DESKRIPTIVE STATISTIK

Stichprobenumfang:
..wird üblicherweise mit n angegeben und entspricht der Gesamtzahl der erhobenen Daten eines Datensatzes, oder auch die Gesamtzahl der durchgeführten Versuche. Im allgemeinen nimmt mit zunehmendem Stichprobenumfang auch die Genauigkeit einer Aussage zu, allerdings ist dies meist auch mit zusätzlichem Aufwand verbunden.

Wertevorrat:
In der mathematischen Notation ist das ein wichtiger Begriff. Darunter versteht man die Anzahl der möglichen Ergebnisse eines Versuchs. Z.B. der Wertevorrat des Experimentes „Würfel werfen" ist {1,2,3,4,5,6}.

LAGEPARAMETER

Arithmetisches Mittel $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ oder $\bar{x} = \sum_{i=1}^{n} x_i . h_{(x_i)}$

$h(x_i)$...Häufigkeit des Auftretens von x_i

Mittlere Abweichung $d = \frac{\sum_{i=1}^{n} x_i - \bar{x}}{n}$,

die mittlere Abweichung ist 0 und daher kein geeignetes Maß um die Streuung der Daten zu quantifizieren

Mittlere absolute Abweichung

vom arithmetischen Mittel

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

vom Median

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

Die absolute Abweichung berechnet sich als der Betrag der Abstände der einzelnen Werte vom Median oder arithmetischen Mittelwert. Sie ist robuster als die

Standardabweichung. In der Regel wird die Abweichung vom Median verwendet, da das arithmetische Mittel weniger resistent gegenüber Ausreißern ist. Bisweilen wird statt der mittleren Abweichungen auch der Median der Abweichungen verwenden, da die Mittelung auch wieder anfällig gegenüber Ausreißern in den Daten ist.

Erwartungswert: Der Erwartungswert ist der wahre arithmetische Mittelwert der zugrundeliegenden Verteilung.

$$E(X) = \mu = \sum x_i . p(x_i), \quad p(x_i) \text{ Wahrscheinlichkeit von } x_i$$

Gewichtetes arithmetisches Mittel: $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

bei mehreren Messungen des Mittelwertes erfolgt die Wichtung mit den Kehrwerten der jeweiligen Varianzen s_i^2

$$\bar{x} = \frac{\sum_{i=1}^{n} \frac{x_i}{s_i^2}}{\sum_{i=1}^{n} \frac{1}{s_i^2}},$$

Geometrisches Mittel: $\bar{x}_G = \sqrt[n]{x_1 x_2 .... x_n}$ *bzw.* $\lg \bar{x}_G = \frac{1}{n} \sum \lg x$, alle x > 0

wenn Einzelwerte vorliegen, die selbst nicht normalverteilt sind, aber ihre Logarithmen.

Harmonisches Mittel:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}, \quad \bar{x}_h = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}, \quad \textit{bzw.} \ \bar{x}_h = \frac{1}{\sum_{i=1}^{n} \frac{w_i}{x_i}} \quad \textit{für} \ \sum_{i=1}^{n} w_i = 1$$

alle $x \neq 0$, w_i sind Wichtungsfaktoren

wird z.B. verwendet zur Berechnung der Durchschnittsleistung, wenn unterschiedliche Arbeitseinheiten in derselben Zeiteinheit erbracht worden sind. Sind die Zeiteinheiten unterschiedlich, dann sind entsprechende Wichtungsfaktoren w_i anzubringen.

Beispiel: Ein Autofahrer fährt staubedingt 10 km mit einer Geschwindigkeit von 20 km/h und danach 10km mit 60 km/h. Was ist die Durchschnittsgeschwindigkeit?
Die Durchschnittsgeschwindigkeit beträgt nicht, wie man irrtümlicherweise meinen könnte (20 + 60)/2 = 40km/h km/h, sondern v = 2/(1/20 + 1/60) = 30km/h. Die Nachrechnung zeigt, das der zurückgelegte Weg von 20km in ½ + 1/6 Stunde tatsächlich einer Geschwindigkeit von 30km/h entspricht.

