Einführung in die induktive Statistik

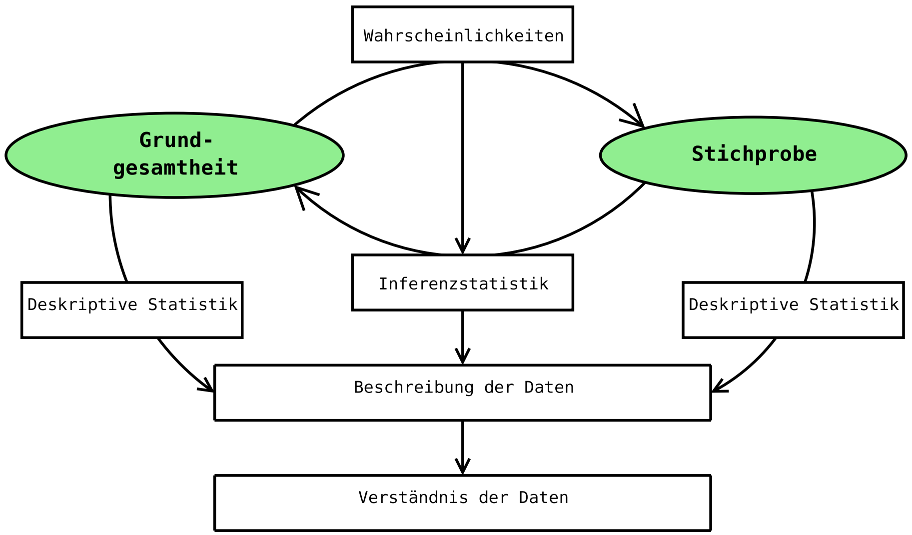
Friedrich Leisch

Institut für Statistik
Ludwig-Maximilians-Universität München

SS 2009

Modellanpassung

Statistik



Statistik

Wahrscheinlichkeitsrechnung: Gesetze bekannt, mit welchen Wahrscheinlichkeiten beobachte ich gewisse Ereignisse?

Statistik: Ein bestimmtes Ereignis ist eingetreten (Daten der Stichprobe), welche Rückschlüsse können wir über die datenerzeugenden Prozesse ziehen?

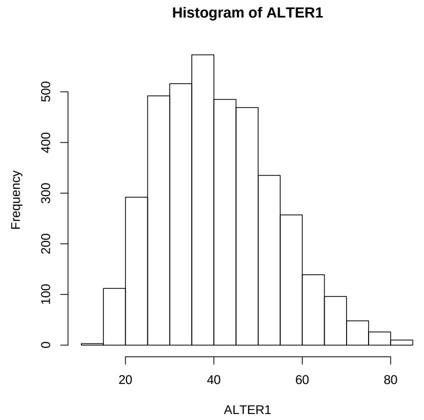
Beispiel: GBÖ Winter 1994

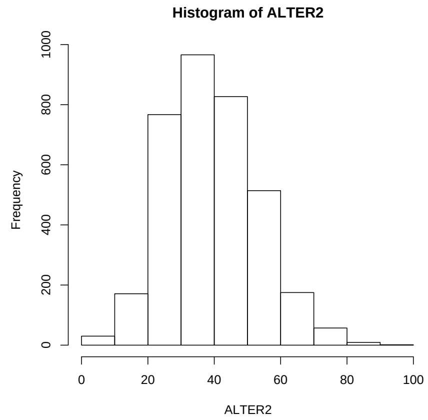
Gäste-Befragung Österreich: wird alle 2 Jahre im Auftrag der österreichischen Fremdenverkehrswerbung durchgeführt (getrennt für Sommer und Winter).

Gäste füllen (recht umfangreiche) Fragebögen direkt am Urlaubsort (Hotel, . . .) aus: $n = 3853$, mehr als Hundert Fragen pro Bogen:

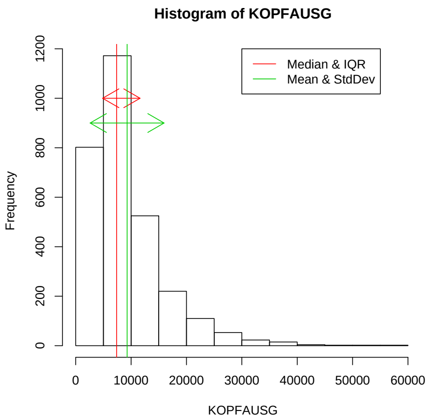
- Alter, Geschlecht, Herkunftsland, Einkommen, Ausgaben?
- Motivation nach Österreich zu kommen?
- Was ist im Urlaub wichtig?
- Aktivitäten im Urlaub?
- . . .

