

# Lecture script to Statistical learning

held by David Petroff  
typeset by Daniel Mayer  
University of Leipzig

November 9, 2017

## 1 Vorbemerkungen

Bei statistischem Lernen geht es darum intelligente Schlüsse aus Daten zu ziehen. Es muss aber nicht unbedingt nur um Daten gehen, wobei der Fokus der Vorlesung auf die Methoden zur Analyse von Daten gelegt wird..

Es wird wenig über Design von Versuchen gehen, also die Art und Konzeption der Datenerhebung zum Beispiel einer klinischen Studie etc. → hier geht es um das Werkzeug der Analyse.

Es wird einige Beispiele aus Petroffs Forschung geben, also aus klinischen Studien, aber es gibt natürlich auch Anwendungen von statistischem Lernen auf ganz anderen Gebieten.

### 1.0.1 beispielhafte anwendungen

Die Frage ob sich Behandlungen A und B unterscheiden

Was sind die Eigenschaften eines diagnostischen Tests (siehe: bedingte Wahrscheinlichkeiten, z.B. die Frage 'wie hoch ist die Wahrscheinlichkeit das jemand tatsächlich Hepatitis A hat, wenn ein Test positiv ausfällt')

oder: 'Gibt es einen Zusammenhang zwischen Krankheiten A und B.'

## 1.1 Wahrscheinlichkeiten

### 1.1.1 Zugänge

Es gibt zwei Zugänge zu Statistik, der eine behandelt relative Häufigkeiten (*frequentistische Statistik*), der andere behandelt das Maß für eine Überzeugung (*Bayes'sche Statistik*)

**frequentistisch** Basiert auf der Idee von wiederholbaren Experimenten (Münzwurf, radioaktiver Zerfall, Schwangerschaft bei Kontrazeptionsmethode A (Verhütung), 5 Jahre überleben nach einer Chemotherapie (aber was definieren wir als Experiment?: Krebsstadium?, Krebsart?, Behandlungsdauer?), Wahrscheinlichkeit eines Regentages

etc.). Wir sehen die Idee der Wiederholbarkeit ist nicht immer einfach festzustellen. in den ersten Vorlesungen folgen wir einem Traditionellen zugang, dadurch bekommt man ein solides fundament.

Dieser Zugang wurde von Kolmogorow gelegt, die entsprechende Axiomatik der klassischen Theorie ist die *Kolmogorow Axiomatik*.

Wir werden aus zeitgründen nicht mathematisch streng sein können.

### 1.1.2 Das Ereignisfeld

Als *Ereignis* bezeichnet man einen möglichen ausgang eines 'Zufallsexperiments' zb: "Zahl liegt oben" beim Münzwurf.

Ein System heißt Ereignisfeld, wenn:

1. es das Sichere und das unmögliche Ereignis enthält
2. A und B Teil eines Systems sind, dann auch
  - (i)  $AB$  (auch  $A \cap B$  geschrieben) " *Produkt*" von A und B bedeutet gleichzeitiges auftreten von A und B
  - (ii)  $A+B$  ( $A \cup B$ ) " *Summe*", mindestens eines der Ereignisse A und B tritt ein
  - (iii)  $A-B$  ( $A \setminus B$ ) " *Differenz*" A tritt ein, während B nicht eintritt.

**Beispiel 1.** Münzwurf-Ereignisfeld  $\{A, B, \Omega, \emptyset\}$

wobei:

A - Zahl oben

B - Wappen Oben

$\Omega$  - Zahl oder Wappen oben

$\emptyset$  weder zahl noch wappen, oder auch: sowohl wappen als auch zahl, umfasst also ALLE unmöglichen Ereignisse

### 1.1.3 Gesetze der Ereignisse

#### Kommutativität

$$A + B = B + A$$

$$AB = BA$$

#### Assoziativität

$$(A + B) + C = A + (B + C)$$

$$(AB)C = A(BC)$$

#### Distributivität

$$A(B + C) = AB + AC$$

$$A + (BC) = (A + B)(A + C)$$

was durch die identitäten klar wird...

## Identitäten

$$A + A = A$$

$$AA = A$$

wir beweisen also das distributivgesetz wie folgt:

$$(A + B)(A + C) = AA + AC + BA + BC = A + BC$$

## 1.2 Wahrscheinlichkeitsbegriff

**Axiom 1.1.** Jedes Ereignis aus dem Ereignisfeld  $F$  ordnet man eine nichtnegative Zahl  $p(A)$  zu, die Wahrscheinlichkeit.

