Statistische Kennzahlen für die Lage

- Nach der passenden grafischen Darstellung der Werte eines Merkmals (algebraische) Charakterisierungen der Verteilung solcher Werte.
- Ziel ist es, die Verteilung durch möglichst wenige Maßzahlen zu beschreiben.

1. Wo liegt die Mitte der Werte?

   Repräsentative Charakterisierung einer Verteilung durch eine Zahl: Lagemaß

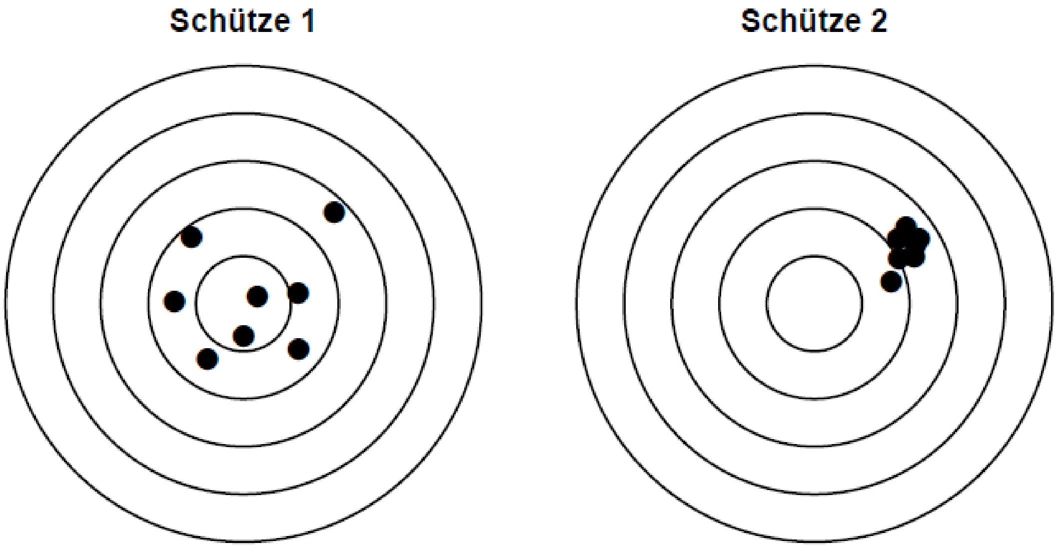
2. Wie streuen die Werte um die Mitte?

   Charakterisierung der Größe der Unsicherheit (= Streuung) der Merkmalswerte: Streuungsmaß

- Später: Vergleich verschiedener Gesamtheiten miteinander mit Hilfe der Maßzahlen.

Statistische Kennzahlen für die Lage

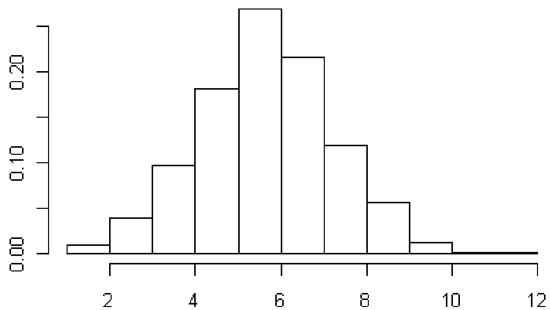
- Beispiel: Welcher Schütze schießt besser?

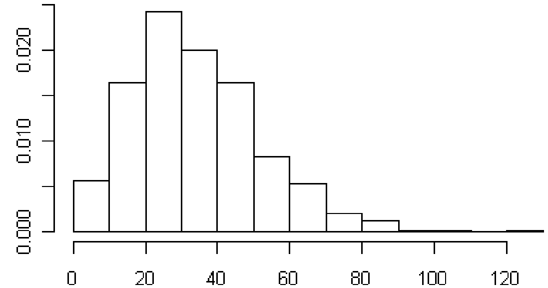


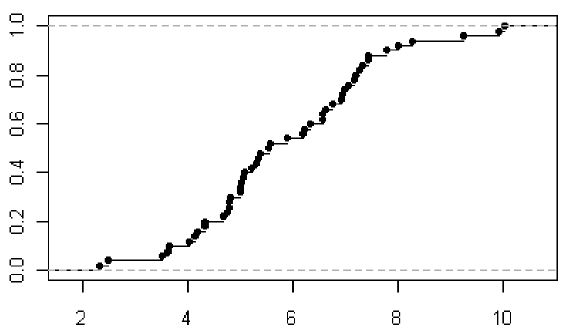
- Schütze 1: Lage gut, Streuung schlecht
- Schütze 2: Lage schlecht, Streuung gut

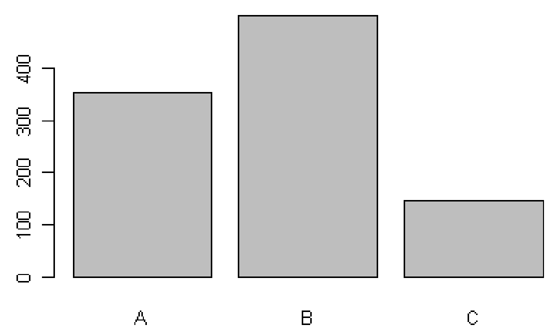
Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung









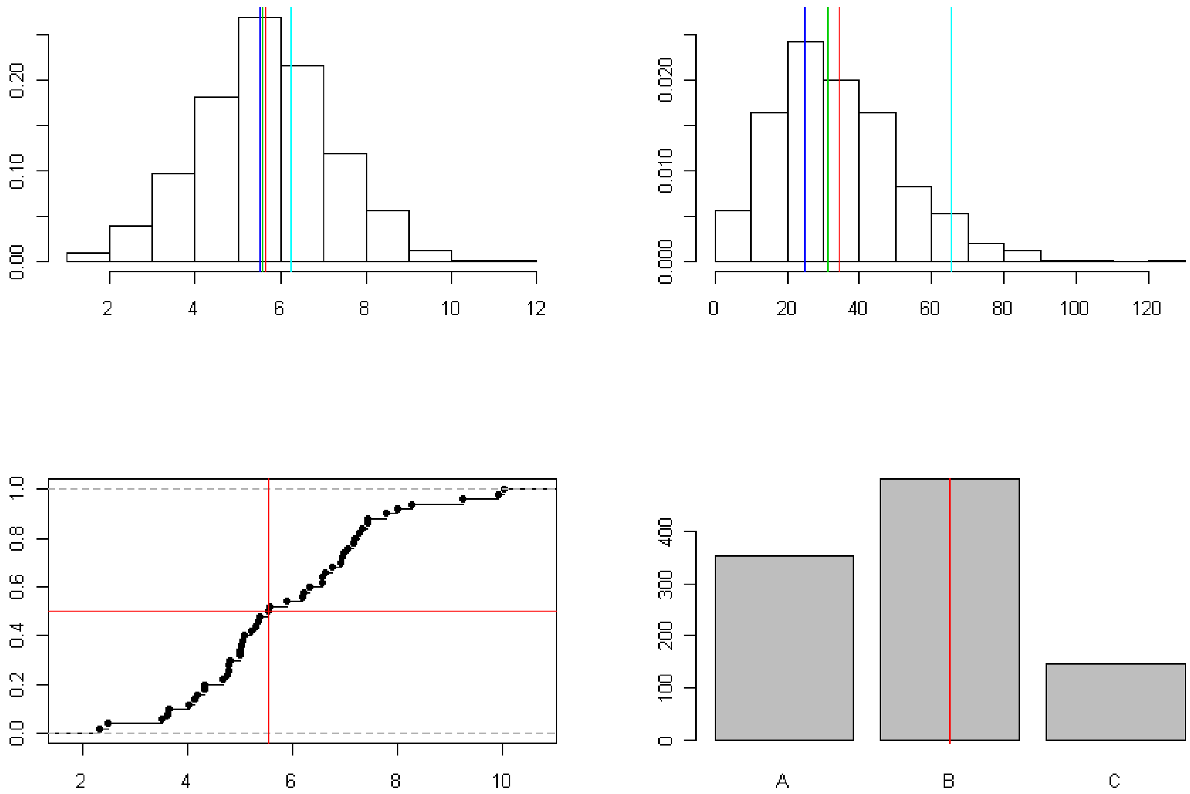
Beispiele

- Histogramm
- Empirische Verteilungs-funktion
- Stabdiagramm

Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung

Jetzt: stärkere Zusammenfassung der Daten auf ihr „Zentrum“



Farbige Linien repräsentieren das Zentrum

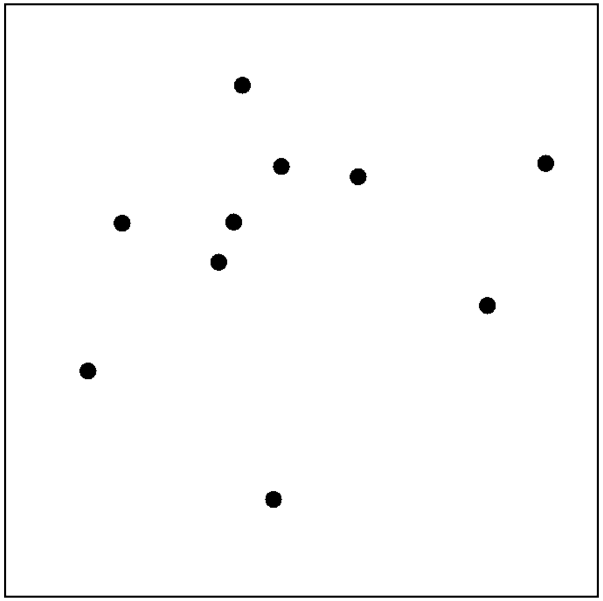
Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung

Jetzt: stärkere Zusammenfassung der Daten auf ihr „Zentrum“

Unterschiedliche Definitionen von „Zentrum“.

