4. EIN METRISCHES MERKMAL

Beispiel: Nitratbelastung in NÖ Trinkwasser
(Quelle, NöSiWAG/WWF,1998)

Angaben (in mg/l - Grenzwert ist 50mg/l) von 526 Meßstellen

23	23	32	32	32	12	12	12	12	12	12	12	23	12	12	12	12
23	23	23	20	20	20	20	20	20	20	4	23	12	12	20	20	12
12	4	1	1	1	1	1	1	9	9	9	9	1	1	1	9	9	9	1	1	1	1	1	1
1	1	1	1	1	1	1	9	9	9	9	9	9	9	9	47	13	14	14	13	13	47
10	47	47	10	47	47	47	47	47	13	47	47	47	47	47	13	15
15	15	47	13	12	15	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
7	7	7	7	7	7	26	18	19	19	19	19	26	26	26	26	26	18	18
18	14	26	26	26	26	26	26	19	26	19	24	24	24	24	1	1	1
9	19	14	14	24	24	24	24	24	1	26	19	19	19	19	19 ...						

Information aus Rohdaten ist unübersichtlich

daher: Methoden um Daten zusammenzufassen
Information übersichtlich darstellen

wesentlicher Begriff:

Häufigkeitsverteilung (oder kurz **Verteilung**):
"welche Zahlen kommen wie oft vor"

- **numerische Methoden:**
  wie kann Verteilung mit (wenigen) Maßzahlen gut und sinnvoll beschrieben werden

- **grafische Methoden:**
  wie kann Verteilung gut dargestellt werden

HÄUFIGKEITSVERTEILUNG:

erhält man durch Abzählen, wie oft eine bestimmte Zahl vorkommt

Bestimmen und **Angeben der Häufigkeiten aller möglicher Werte** (Ausprägungen), die eine Variable annimmt (wie bei kategorialen Daten)

Bsp: Ergebnisse eines Tests bei 20 Studenten

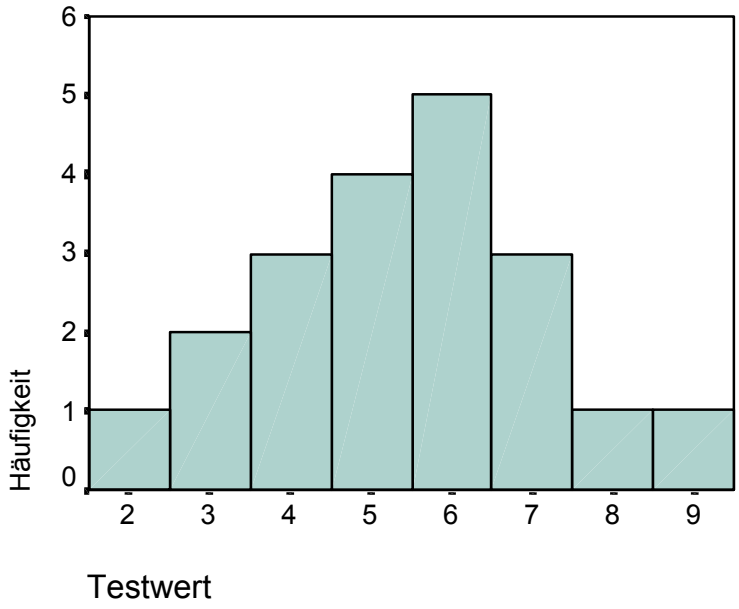
6, 3, 7, 5, 6, 4, 4, 6, 7, 3, 5, 9, 6, 4, 2, 7, 5, 5, 8, 6

Häufigkeitstabelle:

Test-Wert

	Häufigkeit	relative Häufigkeit
2	1	,05
3	2	,10
4	3	,15
5	4	,20
6	5	,25
7	3	,15
8	1	,05
9	1	,05
Gesamt	20	1,00

Histogramm:



HISTOGRAMME:

verwendet man zur Darstellung (stetiger) metrischer Daten

wichtige Idee:
Fläche der Balken ist proportional zu Häufigkeiten
(Höhe und Breite sind nicht egal)

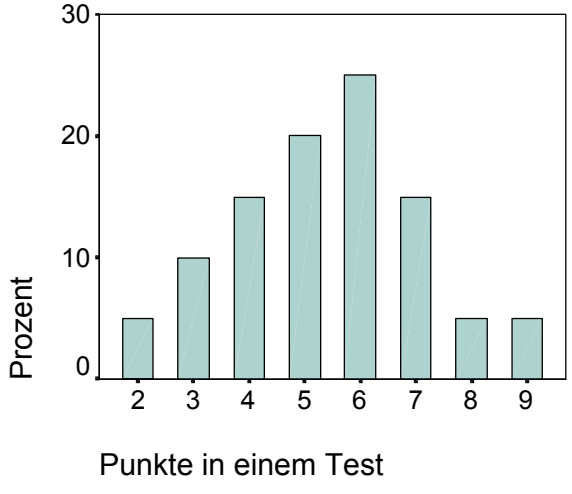
bei Balkendiagrammen:
Höhe der Balken ist proportional zu Häufigkeiten
(Breite ist egal)

dahinter liegt Unterscheidung: diskret - stetig

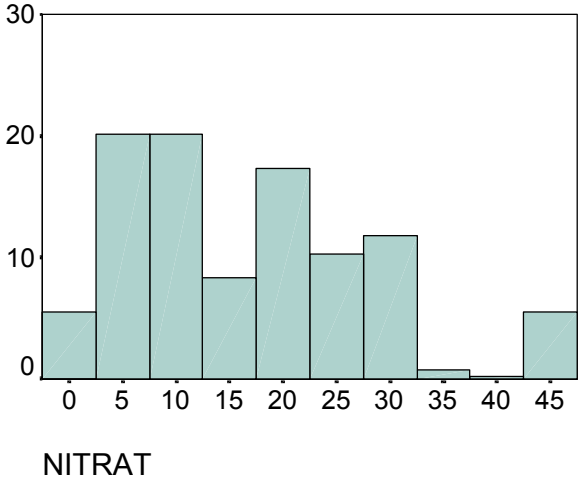
diskrete Daten: wenn Messungen zahlenmäßig nur auf bestimmte Punkte der Zahlengerade beschränkt sind
(z.B. Messung mit ganzen Zahlen: 0,1,2,3,...)

stetige Daten: wenn Messungen im Prinzip beliebig genau erfolgen können, d.h. alle Punkte (in einem bestimmten Bereich) der Zahlengerade kommen in Frage
(Messung mit reellen Zahlen: 3,456 4,78904 etc.)

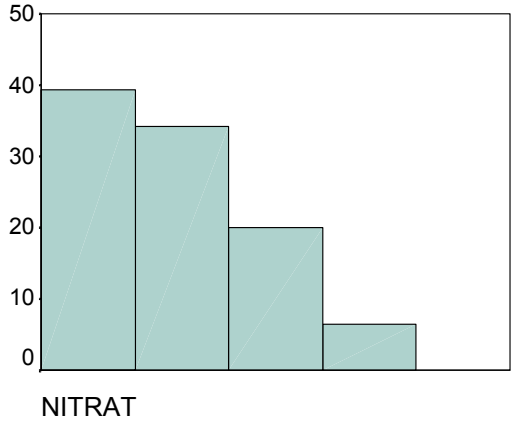
BALKENDIAGRAMM

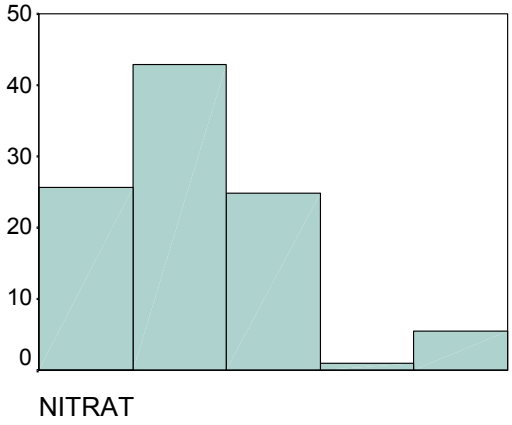


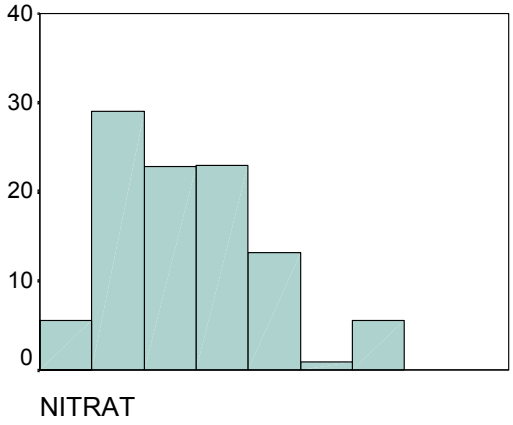
HISTOGRAMM

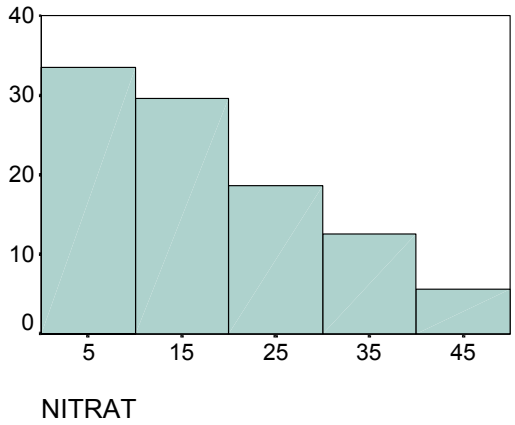


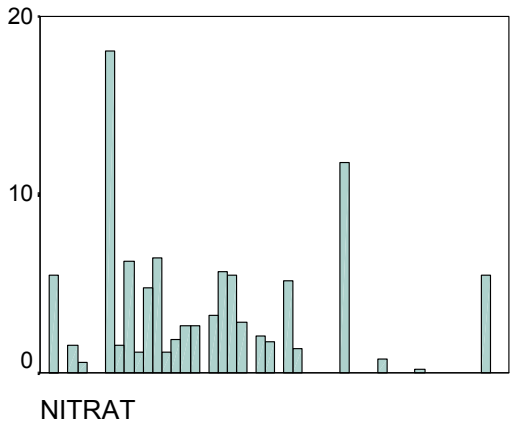
Auswirkung unterschiedlicher Balkenbreite:
Einteilung der Daten in Klassen kann Auswirkungen auf Darstellung haben: (Achtung bei Interpretation)

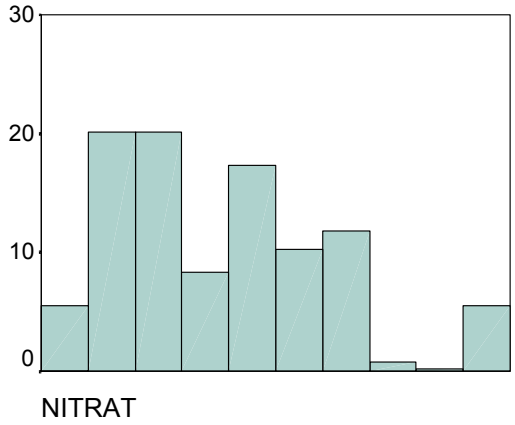












gesamte Fläche ist 1 bzw. 100%

Regeln zur Erstellung eines Histogramms

1. Bei grossem n(>200)

2. Entscheidung wieviele Klassen bzw. Balken
   (üblicherweise zwischen 5 und 20)
   Faustregel:
   Anzahl der Klassen k ist kleinste Zahl so, dass

$$2^k \geq n$$

n	Anzahl Klassen
bis 8	3
9 - 16	4
17-32	5
33-64	6

n	Anzahl Klassen
65 - 128	7
129 - 256	8
257 - 512	9
513 - 1024	10

etc.