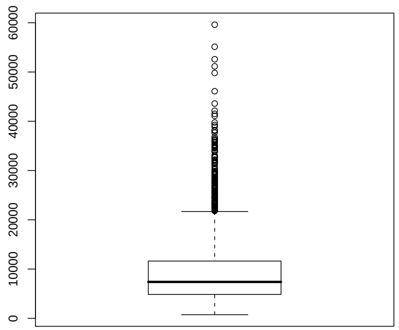
Beispiel GBÖ: Alter



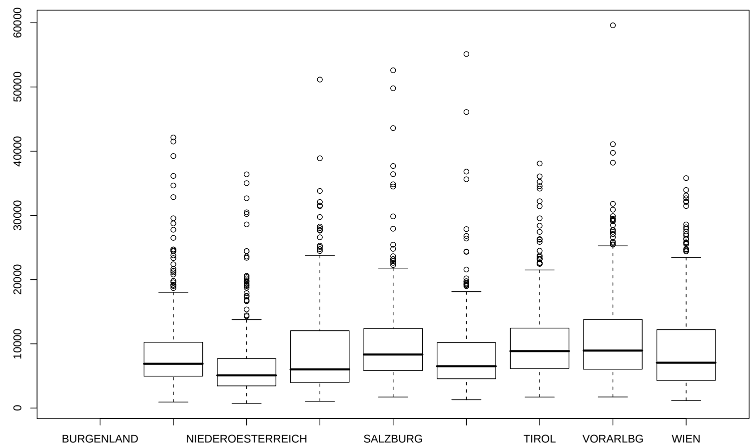


Beispiel GBÖ: Ausgaben

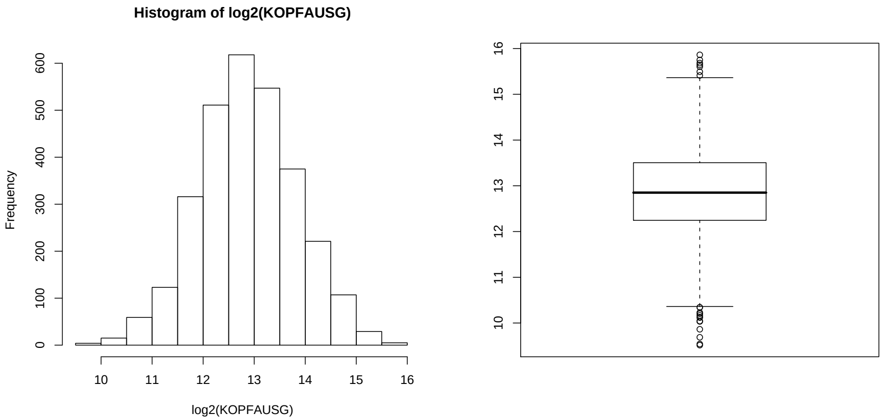




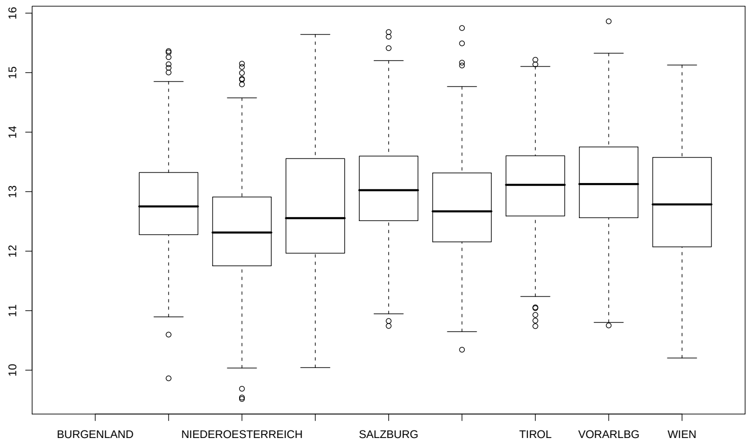
Beispiel GBÖ: Ausgaben



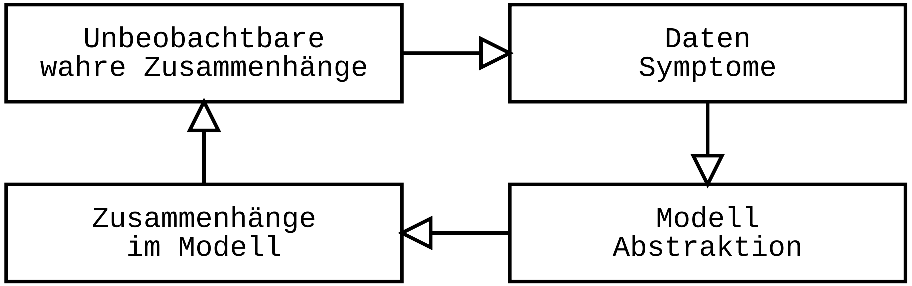
Beispiel GBÖ: log2(Ausgaben)



Beispiel GBÖ: log2(Ausgaben)



Modellbildung



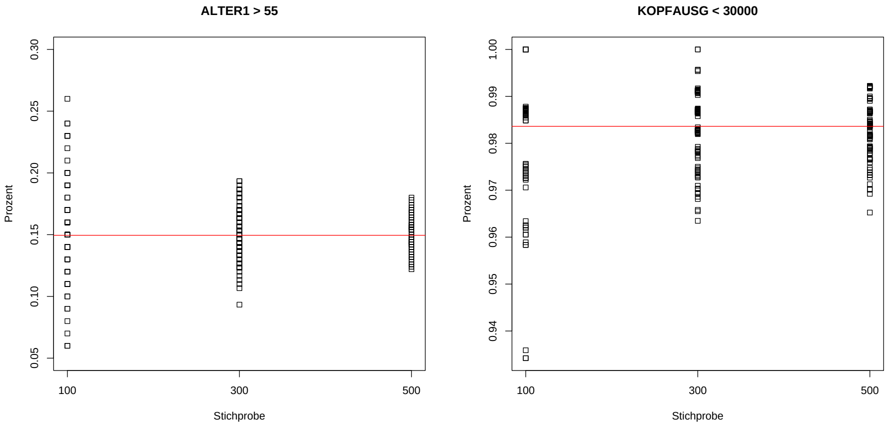
GBÖ: Fragen

1. Wie groß ist die Wahrscheinlichkeit, daß ein zufällig angetroffener (erwachsener) Tourist in Österreich älter als 55 Jahre ist?
   (gesamter Datensatz: 14.9%)

2. Wieviele Touristen gaben 1994 weniger als ATS 30.000,– pro Kopf aus?
   (gesamter Datensatz: 98.3%)

Variabilität der Antworten

Was passiert, wenn uns statt 3800 nur 100–500 Fragebögen zur Verfügung stehen?

