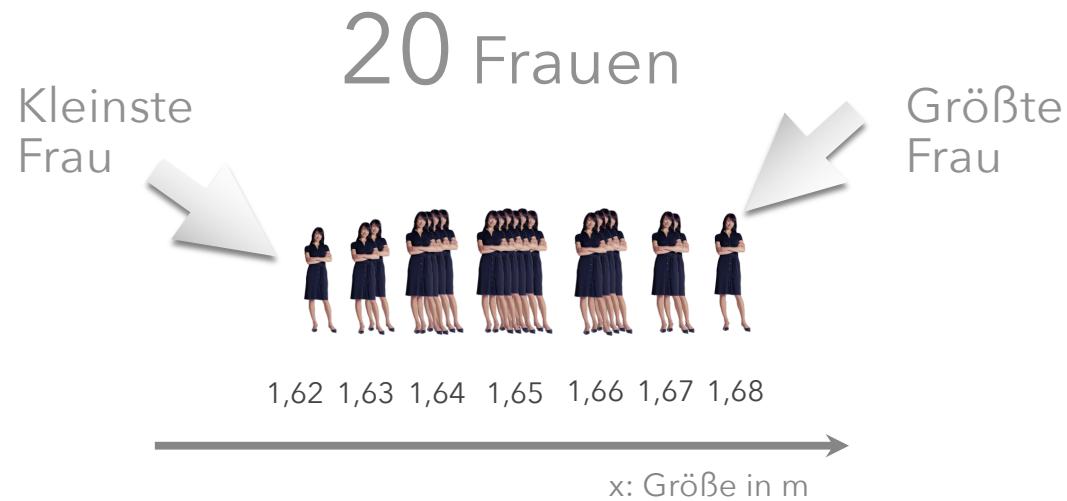


# Deskriptive Statistik

## Streuung

**Frage:** Was können wir noch aus den vorliegenden Informationen ermitteln?

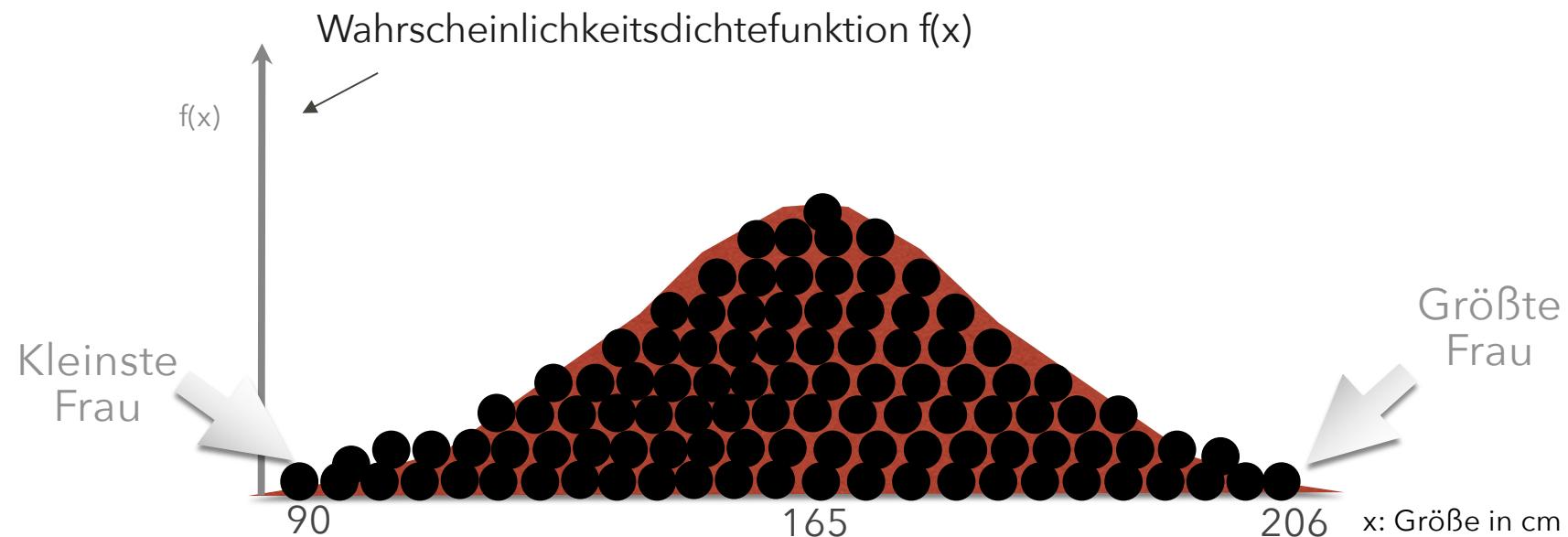


Stellen wir uns nun vor, wir hätten Daten von 1.000 Frauen erhoben...

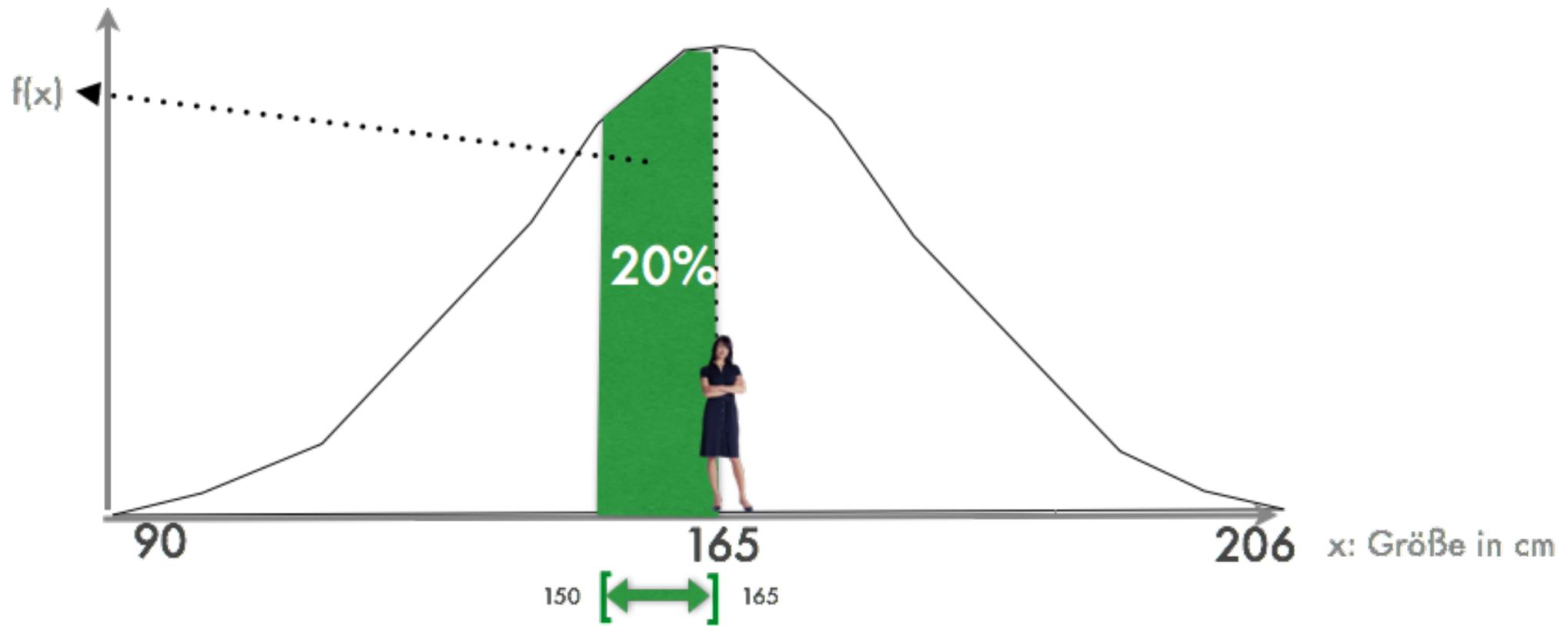
# 1.000 erwachsene deutsche Frauen



Punkte = 1.000 erwachsene  
deutsche Frauen



## Wahrscheinlichkeitsdichtefunktionen



Nun können wir (mit Hilfe der Dichtefunktion) bspw. die Frage beantworten, wie viele Frauen zwischen 150 cm und 165 cm groß sind:  
20% bzw. 200 Frauen

# Wahrscheinlichkeitsdichtefunktion:

Hinweise:

- Die **Wahrscheinlichkeitsdichtefunktion** oder „Dichte“ (engl. probability density function) ist ein Hilfsmittel zur Beschreibung einer **stetigen** Wahrscheinlichkeitsverteilung.