Allgemein: repräsentative Merkmalsausprägung, von der alle beobachteten Werte möglichst wenig abweichen



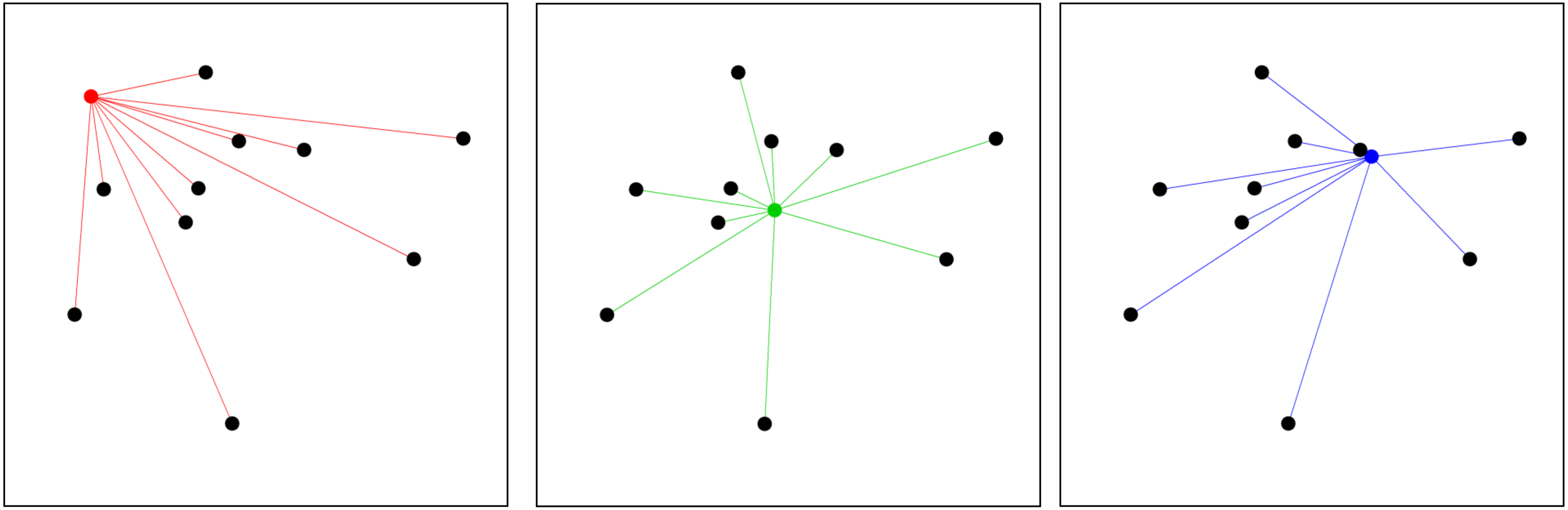
Statistische Kennzahlen für die Lage

Bisher: geringe Informationsverdichtung durch Verteilungsbeschreibung

Jetzt: stärkere Zusammenfassung der Daten auf ihr „Zentrum“

Unterschiedliche Definitionen von „Zentrum“.

Allgemein: repräsentative Merkmalsausprägung, von der alle beobachteten Werte möglichst wenig abweichen



Statistische Kennzahlen für die Lage

- Charakterisierung der Merkmalswerte auf einer Gesamtheit durch eine einzige Zahl: Lagemaße
- Lagemaß = “Mitte der Merkmalswerte”
- Auswahl des geeigneten Lagemaßes hängt vom Skalenniveau ab
- Wichtigste Beispiele:
  - Arithmetisches Mittel: Klassischer Mittelwert
    - Regiert am empfindlichsten auf „Ausreißer“, d.h. wenn für die Verteilung einige ungewöhnlich große oder kleine Werte vorliegen
  - Median: Zentralwert, mittlerer Wert in der geordneten Stichprobe
    - Liegt nicht unbedingt in der Mitte der Merkmalswerte, ist aber dennoch oft ein guter „Repräsentant“
    - Ist nicht unbedingt eindeutig
  - Modalwert: Häufigster Wert in der Stichprobe
    - Ist nicht unbedingt eindeutig
    - Bei stetigen Merkmalen meist erst nach Klassierung geeignet

Statistische Kennzahlen für die Lage

- Lagemaße

  - **Arithmetisches Mittel** = Mittelwert (mean)

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - **Median** = „Zentralwert“ = 50%-Wert: med_x
    Der Median ist derjenige Wert, für den 50% der Merkmalswerte größer oder gleich und 50% kleiner oder gleich sind.
    Der Median ist der mittlere Wert der Rangliste:

$$\mathrm{med}_x := \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ ungerade} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & n \text{ gerade} \end{cases}$$

  - **Modalwert / Modus** = häufigster Wert: mod_x
    Der Modalwert ist derjenige Merkmalswert, der am häufigsten vorkommt.

Statistische Kennzahlen für die Lage

- p-Quantil $Q_p = \tilde{x}_p$
  - Verallgemeinerung des Medians (50%-Wert) auf beliebige Prozentzahlen (100·p%-Werte)
  - Nützliches Mittel zur Beschreibung einer Rangliste $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

Ein **p-Quantil** Q_p, $p \in [0,1]$, *ist eine Zahl, für die* $100 \cdot p\%$ *der Merkmalswerte einer Gesamtheit kleiner oder gleich sind und* $100 \cdot (1-p)\%$ *größer oder gleich.*

Genauer könnte man für Q_p *z.B. Folgendes fordern:*

$Q_p \geq$ *größtem Merkmalswert einer Gesamtheit, der* $\leq 100 \cdot p\%$ *der Merkmalswerte ist und*
$Q_p \leq$ *nächstgrößerem Merkmalswert der Gesamtheit, also*

$$x_{(\lfloor np \rfloor)} \leq Q_p \leq x_{(\lfloor np \rfloor + 1)}.$$

Statistische Kennzahlen für die Lage

Die folgende Berechnungsmethode für Quantile entspricht der obigen Berechnung des Medians.

p-Quantil Berechnung: „Standard“ (Nicht in R, dort `type = 2` wählen.)

$$Q_p := \begin{cases} x_{(j)}, & j := \lceil np \rceil, \quad np \text{ nicht ganzzahlig} \\ \dfrac{x_{(j)} + x_{(j+1)}}{2}, & j := np, \quad np \text{ ganzzahlig} \end{cases}$$

Bezeichnung

- Anstelle von **p-Quantil** sagt man auch $100 \cdot p(\%)$-**Perzentil** oder **(1-p)-Fraktil**.
- 0.25- bzw. 0.75-Quantile heißen auch unteres bzw. oberes **Quartil**: unteres Quartil $q_4 = 0.25$-Quantil; oberes Quartil $q^4 = 0.75$-Quantil.

Statistische Kennzahlen für die Lage

- Nominale Daten
  - Gesucht: x^*, für das Abweichung zwischen x^* und $x_1,\ldots,x_N$ minimal ist
  - Mit nominellen Ausprägungen kann keine sinnvolle Abweichung berechnet werden
  - Dummykodierung führt auf den Modalwert $x(j^*)$

i	**x_i**
1	A
2	C
…	…
N	B

i	**x_i**	**$d_i(1)$**	**$d_i(2)$**	**$d_i(3)$**
1	A	1	0	0
2	C	0	0	1
…	…	…	…	…
N	B	0	1	0
Σ		N_1	N_2	N_3

Statistische Kennzahlen für die Lage

Nominale Daten

Modalwert

Beispiel **Bearbeitungen von Softwareaufgaben**

Bearbeiter(in)		
Ausprägung	**Absolute Häufigkeit**	**Relative Häufigkeit**
Kai	2	0.17
Miriam	3	0.25
Oliver	4	0.33
Tina	3	0.25
	12	1

Die Modalwerte lauten

$x_1(j^*) = \text{Oliver}$

$x_2(j^*) = \text{Export}$

$x_3(j^*) = 1.2$

Aufgabe		
Ausprägung	**Absolute Häufigkeit**	**Relative Häufigkeit**
Abfrage	2	0.17
Export	6	0.5
Verknüpfung	4	0.33
	12	1

Version		
Ausprägung	**Absolute Häufigkeit**	**Relative Häufigkeit**
1.1	3	0.25
1.2	6	0.5
2.0	3	0.25
	12	1

Statistische Kennzahlen für die Lage

Ordinale Daten

$x_1,\ldots,x_N$

$x_i \in W_X, i = 1,\ldots,N$

$W_X = \{x(j) \mid j = 1,\ldots,J\} = \{x(1),\ldots,x(J)\}$

$x(1) < x(2) < \ldots < x(J)$

i	x_i
1	x(3)
2	x(2)
3	x(1)
4	x(1)
5	x(3)

Geordnete Liste →

k	$X_{(k)}$
1	x(1)
2	x(1)
3	x(2)
4	x(3)
5	x(3)

Urliste $x_1,\ldots,x_N$

Geordnete Liste $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}$

$x_{(k)} = x_{i_k}$

mit

$$i_k = \min[\operatorname{argmin}_{i^*}(x_{i^*} \mid i^* \in \{1,\ldots,N\} \backslash \{i_1,\ldots,i_{k-1}\})],$$

$$k = 1,\ldots,N$$

$x_{(k)}$ wird k-ter **Rangwert** genannt, erster und letzter Rangwert $x_{(1)}$ und $x_{(N)}$ heißen **Minimum** und **Maximum**.

