Mathematische und statistische Methoden I

Sprechstunde
jederzeit nach
Vereinbarung und
nach der Vorlesung

Wallstr. 3, 6. Stock,
Raum 06-206

Dr. Malte Persike


persike@uni-mainz.de


lordsofthebortz.de


twitter.com/methodenlehre


tinyurl.com/gplusmethodenlehre

WiSe 2011/2012

Fachbereich Sozialwissenschaften
Psychologisches Institut
Johannes Gutenberg Universität Mainz

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Bortz, S. 41 – 43

Intervallskala

Kreuztabellen

Grafische Darstellung I

Intervalldaten

Grafische Beschreibung: Histogramm

- Das Histogramm stellt die Häufigkeiten vieler Kategorien in einem Säulendiagramm mit **weniger Klassen als Kategorien** dar
- Die Klassen müssen nicht notwendig gleich breit sein
- Für die Klassenbildung beim Histogramm gelten dieselben Faustregeln wie bei den Kreuztabellen
- Die Häufigkeiten können entweder absolute Häufigkeiten (absolutes Histogramm) sein oder relative Häufigkeiten (relatives Histogramm)
- Bei gleichen Klassenbreiten zeigt zumeist die **Höhe einer Säule** die Häufigkeit der Elemente in der Klasse. (wie beim Säulen-/Balkendiagramm)

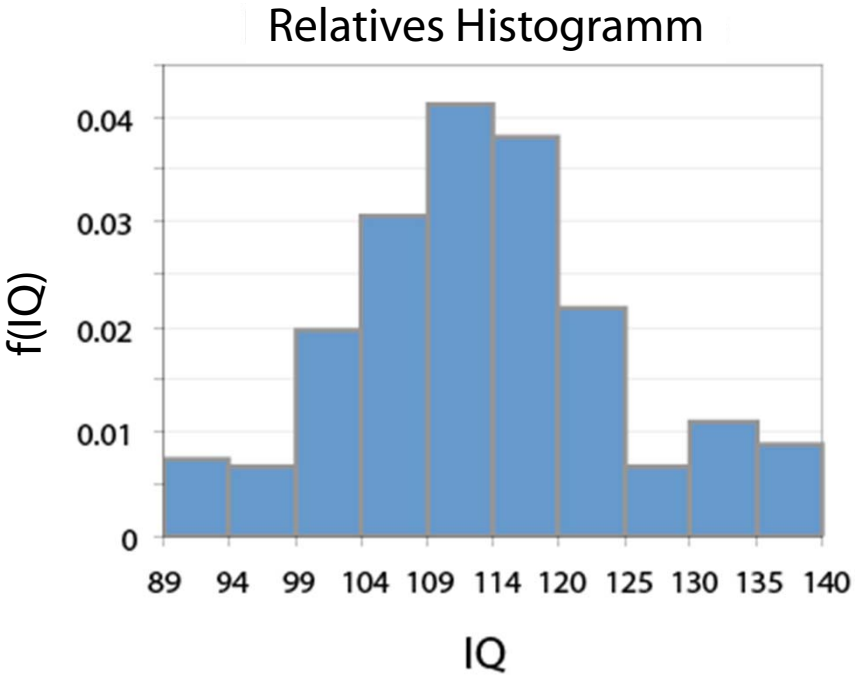
Intervalldaten

Grafische Beschreibung: Histogramm

Beispiel: Verteilung des IQ in diesem Raum.

Student	IQ
1	103
2	110
3	117
4	118
5	125
6	115
7	117
...	...
92	97

→ 92 Werte zwischen 89 und 140



Intervalldaten

Grafische Beschreibung: Histogramm

- **Frage:** Warum darf die **Höhe der Säule** in einem Histogramm nur dann die Häufigkeit der Elemente in den Klassen repräsentieren, wenn diese gleich breit sind?
- **Beispiel:** Säule 1 ist etwas höher als Säule 3, allerdings ist die Klassenbreite unterschiedlich groß
- Aufgrund der **Flächenbewertung** der menschlichen Wahrnehmung scheint Klasse 3 wesentlich mehr Merkmalsträger zu umfassen als Klasse 1

Intervalldaten

Grafische Beschreibung: Histogramm

- **Regel:** Wählt man ungleiche Klassenbreiten, **muss** das Histogramm **normiert** werden (wegen der Flächenbeurteilung der menschlichen Wahrnehmung).
- Wenn nicht die Höhe, sondern die Fläche A_j einer Säule die Häufigkeit repräsentieren soll, gilt für eine Klasse x_j:

  $A = f(x_j)$, und damit $f(x_j) = a_j \cdot d_j$

  (a_j ist die Höhe der Säule, d_j die Klassenbreite)

