Zusammenfassung Einf. in die Statistik (Eberle)

Ayushi Tsydendorzhiev

October 11, 2024

Contents

1 Preliminaries

1.1 Witzige Definitionen

- Zufallsvariable X ist eine messbare Funktion (im Allgemeinen) aus einem Wktsraum in einen messbaren Raum.
- Wahrscheinlichkeitsverteilung P ist ein Maß $\mathcal{A} \to \mathbb{R}_+$

Wir können die Wahrscheinlichkeitsverteilung von X in natürlicher Weise als Bildmaß $\mu_X : A \mapsto P(X^{-1}(A))$ definieren.

- Die Wktverteilung μ_X ist eindeutig durch P festgelegt. Wir können das auf dem Erzeugendensystem von $\mathcal{B}(\mathbb{R})$ überprüfen, da $\mu_X((-\infty, c]) = P(X^{-1}(-\infty, c]) = P(X \leq c)$

Spickzettel:

- Wahrscheinlichkeitsverteilung (Maß) P auf $(\Omega, \mathcal{A})$
- Verteilung von X unter P definiert Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B})$

  $$\mu_X(B) := P(X^{-1}(B))$$

  Das impliziert die Formel $\int x d\mu = \int x f(x) dx$, wobei $f(x)$ die Wktsverteilung der ZV ist.

- Verteilungsfunktion der ZV $X : \Omega \to \mathbb{R}$

$$F_X : \mathbb{R} \to [0,1], c \mapsto P(X \leq c).$$

- u-Quantil ist ein Wert $q \in \mathbb{R} : P(X < q) = F(q-) \leq u$ und $F(q) \geq u$
- verallgemeinerte linksstetige Inverse $\underline{G}(u)$ – das kleinste u-Quantil

$$\inf\{x \in \mathbb{R} : F(x) \geq u\}$$

- verallgemeinerte rechtsstetige Inverse $\overline{G}(u)$ – das größte u-Quantil

$$\inf\{x \in \mathbb{R} : F(x) > u\}$$

- obere / untere Konfidenzschranke $b_\alpha(T(X))/a_\alpha(T(X))$

$$b_\alpha(T(X)) := \max\{N : F_N(T(X)) > \alpha\} \tag{1}$$
$$a_\alpha(T(X)) := \min\{N : F_N(T(X)-) < 1-\alpha\} \tag{2}$$

$$P_N[N \leq b_\alpha(T(X))] \geq 1-\alpha \tag{3}$$
$$P_N[N \geq \alpha_\alpha(T(X))] \geq 1-\alpha \tag{4}$$
$$P_N[N \in [a_{\frac{\alpha}{2}}(T(X)), b_{\frac{\alpha}{2}}(T(X))]] \leq \alpha \tag{5}$$

1.2 *Whimsical Verteilungen*

- Bernoulli-Verteilung BERNOULLI(P)
  - $P(1) = p$, $P(0) = 1-p$
- Geometrische Verteilung GEOM(P,K)
  - Anzahl Misserfolge bis zum ersten Erfolg, $P(X = k) = p(1-p)^k$
- Binomialverteilung BIN(N,P) = Summe von n unabhängigen Zufallsvariablen mit BERNOULLI(P)
  - $P(k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Beta-Verteilung BETA(A,B) = für die Verteilung der Wkten aus der Binomialverteilung, `https://stats.stackexchange.com/questions/47771/what-is-the-intuition-behind-beta-distribution`
- Multinomialverteilung = Bernoulli für mehr als 2 mögliche Ergebnisse.
- Poissonverteilung POISSON(λ) = Grenzwert von BIN(n,p) mit großem n und kleinem $p := \frac{\lambda}{n}$. Hinreichend genaue Approximation von BIN(n,p). Parameter λ heißt Intensität.

  - $P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$

- Hypergeometrische Verteilung HYPERGEOM(M,R,N) = Ziehen n Kugeln aus einer Urne mit insgesamt m Kugeln, r roten und $m - r$ schwarzen; ohne Zurücklegen;
  - k Anzahl gezogener roten und $n - k$ Anzahl gezogener schwarzen Kugeln ohne Zurücklegen.
  - wenn $\frac{n}{m} < 0,05$, approximiert die $\text{BIN}(n, p)$
  - $P(k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}$

1.3 *Eigenschaften für Momente*

- Erwartungswert (der erste zentrale Moment) = $\int_\Omega x f(x) dx$
  - linear und additiv
- Varianz (der zweite zentrale Moment) = $\mathbb{E}\left[(X - m)^2\right] = \int_\Omega (x - m)^2 f(x) dx$
  - nicht-negativ, $\text{Var}[aX] = a^2 \text{Var}[X]$, translationsinvariant
- $\mathbb{E}\left[X^2\right] = \text{Var}[X] + \mathbb{E}\left[X^2\right]^2$

2 *Hypothesentests, Populationsgröße, Konfidenzintervalle*

1. Vorlesung, 09.04.24

2.1 *Hypothesentests*

- p-Wert
  - links, $p_l = F(x)$
  - rechts, $p_r = 1 - F(x-)$
  - beidseitig, $p = 2\min(p_l, p_r)$

p-Wert ist das Maß des halb geschlossenen Intervals $F(x) = \mu((-\infty, a]$ oder $1 - F(x-) = \mu([b, \infty)$.

- Fishers exakter Test

2. Vorlesung, 12.04.24

2.2 *Schätzen der Populationsgröße*

- **German Tank Problem**: es gibt N Panzer mit Nummern 1...N. Wie groß ist N?
  - **Schätzer Y**: empirischer Mittelwert $N = 2 \cdot \hat{E}_n(X) - 1$
  - **Schätzer M**: $N = \max\{X_i : i < n + 1\}$
- Welcher Schätzer ist besser? Systematischer Vergleich.
  - Erwartungswert, Varianz eines Schätzers. Erwartungstreue.

  - Schätzer Y ist erwartungstreu, Schätzer M nicht erwartungstreu
  - $\text{MSE}(Y) = E[(Y-N)^2] = Var(Y)$
- Konfidenzintervall für $\vartheta = N$ unbekannter Parameter, $F_N(x)$ monoton fallend in N.
  - Konfidenzbereich zum Konfidenzniveau 95% bzw. Signifikanzniveau 5%
  - $P_N(N \leq b(M))$ ist nicht die Wkt, dass N unter x liegt, sondern dass N unter x liegt conditioned on N.
  - Wenn wir das Experiment 100mal wiederholen dann in $1-\alpha$ Experimenten wäre die Bedingung erfüllt. Wir können nicht sagen, dass wir unter einem festen Wert mit einer bestimmten Wkt liegen.

3. Vorlesung, 16.04.24

- **Schätzung einer Tierpopulation**. Capture-Recapture-Verfahren:

1. Entnahme Zufallsstichprobe der Größe l und Markierung
2. Nehme Zufallsstichprobe der Größe $n \leq N$
   - H = Anzahl der markierten Tier in der 2. Stichprobe
   - Modell ist $H \sim \text{Hyp}(N, l, n)$
   - Schätzer ist $N \approx \frac{nl}{H}$
   - untere Konfidenzschranke

     $$P_N(F(H) > \alpha) \geq 1-\alpha$$

     kann so umgeschrieben werden, dass F nur von N abhängt, also von dem Wert $a_\alpha(H)$, ab dem die Ungleichung gilt.

     $$= P_N(N \geq a_\alpha(H))$$

3 Statistische Modelle und Verfahren

- Statistische Modelle und Verfahren: schließende Statistik.
  - Ein *statistisches Modell* besteht aus
    - einer Menge $\Omega \neq \emptyset$ mit σ-Algebra $\mathcal{A}$,
    - einer Parametermenge $\Theta \neq \emptyset$
    - einer Familie $\{P_\vartheta : \vartheta \in \Theta\}$ von Wkn auf $(\Omega, \mathcal{A})$
    - einer messbaren Abbildung $X : \Omega \to S$, $(S, \mathcal{B})$ messbarer Raum. Stichprobe ist dann Realisierung dieser Abbildung $X(\omega)$
  - Eine *Statistik* ist eine Abbildung $T(X), T : S \to \mathbb{R}$.
  - Beispiel (Capture-Recapture):
    - $\Theta = \mathbb{N}_{\geq \min(l,n)}$,
    - $\Omega = (\omega_1^{(1)}, \ldots, \omega_1^{(l)}, \omega_2^{(1)}, \ldots, \omega_2^{(n)})$, $P_N = \textsc{Unif}(\Omega_N)$, $\Omega_N = \{w \in \Omega : w_i \leq N \forall i, k\}$,
    - Statistik $H(\omega) = |...|$

