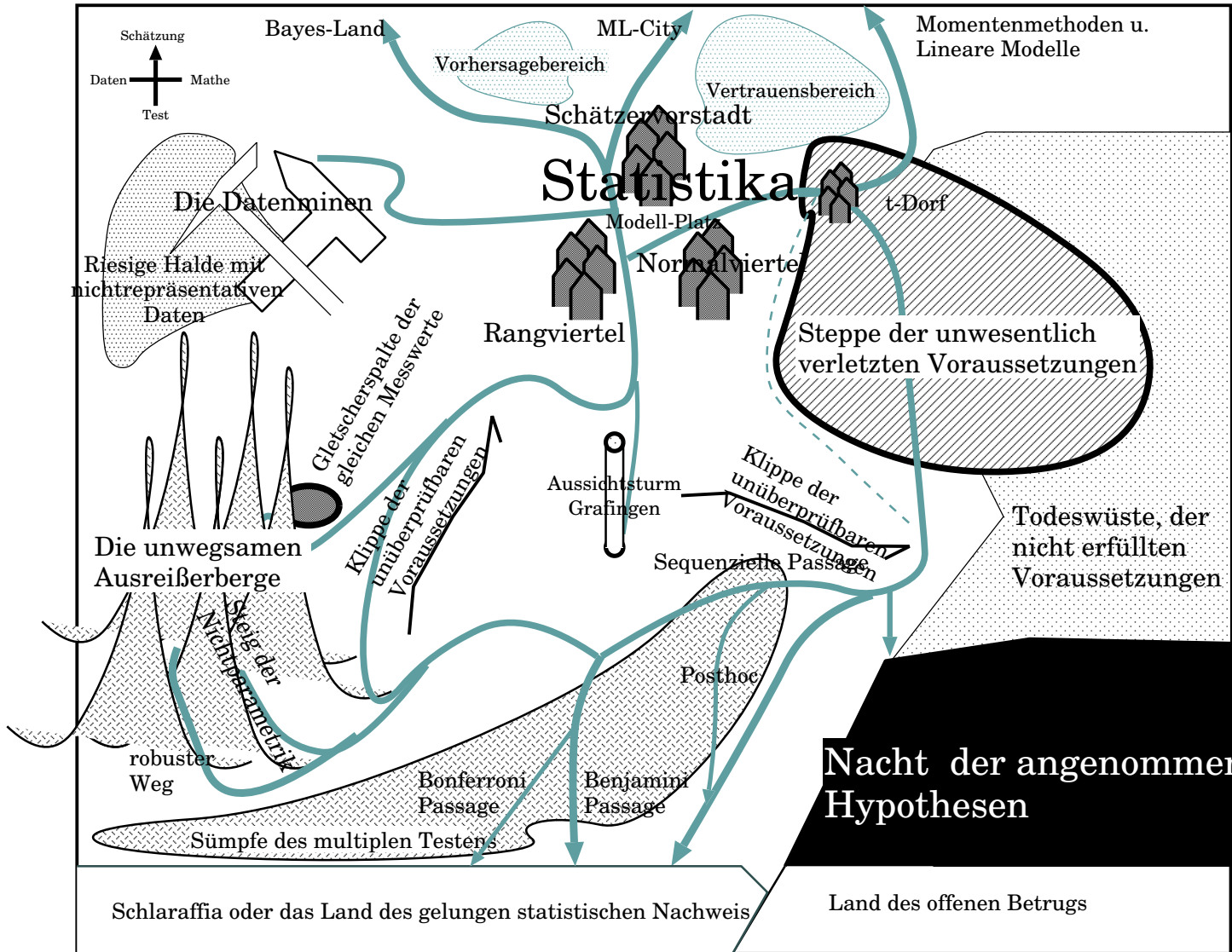


# Stochastik und Statistik

## *Vorlesung 2 (Graphik I)*

K.Gerald van den Boogaart

<http://www.math-inf.uni-greifswald.de/statistik>



# Einteilung der Graphiken und Parameter

		Erste Variable	
		diskret	stetig
zweite Variable	keine	?	?
	diskret	?	?
	stetig	?	S.O.

- stetige Daten
- diskrete Daten
- stetig–stetig
- diskret–diskret
- diskret–stetig

# Lernziele

Zu jeder Graphik lernen wir:

- Für welche Daten eignet sich die Graphik?

Warum lernen wir das?

# Lernziele

Zu jeder Graphik lernen wir:

- Für welche Daten eignet sich die Graphik?
- Wie ist die Graphik aufgebaut?

Warum lernen wir das?

# Lernziele

Zu jeder Graphik lernen wir:

- Für welche Daten eignet sich die Graphik?
- Wie ist die Graphik aufgebaut?
- Was kann man in der Graphik sehen?

Warum lernen wir das?

# Lernziele

Zu jeder Graphik lernen wir:

- Für welche Daten eignet sich die Graphik?
- Wie ist die Graphik aufgebaut?
- Was kann man in der Graphik sehen?
- Woran kann man es erkennen?

Warum lernen wir das?

# Lernziele

Zu jeder Graphik lernen wir:

- Für welche Daten eignet sich die Graphik?
- Wie ist die Graphik aufgebaut?
- Was kann man in der Graphik sehen?
- Woran kann man es erkennen?
- Was übersieht man in der Graphik?

Warum lernen wir das?



# Lernziele

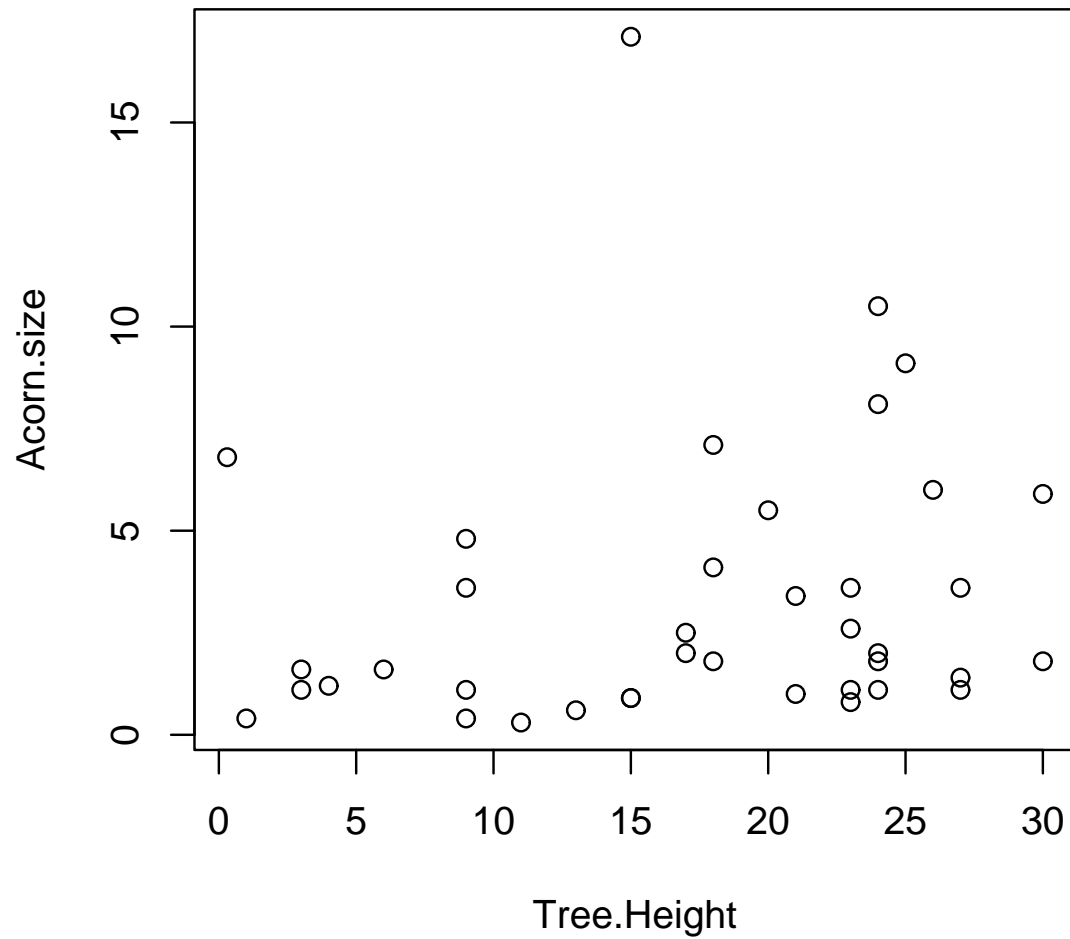
Zu jeder Graphik lernen wir:

- Für welche Daten eignet sich die Graphik?
- Wie ist die Graphik aufgebaut?
- Was kann man in der Graphik sehen?
- Woran kann man es erkennen?
- Was übersieht man in der Graphik?
- Für welche Fragestellungen eignet sich die Graphik?

