Notes - Einführung in die angewandte Statistik

Vorlesung aus dem Wintersemester 2020 / 2021 von Prof. Udo Kamps an der RWTH Aachen

Autor: Jannis Zeller

Letzte Änderung: 4. Oktober 2022

Inhaltsverzeichnis

Disclaimer

Diese Zusammenfassung ist nicht offiziell von der RWTH Aachen oder Dozierenden der betreffenden Lehrveranstaltungen bestätigt oder erprobt. Sie wurde nach bestem Wissen und Gewissen erstellt.

Notation

- **Mächtigkeit einer Menge** A:

$$|A| = \text{Anzhal der Elemente einer Menge}$$

- **Identitätsfunktion**

$$\mathbf{1}_{[a,b]}(y) := \begin{cases} 1, & y \in [a,b] \\ 0, & y \notin [a,b] \end{cases}, \qquad \mathbf{1}_{(a,b)}(y) := \begin{cases} 1, & y \in (a,b) \\ 0, & y \notin (a,b) \end{cases}$$

- Bei der **Nummerierung** der Abschnitte wird die Nummerierung der Vorlesung, bzw. dem der Vorlesung zugrunde liegenden Buch adaptiert.

1 Einführung und Grundbegriffe

1.1 Grundlegende Begriffe

Definition 1.1. Die **Grundgesamtheit / Population / Kollektiv**...
...ist eine Menge von räumlich und zeitlich eindeutig definierten Objekten, die hinsichtlich bestimmter - vom Ziel der Untersuchung abhängender - Kriterien übereinstimmen. Die Objekte einer Grundgesamtheit werden als **statistische Einheiten** (auch Merkmalsträger, Untersuchungseinheiten oder Messobjekte) bezeichnet.
Trennt man eine Grundgesamtheit anhand eines Merkmals (s. u.) so spricht man auch von **Teilgesamtheiten**.

Definition 1.2. Als **Merkmal**...
...$(X, Y, Z,...)$ bezeichnet man eine spezielle Eigenschaft der statistischen Einheiten einer Population, die im Hinblick auf das Ziel einer konkreten statistischen Untersuchung von Interesse ist.
Der Wert eines Merkmals wird als **Merkmalsausprägung** $(x, y, z,...)$ mit Werten im **Wertebereich** bezeichnet.

Definition 1.3. Uni- und Multivariate Merkmale
Wird nur eine Eigenschaft der Merkmalsträger gemessen, spricht man von einem univariatem Merkmal. Handelt es sich um mehrere Eigenschaften handelt es sich um ein multivariates Merkmal.

Definition 1.4. Als **Datum / Messwert / Beobachtungswert**...
...bezeichnet man die an einer statistischen Einheit gemessene Merkmalsausprägung.
Die Liste aller Daten $(x_1, ..., x_n$, etc.) liefert die/den **Urliste / Datensatz**.
Die Anzahl der Messwerte liefert den **Stichprobenumfang**.

1.2 Klassifikation von Merkmalen

Definition 1.5. Als **Skala**...
...bezeichnet man eine Vorschrift, die jeder statistischen Einheit der Stichprobe einen Beobachtungswert zuordnet.

Definition 1.6. Klassifikation von Merkmalen

$$\textbf{Merkmalstyp} \begin{cases} \text{qualitativ} & \begin{cases} \text{nominal } (\textit{Geschlecht}) \\ \text{ordinal } (\textit{Schulnoten}) \end{cases} \\ \text{quantitativ / metrisch} & \begin{cases} \text{diskret } (\textit{Augenzahlen Würfel}) \\ \text{stetig } (\textit{Temperatur}) \end{cases} \end{cases}$$

Als **dichotomes Merkmal** bezeichnet man ein nominales Merkmal mit genau zwei Ausprägungen.

Als **quasi stetige Merkmale** werden Merkmale bezeichnet, deren Ausprägung aus Gründen der Messgenauigkeit (*Zeit*) oder aufgrund der zugehörigen Messeinheit (*Währung*) nur diskret messbar ist, die aber aufgrund der Feinheit der Abstufungen als stetig angesehen werden.

Achtung: Ein Merkmal kann je nach Klassifikation mit unterschiedlichen Skalenniveaus gemessen werden (z. B. Körpergröße: kleiner als 1.80 m vs. größer als 1.80 m / klein, mittel, groß / Intervalle / Zentimeterangabe...

Als **klassiertes Merkmal** bezeichnet man Merkmale, deren Ausprägungen Intervalle sind.

2 Tabellarische und graphische Darstellung

2.1 Absolute und Relative Häufigkeiten

Definition 2.1. Absolute Häufigkeit
Die Anzahl n_j des Auftretens einer Merkmalsausprägung u_j der Urliste $x_1, ..., x_n$ heißt absolute Häufigkeit der Beobachtung u_j, $j \in \{1, ..., m\}$. Es gilt also:

$$n_j = \left|\left\{i \in \{1, ...n\} | x_i = u_j\right\}\right|$$

Regel 2.1. ***Summe der absoluten Häufigkeiten***
Für die absoluten Häufigkeiten $n_1, ..., n_m$ der verschiedenen Ausprägungen $u_1, ..., u_m$ einer Urliste $x_1, ..., x_n$ gilt stets:

$$\sum_{j=1}^{m} n_j = n$$

Definition 2.2. Relative Häufigkeit
Sei n_j, $j \in \{1, ..., m\}$ die absolute Häufigkeit einer Merkmalsausprägung in der Urliste $x_1, ..., x_n$. Dann heißt

$$f_j = \frac{n_j}{n}$$

die relative Häufigkeit der Mermalsausprägung u_j.

Aus 2.1 folgt sofort:

$$\sum_{j=1}^{m} f_j = 1\,.$$

- Tabellarische Darstellungen von Häufigkeiten heißen **Häufigkeitstabelle.**
- Das m-Tupel $(f_1, ..., f_m)$ heißt **Häufigkeitsverteilung** des Merkmals.
- Man spricht von **kumulierten Häufigkeiten**, wenn Mermalsausprägungen zusammengefasst, also Häufigkeiten addiert werden.

Definition 2.3. Illustrationen relativer Häufigkeiten

- **Stabdiagramm / Säulendiagramm**
  Bei einem Stabdiagramm / Säulendiagramm ist die Höhe des Stabes / der Säule, der eine relative Häufigkeit eines Merkmals illustriert, proportional zu diesem Merkmal. Die Abszisse ist dabei nicht zwingend metrisch, wenn das Merkmal metrisch ist.
- **Kreisdiagramm**
  Bei einem Kreisdiagramm ist die Fläche des Kreissegmentes bzw. der **Winkel** des Segments proportional zur relativen Häufigkeit.

$\Rightarrow$ Um auch für **stetige Merkmale** sinnvolle Häufigkeitstabellen und Diagramme erstellen zu können, müssen die Daten meist klassiert werden.