# Wahrscheinlichkeitsdichtefunktion:

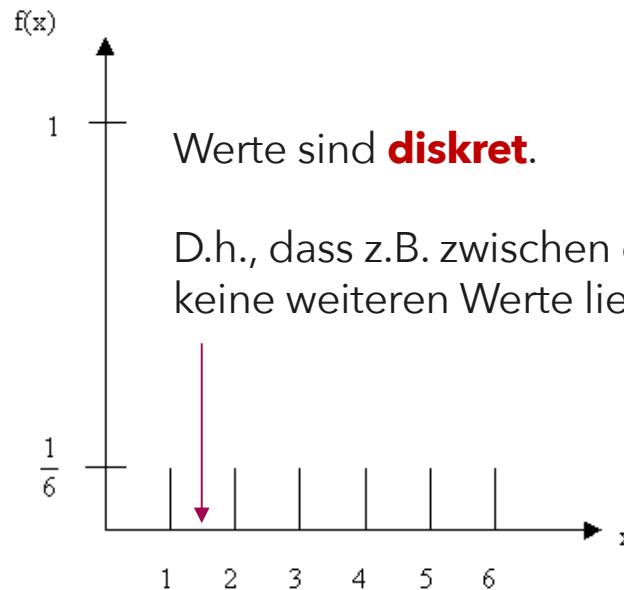
Hinweise:

- **Stetig** sind solche Merkmale, die theoretisch unendlich viele Ausprägungen aufweisen können (z.B. Körpergröße, Länge, Gewicht, Zeit).
- Das Gegenteil von stetig ist **diskret**.
- **Diskret** sind solche Merkmale, die nur endlich viele Ausprägungen annehmen können. Insbesondere sind alle Merkmale diskret, deren Werte man durch Zählen ermitteln kann (z.B. Seiten eines Würfels)

# Wahrscheinlichkeitsfunktion:



$$\begin{aligned}f(1) &= P(X=1) = \frac{1}{6} \\f(2) &= P(X=2) = \frac{1}{6} \\f(3) &= P(X=3) = \frac{1}{6} \\f(4) &= P(X=4) = \frac{1}{6} \\f(5) &= P(X=5) = \frac{1}{6} \\f(6) &= P(X=6) = \frac{1}{6}\end{aligned}$$



- Bei diskreten Werten können Wahrscheinlichkeitsfunktionen ermittelt werden
- Bsp.: Wahrscheinlichkeitsfunktion eines Würfels (Merkmal: Augenzahl)

# Wahrscheinlichkeitsfunktion:

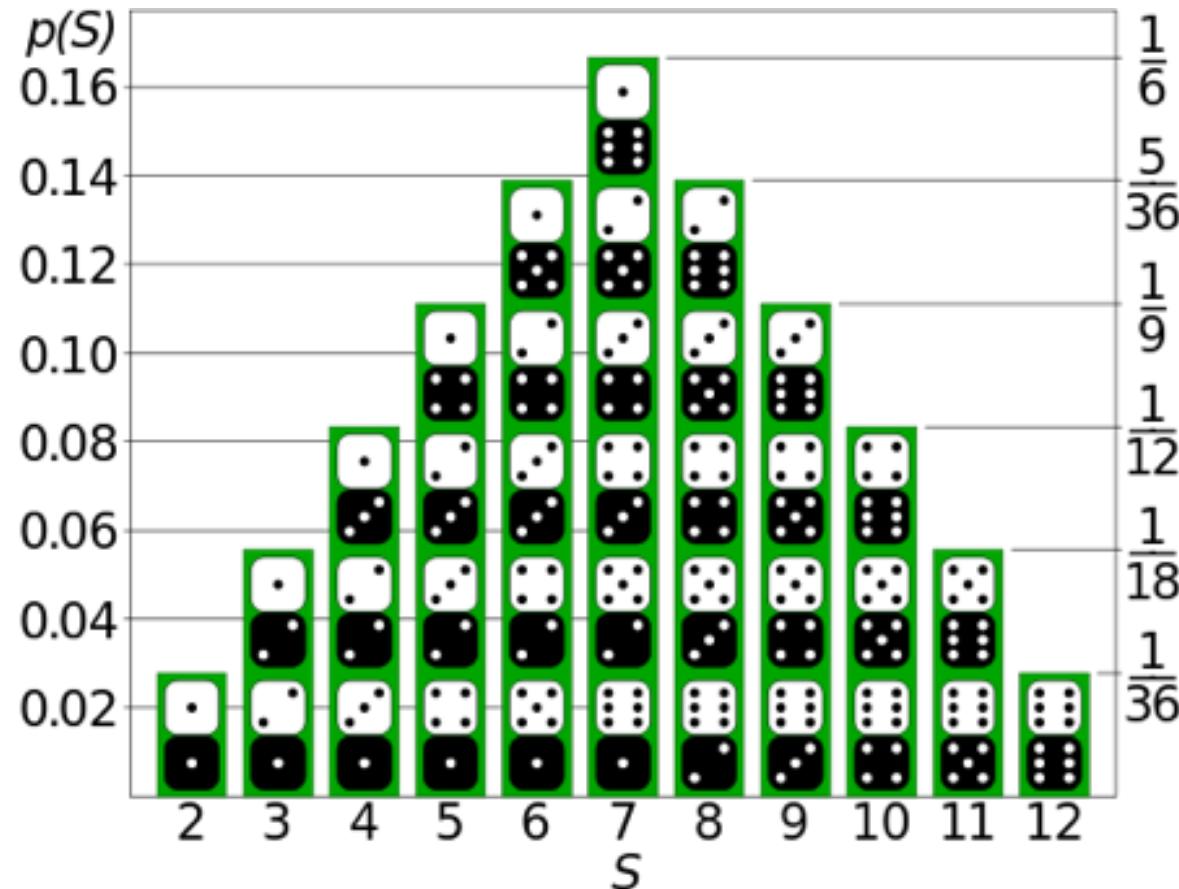
Frage: wie sieht die Wahrscheinlichkeitsfunktion von zwei Würfelwürfen mit zwei Würfeln aus (als Säulendiagramm dargestellt)?:

Die Augen sind das **Merkmal**.

