

# 2. Merkmale / Zufalls**variablen**

- **2.1 Grundbegriffe**

- Typen von Merkmalen bzw. Zufallsvariablen
- Häufigkeits- bzw. Wahrscheinlichkeits**verteilung**
- **Kumulierte** Häufigkeits- bzw. Wahrscheinlichkeitsverteilung

- 2.2 Kennzahlen diskreter Merkmale / Zufallsvariablen

- Arithmetischer Mittelwert / Erwartungswert
- Andere Mittelwerte: geometrischer / harmonischer Mittelwert
- Median, Quantil, Modus
- Varianz / Standardabweichung

- 2.3 **Stetige** Merkmale / Zufallsvariablen

- Wahrscheinlichkeitsdichten / Dichtefunktion
- Übertragung der diskreten Kennzahldefinitionen

- 2.4 Wichtige **Standardverteilungen**:

- Gleichverteilung
- Binomialverteilung, Poissonverteilung
- Exponentialverteilung
- Normalverteilung

## 2. Merkmale / Zufallsvariablen - Motivation -

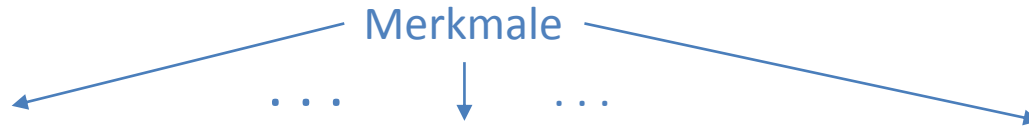
Ihr Freund bietet Ihnen folgendes Spiel an:

Sie werfen zwei Würfel. Wenn die Augensumme größer als 9 ist gewinnen Sie 1€, wenn sie kleiner als 4 ist, verlieren Sie 1€, und in den übrigen Fällen verlieren Sie 10 Cent.

Ist das Spiel für Sie vorteilhaft?

→ Es genügt nicht, die Gewinnwahrscheinlichkeit zu betrachten. Alle möglichen Ergebnisse müssen mit ihren Wahrscheinlichkeiten betrachtet werden.

# Merkmale und Merkmalstypen



Kundenrnr	Alter	Umsatz 2009	Wohnort	Zufrieden- heit	Anzahl Anrufe	Größe [cm]
1	54	1839,73 €	Karlsruhe	1	1	175
2	23	107,67 €	Rüppurr	2	2	182
3	30	1220,43 €	Karlsruhe	3	0	167
4	28	698,24 €	Ettlingen	1	0	171
5	43	1915,62 €	Ettlingen	3	1	179
6	51	1509,81 €	sonstige	2	0	166
7	47	1766,03 €	Rüppurr	2	2	188
8	19	558,50 €	Karlsruhe	1	0	173
9	35	1837,36 €	Rüppurr	3	0	178
10	27	1742,11 €	Rüppurr	3	1	165
11	25	102,96 €	Ettlingen	1	0	185
12	31	4,47 €	Karlsruhe	1	1	180
13	45	424,14 €	Karlsruhe	2	0	176
14	22	1945,67 €	Karlsruhe	1	0	171
15	18	939,74 €	Rüppurr	3	0	173
16	50	1060,89 €	sonstige	3	0	179
17	38	1706,42 €	Ettlingen	2	0	188
18	24	971,48 €	Ettlingen	1	0	193
19	19	127,42 €	sonstige	2	0	176
20	41	1199,33 €	sonstige	4	0	180

Zufriedenheit:  
 1 = begeistert  
 2 = sehr zufrieden  
 3 = zufrieden  
 4 = unzufrieden

# Diskrete und stetige Merkmale

## Definition 25.1

Ein **Merkmal** in einem **Datensatz** ist eine Variable ( $\triangleq$  Spalte) und heißt

- **diskret**, wenn es endlich viele oder abzählbar unendlich viele Ausprägungen gibt.  
(Bsp: Geschlecht, Nationalität, Anz. Anrufe)
- **stetig**, wenn es alle Werte in einem reellen Intervall als Ausprägungen annehmen kann. Das Intervall kann auch (halb-)offen sein.  
(Bsp: Laufzeit)

**Quasistetige** Merkmale sind diskrete Merkmale mit so feiner Abstufung, dass man sie als stetig behandelt. (Bsp: Umsatz)

Die Unterscheidung zwischen diskret und stetig hat damit zu tun, ob es für das Merkmal Sinn macht zu fragen, wie oft exakt eine bestimmte Ausprägung auftritt:

**Beispiel:** Wie viele Kunden haben in 2009 getätigt:

- genau 2 Anrufe (ok, diskretes Merkmal)
- genau 123,21€ Umsatz (ziemlich sinnlos, da (quasi-)stetiges Merkmal)

# Merkmaltypen

## Definition

Ein **Merkmal** heißt

- **Nominal**, wenn nur „gleich“ und „ungleich“ unterscheidbar ist, es aber keine sinnvolle Rangordnung gibt (Bsp: Kundennummer, Hersteller)
- **Ordinal**, wenn sich die Ausprägungen anordnen lassen, die Abstände aber nicht vergleichbar sind. (Bsp: Zufriedenheit, Priorität)
- **Intervallskaliert**, wenn die Ausprägungen Zahlen sind, Differenzen aussagekräftig sind, der Nullpunkt aber willkürlich ist, d.h. Multiplizieren und Dividieren keinen Sinn macht.  
(Bsp. Datum: Datumsdifferenzen sind vergleichbar. Die Aussage „Datum B ist doppelt so groß ist wie Datum A“ macht aber keinen Sinn.)
- **Verhältnisskaliert**, wenn es intervallskaliert ist, und ein sinnvoller absoluter Nullpunkt existiert, d.h. Quotienten (=Verhältnisse) Sinn machen.  
(Bsp: Dauer, Alter, Kosten, Anzahl an Anfragen)

metrisch

# Beispiel (Merkmalstypen)

Welche der folgenden Merkmale sind

- nominal, ordinal, intervallskaliert bzw. verhältnisskaliert? (jeweils kurze Begründung)
  - diskret bzw. stetig?
- Priorität* von Datenpaketen in einem WAN (Ausprägungen: 1, 2, 3)
  - Kommunikationsprotokoll* im Netzwerkverkehr (Ausprägungen: http, https, ftp, ...)
  - Laufzeit* von Datenpaketen in einem WAN. (In Millisekunden, eine Dezimalstelle)
  - Jahresdurchschnittstemperatur* eines Ortes. (In Grad Celsius)

**Lösung:**

**Priorität: Ordinal** denn es gibt höhere oder niedrigere Prioritäten. Man kann aber nicht sagen, dass der Unterschied zwischen Prio 1 und Prio 2 gleich ist wie der zwischen Prio2 und Prio 3. **Diskret**

**Protokoll: Nominal**, denn es gibt keine Rangordnung. **Diskret**.

**Laufzeit: Verhältnisskaliert**, sowohl Differenzen (5 ms kürzer) als auch Faktoren (doppelt so lang) machen Sinn. **Stetig** (genauer: quasistetig).

**Temperatur: Intervallskaliert**. Es macht Sinn zu sagen „der Unterschied zwischen 20 Grad und 10 Grad ist genauso groß wie zwischen 30 Grad und 20 Grad, nämlich 10 Grad (braucht z.B. die gleiche Energie, um Wasser aufzuheizen). Es macht aber keinen Sinn zu sagen, „2 Grad ist doppelt so warm wie 1 Grad, ebenso wie 20 Grad doppelt so warm ist wie 10 Grad“, da der Nullpunkt willkürlich gewählt ist. **Stetig**.

# Zufallsvariable

## Definition

Eine **Zufallsvariable**  $X$  ist ein Merkmal, dessen Wert durch ein Zufallsexperiment **eindeutig** bestimmt wird. Sie werden meist mit Großbuchstaben, wie  $X, Y, Z$  bezeichnet.

## Beispiele

$Z$  = Anzahl Anfragen an einem Server, die morgen ankommen

$Y$  = Laufzeit eines Jobs

$U$  = morgiger Umsatz eines Geschäfts

$X$  = Anzahl Würfe eines 1€-Stücks, bis das erste Mal „Kopf“ kommt“

Dann gilt:  $P(X = 1) = 1/2$  ,  $P(X = 2) = 1/4$

Mögliche Werte (= **Realisationen**) von Zufallsvariablen werden mit Kleinbuchstaben wie  $x$  oder  $x_i$  bezeichnet.

Die Formel in obigem Beispiel lautet also:  $P(X = x) = 0.5^x$

## Bemerkung:

(Un-)gleichungen, in denen Zufalls**variable** vorkommen, beschreiben Zufalls-**Ereignisse**, z.B. „ $X = 3$ “, „ $X > 3$ “, „ $X = x_5$ “

# Zufallsvariable

Zufalls-**Ereignisse** sind **binär**; sie treten entweder ein oder nicht.

Vom Zufall bestimmt Zahlen heißen dagegen Zufallsvariablen:

## Definition

Eine **Zufallsvariable** ist ein metrisches Merkmal, dessen Wert durch ein Zufallsexperiment bestimmt wird. Sie werden meist mit Großbuchstaben, wie  $X$ ,  $Y$ ,  $Z$  bezeichnet.

(Sie beziehen sich auf ein genau festgelegtes Zufallsexperiment und stehen abkürzend für Beschreibungen wie „Anzahl Würfe eines 1€-Stücks, bis das erste Mal ‚Kopf‘ kommt“.)

## Beispiele

$Z$  = Anzahl Anfragen an einem Server  $s$ , die morgen ankommen

$Y$  = Laufzeit eines Jobs

$X$  = Anzahl Würfe eines 1€-Stücks, bis das erste Mal ‚Kopf‘ kommt

Konkrete mögliche Werte (= **Realisationen**) dieser Zufallsvariablen werden mit Kleinbuchstaben wie  $x$  oder  $x_i$  bezeichnet. (Man verwendet sie an Stelle von Konstanten, wenn man eine Formel schreiben will, die für jeden möglichen Wert dieser Konstanten gilt.)

## Beispiel:

$$P(X = 1) = 0.5$$

$$P(X = x) = 0.5^x$$

(Un-)gleichungen, in denen Zufallsvariable vorkommen, beschreiben Zufalls-Ereignisse:

**Beispiele:** „ $X = 3$ “ , „ $X > 3$ “ , „ $X = x_5$ “



# Diskrete Zufalls-Variablen

## Definition

Eine **Zufallsvariable**  $X$  heißt **diskret**, wenn sie nur endlich oder abzählbar unendlich viele Werte  $x_1, x_2, \dots$  annehmen kann. Die Wahrscheinlichkeit, dass der Wert  $x_i$  auftritt, schreibt man als  $P(X = x_i)$ , oder oft auch abkürzend  $p_i$ .