3. Klassenbreite festlegen
   Faustregel:

$$\textit{Klassenbreite} = \frac{\text{max - min}}{\text{Anzahl der Klassen}}$$

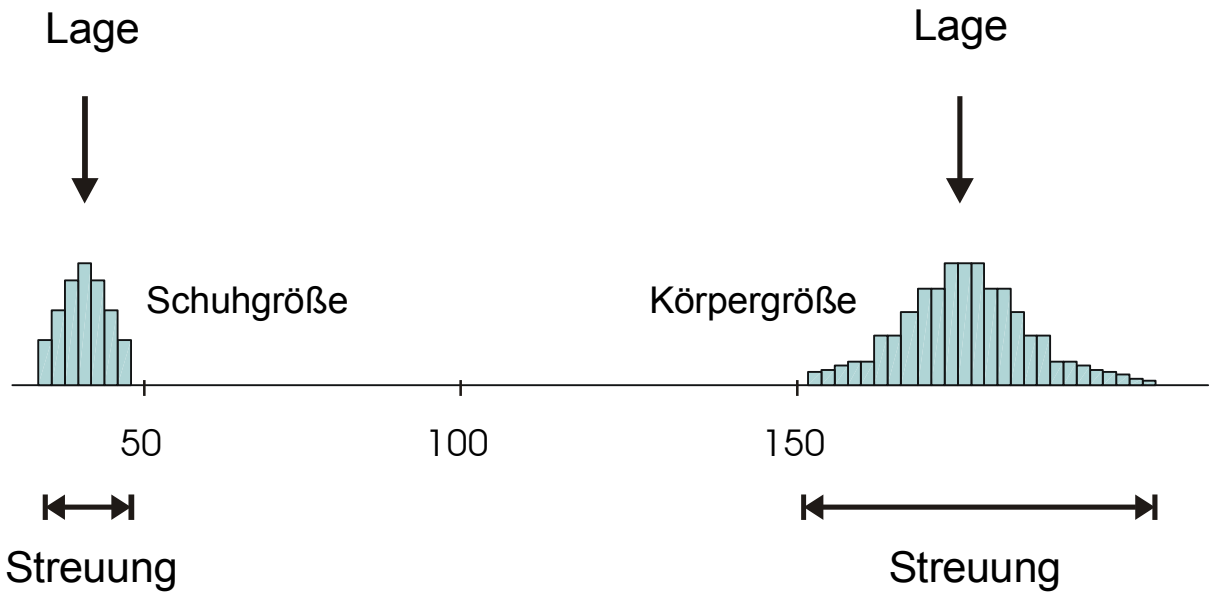
   aufrunden auf Zahl, mit der man gut arbeiten kann

4. Höhe der Balken entspricht der Häufigkeit bzw. der relativen Häufigkeit für die jeweilige Klasse

5. manchmal auch ungleiche Klassenbreiten
   dann muss darauf geachtet werden, dass die Höhe des Balkens stimmt (Fläche proportional zu Häufigkeit)

NUMERISCHE ZUSAMMENFASSUNGEN und GRAFISCHE DARSTELLUNGEN:

Statistische Maßzahlen:



LAGEMASSE:
Mittelwert ($\bar{x}$)
Median ($\tilde{x}$ oder Q_2)
Modus

STREUUNGSMASSE
Varianz (s^2)
Standardabweichung (s)
Interquartilabstand bzw. Q_1,Q_3
Spannweite (“Range“)

Grafische Darstellungen:

Histogramm
Stem-and-Leaf Plot
Box Plot

LAGEMASSE

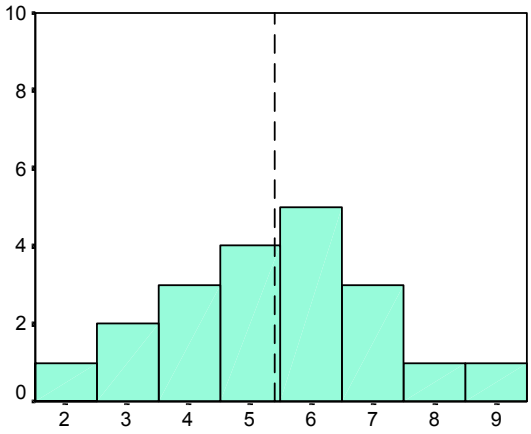
MITTELWERT (ARITHMETISCHES MITTEL)

$$Mittelwert = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\text{Summe aller Werte}}{\text{Anzahl der Werte}}$$

eingipfelige symmetrische Verteilung

Beispiel: Punkteanzahl in einem Test von n =20 Studenten:

x_i : 2, 3,3, 4,4,4, 5,5,5,5, 6,6,6,6,6, 7,7,7, 8, 9



$\bar{x} = 5,4$

- hier ist Mittelwert ein gutes Maß zur Beschreibung und Zusammenfassung der Daten
- etwa die Hälfte der Werte größer bzw. kleiner als $\bar{x}$

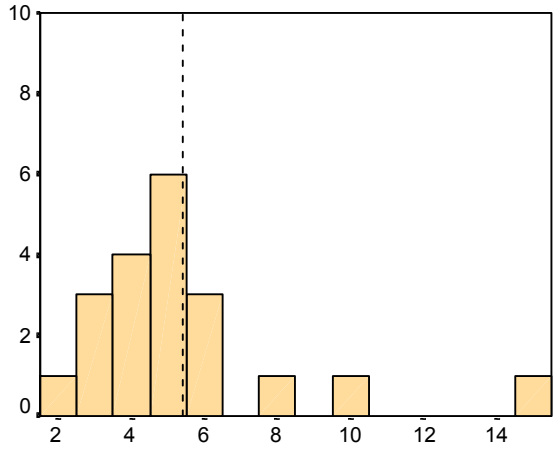
die Form der Verteilung bestimmt ob Mittelwert ein gutes Maß zur Beschreibung ist

In den folgenden Fällen ist Verwendung von $\bar{x}$ weniger sinnvoll bis unsinnig:

schiefe Verteilungen

Beispiel: Anzahl Angestellter in *n* =20 Gemüsegeschäften

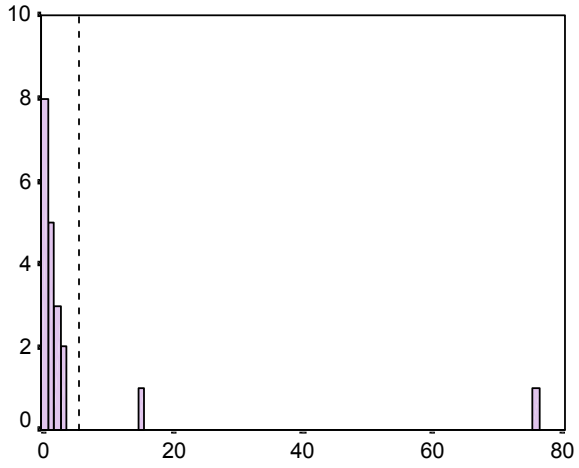
2, 3,3,3, 4,4,4,4, 5,5,5,5,5,5, 6,6,6, 8, 10, 15



$\bar{x} = 5,4$

Beispiel: Anzahl der Tage im Krankenstand von *n* = 20 Arbeitern:

0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 15, 76

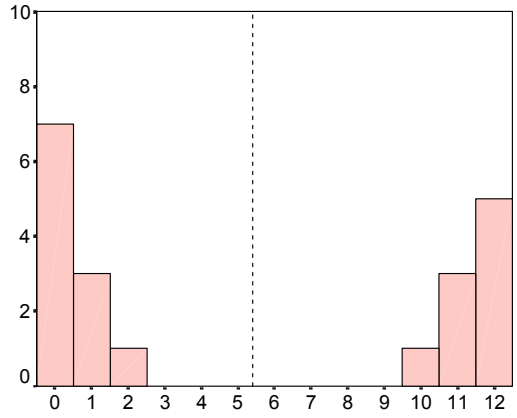


$\bar{x} = 5,4$

U-förmige Verteilungen

Beispiel: Anzahl von Ausgaben einer Monatszeitschrift gelesen von *n* = 20 Personen:

0,0,0,0,0,0,0, 1,1,1, 2, 10, 11,11,11, 12,12,12,12,12



$\overline{x} = 5,4$

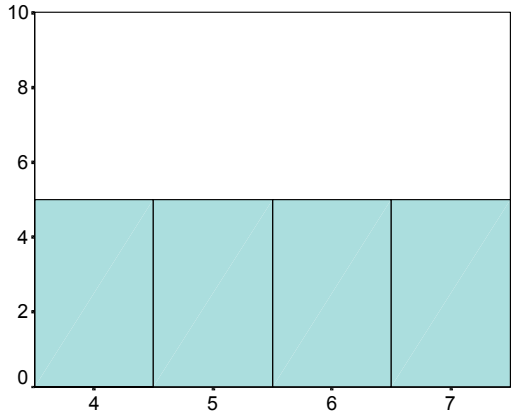
- Verteilung zwar symmetrisch, hat aber 2 Gipfel;
- Mittelwert repräsentiert keinen typischen Wert (wie auch bei schiefen Verteilungen)

⇨ Zusammenfassung der Daten müßte komplexer sein, als einfach den Mittelwert anzugeben

z. B. „etwa die Hälfte der Leute liest keine oder fast keine Ausgabe, während die andere Hälfte alle oder nahezu alle Ausgaben liest“)

Gleichverteilung

Beispiel: Anzahl von PKWs, die während einer Grünphase über eine Kreuzung fahren (*n* =20 Grünphasen)



- Verteilung zwar symmetrisch, hat aber keinen Gipfel
- Mittelwert repräsentiert keinen typischen Wert

<u>im wesentlichen gibt es 2 Arten von Daten:</u>

(1) einigermaßen symmetrische Verteilung mit einem Gipfel in der Mitte
Mittelwert ist gutes Maß zur Beschreibung, da er typisch für die Daten ist

(2) schiefe, U-förmige mehrgipfelige oder Gleichverteilung
es gibt keine typischen Werte
Mittelwert ist schlechtes Maß

MEDIAN

definiert als ”Wert in der Mitte”:

50% der Werte sind KLEINER (bzw. $\leq$) als der Median
50% der Werte sind GRÖSSER (bzw. $\geq$) als der Median

$\tilde{x}$ ... Median

Werte	17	17	18	18	18	**19**	19	19	19	20	21
Rangplätze	1	2	3	4	5	6	7	8	9	10	11

m ... mittlerer Rangplatz

Berechnung: (nach J. Tukey)

1) Sortieren der Werte der Größe nach
   (Rangreihen: vergeben von Rangplätzen)
2) Berechnen des Rangplatzes d. Medians $m = \frac{n+1}{2}$
3) Bestimmen des Wertes d. Medians

wenn *m* eine ,5 Zahl ist, dann nimmt man den Mittelwert der beiden mittleren Zahlen:

$\tilde{x} = 18{,}5$

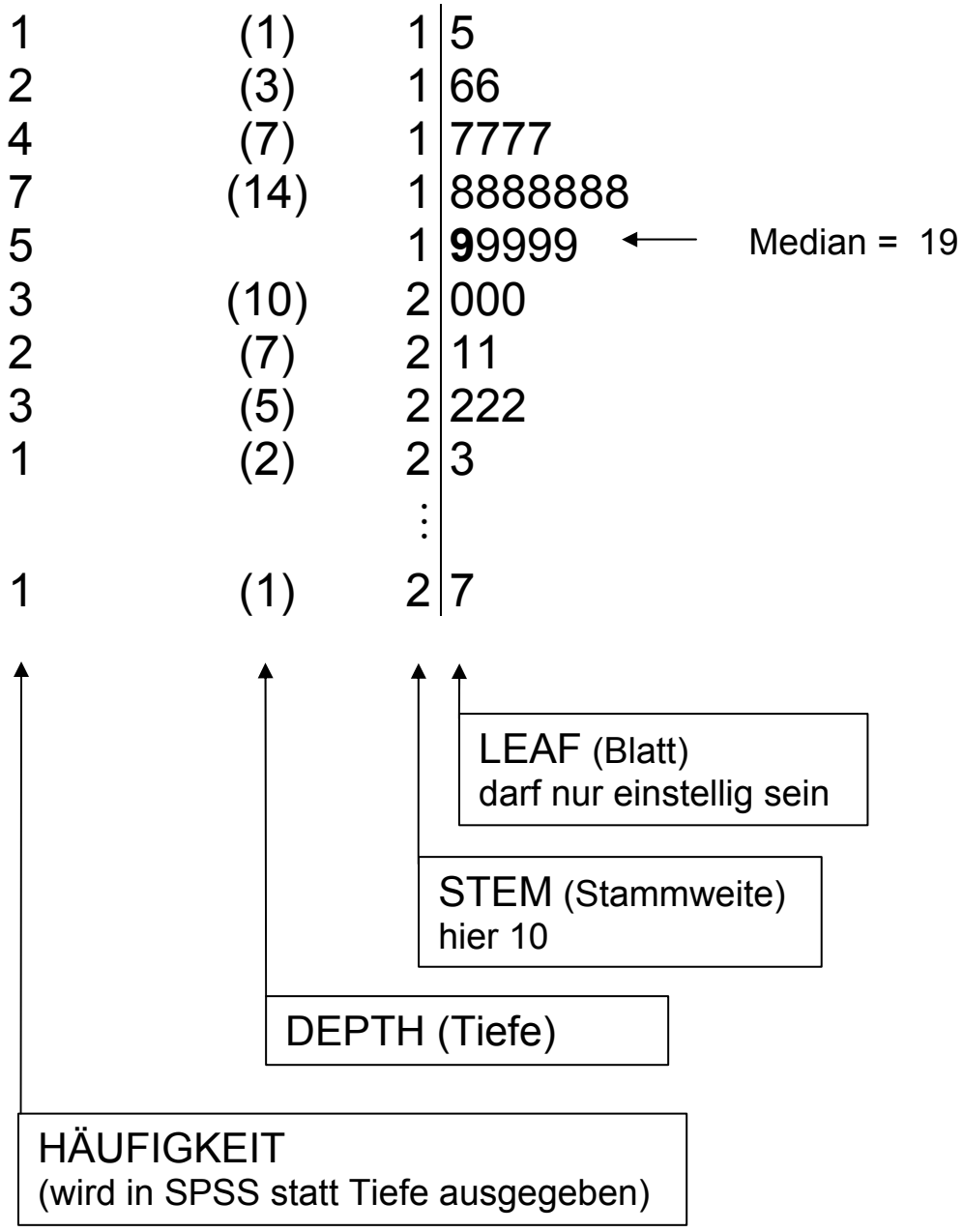
Werte	17	17	18	18	**18**	**19**	19	19	19	20
Rangplätze	1	2	3	4	5	6	7	8	9	10

m = 5,5

Beispiel: Berechnen Sie den Median für folgende Werte:

17	19	22	21	27	22	18	18	16	18
15	16	17	20	19	19	19	20	18	18
23	20	18	21	18	17	19	22	17	

STEM–and–LEAF PLOT (John Tukey)



VORGANGSWEISE ZUR ERSTELLUNG EINES STEM-and-LEAF PLOTS:

1. Entscheiden: was ist Stem, was ist Leaf
   Leaf nur eine Stelle
   Stem (meistens auch nur eine Stelle)
   Rest abschneiden (ist das Gleiche wie runden)
   (meistens verwendet man insgesamt nur 2 Stellen)

2. Entscheiden: wieviele Stems
   zuviele und zuwenige sind schlecht
   Faustregel: zwischen $\sqrt{n}$ und $2\sqrt{n}$
   1-,2-, 5- oder 10-zeilig für gleichen Stem
   wenn positive und negative Zahlen vorkommen:
   dann 2 Stems für 0: +0,-0

3. Schritte 1 und 2 kann man am besten ausführen, wenn man zuerst den größten und kleinsten Wert bestimmt

4. alle Stems auflisten, noch ohne Daten

5. (erste) Zahl aus Datensatz in Stem und Leaf aufteilen
   Leaf an richtiger Stelle einfügen
   das Gleiche mit den anderen Zahlen
   Achtung: Leaves untereinander schreiben (sonst kein Histogramm)

6. am Schluß, wenn notwendig nocheinmal zeichnen, Leaves sortieren

häufige Fehler:

- Leaf mehr als eine Stelle: z.B. 5|21,44,24 statt 5|242
- Stem wird wiederholt: 15|5 ist nicht 1|5

MODUS ("Gipfel" einer Verteilung)

ist der häufigste Wert

z. B.: 0, 0, 1, 1, 1, 1, 2, 2, 3, 5, 19

Modus = 1

manchmal mehr als 1 Modus:

z. B.: 0, 0, 1, 1, 1, 2, 2, 3, 3, 3, 5, 19

Modus Modus

lokaler Modus bei 1 und bei 3

man spricht von "bimodalen" und "multimodalen" Verteilungen, wenn eine Verteilung 2 oder mehrere Gipfel hat

manchmal kein Modus:

z. B.: 0, 1, 3, 4, 6, 7, 9, 11, 15

bei Gleichverteilung

VERGLEICH ZWISCHEN MITTELWERT, MEDIAN und MODUS

Mittelwert: Summe der Werte / Anzahl der Werte
Median: mittlerer Daten-Punkt
Modus: häufigster Wert

bei eingipfeligen, einigermaßen symmetrischen Verteilungen sind Werte bei allen dreien annähernd gleich

bei anderen Formen von Verteilungen sind Werte unterschiedlich
(meist Mittelwert $\neq$ Modus $\cong$ Median)

Beispiel:	Mittelwert	Median	Modus
symm. Verteilung (Testpunkte)	5.4	5.5	6
schiefe Verteilung (Krankenstände)	5.4	1	0
U-förmige Verteilung (Monatsmagazin)	5.4	1.5	0 (bzw.12)

⇨ bei Unterschieden:

starke Abweichung von symmetrischer und/oder eingipfeliger Verteilung

Skalenniveau und Lagemaße

	Mittelwert	Median	Modus
Nominalskala	-	-	ok
Rangskala	-	ok	ok
Intervall-(Rational)skala	ok	ok	ok

Vorteile des MITTELWERTES:

1) kann ausgerechnet werden, ohne Daten vorher sortieren zu müssen

2) wenn man $\bar{x}$ und n kennt, kann man Summe der Einzelwerte ausrechnen (manchmal nützlich)

$$\bar{x} = \frac{1}{n}\sum x_i \rightarrow n \cdot \bar{x} = \sum x_i$$

(bei Median und Modus nicht möglich)

3) Mittelwert kann bei weiteren Analyseschritten verwendet werden, z. B. bei Zusammenfassen mehrerer Datensätze (s. gewichtete Mittelwerte)

„AUSREISSER“ (*outlier*)

Manchmal gibt es einen oder mehrere Werte, die weit weg vom Rest liegen

Beispiel: Krankenstände

Mittelwert wäre 1,7 statt 5,4 wenn man die Beobachtung “76“ weggelassen hätte

wie entstehen Ausreisser und was macht man ?

möglicherweise Fehler:	korrigieren oder weglassen
möglicherweise reliabel:	weglassen und getrennt darüber berichten

Median und Modus sind robust gegenüber Ausreissern
(s. Vergleich Mittelwert gegenüber Median bzw. Modus)

GEWICHTETE MITTELWERTE

bei Kombination von Datensätzen 2 Möglichkeiten einen Gesamtmittelwert zu bestimmen:

		n	$\sum x_i$	$\bar{x}$
Daten1:	0,1,1,2,2,3,3,4	8	16	2
Daten2:	3,4,5	3	12	4
kombiniert:	0,1,1,2,2,3,3,3,4,4,5	11	28	2.55

(1) Mittelwert der einzelnen Mittelwerte: $\left(\frac{\bar{x}_1 + \bar{x}_2}{2}\right)$

(2) Mittelwert aller individuellen Werte

kann unterschiedlich sein: für (1): $\bar{x} = 3$
für (2): $\bar{x} = 2{,}55$

Berechnung des gewichteten Mittels:
wenn man Mittelwerte und Stichprobengrößen (n) kennt

$$\bar{x} = \frac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2}{(n_1 + n_2)} = \frac{8 \cdot 2 + 3 \cdot 4}{8 + 3} = \frac{28}{11} = \underline{\underline{2.55}}$$

„gewichtet“: weil jeder Mittelwert mit seinem Anteil an Gesamtbeobachtungszahl “gewichtet“ wird

STREUUNGSMASSE

um Daten bzw. Verteilungen zu beschreiben genügt es nicht nur eine Maßzahl für die Lage anzugeben

z. B.: „das durchschnittliche Alter der Studenten im HS ist 20,5“

auch die Streuung ist wichtig

oft genügt:
z.B.: „die meisten Studenten sind zwischen 18 und 21 Jahre alt“

$\Rightarrow$ manchmal präzisere Angaben notwendig:

2 Arten von Maßzahlen für Streuung:

- wie sehr weichen Daten vom Mittelwert ab

  - mittlere absolute Abweichung
  - Varianz
  - Standardabweichung

- in welchem Bereich liegen die Daten, bzw. zwischen welchen Werten liegen die Daten bzw. 50% der Daten

  - Spannweite („range“)
  - Inter-Quartil-Abstand

Streuungsmaße nur bei eingipfeligen symmetrischen Verteilungen gut interpretierbar

sonst sollte man zusätzliche Angaben machen

MITTLERE ABSOLUTE ABWEICHUNG

Beispiel: Anzahl von Angestellten, die an 5 Werktagen zu spät ins Büro kommen

x_i: 11, 2, 5, 1, 6 $\qquad \overline{x} = 25/5 = 5$

Abweichungen vom Mittelwert: $\qquad d_i = x_i - \overline{x}$

d_i: 6, -3, 0, -4, 1

Mittelwert dieser Abweichungen: $\frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})$,

ist sinnlos, da immer 0 und 0 ist kein Maß für Abweichung

besser: Mittelwert der **absoluten** Abweichungen

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

z. B.: $MAA = (6+3+0+4+1)/5 = \underline{\underline{2.8}}$

“die Werte liegen im Durchschnitt 2.8 Einheiten vom Mittwert 5 entfernt“

Vorteile:
- einfach verständliches Maß
- bei kleinen Datensätzen leicht berechenbar

Nachteile:
- bei großen Datensätzen Berechnung lästig
- für spätere statistische Analysen nicht verwendbar

VARIANZ UND STANDARDABWEICHUNG

Standardabweichung (s) ist Wurzel aus Varianz (s^2)

vom Konzept her nicht so offensichtlich,
aber leichter zu berechnen
können weiterverwendet werden (spätere statistische Analyse)

beide Maße gehen von quadrierten Abständen aus: $(x_i - \overline{x})^2$
(diese auch immer positiv)

Abweichungen:	6	-3	0	-4	1
quadrierte Abweichungen:	36	9	0	16	1

- **Varianz** (ist Mittelwert der quadrierten Abweichungen)

$$Var(x) = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad s^2 = 62/5 = 12.4$$

(große Abweichungen vom Mitelwert gehen stärker ein)

Varianz mißt Streuung in „quadrierten Einheiten“
um zu den ursprünglichen Einheiten zurückzukehren:

- **Standardabweichung:** ($\Rightarrow$ Wurzel aus Varianz)

$$s = \sqrt{Var(x)} = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad s = \sqrt{12,4} = 3.5$$

(Standardabweichung ist etwas größer als MAA)

Verschiebungssatz:

einfachere Berechnung der Varianz :

Merkspruch:
„Mittelwert der Quadrate minus Quadrat des Mittelwerts“

$$s^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2 = \frac{1}{n}\sum_i x_i^2 - \overline{x}^2$$

Mittelwert der Quadrate

Quadrat des Mittelwerts

Herleitung:

$$s^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2$$

$$= \frac{1}{n}\sum_i \left(x_i^2 - 2x_i\overline{x} + \overline{x}^2 \right)$$

$$= \frac{1}{n}\left[\sum x_i^2 - 2\overline{x}\sum_i x_i + n\overline{x}^2 \right]$$

$$= \frac{1}{n}\left[\sum x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 \right]$$

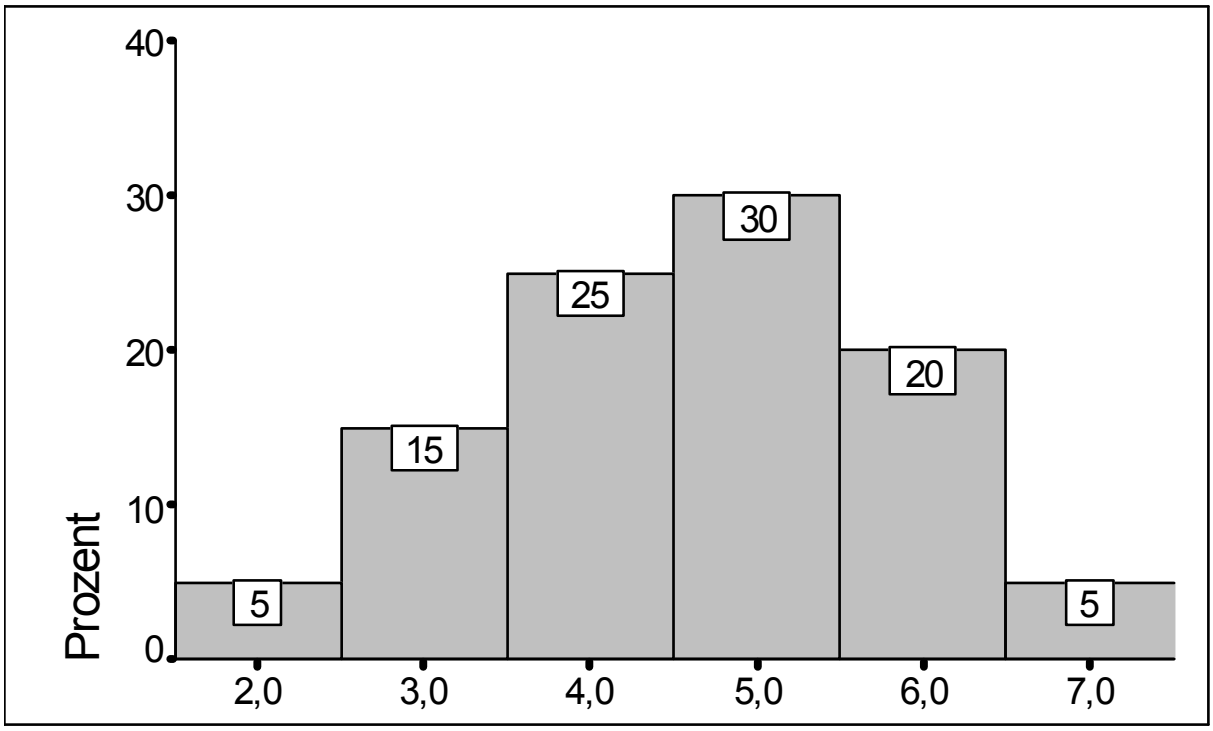
$$= \frac{1}{n}\left[\sum x_i^2 - n\overline{x}^2 \right]$$

$$= \frac{1}{n}\sum x_i^2 - \overline{x}^2$$

BERECHNUNG von MITTELWERT und VARIANZ mit RELATIVEN HÄUFIGKEITEN

Beispiel:

Häufigkeitsverteilung der Wühldauer (in Minuten) von Schweinen in einem neuen Stall.



wie lassen sich Mittelwert und Varianz berechnen ?

