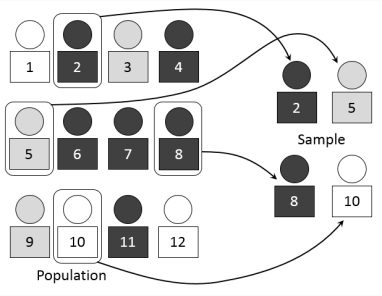
Statistik

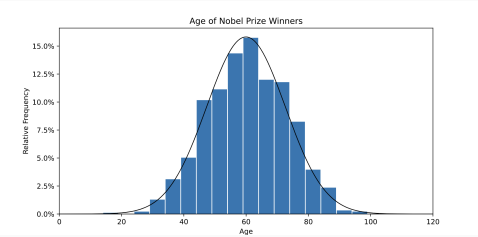
Vorlesung 8 - Parameterschätzung Teil 1: Stichproben und Schätzer

Prof. Dr. Sandra Eisenreich

Hochschule Landshut

Motivation





Grundsituation der schließenden Statistik:

- geg.: stichprobenartige Daten,
- ges.: Informationen über die Grundgesamtheit
- Wie? Bestimme die zugrundeliegende Verteilung und die Parameter davon (z.B. μ, σ, p, λ...)

Beispiel

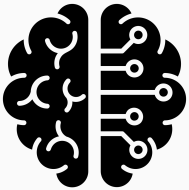
- Von 100.000 Bauteilen werden stichprobenartig 1.000 überprüft, 2 davon defekt.
- Verteilung von X= "Anzahl von niO Bauteilen":

$$b_{100.000,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}, p \text{ unbekannt}$$

- "statistisches Modell" ="$b_{100.000,p}$, p gesucht"

Wie können wir p bestimmen? Intuitiv: Schätzwert= $p = \frac{2}{1000}$.
(warum?? - das lernen wir in diesem Kapitel).

Motivation: Probabilistic ML = Lernen einer Verteilung aus Daten



- ML Modell: angenommene Wahrscheinlichkeitsverteilung mit vielen vielen Parametern (z.B. $N(\mu = f(x), \sigma^2)$, wobei der $f(x)$ von vielen Parametern abhängt)
- Vorhersage: der wahrscheinlichste Wert
- Training des ML-Modells: aus Daten = Stichproben die besten Parameter bestimmen.

Wie? $\rightarrow$ Parameterschätzung.

Agenda

1. Stichproben
2. Grundlagen der Parameterschätzung: Statistisches Modell und Punktschätzer
3. Grundlagen der Parameterschätzung: Gütemaße und Eigenschaftern von Schätzern

Stichproben

Stichprobe

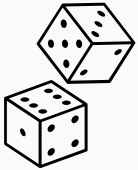
Definition

n Beobachtungswerte $x_1, x_2, \ldots, x_n$ heißen Stichprobe vom Umfang n, die Werte x_i heißen Stichprobenwerte. Wird eine Stichprobe durch ein Zufallsexperiment gewonnen, so heißt sie (Zufalls-)Stichprobe.

Notation: wird ein Experiment n-mal unabhängig durchgeführt, so ist

- X_i = Ausgang des i-ten Experiments
- Ausgang von allen = Zufallvektor $(X_1, \ldots, X_n)$
- Realisierung = Stichprobenwerte $(x_1, \ldots, x_n)$

Beispiel - Würfelexperiment

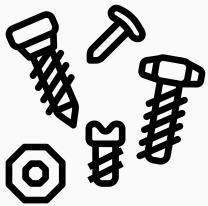


10-mal würfeln, $X_i =$ Ergebnis des i-ten Wurfs

Stichproben sind z.B.

$$S_1 = (2, 5, 2, 6, 2, 5, 5, 6, 4, 3) \quad \text{und}$$
$$S_2 = (4, 6, 2, 2, 6, 1, 6, 4, 4, 1)$$

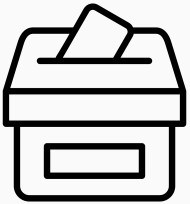
Beispiele - Defekte und Wahlumfragen



- In einer Kiste sind 1000 Schrauben. Stichprobe: 100,

$$X_i = \text{Zustand der Schraube } i \in \{0, 1 = \text{defekt}\}$$

Stichprobe ist Realisierung des Zufallsvektors $(X_1, X_2, \ldots, X_{100})$, z.B.: $(0, 1, 0, 0, 0, 0, 1, 1, 0, 0, \ldots, 1)$



- Wahlumfrage: Stichprobe besteht aus k zufällig ausgewählten Wahlberechtigten.

$$X_i = \text{Wahlverhalten des Wählers } i.$$

Wahlumfrage ist Realisierung von $(X_1, X_2, \ldots, X_n)$.