## Beispiel 27.3

Zwei Würfel werden geworfen. Wir betrachten die Zufallsvariablen  $X :=$  „Augensumme“ und  $Y :=$  „Produkt der Augenzahlen“.

Bestimmen Sie

- a)  $P(X = 5)$
- b) die gesamte Wahrscheinlichkeitsverteilung von  $X$
- c)  $P(X \leq 4)$       d)  $P(X > 4)$
- e)  $P(Y = 5)$


# Häufigkeitsverteilung

## Definition

Die **Häufigkeitsverteilung** eines Merkmals  $X$  eines vorliegenden Datensatzes ist eine **Funktion**, die allen möglichen Realisationen  $x_i$  von  $X$  die zugehörigen (relativen oder absoluten) Häufigkeiten zuordnet.

## Beispiel:

Häufigkeitsverteilung der *Noten* in der Mathe 2 Klausur im Sommersemester 2016.



The diagram consists of the word 'Merkmal' centered above a horizontal blue bracket. The bracket extends from the left edge of the word 'Noten' to the right edge of the word 'Klausur' in the example text below.

## Definition

Unter der **Wahrscheinlichkeitsverteilung** (auch **Zähldichte**, **W-Verteilung** oder **Wahrscheinlichkeitsfunktion**) einer diskreten Zufallsvariablen  $X$  versteht man eine **Funktion**, die jedem möglichen Wert  $x_i$  von  $X$  die zugehörige Wahrscheinlichkeit zuordnet.

## Beispiel:

Wahrscheinlichkeitsverteilung der *Anzahl Richtige* beim Lotto.

Eine Wahrscheinlichkeits- bzw. Häufigkeitsverteilung gibt also **für alle** möglichen Werte von  $X$  an, mit welcher Wahrscheinlichkeit sie auftreten. Man kann sie als Tabelle, Formel oder Diagramm angeben.

# Beispiel einer W-Verteilung

## Wahrscheinlichkeitsverteilung der Zufallsvariablen $X = \text{Anzahl Richtige im Lotto}$

Als Tabelle:

Richtige	0	1	2	3	4	5	6
Wahrsch.:	44%	41%	13%	1.8%	$9.7 \cdot 10^{-4}$	$1.8 \cdot 10^{-5}$	$7.2 \cdot 10^{-8}$

Als Formel:

$$P(X = x) = \frac{\binom{6}{x} \binom{43}{6-x}}{\binom{49}{6}} \quad (x \in \{0; 1; \dots; 6\})$$

### Lösung zu 27.3 (vor-vorige Folie)

- a) Das Ereignis „Augensumme ist 5“ wird durch  $X = 5$  beschrieben. Es hat die Wahrscheinlichkeit  $P(X = 5) = \frac{4}{36}$  (denn es gibt vier Wurffolgen  $(1, 4)$ ,  $(2, 3)$ ,  $(3, 2)$ ,  $(4, 1)$  mit der Augensumme 5).
- b) Wir stellen die möglichen Werte  $x_i$ , die  $X$  annehmen kann, gemeinsam mit den zugehörigen Wahrscheinlichkeiten  $p_i = P(X = x_i)$  zusammen:

$x_i$	2	3	4	5	6	7	8	9	10	11	12	
$p_i$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\sum p_i = 1$

Die Summe über alle Wahrscheinlichkeiten ist gleich 1, denn  $X$  nimmt mit Sicherheit eine der Zahlen 2, ..., 12 an.

- c) Das Ereignis  $X \leq 4$  tritt ein, wenn  $X = 2$ ,  $X = 3$  oder  $X = 4$  ist. Daher ist die zugehörige Wahrscheinlichkeit (da diese Ereignisse unvereinbar sind):

$$P(X \leq 4) = P(X = 2) + \dots + P(X = 4) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36} = 16.7\%.$$

- d) Analog wie zuvor ist

$$P(X > 4) = P(X = 5) + \dots + P(X = 12) = \frac{4}{36} + \dots + \frac{1}{36} = \frac{30}{36} = 83.3\%.$$

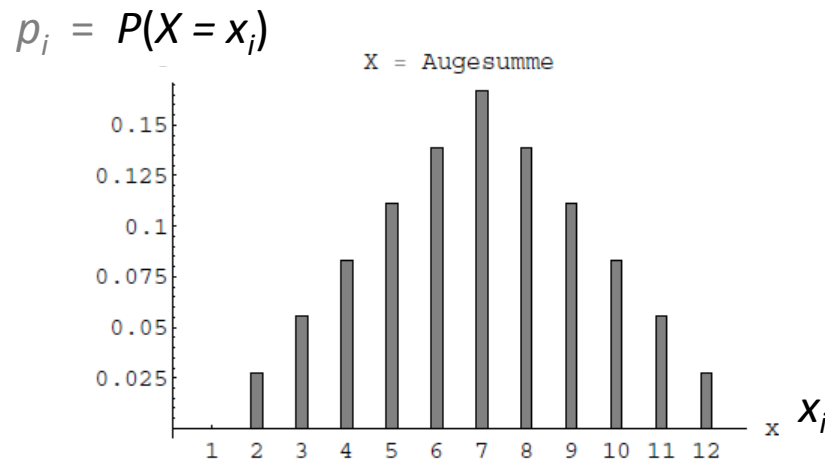
Wir hätten das aber auch einfacher berechnen können:  $X > 4$  ist ja das Gegenereignis von  $X \leq 4$ , daher ist

$$P(X > 4) = 1 - P(X \leq 4) = 1 - \frac{6}{36} = \frac{30}{36} = 83.3\%.$$



# Darstellung einer diskreten Wahrscheinlichkeitsverteilung als Stabdiagramm

Die Wahrscheinlichkeiten  $p_i$  einer diskreten Zufallsvariablen entsprechen den relativen Häufigkeiten  $f_i$  der beschreibenden Statistik. So kann die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen auch graphisch in einem **Stabdiagramm** dargestellt werden: Dabei werden die Realisationen  $x_i$  der Zufallsvariablen auf der horizontalen und die zugehörigen Wahrscheinlichkeiten  $p_i$  als Stäbe auf der vertikalen Achse aufgetragen. Abbildung 27.1 zeigt die Wahrscheinlichkeitsverteilung zu Beispiel 27.3.



**Abbildung 27.1.** Wahrscheinlichkeitsverteilung von  $X = \text{Augensumme zweier Würfel}$ .

# Verteilungsfunktion

## Definition

Die *empirische Verteilungsfunktion* eines mindestens ordinalen Merkmals  $X$  in einem Datensatz ist eine **Funktion**, die allen möglichen Realisationen  $x_i$  die (relativen oder absoluten) **Häufigkeiten von Werten**  $\leq x_i$  zuordnet.

## Definition

Unter der *Verteilungsfunktion*  $F_X$  einer mindestens ordinalen Zufallsvariablen  $X$  versteht man eine **Funktion**, die allen möglichen Werten  $x_i$  die Wahrscheinlichkeiten  $P(X \leq x_i)$  zuordnet. Die Funktion wird meist mit  $F_X$  oder auch nur mit  $F$  bezeichnet:  $F_X(w) := P(X \leq w)$

Sind  $x_1, x_2, \dots, x_n$  die möglichen Werte von  $X$ , so gilt:

$$F(w) = \sum_{x_i: x_i \leq w} P(X = x_i) \quad (\text{die Summe läuft über alle Realisationen } x_i, \text{ die } \leq w \text{ sind})$$

## Beispiel:

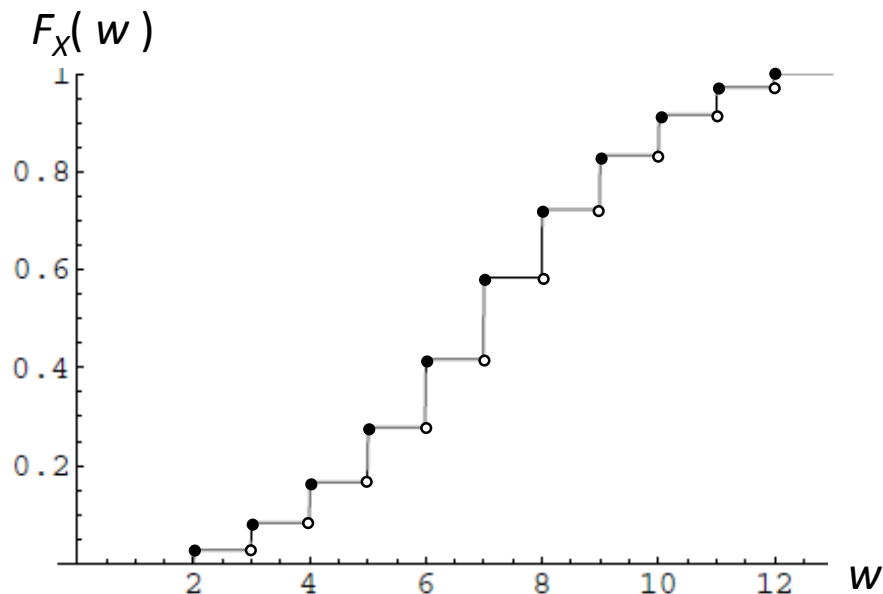
Stellen Sie die Verteilungsfunktion für die Augensumme von zwei Würfeln tabellarisch und als Graphen dar.

f) Bestimmen Sie  $P(X \leq 4), P(X < 4), P(X > 4), P(X \geq 4), P(4 < X \leq 9)$