Handelt es sich bei diesen Betrachtungen um unterschiedlich lange Wegstrecken, so sind die Geschwindigkeitsbeiträge entsprechend zu gewichten. Die Fahrt nach Wien bei einer Gesamtstrecke von 260km, wovon 250 km mit dem Zug in 3 Stunden und die restlichen 10 km in der Stadt in einer Stunde zurückgelegt werden, ergibt eine Durchschnittsgeschwindigkeit von v = 260/4 = 675 km/h.

Die Wichtungsfaktoren ergeben sich aus den einzelnen Streckenabschnitten, w_1=250; w_2=10, bzw. normiert w_1=250/260; w_2=10/260 (Nachrechnen !).

Modus oder Modalwert

Der Modus oder Modalwert ist der häufigste Wert einer Häufigkeitsverteilung. Da eine Verteilung mehrgipfelig sein kann, können einer Verteilung auch mehrere Modi zugeordnet sein

Perzentile und Quartile

Das p.100-te Perzentil ist der Wert mit der Ordnungszahl p (n+1) mit 0 < p < 1. Das 50igste Perzentil, oder das 0,5 Quantil, ist der Median. So ist z. B. das 50. Perzentil für n=80 mit 0,5 (80+1) = 40,5 der Mittelwert aus dem 40. und 41. Wert der aufsteigend geordneten 80 Meßwerte, das 10. Perzentil ware dann dementsprechend 0,1(80+1) = 8,1 der achte Wert.
Das 1. Quartil $Q_{0,25}$ ist somit das 25. Perzentil, der Median das 50. Perzentil, das 3. Quartil $Q_{0,75}$ das 75. Perzentil. Analoges gilt für Dezile.

(Inter-)Quartilabstand ((inter-)quartile range)

Hat man die Quartile $Q_{0,25}$ und $Q_{0,75}$ berechnet, so bezeichnet man deren Differenz als Quartilabstand (QR) oder Interquartilabstand (IQR):

$$\text{IQR oder } Q_R = Q_{0,75} - Q_{0,25}$$

Innerhalb des QR kommen 50% aller Messwerte zu liegen, er ist - wie auch der Median bzw $Q_{0,50}$ - unempfindlich gegenüber Ausreißern.

Median oder Zentralwert: $\tilde{x}$ (x Schlange)...

ist der mittlere Wert einer geordneter Datenreihe. Bei ungerader Anzahl ist der Median der mittlere Wert, bei gerader Anzahl der Datenpunkte ist der Median der arithmetische Mittelwert der beiden in der Mitte stehenden Einzelwerte. Der Median wird durch extrem liegende Werte kaum beeinflusst im Gegensatz zum arithmetischen Mittelwerte, der stark beeinflusst wird. Die Ausreißerunempfindlichkeit des Medians nennt man **Robustheit**. Wenn eine Klasseneinteilung vorliegt, und die einzelnen Datenpunkte nicht bekannt sind, dann wird der Median durch lineare Interpolation geschätzt

Median für klassierte Daten: wird durch lineare Interpolation ermittelt

$$\tilde{x} = \tilde{U} + b\left(\frac{n/2 - \left(\sum f\right)_{\tilde{U}}}{f_{Median}}\right)$$

$\tilde{U}$	untere Klassengrenze der Medianklasse
b	Klassenbreite
n	Anzahl der Werte
$\left(\sum f\right)_{\tilde{U}}$	Summe der abs. Häufigkeitswerte aller Klassen unterhalb der Medianklasse
f_{Median}	Anzahl der Werte in der Medianklasse

STREUUNGSPARAMETER

Unter Streuung fasst man in der deskriptiven Statistik verschiedene Maßzahlen zusammen, die der Einschätzung der Streubreite von Stichprobenwerten um ihren Mittelwert dienen. Die verschiedenen Berechnungsmethoden unterscheiden sich prinzipiell durch ihre Beeinflussbarkeit bzw. Empfindlichkeit gegenüber Ausreißern.

Spannweite R (range)

Die Spannweite ist das einfachste Streuungsmaß; Sie entspricht der Distanz zwischen dem größten und dem kleinsten Messwert:

$$R = x_{max} - x_{min}$$

R ist aufgrund der Tatsache, dass - unabhängig von der Stichprobengröße - nur zwei Werte (die so genannten Extremwerte) berücksichtigt werden, nicht robust gegenüber Ausreißern.