3.1 Grundlegende Modelle

- Grundlegende Modelle:
  - Bernoulli-Modell, Schätzer von Wkeiten: $\Theta = [0, 1]$, $X = (X_1, \ldots, X_n)$, $X_i \sim \textsc{Bernoulli}(\vartheta)$, Statistik z. B. rel. Häufigkeit $\overline{X}_n = \frac{1}{n} \sum_i^n X_i$.
  - Gauß-Modell, Parameterschätzung: $\Theta = \mathbb{R} \times \{0, \infty\}$ (Mittelwert und Varianz), $X = (X_1, \ldots, X_n)$, $X_i \sim \mathcal{N}(m, v)$, Staistiken z. B. Stichprobenmittelwert, Stichprobenvarianz, etc
    - Oft einfach angenommen, da CLT unter relativ einfachen Annahmen gilt (Mittelwert von vielen kleinen ZV ist normalverteilt)
  - Nichtparametrische Schätzung der Verteilung bzw. Verteilungsfunktion: $\Theta = \mathcal{P}(\mathbb{R})$ = alle WVn auf $(\mathbb{R}, \mathcal{B}(\mathbb{R})$, unendlichdimensional. $X = (X_1, \ldots, X_n), X_i \sim \mu$ unabhängig unter P_μ. Statistiken z. B. empirische Verteilung, Verteilungsfunktion klar (Summe aller von links auftretenden Ereignissen...)
  - Nichtparametrische Dichteschätzung: $\Theta = \{f : \mathbb{R} \to [0, \infty) : \int f(x)dx = 1, \exists$ schwache Ableitung $f'' \in L^2(\mathbb{R})\}$. Das heißt einfach, dass f zweite Stammfunktion einer quadratintegrierbaren Funktion g ist. Statistik: Faltung mit Gauß-Glockenkurven um jeden Punkt x_i mit einer gewissen Varianz v_i. Schätzwert

$$\hat{f}_n(x) = \frac{1}{n} \sum_i^n \varphi_\vartheta(x - X_i), \varphi_\vartheta(x) = \frac{1}{\sqrt{2\pi h}} e^{-\frac{y}{}}.$$

- Regression $(X_1, Y_1), \ldots, (X_n, Y_n) : \Omega \to \mathbb{R}^d \times \mathbb{R}$. Die Annahme ist $Y_i = f(X_i) + \sqrt{v}\varepsilon$, nichtparametrisch $\Theta = \{(v, f) : v \in (0, \infty), f : \mathbb{R}^d \to \mathbb{R}\}$, lineares Modell $f(x) = w^T x$, $\Theta = \{(v, w) : v \in)(0, \infty), w \in \mathbb{R}^d\}$
- neuronale Netzwerke mit einer bestimmten Architektur...

3.2 *Grundlegende Verfahren*

- Grundlegende Verfahren:
  - **(Punkt-)Schätzer I:** Gegeben ist $g : \Theta \to \mathbb{R}$, gesucht wird $g(\vartheta)$, z. B. der Erwartungswert m
    - Ein *Schätzer* für $g(\vartheta)$ ist eine Abbildung $\hat{g} = T(X)$, $T : S \to \mathbb{R}$ (also eine Statistik).
  - Wir können verschiedene Schätzer hinsichtlich ihrer Qualität vergleichen:
    - Der *systematische Fehler* von $\hat{g}$ ist $\text{Bias}_\vartheta(\hat{g}) = E(\hat{g}) - g(\vartheta)$. Schätzer ist *erwartungstreu*, falls $\text{Bias}_\vartheta = 0 \forall \vartheta \in \Omega$.
    - Der mittlere quadratische Fehler von $\hat{g}$ ist

$$\text{MSE}_\vartheta(\hat{g}) = E_\vartheta((\hat{g} - g(\vartheta))^2 = \text{Var}[\hat{g}] + Bias^2(\hat{g}),$$

$$\text{RMSE} = \sqrt{\text{MSE}}.$$

Schätzer schätzen unbekannten Parameter ϑ mit Input Stichprobe X, sind also insb. eine Statistik und insb. eine Zufallsvariable.

4. Vorlesung, 19.04.24

- **Schätzer II:** Anwedung und Beispiele
  - Modell: beobachten $X = (X_1, \ldots, X_n)$, $X_i \sim \mu$ i.i.d. unter P_μ
  - Schätzer vom **Erwartungswert** m: $m(\mu) = \int f(x)\mu(dx)$, Varianz $v(\mu) = \int (x - m(\mu))^2 \mu(dx)$.
    - Erwartungswert ist eine Funktion $G : \mu \to m(\mu)$ im obigen Sinne.
    - Beispiel: empirischer Mittelwert, siehe oben. MSE, siehe oben.
  - Schätzer von der **Varianz** v: Stichprobenvarianz $= \frac{1}{n}\sum(X_i - \overline{X_n})^2$
    - nicht erwartungstreu, da $n - 1$ Freiheitsgrade
    - $V^* = \frac{1}{n-1}\sum(X_i - \overline{X_n})^2$ erwartungstreu
  - Schätzer von der **Dichte** μ: empirische Verteilung $\hat{\mu}_n$
    - relative Häufigkeit von Werten in B, $\hat{\mu}_n(B) = \frac{1}{n}\sum_i 1_B(x_i)$ (erwartungstreu)
- Vorsicht: Schätzer ohne Konfidenzbereich sind nutzlos.

3.3 Konfidenzbereiche

- Konfidenzbereiche
  - Sei $\alpha \in (0,1)$. Ein Konfidenzbereich für $g(\vartheta)$ mit Konfidenzniveau $1-\alpha$ (oder mit Sicherheit $1-\alpha$) ist eine Abbildung $C(X) : S \to \mathcal{P}(G), x \mapsto C(x) \subseteq G$, wobei S = Stichprobe und $\mathcal{P}(G)$ = die Potenzmenge von $G(\mathbb{R})$, mit
$$P_\vartheta(g(\vartheta) \in C(X)) \geq 1-\alpha \qquad \forall \vartheta \in \Theta$$
  - Konkret für $G = \mathbb{R}$ gilt $C(X) = [a(X), b(X)]$, wobei $a(X), b(X)$ Statistiken sind.

Kochrezepte für Konfidenzbereiche.

- über die **Verteilungsfunktion**:
  - Es ist eine Statistik $T(x) : \Omega \to \mathbb{R}$ gegeben. Sie ist eine Zufallsvariable. Betrachte ihre Verteilungsfunktion $F_\vartheta(c) = P_\vartheta(T(x) \leq c)$. Die Idee war, wenn $P(T(X) \leq \alpha)$, dann verwerfen wir die Nullhypothese.
  - $C(x) := \{g(\vartheta) : \vartheta \in \Theta, F_\vartheta(F(x)) > \alpha\}$
    $\implies P_\vartheta(g(\vartheta) \in C(X)) \geq 1-\alpha \forall \vartheta \in \Theta$
  - Wir schließen die Parameterwerte aus, für die $T(x)$ verdächtig klein ist.
  - Analog für beidseitige $(1-\frac{\alpha}{2})$ Konfidenzbereiche
- Standard-Schul-Beispiel: $\Theta = [0,1] \ni p$
  - $X_1, \ldots, X_n$ i.i.d. $\sim$ Bernoulli(p) unter P_p
  - $H_n = \sum_i X_i$ unter P_p Häufigkeit von 1
  - $F_{n,p}(c) = P_p(H_n \leq c) = \sum_{k=0}^{c} \binom{n}{k} p^k (1-p)^{n-k}$ für $c = 0 \ldots n$
  - Ableitung von $F_{n,p}$ ausrechnen, ist streng monoton fallend, $F(p)$ als Funktion von p plotten $F(1) = 0, F(0) = 1$, und dann suchen $F(c) = \alpha$. Dieses c legt Konfidenzintervall $C(X) = [0, c]$ zum Niveau $1-\alpha$ fest.
- über die **Likelihood**:
  - Idee: wir schließen Parameterwerte aus, unter denen X eher unwahrscheinlich (unlikely) ist.
  - Annahme: entweder Massen- oder Dichtefunktion. S abzählbar, $f_\vartheta(x) = P_\vartheta(X = x)$ Massenfunktion oder $S = \mathbb{R}^d$, $P_\vartheta(X \in B) = \int_B f_\vartheta(x)dx$ Dichtefunktion. Dabei heißt $\vartheta \mapsto f_\vartheta(x)$ eine **Likelihood-Funktion**.

  - Bestime c_ϑ möglichst klein, so dass $P(g(\vartheta) \in C(X)) \geq P(f_\vartheta(X) \geq c_\vartheta) \geq 1-\alpha$ gerade noch geht.

- über eine **Pivot-Statistik**:
  - Ein Pivot für $g(\vartheta)$ ist eine Statistik $T(X, g(\vartheta))$, deren Verteilung nicht von ϑ abhängig ist. Sehr robust.
  - **Beispiel 1**: Gauß-Modell mit fester Varianz, schätze Mittelwert m.
  - Ziehe i.i.d. $X_1, \ldots, X_n$, dann $\overline{X} \sim N(m, \frac{v}{n})$. Dann ist $Z = \frac{\overline{X}-m}{\sqrt{v/n}}$ wie $N(0,1)$ verteilt und somit *die Verteilung* von Z von m nicht abhängt. Damit können wir ein Konfidenzintervall konstruieren.
  - **Beispiel 2**: Gauß-Modell mit m, v unbekannt, sprich normalverteilt mit Mittelwert m und Varianz v.
  - $T_n = \frac{\overline{X}_n - m}{\sqrt{V_n^*/n}}$ ist ein Pivot für m, Studentsche t-Statistik.
  - Wir normieren X_i durch $Y_i := \frac{X_i - m}{\sqrt{r}}$, dann ist $Y_i \sim N(0,1)$.
  - Rechne $T_n = \sqrt{n}\frac{\frac{\overline{X}_n - m}{\sqrt{r}}}{\sqrt{\frac{V_n^*}{r}}} = \sqrt{n}\frac{\overline{Y}_n}{\sqrt{\frac{1}{n-1}\sum_i^n (Y_i - \overline{Y}_n)^2}}$ und die Verteilung hängt nicht von m und v ab.
  - Damit $P(|\overline{X}_n - m| \geq c\sqrt{\frac{V_n^*}{n}}) = P(|T_n| \geq c) = 2P(T_n \geq c) = 2(1 - F_{T_n}(c)) \leq \alpha \forall m, v$ falls $c \geq F_{T_n}^{-1}(1 - \frac{\alpha}{2})$
  - Habe ein $(1-\alpha)$ Konfidenzintervall $\overline{X}_n \pm q_{1-\frac{\alpha}{2}}\sqrt{\frac{V_n^*}{n}}$. Die Verteilung von T_n heißt Studentsche t-Verteilung mit $n-1$ Freiheitsgraden, kann man explizit ausrechnen.