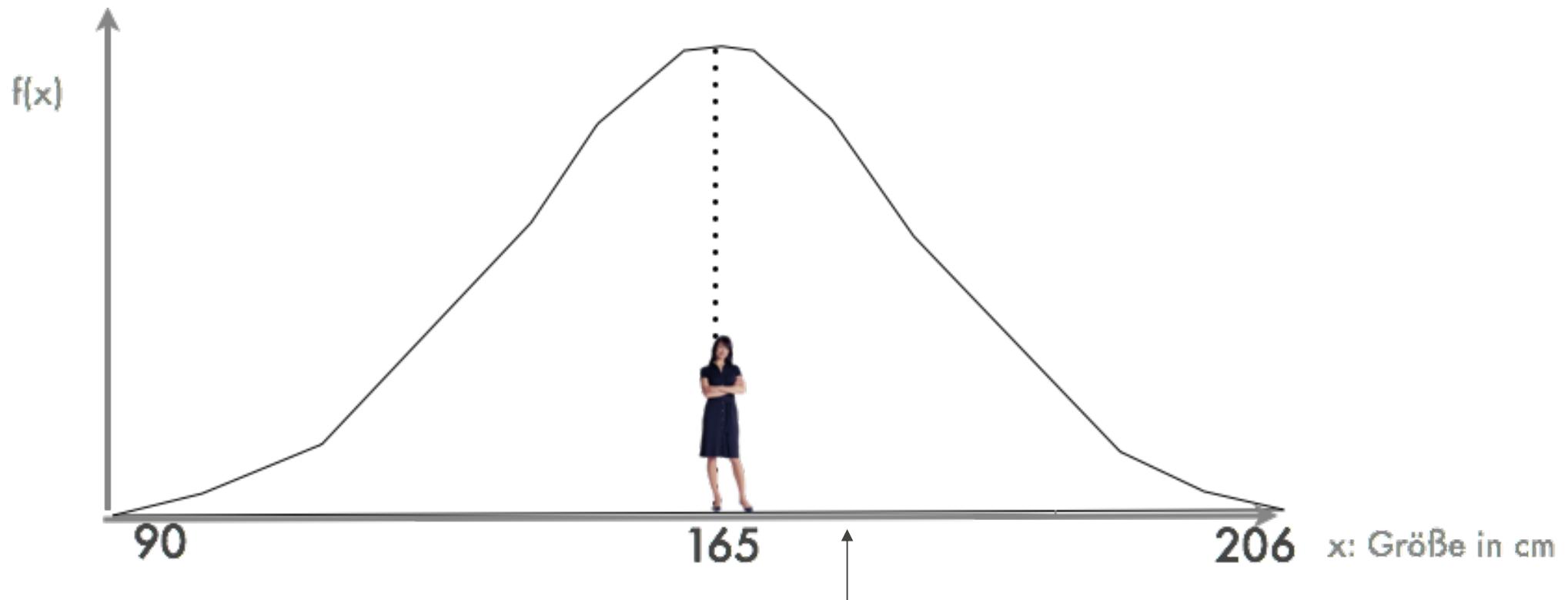
Deren Anzahl ist die **Merkmalsausprägung**.

Die Funktion gibt die Wahrscheinlichkeit des Auftretens einer bestimmten Ausprägung an.