Beispiel: Punkte im Test von 20 Studenten

- **Daten** x_i**:** 2, 3,3, 4,4,4, 5,5,5,5, 6,6,6,6,6, 7,7,7, 8, 9

  (Index ***i***: ***i*** bezeichnet die i -te Person, ingesamt n)

- **unterschiedliche Werte** x_j^***:** 2 3 4 5 6 7 8 9

  (Index ***j***: ***j*** bezeichnet die j -te Zahl, die vorkommt, ingesamt J verschiedene Zahlen)

  - h_j ... absolute Häufigkeit der j -ten Zahl
    (wie oft kommt die j -te Zahl vor)

  - r_j ... relative Häufigkeit der j -ten Zahl $(r_j = h_j / n)$
    (Anteil der j -ten Zahl)

- **Berechnung des Mittelwerts**

$$\bar{x} = (2+3+3+4+4+4+5+5+5+5+6+6+6+6+6+7+7+7+8+9)/20$$
$$= (2\cdot 1 + 3\cdot 2 + 4\cdot 3 + 5\cdot 4 + 6\cdot 5 + 7\cdot 3 + 8\cdot 1 + 9\cdot 1) / 20$$
$$= (x_1 \cdot h_1 + x_2 \cdot h_2 + \cdots + x_J \cdot h_J) / n$$
$$= \frac{1}{n} \sum_j x_j \cdot h_j$$
$$= \sum_j x_j \cdot \frac{h_j}{n}$$
$$= \sum_j x_j \cdot r_j$$

Tabelle für Mittelwert und Varianz:

j	x_j^*		h_j	r_j	$x_j^* \cdot r_j$	x_j^{*2}	$x_j^{*2} \cdot r_j$
1	2	I	1	0.05	0.10	4	0.20
2	3	II	2	0.10	0.30	9	0.90
3	4	III	3	0.15	0.60	16	2.40
4	5	IIII	4	0.20	1.00	25	5.00
5	6	IIIII	5	0.25	1.50	36	9.00
6	7	III	3	0.15	1.05	49	7.35
7	8	I	1	0.05	0.40	64	3.20
8	9	I	1	0.05	0.45	81	4.05
$\sum$					5.40		32.10
					[1]		[2]

- **MITTELWERT:**

$$\bar{x} = \sum_{i=1}^{n} x_i = \sum_{j=1}^{J} x_j^* \cdot r_j = 5.40$$

- **VARIANZ:**
  (Mittelwert der Quadrate [2] minus Quadrat des Mittelwerts [1])

$$s^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 = [2] - [1]^2 = 32,10 - 5,4^2 = 2.94$$

  Mittelwert der Quadrate:

$$\frac{1}{n}\sum_i x_i^2 = \frac{1}{n}\sum_{j=1} x_j^{*2} \cdot h_j = \sum_j x_j^{*2} r_j = 32.10$$

Übungsaufgabe: *berechnen Sie* $\bar{x}$ *und* s^2 *für die Wühldauer von Schweinen aus dem obigen Histogramm*

SPANNWEITE („range“)

ist größter Wert minus kleinstem Wert

Beispiel: x_i : $11,2,5,1,6$ $\quad \bar{x} = 5$

$\Rightarrow$ Spannweite $= 11-1 = 10$

wird in der Praxis nicht als Streuungsmaß verwendet,
eher zur Beschreibung der Größe des Wertebereiches der Variable X

Wertebereich (auch "Träger" genannt):
jener Bereich auf der Zahlengerade, in dem Werte von X vorkommen

INTERQUARTILABSTAND (IQD)

Quantile: sind jene Werte, die eine Häufigkeitsverteilung in mehrere gleich große Stücke teilen

Quartile: teilen eine Häufigkeitsverteilung in **4** Teile

Q_1: unteres Quartil ... jener Wert, unter dem 25% aller Werte liegen

Q_2: Median 50% der Werte sind kleiner

$Q_{3:}$ oberes Quartil ... jener Wert, unter dem 75% der Werte liegen

Beispiel:

Daten: 16, 17,17, 18,18,18, 19,19,19, 20,20, 22

1	6
1	77 ------- Q_1=17.5
1	888 ------- Q_2=18.5
1	999 ------- Q_3=19.5
2	00
2	
2	2

Perzentile: teilen die Häufigkeitsverteilung in 100 Teile (x_p ... p - Perzentil)

$x_{0.25} \triangleq Q_1$ ... 0,25 – Perzentil (unteres Quartil)
$x_{0.50} \triangleq Q_2$ ... 0,50 – Perzentil (Median)
$x_{0.75} \triangleq Q_3$ ... 0,75 – Perzentil (oberes Quartil)

Berechnung der Quartile: (nach John Tukey)

1) Sortieren der Werte der Größe nach
2) Berechnen des Rangplatzes d. Medians $m = \frac{n+1}{2}$
3) Berechnen der Tiefe der Quartile: $m' = \frac{\lfloor m \rfloor + 1}{2}$

   $\lfloor .... \rfloor$ heißt abrunden z.B. $\lfloor 12.5 \rfloor = 12$

4) in die Rangreihe von oben und von unten hineinzählen bis zur Stelle m' - dies ist die Stelle für Q_1 bzw. Q_3

5) wenn m' keine ganze Zahl ist, Vorgangsweise wie bei Median (d.h. Mittelwert der benachbarten Zahlen)

(es gibt vier verschiedene Situationen, je nachdem ob *n* eine gerade und *m* eine ganze Zahl ist)

alternative Methode

Rangplätze von Q_1 , Q_2 , und Q_3 ergeben sich aus $(n+1)\cdot p$ wobei *p* das p-te Perzentil ist

Stelle für Q_1 ... $(n+1)\cdot 0{,}25$
Q_2 ... $(n+1)\cdot 0{,}50$
Q_3 ... $(n+1)\cdot 0{,}75$

wieder: wenn $(n+1)\cdot p$ keine ganze Zahl ist:
Quartil ist Mittelwert der beiden benachbarten Zahlen

Interquartilabstand: ("inter-quartile distance")

$$IQD = Q_3 - Q_1$$

mittlerer Quartilabstand:

$$MQA = \frac{Q_3 - Q_1}{2}$$

- diese Maße nicht so von extremen Werten (wie min., max. sein können) abhängig
- beschreiben Intervall in dem 50% der Werte rund um den Median liegen
- als Beschreibungsmaß nur bei symmetrischen, eingipfeligen Verteilungen geeignet

meistens ist es besser alle Quartile (Q_1 , Q_2 , Q_3) anzugeben

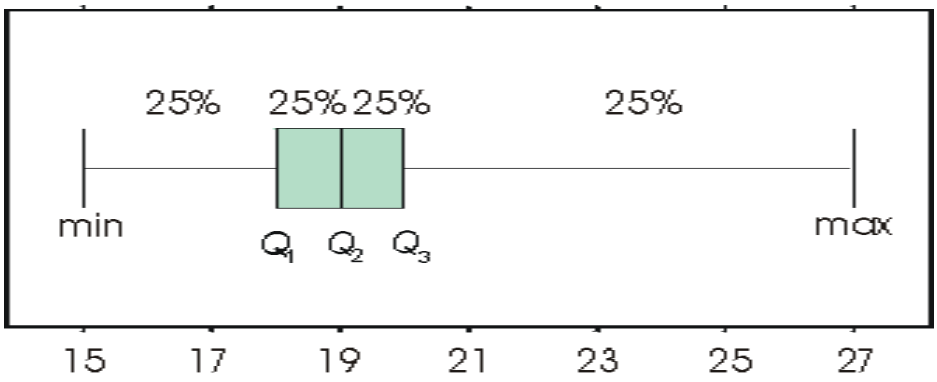
Skalenniveau und Streuungsmaße

	MAA	Varianz	IQD	Range
(Nominalskala)	-	-	-	(ok)
Rangskala	-	-	?	ok
Intervall-(Rational)skala	ok	ok	ok	ok

BOX PLOT

- gibt Verteilung gut wieder
- zeigt Lage- und Streuungsmaß
- günstig bei Vergleich mehrerer Datensätze

einfache Version:



ergibt sich aus Minimum,Maximum und Quartilen:

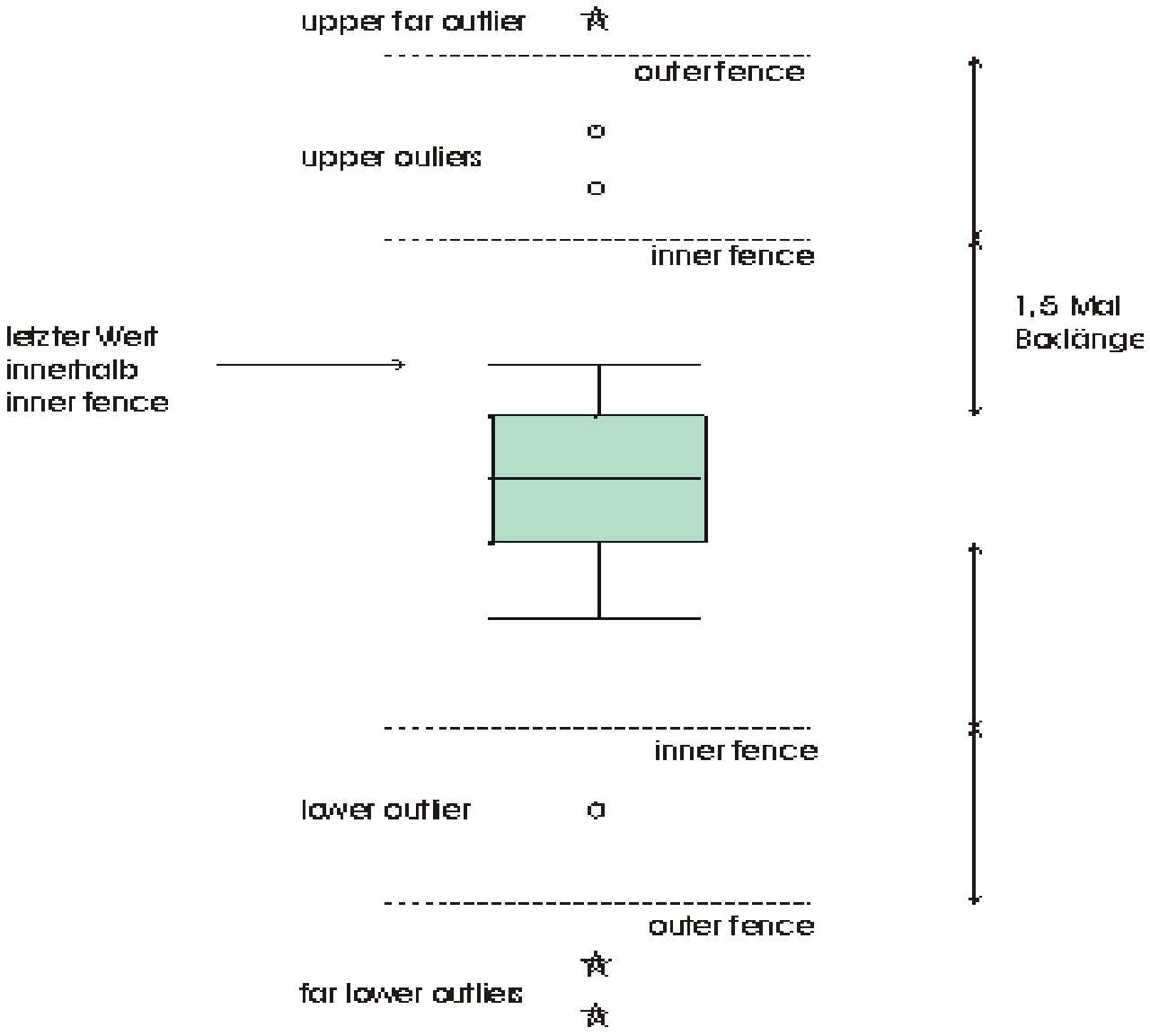
Beispiel:

(1)	**1**	**5**
(3)	1	66
(7)	1	7777
(14)	**1**	**8**888888
	1	**9**9999
(10)	**2**	00**0**
(7)	2	11
(5)	2	222
(2)	2	3
	⋮	
(1)	**2**	**7**

$$n = 29$$
$$m = (29+1)/2 = 15$$
$$m' = (15+1)/2 = 8$$

$$\min = 15$$
$$Q_1 = 18$$
$$Q_2 = 19$$
$$Q_3 = 20$$
$$\max = 27$$

Anatomie eines Box Plots



WICHTIGE FRAGESTELLUNGEN
(BEI EINEM METRISCHEN MERKMAL)

- In welchem Bereich kann man einen Mittelwert in der Grundgesamtheit erwarten ?

Diese Frage stellt man, wenn man abschätzen möchte, wie groß die Schwankungen sein können, denen ein Durchschnittswert unterliegen kann, wenn man Daten nur aus Stichproben hat, aber über die Grundgesamtheit etwas aussagen möchte

Beispiel: In welchem Bereich kann *NPDC* die Einnahmen aus dem Verkauf von *CARIDEX* erwarten ?

- Ist ein Mittelwert anders (kleiner, größer, oder ungleich) als eine bestimmte Vorgabe ?

Diese Frage stellt man, wenn man prüfen möchte, ob die Größe eines Durchschnittswertes in der Grundgesamtheit einer bestimmten Annahme entspricht.

Beispiel: Erreicht man mit einer bestimmten Diät eine durchschnittliche Gewichtsabnahme von mindestens 5 kg?

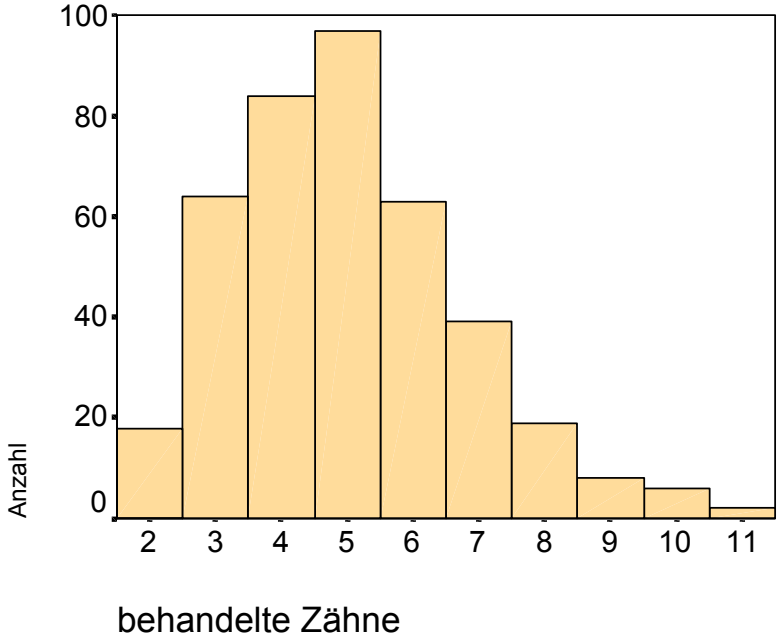
FRAGESTELLUNG 1:

In welchem Bereich kann man einen Mittelwert in der Grundgesamtheit erwarten ?

