

Mathematik und Statistik

Kapitel 3: Wahrscheinlichkeitsrechnung/Statistik

Prof. Dr. Jan Gertheiss
Fakultät für Agrarwissenschaften
Georg-August-Universität Göttingen

Wintersemester 2013/2014

Gliederung

1. Deskriptive/Univariate Statistik
 - ▶ Arten von Merkmalen
 - ▶ Graphische Beschreibung von Daten
 - ▶ Einfache Kenngrößen (Mittelwert, Varianz, etc.)
2. Kombinatorik und elementare Wahrscheinlichkeitsrechnung
 - ▶ Kombinatorik und Laplace-Wahrscheinlichkeiten
 - ▶ Binomialverteilung
 - ▶ Diskrete und stetige Zufallsgrößen
 - ▶ Normalverteilung
 - ▶ (Un)abhängige Zufallsgrößen
3. Zusammenhangsanalyse und Regression
 - ▶ Statistik in Kontingenztabellen
 - ▶ Korrelationen
 - ▶ Mittelwerts-Vergleiche, t-Test, Tests auf Unabhängigkeit
 - ▶ Lineare Regression, Varianzanalyse
 - ▶ Ausblick auf weitere Regressionsmodelle

Statistische Einheiten, Merkmale, Gesamtheiten

- ▶ *Statistische Einheiten*: Objekte (oder Subjekte), an denen interessierende Größen erfasst werden.
- ▶ *Grundgesamtheit*: Menge aller für die konkrete Fragestellung relevanten statistischen Einheiten.
- ▶ *Teilgesamtheit*: Teilmenge der Grundgesamtheit.
- ▶ *Stichprobe*: Tatsächlich untersuchte Teilmenge der Grundgesamtheit.
- ▶ *Merkmal*: Interessierende Größe, *Variable*.
- ▶ *Merkmalsausprägung*: Konkreter Wert des Merkmals für eine statistische Einheit.

Merkmaltypen

Einteilung nach der Anzahl möglicher Ausprägungen:

- ▶ *Diskrete* Merkmale: Endlich (oder abzählbar unendlich) viele Ausprägungen.
- ▶ *Stetige* Merkmale: Alle Werte eines Intervalls sind mögliche Ausprägungen.

Eine weitere Einteilung erfolgt anhand des sog. *Skalenniveaus*:

- ▶ *Nominal-skalierte* Merkmale: Ausprägungen sind lediglich Namen, keine Ordnung möglich.
- ▶ *Ordinal-skalierte* Merkmale: Ausprägungen können geordnet werden, aber Abstände nicht interpretiert werden.
- ▶ *Intervall-skalierte* Merkmale: Ausprägungen sind Zahlen, Interpretation der Abstände möglich, aber kein fester Nullpunkt.
- ▶ *Verhältnis-skalierte* Merkmale: Ausprägungen besitzen sinnvollen absoluten Nullpunkt.

Mehr zu Skalenniveaus

Skalenniveau	sinnvoll interpretierbare Berechnungen			
	auszählen	ordnen	Differenzen bilden	Quotienten bilden
nominal	ja	nein	nein	nein
ordinal	ja	ja	nein	nein
intervall	ja	ja	ja	nein
verhältnis	ja	ja	ja	ja

Eine Zweiteilung ergibt sich als:

- ▶ *qualitative/kategoriale* Merkmale: Endlich viele Ausprägungen, höchstens Ordinalskala.
- ▶ *quantitative/metrische* Merkmale: Ausprägungen geben eine Intensität wieder, Intervall- oder Verhältnisskala.

Univariate Beschreibung von Daten

Häufigkeiten

Wir gehen von einer Erhebung vom Umfang n aus, bei der an n Untersuchungseinheiten die Werte x_1, \dots, x_n des Merkmals X beobachtet/gemessen wurden (die sog. *Rohdaten*).

- ▶ Die Daten werden nach den vorkommenden Ausprägungen durchsucht. Diese werden mit a_1, a_2, \dots, a_k , $k \leq n$, bezeichnet.
- ▶ Dann lassen sich absolute und relative Häufigkeiten berechnen:

$h(a_j) = h_j$ absolute Häufigkeit der Ausprägung a_j ,
d.h. Anzahl der x_i aus x_1, \dots, x_n mit $x_i = a_j$

$f(a_j) = f_j = h_j/n$ relative Häufigkeit von a_j

h_1, \dots, h_k absolute Häufigkeitsverteilung

f_1, \dots, f_k relative Häufigkeitsverteilung

Univariate Beschreibung von Daten

Häufigkeiten

Etwa für ein kategoriales Merkmal mit k Kategorien lassen sich die Häufigkeiten gut in einer Häufigkeitstabelle zusammenfassen:

Ausprägung	a_1	a_2	\dots	a_k
absolute Häufigkeit	h_1	h_2	\dots	h_k
relative Häufigkeit	f_1	f_2	\dots	f_k

Eine derartige Häufigkeitstabelle lässt sich auch graphisch gut veranschaulichen.

Univariate Beschreibung von Daten

Einfache Graphiken für diskrete Merkmale

- ▶ **Stabdiagramm:** Trage über a_1, \dots, a_k jeweils einen zur x -Achse senkrechten Strich (Stab) mit Höhe h_1, \dots, h_k (oder f_1, \dots, f_k) ab.
- ▶ **Säulendiagramm:** wie Stabdiagramm, aber mit Rechtecken statt Strichen.
- ▶ **Balkendiagramm:** wie Säulendiagramm, aber mit vertikal statt horizontal gelegter x -Achse.
- ▶ **Kreisdiagramm:** Flächen des Kreissektors sind proportional zu den Häufigkeiten, d.h. Winkel des Kreissektors $j = f_j \cdot 360^\circ$.

Univariate Beschreibung von Daten

Histogramm

Für die Veranschaulichung eines mindestens ordinal-skalierten Merkmals, das in vielen Ausprägungen vorliegt, insbes. für stetige Merkmale, bietet sich das sog. *Histogramm* an.

- ▶ Gruppiere die Daten in k benachbarte Intervalle $[c_0, c_1), [c_1, c_2), \dots, [c_{k-1}, c_k)$.
- ▶ h_j und f_j bezeichnen nun die absoluten bzw. relativen Häufigkeiten von Daten im Intervall $[c_{j-1}, c_j)$.
- ▶ Zeichne über den Klassen $[c_0, c_1), [c_1, c_2), \dots, [c_{k-1}, c_k)$ Rechtecke mit
Breite: $d_j = c_j - c_{j-1}$
Höhe: gleich (oder proportional zu) h_j/d_j bzw. f_j/d_j
Fläche: gleich (oder proportional zu) h_j bzw. f_j
- ▶ Am Histogramm lässt sich auch gut der Typ der Verteilung ablesen: symmetrisch/linkssteil/rechtssteil; unimodal/bimodal/multimodal.

Univariate Beschreibung von Daten

Empirische Verteilungsfunktion

Die *empirische Verteilungsfunktion* beantwortet die Frage “Welcher Anteil der Daten ist kleiner oder gleich einem bestimmten Wert x ?”

- ▶ Absolute kumulierte Häufigkeitsverteilung

$$H(x) = \text{Anzahl der Werte } x_i \text{ mit } x_i \leq x,$$

bzw. mit geordneten Ausprägungen $a_1 < \dots < a_k$ und deren Häufigkeiten

$$H(x) = h(a_1) + \dots + h(a_j) = \sum_{i: a_i \leq x} h_i.$$

Dabei ist a_j die größte Ausprägung, für die noch $a_j \leq x$ gilt, so dass also $a_{j+1} > x$.

Univariate Beschreibung von Daten

Empirische Verteilungsfunktion

- ▶ Empirische Verteilungsfunktion

$$F(x) = H(x)/n = \text{Anteil der Werte } x_i \text{ mit } x_i \leq x,$$

bzw.

$$F(x) = f(a_1) + \dots + f(a_j) = \sum_{i: a_i \leq x} f_i,$$

wobei $a_j \leq x$ und $a_{j+1} > x$.

Einfache Kenngrößen

Lagemaße

- **Arithmetisches Mittel:** Das arithmetische Mittel wird aus der Urliste x_1, \dots, x_n durch

$$\bar{x} = \frac{1}{n}(x_1 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

berechnet. Für Häufigkeitsdaten mit Ausprägungen a_1, \dots, a_k und relativen Häufigkeiten f_1, \dots, f_k gilt

$$\bar{x} = a_1 f_1 + \dots + a_k f_k = \sum_{j=1}^k a_j f_j.$$

Beachte: Das arithmetische Mittel ist nur für **metrische** Merkmale sinnvoll! Für qualitative (nominale oder ordinale) Merkmale ist es ungeeignet (Ausnahme: binäre Merkmale)!