Motivation - Mittelwert und Varianz einer Stichprobe

Erinnerung: Erwartungswert einer Zufallsvariable = "physikalischer Schwerpunkt"

Vermutung 1: Schätzwert für den Erwartungswert einer Stichprobe = Mittelwert $\overline{x}$

Erinnerung: Varianz = erwartete quadratischen Abweichung vom Erwartungswert.

Vermutung 2: Schätzwert für die Varianz einer Stichprobe:

- berechne die quadratischen Abweichungen $(x_i - \overline{x})^2$
- Berechne den Erwartungswert dieser Werte, also ihren Mittelwert: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Vermutung 1 ist richtig, Vermutung 2 fast!

Mittelwert und Varianz einer Stichprobe

- der Mittelwert oder das arithmetische Mittel der Stichprobe $(x_1, \ldots, x_n)$:

$$\overline{x} := \frac{1}{n}(x_1 + \ldots + x_n),$$

- die mittlere quadratische Abweichung/mean squared error der Stichprobe (auch mittlere Summe der Quadrate genannt):

$$m := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- die (korrigierte) Varianz s^2 und die Standardabweichung s der Stichprobe:

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, s := \sqrt{s^2}$$

Mittelwert und Varianz als Zufallsvektoren

$(x_1, \ldots, x_n)$ = eine Realisierung des Zufallsvektors $(X_1, X_2, \ldots, X_n)$. Dann sind $\overline{x}, m, s^2$ der vorigen Seite realisierungen der Zufallsvariablen

$$\overline{X} := \frac{1}{n}(X_1 + \ldots + X_n) := \frac{1}{n} S_n$$

$$MSE := \frac{1}{n}\left((X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2\right)$$

$$S^2 := \frac{1}{n-1}\left((X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2\right) = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} X_i^2\right) - n\overline{X}^2\right)$$

Dies sind alles Funktionen in den Zufallsvariablen X_i.

Unabhängig und identisch verteilt (iid)

Definition

Eine Stichprobe $(x_1, \ldots, x_n)$ heißt unabhängig und identisch verteilt oder iid (independent and identically distributed), falls die zugrundeliegenden Zufallsvariablen unabhängig sind und dieselbe Wahrscheinlichkeitsverteilung bzw. -dichte besitzen.

Beispiele: Ein n-stufiges Bernoulli-Experiment, n-mal Würfeln, ...

Im Machine Learning geht man oft davon aus, dass die zugrundeliegenden Daten iid sind.

Grundlagen der Parameterschätzung: Statistisches Modell und Punktschätzer

Modellrahmen der schließenden Statistik

Im Bauteil-Beispiel sucht man die wahre Verteilung aus folgendem sogenannten statistische Modell:

- die Familie von Wahrscheinlichkeitsmaßen $b_{100.000,p}$
- mit unbekanntem Parameter p
- aus dem Parameterraum $[0, 1]$.

Allgemein:

gegeben: Zufallsvektor $X = (X_1, \ldots, X_n)$ dessen Ergebnis einer Stichprobe $(x_1, \ldots, x_n)$ in einem Stichprobenraum $\mathcal{X} \subset \mathbb{R}^n$ ist.

gesucht: das X zugrunde liegende Wahrscheinlichkeitsmaß P.

Statistisches Modell

gegeben:

- ein Stichprobenraum $\mathcal{X} \subset \mathbb{R}^n$.
- eine Familie von Wahrscheinlichkeitsmaßen P_θ abhängig von einem oder mehreren
- Parametern θ aus einem geeigneten Parameterraum $\Theta \subset \mathbb{R}$, s.d.
- s.d. $\theta \mapsto P_\theta$ injektiv ist (unterschiedliche Parameter $\rightarrow$ unterschiedliche Wahrsch.)

gesucht: Der "wahre" Parameter θ.

Definition

Man nennt $(\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ ein statistisches Modell.

