Datenanalyse

Ausgearbeitete Fragen/ Zusammenfassung

(Dieses Dokument basiert auf Prüfungsfragen von Prof Filzmoser, teilweise aus dem Skript erarbeitet, teilweise aber aus anderen Quellen da das Skript nicht ausreicht)

Allgemeine Tipps zur Prüfung: die Prüfung ist nicht zu unterschätzen!! man muss sich wirklich mit dem Stoff auskennen! Regel Nummer 1 = Formeln auswendig können (leider) und dann jede einzelne Variable erklären können- und zusätzlich ein allgemeines Verständnis für ein Kapitel haben (Wofür macht man das überhaupt, Sinn dahinter)

Ansonsten ist Professor Filzmoser sehr nett, auch wenn man nichts kann wird er nicht gemein sondern stellt noch mehr Fragen, merkt aber wenn man wirklich keine Ahnung hat

## Histogramm

**Fragen:**

- Histogrammfunktion

- Wahl der Intervalllänge nach verschiedenen Methoden (Wichtig: Histogramm-Funktion aufschreiben können + Indikatorfunktion erklären können)- welche Parameter fließen in die Funktion ein

- was ist Ziel eines Histogramms

- Wieso werden Intervallängen angepasst

- Zusammenhang zur DichteFkt erklären

Frage aus dem Fach “Explorative Datenanalyse”:

* Histogramme (hier wurde auch ganz grob gefragt, welche Variablen wohl in die verschiedenen Kriterien für "optimale" Klassenbreiten hineinspielen, und was "optimal" überhaupt bedeutet)

## Histogramm-Funktion





Indikatorfunktion:



Die Histogramm-Funktion ist hier mit relativen Häufigkeiten definiert, kann aber auch mit absoluten Häufigkeiten dargestellt werden. Die Division durch die Intervallbreite macht vor allem Sinn bei nicht äquidistanten Intervallgrenzen, weil dann die einzelnen Balken des Histogramms flächenmäßig richtig dargestellt werden.

Ein Histogramm ist die graphische Darstellung der Häufigkeitsverteilung von Messwerten. Man geht dabei von den nach ihrer Größe geordneten Daten aus und teilt sie in eine bestimmte Anzahl von Klassen auf. Diese können, müssen aber nicht gleich breit sein. Über jeder Messwertklasse wird ein Rechteck errichtet, dessen Fläche proportional zur Häufigkeit dieser Klasse ist.

Der absolute Wert entspricht der Anzahl an Werten, die zu einer Klasse gehören. Der relative Wert hingegen drückt aus, wie viel Prozent der Werte einer Klasse angehören. Je nach Anwendungsfall kann sowohl das Arbeiten mit absoluten als auch mit relativen Werten Vorteile mit sich bringen. [<https://www.quality.de/lexikon/histogramm-definition/>]

Wo werden Histogramme speziell verwendet?

In der digitalen Bildverarbeitung

## Ziel von Histogrammen

Ein Histogramm liefert Dir die grafische Darstellung der absoluten oder relativen Häufigkeitsverteilung eines quantitativen, klassierten Merkmals in einem speziellen Säulendiagramm: Der Flächeninhalt der einzelnen (aneinandergrenzenden) Säulen gibt die Häufigkeit der jeweiligen Klassen wider, die Höhe der Säulen steht für die Häufigkeitsdichte der Klasse. Damit erhältst Du also mehr Informationen als es ein Säulendiagramm zu liefern vermag.

## Wahl der Intervalllänge:

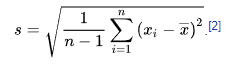
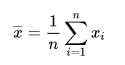
Intervalllänge nach Sturges:

Diese Wahl ist für Daten gedacht, die aus einer normalverteilten Grundgesamtheit kommen.

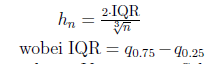
Intervalllänge nach Scott: empfindlicher gegenüber Ausreißern



empirische Standardabweichung:

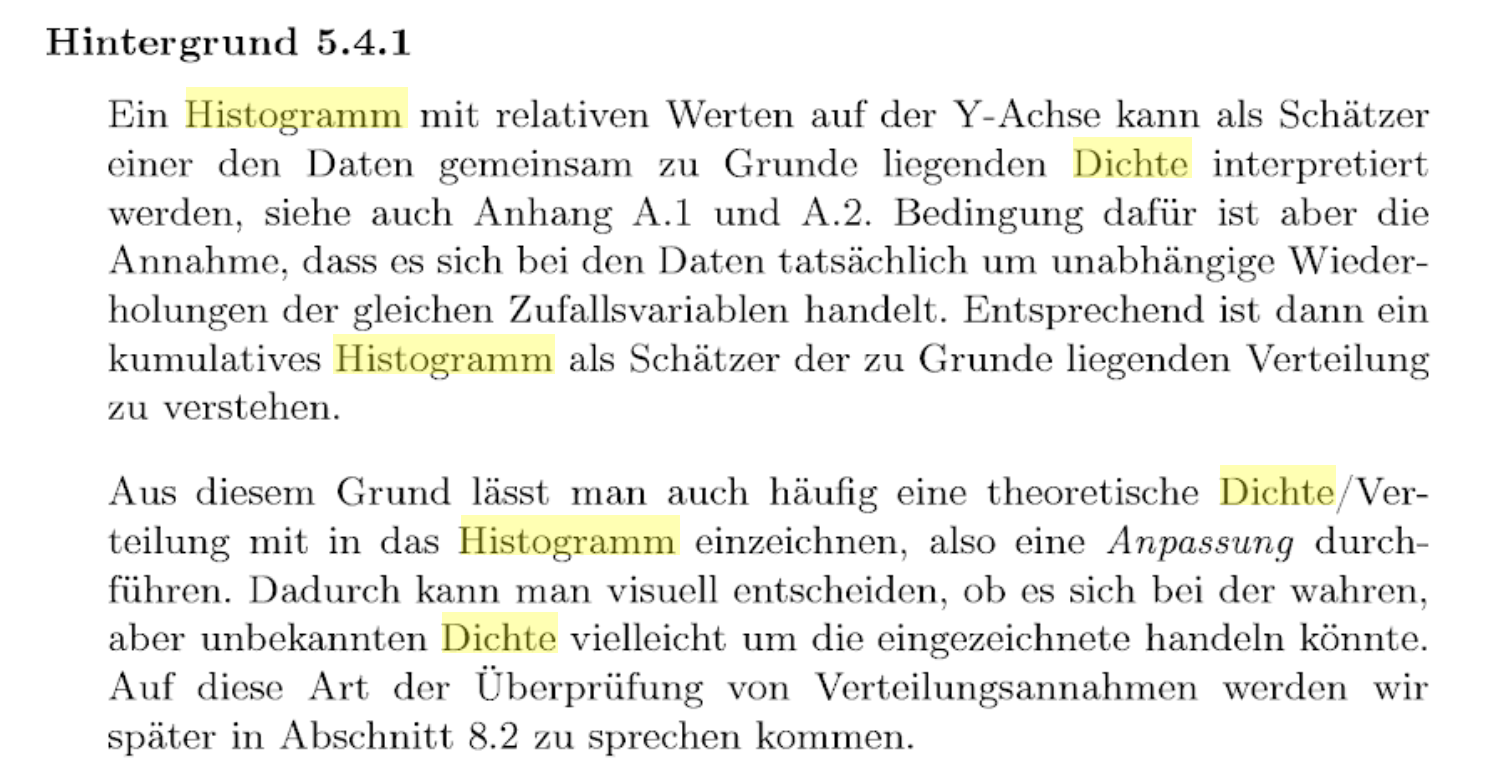


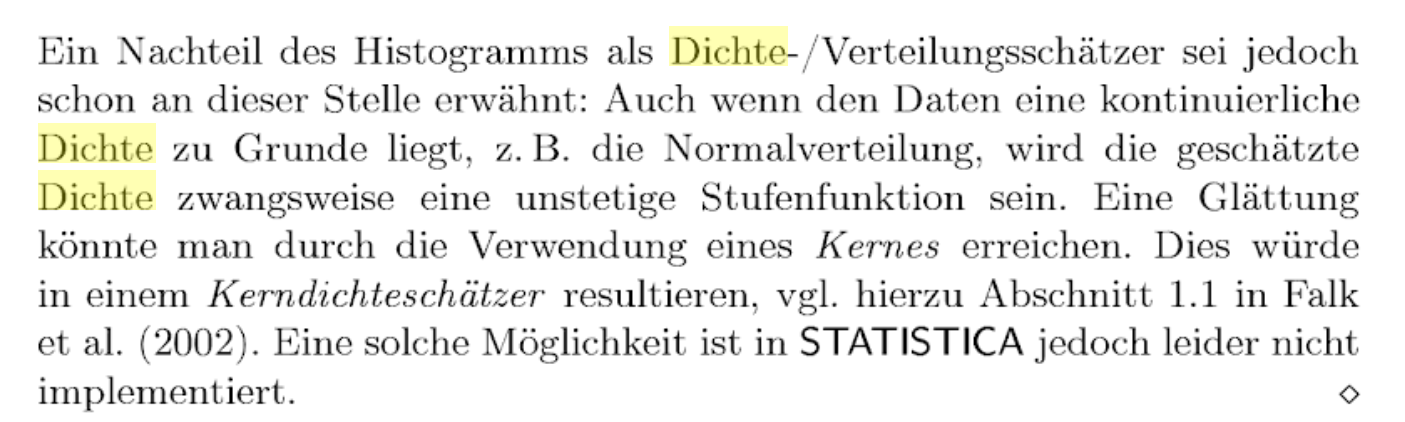
Intervallänge nach Freedman und Diaconis: Diese Regel sollte daher von Ausreißern weniger beeinflusst sein