2.2 Empirische Verteilungsfunktion

Ziel:

Graphische Darstellung für Häufigkeiten quantitativer Merkmale

$\Rightarrow$ Ausprägungen werden auf der Abszisse abgetragen.

$\Rightarrow$ kumulierte Häufigkeiten werden auf der Ordinate abgetragen.

$\Rightarrow$ **Empirische Verteilungsfunktion** an der Stelle x = Anteil der Beobachtungen kleiner gleich einem Wert x.

Definition 2.4. Die **empirische Verteilungsfunktion**...
$F_n : \mathbb{R} \to [0, 1]$ wird für $x_1, ..., x_n \in \mathbb{R}$ definiert durch

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(x_i), \quad x \in \mathbb{R}\,.$$

Definition 2.5. Rangwertreihe
Für Beobachtungswerte $y_1, ..., y_r$ eines metrisch skalierten Merkmals heißt die aufsteigend geordnete Auflistung der Beobachtungswerte

$$y_{(1)} \leq y_{(2)} \leq ... \leq y_{(r)}$$

Rangwertreihe.

Regel 2.2. ***Empirische Verteilungsfunktion und Rangwertreihe***
Liegen im Datensatz $x_1, ...x_n$ insgesamt m verschiedene Ausprägungen $u_{(1)} < ... < u_{(m)}$ mit zugehörigen relativen Häufigkeiten $f_{(1)}, ..., f_{(m)}$ vor, so gilt:

$$F_n(x) = \begin{cases} 0, & x < u_{(1)} \\ \sum_{j=1}^{k} f_{(j)}, & u_{(k)} \leq x < u_{(k+1)},\ k \in \{1, ..., m-1\} \\ 1, & x \geq u_{(m)} \end{cases} .$$

Regel 2.3. ***Eigenschaften der empirischen Verteilungsfunktion***
Allgemeine Eigenschaften

(*i*) *Die empirische Verteilungsfunktion ist eine Treppenfunktion.*

(*ii*) *Die Sprunghöhe an der Stelle $u_{(j)}$ ist die relative Häufigkeit $f_{(j)}$.*

(*iii*) *Die empirische Verteilungsfunktion ist monoton wachsend.*

Für reelle Zahlen x, y mit $x < y$ beschreiben

(*iv*) *$F_n(x)$ den Anteil der Beobachtungswerte im Intervall $(-\infty, x]$.*

(*v*) *$1 - F_n(x)$ den Anteil der Beobachtungswerte im Intervall (x, ∞).*

(*vi*) *$F_n(y) - F_n(x)$ den Anteil der Beobachtungswerte im Intervall $(x, y]$.*

3 Klassierte Daten und Histogramm

3.1 Klassieren von Daten

Ziel beim Übergang vom Stab-/Säulendiagramm zu Histogramm:

- Kumulieren von Daten zur Konstruktion einer informativen Grafik

$\Rightarrow$ Abszisse als metrische Achse

$\Rightarrow$ Klassieren von Daten

Definition 3.1. Klassen

- Der Wertebereich $[a, b]$ eines Merkmals wird in $M \in \mathbb{N}$ Klassen $K_1, ..., K_M$ unterteilt.

$\Rightarrow$ Die $K_j = (v_{j-1}, v_j]$ mit $v_0 = a$ und $v_M = b$ bilden eine Zerlegung von $[a, b]$.

$\Rightarrow$ Man spricht bei $n(K_j) = \sum_{i:u_i \in K_j} n_i = n(K_j) = \sum_{i=1}^{n} \mathbf{1}_{K_j}(x_i)$ auch von der absoluten Klassenhäufigkeit (analog relative Klassenhäufigkeit).

$\Rightarrow$ Man nennt $f(K_1), ..., f(K_M)$ auch Häufigkeitsverteilung des Merkmals zur Klasseneinteilung $K_1, ..., K_M$ oder **klassierte Häufigkeitsverteilung**.

- Man nennt $b_j = v_j - v_{j-1}$ auch die Klassenbreite(n).

Möglichkeiten des Klassierens von Daten

- Klassierung **nachträglich** zu Auswertungszwecken, z. B. bei der Erstellung eines Histograms
- Klassierung **während** der Erhebung der Daten (z. B. Gehälter direkt klassiert abfragen)

3.2 Histogramm

Definition 3.2. Ein **Histogramm**...
wird auf folgende Weise konstruiert: Die Fläche des Rechtecks über dem Intervall K_j ist proportional zur relativen Häufigkeit $f(K_j)$, d. h.

$$f(K_j) = b_j \cdot h_j = (v_j - v_{j-1}) \cdot h_j, \quad j = 1, ..., M,$$

wobei h_j die Höhe des Rechtecks bezeichne. h_j berechnet sich dementsprechend gemäß:

$$h_j = \frac{f(K_j)}{b_j}, \quad j = 1, ..., M$$

Das Histogramm als Flächendiagramm

- Das Histogramm ist ein Flächendiagramm.
- Nur bei einer Klassierung in gleich Breite Intervalle ($b_j = b \in \mathbb{R} \; \forall \; j \in \{1, ..., M\}$) ist die Höhe der Histogrammbalken selbst ein Maß für die Häufigkeiten. Es ergibt sich dann aber ggf. ein Histogramm mit anderem Maßstab auf der Ordinate.
- Anteile von Beobachtungen in gleichbreiten (Teil-)Intervallen sind bei einem Histogramm vergleichbar.

Breite und Normierung eines Histogramms

- Histogramme, die wie oben konstruiert werden haben immer eine Flächensumme der Säulen von 1.
- Häufig wird mit einem Faktor $c > 0$ skaliert, d. h. $h_j = c \cdot f(K_j)/b_j$.
- Für $c = n$ kann man offenbar auch $f(K_j)$ durch $n(K_j)$ ersetzen.

Faustregeln bei der Erstellung von Histogrammen

1. Bei n Beobachtungen höchstens $\sqrt{n}$ Klassen, oder...
2. ...bei n Beobachtungen höchstens $10 \cdot \log_{10}(n)$ Klassen.

3.3 (Approximierende) empirische Verteilungsfunktion für klassierte Daten

Problematik

- Ziel: Erstellung einer (approximierenden) empirischen Verteilungsfunktion für klassierte Daten.
- Problem: Wahrscheinlichkeit eines Wertes kleiner oder gleich einer gegebenen Zahl x ist innerhalb einer Klasse nicht (mehr) exakt bekannt.
- Lösung: **Proportionalitätsprinzip** ⇒ Daten innerhalb einer Klasse werden als gleichverteilt betrachtet:

  In diesem Fall ist die Häufigkeit eines Wertes aus einem Intervall $I := (\alpha, \beta] \subset K_j$ nur von der Intervalllänge $L := \beta - \alpha$ abhängig:

$$f(I) = f(K_j) \cdot \frac{L}{v_j - v_{j-1}} = h_j \cdot (\beta - \alpha)$$

⇒ Intervalle gleicher Länge haben in diese Approximation immer die gleiche (relative) Häufigkeit.