Warum lernen wir das?

# Vorbereitung: Darstellung des Wertes durch die Lage

# Streudiagramm

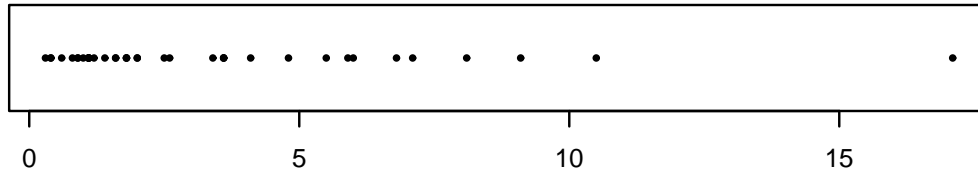


# Graphiken für stetige Daten

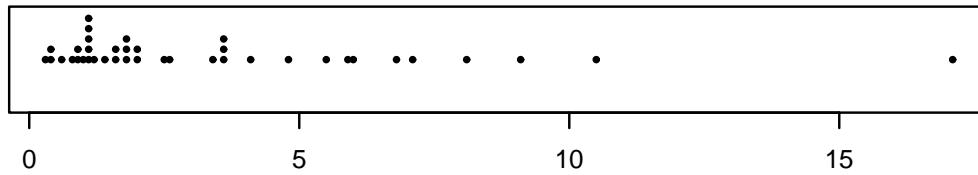
- Punktdiagramm (stapeln, verzittern)
- Histogramm
- Kastendiagramm / Boxplot
- Q Q-Plots (Quantils-Quantils Plot)
- (Empirische Verteilungsfunktion)

# Punktdiagramm

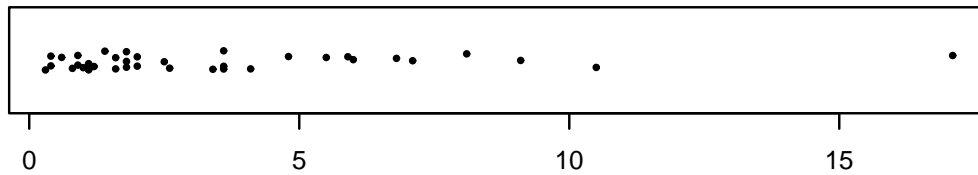
Punktdiagramm



gestapeltes Punktdiagramm



verzittertes Punktdiagramm

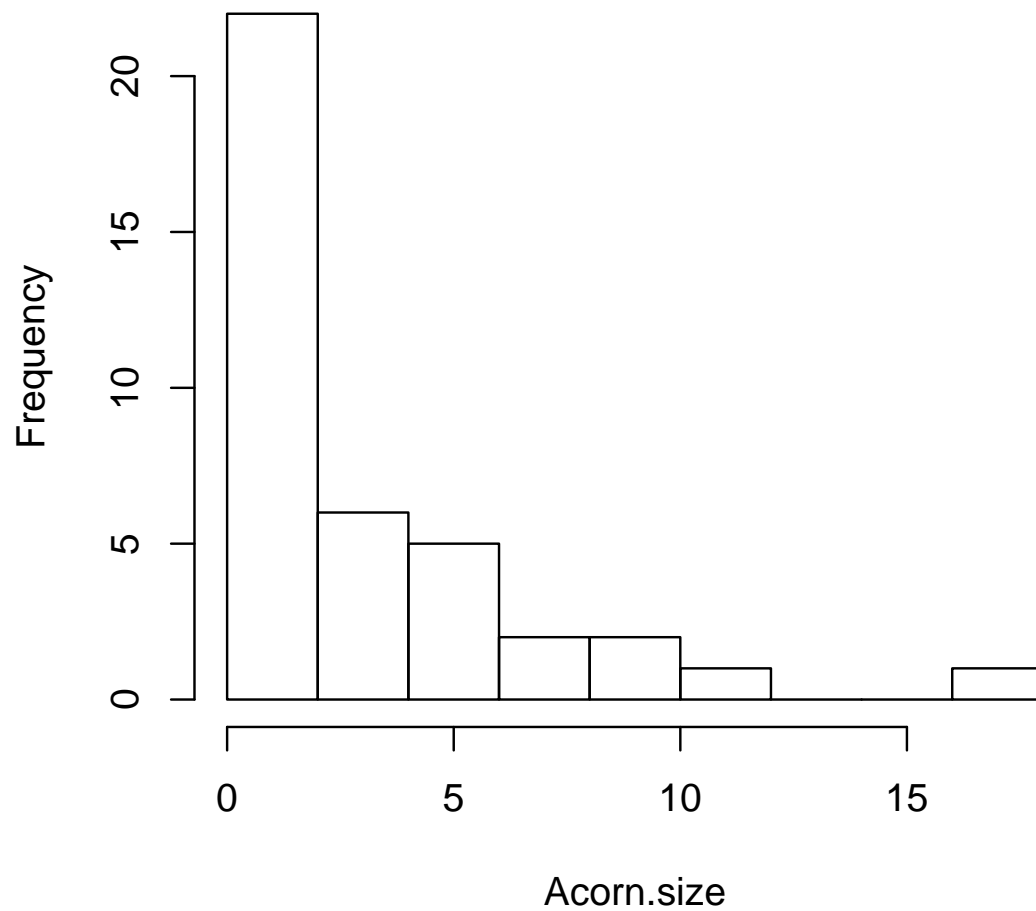


# Punktdiagramm

- Vollständig bis auf Überdeckung
- Verzittern und Stapeln
- Was “sieht” man?

