Lecture script to Statistical learning

held by David Petroff
typeset by Daniel Mayer
University of Leipzig

October 24, 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 Vorlesunng 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. $\rightarrow$ 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 Krankheien A un 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 Jahres 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*", mindestenns eines der Ereignisse A und B tritt ein
   (iii) A-B ($A \quad B$) " *Differenz*" A tritt ein, während B nicht eintritt.

Beispiel 1. Münzwurf-Ereignisfeld $\{A, B, \Omega, \emptyset\}$
wobei:
A - Zahl oben
B - Wappen Oben
Ω - 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 $P(A_1, A_2, \cdots, A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$, und es gelten folgende Eigenschaften für Wahrscheinlichkeiten:*

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

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

(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) $P(A_i + A_2 + \cdots + A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n)$

1.2.1 Bedingte Wahrscheinlichkeiten

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

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

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

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

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:

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

mit $\bigcup B_i = \Omega$ *(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 moglichen Ergebnisse eines diagnostischen Tests (bei Tests wie einem Schwangerschaftsstest macht Binarität 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 (Wahrscheinlichhkeit krank zu sein), $P(T^+|D^+)$ die Sensitivität, sowie $P(T^-|D^-)$ als die Spezifität. $P(D^+|T^+)$ heißt der Positiv predictive value (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 Wahrscheinlichhkeit, dass ein negativer Test tatsächlich bedeutet, dass der Patient gesund ist.

1.3 Zufallsvariablen und Verteilungsfunktionen

qualitative beschreibung aus Gnedenko:
'eine Zufallsgröße, (auch Zufallsvaiable) ist eine Größe, deren Wert vom Zufall abhängen, und für die eine Wahrscheinlichkeitsverteilungsfunktion existiert'
Jedem Elemetarereignis (unzerteilbar) $\omega \in \Omega$, wird eine reele Zahl zugeordnet:
$X = X(\omega) : \Omega \to \mathbb{R}$.
$F_x(t) := P(X < t)$ wird als Verteilungsfunktion der Zufallsgröße x definiert. Sie ist monoton nicht fallend, linksseitig stetig und gehorcht den Bedingungen: $F(-\infty) = 0 \quad F(\infty) = 1$
umkehrung: jede solcher funktionen lässt sich als Verteilungsfunktion einer Zufallsgröße deuten.

1.4 Wichtige Verteilungsfunktionen

Binomialverteilung

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

wobei $q := 1 - p$

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

Poisson Verteilung

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

Normalverteilung

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

1.5 1.5 Erwartungswer, 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 missing)

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

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:

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

wobei $p(x)$ die Wahrscheinlichkeitsdichte ist.

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

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

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:

$$x = \sigma x' + a, \quad dx = \sigma dx' \tag{9}$$

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

ungerade funktion ergibt 0

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

1.5.1 Varianz (auch Dispersion)

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

diskret:

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

Stetig:

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

repetition of last class:

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

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

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

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

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

Beispiel 5. Würfel:

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

$$V(X) = \int \left[x - E(X)\right]^2 p(x) dx \tag{17}$$

Beispiel 6. Uniformfverteilung $[a, b]$

$$V(X) = \frac{1}{b-a}\int_a^b x^2 dx - \left(\frac{b+a)}{2}\right)^2 = \frac{b^3-a^3}{3(b-a)} - \frac{(b+y)^2}{4} = \frac{(b-a)(b^2+a^2+ab)}{3(b-a)} - \frac{(b-a)^2}{4} \tag{18}$$

$$= \frac{1}{12}(4b^2+4a^2+4ab-4b^2-6ab-3a^2) = \frac{1}{12}(b^2+a^2-2ab) = \frac{b-a)^2}{12} \tag{19}$$

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

$$m_k := E(X^k) \quad \text{diskret also } \sum_i (x_i)^k p_i \quad \text{und stetig: } \int x^k p(x) dx \tag{20}$$

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

$$\mu_k := E\left[(X - m_1)^k\right] \tag{21}$$

Die Varianz ist also das zweite Zentralmoment:

$$V(X) = \mu_2 = m_2 - (m_1)^2 \tag{22}$$

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

1.6 1.6 Korrelation

Eine Erweiterung dieser momente stellt die *Kovarianz* dar:

$$b(X,Y) := E\left[(X - E(X))\,(Y - E(Y))\right] \tag{23}$$

Sie ist das gemischte Zentralmomente zweiter Ordnung.
Es gilt Offensichtlich:

$$b(X,X) = V(X) \tag{24}$$

Die normierte Größe $\rho(X,Y) := \rho_{X,Y} = \frac{b(X,Y)}{\sqrt{V(X)V(Y)}}$ bezeichnet man als Korrelationskoefffizient. 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 nichtlienar sein, aber generelle unabhängige funktionen sind natürlich aud linear unabhängig)