# Lösung zum Beispiel

f) Verteilungsfunktion der Zufallsvariablen  $X$ : „Augensumme von zwei Würfeln“:

w:	2	3	4	5	6	7	8	9	10	11	12
$P(X = w)$ :	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
$F_X(w) = P(X \leq w)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36



$$g) P(X \leq 4) = F(4) = \frac{6}{36}$$

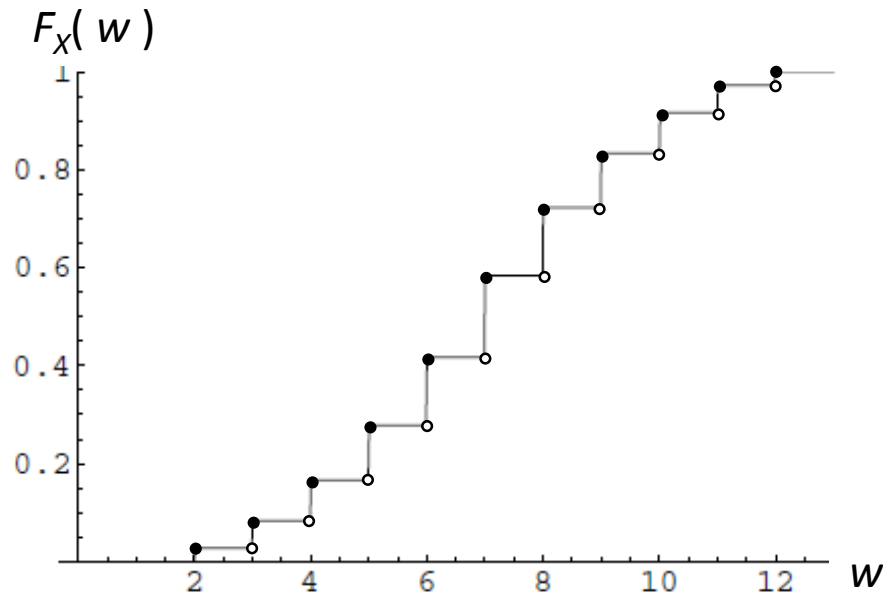
$$P(X < 4) = F(3) = \frac{3}{36}$$

$$P(X > 4) = 1 - F(4) = \frac{30}{36}$$

$$P(X \geq 4) = 1 - F(3) = \frac{33}{36}$$

$$P(4 < X \leq 9) = F(9) - F(4) = \frac{24}{36}$$

# Eigenschaften der Verteilungsfunktion



## Satz 27.5

- Jede Verteilungsfunktion wächst **monoton** von 0 bis 1.
- An Sprungstellen ist sie rechtsseitig stetig, d.h. der obere Wert ist der Funktionswert
- Die **Höhe der Sprungstelle** an einer Stelle  $x_0$  **ist die Wahrscheinlichkeit**, mit der das Ereignis  $X = x_0$  eintritt.



## 2.1 Was Sie gelernt haben sollten

### Merkmale und Zufallsvariablen:

- Typen von Merkmalen bzw. Zufallsvariablen.  
Welche Berechnungen machen dann jeweils Sinn?
- Notation
- Interpretieren bzw. Ermitteln von
  - Häufigkeits- bzw. Wahrscheinlichkeits**verteilung**
  - **Kumulierter** Häufigkeits- bzw. Wahrscheinlichkeitsverteilung

# 2. Quantitative Merkmale / Zufallsvariablen

- **2.1 Grundbegriffe**
  - Typen von Merkmalen bzw. Zufallsvariablen
  - Häufigkeits- bzw. Wahrscheinlichkeits**verteilung**
  - **Kumulierte** Häufigkeits- bzw. Wahrscheinlichkeitsverteilung
- **2.2 Kennzahlen diskreter Merkmale / Zufallsvariablen**
  - Arithmetischer **Mittelwert / Erwartungswert**
  - **Andere Mittelwerte**: geometrischer / harmonischer Mittelwert
  - **Median, Quantil, Modus**
  - **Varianz / Standardabweichung**
- **2.3 Stetige Merkmale / Zufallsvariablen**
  - Wahrscheinlichkeitsdichten / Dichtefunktion
  - Übertragung der diskreten Kennzahldefinitionen
- **2.4 Wichtige Standardverteilungen:**
  - Gleichverteilung
  - Binomialverteilung, Poissonverteilung
  - Exponentialverteilung
  - Normalverteilung

# Arithmetischer Mittelwert / Durchschnitt

## Definition 25.6 (arithmetischer Mittelwert)

Der **arithmetische Mittelwert** der Zahlen  $w_1, \dots, w_n$  ist definiert als

$$\bar{w} = \frac{1}{n} (w_1 + \dots + w_n) = \frac{1}{n} \sum_{i=1}^n w_i$$

Wenn von *dem* **Mittelwert** oder dem **Durchschnitt** gesprochen wird, ist meist der arithmetische Mittelwert gemeint.

Der arithmetische Mittelwert macht ***nur für mindestens intervallskalierte Merkmale*** Sinn.

Der Mittelwert liegt immer zwischen dem größten und kleinsten Einzelwert.

Wenn man Zahlen, die **addiert** werden sollen, durch ihr arithmetisches Mittel ersetzt, ändert sich das Ergebnis nicht. **Für andere Rechenoperationen gilt das nicht.**

# Arithmetischer Mittelwert

Man kann den arithmetischen Mittelwert direkt aus der Häufigkeitsverteilung der Werte ausrechnen. Das ist kürzer als aus den Einzelwerten:

## Satz 25.6

Sei  $w_1, \dots, w_n$  die Werteliste eines Merkmals, in der die Ausprägungen  $x_1, \dots, x_m$  mit den absoluten Häufigkeiten  $h_1, \dots, h_m$  bzw. den relativen Häufigkeiten  $f_1, \dots, f_m$  auftreten. (es kommen also  $m$  verschiedene Werte vor)

Die Berechnung des arithmetischen Mittelwertes kann auf folgende gleichwertige Arten erfolgen:

$$\bar{w} = \frac{1}{n} \cdot \sum_{i=1}^n w_i = \frac{1}{n} \cdot \sum_{i=1}^m x_i \cdot h_i = \sum_{i=1}^m x_i \cdot f_i$$

# Mittelwert

## Beispiel

Ein Server hat Jobs zu bearbeiten, deren Rechenzeit  $r$  in Minuten wie folgt ist:

1, 10, 5, 10, 10, 5, 1

Berechnen Sie die mittlere Rechenzeit pro Job.

a) Die mittlere Rechenzeit pro Job beträgt nach Def. 25.6:

$$\bar{r} = \frac{1 + 10 + 5 + 10 + 10 + 5 + 1}{7} = \frac{42}{7} = 6$$

b) Alternativ kann man das selbe Ergebnis nach Satz 25.6 auch aus der Häufigkeitstabelle berechnen:

$x_i$	1	5	10
$f_i$	$2/7$	$2/7$	$3/7$

$$\bar{r} = \frac{2}{7} \cdot 1 + \frac{2}{7} \cdot 5 + \frac{3}{7} \cdot 10 = 6$$

# Arithmetischer Mittelwert

Bemerkung:

Will man die Gesamtrechenzeit aller Jobs berechnen, so kann man sie wahlweise elementar berechnen :

$$1 + 10 + 5 + 10 + 10 + 5 + 1 = 42$$

oder aus dem Mittelwert 6 und der der Anzahl Summanden 7:

$$6 \cdot 7 = 42$$

## Satz

Wenn man  $n$  Werte jeweils durch ihren arithmetischen Mittelwert ersetzt, ändert sich an der Summe nichts.

Für andere Rechenoperationen gilt das im Allgemeinen nicht!

$$1 \cdot 10 \cdot 5 \cdot 10 \cdot 10 \cdot 5 \cdot 1 = 25\,000, \text{ aber } 6^7 = 279\,936$$

# Gewichteter Mittelwert

## Beispielaufgabe:

Ich habe mein Vermögen von 400€ auf 2 Konten mit 100€ bzw. 300€ und Zinssätzen  $x_1$  bzw.  $x_2$  verteilt.

Was ist der mittlere Zinssatz, den ich auf mein Vermögen erziele?

Die Zinszahlung beträgt  $100 \text{ €} \cdot x_1 + 300 \text{ €} \cdot x_2 = 400 \text{ €} \cdot \left( \frac{100}{400} x_1 + \frac{300}{400} x_2 \right)$

mittlerer Zinssatz

Das nennt man den **gewichtete Mittelwert** von  $x_1$  und  $x_2$ , mit Gewichtungsfaktoren  $g_1 = 1/4$  und  $g_2 = 3/4$  :

## Definition:

Der **gewichtete Mittelwert** der Zahlen  $w_1, \dots, w_n$  mit Gewichtungsfaktoren  $g_1, \dots, g_n$  ( $g_1 + g_2 + \dots + g_n = 1$ ,  $g_i \geq 0$ )

ist definiert als 
$$\sum_{i=1}^n g_i \cdot w_i$$

# Geometrischer Mittelwert

## Definition (geometrischer Mittelwert)

Der **geometrische Mittelwert** der Zahlen  $w_1, \dots, w_n$  ist

$$\overline{w}_{geom} := \sqrt[n]{w_1 \cdot \dots \cdot w_n} = \sqrt[n]{\prod_{i=1}^n w_i}$$

Nur sinnvoll für (mindestens) **verhältnisskalierte** Merkmale  
(da sonst das Produkt keinen Sinn ergibt)

**Satz** Wenn man Zahlen, die **multipliziert** werden sollen durch ihr geometrisches Mittel ersetzt, ändert sich das Ergebnis nicht.

Der geometrische Mittelwert macht also Sinn, wenn es auf das Produkt der Werte ankommt.



# Allgemeines Konzept des Mittelwertes

## Definition

Zu einer Liste von  $n$  Zahlenwerten in einem bestimmten Kontext bezeichnet man einen Wert als **Mittelwert**, wenn er folgende Eigenschaft besitzt:

Statt sich mit den vielen verschiedenen Werten auseinanderzusetzen, will man vereinfachend **alle Werte in der Liste durch den Mittelwert ersetzen** können, **ohne dass sich das Ergebnis dadurch ändert**.

Auf welches Ergebnis es dabei ankommt, ist kontextabhängig.

## Beispiel:

Bei Gewinnen aus mehreren Glücksspielen kommt es auf die Summe an. Also ist der arithmetische Mittelwert relevant.

Bei Wachstumsfaktoren für aufeinanderfolgende Zeiträume ist das Gesamtwachstum deren Produkt. Also ist der geometrische Mittelwert relevant.

# Welcher Typ von Mittelwert passt wann?

Es gibt also nicht nur den arithmetischen Mittelwert, sondern verschiedene Typen von Mittelwerten. Wie entscheidet man, wann welcher passt?:

Wenn man einen **Mittelwert** von  $w_1, \dots, w_n$  berechnet, so meist zu folgendem Zweck:

Man sucht einen **Ersatzwert**, um sich nicht mit allen  $n$  Einzelwerten befassen zu müssen. Tut man so, als wäre in allen Datensätzen dieser Ersatzwert, so soll der dadurch entstehende **Fehler** möglichst klein sein.

Dabei muss man aus dem Kontext **beurteilen, auf welche Zielgröße  $z$  es ankommt** (das hat nichts mit Mathematik zu tun): Wie viel Geld man am Ende hat, wie viele Bakterien es am Ende gibt, ... Dann findet man den passenden Mittelwert wie folgt:

1. Man setzt an, wie man die Zielgröße  $z$  zum einen aus den verschiedenen Einzelwerten berechnen würde (z.B.  $z = g(w_1, \dots, w_n)$ )
2. Man setzt an, wie man  $z$  berechnen würde, wenn man für alle Datensätze den unbekannten Ersatzwert  $\bar{w}$  verwendet (also  $z = g(\bar{w}, \dots, \bar{w})$ ).
3. Der Ersatzwert, für den beides gleich wird, ist der gesuchte Mittelwert, d.h. man löst die Gleichung  $g(w_1, \dots, w_n) = g(\bar{w}, \dots, \bar{w})$  nach  $\bar{w}$  auf.