  Somit ist die Höhe einer Säule $a_j = f(x_j) / d_j$
- Dies gilt auch für die Darstellung mit absoluten Häufigkeiten $h(x_j)$

  Dann ist die Höhe einer Säule $a_j = h(x_j) / d_j$

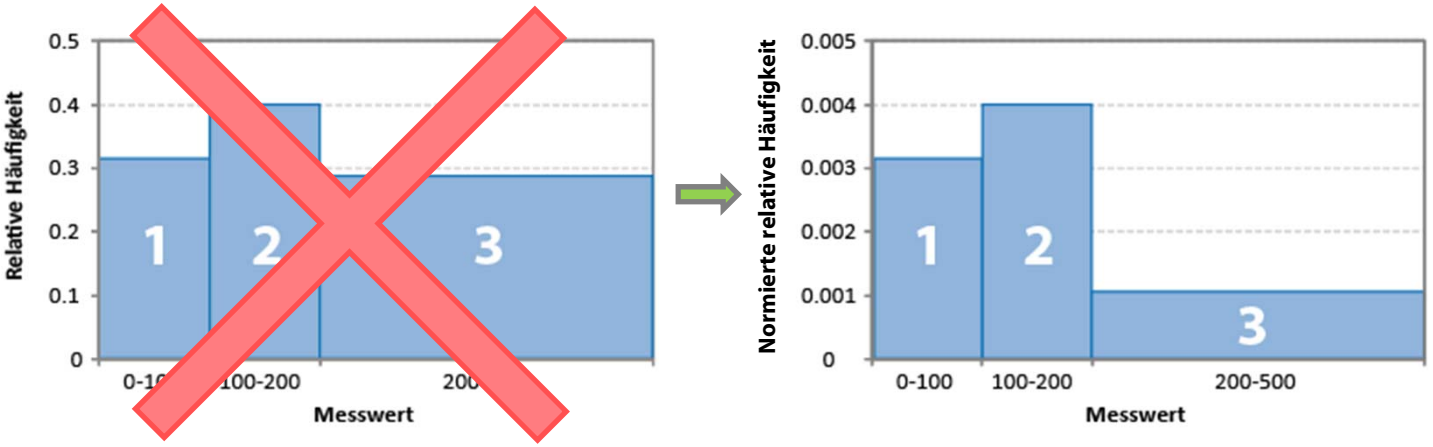
Intervalldaten

Grafische Beschreibung: Histogramm

- **Regel:** Wählt man ungleiche Klassenbreiten, **muss** das Histogramm **normiert** werden (wegen der Flächenbeurteilung der menschlichen Wahrnehmung).
- Wenn nicht die Höhe, sondern die Fläche A_j einer Säule die Häufigkeit repräsentieren soll, gilt für eine Klasse x_j:

$$A = f(x_j), \text{ und damit } f(x_j) = a_j \cdot d_j \text{ bzw. } a_j = f(x_j) / d_j$$

(a_j ist die Höhe der Säule, d_j die Klassenbreite)



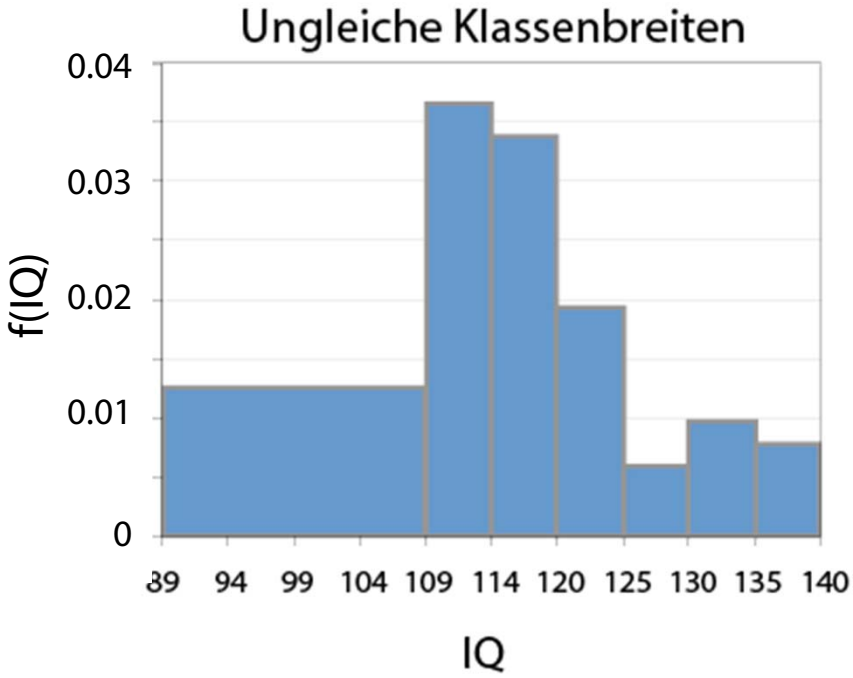
Intervalldaten

Grafische Beschreibung: Histogramm

Beispiel: Verteilung des IQ in diesem Raum.

