M24 Statistik 1: Wintersemester 23/24

Vorlesung 04: Darstellung von Daten

Prof. Matthias Guggenmos

Health and Medical University Potsdam

Darstellung von Lage- und Streuungsmaßen in Text und Bild

Der Forschungsprozess

Auf Basis einer neuen/aktualisierten Theorie ergeben sich häufig neue Fragen und Hypothesen

Wundern, Raten, Fragen
(best case: bereits eine Theorie)

↓

Hypothese

↓

Operationalisierung

↓

Experiment / Studie

Player	Minutes	Points	Rebounds
A	41	20	6
B	30	29	7
C	22	7	7
D	26	3	3
E	20	19	8

↓

Daten

↓

Auswertung / Ergebnis

↓

Schlussfolgerung / Interpretation

↓

Aufstellen oder Update einer Theorie

Angabe von Lage- und Streuungsmaßen in wissenschaftlichen Arbeiten

- Handelt es sich um einzelne Werte, können diese übersichtlich im Fließtext berichtet werden:

> Respondents' mean rating for self-estimated musicality was 6.1 (SD = 2.4), and the mean rating for importance of music in their life was 8.2 (SD = 1.8). Thus, partici-

- Handelt es sich um eine größere Anzahl von Werten (z.B. bei mehreren Bedingungen) bietet sich eine Darstellung in Tabellenform an.
- Auch bei dieser Darstellung sollten Lage- *und* Streuungsmaße angegeben werden.
- In wissenschaftlichen Manuskripten werden Tabellen häufig nach den APA-Richtlinien (7. Edition, 2020) formatiert:

Musikstil	*M*	*SD*
Klassik	2,9	1,1
Rock	4,1	0,8
Rap	3,3	1,2

Siehe Link[1] für Informationen zur Formatierung von Tabellen im APA-Stil

Beispieltabelle

TABLE 4 *Mean ratings (and SDs) for the functions of music within the six dimensions of musical styles*

	Preference dimensions (genres)						
Statement	All styles (N = 503)	Sophisticated (n = 38)	Electronic (n = 30)	Rock (n = 218)	Rap (n = 37)	Pop (n = 88)	Beat, folk, & country music (n = 10)
Puts me in a good mood	8.2 (1.3)	8.3 (1.6)	8.4 (1.0)	8.3 (1.2)	8.0 (1.9)	7.8 (1.4)	8.7 (0.5)
Helps me chill and tune out	6.9 (2.3)	7.6 (2.1)	6.9 (2.6)	6.8 (2.4)	6.4 (2.9)	7.2 (1.7)	7.3 (2.2)
Energizes me	6.8 (2.0)	6.2 (2.3)	7.3 (1.8)	6.8 (1.9)	7.2 (2.0)	6.4 (2.1)	7.6 (1.8)
Lets me appreciate as art	6.5 (2.6)	8.1 (1.5)	6.4 (2.6)	6.7 (2.5)	5.7 (3.1)	5.7 (2.9)	6.3 (2.9)
Enables me to reminisce	6.4 (2.5)	5.2 (2.9)	6.1 (2.7)	6.8 (2.2)	6.0 (2.3)	6.6 (2.4)	7.8 (1.1)
Enables me to better understand my thoughts and feelings	6.0 (2.5)	5.8 (3.0)	5.5 (2.8)	6.2 (2.4)	5.0 (3.1)	5.9 (2.3)	6.5 (2.0)
Is what I listen to as background music	5.7 (2.8)	5.4 (2.8)	6.2 (2.3)	5.5 (2.8)	5.5 (2.9)	6.2 (2.7)	5.0 (3.1)
Is what I like to dance to	5.6 (3.2)	2.8 (2.9)	7.5 (2.8)	5.7 (3.0)	7.4 (2.3)	5.4 (3.0)	5.5 (3.2)
Expresses my identity	5.5 (2.7)	5.5 (2.3)	5.2 (2.4)	6.0 (2.6)	4.7 (2.9)	4.9 (2.7)	6.3 (2.8)
Expresses my values	5.5 (2.6)	5.7 (2.6)	4.1 (2.9)	6.0 (2.5)	5.2 (2.4)	5.1 (2.5)	6.3 (2.5)
Lets me forget my problems	5.5 (2.6)	5.8 (2.3)	6.0 (2.4)	5.3 (2.7)	4.5 (2.9)	5.1 (2.4)	6.6 (1.8)
Helps me feel close to others	5.4 (2.7)	5.0 (3.1)	5.2 (2.8)	5.6 (2.6)	5.1 (2.9)	5.3 (2.5)	5.3 (2.8)
Makes me feel ecstatic	4.9 (3.2)	3.8 (3.1)	6.5 (3.1)	5.3 (3.0)	4.3 (3.2)	3.6 (3.0)	6.8 (1.8)
Lets me experiment with different sides of my personality	4.8 (3.0)	3.6 (2.8)	4.7 (3.4)	5.1 (2.9)	4.8 (3.1)	4.4 (2.9)	4.6 (3.7)
Helps me meet people	4.4 (2.8)	3.6 (3.0)	5.1 (3.0)	4.7 (2.8)	4.3 (2.8)	3.8 (2.8)	4.0 (2.7)
Makes me identify with the artists	3.4 (2.9)	3.5 (3.3)	3.3 (3.2)	3.6 (2.8)	3.2 (2.9)	2.7 (2.5)	4.3 (3.4)
Gives me information	3.3 (2.8)	3.5 (3.1)	1.9 (2.3)	3.6 (2.7)	2.8 (2.8)	2.8 (2.4)	4.1 (3.0)