Bsp.: CARIDEX, Anzahl behandelter Zähne pro Woche

Gesucht ist der Mittelwert der in einer Woche mit Caridex behandelten Zähne **für alle** in Frage kommenden Zahnärzte.
In welchem Bereich kann man diesen Mittelwert erwarten ?

repräsentative Stichprobe von 400 Zahnärzten ergab



$\overline{x} = 4{,}015$

obige Frage anders gestellt:

Wenn ich einen Mittelwert aus einer Stichprobe errechnet habe, kann ich dann daraus schließen, wo der unbekannte Mittelwert der Population liegt ?

Exkurs: Inferenzstatistik - zentraler Grenzwertsatz

INFERENZSTATISTIK - Überblick

Beschreibende Statistik:
Methoden, um Information aus Daten zu gewinnen und zusammenzufassen

- Grafische Methoden
- Numerische Methoden

Inferenzstatistik: (schließende Statistik)
Methoden, um Information aus Stichproben so zu nutzen, daß Rückschlüsse auf eine Population gezogen werden können

Schätzen, Prognose
Testen (Prüfen von Fragestellungen)
Modellieren (Aufspüren von Zusammenhängen, Entwicklung von Erklärungsmustern)

meist nicht möglich die gesamte Population zu untersuchen - zu aufwendig oder zu kostspielig

einfacher: Stichprobe ziehen - Schlußfolgerungen aufgrund von Statistiken

aber: Schätzungen und Schlußfolgerungen aufgrund von Stichprobendaten nicht immer korrekt

daher: Maße für die Zuverlässigkeit von Schlußfolgerungen

Konfidenzniveau:
Prozentsatz von Fällen in denen eine Schätzung korrekt ist

Signifikanzniveau:
Prozentsatz der Fälle in denen eine Entscheidung (beim Hypothesentesten) richtig ist

Methoden der Inferenzstatistik

- Methoden zur ***Schätzung*** oder ***Vorhersage*** von Werten in einer Population (s. Fragestellung 2)

  Beispiele:
  - Wieviel Prozent der Österreicher sind für die Einführung des Euro ?
  - Wie hoch ist der Durchschnittsverbrauch von Milch/Tag bei Kindern unter 10 ?

- Methoden zur ***Überprüfung*** (Testen) ***von Fragestellungen*** (Hypothesen), die die Population betreffen (s.Fragestellung 3)

  Beispiele:
  - Unterscheiden sich Arbeiter und Angestellte bezüglich der durchschnittlichen Anzahl von Krankenstandstagen pro Jahr ?
  - Besteht ein Zusammenhang der Intelligenzquotienten bei verheirateten Paaren ?

- ***Modelle*** zum Auffinden von Zusammenhängen bzw. von Erklärungsmustern
  (s. Kapitel 5: Zwei metrische Merkmale)

  Beispiele:
  - Hängen die Anzahl eingenommener Vitamintabletten und die Anzahl von Krankenstandstagen zusammen ?
  - Lässt sich der Benzinverbrauch durch das Gewicht, Hubraum, PS, oder Herstellerland eines Autos erklären?

Grundidee der Inferenzstatistik:

Population (Grundgesamtheit):
Charakteristika (Maßzahlen) in einer Population heißen ***Parameter***

Beispiele:
- Durchschnittsalter aller Studenten an WU,
- Anteil der Euro-Befürworter in Österreich

Stichprobe:
Charakteristika in Stichprobe heißen ***Statistiken***

Beispiele:
- Mittelwert, Prozentsatz, Odds-Ratio, ...

wie stehen Statistiken und Parameter miteinander in Beziehung ?

3 Arten von Verteilungen:

- **Verteilung einer Population** („population distribution“)

  Häufigkeiten für alle Werte aus einer Population

- **Verteilung in einer Stichprobe** („sample distribution“)

  Häufigkeiten für die Werte aus einer Stichprobe

- **Verteilung von Statistiken** („sampling distribution“)

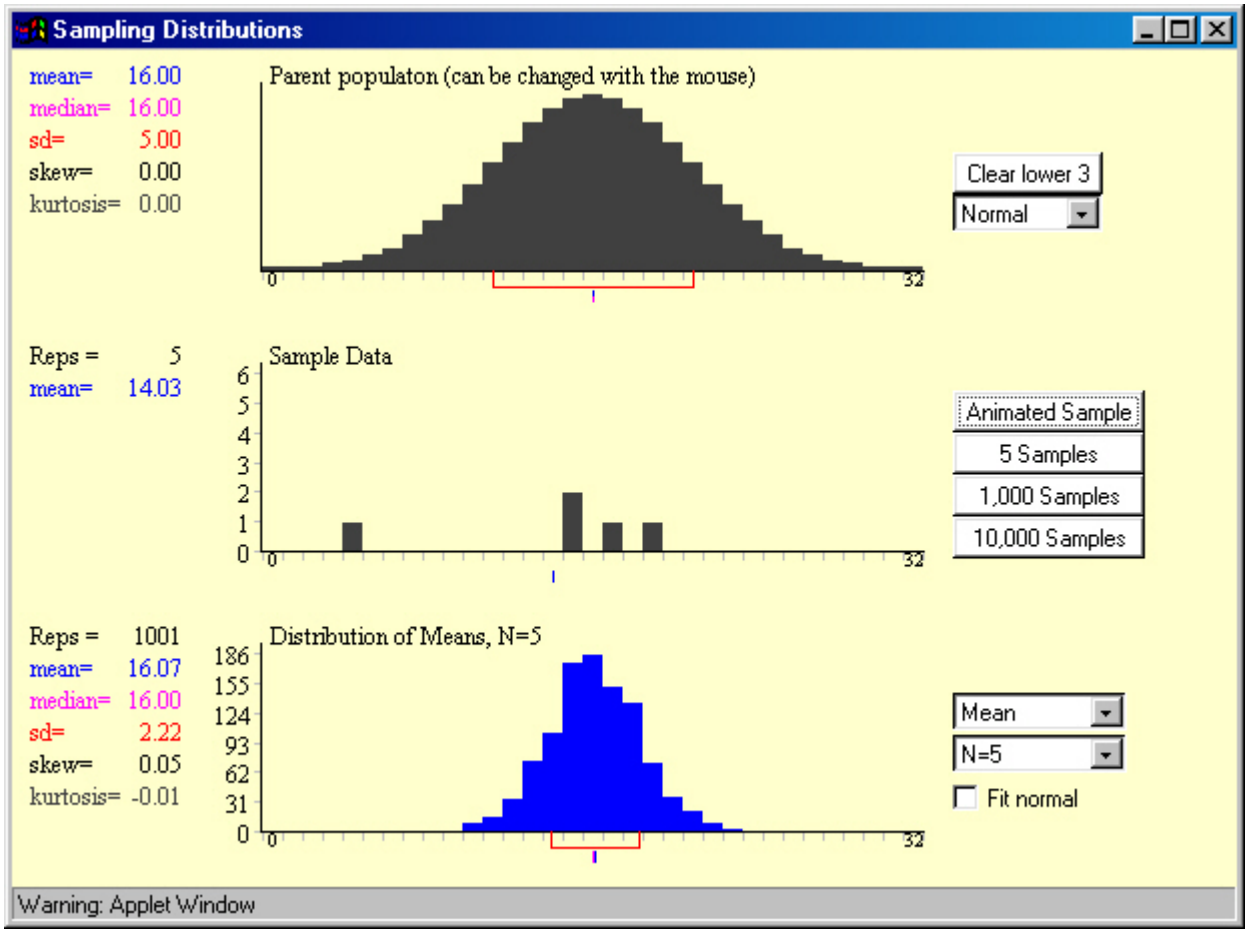
  Häufigkeiten, mit der bestimmte Kennwerte (Statistiken) auftreten (berechnet jeweils aus den Werten einer Stichprobe), wenn man viele Stichproben aus derselben Population gezogen hat

Beispiel:

- man zieht eine Stichprobe aus einer Population und berechnet den Mittelwert
- man zieht eine zweite Stichprobe aus einer Population und berechnet wieder den Mittelwert
- man zieht ein dritte Stichprobe, berechnet den Mittelwert, etc.
- das macht man viele Male
- schließlich zeichnet man ein Histogramm aller so berechneten Mittelwerte
- das Resultat sieht immer aus wie eine **Normalverteilung**, egal wie die Verteilung in der Population war (dies gilt auch für andere Statistiken, wie Anteile, Regressionskoeffizienten, etc.)

diese Gesetzmäßigkeit nennt man:

zentraler Grenzwertsatz (siehe Seite 4-42)



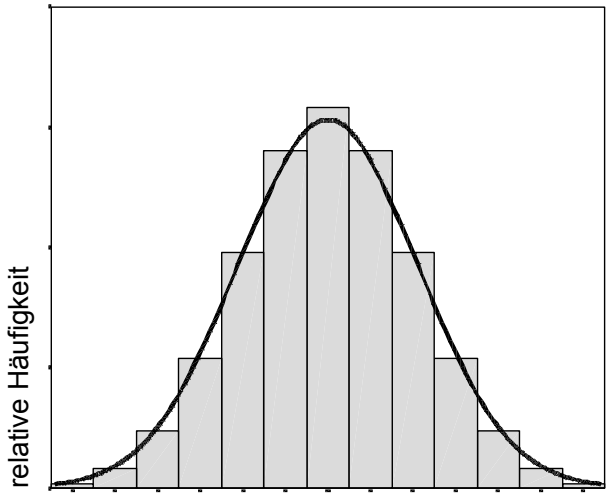
Exkurs:

NORMALVERTEILUNG

ist eine theoretische Verteilung

manchmal kann Verteilung von Daten mittels mathematischer Formeln beschrieben werden

man kann Daten beschreiben mittels weniger Kenngrößen (Parameter), wie z.B. Mittelwert und Varianz



- gibt relative Häufigkeit für bestimmt Werte von x an
- mathematische Formel:

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

hat zwei wichtige Kenngrößen:

μ... (sprich *„mü“*) ....ist Mittelwert
σ... (sprich *„sigma“*) ....ist Standardabweichung

Formel kompliziert, aber ohne direkte Bedeutung für praktische Arbeit
sie besagt:

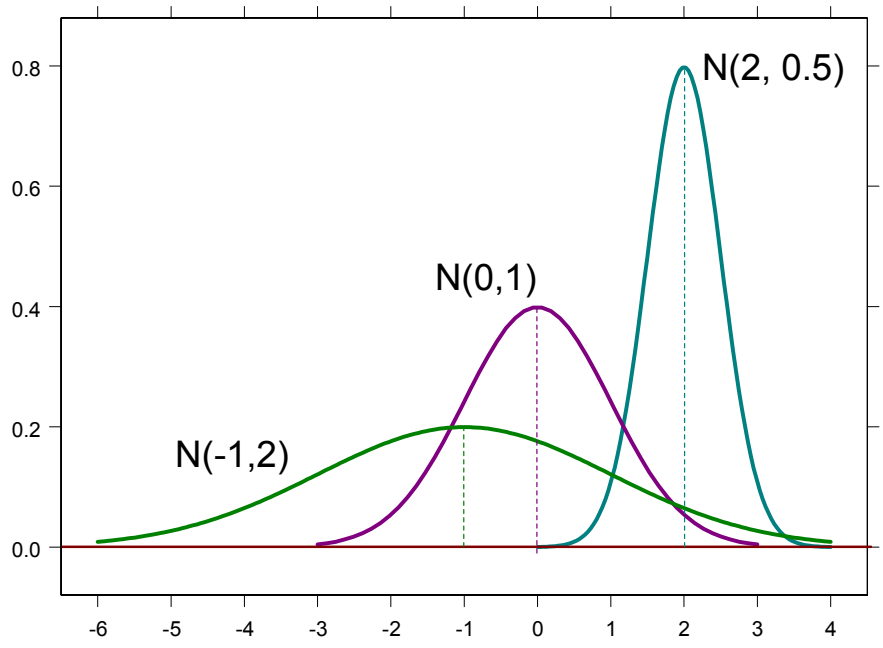
$\Rightarrow$ die Häufigkeit der Werte x ist proportional wie weit x vom Mittelwert μ entfert ist, relativ zur Größe der Standardabweichung (je weiter weg, umso kleiner die Häufigkeit)

Eigenschaften der Normalverteilung:

- ist stetige Verteilung, geht von -∞ bis +∞

- wird allein durch Mittelwert μ *und Varianz* σ^2 festgelegt
  oft genügt es zu sagen: Daten folgen einer Normalverteilung mit bestimmtem Mittelwert und bestimmter Standardabweichung um diese gut zu beschreiben
  z. B.:Daten folgen einer Normalverteilung mit $\mu = 9,\ \sigma = 3$

- es gibt viele Normalverteilungen - je nach μ und σ^2



- man schreibt auch kurz: $X \sim N\left(\mu, \sigma^2\right)$

  sprich: *"die Variable X ist normalverteilt mit* μ *und* σ^2*"*
  ("$\sim$" steht für *"ist verteilt"* und "*N*" für *"Normalverteilung"*)
  Bsp: X ist normalverteilt mit Mittelwert 2 und Varianz 0.5