Varianz der Verteilung bzw. Grundgesamtheit

$$V(X) = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

μ ist der Erwartungswert der Verteilung, n der Umfang der Grundgesamtheit

Experimentelle Varianz.

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

Diese ist ein Schätzwert der Varianz aus dem Mittelwert der Stichprobe. n ist der Stichprobenumgang

Standardabweichung

Die Standardabweichung ist die Wurzel aus der Varianz; Damit hat sie die gleiche Dimension wie die Daten. Sie ist das bevorzugte Streuungsmaß bei quantitativ stetigen, symmetrisch verteilten Daten. Die Standardabweichung darf nur bei

quantitativen Merkmalen berechnet werden. Bei normalverteilten Daten liegen etwa 2/3 der Messwerte innerhalb des Intervalls $\mu \pm \sigma$ und etwa 95 % innerhalb des Intervalls $\mu \pm 2\sigma$

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

Mit der Division durch n-1 wird die Schätzung der Standardabweichung erwartungstreu.

Für die praktische Berechnung eignet sich folgende Formel besser:

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

Bei klassierten Daten (k Klassen) rechnet man mit den Klassenmitten x_j und den Klassenhäufigkeiten n_j. n ist die Gesamtzahl aller Beobachtungen.

$$s = \sqrt{\frac{\sum_{j=1}^{k} n_j x^2{}_j - \frac{(\sum_{j=1}^{k} n_j x_j)^2}{n}}{n-1}}$$

Standardabweichung oder Standardfehler des Mittelwertes (SEM = standard error of the mean)

$$SEM = s(\bar{x}) = \frac{s}{\sqrt{n}}$$

Schiefe (Skewness)

Die Schiefe entspricht dem 3. zentralen Moment. Berechnet wird sie bei quantitativen Merkmalen nach der Formel:

$$g_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Bei symmetrischen Verteilungen hat die Schiefe den Wert 0. Falls g1 > 0, ist die Verteilung rechtsschief (oder linksgipfelig). Falls g1 < 0, ist die Verteilung linksschief (oder rechtsgipfelig).

Variationskoeffizient CV (coefficient of variation)

CV = s / AM; Der CV ist die relative Standardabweichung bezogen auf den Mittelwert, d.h. das Verhältnis s/AM. Meist wird der CV in % angegeben.

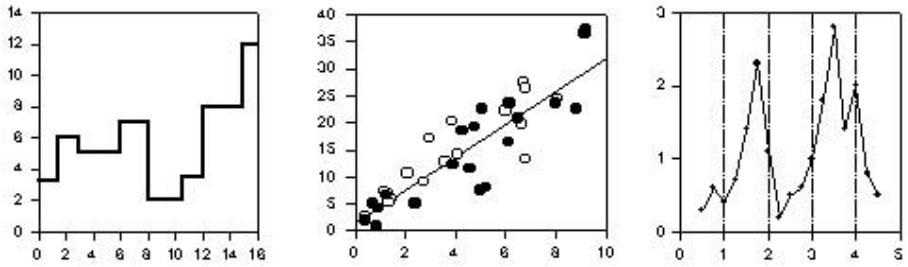
GRAPHISCHE DARSTELLUNG

(teilweise aus der Hilfe von Sigma Plot)