# Erwartungswert

Vom *Mittelwert* spricht man nur in Bezug auf konkret vorliegende Datensätze. Das Gegenstück zum arithmetischen Mittelwert ist für Wahrscheinlichkeitsverteilungen der *Erwartungswert*:

## Definition 27.19

Sei  $X$  eine diskrete, **metrische** Zufallsvariable mit Realisationen  $x_1, \dots, x_m$ . Der **Erwartungswert** von  $X$  ist eine reelle Zahl, die definiert ist durch

$$E(X) := \sum_{i=1}^m x_i \cdot P(X = x_i)$$

Wobei die Summe über **alle möglichen Werte**  $x_i$  läuft, die  $X$  annehmen kann. Oft wird die Bezeichnung  $\mu$  für den Erwartungswert verwendet.

## Beispiel:

Der Erwartungswert der Augenzahl  $X$  eines Würfels ist

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

**Bemerkung:** Der Erwartungswert  $E(X)$  ist ein *gewichteter Mittelwert* der möglichen Werte von  $X$  – gewichtet nach deren Wahrscheinlichkeit.

# Interpretation von Erwartungswerten

**Satz 27.45 (Gesetz der großen Zahlen, vereinfachte Fassung)**

Sei  $X$  eine Zufallsvariable. Für eine genügend große Zahl von Wiederholungen des zugehörigen Zufallsexperimentes nähert sich der arithmetische Mittelwert  $\bar{X}$  der beobachteten Ergebnisse mit an Sicherheit grenzender Wahrscheinlichkeit beliebig genau an den Erwartungswert  $E(X)$  an.

Zwischen Mittelwert und Erwartungswert gilt also eine analoge Beziehung wie zwischen relativer Häufigkeit und Wahrscheinlichkeit.

# Interpretation von Erwartungswerten

## **Satz 27.45 (Gesetz der großen Zahlen)**

Sei  $X$  eine Zufallsvariable. Führt man  $n$  unabhängige Wiederholungen des Zufallsexperimentes durch und beobachtet Werte  $X_1, \dots, X_n$ , so kann man ihren arithmetischen **Mittelwert**  $\bar{X}$  berechnen.

Es gilt: Für genügend großes  $n$  werden die Abweichungen zwischen  $E(X)$  und  $\bar{X}$  beliebig klein.

Genauer: Für jede Genauigkeit  $\varepsilon > 0$ , die man vorgibt, gilt:

$$\lim_{n \rightarrow \infty} P(|E(X) - \bar{X}| < \varepsilon) = 1$$

Man sagt auch,  **$\bar{X}$  konvergiert stochastisch gegen  $E(X)$**  für wachsenden Stichprobenumfang  $n$ .

Der **Erwartungswert** einer Zufallsvariablen  $X$  gibt also an, welchen **Mittelwert** man „im langfristigen Mittel“, d.h. bei einer genügend großen Anzahl an Versuchen annähernd erhalten wird.

# Erwartungswert einer diskreten Zufallsvariablen

## Beispiel 27.20

Bei einem Glücksspiel dürfen Sie zwei Würfel werfen und bekommen die Augensumme (Zufallsvariable  $X$ ) in Euro ausbezahlt. Wie teuer darf der Einsatz maximal sein, damit Sie im langfristigen Mittel (d.h. wenn Sie sehr oft spielen) kein Geld verlieren?

## Lösung:

Die Wahrscheinlichkeitsverteilung der Zufallsvariable „Augensumme von zwei Würfeln“ wurde bereits in Aufgabe 27.3 berechnet. Damit erhält man für den Erwartungswert von  $X$ :

$$E(X) = \frac{1}{36} \cdot 2\text{€} + \frac{2}{36} \cdot 3\text{€} + \dots + \frac{6}{36} \cdot 7\text{€} + \dots + \frac{1}{36} \cdot 12\text{€} = 7\text{€}$$

Das Spiel darf also maximal 7€ Einsatz kosten, sonst verlieren Sie im langfristigen Mittel Geld.

# Beispielaufgaben (Mittelwert)

Welche Schlussfolgerungen sind richtig welche falsch? Korrigieren Sie falls möglich.

1. Die 100 Mitarbeiter eines Unternehmens haben 2013 im (arithmetischen) Mittel jeweils 500 € Telefonkosten verursacht.  
*Dann beträgt die Gesamt-Telefonrechnung 50 000 €.*
2. Ich habe mein Vermögen von 600 € auf 3 Konten mit 100 €, 200 € bzw. 300 € verteilt. Der (arithmetische) Mittelwert der Jahreszinssätze beträgt 3 % .  
*Dann bekomme ich pro Jahr insgesamt  $600 \text{ €} \cdot 0.03 = 18 \text{ €}$  Zinsen.*
3. Die Wertzuwachsraten einer Aktie seien **+100%** für 2007, **-91%** für 2008, und **+0%** für 2009 (negative Werte = Verlust).
  - a) Dann ist die mittlere jährliche Wertzuwachsrate 3%.
  - b) Dann steht die Aktie am Ende um 9% höher als am Anfang

**Merke: Man kann nicht immer die Einzelwerte einfach durch ihren arithmetischen Mittelwert ersetzen.**

# Lösung zu den Beispielaufgaben (Mittelwert)

1.  $GK$  := Gesamtkosten,  $K_i$  := Kosten des  $i$ -ten Mitarbeiters  
Bekannt:  $n = 100$  Mitarbeiter,  $\bar{K} = 500$  €  
Es gilt:  $GK = K_1 + \dots + K_n$  und  $\bar{K} = (K_1 + \dots + K_n) / n$ , also  
 $GK = \bar{K} \cdot n = 500 \text{ €} \cdot 100 = \underline{50\,000 \text{ €}} \rightarrow$  Schluss war korrekt
2.  $Z$  := Gesamtzinsszahlung,  $Z_i$  = Zinsszahlung auf Konto  $i$ ,  
 $\bar{z}$  := durchschnittl. Zinssatz,  $z_i$  = Zinssatz auf Konto  $i$ ,  
Bekannt:  $\bar{z} = 0.03$   
Ich erhalte an Zinsen:  $Z = Z_1 + \dots + Z_3 = 100 \text{ €} \cdot z_1 + 200 \text{ €} \cdot z_2 + 300 \text{ €} \cdot z_3$   
Mit dem mittleren Zinssatz gerechnet:  $600 \text{ €} \cdot \bar{z} = 600 \text{ €} \cdot (z_1 + z_2 + z_3) / 3 =$   
 $= 200 \text{ €} \cdot z_1 + 200 \text{ €} \cdot z_2 + 200 \text{ €} \cdot z_3$   
Mit dem mittleren Zinssatz kommt also etwas falsche heraus.  
 $\rightarrow$  Schluss nicht korrekt
3.  $X$  := „Wert der Aktie Ende 2009“ / Wert der Aktie Anfang 2007“  
Der Gesamt-Wertsteigerungsfaktor ist  $X = (1+1) \cdot (1-0.91) \cdot (1+0) = 0.18$   
Pro Jahr ist das im Mittel  $\sqrt[3]{0.18} \approx 0.56$   
Die Aktie ist insgesamt auf 18% ihres Wertes gefallen, denn sie hat sich zunächst verdoppelt, dann fast gezehntelt. Pro Jahr hat sie also im Mittel knapp die Hälfte verloren. Also war der Schluss nicht korrekt.



# Median / Quantile

## Definition (Median, vereinfachte Fassung)

Der Median  $m$  einer Werteliste ist ein Wert mit der Eigenschaft, dass ca. 50% der Werte  $\leq m$  sind.

Der Median  $m$  einer Zufallsvariablen  $X$  ist ein Wert mit der Eigenschaft, dass  $P(X \leq m)$  ca. 50% beträgt.

**Beispiel:** Der Median des Brutto-Jahresverdienstes von IT-Beratern 2013 betrug **58 010 €**.

## Definition (Quantil, vereinfachte Fassung)

Für einen Wert  $p$  zwischen 0 und 1 ist das  $p$ -Quantil  $q$  einer Werteliste ein Wert mit der Eigenschaft, dass ca. ein Anteil von  $p$  der Werte  $\leq q$  sind.

Analog ist das  $p$ -Quantil  $q$  einer Zufallsvariablen  $X$  ein Wert mit der Eigenschaft, dass  $P(X \leq q)$  in etwa gleich  $p$  ist.

**Beispiel:** Das 25%-Quantil des Brutto-Jahresverdienstes von IT-Beratern 2013 betrug **48 063 €**.

**Median und Quantile machen nur für mindestens ordinale Merkmale Sinn.**

Nicht immer ist die geforderte Wahrscheinlichkeit von 50% bzw.  $p$  exakt erreichbar. Deshalb sind die präzisen Definitionen komplizierter. (-> nächste Folie)

# Beispiel

## Beispiel 1

5 Personen haben folgende Bargeldmengen bei sich: 15€, 4€, 1000 €, 34€, 18€ .  
bestimmen Sie den Median der bargeldmengen.

**Lösung:** Median = **18 €**. (3 von 5 Personen haben  $\leq 18€$  und 3 von 5 haben  $\geq 18€$ )

- Median ist **aufwändiger** zu berechnen als der Mittelwert. (Liste muss sortiert werden)
- Median ist **robuster gegen Ausreißer** als der Mittelwert.

## Beispiel 2

Bestimmen Sie den Median des Merkmals X  
aus nebenstehender **Häufigkeitstabelle**:

X	1	2	3	4	5	7
Häufigk.	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

**Lösung:**

50 % sind  $\leq 2$ , 50%  $\geq 3 \Rightarrow$  Median = 2.5 (Mitte von [2;3])

## Beispiel 3

Bestimmen Sie das 30%-Quantil der Bargeldmengen aus Beispiel 1 und von X aus Beispiel 2.

**Lösung:**

- 1) **15€** , denn es haben sowohl  $2/5 = 40\%$ , also mindestens 30%  $\leq 15€$   
als auch  $4/5 = 80\%$ , also mindestens 30%  $\geq 15€$
- 2) **2** , denn sowohl sind  $4/8 = 50\%$  , also mindestens 30% sind  $\leq 2$  ,  
als auch  $7/8$ , also über 30% sind  $\geq 2$

# Median (präzisere Definition)

## Definition (Median)

Der **Median  $m$**  einer Liste  $w_1, \dots, w_n$  von Werten ist ein Wert mit der Eigenschaft, dass (mindestens) die Hälfte der Listenelemente einen Wert  $\leq m$  und (mindestens) die Hälfte der Listenelemente einen Wert  $\geq m$  hat.

