Aus Wikipedia:

Deskriptive Statistik (auch beschreibende Statistik oder empirische Statistik): Vorliegende Daten werden in geeigneter Weise beschrieben, aufbereitet und zusammengefasst. Mit ihren Methoden verdichtet man Daten zu Tabellen, graphischen Darstellungen und Kennzahlen.

Induktive Statistik (auch mathematische Statistik, schließende Statistik, beurteilende Statistik oder Inferenzstatistik): In der induktiven Statistik leitet man aus den Daten einer Stichprobe Eigenschaften einer Grundgesamtheit ab. Die **Wahrscheinlichkeitstheorie** liefert die Grundlagen für die erforderlichen Schätz- und Testverfahren.

Die **Wahrscheinlichkeitstheorie** oder auch **Wahrscheinlichkeitsrechnung** ist ein Teilgebiet der Mathematik, das aus der Formalisierung, der Modellierung und der Untersuchung von Zufallsgeschehen hervorgegangen ist.

Inhalt der Vorlesung (vgl. Modulhandbuch):

Deskriptive Statistik

- Visualisierung von Verteilungen
- Skalen
- Lageparameter
- Streuungsparameter
- Bivariate Zusammenhänge (Visualisierung, Kovarianz und Korrelation)
- Lineare Regression

Wahrscheinlichkeitstheorie

- Wahrscheinlichkeitsräume
- Laplace-Wahrscheinlichkeit und Kombinatorik
- Bedingte Wahrscheinlichkeit und Satz von Bayes
- Zufallsvariable
- Diskrete Verteilungen
- Stetige Verteilungen
- Normalverteilung und zentraler Grenzwertsatz

In den **Übungen** wird zunächst mehr die deskriptive Statistik im Fokus stehen, in der Vorlesung dann bald die Wahrscheinlichkeitstheorie.

Dabei werden in den Übungen teilweise ergänzende und vertiefende Topics der deskriptiven Statistik behandelt, nicht notwendig synchron zur Vorlesung. Diese sind aber **ebenso Prüfungsstoff**.

Sie werden zu dieser Vorlesung Teile getippt in Moodle vorfinden, andere Abschnitte werden dann auch wieder in reiner Tafelarbeit präsentiert (womit ich eigentlich die besseren Erfahrungen gemacht habe). Teilweise wird es Mischformen geben, auch mit eingescannten handschriftlichen Teilen oder auch mit kopierten anderweitigen Quellen. Ihre Moodle-Unterlagen werden also kein komplettes Vorlesungsskript zum Selbststudium darstellen. Es lohnt sich, in die Vorlesung zu kommen, oder sich notfalls eine Mitschrift zu organisieren.

1 Deskriptive Statistik

1.1 Skalen, Lageparameter und Streuungsmaße

Die Visualisierung von Verteilungen wird in den Übungen genauer behandelt.

Im Folgenden wird oft zitiert aus

ARENS, Tilo, 2015, *Mathematik*, 3. Auflage, Berlin, Springer Spektrum. ISBN 978-3-642-44918-5 (Auch als eBook vorhanden)

So auch das folgende Beispiel aus Kap. 36:

„An einem kalten Sonntag am 10. Dezember 2006 stürzte der schwer bepackte Radfahrer Hans Meier mit seinem Fahrrad auf der regennassen Goethestraße in Berlin Mitte. Nach kurzer Behandlung in einem Krankenwagen konnte er seine Fahrt fortsetzen." So könnte eine Erzählung beginnen. Nehmen wir die Beschreibung als korrekt an und fragen nach ihrem statistischen Gehalt. Versuchen wir, das Wesentliche herauszuholen. Doch was wichtig ist, hängt von uns und unseren Interessen ab: Wollen wir etwas aussagen über das Wetter im Dezember, über Hans Meier, über Unfälle in der Goethestraße oder über Unfälle mit Fahrrädern?