Einfache Kenngrößen

Lagemaße

- ▶ **Median:** Für ungerades n ist der Median x_{med} die mittlere Beobachtung der geordneten Urliste $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, für gerades n das arithmetische Mittel der beiden in der Mitte liegenden Beobachtungen, d.h.:

$$x_{med} = \begin{cases} x_{((n+1)/2)} & \text{für } n \text{ ungerade} \\ \frac{1}{2} (x_{(n/2)} + x_{(n/2+1)}) & \text{für } n \text{ gerade} \end{cases}$$

Eigenschaften des Medians:

Mindestens 50 % der Daten sind kleiner oder gleich x_{med} .

Mindestens 50 % der Daten sind größer oder gleich x_{med} .

- ▶ **Modus** x_{mod} : Ausprägung mit der größten Häufigkeit.

Einfache Kenngrößen

Lagemaße

- ▶ **Lageregeln:**

Symmetrische Verteilungen: $\bar{x} \approx x_{med} \approx x_{mod}$

Linkssteile Verteilungen: $\bar{x} > x_{med} > x_{mod}$

Rechtssteile Verteilungen: $\bar{x} < x_{med} < x_{mod}$

- ▶ **Geometrisches Mittel:** Das geometrische Mittel zu den Faktoren x_1, \dots, x_n ist

$$\bar{x}_{geom} = (x_1 \cdot \dots \cdot x_n)^{1/n}.$$

Das geometrische Mittel kommt vor allem bei der Berechnung von Wachstumsraten zu Einsatz.

Einfache Kenngrößen

Quantile und Boxplot

Quantile

Jeder Wert x_p mit $0 < p < 1$, für den mindestens ein Anteil p der Daten kleiner/gleich x_p und mindestens ein Anteil $1 - p$ größer/gleich x_p ist, heißt *p-Quantil*. Es muss also gelten

$$\frac{\text{Anzahl } x\text{-Werte} \leq x_p}{n} \geq p \quad \text{und} \quad \frac{\text{Anzahl } x\text{-Werte} \geq x_p}{n} \geq 1 - p.$$

Damit gilt für das p -Quantil:

$$\begin{aligned} x_p &= x_{([np]+1)}, && \text{wenn } np \text{ nicht ganzzahlig,} \\ x_p &\in [x_{(np)}, x_{(np+1)}], && \text{wenn } np \text{ ganzzahlig.} \end{aligned}$$

Dabei ist $[np]$ die zu np nächste kleinere ganze Zahl.

Einfache Kenngrößen

Quantile und Boxplot

Wichtig sind vor allem:

- ▶ Unteres Quartil = 25%-Quantil = $x_{0.25}$,
- ▶ Median = 50%-Quantil = $x_{0.5}$,
- ▶ Oberes Quartil = 75%-Quantil = $x_{0.75}$,
- ▶ Dezile, d.h. $p = 0.1, 0.2, \dots, 0.9$,
- ▶ 5% sowie 95%-Quantile.

Interquartilsabstand (IQR): Die Distanz

$$d_Q = x_{0.75} - x_{0.25}$$

heißt Interquartilsabstand (interquartile range).

Fünf-Punkte-Zusammenfassung einer Verteilung:

$$x_{\min}, x_{0.25}, x_{\text{med}}, x_{0.75}, x_{\max}$$

Einfache Kenngrößen

Quantile und Boxplot

Box-Plot:

1. $x_{0.25}$ = Anfang der Box
 $x_{0.75}$ = Ende der Box
 d_Q = Länge der Box
2. Der Median wird durch einen Punkt oder Strich in der Box markiert.
3. Zwei Linien ("whiskers") außerhalb der Box gehen bis zu x_{min} und x_{max} .

Modifizierter Box-Plot:

- ▶ Die Linien außerhalb der Box werden nur bis zu x_{min} bzw. x_{max} gezogen, falls x_{min} bzw. x_{max} innerhalb der sog. Zäune $[z_u, z_o]$ liegen.
- ▶ Andernfalls gehen die Linien nur bis zum kleinsten bzw. größten Wert innerhalb der Zäune.
- ▶ Außerhalb liegende Werte werden individuell eingezeichnet, z.B. als Punkte.
- ▶ Mögliche Zäune sind z.B. $z_u = x_{0.25} - 1.5d_Q$ und $z_o = x_{0.75} + 1.5d_Q$.

Einfache Kenngrößen

Varianz und Standardabweichung

Varianz und Standardabweichung quantifizieren die Streuung der Daten.

- Die (*empirische*) *Varianz* der Werte x_1, \dots, x_n ist

$$\tilde{s}^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2] = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Für Häufigkeitsdaten gilt

$$\tilde{s}^2 = \frac{1}{n}[(a_1 - \bar{x})^2 f_1 + \dots + (a_k - \bar{x})^2 f_k] = \frac{1}{n} \sum_{j=1}^k (a_j - \bar{x})^2 f_j.$$

- Die sog. *Stichprobenvarianz* ist

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

- Die *Standardabweichung* \tilde{s} bzw. s ist die (positive) Wurzel aus der Varianz, d.h.

$$\tilde{s} = +\sqrt{\tilde{s}^2} \quad \text{bzw.} \quad s = +\sqrt{s^2}$$

Elementare Wahrscheinlichkeitsrechnung

Einführung

Daten können oft als das Ergebnis eines *Zufallsvorgangs* aufgefasst werden.

Zufallsvorgang:

- ▶ Ein Zufallsvorgang führt zu einem von mehreren sich gegenseitig ausschließenden Ergebnissen.
- ▶ Es ist vor der Durchführung ungewiss, welches Ergebnis tatsächlich eintreten wird. Es lassen sich nur Wahrscheinlichkeiten angeben.

Um die möglichen Ausgänge eines Zufallsvorgangs näher zu beschreiben, sind einige Grundbegriffe der Mengenlehre nützlich.

Eine *Menge* ist eine Zusammenfassung verschiedener Objekte zu einem Ganzen. Die einzelnen Objekte werden Elemente genannt.

Eine Menge ohne Elemente ist die sog. *leere Menge*, i.Z.: \emptyset .

Elementare Wahrscheinlichkeitsrechnung

Grundlegende Begriffe der Mengenlehre

- ▶ Die Eigenschaft “ x ist ein Element der Menge A ” stellt man in Zeichen dar als: $x \in A$, sonst: $x \notin A$.
- ▶ A ist Teilmenge von B , i.Z.: $A \subset B$, wenn jedes Element von A auch in B ist.
- ▶ Die Schnittmenge $A \cap B$ ist die Menge aller Elemente, die sowohl in A als auch in B sind; i.Z.: $A \cap B = \{x : x \in A \text{ und } x \in B\}$.
- ▶ Die Vereinigungsmenge $A \cup B$ ist die Menge aller Elemente, die in A oder B sind; i.Z.: $A \cup B = \{x : x \in A \text{ oder } x \in B\}$.
- ▶ Die Differenzmenge $A \setminus B$ ist die Menge aller Elemente, die in A aber nicht in B sind; i.Z.: $A \setminus B = \{x : x \in A \text{ und } x \notin B\}$.
- ▶ Für $A \subset \Omega$ ist die Komplementärmenge \bar{A} von A bzgl. Ω die Menge aller Elemente von Ω , die nicht in A sind, i.Z.: $\bar{A} = \Omega \setminus A$.
- ▶ Die Potenzmenge $\mathcal{P}(A)$ ist die Menge aller Teilmengen von A ; i.Z.: $\mathcal{P}(A) = \{M : M \subset A\}$.
- ▶ Die Mächtigkeit von A , i.Z.: $|A|$, gibt an, wieviele Elemente in A enthalten sind.

Elementare Wahrscheinlichkeitsrechnung

Zufallereignisse

- ▶ Der *Ergebnisraum* $\Omega = \{\omega_1, \dots, \omega_n\}$ ist die Menge aller möglichen Ergebnisse ω_i , $i = 1, \dots, n$, eines Zufallsvorgangs.
- ▶ Teilmengen von Ω heißen (Zufalls-)Ereignisse.
- ▶ Die einelementigen Teilmengen von Ω , d.h. $\{\omega_1\}, \dots, \{\omega_n\}$, werden als *Elementarereignisse* bezeichnet.