**Axiom 1.2.**  $P(\Omega) = 1$

**Axiom 1.3.** Sind Ereignisse  $A_i$  unvereinbar, ie  $A_i A_j = \emptyset$  für  $i \neq j$ , so ist  $P(A_1, A_2, \dots, A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$ , und es gelten folgende Eigenschaften für Wahrscheinlichkeiten:

(a)  $P(\emptyset) = 0$

(b)  $P(\bar{A}) = 1 - P(A)$ ,  $\bar{A} := \Omega - A$

(c)  $0 \leq P(A) \leq 1$

(d) Für  $A \subset B$  ( $A$  ist teilmenge von  $B$ ) folgt  $P(A) \leq P(B)$

(e)  $P(A + B) = P(A) + P(B) - P(AB)$

(f)  $P(A_1 + A_2 + \dots + A_n) \leq P(A_1) + P(A_2) + \dots + P(A_n)$

### 1.2.1 Bedingte Wahrscheinlichkeiten

Die Wahrscheinlichkeit von  $A$  unter der Bedingung dass  $B$  eingetreten ist schreibt man  $P(A|B)$

$$P(A|B) := \frac{P(AB)}{P(B)} \quad (1)$$

Motivation: gegeben seien  $n$  unvereinbare gleichwahrscheinliche Ereignisse  $A_1, A_2, \dots, A_n$  mit  $m$  günstig für  $A$ ,  $k$  günstig für  $B$ , und  $r$  günstig für  $AB$ :

$$P(A|B) = \frac{r}{k} = \frac{r/n}{k/n} = \frac{P(AB)}{P(B)} \quad (2)$$

**Beispiel 1.** Zwei würfel werden geworfen. Wie groß ist die Wahrscheinlichkeit, die Summe 8 zu erhalten (Ereignis  $A$ ), falls bekannt ist, dass die summe grade ist (Ereignis  $B$ )

$$P(A) = 5/36 \quad P(B) = 1/2, \quad P(AB) = 5/36, \quad P(A|B) = \frac{P(AB)}{P(B)} = 5/18$$

### 1.2.2 Bayes'sche Formel

Seien  $A_1, A_2, \dots, A_n$  unvereinbar, So kann man die bedingte Wahrscheinlichkeit schreiben als:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)} \quad (3)$$

mit  $\bigcup A_i = \Omega$  (das wurde nachträglich eingefügt)

### 1.2.3 Diagnostische Verfahren - Anwendung von bedingter Wahrscheinlichkeit

Es seien  $D^+, D^-$  zwei Mögliche Krankheitszustände (Diseases, wobei krank  $D^+$  ist) und  $T^+, T^-$  die zwei möglichen Ergebnisse eines diagnostischen Tests (bei Tests wie einem Schwangerschaftstest macht Binariät Sinn, bei Tests wie dem von Leberwerten, ist die Binariät (ob sinnvoll oder nicht), durch eine Grenzziehung hergestellt.) So bezeichnet man

$P(D^+)$  als die **Prävalenz** (Wahrscheinlichkeit krank zu sein),

$P(T^+|D^+)$  die **Sensitivität**, sowie

$P(T^-|D^-)$  als die **Spezifität**.

$P(D^+|T^+)$  heißt der **positiv prediktiver Wert** (PPV) also die Wahrscheinlichkeit das der Patient krank ist wenn der test positiv ausfällt, sowie

$P(D^-|T^-)$  der **negativ prediktiver Wert** (NPV), also die Wahrscheinlichkeit, dass ein negativer Test tatsächlich bedeutet, dass der Patient gesund ist.

## 1.3 Zufallsvariablen und Verteilungsfunktionen

qualitative beschreibung aus Gnedenko:

'eine Zufallsgröße, (auch Zufallsvariable) ist eine Größe, deren Wert vom Zufall abhängen, und für die eine Wahrscheinlichkeitsverteilungsfunktion existiert'

Jedem Elementarereignis (unzerteilbar)  $\omega \in \Omega$ , wird eine reelle Zahl zugeordnet:

$X = X(\omega) : \Omega \rightarrow \mathbb{R}$ .