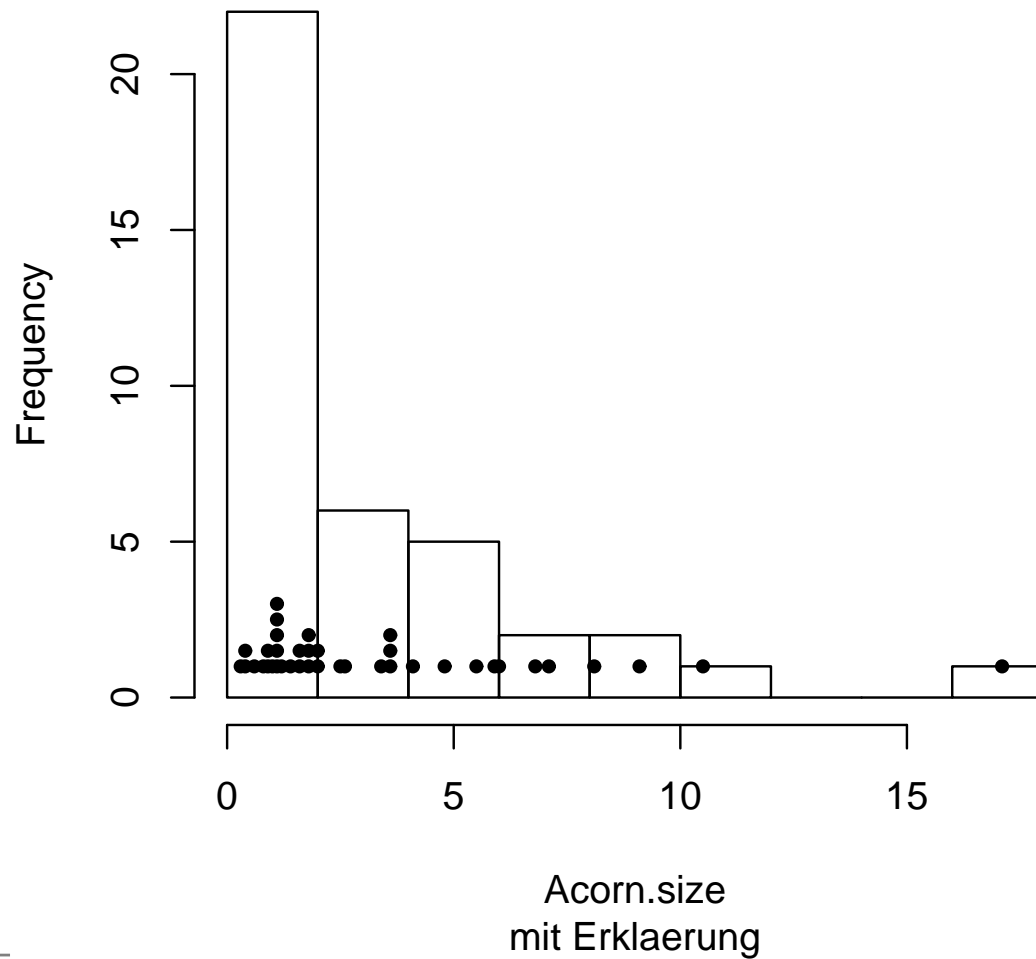
# Histogramm

Histogram of Acorn.size



# Histogramm

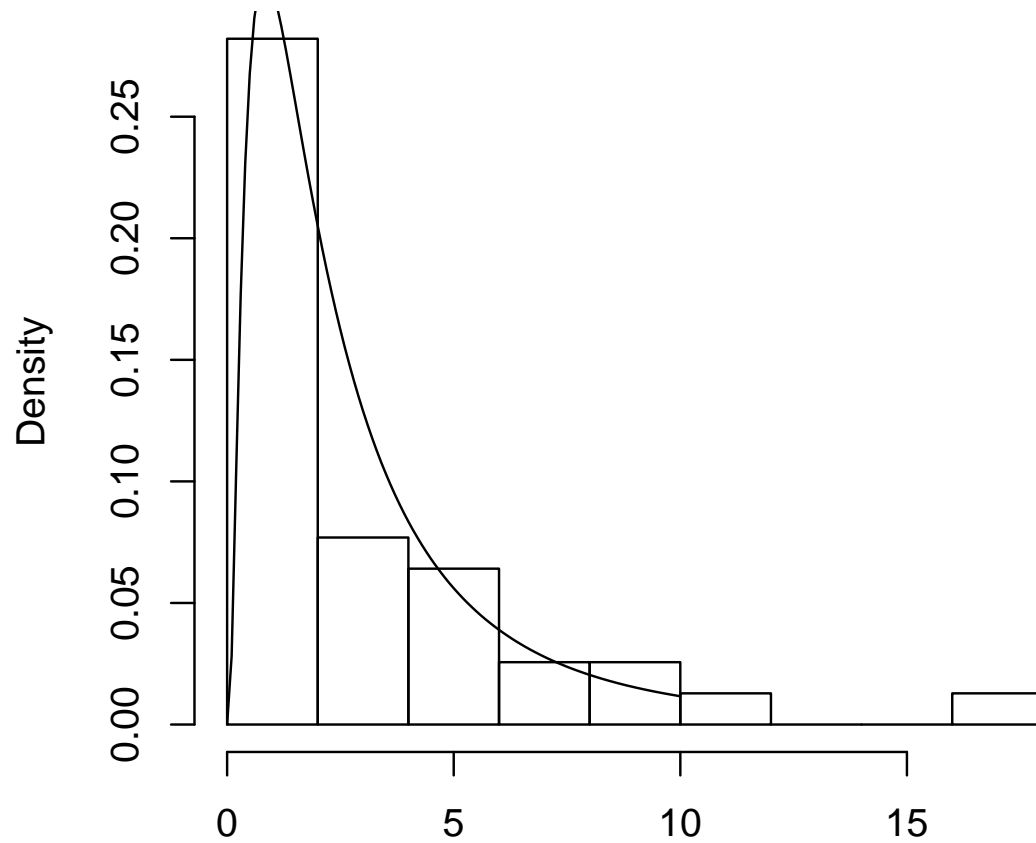
Histogram of Acorn.size





# Histogramm

Histogram of Acorn.size



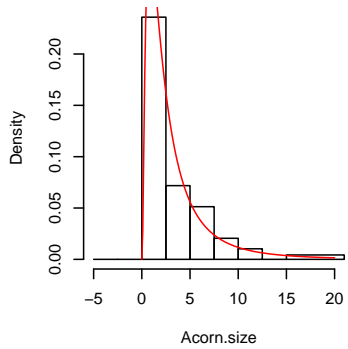
Acorn.size  
als Dichteschätzung

# Histogramm

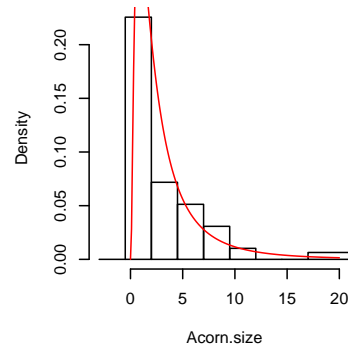
- Stellt Anzahl von Datenpunkten im Intervall dar.
- Stellt die Dichte (Datenpunkte pro Punkt und Einheitslänge) der Punkte dar.
- Balkenhöhe ist zufällig.
- Variation von Balkenanfang und Balkenanzahl führt zu verschiedenen Eindrücken.
- Zu kleine Balken  $\Rightarrow$  “Zufallsflimmer”
- Zu große Balken  $\Rightarrow$  Information zu sehr zusammengefaßt.
- Extreme Ausreißer eventuell am linken oder rechten Rand erkennbar.

# Einfluß des Balkenanfangs

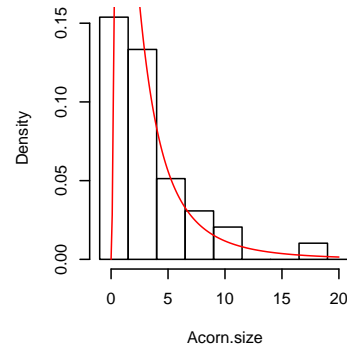
Histogram of Acorn.size



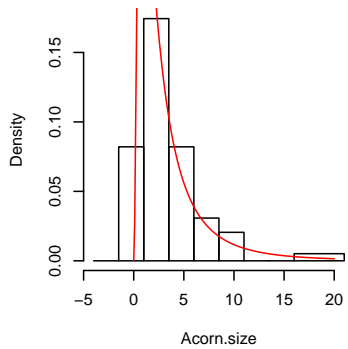
Histogram of Acorn.size



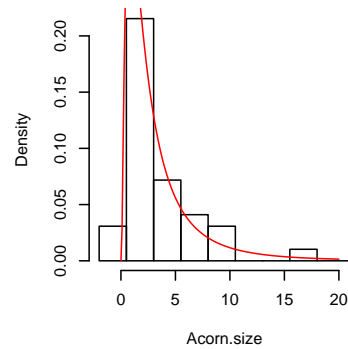
Histogram of Acorn.size



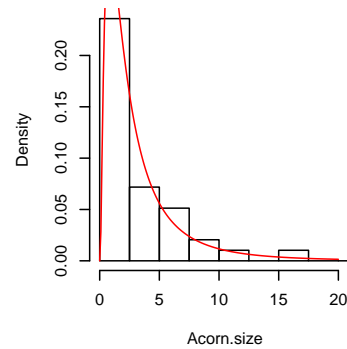
Histogram of Acorn.size



Histogram of Acorn.size



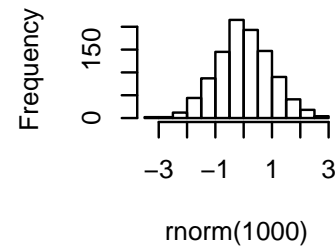
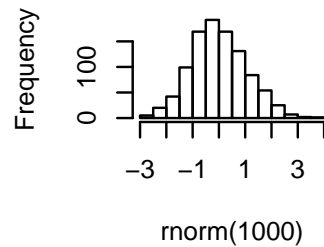
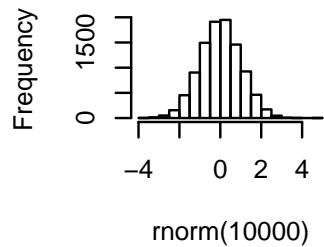
Histogram of Acorn.size