Als Grundausrüstung für ein mathematisches Modell, in dem wir Statistik treiben können, haben wir die Sprache der Mengen und Abbildungen (Funktionen). Die Statistik macht Aussagen über **Grundgesamtheiten**, das sind dann einfach Mengen Ω, gestützt auf Beobachtungen an einzelnen Elementen $\omega \in \Omega$.
Im obigen Beispiel könnte die Grundgesamtheit Ω die Menge aller Patienten sein, die am 10.12.06 in einem bestimmten Einsatzwagen untersucht wurden.
Gegenstand einer statistischen Erhebung sind **Merkmale** und deren **Ausprägungen**.

Definition: Ein **Merkmal** X ist eine Abbildung $X: \Omega \to \mathrm{M}$ von der Menge Ω (also der Grundgesamtheit) in eine Menge M (die oft aus Zahlen besteht). Die einzelnen Bilder $X(\omega)$ sind dann die sogenannten **Ausprägungen** des Merkmals X. Merkmale werden meist mit lateinischen Großbuchstaben, deren Ausprägungen mit kleinen lateinischen Buchstaben bezeichnet: $X(\omega) = x$

Im Beispiel könnte sein X = „Alter in Jahren", z.B. $X(Hans\ Meier) = 23$, $X(Anne\ Müller) = 22$, und so fort. Allgemein erhalten wir zu n Elementen $\omega_1, \omega_2, \dots, \omega_n$ die Ausprägungen
$x_1 = X(\omega_1),\ x_2 = X(\omega_2), \dots, x_n = X(\omega_n)$

Anmerkung: Hier ist schon die erste Gelegenheit zum Aussteigen: Der Grund, warum Viele mit Wahrscheinlichkeitsrechnung und Statistik Probleme haben, ist schlicht die ungewohnte Verwendung von Buchstabensorten im mathematischen Formalismus: Sie haben gelernt: Eine Abbildung $f: \mathrm{X} \to \mathrm{Y}$ von einer Menge X in eine Menge Y ordnet jedem Element $x \in \mathrm{X}$ genau ein sogenanntes Bildelement $y = f(x)$ aus der Menge Y zu. Wir denken dabei an Beispiele wie $f: \mathbb{R} \to [-1; 1]$, $f(x) = \sin(x)$.
In der Sprache der Mathematik besteht zwischen $f(x) = y$ und $X(\omega) = x$ kein begrifflicher Unterschied, und $X(Hans\ Meier) = 23$ ist dann auch so etwas wie anderswo $f\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2}$. Und auch in der Wahrscheinlichkeitsrechnung werden wir die Symbolik $X: \Omega \to \mathbb{R}$ und $X(\omega) = x$ verwenden. Dort ist dann Ω die Menge aller möglichen Ergebnisse ω eines Zufallsexperiments und die Abbildung X eine sogenannte Zufallsvariable, wie wir später sehen werden.

Bei den Merkmalsausprägungen $x_1, x_2, \dots, x_n$ können natürlich einige Werte öfters vorkommen. Wir erhalten eine **Häufigkeitsverteilung** des Merkmals X, wenn wir zu jedem der vorkommenden verschiedenen (!) Werte a_j jeweils die Häufigkeit n_j ermitteln, mit der dieser Wert in der Liste $x_1, x_2, \dots, x_n$ auftritt. So können wir zum Beispiel aus der Reihe der beobachteten Ausprägungen 2,3,4,1,3,5,1,3,5,2,1,3 eines Merkmals (Wir denken zum Beispiel an die Notenliste von 12 Kandidaten) in einer Tabelle

Wert	1	2	3	4	5
Anzahl	3	2	4	1	2

die Häufigkeitsverteilung dieses Merkmals darstellen.