- approximative Konfidenzintervalle ($\to$ Normalapproximation)
  - Beispiel: Bernoulli-/Binomialmodell, $\Theta = [0,1] \ni p$
    - $H_n = X_1 + \ldots + X_n \sim Bin(n,p)$, $X_i \sim Bernoulli(p)$
    - $E(H_n) = np, Var(H_n) = np(1-p), \hat{p}_n = \frac{H_n}{n}$
    - Zentraler Grenzwertsatz / de Moivre-Laplace: $\frac{H_n - np}{\sqrt{np(1-p)}} \in (-c, c) \longrightarrow N(0,1)(-c,c) = 2(\Phi(c) - \frac{1}{?}) \implies$ für große n gilt näherungsweise $P(|p - \hat{p}_n| < c\sqrt{\frac{p(1-p)}{n}}) \approx 2(\Phi(c) - \frac{1}{?}) \geq 1 - \alpha$. Das Problem ist die Abhängigkeit von p in der Schranke.
    - Idee 1: ersetze p durch Schätzer $\hat{p}$, praktisch aber ungenau.
    - Idee 2: $p(1-p)$ durch $\frac{1}{4}$ abschätzen, sicher
    - Idee 3: Ungleichung auflösen

This is what's known as Z-Test.

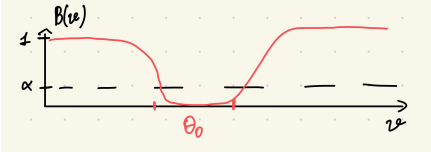
5. Vorlesung, 23.04.24

This is what's known as Student's t-Test.

3.4 *Hypothesentests*

- Gegeben ist $(\Omega, \mathcal{A}, (P_\vartheta)_{\vartheta \in \Theta})$ mit $X : \Omega \to S$
  - Nullhypothese $H_0 : \vartheta \in \Theta_0 \subseteq \Theta$

  - Alternative $H_1 : \vartheta \in \Theta_1$ und $\Theta_0 \cap \Theta_1 = \emptyset$
  - Was ist ein Hypothesentest formal?
    * Ein *Hypothesentest für* H_0 ist gegeben durch eine messbare Funktion $\varphi : S \to \{0,1\}$, bzw. $[0,1]$ (randomisierter Test, verwerfe mit Wahrscheinlichkeit aus diesem Bereich).
    * $R = \{x \in S : \varphi(x) > 0\}$ heißt *der Verwerfungsbereich*.
  - $\beta(\vartheta) := E_\vartheta(\varphi(x)) = P_\vartheta(X \in R)$ falls nicht randomisiert, heißt *Macht* bzw. *Gütefunktion* des Tests.
  - der Test hat *Signifikanzniveau* $\alpha \in (0,1)$, falls $\forall \vartheta \in \Theta_0 : \beta(\vartheta) \leq \alpha$



  - Fehler 1. Art – verwerfen die H_0, obwohl sie gilt, soll um egal welche Kosten $\leq \alpha$ bleiben.
  - Fehler 2. Art – verwerfen die H_0 nicht, obwohl falsch, Wahrscheinlichkeit $1 - \beta(\vartheta), \vartheta \in \Theta_1$. Die Wahrscheinlichkeit soll möglichst klein sein.

- Äquivalenz Hypothesentest $\iff$ Konfidenzbereich
  - Sei $C(X)$ ein $(1-\alpha)$-Konfidenzbereich für ϑ. Definiere $\varphi(X) := 0$, falls $\vartheta_0 \in C(X)$ und $\varphi(X) = 1$, falls $\vartheta_0 \notin C(X)$.

- Zusammenfassung:
  - Verteilungsfunktion – p-Werte
  - Likelihood –
  - Pivot – Hypothesentest
  - Normalapproximation

- Beispiel: Gauß-Modell, $X_1, \ldots, X_n \sim N(m, v)$
  - $H_0 : m = m_0$
  - $(1-\alpha)$-Konfidenzbereich : $C(X) = \{m \in \mathbb{R} : \left|\frac{\overline{X}_n - m}{\sqrt{V_n^*/n}}\right| < q_{1-\alpha}\}$ definiert einen Hypothesentest zum Niveau α, wenn $\varphi(X) = 1$, falls $|T_n(m_0)| \geq q_{1-\frac{\alpha}{2}}$. Das ist genau der Studentsche t-Test.
  - $H_0 : m = m_0$, $H_1 : m > m_0$
  - $C'(X) = \{m \in \mathbb{R} : T_n(m) < q_{1-\alpha}\}$

4 Likelihood

Annahmen: entweder S abzählbar, $f_\vartheta(x)$ Massenfunktion oder $S = \mathbb{R}^d$ und $f_\vartheta(x)$ Dichtefunktion

- Likelihood-Funktion $L(\vartheta; x) = f_\vartheta(x)$ Dichtefunktion von X unter ϑ; X ist fest.
- log-Likelihood $l(\vartheta; x) = \log f_\vartheta(x)$ falls $f_\vartheta(x) > 0$

Beispiel: Produktmodell $X = (X_1, \dots, X_n)$, X i.i.d. mit Dichtefunktion g_ϑ. Dann ist die Likelihood $L(\vartheta; x) = \prod_i^n g_\vartheta(x_i)$, log-Likelihood $l(\vartheta; x) = \sum_i^n \log g_\vartheta(x_i)$.

4.1 Maximum-Likelihood-Prinzip

Konstruiere einen Schätzer $\hat{\vartheta} = T(X)$ für ϑ: wähle den Parameterwert $T(x)$, für den der Beobachtungswert x am plausibelsten ist, d.h.

$$L(T(x); x) = \max_{\vartheta \in \Theta} L(\vartheta_i; x) \qquad x \in S$$

- Der Schätzer $\hat{\vartheta} = T(X)$ heißt *Maximum-Likelihood-Schätzer* (MLS) für ϑ, falls $T(X) \in \operatorname{argmax}_{\vartheta \in \Theta} L(\vartheta; x) \forall x \in S$
- **Beispiel 1**: German tank problem
  - $L(N; x) = f_N(x) = \frac{1}{|\Omega_N|}$ für $x \in \Omega_N$ und 0 sonst. Das ist maximal für $N = \max\{x_1, \dots, x_n\}$. Unser MLE ist $\hat{N} = \max\{x_1, \dots, x_n\}$.

6. Vorlesung, 26.04.24

- **Beispiel 2**: *Bernoulli*(p), $L(p; X) = P_p(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$,
  - $l(p, x) = \log(p) \sum_i x_i + \log(1-p) \sum_i (1 - x_i)$, optimiere nach p (Ableitung $= 0$).
- **Beispiel 3**: Schätzen einer unbekannten WV
  - Θ = alle WVn auf $\hat{a} \in \mathbb{R}^n = \{\overline{p} : \sum p_i = 1\} \in \mathbb{R}^n$ ein Standard-Simplex, $X_1, \dots, X_n$ unabängig mit Verteilung p unter P_p.
  - $L(p; x) = P(X_1 = x_1, \dots, X_n = x_n) = \prod_i p_{x_i} = \prod_{l=1}^k p_l^{h_l}$, wobei h_l = Häufigkeit von p_l.
  - **Behauptung:** MLE $\hat{p}$ für p ist $\hat{p}_l = \frac{H_l}{n}$ = relative Häufigkeit von a_l mit Stichprobe $X_1, \dots, X_n$. Empirische Verteilung ist MLE für zugrundeliegende WV! Beweis mit Lagrange.
- **Beispiel 4**: Gauß-Modell
  - $\Theta = \mathbb{R} \times (0, \infty)$,

- $L(m, v; x) = -\frac{n}{2}\log(2\pi v) - \frac{1}{2v}\sum(x_i - \overline{x})^2 - \frac{n}{2v}(\overline{x}_n - m)^2$
- Schätzen von m bei bekannter Varianz v, MLE $\hat{m} = \overline{X}_n$
- Schätzen von v bei bekanntem Mittelwert m, MLE empirische Varianz
- Schätzen von $\vartheta = (m, v)$
  - $m = \overline{X}_n$, $v = \frac{1}{n}\sum_i (x_i - \overline{x}_n)^2$
  - MLE $\hat{\vartheta} = (\overline{X}_n, V_n)$ (nicht erwartungstreu in v)

4.2 *Suffiziente Statistiken*

Datenreduktion, beinhaltet alle Informationen, die für Rückschlüsse auf ϑ relevant sind. Idee: Wahlprognose $X_1, \ldots, X_n \sim Bernoulli(p)$. Für Rückschlüsse auf p reicht es, $\sum X_i$ zu wissen.

- Eine Statistik $T(X)$ heißt *suffizient* für ϑ, falls

  $$L(\vartheta, x) = f_\vartheta(x) = g_\vartheta(T(x))h(x) \quad \forall \vartheta \in \Theta, x \in S$$

  mit messbaren Funktionen $g_\vartheta : \mathbb{R} \to [0, \infty]$, $h : S \to [0, \infty]$. Likelihood hängt nur durch $T(X)$ von ϑ ab.