Balkendiagramm mit Fehlerbalken

- Der **Fehlerbalken** ist die klassische Wahl zur Darstellung der Streuung von Daten.
- Die Ausdehnung des Fehlerbalkens entspricht dem Lagemaß (z.B. Mittelwert) plus/minus dem Streuungsmaß (z.B. Standardabweichung).
- Der gesamte Fehlerbalken hat also die Ausdehnung $2 \cdot \text{Streuungsmaß}$

- Klassische Kombinationen von Lagemaß und Streuungsmaß sind:
  - $\text{Mittelwert} \pm \text{Standardabweichung}$
  - $\text{Mittelwert} \pm \text{Standardfehler}$ (dazu kommen wir noch)
  - $\text{Mittelwert} \pm \text{Konfidenzintervall}$ (dazu kommen wir noch)
  - $\text{Median} \pm \text{IQR}$
  - $\text{Median} \pm \text{Median-Abweichung}$ (das behandeln wir nicht)

Beispiel

Musikstil	*M*	*SD*
Klassik	2,9	1,1
Rock	4,1	0,8
Rap	3,3	1,2

Einzelner Punkt statt Balken

- Statt eines Balkens kann der Mittelwert auch durch einen einzelnen Punkt repräsentiert werden – manche empfinden das als eleganter:

The case against bar plots

- Trotz ihrer hohen Verbreitung haben Balkendiagramme (Barplots) eine Reihe von Nachteilen:
  - Sie geben kaum Information über die spezifische Verteilung der Daten und mögliche Ausreißer
  - Die intransparente Darstellungsweise verdeckt häufig, dass 1) Daten unrealistisch sind oder 2) Ausreißer das Lagemaß verzerren oder 3) die Verteilung der Daten unpassend für das verwendete Lagemaß sind.
  - Sie legen den Fokus auf irrelevante Bereiche der Skala (siehe Abbildung unten)

Durch die tatsächliche Verteilung der Datenpunkte im rechten Plot wird klar, dass im Balkendiagramm ein vergleichsweise starker Fokus auf Bereiche gelegt wird, in den gar keine Daten enthalten sind ("Zone of Irrelevance"), und andererseits Extremwerte, insbesondere oberhalb des Fehlerbalkens, visuell völlig unrepräsentiert sind ("Zone of Invisibility"). Der Fehlerbalken ist außerdem leicht mit der Illusion verbunden, dass sich alle Datenpunkte innerhalb des angezeigten Bereiches befinden ("Within-the-bar bias").[2].

The case against bar plots

- Alle Datenverteilungen haben den gleichen Mittelwert und Standardabweichung

Illustration der Tatsache, dass unterschiedlichste Verteilungen von Daten im selben Balkendiagramm münden, hier am Beispiel von Mittelwert und Standardabweichung. Alle Verteilungen haben exakt den gleichen Mittelwert und Standardabweichung, sowohl entlang der y-Achse, als auch entlang der x-Achse (hier sind aber nur die Mittelwerte bezogen auf die y-Achse dargestellt)[3].

The case against bar plots

Alberto Cairo's "DataSaurus"[4].

