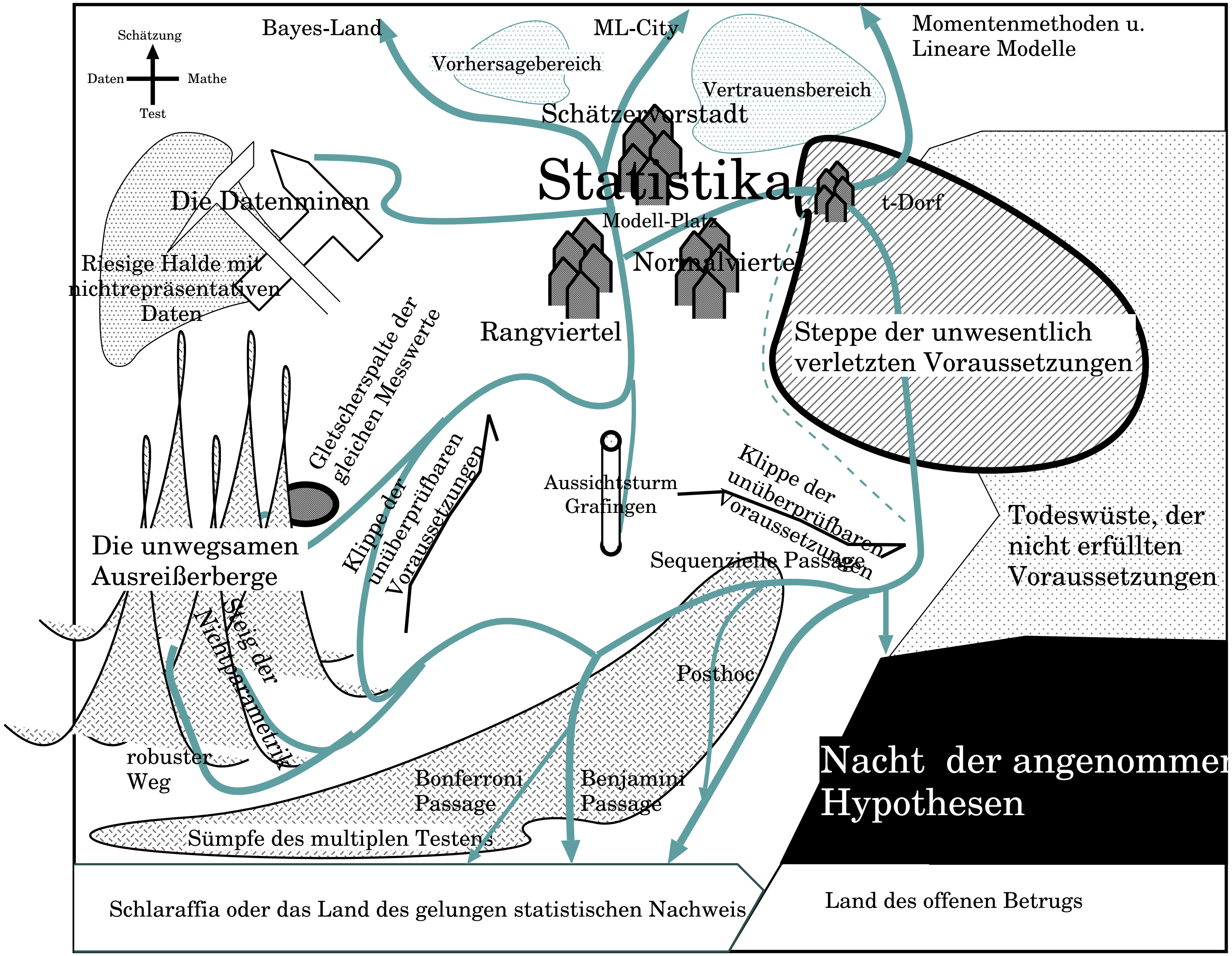
Datenanalyse und Statistik

Vorlesung 2 (Graphik I)

K.Gerald van den Boogaart
http://www.stat.boogaart.de


Schätzung
Daten
Mathe
Test
Bayes-Land
ML-City
Momentenmethoden u.
Lineare Modelle
Vorhersagebereich
Vertrauensbereich
Schätzervorstadt
Statistika
Die Datenminen
Modell-Platz
t-Dorf
Riesige Halde mit
nichtrepräsentativen
Daten
Normalviertel
Rangviertel
Steppe der unwesentlich
verletzten Voraussetzungen
Gletscherspalte der
gleichen Messwerte
Klippe der
unüberprüfbaren
Voraussetzungen
Aussichtsturm
Grafingen
Klippe der
unüberprüfbaren
Voraussetzungen
Todeswüste, der
nicht erfüllten
Voraussetzungen
Die unwegsamen
Ausreißerberge
Sequenzielle Passagen
Steig der
Nichtparametrik
Posthoc
robuster
Weg
Bonferroni
Passage
Benjamini
Passage
Nacht der angenommen
Hypothesen
Sümpfe des multiplen Testens
Schlaraffia oder das Land des gelungen statistischen Nachweis
Land des offenen Betrugs

Einteilung der Graphiken und Parameter

		Erste Variable	
		diskret	stetig
	keine	?	?
zweite Variable	diskret	?	?
	stetig	wie diskret-stetig	?

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

Lernziele

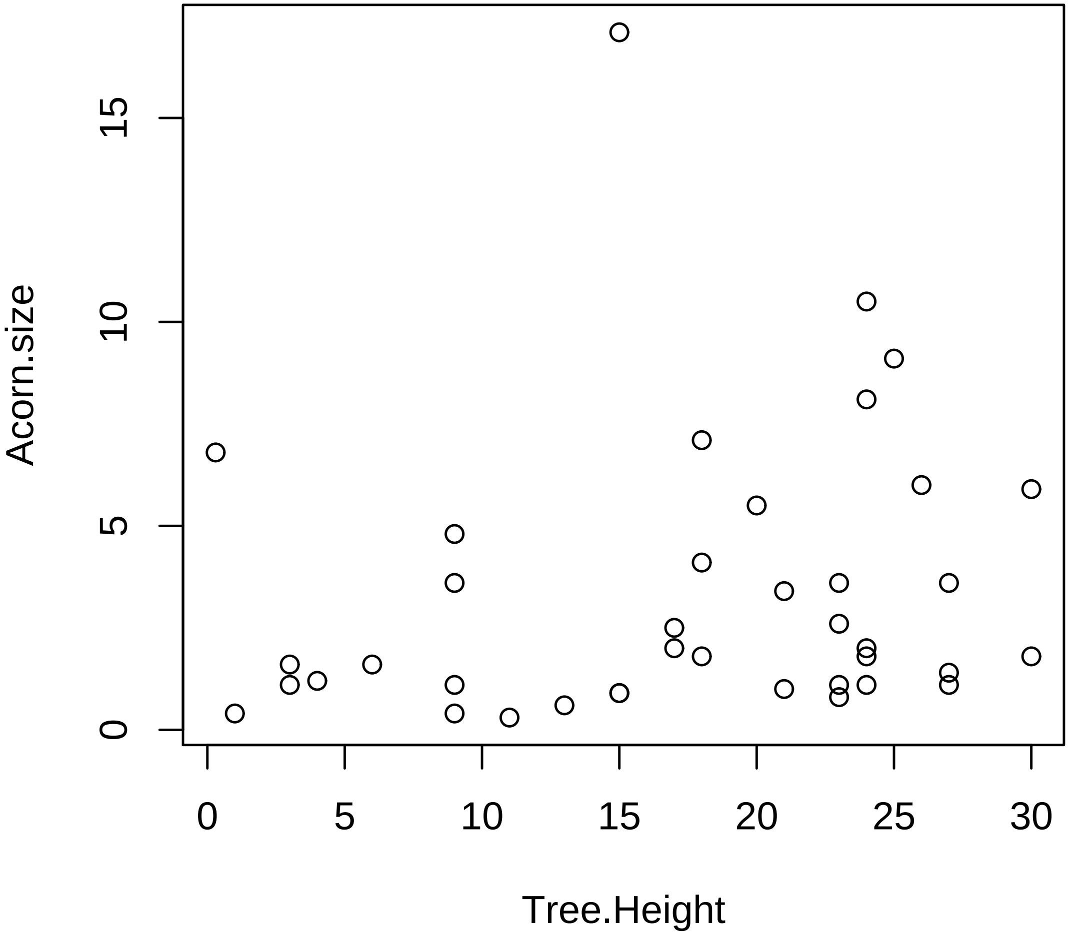
Zu jeder Graphik lernen wir:

- 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

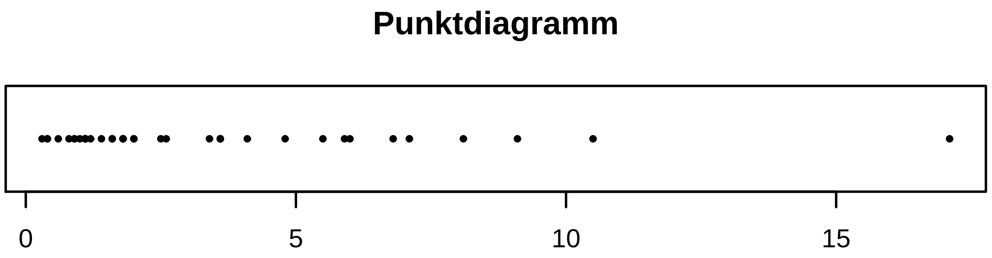
Streudiagramm

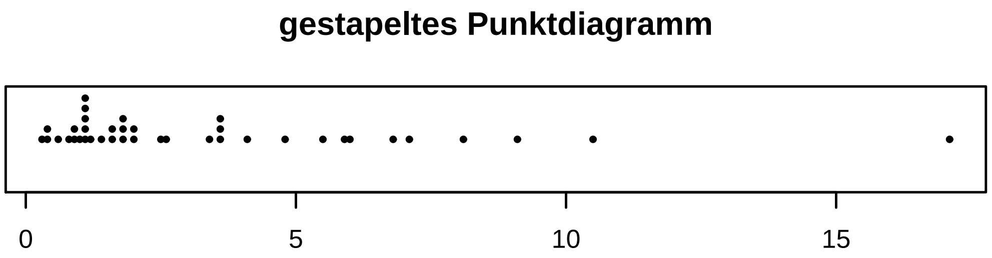


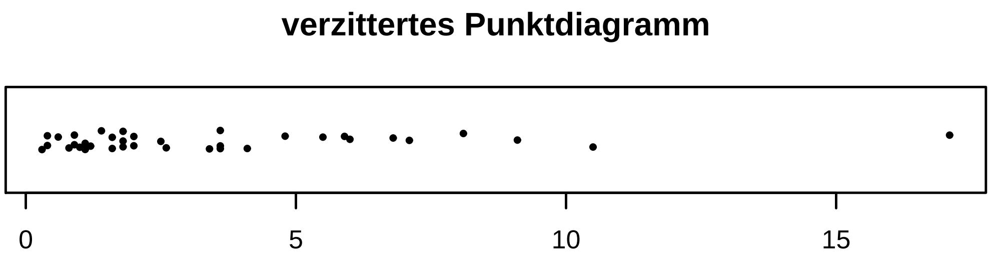
Graphiken für stetige Daten

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

Punktdiagramm



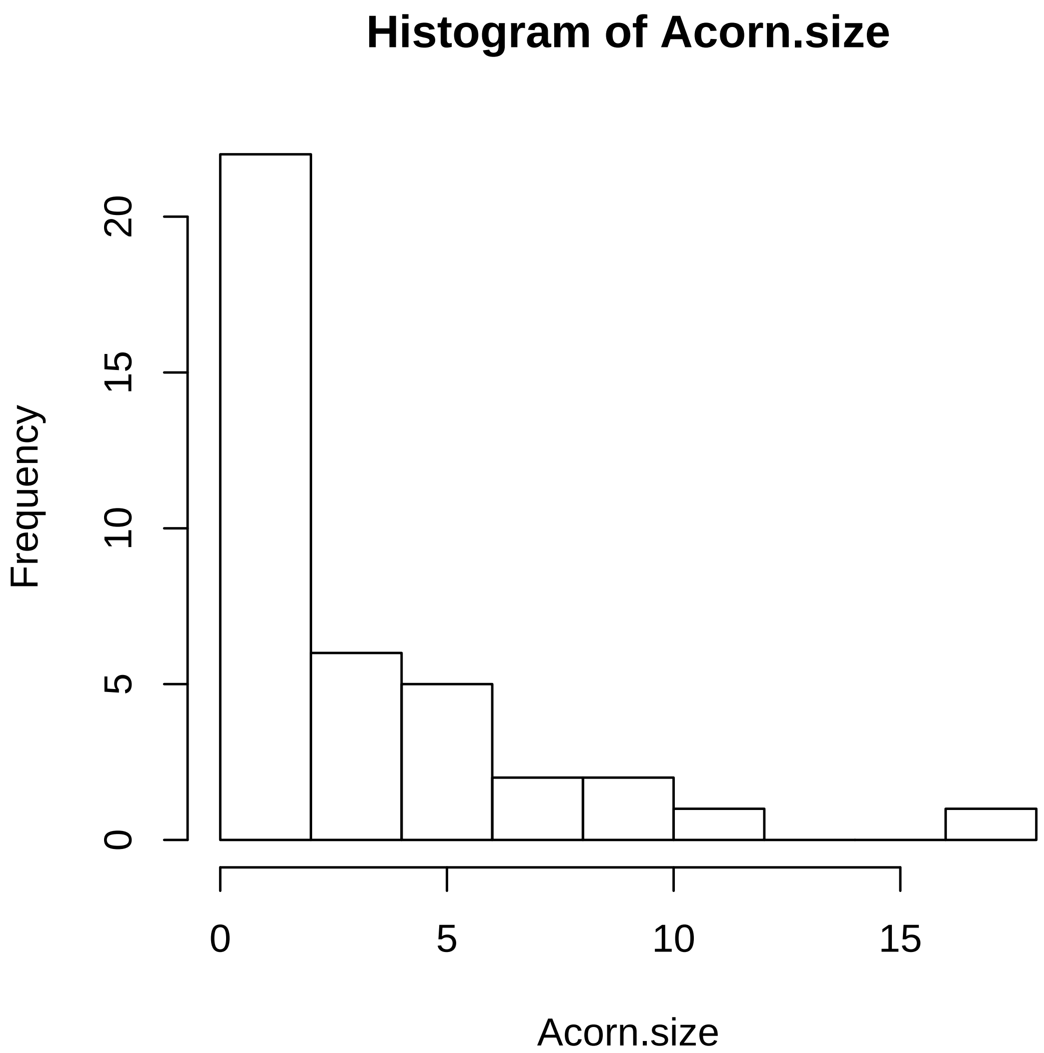




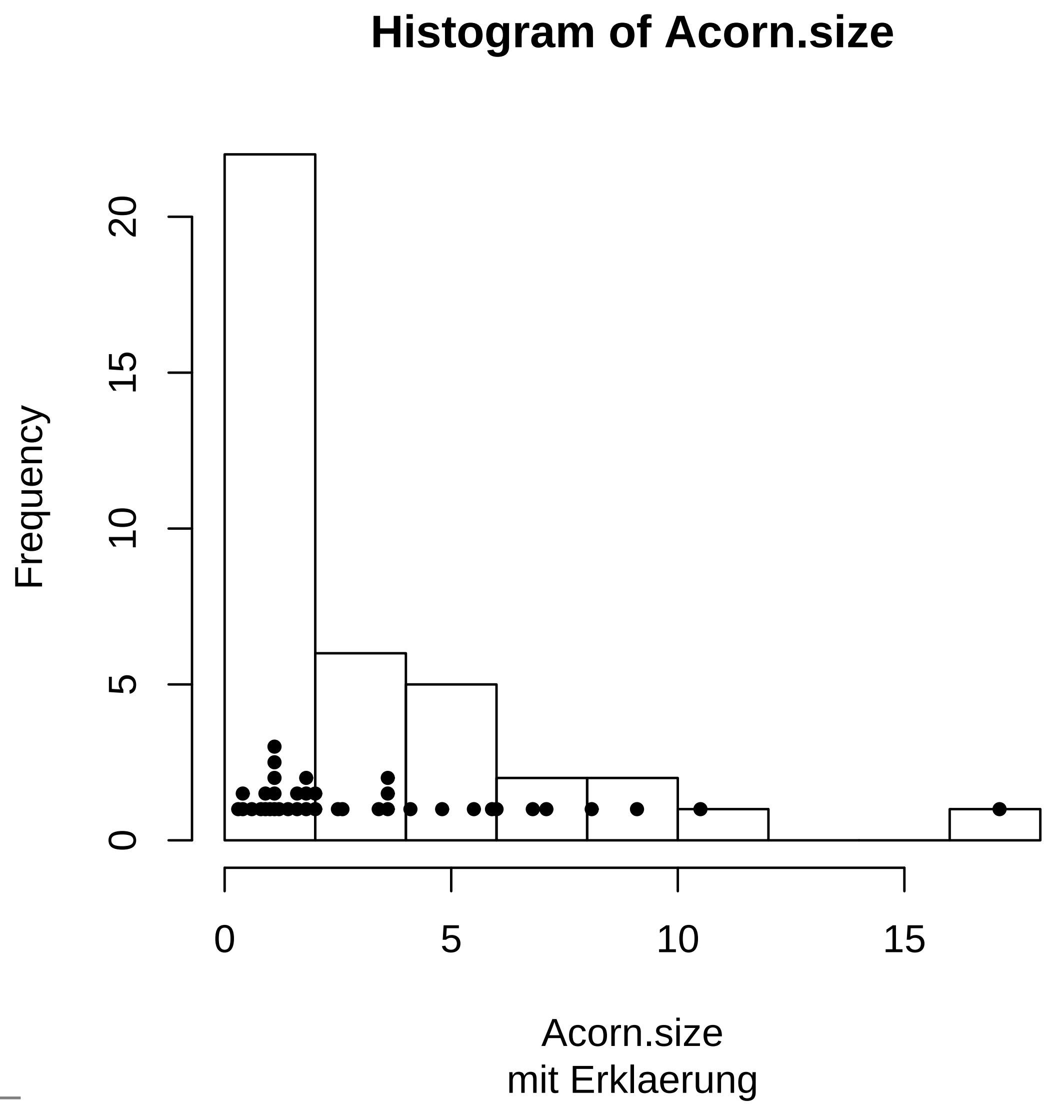
Punktdiagramm