Student	IQ
1	103
2	110
3	117
4	118
5	125
6	115
7	117
...	...
92	97

→ 92 Werte zwischen 89 und 140



Intervallskala

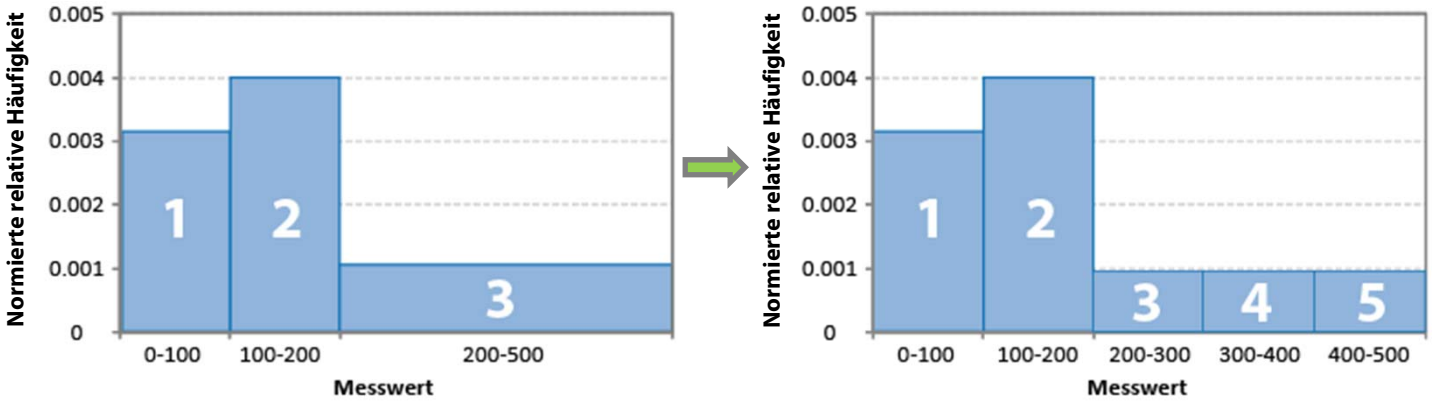
Kreuztabellen

Grafische Darstellung I

Intervalldaten

Grafische Beschreibung: Histogramm

- **Problem:** Ein normiertes Histogramm ist in Bezug auf die y-Achse nur **schwer interpretierbar**.
- Um die relative/absolute Häufigkeit einer Klasse zu bestimmen, muss – außer bei einer Klassenbreite von 1 – stets gerechnet werden
- Dies führt bei Histogrammen **mit gleicher Klassenbreite** zu unnötigem Interpretationsaufwand