Der **Median einer Wahrscheinlichkeitsverteilung** einer Zufallsvariablen  $X$  ist ein Wert  $m$  mit der Eigenschaft, dass  $P(X \leq m) \geq 50\%$  und gleichzeitig  $P(X \geq m) \geq 50\%$ . (Bei Mehrdeutigkeit nimmt man die Mitte des in Frage kommenden Bereiches.)

Der Median macht für mindestens ordinale Merkmale Sinn.

## Beispiel 1

5 Personen haben folgende Bargeldmengen bei sich: 15€, 4€, 1000 €, 34€, 18€

→ Median = 18 €. (3 von 5 Personen haben  $\leq 18\text{€}$  und 3 von 5 haben  $\geq 18\text{€}$ )

- Median ist **aufwändiger** zu berechnen als der Mittelwert. (Liste muss sortiert werden)
- Median ist **robuster gegen Ausreißer** als der Mittelwert.

## Beispiel 2

Bestimmen Sie den Median eines Merkmals aus nebenstehender **Häufigkeitstabelle**:

Wert	1	2	3	4	5	7
Häufigk.	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Lösung: 50 % sind  $\leq 2$ , 50%  $\geq 3 \Rightarrow$  Median = 2.5 (Mitte von [2;3])

# Quantile (präzise Definition)

## Definition 25.10

Das  **$p$ -Quantil  $q_p$**  einer Liste  $w_1, \dots, w_n$  von Werten ist ein Wert mit der Eigenschaft, dass (mindestens) ein Anteil  $p$  der Listenelemente  $\leq q_p$  ist und (mindestens) ein Anteil  $(1-p)$  der Listenelemente  $\geq q_p$  ist.

Ein  **$p$ -Quantil  $q_p$**  der *Wahrscheinlichkeitsverteilung* einer Zufallsvariablen  $X$  ist ein Wert  $q_p$  mit der Eigenschaft, dass  $P(X \leq q_p) \geq p$  und  $P(X \geq q_p) \geq (1 - p)$ .

(Bei Mehrdeutigkeit nimmt man die Mitte des in Frage kommenden Bereiches.)

Das 50%-Quantil ist der Median.

Die Berechnung von Quantilen erfolgt analog zum Median.

## Beispiele

Welches (niedrige) Einkommen wird nur noch von 10% der Bevölkerung unterschritten?  
(**10%-Quantil** der Einkommen)

Welche (hohe) Einkommen wird nur von 10% der Bevölkerung überschritten?  
(**90%-Quantil** der Einkommen)

Für welche Last muss ich mein System dimensionieren, damit es bei schwankender Last in 99 % der Fälle nicht überlastet wird? (**99%-Quantil** der Last)

# Wiederholung Median / Quantile

## Anschauliche Definition:

**Median:** Ein Wert  $m$ , für den  $P(X \leq m) = 50\%$  gilt. Wenn das nicht geht, der kleinste Wert  $m$ , für den  $P(X \leq m) \geq 50\%$

**90%-Quantil:** Ein Wert  $q$ , für den  $P(X \leq q) = 90\%$  gilt. Wenn das nicht geht, der kleinste Wert  $q$ , für den  $P(X \leq q) \geq 90\%$

**Beispiel:** Wahrscheinlichkeitsverteilung der Zufallsvariablen  $X$ :

x:	-5	-1	0	2	5	7	11	12	14	50
P( X = x):	3%	8%	12%	13%	17%	9%	14%	13%	9%	2%

$P(X \leq -5) = 3\%$ ,  $P(X \leq -1) = 11\%$ ,  $P(X \leq 0) = 23\%$ ,  $P(X \leq 2) = 36\%$ ,  $P(X \leq 5) = 53\%$ ,  
also ist 5 der Median von  $X$ .

$P(X \leq 7) = 62\%$ ,  $P(X \leq 11) = 76\%$ ,  $P(X \leq 12) = 89\%$ ,  $P(X \leq 14) = 98\%$ ,  $P(X \leq 50) = 100\%$ ,  
also ist 14 das 90%-Quantil von  $X$ .

# Quantile aus der Verteilungsfunktion ablesen

Man kann Quantile direkt aus der kumulierten Verteilungsfunktion ablesen:

Fortsetzung des Beispiels von Folie 2-9, 2-11 und 2-12:

$X :=$  „Augensumme von zwei Würfeln“

Was ist das 15%-Quantil von  $X$ ?

Für welchen Wert  $q$  ist  $P(X \leq q) \approx 15\%$ , also  $F(q) \approx 15\%$ ?

-> „Umkehrfunktion“ von  $F$  nutzen bzw.  $F$  rückwärts ablesen.

-> Antwort: **Vier**. (Genauer: 4 ist die kleinste Zahl  $q$  mit  $P(X \leq q) \geq 15\%$ .)

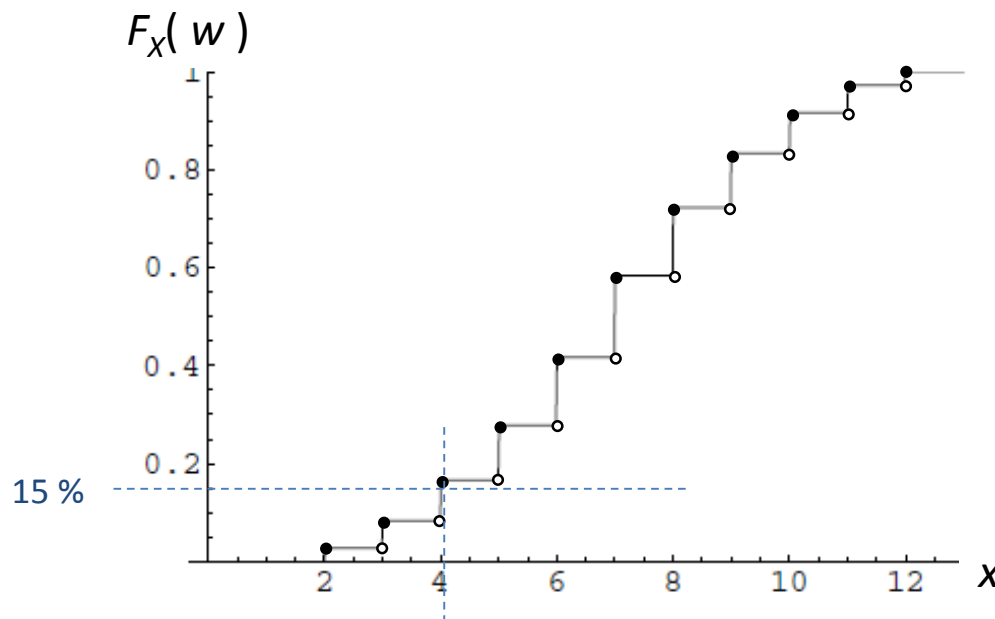


Abb. 27.2: Verteilungsfunktion von  $X$ =„Augensumme von 2 Würfeln“

# Modalwert / Modus

**Definition** (häufigster/wahrscheinlichster Wert)

Der **Modalwert** oder **Modus** einer Liste  $w_1, \dots, w_n$  von Werten ist ein Wert, dessen Häufigkeit größer oder gleich der Häufigkeit jedes anderen Wertes ist.

Der **Modalwert** oder **Modus** einer **Wahrscheinlichkeitsverteilung** einer Zufallsvariablen  $X$  ist ein Wert  $m$  mit der Eigenschaft, dass  $P(X = m) \geq P(X = x)$  für jedes andere  $x$ .

## Beispiel

Bei einer Umfrage in einem Wahlbezirk ergibt sich folgende Häufigkeitstabelle:

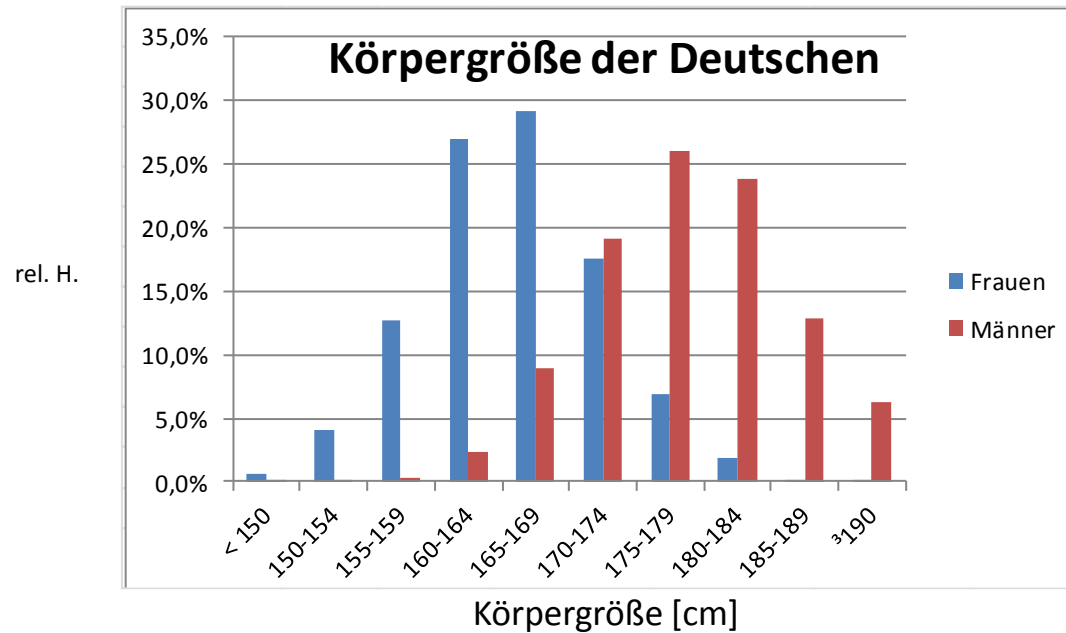
Linke	Grüne	SPD	FDP	CDU
5%	20%	25%	5%	35%

Was ist der Modalwert? → CDU

Was ist der Modalwert, wenn SPD und Grüne einen gemeinsamen Kandidaten aufstellen? → Grüne+SPD

**Also: Wenn man die Klasseneinteilung ändert, d.h. Ausprägungen zusammenlegt oder aufsplittet, kann sich der Modalwert stark ändern.**

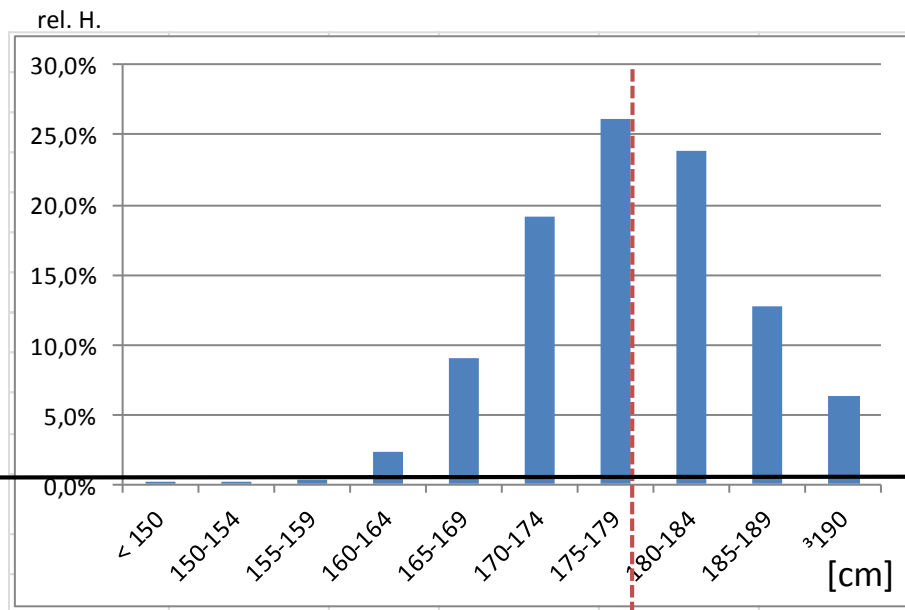
# Beispiel (symmetrische Verteilung)





# Beispiel (symmetrische Verteilung)

## Körpergröße von Männern in Deutschland:



Quelle: nach SOEP 2006

Eine solche Verteilung heißt Normalverteilung.

