Vorlesung 10

Inhalt

- Kenngrößen von Daten
- Median, Quantile, Mittelwert, Stichprobenvarianz
- Normalverteilung

Lernziele

- Wichtige Kenngrößen von Daten kennen und berechnen können
- Die Normalverteilung kennen

Benötigte Vorkenntnisse

- Häufigkeiten, Daten, Funktionen

Daten

Helgoländer Tiefe Rinne,
Fang vom 6.9.1988

Carapaxlänge (mm):
Nichteiertragende Weibchen ($n = 215$)

2,9	3,0	2,9	2,5	2,7	2,9	2,9	3,0
3,0	2,9	3,4	2,8	2,9	2,8	2,8	2,4
2,8	2,5	2,7	3,0	2,9	3,2	3,1	3,0
2,7	2,5	3,0	2,8	2,8	2,8	2,7	3,0
2,6	3,0	2,9	2,8	2,9	2,9	2,3	2,7
2,6	2,7	2,5	.	.	.	.	.

Information aus den Daten

Es ist oft möglich, das Wesentliche an einer Stichprobe mit ein paar Zahlen zusammenzufassen.

Wesentlich:

- Wie groß? Lageparameter
- Wie variabel? Streuungsparameter

Median

Der Median:

die Hälfte der Beobachtungen sind kleiner,
die Hälfte sind größer*.

Der Median ist
das 50%-Quantil
der Daten.

*Diese „Definition“ genügt für die meisten praktischen Fälle (und ist intuitiv sehr plausibel), die mathematisch präzise Definition siehe die folgende Folie.

Nachtrag:

Der Median:

die Hälfte der Beobachtungen sind kleiner,
die Hälfte sind größer*.

Der Median ist das 50%-Quantil der Daten.

*Eine mathematisch präzise Definition:

Seien n der Größe nach geordnete Beobachtungswerte $y_1 \leq y_2 \leq \cdots \leq y_n$ gegeben, dann ist (der/ein) Median m ein Wert, so dass höchstens $n/2$ Werte $\geq m$ und höchstens $n/2$ Werte $\leq m$ sind.

Falls n ungerade ist, sagen wir $n = 2k + 1$, so ist durch diese Forderung $m = y_{k+1}$ eindeutig festgelegt, denn

$$\underbrace{y_1, y_2, \ldots, y_k}_{k \text{ Werte}} \leq y_{k+1} \leq \underbrace{y_{k+2}, y_{k+2}, \ldots, y_{2k+1}}_{k \text{ Werte}}.$$

Falls n gerade ist, sagen wir $n = 2k$, so erfüllen y_k, y_{k+1} und ggfs. auch jeder Wert zwischen y_k und y_{k+1} diese Forderung. Wenn ein konkreter Wert verlangt wird, nimmt man dann oft pragmatisch $(y_{k+1} + y_k)/2$

Die Quartile

Das erste Quartil, Q_1:
ein Viertel der Beobachtungen sind kleiner, drei Viertel sind größer.

Q_1 ist das 25%-Quantil der Daten.

Das dritte Quartil, Q_3:
drei Viertel der Beobachtungen sind kleiner, ein Viertel sind größer.

Q_3 ist das 75%-Quantil der Daten.

Auch hier: Diese „Definition“ genügt für die meisten praktischen Fälle (und ist intuitiv sehr plausibel), für eine mathematisch präzise Definition siehe die folgende Folie.

Nachtrag:

Erstes Quartil, Q_1: drei Viertel der Beobachtungen sind kleiner, ein Viertel sind größer*.

Drittes Quartil, Q_3: drei Viertel der Beobachtungen sind kleiner, ein Viertel sind größer†.

*Präziser kann man Folgendes fordern: Q_1 ist eine Zahl, so dass

- höchstens 25% der Beobachtungswerte $< Q_1$ und
- höchstens 75% der Beobachtungswerte $> Q_1$ sind.

†Präziser kann man Folgendes fordern: Q_3 ist eine Zahl, so dass

- höchstens 75% der Beobachtungswerte $< Q_3$ und
- höchstens 25% der Beobachtungswerte $> Q_3$ sind.

Quartilabstand, Quantile

Quartilabstand $Q_3 - Q_1$ ist ein Streuungsparameter

Das $p-$Quantil der Daten Q_p: Ein Anteil p der Beobachtungen sind kleiner, ein Anteil $1 - p$ ist größer.

Median und Quartile sind spezielle Quantile.

Mittelwert und Standardabweichung

Am häufigsten werden benutzt:

Lageparameter

Der Mittelwert (engl. mean) $\overline{x}$

Streuungsparameter

Die Standardabweichung s

Mittelwert und Standardabweichung

Am häufigsten werden benutzt:

Lageparameter

Der Mittelwert (engl. mean) $\overline{x}$

Streuungsparameter

Die Standardabweichung s

Notation:

Wenn die Beobachtungen $x_1, x_2, x_3, \ldots, x_n$ heißen, schreibt man oft $\overline{x}$ für den Mittelwert.

Definition: Mittelwert= $\frac{\text{Summe der Messwerte}}{\text{Anzahl der Messwerte}}$

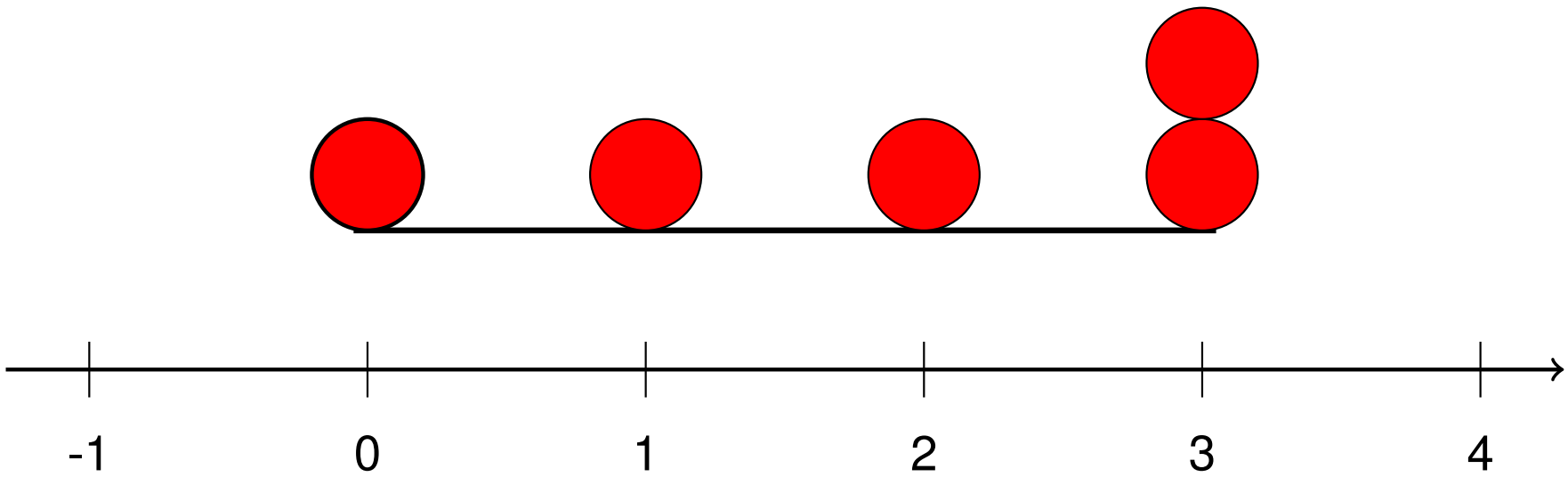
Der Mittelwert von $x_1, x_2, \ldots, x_n$ als Formel:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Beispiel: $x_1 = 3$, $x_2 = 0$, $x_3 = 2$, $x_4 = 3$, $x_5 = 1$
$\overline{x} = \text{Summe}/\text{Anzahl} = (3 + 0 + 2 + 3 + 1)/5 = 9/5 = 1{,}8$