Zusammenhang Dichtefunktion: (<https://www.mathebibel.de/dichtefunktion>)

Als letztes erstellen wir ein Histogramm mit eingezeichneter Dichtefunktion einer Normalverteilung. Eine solche Graphik wir häufig gezeichnet um zu überprüfen ob Daten mit der Normalverteilung übereinstimmen





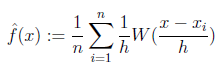
# Dichteschätzung

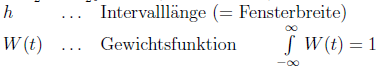
**Fragen:**

- Dichteschätzung & warum gesichert ist, dass bei den gegebenen Eigenschaften der Schätzer auch wirklich eine Dichtefunktion rauskommt (dazu muss man halt die Eigenschaften der Schätzer und die Eigenschaften der Dichtefunktion kennen)

- Eindimensionale Dichteschätzung: Was tut man generell? Welche Eigenschaften hat die Dichte? (Fläche unter der Dichte = 1) Wie wird gewährleistet, dass diese Fläche gleich 1 ist? Kann eine Dichte negativ sein, wenn ja/nein: warum? (Wahl der Gewichtsfunktion: Bei Boxcar und Kosinus-Gewichtsfunktion wird so gewichtet, dass die Gewichte positiv sind.) Eindimensionale Dichteschätzung

* Dichtefunktion kann nicht negativ sein
* Dichte= Ableitung der Verteilungsfunktion
* Bedingung, dass Dichte nicht negativ sein kann, bedeutet dass die Verteilungsfunktion monoton wachsend sein muss
* Fläche unter der Dichtefunktion (ihr Integral) hat den Inhalt = 1
* Summe aller ihrer einzelnen Wahrscheinlichkeiten ergibt 1
* Aus der Dichtefunktion selbst lassen sich keine Wahrscheinlichkeiten ablesen!





* steht h für einen frei wählbaren Parameter der bestimmt wie breit das Intervall ist [x-0.5h, x+0.5h]
* h,die Fensterweite spielt eine wichtige Rolle, wird gelegentlich auch als “smooting factor” bezeichnet

H bestimmt wie viele Objekte Einfluss auf die Dichte eines Punktes x nehmen können.

**Je kleiner h** gewählt wird, umso **weniger Objekte** befinden sich im **Radius h** um x, und umso **weniger Objekte** bestimmen die **Dichte** von X.

Wenn h jedoch **groß gewählt** wird, beeinflussen **viele Objekte** gleichzeitig die **Dichte** des Punktes x.

Nehmen nur **wenige Objekte Einfluss** auf die Dichtebestimmung von x, so **variieren die Dichten** benachbarter Punkte **stärker**, als wenn viele Objekte in die Dichtebestimmung eingehen. Bei **vielen Objekten** nivellieren sich die Einflüsse der Einzelnen Objekte gegenseitig, so dass der **Einfluss** eines **einzelnen Objektes** **nicht zu hoch** wird.

Insofern sind Dichtefunktionen mit **großem h** **runder** und besitzen nicht so viele Modi wie welche mit einem kleinen h.

Infos über h von hier:

<https://www.spektrum.de/lexikon/mathematik/dichteschaetzung/1829>

<https://books.google.at/books?id=_XUdBgAAQBAJ&pg=PA438&lpg=PA438&dq=dichtesch%C3%A4tzung+boxcar+cosinus&source=bl&ots=t4QBxO3tFI&sig=L1mFGLK0HMYrtoaT8YVYlta8iQ8&hl=de&sa=X&ved=0ahUKEwiMmZnMvaHaAhXPKFAKHcuFDJMQ6AEIRjAJ#v=onepage&q=dichtesch%C3%A4tzung%20boxcar%20cosinus&f=false>

* Für die Gewichtsfunktion w(u) kann jede beschränkte, symmetrische und nichtnegative Funktion herangezogen werden, für die gilt

hier einige am häufigsten verwendete Gewichtfunktionen:

### Boxcar Function (Rechteck Gewichtsfunktion)

* Als Ergebnis erhält man eine Treppenfunktion



### 

### Cosinus Gewichtsfunktion



# Wahrscheinlichkeitsnetz und empirische Verteilungsfunktion

**Fragen:**

- worum geht’s

- Ziel

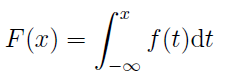
- womit werden die achsen beschriftet

- warum ist eine annähernde gerade ein indiz für normalverteilung?

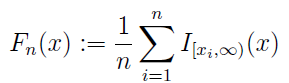
Ziel:

man kann die Daten eines statistischen Merkmals daraufhin untersuchen, ob ihnen eine bestimmte Wahrscheinlichkeitsverteilung zu Grunde liegt.

### **theoretische Verteilungsfunktion**



### **empirische Verteilungsfunktion**

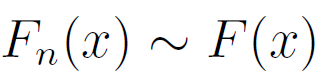


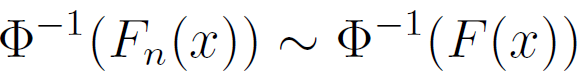
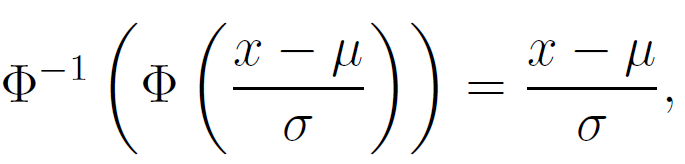
* bei wenig Werten können die Abstände zur theoretischen Verteilungsfunktion groß sein, während bei mehreren Werten die empirische kaum mehr von der theoretischen Verteilungsfunktion unterscheidbar ist.
* die empirische Verteilungsfunktion konvergiert gegen die theoretische Verteilungsfunktion

Beim Plot der empirischen Verteilungsfunktion ist oft schwer nachvollziehbar, ob Daten aus einer speziellen Verteilung (z.B.) Normalverteilung kommen.

Man könnte dazu aber die vertikale Achse verzerren, indem man die Skalierung ändert.

Im Wahrscheinlichkeitsnetz wird die vertikale Achse zwischen 0 und 1 nicht in gleich große Intervalle geteilt sondern es werden die Abstände proportional zu aufgetragen, wobei die Inverse der Verteilungsfunktion der Standard- Normalverteilung ist.

Wenn die Daten dann ungefähr normalverteilt sind so wird  sein und

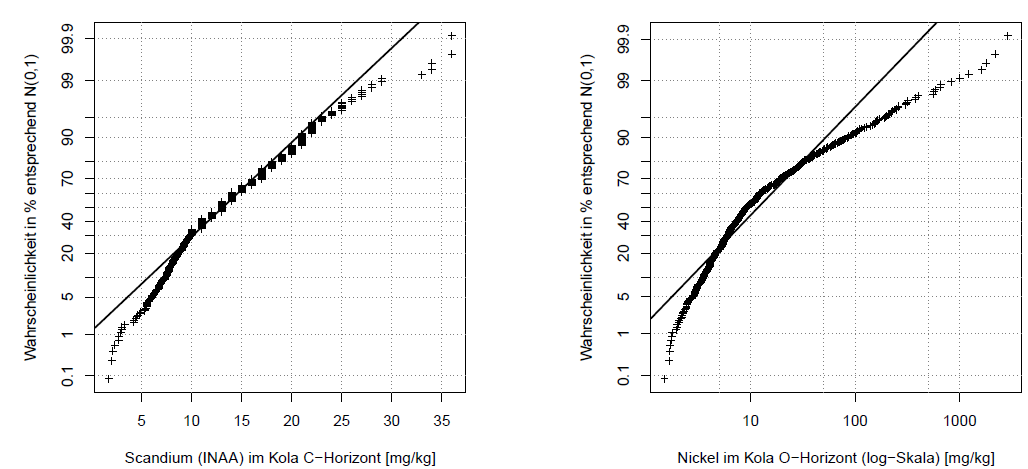
 = 

sodass die Punkte ungefähr auf einer Geraden liegen.