Elementare Wahrscheinlichkeitsrechnung

Axiome und Rechenregeln

Die Wahrscheinlichkeit P von Ereignissen lässt sich als Abbildung auffassen. Dabei ordnet P jedem Ereignis A , das eine Teilmenge von Ω ist, eine Zahl zwischen 0 und 1 zu, also:

$$P : \{A : A \subset \Omega\} \rightarrow [0, 1]$$

Es gelten dabei die Axiome von Kolmogoroff:

1. $P(A) \geq 0$.
2. $P(\Omega) = 1$.
3. Falls $A \cap B = \emptyset$, so ist $P(A \cup B) = P(A) + P(B)$.

Hieraus lassen sich noch weitere Rechenregeln ableiten:

- ▶ $0 \leq P(A) \leq 1$ für $A \subset \Omega$,
- ▶ $P(\emptyset) = 0$,
- ▶ $P(A) \leq P(B)$, falls $A \subset B$ und $A, B \subset \Omega$,
- ▶ $P(\bar{A}) = 1 - P(A)$ mit $\bar{A} = \Omega \setminus A$,
- ▶ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Elementare Wahrscheinlichkeitsrechnung

Bedingte Wahrscheinlichkeit und Unabhängigkeit

- **Bedingte Wahrscheinlichkeit:** Seien A und B zwei Ereignisse (und gelte $P(B) > 0$), dann ist die *bedingte Wahrscheinlichkeit* von “ A unter der Bedingung/gegeben B ” durch

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Hieraus ergibt sich auch

$$P(A \cap B) = P(A|B) \cdot P(B).$$

- **Unabhängigkeit:** Zwei Ereignisse A und B werden *unabhängig* genannt, wenn

$$P(A|B) = P(A),$$

oder äquivalent

$$P(A \cap B) = P(A) \cdot P(B)$$

Elementare Wahrscheinlichkeitsrechnung

Satz von der totalen Wahrscheinlichkeit und Bayes-Formel

- **Satz von der totalen Wahrscheinlichkeit:** Gegeben seien paarweise disjunkte¹ Ereignisse A_j , $j = 1, \dots, k$, und es gelte $\bigcup_{j=1}^k A_j = \Omega$. Dann ist

$$P(B) = \sum_{j=1}^k P(B \cap A_j) = \sum_{j=1}^k P(B|A_j) \cdot P(A_j).$$

- **Formel von Bayes:**

$$P(A_i|B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^k P(B|A_j) \cdot P(A_j)}$$

Insbesondere gilt:

$$P(A|B) = \frac{P(B \cap A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

¹d.h. $A_j \cap A_r = \emptyset$ für $j \neq r$.

Laplace-Wahrscheinlichkeiten

Bei einem sog. *Laplace-Experiment* geht man davon aus, dass alle Ergebnisse (d.h. Elementarereignisse) **gleichwahrscheinlich** sind.

Daher lässt sich die Wahrscheinlichkeit des Ereignisses A berechnen als:

$$P(A) = \frac{\text{Anzahl der für } A \text{ günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}} = \frac{|A|}{|\Omega|}$$

Beispiele:

- ▶ Münz-Wurf,
- ▶ Würfel-Wurf,
- ▶ Ziehung der Lottozahlen, etc.

Um die Mächtigkeit $|A|$ und $|\Omega|$ zu bestimmen, benötigt man oft die Regeln der Kombinatorik.

Kombinatorik

oder die Kunst des Abzählens

Um die Anzahl möglicher Ergebnisse eines Zufallsvorgangs zu bestimmen, stellt man sich diesen oft als Urnenmodell vor:

- ▶ In der Urne befinden sich N nummerierte Kugeln.
- ▶ Es werden n Kugeln gezogen, entweder mit oder ohne Zurücklegen.
- ▶ Wir berechnen die Anzahl möglicher Versuchsausgänge. Dabei wird noch unterschieden, ob die Reihenfolge, mit der die Kugeln gezogen werden, beachtet wird oder nicht.

Kombinatorik

oder die Kunst des Abzählens

Im Urnenodell ergibt sich dann:

	ohne Zurücklegen	mit Zurücklegen
Reihenfolge berücksichtigt	$\frac{N!}{(N-n)!}$	N^n
Reihenfolge nicht berücksichtigt	$\binom{N}{n}$	$\binom{N+n-1}{n}$

Dabei ist die Fakultät $N! = N \cdot (N-1) \cdot (N-2) \cdot \dots \cdot 2 \cdot 1$, und der Binomialkoeffizient

$$\binom{N}{n} = \frac{N!}{(N-n)! \cdot n!}.$$

Binomialverteilung

Wir betrachten n unabhängige Wiederholungen eines Einzelversuchs mit dichotomen Ausgang, d.h.: “Treffer/Niete”, “Erfolg/Misserfolg”, 1/0, etc.

- ▶ Die Erfolgswahrscheinlichkeit bei jedem Einzelversuch sei p . Die Misserfolgswahrscheinlichkeit ist somit $q = 1 - p$.
- ▶ Die sog. Zufallsvariable X bezeichne die Gesamtzahl an Erfolgen bei n Einzelversuchen.
- ▶ Dann folgt X einer *Binomialverteilung* mit Parametern n und p und es gilt

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \text{ falls } x = 0, 1, \dots, n,$$

und $P(X = x) = 0$ sonst.

Diskrete und Stetige Zufallsvariablen

Definitionen

- ▶ **Zufallsvariable:** Eine Variable oder ein Merkmal X , dessen – über Zahlen repräsentierte – Werte oder Ausprägungen die Ergebnisse eines Zufallsvorgangs sind, heißt *Zufallsvariable* (oder auch *Zufallsgröße* X).
- ▶ Die Zahl $x \in \mathbb{R}$, die X bei einer Durchführung des Zufallsvorgangs annimmt, heißt *Realisierung* oder Wert von X .
- ▶ **Diskrete Zufallsvariable:** Eine Zufallsvariable X heißt *diskret*, falls sie nur endlich oder abzählbar unendlich viele Werte $x_1, x_2, \dots, x_k, \dots$ annehmen kann. Die *Wahrscheinlichkeitsverteilung* von X ist durch die Wahrscheinlichkeiten

$$P(X = x_i) = p_i, \quad i = 1, 2, \dots, k, \dots,$$

gegeben. Die Menge $\mathcal{T}_X = \{x_1, x_2, \dots, x_k, \dots\}$ heißt *Träger* von X .

Diskrete und Stetige Zufallsvariablen

Definitionen

- ▶ **Verteilungsfunktion einer diskreten Zufallsvariable:**

$$F(x) = P(X \leq x) = \sum_{i: x_i \leq x} P(X = x_i) = \sum_{i: x_i \leq x} p_i$$

- ▶ **Unabhängigkeit von diskreten Zufallsvariablen:** Zwei diskrete Zufallsvariablen X und Y mit den Trägern $\mathcal{T}_X = \{x_1, x_2, \dots, x_k, \dots\}$ und $\mathcal{T}_Y = \{y_1, y_2, \dots, y_l, \dots\}$ heißen *unabhängig*, wenn für beliebige $x \in \mathcal{T}_X$ und $y \in \mathcal{T}_Y$

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y).$$

Dabei beschreibt $P(X = x, Y = y)$ die *gemeinsame Verteilung* von X und Y .

Diskrete und Stetige Zufallsvariablen

Definitionen

- ▶ **Stetige Zufallsvariablen und Dichten:** Eine Zufallsvariable X heißt *stetig*, wenn es eine Funktion $f(x) \geq 0$ gibt, so dass für jedes Intervall $[a, b]$

$$P(a \leq X \leq b) = \int_a^b f(x) dx.$$

Die Funktion $f(x)$ heißt (*Wahrscheinlichkeits-*)*Dichte* von X .

Es gilt die Normierungseigenschaft $\int_{-\infty}^{+\infty} f(x) dx = 1$, d.h. die Gesamtfläche zwischen x -Achse und der Dichte $f(x)$ ist gleich 1.

- ▶ **Wahrscheinlichkeiten stetiger Zufallsvariablen:** Für stetige Zufallsvariablen X gilt

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

und

$$P(X = x) = 0 \quad \text{für jedes } x \in \mathbb{R}.$$

Diskrete und Stetige Zufallsvariablen

Definitionen

► Verteilungsfunktion einer stetigen Zufallsvariable:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

Eigenschaften:

1. $F(x)$ ist stetig und monoton wachsend mit Werten im Intervall $[0, 1]$.
2. Für die Grenzen gilt

$$F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0,$$

$$F(+\infty) = \lim_{x \rightarrow +\infty} F(x) = 1.$$

3. Für Werte von x , an denen $f(x)$ stetig ist, gilt

$$F'(x) = \frac{d}{dx} F(x) = f(x).$$

4. Für Intervalle erhält man

$$P(a \leq X \leq b) = F(b) - F(a),$$

$$P(X \geq a) = 1 - F(a).$$

Diskrete und Stetige Zufallsvariablen

Definitionen

- ▶ **Unabhängigkeit von stetigen Zufallsvariablen:** Zwei stetige Zufallsvariablen X und Y heißen *unabhängig*, wenn für alle $x \in \mathbb{R}$ und $y \in \mathbb{R}$

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y) = F_X(x) \cdot F_Y(y).$$

Dabei ist F_X bzw. F_Y die Verteilungsfunktion von X bzw. Y . Wie zuvor beschreibt $P(X = x, Y = y)$ die *gemeinsame Verteilung* von X und Y .