Intervallskala

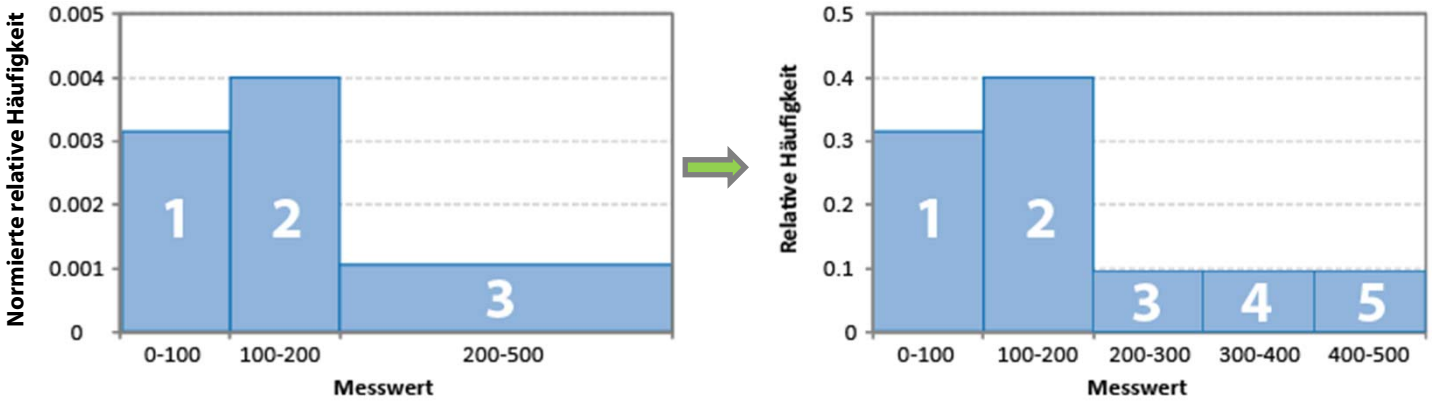
Kreuztabellen

Grafische Darstellung I

Intervalldaten

Grafische Beschreibung: Histogramm

- **Problem:** Ein normiertes Histogramm ist in Bezug auf die y-Achse nur **schwer interpretierbar**.
- Um die relative/absolute Häufigkeit einer Klasse zu bestimmen, muss – außer bei einer Klassenbreite von 1 – stets gerechnet werden
- Bei gleichen Klassenbreiten wird ein Histogramm daher zumeist **wie ein Säulendiagramm** skaliert.