Der Box-Plot

- Eine bekannte Darstellungsform für den Median ist der **Boxplot**
- In einer gängigen Variante zeigt der Boxplot drei Informationen an:
  1. **Median (als einfache Linie)**
  2. **Box:** die "mittleren 50% der Daten" (25% über dem Median, 25% unter dem Median)
  3. **Antennen ("Whiskers"):** zeigen Grenzen der Ausreißer-Definition an (Ausreißer = alle Punkte unter- und überhalb der Antennen). Häufig ist hier das Kriterium, dass die Daten im Bereich $[Q_1 - 1.5 \cdot IQR; Q_3 + 1.5 \cdot IQR]$ liegen müssen.

- Der Boxplot gibt eine schnelle Übersicht über wesentliche Kennwerte eines Datensatzen
- Zu beachten ist, dass zahlreiche Varianten des Boxplots existieren
  - Bei einer weiteren häufigen Variante zeigen die Antennen das absolute Maximum und Minimum der Daten an (also *inklusive* möglicher Ausreißer)

Moderne Darstellungsformen

Raincloud-Plot

Bildnachweis[5]

Klassische Darstellung mit Lage- und Streuungsmaß

Bildnachweis[6]

- Moderne Statistik- und Plotsoftware ermöglicht heutzutage eine verbesserte und transparentere Darstellung von Daten:
  - Anzeige einzelner Datenpunkte, meist getrennt durch “Jitter” (d.h. leichte horizontale oder vertikale Versetzung mit zufälligen Abständen, um Überschneidung der Datenpunkte zu reduzieren)
  - Anzeige der Verteilung mittels (interpolierter) Histogramme (wichtige Information für die Auswahl geeigneter Lage- und Streumaße, aber auch statistischer Tests)
  - *Zusätzliche* Anzeige von Kennwerten

Problematische Abbildungen

Abschneiden der y-Achse

- Das Abschneiden der y-Achse verzerrt häufig die Stärke von Effekten.

Kundenentwicklung wie von einer deutschen Bank dargestellt

Kundenentwicklung mit Beginn der y-Achse bei 0

- Die grundsätzliche Empfehlung ist daher, die y-Achse bei 0 beginnen zu lassen.
- Es gibt aber Ausnahmen:
  - Vorliegen eines anderen natürlichen Referenzwertes (z.B. IQ-Wert 100, wenn alle Werte über 100 liegen).
  - Wären *tatsächlich vorhandene Unterschiede* zwischen Balken verschiedener Bedingungen überhaupt nicht mehr wahrnehmbar, kann ein Abschneiden der y-Achse sinnvoll sein (oder eine Logarithmus-Skala!).
  - In manchen Fällen können Messwerte niemals unter einen Mindestwert fallen. Beispielsweise sind motorische Reaktionszeiten physiologisch bedingt fast immer über 100ms – in diesem Fall ist der Bereich 0-100ms “Totraum” und kann sinnvollerweise weggelassen werden.
  - Im Idealfall wird das Abschneiden der y-Achse durch einen “Bruch” angezeigt (siehe nächste Folie).

Beispiel für gebrochene y-Achse

Durch den Bruch der y-Achse wird betont, dass die Balken zur besseren Übersichtlichkeit “abgeschnitten” wurden. So werden fälschliche Wahrnehmungen und Interpretationen durch die abgeschnittene Achse eher vermieden. Bildnachweis[7]

Gleiche Intervalle auf x-Achse

- Ungleichmäßige Intervalle auf der x-Achse verzerren die Daten von Liniendiagrammen

Worst plot ever

Abbildung in der Financial Times[8]

Tatsächliche Rohdaten[9]

Verteilungen

Empirische Verteilungen

- Eine **Häufigkeitsverteilung** oder auch kurz **Verteilung** gibt zu verschiedenen Werten eines Merkmals an, wie häufig dieser Wert vorkommt.
- Die Häufigkeit kann entweder als **absolute** oder **relative Häufigkeit** angegeben werden.

- **Empirische Verteilungen** geben an, wie die tatsächlich gemessenen Daten einer Stichprobe verteilt sind. Das AirBnB-Beispiel zeigt empirische Verteilungen.
- **Theoretische Verteilungen** sind durch eine Funktion definiert, die die Verteilung von Daten mathematisch beschreibt.
  - Im Gegensatz zu empirischen Verteilungen geben theoretische Verteilungen die (erwartete) Häufigkeit *zu jedem möglichen Wert* des Merkmals an.