# Wahrscheinlichkeitsfunktion:

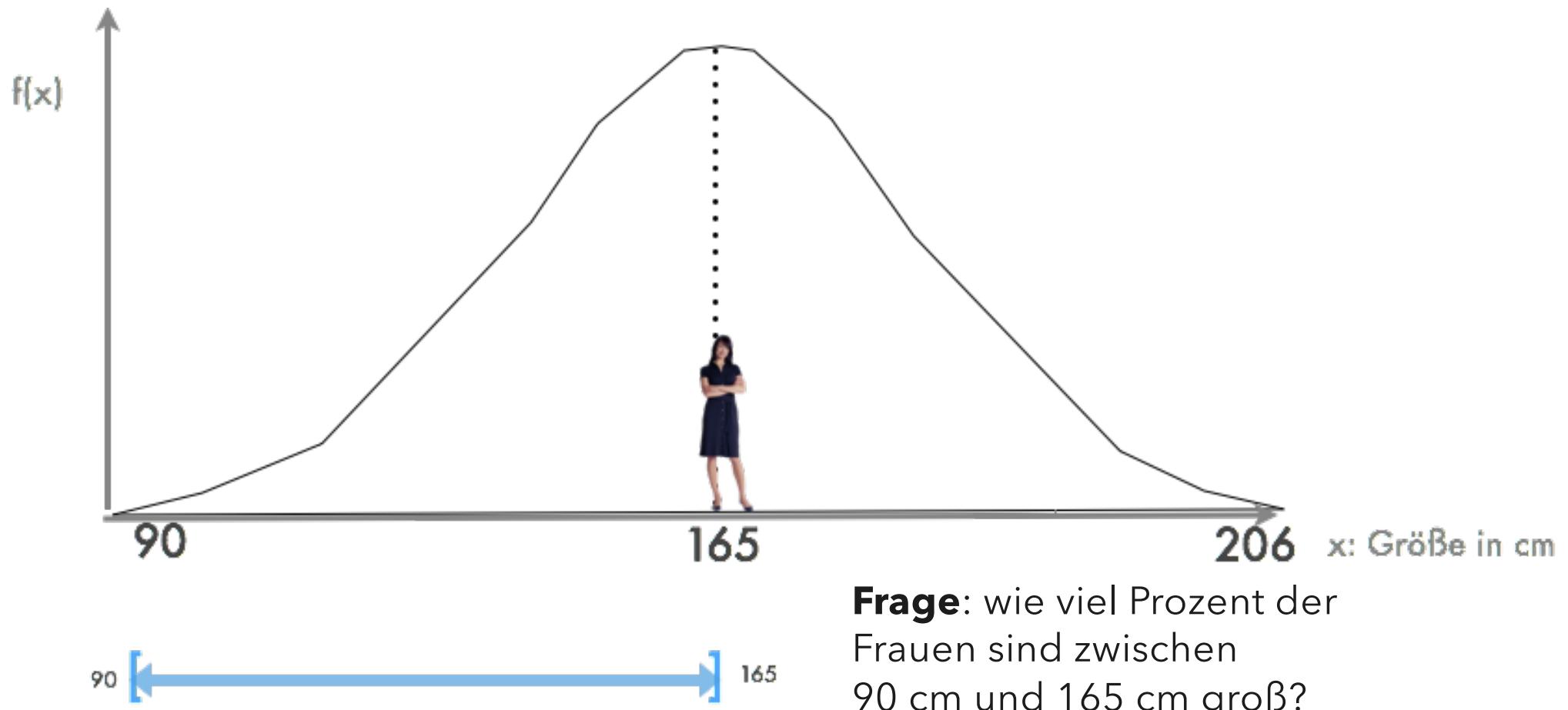


# Wahrscheinlichkeitsdichtefunktion:

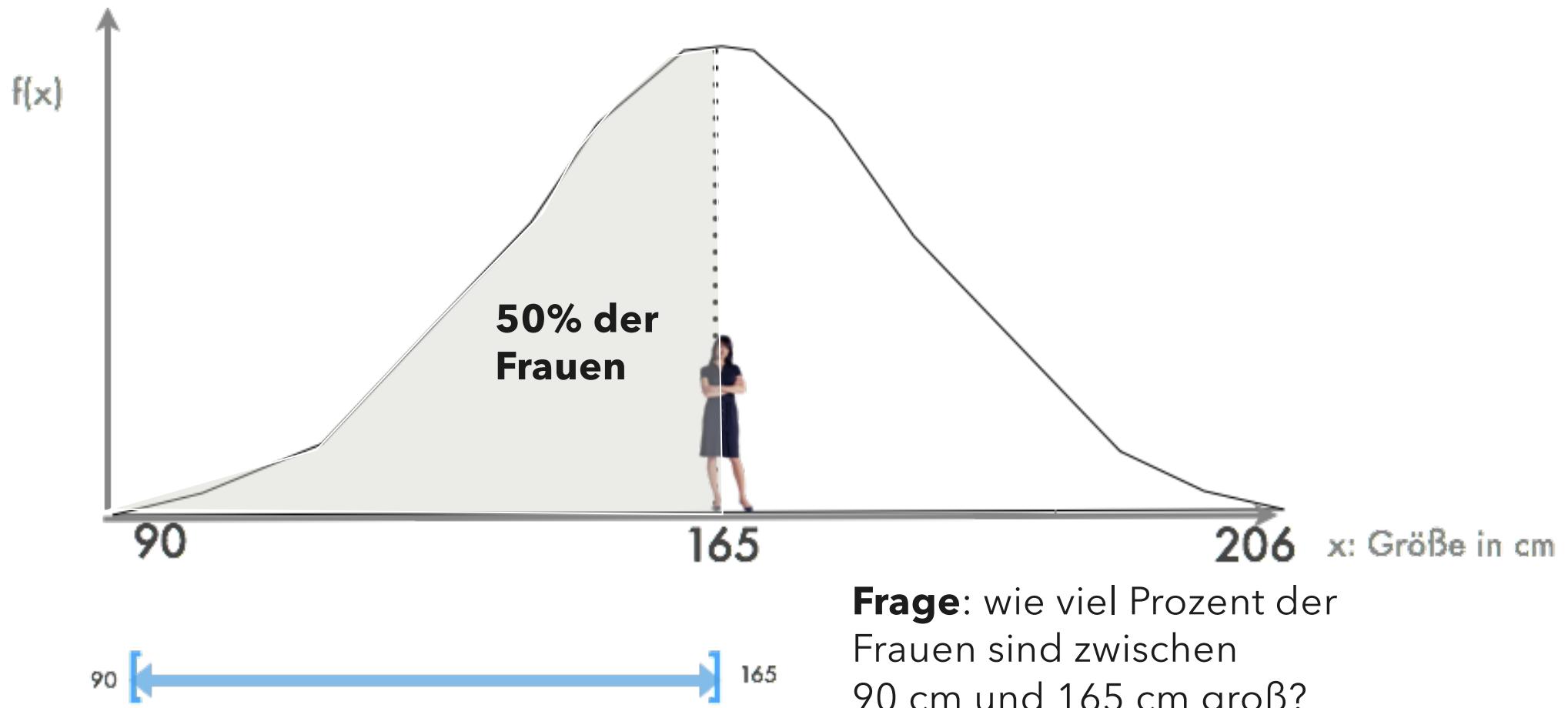


Körpergröße ist **stetig** (es kann theoretisch jeder beliebige Wert vorkommen – also theoretisch unendlich viele)

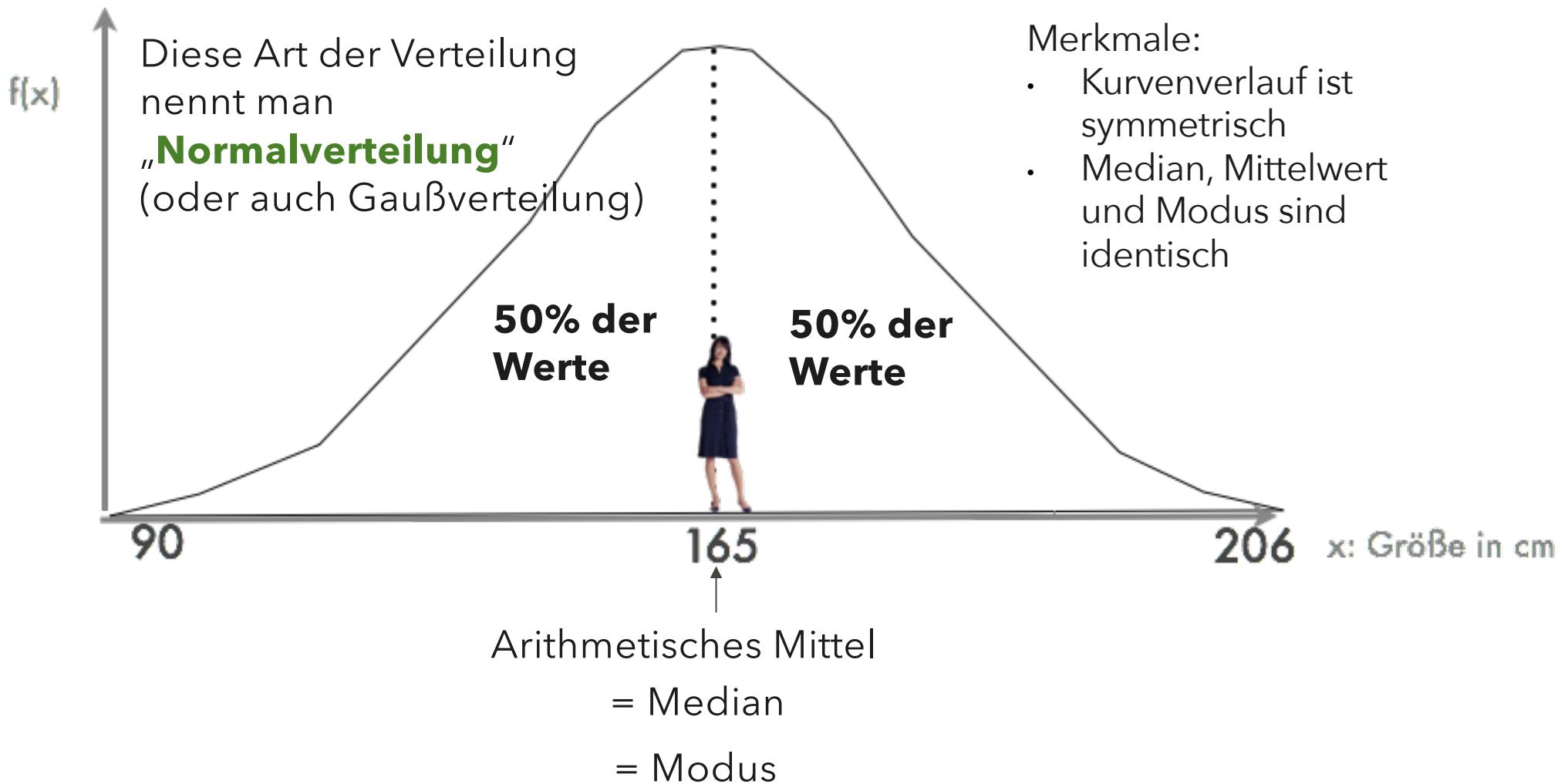
# Wahrscheinlichkeitsdichtefunktion:



# Wahrscheinlichkeitsdichtefunktion:



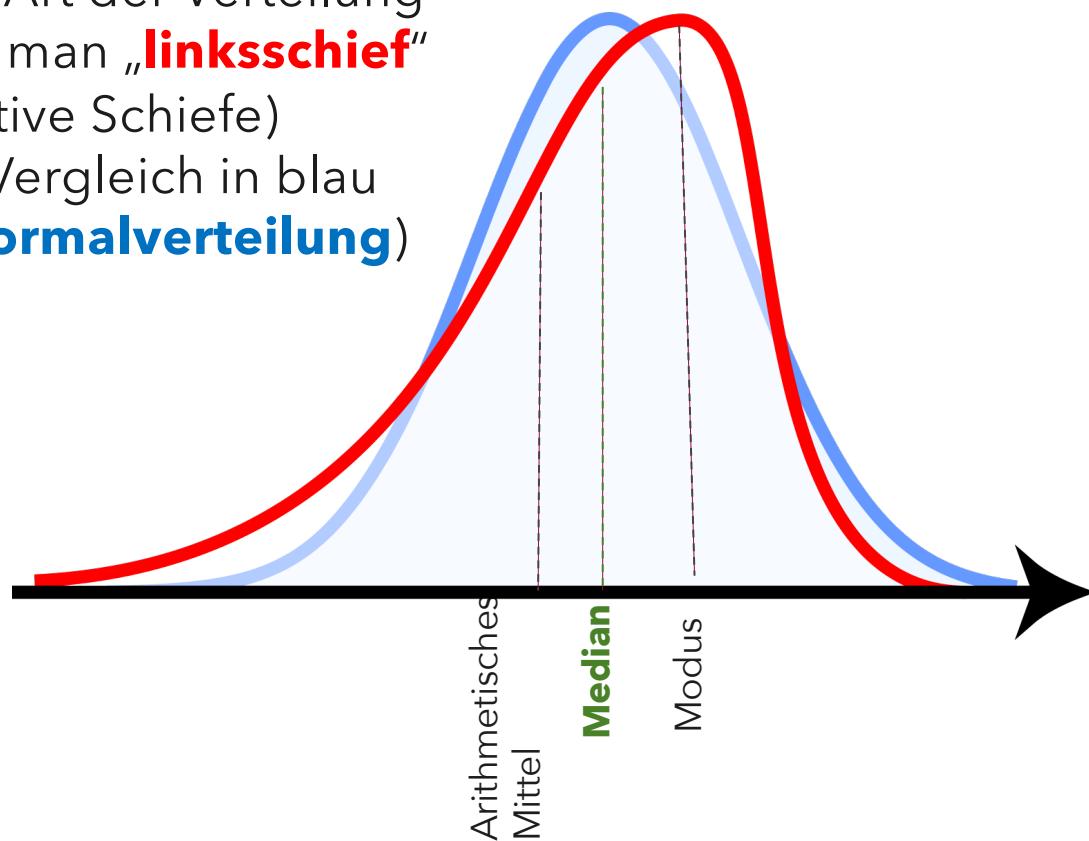
# Normalverteilung



# Linksschiefe Verteilung

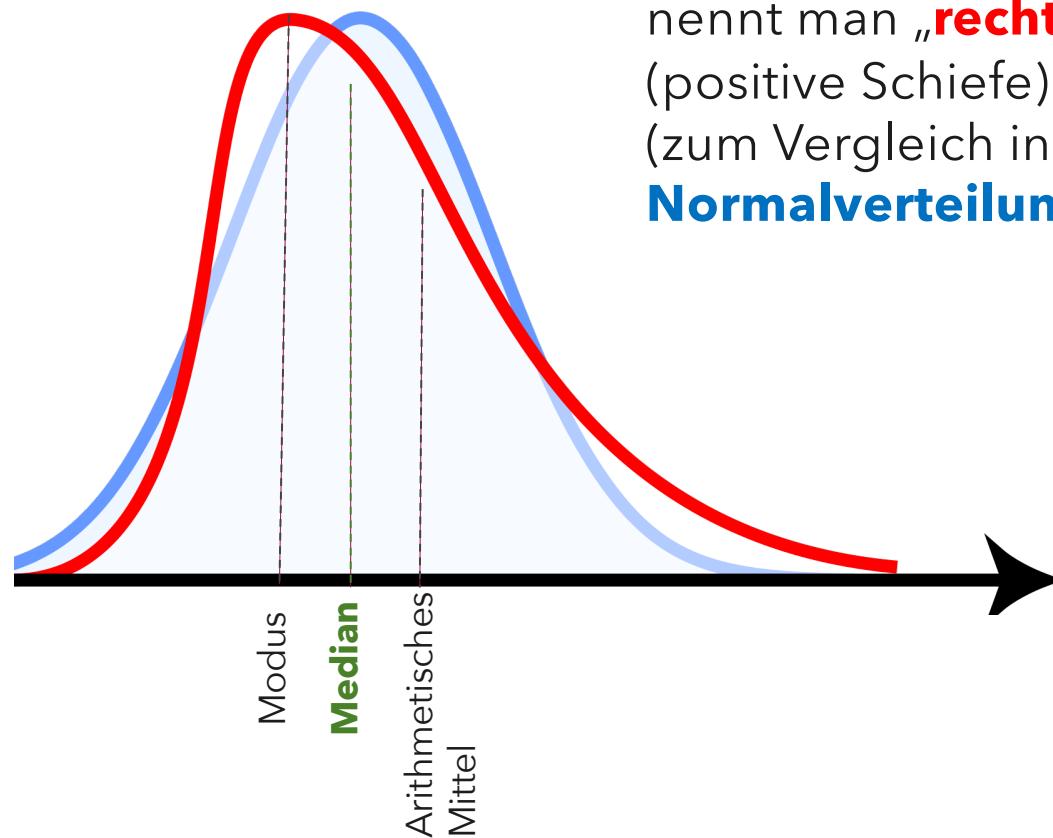
≠ Normalverteilung

Diese Art der Verteilung  
nennt man „**linksschief**“  
(negative Schiefe)  
(zum Vergleich in blau  
die **Normalverteilung**)



# Rechtsschiefe Verteilung

≠ Normalverteilung



Diese Art der Verteilung  
nennt man „**rechtsschief**“  
(positive Schiefe)  
(zum Vergleich in blau die  
**Normalverteilung**)

# Logik zur Auswahl der Lagemaße

Deskriptive Statistik		
Schritte zur Ermittlung der passenden Kennzahlen		Statistische Kennzahl
(1) Welches Skalenniveau liegt vor?	(2) Welche Verteilung liegt vor?	--> Lagemaße
Nominal	<i>Verteilung nicht vorhanden</i>	Modus
Ordinal	<i>Verteilung nicht relevant</i>	Modus Median
Metrisch	Fall 1: Daten sind <u>nicht</u> normalverteilt	Modus Median
	Fall 2: Daten sind normalverteilt	Modus Median Mittelwert

# Wiederholungsfragen

**Bitte geben Sie jeweils an, ob die Aussage richtig oder falsch ist:**

Markieren Sie dafür das Kästchen vor der Ziffer: Richtig Aussage  / Falsche Aussage:  .

1.  Bei nominalskalierten Variablen ist es sinnvoll, einen Mittelwert zu berechnen.
2.  Der Modalwert ist der am häufigsten vorkommende Wert.
3.  Bei metrischen Merkmalen können wir nur den Modalwert berechnen.
4.  Diskret sind solche Merkmale, die nur endlich viele Ausprägungen annehmen können.
5.  Stetig sind solche Merkmale, die überabzählbar viele Ausprägungen aufweisen können (z.B. Länge, Gewicht, Zeit).
6.  Die Normalverteilung ist ein wichtiger Typ stetiger Wahrscheinlichkeitsverteilungen.
7.  Der Median kann auch dann berechnet werden, wenn bei metrischen Merkmalen keine Normalverteilung vorliegt
8.  Das arithmetische Mittel kann auch bei ordinalen Merkmalen berechnet werden.

# Deskriptive Statistik

Streuungsmaße

# Streuungsmaße: Spannweite

- Das einfachste Streuungsmaß ist die Spannweite.
- Sie wird als Differenz zwischen größtem und kleinstem Wert berechnet.
- Spannweite = Maximum - Minimum
- Unser Beispiel: Spannweite =  $1,68 \text{ m} - 1,62 \text{ m} = 0,06 \text{ m}$
- **Negativ:** Basiert nur auf den beiden Extremwerten der Erhebung. Sagt nichts über dazwischen liegende Werte aus.

Merkmalsausprägung	Häufigkeit	Relative Häufigkeit
1,62	1	5 %
1,63	2	10 %
1,64	4	20 %
1,65	6	30 %
1,66	4	20 %
1,67	2	10 %
1,68	1	5 %
	20	100 %

# Streuungsmaße: Varianz

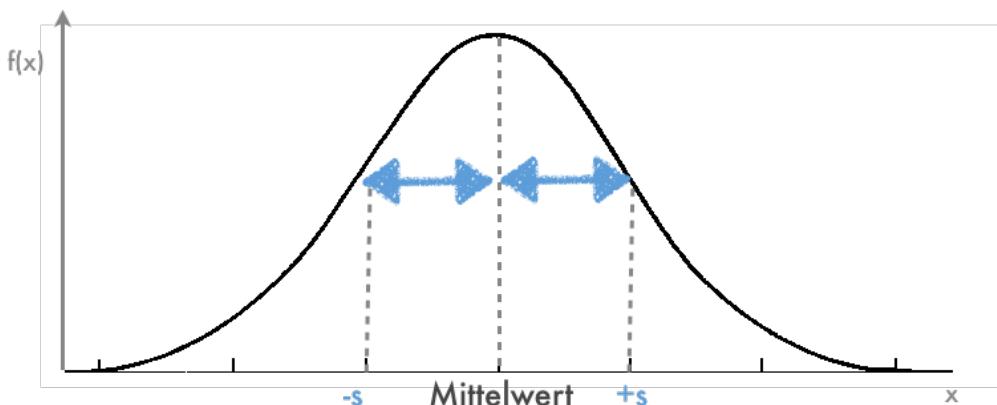
- Die empirische Varianz ( $s^2$ ), auch Stichprobenvarianz genannt, ist ein Streuungsmaß, welches die Verteilung von Werten um den Mittelwert kennzeichnet.
- Sie misst, wie dicht die Werte einer Häufigkeitsverteilung um den Mittelwert streuen.