Für die Normalverteilung sind Mittelwert, Median und Modalwert Identisch.

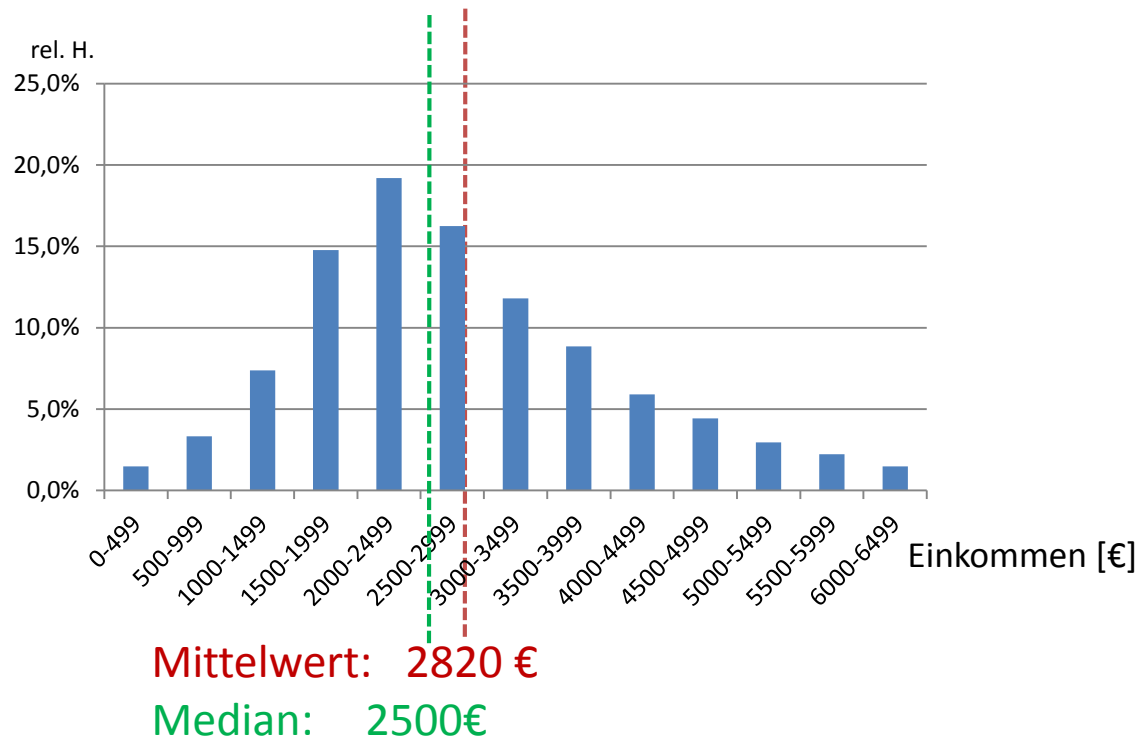
Mittelwert: 178 cm

Median: 178 cm

Modalwert: 178 cm

# Beispiel (rechtsschiefe Verteilung)

Haushaltseinkommen in einer (nicht repräsentativen) Stichprobe:



Eine Verteilung heißt **rechtsschief**, wenn sie rechts flacher abfällt als links. Bei rechtsschiefen Verteilungen ist der Mittelwert größer als der Median. Rechtsschiefe Verteilungen entstehen z.B. oft dann, wenn die Werte links eine Untergrenze haben (im Beispiel 0 €), rechts aber keine Obergrenze.

# Streuungskennwerte

## Definition

Kennzahlen, die etwas über den „**typischen Wert**“ eines Merkmals aussagen, werden als **Lagekennwert** (auch **Lagemaß**) bezeichnet.  
(Beispiele: Mittelwerte, Median, Modus)

Ein **Streuungskennwert** (auch **Streuungsmaß**) ist eine Kennzahl, die etwas darüber aussagt, wie stark die Werte eines Merkmals **schwanken**. (Beispiele: Varianz, Standardabweichung, Spannweite)

# Quartilsabstand

## Definition

Unter dem 1. **Quartil** versteht man das 25%-Quantil, unter dem 3. **Quartil** das 75% Quantil, und unter dem **Quartilsabstand** die Differenz zwischen beiden.

Quartile sind also bestimmte Quantile.

Der Quartilsabstand ist ein gängiger **Streuungskennwert**.

# Stichproben-Varianz und -Standardabweichung

Die **wichtigsten Streuungskennwerte** sind Varianz und Standardabweichung:

## Definition

Sei  $w_1, \dots, w_n$  die Werteliste eines Merkmals. Dann ist die **unkorrigierte Stichprobenvarianz  $s^2$**  des Merkmals definiert als

$$s^2 := \frac{1}{n} \cdot \sum_{i=1}^n (w_i - \bar{w})^2$$

und die **unkorrigierte Stichproben-Standardabweichung  $s$**  als

$$s := \sqrt{s^2}$$

Beide sind sog. **Streuungsmaße**, d.h. sagen etwas darüber aus, wie stark die Werte des Merkmals schwanken.

**Bemerkung:** Es gibt auch eine **korrigierte** Stichprobenvarianz, die später noch eingeführt wird. Sie ist definiert als

$$\frac{1}{n-1} \cdot \sum_{i=1}^n (w_i - \bar{w})^2$$

# Stichproben-Varianz und -Standardabweichung

## Satz

Treten in der Werteliste  $w_1, \dots, w_n$  eines Merkmals die Ausprägungen  $x_1, \dots, x_m$  mit den absoluten Häufigkeiten  $h_1, \dots, h_m$  bzw. den relativen Häufigkeiten  $f_1, \dots, f_m$  auf ( $m \leq n$ ), dann lässt sich die unkorrigierte Stichprobenvarianz auch berechnen als

$$s^2 = \frac{1}{n} \cdot \sum_{j=1}^m h_j \cdot (x_j - \bar{w})^2 = \sum_{j=1}^m f_j \cdot (x_j - \bar{w})^2$$

Diese Formel eignet sich also, wenn von einem Merkmal eine Häufigkeitstabelle oder ein Häufigkeitsdiagramm vorliegt.

# Berechnung der Stichprobenvarianz (1)

## Beispiel 1

In einer Umfrage unter den 20 Vertriebsmitarbeitern einer Abteilung wurde ermittelt, wie viele Kundengespräche sie jeweils an einem bestimmten Tag geführt haben:

3; 2; 4; 1; 3; 2; 0; 1; 2; 2; 1; 3, 4; 1; 2; 1; 1; 0; 3; 2

Bestimmen Sie Mittelwert und unkorrigierte Standardabweichung der Anzahl Kundengespräche

**Lösungsvariante 1** (nach VL 2-23):

$$\overline{w} = \frac{1}{20} (3 + 2 + 4 + \dots + 3 + 2) = \underline{\underline{1.9}}$$

$$s^2 = \frac{1}{20} ((3 - 1.9)^2 + (2 - 1.9)^2 + (4 - 1.9)^2 + \dots + (3 - 1.9)^2 + (2 - 1.9)^2) = 1.29$$

$$s = \sqrt{1.29} \approx \underline{\underline{1.14}}$$

# Berechnung der Stichprobenvarianz (2)

## Beispiel 1

In einer Umfrage unter den 20 Vertriebsmitarbeitern einer Abteilung wurde ermittelt, wie viele Kundengespräche sie jeweils an einem bestimmten Tag geführt haben:

3; 2; 4; 1; 3; 2; 0; 1; 2; 2; 1; 3, 4; 1; 2; 1; 1; 0; 3; 2

Bestimmen Sie Mittelwert und unkorrigierte Standardabweichung der Anzahl Kundengespräche

## Lösungsvariante 2 (nach VL 2-24):

Man ermittelt folgende Häufigkeitstabelle:

Gespräche	0	1	2	3	4
$h$	2	6	6	4	2
$f$	10%	30%	30%	20%	10%

$$\bar{w} = \frac{1}{20} (2 \cdot 0 + 6 \cdot 1 + 6 \cdot 2 + 4 \cdot 3 + 2 \cdot 4) = \underline{\underline{1.9}}$$

$$\text{oder } \bar{w} = 0.1 \cdot 0 + 0.3 \cdot 1 + 0.3 \cdot 2 + 0.2 \cdot 3 + 0.1 \cdot 4 = \underline{\underline{1.9}}$$

$$s^2 = \frac{1}{20} (2 \cdot (0 - 1.9)^2 + 6 \cdot (1 - 1.9)^2 + 6 \cdot (2 - 1.9)^2 + 4 \cdot (3 - 1.9)^2 + 2 \cdot (4 - 1.9)^2) = 1.29$$

$$\text{o. } s^2 = 0.1 \cdot (0 - 1.9)^2 + 0.3 \cdot (1 - 1.9)^2 + 0.3 \cdot (2 - 1.9)^2 + 0.2 \cdot (3 - 1.9)^2 + 0.1 \cdot (4 - 1.9)^2 = 1.29$$

$$s = \sqrt{1.29} \approx \underline{\underline{1.14}}$$



# Stichprobenvarianz (3)

Oft bequemer zum Berechnen der Stichprobenvarianz ist der

## Verschiebungssatz

Sei  $w_1, \dots, w_n$  eine Urliste mit Umfang  $n$ . Dann ist die unkorrigierte Stichprobenvarianz gleich

$$s^2 = \overline{w^2} - (\overline{w})^2$$

d.h. 
$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^n (w_i)^2 - \left( \frac{1}{n} \cdot \sum_{i=1}^n w_i \right)^2$$

Vorteil bei sehr großen Stichproben, die nicht in den Hauptspeicher passen:  
Mittelwert und Standardabweichung können in einem einzigen Durchlauf durch die Stichprobe berechnet werden.