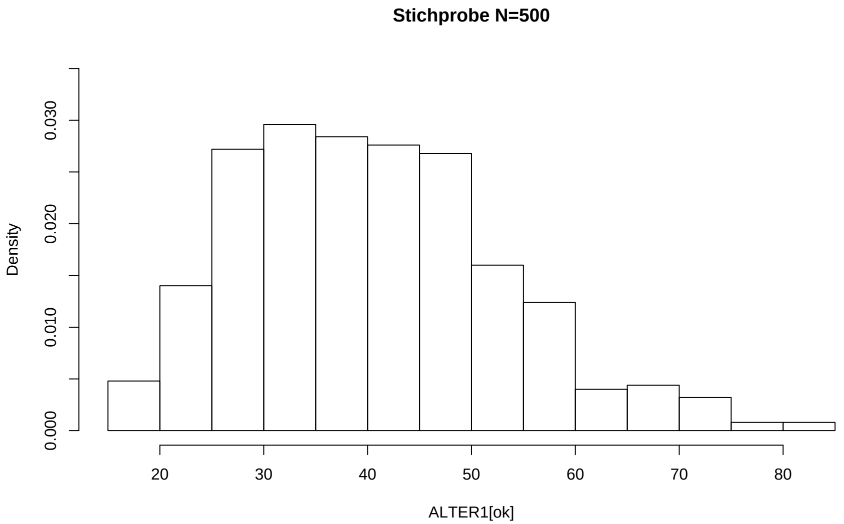


Modellanpassung

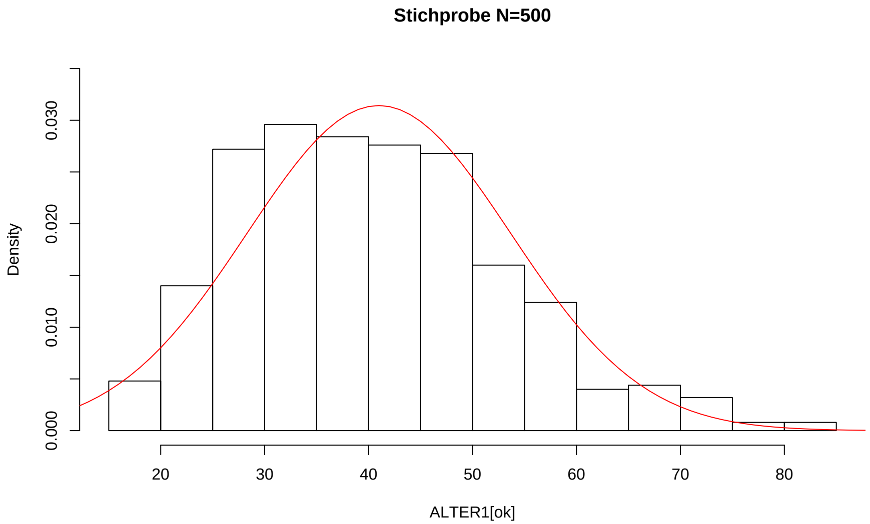
Alternative Lösungsmöglichkeit:

- Wir schätzen die Anzahl der Personen, die älter als 55 sind, nicht direkt, sondern versuchen ein Modell zu entwickeln.
- Der Einfachheit halber unterstellen wir für das Alter eine Normalverteilung (übliche Schätzer für Mittelwert und Standardabweichung).
- Für die Wahrscheinlichkeit, daß eine Person älter als 55 ist, benutzen wir die entsprechende Verteilungsfunktion.