Statistische Kennzahlen für die Lage

Ordinale Daten

$x_1, \ldots, x_N$

$x_i \in W_X, i = 1, \ldots, N$

$W_X = \{x(j) \mid j = 1, \ldots, J\} = \{x(1), \ldots, x(J)\}$

$x(1) < x(2) < \ldots < x(J)$

$x_{(k)}$ wird k-ter **Rangwert** genannt, erster und letzter Rangwert $x_{(1)}$ und $x_{(N)}$ heißen **Minimum** und **Maximum**.

i	**x_i**	**$R(x_i)$**
1	x(3)	4.5
2	x(2)	3
3	x(1)	1.5
4	x(1)	1.5
5	x(3)	4.5

← Ränge

k	**$X_{(k)}$**
1	x(1)
2	x(1)
3	x(2)
4	x(3)
5	x(3)

$$R(x_i) = \frac{1}{\# K^*} \sum_{k^* \in K^*} k^* \quad \text{mit } K^* = \{k^* \mid x_{(k^*)} = x_i\}$$

$R(x_i)$ ist der **Rang** von x_i.

Gesucht: $x_{(k^*)}$, für das Abweichung zwischen $x_{(k^*)}$ und $x_1, \ldots, x_N$ minimal ist.

Statistische Kennzahlen für die Lage

Ordinale Daten

Beispiel **Bearbeitungen von Softwareaufgaben**

i	Version_i
1	1.1
2	1.2
3	1.1
4	1.2
5	2.0
6	1.2
7	1.2
8	1.2
9	1.2
10	1.1
11	2.0
12	2.0

Geordnete Liste →

k	$\text{Version}_{(k)}$
1	1.1
2	1.1
3	1.1
4	1.2
5	1.2
6	1.2
7	1.2
8	1.2
9	1.2
10	2.0
11	2.0
12	2.0

Ränge →

$$\frac{1}{3}\sum_{s=1}^{3} s = 2$$

$$\frac{1}{6}\sum_{s=4}^{9} s = 6.5$$

$$\frac{1}{3}\sum_{s=10}^{12} s = 11$$

i	Version_i	$\text{R}(\text{Version}_i)$
1	1.1	2
2	1.2	6.5
3	1.1	2
4	1.2	6.5
5	2.0	11
6	1.2	6.5
7	1.2	6.5
8	1.2	6.5
9	1.2	6.5
10	1.1	2
11	2.0	11
12	2.0	11

Statistische Kennzahlen für die Lage

Quantitative Daten

$x_1, \ldots, x_N$

$x_i \in W_X, i = 1, \ldots, N$

$W_X = \{x(j) \mid j = 1, \ldots, J\} = \{x(1), \ldots, x(J)\}$

bzw. $W_X = (-\infty, \infty)$

Der Median minimiert die Summe der absoluten Abweichungen.

$$\Delta_a(x) = \sum_{i=1}^{N} |x_i - x|$$

Der Mittelwert minimiert die Summe der quadratischen Abweichungen.

$$\Delta(x) = \sum_{i=1}^{N} (x_i - x)^2$$

Statistische Kennzahlen für die Lage

Quantitative Daten

Generell gilt: $\Delta(x) = \sum_{i=1}^{N}(x_i - x)^2$ ist minimal für $x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Beweis $\forall x \in \Re$:

$$\Delta(x) = \sum_{i=1}^{N}(x_i - x)^2 = \sum_{i=1}^{N}[(x_i - \bar{x}) + (\bar{x} - x)]^2$$

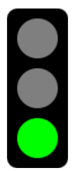
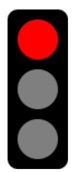
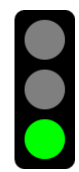
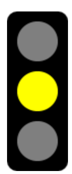
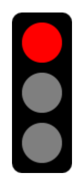
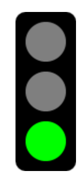
$$= \sum_{i=1}^{N}(x_i - \bar{x})^2 + 2(\bar{x} - x)\underbrace{\sum_{i=1}^{N}(x_i - \bar{x})}_{=0\ (*)} + \underbrace{\sum_{i=1}^{N}(\bar{x} - x)^2}_{=N(\bar{x}-x)^2}$$

$$= \Delta(\bar{x}) + \underbrace{N(\bar{x} - x)^2}_{\geq 0} \geq \Delta(\bar{x})$$

$$(*) \quad \sum_{i=1}^{N}(x_i - \bar{x}) = \sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{l=1}^{N} x_l\right) = \sum_{i=1}^{N} x_i - \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{N} x_l = \sum_{i=1}^{N} x_i - \frac{1}{N} \cdot N \sum_{l=1}^{N} x_l$$

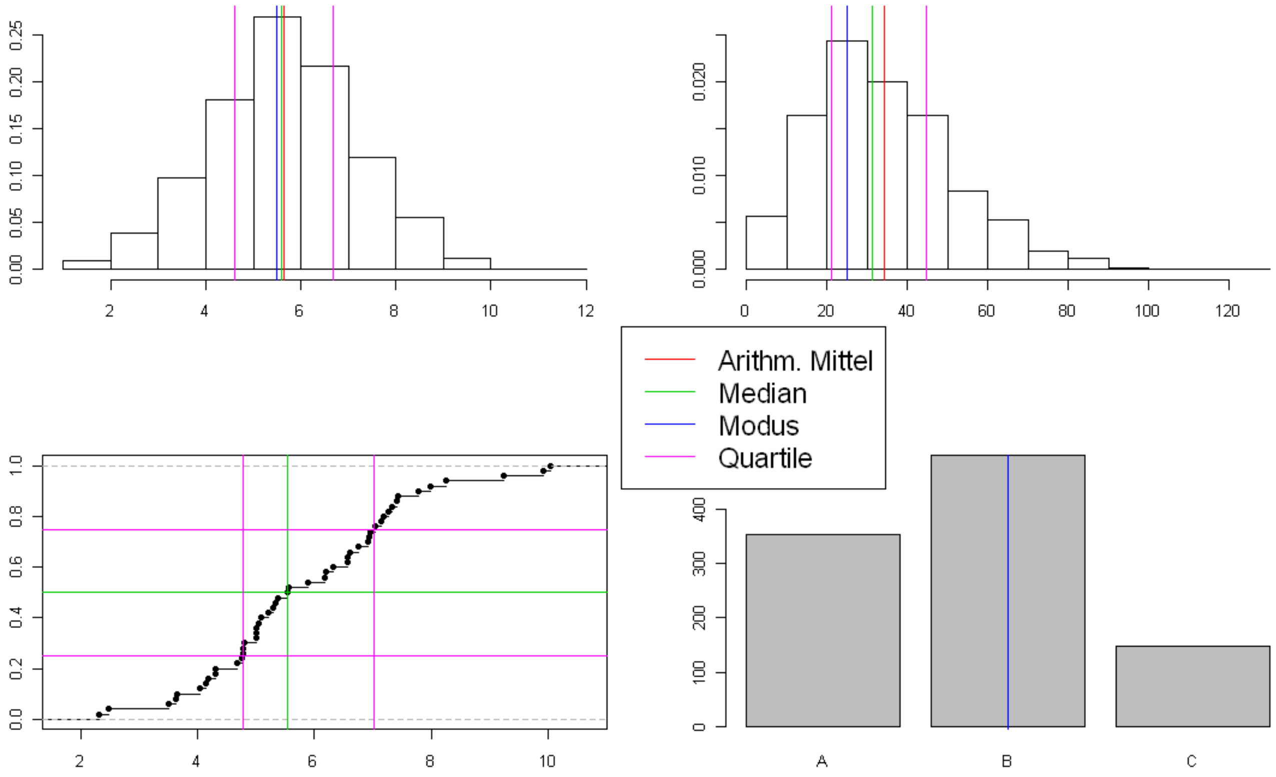
Statistische Kennzahlen für die Lage

Zusammenfassung: Welche Maßzahlen sind bei welchem Skalenniveau geeignet?