* Man liest dann bei der Wahrscheinlichkeit 50% eine Schätzung für  ab und bei 84% eine Schätzung für 
* Im Falle einer Normalverteilung sollten die Abweichungen von der Geraden “rein zufällig” sein und nicht systematisch
* Je mehr Beobachtungen, desto weniger Abweichungen
* Aufgrund der Skalierung der horizontalen Achse kann man auch die Rückschlüsse auf die originalen Datenwerte machen.

warum?

gerundete Werte besser ersichtlich, weil jeder einzelne Wert eingetragen ist



Achsenbeschriftung: y-Achse = Wahrscheinlichkeiten in % entsprechend N(0,1)

# QQ-Plots

Nachdem die vertikale Achse eigentlich den Quantilen der Standard-Normalverteilung entspricht, könnte man anstelle der Skalierung mit den Wahrscheinlichkeiten auch eine Skalierung mit den Quantilen vornehmen. Die Struktur des Plots würde sich nicht verändern, bloß die Skalierung der vertikalen Achse. Dies wird im folgenden Abschnitt als Q-Q Plot eingeführt.

Mit Quantile-Quantile Plots kann man zwei Verteilungen unmittelbar miteinander vergleichen.

Meist ist eine der Verteilungen eine hypothetische (z.B. Normalverteilung) und die andere eine Verteilung vorhandener Daten.

* Vergleichsbasis sind Quantile der Verteilung

-> wenn Abweichungen von der Geradenform zeigen, sind mehr oder weniger große Unterschiede zwischen den beiden Verteilungsfunktionen

## Ursachen für Unterschiede

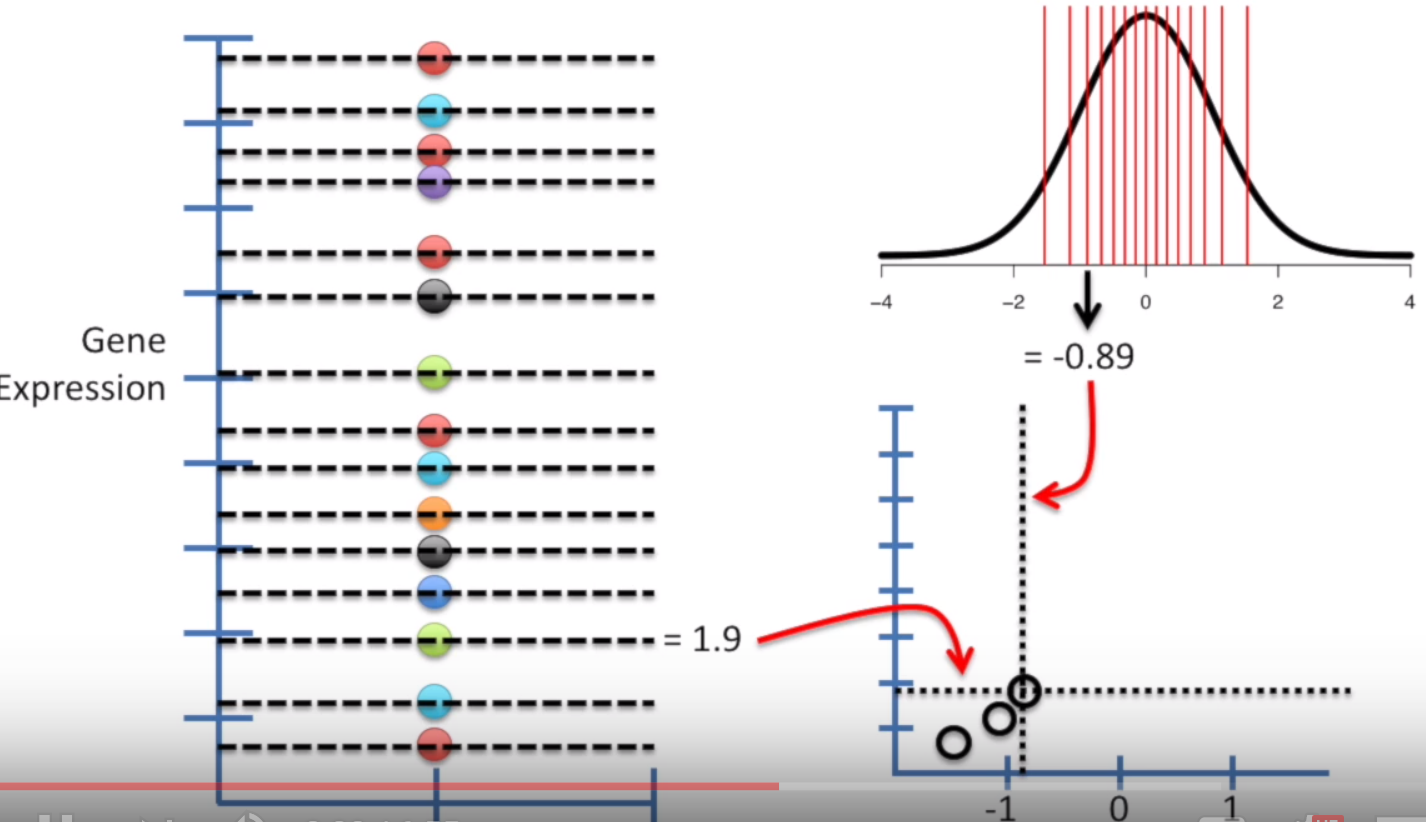
* Ausreißer
* Krümmung an den Enden: empirische Verteilungsfunktion hat mehr Masse in den Schwänzen der Verteilung als theoretische

* Krümmung an den Enden: empirische Verteilungsfunktion hat weniger Masse in den Schwänzen der Verteilung als theoretische
* Kovexe/ konkave Gestalt: falls X(theoretische VF) symmtrisch verteilt : Y (empirische VF) ist linksschief, wenn die unteren Quantilen weiter vom Median entfernt sind als die oberen
* Konvexe oder konkave Gestalt: Falls X symmetrisch verteilt ist: Y ist rechtsschief,wenn die oberen Quantile weiter vom Median entfernt sind als die unteren Quantile
* horizontale Segmente: gerundete Werte oder diskrete Verteilung
* Plateaus: Zusammensetzung von 2 oder mehreren Verteilungen, oder Cluster in Daten

**Anmerkung:**

* QQ-Plots reagieren empfindlich auf (auch zufällige) Abweichungen nahe den Rändern bei Verteilungen, deren Definitionsbereich bis 
* für schöne, lineare Gestalt= größere Anzahl an Datenwerten benötigt

Man kann durch QQ- Plots auch auf Symmetrie prüfen



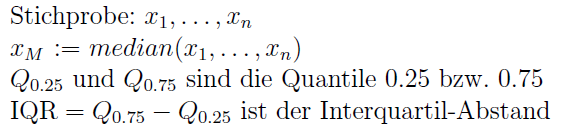
# Boxplot

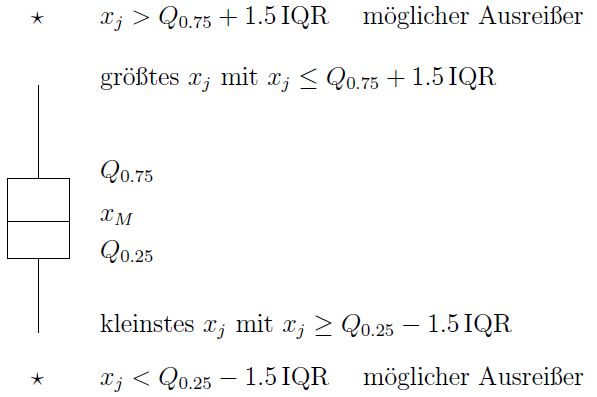
### Zweck

Darstellung wichtiger Kenndaten von Verteilungen eindimensionaler Zufallsgrößen. Vergleich mehrerer Stichproben, von Datengruppen, usw.

Boxplots eignen sich unmittelbar dazu, Ausreißer in einer Messreihe zu erkennen

### Definition



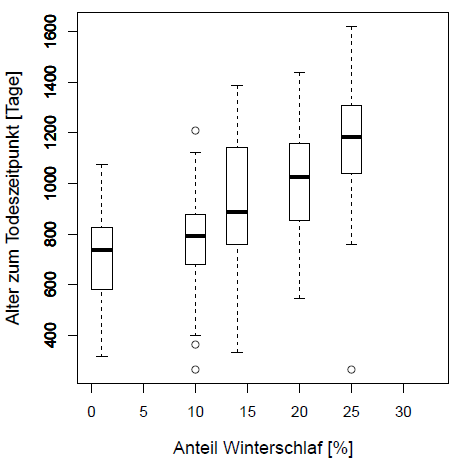
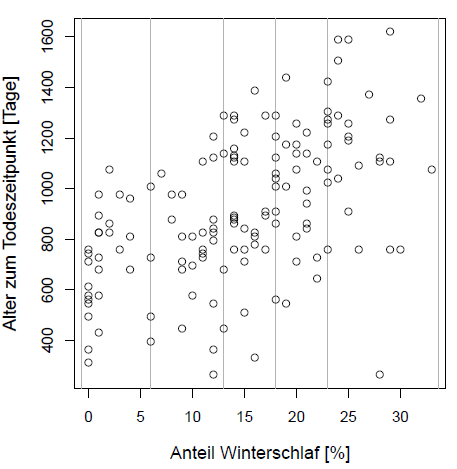


Boxplots sind durch ihre Konstruktion sehr informativ:

* Robustheit durch Verwendung von Median und Quartilen
* Schiefe zu beurteilen nach der Lage des Medians zu den Quartilen und nach dem Auftreten von Ausreißern
* Beurteilung der Wölbung und des Ausreißeranteils
* Boxplots können Kerben (notched) haben: als Konfidenzinertervalle

Streifen Boxplots:

Es ist oft schwierig aus einer 2-dimensionalen Punktwolke Trends abzulesen, als Hilfe unterteilt man die Daten und stellt sie mittels Boxplots dar.