$F_x(t) := P(X < t)$  wird als Verteilungsfunktion der Zufallsgröße  $x$  definiert. Sie ist monoton nicht fallend, linksseitig stetig und gehorcht den Bedingungen:  $F(-\infty) = 0$   $F(\infty) = 1$

umkehrung: jede solcher funktionen lässt sich als Verteilungsfunktion einer Zufallsgröße deuten.

## 1.4 Wichtige Verteilungsfunktionen

### Binomialverteilung

$$P_n(m) = \binom{n}{m} p^m q^{n-m} \quad (4)$$

wobei  $q := 1 - p$

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \sum_{k < x} P_k & \text{for } 0 < x \leq n \\ 1 & \text{for } x > n \end{cases} \quad (5)$$

### Poisson Verteilung

$$P_n = \frac{\lambda^n e^{-\lambda}}{n!}, \quad \lambda > 0$$
$$F_x(t) = \sum_{k=0}^t \frac{\lambda^k}{k!} e^{-\lambda}, \quad t \in \mathbb{R}, n \in \mathbb{N} \quad (6)$$

### Normalverteilung

$$F(x) = \Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x x e^{-\frac{(z-a)^2}{2\sigma^2}} dz \quad \sigma > 0 \quad (7)$$

## 1.5 Erwartungswert, Varianz und weitere Momente

Erwartungswert  $E(X)$  einer Zufallsgröße.

diskret:

$$E(X) = \sum_i x_i P_i$$

#### Beispiel 1. Würfel

$$E(X) = \frac{1}{6} \sum_i i = \frac{21}{6} = 7/2$$

#### Beispiel 2. Binomialverteilung

$$E(X) = \sum_{k=0}^n k P_n(k) = \sum k \binom{n}{k} p^k (1-p)^{n-k}$$

Nebenrechnung:

$$k \binom{n}{k} = \frac{kn!}{k!(n-k)!} = \frac{n(n-1)!}{(k-1)!((n-1)-(k-1))!} = n \binom{n-1}{k-1}$$

Deshalb:

$$E(x) = n \sum_{k=1}^n \binom{n-1}{k-1} p^k (1-p)^{n-k} = np \sum_{k=i}^n p^{k-1} (1-p)^{n-k}$$

neue indizes:  $k' = k - 1, \quad n = n - 1$

$$E(X) = np \underbrace{\sum_{k'=0}^{n'} \binom{n}{k'} p^{k'} (1-p)^{n'-k'}}_{=1}$$

thusly:

$$E(X) = np$$

(die varianz braucht eine ähnliche herleitung)

stetiger fall:

$$E(X) = \int x p(x) dx$$

wobei  $p(x)$  die Wahrscheinlichkeitsdichte ist.

**Beispiel 3.** Wahrscheinlichkeitsverteilung auf dem intervall  $[a, b]$

$$E(X) = \frac{1}{b-a} \int_a^b x dx = \left[ \frac{1}{2(b-a)} x^2 \right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{1}{2}(b+a) \quad (8)$$

**Beispiel 4.** Normalverteilung:

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-(x-a)^2/2\sigma^2} dx$$

substitute  $x' = \frac{x-a}{\sigma}$  thus:

$$x = \sigma x' + a, \quad dx = \sigma dx' \quad (9)$$

$$E(X) = \frac{1}{\sqrt{2\pi}} \int (\sigma x' + a) e^{-x'^2/2} dx' \quad (10)$$

ungerade funktion ergibt 0

$$E(X) = \frac{a}{\sqrt{2\pi}} \int e^{-x'^2/2} dx' = a$$

### 1.5.1 Varianz (auch Dispersion)

$$V(X) = E[(X - E(X))^2]$$

diskret:

$$V(X) = \sum_i [X_i - E(X)]^2 P(X_i)$$

Stetig:

$$V(X) = \int (x - E(X))^2 p(x) dx$$

**repetition of last class:**

somehow cryptic, there might be something missing...