Skalenniveau → **↓ Lagemaß**	**Nominal**	**Ordinal**	**Quantitativ**
Modus		 – Informations-verlust	 – Nur für klassierte Daten
Median		 – Geringe Aussagekraft für kleine J	 + Robust – Informations-verlust – Hohe Streubreite
Arithmetisches Mittel	 – Nur für J = 2		 – Ausreißeranfällig + Informations-nutzung + Geringe Streubreite

Statistische Kennzahlen für die Streuung

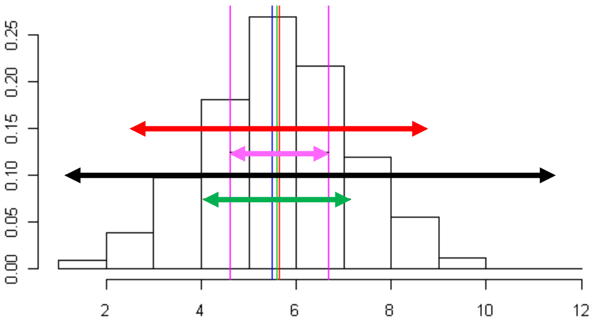
Bisher: Beschreibung von Häufigkeitsverteilung und Lage

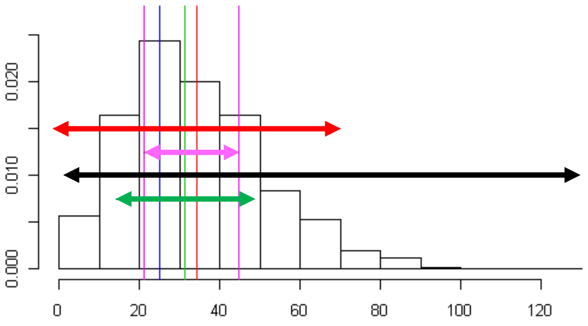


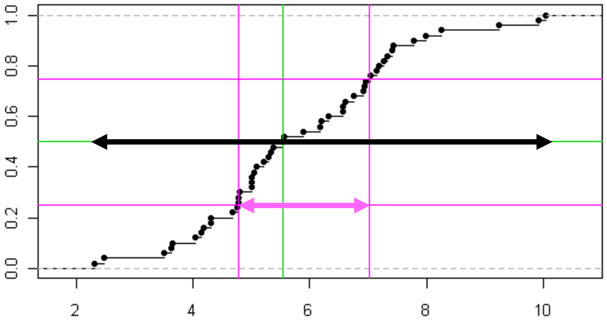
Statistische Kennzahlen für die Streuung

Bisher: Beschreibung von Häufigkeitsverteilung und Lage

Jetzt: Beschreibung der mittleren Variation um die Lage





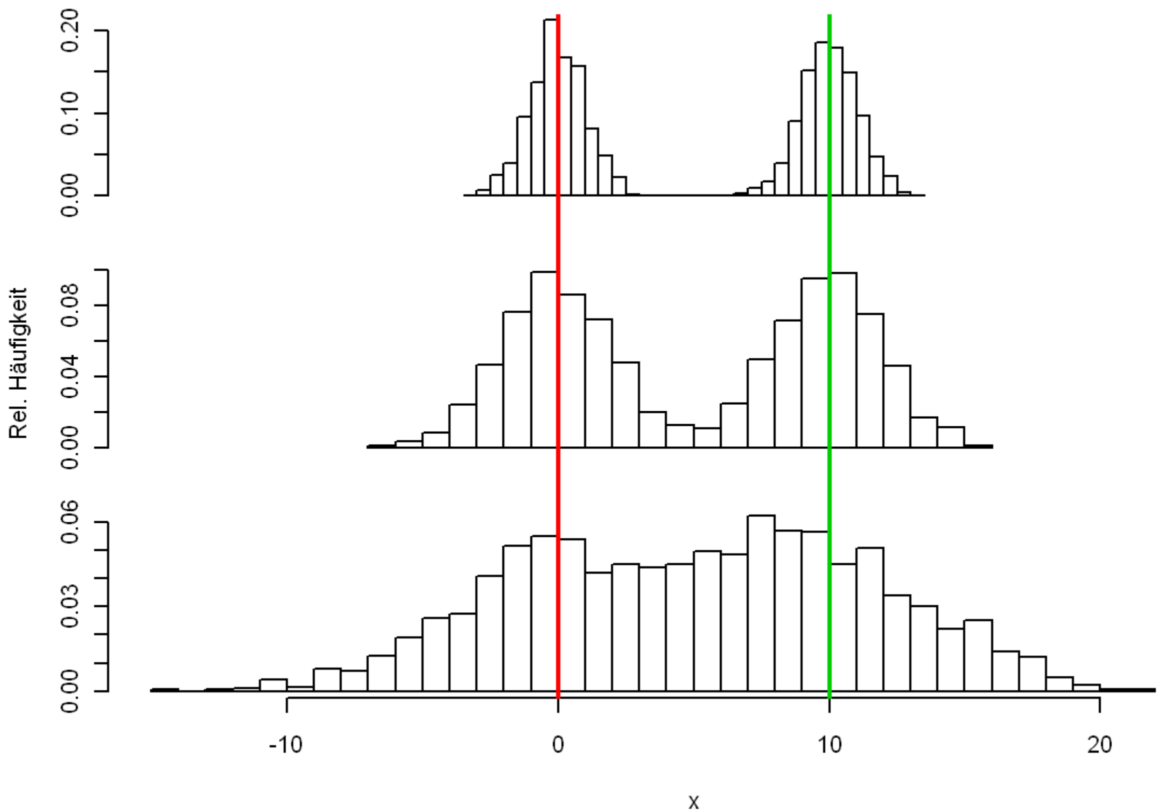


Allgemein: Streuung desto höher, je schlechter sich konkrete Werte vorhersagen lassen.

Statistische Kennzahlen für die Streuung

Bisher: Beschreibung von Häufigkeitsverteilung und Lage

Jetzt: Beschreibung der mittleren Variation um die Lage



Statistische Kennzahlen für die Streuung

- Streuungsmaße

  - empirische **Varianz**: „Durchschnitt" der quadrierten Abweichungen vom arithmetischen Mittel

$$\mathrm{var}_x = s_x^2 := \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)} = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{(n-1)}$$

  - **Standardabweichung**: Wurzel aus der **Varianz**

$$s_x := \sqrt{\mathrm{var}_x}$$

  - **Quartilsdifferenz** (interquartile range)

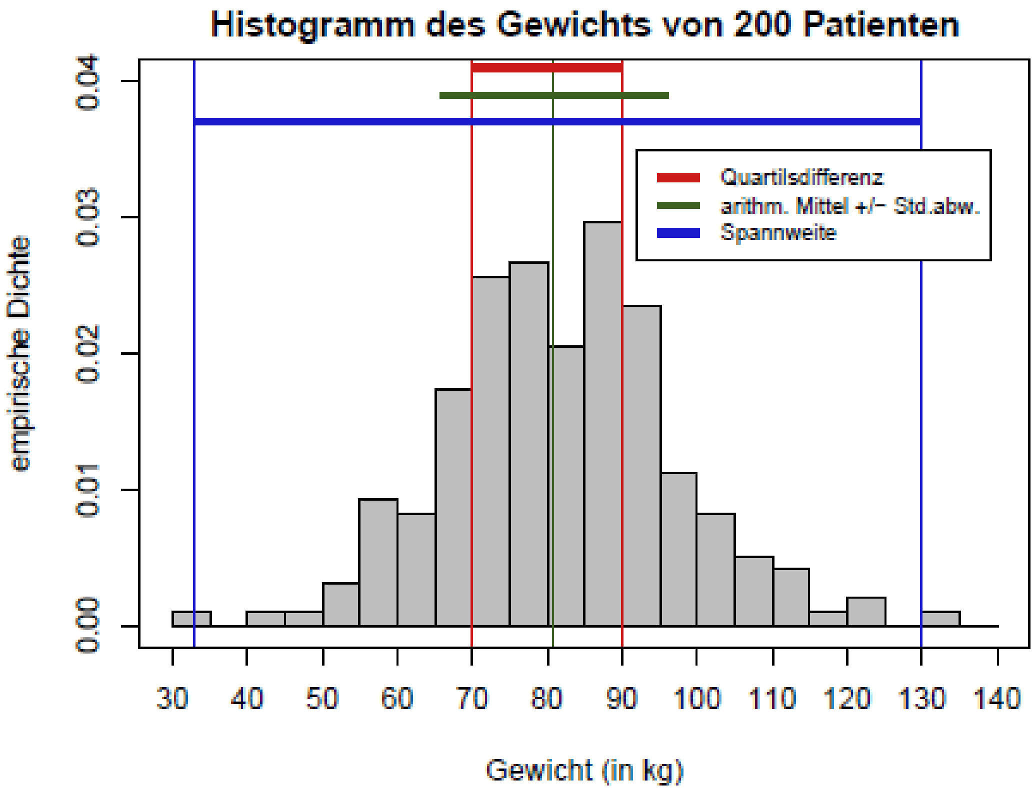
$$qd_x := q^4 - q_4$$

  - **Spannweite** (range)

$$R_x := \max(x) - \min(x) = x_{(n)} - x_{(1)}$$

Statistische Kennzahlen für die Streuung

- Beispiel 1, Gewicht von 200 Patienten
  $s_x = 15.14$ kg, $qd_x = 20$ kg, $R_x = 97$ kg



