CHARLOTTE FRESENIUS
HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Statistik I

Einheit 4: Visualisierung

07.11.2024 | Prof. Dr. Stephan Goerigk

Wiederholung Einheit 3 – Was können wir schon?

- Kenntnis der **Skalennveaus**
- Kenntnis univariater Deskriptivstatistiken
- Wissen, wann welche Deskriptivstatistiken gerechnet werden sollten

Kompetenzen:

- Berechnen von **absoluten/relativen Häufigkeiten** und Darstelltung in **Häufigkeitstabelle**
- Berechnen von Summen und Notation mit **Summenzeichen**.
- Berechnen von Modus, Mittelwert (mehrere Gruppen, aus Häufigkeitstabelle), Median, Spannweite, Varianz, Standardabweichung, Quantile, Quartilabstand

Agenda für Heute:

Ziel: Effiziente Beschreibung von 2 Variablen (bivariate Statistik)

Was wir danach kennen werden:

- Eigenschaften unterschiedlicher statistischer Graphen
- Situationen in denen diese eingesetzt werden sollten

Was wir danach können werden:

- Graphen für 2 oder mehr Variablen erstellen

Visualisierung

Graphen

Wann nutzen wir Graphen?

Univariat:

- Verteilungen einer Variable visualisieren
    - Histogramm (numerische Variable; zeigt Häufigkeiten)
    - Balkendiagramm (kategoriale Variable; zeigt Häufigkeiten)
    - Boxplot (zumindest ordinale Variable; zeigt Median und IQR)

Multivariat:

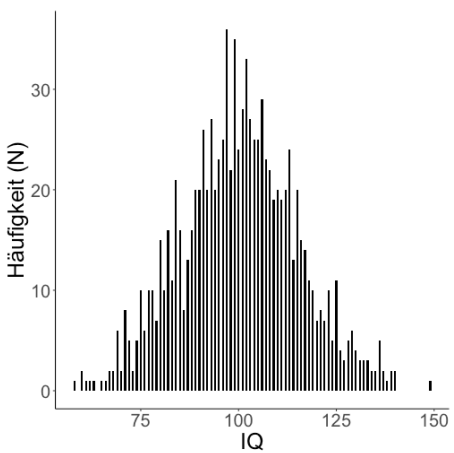
- Ausprägung einer Variable innerhalb Kategorien einer anderen Variable (Gruppenvergleich)
    - Balkendiagramm (Höhe des Balkens = stat. Kennwert einer Gruppe, z.B. Mittelwert)
    - Boxplot (zumindest ordinale Variable; zeigt Median und IQR einer Gruppe)
- Gemeinsames Variieren von 2 (kontinuierlichen) Variablen
    - Streudiagramm (Punkt = Messwertpaar 2er Variablen für 1 Person im Datensatz)

Visualisierung

Graphen

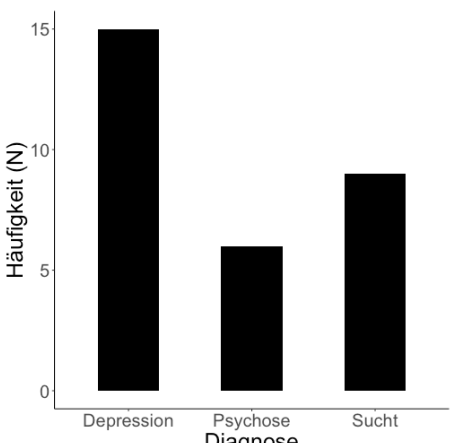
Univariat (haben wir bereits gelernt)

Histogramm für numerische Variablen:



- **X-Achse:** Ausprägungen der Variablen
- **Y-Achse:** Absolute/relative Häufigkeiten dieser Ausprägungen

Balkendiagramm für kategoriale Variablen:



- **X-Achse:** Ausprägungen der Variablen
- **Y-Achse:** Absolute/relative Häufigkeiten dieser Ausprägungen

