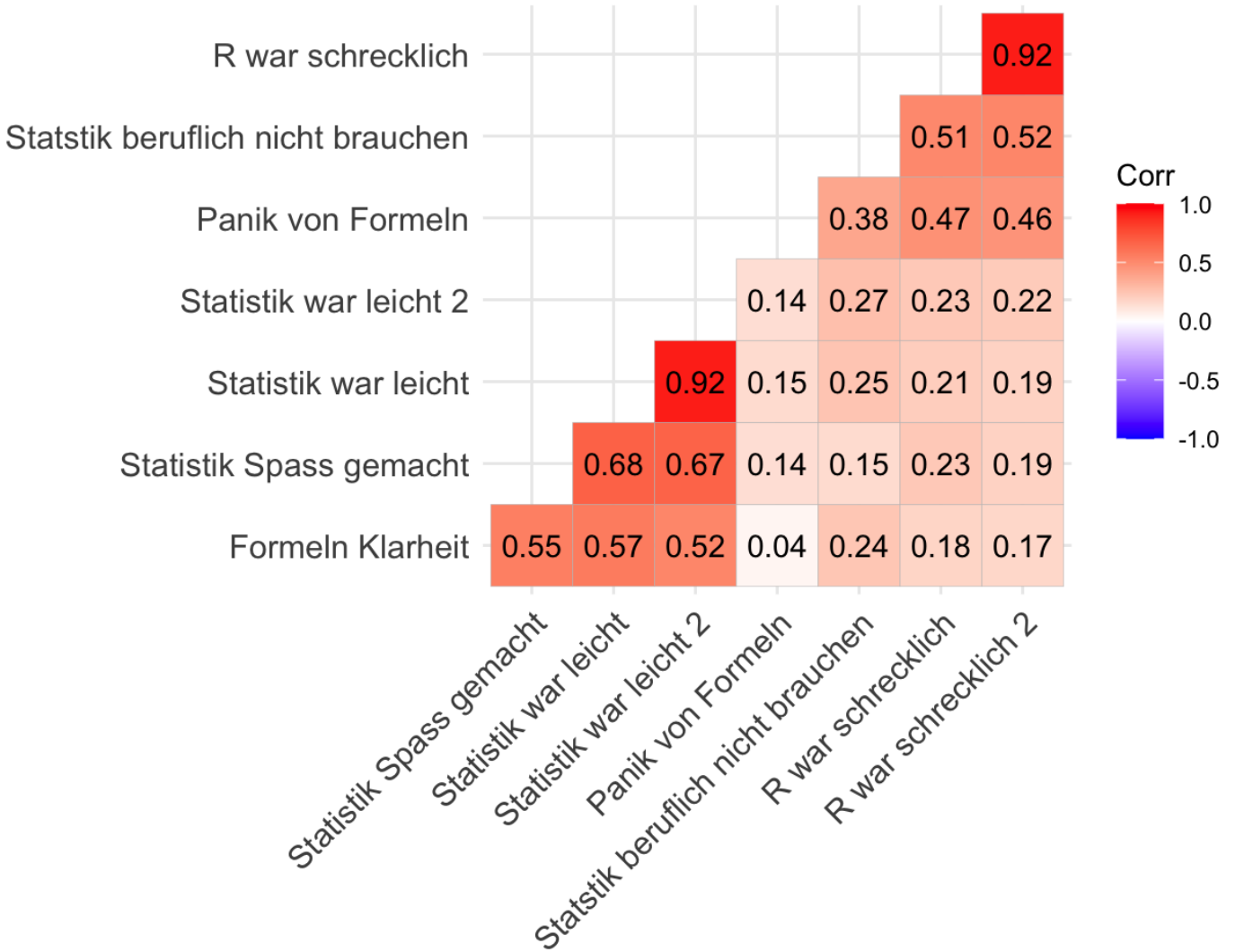
Korrelationen

- R-Code anzeigen



Korrelation doppelte Frage

- R-Code anzeigen



3 Was bisher geschah

3.1 Univariate Statistik

3.1.1 Mittelwert

Der Mittelwert ist das Gleiche wie ein Durchschnitt.

$$\overline{x} = \frac{1}{n} \sum_{i}^{n} (x_i)$$

$$\overline{y} = \frac{1}{n} \sum_{i}^{n} (y_i)$$

Mittelwert wird geschrieben als Kennwert $\overline{x}$ und Parameter μ.

