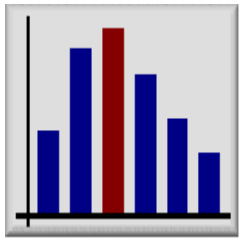
Verteilungsanalyse

Johannes Hain

Lehrstuhl für Mathematik VIII – Statistik



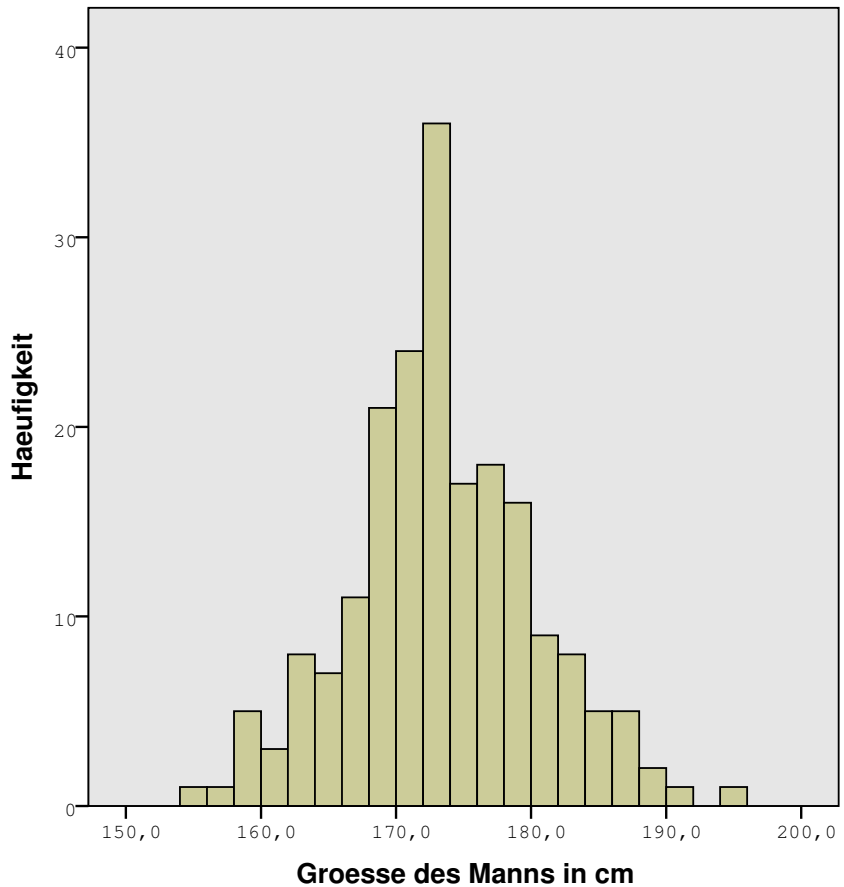
Datentypen

Als Sammeln von **Daten** bezeichnet man in der Statistik das Aufzeichnen von Fakten. Erhobene Daten klassifziert man in unterschiedliche **Skalenniveaus**:

- **kategoriale (= nominal skalierte) Daten:** Gröbstes Skalenniveau; klassifiziert Daten nur in verschiedene Kategorien ohne Ordnung.
  Beispiele: Farben, Städte, Automarken
- **Metrische Daten** sind Messungen, die durch Zahlen sinnvoll interpretiert werden können. Man unterscheidet hierbei noch die beiden folgenden Skalenniveaus:
  - **ordinalskalierte Daten:** Daten liegt interne Ordnung zugrunde, sodass Bildung einer Reihenfolge möglich ist.
    Beispiele: Schulnoten, Schulabschlüsse
  - **intervallskalierte Daten:** Daten besitzen lückenlosen Wertebereich, Abstände zwischen den einzelnen Daten sind von Bedeutung und interpretierbar.
    Beispiele: Körpergröße, Temperatur

Verteilungsanalyse metrischer Daten

Die Verteilung von **kategorialen** Daten veranschaulicht man sich z.B. mit Hilfe von Balkendiagrammen. Dies ist bei **metrischen** Daten wegen des stetigen Wertebereichs (meist) nicht möglich. Die Verteilung wird in diesem Fall mit einem **Histogramm** dargestellt:



Histogramme

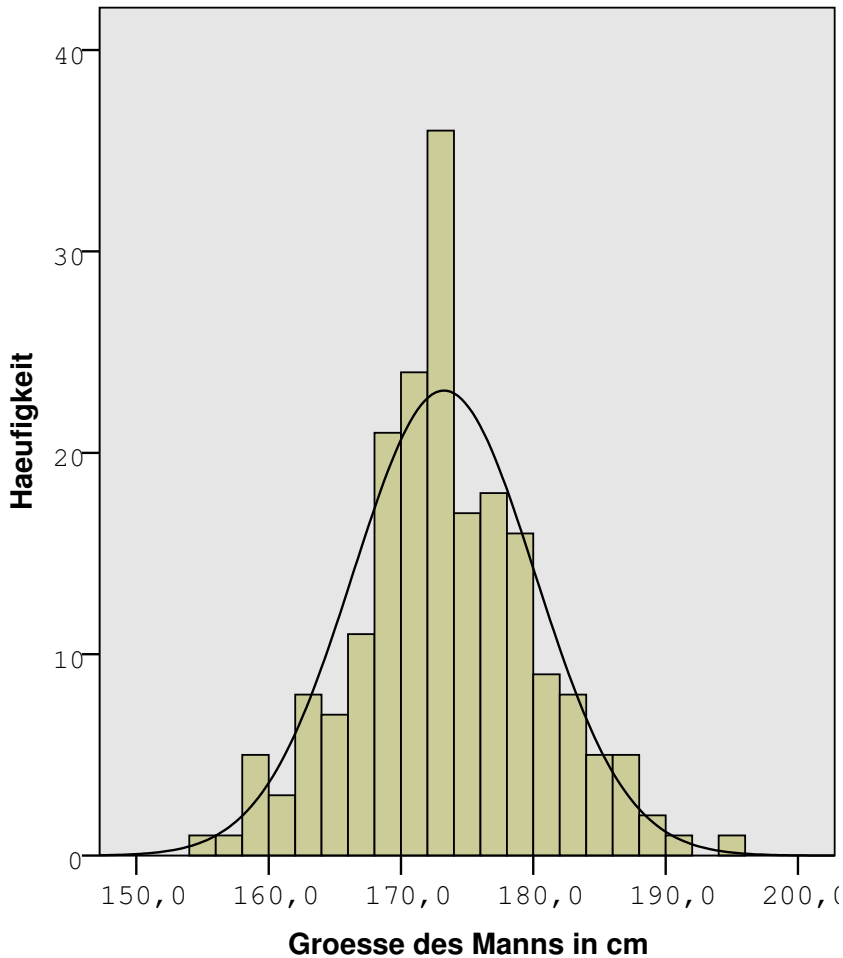
Erstellung eines Histogramms in SPSS

- → *Analysieren*
- → *Deskriptive Statistiken*
- → *Häufigkeiten*
- → Wähle die zu untersuchende Variable aus und gehe auf das Feld *Diagramme*
- → Wähle als *Diagrammtyp* das Feld *Histogramme* aus

Als Alternative kann man sich Histogramme auch ausgeben lassen unter:

Diagramme → *Veraltete Diaglogfelder* → *Histogramme*

Versucht man nun eine Kurve durch das Histogramm zu legen, die die Lage der Balken möglichst gut approximiert, ergibt sich folgendes Bild:



Histogramme

Histogramme mit Normalverteilungskurve in SPSS

→ Gehe vor wie bei der Erstellung eines Histogramm, beschreiben auf Folie 4

→ Wähle zusätzlich noch das Feld *Mit Normalverteilungskurve* aus

Als Alternative geht dies auch unter:

Diagramme → *Veraltete Diaglogfelder* → *Histogramme*

Im daraufhin erscheinenden Dialogfeld setzt man im Feld *Normalverteilungskurve anzeigen* ein Häckchen.

Die eingezeichnete Approximationskurve ist die sogenannte **Dichte der Normalverteilung**. Wir verallgemeinern

Definition: Dichte

Die Dichte einer Verteilung f_X ist eine Funktion, mit der sich die Wahrscheinlichkeit berechnen lässt, dass eine Zufallsvariable vom stetigen Typ in ein gewisses Intervall fällt.

Übersetzung ins Mathematische:

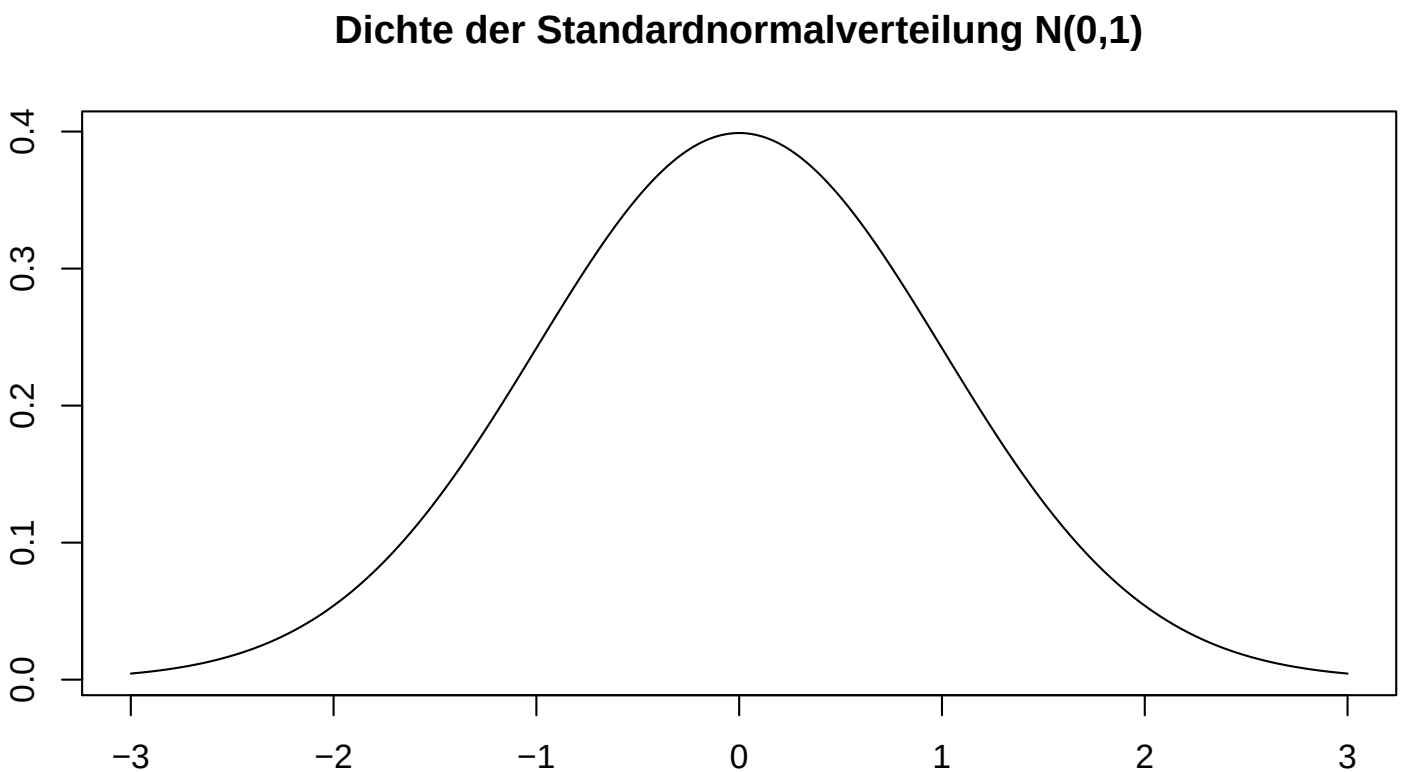
Eine Funktion f_X heißt Dichte einer Zufallsvariable X, falls gilt