2D Plot Types

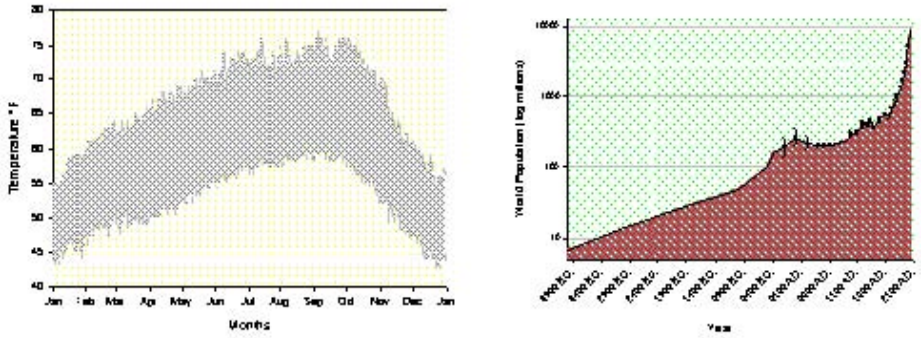
Scatter, Line, and Line/Scatter Plots

Scatter, line, and line/scatter plots graph data as symbols, as lines only with no symbols, or as symbols and lines. Line shapes can be straight segments, splines, or steps. Add drop lines to either axis to any of these plot types, and add error bars to plots with symbols. Draw linear or polynomial regressions with confidence and prediction intervals for each curve.



Area Plots

Using area plots, you can fill an area under a curve with a color making the curve easier to see. You can orient the fill up, down, left, or right. If your curve is a closed polygon, you can also fill the polygon. You can have multiple curves (plots) on a page, so you can stack Area Plots.

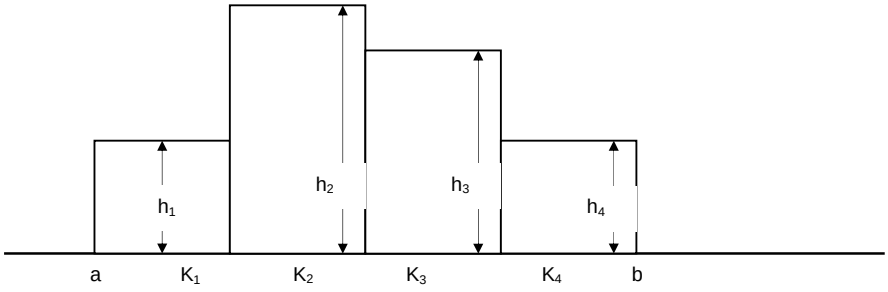


Histogramme (Bar Charts)

Histogramm

Über eine Stichprobe mit n reellen Messwerte $x_1,....,x_n$ wird zunächst ein Intervall [a,b] gelegt, das alle Meßwerte enthalte. Dieses Intervall wird durch k in gleich große

Teilintervalle $K_1,....,K_n$ der Länge $L = (b-a)/k$ zerlegt. Über jeder Klasse K_j wird ein Rechteck der Höhe h_j errichtet.



Für die Höhe h_j des Intervalls K_j sind üblich:

n_j	absolute Häufigkeit = Anzahl der x_i in der Klasse K_j
n_j/n	relative Häufigkeit
n_j/nL	Dichteschätzung

Mit der Angabe von n_j/nL auf der Ordinate wird die durch das Histogramm definierte Treppenfunktion eine Dichtefunktion, d.h. $\int_a^b h(x)dx = 1$. Das erlaubt den Vergleich mit über das Histogramm gelegten theoretischen Dichten.

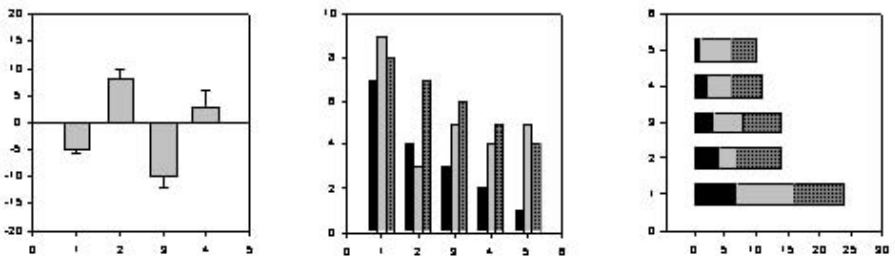
Wahl des Intervalls [a,b] und der Klassenzahl k: In vielen Statistikprogrammen gibt es eine automatische Bestimmung von [a,b] und der Klassenzahl k, nach nicht immer ganz klaren Regeln. Die Klassenzahl wird bei geringen Stichproben meist zu groß gewählt. Für die manuelle Festlegung der Klassenzahl gibt es folgende einfache Regel:

$[a,b] = [\min(x_i), \max(x_i)]$,

$$k = \begin{cases} \lfloor 2\sqrt{n} \rfloor & \text{für } n \le 100 \\ \lfloor 10\log(n) \rfloor & \text{für } n > 100 \end{cases}$$

Bar charts plot data either as vertical or horizontal bars. They originate from zero in either a positive or negative direction. Simple bar charts plot each row of data as a separate bar, and grouped bar charts plot multiple columns of data by grouping data in the same rows. Stacked bar charts plot data as segments of a bar; each data point is drawn as a bar segment starting where the previous data point ended.