3.1.2 Varianz

Die Varianz ist er Mittelwert (Durchschnitt) der quadrierten Abweichungen vom Mittelwert.

$$\sigma^2 = V = \frac{1}{n} \sum_{i}^{n} (x_i - \mu)^2$$

Die Varianz als unbekannter Parameter wird als σ^2 gekennzeichnet.

3.1.3 Standardabweichung

Die Standardabweichung ist die Wurzel der Varianz. Gefühlt ist die Standardabweichung sowas wie die durchschnittliche Abweichung (Beträge) vom Mittelwert (eben durch die Quadrierung und Rückrechnung über die Wurzel nicht ganz dasselbe).

$$s_x = \hat{\sigma}_x = \sqrt{\frac{1}{n-1}\sum_{i}^{n}(x_i - \overline{x})^2}$$

Die Standardabweichung als unbekannter Parameter wird als σ_x gekennzeichnet und der Kennwert als s_x.

Standardabweichung im Vergleich

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

<detached>

Standardfehler

Definition und Eigenschaften

Die Standardabweichung einer Stichprobenkennwerte-Verteilung nennt man *Standardfehler (standard error, SE)*. Wenn es sich um eine Verteilung von *Mittelwerten* handelt *Standardfehler des Mittelwerts* ($s_{\overline{x}}$).

Der Standardfehler gibt an, wie gut $\overline{x}$ den Populationsparameter μ schätzt.

Eigenschaften

- Der Standardfehler jedes Kennwertes nimmt mit grösser werdendem n ab 👉 $\sigma_{\overline{x}}$ ist umgekehrt proportional zu $\sqrt{n}$.
- Je stärker das gemessene Merkmal in der Population streut, desto grösser der Standardfehler 👉 $\sigma_{\overline{x}}$ ist proportional zu σ.

Die Formel:

$$\sigma_{\overline{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Effekte des Stichprobenumfangs

Standardisierung

Standardisierung

Standardisierung von Verteilungen und Kennwerten macht alles vergleichbar.

- $z_i = \frac{x_i - \overline{x}}{s}$.
- Eine z-Verteilte Grösse hat immer: $\overline{x} = 0$ und $s = 1$.
- Standardisierte Verteilungen und Kennwerte sind vergleichbar.

Konfidenzintervalle für Mittelwerte

Konfidenzintervalle

Konfidenzintervalle geben einen Wertebereich an, in dem die Parameter (GG) der Stichprobenkennwerte mit einer angebbaren Wahrscheinlichkeit liegen.