$$P(a < X < b) = \int_a^b f_X(t)\, dt.$$

Die Dichtefunktion der Normalverteilung lautet:

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Beispiel: Für $\mu = 0$ und $\sigma^2 = 1$ ergibt sich die **Standardnormalverteilung**, $N(0, 1)$:



Es existieren in der Statistik aber noch viele andere Wahrscheinlichkeitsverteilungen, z.B.

- die **Poissonverteilung:** $f_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!}$

$\rightarrow$ Anzahl der Selbstmorde pro Tag, Anzahl der Störfälle in einem Kernkraftwerk, usw.

- die **Exponentialverteilung:** $f_\lambda(x) = \lambda e^{-\lambda x}$

$\rightarrow$ Zeit zwischen zwei Meteoriteneinschlägen, Lebensdauer von elektronischen Bauelementen, usw.

- die **Lognormalverteilung:** $f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right)$

$\rightarrow$ Aktienkurse, Brutto-/Nettoeinkommen einer Bevölkerung, usw.

Kenngrößen der Normalverteilung

Die Normalverteilung wird charakterisiert durch zwei wichtige Kenngrößen: den **Erwartungswert** und die **Varianz**.

Interpretation des Erwartungswertes

Der Erwartungswert einer Zufallsvariablen, $E(X)$, beschreibt denjenigen Wert, den man bei sehr häufiger Wiederholung von X im Mittel beobachten wird. (Dies bezeichnet man auch als das **Gesetz der großen Zahlen**.)

Definition der Varianz

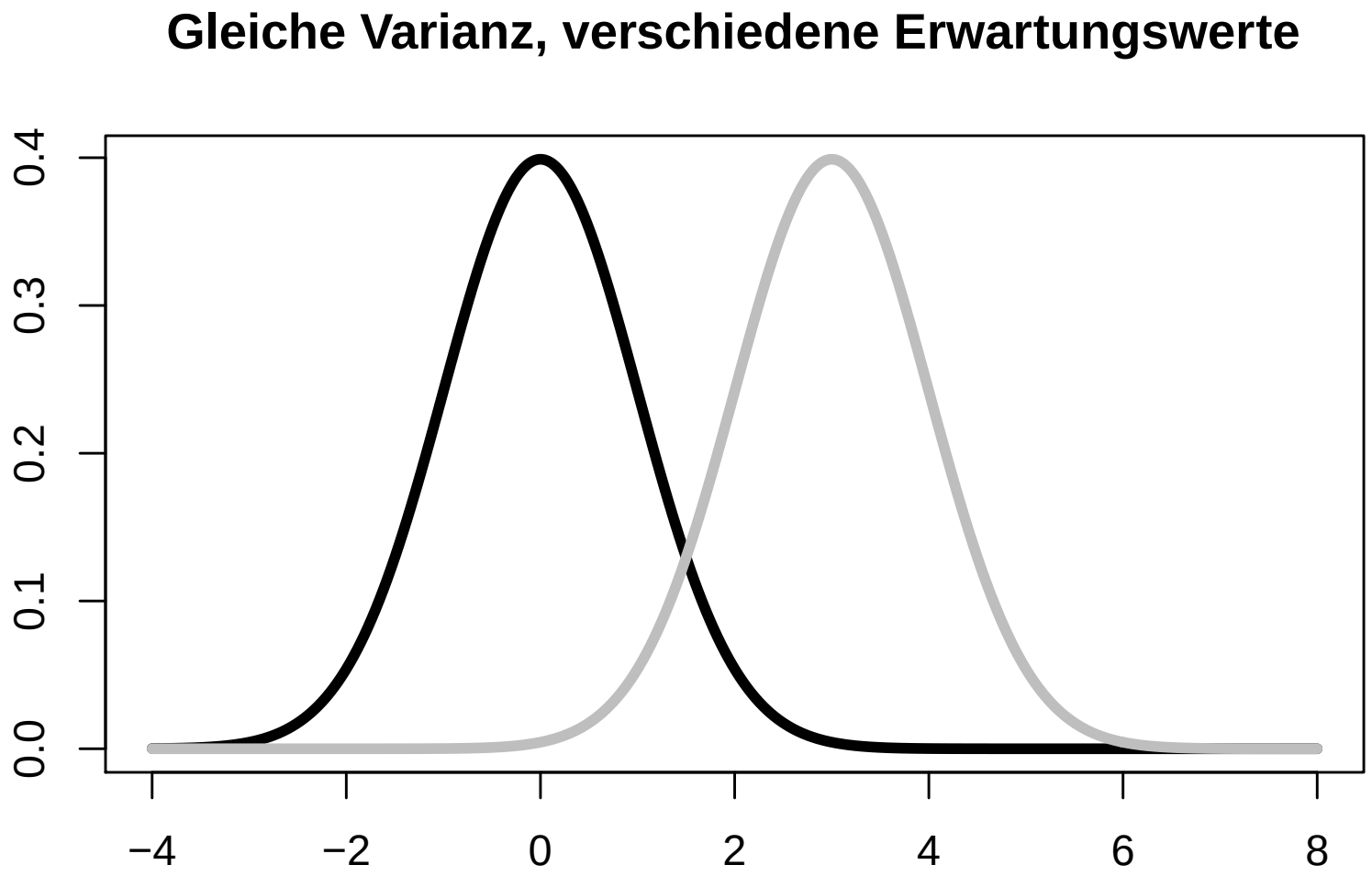
Die Varianz σ^2 einer Zufallsvariablen definiert sich als die mittlere quadratische Abweichung vom Erwartungswert, d.h.