Statistische Kennzahlen für die Streuung

- Streuungsmaße

  - **Variationskoeffizient** (relative Standardabweichung)

$$v_x := \frac{s_x}{\bar{x}}$$

  - **Mittlere absolute Medianabweichung,**
    MD (von „Mean Deviation from the median“)

$$\mathrm{md}_x := \frac{1}{n} \sum_{i=1}^{n} |x_i - \mathrm{med}_x|$$

  - **Mediane absolute Medianabweichung,**
    MAD (von „Median Absolute Deviation“)

$$\mathrm{mad}_x := \mathrm{med}(|x_i - \mathrm{med}_x|)$$

Statistische Kennzahlen für die Streuung

Nominale Daten

$x_1, \ldots, x_N$

$x_i \in W_X, i = 1, \ldots, N$

$W_X = \{x(j) \mid j = 1, \ldots, J\} = \{x(1), \ldots, x(J)\}$

i	**x_i**
1	A
2	C
…	…
N	B

Rechnen nur sinnvoll mit Dummyvariablen bzw. Häufigkeiten

i	**x_i**	**$d_i(1)$**	**$d_i(2)$**	**$d_i(3)$**
1	A	1	0	0
2	C	0	0	1
…	…	…	…	…
N	B	0	1	0
Σ		N_1	N_2	N_3

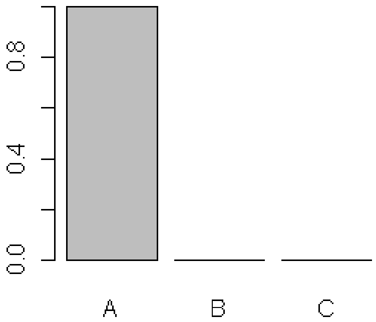
Statistische Kennzahlen für die Streuung

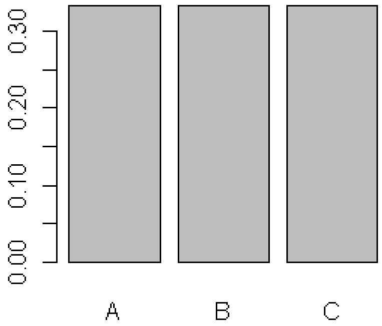
Nominale Daten

Allgemein: Streuung ist desto höher, je schlechter sich konkrete Werte vorhersagen lassen.

Nominale Merkmalsausprägungen lassen sich um so besser vorhersagen, je häufiger eine bestimmte Kategorie vorkommt.

Geringste Streuung , falls es ein j gibt mit f_j = 1. →





← Höchste Streuung , falls f_j = 1/J, j=1,…,J.

Statistische Kennzahlen für die Streuung

Nominale Daten

Geringste Streuung , falls es ein j gibt mit f_j = 1.

Höchste Streuung , falls f_j = 1/J, j=1,…,J.

Simpson‘s D

$$D = 1 - \sum_{j=1}^{J} f_j^2$$

D entspricht dem Anteil von Paaren mit unterschiedlichen Merkmalsausprägungen an allen aus der Urliste bildbaren Beobachtungspaaren:

$$D = \frac{\#\{(i,k) \in \{1,\ldots,N\} \times \{1,\ldots,N\} \mid x_i \neq x_k\}}{N^2}$$

Beispiel

i	$\mathbf{x_i}$
1	A
2	B
3	A
4	C

$$D = 1 - \left(\frac{\boxed{2^2} + \boxed{1^2} + \boxed{1^2}}{4^2} \right) = 1 - \frac{6}{16} = \frac{5}{8}$$

$$= \frac{\#\{(1,2),\ (1,4),\ (2,1),\ (2,3),\ (2,4),\ (3,2),\ (3,4),\ (4,1),\ (4,2),\ (4,3)\}}{4^2}$$

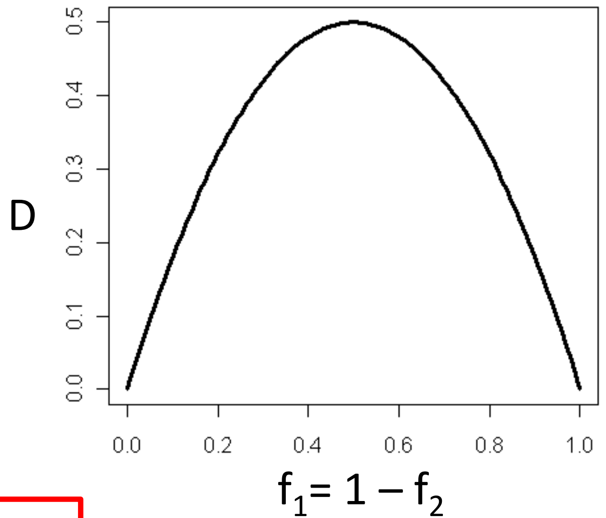
Statistische Kennzahlen für die Streuung

Nominale Daten

Geringste Streuung , falls es ein j gibt mit $f_j = 1$.

Höchste Streuung , falls $f_j = 1/J$, $j=1,\ldots,J$.

Beispiel J=2



Simpson's D

$$D = 1 - \sum_{j=1}^{J} f_j^2$$

$$\boxed{0} \leq D \leq \boxed{1 - \frac{1}{J}}$$

$$\boxed{D = 0 \quad \text{für} \quad \max[(f_1, \ldots, f_J)] = 1}$$

$$\boxed{D = 1 - \frac{1}{J} \quad \text{für} \quad f_1 = \ldots = f_J = \frac{1}{J}}$$

Statistische Kennzahlen für die Streuung

Nominale Daten

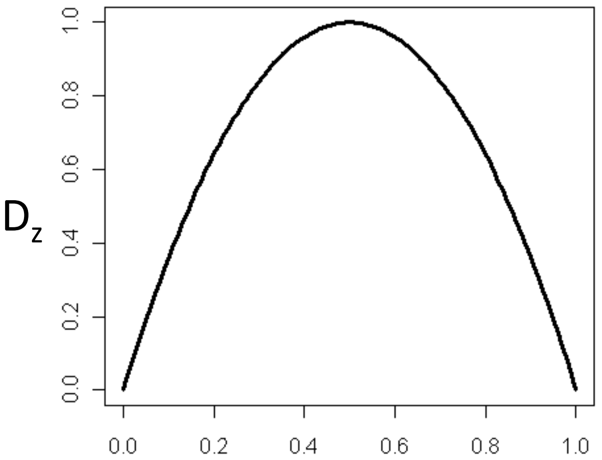
Geringste Streuung , falls es ein j gibt mit $f_j = 1$.

Höchste Streuung , falls $f_j = 1/J$, $j=1,\dots,J$.

Simpson's D_z (Normierte Version)

$$D_z = \frac{J(1-\sum_{j=1}^{J} f_j^2)}{J-1}$$

Beispiel J=2



$$\boxed{0} \le D_z \le \boxed{1}$$

$$\boxed{D_z = 0 \quad \text{für} \quad \max[(f_1,\dots,f_J)] = 1}$$

$$\boxed{D_z = 1 \quad \text{für} \quad f_1 = \dots = f_J = \frac{1}{J}}$$

Statistische Kennzahlen für die Streuung

Nominale Daten

Informationstheorie: Ein Ereignis liefert desto mehr Information, je geringer seine Eintrittswahrscheinlichkeit ist.

Kodierung der Elementarereignisse in Bits, Beispiel **Kaffeebestellung:**

1. Bit:	0 = Tasse	1 = Kännchen
2. Bit:	0 = Schwarz	1 = mit Milch
3. Bit:	0 = Süßstoff	1 = Zucker

8 Mögliche Bestellungen: 000, 001, 010, 011, 100, 101, 110, 111

Beträgt die Wahrscheinlichkeit einer Teilmenge dieser Bestellungen p = 1/8, wird genau eine Bestellung ausgewählt und man erhält Information über alle $3 = -\log_2(1/8)$ Bits, falls die Teilmenge ausgewählt wird.

Wird dagegen die Menge möglicher Bestellungen auf 50%, z.B. alle Bestellungen mit Kännchen eingegrenzt, also p = 1/4, so erhält man Information über $2 = -\log_2(1/4)$ Bits.

Statistische Kennzahlen für die Streuung

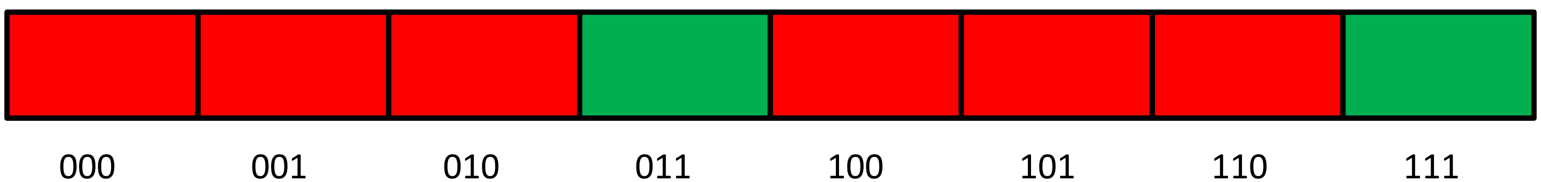
Nominale Daten

Die Information einer Merkmalsausprägung x(j) in Bits kann also allgemein definiert werden durch – $\log_2(f_j)$.