Statistisches Modell im Bernoulli-Experiment

$S_n = X_1 + \cdots + X_n$: Anzahl von Treffern in den n Experimenten

- Stichprobenraum: $\mathcal{X} := \{0,1\}^n$
- Wahrscheinlichkeitsmaß: für $x = (x_1, \ldots, x_n) \in \mathcal{X}$ mit $x_1 + \cdots + x_n = k$ gilt

$$P_\theta(S_n = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}, \quad k = 0, 1, \ldots, n.$$

- Parameter=Trefferwahrscheinlichkeit $\theta = p$ aus dem Parameterraum $\Theta := [0,1]$

Frage: Was ist ein guter Schätzwert für p? $\rightarrow$ die relative Trefferhäufigkeit.

relative Trefferhäufigkeit

In einem Bernoulli-Experiment sei X_i der Ausgang des i-ten Experiments, und $S_n = X_1 + \ldots X_n$ die Anzahl von Treffern. $(x_1, \ldots, x_n)$ sei eine Stichprobe von $(X_1, \ldots, X_n)$. Dann ist:

- die Zufallsvariable relative Trefferhäufigkeit gegeben durch $\frac{S_n}{n}$, und
- die relative Trefferhäufigkeit der Stichprobe (die Realisierung von $\frac{S_n}{n}$) gegeben durch

$$\frac{\sum_{i=1}^{n} x_i}{n}.$$

Beispiel: Statistisches Modell bei Ziehen ohne Zurücklegen: Qualitätskontrolle

Beispiel eines Settings "Ziehen ohne Zurücklegen":

- Warensendung mit N Sendungen liegt vor; θ davon seien defekt und $N - \theta$ intakt
- N sei bekannt und θ unbekannt
- der Sendung werden zufällig n Bauteile entnommen und auf Intaktheit geprüft

Was ist das Statistische Modell? Was wäre instinktiv ein guter Schätzwert für θ?

Ergebnis: Statistisches Modell bei Ziehen ohne Zurücklegen: Qualitätskontrolle

- Stichprobenraum: $\mathcal{X} := \{0,1\}^n$
- Wahrscheinlichkeitsmaß: Die Anzahl $S_n := X_1 + \cdots + X_n$ der defekten Bauteile der Stichprobe besitzt die hypergeometrische Verteilung $\mathrm{Hyp}(n, \theta, N-\theta)$, also für $x = (x_1, \ldots, x_n) \in \mathcal{X}$ mit $x_1 + \cdots + x_n = k$ besitzt

$$P_\theta(S_n = x) = \frac{\theta(\theta-1)\cdots(\theta-k+1)(N-\theta)(N-\theta-1)\cdots(N-\theta-(n-k)+1)}{N(N-1)\cdots(N-n+1)}$$

- Parameter: θ
- Parameterraum: $\Theta = \{0, 1, \ldots, N\}$

guter Schätwert für θ:

$$E(S_n) = n \cdot p = n \cdot \frac{\theta}{N} = k \Rightarrow \theta = k \cdot \frac{n}{N}$$

Beispiel: Statistisches Modell für die Normalverteilung

In der Produktion wird stichprobenartig die Länge jedes 20. Bauteils gemessen und dessen Abweichung X_i vom Sollmaß festgestellt. Dann sind alle X_i normalverteilt.

- Was sind Erwartungswert (sollte 0 sein) und Varianz?
- Was sind Schätzwerte für μ und σ^2?

Ergebnis: Statistisches Modell für die Normalverteilung

- Stichprobenraum: $\mathcal{X} = \mathbb{R}^n$
- Parameter: $\theta = (\mu, \sigma^2)$ Erwartungswert und Varianz
- Wahrscheinlichkeitsmaß: $X = (X_1, \ldots, X_n)$ mit $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Für $x = (x_1, \ldots, x_n) \in \mathcal{X}$ gilt für die Wahrscheinlichkeitsdichte f zu P

$$f_\theta(X = x) = \prod N(x_i | \mu, \sigma^2)$$

- Parameter: μ und σ^2
- Parameterraum: $\Theta := \mathbb{R} \times \mathbb{R}_{\geq 0}$

guter Schätwert für μ, σ^2: Mittelwert und Varianz der Stichproben, also die Realisierung von $\overline{X} = \frac{S_n}{n}$ und $S^2 = \frac{1}{n-1}\left((X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2\right)$. = Punktschätzer

Ein Zufallsexperiment wird unabhängig n-mal wiederholt wird, X_i = Ausgang des i-ten Experiments. Sei $(x_1, \ldots, x_n)$ eine Stichprobe.

- Es sei f eine Funktion, s.d. $\widetilde{\theta} = f(x_1, \ldots, x_n)$ ein ein Schätzwert für θ ist.
- Dann heißt die Zufallsvariable

$$T = f(X_1, \ldots, X_n)$$

Schätzfunktion oder Punktschätzer für den Parameter θ.