- Gilt $h(x) > 0$,dann folgt daraus für MLE:

  $$\text{argmax}_{\vartheta\in\Theta} L(\vartheta_i; x) = \text{argmax}_{\vartheta\in\Theta} g_\vartheta(T(X)).$$

  und wenn das Maximum eindeutig ist (z. B. falls l strikt konkav), dann ist MLE eine Funktion von $T(x)$.

- **Lemma**: $T(X)$ ist suffizient, wenn $P_\vartheta(X = x|T(x) = t)$ nicht von ϑ abhängt.

- Als Zweistufenmodell
  - Ziehe $T(X)$ aus Verteilung von $T(X)$ (hängt von ϑ ab).
  - Ziehe X aus Verteilung von X gegeben $T(X)$ (hängt nicht von ϑ ab).

Die Idee: je größer die betrachtete Klasse von Wktsverteilungen, desto mehr Informationen muss die suffziente Statistik enthalten.

- **Satz**: Rao-Blackwell. Ein beliebiger Schätzer $\hat{g}$ kann verbessert werden, wenn man stattdessen die bedingte Erwartung $\tilde{g} := \mathbb{E}[\hat{g}|T(X)]$ betrachtet.

4.3 *Exponentielle Familien*

7. Vorlesung, 30.04.24

- Definition: $f_\vartheta(x) = e^{c(\vartheta)\cdot T(x)+d(\vartheta)+U(x)}1_A(x)$
  - in natürlicher Form, $f_\vartheta(x) = e^{\vartheta\cdot T(x)+d(\vartheta)+U(x)}1_A(x)$,
  - alternative Schreibweise, $f_\vartheta(x) = \frac{1}{\mathcal{Z}(\vartheta)}e^{c(\vartheta)\cdot T(x)}h(x)$, wobei
    - Normierungskonstante $\mathcal{Z}(\vartheta) := e^{-d(\vartheta)}$,
    - Referenzdichte $h(x) = e^{U(x)}1_A(x)$
- Beispiele: Exponentialverteilungen, Bernoulli-, Binomial-, Poisson-Verteilungen, Normalverteilungen.
- In einer exponentiellen Familie ist $T(X)$ eine suffiziente Statistik.
- Wichtige Eigenschaft: Stabilität under Produktbildung.

8. Vorlesung, 03.05.24

- Berechnung von Momenten:
  - Momenterzeugende Funktion $M(s) := \mathbb{E}\left[e^{sT(x)}\right] = \int e^{sT(x)}f_\vartheta(x)dx = e^{d(\vartheta)-d(s+\vartheta)} < \infty$
  - Mit Algebra $M(s) = \frac{\mathcal{Z}(s+\vartheta)}{\mathcal{Z}(\vartheta)} = e^{d(\vartheta)-d(s+\vartheta)} < \infty$
    - eine Bedingung an $M(s)$: muss in der Umgebung von 0 endlich sein
    - $\mathbb{E}[T(X)] = \mathbb{E}\left[\nabla_s e^{sT(X)}|_{s=a}\right] = \nabla M(0) = -\nabla d(\vartheta)$
    - $\mathbb{E}\left[T_i(X)T_j(X)\right] = \dots$
    - $\mathrm{Cov}[T_i(X)T_j(X)] = \dots$
- **Satz**: $\hat{\vartheta}$ ist MLS $\iff \mathbb{E}[T(X)] = T(X)$
- Berechnung des MLS:
  - $l(\vartheta, X) = \log f_\vartheta(X) = d(\vartheta) + \vartheta T(X) + \log h(x)$
  - $\nabla l(\vartheta) = \nabla d(\vartheta) + T(X) = 0 \iff \nabla(\vartheta) = -T(X) \iff \mathbb{E}[T(X)] = T(X)$. Der MLE ist genau der Parameterwert ϑ, unter dem $T(X)$ dem Erwartungswert entspricht.

4.4 *Likelihood-Quotienten-Test*

Gegeben ist $H_0 : \vartheta \in \Theta_0$, $H_1 : \vartheta \in \Theta_1$.

- Ein Hypothesentest ist eine Funktion $\varphi : S \to [0,1]$ mit Verwerfenswahrscheinlichkeit $\varphi(x)$.
- Der Verwerfungsbereich ist $C(X) = \{x \in S : \varphi(x) = 1\}$.
- Die Gütefunktion $G(\vartheta) = \mathbb{E}[\varphi(X)]$, Erwartungswert der Wkt, die Nullhypothese zu verwerfen.

- Der Test hat Signifikanzniveau α, falls $G(\vartheta) \leq \alpha \qquad \forall \alpha \in \Theta_0$.
- Die Macht des Hypothesentests ist $G(\vartheta)$ eingeschränkt auf Θ_1. Kleinere Macht $\implies$ größere Wkt, die Nullhypothese fälschlicherweise nicht zu verwerfen, Fehler 2. Art.

Einfacher Likelihood-Quotienten-Test, feste ϑ_0 und ϑ_1:

- Entscheidungsregel: relative Dichte

$$R(x) = \frac{L(\vartheta_1; x)}{L(\vartheta_0; x)} > c,$$

- Entscheidungsfunktion $\varphi = 1_{R(x)>c} + p \cdot 1_{R(x)=c}$
- **Neyman-Pearson-Lemma**: der LQT ist der mächtigste Test zum Niveau α.

9. Vorlesung, 06.05.24

Θ_0 und Θ_1 sind jetzt ein Bereich, keine feste Parameter

Monotone Likelihood-Quotienten, $\vartheta_0 \in \Theta_0$ und $\vartheta_1 \in \Theta_1$:

- nach Definition gilt:

$$R(x) = \frac{L(\vartheta_1; x)}{L(\vartheta_0; x)} = \frac{\mathcal{Z}(\vartheta_0)}{\mathcal{Z}(\vartheta_1)} e^{(c(\vartheta_1)-c(\vartheta_0))T(x)},$$

- **Satz**: für das einseitige Testproblem $H_0 : \vartheta \leq \vartheta_0$, $H_1 : \vartheta > \vartheta_0$ und falls der Likelihood-Quotient streng monoton wachsend in $T(X)$ ist, dann ist LQT $\varphi(X)$ der gleichmäßig *mächtigste* Test,

$$\mathbb{E}[\psi(X)] \leq \mathbb{E}[\varphi(X)]$$

Allgemeine Likelihood-Quotienten-Tests, (ad hoc)

$$R(x) = \frac{\sup\{L(\vartheta; x) : \vartheta \in \Theta_1\}}{\sup\{L(\vartheta; x) : \vartheta \in \Theta_0\}}$$

4.5 Studentsche Konfidenzintervalle und t-Test

- Einführung: Gauß-Modell mit unbekanntem Mittelwert und Varianz, Schätzung von m.
  - Empirischer Mittelwert $\overline{X_n}(\omega)$
  - Varianz bekannt: Pivot
  - Varianz unbekannt: Studentscher t-Test
- $\overline{X}_n \sim N(m, \frac{v}{n})$ und $\frac{n-1}{v} V_n \sim \chi^2(n-1)$
- Verwende Studentsche t-Statistik mit $n-1$ Freiheitsgraden,

$$T_{n-1}(X) = \frac{\sqrt{n} \cdot (\overline{X}_n - m)}{\sqrt{V_n}}.$$

10. Vorlesung, 10.05.24

- Lemma 2.16:

$$f_{\chi^2(n)} \propto x^{\frac{n}{2}-1} e^{-\frac{x}{2}} 1_{(0,\infty)}(x)$$

und

$$f_{t(n)}(x) \propto (1 + \frac{x^2}{n})^{-\frac{n+1}{2}} (\sim e^{\frac{x^2}{2}} \text{ für } n \to \infty).$$

Fazit: wir können die Varianz mit der Stichprobenvarianz ersetzen und dabei einen nicht allzu großen Fehler machen.

4.6 Anwendung auf Konfidenzintervalle & Tests

- ein Test ist **unverfälscht** zum Niveau α, wenn $P(T(X) > c) \leq \alpha$ wenn H_0 wahr und $P(T(X) > c) \geq \alpha$ wenn H_1 wahr. In anderen Worten, $\mathbb{E}[\varphi(\vartheta_0)] \leq \alpha \leq \mathbb{E}[\varphi(\vartheta_1)]$
- Studentscher t-Test ist ein LQT für Gauß-Modell und ist der beste unverfälschte Test i. S. v. oben.

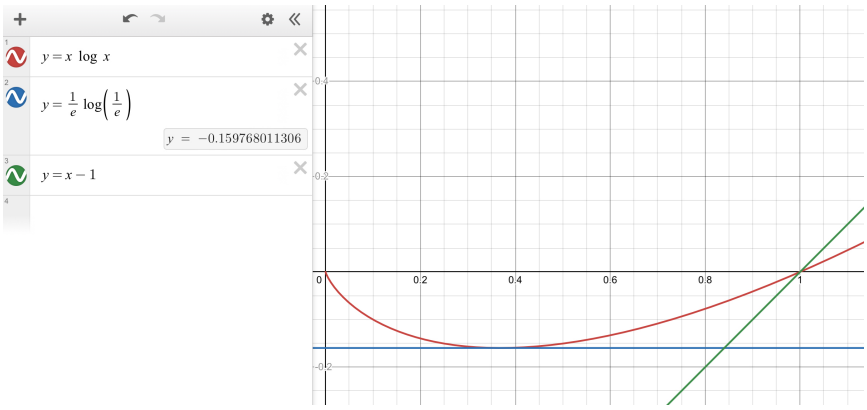
5 *Entropie und Information*

11. Vorlesung, 14.05.24

Es geht um die stetige streng konvexe Funktion

$$u(x) = \begin{cases} x \log x & x > 0 \\ 0 & x = 0 \end{cases}, \qquad u'(x) = 1 + \log x, \qquad u''(x) = \frac{1}{x}$$

Es gilt $u(x) \geq x - 1$ und $\min u(x) = \frac{1}{e} \log(\frac{1}{e}) = -\frac{1}{e}$.