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

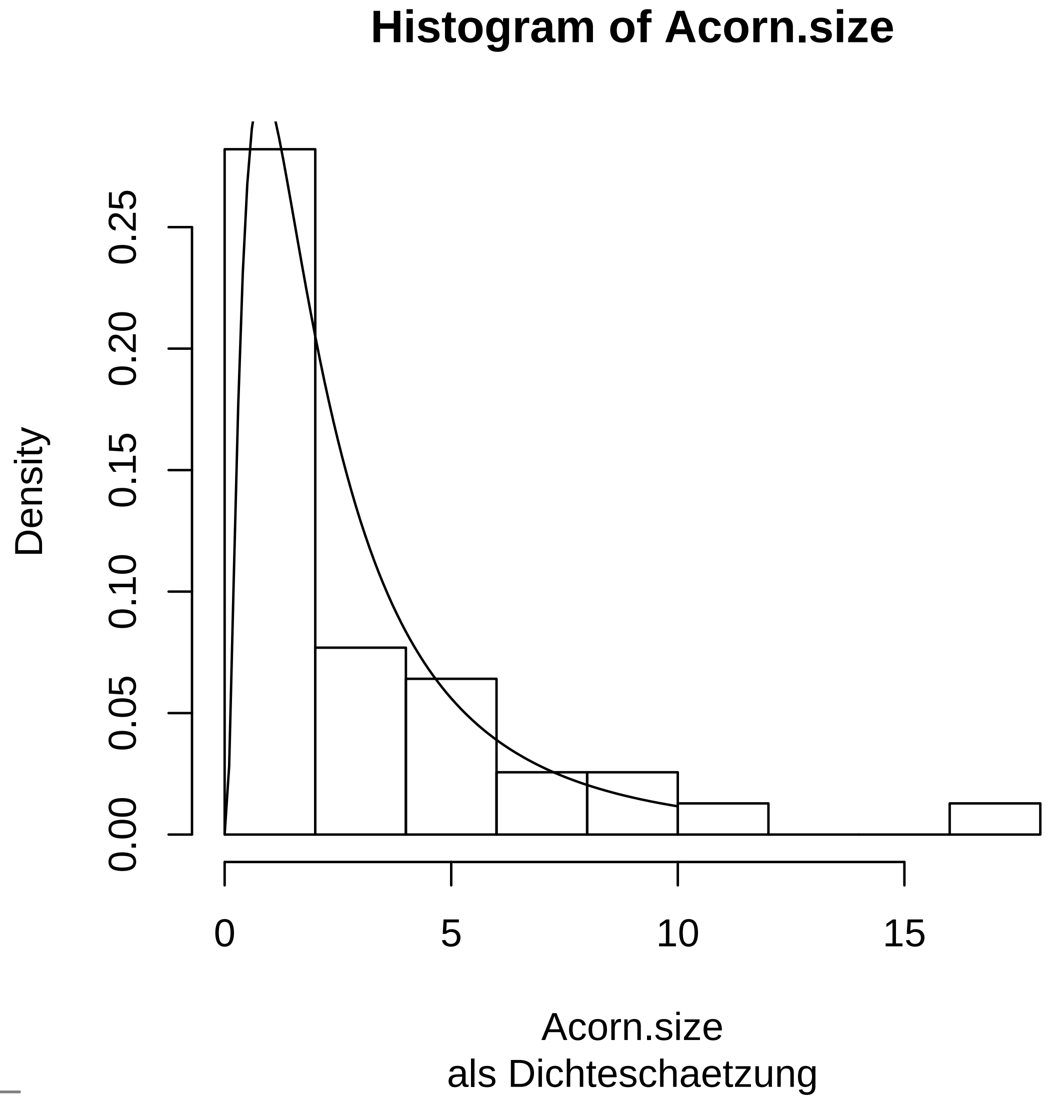
Histogramm


Histogram of Acorn.size
Frequency
0
5
10
15
20
0
5
10
15
Acorn.size

Histogramm


Histogram of Acorn.size
Frequency
0
5
10
15
20
0
5
10
15
Acorn.size
mit Erklaerung

Histogramm

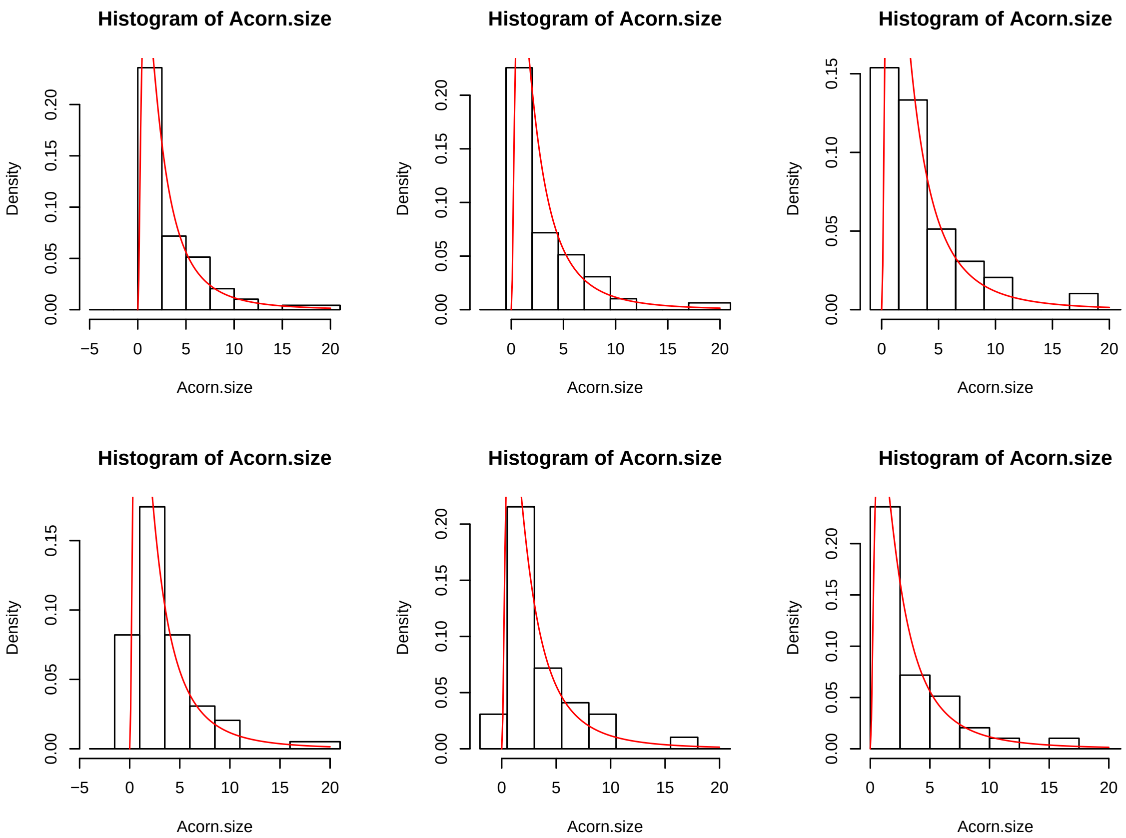


Histogram of Acorn.size
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
Acorn.size
als Dichteschaetzung