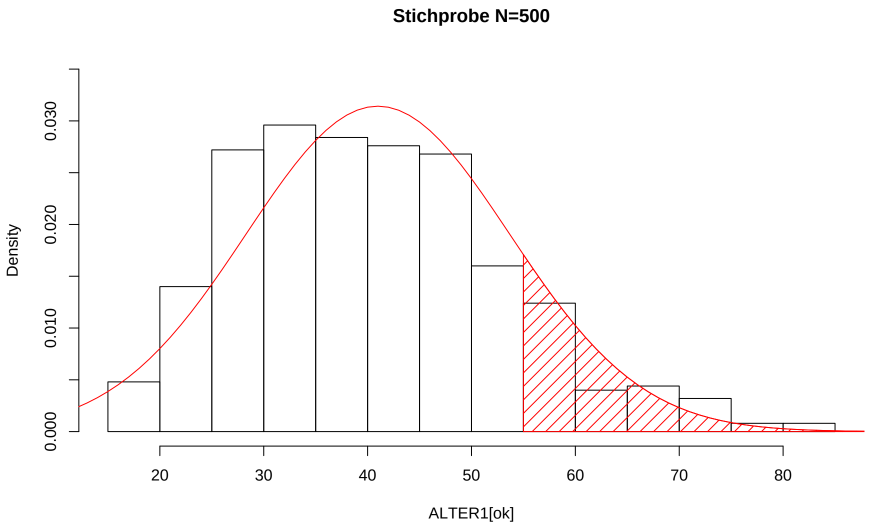
Modellanpassung



Modellanpassung

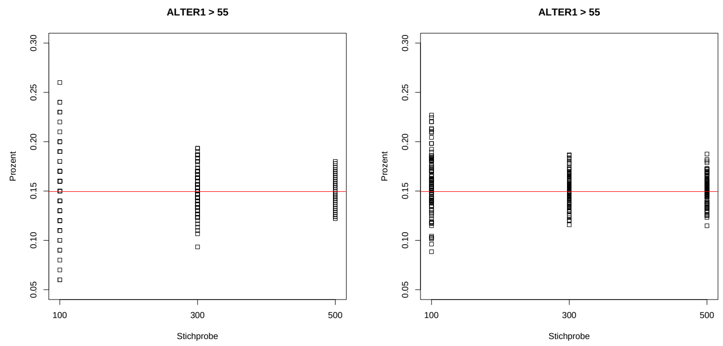


Modellanpassung



Vergleich der Variabilität

Vergleich Modell (rechts) mit direktem Schätzen der Prozentwerte (links): geringere Streuung der Schätzwerte!

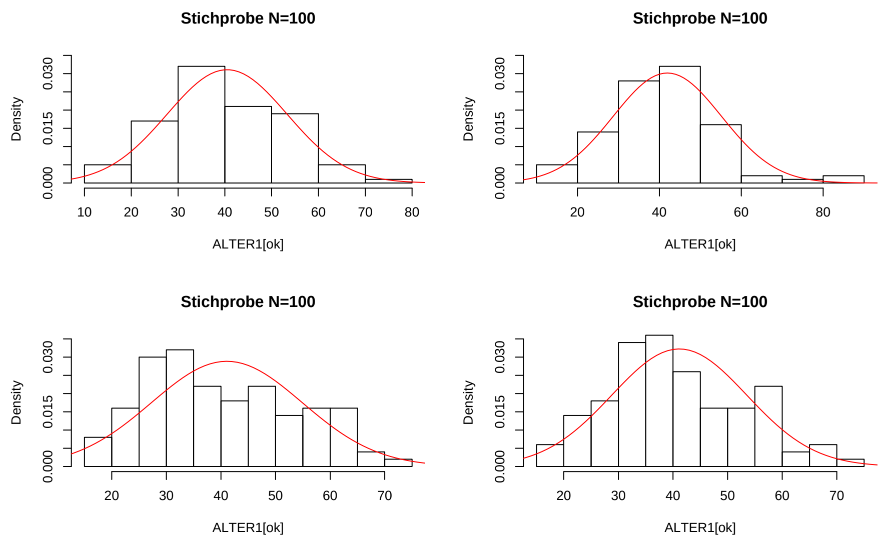


Warum Modell manchmal besser?

Warum funktioniert das Schätzen der Prozentwerte bei kleinen Stichproben über das Modell besser, obwohl das Alter offensichtlich nicht wirklich normalverteilt ist?

- Schätzung von Dichte und Quantilen am Rand einer Verteilung generell schlecht konditioniert (Abhängigkeit von wenigen Beobachtungen).
- Das Normalverteilungsmodell verdichtet die Stichprobe auf zwei Kennwerte (Mittel, Varianz), dabei werden Zufallschwankungen ausgeglichen.
- Die höhere Präzision beim Schätzen von Mittel und Varianz gleicht den Fehler der Modellierung mehr als aus.
- Kenntnis der zugrunde liegenden Struktur besser als voreilige Schlüsse aus Symptomen.

Warum Modell manchmal besser?



Modellanpassung

Um ein statistisches Modell an Daten anzupassen (derzeit einfach nur eine Verteilung an ein Merkmal einer Stichprobe), sind folgende Fragen zu beantworten:

1. Welches Modell, d.h., welche Verteilung?

2. Wie kann ich die Parameter des Modells schätzen?

3. Passen meine Daten zu dem Modell?

Je nach Datenlage kann sich die Reihenfolge ändern bzw. müssen die Fragen mehrmals iterativ durchlaufen werden.

Inferenzstatistik

Ziel der statistischen Modellierung ist in der Regel, Aussagen aus einer Stichprobe auf die Grundgesamtheit zu verallgemeinern, bzw. auf Eigenschaften der Grundgesamtheit zurückzuschließen.

Dies wird auch schließende Statistik, Inferenzstatistik oder induktive Statistik genannt.

Grundlage ist in den meisten Fällen das Treffen wahrscheinlichkeitstheoretischer Annahmen, wie die Stichprobe zustande gekommen ist.

Aus dem Bild der Grundgesamtheit (= Stichprobe), werden mögliche Urbilder (=Eigenschaften der Grundgesamtheit) rekonstruiert bzw. mit Wahrscheinlichkeiten quantifiziert.