$$iP_n(i) = npP_{n'}(i') \quad (11)$$

$$E(X) = \sum_{i=1}^n iP_n(i) \quad (12)$$

$$= \sum_{i=1}^n iP_n(i) = np \sum_{i'=0}^{n'} P_{i'} = np \quad (13)$$

$$V(x) = \sum_{i=0}^n (i - np)^2 P_n(i) = (np)^2 \underbrace{\sum_{i=0}^n P_n(i)}_{=1} - 2np \underbrace{\sum_{i=0}^n iP_n(i)}_{=np} + \sum_{i=1}^n i^2 P_n(i) \quad (14)$$

$$\begin{aligned}
\Rightarrow V(X) &= \sum_{i=1}^n i^2 P_n(i) - (np)^2 = \sum_{i=1}^n (i-1+1)iP_n(i) - (np)^2 \\
&= np \sum_{i'=0}^{n'} (i'+1)P_{n'}(i') - (np)^2 \\
&= np \left( \sum_{i'=0}^{n'} i' P_{n'}(i') + \sum_{i'=0}^{n'} P_{n'}(i') \right) - (np)^2 \\
&= np(n'p + 1) - (np)^2 = np((n-1)p + 1) - (np)^2 = np(1-p)
\end{aligned} \tag{15}$$

**Beispiel 5.** Würfel:

$$V(X) = \frac{1}{6} \sum_{i=1}^6 \left(i - \frac{7}{2}\right)^2 = \frac{1}{3} \sum_{i=1}^3 \left(i - \frac{7}{2}\right)^2 = \frac{1}{3} \left[ \left(\frac{5}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right] = \frac{35}{12} \tag{16}$$

**Beispiel 6.** Uniformfverteilung  $[a, b]$

$$\begin{aligned}
V(X) &= \frac{1}{b-a} \int_a^b x^2 dx - \left( \frac{(b+a)}{2} \right)^2 = \\
&\quad \frac{b^3 - a^3}{3(b-a)} - \frac{(b+a)^2}{4} = \\
&\quad \frac{(b-a)(b^2 + a^2 + ab)}{3(b-a)} - \frac{(b-a)^2}{4} \tag{17}
\end{aligned}$$

$$= \frac{1}{12} (4b^2 + 4a^2 + 4ab - 4b^2 - 6ab - 3a^2) = \frac{1}{12} (b^2 + a^2 - 2ab) = \frac{(b-a)^2}{12} \tag{18}$$

### 1.5.2 Gewöhnliches Moment

Wir bezeichnen  $m_k$  als das gewöhnliche Moment (oder auch Anfangsmoment) k-ter Ordnung

$$m_k := E(X^k) \quad \text{diskret also } \sum_i (x_i)^k p_i \quad \text{und stetig: } \int x^k p(x) dx \tag{19}$$

Das Zentrale moment (auf das Znetrum  $E(X)$  bezogen) k'ter ordnung ist

$$\mu_k := E \left[ (X - m_1)^k \right] \tag{20}$$

Die Varianz ist also das zweite Zentralmoment:

$$V(X) = \mu_2 = m_2 - (m_1)^2 \tag{21}$$

man kann immer  $\mu_k$  durch  $m_l$  ( $l \leq k$ ) ausdrücken

## 1.6 1.6 Korrelation

Eine Erweiterung dieser Momente stellt die *Kovarianz* dar:

$$b(X, Y) := E[(X - E(X))(Y - E(Y))] \quad (22)$$

Sie ist das gemischte Zentralmomente zweiter Ordnung.

Es gilt offensichtlich:

$$b(X, X) = V(X) \quad (23)$$

Die normierte Größe  $\rho(X, Y) := \rho_{X,Y} = \frac{b(X,Y)}{\sqrt{V(X)V(Y)}}$  bezeichnet man als Korrelationskoeffizient.

Es gilt:  $-1 \leq \rho \leq 1$ .

Für  $X = Y$  gilt  $\rho = 1$  und

für  $X = -Y$  gilt  $\rho = -1$ .