Geometrische Bedeutung des Mittelwerts: Schwerpunkt

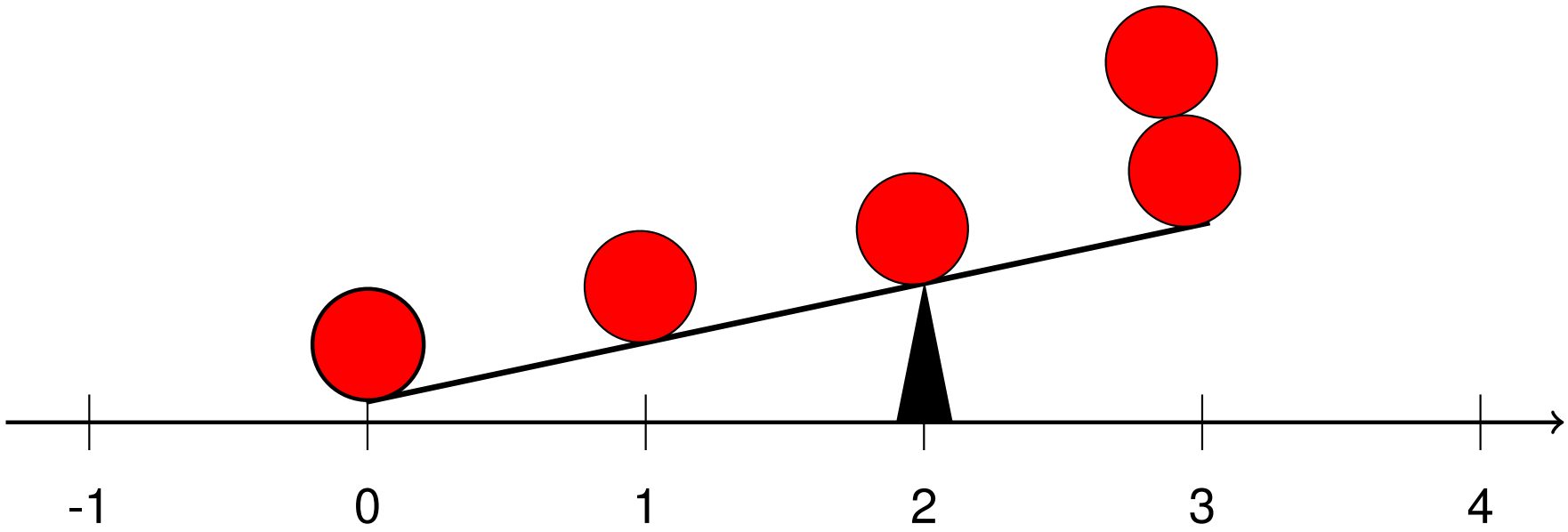
Wir stellen uns die Beobachtungen $x_1 = 3$, $x_2 = 0$, $x_3 = 2$, $x_4 = 3$, $x_5 = 1$ als gleich schwere Gewichte auf einer Waage vor:



Wo muß der Drehpunkt sein, damit die Waage im Gleichgewicht ist?

Wo muß der Drehpunkt sein, damit die Waage im Gleichgewicht ist?

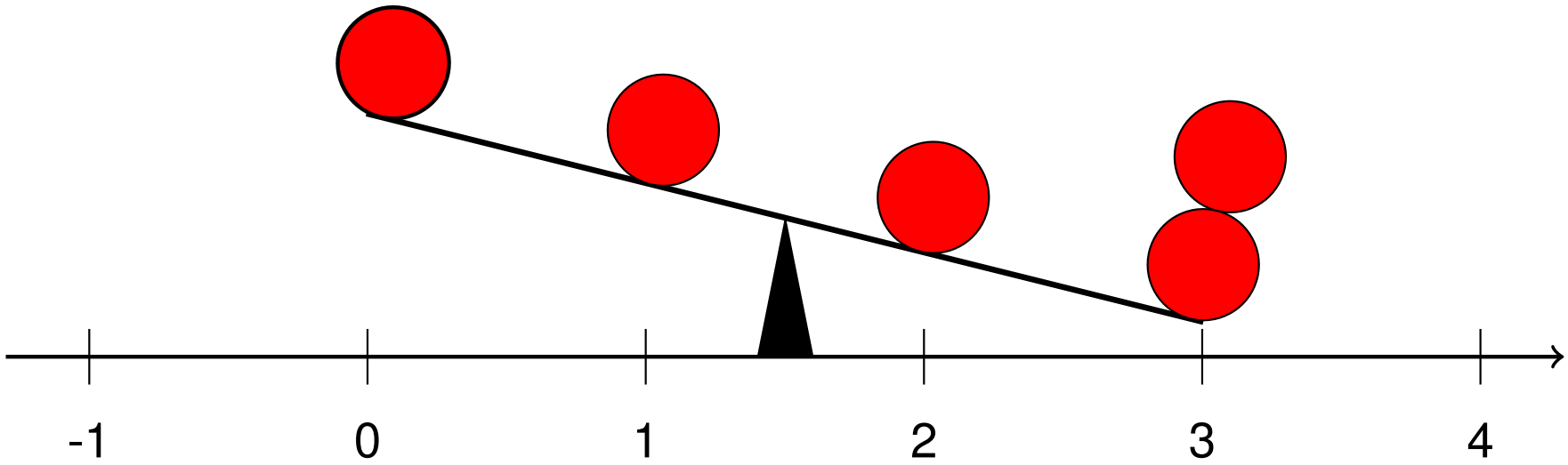
$$\overline{x} = 2\ ?$$



zu groß!