- wichtigste ist ***N* (0,1)** - "**Standardnormalverteilung**":
  hat Mittelwert 0 und Varianz 1(Standardabweichung auch 1)
- eine standardnormalverteilte Variable wird mit ***Z*** bezeichnet

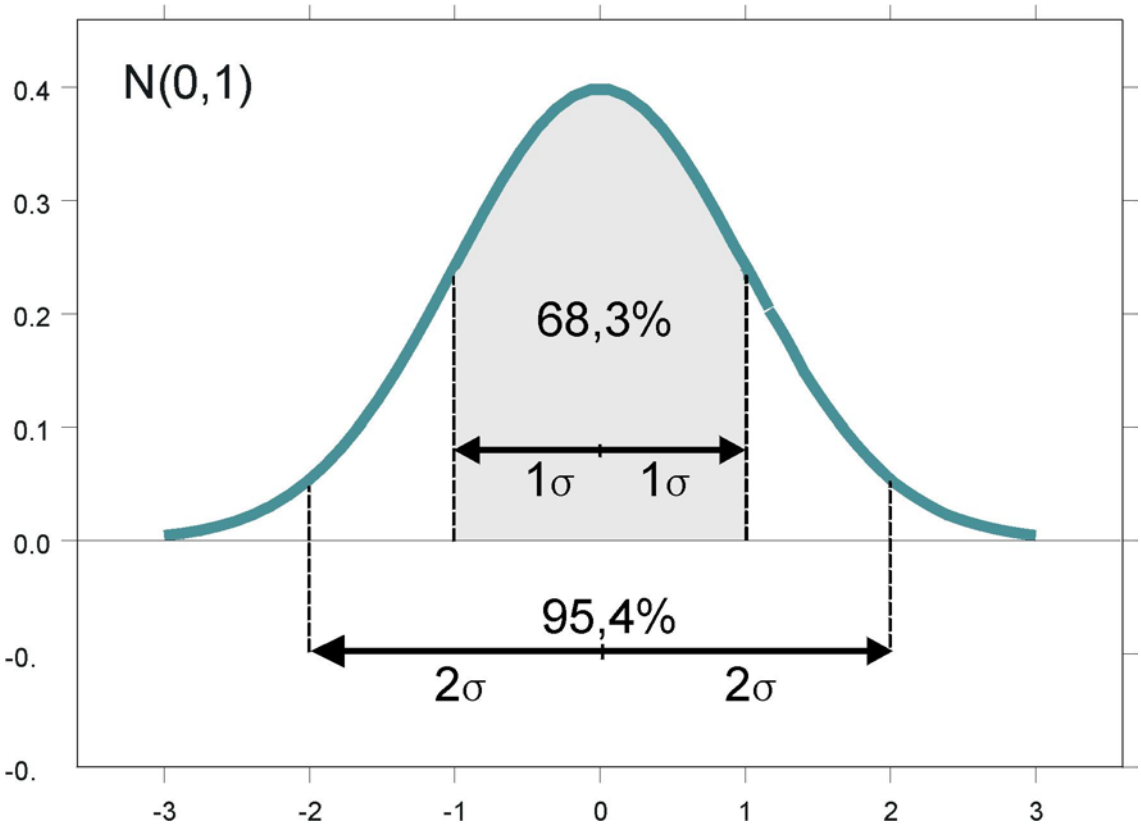
Eigenschaften der Normalverteilung (Fortsetzung):

- eingipfelig, symmetrisch
  (Mittelwert, Median und Modus sind gleich)

- Normalverteilung paßt daher oft sehr gut für symmetrische, eingipfelige Daten (z. B. IQ, Körpergröße)

- Normalverteilung ist ein *Modell*
  (vereinfachtes Abbild der "Realität")

- wenn man Mittelwert und Varianz kennt, kann man die relative Häufigkeit angeben, mit der Werte eines bestimmtes Intervalls vorkommen



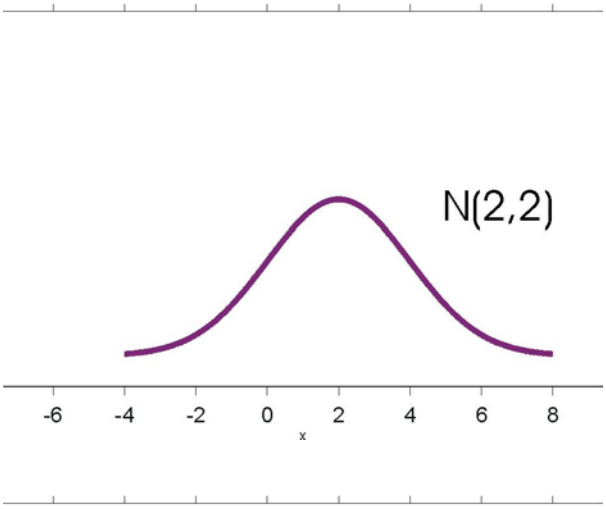
ca. 68% der Daten liegen $\pm 1\sigma$ vom Mittelwert
ca. 95% der Daten liegen $\pm 2\sigma$ vom Mittelwert

STANDARDISIEREN - STANDARDNORMALVERTEILUNG

nur selten haben Daten Mittelwert 0 und Varianz 1
man kann aber jede beliebige Variable X in eine Variable Z so transformieren, dass dann Mittelwert 0 und Varianz 1 ist

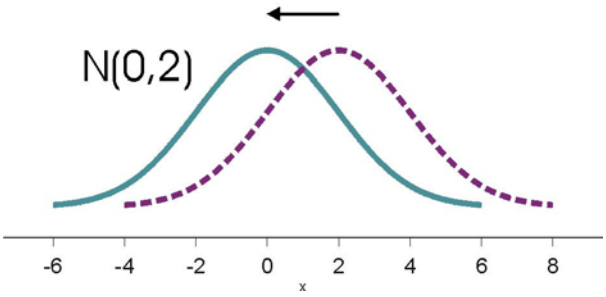
Methode heißt ***standardisieren:*** $z = \frac{x - \mu}{\sigma}$

Beispiel:



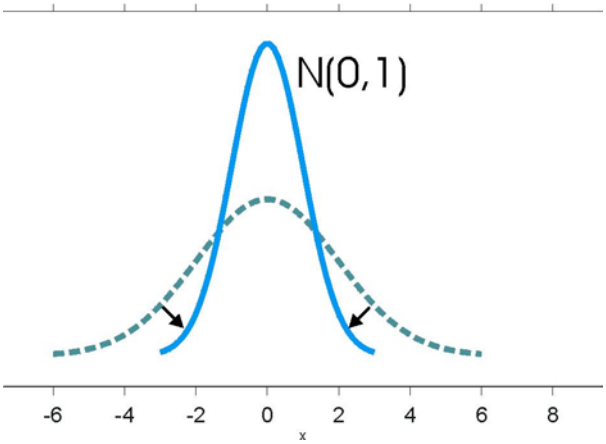
$$X \sim N(2,2)$$

↓



$$X - 2 \sim N(0,2)$$

↓



$$Z = \frac{X - 2}{\sqrt{2}} \sim N(0,1)$$

ZENTRALER GRENZWERTSATZ

gegeben sei eine beliebige Verteilung in der Grundgesamtheit mit Mittelwert μ und Varianz σ^2

zentraler Grenzwertsatz:

Mittelwerte aus Stichproben der Größe *n* sind approximativ **normalverteilt**, wenn *n* ausreichend groß ist

wobei:

der Mittelwert der Stichprobenmittelwerte ist gleich dem Mittelwert in der Grundgesamtheit:

$$\mu_{\bar{X}} = \mu$$

die Standardabweichung der Stichprobenmittelwerte verkleinert sich um den Faktor $\sqrt{n}$, d.h.

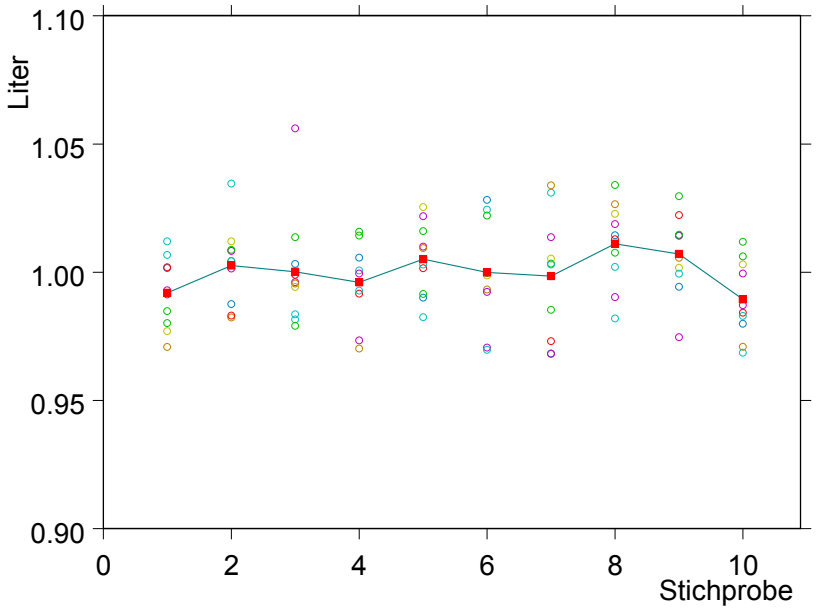
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- dies gilt nicht nur für Stichprobenmittelwerte sondern auch für andere Statistiken wie etwa Anteile, Korrelations-oder Regressionskoeffizienten

- was ausreichend groß ist hängt von der Form der Verteilung in der Population ab:
  je weniger eingipfelig und symmetrisch umso höheres *n* notwendig (relativ sichere Faustregel: *n* =30)

Variabilität der Verteilung in Stichproben und bei der Verteilung von Statistiken

Bsp:
- Getränk wird in 1 Liter Flaschen abgefüllt
- Maschine arbeitet mit $\mu = 1$ und $\sigma = 0.02$
- über 10 Tage wird je eine Stichprobe von 10 Flaschen geprüft und der Mittelwert bestimmt (Linie)



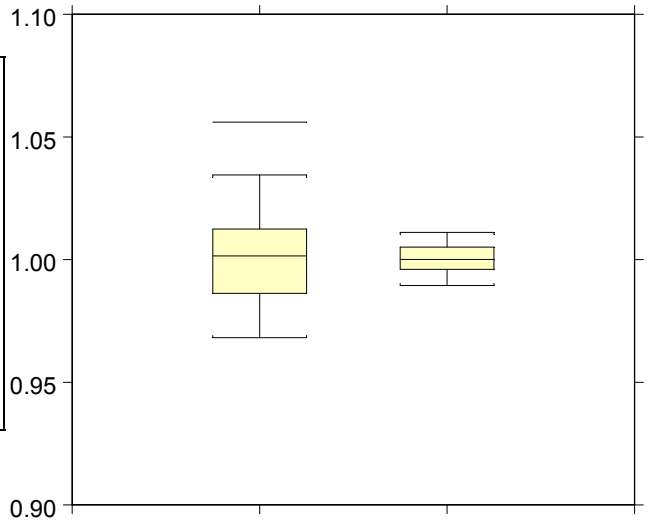
Verteilung der Stichprobenmittelwerte

hat viel weniger Streuung als die Verteilung in der Population, nämlich um den Faktor $\frac{1}{\sqrt{n}}$

Verteilung aller Werte

$\bar{x} = 1.00$

$s = 0.018$



Verteilung der Mittelwerte

$\bar{x}_{\bar{X}} = 1.00$

$s_{\bar{X}} = 0.007$

KONFIDENZINTERVALLE

wir wissen nun wie groß die Variabilität von Mittelwerten sein kann und welche Form ihre Verteilung hat

weiters wissen wir: (s. Eigenschaften der Normalverteilung)
ca. 68% von Stichprobenmittelwerten liegen $\pm 1\sigma$ von μ
ca. 95% von Stichprobenmittelwerten liegen $\pm 2\sigma$ von μ

(tatsächlich 95% sind es bei $\pm 1{,}96\sigma$)

Fragestellung 1 (*Caridex*) war:

Wenn ich einen Mittelwert errechnet habe, kann ich dann daraus schließen wo das unbekannte μ liegt ?