Gerade beim Vorliegen großer Datenmengen (großem n) ist ein gutes Visualisieren von Bedeutung. Die Art der darstellenden Grafiken, und auch die Extrakte charakteristischer Daten oder Zahlenwerte aus den Merkmalsausprägungen $x_1, x_2, \dots, x_n$ ist abhängig von deren Struktur. Der häufigste Fall wird sein, dass die x_i Zahlenwerte sind. Man spricht dann von einem **quantitativen Merkmal** (kardinalem Merkmal). Man sagt auch, das Merkmal sei **kardinal skaliert**. Manchmal kommt es bei den Ausprägungen $x_1, x_2, \dots, x_n$ nur auf die Vergleichbarkeit (Position auf einer Rangskala) an. Im Eingangsbeispiel könnte $X(\omega)$ das subjektive Befinden des Patienten ω auf einer Skala von 1 bis 10 sein. Man spricht dann von einem **ordinalen** (**ordinal skaliertem**) **Merkmal**. Manchmal haben die Merkmalsausprägungen aber auch gar keine intrinsische Struktur (außer gleich/ungleich). Beispielsweise $X(\omega)$ = Lohnsteuerklasse der Person ω, oder $X(\omega)$ = Geschlecht der Person ω, oder $X(\omega)$ = Name der Krankenversicherung der Person ω. In diesem Fall spricht man von einem **nominalen Merkmal**.

Aufgabe

Treffen Sie die richtigen Zuordnungen von Merkmalen zu Skalentypen durch Ankreuzen in der folgenden Tabelle

	nominal	ordinal	quantitativ
Buchbesprechungen auf einer Web-Seite (1-5 Sterne)			
Geschwindigkeit			
Farbe eines Autos			
Blutgruppe (A,B,AB oder 0)			
Kleidergröße (S,M,L,XL)			
Klausurnoten (1,0;1,3;1,7;2,0;2,3; … 4,0;5,0)			
Bundesland			
Datum			
Temperatur			
Produzierte Mengen Tee in Indien pro Jahr			

Je nach Merkmalstyp gibt es verschiedene Arten der Darstellung (Visualisierung), wie zum Beispiel Strich- Balken- oder Kuchendiagramme (Kreissektorendiagramme) → Übungen.

Lageparameter eines Datensatzes (einer Verteilung) $x_1, x_2, \dots, x_n$:

Definitionen:

1) Der (ein) **Modus** oder **Modalwert** x_{mod} ist die (eine) Ausprägung mit maximaler Häufigkeit.

2) Bei einem der Größe nach geordneten Datensatz $x_1 \leq x_2 \leq \dots \leq x_n$ teilt der **Median**

$$x_{\mathrm{med}} := \begin{cases} x_{\frac{n+1}{2}} & falls\ n\ ungerade \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & falls\ n\ gerade \end{cases}$$

den Datensatz in zwei gleich große Hälften: Bei ungeradem n ist der Median x_{med} also der Wert genau in der Mitte der (sortierten Liste), bei geradem n der Durchschnitt der beiden Werte in der Mitte.

3) Das **arithmetische Mittel** ist $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$

Beispiel mit $n = 12$: Bei dem Datensatz (2,3,4,1,3,5,1,3,5,2,1,3) ist x_{mod} = 3, weil die 3 am öftesten vorkommt (4 mal). Zur Bildung des Medians sortieren wir zunächst: (1,1,1,2,2,3,3,3,3,4,5,5). Man sagt auch: Aus der **Urliste** machen wir eine **Rangliste**. Die mittleren beiden Werte sind die erste und die zweite 3. Damit 3 auch deren Mittelwert, also x_{med} = 3.
Das arithmetische Mittel $\bar{x}$ des ganzen Datensatzes berechnen wir am besten aus seiner Häufigkeitsverteilung

Wert	1	2	3	4	5
Anzahl	3	2	4	1	2

$\bar{x} = \frac{1}{12}(3 \cdot 1 + 2 \cdot 2 + 4 \cdot 3 + 1 \cdot 4 + 2 \cdot 5) = 2{,}75$