Der Informationsgehalt des gesamten Merkmals x ergibt sich durch die **Entropie** genannte erwartete Information H(x) von x:

$$H(x) = -\sum_{j=1}^{J} f_j \log_2 f_j$$

Beispiel **Kaffeebestellung**: Sei x_F die Antwort auf eine bestimmte Frage F



F = „Möchten Sie Ihren Kaffee mit Milch und Zucker?“

$x_F(0)$ = „Nein“, $x_F(1)$ = „Ja“, $f_1 = 6/8$, $f_2 = 2/8$,

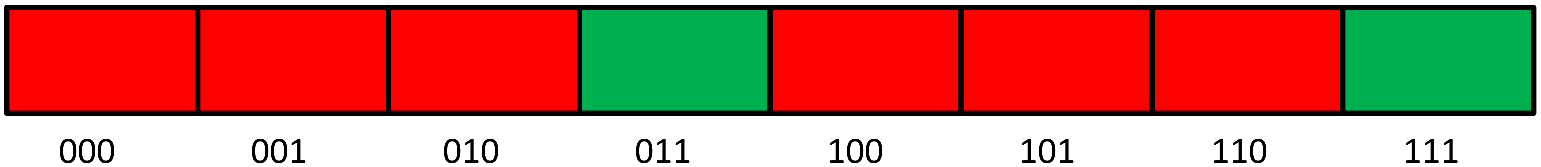
$$H(x_F) = -(6/8)\cdot\log_2(6/8) - (2/8)\cdot\log_2(2/8) = \boxed{0.8113}$$

Statistische Kennzahlen für die Streuung

Nominale Daten

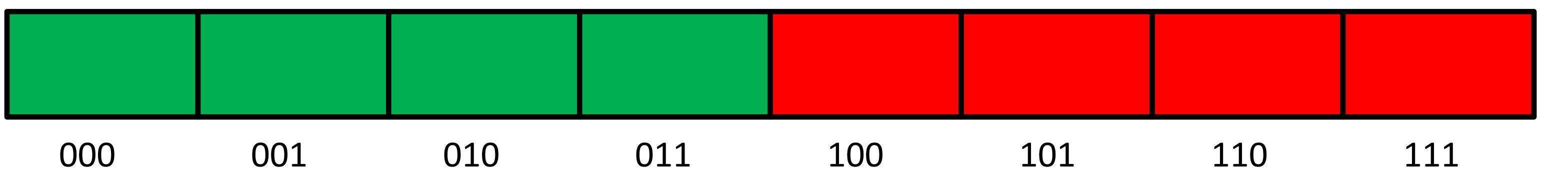
Entropie von x: $H(x) = -\sum_{j=1}^{J} f_j \log_2 f_j$

Beispiel **Kaffeebestellung**



F = „Möchten Sie Ihren Kaffee mit Milch und Zucker?“

$x_F(0)$ = „Nein“, $x_F(1)$ = „Ja“, $f_1 = 6/8$, $f_2 = 2/8$, $H(x_F) = \boxed{0.8113}$



F = „Möchten Sie Ihren Kaffee in der Tasse?“

$x_F(0)$ = „Nein“, $x_F(1)$ = „Ja“, $f_1 = 4/8$, $f_2 = 4/8$,

$H(x_F) = -(4/8)\cdot\log_2(4/8) - (4/8)\cdot\log_2(4/8) = \boxed{1}$

Statistische Kennzahlen für die Streuung

- Entropie gibt also die Information an, die man im Mittel durch Kenntnis der tatsächlichen Ausprägung erhält, wenn man vorher nur die Verteilung kannte. Ist diese hoch, konnte man den Wert vorher schlecht vorhersagen => hohe Streuung.
- Ist der Informationszugewinn gering, konnte man vorher schon gut prognostizieren.
- Beispiel „Wer wird Millionär“
  - Kandidat ist sicher = geringe Streuung, keine weitere Information durch Joker
  - Kandidat ist unsicher = hohe Streuung, erhofft Informationsgewinn durch Publikumsjoker
  - Ist hier die Streuung hoch, weiterer Informationsgewinn durch Einzelbefragungsjoker

Statistische Kennzahlen für die Streuung

Nominale Daten

Entropie von x: $$H(x) = -\sum_{j=1}^{J} f_j \log_2 f_j$$

Die Entropie ist ein sinnvolles Maß für die Streuung, denn sie erfüllt die Forderungen:

Geringste Streuung , falls es ein j gibt mit $f_j = 1$.

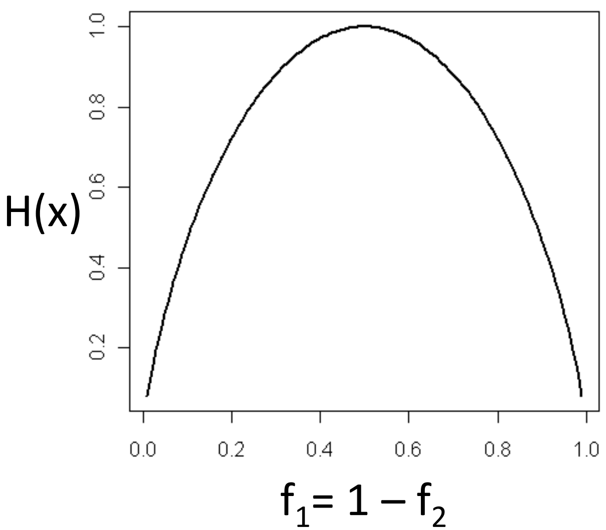
Höchste Streuung , falls $f_j = 1/J$, $j=1,\dots,J$.

$$\boxed{0} < H(x) \le \boxed{\log_2(J)}$$

$$\lim[H(x)] = 0 \quad \text{für} \quad \max[(f_1,\dots,f_J)] \to 1$$

$$H(x) = \log_2(J) \quad \text{für} \quad f_1 = \dots = f_j = \frac{1}{J}$$

Beispiel J=2



Statistische Kennzahlen für die Streuung

Nominale Daten

Normierte Entropie von x: $H_n(x) = -\sum_{j=1}^{J} f_j \frac{\log_2 f_j}{\log_2 J}$

Die Entropie ist ein sinnvolles Maß für die Streuung, denn sie erfüllt die Forderungen:

Geringste Streuung , falls es ein j gibt mit f_j = 1.

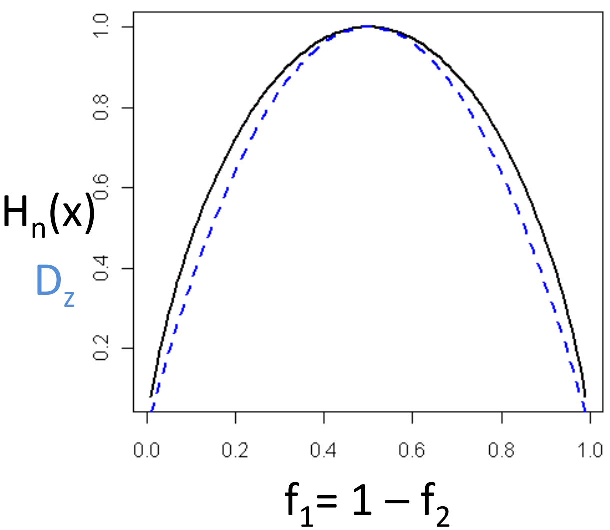
Höchste Streuung , falls f_j = 1/J, j=1,...,J.

$\boxed{0} < H_n(x) \leq \boxed{1}$

$\lim[H(x)] = 0$ für $\max[(f_1,...,f_J)] \to 1$

$H(x) = 1$ für $f_1 = ... = f_J = \frac{1}{J}$

Beispiel J=2



Statistische Kennzahlen für die Streuung

Nominale Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

Merkmal	D_z	$H_n(x)$
Bearbeiter(in)	$4/3\cdot(1-0.17^2-0.25^2-0.33^2-0.25^2)$ = **0.9815**	$[-0.17\cdot\log_2(0.17)-0.25\cdot\log_2(0.25)-0.33\cdot\log_2(0.33)-0.25\cdot\log_2(0.25)]\,/\log_2(4)$ = **0.9796**
Aufgabe	$3/2\cdot(1-0.17^2-0.5^2-0.33^2)$ = **0.9167**	$[-0.17\cdot\log_2(0.17)-0.5\cdot\log_2(0.5)-0.33\cdot\log_2(0.33)]/\log_2(3)$ = **0.9206**
Version	$3/2\cdot(1-0.25^2-0.5^2-0.25^2)$ = **0.9375**	$[-0.25\cdot\log_2(0.25)-0.5\cdot\log_2(0.5)-0.25\cdot\log_2(0.25)]/\log_2(3)$ = **0.9464**

Bearbeiter(in)		
Ausprägung	**Absolute Häufigkeit**	**Relative Häufigkeit**
Kai	2	0.17
Miriam	3	0.25
Oliver	4	0.33
Tina	3	0.25
	12	1

Aufgabe		
Ausprägung	**Absolute Häufigkeit**	**Relative Häufigkeit**
Abfrage	2	0.17
Export	6	0.5
Verknüpfung	4	0.33
	12	1

Version		
Ausprägung	**Absolute Häufigkeit**	**Relative Häufigkeit**
1.1	3	0.25
1.2	6	0.5
2.0	3	0.25
	12	1

Statistische Kennzahlen für die Streuung

Ordinale Daten

$x_1,\ldots,x_N$

$x_i \in W_X, i = 1,\ldots,N$

$W_X = \{x(j) \mid j = 1,\ldots,J\} = \{x(1),\ldots,x(J)\}$

$x(1) < x(2) < \ldots < x(J)$

i	**x_i**
1	x(3)
2	x(2)
3	x(1)
4	x(1)
5	x(3)

Geordnete Liste →

k	**$X_{(k)}$**
1	x(1)
2	x(1)
3	x(2)
4	x(3)
5	x(3)

Simpson‘s D und H(x) sind anwendbar, allerdings wird Information der Kategorienordnung nicht genutzt.