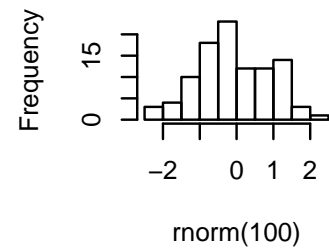
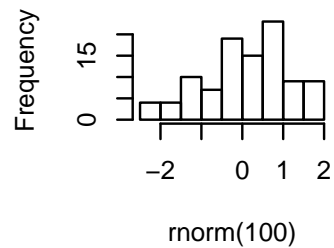
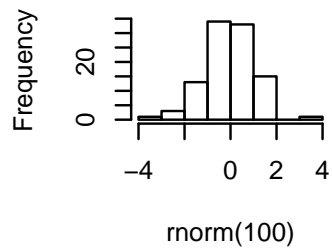
# Beschreibung der Verteilungsform und Normalverteilung als Referenzverteilung

# Normalverteilung

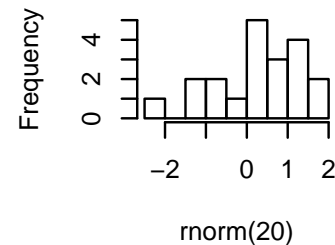
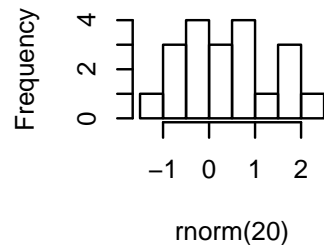
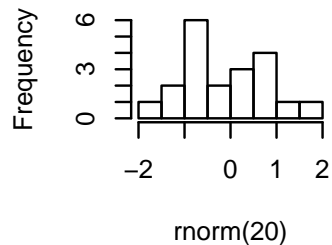
Histogram of rnorm(10000) Histogram of rnorm(1000) Histogram of rnorm(1000)



Histogram of rnorm(10) Histogram of rnorm(10) Histogram of rnorm(10)

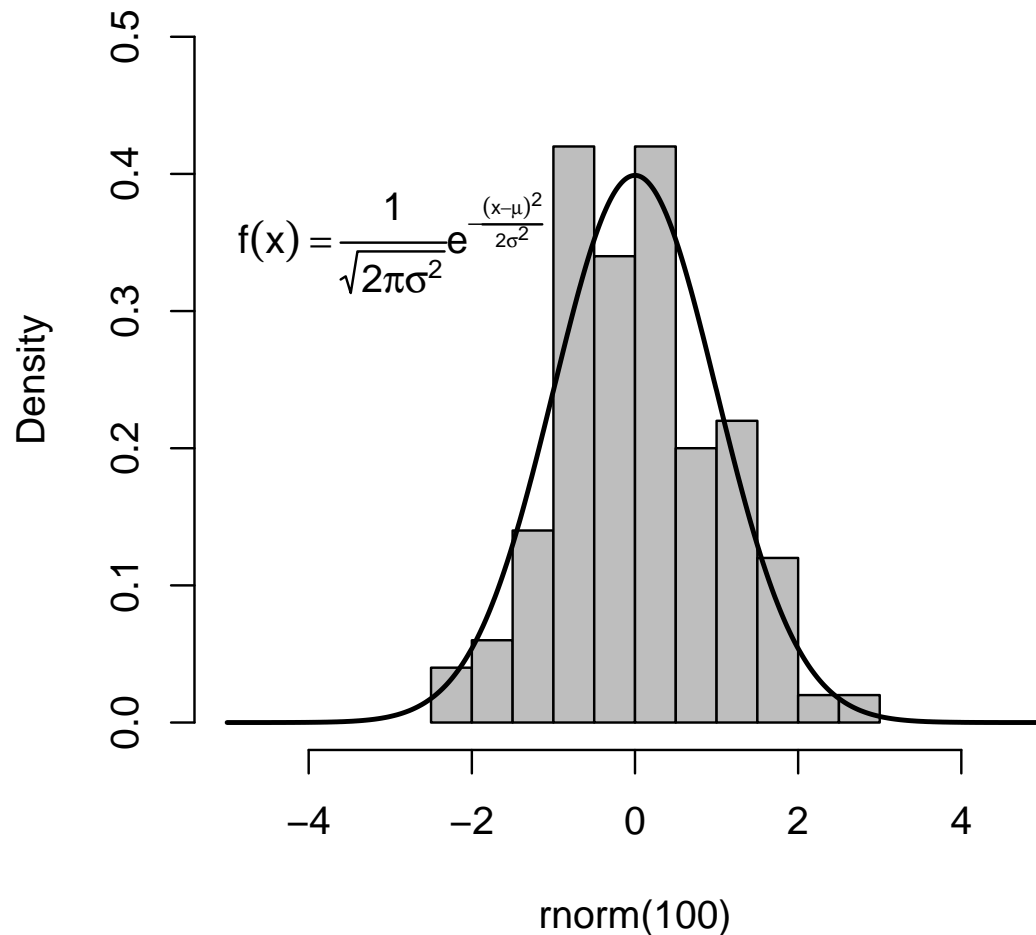


Histogram of rnorm(20) Histogram of rnorm(20) Histogram of rnorm(20)



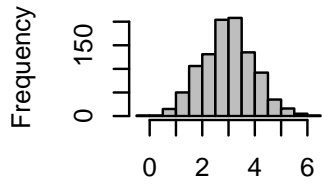
# Dichte der Normalverteilung

Histogramm und Dichte  
einer Normalverteilung



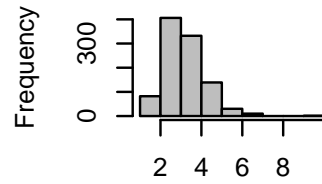
# Verteilungseigenschaften

**symmetrisch  
eingipflig**



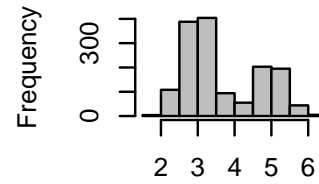
`rnorm(1000, mean = 3)`

**rechtsschief**

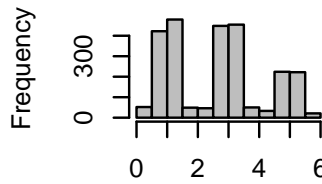


`rlnorm(1000, mean = log(3), sd = mean = 3, sd = 0.4), rnorm(500, r`

**zweigipflig/bimodal**

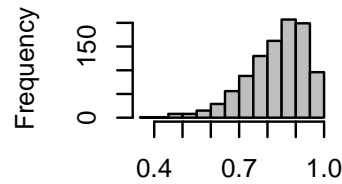


**multimodal**



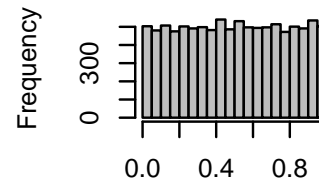
`norm(500, 3, 0.3), rnorm(500, 5, 0.3), rno`