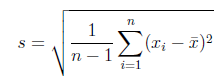


# Robuste univariate Schätzer

Bei **guter Datenqualität** (keine Ausreißer, keine Rundungseffekte) bietet sich das arithmetische Mittel als Schätzer für Lokation an



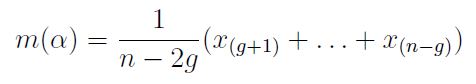
und die empirische Standardabweichung als Schätzer für Streuung:



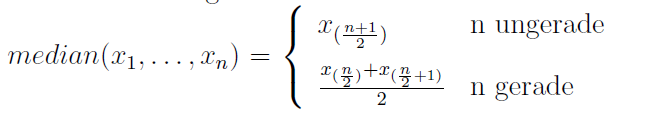
Wenn aber Ausreißer vorkommen ist es besser **robuste Schätzer** zu verwenden

### Gestutze Mittel (alpha-trimmed mean)



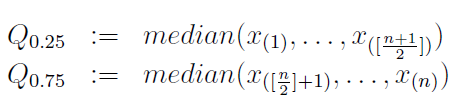


### Median

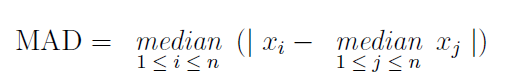


### IQR (Interquartile Range)



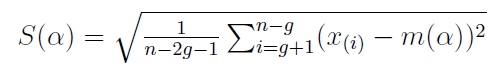


### MAD (Median Absolute Deviation) (medmed) <http://www.statisticshowto.com/median-absolute-deviation/>

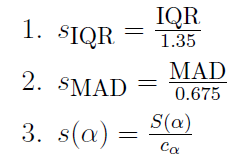


### gestutzte Streuung

definiert als:



Konsistene Schätzer für die Standardabweichung (da IQR, MAD und S(alpha) keine konsistenten Schätzer sind )(Ein Schätzer ist konsistent, wenn er für immer größere Stichproben immer genauer wird.)



Anmerkung: Als Schätzfunktion für die Varianz wird das Quadrat der angeführten Schätzfunktionen genommen.

- MAD: <http://www.statisticshowto.com/median-absolute-deviation/>

- : Boxplot, IQR, Schätzwerte

# Univariate Ausreißererkennung

Fragen:

- Univariate Ausreißererkennung, wie funktioniert das?

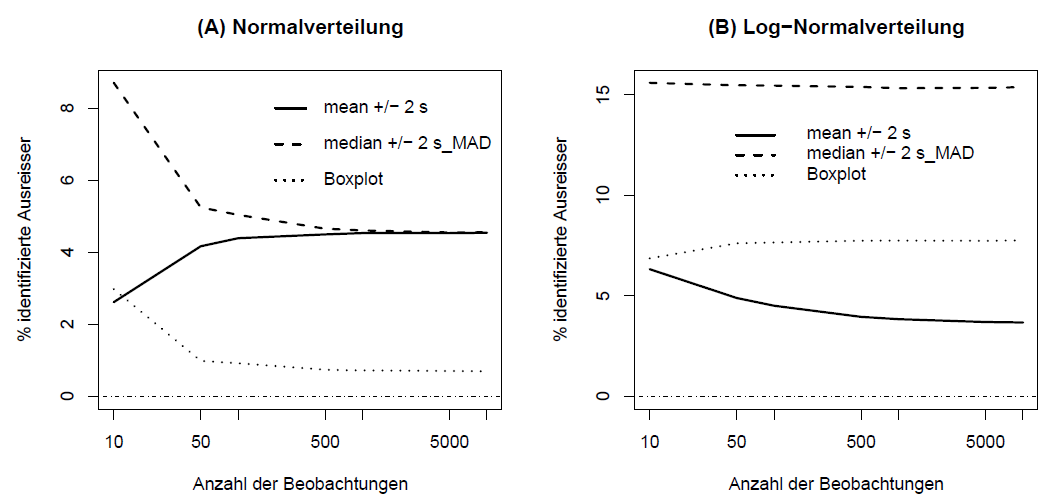
Boxplots eignen sich Ausreißer in einer Messreihe zu erkennen.

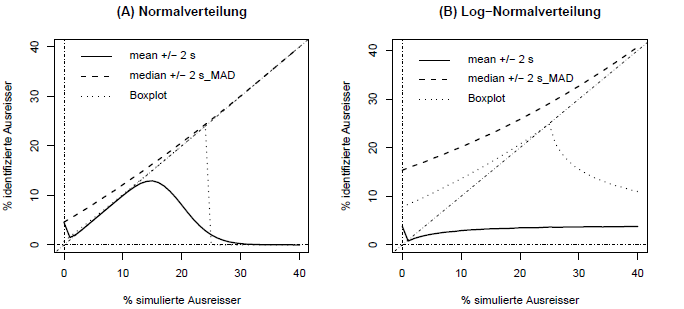
Man kann aber noch andere Regeln zur Identifikation von univariaten Ausreißern

herleiten , die auf robuster Lokations- und Streuungsschätzung basieren.

Bei der Normalverteilung liegen im Bereich Mittel +- 2 mal Standardabweichung die

inneren 95% der Werte.

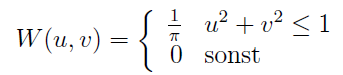


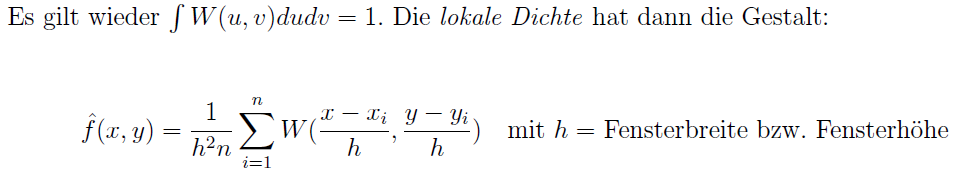


# Zweidimensionale Dichteschätzung

Grundprinzip gleich: Teilbereich auswählen (quadratisch,kreisförmig), in dem eine vorgegebene Gewichtsfunktion evaluiert wird

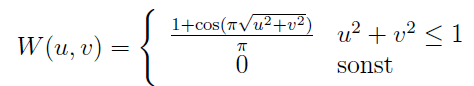
**Boxcar Funktion:**





Glattere Dichten können mit cosinus-Gewichtsfunktion dargestellt werden

## Cosinus-Funktion



Darstellung durch Graustufen, Farbstufen, Isolinien, Gebirge

# Robuste Schätzung linearer Trends & Regression

**Fragen:**

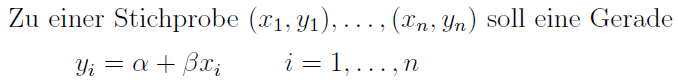
- Kriterium für klassische Regression

- warum robuste Regression

- paar robuste Regressionsmethoden freier Wahl erklären (Formeln der robusten Regression aufschreiben können)

- LOWESS (hier wäre es gut, die beiden Gewichtungsfunktionen zu kennen und außerdem, wie Regression einfließt. Natürlich muss man nicht die ganze Seite aus dem Skriptum 1:1 runterratschen.)

- Warum ist LS-Schätzer nicht robust?



geschätzt werden, die annähernd durch die Punkte der Stichprobe geht.

Die Schätzwerte alpha und beta sollen unempfindlich gegenüber Ausreißern sein.

Residuen:

Übliche Regression ist dafür nicht geeignet:

LS Kriterium: (Kriterium für klassische Regression)



Abweichende Werte: großes Residuenquadrat

Minimum von kriterium wird kleiner, wenn Residuen gleichmäßiger aufgeteilt werden

### Warum ist LS-Schätzer nicht robust?

weil bei LS-Regression die Summe der quadrierten Residuen minimiert wird- nicht so robust wie wenn man den Median statt der Summe nimmt

### **Robuste Gerade nach Tukey**

Daten: Stichprobe ****

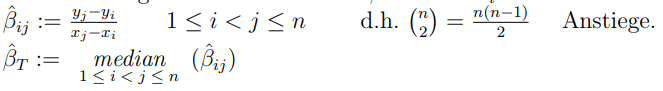
1. Sortieren nach den x-Werten
2. Einteilung in 3 Gruppen
3. Berechnung der Mediane für x-Werte und y-Werte in den einzelnen Gruppen

Schwachstellen des Verfahrens:  
• eventuell sehr langsame Konvergenz bei “schlechter" Konfiguration der Daten.  
• manchmal keine Konvergenz -> kann aber durch geringfügige Modifikation des Algorithmus repariert werden.