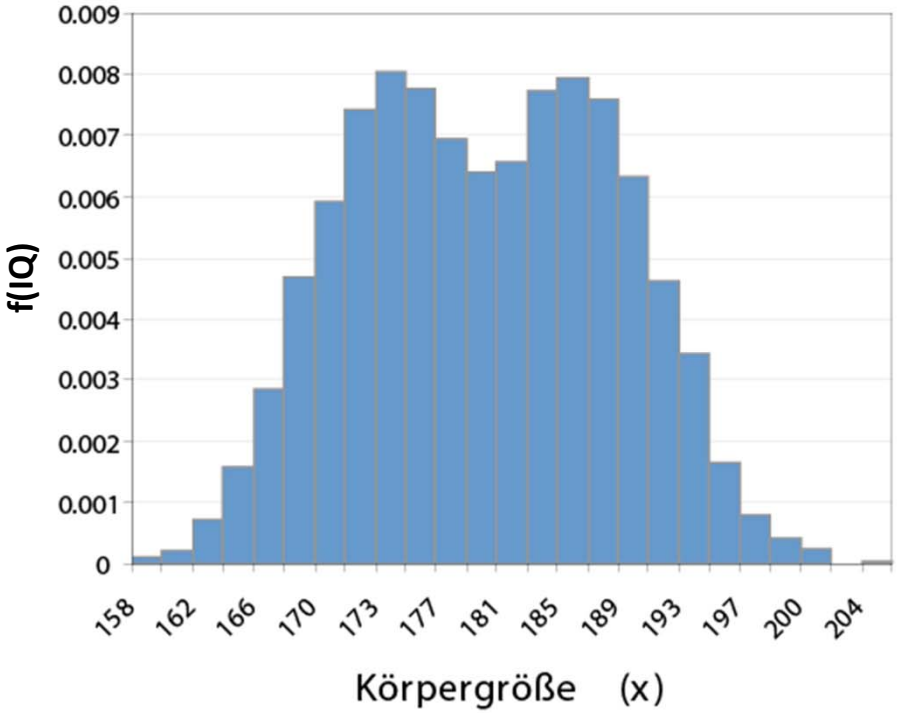


Intervalldaten

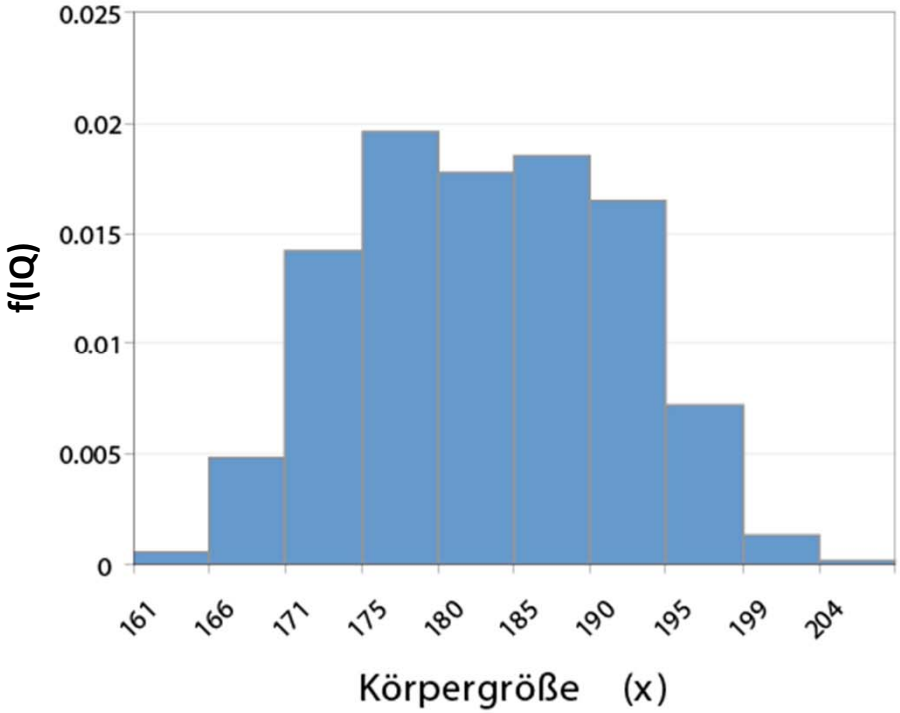
Grafische Beschreibung: Histogramm

Achtung: Die Wahl der Klassenanzahl kann für die Aussage entscheidend sein.
Beispiel: Körpergrößen an der Geisteswissenschaftlichen Fakultät der Uni Mainz

Klassenanzahl: 25



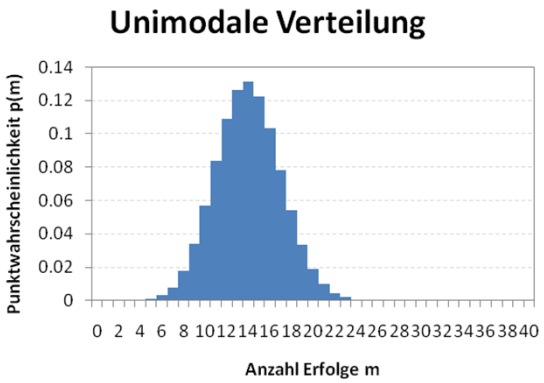
Klassenanzahl: 10

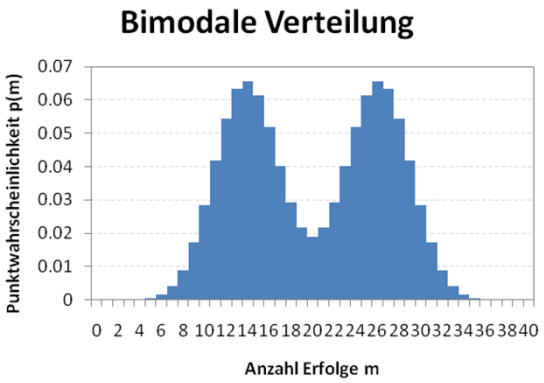


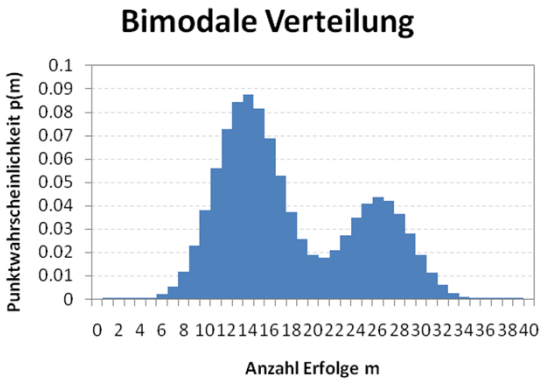
Intervalldaten

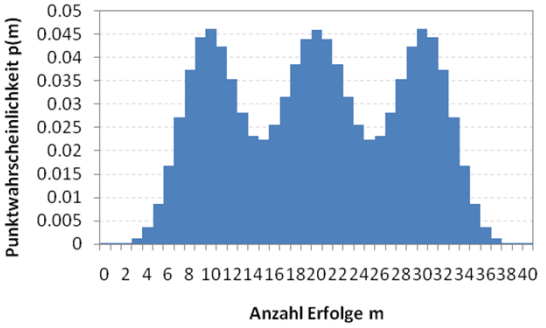
Grafische / verbale Beschreibung: Modalität

- Je nach Anzahl der (lokalen) Maxima unterscheidet man **uni-**, **bi-** und **multimodale** Verteilungen.