Wo muß der Drehpunkt sein, damit die Waage im Gleichgewicht ist?

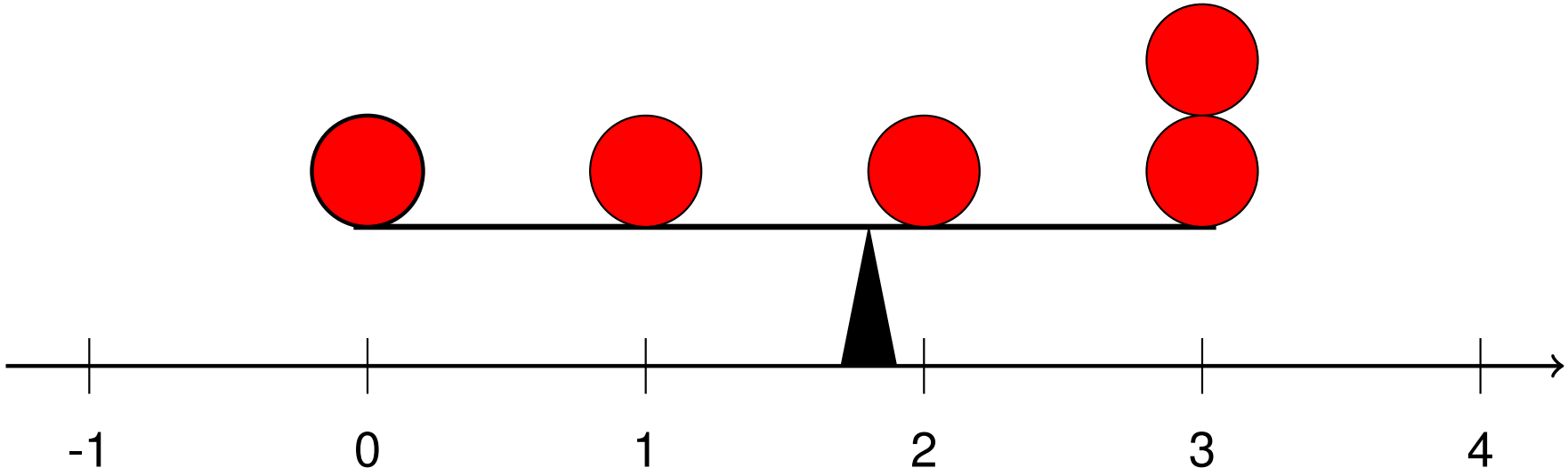
$\overline{x} = 1.5$?



zu klein!

Wo muß der Drehpunkt sein, damit die Waage im Gleichgewicht ist?

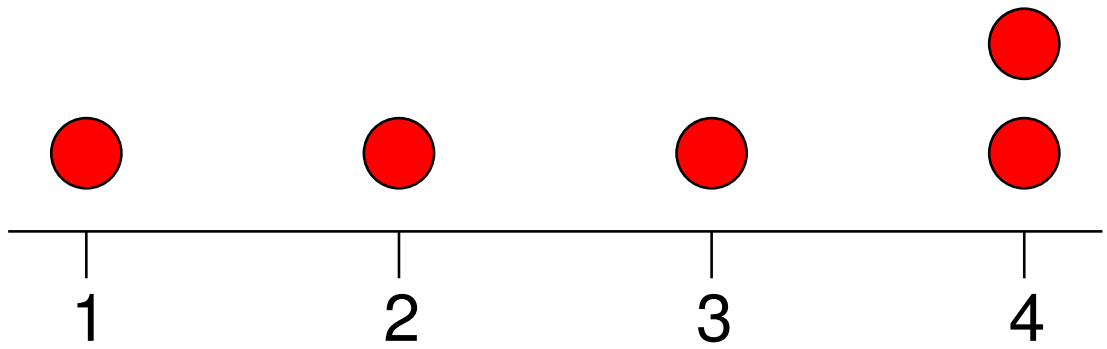
$\overline{x} = 1.8$?



richtig!

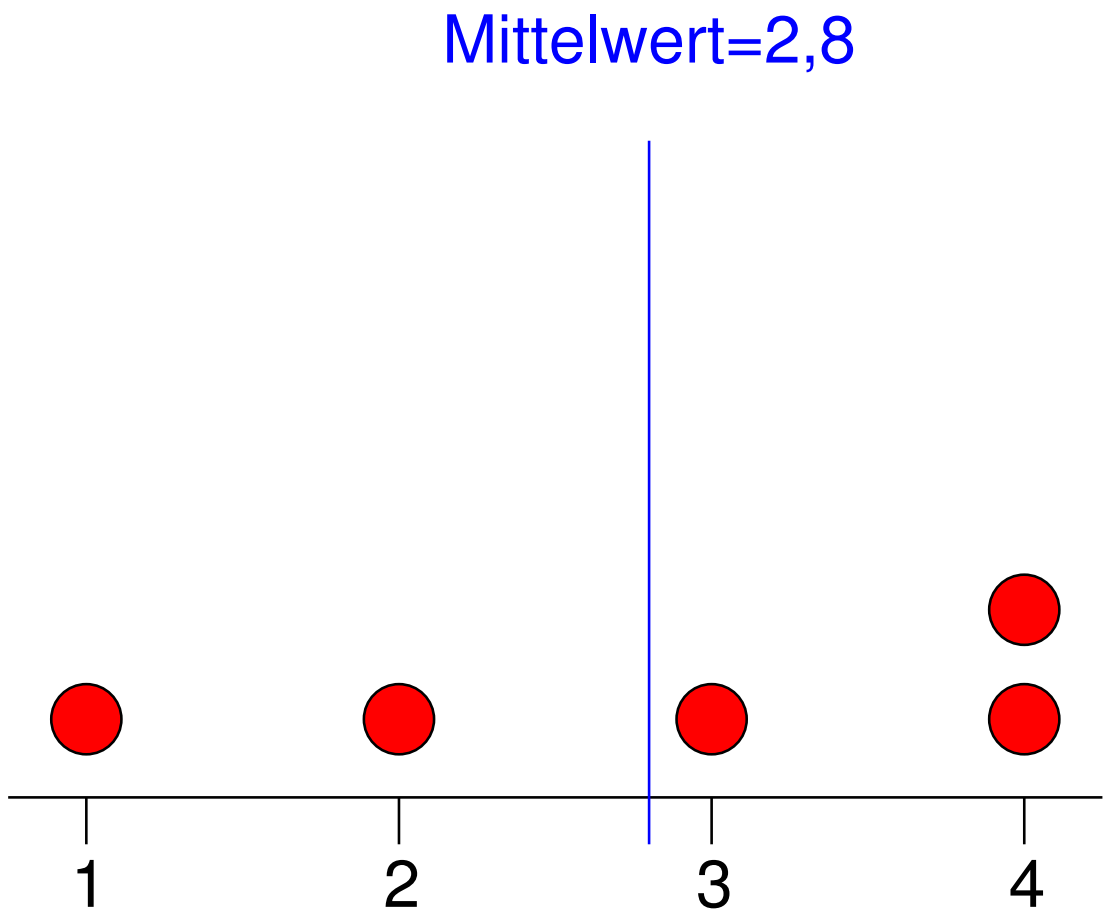
Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?



Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?

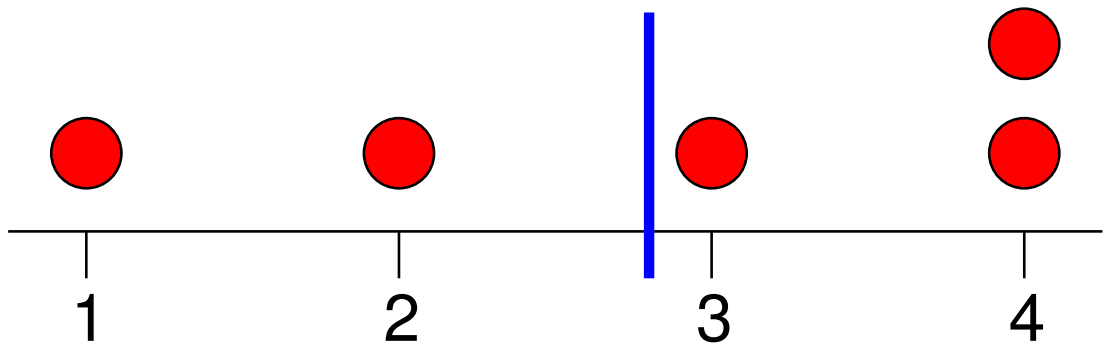


Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?

typische

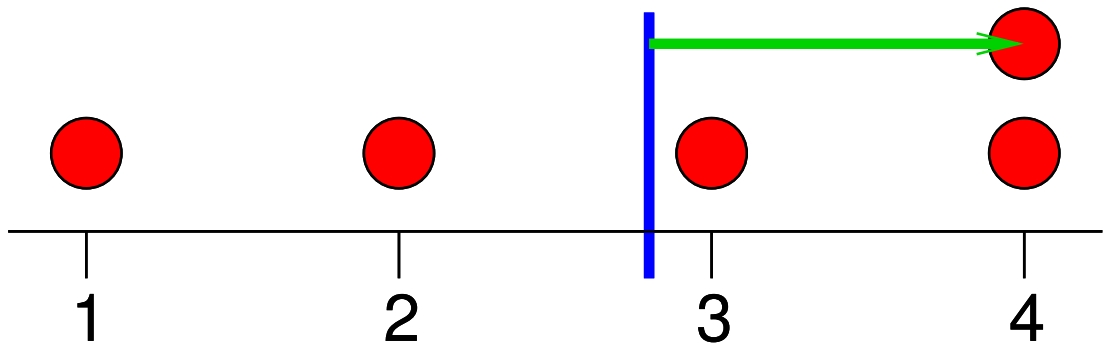
Abweichung =?



Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?

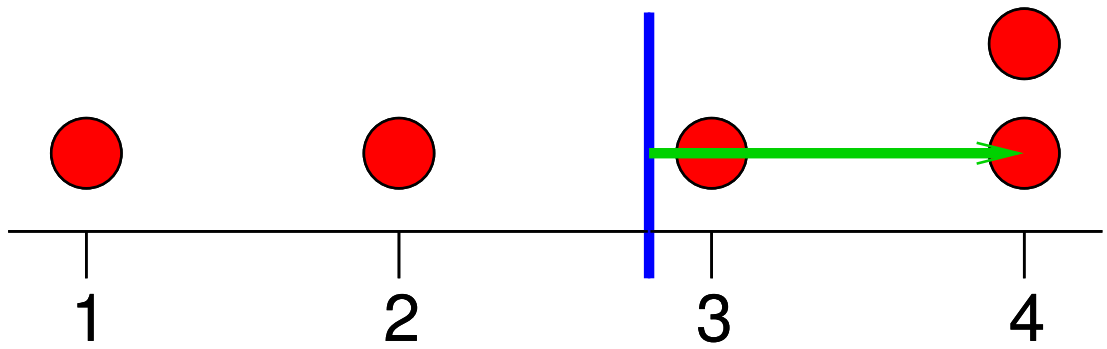
Abweichung $= 4 - 2{,}8 = 1{,}2$



Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?

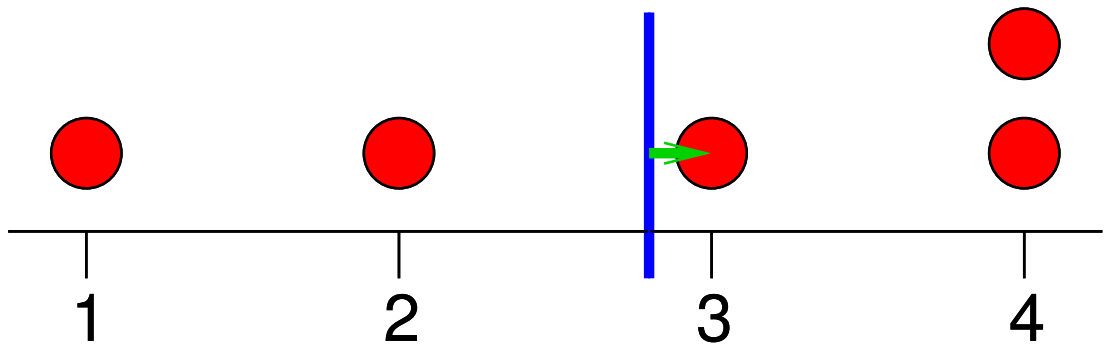
Abweichung $= 4 - 2{,}8 = 1{,}2$



Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?