Bemerkungen:

a) Ein Datensatz kann mehrere Modi haben (bei gleich häufigen Ausprägungen). Es sind also multimodale, bimodale und unimodale Verteilungen möglich. Im obigen Beispiel lag Unimodalität vor. Der Modus ist vor allem bei nominalen Merkmalen sinnvoll (am häufigsten nachgefragte Schuhgröße, häufigste Unfallursache, Partei mit den meisten Stimmen).

b) (Ohne Beweis) Der Median ist diejenige Zahl z, für die die Summe der absoluten Abweichungen $a(z) \coloneqq \sum_{i=1}^{n} |x_i - z|$ minimal wird.

c) Das arithmetische Mittel $\bar{x}$, also der Durchschnitt, ist der bekannteste und auch wichtigste Lageparameter. Bei nominalen Merkmalen (sofern sie überhaupt Zahlenwerte haben) macht $\bar{x}$ natürlich keinen Sinn.

d) Die Summe aller Abweichungen von Mittelwert ist null: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
Beweis → Tafel

e) Der Mittelwert ist diejenige Zahl z, für die die Summe der Abstandsquadrate $q(z) \coloneqq \sum_{i=1}^{n}(x_i - z)^2$ minimal wird. Beweis → Tafel

e) Die Berechnung von $\bar{x}$ im obigen Beispiel anhand der Häufigkeitsverteilung geht ganz allgemein so: Sind $a_1, a_2, \dots, a_k$ die **verschiedenen** Werte, welche im Datensatz $x_1, x_2, \dots, x_n$ vorkommen, und zwar mit den Häufigkeiten $n_1, n_2, \dots, n_k$, $\sum_{j=1}^{k} n_j = n$, dann ist $\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j a_j$.
Sind jeweils $h_j \coloneqq \frac{n_j}{n}$ die sogenannten **relativen Häufigkeiten** der verschiedenen Werte a_j, $\sum_{j=1}^{k} h_j = 1$, dann ist $\bar{x} = \sum_{j=1}^{k} h_j a_j$.

f) Wendet man auf die Daten x_i eine lineare Transformation (korrekter müsste es eigentlich heißen „affin lineare Transformation“) $y_i = f(x_i) \coloneqq ax_i + b$ (mit festen $a, b \in \mathbb{R}$) an, so ist $\bar{y} = f(\bar{x})$.
Beweis → Tafel
Z.B. könnten wir Messwerte x_i, welche in Grad Celsius vorliegen, mittels $y_i \coloneqq 1{,}8x_i + 32$ auf Grad Fahrenheit umrechnen. Obige Formel besagt dann, dass wir für den Mittelwert $\bar{y}$ in Grad Fahrenheit nicht erst alle x_i einzeln auf y_i umrechnen und daraus den Durchschnittswert $\bar{y}$ zu berechnen brauchen, sondern wir können stattdessen auch einfach den Mittelwert $\bar{x}$ von Celsius auf Fahrenheit umrechnen: $\bar{y} = 1{,}8\bar{x} + 32$.

Aufgabe (Arens Kap. 36)

Von einer Fußballmannschaft (11 Mann) sind 4 Spieler jünger als 25 Jahre, 3 sind 25, der Rest (4 Spieler) ist älter. Das Durchschnittsalter liegt bei 28 Jahren. Wo liegt der Median? Wie ändern sich Median und Mittelwert, wenn für den 40-jährigen Torwart ein 18-jähriger eingewechselt wird?

Streuungsparameter eines Datensatzes (einer Verteilung) $x_1, x_2, \dots, x_n$:
Sie quantifizieren, wie stark die Ausprägungen um ihr „Zentrum“ herum „streuen“.