Die Verteilungsfunktion wird dann approximiert, indem zusammengesetzt wird:

> Anteil von Beobachtungen, die kleiner oder gleich x sind =
> Anteil von Beobachtungen kleiner oder gleich v_{j-1} +
> + Anteil von Beobachtungen im Intervall $(v_{j-1}, x]$,

wobei auf den zweiten Summand die entsprechende Approximation angewandt wird.

Definition 3.3. Die **(approximierende) empirische Verteilungsfunktion** zur Klasseneinteilung $K_1, ..., K_M$ eines Datensatzes $x_1, ..., x_n$ mit zugehöriger Häufigkeitsverteilung $f(K_1), ..., f(K_M)$ und klassengrenzen $v_0 < v_1 < ... < v_M$ ist definiert durch:

$$F_n^*(x) : \mathbb{R} \to [0,1],\ x \mapsto \begin{cases} 0, & x \leq v_0, \\ \sum_{i=1}^{j-1} f(K_i) + f(K_j)\frac{x - v_{j-1}}{b_j}, & v_{j-1} < x \leq x_j,\ j = 1, ..., M \\ 1, & x > v_M \end{cases}$$

Regel 3.1. ***Eigenschaften einer (approximierenden) empirischen Verteilungsfunktion für klassierte Daten***

(*i*) *F_n^* ist stetig, monoton wachsend und stückweise linear.*

(*ii*) *Die Steigung von F_n^* in K_j ist $f(K_j)/b_j$.*

(*iii*) *Ist H_x die Fläche des entsprechenden Histogramms über $(-\infty, x]$, so gilt:*

$$H_x = F_n^*(x), \quad x \in \mathbb{R}$$

4 Lage- und Streuungsmaße

Die Verwendung von Lage- und Streuungsmaßen hängt entscheidend von der zugrundeliegenden Skala des relevanten Merkmals ab. Wie gehen hier „von unten (nominal) nach oben (stetig) “ vor:

4.1 Modus und Rang

Definition 4.1. Modus (ab Nominalskala)
Jede Ausprägung u_{j^*}, deren absolute, bzw. relative Häufigkeit die Eigenschaft

$$n_{j^*} = \max\{n_1, ...n_m\} \quad \text{bzw.} \quad f_{j^*} = \max\{f_1, ..., f_m\}$$

erfüllt wird als Modus oder Modalwert bezeichnet. Man schreibt: $u_{j^*} =: x_{\text{mod}}$. Der Modus ist also der am häufigsten Auftretende Wert.

Definition 4.2. Rang (ab Ordinalskala)
Sei $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ die Rangwertreihe eines ordinalen Datensatzes.

(*i*) Kommt eine Beobachtungswert x_j genau einmal in der Urliste vor, so heißt dessen Position in der Rangwertreihe **Rang von** x_j. Man schreibt: $\mathrm{R}(x_j)$.

(*ii*) Tritt ein Beobachtungswert x_j s-mal ($s \in \mathbb{N}$) in der Urliste auf, das heißt für die Rangwertreihe gilt:

$$x_{(r-1)} < \underbrace{x_{(r)} = x_{(r+1)} = ... = x_{(r+s-1)}}_{=x_j \ (s\text{-mal})} < x_{(r+1)},$$

so wird mit dem Begriff Rang von x_j das arithmetische Mittel aller Positionen in der Rangwertreihe mir Wert x_j bezeichnet, d. h.

$$\mathrm{R}(x_j) = \frac{r + (r+1) + ... + (r+s-1)}{x} = r + \frac{s-1}{2}.$$

Das mehrfache Auftreten eines Wertes in der Urliste wird als **Bindung** bezeichnet.

4.2 Quantile bis zur Ordinalskala

Der folgende Median ist der Wert der Urliste, für den mindestens 50% aller Beobachtungswerte kleiner oder gleich und mindestens 50% aller Beobachtungswerte größer oder gleich dem Median sind.

Definition 4.3. Median
Sei $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ die Rangwertreihe eines ordinalen Datensatzes. Dann wird der Median $\tilde{x}_{0.5}$ definiert durch:

$$\tilde{x}_{0.5} \begin{cases} = x_{\left(\frac{n+1}{2}\right)}, & n \text{ ungerade}, \\ \in \left\{ x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n+1}{2}\right)} \right\}, & n \text{ gerade} \end{cases}.$$

Der Median wird auf folgende Weise verallgemeinert: Das p-Quantil ist der Wert der Urliste, für den mindestens $p \cdot 100\%$ aller Beobachtungswerte kleiner oder gleich und mindestens $(1-p) \cdot 100\%$ aller Beobachtungswerte größer oder gleich dem p-Quantil sind.

Definition 4.4. p-Quantile
Für $p \in (0,1)$ wird das p-Quantil $\tilde{x}_p$ des Datensatzes $x_1, ..., x_n$ definiert durch:

$$\tilde{x}_p \begin{cases} = x_{(k)}, & \text{falls } np < k < np+1, \quad np \notin \mathbb{N}, \\ \in \left\{ x_{(k)}, x_{(k+1)} \right\}, & \text{falls } k = np \in \mathbb{N} \end{cases}.$$

Zusätzliche Bezeichnungen

- Für $p = 0.25$ nennt man $\tilde{x}_p$ das untere Quartil.
- Für $p = 0.75$ nennt man $\tilde{x}_p$ das obere Quartil.
- Für $p = k/10$ nennt man $\tilde{x}_p$ das k-te Dezentil ($k = 1, ..., 9$).
- Für $p = k/100$ nennt man $\tilde{x}_p$ das k-te Perzentil ($k = 1, ..., 99$).

4.3 Quantile für metrische Skalen

Für metrische Daten werden Quantile anders definiert:

Definition 4.5. Median und Qantile
Sei $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ die Rangwertreihe eines metrischen Datensatzes.