5.1 *Entropie*

Die **Entropie** $H(\mu)$ von **diskreter** Wktsverteilung μ ist der mittlere Informationsgewinn beim Ziehen einer Stichprobe, also

$$H(\mu) = -\sum_{x \in S, f(x) \neq 0} f(x) \log f(x) = -\sum_{x \in S, f(x) \neq 0} u(f(x))$$

- Maß für Überraschung oder Informationszuwachs beim Eintritt von Ereignissen der Wkt $p \in [0,1]$ definieren wir als $-\log f(x)$, $f(x)$ ist die Massenfunktion für μ auf S.

$$H(x) = \underbrace{-\log(f(x))}_{\text{Inform-zuwachs}} \cdot \underbrace{f(x)}_{\text{Wkt. von } x}$$

- Log, weil bei n unabhängigen Stichproben wird aus der Information der Probe $-\log(\prod f_i(x_i))$ ergibt sich der Informationszuwachs als Summe einzelner Beobachtungen $-\sum \log f_i(x_i)$.
- Die Entropie ist immer positiv, da $f(x) \log f(x)$ immer negativ für $f(x) \in [0,1]$. Das gilt nur im abzählbaren Fall!
- Die Entropie ist eine strikt konkave Abbildung auf dem Raum der Wahrscheinlichkeitsverteilungen $WV(S)$.

Beispiele:

- Bernoulli(p), Einsetzen ergibt $H(Bernoulli(p)) = -p\log p - (1-p)\log(1-p)$, Minimum bei $p = 0$ oder $p = 1$, Maximum bei $p = 0.5$. Sprich, die Verteilung hat minimale Entropie bei $p = \{0,1\}$ und maximale Entropie bei $p = 0.5$.
- **Entropieminima**: Das Dirac-Maß hat minimale Entropie, da $H(\mu) = 0 \iff f(x) \in \{0,1\}$.
- **Entropiemaximum**: Die Gleichverteilung hat maximale Entropie wenn $|S|$ endlich ist. Es gilt $H(Unif(S)) = -\log(1/|S|) = \log(|S|) \geq H(\mu)$ für alle anderen μ. Das heißt, Gleichverteilung ist ein ungeordneter Zustand.
- Entropie kann unendlich sein, falls S wie oben unendlich ist.
- Entropie der Produktverteilung $H(\mu_1 \otimes \ldots \otimes \mu_n) = \sum_i H(\mu_i)$. Beweis durchs Nachrechnen.

Was bedeutet Entropie statistisch?

- Nehme n unabhängige Stichproben unter μ. Betrachte relative Häufigkeit $\hat{p}(x_i) = \frac{\#x_i}{n}$ und die empirische Wkt $p(x_1,\ldots,x_n) = \prod_i \hat{p}(x_i)$ als empirische Likelihood-/Dichtefunktion $L(\mu; x_1 \ldots x_n) = \prod_i \hat{p}(x_i)$. Gemittelt ergibt sich der mittlere Informationszuwachs als

  $$\overline{H(X_i)} = -\frac{1}{n}\log(\prod_i \hat{p}(X_i)) = -\frac{1}{n}\sum_i \log \hat{p}(X_i).$$

  Die Entropie $H(\hat{p})$ ist der mittlere Informationsgewinn bei n Stichproben. Wie verhält sich diese Größe asymptotisch, wenn $n \to \infty$?
- Satz von Shannon-McMillan I:

  $$-\frac{1}{n}L(\mu; X_1 \ldots X_n) \overbrace{\to}^{n\to\infty} H(\mu) \text{ almost surely}$$

  Beweis durch GGZ, $\frac{1}{n}l(\mu; X) = \frac{1}{n}\sum_i \log(f(X_i))$, Summanden immer negativ, GGZ anwendbar, also konvergiert die Summe gegen den Erwartungswert $\int \log f d\mu = \sum f(x)\log(f(x)) = -H(\mu)$. **Entropie ist** $\mathbb{E}[H(X)]$!!!

In der Statistik geht es darum, zwei Wktsverteilungen voneinander zu unterscheiden. Das lässt sich durch relative Entropie quantifizieren.

5.2 *Relative Entropie*

Gegeben sind zwei Wktsverteiliungen μ, ν. Die **relative Entropie** $H(\mu|\nu)$ für absolutstetige Wahrscheinlichkeitsmaße ist nicht symmetrisch und definiert **bezüglich** ν.

$$H(\mu|\nu) = \int_{\mathbb{R}^d} \frac{f(x)}{g(x)} \log\left(\frac{f(x)}{g(x)}\right) dx.$$

- μ muss absolutstetig bzgl. ν sein, Notation $\mu \ll \nu$
- dann existiert relative Dichte $\omega = \frac{d\mu}{d\nu}$
  - $\mu[B] = \int_B \omega d\nu \quad \forall B \in \mathcal{B}$ (Lebesgue-Stieltjes?). Kurzschreibweise "$\mu = \omega d\nu$". Das ist eindeutig bis auf ν-Nullmengen (man kann eventuell das Maß auf einer Nullmenge ändern, aber ist grds. egal). Die Dichte ist größer gleich Null und strikt größer als Null μ-f.ü, nicht ν-f.ü.
  - In dieser Vorlesung sind S abzählbar oder $\mathbb{R}^d$, μ, ν mit Massenfunktion/Dichte f, g (sie existieren).

μ hat Verteilung f, ν hat Verteilung g.

- wenn Nullmengen von ν in Nullmengen von μ enthalten sind, $\{g = 0\} \subseteq \{f = 0\}$, dann $\mu \ll \nu$ und es gilt für Dichte

$$\omega(x) = \frac{d\mu}{d\nu}(x) = \begin{cases} \frac{f(x)}{g(x)} = \frac{L(x,\mu)}{L(x,\nu)} & g(x) > 0 \\ \text{beliebig} & g(x) = 0 \text{ (Nullmenge)} \end{cases}$$

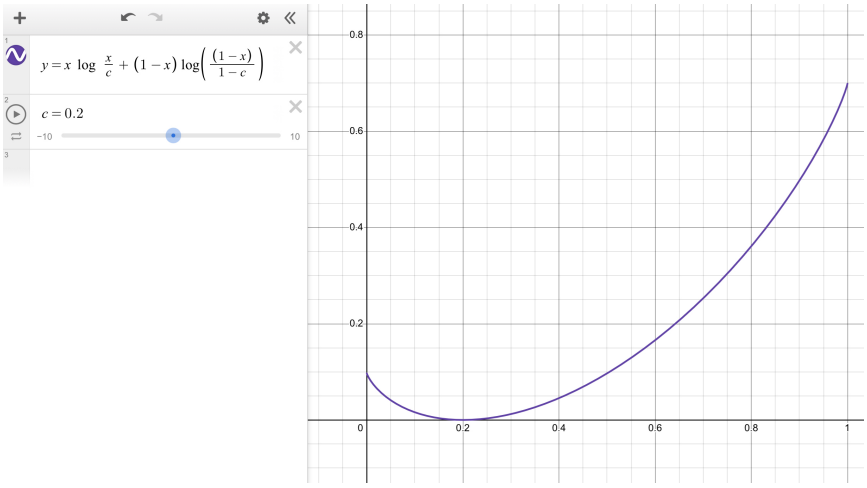
- Beim Entfernen von g Nullmengen ändert sich $\int f(x)dx$ nach Voraussetzung nicht. Dann kann man $f(x)$ durch $\frac{g(x)}{g(x)}$ ergänzen und $\omega(x)g(x)$ bekommen. Das ist aber genau die relative Dichte $\int \omega d\nu$.
- Rechenregeln:
  - $\int h d\mu = \int h \frac{d\mu}{d\nu} \quad \forall h \geq 0$ messbar,
  - $m \ll \nu \ll \eta$ impliziert $\mu \ll \eta$ mit $\frac{d\mu}{d\eta} = \frac{d\mu}{d\nu} \cdot \frac{d\nu}{d\eta}$
  - Umgekehrt, wenn $m \ll \nu$ und $\frac{d\mu}{d\nu} > 0$ ν-f.ü., dann $\nu \ll \mu$ mit $\frac{d\nu}{d\mu} = \frac{1}{\frac{d\mu}{d\nu}}$.
- **Relative Entropie** von μ bzgl. ν ist definiert als

$$H(\mu|\nu) := \int \log\left(\frac{d\mu}{d\nu}\right) d\mu = \int u \frac{d\mu}{d\nu} d\nu \quad \text{wenn } \mu \ll \nu$$

  und ist auch als Kullback-Leibler-Divergenz von μ bzgl ν bekannt.
- **Beispiel:** Sei ν unser Modell, μ tatsächliche Verteilung. Dann ist $H(\mu|\nu) = \int \log \frac{f}{g} d\mu = \int (-\log g - (-\log f)) d\mu$
  - $-\log g$ ist Überraschung in unserem Modell beim Ziehen einer Stichprobe,
  - $-\log f$ ist Überraschung beim korrekten Modell,
  - die Differenz von beiden ist die Überraschung daraus, dass wir aus dem falschen (Alternativhypothese) Modell eine Stichprobe gezogen haben und das Integral ist ein Mittelwert darüber,
  - hätten wir ein korrektes Modell, dann wäre die Überraschung gleich Null.