Use the Graph Properties dialog box to modify bar width, bar fill colors, and bar fill patterns. Add error bars to simple and grouped bar charts.



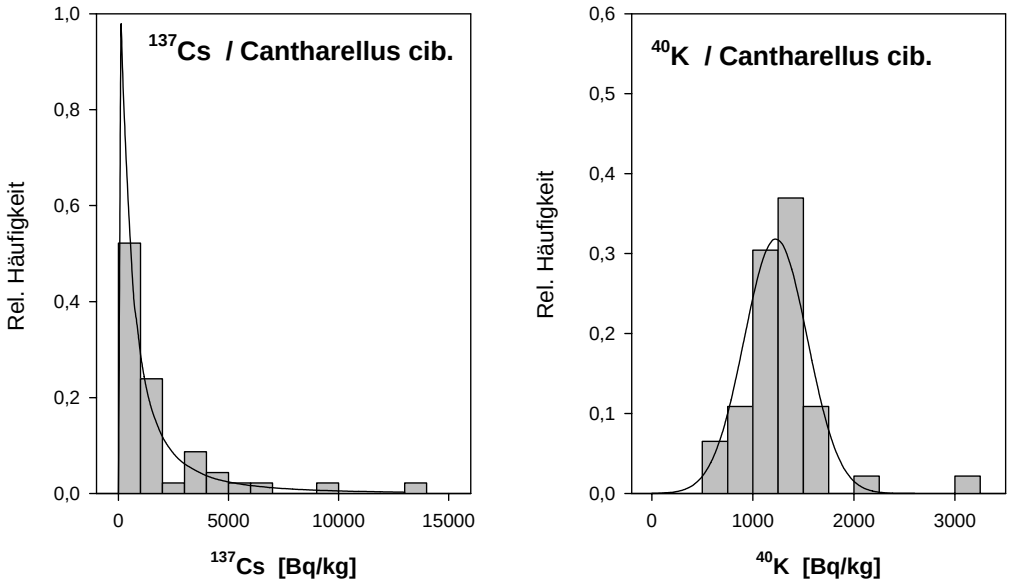
Nachstehend sind zwei Beispiele für Häufigkeitsverteilungen von ^{137}Cs und ^{40}K in Cantharellus cibarius (Pfifferling) mit den zugehörigen Modellen angeführt.

Normalverteilung: ^{40}K

Die experimentellen Daten von ^{40}K sind unterteilt in Klassen mit einer Klassenbreite von 500 Bq/kg. Der Stichprobenumfang beträgt n = 45, die Faustregel Anzahl Klassen = $\lfloor 2\sqrt{n} \rfloor$ lässt sich bei dem vorgegebenen Maximalwert also gut anwenden. Neben der Faustregel für die Anzahl der Klassen ist immer auch der Zahlenbereich zu berücksichtigen, in dem die Daten liegen. Die Klassenbreite sollte sich gut in das Dezimalsystem einfügen, z.B. wäre eine Klassenbreite von 300 o.ä nicht empfehlenswert. Die Höhe der Säulen im Histogram gibt die relative Häufigkeit der jeweiligen Klasse wieder und sie entspricht damit dem Integral über die Häufigkeit in der jeweiligen Klasse. Empfehlenswert für die Skalierung der Ordinate ist die Wahl der relativen Häufigkeit und nicht der absoluten Häufigkeit, weil auf diese Weise die Verteilung mit „einem Blick" besser erfasst werden kann. Die absolute Häufigkeit kann evtl. als zusätzliche Achse dargestellt werden. Die experimentelle Häufigkeitsverteilung von ^{40}K lässt sich mit einer Normalverteilung gut approximieren, wenn der Ausreißer in der Klasse [3000,3500] weggelassen wird. Die experimentell ermittelten Schätzungen für die Parameter μ und σ der Normalverteilung sind $\bar{x}$ = 1327 Bq/kg (als Schätzung für μ) und s = 313 Bq/kg (als Schätzung für σ). Die Dichtefunktion (pdf) $\phi(x)$ der Normalverteilung kann jetzt über das Histogram gelegt werden:

$$\phi(x) = \frac{1}{2\sqrt{\pi}\sigma} e^{-\frac{(\mu - x)^2}{2\sigma^2}}$$

Für die Skalierung auf der Ordinate ist noch zu berücksichtigen, dass die Säulen jeweils die kumulierte Häufigkeit über eine Klasse darstellen und nicht die mittlere Häufigkeit in einer Klasse! Sie sind also um den Faktor kb = Klassenbreite überhöht! Um die Normalverteilung (gilt für andere Verteilungen sinngemäß) entsprechend anzupassen, ist also nicht die Dichtefunktion darzustellen sondern eine angepasste Dichtefunktion $\phi^*(x) = \phi(x) \cdot kb$, mit kb der Klassenbreite



Log-Normalverteilung: ^{137}Cs

Die experimentell ermittelten ^{137}Cs Konzentrationen sind nicht wie die ^{40}K normalverteilt, sondern log-normalverteilt (nach Komogorov-Smirnov Test). Eine Log-Normalverteilung liegt vor, wenn die Logarithmen der Beobachtungen normalverteilt sind. Zur Ermittlung der Parameter μ und σ werden die natürlichen Logarithmen der Originaldaten verwendet und daraus $\bar{x}$ als Schätzung für μ, und s als Schätzung für σ berechnet. Im vorliegenden Histogramm ist eine Klassenbreite von 1000 Bq/kg gewählt, daraus ergeben sich insgesamt 14 Klassen, was der Faustregel zur Festlegung der Klassenzahl sehr gut entspricht. Über das Histogramm wird wieder die Dichtefunktion gelegt, in diesem Fall die Dichtefunktion $\phi(x)$ der Log-Normalverteilung:

$$\phi(x) = \frac{1}{\sqrt{2\pi}\ \sigma\ x} e^{-\frac{(\mu - \ln x)^2}{2\sigma^2}}$$

Die Darstellung im Histogram erfolgt in gleicher Weise wie für die Normalverteilung, in dem die Daten in Klassen eingeteilt und dargestellt werden. Analog der Skalierungskorrektur in der Normalverteilung erfolgt auch für die Normalverteilung die Anpassung der Dichtefunktion zur Darstellung nach $\phi^*(x) = \phi(x) \cdot kb$, mit kb der Klassenbreite.

Box and Whiskers ("Schachtel – Schnurhaar")

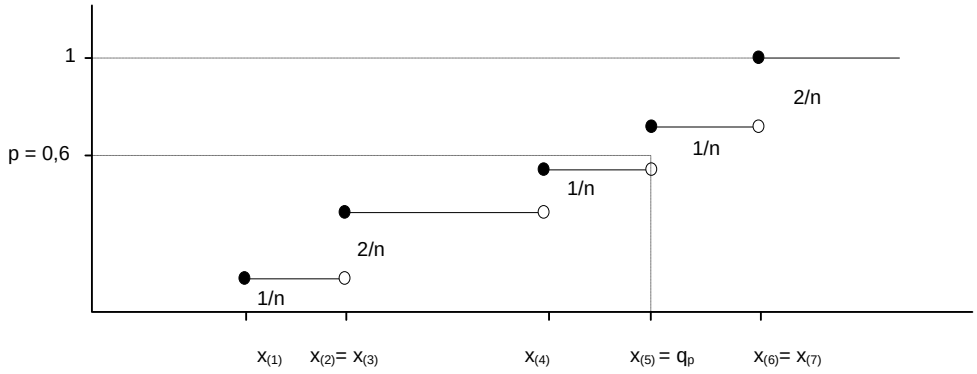
Empirische Quantile: In einer geordneten Stichprobe $x_{(1)} \le x_{(2)} \le x_{(3)\ldots..} \le x_{(n)}$ mit $0 \le p \le 1$ heißt

$$q_p = \begin{cases} 1/2.(x_{(np)} + x_{(np)+1}) & \textit{falls np ganzzahlig} \\ x_{(\lfloor np \rfloor + 1)} & \textit{falls np nicht ganzzahlig} \end{cases}$$