$$\sigma^2 := Var(X) := E\left((X - E(X))^2\right).$$

Die **Standardabweichung** σ ist definiert durch: $\sigma := \sqrt{Var(X)}$.

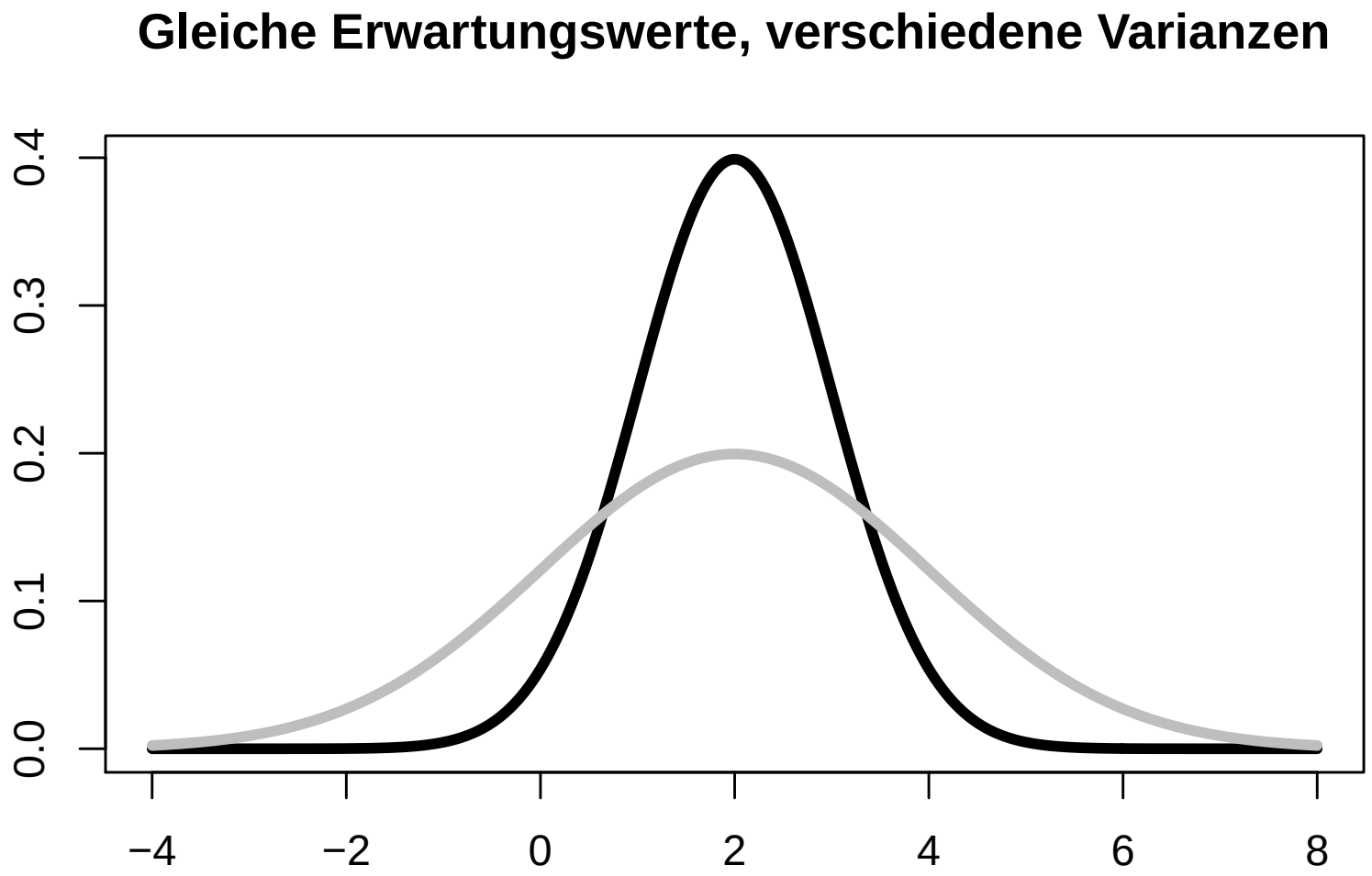
Kenngrößen von Zufallsvariablen

Den Erwartunswert nennt man auch **Lageparameter** der Verteilung:



Kenngrößen von Zufallsvariablen

Die Varianz nennt man auch **Streuungsparameter** einer Verteilung:



Empirische Kenngrößen

Dilemma in der Statistik

Die Kenngrößen einer Zufallsvariablen sind von zentraler Bedeutung, aber unbekannt!