$$\text{KI: } \overline{X} \pm z_1 \cdot SE$$

$$\text{KI: } \overline{X} \pm z_1 \cdot \frac{s_x}{\sqrt{n}}$$

$$\text{KI}_{l.05} = \overline{x} - 1.96 \cdot \frac{s_x}{\sqrt{n}}$$

$$\text{KI}_{r.05} = \overline{x} + 1.96 \cdot \frac{s_x}{\sqrt{n}}$$

3.2 Bivariate Statistik

3.2.1 Kovarianz und Korrelation

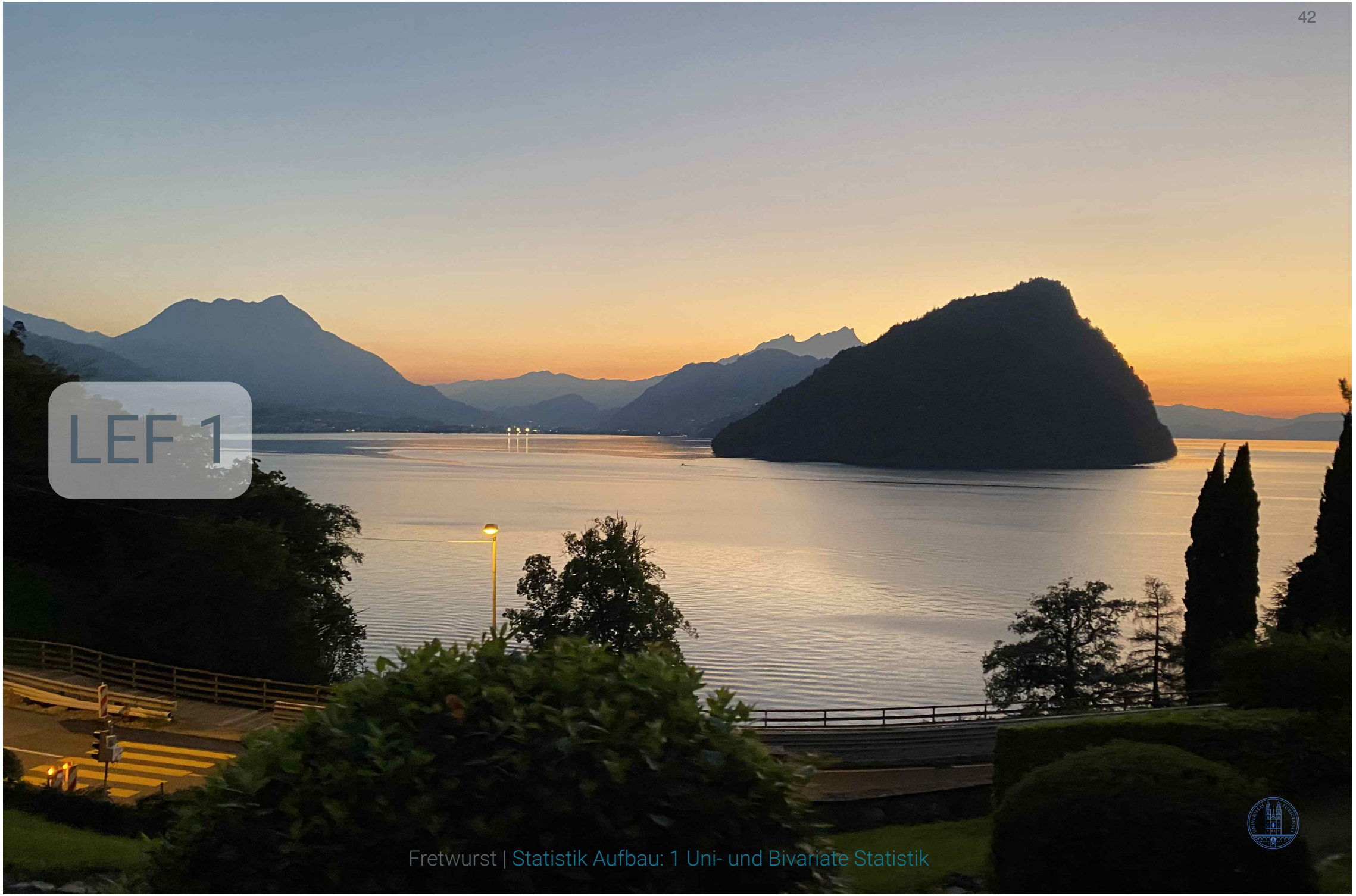
$$cov = C = \frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$r = \frac{\sum_{i}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n \cdot s_x \cdot s_y}$$

3.3 Hypothesentesten

Testen gegen die Nullhypothese

1. Könnte in der Auswahlgesamtheit der wahre Wert auch 0 sein, oder ein anderes Vorzeichen haben?
2. Die Nullhypothese ist eine statistische Hypothese gegen Falschentscheidungen aufgrund von Zufallsziehungen.
3. Nullhypothesen werden anhand von bekannten Verteilungen getestet.