(i) Der Median ist definiert durch:

$$\tilde{x}_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ ungerade}, \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right), & n \text{ gerade} \end{cases}.$$

(*ii*) Für $p \in (0,1)$ ist das p-Quantil $\tilde{x}_p$ gegeben durch

$$\tilde{x}_p = \begin{cases} x_{(k)}, & \text{falls } np < k < np+1, \quad np \notin \mathbb{N}, \\ \frac{1}{2}\left\{x_{(k)} + x_{(k+1)}\right\}, & \text{falls } k = np \in \mathbb{N} \end{cases} .$$

Die „wörtliche“ Definition von oben, würde jeder Wert aus dem Intervall $[x_{(n/2)}, x_{(n/2+1)}]$ (für Median), bzw. $[x_{(k)}, x_{(k+1)}]$ für $k = np \in \mathbb{N}$ (für Quantile) erfüllen. Hier wird die Intervallmitte gewählt. Alternativ wird manchmal die linke Intervallgrenze gewählt:

$$\tilde{x}_p = \min\{x | F_n(x) \geq p\} .$$

4.4 Eigenschaften von Median und arithmetischem Mittel

Definition 4.6. Arithmetisches Mittel
Sei $x_1, ..., x_n$ ein Datensatz aus Beobachtungswerten eines metrischen Merkmals. Dann ist das arithmetische Mittel $\bar{x}_n$ definiert durch

$$\bar{x} := \bar{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{m} f_j u_j .$$

Regel 4.1. *Minimalität des Medians*
Für $t \in R$ sei die Funktiong definiert durch

$$g(t) = \sum_{i=1}^{n} |x_i - t| .$$

Dann nimmt die Funktion g ihr Minimum an für den Median, d. h. für $t = \tilde{x}_{0.5}$:

$$\boxed{g(t) \geq g(\tilde{x}_{0.5}) \quad \text{für alle } t \in \mathbb{R} .}$$

Regel 4.2. *Minimalität des arithmetischen Mittels*
Für $t \in R$ sei die Funktion f definiert durch

$$f(t) = \sum_{i=1}^{n} (x_i - t)^2 .$$

Dann nimmt die Funktion g ihr Minimum an für den Median, d. h. für $t = \bar{x}$:

$$\boxed{f(t) \geq f(\bar{x}) \quad \text{für alle } t \in \mathbb{R} .}$$

Definition 4.7. Gewichtetes arithmetisches Mittel
Seien $x_1, ..., x_n \in \mathbb{R}$ ein metrischer Datensatz und $g_1, ..., g_n \geq 0$ mit $sum_{i=1}^{n} g_i = 1$. Das bezüglich $g_1, ..., g_n$ gewichtete arithmetische Mittel $\bar{x}_g$ von $x_1, ..., x_n$ ist definiert

durch

$$\boxed{\bar{x}_g = \sum_{i=1}^{n} g_i x_i \,.}$$

Regel 4.3. ***Zusammengesetzte arithmetische Mittel***
Gegeben seien zwei Datensätze x_i $(i = 1, ..., n_1)$ und y_j $(j = 1, ..., n_2)$ mit arithmetischen Mitteln $\bar{x}$ und $\bar{y}$. Das arithmetische Mittel $\bar{z}$ aller $n_1 + n_2$ Beobachtungswerte des zusammengesetzten (gepoolten Datensatzes) $z_1 = x_1, ..., z_{n_1} = x_{n_1}, z_{n_1+1} = y_1, ..., z_{n_1+n_2} = y_{n_2}$ lässt sich bestimmen als

$$\bar{z} = \frac{n_1}{n_1 + n_2}\bar{x} + \frac{n_2}{n_1 + n_2}\bar{y} \,.$$

Mithilfe dieser Regel kann man auch einzelne Werte anfügen ($n_2 = 1$).

4.5 Streuungsmaße

Definition 4.8. Spannweite und Quartilsabstand
Für einen metrischen Datensatz $x_1, ..., x_n$ ist die Spannweite R definiert als:

$$R = x_{(n)} - x_{(1)} \,.$$

Der Quartilsabstand Q ist definiert als

$$Q = \tilde{x}_{0.75} - \bar{x}_{0.25} \,.$$

Natürlich gilt immer $Q \leq R$.

Definition 4.9. Empirische Varianz
Für $x_1, ..., x_n$ heißt

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

empirische Varianz s_n^2 von $x_1, ..., x_n$ (kurz: s^2). Dabei wird $s := s_n = \sqrt{s_n^2}$ als **empirische Standardabweichung** bezeichnet.

Regel 4.4. ***Steiner-Regel, Verschiebungssatz***
Für $a \in \mathbb{R}$ und einen Datensatz $x_1, ..., x_n$ gilt:

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

bzw. für $a = 0$:

$$\frac{1}{n}\sum_{i=1}^{n} s^2 + \bar{x}^2 \quad \text{oder} \quad s^2 = \bar{x^2} - \bar{x}^2 \,.$$

Regel 4.5. ***Zusammengesetzte empirische Varianz***
Gegeben seien zwei Datensätze x_i $(i = 1, ..., n_1)$ und y_j $(j = 1, ..., n_2)$ mit empirischen Varianzen s_x^2 und s_y^2. Die empirische s_z^2 aller $n_1 + n_2$ Beobachtungswerte des zusammengesetzten (gepoolten Datensatzes) $z_1 = x_1, ..., z_{n_1} = x_{n_1}, z_{n_1+1} = y_1, ..., z_{n_1+n_2} = y_{n_2}$ lässt sich bestimmen als

$$s_z^2 = \frac{n_1}{n_1+n_2}s_x^2 + \frac{n_2}{n_1+n_2}s_y^2 + \frac{n_1}{n_1+n_2}(\bar{x} - \bar{z})^2 + \frac{n_2}{n_1+n_2}(\bar{y} - \bar{z})^2 \,.$$

Definition 4.10. Mittlere absolute Abweichung
Für $x_1, ..., x_n$ heißt

$$d = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}_{0.5}|$$

mittlere absolute Abweichung d von $x_1, ..., x_n$.

Regel 4.6. *Für einen Datensatz gilt immer:*

$$d \leq s \,.$$

Regel 4.7. ***Eigenschaften von Streuungsmaßen: Häufigkeiten***
Gegeben sei die Häufigkeitsverteilung $f_1, ..., f_m$ mit zugehörigen Ausprägungen $u_1, ..., u_m$.

(i) Spannweite:

$$R = \max_{j\in J}\{u_j\} - \min_{j\in J}\{u_j\} \,, \quad \text{mit} \quad J = \{i \in \{1, ..., m\} | f_i > 0\} \,.$$

(ii) Varianz:

$$s^2 = \sum_{j=1}^{m} f_j(u_j - \bar{x})^2 \,.$$

(iii) Mittlere absolute Abweichung:

$$d = \sum_{j=1}^{m} f_j \, |u_j - \tilde{x}_{0.5}| \,.$$

4.6 Transformation von Daten

Regel 4.8. ***Linear transformierter Datensatz***
Seien $a, b \in \mathbb{R}$ und $y_1, ..., y_n$ ein linear transformierter Datensatz von $x_1, ..., x_n$, d. h.:

$$y_i = a + b\,x_i, \quad i = 1, ..., n\,.$$

Dann gilt:

(*i*) $\tilde{y} = a + b\,\tilde{x}$

(*ii*) $\bar{y} = a + b\,\bar{x}$

(*iii*) $s_y^2 = b^2 s_x^2$

(*vi*) $d_y = |b|\ d_x$.