Diskrete und Stetige Zufallsvariablen

Erwartungswert

- ▶ **Erwartungswert einer diskreten Zufallsvariable:** Der *Erwartungswert* einer diskreten Zufallsvariable X mit den Werten x_1, \dots, x_k, \dots und der Wahrscheinlichkeitsverteilung p_1, \dots, p_k, \dots ist

$$\mu = E(X) = x_1 p_1 + \dots + x_k p_k + \dots = \sum_{i \geq 1} x_i p_i.$$

- ▶ **Erwartungswert einer stetigen Zufallsvariable:** Der *Erwartungswert* einer stetigen Zufallsvariable X mit Dichte $f(x)$ ist

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f(x) dx.$$

- ▶ **Zwei wichtige Eigenschaften des Erwartungswerts:**
 1. Lineare Transformationen: Für $Y = aX + b$ (mit Konstanten a, b) ist $E(Y) = E(aX + b) = aE(X) + b$.
 2. Additivität: Für zwei Zufallsvariablen X und Y ist $E(X + Y) = E(X) + E(Y)$.

Diskrete und Stetige Zufallsvariablen

Erwartungswert und arithmetisches Mittel

- ▶ Der Erwartungswert $E(X) = \mu$ charakterisiert das Verhalten einer Zufallsvariable/eines Zufallsexperiments.
- ▶ Das arithmetische Mittel \bar{x} beschreibt den Schwerpunkt der tatsächlich beobachteten Daten.
- ▶ Das arithmetische Mittel \bar{x} kann aber verwendet werden, um den Erwartungswert einer zu Grunde liegenden Zufallsvariable/Verteilung zu schätzen.
- ▶ Für unabhängige Wiederholungen X_1, \dots, X_n von X und beliebig kleines c gilt nämlich (Gesetz der großen Zahlen):

$$P(|\bar{X}_n - \mu| \leq c) \rightarrow 1 \quad \text{für } n \rightarrow \infty,$$

wobei

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n).$$

Man sagt: \bar{X}_n konvergiert nach Wahrscheinlichkeit gegen μ .

Beachte: \bar{X}_n ist auch eine Zufallsgröße.

Diskrete und Stetige Zufallsvariablen

Varianz

- ▶ **Varianz und Standardabweichung einer diskreten Zufallsvariable:** Die *Varianz* einer diskreten Zufallsvariable X mit $\mu = E(X)$ ist

$$\sigma^2 = \text{Var}(X) = (x_1 - \mu)^2 p_1 + \dots + (x_k - \mu)^2 p_k + \dots = \sum_{i \geq 1} (x_i - \mu)^2 p_i.$$

Die *Standardabweichung* ist $\sigma = +\sqrt{\text{Var}(X)}$.

- ▶ **Varianz und Standardabweichung einer stetigen Zufallsvariable:** Die *Varianz* einer stetigen Zufallsvariable X mit Dichte $f(x)$ und $\mu = E(X)$ ist

$$\sigma^2 = \text{Var}(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx.$$

Die *Standardabweichung* ist $\sigma = +\sqrt{\text{Var}(X)}$.

Diskrete und Stetige Zufallsvariablen

Varianz

► Wichtige Eigenschaften von Varianzen:

1. Varianz als erwartete quadratische Abweichung:

$$\text{Var}(X) = E((X - \mu)^2).$$

2. Verschiebungsregel:

$$\text{Var}(X) = E(X^2) - (E(X))^2 = E(X^2) - \mu^2.$$

3. Lineare Transformationen: Für $Y = aX + b$ (mit Konstanten a, b) ist

$$\text{Var}(Y) = \text{Var}(aX + b) = a^2 \text{Var}(X).$$

4. Varianz der Summe von unabhängigen Zufallsvariablen: Falls X und Y unabhängig sind, gilt

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Diskrete und Stetige Zufallsvariablen

Modus, Median und Quantile für stetige Zufallsvariablen

- ▶ **Modus:** Jeder Wert, für den $f(x)$ ein Maximum besitzt, ist *Modus*, kurz x_{mod} . Falls das Maximum eindeutig ist und $f(x)$ keine weiteren lokalen Maxima besitzt, heißt $f(x)$ unimodal.
- ▶ **Median and Quantile:** Für $0 < p < 1$ ist das *p-Quantil* x_p die Zahl auf der x -Achse, für die

$$F(x_p) = p.$$

Der *Median* x_{med} ist das 50%-Quantil, d.h.

$$F(x_{med}) = \frac{1}{2}.$$

Für streng monotone Verteilungsfunktionen $F(x)$ sind p -Quantil und Median eindeutig bestimmt.

Normalverteilung

Die Normalverteilung ist die wichtigste (stetige) Verteilung.

- ▶ Eine Zufallsvariable X heißt *normalverteilt* mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, kurz $X \sim N(\mu, \sigma^2)$, wenn sie die Dichte

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

besitzt. Es gilt

$$E(X) = \mu, \quad \text{Var}(X) = \sigma^2.$$

Die Normalverteilung wird auch als Gauß-Verteilung und die Dichtekurve als Gauß-Kurve bezeichnet.

- ▶ Speziell für $\mu = 0$, $\sigma^2 = 1$ erhält man die *Standardnormalverteilung* $N(0, 1)$ mit der Dichte

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

Die Verteilungsfunktion der Standardnormalverteilung ist $\Phi(x)$.

Normalverteilung

- ▶ **Standardisierung:** Ist X eine $N(\mu, \sigma^2)$ -verteilte Zufallsvariable, so ist die *standardisierte Zufallsvariable*

$$Z = \frac{X - \mu}{\sigma}$$

standardnormalverteilt, d.h. $Z \sim N(0, 1)$.

- ▶ Für die Verteilungsfunktion $F(x)$ einer $N(\mu, \sigma^2)$ -verteilten Zufallsvariable X gilt

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \Phi(z) \quad \text{mit} \quad z = \frac{x - \mu}{\sigma}.$$

Die Funktion Φ lässt sich nicht in geschlossener Form angeben, ihre Werte sind aber tabelliert.

Zusammenhangsanalyse

Einführung

- ▶ Zwei kategoriale Merkmale X und Y sind gegeben. Wir möchten eine Maßzahl, die den Zusammenhang von X und Y misst.
- ▶ Zwei metrische Merkmale X und Y sind gegeben. Wir möchten eine Maßzahl, die den Zusammenhang von X und Y misst.
- ▶ Wir betrachten ein metrisches Merkmal X in verschiedenen Gruppen, die z.B. durch ein diskretes Merkmal Y definiert werden. Wir möchten untersuchen, wie sich der Erwartungswert von X in den einzelnen Gruppen verhält.
- ▶ Im einfachsten Fall liegt nur eine Gruppe vor, und wir möchten den Erwartungswert von X mit einem hypothetischen Erwartungswert vergleichen.
- ▶ Wir betrachten eine abhängige Variable und möchten untersuchen wie diese durch andere Größen beeinflusst wird.

Kontingenztafeln

Einführung

- ▶ Wie betrachten zwei diskrete Merkmale X und Y mit den möglichen Ausprägungen a_1, \dots, a_k für X und b_1, \dots, b_m für Y .
- ▶ In der Urliste liegen für jedes Objekt die gemeinsamen Messwerte vor, d.h. man erhält die Tupel $(x_1, y_1), \dots, (x_n, y_n)$.
- ▶ Eine $(k \times n)$ -Kontingenztafel der absoluten Häufigkeiten besitzt die Form

	b_1	\dots	b_m	
a_1	h_{11}	\dots	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	\dots	h_{2m}	$h_{2\cdot}$
\vdots	\vdots		\vdots	\vdots
a_k	h_{k1}	\dots	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	\dots	$h_{\cdot m}$	n

Dabei bezeichnen

$h_{ij} = h(a_i, b_j)$ die absolute Häufigkeit der Kombination (a_i, b_j) ,
 $h_{1\cdot}, \dots, h_{k\cdot}$ die Randhäufigkeiten von X und
 $h_{\cdot 1}, \dots, h_{\cdot m}$ die Randhäufigkeiten von Y .