Falls  $X$  und  $Y$  unabhängig sind, dann gilt  $\rho = 0$

(aber nicht notwendigerweise umgekehrt, die Abhängigkeit könnte nichtlinear sein, aber generelle unabhängige Funktionen sind natürlich auch linear unabhängig)

## Anwendung auf Wahrscheinlichkeiten

$$E(X) = p_x \quad (24)$$

$$V(X) = p_x(1 - p_x) \quad (25)$$

$$\rho_{x,y} = \frac{p_{x,y} - p_x p_y}{\sqrt{p_x(1 - p_x) + p_x(1 - p_y)}} \quad (26)$$

$$\Rightarrow p_{xy} = p_x p_y + \rho_{x,y} \sqrt{p_x(1 - p_x)p_y(1 - p_y)} \quad (27)$$

grenzfälle:  $\rho = 0$ :  $p_{xy} = p_x p_y$

$\rho = 1$ : dann  $p_{xy} = (p - x^2 + p_x(1 - p_x)) = p_x$

$\rho = -1$  ( $p_x = 1 - p_y$ )  $\Rightarrow p_{xy} = P_x(1 - p_x) - p_x(1 - p_x) = 0$

## 1.7 1.7 Einige Wichtige Sätze der Wahrscheinlichkeitstheorie

**Gesetz der Großen Zahlen** Bernoulli: Für alle  $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{\mu}{n} - p\right| < \epsilon\right\} = 1 \quad (28)$$

Wobei  $\mu$  Anzahl der Ereignisse,  $n$  Anzahl der Versuche,  $p$  Wahrscheinlichkeit des Ereignisses

(streng mathematisch ist das nicht korrekt sondern bedarf erst noch einem Beweis den Borel wesentlich später gemacht hat, siehe Literatur)

Tschepyschew:

(man kennt ihn in der Informatik wegen der Tschebischew polynome)

für alle  $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i)\right| < \epsilon\right\} = 1 \quad (29)$$

für eine Folge paarweise verschiedener unabhängiger Zufallsgrößen  $\{X_i\}_{i=1,2,\dots,n}$  mit gleichmäßig beschränkter Varianz:  $\forall i \quad V(X_i) \leq C$



### 1.7.1 Lokaler Grenzwert von Moivre Laplace

Sei  $0 < p < 1$  die Wahrscheinlichkeit eines Ereignisses, dann wissen wir: In  $n$  Versuchen gilt,  $P(n) = \binom{n}{m} p^m (1-p)^{n-m}$  so gilt:

$$\lim_{n \rightarrow \infty} \frac{\sqrt{np(1-p)} P_n(m)}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} \rightarrow 1 \quad \text{mit } x = \frac{m - np}{\sqrt{np(1-p)}} \quad (30)$$

man normiert die Binomialverteilung und im Grenzfalle wird sie zur Normalverteilung.

### 1.7.2 Zentraler Grenzwertsatz

Sei  $S_n = \sum_{i=1}^n X_i$  mit  $E(X_i) < \infty$ ,  $V(X_i) = \sigma^2 < \infty$   
so gilt für jedes  $t$

$$\lim_{n \rightarrow \infty} P\left(\frac{S_n - nE(X_i)}{\sqrt{n}\sigma} < t\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{x^2}{2}} dx \quad (31)$$

also die Folge der Verteilung der standardisierten Zufallsgrößen konvergiert gegen die Standardnormalverteilung, das heißt  $\mu=0$ ,  $\sigma=1$

## 2 Deskriptive Statistik

Eine Beschreibung von Daten und Kohorten ist zentral für das Verständnis einer Arbeit (Veröffentlichung)

Ziel ist es mit wenigen Kenngrößen das Wesentliche zu charakterisieren.

dazu gibt es 'Punktschätzer' für Erwartungswerte und Konfidenzintervalle (KI engl. CI) als Maß für die Genauigkeit der Schätzung.

was ist ein Konfidenzintervall

$$KI[a, b] : P(a \leq \Theta \leq b) = 1 - \alpha \quad (32)$$

Man will schätzen wie groß die Wahrscheinlichkeit eines Erwartungswertes ist.

### 2.1 Ein Merkmal

Nominale und Ordinale Größen Es gibt Größen, die sich ordnen lassen (**nominale Größe**) wie zum Beispiel das Alter oder die Größe, aber auch Eigenschaften, die sich *nicht* ordnen lassen, wie zum Beispiel bei einer genetischen Arbeit die Herkunft von Menschen (**nominale Größe**)

absolute und relative Häufigkeiten. (z.B. Häufigkeitstabellen)

meist ist es gut etwas graphisch darzustellen:

Balkendiagramme, oft mit Konfidenzintervall oder Standardfehler (KI und SE (Standarderror))

Kreisdiagramm (in den Fachzeitschriften verpönt, weil man schlecht einschätzen kann, ob etwas 20 oder 30 Prozent ist, klarer ist hier ein Balkendiagramm).