Anwendung auf Wahrscheinlichkeiten

$$E(X) = p_x \tag{25}$$

$$V(X) = p_x(1 - p_x) \tag{26}$$

$$\rho_{x,y} = \frac{p_{X,Y} - p_x p_y}{denom\sqrt{p_x(1-p_x) + p_x(111 - p(y)}} \tag{27}$$

$\Rightarrow$

$$p_{XY} = p_x P_y + \rho_{X,Y}\sqrt{p_x(1-p_x)p_y(1-p_y)} \tag{28}$$

grenzfälle: $\rho = 0$: $p_{XY)} = P_x P_Y$
$\rho = 1$: dann $p_{XY} = (p - x^2 + p_x(1-p_x) = p_x$
$\rho = -1 \quad (p_X = 1 - P_Y) \Rightarrow P_{XY} = P_x(1-p_x) - p_x(1-p_X) = 0$

1.7 1.7 Einige Wichtige Sätze der Wahrscheinlichkeitstheoorie

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

$$\lim_{n\to\infty} P\left\{|\frac{\mu}{n} - p| < \epsilon\right\} = 1 \tag{29}$$

wobei μ anzahl der ereignisse, n anzahl der Versuche, p Wahrscheinlichkeit des Ereignisse
(streng mathematisch ist das nicht korrekt sondern bedarf erst noch einem bewies den borel wesentlich später gemacht hat, siehe literatur)
Tschepyschew:
(man kennt ihn in der informatik von den Tschebischew polynomen)
für alle $\epsilon > 0$

$$\lim_{n\to\infty} P\left\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)| < \epsilon e\right\} = 1 \tag{30}$$

für eine Folge paarweise verschiedener unabhängiger Zufallsgrößen

$$\{X_i\}_{i=1,2,\ldots,n} \tag{31}$$

mit gleichmäßig beschränkter varianz:

$$\forall i \quad V(X_i) \leq C \tag{32}$$

1.7.1 Lokaler Grenzwert von Movre Laplace

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

$$\lim_{n\to\infty} \frac{\sqrt{np(1-p)}P_n(m)}{\frac{1}{2\pi}e^{-\frac{x^2}{2}}} \to 1 \quad \text{mit } x = \frac{m-np}{\sqrt{np(1-p)}} \tag{33}$$

man normiert die binomialverteilung und im grenzfall wird sie zutt normalverteilung.

1.7.2 Zentraler Grenzwertsatz

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

$$\lim_{n\to\infty} P(\frac{S_n - nE(X_i)}{\sqrt{\pi}\sigma} < t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx \tag{34}$$

also die folge der verteilung der standartisierten zufallsgrößen konvergiert gegen die Standartnormalverteilung, das heißt a=0, $\sigma = 1$

2 Deskriptive Statistik

Eine Beschreibung vun daten und Kohorten ist zentral für das Verstäändniss ener Arbeit (Veröffentlichung)
Ziel ist es it wenigen Kenngrößen ddas wesentliche zu charactersisieren.
dazu gibt es 'punktschötzer' für erwartungswerte und Konfidenzintervalle(KI engl. CI) als maß für die genauigkeit der Schätzung.
was ist ein konfidenzintervall

$$(1-\alpha) \quad KI[a,b]: \quad P(a \leq \Theta \leq b = 1-\alpha \tag{35}$$

man will schätzen wie wahrscheinlichkeit eines erwartungswertes ist.

Beispiel 1. wir messen die (menge) von haemoglobin bei 1 Mio menschen. dann gehe ich zum gefrierschrank und wähle 100 proben und ich messen einen wert der etwas anders ist als der bei den 1Mio, gebe eine konfidenzintervall von der zweiten messung an und mache das mit noch mehr proben , dann liegt manchmal das konfidenzmaß nicht über dem erwaartungswert.

2.1 Ein Merkmal

Nominale und Ordinale Größen

zb bei einer genetischen arbeit: menschen verschiedener herkunft lassen sich nicht ordnen: kategoriale größe $\rightarrow$ nominale größe

absolute und relative häufigkeiten. (zb Häufigkeitstabellen)
meist ist es gut etwas graphisch darzustellen: Balkendiagramme, oft mit konfidenzintervall oder standartfehler (KI und SE(standarterror)) kreisdiagramm(in den fachzeitschriften verpönt, weil man feststellt das man falsch einschätzen kann ob etwas 20 oder 30 prozent ist, klarer im balkendiagramm.

2.1.1 Metrische Daten

Lagemaß:

als Mittelwert (arithmetsich ($\frac{1}{n}\sum_{i=1}^{n} x_i$) oder geometrisch(dh log-Skala)($\left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}}$)
- übliche und 'robuste' Methoden (zb wenn ein wert stark abweicht sonst aber alles ähnlich ist, ist der mittelwert mit der log scala
- Median und andere Quantile(verschiedene Schätzverfahren)

2.1.2 streumaß

standartabweichung ("sample Method") $sd^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
mit $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- 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 grüßen: Korrelationskoeffizient (min KI) , Streudiagramm

2.3 Simplsons Paradoxon

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

Beispiel 1.