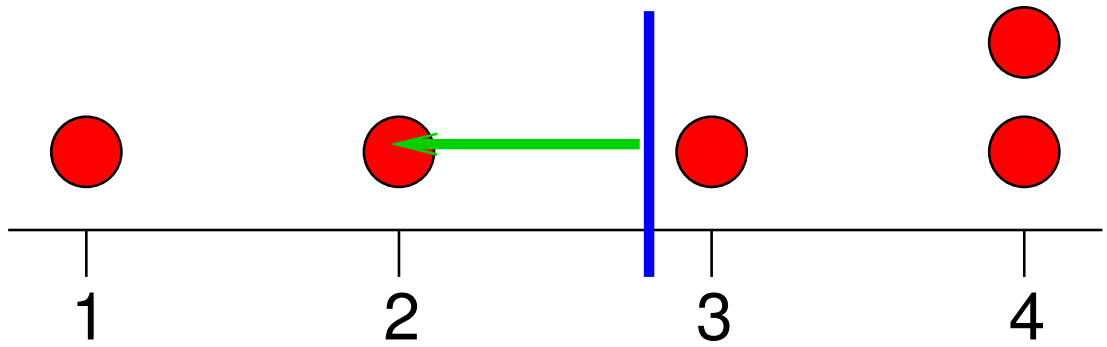
Abweichung $= 3 - 2{,}8 = 0{,}2$



Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?

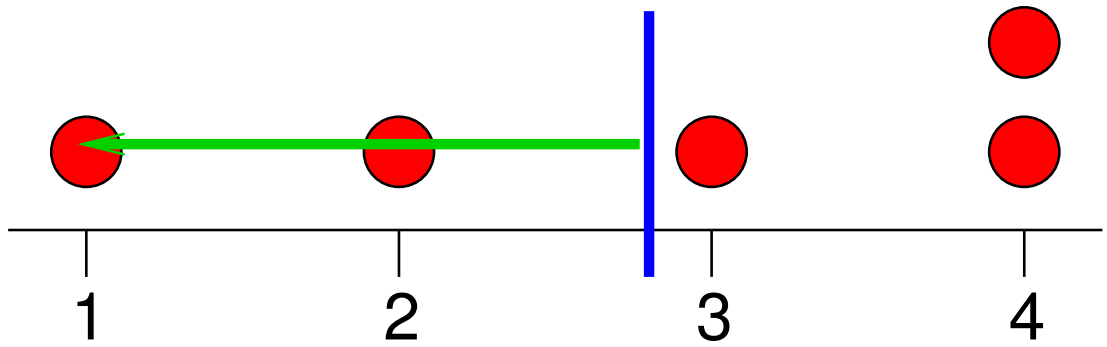
Abweichung = 2 – 2,8 = –0,8



Standardabweichung

Wie weit weicht eine typische Beobachtung vom Mittelwert ab ?

Abweichung = 1 – 2,8 = –1,8



Standardabweichung

Die Standardabweichung σ (“sigma”) [auch *SD* von engl. *standard deviation*] ist ein gewichtetes Mittel der Abweichungsbeträge, und zwar

$$\sigma = \sqrt{\text{Summe}(\text{Abweichungen}^2)/n}$$

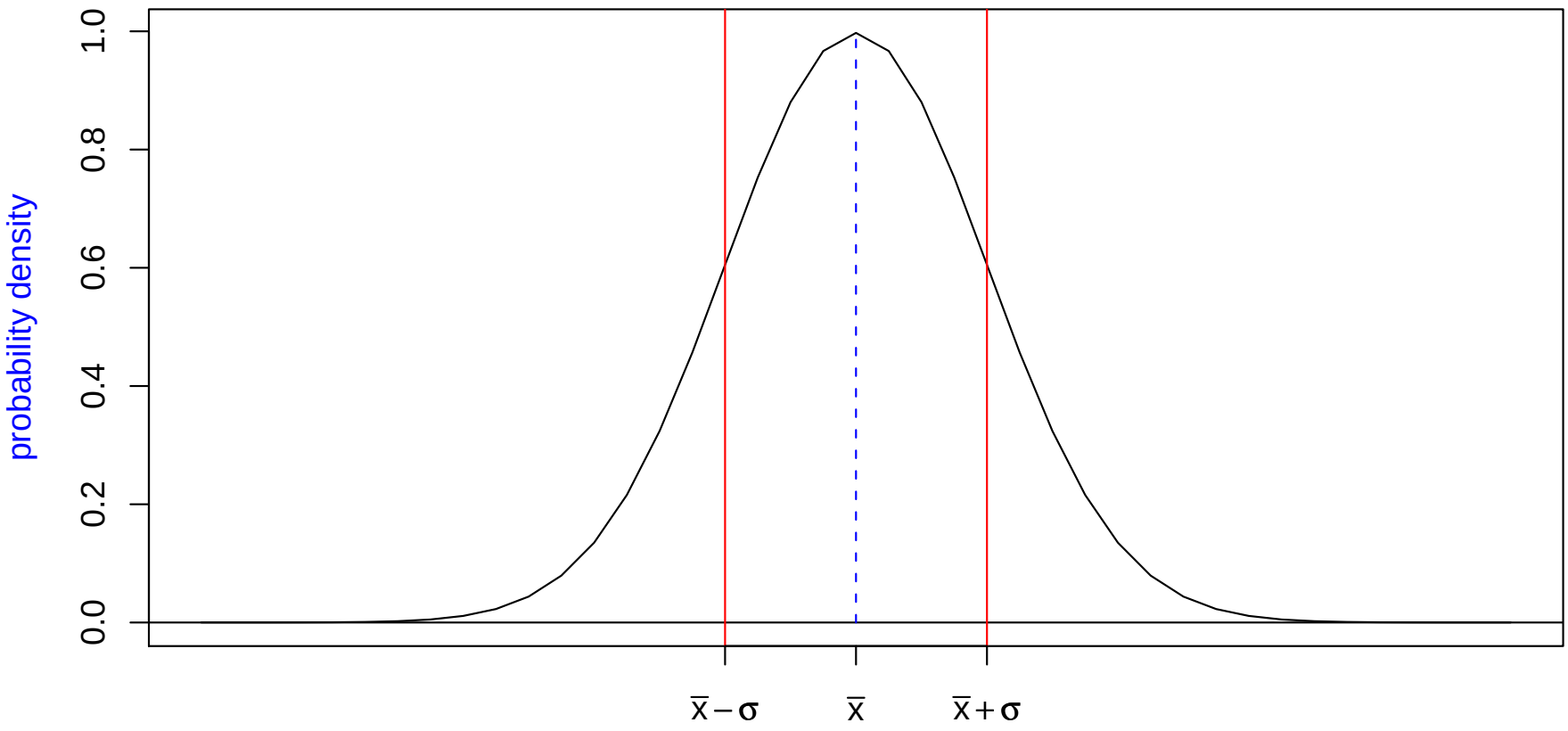
Die Standardabweichung von $x_1, x_2, \ldots, x_n$ als Formel:

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

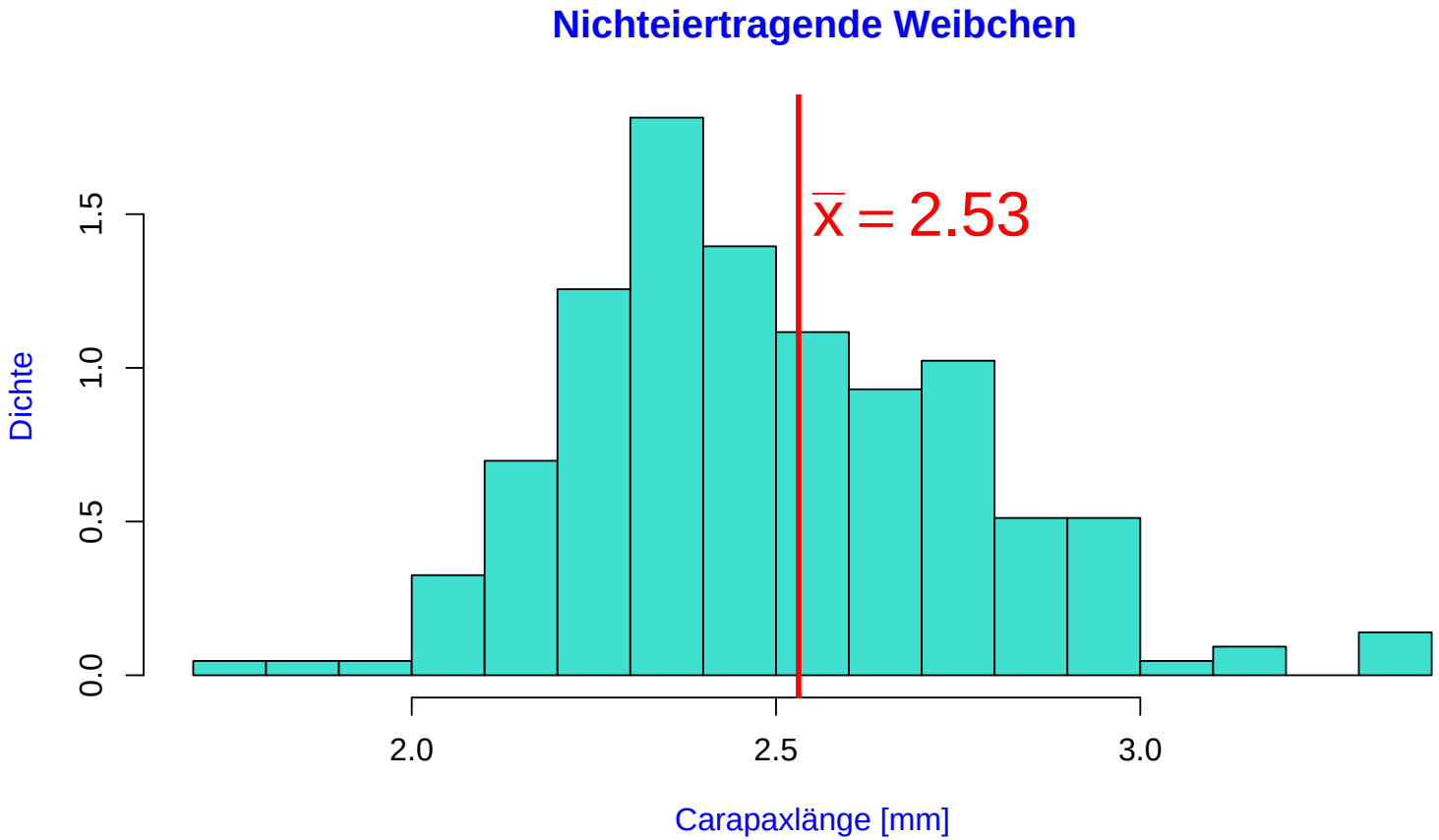
$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ heißt Varianz.

Faustregel für die Standardabweichung

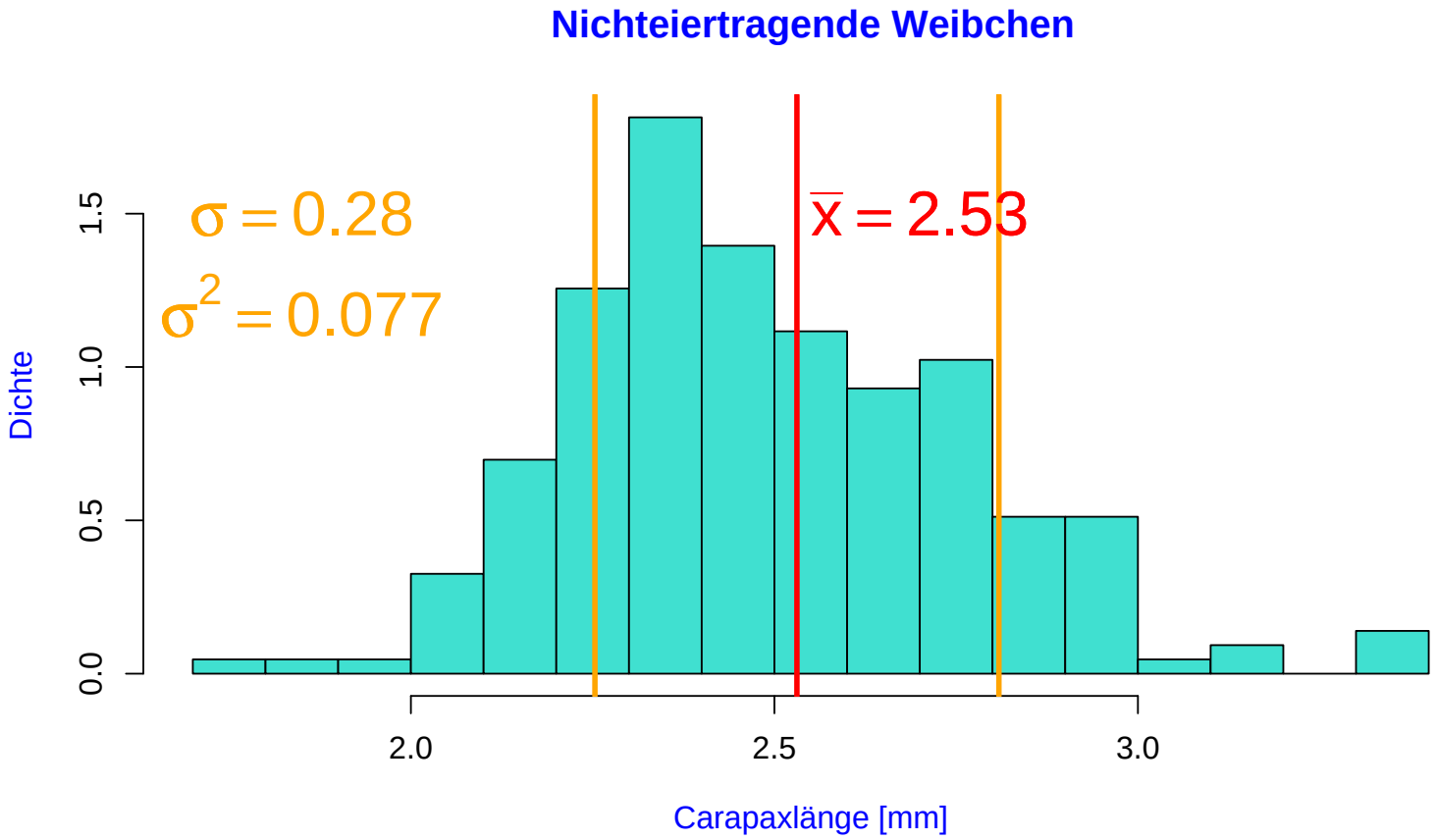
Bei ungefähr glockenförmigen (also eingipfligen und symmetrischen) Verteilungen liegen ca. 2/3 der Verteilung zwischen $\overline{x} - \sigma$ und $\overline{x} + \sigma$.