Schätzen von Parametern

Schätzer für Parameter

X Merkmal bzw. Zufallsvariable
Parameter:

- Kennwerte einer (unbekannten) Verteilung, z.B.

$$E(X), Var(X), \text{ Median, } \rho(X,Y), \ldots$$

- (unbekannte) Parameter eines Verteilungstyps, z.B.

$$\lambda \text{ bei } Po(\lambda);\ \mu, \sigma^2 \text{ bei } N(\mu, \sigma^2);\ \pi \text{ bei } B(n, \pi), \ldots$$

$X_1, \ldots, X_n$ Stichprobenvariablen, hier: i.i.d. wie X
$x_1, \ldots, x_n$ Stichprobenwerte

Schätzer für Parameter

Generelle Notation	Beispiel
$X \sim F(x\|\theta)$	$\theta = \mu = E(X)$
θ unbekannter Parameter(-vektor)	$X \sim N(\mu, \sigma^2), \theta = (\mu, \sigma^2)$
$\theta \in \Theta$ Parameterraum	$\Theta = \mathbb{R}$ bzw. $\Theta = \mathbb{R} \times \mathbb{R}_+$

Gesucht: Schätzer bzw. Schätzwert für

$\theta: \quad \hat{\theta} \equiv t = g(x_1, \ldots, x_n) \qquad \mu: \quad \bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$

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

Schätzer für Parameter

Definition: Schätzer (Schätzfunktion, Schätzstatistik)

Zufallsvariable $T = g(X_1, \ldots, X_n)$

(Deterministische) Funktion der Stichprobenvariablen $X_1, \ldots, X_n$ heißt *Schätzer*.

Schätzwert $t = g(x_1, \ldots, x_n)$ ist Realisierung von T in der Stichprobe.

Beispiele für Schätzer/Schätzwerte

1. Arithmetisches Mittel

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) \quad \text{Schätzer für } \mu = E(X)$$
$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) \quad \text{Schätzwert für } \mu = E(X)$$

Beispiele für Schätzer/Schätzwerte

2. Spezialfall: X binär

$$P(X = 1) = \pi = E(X), \quad P(X = 0) = 1 - \pi$$
$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) = \frac{H}{n} \quad \text{für } \pi = E(X),$$

wobei H die absolute Häufigkeit von Einsen in der Stichprobe ist.

$$\bar{X} = \frac{H}{n} \quad \text{relative Häufigkeit}$$

ist Schätzer für $\pi = P(X = 1)$.

Beispiele für Schätzer/Schätzwerte

3. Stichprobenvarianz

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{für } \sigma^2 = Var(X)$$

4. Oder: Empirische Varianz

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

Fragen: Wie "gut" sind solche Schätzer ? Welcher Varianzschätzer ist „besser"?

Erwartungstreue

Beispiel: $\bar{X}$ Schätzer für $\mu = E(X)$

X Zufallsvariable mit $\mu = E(X)$; $X_1, \ldots, X_n$ i.i.d. wie X.

μ unbekannter, aber fester Wert

$\Rightarrow E(\bar{X}) = E(\frac{1}{n}(X_1 + \ldots + X_n)) = \frac{1}{n}(\underbrace{E(X_1)}_{\mu} + \ldots + \underbrace{E(X_n)}_{\mu}) = \mu$

Also: Unabhängig davon, welchen wahren (aber unbekannten) Wert μ tatsächlich besitzt, gilt

$$E(\bar{X}) = \mu\,.$$

D.h.: Der erwartete Wert von $\bar{X}$ ist μ. Damit:

Keine systematische "Verzerrung" beim Schätzen.

Erwartungstreue

Definition: Erwartungstreue und Verzerrung

- $T = g(X_1, \ldots, X_n)$ heißt *erwartungstreu (unverzerrt)* für θ:$\Leftrightarrow$

$$E(T) = \theta \qquad \text{für alle } \theta \in \Theta$$

- T heißt verzerrt:$\Leftrightarrow$ $\quad E(T) \neq \theta$

  $E(T) - \theta$ heißt *Verzerrung (Bias)*.

- $T = g(X_1, \ldots, X_n)$ heißt *asymptotisch erwartungstreu*:$\Leftrightarrow$

$$\lim_{n\to\infty} E(T) = \theta$$

Erwartungstreue

Beispiele:

- $E(\bar{X}) = \mu$, d.h. $\bar{X}$ für μ unverzerrt

- $\frac{H}{n}$ für π unverzerrt

- $E(\tilde{S}^2) = E(\frac{1}{n}\sum_i (X_i - \bar{X})^2) = \frac{n-1}{n}\sigma^2$

  $\Rightarrow \tilde{S}^2$ verzerrt

  $Bias(\tilde{S}^2) = E(\tilde{S}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$

  $\tilde{S}^2$ asymptotisch ewartungstreu, da Verzerrung $-\frac{\sigma^2}{n} \to 0$ für $n \to \infty$

- $E(S^2) = \sigma^2$, $\quad S^2$ unverzerrt

Erwartungstreue

Beweis Verzerrung von $\tilde{S}^2$:

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] \overset{\text{Linearität von } E}{\overset{\downarrow}{=}} \frac{1}{n}\sum_{i=1}^{n}\underbrace{E\left[(X_i-\mu)^2\right]}_{Var(X_i)=\sigma^2} = \frac{1}{n}\cdot n\sigma^2 = \sigma^2$$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2 \overset{\text{ausquadrieren}}{\overset{\downarrow}{=}} \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + \frac{2}{n}(\mu-\bar{X})\cdot\sum_{i=1}^{n}(X_i-\mu) + \frac{1}{n}n(\bar{X}-\mu)^2 = \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\bar{X}-\mu)\underset{\uparrow \sum X_i = n\bar{X}}{\cdot}(\bar{X}-\mu)\cdot n + (\bar{X}-\mu)^2 = \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\bar{X}-\mu)^2
\end{aligned}$$