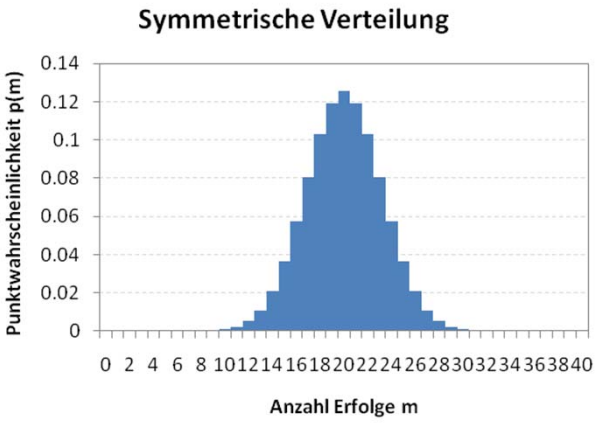


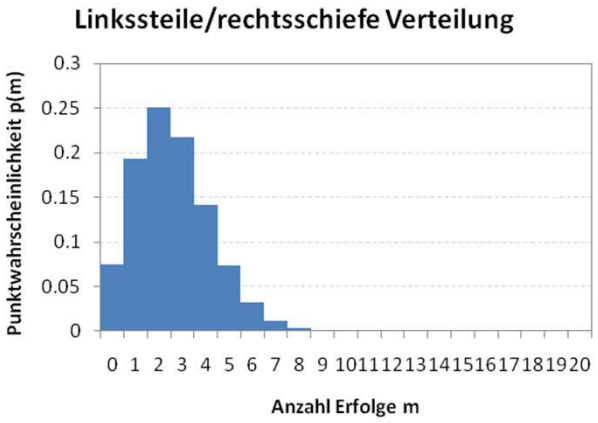


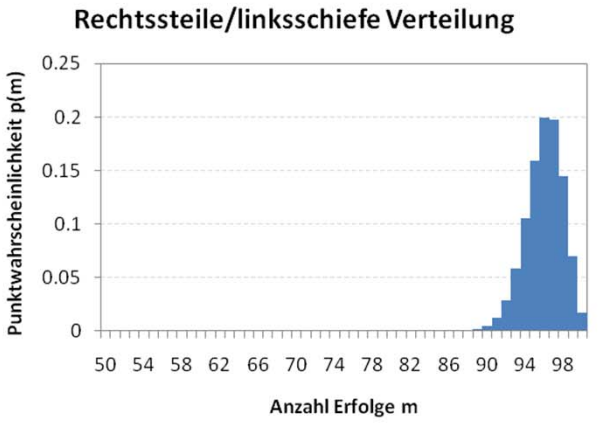
Intervalldaten

Grafische / verbale Beschreibung: Schiefe

- **Symmetrische Verteilungen:** Häufigkeiten für die Ausprägungen einer Zufallsvariablen verlaufen (annähernd) gleichartig um den Mittelwert.
- **Linkssteile / rechtsschiefe Verteilungen:** Häufigkeiten laufen rechts des Mittelwertes flacher aus.
- **Rechtssteile / linksschiefe Verteilungen:** Häufigkeiten laufen links des Mittelwertes flacher aus.







Bortz, S. 62

Intervalldaten

Grafische Beschreibung: Empirische Verteilungsfunktion

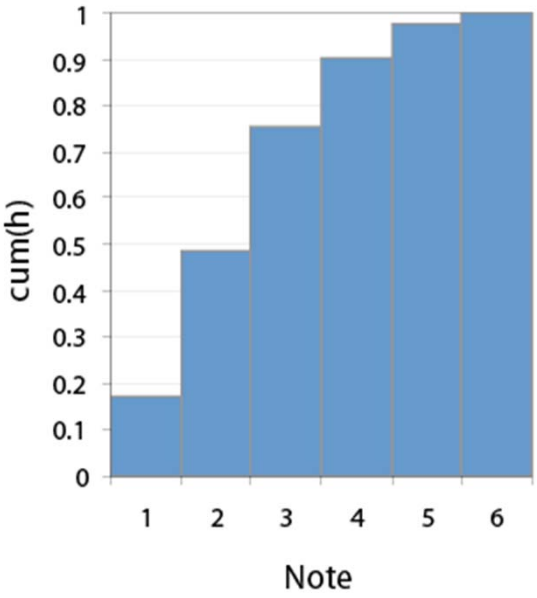
Die empirische Verteilungsfunktion bei c Klassen ist

$$F(X \leq x_j) = F(x_j) = \sum_{c=1}^{j} f(x_c)$$

mit $j = 1 \ldots k$

Zur grafischen Darstellung werden also die empirischen relativen Häufigkeiten aufsummiert

Note x	h(x)	f(x)
1	7	0.17
2	13	0.32
3	11	0.27
4	6	0.15
5	3	0.07
6	1	0.02



Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Kennwerte

- Maße der zentralen Tendenz
  - Mittelwert
- Streuungsmaße (Dispersionsmaße)
  - Mittlere Differenz
  - (Abweichungs-)Quadratsumme
  - Varianz
  - Standardabweichung

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Mittelwert

- Der **Mittelwert** ist bei n Beobachtungen $x_1 \ldots x_n$ definiert als

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots x_N) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Ist durch „extreme" Werte beeinflussbar (ausreißerempfindlich)

- Ist der Schwerpunkt der Beobachtungen, d.h.