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

- Varianz = Summe der quadrierten Abweichungen aller Messwerte vom arithmetischen Mittel geteilt durch die Anzahl der Freiheitsgrade n-1 (diese Thematik wird zu einem späteren Zeitpunkt behandelt).
- **Negativ**: der Wert ist schwer interpretierbar.

# Streuungsmaße: Standardabweichung

- Die Standardabweichung ( $s$ ) ist ein Maß für die Streubreite der Werte um dessen Mittelwert und wird als Wurzel der Varianz berechnet.
- Sie ist die durchschnittliche Entfernung aller gemessenen Ausprägungen eines Merkmals vom Mittelwert (nach „oben“  $+s$  und „unten“  $-s$ ).



# Streuungsmaße: Standardabweichung

- Man errechnet die Summe der quadratischen Abweichungen aller Messwerte vom Mittelwert, teilt diese durch die um 1 verminderte Fallzahl n (die „Freiheitsgrade“ – dies wird in anderen Folien noch erläutert) und zieht hieraus die Wurzel:

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

- Je stärker einzelne Messwerte von ihrem Mittelwert abweichen, desto größer wird die Standardabweichung.

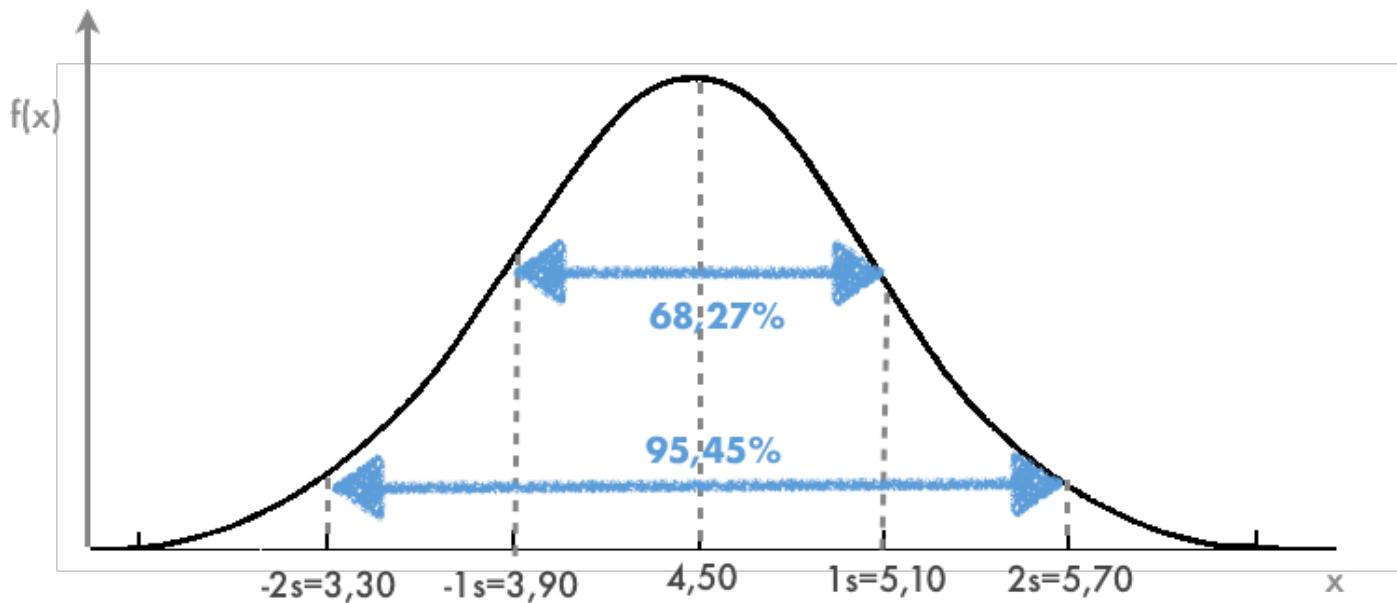
# **Streuungsmaße:** Standardabweichung

- Der Vorteil der Standardabweichung gegenüber der Varianz ist, dass die Standardabweichung die gleiche Messeinheit wie die ursprüngliche Variable hat.
- **Positiv:** Werte können gut interpretiert werden.

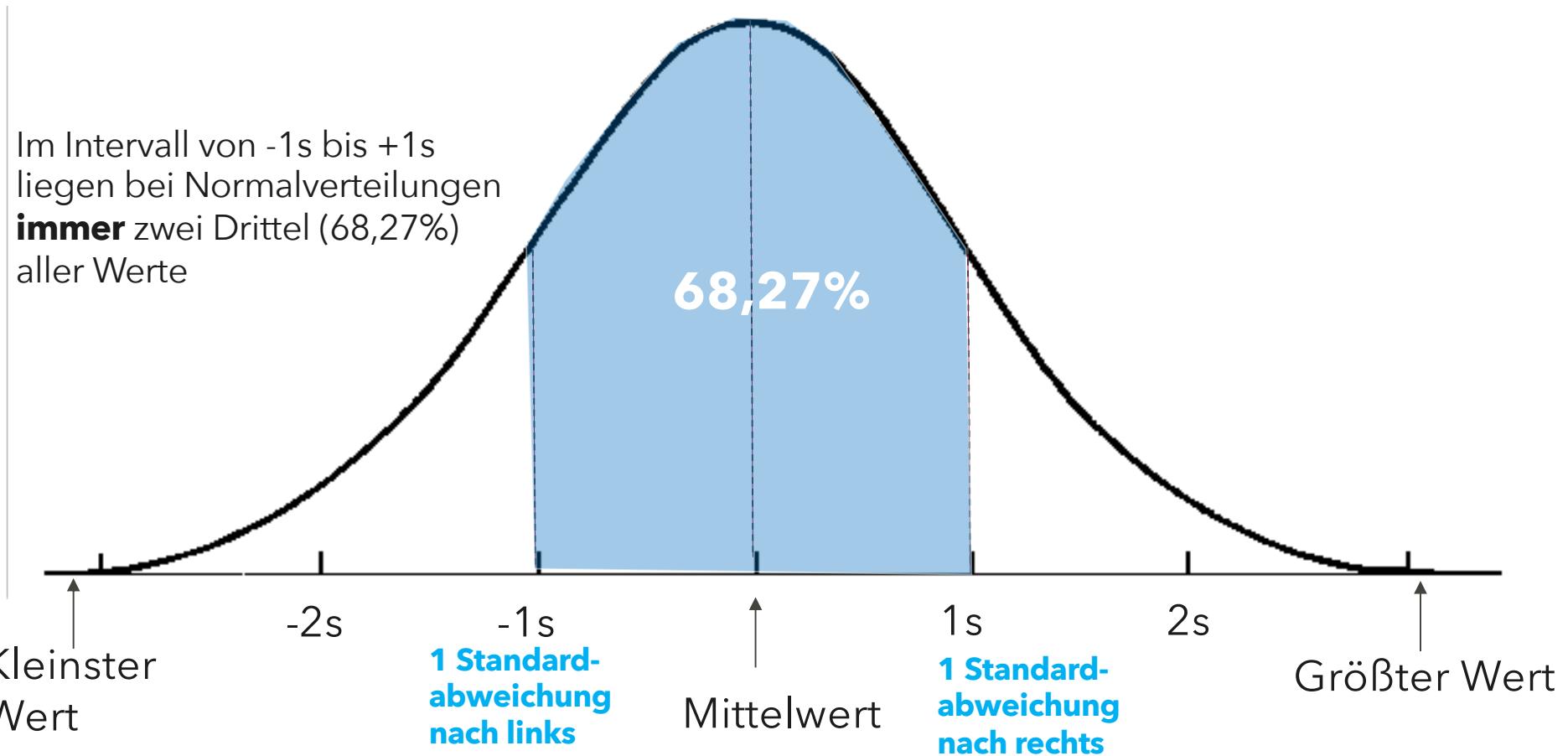
# **Streuungsmaße:** Standardabweichung

- Beispiel: Gefragt wurden 1.000 Personen, wie viel Euro sie im Schnitt ausgeben, wenn sie mittags Essen gehen (Normalverteilung liegt vor)
- Der Mittelwert liegt bei 4,50 Euro, die Standardabweichung bei  $s = 0,60$ .
- Das heißt, dass die durchschnittliche Entfernung aller Antworten zum Mittelwert 60 Cent beträgt.
- Aufgrund der beschriebenen Faustformel lässt sich ableiten, dass rund 68 Prozent aller Befragten der Stichprobe mittags zwischen 3,90 Euro und 5,10 Euro ausgeben ( $4,50 \pm 0,60$  Euro). Rund 95 Prozent geben zwischen 3,30 Euro und 5,70 Euro aus ( $4,50 \pm 2$  mal  $0,60$  Euro).