Erwartungstreue

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] - \underbrace{E(\bar{X}-\mu)^2}_{Var(\bar{X})=\frac{\sigma^2}{n}} =$$

$$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

$$E(S^2) = E(\tfrac{n}{n-1}\tilde{S}^2) = \tfrac{n}{n-1}\cdot E(\tilde{S}^2) = \tfrac{n}{n-1}\cdot\tfrac{n-1}{n}\sigma^2 = \sigma^2$$

$\rightarrow$ S^2 unverzerrter Schätzer für σ^2

Varianz von Schätzern

Definition: Varianz und Standardabweichung eines Schätzers

$$\begin{aligned}
T &= g(X_1,\ldots,X_n) \text{ Schätzer} \\
Var(T) &= Var\{g(X_1,\ldots,X_n)\} \text{ Varianz von } T \\
\sigma_T &= +\sqrt{Var(T)} \text{ Standardabweichung von } T
\end{aligned}$$

Bemerkung:

Exakte analytische Formeln nur in einfachen Fällen angebbar; oft Approximation für großes n.

Beispiel: $\bar{X}$

$\sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{\sigma^2}{n}$, $\quad \sigma^2 = Var(X)$, $\quad \sigma^2$ aber unbekannt

Schätzer: $\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$

Varianz von Schätzern

Definition: Erwartete quadratische Abweichung, Mean Square Error

$$MSE(T) = E\{(T-\theta)^2\} = Var(T) + (Bias(T)^2)$$

Bemerkung:

Der Mean Square Error $MSE(T)$ fasst als Erwartungswert der quadratischen Abweichung $(T-\theta)^2$ des Schätzers T vom zu schätzenden Parameter θ die Varianz und die quadrierte Verzerrung in einem gemeinsamen Gütekriterium für T zusammen.

Varianz von Schätzern

Definition: Konsistenz

- T heißt (MSE-)konsistent für $\theta :\Leftrightarrow MSE(T) \to 0$ für $n \to \infty$

- T heißt (schwach) konsistenz für $\theta :\Leftrightarrow P(|T - \theta| < \epsilon) \to 1 \ \forall \epsilon > 0$ und für $n \to \infty$

Bemerkungen:

- Damit $MSE(T) = Var(T) + (Bias(T))^2 \to 0$ geht, muss $Var(T) \to 0$ und $Bias(T) \to 0$ gelten.

- Aus MSE-Konsistenz folgt schwache Konsistenz.

Effiziente Schätzstatistiken

$MSE(T)$ Maß für Güte von T

$Var(T)$ Maß für Varianz von T

$\Rightarrow$ Man kann zwei Schätzer T_1, T_2 bzgl. MSE (oder auch Var) vergleichen.

Definition: T_1 (MSE-)effizienter als $T_2 :\Leftrightarrow$

$$MSE(T_1) \le MSE(T_2)$$

Bei erwartungstreuen Schätzern T_1, T_2:

$$Bias(T_1) = Bias(T_2) = 0 \quad \Rightarrow \quad MSE(T_i) = Var(T_i),\ i = 1, 2$$
$$\Rightarrow T_1 \text{ effizienter als } T_2 \quad \Leftrightarrow \quad Var(T_1) \le Var(T_2)$$

Effiziente Schätzstatistiken

Definition:

Ist ein Schätzer T besser als alle zur “Konkurrenz” zugelassenen anderen Schätzer $\tilde{T}$, so heißt T (MSE-)effizient für θ.

Beispiele:

- $\bar{X}$ für μ, unter allen erwartungstreuen Schätzern für μ.

- $\bar{X}$ für μ, falls X normalverteilt ist;
  alternative Schätzer dürfen dann auch verzerrt sein!

Parameterschätzung

Momentenschätzer

Die einfachste Form der Parameterschätzung ist die sogenannte Momentenmethode:

1. Moment: Erwartungswert $\mathbb{E}X$
2. Moment: Varianz $\mathbb{E}(X - \mathbb{E}X)^2$
3. Moment: Schiefe $\mathbb{E}(X - \mathbb{E}X)^3$
k. Moment: $\mathbb{E}(X - \mathbb{E}X)^k$

Viele Verteilungen haben Momente als Parameter, diese können aus der Stichprobe mittels

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{bzw.} \qquad \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

geschätzt werden.

Falls erwartungstreue Schätzer gewünscht werden, muß die Normierung für das 2. und alle höheren Momente mit $1/(n-1)$ erfolgen.

Maximum Likelihood

Momentenschätzer funktionieren nicht, falls die Parameter der Verteilung nicht (einfach oder sinnvoll) als Funktion der Momente dargestellt werden können.

Bsp: Gleichverteilung

Ein allgemeines Prinzip zur Konstruktion von Schätzfunktionen ist die Maximum-Likelihood-Methode (kurz ML-Schätzung): Wir wählen die Parameter, bei denen die gesehene Stichprobe maximale Wahrscheinlichkeit hat.

Momentenschätzer werden meistens nur dann benutzt, wenn sie gleichzeitig auch ML-Schätzer sind.