Histogramm

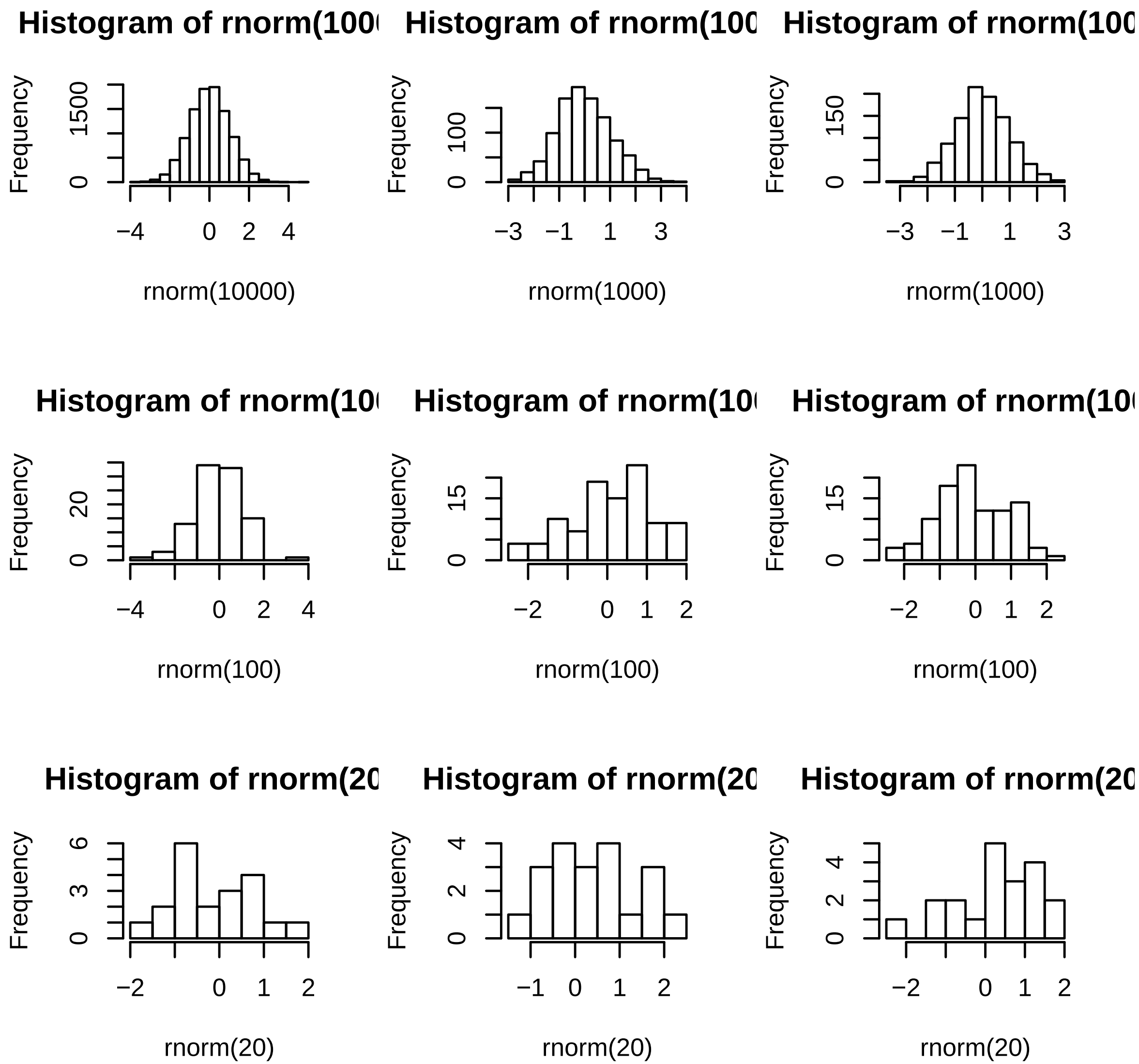
- Stellt Anzahl von Datenpunkten im Intervall dar.
- Stellt die Dichte (Datenpunkte pro Punkt und Einheitslänge) der Punkte dar.
- Balkenhöhe ist zufällig.
- Variation von Balkenanfang und Balkenanzahl führt zu verschiedenen Eindrücken.
- Zu kleine Balken $\Rightarrow$ “Zufallsflimmer”
- Zu große Balken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