**linksschief,  
eingeschaenkt**



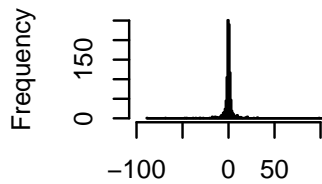
`rbeta(1000, 10, 2)`

**Gleichverteilung  
auf [0,1]**



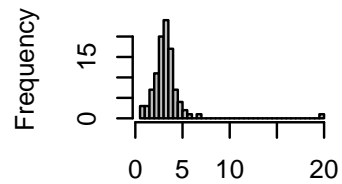
`rbeta(10000, 1, 1)`

**Schwere  
Verteilungsschwaenz**



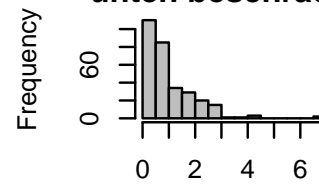
`rcauchy(1000)`

**Ausreisser**



`c(rnorm(100, mean = 3), 20)`

**rechtsschief  
monoton fallend  
unten beschaenkt**



`rexp(300)`

# Kenngrößen und Parameter

- Lage
- Streuung
- Form
- Verteilung

Kenngrößen und Parameter sind konventionelle Zusammenfassungen der Daten in einzelne Zahlen, die jeweils einen bestimmten Aspekt quantitativ erfassen.



# Lageparameter

- Lage
  - Mittelwert (geometrisch und arithmetisch)
  - Median
  - Modus
  - Quantile (Quartile, Dezentile)
- Streuung
- Form
- Verteilung

# (arithmetischer) Mittelwert

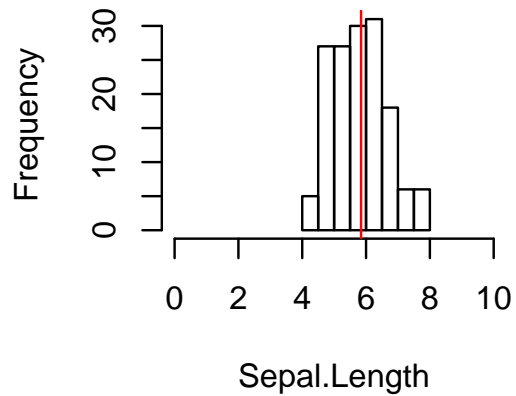
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

```
> mean(iris$Sepal.Length)
```

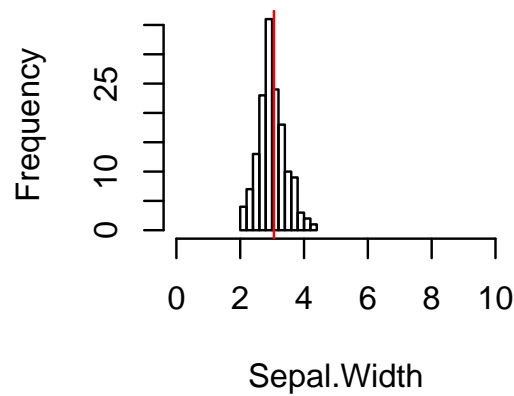
```
[1] 5.843333
```

# Mittelwert

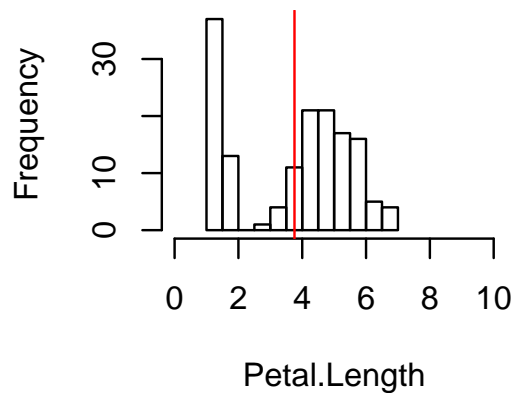
## Histogram of Sepal.Length



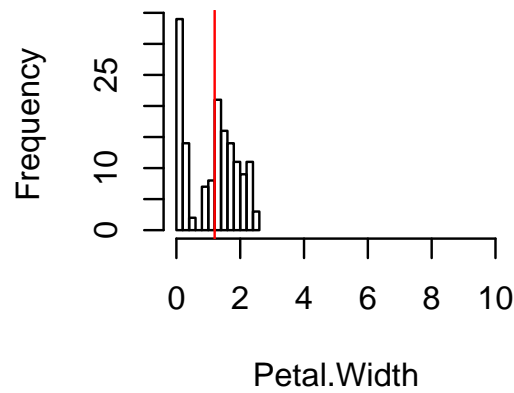
## Histogram of Sepal.Width



## Histogram of Petal.Length



## Histogram of Petal.Width



# (geometrischer) Mittelwert

Für die ratio-Skala gibt es noch den geometrischen Mittelwert

$$\bar{x} = \sqrt[n]{\prod_{i=1}^n x_i} = (x_1 x_2 \cdots x_n)^{\frac{1}{n}}$$

```
> exp(mean(log(iris$Sepal.Length)))
```

```
[1] 5.78572
```

# Median

Der Median ist der mittlere Wert:

```
> median(c(4, 5, 1, 3, 6, 7, 8))
```

```
[1] 5
```

```
> median(c(4, 1, 3, 6, 7, 8))
```

```
[1] 5
```

```
> median(iris$Sepal.Length)
```

```
[1] 5.8
```