Definition 4.11. Zentrierung
Für Beobachtungswerte $x_1, ..., x_n \in \mathbb{R}$ eines metrischen Merkmals heißt die lineare Transformation $y_i = x_i - \bar{x},\ i = 1, ..., n$ Zentrierung. Die transformierten Daten werden als zentriert (oder als **Residuen**) bezeichnet.

Definition 4.12. Standardisierung
Für $x_1, ...x_n$ mit $s_x > 0$ heißt die lineare Transformation

$$z_i = \frac{x_i - \bar{x}}{s_x}, \quad i = 1, ..., n$$

Standardisierung. Die $z_1, ..., z_n$ werden als standardisiert bezeichnet.

Regel 4.9. *Sind die $z_1, ..., z_n$ standardisiert, so gilt:*

$$\bar{z} = 0, \qquad s_z = s_z^2 = 1\,.$$

5 Zusammenhangsmessung

Ausgangssituation

- Gemessen mehrere Merkmale eines Objekts, d. h. Paare von Merkmalen X und Y.
- Ein Paar (X, Y) heißt auch bivariates Merkmal mit den Komponenten X und Y.

- Im Folgenden werden verschiedene Zusammenhangsmaße für den Fall X und Y beide nominal und X und Y beide metrisch betrachtet.
- Es werden anstatt der Urliste hier meist direkt die Merkmalsausprägungen betrachtet und bereits mit x_i usw. bezeichnet.

5.1 Kontingenztafeln

Unter einer Kontingenztafel versteht man eine Tabelle, die die gemeinsamen Häufigkeiten der Merkmalsausprägungen beinhaltet. Spalten- und Zeilensummen liefern dann die Häufigkeiten für die jeweilige Ausprägung unabhängig vom jeweils anderen Merkmal. Selbiges gilt für die relativen Häufigkeiten.

In dieser Schreibweise ist

n_{ij} die absolute Häufigkeit des Tupels (x_i, y_j) in der Urliste

und

$f_{ij} = \frac{n_{ij}}{n}$ die zugehörige relative Häufigkeit.

Das Konzept lässt sich entsprechend durch Erhöhung der Anzahl an Indizes verallgemeinern. Allerdings wird die tabellarische Darstellung dann schwieriger bzw. umständlicher.

5.2 Bedingte Häufigkeiten

Definition 5.1. Bedingte Häufigkeit
Sei $n_{\bullet j} > 0$. Der Quotient

$$f_{X=x_i|Y=y_j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}$$

heißt bedingte Häufigkeit (von $X = x_i$ unter der Bedingung $Y = y_i$). Die zugehörige Häufigkeitsverteilung

$$f_{X=x_1|Y=y_j}, \ \cdots, f_{X=x_p|Y=y_j}$$

wird als bedingte Häufigkeitsverteilung (von X unter der Bedingung $Y = y_j$ bezeichnet.

Selbiges geht natürlich umgekehrt.

Regel 5.1. *Für die bedingte Häufigkeitsverteilung eines Datensatzes* (x_i, y_j), $i =$

$1, ..., p,\ j = 1, ..., q$ *gilt*

$$\sum_{i=1}^{p} f_{X=x_i|Y=y_j} = \sum_{j=1}^{q} f_{Y=y_j|X=x_1} = 1\,.$$

5.3 χ^2-Größe: Quantifizierung eines Zusammenhangs

Definition 5.2. χ^2-Größe
Bei positiven Randhäufigkeiten $n_{i\bullet}$, $n_{\bullet j}$ wird die χ^2-Größe χ^2 definiert durch:

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - v_{ij})^2}{v_{ij}}, \quad \text{mit} \quad v_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n},$$

für $i = 1, ..., p$, $j = 1, ..., q$.

Offensichtlich gilt $\chi^2 \geq 0$.

Definition 5.3. Empirische Unabhängigkeit
Merkmale X und Y heißen empirisch unabhängig, wenn gilt:

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}$$

für alle $i = 1, ..., p$ und für alle $j = 1, ..., q$.

Regel 5.2. *Die empirische Unabhängigkeit von X und Y ist äquivalent zu*

$$f_{ij} = f_{i\bullet} f_{\bullet j}$$

für alle $i = 1, ..., p$ und für alle $j = 1, ..., q$. Daraus folgt:

$$f_{X=x_i|Y=y_j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n} = f_{i\bullet}$$

und vice versa.

Inhaltlich bedeutet dies:

Bei empirischer Unabhängigkeit ist die bedingte Häufigkeit von x_i und y_j gleich der relativen Häufigkeit der x_i im Datensatz.

Mit der Eigenschaft

$$v_{ij} = n f_{i\bullet} f_{\bullet j}$$

der χ^2-Größe lässt sich interpretieren:

Die χ^2-Größe vergleicht die Kontingenztafel mit der Kontingenztafel bei empirischer Unabhängigkeit.

Hier wird auch klar, wieso im Falle $n_{i\bullet} = 0$ für ein i oder $n_{\bullet j}$ für ein j die alternative χ^2-Größe lautet:

$$\chi^2 = \sum_{i,j:\ v_{ij}>0} \frac{(n_{ij} - v_{ij})^2}{v_{ij}},$$

denn dann haben die Tafeln n_{ij} und v_{ij} in der betreffenden Zeile oder Spalte beide eine Nullzeile oder -spalte.

Regel 5.3. *Für die χ^2-Größe gilt:*

$$\boxed{\chi^2 = 0 \quad \Leftrightarrow \quad X \text{ und } Y \text{ sind empirisch unabhängig.}}$$

Regel 5.4. *Für die χ^2-Größe gilt:*

$$\boxed{\chi^2 = n\left(\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}}\right) - n}$$

Assoziationsmaße für die 2×2-Kontingenztafel (Vierfeldertafel)

Für $p = q = 2$ gilt:

$$\boxed{\chi^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}$$

Aufgrund des vergleichsweise einfachen Nenners betrachtet man für Vierfeldertafeln häufig:

Definition 5.4. Assoziationskoeffizient von Yule
Für ein bivariates Merkmal mit $p = q = 2$ (Bezeichnungen von oben) ist der Assoziationskoeffizient von Yule:

$$\boxed{A = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} \in [-1, 1]}$$

Regel 5.5. *Für den Assoziationskoeffizient von Yule gilt:*

(*i*) *Mit dem Kreuzproduktverhältnis* $c = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ *gilt* $A = \frac{c-1}{c+1}$.

(*ii*) *empirische Unabhängigkeit* $\Leftrightarrow A = 0$

(*iii*) *Hat ein Eintrag der Vierfeldertafel den Wert Null, so gilt* $|A| = 1$.