Einfluß des Balkenanfangs

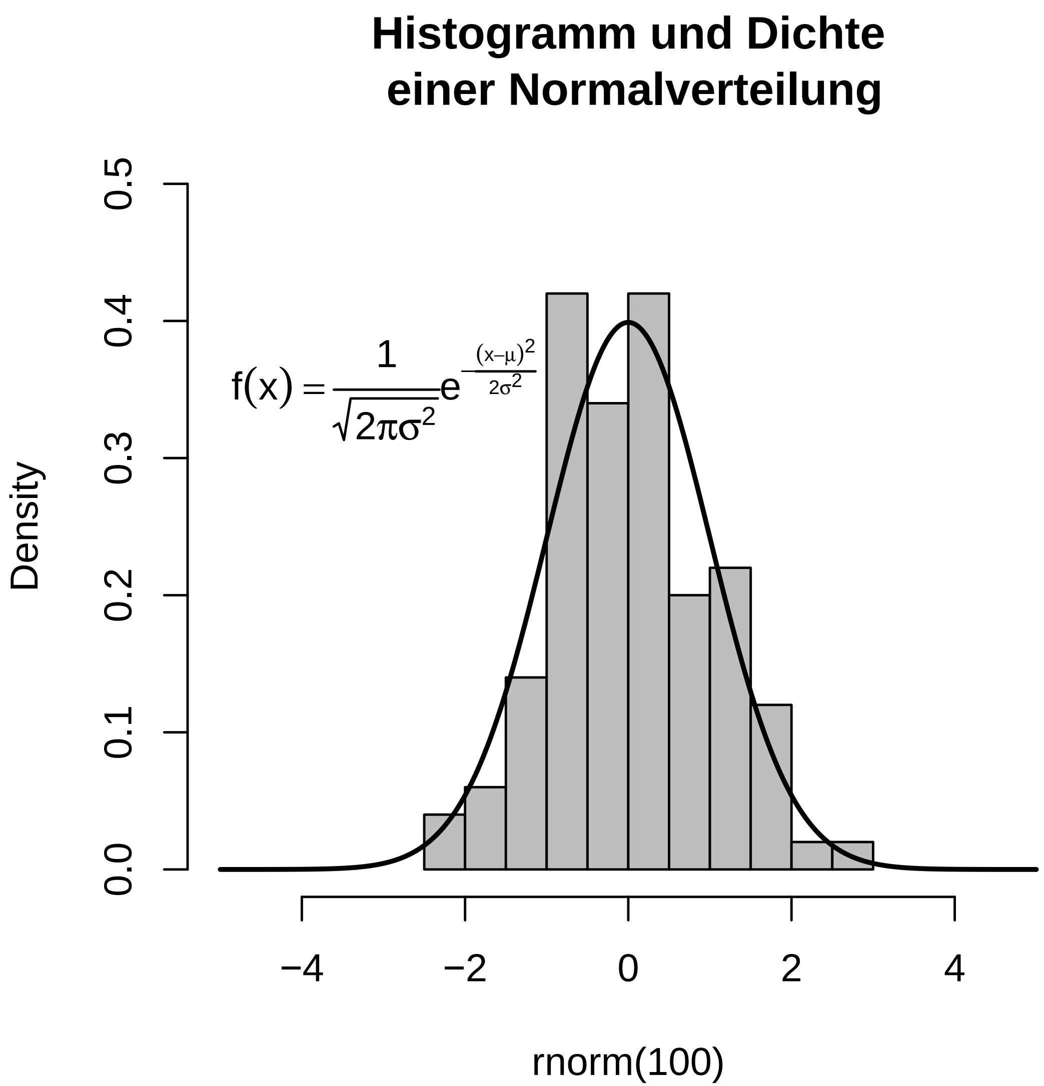


Beschreibung der Verteilungsform und Normalverteilung als Referenzverteilung

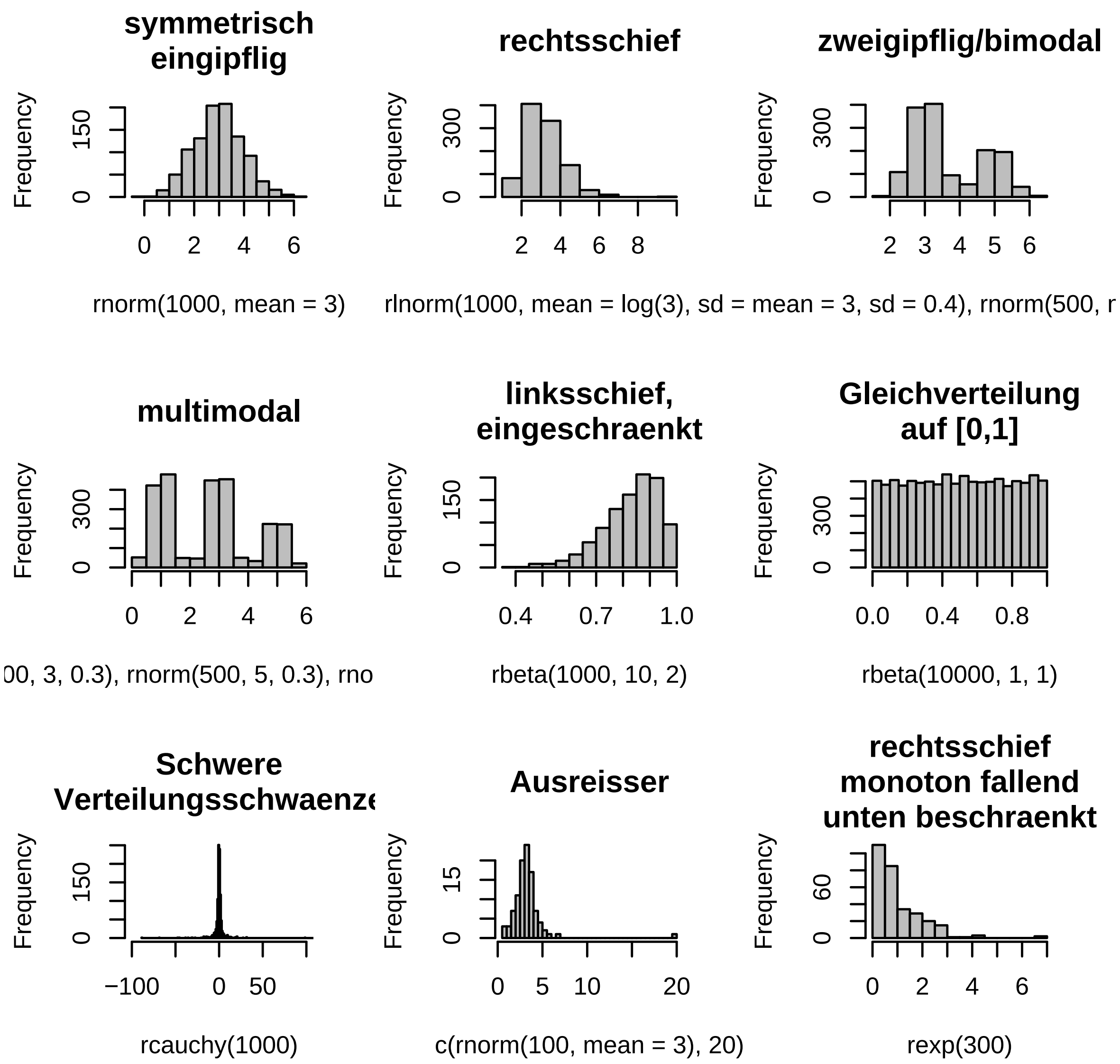
Normalverteilung



Dichte der Normalverteilung



Verteilungseigenschaften



Kenngrößen und Parameter

- Lage
- Streuung
- Form
- Verteilung

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

Lageparameter

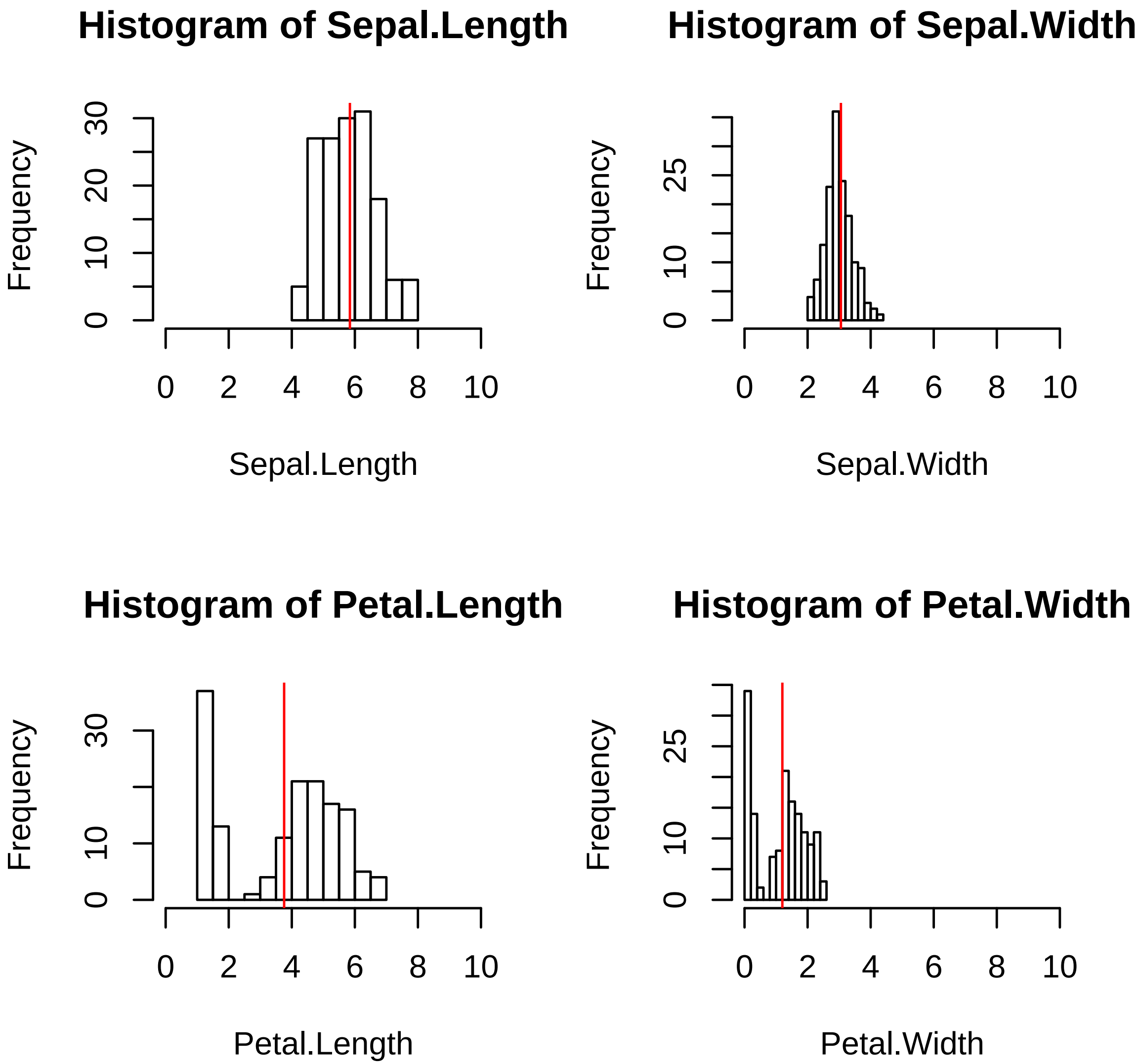
- Lage
  - Mittelwert (geometrisch und arithmetisch)
  - Median
  - Modus
  - Quantile (Quartile, Dezentile)
- Streuung
- Form
- Verteilung

(arithmetischer) Mittelwert

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

```
> mean(iris$Sepal.Length)
[1] 5.843333
```

Mittelwert



Histogram of Sepal.Length
Frequency
0 10 20 30
0 2 4 6 8 10
Sepal.Length
Histogram of Sepal.Width
Frequency
0 10 25
0 2 4 6 8 10
Sepal.Width
Histogram of Petal.Length
Frequency
0 10 30
0 2 4 6 8 10
Petal.Length
Histogram of Petal.Width
Frequency
0 10 25
0 2 4 6 8 10
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
```

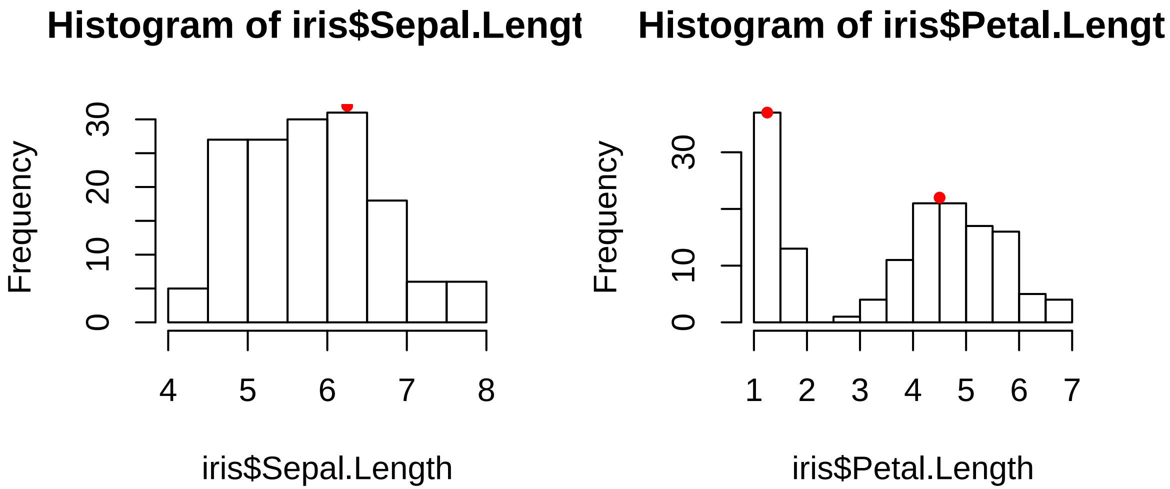
Median

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
```