### **Robuste Gerade nach Theil**

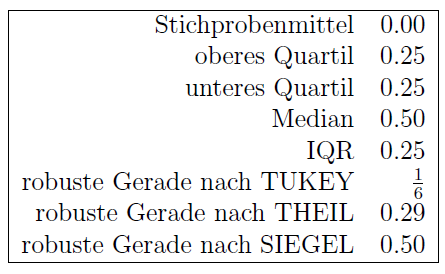
Mediane paarweiser Anstiege werden berücksichtigt

alle x müssen unterschiedlich sein



Maß für Grad der Robustheit:

**Bruchpunkt** = Anteil der Daten, der durch beliebig große bzw. beliebig kleine Werte ersetzt werden kann, ohne dass dadurch der Schätzer unsinnige Resultate liefert.



### **Robuste Gerade nach Siegel**

basiert ebenfalls auf Medianen von paarweisen Anstiegen, diese werden aber wiederholt berechnet:

### **LMS (Least Median of Squares Regression)**

**Fragen:**

- Formel aufschreiben. Warum nimmt man dem Median und nicht die Summe im Vergleich zu LS-Regession?

Bei **LS-Regression** wird die Summe der quadrierten Residuen minimiert, bei **LMS** ersetzt man die Summe durch den Median um starke Robustheit zu erzielen. (Bruchpunkt 50%)

LMS:



### **LTS (Least Trimmed Sum of Squares)**

**Fragen:**

- Formel aufschreiben. Welche Werte/Punkte werden weggelassen und warum? (Große Residuenquadrate haben größeren Einfluss)

Bei LTS wird die Summe der kleinsten quadrierten Residuen minimiert.

Seien die der größe nach geordneten quadrierten Residuen, dann lautet LTS:

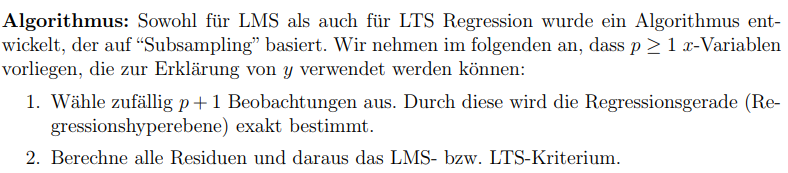


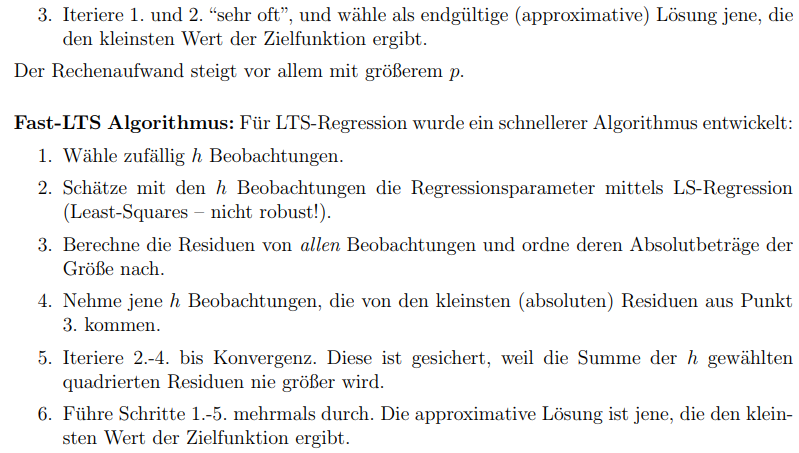
mit 

je nach Wahl von **h** erhält man eine Methode mit Bruchpunkt zwischen 0 und 50%

Mit geringerem Bruchpunkt leidet zwar die Robustheit, aber die Effizienz (Präzision des Schätzers) wird erhöht.

Algorithmus:





# Glättung und Schätzung nichlinearer Trends

## Nichtlineare Glätter für äquidistante (Zeit-)Punkte

Ziel:Glättung der Zeitreihe, die bis zu einem gewissen Grad Robust ist gegenüber Störungen im Signal.

Mit der Glättung sollen Muster in der Zeitreihe erkannt werden.

Zeitreihe= glatte Kurve + Residuen

## LOWESS

# LOWESS(LOcally WEighted regression Scatter plot Smoothing)

**Fragen:**

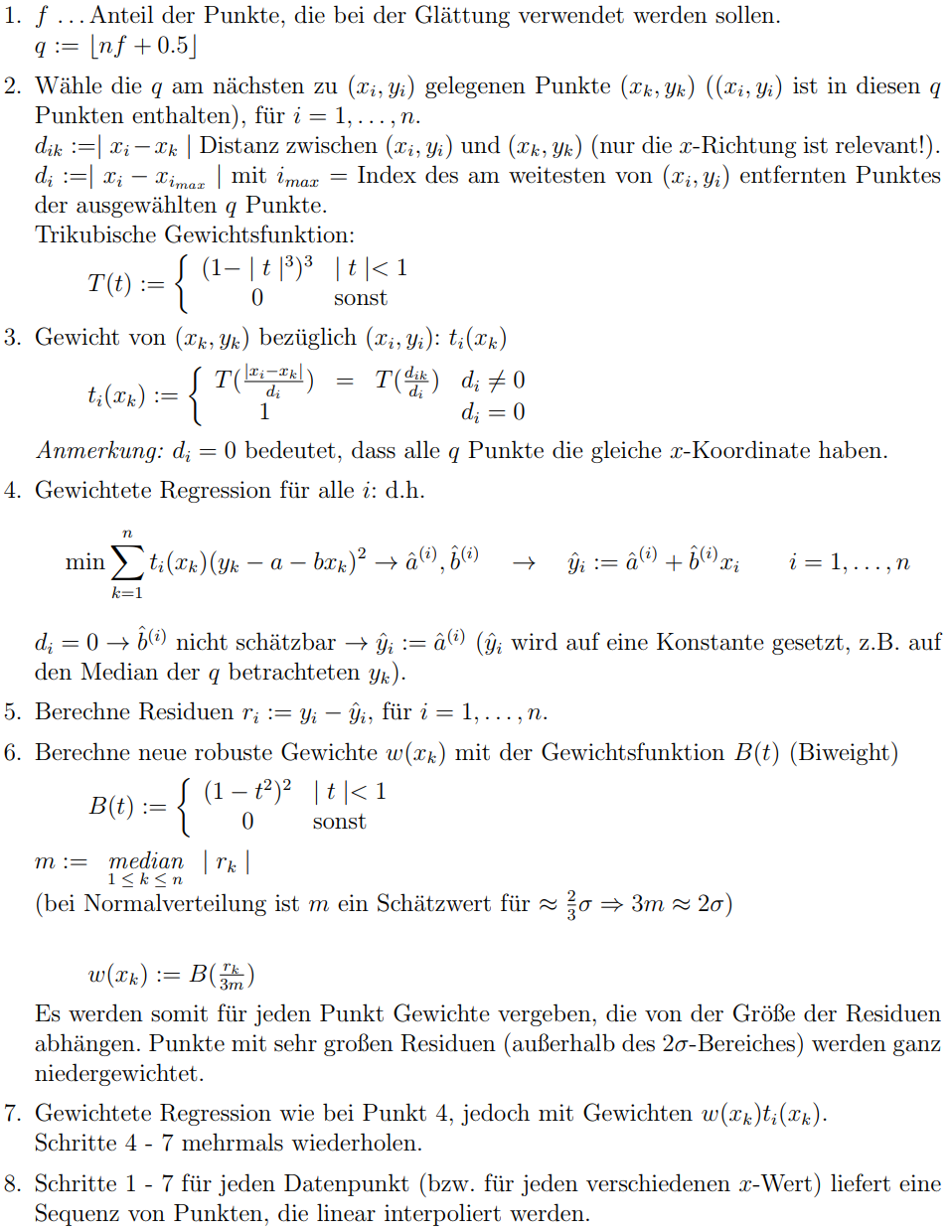
- hier wäre es gut, die beiden Gewichtungsfunktionen zu kennen und außerdem, wie Regression einfließt. Natürlich muss man nicht die ganze Seite aus dem Skriptum 1:1 runterratschen.)

- Algorithmus - Ungefähr aufschreiben was jeder Schritt macht, detailliert nachgefragt wie Distanzen berechnet werden, wie gewichtet wird, trikubische Fkt aufschreiben (Was ist hier klein t?)

LOcally WEighted regression Scatter plot Smoothing ist ein auf Scatterplots anwendbares Glättungsverfahren, das Trends in den Daten erkennen lässt.

Es arbeitet wie ein nichtlineares Regressionsverfahren: x-Daten vorgegeben und y-Daten geglättet.