Normalverteilung

- Eine der wichtigsten theoretischen Verteilungen in der Psychologie ist die **Normalverteilung** (auch **Gauß-Verteilung**).
- Aufgrund ihrer Form wird sie umgangssprachlich auch als **Glockenkurve** bezeichnet.
- Mathematische Definition:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- μ kennzeichnet den Mittelwert und σ die Standardabweichung der Verteilung.

Zur Wiederholung: müssen statistische Kennwerte wie Mittelwert und Standardabweichung nicht aus Stichproben geschätzt werden, sondern sind als bekannt angenommen, werden sie häufig mit griechischen Buchstaben bezeichnet.

$$\bar{x} \longrightarrow \mu$$
$$s \longrightarrow \sigma$$

- Die Konstante $\frac{1}{\sigma\sqrt{2\pi}}$ sorgt dafür, dass die Fläche unter der Verteilung gleich 1 ist.

Normalverteilung: die Parameter μ und σ

- Die Parameter μ und σ sind die **Formparameter** (engl. *shape parameter*) der Normalverteilung — mit diesen zwei Parametern ist die Verteilung eindeutig definiert.
- Als Lageparameter verschiebt μ die Verteilung auf der x-Achse:

- Als Streuungsparameter verändert σ die Breite der Verteilung:

Was ist "normal" an der Normalverteilung bzw. warum ist die Normalverteilung so häufig?

- Auch wenn die Historie des Terms **Normalverteilung** umstritten ist[10] bringt er zum Ausdruck, dass es sich um eine empirisch sehr häufig beobachtete Verteilung handelt.
- Die Erklärung für die Häufigkeit der Normalverteilung liefert der **zentrale Grenzwertsatz**:

Definition

Zentraler Grenzwertsatz: bei einer additiven Überlagerung vieler kleiner unabhängiger Zufallseffekte zu einer aggregierten Zufallsvariable Z, nähert sich für $n \to \infty$ die Verteilung von Z der Normalverteilung an. (Beweis[11] — kein Klausurstoff)

- Fast alle psychologischen Phänomene sind Ausdruck einer Überlagerung vieler kleiner Zufallseffekte:
  - Genetische Zufallseffekte
  - Entwicklungsbiologische Zufallseffekte (z.B. im Mutterleib)
  - Zufällige Umwelteinflüsse (Familie, sozialer Kontext, Klima)
- Auch Messfehler wären ein solcher Zufallseffekt — gleichzeitig sind Verteilungen in der Psychologie aufgrund der genannten Effekte *auch ohne Messfehler* häufig normalverteilt.
- Obwohl die Vielzahl der Zufallseffekte Vorhersagen in der Psychologie enorm erschwert, haben sie **aus statistischer Sicht auch einen Vorteil**: wir können häufig (nicht immer!) annehmen, dass psychologische Merkmale einer Normalverteilung in der Bevölkerung folgen.

Normalverteilungen in freier Wildbahn

Eine kurze Geschichte der Normalverteilung

- Die Normalverteilung wurde mehrmals, zum Teil unabhängig, und für verschiedene Zwecke hergeleitet.
- Die erste Herleitung der Normalverteilung stammt aus dem Jahr **1733** von **Abraham de Moivre**, der nach einer **Approximationsfunktion für die Binomialverteilung** $f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$ suchte, da ihm die Berechnung der Fakultäten (z.B. $n!$) bei großen Zahlen mühselig wurde.
  - Tatsächlich ist diese Herleitung bereits ein Spezialfall des zentralen Grenzwertsatzes, da die Summe von binären Zufallsvariablen behandelt wird (z.B. wie oft “Kopf” bei “Kopf oder Zahl”).
  - Link zu einer Herleitung[12]
- Der **zentrale Grenzwertsatz** in seiner allgemeinen Form (beliebige Verteilungen) wurde **1778** von **Pierre-Simon Laplace** hergeleitet.
  - Link zu einer Herleitung[13]
- Im Jahr **1808 gelang Robert Adrain** der Nachweis, dass die Normalverteilung eine **valide Beschreibung von zufälligen Messfehlern** ist.