Visualisierung

Graphen

Bivariat

- Wir wissen nun, wie wir Häufigkeiten (N) in Histogrammen und Balkendiagrammen dastellen können
- Manchmal wollen wir jedoch auch andere Statistiken darstellen (z.B. Mittelwert)
- Häufiger Anwendungsfall: Vergleich der Ausprägung einer Variable in 2 Gruppen

---

Beispiel: Grübelneigung in Gruppe A (Depression) und B (Keine Depression)

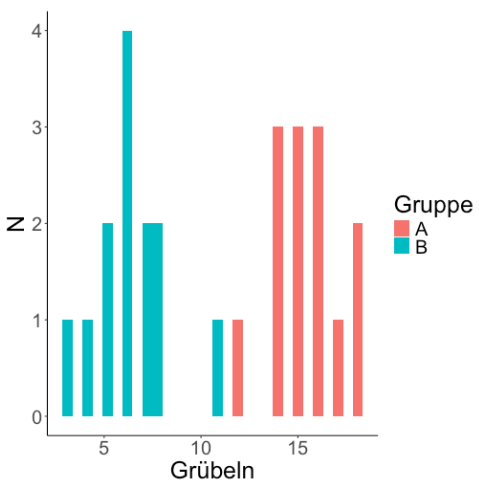
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Gruppe	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B
Grübeln	14	15	18	15	15	18	16	12	14	14	17	16	16	7	6	11	8	3	8	6	5	7	5	6	6	4

Visualisierung

Graphen

Bivariat (nun haben wir 2 Variablen, Grübeln + Gruppe)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Gruppe	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B
Grübeln	14	15	18	15	15	18	16	12	14	14	17	16	16	7	6	11	8	3	8	6	5	7	5	6	6	4



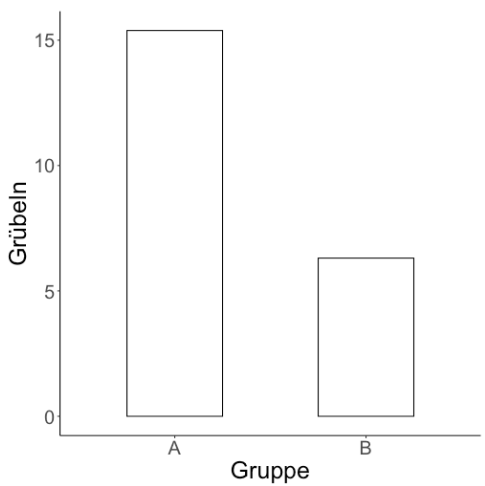
- 2 Gruppen auch im Histogramm darstellbar
- Y-Achse = N
- X-Achse = Merkmalsausprägungen
- Gruppen als Farben codiert
- Maß der zentralen Tendenz abschätzbar, aber nicht eindeutig

Visualisierung

Graphen

Bivariat (nun haben wir 2 Variablen, Grübeln + Gruppe)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Gruppe	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B
Grübeln	14	15	18	15	15	18	16	12	14	14	17	16	16	7	6	11	8	3	8	6	5	7	5	6	6	4



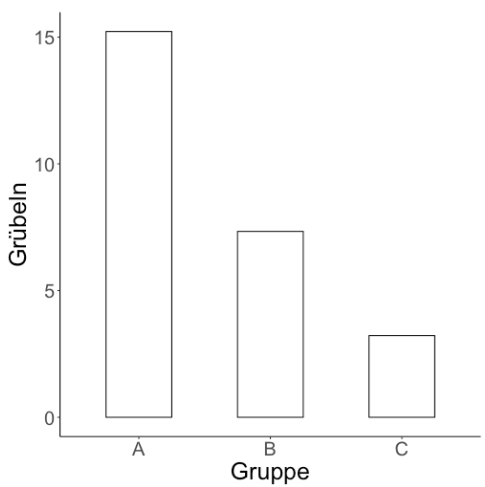
- Balkendiagramm ist übersichtlicher
- Y-Achse = numerisch (oft AV)
- X-Achse = kategorial (oft UV)
- Balkenhöhe = Mittelwert