5.4 Kontingenzkoeffizient nach Pearson

Regel 5.6. ***Obere Schranke und Vollständige Abhängigkeit***

(*i*) *Es gibt eine obere Schranke für die* χ^2*-Größe:*

$$\chi^2 \leq n \cdot \min\{p-1, q-1\}$$

(*ii*) *Für die* χ^2*-Größe gilt* $\chi^2 = n \cdot \min\{p-1, q-1\}$ *genau dann, wenn eine der folgenden Bedingungen erfüllt ist:*

1. $p < q$ *und in jeder Spalte sind die Häufigkeiten in genau einem Feld konzentriert.*
2. $p = q$ *und in jeder Zeile und in jeder Spalte sind die Häufigkeiten in genau einem Feld konzentriert.*
3. $p > q$ *und in jeder Zeile sind die Häufigkeiten in genau einem Feld konzentriert.*

Da die χ^2-Größe für feste Dimensionen offenbar unbeschränkt ist, konstruiert man zur Zusammenhangsmessung weitere, beschränkte Größen.

Definition 5.5. Kontingenzkoeffizient nach Pearson
Der Kontingenzkoeffizient C nach Pearson ist definiert durch

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Regel 5.7. ***Eigenschaften des Kontingenzkoeffizient nach Pearson***
Für den Kontingenzkoeffizient nach Pearson gilt

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}} \quad \text{mit} \quad \phi^2 = \frac{\chi^2}{n} = \sum_{i,j} \frac{(f_{ij} - f_{i\bullet} f_{j\bullet})^2}{f_{i\bullet} f_{j\bullet}}$$

und

$$0 \leq C \leq \sqrt{\frac{\min\{p-1, q-1\}}{\min\{p, q\}}} < 1$$

Definition 5.6. Korrigierter Kontingenzkoeffizient nach Pearson
Der korrigierte Kontingenzkoeffizient C_* nach Pearson ist definiert durch:

$$C_* = C \cdot \sqrt{\frac{\min\{p, q\}}{\min\{p, q\} - 1}}.$$

Regel 5.8. *Für den korrigierten Kontingenzkoeffizient nach Pearson C_* gilt $0 \leq C_* \leq 1$. Außerdem gilt:*

(*i*) $C_* = 0 \Leftrightarrow X$ *und* Y *empirisch unabhängig.*

(*ii*) $C_* = 1 \Leftrightarrow$ *eine der folgenden Bedingungen für die zugehörige Kontingenztafel ist erfüllt:*

1. $p < q$ *und in jeder Spalte sind die Häufigkeiten in genau einem Feld konzentriert.*
2. $p = q$ *und in jeder Zeile und in jeder Spalte sind die Häufigkeiten in genau einem Feld konzentriert.*
3. $p > q$ *und in jeder Zeile sind die Häufigkeiten in genau einem Feld konzentriert.*

Interpretation der Assoziationsmaße

- Assoziationsmaße liefern lediglich *Anhaltspunkte* für die Stärke eines Zusammenhangs.
- Aussagen über ein explizites Änderungsverhalten der Merkmale untereinander sind nicht möglich.

$\Rightarrow$ Kausale Zusammenhänge sind so nicht nachweisbar!

5.5 Zusammenhangsmessung bei metrischen Merkmalen

Ziel ist es für bivariate metrische Merkmale sinnvolle(re) Darstellungsmöglichkeiten und Zusammenhangsmaße zu entwickeln.

Streudiagramm

Eine einfache mögliche Darstellung ist das Streudiagramm, in dem einfach alle Beobachtungspaare $(x_1, y_1), ..., (x_n, y_n)$ als Punkte eingetragen werden. Bei einer auffälligen

Systematik in diesem Plot lässt sich ein Zusammenhang vermuten. Das weitere Ziel ist, diesen Zusammenhang auch zu quantifizieren.

Wichtig: Es wird keinerlei Aussage getroffen, welcher Art dieser Zusammenhang ist (z.B. kausal oder wechselseitig).

Definition 5.7. Empirische Kovarianz
Seien $(x_1, y_1), ..., (x_n, y_n)$ Messwerte eines bivariaten, quantitativen Merkmals (X, Y). Dann heißt

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

(empirische) Kovarianz der Merkmale X und Y.

Es gilt offenbar

$$s_{xx} = s_x^2 \,.$$

Regel 5.9. *Seien $(x_1, y_1), ..., (x_n, y_n)$ Messwerte eines bivariaten, quantitativen Merkmals (X, Y) mit Kovarianz x_{xy}. Die gemäß*

$$x_i^* = a + b\, x_i \quad \text{und} \quad y_i^* = c + d\, y_i$$

mit $a, b, c, d \in \mathbb{R}$ linear transformierten Daten $(x_1^*, y_1^*), ..., (x_n^*, y_n^*)$ haben die Kovarianz

$$s_{x^*y^*} = bd\, s_{xy} \,.$$

Regel 5.10. *Für die empirische Kovarianz s_{xy} gilt:*

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y} = \bar{xy} - \bar{x} \cdot \bar{y} \,.$$

Definition 5.8. Bravais-Pearson-Korrelation
Seien $(x_1, y_1), ..., (x_n, y_n)$ Messwerte eines bivariaten, quantitativen Merkmals (X, Y) und $s_x, s_y > 0$. Dann ist der Bravais-Pearson-Korrelationskoeffizient definiert durch:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \,.$$

Transformiert man die Daten, wie in 5.9, so folgt:

$$r_{x^*y^*} = \frac{bc}{|bc|} r_{xy}\,.$$

Regel 5.11. ***Eigenschaften der Bravais-Pearson-Korrelation***

(*i*) *r_{xy} ist symmetrisch in X und Y, d. h. $r_{xy} = r_{yx}$.*

(*ii*) *Es gilt (Cauchy-Schwarz):*

$$\boxed{-1 \leq r_{xy} \leq 1\,.}$$

(*iii*) *Falls $r_{xy} = 1$ $(= -1)$ gilt: Es existiert eine positive (negative) Zahl b und eine reelle Zahl a mit*

$$y_i = a + b\,x_i$$

Es gibt also einen linearen Zusammenhang zwischen X und Y derart, dass

> Wenn X um eine Einheit steigt, dann steigt (fällt) Y ebenfalls und zwar um b ($|b|$) Einheiten.

*Dieser Ordnungsverhalten nennt man **gleichsinnig (gegensinnig)**.*

Definition 5.9. Bezeichnungen für die Korrelation

a) positiv korreliert, falls $r_{xy} > 0$

b) unkorreliert, falls $r_{xy} = 0$

c) negativ korreliert, falls $r_{xy} < 0$

d) schwach korreliert, falls $0 \leq |r_{xy}| < 0.5$

e) stark korreliert, falls $0.8 \leq |r_{xy}| \leq 1$

Wichtig: Beispiele, wie der parabelförmige Zusammenhang (im Scatterplot) zeigen, dass $r_{xy} = 0$ bzw. sehr kleine Korrelationen *nicht* bedeuten, dass es *keinen Zusammenhang* zwischen den Merkmalen X und Y gibt. $r_{xy} = 0$ besagt lediglich, dass es *keinen linearen Zusammenhang* zwischen X und Y gibt.