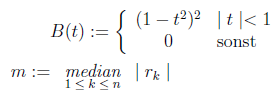
Algorithmus:



**Trikubische Gewichtsfunktion:**



**Gewichtsfunktion Biweight:**



Locally Weighted regression Scatter plot Smoothing

= auf Scatterplots anwendbares Verfahren das Trends erkennen lässt

arbeitet wie ein nichtlineares Regressionsverfahren= x-Daten vorgegeben, und entsprechenden y-Daten geglättet

<https://www.youtube.com/watch?v=Vf7oJ6z2LCc>

### Upper and Lower Smoothing

= bietet zusätzlich Streuungsinformation

### Pairs of Middle Smoothing

Nachdem es nicht egal ist ob LOWESS auf Daten (xi,yi) oder auf (yi,xi) angewendet wird, können beide Varianten versucht werden.

Scatterplot Matrix/ Draftman’s Display

= es werden immer Paare gegenübergestellt

# Zeitreihen

**Fragen:**

- MA, AR, ARMA, ARIMA-Modelle, Autokovarianz, Autokorrelation (hier war er besonders darauf bedacht, dass ich die richtigen Formeln produziere).

- Was bedeuten die u bei MA? (Restkomponenten) Was wird bei ARIMA ersetzt? (xt durch Differenzen)

- Zeitreihenmodelle aufschreiben und erklären

- Was bedeuten die u bei MA? (Restkomponenten) Was wird bei ARIMA ersetzt? (xt durch Differenzen)

Trend einer Zeitreihe= Zeitreihe steigt oder fällt über einen gewissen Zeitraum

## Exponentielles Glätten

(keine konkreten Fragen im Vowi)

<https://www.youtube.com/watch?v=Fqge2HDH2Co>

Simple Exponential Smoothing

<https://www.youtube.com/watch?v=Fqge2HDH2Co>

Holt’s exponential Smoothing

<https://www.youtube.com/watch?v=DUyZl-abnNM>

Die Prognose von zukünftigen Werten könnte so gemacht werden, dass man die beobachteten Werte gewichtet nach ihrer Aktualität berücksichtigt

xt… geglätteten Werte

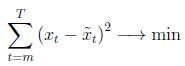
alpha.... Glättungsfaktor 0< alpha<1



Je kleiner genommen wird, desto weniger werden die aktuellsten Werte berücksichtigt, und daraus folgt, dass die Sequenz der glatter wird.

* Alpha liegt zwischen 0 und 1

alpha kann praktisch so bestimmt werden, dass für die vorhandenen Daten die Summe der quadrierten Residuen minimiert wird:



Gewichte für vergangene Werte fallen exponentiell ab

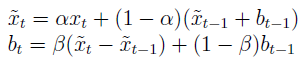
Prognose aller künftigen Werte Hängt hier nicht vom Horizont h ab.

* beschränkt auf Zeitreihen ohne Trend und Saisoneinflüssen

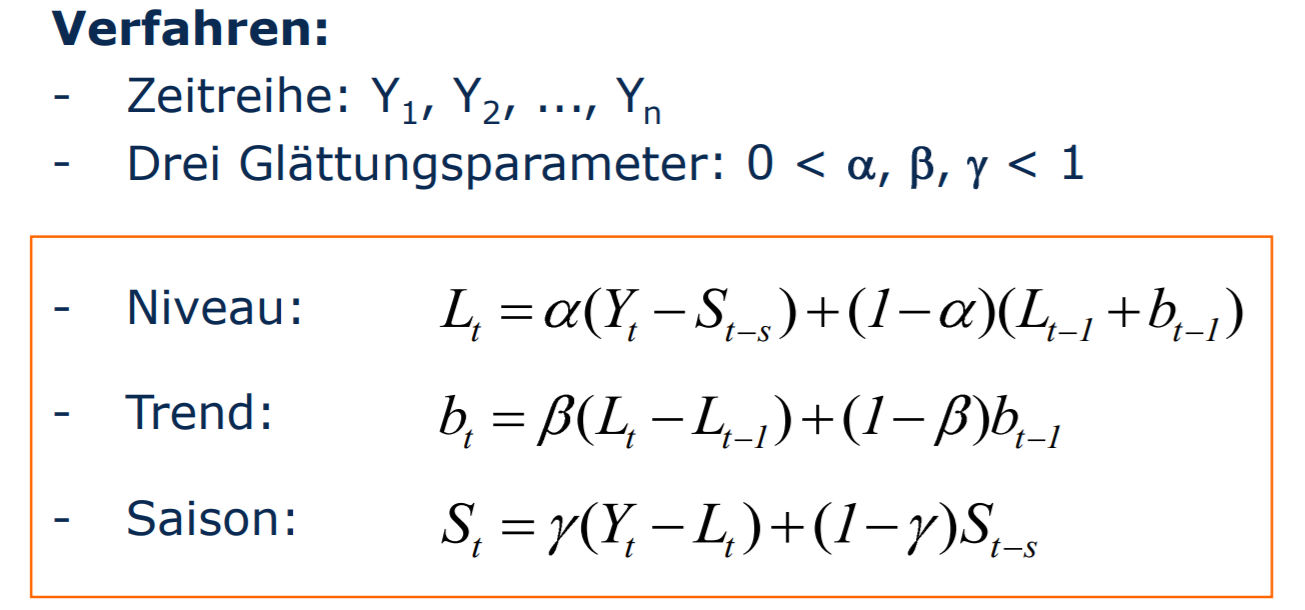
**Glättung nach Holt-Winters**

Trendvariable wird berücksichtigt

an jedem Zeitpunkt t erfasst bt den lokalen Anstieg der Zeitreihe



### Verfahren ermöglicht es uns, Prognosen für Zeitreihen mit Trend- und Saisonkomponente zu machen.



## Modellierung von Zeitreihen

### Moving Average Modell (MA)

stationäre Zeitreihe



a und 0…. sind die unbekannten Parameter

ut… white noise

(Erklärung white noise= stationäre Zeitreihen sind dadurch charakterisiert,

dass sie gleichen Erwartungswert (Mittel) und gleiche Varianz fur alle t haben, und dass ihre Autokovarianz gleich für alle t und jedes k > 0 ist. Sie werden als white

noise (weies Rauschen) bezeichnet. Man kann auf die Eigenschaft white noise testen mit der Q-Statistik, auch Ljung-Box Statistik genannt)

Wenn im Korrelogramm die Korrelationen stark abfallen, und nach lag q nicht mehr signikant sind, so ist das ein Indiz für einen MA(q) Prozess

### Autoregressives Modell (AR)



 und a … unbekannte Parameter

Die Autokorrelationen gehen langsam gegen 0, manchmal weisen sie eine sinus-förmige Struktur auf.

### ARMA (Autoregressive-Moving-Average) Model

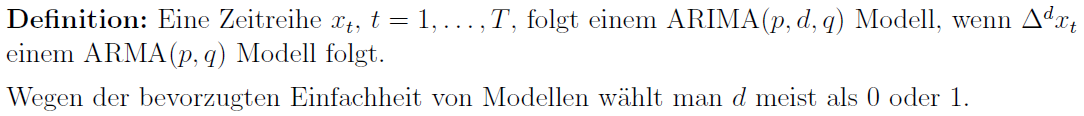
Wenn weder im Korrelogramm noch im partiellen Korrelogramm eine “Implosion” auf 0 ab einem bestimmten lag sichtbar wird, dann kann eine Mischung aus AR und MA sinnvoll sein. Diese bezeichnet man als ARMA.

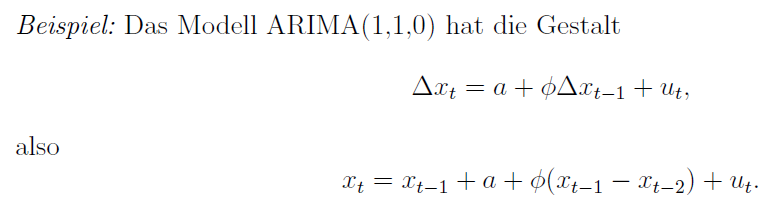


Einfachen Modellen sollte der Vorzug gegeben werden, also p und q sollten klein gewählt werden.

### ARIMA Modell (AutoRegressive Integrated Moving Average)

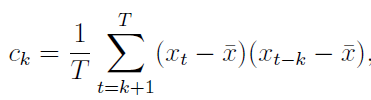
* setzt eine stationäre Zeitreihe voraus
* liegt ein Trend vor, so spricht man von nicht-stationären oder integrierten, also trendbehafteten Zeitreihe (saisonal inkludiert)
* Trends können durch Bildung von Differenzen eliminiert werden
* Differenz-Operator 





### Autokovarianz

Abhängigkeiten werden mit der Autokovarianz analysiert.