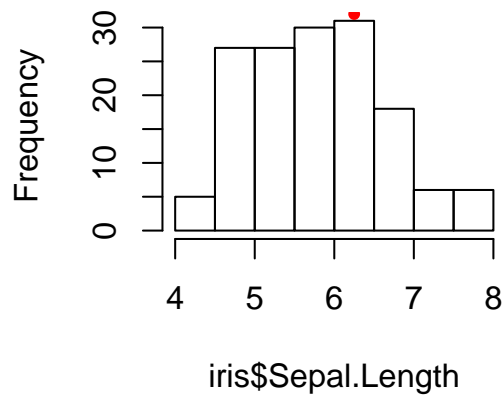
```
> sapply(iris[, 1:4], median)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.80	3.00	4.35	1.30

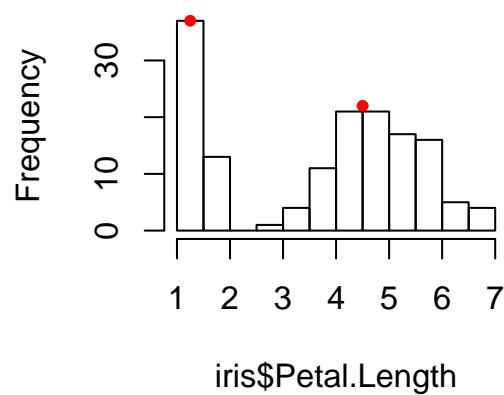
# Modus

Der Modus oder Modalwert bezeichnet den Bereich mit der größten Punktdichte.

Histogram of iris\$Sepal.Length

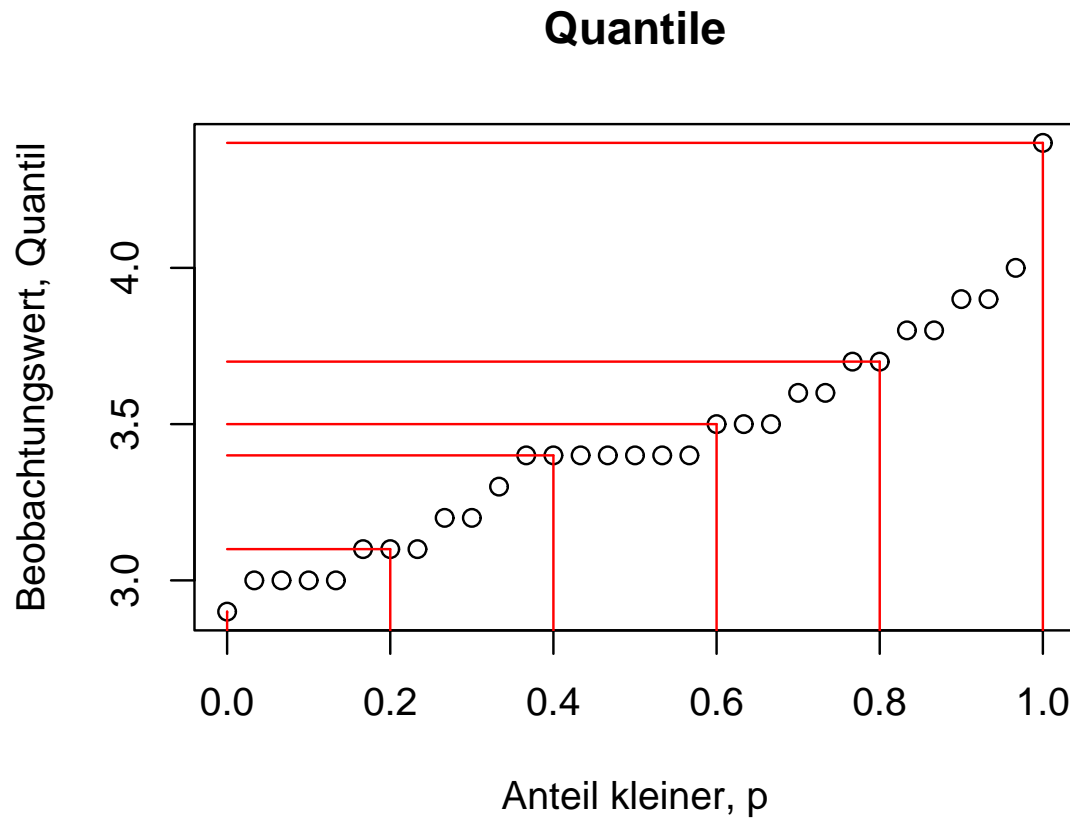


Histogram of iris\$Petal.Length



# Quantile

Das (empirische)  $p$ -Quantil  $\hat{q}_p$  ist der Wert für den der Anteil  $p$  des sortierten Datensatzes kleiner ist.



Details in den Übun-

gen.

# Spezielle Quantile

- $\frac{1}{2}$ -Quantil ist der Median
- $\frac{1}{4}$ -Quantil heißt auch **erstes Quartil**
- $\frac{3}{4}$ -Quantil heißt auch **drittes Quartil**
- $\frac{n}{10}$ -Quantil heißt auch **n-tes Dezantil**
- 0-Quantil heißt auch Minimum (sehr zufällig!!!)
- 1-Quantil heißt auch Maximum (sehr zufällig!!!)



# Streuparameter

- Lage
- Streuung
  - Varianz
  - Standardabweichung
  - IQR
  - Variationkoeffizient
  - geometrische Standardabweichung
- Form
- Verteilung

# Streuparameter für die reelle Skala

- Varianz

$$\widehat{var}(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

# Streuparameter für die reelle Skala

- Varianz

$$\widehat{var}(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Standardabweichung

$$\widehat{sd}(X) = \sqrt{\widehat{var}(X)}$$

# Streuparameter für die reelle Skala

- Varianz

$$\widehat{var}(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

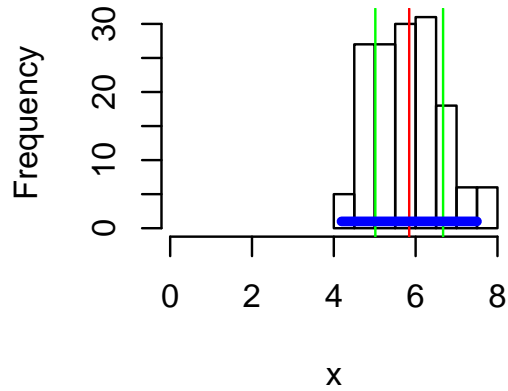
- Standardabweichung

$$\widehat{sd}(X) = \sqrt{\widehat{var}(X)}$$

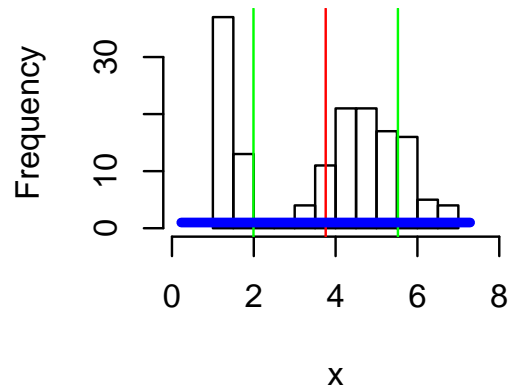
- Interquartilsabstand

$$\widehat{IQR}(X) = q_{0.75} - q_{0.25}$$

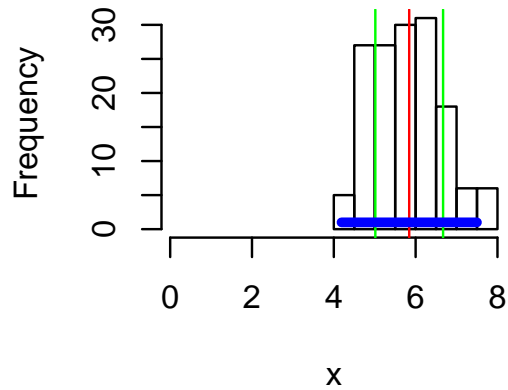
**classical**  
mean= 5.84 sd= 0.83



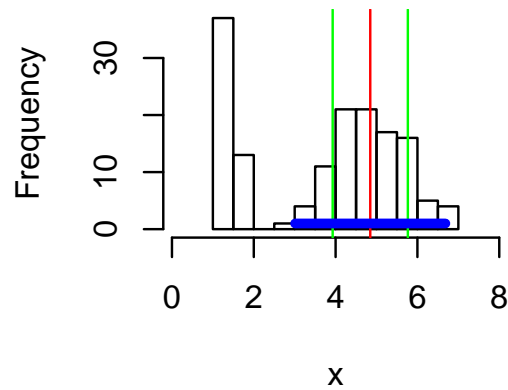
**classical**  
mean= 3.76 sd= 1.77