Statt mit den Einzelwerten  $w_1, \dots, w_n$  der Urliste kann natürlich analog zum von drei Folien zurück Satz behandeln auch mit Häufigkeiten gearbeitet werden.

# Varianz / Standardabweichung von Zufallsvariablen

## Definition 27.30

Sei  $X$  eine diskrete Zufallsvariable mit Erwartungswert  $\mu = E(X)$  und Realisationen  $x_1, \dots, x_m$ .

Die **Varianz** von  $X$  ist definiert durch

$$V(X) := \sum_{i=1}^m ((x_i - \mu)^2 \cdot P(X = x_i))$$

Für die Varianz einer Zufallsvariablen  $X$  verwendet man meist die Bezeichnung  $V(X)$  oder  $\sigma^2(X)$ .

Die **Standardabweichung** von  $X$  ist definiert als  $\sigma(X) := \sqrt{V(X)}$

## Beispiel 27.32 Varianz einer diskreten Zufallsvariablen

Berechnen Sie die Varianz und die Standardabweichung der Zufallsvariablen  $X = \text{Augensumme beim Wurf von zwei Würfeln}$  (Beispiel 27.3).

# Varianz einer Zufallsvariablen (Beispiel)

## Beispiel 27.32 Varianz einer diskreten Zufallsvariablen

Berechnen Sie die Varianz und die Standardabweichung der Zufallsvariablen  $X = \text{Augensumme beim Wurf von zwei Würfeln}$  (Beispiel 27.3).

### Lösung:

Die Wahrscheinlichkeitsverteilung von  $X$  wurde schon in Beispiel 27.3 berechnet:

$x_i$	2	3	4	5	6	7	8	9	10	11	12	
$p_{i;}=P(X=x_i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	$\Sigma p_i = 1$

Aus Beispiel 27.20 ist bekannt , dass  $E(X) = 7$

Also gilt

$$V(X) = (2-7)^2 \cdot \frac{1}{36} + (3-7)^2 \cdot \frac{2}{36} + (4-7)^2 \cdot \frac{3}{36} + \dots + (12-7)^2 \cdot \frac{1}{36} \approx \underline{\underline{5.83}}$$

$$\sigma(X) = \sqrt{V(X)} \approx \underline{\underline{2.42}}$$

# Vertiefung: Ungleichung von Tschebyscheff

**Satz 27.40 (Ungleichung von Tschebyscheff)** Für eine Zufallsvariable mit Erwartungswert  $E(X) = \mu$  und Varianz  $\text{Var}(X) = \sigma^2$  gilt für beliebiges  $c > 0$  die Ungleichung

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

In Worten: Die Wahrscheinlichkeit, dass  $X$  um  $c$  oder mehr vom Erwartungswert abweicht, ist unabhängig von der Verteilung höchstens gleich  $\frac{\sigma^2}{c^2}$ .

Nur mit Kenntnis der Varianz kann man also schon das Risiko großer Abweichungen vom Erwartungswert nach oben hin eingrenzen.

## Beispiel:

$\sigma = 2$ , dann ist die Wahrscheinlichkeit dafür, dass  $X$  um mehr als 4 vom Erwartungswert abweicht kleiner als  $2^2/4^2 = 1/4$

Eine exakte Berechnung unter Berücksichtigung der genauen Wahrscheinlichkeitsverteilung wird aber meist genauere Werte liefern, d.h. das Risiko als viel geringer einstufen, als die Ungleichung von Tschebyscheff.

# Beweis der Ungleichung von Tschebyscheff

$$P(|X - \mu| \geq c) = \sum_{i: |x_i - \mu| \geq c} p_i \leq \sum_{i: |x_i - \mu| \geq c} \frac{(x_i - \mu)^2}{c^2} p_i \leq \sum_i \frac{(x_i - \mu)^2}{c^2} p_i = \frac{\text{Var}(X)}{c^2}.$$

Die erste Ungleichung gilt, da  $1 \leq \frac{(x_i - \mu)^2}{c^2}$  für  $|x_i - \mu| \geq c$ . Die zweite Ungleichung gilt, da die Anzahl der Elemente, über die summiert wird, vergrößert wurde.

# Rechengesetze für Erwartungswert und Varianz

## Beispiel:

In einer Beratungsfirma bestehen die Projektteams im Durchschnitt aus 5 Mitarbeitern und die durchschnittliche Projektdauer beträgt 70 Arbeitstage.

Was kann man daraus über die im Schnitt pro Projekt benötigte Anzahl an Personentagen schließen?

# Rechengesetze für Erwartungswert und Varianz

## Beispiel:

In einer Beratungsfirma bestehen die Projektteams im Durchschnitt aus 5 Mitarbeitern und die durchschnittliche Projektdauer beträgt 70 Arbeitstage.

Was kann man daraus über die im Schnitt pro Projekt benötigte Anzahl an Personentagen schließen?

Nichts, wie man an folgendem Beispiel sehen kann:

**D:** Projektdauer, **M:** #Mitarbeiter, **T:** #Personentage

D \ M:	1	9
10	50%	0
130	0	50%

D \ M:	1	9
10	25%	25%
130	25%	25%

in beiden Fällen ist  $E(M)=5$  und  $E(D)=70$ , und trotzdem

$$\begin{aligned} E(T) &= \\ &= 1 \cdot 10 \cdot 0.5 + 9 \cdot 130 \cdot 0.5 = \\ &= \mathbf{590} \end{aligned}$$

$$\begin{aligned} E(T) &= \\ &= 1 \cdot 10 \cdot 0.25 + 9 \cdot 10 \cdot 0.25 + \\ &\quad 1 \cdot 130 \cdot 0.25 + 9 \cdot 130 \cdot 0.25 = \\ &= \mathbf{350} \end{aligned}$$

# Rechengesetze für Erwartungswert und Varianz

(Vorgriff, kommt später noch ausführlicher)

**Satz: 27.28** Für Zufallsvariablen  $X, Y$  des selben Zufallsexperimentes, und Konstante  $a, b$  (die also nicht vom Ausgang des Experimentes abhängen) gilt:

$$(E1) \quad E(aX + b) = a \cdot E(X) + b$$

$$(V1) \quad V(aX + b) = a^2 \cdot V(X)$$

$$(E2) \quad E(X + Y) = E(X) + E(Y)$$

$$(V2) \quad V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$$

$$(E3) \quad E(X \cdot Y) = E(X) \cdot E(Y) + Cov(X, Y)$$

$$(C1) \quad Cov(aX, Y) = Cov(X, aY) = a \cdot Cov(X, Y)$$

$$(C2) \quad Cov(X, X) = V(X)$$

**Falls  $X$  von  $Y$  unabhängig** ist, vereinfachen sich die Formeln, da dann  $Cov(X, Y) = 0$ .

E1, E2, V1 und V2 gelten analog auch für mehr als zwei Zufallsvariablen  $X_1, \dots, X_n$ , insbesondere:

$$(E4) \quad E(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n)$$

**Falls** die  $X_1, \dots, X_n$  paarweise voneinander stochastisch **unabhängig** sind gilt zusätzlich:

$$(V3) \quad V(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \dots + a_n^2V(X_n)$$

**Vorsicht: Für die Standardabweichung gelten diese Rechenregeln nur indirekt über die zugehörige Varianz!**

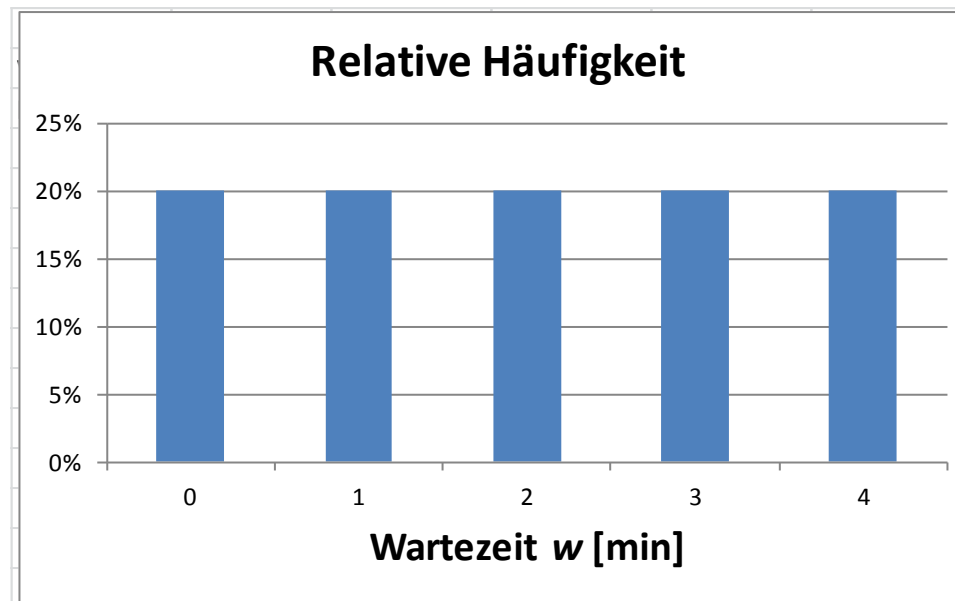


# Was Sie gelernt haben sollten

- Standardabweichungen zu einer Wahrscheinlichkeitsverteilung oder Werteliste berechnen
- Standardabweichungen interpretieren (z.B. aus Wahrscheinlichkeitsverteilungen schätzen)
- Den Erwartungswert von Zufallsvariablen, die per Formel aus anderen Zufallsvariablen berechnet werden, berechnen. (Mit und ohne Benutzung der Rechengesetze)
- Die Standardabweichung von Zufallsvariablen, die per Formel aus anderen Zufallsvariablen berechnet werden, berechnen. (Mit und ohne Benutzung der Rechengesetze)