Kenngrößen für Kontingenztafeln

Wir möchten eine Maßzahl, die angibt wie stark X und Y voneinander abhängen.

- ▶ χ^2 -Koeffizient:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{\left(h_{ij} - \frac{h_{i.} \cdot h_{.j}}{n}\right)^2}{\frac{h_{i.} \cdot h_{.j}}{n}}, \quad \chi^2 \in [0, \infty)$$

- ▶ χ^2 groß, wenn X und Y voneinander abhängen,
- ▶ χ^2 klein, wenn X und Y nicht voneinander abhängen.

- ▶ Kontingenzkoeffizient:

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

mit Wertebereich $K \in [0, \sqrt{(M-1)/M}]$, wobei $M = \min\{k, m\}$.

Korrigierte Version $K^* = K / \sqrt{(M-1)/M} \in [0, 1]$.

Kenngrößen für Kontingenztafeln

Spezialfall (2×2) -Tafel

	$Y = 1$	$Y = 2$	
$X = 1$	h_{11}	h_{12}	$h_{1\cdot}$
$X = 2$	h_{21}	h_{22}	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	n

- ▶ Unter einer *Chance* (“odds”) versteht man das Verhältnis zwischen dem Auftreten von $Y = 1$ und $Y = 2$. In der Subpopulation $X = r$ ergibt sich die (empirische) *bedingte Chance*

$$\gamma(1, 2 | X = r) = \frac{h_{r1}}{h_{r2}}.$$

- ▶ Das *Kreuzproduktverhältnis* (*relative Chance* oder *Odds Ratio*) ist

$$\gamma = \frac{\gamma(1, 2 | X = 1)}{\gamma(1, 2 | X = 2)} = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11}h_{22}}{h_{21}h_{12}}.$$

Kovarianzen und Korrelationen

Wir betrachten zwei metrische Merkmale/Zufallsvariablen X und Y , von denen an Objekten $i = 1, \dots, n$ Beobachtungen der Form $(x_1, y_1), \dots, (x_n, y_n)$ vorliegen.

- ▶ Die *Kovarianz* der Zufallsvariablen X und Y ist

$$\text{Cov}(X, Y) = E([X - E(X)][Y - E(Y)]).$$

- ▶ Die *Korrelationskoeffizient* der Zufallsvariablen X und Y ist

$$\rho = \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Kovarianzen und Korrelationen

- ▶ Der Korrelationskoeffizient hat Werte zwischen -1 und $+1$ und misst den linearen Zusammenhang von X und Y . Er misst sowohl Richtung als auch Stärke des Zusammenhangs.
- ▶ Zwei Zufallsvariablen X und Y heißen *unkorreliert*, wenn gilt

$$\rho(X, Y) = 0.$$

Wenn $\rho(X, Y) \neq 0$, heißen X und Y *korreliert*.

- ▶ Sind zwei Zufallsvariablen *unabhängig*, so sind sie auch unkorreliert, d.h. es gilt $\rho(X, Y) = 0$. Die Umkehrung gilt im Allgemeinen jedoch nicht.

Kovarianzen und Korrelationen

- ▶ Die beobachteten Daten $(x_1, y_1), \dots, (x_n, y_n)$ können in Form eines sog. *Streudiagramms* (scatter plot) als Punkte in ein x - y -Koordinatensystem eingezeichnet werden.
- ▶ Die empirische Version von ρ ist der *Bravais-Pearson-Korrelationskoeffizient*

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

mit Wertebereich $-1 \leq r \leq 1$.

- $r > 0$ positive Korrelation, gleichsinniger linearer Zusammenhang, Tendenz: Werte (x_i, y_i) um eine Gerade positiver Steigung.
- $r < 0$ negative Korrelation, gegensinniger linearer Zusammenhang, Tendenz: Werte (x_i, y_i) um eine Gerade negativer Steigung.
- $r = 0$ keine Korrelation, unkorreliert, kein linearer Zusammenhang

Mittelwertsvergleiche

Einführung

Gegeben seien Merkmale/Zufallsvariablen X und Y . Wir überprüfen Hypothesen über $E(X)$ und $E(Y)$. Im einfachsten Fall nur bzgl. $E(X)$.

Allgemeine Form der Hypothesen über $\mu = E(X)$:

- (a) $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$ zweiseitige Alternative H_1
- (b) $H_0 : \mu = \mu_0, H_1 : \mu < \mu_0$ einseitige Alternative H_1
- (c) $H_0 : \mu = \mu_0, H_1 : \mu > \mu_0$ einseitige Alternative H_1

Bemerkungen:

- ▶ Verschiedene Tests unterscheiden sich durch Annahmen über X .
- ▶ Gauß- und (Student) t-Test sind die bekanntesten Tests zum Überprüfen obiger Hypothesen

(Exakter) Gauß-Test

- ▶ Annahmen: $X \sim N(\mu, \sigma^2)$ mit bekannter Varianz σ^2 , Stichprobenvariablen X_1, \dots, X_n i.i.d. (independent identically distributed) wie X .
- ▶ Hypothesen über $\mu = E(X)$: (a), (b), (c) wie angegeben.
- ▶ Idee für Test: Falls H_0 richtig ist: $E(X) = \mu_0$. Bilde arithmetisches Mittel \bar{x} zu den Stichprobenwerten x_1, \dots, x_n . Lehne H_0 ab, falls Abweichung zwischen μ_0 und \bar{x} zu groß.
- ▶ Frage: Wie groß sind die kritischen Werte für diese Abweichung zu wählen?

(Exakter) Gauß-Test

Diskussion für Hypothesenpaar (c)

$$H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0$$

Übergang von \bar{X} zu standardisierter Teststatistik

$$Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}$$

$$\text{Unter } H_0 \text{ gilt: } \bar{X} \sim N(\mu_0, \frac{\sigma^2}{n}) \quad \Rightarrow \quad Z \sim N(0, 1)$$

$$\text{Testvorschrift für } Z: H_0 \text{ ablehnen} \quad \Leftrightarrow \quad Z > k$$

(Exakter) Gauß-Test

Frage: Wie ist der kritische Wert k zu wählen?

Prinzip: Die Wahrscheinlichkeit für den

Fehler 1. Art: H_0 wird abgelehnt, obwohl H_0 richtig ist

soll (höchstens) gleich einem (kleinen) vorgegebenen *Signifikanzniveau* α sein (z.B. $\alpha = 0.1, 0.05, 0.01$). D.h.

$$P(H_0 \text{ ablehnen} \mid H_0 \text{ richtig}) = \alpha$$

(Exakter) Gauß-Test

Beim exakten Gauß-Test ist dies äquivalent zu

$$P(Z > k \mid \mu = \mu_0) = \alpha$$

$$\Leftrightarrow k = z_{1-\alpha} \text{ (mit } z_{1-\alpha} = (1 - \alpha)\text{-Quantil der Standardnormalverteilung)}$$

Testvorschrift: H_0 ablehnen, falls $Z > z_{1-\alpha}$

(Exakter) Gauß-Test

Bemerkungen:

- ▶ Neben dem Fehler 1. Art gibt es den
Fehler 2. Art: H_0 wird nicht abgelehnt, obwohl H_1 richtig ist.
Es gilt: Je kleiner (größer) das Signifikanzniveau α gewählt wird, desto größer (kleiner) wird die Wahrscheinlichkeit für einen Fehler 2. Art.
- ▶ Falls $H_0 : \mu \leq \mu_0$ gilt, also auch $\mu < \mu_0$ als Nullhypothese möglich ist, folgt

$$P(\text{Fehler 1. Art}) = P(Z > z_{1-\alpha} \mid \mu \leq \mu_0) \leq \alpha$$

(Exakter) Gauß-Test

Ein-Stichproben-Fall

Hypothesenpaar (b)

$H_0 : \mu = \mu_0, H_1 : \mu < \mu_0$ symmetrisch zu (c).

⇒ **Testvorschrift:** H_0 ablehnen, falls $Z < -z_{1-\alpha}$

Hypothesenpaar (a)

$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$

Unter $H_0 : Z \sim N(0, 1)$

Große Abweichungen von \bar{X} von μ_0 **sowohl nach oben als auch unten** sollen zur Ablehnung von H_0 führen.

⇒ **Testvorschrift:** H_0 ablehnen, falls $|Z| > z_{1-\frac{\alpha}{2}}$

(Exakter) Gauß-Test

Ein-Stichproben-Fall

Alternative Formulierungen der Testentscheidungen:

- (1) Mit p -Werten (“Überschreitungswahrscheinlichkeiten”),
- (2) Mit Hilfe von Konfidenzintervallen (vgl. später).