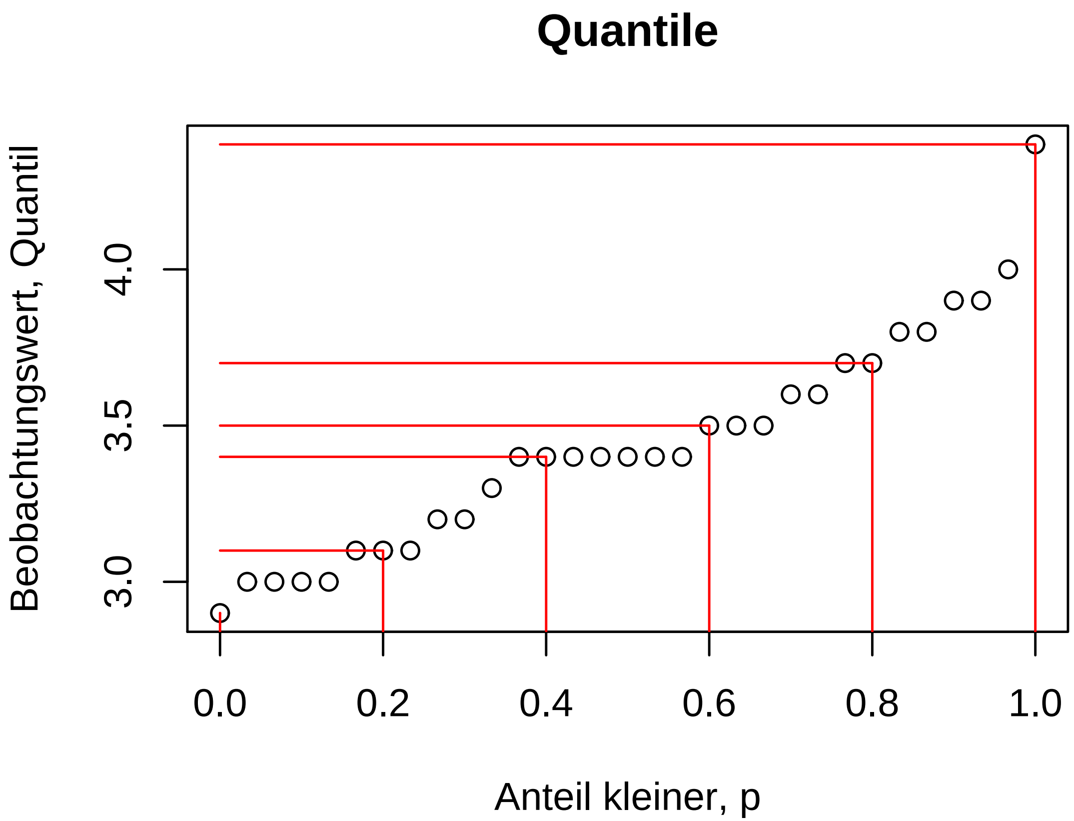
Modus

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



Quantile

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



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 Dezentil**
- 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 relle Skala

- Varianz

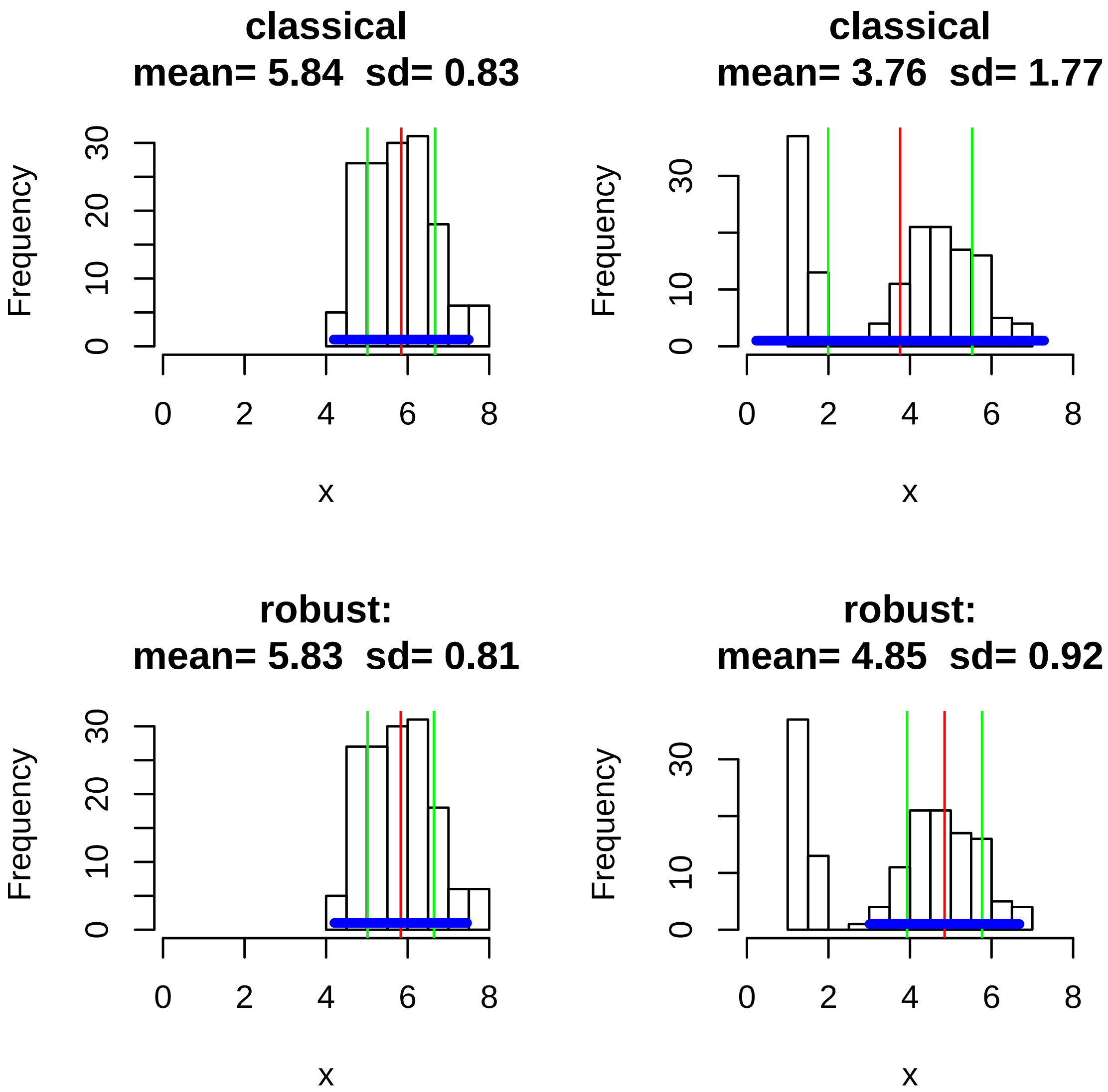
$$\widehat{var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Standardabweichung

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

- Interquartilsabstand

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


classical
mean= 5.84 sd= 0.83
classical
mean= 3.76 sd= 1.77
robust:
mean= 5.83 sd= 0.81
robust:
mean= 4.85 sd= 0.92
Frequency
x
0
2
4
6
8
0
10
20
30