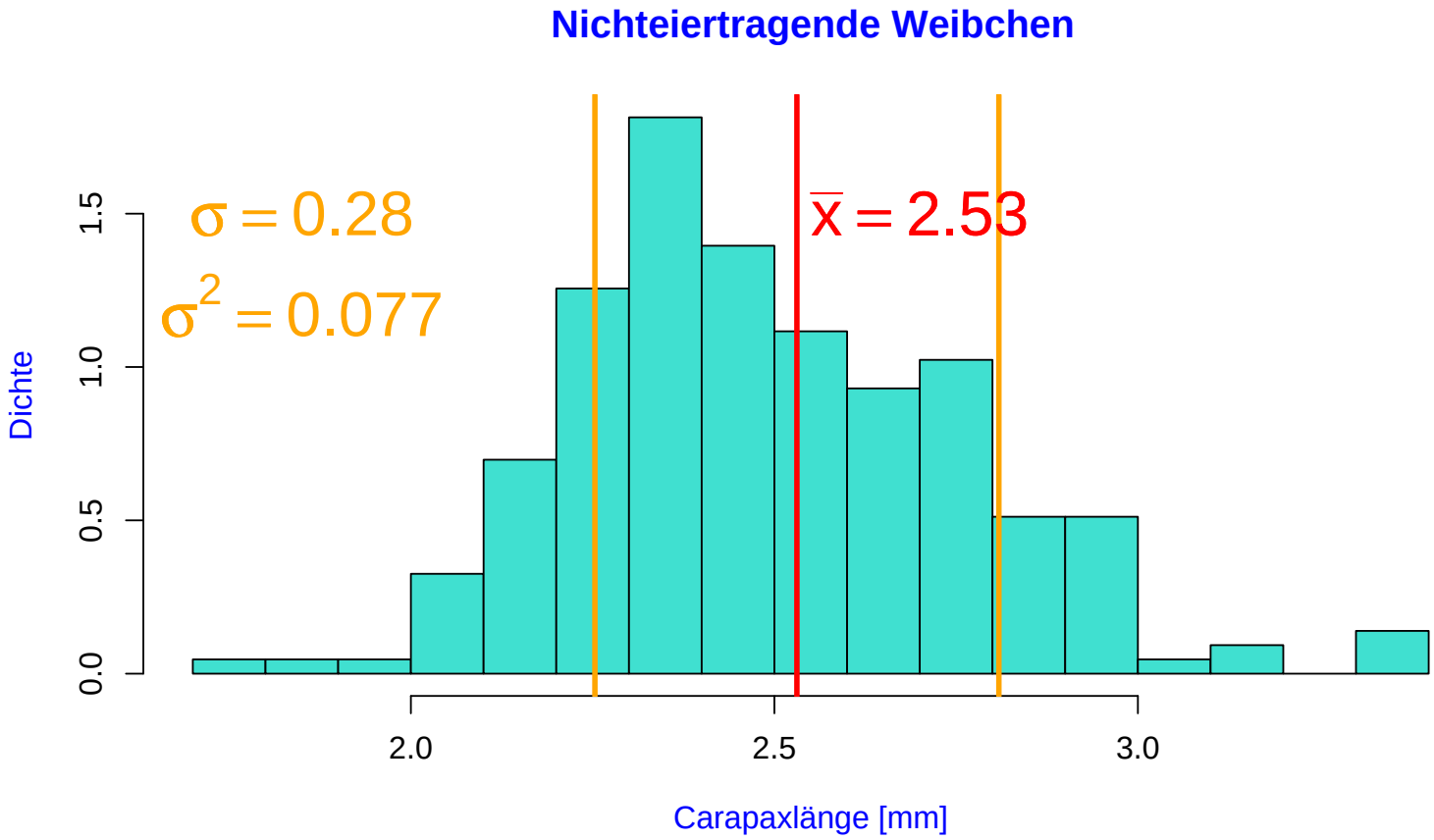
Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



Standardabweichung der Carapaxlängen nichteiertragender Weibchen vom 6.9.88



Hier liegt der Anteil zwischen $\overline{x} - \sigma$ und $\overline{x} + \sigma$ bei 72%.

Varianz der Carapaxlängen nichteiertragender Weibchen vom 6.9.88

Alle Carapaxlängen im Meer: $\mathcal{X} = (X_1, X_2, \ldots, X_N)$.
Carapaxlängen in unserer Stichprobe: $\mathcal{S} = (S_1, S_2, \ldots, S_{n=215})$
Stichprobenvarianz:

$$\sigma_{\mathcal{S}}^2 = \frac{1}{n}\sum_{i=1}^{215}(S_i - \overline{S})^2 \approx 0{,}0768$$

Können wir 0,0768 als Schätzwert für die Varianz σ_X^2 in der ganzen Population verwenden?