Anmerkungen zur Korrelation

- Mit r_{xy} sind aufgrund der Symmetrie keine kausalen Zusammenhänge nachweisbar. Die „Richtung des Zusammenhangs“ kann nur auf Basis des Sachkontextes ermittelt werden.

- Als **Scheinkorrelationen** bezeichnet man Korrelationen zwischen zwei Merkmalen X und Y die durch eine dritte Variable Z induziert wird (Schuhgröße, Körpergewicht, Körpergröße).
- Es gibt z. T. auch „**unsinnige Korrelationen**“, die durch Fluktuation oder eine mangelhafte Datenlage erzeugt werden (Störche, Geburtenrate).

6 Regressionsanalyse

6.1 Hinführung

Idee

- Merkmal Y wird als Funktion des Merkmals X aufgefasst:

$$Y = f(X) \quad \text{mit einer Funktion } f : \mathbb{R} \to \mathbb{R}\,.$$

- Oft wird f zumindest teilweise unbekannt sein oder von unbekannten Parametern abhängen.
- Ein solcher Zusammenhang wird im Folgenden **unterstellt.**

Ziel

Mittels eines Datensatzes $(x_1, y_1), ..., (x_n, y_n)$ Aussagen über die Funktion f zu treffen.

Definition 6.1. Regressionsbegriffe

- Merkmal X heißt **Regressor** oder **erklärende Variable**,
- Y wird als **Regressand** bzw. als **abhängige Variable** bezeichnet.
- f heißt **Regressionsfunktion**.
- Die Werte $\widehat{y}_i = f(x_i)$, $i = 1, ..., n$, heißen **Regressionswerte**.

Anmerkungen

- I. Allg. gilt: $f(x_i) = \widehat{y}_i \neq y_i$.
- Dies kann durch **(natürliche) Schwankungen** in den Eigenschaften der Objekte oder durch **Messfehler** und Messungenauigkeiten hervorgerufen sein.

$\Rightarrow$ **Regressionsmodell**

$$\boxed{Y = f(X) + \varepsilon}$$

mit dem **Fehlerterm** ε, der alle möglichen Fehlerarten repräsentiert.

- Auf Datenebene:

$$\varepsilon_i = y_i - f(x_i) \quad \text{Fehler der } i\text{-ten Messung}$$
$$y_i = f(x_i) + \varepsilon_i,\ i \in \{1, ..., n\}\,.$$

Problemstellung:

Passe die Regressionsfunktion f in einer **Klasse $\mathcal{H}$ von Funktionen** möglichst gut an die vorliegenden Daten an.

Ein Beispiel für eine Klasse von Funktionen sind z. B. die Polynome vom Grad p:

$$\mathcal{H} = \left\{ f_{a_0,\ldots,a_p} | f_{a_0,\ldots,a_p}(x) = \sum_{k=0}^{p} a_p x^p,\ x \in \mathbb{R},\ \mathbf{a} \in \mathbb{R}^p \right\}$$

6.2 Methode der kleinsten Quadrate

Idee:

Suche Funktion $\widehat{f} \in \mathcal{H}$, so dass

$$Q(f) = \sum_{i=1}^{n} \Big(y_i - f(x_i) \Big)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

minimal ist.

Parametrische Klasse $\mathcal{H}$

- Für parametrische Funktionen $f_{b_1,\ldots,b_j}$ aus einer Klasse $\mathcal{H}$, die durch j Parameter $b_1, \ldots, b_j \in \mathbb{R}$ $(j \in \mathbb{N})$ beschrieben wird, hat Q die Form:

$$Q(b_1, \ldots, b_j) = \sum_{i=1}^{n} \Big(y_i - f_{b_1,\ldots,b_j}(x_i) \Big)^2 .$$

- Suche also das Tupel $\widehat{\mathbf{b}} := (\widehat{b}_1, \ldots, \widehat{b}_j) \in \mathbb{R}^j$ mit:

$$Q(\widehat{\mathbf{b}}) \leq Q(\mathbf{b}) \quad \text{für alle } \mathbf{b} \in \mathbb{R}^j .$$

6.3 Modell der linearen Einfachregression

Definition 6.2. Modell der linearen Einfachregression
Ist f eine lineare Funktion, $f(x) = a + b\,x,\ x \in \mathbb{R}$, so heißt

$$Y = a + b\,X \quad (+\varepsilon)$$

Modell der linearen Einfachregression. Die Regressionsfunktion f heißt auch **Regressionsgerade**.

Die Quadratsumme lautet hier dementsprechend:

$$Q(a,b) = \sum_{i=1}^{n} \Big(y_i - (a + b\,x_i)\Big)^2 .$$

Die Minimierung dieser Funktion der Parameter a und b ist mithilfe von partiellen Ableitungen analytisch möglich und führt zurück auf bekannte Größen:

Regel 6.1. ***Lösung der linearen Einfachregression***[a]

- *Für $s_x^2 > 0$ gilt für die Lösung einer linearen Einfachregression:*

$$\boxed{\widehat{b} = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad \widehat{a} = \bar{y} - \widehat{b} \cdot \bar{x}\,.}$$

- *Die minimale Quadratsumme lautet dann:*

$$Q(\widehat{a}, \widehat{b}) = n s_y^2 (1 - r_{xy}^2)\,.$$

[a] Burkschat, Cramer, Kamps 2012, S. 304-305

Im Fall $s_x^2 = 0$ gilt $x_i = \bar{x}$ für alle $i = 1, ..., n$. Das heißt, die Werte liegen in einem x-y-Streudiagramm auf einer Senkrechten. Die Methode der kleinsten Quadrate liefert dann, dass jede lineare Funktion durch $(\bar{x}, \bar{y})$, d. h. $g_b(x) = \bar{y} + b(x - \bar{x})$ für $b \in \mathbb{R}$ beliebig optimal ist.