Man behilft sich durch die Berechnung von **Schätzern** basierend auf der Stichprobe $X_1, \ldots, X_n$:

- Schätzer des Erwartungswertes μ: $\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$
- Schätzer der Varianz σ^2: $S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- Schätzer der Standardabweichung σ: $S := \sqrt{S^2}$

Die Schätzer hängen von der zufälligen Stichprobe ab, sind also selbst wiederum zufällig sind. Man unterliegt beim Schätzen einer theoretischen Kenngröße also stets einer gewissen **Unsicherheit**.

Empirische Kenngrößen

Berechnung der empirischen Schätzer in SPSS

- → *Analysieren*
- → *Deskriptive Statistiken*
- → *Deskriptive Statistik...*
- → Ziehe die zu untersuchenden Variablen in das Feld *Variable(n):* und bestätige mit OK.

Robuste Kenngrößen

Nachteil von $\bar{X}$ und S^2

Die beiden Schätzer $\bar{X}$ und S^2 von Mittelwert und Varianz haben allerdings einen praktischen Nachteil: sie sind sehr anfällig gegenüber **Ausreißern**.

Beispiel:
Der Datensatz `Milliardaer.sav` zeigt sehr deutlich, wie ein einziger Ausreißer den Mittelwert und die Standardabweichung verändern kann. Durch die Hinzunahme eines einzigen extremen Wertes verschiebt sich der Mittelwert und die Standardabweichung sehr stark, obwohl sich die Daten kaum geändert haben.

Die Motivation nach Lokations- und Dispersionsparametern, die weniger ausreißeranfällig sind, wird in diesem Beispiel klar.

Robuste Kenngrößen

Lokations- und Dispersionsparameter, die weit weniger sensibel auf einzelne Ausreißer in einem Datensatz reagieren sind:

- der **Median**:
  Dieser ist ein Maß für das Zentrum der Verteilung; links und rechts des Medians befinden sich genau 50% der Beobachtungen.
- der **Interquartilabstand (IQR)**:
  Der IQR ist Maß für die Streuung der Daten und gibt die Breite des Bereichs an, in dem genau die mittleren 50% der Beobachtungen liegen.

Lage- und Streuungsparameter die ausreißerunanfällig sind bezeichnet man auch als **robuste** Maße.

Robuste Kenngrößen

Die Berechnung von Median und IQR ist in SPSS ein wenig umständlich:

Berechnung von Median und IQR in SPSS

- → *Analysieren*
- → *Deskriptive Statistiken*
- → *Explorative Datenanalyse*
- → Ziehe die zu untersuchenden Variablen in das Feld *Abhängige Variablen* (ggfs. kann man im Feld *Faktorenliste* noch eine Gruppierungsvariable bestimmen)
- → Wähle im Feld *Anzeige* die Option *Statistiken* aus
- → Klicke das Feld *Optionen* an und wähle dann die Option *Paarweiser Fallausschluss*

Das ursprüngliche Ziel zu Beginn war die Analyse der Verteilung von Daten sowie die Bestimmung der Wahrscheinlichkeitsverteilung einer Messgröße.

Für unsere Zwecke betrachten wir eine vereinfachte Fragestellung:

Fragestellung bei der Verteilungsananlyse

Sind die vorliegenden stetigen Daten normalverteilt oder sind sie nicht normalverteilt?

$\rightarrow$ Wie geht man hier vor?

Verteilungsanalyse: Allgemeine Situation

Um Aussagen über die Verteilungseigenschaften von Daten zu machen, kann man sowohl grafische Hilfsmittel heranziehen, als auch Hypothesentests durchführen. Man sollte aber stets **beide** Möglichkeiten betrachten!

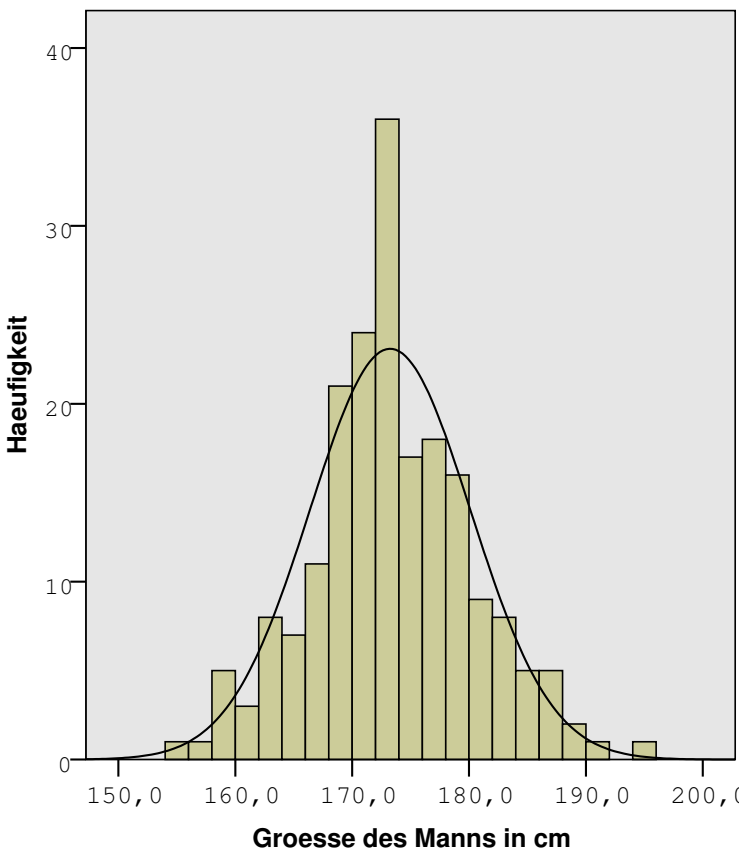
Die beiden wichtigsten grafischen Hilfsmittel zur Verteilungsanalyse sind:

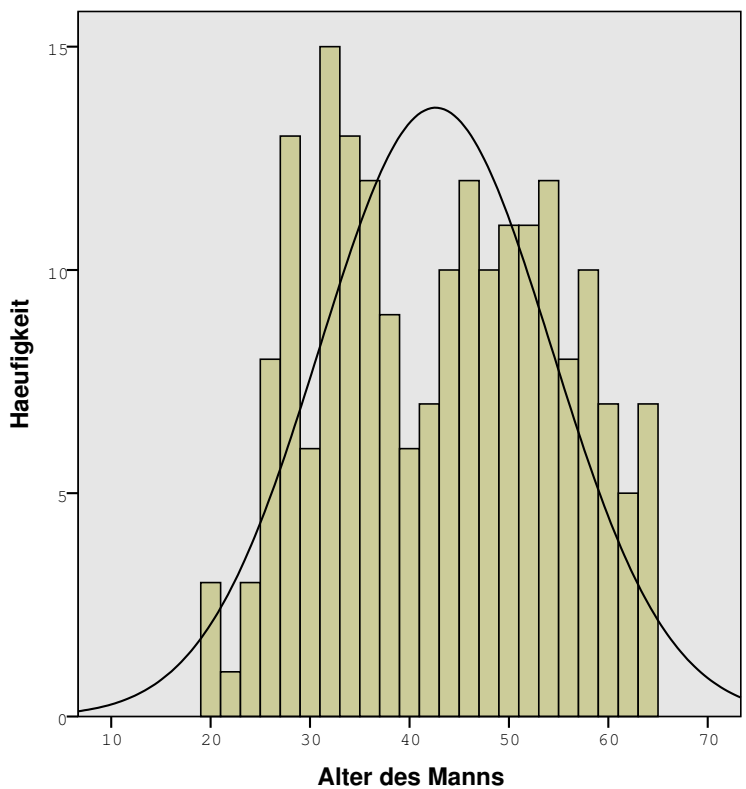
- **Histogramm** und
- **Boxplot**.

Es existieren noch weitere grafische Hilfsmittel wie beispielsweise der **Normal-Probability-Plot (Q-Q-Plot)** oder das **Stamm-Blatt-Diagramm**. Die beiden oben genannten Darstellungen der Daten sind aber die gebräuchlichsten, weshalb auf die Einführung weiterer Darstellungen verzichtet wird.

Histogramme

Wie oben beschrieben kann man mittels eines Histogramms erkennen, ob die Daten normalverteilt sind. Je nach dem wie gut die Anpassung an die theoretische Normalverteilungsdichte spricht dies eher für oder gegen einer Normalverteilung.





Boxplots

Ein weiteres wichtiges grafisches Hilfsmittel zur Beschreibung eines Datensatzes ist der **Box-Whisker-Plot**, kurz **Boxplot**.

Konstruktion eines Boxplots

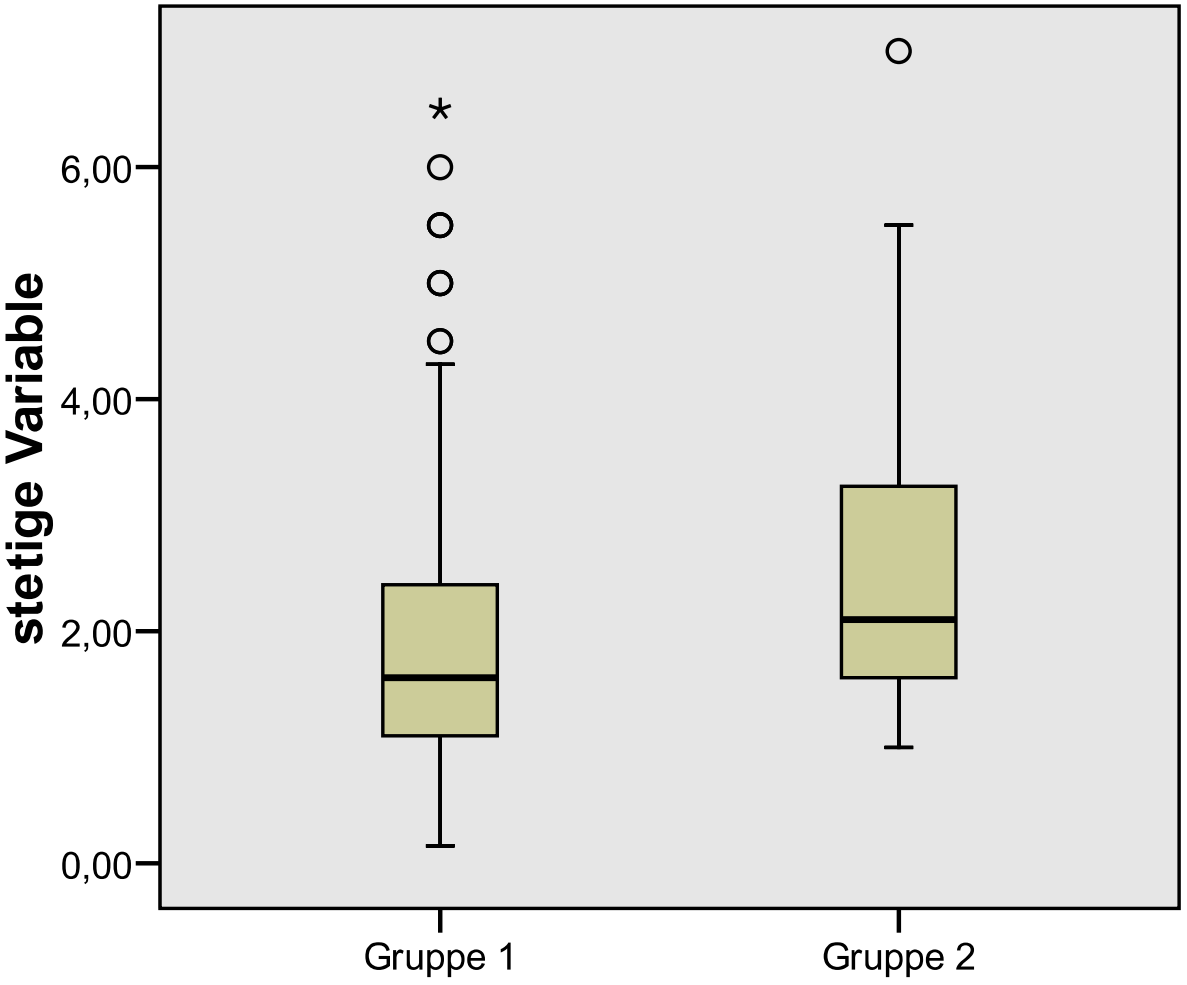
Ein Boxplot basiert auf dem Interquartilabstand (IQR), der genau die Werte in der „Box“ umfasst. Der Balken in der Mitte der Box ist der Median. Die Whisker beschreiben die Lage der Daten in den Außenbereichen und enden an den Stellen $\pm 1.5 \cdot IQR$. Alle Werte unter- und überhalb davon werden als Ausreißer gekennzeichnet.