### Autokorrelation



# Parameterschätzung im Mehrdimensionalen (multivariate Schätzer)

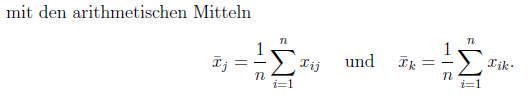
### Kovarianz



E… Erwartungswert (meist geschätzt durch arithmetisches Mittel)

klassische Schätzung für Kovarianz wird als Stichprobenkovarianz bezeichnet und ist definiert als :





wenn man j=k setzt erhält man die Stichprobenvarianz

Korrelation zwischen Zufallsgrößen xj und xk ist definiert als:



Nimmt Größe im Intervall [-1,1] an.

* bei einem Wert von 1 bzw -1 enthalten xi und xk die gleichen Information
* bei -1 besteht ein umgekehrter Zusammenhang
* bei 0 besteht kein Zusammenhang

klassische Schätzung der Korrelation wird als Stichprobenkorrelation bezeichnet und ist definiert durch:



misst die lineare Beziehung zwischen xi und xk

wenn rjk = 0 besteht kein linearer Zusammenhang

Robuste Schätzung von Kovarianz und Korrelation

Man gibt den Beobachtungen, die vom Hauptteil abweichen, weniger Gewicht, und diese fließen dann weniger ein

Spearman Rang-Korrelation

**MCD (Minimum Covariance Determinant)**

= Determinante der Kovarianzmatrix wird minimiert

* es werden nicht alle sondern nur h<n Beobachtungen genommen von denen die Stichproben- Kovarianzmatrix berechnet wird
* Als Nebenprodukt erhält man eine robuste Lokationsschätzung mit dem arithmetischen Mittel der h Beobachtungen

Distanz und Ähnlichkeit

Euklidische Distanz

Manhattan Distanz

Minkowski Distanz

Mahalanobis

=wichtig weil sie Kovarianzstruktur mitberücksichtigt

# Multivariate Ausreißererkennung

**Fragen:**

- Formel der Mahalanobis-Distanz aufschreiben können, was ist der "Vorteil" gegenüber Euklidischer-Distanz, wie werden "Ausreißer" erkannt

### Mahalanobis-Distanz

besonders wichtiges Distanzmaß, weil sie Kovarianzstruktur mitberücksichtigt

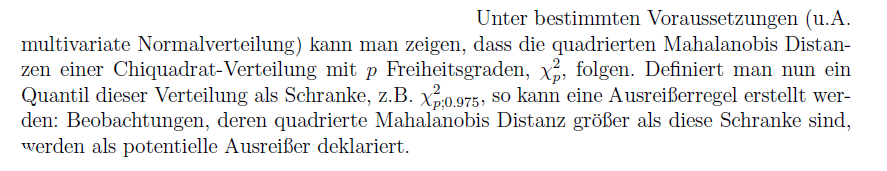


.... Kovarianzmatrix

... Zentrum

Sie zeigt an, wie weit die Beobachtung von diesem Zentrum entfernt ist, relativ zur zugrunde liegenden Kovarianzstruktur.

Beobachtungen mit großer Mahalanobis Distanz = Ausreißer



Man muss bei der Formel sowohl Zentrum als auch Kovarianzmatrix aus den Daten schätzen.

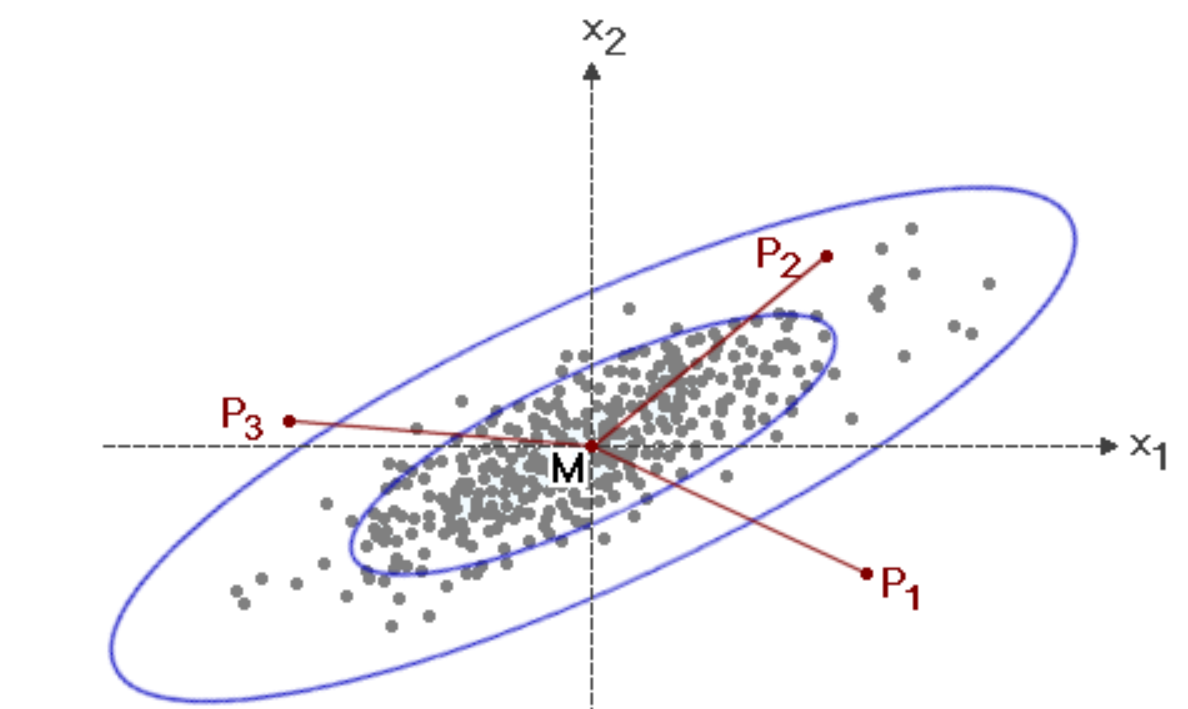
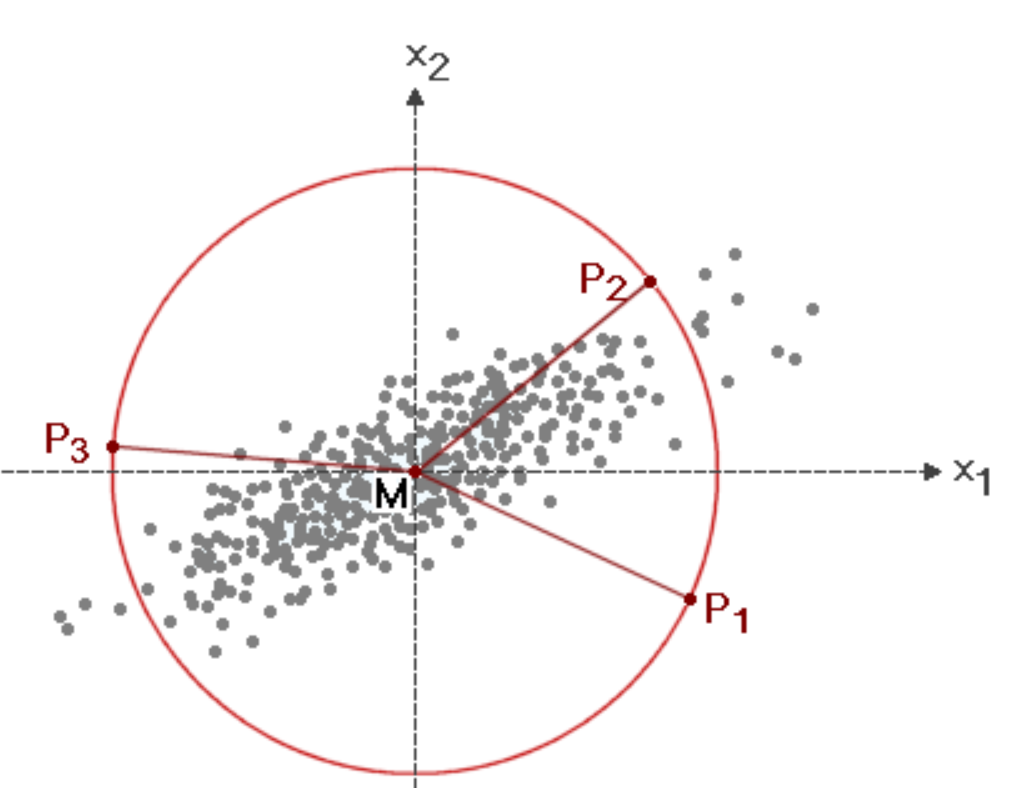
Man benötigt robuste Schätzer für Lokation und Kovarianz = MCD Schätzer.

Mahalanobis Distanz zeigt an wie weit die Beobachtung von diesem Zentrum entfernt ist, relativ zur zugrunde liegenden Kovarianzstruktur.

Euklidische Distanz kann nicht beschreiben wie weit man an den Rand der Punktewolke kommt.

Betrachtet man die Distanz in mehrdimensionalen Räumen, so stellt man fest, dass die "klassische" euklidische Distanz irreführend sein kann.