Visualisierung

Graphen

Bivariat - nun haben wir 2 Variablen, Grübeln (numerisch) + Gruppe (Kategorie)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Gruppe	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	C	C	C	C	C	C	C	C	C
Grübeln	14	15	18	15	15	18	16	12	14	6	9	8	8	7	6	11	8	3	5	3	2	4	2	3	3	1	6



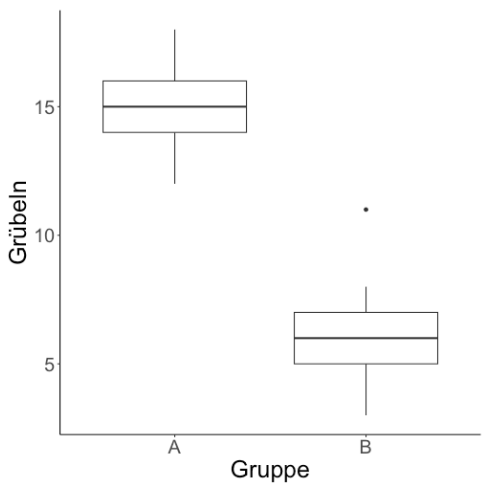
- Balkendiagramm ist übersichtlicher
- Y-Achse = numerisch (oft AV)
- X-Achse = kategorial (oft UV)
- Balkenhöhe = Mittelwert
- Auch 3 oder mehr Gruppen sind darstellbar
- Nicht verwechseln: Immer noch ein bivariat (aber Gruppe hat 3 Stufen)

Visualisierung

Graphen

Bivariat (nun haben wir 2 Variablen, Grübeln + Gruppe)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Gruppe	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B
Grübeln	14	15	18	15	15	18	16	12	14	14	17	16	16	7	6	11	8	3	8	6	5	7	5	6	6	4



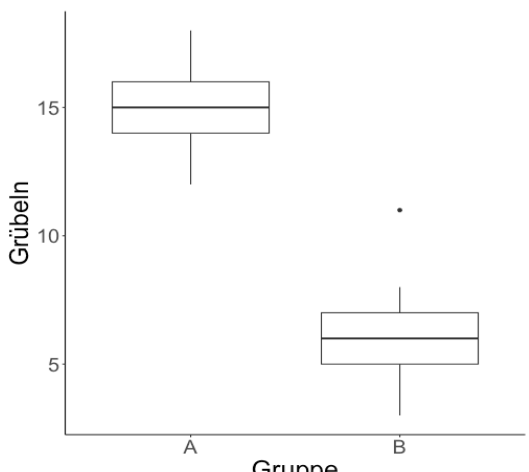
Boxplot:

- Alternative zum Balkendiagramm
- Y-Achse = numerisch (oft AV)
- X-Achse = kategorial (oft UV)
- Mittellinie = Median
- Median → auch für Ordinalskala geeignet

Visualisierung

Graphen

Bivariat (nun haben wir 2 Variablen, Grübeln + Gruppe)



Bestandteile Boxplot:

- Mittellinie = Median
- Box = Quartilabstand (IQR, 50% aller Werte)
- Whiskers = maximale Länge von IQR mal 1.5
- Punkte außerhalb der Whiskers = Ausreißer
- Y-Achse = numerisch (oft AV)
- X-Achse = kategorial (oft UV)
- 1 Box pro Gruppe (kann theoretisch auch ungruppiert dargestellt werden)

Visualisierung

Graphen

Anleitung zum Boxplot erstellen:

1. Berechne den Median, der den Datensatz in zwei Hälften unterteilt
2. Berechne IQR und somit die Länge der Box, indem du die Mediane der beiden Datensatz-Hälften berechnest
3. Überlege dir eine sinnvolle Skalierung und zeichne die Box mit dem Median darin ein (keine Sorge, die Box ist aufgrund der Rangskalierung nicht immer symmetrisch)
4. Berechne die **potentielle** maximale Länge der Whisker, indem du IQR mit 1.5 multiplizierst
5. Zeichne die Whisker ein – Achtung: Whisker werden nur bis zum letzten Wert eingezeichnet, der noch innerhalb der potenziellen maximalen Länge der Whisker liegt!
6. Zeichne eventuelle Ausreißer und Extremwerte ein – mit Punkten/Sternchen