#### 2.1.1 Metrische Daten

Lagemaß:

als Mittelwert (arithmetisch ( $\frac{1}{n} \sum_{i=1}^n x_i$ ) oder geometrisch (d.h. log-Skala) ( $[\prod_{i=1}^n x_i]^{\frac{1}{n}}$ ))

- übliche und 'robuste' Methoden (z.B. wenn ein Wert stark abweicht, sonst aber alles ähnlich ist, ist der Mittelwert mit der log-Skala)

- Median und andere Quantile (verschiedene Schätzverfahren)

## 2.1.2 streumaß

standardabweichung ("sample Method")  $sd^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

mit  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

- Interquartilabstand (enl IRQ) 25 und 75 perzentil

- Spannweite (int nicht wirklich ein streumaß, wird aber oft zusätzlich verwendendet)

graphisch : Histogramm, Boxplot

## 2.2 Zusammenhang Zweier Merkmale

bei nominalen Größen arbeitet man oft mit Kontingenztafel: odds ration, relatives risiko. Graphisch auch: forest plots bei metrischen größen: Korrelationskoeffizient (min KI) , Streudiagramm

## 2.3 Simplsions Paradoxon

grundidee: ein effekt den ma in der gesamtgruppe sieht muss nicht "echt" sein, er kann in subgruppen anders ausfallen.

Beispiel 1.		A	B
	Erfolg	70 (30%)	50 (22%)
	Misserfolg	160 (70 %)	182 (82%)
		230	232

Hier würden wir sagen dass Gruppe A besere Ergebnisse hat als Gruppe B.

Wenn wir aber nun den Datensatz genauer betrachten und zwischen Männern und Frauen unterscheiden, ergibt sich ein anderes Bild:

		A	B
Männer	Erfolg	7 (20%)	45 (20%)
	Misserfolg	28 (80 %)	45 (20 %)
Frauen	Erfolg	63 (32 %)	5 (33%)
	Misserfolg	132 (68%)	10 (67%)

Hier sehen wir, dass die Aussage für Männer auf alle Fälle Falsch ist, es für Frauen keine Unterschied zwischen den beiden Gruppen gibt.

## 3 Statistisches Testen

### 3.1 Die Logik des Testens

Die Analogie zum beweis durch widerspruch kann hilfreich sein, hier ein beispiel:

was ist nun der zusammenhang zwischen dem ergebnis des testens und der wahrheit der 0 hypothese?

### 3.2 Der T-test

Vergleich zweier Mittelwerte  $H_0 : \mu_1 = \mu_2$

wir schätzen die "t-statistik" (annahme gleicher varianz)

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{wobei} \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

statistisches testen	Beweis durch Widerspruch
annahme $H_0 : \mu_1 = \mu_2$ i.e. Mittelwert der Gruppe 1 ist gleich dem Mittelwert der Gruppe 2	Annahme $\sqrt{2}$ ist rational
man glaubt nicht an Annahme	man glaubt nicht an Annahme
Folge: man nimmt an dass Annahme stimmt	wenn die Annahme stimmt:...
kommt etwas sehr unwahrscheinliches raus so ist die annahme nicht plausibel (korrelation < 5% $H_0$ wird abgelehnt	kommt man auf einen Widerspruch, so so muss die Annahme Falsch sein
kommt was plausibles raus, so weiß man wenig über die annahme, das Konfidenzintervall kann helfen	kommt man nicht auf einen Widerspruch, so weiß man ein wenig mehr über die Annahme

	$H_0$ stimmt	$H_0$ Stimmt nicht
$H_0$ abgelehnt	Typ I Fehler, $\alpha$ (korrelation $\alpha \leq 0.05$ )	Power: $1 - \beta$
$H_0$ nicht abgelehnt	alles so wie gewollt	Typ II Fehler, $\beta$ , Planung: $\beta = 0.1$ oder $0.2$ anstrebt ...Sicherheit über $\beta$ hat man nicht

mit  $n_i$  der stichprobengröße der  $i$ -ten gruppe  $\bar{x}_i$  Mittelwert  $\frac{1}{n_i} \sum_{j=1}^n X_j^{(i)}$

$$s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^n (X_j^{(i)} - \bar{X}_i)^2$$