- **Interpretation 1**: Ist S abzählbar und ν ein Zählmaß (keine Wktsverteilung), dann gilt $H(\mu|\nu) = \sum f(x)\log(f(x)) = -H(\mu)$. Relative Entropie bzgl. des Zählmaßes entspricht also Minus der Entropie.
- **Interpretation 2**: Ist S endlich und $\mu = Unif(S)$, dann gilt $H(\mu|\nu) = \sum f(x) \cdot \log(\frac{f(x)}{1/|S|}) = \sum f(x)(\log(f(x)) + \log(|S|) = -H(\mu) + \log(|S|)$. Entropie der Wktsverteilung ist gleich der Entropie von der Wktsverteilung verschoben um eine Konstante. Insbesondere gilt $H(\mu|\nu) \geq 0 \implies H(\mu) \leq \log(|S|) = H(Unif(S))$. Also hat Gleichverteilung wirklich eine maximale Entropie.
- Eigenschaften der relativen Entropie:
  - $H(\mu|\nu) \geq 0$ und Null gdw $\mu = \nu$
    - $H(\mu|\nu) = \int u\frac{d\mu}{d\nu}d\nu$ (interpretiere als $x\log x$) $\geq \int \frac{d\mu}{d\nu}d\nu - \int 1 d\mu = 0$ nach der Ungleichung $u(x) \geq x - 1$
    - Die Gleichheit in der Abschätzung oben oben gilt gdw $u(\frac{d\mu}{d\nu}) = \frac{d\mu}{d\nu} - 1$ ν-f.ü. $\iff \frac{d\mu}{d\nu} = 1$ ν-f.ü. $\iff \mu = 0$.
  - $H(\mu_1 \otimes \ldots \otimes \mu_n|\nu_1 \otimes \ldots \otimes \nu_n) = \sum_{i=1}^n H(\mu_i|\nu_i)$
    - ohne Beweis, Übung (LOL)

Beispiel: $H(Bernoulli(a)|Bernoulli(b)) = a\log a/p + (1-a)\log\frac{1-a}{1-p}$



Beispiel: Normalverteilungen.

Prof: "Was wir hier sehen, ist so eine Art Abstandsbegriff für Wktsverteilungen, der nicht symmetrisch ist."

5.3 *Relative Entropieminimierung unter Nebenbedingung*

Anwendung: Wir wollen für eine Wktsverteilung μ_0 (Referenz) relative Entropie minimieren und eine andere Wktsverteilung finden, gegeben einen festen Abstand $c \in \mathbb{R}$. Mathematisch gilt $H(\mu|\mu_0) = \min \int T d\mu \geq m$. Das ist genau der Fall, wenn die Verteilung aus der exponentiellen Familie ist.

Exponentielle Familie: wie bereits vor zwei Wochen gemacht. Wichtig ist der Erwartungswert $m(\vartheta) = \int T d\mu_\vartheta$, auch $\mu_\vartheta \ll \mu_0$ mit $\frac{d\mu_\vartheta}{d\mu_0}(x) = \frac{Z(0)}{Z(\vartheta)} e^{\nu T(x)}$

- Gibbssches Variationsprinzip: für $\vartheta \geq 0$ ist μ_ϑ die eindeutige Lösung von dem Minimierungsproblem mit $m = m(\vartheta)$.

Prof: "In der Thermodynamik ist die Aussage, dass diese ... Verteilung minimiert die relative Entropie unter den gegebenen Nebenbedingungen."

12. Vorlesung, 17.05.24

Recap: relative Dichte wohldefiniert wenn Nenner nicht o, relative Dichte als Quotient, relative Entropie als Integral, Bezug zu exponentiellen Familien wenn wir Entropie minimieren, global einfach die Verteilung selbst, lokal (mit bestimmten Erwartungswert) immer eine exponentielle Familie. Das möchten wir beweisen.

- Beweis Gibbssches Variationsprinzip.
  - Definiere relative Entropie von μ bzgl. μ_0, $F(\mu) = H(\mu|\mu_0) - \vartheta \int T d\mu$.
  - $H(\mu_\vartheta|\mu_0) = \min\{H(\mu|\mu_0) : \mu WV mit \int T d\mu = \int T d\mu_\vartheta\}$
  - $H(\mu|\mu_0) = \int \log \frac{d\mu}{d\mu_0} d\mu = \int \log \frac{d\mu}{d\mu_\vartheta} d\mu + \int \log \frac{d\mu_\vartheta}{d\mu_0} d\mu \ldots$

5.4 *Anwendungen in der Statistik*

Wir werden subexponentielle Terme ignorieren. Wenn $a_n, b_n > 0$ zwei Folgen sind, dann sind sie asymptotisch äquivalent, $a_n \simeq b_n$, wenn

$$\lim_{n\to\infty} \left(\frac{a_n}{b_n}\right)^{\frac{1}{n}} = 1,$$

oder,

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{a_n}{b_n}\right) = 0.$$

Prof: wir interessieren uns nur für Abweichungen, die groß genug sind...

- Beispiel: $n^\alpha e^{cn} \simeq n^\beta e^{dn} \iff c = d$.
- Setup: $X_1, \ldots, X_n$ i.i.d., mächtigster Test für $X_i \sim \nu$ vs. $X_i \sim \mu$ ist LQT,
$$\lambda_n(X_1, \ldots, X_n) = \frac{L_n(\mu, X)}{L_n(\nu, X)} = \prod_i w(X_i)$$
das Produkt der relativen Dichte der zugrundeliegenden Veteilungen. Uns interessiert das asymptotische Verhältnis des LQTs für $n \to \infty$...
- Satz von Shannon-McMillan II:
$$Z_n \simeq e^{nH(\mu|\nu)}, \quad Z_n \simeq e^{-nH(\nu|\mu)}$$
Relative Entropie misst die asymptotische Unterscheidbarkeit von μ und ϑ.

- Konsistenz: für $n \to \infty$ die Folge $\hat{\vartheta}_n := \hat{\vartheta}(X_1, \ldots, X_n)$ konvergiert zu ϑ.
- Asymptotische Macht von LQT
  - zu einem festen Niveau α gilt, dass $\lim_{n\to\infty} \frac{1}{n} \log c_n = H(\nu|\mu)$,
  - $\lim_{n\to\infty} \frac{1}{n} \log \mu^n(w_n \leq c_n) \to -H(\nu|\mu)$

5.5 *Fisher-Information*

- Regularitätsannahmen:
  - $\vartheta \neq \vartheta' \implies \mu_\vartheta \neq \mu_{\vartheta'}$
  - $\forall x \in S, \vartheta \in \Theta$: $f_\vartheta(x) > 0$
  - $f_\vartheta(x) \in C^3(S)$
  - $\int \frac{\partial}{\partial\vartheta} f_\vartheta(x) dx = \frac{\partial}{\partial\vartheta} \int f_\vartheta(x) dx = 0$
- Definiere Fisher-Information des parametrischen Modells als

$$I(\vartheta) = \int \frac{1}{f_\vartheta(x)} |\frac{\partial f_\vartheta(x)}{\partial\vartheta}|^2 dx = \int l'(x)^2 f_\vartheta(x) dx$$

13. Vorlesung, 28.05.24

- Zusammenhang mit relativer Entropie – infinitesimale Änderung
- Cramer-Rao-Schranke: für den MSE jedes erwartungstreuen Schätzers gilt

$$MSE_\nu(\hat{g}) = Var_\nu(\hat{g}) \geq \frac{g'(\nu)^2}{I(\nu)}.$$

  - (Konstante Schätzer haben Varianz 0, die Erwartungstreue ist also notwendig.)
  - Beweis: Cauchy-Schwartz für Kovarianz.

Beispiele: Bernoulli, $\hat{p} = X$ mit Varianz $p(1-p) = \frac{1}{I(p)}, g(p) = p$. Optimal! Normalverteilung, $\hat{m} = \overline{X}, Var(\hat{m}) = v = \frac{1}{I(m)}$. Optimal! Allgemein: unter welchen Bedingungen haben wir Gleichheit? Wenn die Cauchy-Schwartz-Ungleichung im Beweis scharf ist bzw. Gleichheit gilt. Das ist genau dann der Fall, wenn $l'(\vartheta, x)$ als affine Funktion von $T(X)$ darstellbar ist, $a(\vartheta)T(X) + b(\vartheta)$.

- Anwendung auf Produktmodell der Normalverteilung, der beste Schätzer im Sinne von oben hat Effizienz $\Omega(1/n)$.

Beispiele: Wieder Normalverteilung, Gleichverteilung auf $(0, \vartheta)$. Wir haben vor vier Wochen einen supereffizienten Schätzer mit Varianz $O(n^{-2})$ ausgerechnet. Was läuft hier falsch? Die Dichte in diesem Fall

ist nicht differenzierbar und die Regularitätsannahme ist damit nicht erfüllt.