Abraham de Moivre

Pierre-Simon Laplace

Robert Adrain

Eine kurze Geschichte der Normalverteilung

Carl Friedrich Gauß

- Erst im Jahr **1809 tritt Carl Friedrich Gauß** auf die Bildfläche und erbringt ebenfalls den Beweis für die Allgemeingültigkeit der Normalverteilung für Messfehler (daher wird die Normalverteilung auch zuweilen **Fehlergesetz** genannt).
- Ein wichtige Motivation für Gauß' Arbeit an Fehlerverteilungen waren astronomische Messungen, die häufig ungenau waren und viele Messwiederholungen erforderten. Eine zentrale Frage war: welcher Geseztmäßigkeit folgen diese Messfehler?
- Der Beweis von Gauß war von besonderer Eleganz und basierte auf drei Annahmen:
  1. Messfehler sind symmetrisch (Fehler in -X und +X Richtung sind gleich wahrscheinlich)
  2. Kleinere Fehler sind häufiger als größere Fehler
  3. Der Mittelwert ist der beste Schätzer für den Lageparameter der wahren theoretischen Verteilung, oder anders gesagt: der beste Schätzer für die unbekannten Fehler $x_i - \mu$ (μ ist nicht bekannt!) ist $x_i - \bar{x}$.

- Mit diesen wenigen Annahmen konnte Gauß zeigen, dass *die Funktion die mit größter Wahrscheinlichkeit Fehler dieser Art erzeugt*[14], die Normalverteilung ist. Herleitung[15]
- Eine **heute populäre und intuitive Herleitung** basiert auf einem Gedankenexperiment, in dem Würfe von Dartpfeilen auf ein Bull's Eye betrachtet werden, wobei Wurffehler in x- und y-Richtung als unabhängig angenommen werden. Sie geht auf **John Herschel (1850)** zurück. Instruktives Video zur Herleitung[16]

John Herschel

Charakteristiken von Verteilungen

Bildnachweis[17]

Vorschau: p-Wert

- In Ihren Übungen mit JASP wird Ihnen u.U. ein Wert bereits jetzt begegnen: der **p-Wert**
- Der p-Wert ist von zentraler Bedeutung in der psychologischen Forschungsliteratur (for the better or the worse).
- Der p-Wert ist ein **Signifikanz-Maß**: er gibt vereinfacht gesprochen einen Hinweis darauf, wie wahrscheinlich es ist, dass ein gefundener Effekt (z.B. Mittelwertsunterschied zwischen zwei Gruppen) auf bloßem Zufall basiert.
- **Je kleiner der p-Wert, desto höher die statistische Signifikanz**, desto sicherer können wir uns also sein, dass ein Effekt nicht nur auf einer zufälligen Schwankung von Messfehlern basiert.
- Als Konvention hat sich etabliert, dass bei p-Werten kleiner 0,05 Effekte als **statistisch signifikant** gewertet werden.
- Den p-Wert werden wir noch ausführlich in den Vorlesungen zur Inferenzstatistik behandeln.

https://xkcd.com/1478/

Fußnoten

1. https://apastyle.apa.org/style-grammar-guidelines/tables-figures/tables
2.

Weissgerber TL, Winham SJ, Heinzen EP, Milin-Lazovic JS, Garcia-Valencia O, Bukumiric Z, Savic MD, Garovic VD, Milic NM (2019) Reveal, Don’t Conceal: Transforming Data Visualization to Improve Transparency. Circulation 140:1506–1518.

3. https://www.biztory.com/blog/bar-charts-the-good-the-bad-and-the-ugly
4. https://blog.revolutionanalytics.com/2017/05/the-datasaurus-dozen.html
5. https://jasp-stats.org/2021/10/05/raincloud-plots-innovative-data-visualizations-in-jasp/
6. https://jasp-stats.org/2021/10/05/raincloud-plots-innovative-data-visualizations-in-jasp/
7. https://support.goldensoftware.com/hc/en-us/articles/360019762133-Using-Break-Axis-in-Grapher
8. https://twitter.com/jonatanpallesen/status/1694966308465439117
9. https://twitter.com/jonatanpallesen/status/1694966308465439117
10. https://stats.stackexchange.com/questions/430621/why-is-the-normal-distribution-called-normal
11. https://alanhdu.github.io/posts/2019-10-21-normal-distribution-derivation/
12. http://www.stat.yale.edu/~pollard/Courses/241.fall2014/notes2014/Bin.Normal.pdf
13. https://towardsdatascience.com/central-limit-theorem-proofs-actually-working-through-the-math-a994cd582b33
14. In diesem Kontext erfand Gauß direkt auch das Prinzip der Maximum-Likelihood-Schätzung
15. https://notarocketscientist.xyz/posts/2023-01-27-how-gauss-derived-the-normal-distribution/
16. https://www.youtube.com/watch?v=cy8r7WSuT1I
17. https://matheguru.com/stochastik/schiefe-linksschief-rechtsschief-symmetrisch.html