Streuparameter für die ratio Skala

- Variationskoeffizient

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

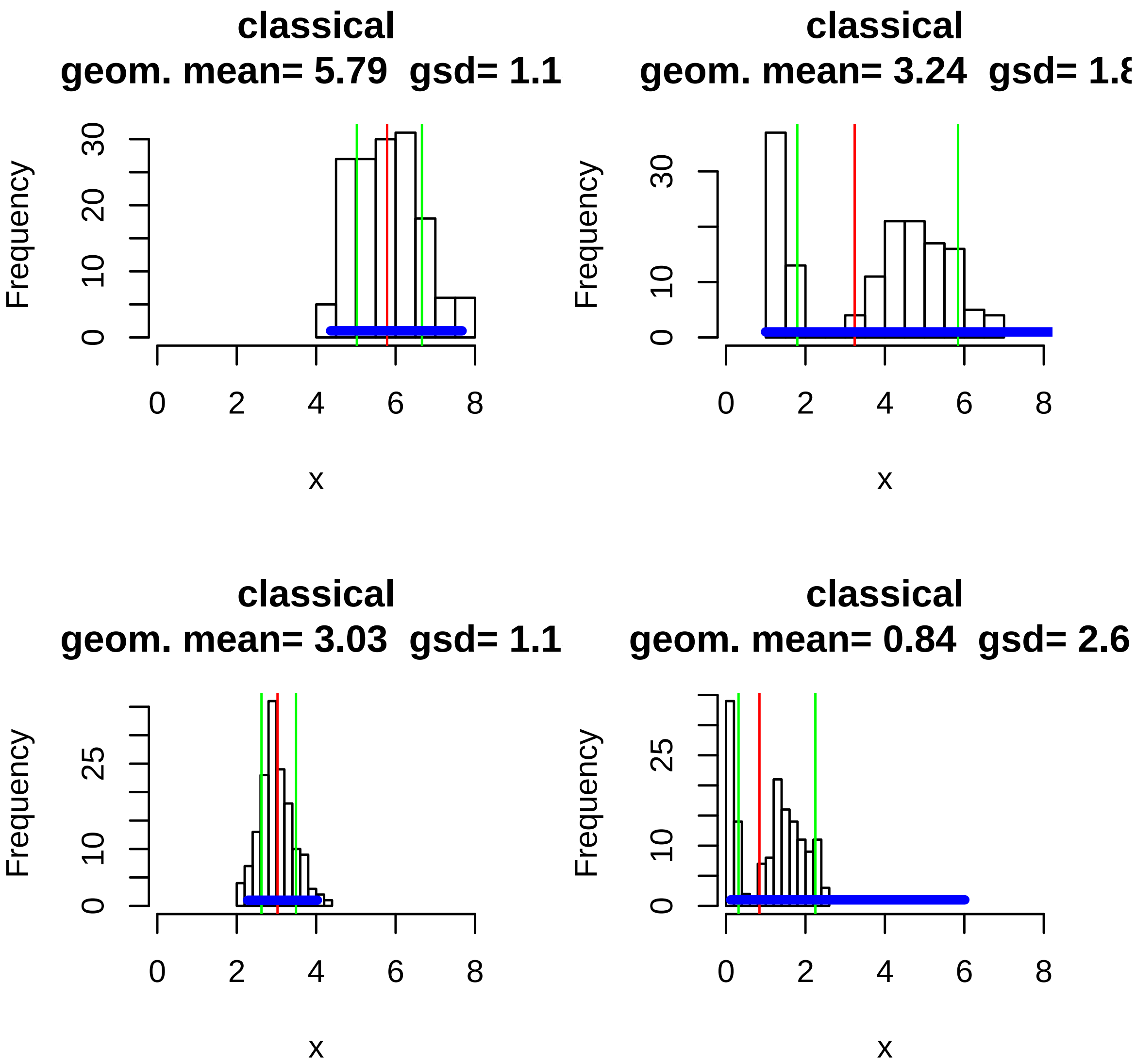
- Standardabweichung des Logarithmus

$$\widehat{sd}(\ln(X))$$

- Geometrische Standardabweichung

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

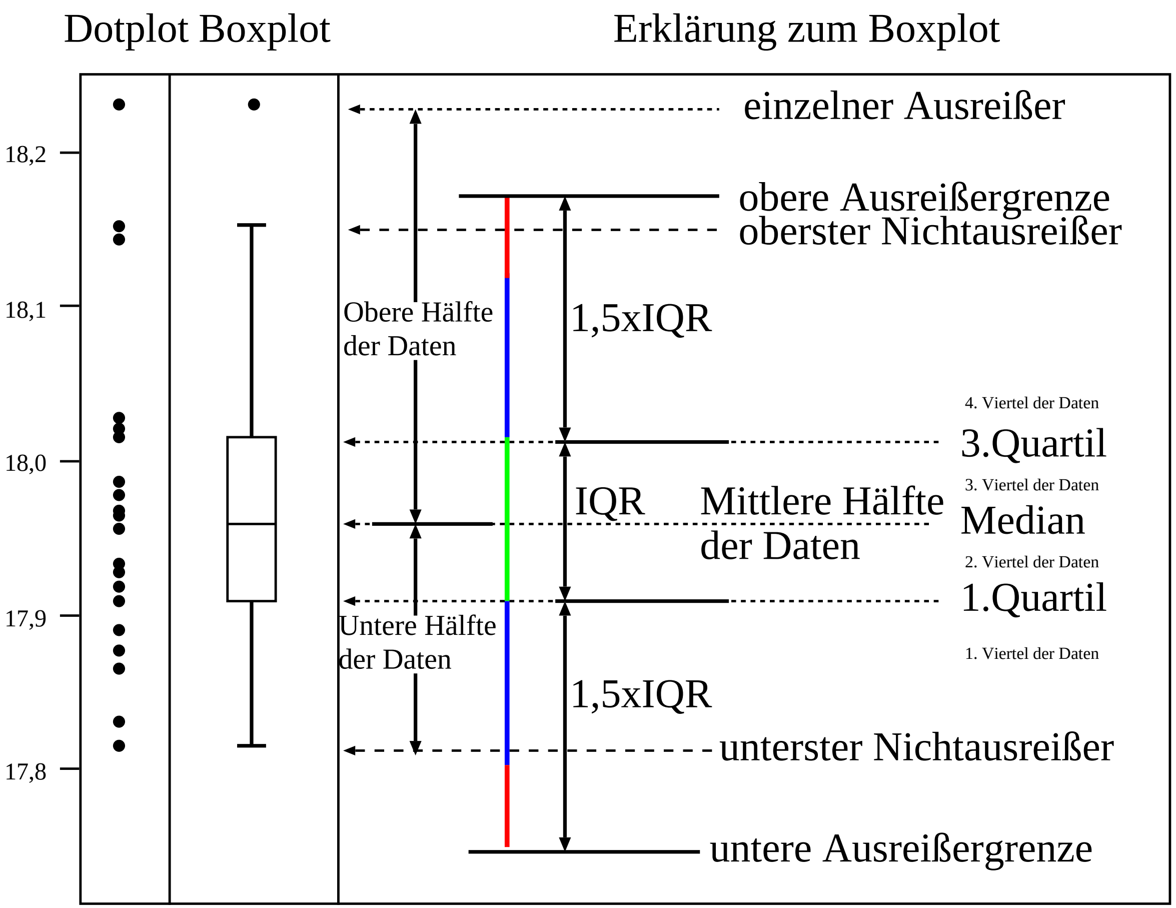
Blick mit der Ratioskala



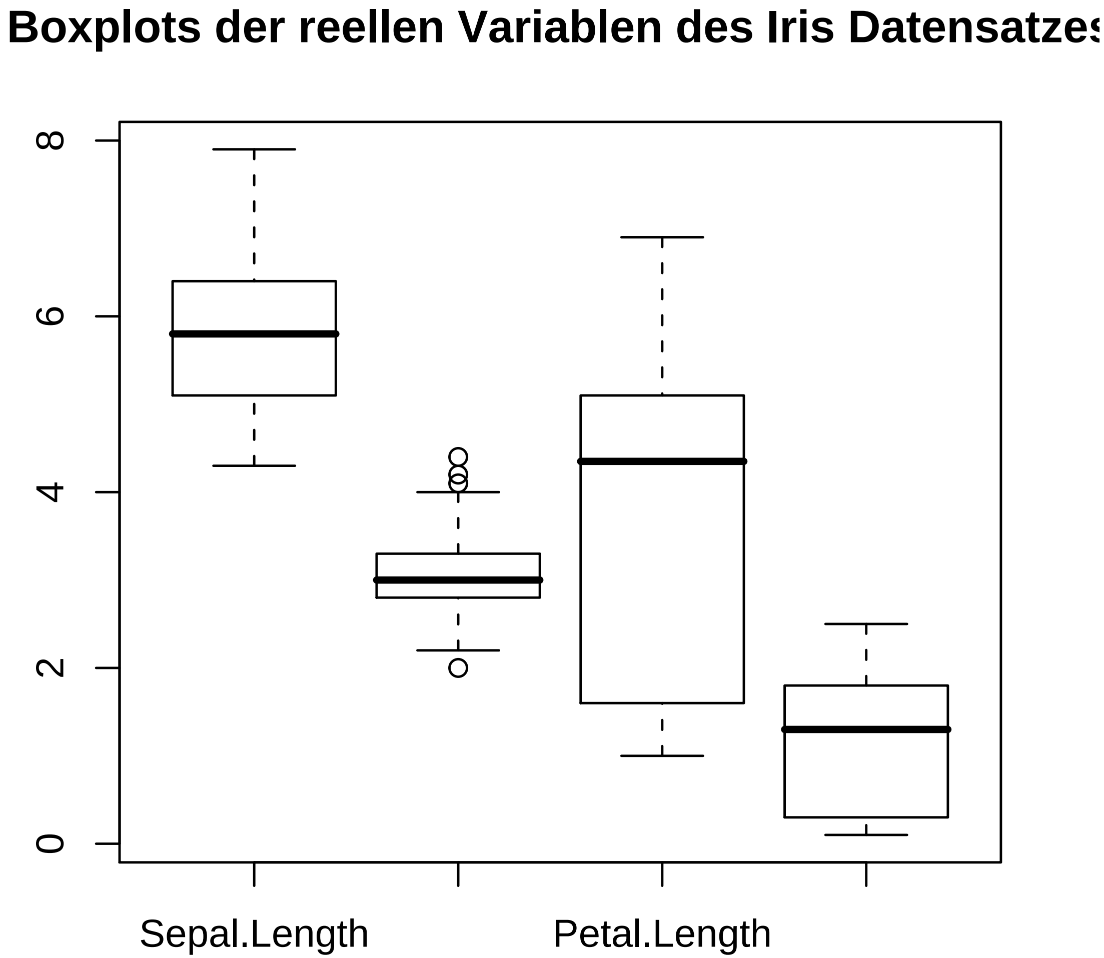
Weitere Parameter

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

Kastendiagramm/Boxplot



Kastendiagramme



Interpretation

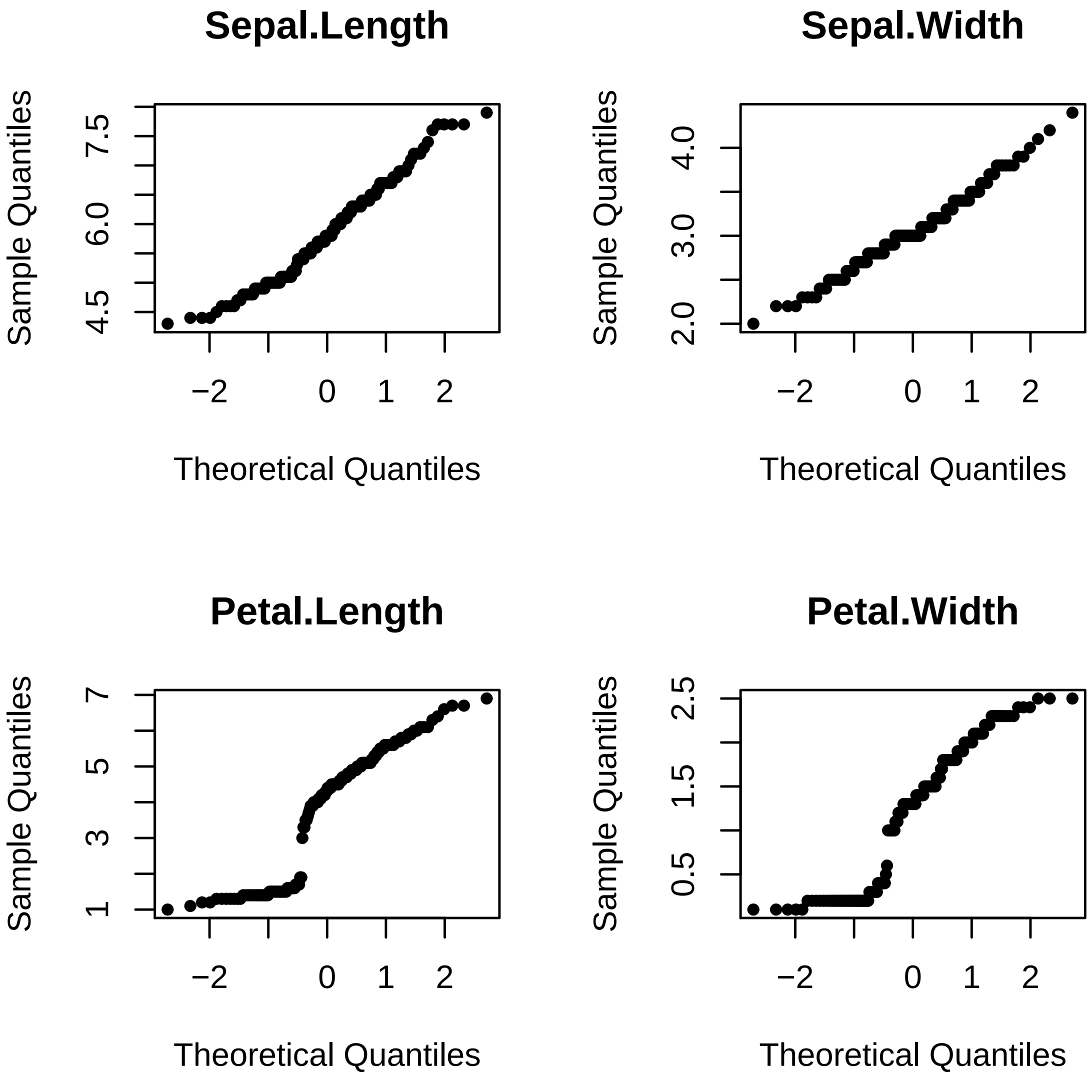
- Ausreißer
- Stichprobenlage / Median
- Stichprobenstreuung / IQR
- Symmetrie und Schiefe der Verteilung
- eventuell extreme Werthäufungen

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)
- Individum fehlerhafterweise in der Stichprobe (z.B. andere Art)
- Anthropogene Überprägung (das verlorene Geldstück mit hohem Kupfergehalt.)