Definitionen:

1) Bei einem der Größe nach geordneten Datensatz $x_1 \leq x_2 \leq \dots \leq x_n$ ist die Differenz zwischen größtem und kleinsten Wert $x_n - x_1$ die **Spannweite**. Der **Quartilsabstand** errechnet sich aus der Differenz $q_o - q_u$ aus oberem und unteren Quartil: Dazu noch eine Definition: Der Median teilt zunächst die sortierte Liste in zwei gleich große sortierte Listen. Das **untere Quartil** q_u ist dann der Median der unteren Liste, entsprechend das **obere Quartil** q_o der Median der oberen Liste.

⚠ Es gibt unterschiedliche Definitionen (und Bezeichnungen) für die Begriffe „unteres Quartil“ und „oberes Quartil“ (und manchmal wird q_u mit $q_{0,25}$ und q_o mit $q_{0,75}$ bezeichnet, oder es wird z.B. auch das untere Quartil als „erstes Quartil“, der Median als „zweites Quartil“ und das obere Quartil als „drittes Quartil“ bezeichnet). Entsprechend kann man bei der Berechnung des Quartilsabstandes zu unterschiedlichen Ergebnissen kommen. Wir werden im weiteren Verlauf dieser Vorlesung den Begriff des Quartilsabstands ohnehin nicht mehr benötigen.

2) Erwähnt seien hier noch die **mittlere absolute Abweichung vom Median** $\frac{1}{n}\sum_{i=1}^{n}|x_i - x_{med}|$ und der **Median der absoluten Abweichungen vom Median**, oft bezeichnet mit MAD. Dies ist also der Median des Datensatzes $|x_1 - x_{med}|, |x_2 - x_{med}|, \dots, |x_n - x_{med}|$ (den man zweckmäßigerweise auch wieder sortieren kann). Das wichtigste und auch theoretisch am besten erforschte Streuungsmaß ist jedoch

3) die **Varianz** $var(\boldsymbol{x}) \coloneqq \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Die Wurzel aus der Varianz heißt **Standardabweichung** $s \coloneqq \sqrt{var(\boldsymbol{x})}$. Die Varianz wird deshalb oft auch mit s^2 bezeichnet.

Beispiel: Wir nehmen wieder den Datensatz (1,1,1,2,2,3,3,3,3,4,5,5) aus dem vorigen Beispiel mit $x_{\text{med}} = 3$ und $\bar{x} = 2{,}75$. Die Spannweite ist 5 - 1 = 4. Die untere Hälfte (1,1,1,2,2,3) hat als Median die Zahl 1,5. Die obere Hälfte (3,3,3,4,5,5) hat als Median die Zahl 3,5. Dies ergibt den Quartilsabstand 3,5 – 1,5 = 2.
Um Missverständnissen vorzubeugen hier noch ein weiteres Beispiel, diesmal mit $n = 15$:
(2,2,2,3,3,3,4,4,5,5,6,6,7,7,8) Hier teilt der Median 4 (das ist die zweite 4 in der Reihe, die also genau in der Mitte steht) die Liste in die untere Hälfte (2,2,2,3,3,3,4) mit Median 3 (steht wieder genau in der Mitte) und die obere Hälfte (5,5,6,6,7,7,8) mit einer 6 in der Mitte. Der Quartilsabstand wäre somit 6 – 3 = 3.
Bleiben wir beim ersten Beispiel. Dann ist die mittlere absolute Abweichung vom Median $\frac{1}{12}(3 \cdot |1-3| + 2 \cdot |2-3| + 4 \cdot |3-3| + 1 \cdot |4-3| + 2 \cdot |5-3|) = \frac{13}{12} \approx 1{,}08\overline{3}$. Zur Berechnung von MAD sortieren wir die absoluten Abweichungen vom Median 3: (2,2,2,1,1,0,0,0,0,1,2,2). Das ergibt die Rangliste (0,0,0,0,1,1,1,2,2,2,2,2) mit Median 1 (als Mittelwert der beiden mittleren Einsen), also MAD = 1.
Als Varianz bekommen wir $\frac{1}{12}(3 \cdot (1-2{,}75)^2 + 2 \cdot (2-2{,}75)^2 + 4 \cdot (3-2{,}75)^2 + 1 \cdot (4-2{,}75)^2 + 2 \cdot (5-2{,}75)^2) = 1{,}8541\overline{6}$ und als Standardabweichung $s \approx 1{,}3617$.