( $T = \frac{\Delta x}{SE}$  ist die grundlegende Struktur

unter  $H_0$  : T hat "t-Verteilung" mit freiheitsgraden  $f = n_1 + n_2 - 2$

Im gegensatz zum normalen T-Test gibt es eine Variante des T-Tests, den Welch test, der keine Annahme über die Gleichheit der Varianz annimmt.  
T-Verteilung mit f Freiheitsgraden

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad f = \frac{(\tilde{s}_1^2 + \tilde{s}_2^2)^2}{\frac{\tilde{s}_1^4}{n_1-1} + \frac{\tilde{s}_2^4}{n_2-1}} \quad \tilde{s}_i = \frac{s_i}{n_i}$$

Konfidenzintervall für  $\Delta\mu$ :

$$\Delta\bar{x} \pm \underbrace{t_{\alpha/2, f}}_{\approx 2 \text{ für } \alpha=0} SE$$

pWert : wahrscheinlichkeit den wert T zu beobachten unter  $H_0$   
ist der p-Wert  $p_0$  so beinhaltet ein  $(1 - p_0)$ -KI gerade so den Wert Null.  
zB ist  $P = 0.05 \Rightarrow$  das 95% KI interval erreicht die 0 grade so.  
gegeben sei:  
wobei wir die notation  $n_{.j} = \sum_i n_i j$  verwenden

	A	B	
I	$n_{11}$	$n_{12}$	$n_{1\cdot}$
II	$n_{21}$	$n_{22}$	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

$\frac{n_{11}}{n_{21}}$  schätzt das Odds (die chance von "I" im Vergleich zu "II" bei der gruppe A)

$$\hat{OR} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}}$$

schätzt das odds ration (chancenverhältnis) KI:  $\hat{OR}e^{Z_{\alpha/2}SE}$  hier sit die frage ob f im intervall (null auf der log scala)

(fisher test  $\Leftrightarrow$  KI von OR (mit anderer schätzmethode allerdings)

Fisher-Test heißt "exakt", da ein strenger wert aus kombinatorik berechnet wird

$$P = \frac{\binom{n_{1\cdot}}{n_{11}} \binom{n_{2\cdot}}{n_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

wir sind skeptisch, da wir eine symmetrische formel erwarten aber der nenner nicht symmetrisch ist, wir stellen aber fest das

$$\binom{n_{\cdot\cdot}}{n_{\cdot 1}} = \binom{n_{11} + n_{22} + n_{12} + n_{21}}{n_{11} + n_{21}}$$

Teil von Kristin Reiche

## 4 Lineare regressiionsmodelle

- einfache lineare regressionsmodelle
- Multivariable lineare regressionsmoldelle
- Voraussetzungen für lineare Regressionsmodelle
- Generalisierte lineare regressionsmodelle (GLM)
- Auswertung von expressionsmodellen

Answendungen sind:

Molekularbiologische Hochdurchsatzdaten Das ziel der statistischen nmethode ist das Messen der Werte einer Zielvariablen in abhängigkeit von unabhängigen variablen ( sogenannten kovariablen)

**Definition 1.** Ein statistisches Modell stellt eine Zielvariable die meist mit  $Y$  angegeben wird in Beziehung zu einer oder mehreren Kovariaten(Kovariablen)

Zielvariable = Modell(Kovariante) + Fehler

$$Y = f(X) + \varepsilon$$

$Y$  ist hier die Zielvariable oder auch abhängige variable.

$X$  sind Kovariaten oder auch unabhängige variablen.

$f(X)$  ist die unbekannte Funktion die den systematischen effekt von  $X$  auf  $Y$  modelliert

$\varepsilon$  ist der Zufällige fehler . gibt den anteil der Varianz von  $Y$  an der nicht durch  $f(x)$  erklärt wird

⇒ Statistische modelle zerlegen die Zielvariablen in einen systematischen ( $f(X)$ ) und zufälligen Teil ( $\varepsilon$ ).

#### 4.1 Anwendungen von statistischen Modellen

**Inferenz:** Ziel ist es die Art des Zusammenhanges zwischen  $X$  und  $Y$  zu verstehen  
Genauer: Wie ändert sich  $Y$  als funktion von  $X$ .