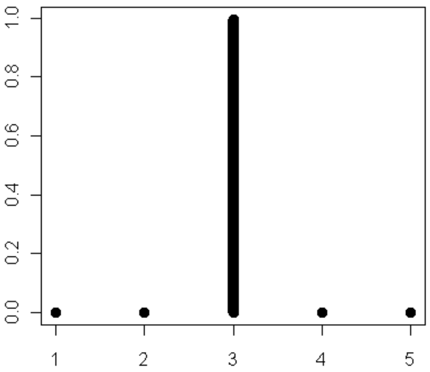
Statistische Kennzahlen für die Streuung

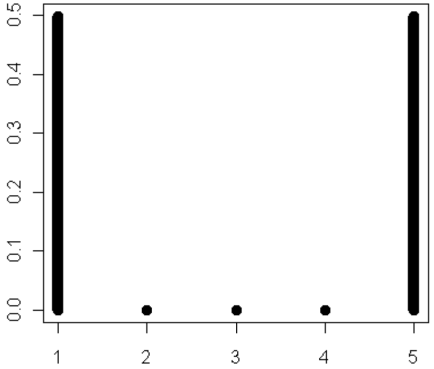
Ordinale Daten

Allgemein: Streuung desto höher, je schlechter konkrete Werte sich vorhersagen lassen.

Werte lassen sich umso besser vorhersagen, je stärker sie sich um den Median verdichten.

Geringste Streuung für $N(\tilde{x}_{0.5}) = N$ →





← Höchste Streuung für $N(\tilde{x}_0) = N(\tilde{x}_1) = N/2$

Nicht mehr höchste Streuung bei ausgeglichener Belegung, da die Kategorien unterschiedlich weit von der Mitte entfernt sind. Höchste Streuung bei maximaler Entfernung zur Mitte, also bei gleichmäßiger Konzentration an Minimum und Maximum.

Statistische Kennzahlen für die Streuung

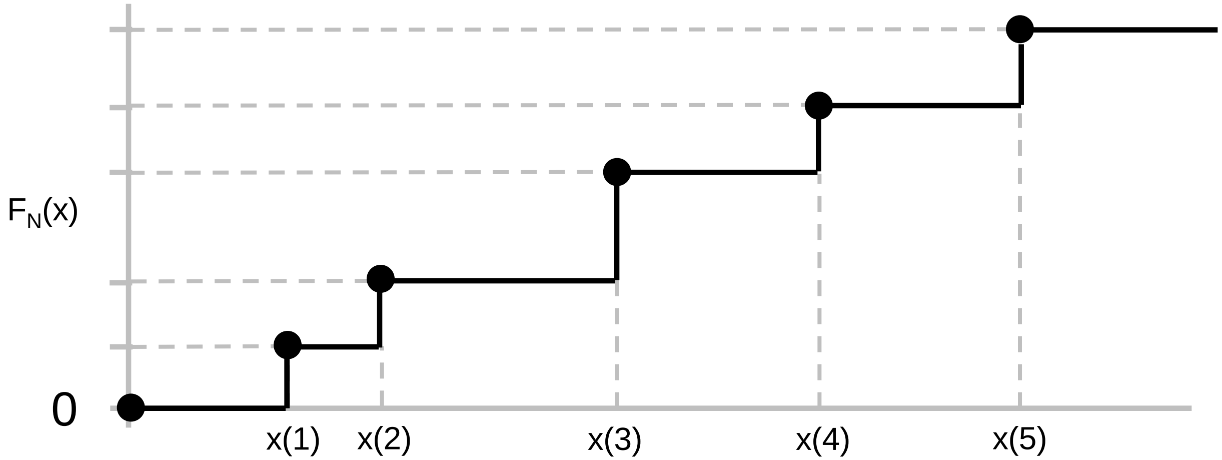
Ordinale Daten

Geringste Streuung für $N(\tilde{x}_{0.5}) = N$

Höchste Streuung für $N(\tilde{x}_0) = N(\tilde{x}_1) = N/2$

Dispersionsindex nach Leti

$$D_L = \sum_{j=1}^{J-1} F_N[x(j)] \cdot \left(1 - F_N[x(j)]\right)$$



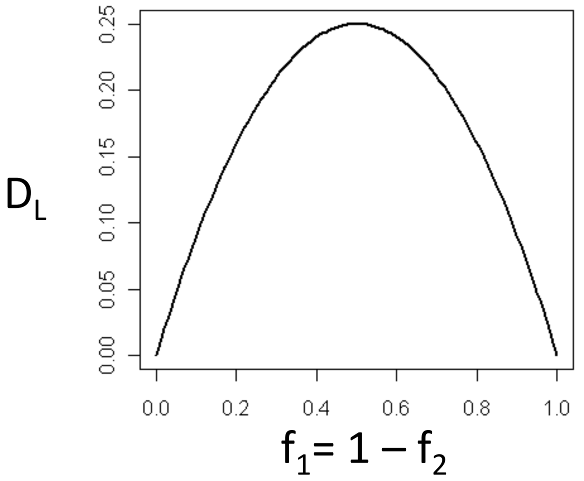
Statistische Kennzahlen für die Streuung

Ordinale Daten

Geringste Streuung für $N(\tilde{x}_{0.5}) = N$

Höchste Streuung für $N(\tilde{x}_0) = N(\tilde{x}_1) = N/2$

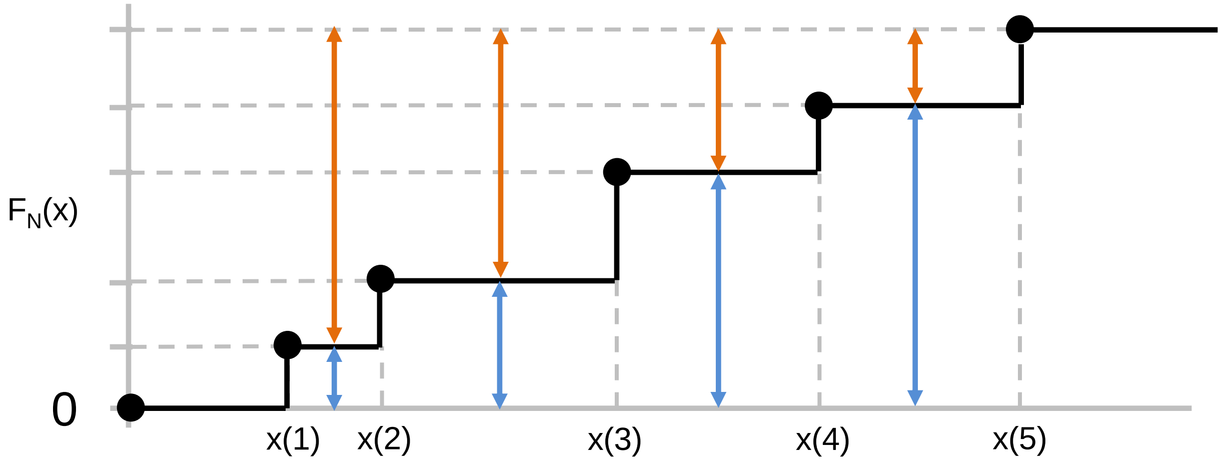
Beispiel J=2



Dispersionsindex nach Leti

$$D_L = \sum_{j=1}^{J-1} \boxed{F_N[x(j)]} \cdot \boxed{\left(1 - F_N[x(j)]\right)}$$

$$\boxed{0} \leq D_L \leq \boxed{\frac{J-1}{4}}$$



Statistische Kennzahlen für die Streuung

Ordinale Daten

Geringste Streuung für $N(\tilde{x}_{0.5}) = N$

Höchste Streuung für $N(\tilde{x}_0) = N(\tilde{x}_1) = N/2$

Normierter Dispersionsindex nach Leti

$$D_{Lz} = \frac{4}{J-1} \sum_{j=1}^{J-1} F_N[x(j)] \cdot \left(1 - F_N[x(j)]\right)$$

$$\boxed{0} \leq D_{Lz} \leq \boxed{1}$$

Für J=2 gilt $D_z = D_{Lz}$,
d.h. normierte Versionen von Simpson und Leti sind äquivalent

Beweis:

$$\boxed{D_{Lz}} = \frac{4}{2-1} \sum_{j=1}^{1} F_N[x(j)](1 - F_N[x(j)])$$

$$= 4 \cdot (f_1(1-f_1)) = 2\,(2f_1 - 2f_1^2)$$

$$= 2(1 - f_1^2 - 1 + 2f_1 - f_1^2) = 2(1 - [f_1^2 + (1-f_1)^2])$$

$$= \frac{2\left(1 - \sum_{j=1}^{2} f_j^2\right)}{2-1} = \boxed{D_z}$$

$\square$

Statistische Kennzahlen für die Streuung

Quantitative Daten

$x_1, \ldots, x_N$

$x_i \in W_X, i = 1, \ldots, N$

$W_X = \{x(j) \mid j = 1, \ldots, J\} = \{x(1), \ldots, x(J)\}$

bzw. $W_X = (-\infty, \infty)$

Allgemein: Streuung desto höher, je schlechter konkrete Werte sich vorhersagen lassen.