Q Q-Plots



Sepal.Length
Sample Quantiles
4.5
6.0
7.5
−2
0
1
2
Theoretical Quantiles
Sepal.Width
Sample Quantiles
2.0
3.0
4.0
−2
0
1
2
Theoretical Quantiles
Petal.Length
Sample Quantiles
1
3
5
7
−2
0
1
2
Theoretical Quantiles
Petal.Width
Sample Quantiles
0.5
1.5
2.5
−2
0
1
2
Theoretical Quantiles

Interpretation Q Q-Plot

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

Exkurs: Bindungen

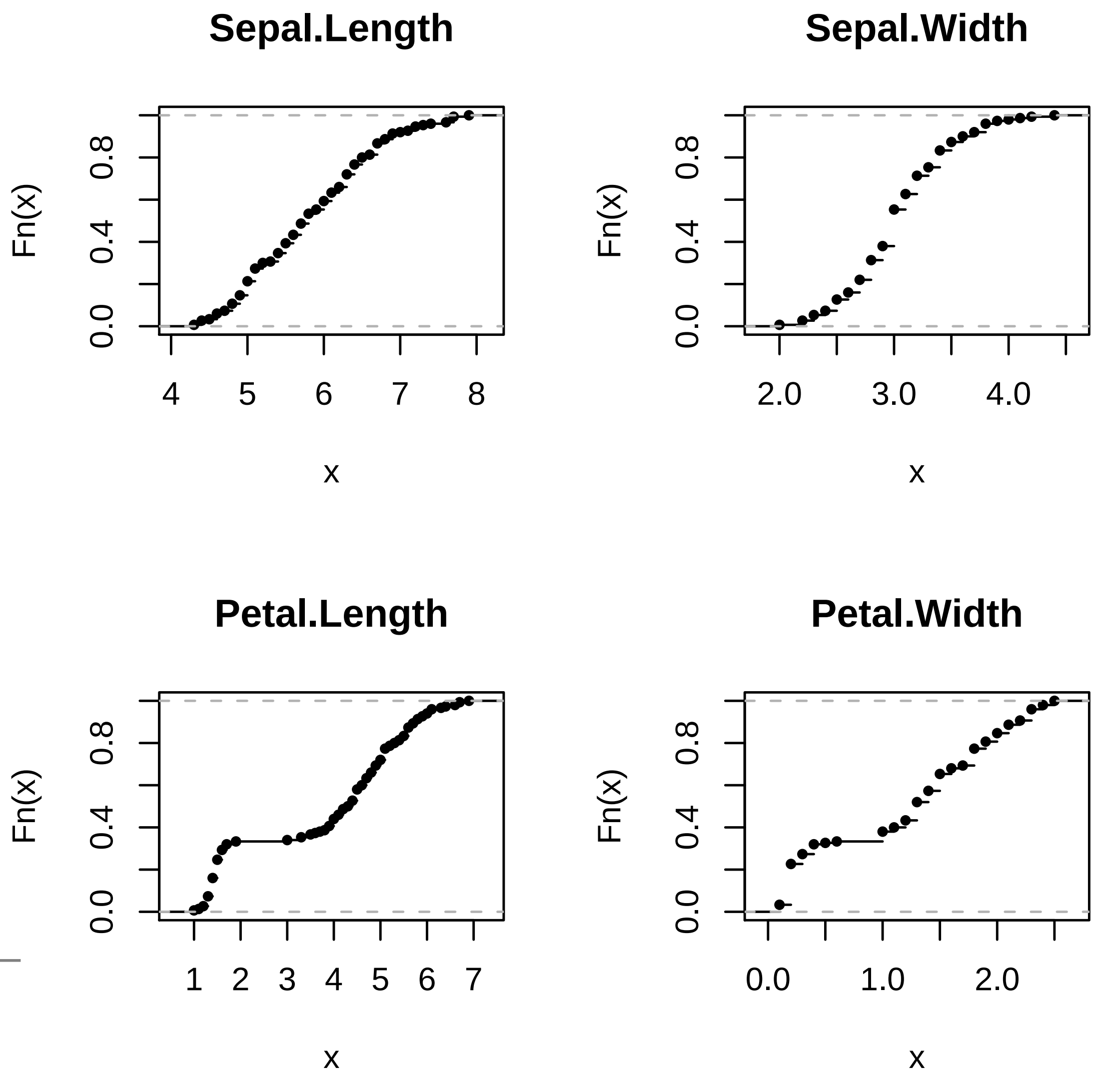
Definition: Von einer **Bindung** spricht man, wenn ein Datenwert in einer stetigen Variable zwei oder mehrfach auftritt.
Mögliche Ursachen:

- 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

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



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 varieren) / Verteilungsform
- **Kastendiagramm** / Ausreißer, Streung, Lage, Symmetrie
- **Q Q-Plot** / Vergleich mit Verteilung
- **Empirische Verteilungsfunktion** / Quantile