$\Longrightarrow$ Der Vorteil des Boxplots besteht darin, dass man nicht nur über die Lokation der Daten, sondern auch über die Streuung der Daten (=Dispersion) auf einen Blick informiert wird.

$\Longrightarrow$ Sind die Daten beispielsweise nicht symmetrisch, können die Whisker unterschiedlich lang sein, sowie der Median nicht in der Mitte der Box liegen.

Boxplots

Beispiel für einen Boxplot:



Boxplots

Erstellung eines Boxplots in SPSS

- → *Analysieren*
- → *Deskriptive Statistiken*
- → *Explorative Datenanalyse*
- → Wähle das Feld *Diagramme* aus und wähle im Feld *Boxplots* die gewünschte Option (z.B. *Faktorstufen zusammen*)

Als Alternative kann man sich Boxplots auch ausgeben lassen unter:

Diagramme → *Veraltete Diaglogfelder* → *Boxplot*

Normalverteilungstests

Neben den grafischen Hilfsmittel gibt es auch inferenzstatistische Möglichkeiten, Aussagen darüber zu machen, ob die Daten einer Normalverteilung folgen.

In SPSS sind die beiden Standardtests hierfür:

- **Kolmogorov-Smirnov-Test**
- **Shapiro-Wilk-Test**

Zu bevorzugen ist jedoch stets der **Shapiro-Wilk-Test**. Um zu verstehen wie ein statistischer Test durchgeführt wird und wie man ein Testergebnis korrekt interpretiert, behandeln wir zunächst die Grundlagen von statistischen Hypothesentests.

Induktive Statistik

Neben der deskriptiven und der explorativen Statistik, ist das dritte große Teilgebiet der Statistik die **induktive Statistik** (auch **schließende Statistik** genannt).

Gegenstand der induktiven Statistik

Es wird versucht mit Hilfe einer Stichprobe auf Eigenschaften der Grundgesamtheit zu schließen. Diese Grundgesamtheit ist im Allgemeinen sehr viel größer als der Umfang der Stichprobe.

Die Methoden der induktiven Statistik bezeichnet man auch auch als **Testverfahren**. Dabei wird eine zu überprüfende Hypothese, auch **Nullhypothese** (oder H_0) aufgestellt, die mit einem **Test** auf Korrektheit überprüft wird.

Merke:
Nullhypothesen sind Präzisierungen der zu untersuchenden Fragestellung.

Beispiele für Nullhypothesen:

H_0 : Die Zufallsvariable X ist nach *irgendeiner* Normalverteilung $N(\mu, \sigma^2)$-verteilt, wobei μ und σ^2 beliebig seien.

H_1 : Die Zufallsvariable X ist nicht normalverteilt.

H_0 : Männer und Frauen haben einen gleich hohen IQ-Wert.

H_1 : Der IQ-Wert von Männern und Frauen ist nicht gleich.

H_0 : In der Firma XY verdienen Frauen genauso viel oder mehr als Männer.

H_1 : In der Firma XY verdienen Frauen weniger als Männer.

Hypothesentesten

Fassen wir zusammen:

- Zu einer aufgestellten Nullhypothese H_0 wird auch immer eine inhaltlich komplementäre **Alternativhypothese** H_1 formuliert.
- Die Nullhypothese H_0 stellt dann die Basis dar, von der aus entschieden wird, ob die Alternativhypothese H_1 akzeptiert werden kann oder nicht.

$\Longrightarrow$ Die eigentlich zu prüfende Hypothese muss also in die Alternativhypothese H_1 gesteckt werden!!

Achtung: Warum ist die Formulierung von

H_0 : Wohlhabende Kinder und sozial schwache Kinder unterschieden sich nicht in ihren Lesefähigkeiten.

H_1 : Wohlhabende Kinder können besser lesen als sozial schwache Kinder.

statistisch nicht korrekt?

Grundlegende Idee zur Überprüfung von H_0

Anhand einer gegebenen Stichprobe $X_1, \ldots, X_n$ von unabhängig und identisch verteilten Zufallsvariablen wird ein konkreter Wert, die sog. **Teststatistik** $T = T(X_1, \ldots, X_n)$ berechnet. Anhand von T und seiner Verteilung wird dann eine Entscheidung getroffen.