14. Vorlesung, 31.05.24

Anwendung von der Informationsungleichung: Lokationsmodell, $f_\vartheta(x) = g(x-\vartheta)$, g ist stetig und symmetrisch. Sei $\hat{\vartheta}$ ein erwartungstreuer Schätzer für ν, dann nach Informationsungleichung gilt

$$Var_\vartheta(\hat{\vartheta}_n) \geq \frac{1}{nI(\vartheta)} \forall \vartheta \in \Theta$$

Es gilt auch $I(\vartheta) = \mathbb{E}\left[l'(\vartheta, X)^2\right] = \int \frac{g'(x-\vartheta)^2}{g(x-\vartheta)} dx = \int \frac{g'(x)^2}{g(x)} dx \forall \vartheta$

- $I(\vartheta)$ groß $\iff$ hohe Unterscheidbarkeit $\iff$ kleine Varianz möglich

Definiere *asymptotische Effizienz* als $Var(\hat{\vartheta}_n) \to \frac{1}{nI(\vartheta)}, n \mapsto \infty$.

- Stichprobenmittlwert...
  - effizient bei der Normalverteilung, nicht asymptotisch effizient bei der Doppelexponentialverteilung, sehr schlecht bei heavy tails
- Stichprobenmedian...
  - nicht asymptotisch effizient bei der Normalverteilung, asymptotisch effizient bei der Doppelexponentialverteilung, akzeptabel bei heavy tails. Median ist sehr robust!

Unterschied liegt darin, dass der MLE für NV und DEV genau der Mittlwert/Median ist.
Asymptotische Normalität von MLE: Satz von Fisher, Wilks, Wald und Beweis(-skizze).

6 *Empirische Verteilung*

15. Vorlesung, 04.06.24

Nicht-parametrische Modelle. Die empirische Wahrscheinlichkeitsverteilung ist eine Zufallsvariable mit Werten in dem Raum von Wktsverteilungen $WV(S)$, da sie von der Stichprobe abhängt.

- empirische Verteilung L_n

Die empirische Verteilung L_n ist eine zufällige Wktsverteilung, d. h. L_n ist eine Zufallsvariable mit Werten in $WV(S)$. Für jede disjunkte Zerlegung von $S = B_1 \cup B_2 \cup \ldots \cup B_n$ gilt $(H(B_1), H(B_2), \ldots, H(B_n)) \sim Mult(n, \mu(B_1), \ldots, \mu(B_n))$.
Wir verwenden $\hat{\mu}_n := L_n$ als Schätzer für μ und es gilt

$$\mathbb{E}[\hat{\mu}_n(B)] = \mu(B)$$

und

$$\mathrm{Var}[\hat{\mu}_n(B)] = \frac{\mu(B)(1-\mu(B))}{n} \leq \frac{1}{4n}.$$

- Satz von Vardarajan: **Konsistenz** von L_n, `https://math.stackexchange.com/questions/2602960/proof-of-weak-convergence-of-empirical-measure`
- Dvoretzky–Kiefer–Wolfowitz Ungleichung: `https://en.wikipedia.org/wiki/Dvoretzky%E2%80%93Kiefer%E2%80%93Wolfowitz_inequality`

6.1 *Plug-in Schätzer*

- statistisches Funktional $g : P \subseteq WV(S) \to \mathbb{R}$
  - Funktion von L_n
- Plug-in Schätzer $\hat{g}_n := g(L_n)$
  - Austausch von der unbekannten Verteilung μ in der Definition (bzw. im Integral) durch die empirische Verteilung L_n und Übergang zum diskreten Fall (also Summen)

6.2 *Bootstrap*

- Bootstrap-Verfahren
- Bootstrap-Stichprobe
  - multinomiales Resampling
- Konfidenzintervall
  - Normalapproximation
  - empirische Quantile: $[2\hat{g}_n - \hat{q}_{1-\frac{\alpha}{2}}(\hat{g}_n^{(1)} \ldots \hat{g}_n^{(B)}), 2\hat{g}_n - \hat{q}_{\frac{\alpha}{2}}(\hat{g}_n^{(1)} \ldots \hat{g}_n^{(B)})]$

16. Vorlesung, 07.06.24

Skipped

6.3 *Anpassungstests*

17. Vorlesung, 11.06.24
18. Vorlesung, 14.06.24

Skipped

- Normalapproximation der Multinomialverteilung
  - degenerierte Normalverteilung $\implies$ Projektion auf Hyperebene
- Chiquadrat-Statistik
- Satz: Die Approximation konvergiert gegen $N(0, I_{k-1})$ für $n \to \infty$
- Satz: Chiquadrat-Statistik konvergiert gegen $\chi^2(k-1)$ für $n \to \infty$

Beweise dazu... sehr lang!

- Anwendung auf Anpassungstests, $H_0 : \mu = \mu_0$
  - χ^2-Test: verwerfe, falls $T = nD_2(L_n|\mu_0) \geq c$,
  - G-Test: verwerfe, falls $G = nH(L_n|\mu_0) \geq \frac{c}{2}$.
  - Beide Tests erfüllen asymptotisch die Niveaubedingung für $n \to \infty$ falls $c = q_{1-\alpha,\chi^2(k-1)}$

$2G$ ist wie T

Ergänzung: hier testen wir, ob die Wahrscheinlichkeitsverteilung die eine Wktverteilung ist. Eigentlich möchten wir testen, ob es eine Verteilung aus gegebener endlichdimensionalen Familie ist $\implies$ der Anpassungstest für parametrische Familien, wobei $H_0 : \mu \in \{\mu_\nu : \vartheta \in \Theta\}$ eine d-dimensionale Mannigfaltigkeit. G-Test mit Parameterschätzung.
Satz von Wilks: unter der Annahme konvergiert $P \circ (2G)^{-1}$ gegen $\chi^2(k-1-d)$.

Wichtig: wir haben nur endlich viele Zustände.

- Alternative zu Anpassungstests: Konfidenzbereich für μ.
  - für Multinomialverteilung bestimme also k Konfidenzbereiche für Binomialverteilung...
  - Vorteil: Kontrolle über jeden einzelnen Parameter
  - Nachteil: für $n = 10$ müssen einzelne Konfidenzbereiche zu $\frac{\alpha}{10}$ bestimmt und geschnitten werden. Brauche mehr Stichprobenwerte, um dasselbe Niveau zu erreichen. ("Bonferroni-Korrektur").

19. Vorlesung, 18.05.24

- Satz von Glivenko-Cantelli
- Wie testet man auf Normalverteilung? Weitere Anpassungstests:
  - Kolmogorov-Smirnov-Test
  - Lilliefors-Test auf Normalverteilung
    - β ist skaleninvariant
    - Probleme mit tails

- Anderson-Darling-Test
- Abweichungen in tails sind nur schwer erkennbar. Alternative: wir verwerfen, falls ein Integral größer ε ist.

Anmerkung: kein Anpassungstest ist perfekt. Mehrere Tests ausprobieren.

6.4 Robuste Verfahren

μ auf $\mathbb{R}$, $X_1, ..., X_n \sim \mu$ i.i.d.

- ein fehlender Datenwert kann $\overline{X}_n$ beliebig verfälschen = nicht robust
- besser ist z. B. der Median $q_{\frac{1}{2}}$, Ordnungsstatistik = robust

Wir charakterisieren Robustheit von T_n bei $(x_1, \ldots, x_n)$ über:

- Bruchpunkt

$$\varepsilon_n(X) = \frac{1}{n} \max_{k \in \{0,1,\ldots,n\}} \{ \sup_{y_i \neq x_i k-mal} |T_n(y_1, \ldots, y_n)| < \infty \}$$

- asymptotischen Bruchpunkt $\varepsilon := \liminf_{n\to\infty} \varepsilon_n(X)$
- die Sensitivitätsfunktion
- die Einflussfunktion

Robuste Konfidenzintervalle für Quantile

- Schätze q_γ mit Plugin-Schätzer $\hat{q}_\gamma = X_{(n\gamma)}$. Bestimme Konfidenzintervall
- Satz: die Wkt, dass das Gamma-Quantil zwischen zwei Ordnungsstatistiken liegt, beträgt die Differenz der Verteilungswerte.
  - Verteilungsunabhängige Konfidenzintervalle! Großer Vorteil.

Robuste Konfidenzintervalle für Tests

- Test für Median von zwei Stichproben
  - Option 1, Mittelwert: t-Test. Probleme: keine Normalverteilung, Ausreißer/Messfehler,
  - Option 2, Median: verwerfe H_0 falls $X_{(k)} > 0$. Pearsons Vorzeichentest. Vorteil: sehr robust, unabhängig von der Verteilung. Nachteil: Größe der X spielt keine Rolle.
  - Option 3, Wilcoxon-Signed-Rank-Test:

20. Vorlesung, 21.06.24

7 Zusammenhang mehrerer Merkmale

7.1 Binäre Merkmale: Chancenquotienten und Vierfeldertafeln

21. Vorlesung, 25.06.24

- $Chance(A) := \frac{P(A)}{P(A^C)} = \frac{P(A)}{1-P(A)}$
  - nichtlineare Transformation $p \mapsto \frac{p}{(1-p)}$
- Chancenquotient $\rho := \frac{p_{11}p_{00}}{p_{10}p_{01}} = \frac{Chance(X=1|Y=1)}{Chance(X=1|Y=0)}$,
  - Aus $\rho = 1$ folgt X, Y unabhängig und $p_{kl} = p_{k+}p_{+l}$
- Schätzer für ρ: Plug-in Schätzer $\hat{\rho}$
- Konfidenzintervall für ρ
  - Modell: $(H_{11}, H_{10}, H_{01}, H_{00}) \sim Mult(N, p_{11}, p_{10}, p_{01}, p_{00})$

22. Vorlesung, 28.06.24
Dümbgen, 7.5

  - Dieselben KI wie beim **Fischers exakten Test** mit Konfidenzintervallen. Idee: gegeben ρ, können wir die bedingte Wkt. $p(H_{11}|H_{1+})$ und $p(H_{11}|H_{+1})$ ausrechnen.