<http://www.statistics4u.com/fundstat_germ/ee_mahalanobis_distance.html>



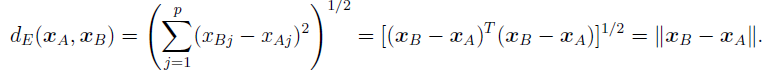
**vereinfacht:** zieht man einen Kreis um die Datenwolke und misst die Dichte der Daten dann sieht man dass diese nicht gleich sind, obwohl die Punkte alle denselben Abstand vom Mittelpunkt haben. Punkt P1 liegt viel weiter weg.

Wenn man nun aber eine Ellipse zeichnet, haben die Punkte auf der Ellipse alle denselben Abstand vom Mittelpunkt. Die Ellipsen konstanter Wahrscheinlichkeit entsprechen einer konstanten Mahalanobis-Distanz.

Anmerkung

Ist die Kovarianzmatrix **C** gleich der Einheitsmatrix, dann sind die Daten nicht korreliert und weisx en gleiche Standardabweichungen auf. Für diesen Fall wird die Mahalanobis-Distanz gleich der euklidischen Distanz.

Euklidische Distanz:& 



# Projektionen mehrdimensionaler Daten

Linearkombinationen von Variablen

## Hauptkomponentenanalyse (PCA)

**Fragen:**

- Formel

- Linearkombination, Definition der Hauptkomponenten, Anzahl der relevanten Hauptkomponenten (das mit dem "Scree Plot" erklären)

- Ziel der Hauptkomponentenanalyse und warum macht man das?

- *(Wenn man nur U = XB hinschreibt, fragt er nach, was das ist. Man muss dann erklären, dass es sich um Matrizen handelt, welche Dimension sie haben und was sie darstellen. Was sind die Spalten der Matrizen? Was davon wird maximiert? Was bedeutet b^T \* Sigma \* b im Lagrange-Problem? Was sind die Lambdas?*

*- Wie berechnet man die Varianz für die Koeffizienten in der Ladungsmatrix B?*

*Antwort: Mit dem Lagrange Problem: *

*und davon ist das die Varianz= *

*Woran sieht man dass die Varianz maximiert wurde?*

**

= zählt zu den wichtigsten Methoden der multivariaten Statistik.

Hauptkomponenten werden häufig als exploratives Werkzeug eingesetzt, das einen Überblick über die Daten liefert

## Formel

Hauptkomponenten werden über Linearkombinationen so definiert:

U= XB

X… n x p Datenmatrix

U… Score Matrix der Dimension n x p (Einzelnen Spalten= Hauptkomponenten, u1,.....up)

B… Ladungsmatrix, p x p Matrix (enthält Koeffizienten die p neue Richtungen definieren), (Koeffizienten b1,...bp zur bestimmung der Hauptkomponenten werden so gewählt, dass Varianz von u maximal ist, außerdem wird gefordert:  fürs 1., fürs zweite dann , usw.)

Geometrisch: man erhält eine Darstellung der Daten im Koordinatensystem, das sogar orthogonal ist

U enthält exakt dieselben Informationen wie X, nur in einer anderen Darstellung

Mit steigender Nummer der Hauptkomponenten nimmt der Informationsgehalt ab.

Dadurch Dimensionsreduktion

* Der Informationsgehalt der höheren Hauptkomponenten nimmt oft rapide ab, so dass sie ohne Verlust an Information weggelassen können.
* Lagrange

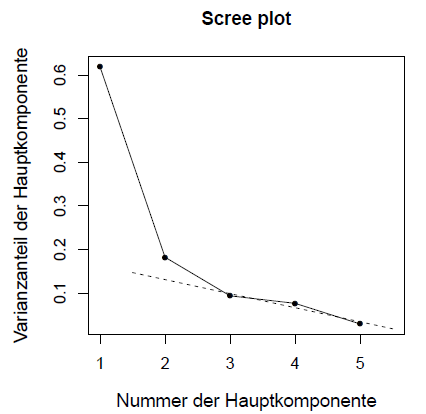
Lambda: Eigenwerte sind somit die Varianzen der Hauptkomponenten, die maximiert wurden.

Maximierung unter Nebenbedingungen: Lagrange Problem



## Ziel der Hauptkomponentenanalyse

* Methode zur linearen Transformation der Variablen, so dass:  
  möglichst wenige neue Variablen die relevante Information  
   beschreiben. (Mass für die Relevanz ist die Varianz)
* die neuen Variablen orthogonal und damit unkorreliert sind
* Ziel der Hauptkomponentenanalyse ist die Dimensionsreduktion bei möglichst wenig Informationsverlust
* Hauptkomponenten nach absteigender Va
* sortiert
* Der resultierende scree plot (“Geröll-Plot") gibt dann einen Hinweis auf das “optimale" k
* Man möchte k dort festlegen, wo der Linienzug “ausflacht”.
* Alle Punkte, die in etwa auf einer Geraden liegen, entsprechen Hauptkomponenten mit irrelevantem Informationsgehalt, können weggelassen werden
* Als Faustregel könnte man k so wählen, dass zumindest 80% der Gesamtvarianz erklärt sind



* Es handelt sich um eine Rotation des Koordinatensystems (Multiplikation mit einer orthogonalen Matrix).

# Weitere multivariate statistische Methoden

## Clusteranalyse

Ziel: Beobachtungen in homogene Gruppen unterteilen

dabei sollen Beobachtungen mit großer Ähnlichkeit in der gleichen Gruppe sein

* Ähnlichkeiten werden über Distanzen bestimmt
* Distanzen hängen von Skalierung ab, deshalb muss man die Variablen zuerst auf Mittel 0 und Varianz 1 standardisieren
* wirkliche Anzahl von Clustern unbekannt, muss später geschätzt werden
* meist auch keine klar getrennten Gruppen

Methoden:

* Partitionierungsmethode
* Hierarchische Clustermethoden
* Fuzzy Clustering
* Modellbasiertes Clustering

## Diskiminanzanalyse

(Keine konkreten Fragen im Vowi)

* multivariate Methode die als Ziel hat Daten zu gruppieren

Unterschied zu Clusteranalyse:

Bei Clusteranalyse -> nicht bekannt welche Beobachtungen zu welchen Gruppen gehören, nicht einmal klar wie viele Gruppen existieren und ob überhaupt eine Gruppenstruktur vorliegt.

Bei Diskriminanzanalyse hingegen kennt man die Klassenzugehörigkeiten der Beobachtungen, zumindest jener Beobachtungen von einem Trainingsdatensatz.

Nun möchte man eine Regel lernen, die es erlaubt, aufgrund der gleichen gemessenen Merkmale (Variablen) neue Beobachtungen den Klassen zuzuordnen. Diese neuen Beobachtungen bilden des sogenannten Testdatensatz.

## Lineare Diskriminanzanalyse (LDA)

Kovarianzen aller Gruppen werden als gleich angenommen

lineare Diskriminanzregel:



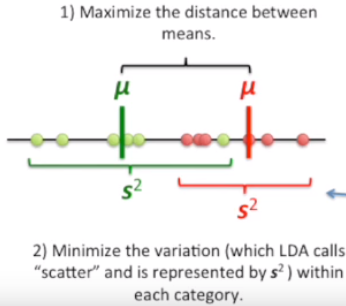
<https://www.youtube.com/watch?v=azXCzI57Yfc>

Möchte wie PCA (Hauptkomponentenanalyse) Dimensionen reduzieren

LDA konzentriert sich aber auf eine maximale Seperation, während PCA sich auf die größte Varianz konzentriert

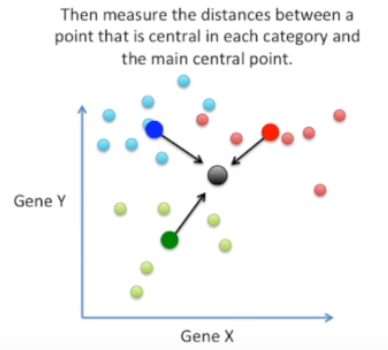
* Annahme dass alle Klassen dieselbe Varianz haben
* Um Achsen zu zeichnen, die die Daten separieren verwenden man 2 Kriterien:

1. Die Distanz zwischen den Mitteln soll maximiert
2. Varianz (scatter) soll minimiert werden in jeder Kategorie



**wenn man mehrere Kategorien hat:**





## Quadratische Diskriminanzanalyse (QDA)



* jede Klasse hat ihre eigenen Kovarianzen

Anmerkungen:

Tukey:

Aber wie weit muss ein Wert von der Masse der Werte abweichen, dass wir ihn als Ausreißer identifizieren? Mit Hilfe des IQR hat Tukey ein Kriterium für Ausreißer angegeben: ein Wert, der um mehr als das 1,5fache des IQR (ÑStep = 1.5 times H-spreadì)von den entsprechenden Quartilen abweicht

Warum nimmt Tukey gerade die 1,5fache IQR als Abweichungsmaß? Auf diese Frage soll Tukey geantwortet haben: “because 1 is too small and 2 is too large”