**Vorhersage:** Ziel ist es den Wert von  $Y$  so genau wie möglich vorherzusagen. (hier ist es nicht unbedingt von Interesse die (exakte) Form von  $f(X)$  zu kennen.

##### 4.1.1 Schätzen der Funktion $f(X)$

$f(X)$  wird anhand einer statistischen lernmethode von einer Menge von trainingsdaten geschätzt.

Die geschätzte Funktion wird mit  $\hat{f}(X)$  angegeben:

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

für trainingsdaten  $(X, Y) = \{(x_1, y_1), \dots, (x_n, y_n)\}$  mit  $n$  Beobachtungen.

wenn trainingsdaten in tupeln angegeben sind sprechen wir von überwachtem lernen.

Wir differenzieren in zwei verschiedene Statistische Lernmethoden:

(i) Parametrische Methoden: Für die funktion  $f(X)$  wird eine bestimmte form angenommen.

(a) Anzahl der kovariablen wird vorab festgesetzt oder mittels vefahren der Modellsektion ausgewählt (erklärung folgt später).

(b) Es werden für die Kovarablen geschätzt

Nachteil ist weniger flexibilität und das das modell oft nicht der wahren form des zusammenhags entspricht.

Vorteil ist das nur Gewichte für Kovariablen geschätzt werden müssen dafür reicht eine geringere stichprobengröße aus.

(ii) Nichtparametrische methoden: es Wird keine Bestimmte form für  $f(X)$  vorab angenommen Es musss die form und die parameter für eine Beliebige komplexe funktion  $f(X)$  anhand der trainigsdaten geschätzt werden.

Diese Modelle sind oft sehr flexibel aber auch weniger gut interpretierbar. Oftmals ist ein größerer Stichprobenumfang notwendig.

##### 4.1.2 lineare Regressionsmodelle

Als Form für  $f(X)$  wird ein (annähernd) linearer Zusammenhang angenommen.  
Zufallsvariable  $Y$  nimmt dabei quantitative werte an

Kovariable  $X_i$  ( mit  $i = 1, \dots, P$  anzahl der Kovariablen) können quantitative oder qualitative Werte annehmen.

### 4.1.3 einfaches lineares regressionsmodell

**Definition 2.** Ein Statistisches Modell was den Wert der Zielvariablen auf basis der werte einer einzigen Kovariable  $X$ , unter der annahme eines linearen Zusammenhanges modelliert.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$\beta_0$ , mittelwert von  $Y$  falls es keinen zusammenhang gibt, sonst Schnittpunkt der y-achse

$\beta_1$  ist der Effekt der Kovariablen  $X$  auf  $Y$ , also der Anstieg in  $Y$  wenn  $X$  sich eine einheit erhöht (regressionskoeffizient).

$\varepsilon$  Fehler  $\varepsilon \approx N(0, \sigma^2)$  (notation: "folgt einer normalverteilung");  $N$  ist die normalverteilung mit in der form  $N(\mu, \sigma^2)$

anteil von  $Y$  der nicht durch  $\beta_0 + \beta_1 X$  erklärt werden kann

### 4.2 annahme für zufällige störgrößen

- (a) Alle störungen haben die gleiche Varianz (Homoskedastizität)

$$Var(\varepsilon_i) = \sigma^2$$

- (b) alle störungen sind um 0 verteilt

$$E(\varepsilon_i) = 0$$

$\Rightarrow$  einflüsse der Störgrößen heben sich im Mittel auf, dh haben keinen systematischen einfluss auf  $Y$

- (c) Störgrößen sind unabhängig untereinander:

$$Cor(\varepsilon_i, \varepsilon_j) = 0 \text{ für } i \neq j$$

### 4.3 Varianzdekomposition

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{erklärbare Varianz}} + \underbrace{\sum_{i=1}^n (\hat{y}_i - y_i)^2}_{\varepsilon = \hat{y}_i - y_i \text{ nicht erklärable Varianz}}$$

#### 4.3.1 schätzung der parameter $\beta_0$ und $\beta_1$

Methode der kleinsten quadrate : reduziere die differenz zwischen werten der zufallsvariablen  $y_i$  und den vorhergesagten werten  $\hat{y}_i$  für alle beobachtungen  $n$ .

$$RSS : \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \rightarrow \min$$

$$\beta_0 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cor(X, Y)}{Var(X)}, \quad \beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$