**robust:**  
mean= 5.84 sd= 0.83



**robust:**  
mean= 4.85 sd= 0.92



# Streuparameter für die ratio Skala

- Variationskoeffizient

$$\hat{v}(X) = \frac{\hat{sd}(X)}{\bar{x}}$$

# Streuparameter für die ratio Skala

- Variationskoeffizient

$$\hat{v}(X) = \frac{\hat{sd}(X)}{\bar{x}}$$

- Geometrische Standardabweichung

$$\exp(\hat{sd}(\ln(X)))$$

# Streuparameter für die ratio Skala

- Variationskoeffizient

$$\hat{v}(X) = \frac{\hat{sd}(X)}{\bar{x}}$$

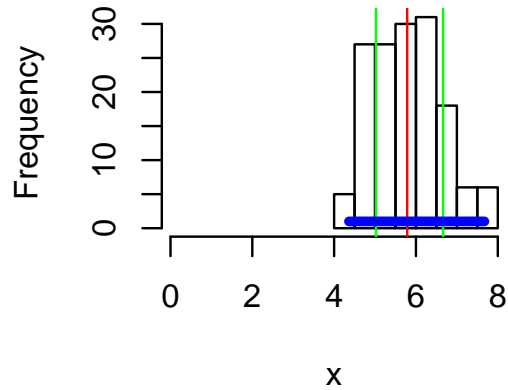
- Geometrische Standardabweichung

$$\exp(\hat{sd}(\ln(X)))$$

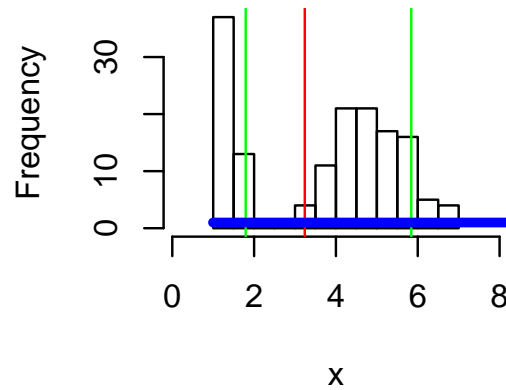


# Blick mit der Ratioskala

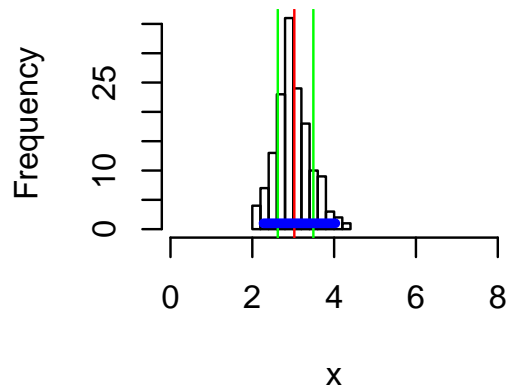
**classical**  
geom. mean= 5.79 gsd= 1.1



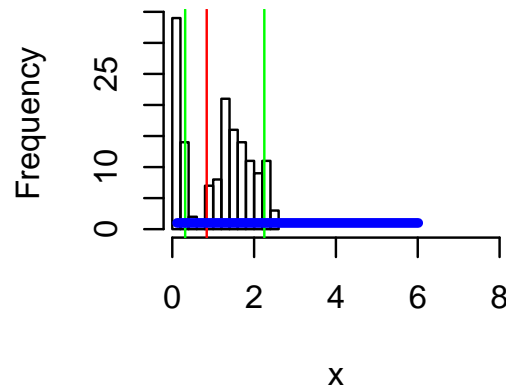
**classical**  
geom. mean= 3.24 gsd= 1.8



**classical**  
geom. mean= 3.03 gsd= 1.1



**classical**  
geom. mean= 0.84 gsd= 2.6



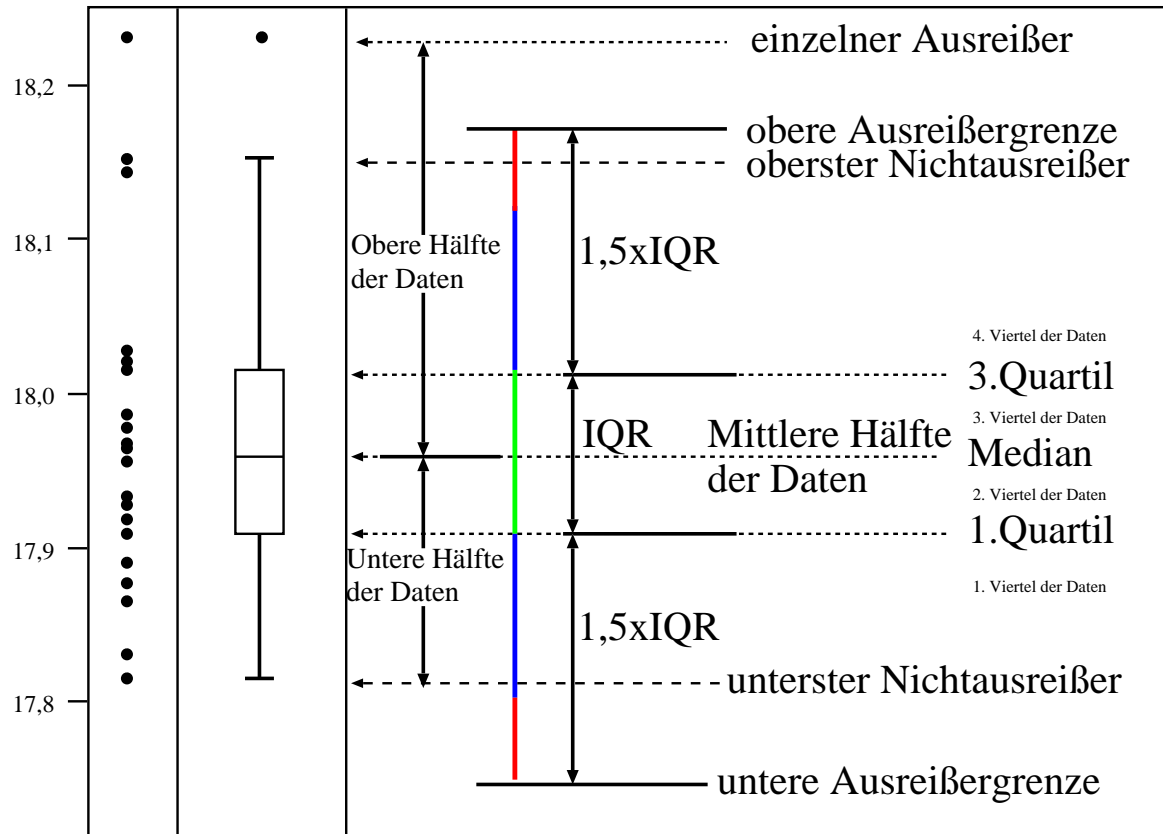
# Weitere Parameter

- Lage
- Streuung
- Form
  - Schiefe
  - Wölbung
  - ...
- Verteilung
  - Hängt vom Verteilungsmodell ab.

# Kastendiagramm/Boxplot

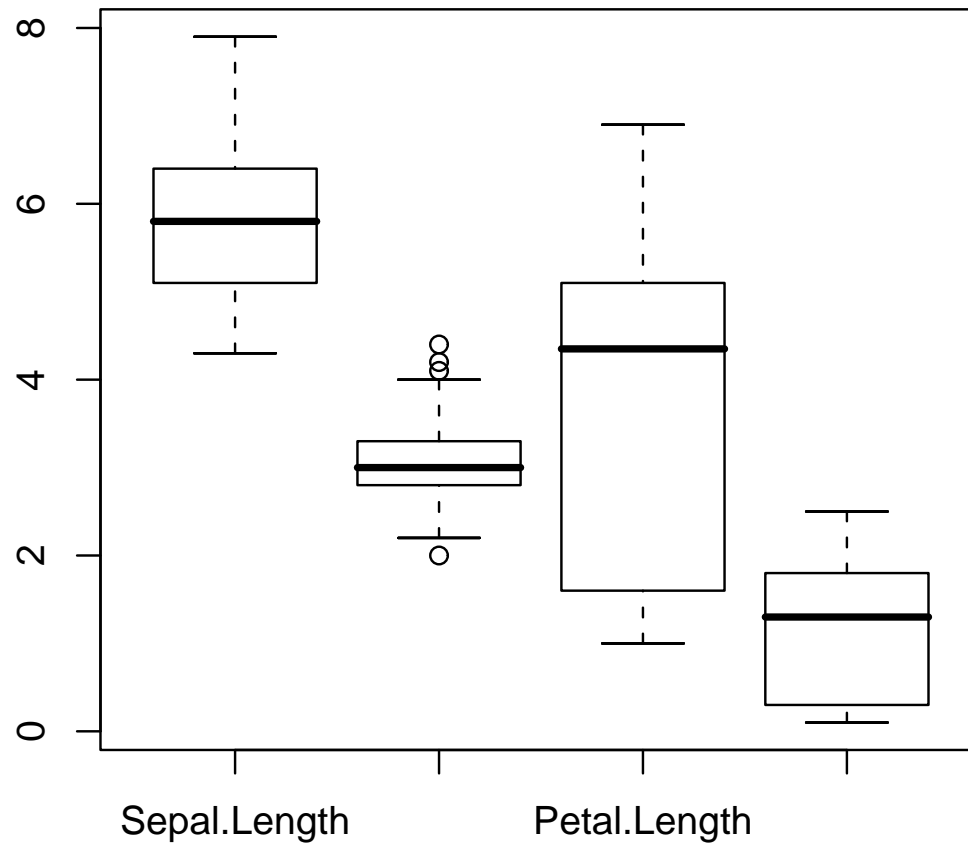
Dotplot Boxplot

Erklärung zum Boxplot



# Kastendiagramme

Boxplots der reellen Variablen des Iris Datensatzes:



# Interpretation

- Ausreißer
- Stichprobenlage / Median
- Stichprobenstreuung / IQR
- Symmetrie und Schiefe der Verteilung
- eventuell extreme Werthäufungen

# Exkurs: Ausreißer

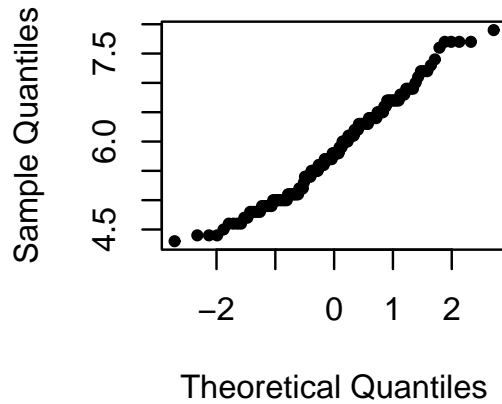
Definition: Ein Ausreißer ist ein Datenpunkt der einen “ungewöhnlich” extremen Wert hat.

Mögliche Ursachen:

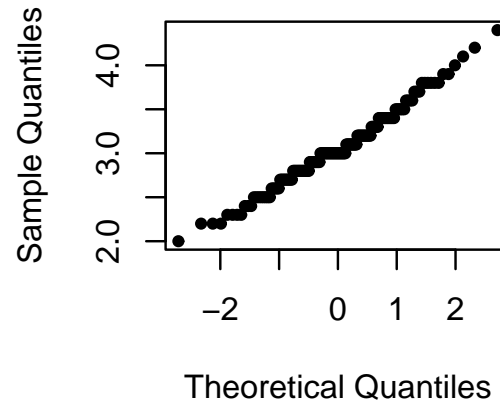
- Zufall (Es gibt halt extreme Werte)
- Schwere Verteilungsschwänze (Ausreißer hier typisch)
- Datenfehler oder Übermittlungsfehler
- Untypischer Spezialfall (der Millionär mit Zweitwohnsitz im armen Bergbauerndorf)