Visualisierung

Graphen

Aufgabe:

Fertige jeweils ein Balkendiagramm und ein Boxplot für folgenden Gruppenunterschied an:

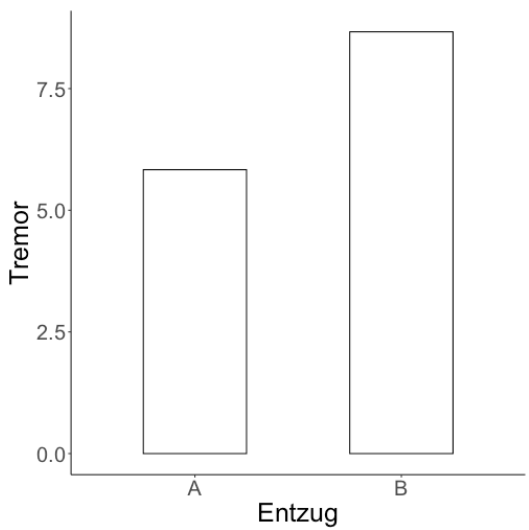
Intensität im Rating des Tremors (Händezittern) zwischen heroinabhängigen Patienten vor dem Entzug (A) und heroinabhängigen Patienten während des Entzugs (B)

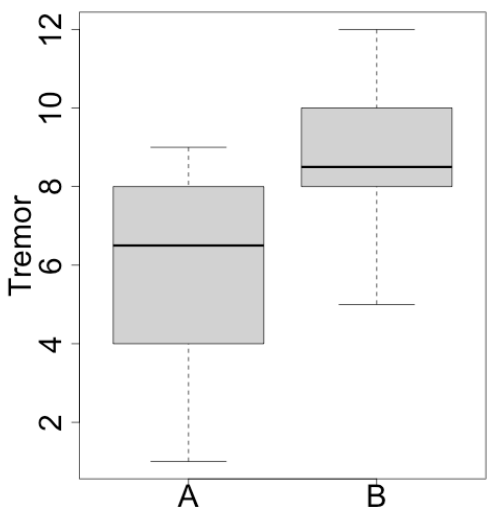
ID	1	2	3	4	5	6	7	8	9	10	11	12
Entzug	A	A	A	A	A	A	B	B	B	B	B	B
Tremor	1	4	6	7	8	9	5	8	12	10	8	9

Visualisierung

Graphen (bivariat)

Lösung:





Visualisierung

Graphen (bivariat)

Lösung für Gruppe A:

$$x_{sortiert} = 1; 4; 6; 7; 8; 9$$

$$Md = \frac{6 + 7}{2} = 6.5$$

$$n \cdot \alpha = 6 \cdot 0.25 = 1.5 \rightarrow I_{0.25} = 2$$

$$n \cdot \alpha = 6 \cdot 0.75 = 4.5 \rightarrow I_{0.75} = 5$$

$$q_{0.25} = x_{(I_{0.25})} = x_{(2)} = 4$$

$$q_{0.75} = x_{(I_{0.75})} = x_{(5)} = 8$$

$$IQR \cdot 1.5 = 6$$

$\rightarrow$ alle Werte < -2 und > 14 sind Ausreißer. Wir zeichnen jedoch nur bis zu den maximal verfügbaren Werten (1 und 9).

Visualisierung

Graphen

Bivariat (2 numerische Variablen)

- Wir können mit Balkendiagramm/Boxplot nun 2 Variablen darstellen: numerisch (Y) + Kategorie (X)
- Was aber, wenn wir 2 numerische Variablen haben?