Definition p -Wert:

- ▶ Der p -Wert ist definiert als die Wahrscheinlichkeit, unter H_0 den beobachteten Prüfgrößenwert oder einen in die Richtung der Alternative extremeren Wert zu erhalten.
- ▶ Ist der p -Wert kleiner oder gleich dem vorgegebenen Signifikanzniveau α , so wird H_0 verworfen. Ansonsten behält man H_0 bei.

(Exakter) Gauß-Test

Ein-Stichproben-Fall

Bemerkungen:

- ▶ Statistische Programmpakete geben in der Regel p -Werte für zweiseitige Tests aus.

Dann: H_0 ablehnen \Leftrightarrow " p -value" $< \alpha$, α vorgegebenes Signifikanzniveau.

- ▶ Vorsicht bei einseitigen Tests zu (b) und (c)! p -Werte müssen für Testentscheidung ggf. modifiziert werden.

Approximativer Gauß-Test

Ein-Stichproben-Fall

Annahmen:

X beliebig verteilt mit $E(X) = \mu$; $Var(X) = \sigma^2$ bekannt.

X_1, \dots, X_n i.i.d. wie X ; Faustregel: $n \geq 30$.

Unter $\mu = \mu_0$ gilt

$$\bar{X} \overset{a}{\sim} N\left(\mu_0, \frac{\sigma^2}{n}\right) \quad \text{bzw.} \quad Z \overset{a}{\sim} N(0, 1)$$

Testvorschrift wie beim exakten Gauß-Test.

Aber: $P(\text{Fehler 1. Art}) \overset{a}{\leq} \alpha$

(Student) t-Test

Ein-Stichproben-Fall

- ▶ Annahmen: Wie beim (exakten) Gauß-Test, aber: σ^2 unbekannt.
- ▶ Hypothesen: (a), (b), (c) wie bisher.
- ▶ Idee: Ersetze σ (beim Gauß-Test) durch

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2},$$

d.h. Teststatistik

$$Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}$$

wird ersetzt durch Teststatistik

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n}$$

(Student) t-Test

Ein-Stichproben-Fall

Man kann zeigen: $X \sim N(\mu_0, \sigma^2) \Rightarrow T \sim t(n-1)$ (Student) t-verteilt mit $n-1$ Freiheitsgraden.

\Rightarrow Herleitung der Testvorschriften wie beim Gauß-Test; ersetze Z durch T und die Dichte ϕ von Z durch Dichte der $t(n-1)$ -Verteilung.

\Rightarrow Ersetze in Testvorschriften Z durch T und z -Quantile durch $t(n-1)$ -Quantile.

Für $n \geq 30$: $t(n-1)$ -Quantile \approx z -Quantile, und T näherungsweise normalverteilt, auch wenn X nicht normalverteilt.

Konfidenzintervalle

Idee: Konstruiere ein Intervall, das mit vorgegebener Wahrscheinlichkeit den wahren Wert μ überdeckt.

Untere und obere Intervallgrenzen

$$G_u = g_u(X_1, \dots, X_n) \quad \text{und} \quad G_o = g_o(X_1, \dots, X_n)$$

bilden $(1 - \alpha)$ -Konfidenzintervall (Vertrauensintervall): \Leftrightarrow

$$P(G_u \leq G_o) = 1, \quad P(G_u \leq \mu \leq G_o) = 1 - \alpha$$

Die Intervallgrenzen G_u und G_o sind Zufallsvariablen! Somit ist auch das Intervall $[G_u, G_o]$ zufällig.

Realisiertes Konfidenzintervall:

$$[g_u, g_o]; \quad g_u = g_u(x_1, \dots, x_n), \quad g_o = g_o(x_1, \dots, x_n)$$

\Rightarrow Warnung: Eine Aussage wie “ μ liegt mit Wahrscheinlichkeit $1 - \alpha$ in $[g_u, g_o]$ ” ist Unsinn!

Konfidenzintervalle

Beispiele:

- ▶ $X \sim N(\mu, \sigma^2)$, σ^2 unbekannt; X_1, \dots, X_n i.i.d. wie X .
($1 - \alpha$)-Konfidenzintervall für μ :

$$\left[\bar{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right],$$

mit $S = \sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X})^2}$ und passendem $t_{1-\frac{\alpha}{2}}$ -Quantil.

- ▶ Konfidenzintervall für μ ohne Normalverteilungsannahme;
approximatives ($1 - \alpha$)-Konfidenzintervall (für $n \geq 30$):

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right],$$

(Falls Varianz σ^2 bekannt, ersetze S durch σ .)

Testentscheidung: Lehne $H_0: \mu = \mu_0$ ab, falls μ_0 nicht im geeigneten ($1 - \alpha$)-Konfidenzintervall enthalten.

Gauß/t-Test

Zwei-Stichproben-Fall

Ziel: Tests zum Vergleich von Parametern zweier Variablen X, Y .

Bezeichnungen und Annahmen:

- ▶ Metrische Merkmale X und Y .
- ▶ Unbekannte Parameter: $E(X) = \mu_X$ und $E(Y) = \mu_Y$.
- ▶ Stichprobenvariablen: X_1, X_2, \dots, X_n und Y_1, Y_2, \dots, Y_m .
- ▶ Annahmen:
 - X_1, \dots, X_n unabhängig und identisch verteilt wie X ,
 - Y_1, \dots, Y_m unabhängig und identisch verteilt wie Y ,
 - $X_1, \dots, X_n, Y_1, \dots, Y_m$ unabhängig.

Gauß/t-Test

Zwei-Stichproben-Fall

Hypothesen:

- ▶ Zweiseitiges Testproblem:

$$(a) \quad H_0 : \mu_X - \mu_Y = \delta_0 \quad \text{vs.} \quad H_1 : \mu_X - \mu_Y \neq \delta_0$$

- ▶ Einseitige Testprobleme:

$$(b) \quad H_0 : \mu_X - \mu_Y \geq \delta_0 \quad \text{vs.} \quad H_1 : \mu_X - \mu_Y < \delta_0$$

$$(c) \quad H_0 : \mu_X - \mu_Y \leq \delta_0 \quad \text{vs.} \quad H_1 : \mu_X - \mu_Y > \delta_0$$

Meistens ist der interessante Fall $\delta_0 = 0$, d.h. z.B. $\mu_X = \mu_Y$?

Gauß/t-Test

Zwei-Stichproben-Fall

Annahmen	Teststatistik	Ablehnbereiche
$X \sim N(\mu_X, \sigma_X^2),$ $Y \sim N(\mu_Y, \sigma_Y^2),$ σ_X^2, σ_Y^2 bekannt.	$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$	(a) $ Z > z_{1-\alpha/2}$ (b) $Z < -z_{1-\alpha}$ (c) $Z > z_{1-\alpha}$
$X \sim N(\mu_X, \sigma_X^2),$ $Y \sim N(\mu_Y, \sigma_Y^2),$ $\sigma_X^2 = \sigma_Y^2$ unbekannt.	$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}}$	(a) $ T > t_{1-\alpha/2}(n+m-2)$ (b) $T < -t_{1-\alpha}(n+m-2)$ (c) $T > t_{1-\alpha}(n+m-2)$
$X \sim N(\mu_X, \sigma_X^2),$ $Y \sim N(\mu_Y, \sigma_Y^2),$ σ_X^2, σ_Y^2 unbekannt.	$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$	(a) $ T > t_{1-\alpha/2}(k)$ (b) $T < -t_{1-\alpha}(k)$ (c) $T > t_{1-\alpha}(k)$
X, Y beliebig verteilt, $n, m \geq 30.$	$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$	(a) $ T > z_{1-\alpha/2}$ (b) $T < -z_{1-\alpha}$ (c) $T > z_{1-\alpha}$

wobei $k = (S_X^2/n + S_Y^2/m)^2 / ((S_X^2/n)^2/(n-1) + (S_Y^2/m)^2/(m-1))$

Gauß/t-Test

Verbundene Stichproben

Bei unabhängigen Stichproben: Separate, unabhängige Stichproben; in getrennten Teilpopulationen.

Jetzt:

X und Y an gleichen Einheiten erhoben; meist Vorher-nachher-Situation bzw. wiederholte Messungen. I.d.R. sind Vergleiche von Lage-Parametern (insbes. Erwartungswerte) interessant.

Gauß/t-Test

Verbundene Stichproben

Annahmen:

Stichprobenpaare $(X_1, Y_1), \dots, (X_n, Y_n)$ unabhängig, aber X_i und Y_i , $i = 1, \dots, n$ jeweils abhängig.