- Individuum fehlerhafterweise in der Stichprobe (z.B. andere Art)
- Anthropogene Überprägung (das verlorene Geldstück mit hohem Kupfergehalt.)

# Q Q-Plots

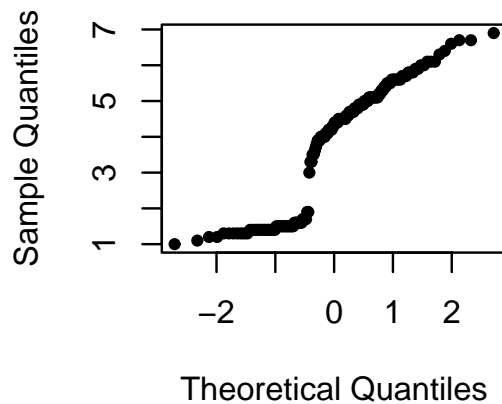
**Sepal.Length**



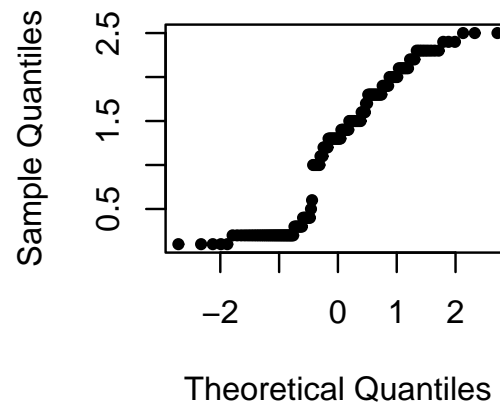
**Sepal.Width**



**Petal.Length**



**Petal.Width**



# Interpretation Q Q-Plot

- Ungefähre Gerade  $\Leftrightarrow$  Verteilungsmodell passend
- “Treppenstufen”  $\Leftrightarrow$  Bindungen (gleiche Werte)
- “Gegen S”  $\Leftrightarrow$  Ausreißer? schwere Verteilungsschwänze?



# Exkurs: Bindungen

Definition: Von einer **Bindung** spricht man, wenn ein Datenwert in einer stetigen Variable zwei oder mehrfach auftritt.

Mögliche Ursachen:

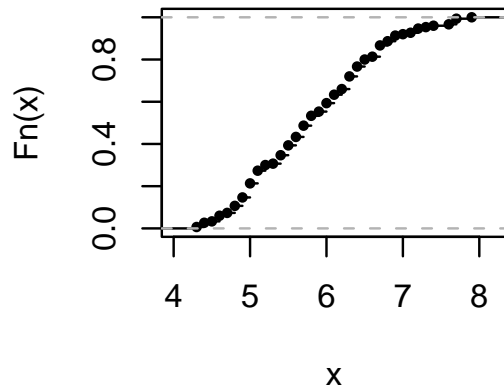
- Rundung
- Ungenau Datenerhebung
- Spezieller Wert hat positive Wahrscheinlichkeit
- Variable nicht wirklich stetig

Manche statistische Verfahren verlieren an zunehmend an Genauigkeit je mehr Bindungen auftreten.

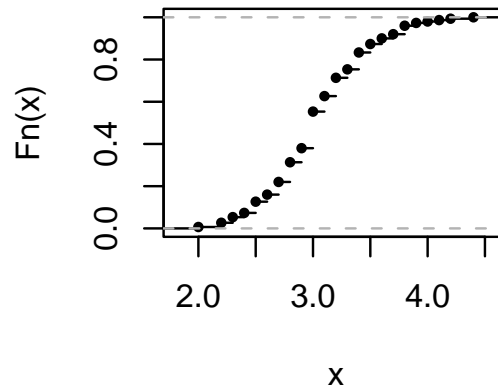
# Empirische Verteilungsfunktion

$$\hat{F}(x) = \text{Anteil des Datensatzes } \leq x$$

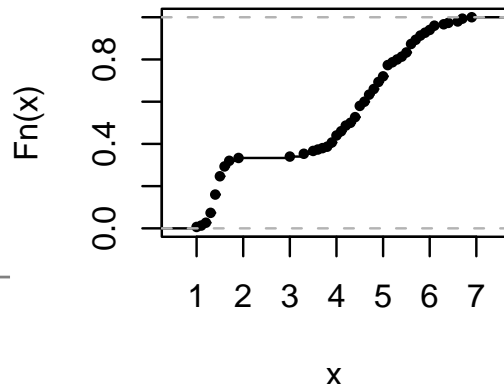
**Sepal.Length**



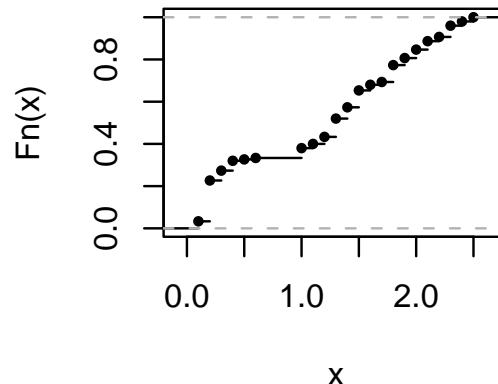
**Sepal.Width**



**Petal.Length**



**Petal.Width**



# Emprische Verteilungsfunktion

- Quantile können leicht abgelesen werden.
- Wahrscheinlichkeiten können leicht abgelesen werden.
- Bindungen erzeugen hohe Sprünge (fast unsichtbar).
- Sonst kann eigentlich nichts abgelesen werden.

# Zusammenfassung zu stetigen Daten

- Lage- und Streuparameter / quantitativ
- Punktdiagramm (stapeln, verzittern) / Daten
- Histogramm (Balken variieren) / Verteilungsform
- Kastendiagramm / Ausreißer, Streuung, Lage, Symmetrie
- Q Q-Plot / Vergleich mit Verteilung
- Empirische Verteilungsfunktion / Quantile