LEF 1

Essayfragen

1. Was ist der Unterschied zwischen unstandardisierten und standardisierten Kennwerten?
2. Welche Masse der zentralen Tendenz kennen Sie?
3. Welche Streumasse kennen Sie?
4. Was kommt raus, wenn man die Covarianz einer Variablen mit sich selbst berechnet?
5. a. Welche Skalenniveaus kennen Sie?
   b. Was macht eine Nominalskala aus?
   c. Was macht eine metrische Skala aus?

MC-Fragen

MC 1.1.

MC 1.1: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Der Mittelwert ist dasselbe wie der Durchschnitt.
○	○	Der Mittelwert einer dichotomen Variablen entspricht dem Prozentsatz der 1-Werte.
○	○	Der Mittelwert wird auch als "Mittel" oder "Arithmetisches Mittel" bezeichnet.
○	○	Je grösser ein Mittelwert ist, desto eher ist er signifikant.

Punkte: 0

MC 1.2.

MC 1.2: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Die Standardabweichung ist die standardisierte Form der Varianz.
○	○	Die Standardabweichung hat n-1 Freiheitsgrade.
○	○	Die Standardabweichung "s" liegt immer zwischen 0 und 1.
○	○	Die Standardabweichung ist die durchschnittliche Abweichung vom Mittelwert.

Punkte: 0

MC 1.3.

MC 1.3: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Die Covarianz ist das Quadrat der Korrelation.
○	○	Die Covarianz ist skalenabhängig und kann daher negativ oder positiv und unendlich klein oder gross sein.
○	○	Die Korrelation "r" liegt immer zwischen 0 und 1.
○	○	Eine Korrelation von genau 0 kann nie signifikant sein.

Punkte: 0

MC 1.4.

MC 1.4: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Bei der bivariaten Korrelation sind r und die Wurzel aus R^2 identisch.
○	○	Das b einer bivariaten Regressionsgeraden liegt immer zwischen -1 und 1.
○	○	Wenn das b nicht signifikant ist, kann nicht mit 95% Wahrscheinlichkeit ausgeschlossen werden, dass es in Wirklichkeit Null ist oder ein zum Stichprobenwert entgegengesetztes Vorzeichen hat.
○	○	Wenn ein b signifikant ist, dann ist auch BETA signifikant.

Punkte: 0

MC 1.5.

MC 1.5: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Statistische Signifikanz bedeutet im Grunde wissenschaftliche Relevanz.
○	○	Bei kleinen Stichproben können Ergebnisse auch schon mal signifikant werden, obwohl die Effekte so kleine sind, dass sie zu vernachlässigen sind.
○	○	Je grösser n, desto schneller wird derselbe Effekt (z.B. Mittelwertunterschied) signifikant.
○	○	Wenn man einmal ein Signifkanzniveau (z.B. 95%) festgelegt hat, sollte man auch dabei bleiben.

Punkte: 0

MC 1.6.

MC 1.6: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Wenn das Konfidenzintervall eines Kennwertes kleiner als .05 ist, dann ist es signifikant.
○	○	Wenn das Konfidenzintervall eines Mittelwertes die 0 nicht einschliesst, dann ist der Mittelwert signifikant von 0 verschieden.
○	○	Wählt man ein höheres Signifikanzniveau (z.B. 99% statt 95%), dann wird das Konfidenzintervall breiter.
○	○	Je breiter ein Konfidenzintervall, desto besser ist ein Kennwert interpretierbar.

Punkte: 0

Für LEF 1: 0 von 12 Punkten, was 0% und etwa einer 1 entspricht.

Take Home

Statistik

- ist ein Modell von Realität
- ist eine mächtige Denkweise
- ist Basis und Kern von Data Science

Aus Statistik Einführung muss sitzen

- Mittelwerte
- Varianz/ Standardabweichung s- Covarianz/ Korrelation

Ausblick

Grundlagen der Modellbildung

Regression mit zwei Unabhängigen

Inferenzstatistische Grundlagen

Punktschätzung, Intervallschätzung, Wahrscheinlichkeitsverteilungen

Grundannahmen von OLS-Schätzungen

mässige Multikollniearität, Homoskedastizität, Modellspezifikation, Linearität