ML für diskretes X

Sei $X_1 = x_1, \ldots, X_n = x_n$ die konkrete Stichprobe.
Gesucht: Schätzwert $\hat{\theta}$ (bzw. T) für θ
Konzept: Bestimme/konstruiere $\hat{\theta}$ so, dass die Wahrscheinlichkeit für Auftreten der Stichprobe maximal wird, d.h. $\hat{\theta}$ so, dass

$$P(X_1 = x_1, \ldots, X_n = x_n|\theta) \to \max_{\theta}.$$

Es ist

$$\begin{aligned} &P(X_1 = x_1, \ldots, X_n = x_n|\theta) = \\ \underset{\substack{\uparrow \\ X_1,\ldots,X_n \text{ i.i.d.}}}{=} &P(X_1 = x_1|\theta) \cdot \ldots \cdot P(X_n = x_n|\theta) = f(x_1|\theta) \cdot \ldots \cdot f(x_n|\theta). \end{aligned}$$

wobei $f(x|\theta)$ wie üblich die Wahrscheinlichkeitsfunktion der diskreten Verteilung bezeichnet.

ML für diskretes X

Definition: Likelihoodfunktion

Bei gegebenen $x_1, \ldots, x_n$ heißt

$$L(\theta) = f(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdot \ldots \cdot f(x_n|\theta)$$

Likelihoodfunktion für θ.

Definition: Likelihood-Prinzip/Maximum-Likelihood-Schätzung

Bestimme $\hat{\theta}$ so, dass

$$L(\hat{\theta}) = \max_{\theta} L(\theta).$$

Bsp: Poisson-Verteilung

$X_1, \ldots, X_4$ i.i.d. $Po(\lambda)$ mit Realisierungen $x_1 = 2, x_2 = 4, x_3 = 6, x_4 = 3$.

$\Rightarrow$ Likelihoodfunktion

$$L(\lambda) = f(x_1|\lambda) \cdots f(x_4|\lambda) = e^{-\lambda}\frac{\lambda^2}{2!}e^{-\lambda}\frac{\lambda^4}{4!}e^{-\lambda}\frac{\lambda^6}{6!}e^{-\lambda}\frac{\lambda^3}{3!}$$
$$= e^{-4\lambda}\lambda^{15}\frac{1}{2!\,4!\,6!\,3!}$$

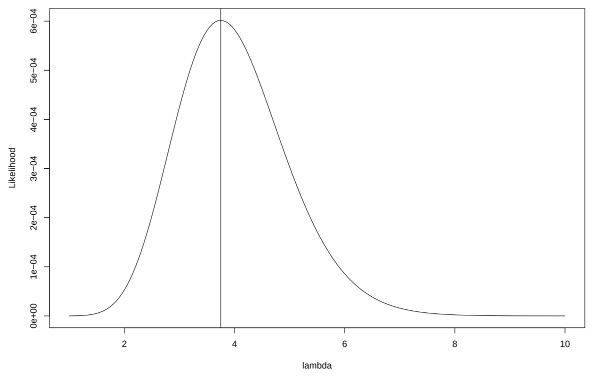
$\Rightarrow$ Log-Likelihoodfunktion

$$\log L(\lambda) = -4\lambda + 15\log\lambda - \log(2!\,4!\,6!\,3!).$$

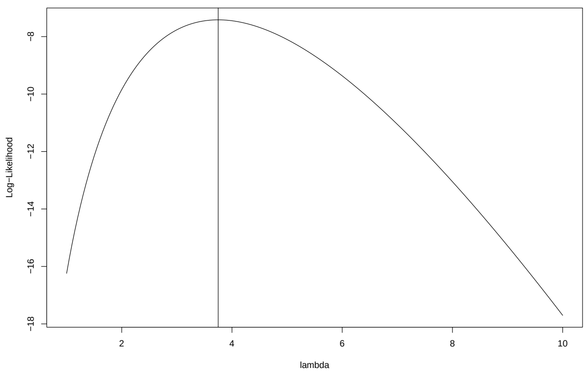
Ableiten und Nullsetzen

$$\Rightarrow \quad \frac{\partial \log L(\lambda)}{\partial\lambda} = -4 + \frac{15}{\hat\lambda} = 0 \Leftrightarrow \hat\lambda = \frac{15}{4}$$

Bsp: Poisson-Verteilung



Bsp: Poisson-Verteilung



ML für stetiges X

Für stetige Zufallsvariablen X mit Dichte $f(x|\theta)$ überträgt man das Konzept in völliger Analogie:
Wähle θ so, dass die Likelihood maximal wird: $L(\theta) = f(x_1|\theta) \cdot \ldots \cdot f(x_n|\theta)$

$$L(\theta) = f(x_1|\theta) \cdot \ldots \cdot f(x_n|\theta) \quad \to \quad \max_\theta$$

Motivation: Versuchen Werte der Dichte „gemeinsam“ groß zu machen.

Oft ist $\hat\theta$ eine (komplizierte, nichtlineare) Funktion von $x_1, \ldots, x_n$:

$$\hat\theta = g(x_1, \ldots, x_n).$$

Setzt man statt der Realisierungen $x_1, \ldots, x_n$ die Stichprobenvariablen $X_1, \ldots, X_n$ ein, wird $T \equiv \hat\theta$ zum Maximum-Likelihood-Schätzer.

ML für stetiges X

Konkrete Berechnung erfolgt meist durch Maximieren der log-Likelihood

$$\begin{aligned}\log L(\theta) &= \log f(x_1|\theta) + \ldots + \log f(x_n|\theta) = \\ &= \sum_{i=1}^{n} \log f(x_i|\theta)\end{aligned}$$

Maxima $\hat{\theta}$ von $L(\theta)$ und $\log L(\theta)$ sind identisch, da log eine streng monotone Transformation ist.