Idee:

Zurückführung auf Ein-Stichproben-Fall durch Übergang zu Differenzen

$$\begin{aligned} D_i &= X_i - Y_i, \quad i = 1, \dots, n \\ \Rightarrow D_1, \dots, D_n &\text{ i.i.d. wie } D = X - Y \end{aligned}$$

Damit: $H_0: E(X) - E(Y) = \delta_0 \Leftrightarrow H_0: E(D) = \delta_0$

\Rightarrow Ein-Stichproben-Tests für Erwartungswert anwendbar.

Tests auf Unabhängigkeit

Erinnerung: Kontingenztafeln

- ▶ Wie betrachten zwei diskrete Merkmale X und Y mit den möglichen Ausprägungen a_1, \dots, a_k für X und b_1, \dots, b_m für Y .
- ▶ In der Urliste liegen für jedes Objekt die gemeinsamen Messwerte vor, d.h. man erhält die Tupel $(x_1, y_1), \dots, (x_n, y_n)$.
- ▶ Eine $(k \times m)$ -Kontingenztafel der absoluten Häufigkeiten besitzt die Form

	b_1	\dots	b_m	
a_1	h_{11}	\dots	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	\dots	h_{2m}	$h_{2\cdot}$
\vdots	\vdots		\vdots	\vdots
a_k	h_{k1}	\dots	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	\dots	$h_{\cdot m}$	n

Dabei bezeichnen

$h_{ij} = h(a_i, b_j)$ die absolute Häufigkeit der Kombination (a_i, b_j) ,
 $h_{1\cdot}, \dots, h_{k\cdot}$ die Randhäufigkeiten von X und
 $h_{\cdot 1}, \dots, h_{\cdot m}$ die Randhäufigkeiten von Y .

Tests auf Unabhängigkeit

Hypothesen

Wir fassen die Merkmale X und Y als Zufallsvariablen auf und fragen uns:

Sind X und Y unabhängig?

Die Unabhängigkeit von X und Y formulieren wir als Nullhypothese

$$H_0 : P(X = a_i, Y = b_j) = P(X = a_i) \cdot P(Y = b_j) \text{ für alle } i, j.$$

Die Abhängigkeit entspricht der Alternativhypothese

$$H_1 : P(X = a_i, Y = b_j) \neq P(X = a_i) \cdot P(Y = b_j) \text{ für mindestens ein Paar } (i, j).$$

Idee: Wir lehnen H_0 ab, und entscheiden uns für H_1 , falls die χ^2 -Statistik

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{\left(h_{ij} - \frac{h_{i.} \cdot h_{.j}}{n} \right)^2}{\frac{h_{i.} \cdot h_{.j}}{n}}$$

einen großen Wert annimmt.

Tests auf Unabhängigkeit

Nochmal: Grundgedanken zu Testtheorie

		Entscheidung für	
		H_0	H_1
H_0 wahr		richtig	Fehler 1. Art (α -Fehler)
H_1 wahr	Fehler 2. Art (β -Fehler)		richtig

Bei einem statistischen Test kontrolliert man die Wahrscheinlichkeit für einen Fehler 1. Art:

Für einen statistischen Test zum Signifikanzniveau α , mit $0 < \alpha < 1$, gilt

$$P(H_1 \text{ annehmen} \mid H_0 \text{ wahr}) \leq \alpha,$$

d.h.

$$P(\text{Fehler 1. Art}) \leq \alpha.$$

Typische Werte für das Signifikanzniveau α sind 0.1, 0.05, 0.01.

Tests auf Unabhängigkeit

Der χ^2 Test

Wir gehen davon aus, dass unabhängige Stichprobenvariablen (X_i, Y_i) , $i = 1, \dots, n$, gruppiert in einer $(k \times m)$ -Kontingenztafel vorliegen.

- ▶ Unter H_0 , d.h. falls die Nullhypothese “ X und Y sind unabhängig” zutrifft, gilt:

Die χ^2 -Statistik folgt für großes n näherungsweise einer χ^2 -Verteilung mit $(k - 1)(m - 1)$ Freiheitsgraden.

- ▶ Wir lehnen daher H_0 ab, falls

$$\chi^2 > \chi^2_{1-\alpha}((k - 1)(m - 1)),$$

wobei $\chi^2_{1-\alpha}((k - 1)(m - 1))$ das $(1 - \alpha)$ -Quantil der χ^2 -Verteilung mit $(k - 1)(m - 1)$ Freiheitsgraden bezeichnet.

Regressionsanalyse

Einführung

- ▶ Ziel: Analyse des Einflusses einer oder mehrerer Variablen X_1, \dots, X_p auf eine Zielvariable Y .
- ▶ Bezeichnungen:
 X_1, \dots, X_p erklärende Variablen (unabhängige Variablen, exogene Variablen, Kovariablen, Regressoren, Prädiktoren)
 Y Zielvariable (zu erklärende Variable, abhängige Variable, endogene Variable, Regressand, Response)
- ▶ Verschiedene Arten von Regressionsmodellen, abhängig vom Typ der Zielvariable Y und der Art des Einflusses von X_1, \dots, X_p .
- ▶ Hier: Y metrisch/stetig.

Lineare Einfachregression

Einführung

Datensituation wie beim Streudiagramm:

(y_i, x_i) , $i = 1, \dots, n$, Beobachtungen für stetige bzw. metrische Merkmale Y und X .

Beispiele:

- ▶ Wachstum von Pflanzen oder Tieren, z.B.
 - ▶ Y Stammumfang eines Baumes, X Alter.
 - ▶ Y Höhe einer Pflanze, X Strahlendosis.
 - ▶ Y Gewicht eines Schweines, X Alter.
 - ▶ ...
- ▶ Wirkung eines Medikaments, z.B. Y Blutdruck, X Dosierung.
- ▶ ...

Lineare Einfachregression

Einführung

- ▶ Zusammenhang zwischen Y und X nicht deterministisch, sondern durch (zufällige) Fehler additiv überlagert.

$$Y = f(x) + \epsilon,$$

wobei f deterministische Funktion, ϵ additiver Fehler.

- ▶ Lineare Einfachregression: f linear, d.h.

$$Y = \alpha + \beta x + \epsilon.$$

- ▶ Primäres Ziel: Schätze α und β aus Daten (y_i, x_i) , $i = 1, \dots, n$.
Unterstelle dabei lineare Beziehung

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

wobei $\alpha + \beta x_i$ systematische Komponente, ϵ_i zufällige Fehler mit $E(\epsilon_i) = 0$.

Weitere Annahmen an die Fehler ϵ_i :

$$\epsilon_i \text{ i.i.d. mit } \sigma^2 = \text{Var}(\epsilon_i)$$

Lineare Einfachregression

Einführung

Standardmodell der linearen Einfachregression:

Es gilt

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n.$$

Dabei sind:

Y_1, \dots, Y_n beobachtbare metrische Zufallsvariablen,

x_1, \dots, x_n gegebene deterministische Werte oder Realisierungen einer metrischen Zufallsvariable X .

$\epsilon_1, \dots, \epsilon_n$ unbeobachtbare Zufallsvariablen, die unabhängig und identisch verteilt sind mit $E(\epsilon_i) = 0$ und $Var(\epsilon_i) = \sigma^2$.

Die Regressionskoeffizienten α, β und die Varianz σ^2 sind unbekannte Parameter, die aus den Daten $(y_i, x_i), i = 1, \dots, n$, zu schätzen sind.

Lineare Einfachregression

Einführung

Bemerkungen:

- ▶ Falls Regressoren nicht deterministisch sondern stochastisch, bedingte Betrachtungsweise, d.h. Modell und Annahmen unter der Bedingung $X_i = x_i$, $i = 1, \dots, n$.
- ▶ Eigenschaften der Zielvariablen:

$$E(Y_i \mid x_i) = E(\alpha + \beta x_i + \epsilon_i) = \alpha + \beta x_i$$

$$\text{Var}(Y_i \mid x_i) = \text{Var}(\alpha + \beta x_i + \epsilon_i) = \text{Var}(\epsilon_i) = \sigma^2$$

$$Y_i \mid x_i, i = 1, \dots, n, \text{ unabhängig}$$

- ▶ Oft zusätzlich Normalverteilungsannahme:

$$\epsilon_i \sim N(0, \sigma^2) \quad \text{bzw.} \quad Y_i \mid x_i \sim N(\alpha + \beta x_i, \sigma^2)$$

Lineare Einfachregression

Schätzen, Testen und Prognose

Ziele:

- ▶ Schätzung von α , β und σ^2 .
- ▶ Testen von Hypothesen über α und v.a. β .
- ▶ Prognose von Y für neuen Wert x des Regressors X .