In unserem Bauteil-Beispiel ist z.B.: $T = \frac{S_n}{n}$

Schätzfunktion / Punktschätzer - mathematischere Definition

Es seien $(\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ ein statistisches Modell und $\tilde{\Theta} \supset \Theta$. Dann heißt jede Abbildung

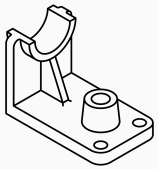
$$T : \mathcal{X} \to \tilde{\Theta}$$

ein Punktschätzer für θ. Für $x \in \mathcal{X}$ heißt der Wert $T(x)$ konkreter Schätzwert für θ zu x.

- die Bezeichnung "Punktschätzer" rührt daher, dass die Schätzwerte einzelne Elemente ("Punkte") von $\tilde{\Theta}$ sind
- im Gegensatz dazu stehen die als Bereichsschätzer bezeichneten Konfidenzbereiche / Konfidenzintervalle (siehe später)

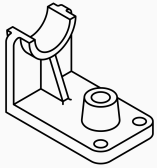
Grundlagen der Parameterschätzung: Gütemaße und Eigenschaftern von Schätzern

Motivation: Schätzfehler



- Schätzer für Anteil von defekten Bauteilen: relative Trefferhäufigkeit $T = \frac{S_n}{n}$
- Stichprobe: 2 von 1000 Bauteilen defekt
- $\Rightarrow \widetilde{p} = \frac{2}{n}$ (Realisierung von T)
- aber: "wahre"Wahrscheinlichkeit $\theta = \frac{3}{1000}$
- $\Rightarrow$ zufälliger Schätzfehler $= \frac{2}{1000} - \frac{3}{1000} =$ Realisierung der ZV $T - \theta$

Motivation: Verzerrung/Bias und Erwartungstreue

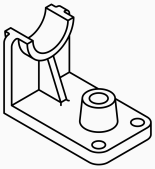


- zufälliger Schätzfehler $= T - \theta$
- Wie berechnet man den mittleren zufälligen Schätzfehler?
- Idee: Viele Stichproben vom Umfang n, berechne den Mittelwert der Schätzfehler (z.B. 2, 3, 4, 3 Defekte von 1000 $\Rightarrow$ Schätzfehler $\frac{-1}{1000}, \frac{0}{1000}, \frac{1}{1000}, \frac{0}{1000} \Rightarrow 0$)

- besser: berechne als mittleren Schätzfehler den Erwartungswert $E_\theta(T) - \theta =$ Verzerrung/Bias
- Falls $E_\theta(T) = \theta$ (kein Bias) heißt T erwartungstreu.

Hinweis: Zum Berechnen von $E_\theta(X)$ braucht man $P_\theta(X = x)$ (hängt von θ ab).

Motivation: mittlere quadratische Abweichung/MSE



- zufälliger Schätzfehler $= T - \theta$
- $\Rightarrow$ quadratische Abweichung $= (T - \theta)^2$
- $\Rightarrow$ mittlere quadratische Abweichung/mean squared error (MSE) ist: $E_\theta[(T - \theta)^2]$.

Hinweis: Der MSE ist ein Standard-Gütemaß für ML-Modelle.

Verzerrung, MSE, erwartungstreu

(a) mittlere quadratische Abweichung /mean squared error (MSE) von T an der Stelle θ:

$$MSE_T(\theta) := E_\theta[(T-\theta)^2]$$

(b) Verzerrung (Bias) von T an der Stelle θ:

$$b_T(\theta) := E_\theta(T) - \theta$$

(c) T heißt erwartungstreu für θ, falls gilt: $E_\theta(T) = \theta$ für jedes $\theta \in \Theta$.

Bias-Variance Tradeoff

Ideal wäre: geringer Bias, geringe Varianz (möglichst genaue Schätzungen). Aber leider hängen beide zusammen:

Theorem (Bias-Variance-Tradeoff)

$$MSE_T(\theta) = Var_\theta(T) + (b_T(\theta))^2$$

Anwendung: Trainiert man ML Modelle, MSE zu minimieren, kann man unter gleichwertigen Modellen nur entweder Bias reduzieren und dabei erhöhrte Varianz in Kauf nehmen, oder umgekehrt.

Motivation: steigende Stichprobenzahlen

Intuitiv sollte klar sein: umfangreichere Stichproben $\Rightarrow$ besserer Schätzer T_n.

- wenn ein Schätzer für $n \to \infty$ erwartungstreu wird, nennt man das asymptotisch erwartungstreu.
- wenn bei steigender Stichprobengröße der Schätzer immer genauer wird (das heißt kleinere Varianz), heißt er konsistent.