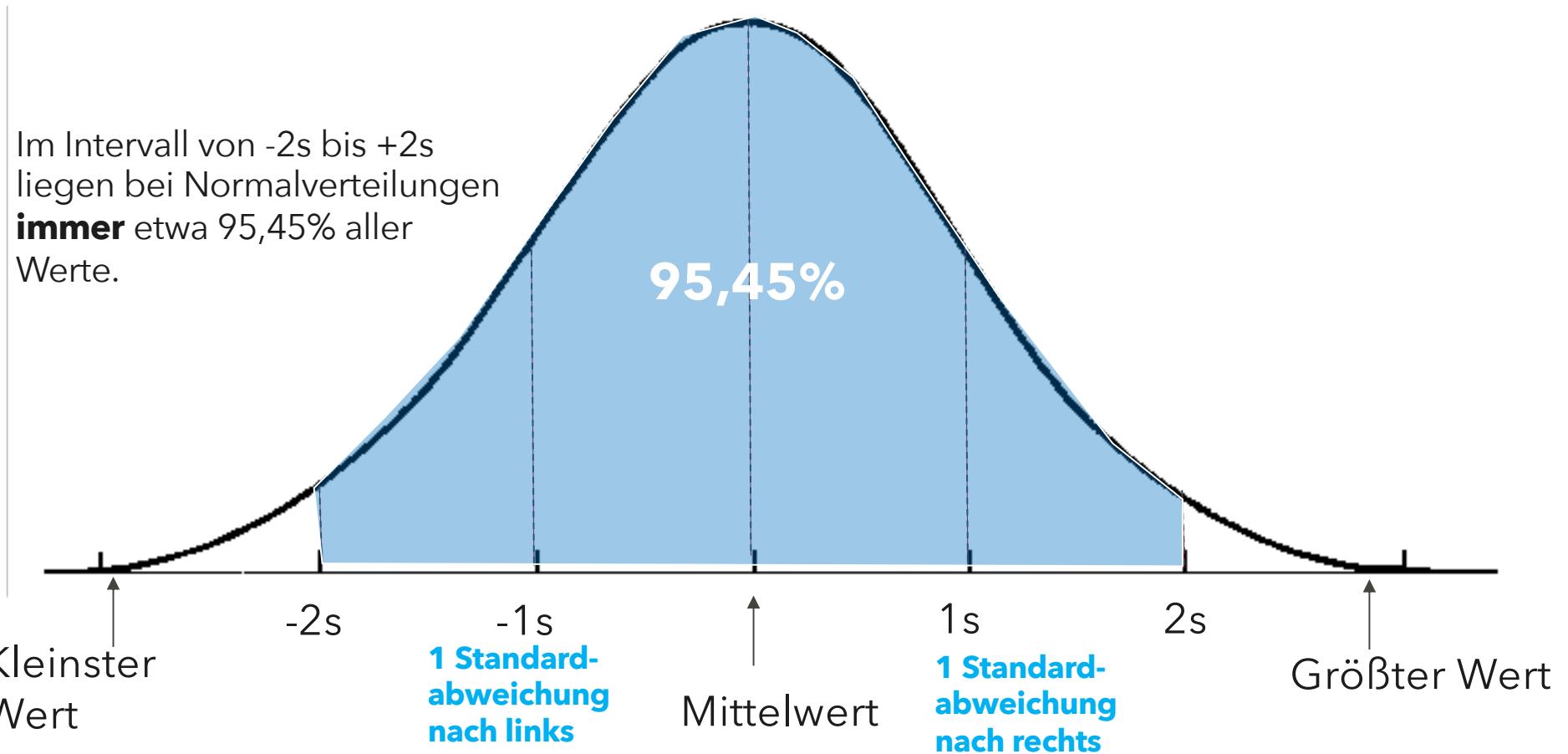
# Streuungsmaße: Standardabweichung



# Streuungsmaße: Standardabweichung



# Streuungsmaße: Standardabweichung



# Streuungsmaße: Standardabweichung

- Für den Vergleich zweier Standardabweichungen kann der **Variationskoeffizient** (VarK) genutzt werden.
- Damit lassen sich bei Verteilungen mit unterschiedlichen Mittelwerten die Variabilität der Messwerte beurteilen.
- Der VarK ist definiert als die relative Standardabweichung, d.h. die Standardabweichung dividiert durch den Mittelwert.
- $\text{VarK} = (\text{Standardabweichung} / \text{Mittelwert})$
- Je stärker einzelne Messwerte von ihrem Mittelwert abweichen, desto größer wird die Standardabweichung.

# Streuungsmaße: Standardabweichung

- Beispiel: Krankenhausstation
- Kosten pro Patient auf Station 1(chirurg.):
  - Mittelwert = 500 Euro ( $s=75$ )
- Kosten pro Patient auf Station 2(intensiv):
  - Mittelwert = 1.000 Euro ( $s=100$ )
- $\text{VarK}(1) = 75/500 = 0,15$  /  $\text{VarK}(2) = 100/1000 = 0,1$
- Interpretation: die Werte von Station 1 variieren stärker als die von Station 2.
- Quelle:PflegeWiki, <http://www.pflegewiki.de/wiki/Variationskoeffizient>

# Streuungsmaße: Standardabweichung

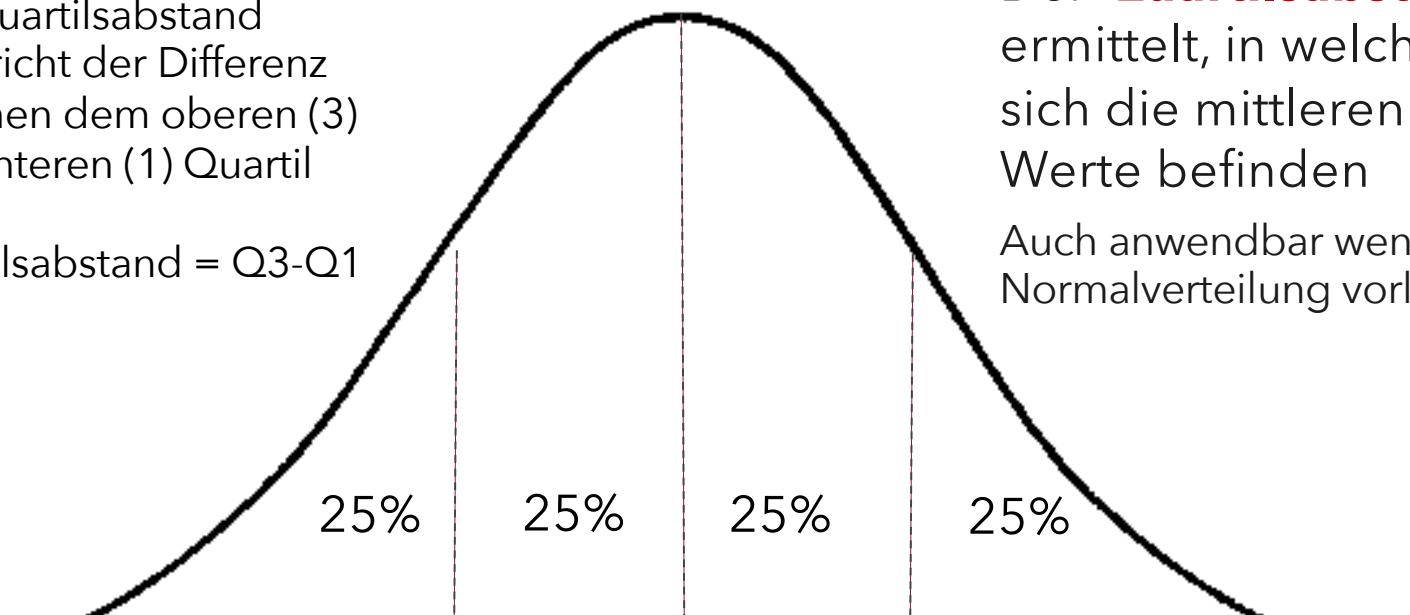
- **Übung:**
- 1.000 Personen wurden befragt, wie hoch ihre monatliche Handyrechnung ist:
- Die gewonnenen Daten sind normalverteilt.
- Der Mittelwert liegt bei 40 Euro
- Die Standardabweichung liegt bei 5 Euro (d.h., dass die durchschnittliche Entfernung aller Antworten zum Mittelwert 5 Euro beträgt)
- Frage: welche Werte haben  $+s_1$  und  $-s_1$ ? Wieviel Prozent der Personen befinden sich zwischen  $-s_1$  und  $+s_1$