Problem: man kennt μ nicht $\rightarrow$ ersetzen durch aktuellen Stichprobenmittelwert ("schätzen")
(man kennt auch σ nicht - später: *t* -Verteilung)

KONFIDENZINTERVALL FÜR MITTELWERTE:
(bei bekanntem Wert der Varianz in Population σ)

$$KI: \quad \bar{x} \pm c \qquad c = z_{(\gamma+1)/2} \frac{\sigma}{\sqrt{n}}$$

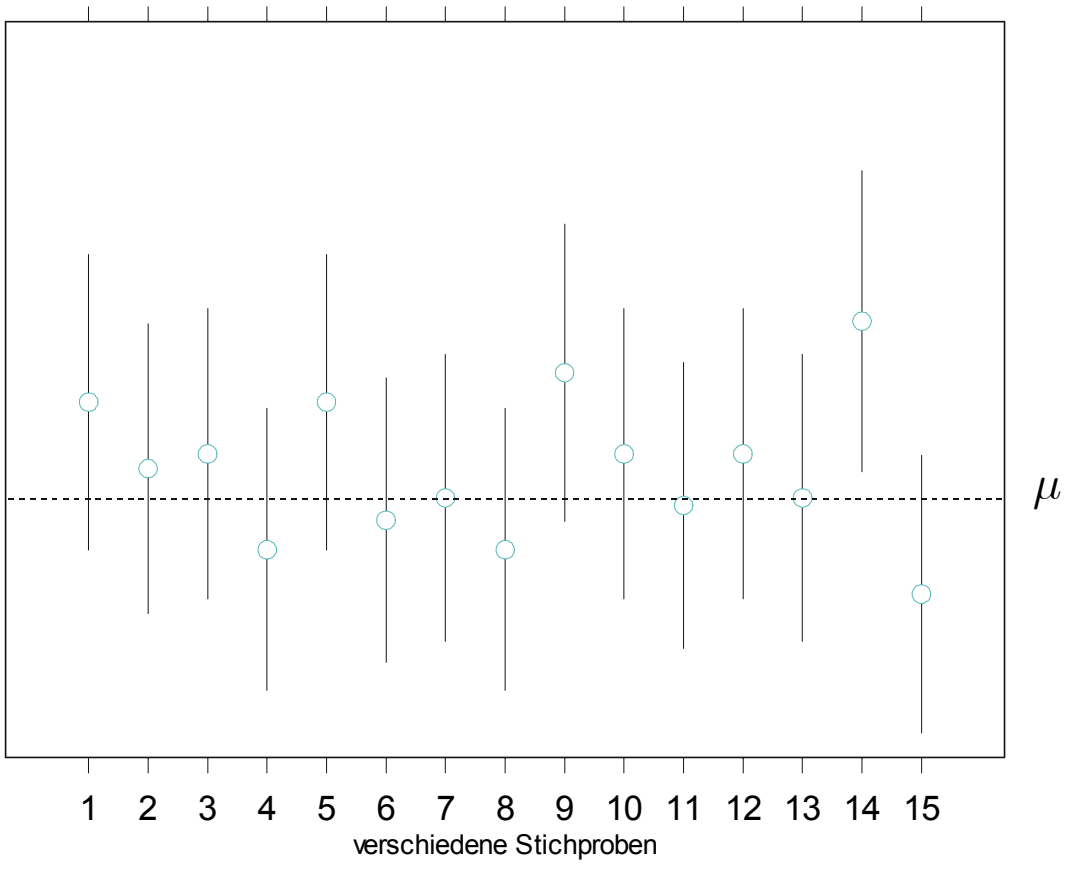
dann die Schlussfolgerung: wenn ich ein Intervall berechnet habe, dann ist es sehr plausibel, dass das unbekannte μ in diesem Intervall liegt

„plausibel“ wird durch das sog. Konfidenzniveau oder Sicherheitsniveau γ ausgedrückt, üblich sind 95% oder 99%

Interpretation von Konfidenzintervallen:

- ziehen sehr vieler Stichproben
- jeweils berechnen des Mittelwerts und des entsprechenden Konfidenzintervalles
- dann: zu einem Anteil von γ würden diese Konfidenzintervalle den tatsächlichen Wert μ aus der Population einschließen
- für eine einzelne Stichprobe kann man keine exakten Angaben machen

Beispiel:
wenn man 15 verschiedene Stichproben erheben würde, könnte man die folgenden 15 Konfidenzintervalle erhalten (die strichlierte Linie zeigt den unbekannten Wert für μ



händische Berechnung eines Konfidenzintervall für einen Mittelwert bei bekannter Varianz

Beispiel: Getränk in 1 Liter Flaschen (1.Stichprobe)
die Abfüllmaschine ist so justiert, dass die Abfüllmenge normalverteilt ist mit $\mu = 1$ und $\sigma = 0.02$

Daten:

1.012	0.993	0.980	0.971	1.002
0.977	1.002	1.007	0.991	0.985

$\overline{x} = 0.992$

95%-Konfidenzintervall:

$$c = z_{0.975} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{0.02}{\sqrt{10}} = 0.012$$

KI: $0.992 \pm 0.012 \quad \rightarrow \quad [\,0.970\,;\,1.004]$

will man ein 99% Konfidenzintervall berechnen, dann ist der z-Wert nicht 1,96 sondern $z_{0.995} = 2{,}58$

Anmerkung:

eine Berechnung mittels der Statistikprogramme SPSS bzw. R ist nur für den (realistischeren) Fall unbekannter Varianz vorgesehen (siehe nächster Abschnitt)

KONFIDENZINTERVALL BEI UNBEKANNTER VARIANZ

normalerweise Varianz in der Grundgesamtheit nicht bekannt

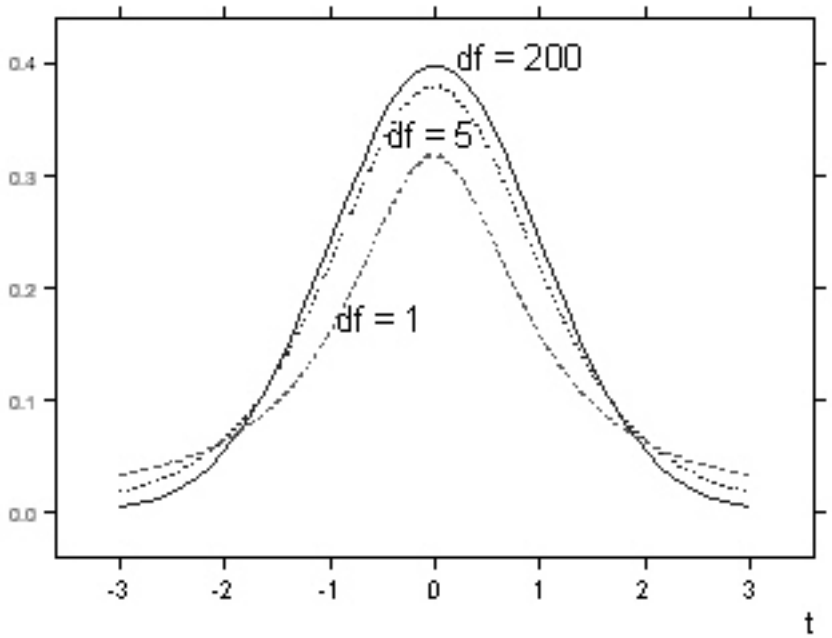
muss aus der Stichprobe geschätzt werden: $s^2 = \hat{\sigma}^2$

$\hat{\sigma}^2$ (sprich "Sigma Quadrat Dach")
"Dach" ("$\wedge$") bedeutet, dass der Parameter geschätzt, d.h. aus Stichprobenwerten errechnet wurde

man verwendet für s^2 dann die Formel $\frac{1}{n-1}\sum(x_i - \bar{x})^2$

man verwendet nicht die Normalverteilung sondern die sogenannte ***t* -Verteilung**

sieht aus wie Normalverteilung, nur breiter
hat Freiheitsgrade *df* (wie χ^2-Verteilung)
wenn $df \rightarrow \infty$, *t* -Verteilung $\rightarrow$ Normalverteilung



KONFIDENZINTERVALL BEI UNBEKANNTER VARIANZ

$$KI: \quad \bar{x} \pm c \qquad c = t_{((\gamma+1)/2;\, df)} \frac{\hat{\sigma}}{\sqrt{n}}$$

die Anzahl der Freiheitsgrade ergibt sich aus:

$$df = n - 1$$

t - Werte für 95% Konfidenzintervalle

df	***t* -Wert**	***df***	***t* -Wert**	***df***	***t* -Wert**	***df***	***t* -Wert**
1	12.71	11	2.20	21	2.08	40	2.02
2	4.30	12	2.18	22	2.07	50	2.01
3	3.18	13	2.16	23	2.07	60	2.00
4	2.78	14	2.14	24	2.06	70	1.99
5	2.57	15	2.13	25	2.06	80	1.99
6	2.45	16	2.12	26	2.06	90	1.99
7	2.36	17	2.11	27	2.05	100	1.98
8	2.31	18	2.10	28	2.05	150	1.98
9	2.26	19	2.09	29	2.05	200	1.97
10	2.23	20	2.09	30	2.04	∞	1.96

zurück zu Fragestellung 1:

Gesucht ist der Mittelwert der in einer Woche mit Caridex behandelten Zähne **für alle** in Frage kommenden Zahnärzte.
In welchem Bereich kann man diesen Mittelwert erwarten ?

aus Stichprobe von n = 400:

$\overline{x} = 4.015$
$\hat{\sigma} = s = 1.76$ (berechnet mit 1/(n-1))

95%-Konfidenzintervall: $KI: \ \overline{x} \pm c \qquad c = t_{(0.975;\, df)} \frac{\hat{\sigma}}{\sqrt{n}}$

df = 399 (df schon sehr groß, daher gleicher Wert wie NV)

$\rightarrow$ t-Wert = 1.96

$$c = 1.96 \frac{1.76}{\sqrt{400}} = 1.96 \cdot 0.088 = 0.172$$

Ergebnis: KI: [3,843; 4,187]

bei einem Konfidenzniveau von 95% ist zu erwarten, dass die durchschnittliche Zahl behandelter Zähne pro Woche für alle in Frage kommenden Zahnärzte zwischen 3,843 und 4,187 liegt

Berechnung mit Statistikprogrammen:

SPSS

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
behandelte Zähne	400	100,0%	0	,0%	400	100,0%

Univariate Statistiken

			Statistik	Standard fehler
behandelte Zähne	Mittelwert		4,01	.09
	95% Konfidenzintervall des Mittelwerts	Untergrenze	3,84	
		Obergrenze	4,19	

Interpretation des Ergebnisses: siehe oben

R

in R werden Konfidenzintervalle für Mittelwerte gemeinsam mit Tests, ob ein Mittelwert einer Vorgabe entspricht (siehe Fragestellung 2), berechnet

```
One Sample t-test

data: CAVITIES
t = 45.5152, df = 399, p-value = < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
3.841581 4.188419
sample estimates:
mean of x
4.015
```

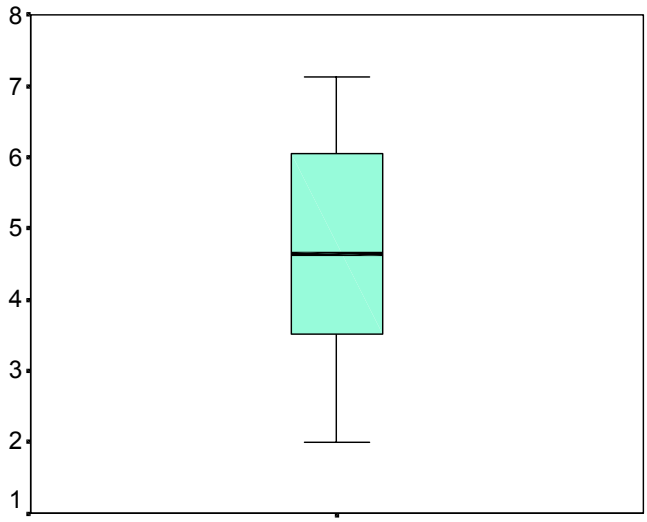
FRAGESTELLUNG 2:

Ist ein Mittelwert anders (ungleich, kleiner, oder größer) als eine bestimmte Vorgabe ?

Diese Frage stellt man, wenn man prüfen möchte, ob die Größe eines Durchschnittswertes in der Grundgesamtheit einer bestimmten Annahme entspricht.

Beispiel: Erreicht man mit einer bestimmten Diät eine durchschnittliche Gewichtsabnahme von mehr als 5 kg?

Stichprobe n = 20 ergab: $\overline{x} = 4.73$



Widerspricht dies der Vorgabe, dass die durchschnittliche Gewichtsabnahme mehr als 5kg beträgt ?

wieder 2 Lösungsansätze:

- wenn Varianz in Population bekannt ist
- wenn Varianz aus Stichprobe geschätzt werden muss

Exkurs: Testen von Hypothesen

TESTEN VON HYPOTHESEN

(Prüfen von Fragestellungen)

Definition:

Hypothese ist eine Vorhersage über einen bestimmten Aspekt einer oder mehrerer Variable

Generelle Problemstellung:

Liefern Stichprobendaten genügend Evidenz, dass die jeweilige Fragestellungen positiv beantwortet werden können

2 Arten von Hypothesen:

- Nullhypothese H_0
- Alternativhypothese H_A

Nullhypothese:

hierbei wird immer ein einzelner Wert für einen Parameter festgelegt

Beispiel:

man will prüfen, ob eine bestimmte Diät innerhalb von 3 Wochen zu einer durchschnittlichen Gewichtsreduktion von 5 kg führt

$$H_0 : \mu = 5$$

Alternativhypothese:

- ist die wichtigere Hypothese (das was man eigentlich wissen will)
- ist logisches Gegenteil der Nullhypothese
- wird akzeptiert, wenn die Nullhypothese verworfen wird
- hier wird ein Bereich für den Wert eines Parameters festgelegt

3 mögliche Formen
Art hängt von Vorwissen oder Interesse ab

Beispiel:

- Konsumentenschutzverein vermutet Schwindel, glaubt, Gewichtsverlust ist geringer als angegeben

$$H_A : \mu < 5$$

- Hersteller ist von Diät überzeugt, möchte zeigen, daß Abnahme durchschnittlich mehr als 5kg beträgt

$$H_A : \mu > 5$$

- Ernährungswissenschaftler möchte nur prüfen, ob die Angabe richtig ist

$$H_A : \mu \neq 5$$

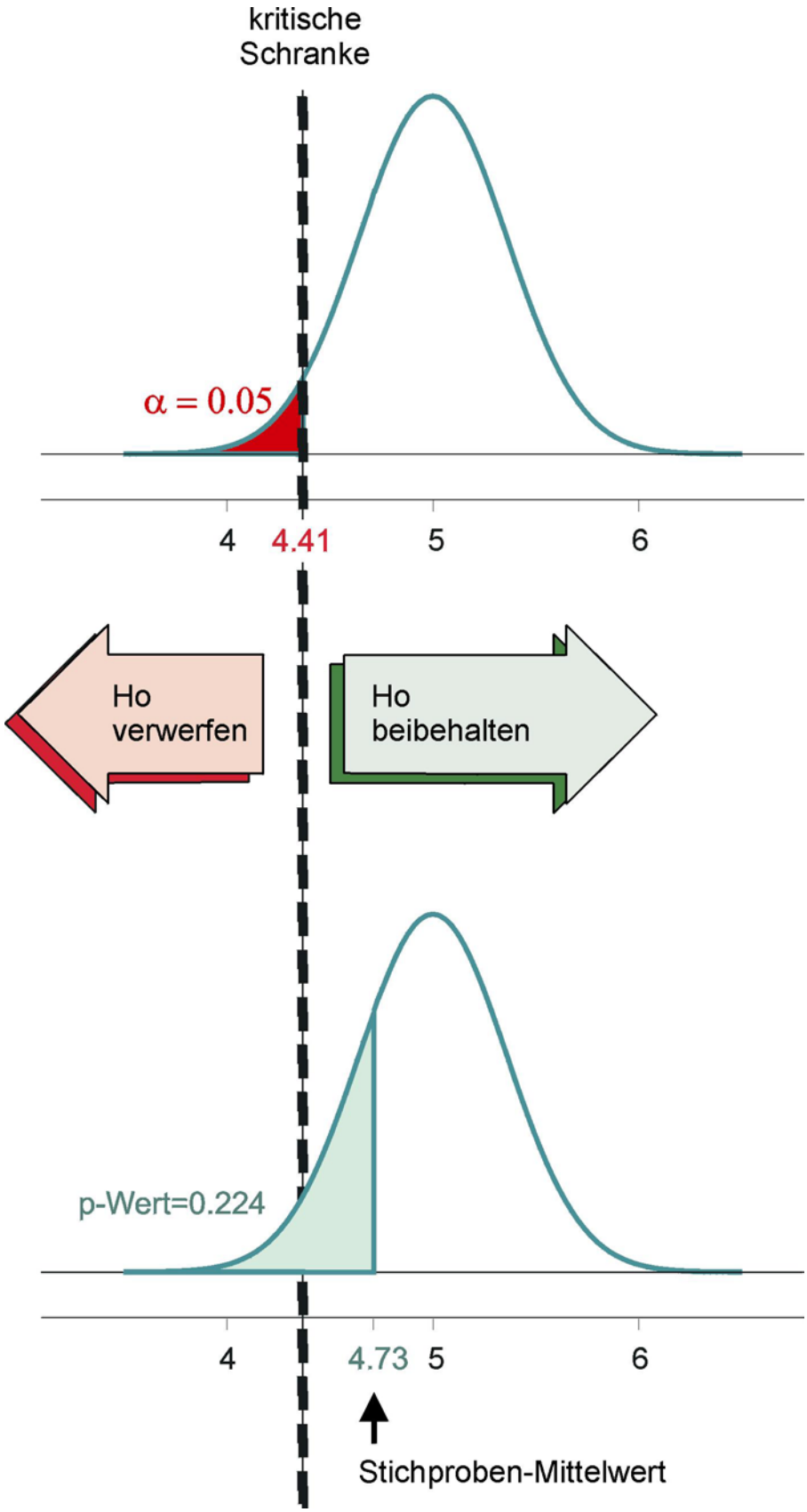
eine Stichprobe von 20 Personen könnte ergeben: $\overline{x} = 4{,}73$
spricht dies für oder gegen die Nullhypothese ?

Entscheidungen nur mit bestimmter Sicherheit:

beim Testen verwendet man das Signifikanzniveau α („Fehlerniveau“) und nicht (wie bei Konfidenzintervallen) das „Sicherheitsniveau“ γ , d.h. $\alpha = 1 - \gamma$, also meistens 0.05

Testen mittels kritischer Schranke bzw. p-Wert:

$H_A : \mu < 5$



Prüfen der Nullhypothese:

mittels p-Wert:

unter der Annahme, dass die Nullhypothese stimmt, kann man ausrechnen, wie wahrscheinlich es ist, einen bestimmten Stichprobenmittelwert (allgemein: eine bestimmte Teststatistik) zu erhalten (oder einen noch extremeren Wert in der Richtung, wie Alternativhypothese formuliert ist)
p-Wert wird auch Überschreitungswahrscheinlichkeit genannt

aufgrund dieser Wahrscheinlichkeit ("p-Wert" oder "Signifikanz" in R bzw. SPSS) wird beurteilt, wie plausibel die Nullhypothese ist

je nach Formulierung der Alternativhypothese gibt es drei verschiedene Situationen (s. Graphik)

bei vorgegebenem Signifikanzniveau α (z.B. $\alpha = 0{,}05$)

"Signifikanz (p-Wert)" < α $\Rightarrow$ Nullhypothese "verwerfen"

sonst: Nullhypothese wird "beibehalten"

mittels kritischem Wert:

kann man den p-Wert nicht ausrechnen, dann beurteilt man die Nullhypothese mittels eines Vergleichs der Teststatistik mit dem kritischen Wert aus der entsprechenden Tabelle:

Teststatistik > krit.Wert $\Rightarrow$ Nullhypothese verwerfen

Elemente eines statistischen Tests (allgemein):

Voraussetzungen:
- Art der Daten (nominal, ordinal, metrisch)
- Verteilung (oft Forderung nach Normalverteilung)
- Zufallsstichproben und Stichproben genügend groß

Hypothesen: (Formulieren von H_0 und H_A)
- Zusammenhangshypothesen
- Unterschiedshypothesen
- ...

Teststatistik (Prüfgröße):
errechneter Wert aus Stichprobe, für den man einen p-Wert berechnen kann

Signifikanzniveau (Irrtumswahrscheinlichkeit): α
mit welcher Sicherheit ($1-\alpha$) soll die Entscheidung getroffen werden

p-Wert bzw. kritischer Wert:
gibt Plausibilität für Teststatistik unter der Nullhypothese an

Schlussfolgerung und Interpretation:

- Angabe, ob H_0 zugunsten einer bestimmten H_A verworfen wird – oder ob sie beibehalten wird
- bei welchem Signifikanzniveau
- Was bedeutet das für die Fragestellung

zum Beispiel aus Fragestellung 2:

Erreicht man mit einer bestimmten Diät eine durchschnittliche Gewichtsabnahme von mehr als 5 kg?
(bei Irrtumswahrscheinlichkeit von α=0.05)

Stichprobe n = 20 ergab: $\bar{x} = 4.73$, $s = 1.6$

2 mögliche Situationen:

- die Varianz σ^2 in der Population ist bekannt:

  Test bei bekannter Varianz: *Einstichproben z-Test*
  Verwenden der Normalverteilung
  in der Praxis eher selten

- die Varianz σ^2 in der Population ist nicht bekannt, und muß daher aus der Stichprobe geschätzt werden

  Test bei unbekannter Varianz: *Einstichproben t-Test*
  Verwenden der *t* - Verteilung

zur Berechnung:

mittels Statistikprogrammen:

- üblicherweise ist in Statistikpaketen, wie auch in SPSS und R, nur die „realistischere“ Möglichkeit unbekannter Varianz direkt implementiert
- wie schon vorher erfolgt die Berechnung auch hier nicht mittels verwendung kritischer Schranken sondern über p-Werte

„händische Berechnung“
erfolgt unter Verwendung von Tabellen, im nächsten Abschnitt sind alle vier Möglichkeiten dargestellt

Beantwortung der Fragestellung 2

Erreicht man mit einer bestimmten Diät eine durchschnittliche Gewichtsabnahme von mehr als 5 kg? (α=0.05)
$H_0 : \mu = 5,\ H_A : \mu > 5$

Stichprobe n = 20 ergab: $\bar{x} = 4.73,\ s = 1.6$

R

```
> t.test(x,mu=5,alternative="greater")

        One Sample t-test

data:  abnahme
t = -0.7414, df = 19, p-value = 0.7662
alternative hypothesis: true mean is greater than 5
95 percent confidence interval:
 4.116767      Inf
sample estimates:
mean of x
 4.734953
```

Ergebnis:

- p-Wert von 0.766 ist größer als 0.05, daher: H_0 beibehalten
- man erreicht mit dieser Diät keine durchschnittliche Gewichtsabnahme von mehr als 5 kg

SPSS

Statistik bei einer Stichprobe

	N	Mittelwert	Standardabweichung
Gewichtsabnahme	20	4.73	1,60

Test bei einer Sichprobe

	Testwert = 5		
	T	df	Sig. (2-seitig)
Gewichtsabnahme	-,741	19	,468

Anmerkungen:

- SPSS gibt nur den p-Wert für eine zweiseitige Alternativhypothese aus, unsere Fragestellung (H_A) war aber einseitig
- den einseitigen p-Wert erhält man durch halbieren des zweiseitigen $\rightarrow$ einseitiger p-Wert = 0.234
- aber Achtung: der Stichprobenmittelwert $\bar{x} = 4.73$ liegt nicht in der Richtung der Fragestellung (Vergleich H_0 mit H_A)
- daher ist der **richtige p-Wert:** 1 – 0.234 = **0.766**

Ergebnis (wie oben)

- p-Wert von 0.766 ist größer als 0.05, daher: H_0 beibehalten

- man erreicht mit dieser Diät keine durchschnittliche Gewichtsabnahme von mehr als 5 kg

händische Berechnung:

"Ein-Stichproben z-Test (bei **bekannter Varianz)":**

TEST mittels kritischem Wert:

1. nachsehen des kritischen Wertes z_{krit} in der NV-Tabelle
   - bei einseitigem Test: bei 1-α (z.B. bei 0.95)
   - bei zweiseitigem Test: bei 1-α/2 (z.B. bei 0.975)
2. berechnen der Teststatistik: $$T = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$$
3. Vergleich kritischer Wert mit Teststatistik

   im zweiseitigen Fall ($H_A : \mu \neq \mu_0$):

   - wenn der Absolutbetrag der **Teststatistik** gleich oder **größer** als der kritische Wert ist , d.h. $|T| \geq z_{krit}$,
     dann **H_o verwerfen,** sonst **H_o beibehalten**

   bei einseitigen Alternativhypothesen (Achtung auf Richtung):

   - $H_A : \mu > \mu_0$
     H_o verwerfen, wenn die **Teststatistik** gleich oder **größer** als der kritische Wert ist , d.h. $T \geq z_{krit}$, sonst **H_o beibehalten**

   - $H_A : \mu < \mu_0$
     H_o verwerfen, wenn die **Teststatistik** gleich oder **kleiner** als der kritische Wert ist , d.h. $T \leq z_{krit}$, sonst **H_o beibehalten**

TEST mit p-value: Wahrscheinlichkeit für H_0 berechnen

Vergleich dieser Wahrscheinlichkeit mit α

- Ist p-Wert („die Wahrscheinlichkeit für H_0“) **größer** als α, dann **H_o beibehalten,** sonst **H_o verwerfen**

Beispiel: Fragestellung 2 (Varianz als bekannt vorausgesetzt)

aus Erfahrung weiß man, dass die Standardabweichung bei solchen Diäten σ = 1.6 beträgt

Test mittels kritischem Wert

- Teststatistik (standardisieren)

$$T = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{4.73 - 5}{\frac{1.6}{\sqrt{20}}} = -0.755$$

- kritischer Wert: bei 0.95 nachsehen ergibt z_{krit} = 1,645

- Vergleich: -0.755 ist kleiner als 1,645 daher Nullhypothese beibehalten, Alternativhypothese nicht akzeptieren: man erreicht mit dieser Diät keine durchschnittliche Gewichtsabnahme von mehr als 5 kg

TEST mit p-value:

Wie groß ist die Wahrscheinlichkeit, dass man ein $\overline{x} = 4.73$ oder einen größeren Wert erhält ($H_A : \mu > 5$), wenn in Wirklichkeit $\mu = 5$ ($H_0 : \mu = 5$) ?

- Teststatistik *T* berechnen: -0.755

  - nachsehen in der Normalverteilungstabelle für *T*: Fläche bis T ist 0.224

  da aber "mehr als 5kg" gefragt war, benötigen wir die Fläche für Werte > T, d.h. p = 1 - 0.224 = 0.776

  - Entscheidung für oder gegen die Nullhypothese: 0.776 ist größer als 0.05 daher: H_O beibehalten

händische Berechnung:

"Ein-Stichproben *t*-Test (bei **unbekannter Varianz)"**

geht im Prinzip gleich, nur:

- s statt σ
  (einsetzen der Stichprobenstandardabweichung als Schätzwert für unbekannte Populationsvarianz)
- *t* - Verteilung statt NV
- Freiheitsgrade müssen berücksichtigt werden: df=n-1

bei händischer Berechnung wird man den Test eher mittels kritischer Schranke durchführen, da die *t* – Verteilung auf Grund der Freiheitsgrade nicht so ausführlich tabelliert sein kann, wie die Normalverteilung

Anmerkung:
in manchen Programme, wie z.B. Excel, sind Funktionen implementiert, die einem für bestimmte Werte der Teststatistik und beliebige Freiheitsgrade die Wahrscheinlichkeiten u.a. der t-Verteilung liefern, dann ist auch Test mittels p-Werten möglich