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$$

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Mittelwert

- Der Mittelwert stimmt häufig mit keiner beobachteten Realisation überein
- Der Mittelwert ist äquivariant gegenüber gewissen (z.B. linearen) Transformationen
- Insbesondere
  1. Addition einer Konstanten a zu allen n Beobachtungen $x_1 \dots x_n$

$$\overline{x + a} = \overline{x} + a$$

  2. Multiplikation aller n Beobachtungen $x_1 \dots x_n$ mit einer Konstanten c

$$\overline{a \cdot x} = a \cdot \overline{x}$$

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Mittelwert

Lageregeln für die Maße der zentralen Tendenz

Bei symmetrischen Verteilungen:

$$\overline{x} \approx x_{med} \approx x_{mod}$$

Bei linkssteilen Verteilungen:

$$\overline{x} > x_{med} \geq x_{mod}$$

Bei rechtssteilen Verteilungen

$$\overline{x} < x_{med} \leq x_{mod}$$

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Mittlere Abweichung

Als **mittlere Abweichung** (MD) von n Beobachtungen $x_1 \ldots x_n$ in einem Datensatz wird die Summe aller **Abweichungsbeträge zum Median** bezeichnet.

$$MD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \tilde{x}\right|$$

Für jeden anderen Wert als für den Median ist der mittlere Abweichungsbetrag größer, d.h.

$$\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \tilde{x}\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|x_i - c\right|$$

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Abweichungsquadratsumme

- Die **Abweichungsquadratsumme** (oder auch: **Fehlerquadratsumme** oder einfach **Quadratsumme**) ist die Summe der quadrierten Abweichungen aller n Beobachtungen $x_1 \ldots x_n$ vom Mittelwert.

$$QS(x) = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Erfasst die Streuung um den Mittelwert
- Nur falls keine Streuung besteht, ist $QS = 0$, d.h. alle beobachteten Werte sind gleich. Sonst: $QS > 0$
- Je größer die Streuung, desto größer ist die QS
- **Problem**: Die Fehlerquadratsumme wird um so größer, je mehr Beobachtungen vorliegen

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Varianz

- Die **Varianz** ist das **mittlere** Abweichungsquadrat aller n Beobachtungen $x_1 \dots x_n$ vom Mittelwert.

$$s^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Erfasst die **mittlere** Streuung um den Mittelwert
- Nur falls keine Streuung besteht, ist $s^2 = 0$, d.h. alle beobachteten Werte sind gleich. Sonst: $s^2 > 0$
- Je größer die Streuung um den Mittelwert, desto größer ist die Varianz
- Ist anfällig gegenüber Ausreißern

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Varianz

- Für jeden anderen Wert als für den Mittelwert ist die Summe der Abweichungsquadrate höher

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(x_i-c\right)^2$$

- Der Mittelwert minimiert also die quadrierten Abweichungen aller Beobachtungen.

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Varianz

Die Formel für die Varianz lässt sich leicht umformen in eine rechnerisch manchmal günstigere Variante:

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\frac{1}{n}\sum_{i=1}^{n}x_i^{\,2}-\overline{x}^2=\overline{x^2}-\overline{x}^2$$

Die Varianz ist also die Differenz des Mittelwerts der quadrierten Daten und dem quadrierten Mittelwert der Daten.

Dies wird auch als **Momentenschreibweise der Varianz** bezeichnet.

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: Standardabweichung

- **Problem**: Die Varianz ist nicht äquivariant zu erlaubten Skalentransformationen

$$s^2(a \cdot x) = a^2 \cdot s^2(x) \quad \text{(mit a = const.)}$$

- Durch Wurzelziehen erhält man die **Standardabweichung** (SD, standard deviation)

$$s(x) = \sqrt{s^2(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Die Standardabweichung ist äquivariant zu den erlaubten Skalentransformationen

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Numerische Beschreibung: s^2 und s

Verhalten von Varianz und Standardabweichung bei Transformationen der n Beobachtungen $x_1 \dots x_n$

1. Die Addition einer Konstanten a zu allen Werten x verändert Varianz und Standardabweichung nicht

$$s^2(x + a) = s^2(x)$$

$$s(x + a) = s(x)$$

2. Die Multiplikation aller Werte x mit einer Konstanten a führt zu einer Erhöhung der Varianz um a^2 und der Standardabweichung um a

$$s^2(a \cdot x) = a^2 \cdot s^2(x)$$

$$s(a \cdot x) = a \cdot s(x)$$

Kreuztabellen

Grafische Darstellung I

Kennwerte

Grafische Darstellung II

Intervalldaten

Mittelwert und Varianz aus kategorisierten Daten

- Liegen intervallskalierte Daten bereits in kategorisierter Form vor (z.B. in einer Häufigkeitstabelle), so können daraus **Mittelwert und Varianz näherungsweise bestimmt** werden.
- Es sei $x_{j,mid} = \frac{OG_j + UG_j}{2}$ die Kategoriemitte der

j-ten von insgesamt k Kategorien mit der Untergrenze UG_j, der Obergrenze OG_j und der Häufigkeit $f(x_j)$