Eigenschaften von Schätzern bei wachsendem Stichprobenumfang

Seien $X_1, X_2, \ldots$ i.i.d. Zufallsvariablen, deren Verteilung von einem reellen Parameter $\theta \in \Theta$ abhängt. Dann heißt die Schätzfolge (T_n)

(a) asymptotisch erwartungstreu für θ, falls

$$\lim_{n\to\infty} E_\theta(T_n) = \theta \quad \forall \theta \in \Theta,$$

(b) konsistent für θ, falls für jedes $\theta \in \Theta$ gilt:

$$\lim_{n\to\infty} P_\theta(|T_n - \theta| \geq \varepsilon) = 0 \quad \forall \varepsilon > 0.$$

Eine Schätzfolge ist konsistent für θ, wenn $\lim_{n\to\infty} Var(T_n) = 0$.

Bernoulli-Experiment

Bernoulliexperiment (Ergebnisse 0,1) mit Trefferwahrscheinlichkeit p.

Theorem

Die relative Trefferhäufigkeit $T_n = T_n(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i$ *ist erwartungstreuer und konsistenter Punktschätzer für das n-stufige Bernoulli-Experiment* $(X_1, \ldots, X_n)$.

Begründung: Bernoulli-Experiment

- $S_n := \sum_{i=1}^{n} X_i$ ist binomialverteilt

$$\Rightarrow E(S_n) = np, Var(S_n) = np(1-p)$$

- damit ist

$$E(T_n) = \frac{1}{n} E(S_n) = p$$

$$Var(T_n) = \frac{1}{n^2} Var(S_n) = \frac{p(1-p)}{n} \Rightarrow \lim_{n \to \infty} Var(T_n) = 0$$

Schätzer für Erwartungswert und Varianz

Theorem

Die Zufallsvariable X beschreibe ein Zufallsexperiment, das n-mal unabhängig wiederholt wird, X_i sei der Ausgang des i-ten Experiments. X habe den Erwartungswert μ und die Varianz σ^2. Dann gilt

1. *erwartungstreuer und konsistenter Punktschätzer für μ:*

$$\overline{X} := \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

2. *erwartungstreuer Punktschätzer für σ^2:*

$$S^2 := \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} X_i^2\right) - n\overline{X}^2\right)$$

Eigenschaft 1

Es ist zu zeigen, dass $E(\overline{X}) = \mu$ (erwartungstreu) und $Var(\overline{X}) = \frac{\sigma^2}{n}$ (konsistent). Da $E(X) = \mu$, ist auch $E(X_i) = \mu$ (immer das gleiche Experiment). Analog: $Var(X_i) = \sigma^2$. Damit ergibt sich:

$$
\begin{aligned}
E(\overline{X}) &= E\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) \\
&= \frac{1}{n}(E(X_1) + E(X_2) + \ldots + E(X_n)) \text{ (Rechenregeln für } E) \\
&= \frac{1}{n}(\mu + \mu + \ldots + \mu) = \frac{n}{n}\mu = \mu \\
Var(\overline{X}) &= Var\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) \\
&= \frac{1}{n^2}(Var(X_1) + Var(X_2) + \ldots + Var(X_n)) \text{ (Rechenregeln für } Var) \\
&= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}
$$

Eigenschaft 2

Zu zeigen ist $E(S^2) = \sigma^2$. Wir verwenden die Regel $E(Y^2) = Var(Y) + E(Y)^2$.

Es war $E(X_i) = \mu$, $Var(X_i) = \sigma^2$. Damit ist $E(X_i^2) = \sigma^2 + \mu^2$.

Außerdem $E(\overline{X}) = \mu$, $Var(\overline{X}) = \frac{\sigma^2}{n}$ (aus 1.). Damit ist $E(\overline{X}^2) = \frac{\sigma^2}{n} + \mu^2$.

Damit ergibt sich

$$
\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\left(\left(\sum X_i^2\right) - n\overline{X}^2\right)\right) = \frac{1}{n-1}\left(\left(\sum E(X_i^2)\right) - nE(\overline{X}^2)\right) \\
&= \frac{1}{n-1}\left(\underbrace{\sum_{i=1}^{n}(\sigma^2 + \mu^2)}_{n(\sigma^2+\mu^2)} - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\
&= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2
\end{aligned}
$$

Literatur

- Hartmann, Peter; Mathematik für Informatiker, Springer-Vieweg; 7. Auflage; 2019
- Henze, Norbert; Stochastik für Einsteiger; Springer; 10. Auflage; 2013
- Jurafsky, D; Martin, J.H; Speech and Language Processing; Third Edition Draft; 2020
- Mitchell, T.M; Machine Learning; The McGraw-Hill Companies, Inc.; 1997
- Witten, I.H.; Frank, E.; Data Mining; Morgan Kaufmann Publishers; Second Edition; 2005