→ Wir bräuchten sehr viele Balken/Boxen (so viele wie Ausprägungen auf X-Achse)

Beispiel:

Zwei numerische Variablen Nachtschlaf (in Stunden) und Leistung in einem Konzentrationstest (0-150 Punkte)

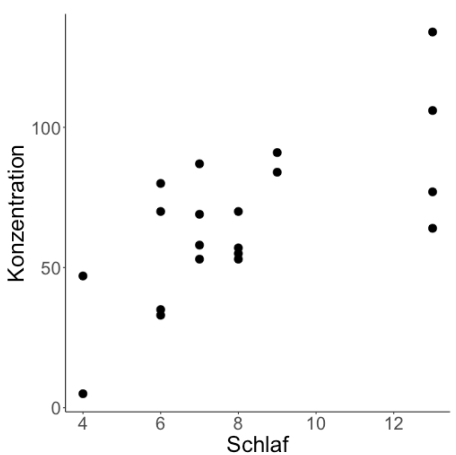
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Schlaf	6	7	13	8	8	13	9	4	6	7	6	7	13	8	8	13	9	4	6	7
Konzentration	80	69	106	70	57	134	84	5	70	53	35	58	77	53	55	64	91	47	33	87

Visualisierung

Graphen

Bivariat (2 numerische Variablen)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Schlaf	6	7	13	8	8	13	9	4	6	7	6	7	13	8	8	13	9	4	6	7
Konzentration	80	69	106	70	57	134	84	5	70	53	35	58	77	53	55	64	91	47	33	87



Streudiagramm:

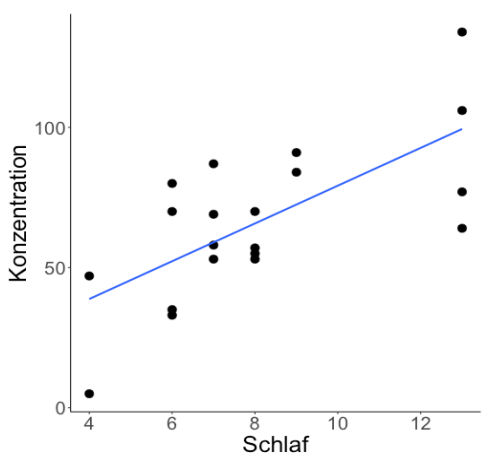
- Punktewolke = Wertepaare (Jeder Punkt = 1 Person)
- Y-Achse: numerische Variable (AV, wenn Wirkrichtung angenommen)
- X-Achse: numerische Variable (UV, wenn Wirkrichtung angenommen)
- Trends lassen sich gut erkennen

Visualisierung

Graphen

Bivariat (2 numerische Variablen)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Schlaf	6	7	13	8	8	13	9	4	6	7	6	7	13	8	8	13	9	4	6	7
Konzentration	80	69	106	70	57	134	84	5	70	53	35	58	77	53	55	64	91	47	33	87



Streudiagramm:

- Später können wir auch lineare Zusammenhänge zwischen den Variablen visualisieren
- hier z.B. positiver Zusammenhang (hohe Werte auf X gehen einher mit hohen Werten auf Y)
- Dafür müssen wir noch lernen, wie eine Regressionsgerade berechnet wird

Visualisierung

Graphen

Aufgabe:

Erstellen Sie ein Streudiagramm für die beobachtete numerische Lerndauer (in Stunden) und Prüfungsleistung (1-60 Punkte)

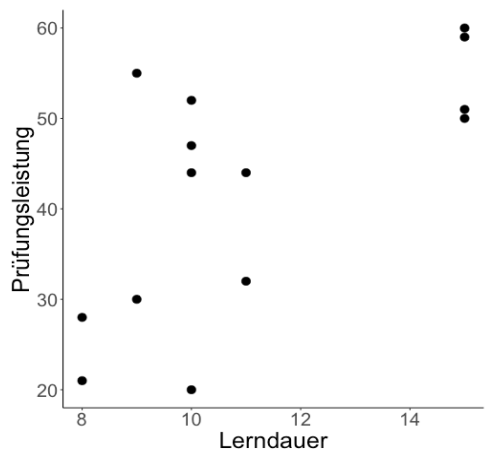
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Lerndauer	8	9	15	10	10	15	11	8	9	15	10	10	15	11
Prüfungsleistung	21	30	51	52	44	59	44	28	55	60	20	47	50	32