# Standardabweichung (Interpretation)

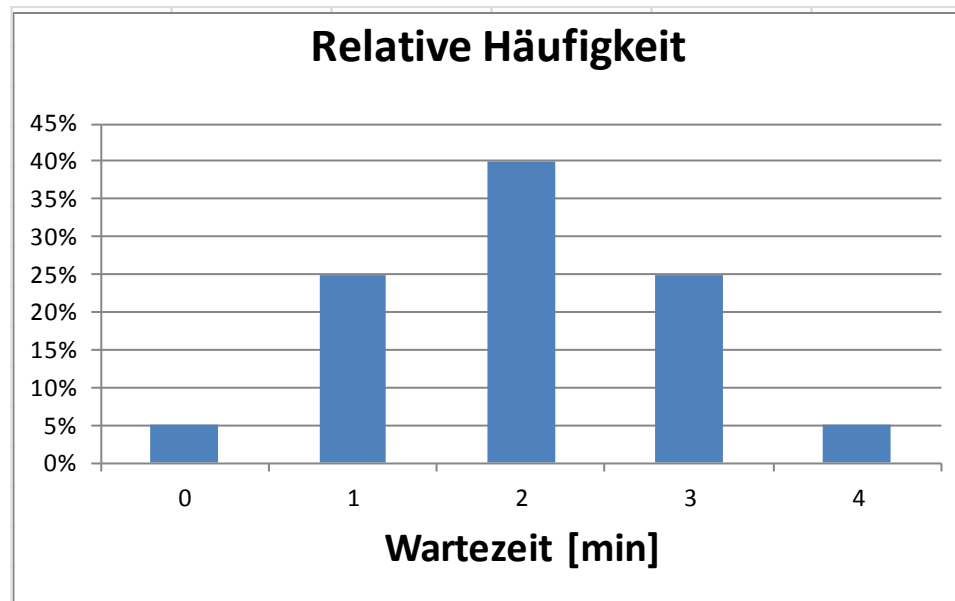
Schätzen Sie die Stichproben-Standardabweichung von  $w$  in folgender Stichprobe:



→ Mit Augenmaß kann man erkennen, dass sie über 1 und deutlich unter 2 sein muss. (Die exakten Werte sind  $\bar{w} = 2 \text{ min}$ ,  $s_w = \sqrt{2} \approx 1.4 \text{ min}$ )  
Achtung: Es geht darum, wie stark die Werte von  $w$  streuen – nicht darum, wie stark sich die Balkenhöhen unterscheiden.

# Standardabweichung (Interpretation)

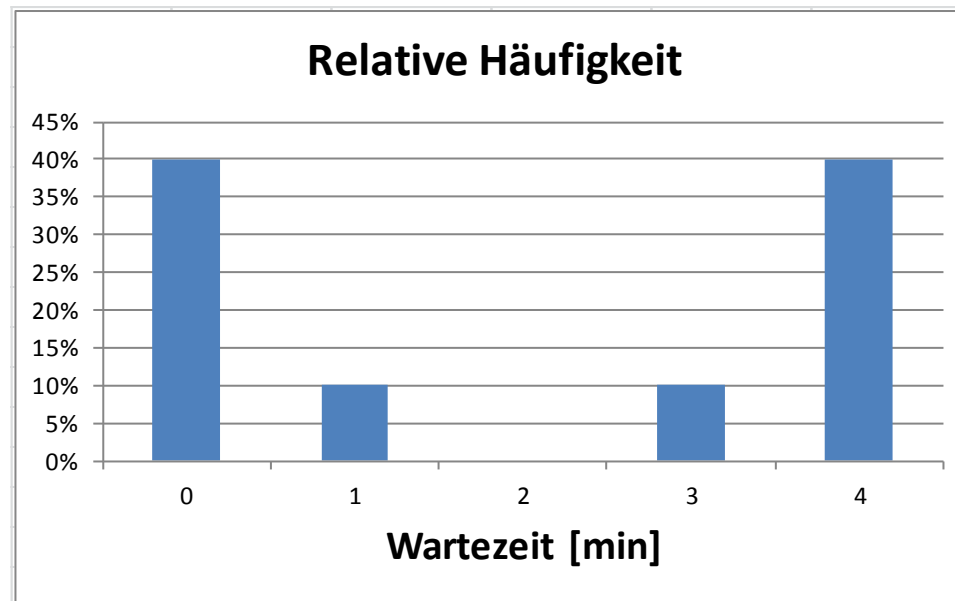
Schätzen Sie die Stichproben-Standardabweichung folgender Stichprobe:



- Auf jeden Fall kleiner als bei Gleichverteilung (vgl. vorhergehende Folie)  
Die Abweichung vom Mittelwert ist viel häufiger 0 als 2, aber durch das Quadrieren haben große Abweichungen stärkeren Einfluss.  
Der exakte Wert ist ca. 0.95 min.

# Standardabweichung (Interpretation)

Schätzen Sie die Stichproben-Standardabweichung folgender Stichprobe:

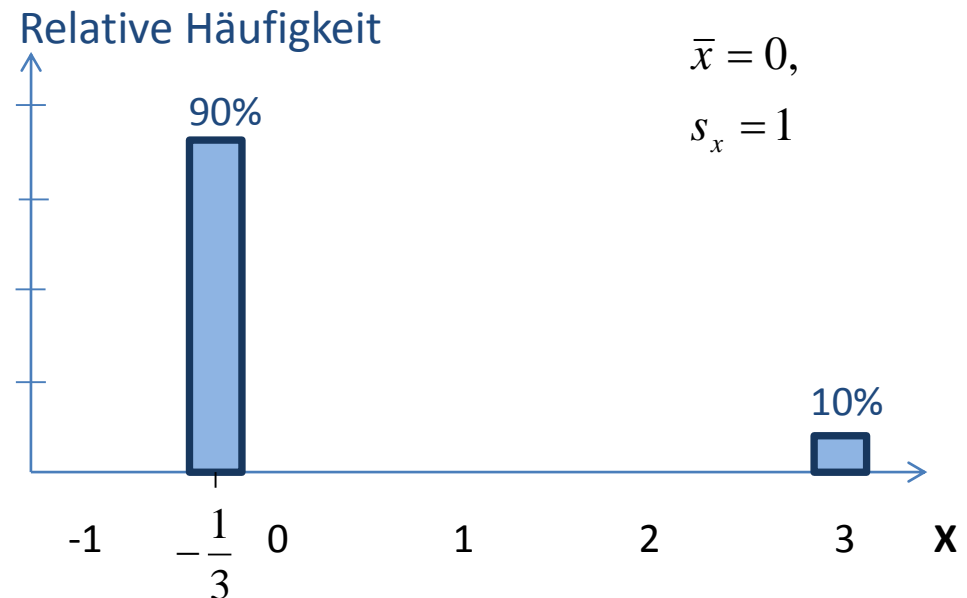
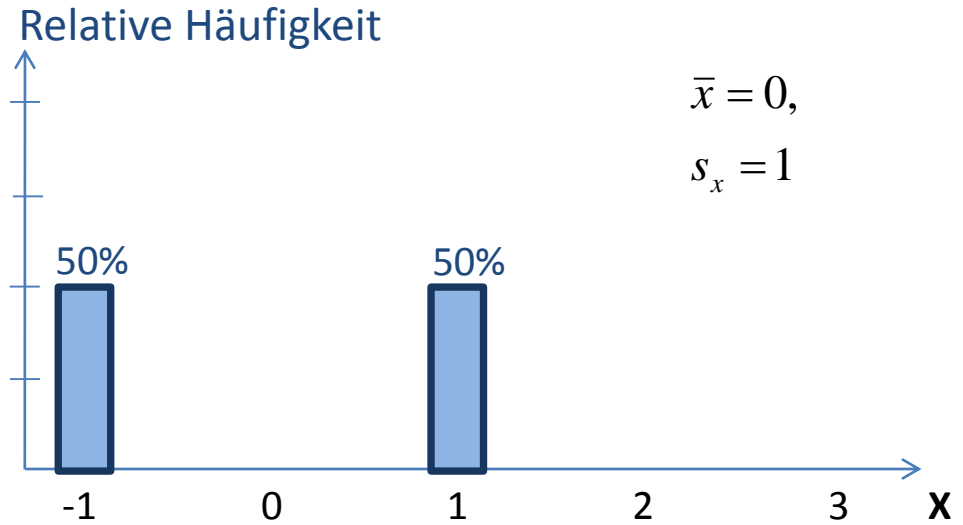


→ Mit Augenmaß kann man erkennen, dass der Wert nahe an 2 aber kleiner als 2 sein muss. Der exakte Wert ist 1.84 Minuten.

# Kennzahlen

Häufigkeitsverteilungen können den selben Mittelwert und die selbe Standardabweichung besitzen, obwohl sie deutlich unterschiedlich sind.

→ Ein paar Kennzahlen zu einem Merkmal enthalten weniger Informationen als die vollständige Häufigkeitsverteilung (z.B. Histogramm).



# Zusammenfassung wichtiger Formeln

## Konkreter Datensatz

Der Datensatz besteht aus  $n$  Werten  $w_1, \dots, w_n$ , darunter aber nur  $m$  verschiedene Ausprägungen  $x_1, \dots, x_m$  mit den relativen Häufigkeiten  $f_1, \dots, f_m$ .

Arithmetischer Mittelwert:

$$\bar{w} = \frac{1}{n} \cdot \sum_{j=1}^n w_j = \sum_{j=1}^m x_j \cdot f_j$$

Stichproben-Varianz:

$$s^2 = \frac{1}{n} \cdot \sum_{j=1}^n (w_j - \bar{w})^2 = \sum_{j=1}^m (x_j - \bar{w})^2 \cdot f_j$$

Standardabweichung:  $s = \sqrt{s^2}$

## Zufallsexperiment

Die Zufallsvariable  $X$  kann  $m$  verschiedene Werte  $x_1, \dots, x_m$  annehmen.

Erwartungswert:

$$\mu = E(X) = \sum_{j=1}^m x_j \cdot P(X = x_j)$$

Varianz:

$$V(X) = \sigma^2(X) = \sum_{j=1}^m (x_j - \mu)^2 \cdot P(X = x_j)$$

Standardabweichung:  $\sigma(X) = \sqrt{V(X)}$

## 2.2 Was Sie bisher gelernt haben sollten

- Für diskrete Merkmale den zur Fragestellung passenden **Typ von Mittelwert** bestimmen
- Für diskrete Merkmale **Häufigkeitsverteilungen** und **Mittelwert** bestimmen.
- **Kumulierte Häufigkeitsverteilung** bestimmen und interpretieren
- Weitere Kennzahlen eines Merkmals bestimmen und interpretieren: **Median, Quantile, Modus, Varianz, Standardabweichung**
- **Erwartungswerte** interpretieren.
- Für diskrete Zufallsvariablen **Wahrscheinlichkeitsverteilung** und **Erwartungswert** bestimmen.
- **Kumulierte Verteilungsfunktion** bestimmen und interpretieren
- Weitere Kennzahlen einer Zufalls-Variable bestimmen und interpretieren: **Median, Quantile, Modus, Varianz, Standardabweichung**