Das Maximum wird i.a. durch Nullsetzen der ersten Ableitung berechnet.

Bsp: Normalverteilung

$X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ mit Realisierungen $x_1, \ldots, x_n$

$$\Rightarrow \quad L(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned}\log L(\mu, \sigma) &= \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &= \sum_{i=1}^{n} \left[-\log\sqrt{2\pi} - \log\sigma - \frac{(x_i-\mu)^2}{2\sigma^2}\right]\end{aligned}$$

Bsp: Normalverteilung

$$\begin{aligned}\Rightarrow \quad & \frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \hat{\mu}}{\hat{\sigma}^2} = 0 \\ & \frac{\partial \log L(\mu, \sigma)}{\partial \sigma} = \sum_{i=1}^{n} \left(-\frac{1}{\hat{\sigma}} + \frac{2(x_i-\hat{\mu})^2}{2\hat{\sigma}^3}\right) = 0 \\ \Rightarrow \quad & \hat{\mu} = \bar{x}, \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\end{aligned}$$

Bayes-Schätzung

Basiert auf subjektivem Wahrscheinlichkeitsbegriff; dennoch enge Verbindung zur Likelihood-Schätzung. Besonders für hochdimensionale, komplexe Modelle geeignet; "Revival" etwa seit 1990.

"Subjektives" Grundverständnis:

- θ wird als Realisierung einer Zufallsvariablen Θ aufgefasst
- Unsicherheit/Unkenntnis über θ wird durch eine priori-Verteilung (stetige oder diskrete Dichte)

  $$f(\theta)$$

  bewertet. Meist: Θ stetige Zufallsvariable; $f(\theta)$ stetige Dichte.

Die Bayes-Inferenz beruht auf der posteriori-Verteilung von Θ, gegeben die Daten $x_1, \ldots, x_n$. Dazu benötigen wir den Satz von Bayes für Dichten.

Bayes-Schätzung

Notation

$f(x, \theta)$	gemeinsame Wahrscheinlichkeitsfunktion oder Dichte
$f(x)$	Randverteilung oder -dichte von X
$f(x \mid \theta)$	bedingte Wahrscheinlichkeitsfunktion bzw. Dichte von X, gegeben $\Theta = \theta$
$f(\theta)$	a priori Wahrscheinlichkeitsfunktion oder a priori Dichte von Θ (d.h. die Randverteilung von Θ)
$f(\theta \mid x)$	a posteriori (oder bedingte) Wahrscheinlichkeitsfunktion oder Dichte von Θ, gegeben die Beobachtung $X = x$

Bayes-Schätzung

Dann gilt folgende Form des **Satzes von Bayes**:

$$f(\theta \mid x) = \frac{f(x, \theta)}{f(x)} = \frac{f(x \mid \theta) f(\theta)}{f(x)}.$$

Θ und X diskret:

$$\Rightarrow \quad f(x) = P(X = x) = \sum_j f(x \mid \theta_j) f(\theta_j),$$

wobei über die möglichen Werte θ_j von Θ summiert wird.

Θ stetig:

$$\Rightarrow \quad f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)} = \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta)\, d\theta}$$

Dabei kann X stetig oder diskret sein.

Bayes-Schätzung

Für Stichprobe $x = (x_1 \ldots, x_n)$ aus $f(x_1, \ldots, x_n \mid \theta)$:

$$f(x) \to f(x_1, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \cdot \ldots \cdot f(x_n \mid \theta) = L(\theta)$$

$\Rightarrow$ **Bayes-Inferenz, Bayesianisches Lernen**:

Die Wahrscheinlichkeitsfunktion oder Dichte von X, gegeben θ, sei

$$f(x \mid \theta)$$

und

$$L(\theta) = f(x_1, \ldots, x_n \mid \theta)$$

die gemeinsame Dichte bzw. Likelihoodfunktion für n unabhängige Wiederholungen von X.

Bayes-Schätzung

Für den unbekannten Parameter wird eine a priori Dichte

$$f(\theta)$$

spezifiziert.

Dann ist die a posteriori Dichte über den Satz von Bayes bestimmt durch

$$f(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1 \mid \theta) \cdots f(x_n \mid \theta) f(\theta)}{\int f(x_1 \mid \theta) \cdots f(x_n \mid \theta) f(\theta)\, d\theta} =$$

$$= \frac{L(\theta) f(\theta)}{\int L(\theta) f(\theta)\, d\theta}.$$

Hinweis: Nenner hängt nicht von θ ab.

Bayes-Schätzung

Bayes-Schätzer

a posteriori Erwartungswert:

$$\hat{\theta}_p = E(\theta \mid x_1, \ldots, x_n) = \int \theta f(\theta \mid x_1, \ldots, x_n)\, d\theta$$

a posteriori Modus oder *maximum a posteriori (MAP)* Schätzer:
Wähle denjenigen Parameterwert $\hat{\theta}_{\mathsf{MAP}}$, für den die a posteriori Dichte maximal wird, d.h.

$$L(\hat{\theta}) f(\hat{\theta}) = \max_{\theta} L(\theta) f(\theta)$$

bzw.

$$\log L(\hat{\theta}) + \log f(\hat{\theta}) = \max_{\theta} \{\log L(\theta) + \log f(\theta)\}\,.$$

Andere Möglichkeiten: Median, ...