Visualisierung

Graphen

Lösung:

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Lerndauer	8	9	15	10	10	15	11	8	9	15	10	10	15	11
Prüfungsleistung	21	30	51	52	44	59	44	28	55	60	20	47	50	32



Visualisierung

Graphen

Bivariat (2 kategoriale Variablen)

Beispiel:

2 Kategoriale Variablen: (1) Entzugsstatus heroinabhängiger Patienten (während des Entzugs = A, vor dem Entzug = B) und (2) Tremor (ja/nein)

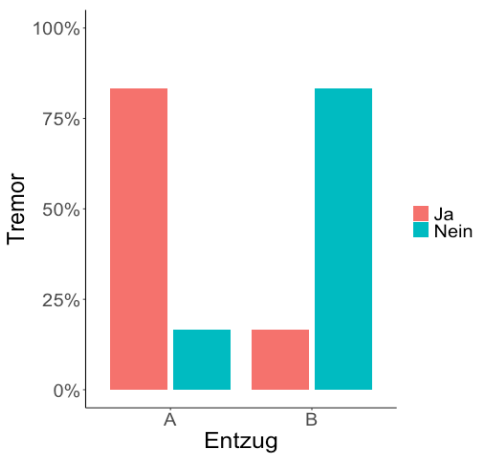
ID	1	2	3	4	5	6	7	8	9	10	11	12
Entzug	A	A	A	A	A	A	B	B	B	B	B	B
Tremor	Ja	Nein	Ja	Ja	Ja	Ja	Nein	Nein	Nein	Ja	Nein	Nein

Visualisierung

Graphen

Bivariat (2 kategoriale Variablen)

ID	1	2	3	4	5	6	7	8	9	10	11	12
Entzug	A	A	A	A	A	A	B	B	B	B	B	B
Tremor	Ja	Nein	Ja	Ja	Ja	Ja	Nein	Nein	Nein	Ja	Nein	Nein



Balkendiagramm:

- Hier funktioniert wieder das Balkendiagramm
- Y-Achse: absolute/relative Häufigkeiten
- X-Achse: kategoriale Variable (UV, wenn Wirkrichtung angenommen)
- Da uns die Achsen für die Darstellung der unterschiedlichen Ausprägungen ausgehen, müssen wir auf Farben zurückgreifen

Visualisierung

Graphen

Zusammenfassung (univariat und bivariat)

Variable1	Variable2	Graph
Numerisch	Keine	Histogramm
Kategorial	Keine	Balkendiagramm (Balken = Häufigkeit)
Kategorial	Numerisch (metrisch)	Balkendiagramm (Balken = Mittelwert)
Kategorial	Numerisch (ordinal)	Boxplot (Linie = Median)
Numerisch	Numerisch	Streudiagramm
Kategorial	Kategorial	Balkendiagramm (Balken = Häufigkeit, Kategorien Variable2 als Farben)

Visualisierung

Graphen

Multivariat (mehr als 2 Variablen)

Beispiel:

2 numerische Variablen (AV = Angst, UV = Verschmutzungsintensität) und 1 kategoriale Variable Diagnose (A = Zwangsstörung, B = Keine Zwangsstörung)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Verschmutzungsintensität	9	10	13	10	10	13	11	7	9	9	12	11	11	10	9	14	11	6	11	9
Diagnose	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B
Angst	45	58	62	53	54	55	72	50	45	69	5	8	1	24	17	3	7	4	15	8