# Streuungsmaße: Quartilsabstand

Der Quartilsabstand entspricht der Differenz zwischen dem oberen (3) und unteren (1) Quartil

$$\text{Quartilsabstand} = Q_3 - Q_1$$

Der **Quartilsabstand** ermittelt, in welchem Bereich sich die mittleren 50% der Werte befinden  
Auch anwendbar wenn keine Normalverteilung vorliegt



Enthält:

**25%**

**50%**

**75%**

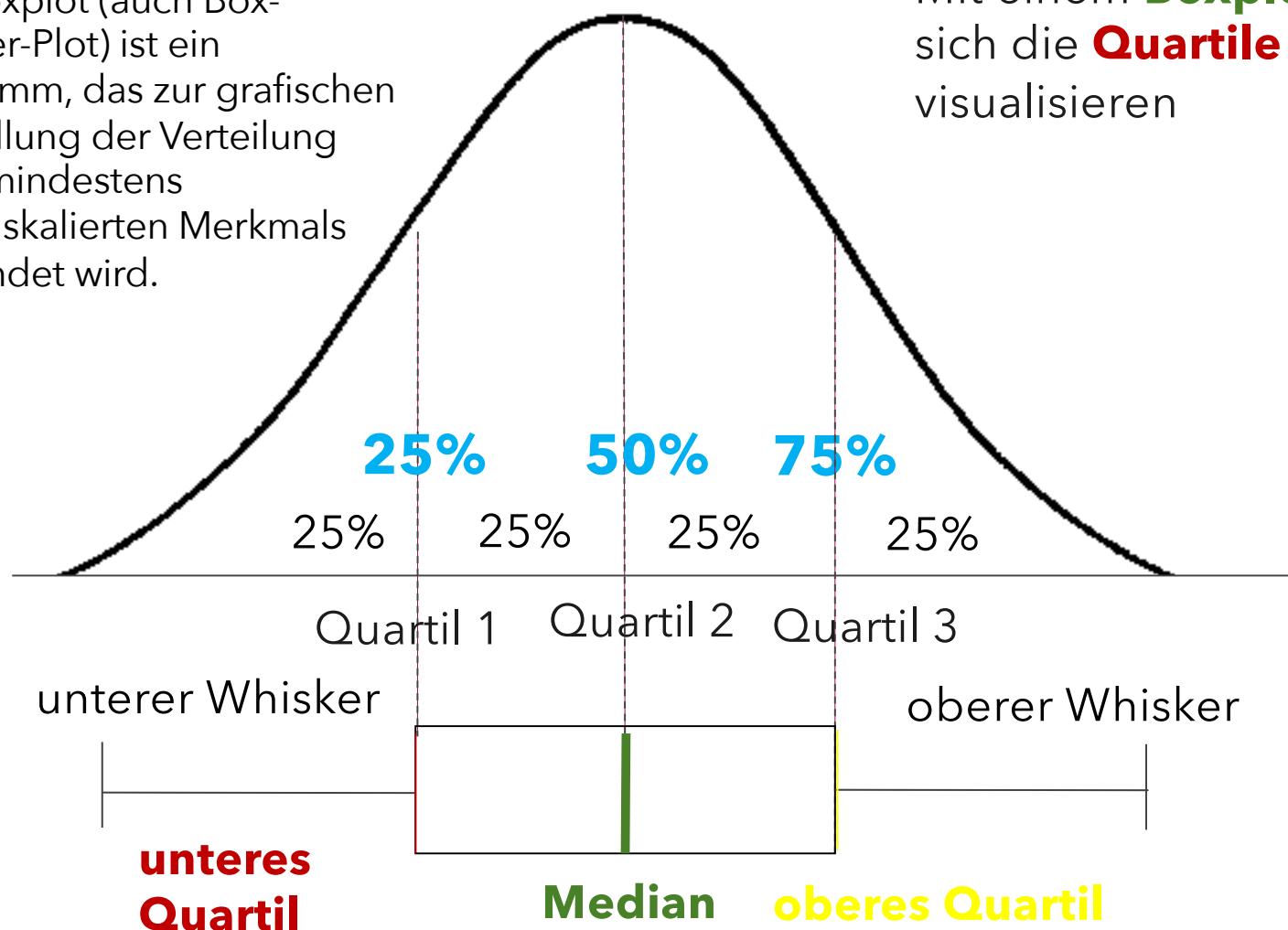
... aller Werte

(Median)

# Streuungsmaße: Quartile & Boxplot

Der Boxplot (auch Box-Whisker-Plot) ist ein Diagramm, das zur grafischen Darstellung der Verteilung eines mindestens ordinalskalierten Merkmals verwendet wird.

Mit einem **Boxplot** lassen sich die **Quartile** gut visualisieren



# Logik zur Auswahl der Streuungsmaße

Deskriptive Statistik		
Schritte zur Ermittlung der passenden Kennzahlen		Statistische Kennzahlen
(1) Skalenniveau	(2) Verteilung	Streuungsmaße
Nominal	<i>Verteilung nicht vorhanden</i>	<i>(es gibt keine Streuung)</i>
Ordinal	<i>Verteilung nicht relevant</i>	Spannweite, Quartilsabstand
Metrisch	Fall 1: Daten sind <u>nicht</u> normalverteilt	Spannweite, Quartilsabstand
	Fall 2: Daten sind normalverteilt	Spannweite, Quartilsabstand, Standardabweichung

# Logik zur Auswahl der Kennzahlen

Deskriptive Statistik			
Schritte zur Ermittlung der passenden Kennzahlen		Statistische Kennzahlen	
(1) Skalenniveau	(2) Verteilung	Lagemaße	Streuungsmaße
Nominal	<i>Verteilung nicht vorhanden</i>	Modus	-
Ordinal	<i>Verteilung nicht relevant</i>	Modus Median	Spannweite Quartilsabstand
Metrisch	Fall 1: Daten sind <u>nicht</u> normalverteilt	Modus Median	Spannweite Quartilsabstand
	Fall 2: Daten sind normalverteilt	Modus Median Mittelwert	Spannweite Quartilsabstand Standardabweichung

# Deskriptive Kennzahlen: 3D-Drucker

- 8 verschiedene 3D-Drucker erzielen pro Stunde jeweils die folgenden Produktionsmengen: 2, 4, 2, 3, 5, 2, 3, 2
- Wir möchten die Drucker nun anhand verschiedener Kennzahlen untersuchen und berechnen den **Modus**, den **Median**, das **arithmetisches Mittel**, die **Spannweite**, der **Quartilsabstand**, die empirische **Varianz** und die **Standardabweichung**



<https://docs.google.com/spreadsheets/d/1S57IvW0M4vzNRRTCK-BYh4-Rzza6OMdOUM32hwIGsoU/edit?usp=sharing>