Ja, können wir machen. Allerdings ist $\sigma_{\mathcal{S}}^2$ im Durchschnitt um den Faktor $\frac{n-1}{n}$ $(= 214/215 \approx 0,995)$ kleiner als $\sigma_{\mathcal{X}}^2$

Varianzbegriffe

Varianz in der Population: $\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$

Stichprobenvarianz: $\sigma_S^2 = \frac{1}{n}\sum_{i=1}^{n}(S_i - \overline{S})^2$

korrigierte Stichprobenvarianz:

$$\begin{aligned} s^2 &= \frac{n}{n-1}\sigma_S^2 \\ &= \frac{n}{n-1} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n}(S_i - \overline{S})^2 \\ &= \frac{1}{n-1} \cdot \sum_{i=1}^{n}(S_i - \overline{S})^2 \end{aligned}$$

Mit “Standardabweichung von $\mathcal{S}$” ist (für Daten oder Stichproben) meistens das korrigierte s gemeint.

Beispiel zur Standardabweichung

i	1	2	3	4	5
x_i	1	3	0	5	1

Beispiel zur Standardabweichung

						Summe
i	1	2	3	4	5	
x_i	1	3	0	5	1	10

$\overline{x} = 10/5 = 2$

Beispiel zur Standardabweichung

						Summe
i	1	2	3	4	5	
x_i	1	3	0	5	1	10
$x_i - \overline{x}$	-1	1	-2	3	-1	0

Beispiel zur Standardabweichung

						Summe
i	1	2	3	4	5	
x_i	1	3	0	5	1	10
$x_i - \overline{x}$	-1	1	-2	3	-1	0
$(x_i - \overline{x})^2$	1	1	4	9	1	16

Beispiel zur Standardabweichung

						Summe
i	1	2	3	4	5	
x_i	1	3	0	5	1	10
$x_i - \overline{x}$	-1	1	-2	3	-1	0
$(x_i - \overline{x})^2$	1	1	4	9	1	16

$$s^2 = \frac{1}{5-1} \sum_{i=1}^{5} \left(x_i - \overline{x}\right)^2 = \frac{16}{5-1} = 4$$

$$s = 2$$

σ versus s : mit n oder $n-1$ berechnen?

Die Standardabweichung σ eines Zufallsexperiments mit n gleichwahrscheinlichen Ausgängen $x_1, \ldots, x_n$ (z.B. Würfelwurf) ist klar definiert durch

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Wenn es sich bei $x_1, \ldots, x_n$ um eine Stichprobe aus einer großen „Population“ handelt (wie meistens in der Statistik), sollten Sie die Formel

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

verwenden.

Warnung

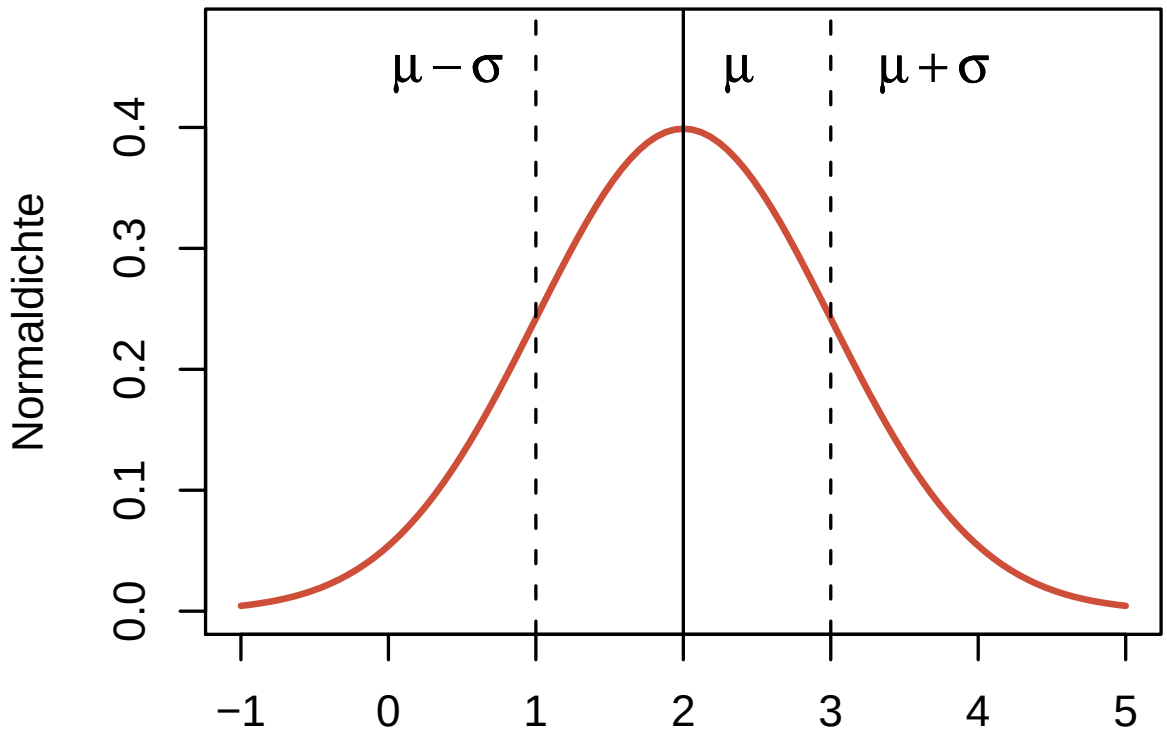
Mittelwert und Standardabweichung. . .

- charakterisieren die Daten gut, falls deren Verteilung glockenförmig ist
- und müssen andernfalls mit Vorsicht interpretiert werden.

Normalverteilung

Glockenförmig: symmetrisch und eingipflig.

Dichte der Normalverteilung: $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$



Die Normalverteilungsdichte heisst auch *Gauß'sche Glockenkurve* (nach Carl Friedrich Gauß, 1777-1855)

Schwankungen des Mittelwerts

Der Mittelwert (=Stichprobenmittel) $\overline{x}$ ist eine Näherung für den "wahren" Mittelwert μ der gesamten Population. Für verschiedene Stichproben schwankt er um den tatsächlichen Wert μ. Die Variabilität von $\overline{x}$ hängt vom Stichprobenumfang n ab.

Der Standardfehler:

$$\frac{s}{\sqrt{n}}$$

ist die geschätzte Standardabweichung des Mittelwertes.

Normalverteilung des Mittelwerts

Der Mittelwert der Stichproben ist stets annähernd normalverteilt, auch wenn die eigentlichen Daten eine ganz andere Verteilung haben (solange man immer dasselbe misst, und die Messungen unabhängig voneinander sind). Dies folgt aus dem zentralen Grenzwertsatz.

Normalverteilung des Mittelwerts

Beispiel: Transpirationsrate von Hirsepflanzen. Blau: theoretische Verteilung der Raten, rot: Verteilung der Stichprobenmittel bei Stichproben der Größe $n = 16$.