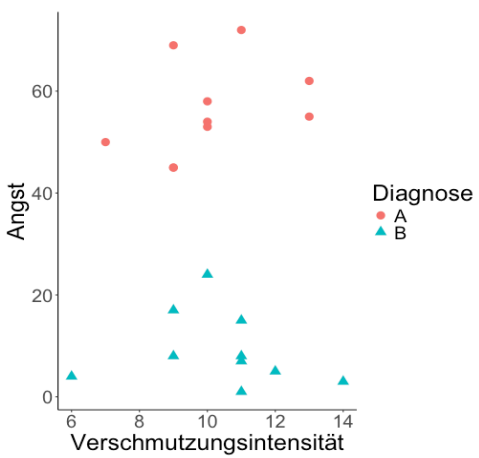
Visualisierung

Graphen

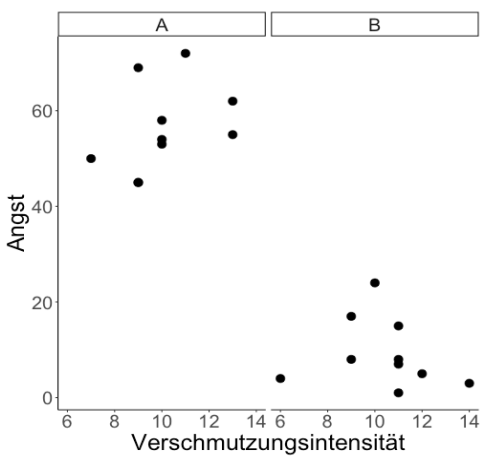
Multivariat (mehr als 2 Variablen)

Mangels weiterer Achsen (als X und Y) behelfen wir uns mit sogenannten Aesthetics (Farben, etc.)

Mit unterschiedlichen Farben und Punktformen:



Mit unterschiedlichen Facetten:

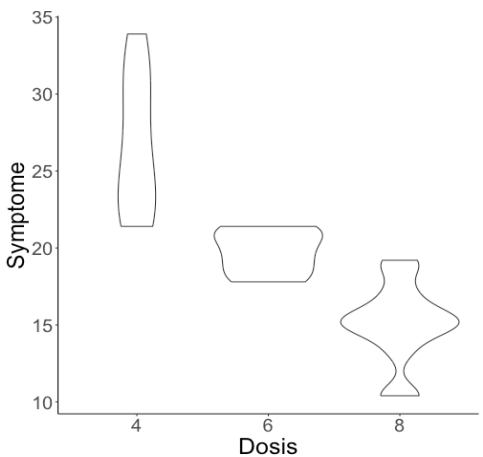


Visualisierung

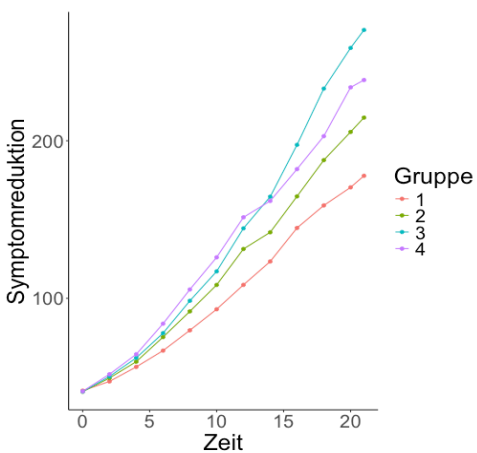
Graphen

Weitere Graphen (nicht prüfungsrelevant)

Violin Plot (ähnliche Verwendung wie Boxplot)



Line Graph (Veränderung über die Zeit bei Messwiederholungen)



Take-aways

- **Graphen** geben uns einen schnellen Überblick über statistische Ergebnisse
- Nützlich zur schnellen **Exploration** von Daten oder als **Visualisierung einer Hypothese** (im Paper)
- Wir können einzelne Verteilungen (**univariat**) oder Kombinationen von Variablen darstellen (**bi-/multivariat**)
- **Balkendiagramm und Boxplot** eigenen sich zur Darstellung von Gruppenunterschieden
- **Vorteile Boxplot:** Median und IQR auch für ordinale Skalen angemessen, zeigt Ausreißer
- **Streudiagramm** eignet sich zur Darstellung von **Merkmalszusammenhängen** zweier numerischer Variablen
- Bei Kombination mehrerer Variablen nutzen wir zusätzlich zu den Achsen **Aethetics** (Farben, Facetten...)