Beispiele für Teststatistiken werden wir bei der Besprechung der Testverfahren viele kennen lernen.

Die populärste Methode zur Hypothesenbeurteilung basierend auf einer Teststatistik T ist die Betrachtung des p-Wertes.

Der p-Wert

Der p-Wert ist die Wahrscheinlichkeit dafür, dass man unter der Nullhypothese H_0 das tatsächlich beobachtete Resultat oder sogar ein noch extremeres erhält.

$\Longrightarrow$ Je unwahrscheinlicher also die Gültigkeit von H_0, desto kleiner wird der p-Wert. Wenn eine gewisse Wahrscheinlichkeitsgrenze unterschritten wird, ist H_0 also so unwahrscheinlich, dass man sich für die Gültigkeit der Alternativhypothese H_1 entscheidet.

$\rightarrow$ Die populärste Grenze für die Wahrscheinlichkeit beträgt 0.05, d.h. ab einem p-Wert von kleiner oder gleich 0.05 wird H_0 abgelehnt.

$\Longrightarrow$ Der p-Wert ist sozusagen also ein Maß für die Glaubwürdigkeit der Nullhypothese.

Ein Signifikanztest gestattet nur **eine** der beiden folgenden Entscheidungen:

$$\textbf{Ablehnung von } H_0 \;=\; \textbf{Annahme von } H_1$$

oder

$$\textbf{Nicht-Ablehnung von } H_0 \;\neq\; \textbf{Annahme von } H_0$$

Dies bedeutet also:

$\Longrightarrow$ Die Nicht-Ablehnung von H_0 darf keinesfalls als ein Nachweis der statistischen Richtigkeit der Nullhypothese fehlinterpertiert werden.

$\Longrightarrow$ Streng genommen bedeutet eine Nicht-Ablehnung von H_0 also eine *Stimmenthaltung*, d.h. *das Stichprobenergebnis ist mit der Nullhypothese vereinbar*.

Bei einer Entscheidung basierend auf einem Signifikanztest hat man niemals absolute Sicherheit – egal wie man sich entscheidet es besteht also immer die Gefahr eine Fehlentscheidung zu treffen:

	H_0 ist wahr	H_0 ist nicht wahr
Entscheidung für H_0	kein Fehler	Fehler 2. Art (β)
Entscheidung für H_1	Fehler 1. Art (α)	kein Fehler

- Bei einem Signifikanztest kann man leider immer nur den Fehler 1. Art kontrollieren. Dieser ist stehts ≤ 0.05.
- Der Fehler 2. Art hingegen kann unter Umständen relativ groß werden.

$\rightarrow$ Dies ist die Begründung für das Vorgehen auf Folie 27, dass die eigentlich zu prüfende Hypothese als H_1 formuliert werden muss.

Rückkehr zum eigentlichen Problem

Nachdem die Grundzüge der Testtheorie behandelt wurden, können wir nun zum Test auf Normalverteilung zurückkehren. In SPSS gibt es zwei Tests auf Normalverteilung, den

- **Kolmogorov-Smirnov-Test** und den
- **Shapiro-Wilk-Test**.

Zu bevorzugen ist jedoch stets der **Shapiro-Wilk-Test**. Die Nullhypothese bei diesen Tests lautet:

$$H_0 : \text{Die Stichprobe ist normalverteilt}$$

Man beachte hierbei, dass man in diesem Fall daran interessiert ist H_0 nicht zu verwerfen – im Idealfall der p-Wert also größer als 0.05 sein sollte!

Normalverteilungstests

Erstellung von Normalverteilungstests in SPSS

- → *Analysieren*
- → *Deskriptive Statistiken*
- → *Explorative Datenanalyse*
- → Wähle das Feld *Diagramme* aus und klicke dort das Feld *Normalverteilungsdiagramm mit Tests* an.

$\Longrightarrow$ Zusammen mit den Normalverteilungstest werden in SPSS immer auch die zugehörigen Q-Q-Plots, sowie die trendbereinigten Q-Q-Plots ausgegeben (siehe oben).

Zusammenfassung

Man hat nun also zwei Möglichkeiten die Verteilungseigenschaften der Daten zu überprüfen:

- **grafisch:** Boxplots, Histogramme, Q-Q-Plots, ...
- **inferenzstatistisch:** Shapiro-Wilk-Test, ...

Dabei ist aber immer zu beachten:

Grundregel bei der Verteilungsanalyse

Man betrachtet aber nie nur eine der beiden Möglichkeiten, sondern immer **beide zusammen**!

Manchmal verrät eine der beiden Möglichkeiten nämlich mehr über die Eigenschaften der Daten als die andere ...

Voraussetzungen von Testverfahren

Zu jedem Testverfahren, gibt es gewisse Voraussetzungen an die Daten, die erfüllt sein müssen um die Aussagekraft des Testverfahrens sicher zu stellen (z.B. muss beim t-Test die Normalverteilungsannahme erfüllt sein).

Man beachte stets

Aussagen in der Statistik sind höchstens so sicher wie die Voraussetzungen dieser Aussagen.

$\Longrightarrow$ Sind die Voraussetzugen eines Testverfahrens nicht oder nur teilweise erfüllt, so muss dies in der entsprechenden vorsichtigen Interpretation des Resultates berücksichtigt werden!

$\Longrightarrow$ Im Zweifelsfall ist es besser auf statistische Tests zu verzichten und sich mit einer einfachen Beschreibung der Daten anhand tabellarischer und grafischer Darstellungen zu begnügen!