Schätzen:

KQ-(Kleinste-Quadrate-)Methode: Bestimme Schätzer $\hat{\alpha}, \hat{\beta}$ so, dass

$$\sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2 \rightarrow \min_{\alpha, \beta}.$$

Lineare Einfachregression

Schätzen, Testen und Prognose

Schätzer für die Steigung β :

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Schätzer für die Konstante/Intercept α :

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

Schätzer für die Varianz σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - (\hat{\alpha} + \hat{\beta}x_i))^2$$

Lineare Einfachregression

Schätzen, Testen und Prognose

Geschätzte Regressionsgerade (Ausgleichsgerade):

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x$$

Geschätzte Fehler, Residuen:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$$

Prognose für neuen Wert x_0 :

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta}x_0$$

Lineare Einfachregression

Schätzen, Testen und Prognose

Bestimmtheitsmaß

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

aus Streuungszerlegung (Quadratsummenzerlegung):

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{SQT} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{SQE} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{SQR}$$

SQT : Gesamtabweichungsquadratsumme in Y -Richtung

SQE : Durch die Regression erklärter Teil von SQT

SQR : Trotz der Regression unerklärt bleibender Teil von SQT

Lineare Einfachregression

Schätzen, Testen und Prognose

Teststatistiken T_{α_0} und T_{β_0} zum Testen von Hypothesen bzgl. α und β :

$$T_{\alpha_0} = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}_{\hat{\alpha}}} \quad \text{und} \quad T_{\beta_0} = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}$$

Hypothesen		Ablehnbereich
$H_0 : \alpha = \alpha_0$	vs. $H_1 : \alpha \neq \alpha_0$	$ T_{\alpha_0} > t_{1-\alpha/2}(n-2)$
$H_0 : \beta = \beta_0$	vs. $H_1 : \beta \neq \beta_0$	$ T_{\beta_0} > t_{1-\alpha/2}(n-2)$
$H_0 : \alpha \geq \alpha_0$	vs. $H_1 : \alpha < \alpha_0$	$T_{\alpha_0} < -t_{1-\alpha}(n-2)$
$H_0 : \beta \geq \beta_0$	vs. $H_1 : \beta < \beta_0$	$T_{\beta_0} < -t_{1-\alpha}(n-2)$
$H_0 : \alpha \leq \alpha_0$	vs. $H_1 : \alpha > \alpha_0$	$T_{\alpha_0} > t_{1-\alpha}(n-2)$
$H_0 : \beta \leq \beta_0$	vs. $H_1 : \beta > \beta_0$	$T_{\beta_0} > t_{1-\alpha}(n-2)$

Multiple lineare Regression

Einführung

Ziel: Erweiterung der linearen Einfachregression für mehrere Kovariablen X_1, \dots, X_p

Daten: $(y_i, x_{i1}, \dots, x_{ip}), i = 1, \dots, n$

Zielvariable Y : metrisch bzw. stetig

Kovariablen: metrisch oder kategorial

- ▶ Metrische Kovariable x kann auch Transformation $x = f(z)$ einer ursprünglichen erklärenden Variablen z sein, z.B. $x = z^2$, $x = \log(z)$, usw.
- ▶ Kategorialer Regressor mit k Kategorien $1, \dots, k$ durch $k - 1$ Dummy-Variablen $x^{(1)}, \dots, x^{(k-1)}$ kodiert; mit k als Referenzkategorie.

Multiple linear Regression

Einführung

Dummy-Kodierung

$$x^{(j)} = \begin{cases} 1, & \text{falls Kategorie } j \text{ vorliegt,} \\ 0, & \text{sonst,} \end{cases}$$

wobei $j = 1, \dots, k - 1$.

$x^{(1)} = \dots = x^{(k-1)} = 0 \quad \Leftrightarrow \quad \text{Referenzkategorie } k \text{ liegt vor.}$

Multiple lineare Regression

Einführung

Standardmodell der multiplen linearen Regression

Es gilt

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n.$$

Dabei sind

Y_1, \dots, Y_n	beobachtbare metrische Zufallsvariablen,
x_{1j}, \dots, x_{nj}	deterministische Werte der Variablen X_j oder Realisierungen von Zufallsvariablen X_j ,
$\epsilon_1, \dots, \epsilon_n$	unbeobachtbare Zufallsvariablen, die unabhängig und identisch verteilt sind mit $E(\epsilon_i) = 0$ und $\text{Var}(\epsilon_i) = \sigma^2$.

Bei Normalverteilungsannahme:

$$\epsilon_i \sim N(0, \sigma^2) \Leftrightarrow Y_i \mid x_{i1}, \dots, x_{ip} \sim N(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2)$$

Multiple linear Regression

Einführung

Matrixnotation

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Y Beobachtungsvektor der Zielvariablen, X Designmatrix

$Y = X\beta + \epsilon$, $E(\epsilon) = 0$; Annahme: Rang von $X = p + 1$

Multiple linear Regression

Schätzen, Testen und Prognose

Schätzer $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)^\top$ nach dem KQ-Prinzip

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 = (Y - X\beta)^\top (Y - X\beta) \rightarrow \min_{\beta}$$

Lösung: KQ-Schätzer

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Gefittete Werte:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}$$

Residuen:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n.$$

Multiple linear Regression

Schätzen, Testen und Prognose

- Einfache Teststatistiken:

$$T_j = \frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma}_j}, \quad j = 0, \dots, p$$

Hypothesen und Ablehnbereiche:

Hypothesen		Ablehnbereich
$H_0 : \beta_j = \beta_{0j}$	vs. $H_1 : \beta_j \neq \beta_{0j}$	$ T_j > t_{1-\frac{\alpha}{2}}(n-p-1)$
$H_0 : \beta_j \geq \beta_{0j}$	vs. $H_1 : \beta_j < \beta_{0j}$	$T_j < -t_{1-\alpha}(n-p-1)$
$H_0 : \beta_j \leq \beta_{0j}$	vs. $H_1 : \beta_j > \beta_{0j}$	$T_j > t_{1-\alpha}(n-p-1)$

- Overall-F-Test zum Testen der Hypothesen

$$H_0 : \beta_1 = \dots = \beta_p = 0,$$

$$H_1 : \beta_j \neq 0 \quad \text{für mindestens ein } j.$$

Multiple lineare Regression

Schätzen, Testen und Prognose

Prognose:

$$\hat{Y}_0 = x_0^\top \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_p x_{0p},$$

mit $x_0 = (1, x_{01}, \dots, x_{0p})^\top$ als neuem Kovariablenvektor.

Varianzanalyse (ANOVA)

Situation: Alle unabhängigen Variablen sind kategorial, die Zielgröße Y ist metrisch/stetig.

- ▶ **Einfaktorielle Varianzanalyse:** Eine unabhängige Variable (Faktor) mit Stufen $i = 1, \dots, I$.

Modell:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, n_i,$$

wobei $\epsilon_{ij} \sim N(0, \sigma^2)$.

Frage: Unterscheidet sich der Erwartungswert von Y zwischen den Faktorstufen, d.h.

$$\mu_1 = \mu_2 = \dots = \mu_I ?$$

- ▶ **Mehrfaktorielle Varianzanalyse:** Betrachte nicht nur einen Faktor sondern mehrere.

Ausblick

Nichtparametrische Regression

Nichtparametrische Regression flexibler als parametrische: Keine parametrische funktionale Form postuliert; nur qualitativ-strukturelle Annahmen.

Beispiel: Additives Modell

$$Y = f_1(X_1) + f_2(X_2) + \beta_1 Z_1 + \dots + \beta_p Z_p + \epsilon$$

mit f_1, f_2, \dots als glatte, unbekannte Funktionen, die aus den Daten “nichtparametrisch” geschätzt werden.

Ausblick

Generalisierte Regression

Situation: Y ist nicht mehr normalverteilt, sondern z.B. binär.

- ▶ Lineares Modell wie bisher nicht mehr tauglich.
- ▶ Spezifiziere (generalisiertes lineares Modell)

$$E(Y|X_1 = x_1, \dots, X_p = x_p) = h(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p),$$

mit bekannter “Responsefunktion” h .

- ▶ Wichtiger Spezialfall für binäres $Y \in \{0, 1\}$: Logistisches Modell

$$\begin{aligned} E(Y|X_1 = x_1, \dots, X_p = x_p) &= P(Y = 1|X_1 = x_1, \dots, X_p = x_p) \\ &= \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}. \end{aligned}$$