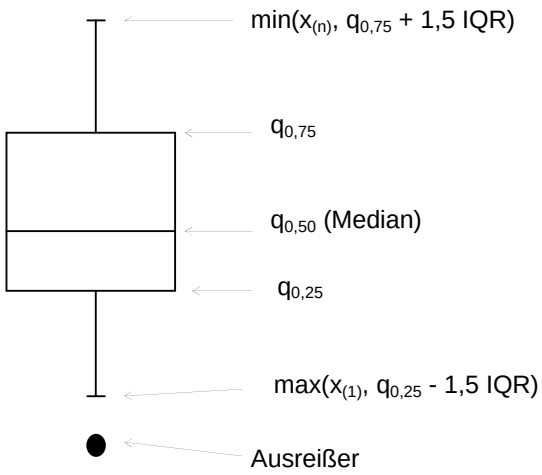
das p-te empirische Quantil oder p-tes empirisches Perzentil.
Bedeutung von q_p: ca. 100.p% der Meßwerte sind kleiner als q_p.
Die empirischen Quantile hängen eng mit der empirischen Verteilungsfunktion zusammen. Diese ist für eine Stichprobe $x_1,\ldots..,x_n$ durch

$$\hat{F}_n(x) = \frac{1}{n} x\{Anzahl\ der\ x_j \le x\} \quad x \in R$$

definiert. $\hat{F}_n$ ist eine Treppenfunktion, die bei jedem Stichproben wert x(j) der geordneten Stichprobe $x_{(1)} \le x_{(2)} \le x_{(3)\ldots..} \le x_{(n)}$ um den Wert 1/n wächst und zwischen den Stichprobenwerten konstant bleibt.



Der Boxplot ist eine graphische Darstellung einer Stichprobe, die auf empirischen Quartilen beruht.



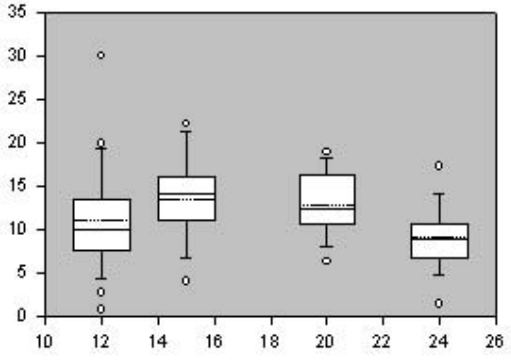
Ausreißer oder Extremwerte

Diese können real oder auch nur Artefakte sein, z.B. Meß- oder Schreibfehler. Es gibt verschiedenste Test um Ausreißer zu identifizieren, die meistens irgendwie auf Annahmen über die zugrunde liegende Verteilung aufbauen. Jedenfalls muß immer sichergestellt werden, dass nicht ein extremer Wert irrtümlicherweise vorschnell entfernt wird, der ein realer Meßpunkt und damit eine wichtige Information sein könnte.

Eine Methode (unter vielen):
x = Ausreißer, falls x außerhalb des Intervalls [Median ± 3 x IQR] liegt.

Box Plots

Box plots graph data as a box representing statistical values. The boundary of the box closest to zero indicates the 25th percentile, a line within the box marks the median, and the boundary of the box farthest from zero indicates the 75th percentile. Whiskers (error bars) above and below the box indicate the 90th and 10th percentiles. In addition, you can graph the mean and outlying points.



***You need a minimum number of data points to compute each set of percentiles**. At least three points are required to compute the 25th and 75th percentiles, five points to compute the 10th percentile, and six points to compute the 5th, 90th, and 95th percentiles. If SigmaPlot is unable to compute a percentile point, that set of points is not drawn.*