Werte lassen sich umso besser vorhersagen, je stärker sie sich um das jeweilige Lagemaß verdichten.

Statistische Kennzahlen für die Streuung

Quantitative Daten

Werte lassen sich umso besser vorhersagen, je stärker sie sich um das jeweilige Lagemaß verdichten.

Lagemaß: Arithmetisches Mittel

Streuungsmaß:

Varianz (mittlere quadratische Abweichung)

$$s_x^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2 \qquad \left(\text{bzw.}\, d_x^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2\right)$$

Standardabweichung

$$s = \sqrt{s_x^2} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2}$$

Statistische Kennzahlen für die Streuung

Quantitative Daten

Von Streuungsparametern abgeleitete Größen für verhältnisskalierte Merkmale

Quartilskoeffizient

$$Q_{koeff} = \frac{2Q}{\tilde{x}_{0.25} + \tilde{x}_{0.75}} = \frac{2(\tilde{x}_{0.75} - \tilde{x}_{0.25})}{\tilde{x}_{0.25} + \tilde{x}_{0.75}} = (\tilde{x}_{0.75} - \tilde{x}_{0.25}) \Big/ \left(\frac{\tilde{x}_{0.25} + \tilde{x}_{0.75}}{2} \right)$$

Variationskoeffizient

$$V_x = \frac{s_x}{\bar{x}}$$

Statistische Kennzahlen für die Streuung

Quantitative Daten: **Berechnung der Varianz aus Häufigkeitsverteilung**

$$s_x^2 = \frac{N}{N\text{-}1}\sum_{j=1}^{J} f_j \cdot [x(j)\text{-}\sum_{k=1}^{K} f_k \cdot x(k)]^2 = \frac{N}{N\text{-}1}\sum_{j=1}^{J} f_j \cdot [x(j)\text{-}\overline{x}]^2$$

Beweis: $$\frac{1}{N\text{-}1}\sum_{i=1}^{N}(x_i-\overline{x})^2 = \frac{1}{N\text{-}1}\sum_{i=1}^{N}(x_{(i)}-\overline{x})^2 = \frac{(x_{(1)}-\overline{x})^2}{N-1} + \frac{(x_{(2)}-\overline{x})^2}{N-1} + \ldots + \frac{(x_{(N)}-\overline{x})^2}{N-1}$$

$$= \underbrace{\frac{(x(1)-\overline{x})^2}{N-1}+\ldots+\frac{(x(1)-\overline{x})^2}{N-1}}_{f_1 \cdot N \text{ mal}} + \underbrace{\frac{(x(2)-\overline{x})^2}{N-1}+\ldots+\frac{(x(2)-\overline{x})^2}{N-1}}_{f_2 \cdot N \text{ mal}} + \ldots + \underbrace{\frac{(x(J)-\overline{x})^2}{N-1}+\ldots+\frac{(x(J)-\overline{x})^2}{N-1}}_{f_J \cdot N \text{ mal}}$$

$$= \frac{N}{N-1} f_1 \cdot (x(1)-\overline{x})^2 + \frac{N}{N-1} f_2 \cdot (x(2)-\overline{x})^2 + \ldots + \frac{N}{N-1} f_J \cdot (x(J)-\overline{x})^2 = \frac{N}{N-1}\sum_{j=1}^{J} f_j \cdot (x(j)-\overline{x})^2 \qquad \square$$

Statistische Kennzahlen für die Streuung

Quantitative Daten: **Varianz von Lineartransformationen**

$$y = ax + b \Rightarrow s_y^2 = a^2 s_x^2$$

Beweis

$$\boxed{s_y^2 =} \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left[a x_n + b - (a\bar{x} + b)\right]^2$$

$$= \frac{1}{N-1}\sum_{n=1}^{N}(a x_n - a\bar{x})^2$$

$$= a^2 \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2 = \boxed{a^2 s_x^2}$$

$$\square$$

$$\bar{y} = \overline{ax+b} = \frac{1}{N}\sum_{n=1}^{N}(ax_n + b) = a\frac{1}{N}\sum_{n=1}^{N}x_n + \frac{bN}{N} = a\bar{x} + b$$

Statistische Kennzahlen für die Streuung

Quantitative Daten: **Verschiebungssatz von Steiner**

$$d_x^2 = \left(\frac{1}{N}\sum_{n=1}^{N}(x_n - b)^2\right) - (\overline{x} - b)^2 \qquad \text{speziell für b=0: } d_x^2 = \overline{x^2} - \overline{x}^2$$

Beweis:

$$\boxed{d_x^2} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2 = \frac{1}{N}\sum_{n=1}^{N}[(x_n - b) + (b - \overline{x})]^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}[(x_n - b)^2 + 2(x_n - b)(b - \overline{x}) + (b - \overline{x})^2]$$

$$= \frac{1}{N}\sum_{n=1}^{N}(x_n - b)^2 + 2(b - \overline{x})\frac{1}{N}\sum_{n=1}^{N}(x_n - b) + \frac{1}{N}\sum_{n=1}^{N}(b - \overline{x})^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}(x_n - b)^2 - 2(\overline{x} - b)^2 + (\overline{x} - b)^2 = \boxed{\frac{1}{N}\sum_{n=1}^{N}(x_n - b)^2 - (\overline{x} - b)^2} \qquad \square$$

Statistische Kennzahlen für die Streuung

Quantitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$$s_{x_4}^2 = \frac{2\cdot(10-13.5)^2 + 1\cdot(11-13.5)^2 + 2\cdot(12-13.5)^2}{11} + \frac{1\cdot(13-13.5)^2 + 2\cdot(14-13.5)^2 + 1\cdot(15-13.5)^2}{11} + \frac{1\cdot(16-13.5)^2 + 1\cdot(17-13.5)^2 + 1\cdot(18-13.5)^2}{11} = \boxed{7}$$

$$\bar{x}_4 = \boxed{13.5}$$

$$V_4 = \frac{\sqrt{7}}{13.5} = \boxed{0.196}$$

k	Anzahl Clicks$_{(k)}$
1	10
2	10
3	11
4	12
5	12
6	13
7	14
8	14
9	15
10	16
11	17
12	18

$$Q_{koeff;4} = \frac{2\cdot 4}{11.5+15.5} = \boxed{0.296}$$

$$\tilde{x}_{4;0.25} = 11.5$$

$$\tilde{x}_{4;0.75} = 15.5$$

$$\boxed{Q_4 = 4}$$

$$\boxed{R_4 = 8}$$

Statistische Kennzahlen für die Streuung

Quantitative Daten: Beispiel **Bearbeitungen von Softwareaufgaben**

$$s_{x_5}^2 = \frac{(3.2-5.075)^2+(3.6-5.075)^2+2\cdot(3.7-5.075)^2}{11} + \frac{(3.9-5.075)^2+(4.2-5.075)^2+(4.5-5.075)^2}{11} + \frac{(4.9-5.075)^2+(6.1-5.075)^2+(6.6-5.075)^2}{11} + \frac{(8.0-5.075)^2+(8.5-5.075)^2}{11} = \boxed{3.24}$$

$$\bar{x}_5 = \boxed{5.075}$$

$$V_5 = \frac{\sqrt{3.24}}{5.075} = \boxed{0.355}$$

k	Bearbei-tungszeit$_{(k)}$
1	3.2
2	3.6
3	3.7
4	3.7
5	3.9
6	4.2
7	4.5
8	4.9
9	6.1
10	6.6
11	8.0
12	8.5

$$Q_{koeff;5} = \frac{2\cdot 2.65}{3.7+6.35} = \boxed{0.527}$$

$$\boxed{\tilde{x}_{5;0.25} = 3.7}$$

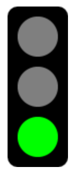
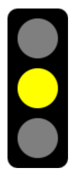
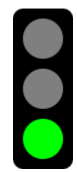
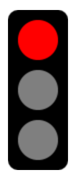
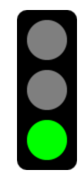
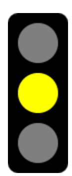
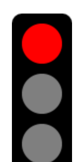
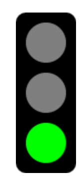
$$\boxed{\tilde{x}_{5;0.75} = 6.35}$$

$$\boxed{Q_5 = 2.65}$$

$$\boxed{R_5 = 5.3}$$

Statistische Kennzahlen für die Streuung

Zusammenfassung: Welche Maßzahlen sind bei welchem Skalenniveau geeignet?

Skalenniveau → ↓ Streuungsmaß	Nominal	Ordinal	Quantitativ
Simpson's D/ Entropie		 – Informations-verlust	 – Nur für klassierte Daten
Leti's D	 – Nur für J = 2		 – Nur für klassierte Daten
MAD/ Spannweite/ Quartilsdifferenz		 – Geringe Aussagekraft für kleine J	 + Robust – Informations-verlust – Hohe Streubreite
Varianz/ Standardabweichung Variationskoeffizient	 – Nur für J = 2		 – Ausreißeranfällig + Informations-nutzung + Geringe Streubreite