Haarfarbe	Augenfarbe		
	grün	nicht grün	
rot	14	57	71
nicht rot	50	471 H_{00}	521
	64	528	592 N

  - Lemma: $p(H_{11}|H_{+1} = l, H_{1+} = n) = C_{n,l,N,\rho}\binom{n}{x}\binom{N-n}{l-x}\rho^x$ und die Konstante C hängt nicht von x ab.
  - Wir können aus diesen Wkten die Konfidenzintervalle für ρ herleiten. Bedingte Wkt => Verteilungsfunktion => Konfidenzschranken über Umkehrfunktion
  - Bei diesem Test typisch ist $H_0 : \rho = 1$, $H_1 : \rho > 1$.
- **Vierfeldertafeln, modifiziertes Modell**.
  - Im Unterschied zur oben sind n_1, n_2 fest vorgegeben.

	Besserung	keine Besserung	
Medikament	H_{11}	H_{12}	n_1
Placebo	H_{21}	H_{22}	n_2
	H_{+1}	H_{+2}	N

  - Modell: $H_{11} \sim Bin(n_1, p_1), H_{21} \sim Bin(n_2, p_2)$ unabhängig.
  - Wie oben.

7.2 Test auf Unabhängigkeit/Assoziation

Gegeben $X_{1\leq k\leq n}$ und $Y_{1\leq l\leq N}$ entscheide, ob die gemeinsame Verteilung das Produkt der Randverteilungen ist, d. h. wenn $p_{kl} = p_k p_l =: \hat{p}_{kl}$ für alle k, l.

- Die Nullhypothese $H_0 : p = \hat{p}$
- Methode 1, multiple Tests: Fishers exakter Test für jeden k und l
- Methode 2, Chiquadrat-Test auf Unabhängigkeit: schätze χ^2-Divergenz

$$\chi^2(p|\overline{p}) = \sum_{k,l} \left(\frac{p_{kl}}{\overline{p}_{kl}} - 1 \right)^2 p_{kl} = \sum_{k,l} \frac{p_{kl}^2}{\overline{p}_{kl}} - 1$$

  Die Schätzung machen wir mit dem Plug-In Schätzer (empirische Verteilung). Unter Nullhypothese ist der Wert Null.

  - Berechnung des p-Werts: entweder exakt (bei kleinen Datensätzen) oder Bootstrap.
  - Alternativ kann man Divergenz durch relative Entropie ersetzen.
- χ^2-Approximation (N groß) mit $(k-1)(l-1)$ Graden.
  - Nachrechnen bestimmter Eigenschaften der χ^2-Verteilung

7.3 Permutationstests

23. Vorlesung, 02.07.24
Dümbgen 8.1-8.3

- (H_0') : **bedingte Austauschbarkeit** von Y gegeben X, wenn für $\pi \in Sym(n)$ gilt $(X, \pi Y) \sim (X, Y)$.
- Lemma: Sei π zufällige Permutation, dann gilt für alle feste x, y, dass Erwartungswert der Häufigkeiten unter π dem naiven Produktmaß entspricht. Der Erwartungswert der Statistik beträgt $\frac{N}{N-1}(k-1)(l-1)$.
- Satz: ...

Ziel: Untersuchen, ob die ZV unter Permutation invariant in der Verteilung sind. Gegeben ist eine Permutation $\pi \in G$, dabei G eine Untergruppe von $Sym(n) \implies G$-Invarianztest.

- Test auf identische Verteilung von X, Y
  - $\pi(X, Y) = (\pi X, \pi Y) \sim (X, Y)$
- Test auf Unabhängigkeit X, Y
  - bedingte Austauschbarkeit $\pi(X, Y) = (X, \pi Y)$
- Test auf Vorzeichensymmetrie
  - Tendenz für positive/negative Vorzeichen

Konstruktion von Tests via G-Symmetrie. Sei T eine beliebige Statistik, $(x, y) \in S^{n+m}$ Datenvektor. Definiere

- linksseitiger p-Wert: $p_l^G(x,y) = \frac{|\{\pi \in G : T(\pi(x,y)) \leq T(x,y)\}|}{|G|} = P(T(\pi(x,y)) \leq T(x,y)) = F_{\{T(\pi(x,y))\}}(T(x,y))$.
- Berechnung von $p_l^G(x,y)$. Abzählen (wenn G klein), Monte-Carlo.
- Satz: unter $(X,Y) \sim \pi(X,y)$ gilt $p_l^G(X,Y) \leq \alpha$ kann höchstens mit Wkt. α gelten.

Beispiel 1: Test auf identische Verteilung, $G = Sym(n+m), T(X_1,...,X_n,Y_1,...,Y_m) = |M_X - M_Y|$, dabei sind M_X, M_Y Mittelwerte/Mediane/... von X und Y.
Beispiel 2: Test auf Unabhängigkeit, $G = Sym(n), (X_n, Y_n)$ aus $\{0,1\} \times \mathbb{R}$ und $\pi(X,Y) = (X, \pi Y), T(X,Y) = \sum_i 1_{X_i=1} R_i$. Dann könnte man erkennen, ob eventuell kleine/große Ranks von Y überwiegen.

8 Regression

8.1 Einfache lineare Regression

24. Vorlesung, 05.07.24

8.2 Lineare Modelle

25. Vorlesung, 09.07.24

- Modell: $Y = \sum_k^d \omega_k X^{(k)} + \varepsilon = A\omega + \varepsilon$
  - Beobachtung $Y : \Omega \to \mathbb{R}^n$
  - unbekannte Parameter $\omega_1, \ldots, \omega_d \in \mathbb{R}$
  - Störgrößen $\varepsilon : \Omega \to \mathbb{R}^n$
  - Design-Matrix $A = (X^{(1)}, \ldots, X^{(d)})$ eine $n \times d$-Matrix mit Rang d
- Beispiele: einfache Regression, multiple lineare Regression, polynomielle Regression, Einstichproben-Lokationsmodell, ...
- Beispiel Mehrstichprobenmodell: Vergleich von p verschiedenen Populationen, z. B. Ertrag bei p verschiedenen Düngersorten.

  $(i = 1, \ldots, p; k = 1, \ldots, n_i)$, n_i Anzahl Stichproben in Gruppe i

  - Response $Y_{ik} = m_i + \varepsilon_{ik}$ von Objekt k in Gruppe i
  - $A = \begin{pmatrix} 1_{n_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1_{n_p} \end{pmatrix}$, $1_{n_i} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{n_i}$
  - Idee: ANOVA, vergleiche die Varianz zwischen den Gruppen mit der Varianz innerhalb der Gruppen, $H_0 : m_1 = \ldots = m_p$.
  - Vergleiche $\hat{V}_{ZG}$ und $\hat{V}_{iG}$
  - Gauß-Fall: $\frac{\hat{V}_{ZG}}{\hat{V}_{iG}} \sim \frac{\frac{v}{p-1}\sum_{i=1}^{p-1} Z_i^2}{\frac{v}{n-p}\sum_{i=p}^{n-1} Z_i^2} = F_{p-1,n-p}$ und $Z_i \sim N(0,1)$ unabhängig.
  - Die Verteilung von $F_{p-1,n-p}$ heißt Fisher-Verteilung mit $p-1$ und $n-p$ Freiheitsgraden.

Schätzer für ω: $Y = A\omega + \varepsilon$

- kann nicht perfekt klappen, $\dim A < \dim Y, d < n$
- $L := Range(A) = \{A\omega : \omega \in \mathbb{R}^d\}$
- $\pi : \mathbb{R}^n \to L$ orthogonale Projektion auf $Range(A)$
- Lemma: $A^\top A \in \mathbb{R}d \times d$ ist invertierbar und $\Pi = A(A^\top A)^{-1}A^\top$
- Ansatz 1: Kleinste Quadrate
  - Minimiere $||Y - A\omega||^2 = ||Y||^2 - 2(A^\top y) \cdot \omega + \omega \cdot A^\top A\omega ! = \min$
  - Berechne den Gradienten $\nabla = -2A^\top y + 2A^\top A\omega ! = 0$
  - $A^\top A\omega = A^\top Y$ Normalengleichungen und $\hat{\omega} = (A^\top A)^{-1}A^\top Y$ Minimum, denn Hesse-Matrix $= 2A^\top A$ ist symmetrisch positiv definit.
  - Bemerkung: $A\hat{\omega} = \Pi Y$ nach Lemma

8.3 Andrere Regressionsverfahren

9 Bayes-Statistik

9.1 Ansatz der Bayesschen Statistik

9.2 Gibbs-Sampling

10 Klausurvorbereitung

Liste A1)

- Verteilungsfunktion
- Konfidenzintervall
- Median
  - `https://math.stackexchange.com/questions/113270/the-median-minimizes-the-sum-of-absolute-deviations-t`
- Chiquadrat-Verteilung
- Empirische Verteilung
- Normalverteilung
  - `https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule`
- Binomialverteilung
- Studentsche t-Statistik
- Exponentialverteilung
  - `https://stats.stackexchange.com/questions/2092/relationship-between-poisson-and-exponential-distribut`