Mittelwert

$$\bar{x} = \sum_{j=1}^{k} f\left(x_j\right) \cdot x_{j,mid}$$

Varianz

$$s^2(x) = \sum_{j=1}^{k} f\left(x_j\right) \cdot \left(x_{j,mid} - \bar{x}\right)^2$$

Kreuztabellen

Grafische Darstellung I

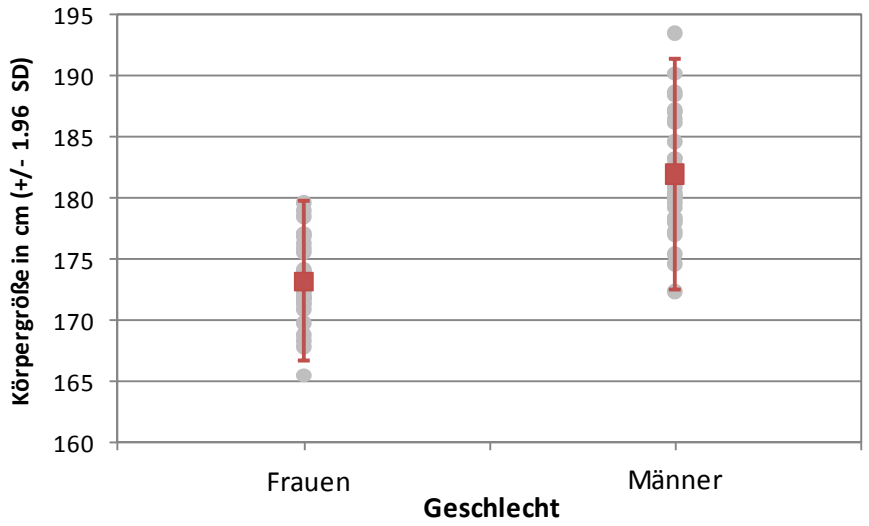
Kennwerte

Grafische Darstellung II

Intervalldaten

Grafische Beschreibung: Fehlerbalkendiagramm

- Das **Fehlerbalkendiagramm** (Error Bar) veranschaulicht Mittelwerte und die Streuung von Daten für mindestens eine Stichprobe.
- Für die Länge der Fehlerbalken existieren verschiedene Konventionen (± 1·SD, ± 1.96·SD, ± 2.58·SD)



Bortz, S. 35 – 37

z-Standardisierung

Transformationsregel

Ziel: Angabe der relativen Lage von Werten in einer Verteilung.

1. Quantile: wie bereits gesehen

2. Angabe einer **normierten Differenz** eines Messwertes zum Mittelwert

Berechnungsvorschrift: Jede Differenz eines Messwertes wird durch die Standardabweichung aller Messwerte geteilt. Die erhaltenen Werte werden als **z-Werte** bezeichnet.

$$z_x = \frac{x - \bar{x}}{s_x}$$

z-Standardisierung

Eigenschaften

- Der z-Wert kann auch als Differenz eines normierten Datenwertes vom normierten Mittelwert betrachtet werden, denn

$$z_x = \frac{x - \bar{x}}{s_x} = \frac{x}{s_x} - \frac{\bar{x}}{s_x}$$

- Der **Mittelwert** von z-Werten ist immer 0
- Die **Standardabweichung** von z-Werten ist immer 1

z-Standardisierung

Skalentransformation

- Mithilfe der z-Transformation können Messdaten mit beliebigem Mittelwert und Standardabweichung in Daten transformiert werden, die einen definierten Mittelwert und Standardabweichung aufweisen.

- **Schritt 1**: z-Standardisierung jedes Datenpunktes
- **Schritt 2**: Transformation jedes Datenpunktes in die neue Skala

$$x_{neu} = \left(z \cdot s_{neu} \right) + \overline{x}_{neu}$$

- Beispiele: Hamburg-Wechsler IQ-Test (MW=100, s=15), IQ-Skala laut IST (MW=100, s=10), Stanine-Skala (MW=5, s=2),

Relevante Excel Funktionen

- Kennwerte
  - ABS()
  - ^-Operator für Quadrierung, POTENZ()
  - WURZEL()
  - MITTELWERT(), MITTELWERTWENN(), MITTELWERTWENNS()
  - MITTELABW()
  - QUADRATESUMME()
  - VAR.P()
  - STABW.N()
  - STANDARDISIERUNG()