Bemerkungen:

a) Die Spannweite als Streuungsmaß ist sehr empfindlich gegenüber „Ausreißern“ (Arens Kap. 36):

In einer Gruppe von 10 Studenten kann die Spannweite des verfügbaren Monatseinkommens vielleicht 50 € betragen. Bei 500 Studenten beträgt die Spannweite vielleicht 1000 €. Ist zufällig der Sohn eines Ölscheichs dabei, dann könnte die Spannweite 1000 000 € betragen.

Der Quartilsabstand gibt an, wie „die mittleren 50%“ streuen. Im Gegensatz zur Spannweite ist er ein sehr robustes Streuungsmaß: Im Extremfall können bis zu 25% der Daten grob falsch sein, ohne dass der Quartilsabstand wesentlich verfälscht wird.

Spannweite und Quartilsabstand machen Sinn bei ordinalen und kardinalen Skalen, aber natürlich nicht bei nominalen Merkmalen.

b) Ähnlich wie bei der Berechnung des arithmetische Mittels $\bar{x}$ lassen sich auch bei der Berechnung der Varianz die relativen Häufigkeiten $h_j \coloneqq \frac{n_j}{n}$, $j = 1, \dots, k$ (der verschiedenen Werte $a_1, a_2, \dots, a_k$, welche im Datensatz $x_1, x_2, \dots, x_n$ vorkommen) verwenden:
$var(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{k} n_j\left(a_j - \bar{x}\right)^2 = \sum_{j=1}^{k} h_j(a_j - \bar{x})^2$,
wie wir es schon im vorigen Rechenbeispiel gemacht haben (Analog kann man natürlich auch bei der Berechnung der mittleren absoluten Abweichung vom Median verfahren).

c) Zur Nomenklatur: $var(\boldsymbol{x})$ wird auch als $var(x_i)$ geschrieben. Mit dem fetten Buchstaben $\boldsymbol{x}$ ist der „Vektor“ $(x_1, x_2, \dots, x_n)$ gemeint. In manchen Lehrbüchern findet man auch die Definition $var(\boldsymbol{x}) \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Auf diesen subtilen Unterschied (der bei großem n nicht ins Gewicht fällt) werden wir später eingehen.

d) Für lineare Transformationen $y_i \coloneqq ax_i + b$ gilt $var(\boldsymbol{y}) = a^2 var(\boldsymbol{x})$
Beweis → Tafel

e) Oft ist es bequemer, $\sum_{i=1}^{n}(x_i - a)^2$ mit einem geeigneten a zu berechnen, anstatt $\sum_{i=1}^{n}(x_i - \bar{x})^2$.
Es gilt der **Verschiebungssatz**: $\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 = var(\boldsymbol{x}) + (\bar{x} - a)^2$.
Beweis → Tafel
Speziell für $a = 0$ ist dann $\frac{1}{n}\sum_{i=1}^{n} {x_i}^2 = var(\boldsymbol{x}) + \bar{x}^2$.

Beispiel: Quadrieren wir die Werte im Datensatz (1,1,1,2,2,3,3,3,3,4,5,5) des vorigen Beispiels, so erhalten wir (1,1,1,4,4,9,9,9,9,16,25,25) mit dem arithmetischen Mittel $\frac{113}{12}$. Subtrahiert man davon $\bar{x}^2 = 2{,}75^2$, so erhält man wieder die Varianz $\frac{113}{12} - 2{,}75^2 = 1{,}8541\overline{6}$