Regel 6.2. ***Eigenschaften der Regressionsgerade***[a]

(*i*) *Es gilt:*

$$Q(\widehat{a}, \widehat{b}) \leq Q(a,b) \quad \text{für alle } a, b \in \mathbb{R}$$

Für $s_x^2 > 0$ ist diese ***Lösung eindeutig.***

(*ii*) *Gilt $s_y^2 > 0$, so gibt r_{xy} das Vorzeichen der Steigung von $\widehat{f}$ vor.*

(*iii*) *Es gilt: $\widehat{f}(\bar{x}) = \widehat{a} + w\widehat{b}\,\bar{x} = \bar{y}$.*

(*iv*) *Es gilt:*

$$\bar{\widehat{y}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{y}_i = \bar{y}\,.$$

(*v*) *Es gilt:*

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0\,.$$

[a] Burkschat, Cramer, Kamps 2012, S. 306-307

Regel 6.3. ***Lineare Transformation***
Seien $s_x^2 > 0$ und $\widehat{f}(x)0\widehat{a} + \widehat{b}\,x$, $x \in \mathbb{R}$ die zu $(x_1, y_1), ..., (x_n, y_n)$ gehörige Regressionsgerade.
Werden die Beobachtungswerte mit $\beta \neq 0, \delta \neq 0, \alpha, \gamma \in \mathbb{R}$, (linear) transformiert gemäß

$$u_i = \beta\, x_i + \alpha\,, \quad v_i = \delta\, y_i + \gamma\,, \quad i \in \{1, ..., n\},$$

so gilt für die Koeffizienten der zu den Daten $(u_1, v_1), ..., (u_n, v_n)$ gehörenden Regressionsgerade $\widehat{g}(u) = \widehat{c} + \widehat{d}\,u,\ u \in \mathbb{R}$:

$$\boxed{\widehat{c} = \delta\,\widehat{a} + \gamma - \frac{\alpha\delta}{\beta}\,\widehat{b}\,, \quad \widehat{d} = \frac{\delta}{\beta}\,\widehat{b}\,.}$$

6.4 Bewertung der Anpassung im Modell $Y = a + b\,X + \varepsilon$

Ziel: Messung der Abweichung von der Geraden

Definition 6.3. Residuen
Die Differenzen $\widehat{e}_i = y_i - \widehat{y}_i$ werden als Residuen bezeichnet, wobei $\widehat{y}_i = \widehat{f}(x_i)$.
Für $\sum \widehat{e}_i^2 > 0$ sind die normierten Residuen definiert durch:

$$\widehat{d}_i = \frac{\widehat{e}_i}{\sqrt{\sum_{i=1}^n \widehat{e}_i^2}}\,.$$

Regel 6.4. ***Eigenschaften der Residuen***

- *Für die normierten Residuen gilt:*

$$-1 \leq \widehat{d}_i \leq 1, \quad i \in \{1, ..., n\}$$
$$\sum_{i=1}^n \widehat{d}_i = 0\,, \quad \sum_{i=1}^n \widehat{d}_i^2 = 1\,.$$

- *Es gilt die Äquivalenz:*

$$\sum_{i=1}^n \widehat{e}_i^2 = 0 \quad \Leftarrow \quad y_i = \widehat{y}_i \ \text{ für alle } i \in \{1, ..., n\}\,.$$

Regel 6.5. ***Streuungszerlegung***[a]

Es gilt:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

bzw.

$$\boxed{s_y^2 = s_{\widehat{y}}^2 + s_{\widehat{e}}^2,}$$

interpretiert als:

Gesamtstreuung = durch Regression erklärte Streuung + Reststreuung

[a] Burkschat, Cramer, Kamps 2012, S. 324-325

Definition 6.4. Bestimmtheitsmaß
Die Güte der Approximation der Regressionsfunktion an die Daten wird gemessen durch das Bestimmtheitsmaß

$$B_{xy} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{s_{\widehat{e}}^2}{s_y^2} = \frac{s_{\widehat{y}}^2}{s_y^2} = r_{xy}^2 \,.$$

Bei linearer Transformation bleibt das Bestimmtheitsmaß erhalten.

Residualplots

Eine weitere häufig genutzte Methode zur Einschätzung der Güte einer linearen Regression ist die Betrachtung der Residualplots. In diesen Plots werden die x_i gegen die $\widehat{d}_i$ oder die $\widehat{e}_i$ aufgetragen. Dabei lassen sich dann leicht Ausreißer und Systematiken identifizieren.

6.5 Exkurs: Nicht-lineare Regressionsfunktion

Die Methode der kleinsten Quadrate lässt sich prinzipiell auf alle (viele) Funktionenklassen anwenden. Allerdings ist es meist nur numerisch möglich die Quadratsumme im Parameterraum zu minimieren.

Transformation auf lineare Zusammenhänge: Y-Komponente

Indem man aber z. B. bei einem Expomentiellen Zusammenhang $Y = a \cdot e^{bX} + \varepsilon$ das Merkmal Y transformiert:

$$\tilde{Y} = \ln\left(a \cdot e^{b \cdot X}\right) = \ln(a) + b\,X = \tilde{a} + b\,X \,.$$

Transformation auf lineare Zusammenhänge: X-Komponente

Ein weiteres häufig genutztes Modell ist

$$Y = a + b\, g(X) + \varepsilon$$

mit einer **bekannten** Funktion g. Dies führt letztlich auf eine Lineare Regression mit dem Datensatz $(g(x_1), y_1), ..., (g(x_n), y_n)$.

6.6 Multiples Regressionsmodell

- **Mehrere** erklärende Variablen $X_1, ..., X_m$ und **eine** abhängige Variable Y
- **Regressionsfunktion**

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$$

- Schätzungen beruhen auf Beobachtungen des Merkmalsvektors $(Y, X_1, ..., X_m)$: $(y_1, x_{11}, ..., x_{1m}), ...,$

$\Rightarrow$ Darstellung mittels Matritzen und Vektoren:

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{=\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}}_{=\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{=\varepsilon}$$

Regel 6.6. *Gilt $n \geq m+1$ und hat $\mathbf{X}$ maximalen Rang, d. h. $\operatorname{rg}(\mathbf{X}) = m+1 = p$, so gilt:*

$$\boxed{\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\, \mathbf{X}'\mathbf{Y}\,.}$$

Regel 6.7. ***Quadratisches Regressionsmodell***
Ein quadratisches Regressionsmodell $Y = a + bX + cX^2 + \varepsilon$ lässt sich mit $X_1 = X$ und $X_2 = X^2$ als multiples Regressionsmodell betrachten. Dann gilt:

$$\begin{pmatrix} \widehat{a} \\ \widehat{b} \\ \widehat{c} \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\, \mathbf{X}' \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Allgemeines Regessionsmodell

Voraussetzungen

- $x_1, ..., x_n \in \mathbb{R}^p$
- $f_j : \mathbb{R}^p \to \mathbb{R},\ 0 \leq j \leq m$ bekannte Funktionen
- $\beta_0, ..., \beta_m \in \mathbb{R}$ unbekannte Parameter

Regressionsmodell

$$Y = \sum_{j=0}^{m} \beta_j f_j(x) + \varepsilon \quad \Rightarrow \quad \mathbf{Y} = \mathbf{X}\,\beta + \varepsilon$$

$\Rightarrow$ Lösung wie multiples Regressionsmodell.