Vorlesung 13b

Maximum-Likelikood-Schätzung

1. Goldene Idee der Statistik:

Daten aufgefasst als

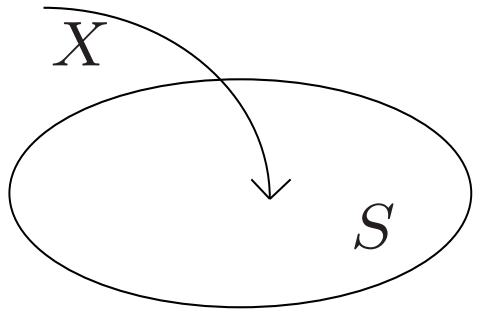
Realisierungen von Zufallsvariablen

Statistisches Modell:

eine Zufallsvariable X, bei deren Verteilung

ein Parameter ϑ frei bleibt:

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$



z.B. $\vartheta = (\mu, \sigma^2)$ bei der Normalverteilung,

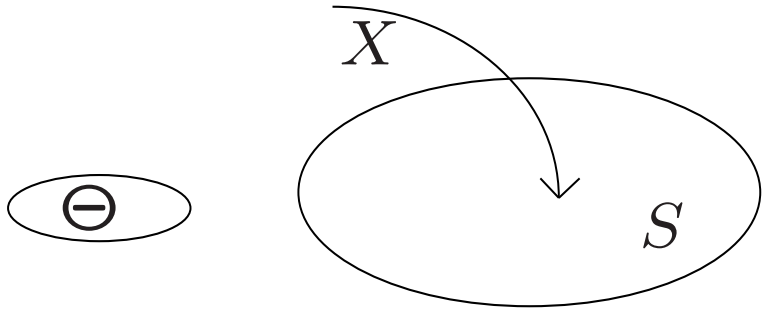
oder $\vartheta = p$ beim Münzwurf.

Statistisches Modell:

eine Zufallsvariable X, bei deren Verteilung

ein Parameter ϑ frei bleibt:

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$



Θ . . . *Menge der Parameter*

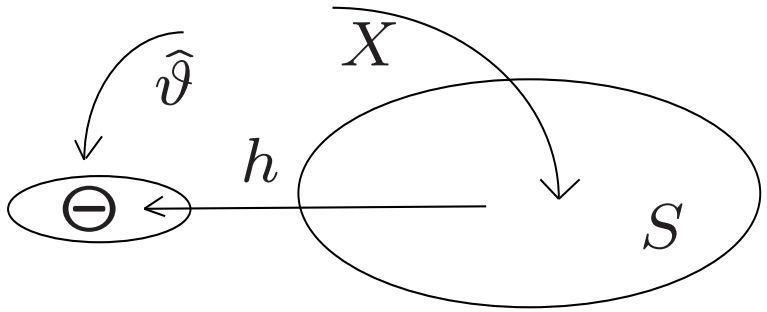
S . . . *Beobachtungsraum*

Statistisches Modell:

eine Zufallsvariable X, bei deren Verteilung

ein Parameter ϑ frei bleibt:

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$



Der Parameter ϑ soll aus den Daten geschätzt werden.

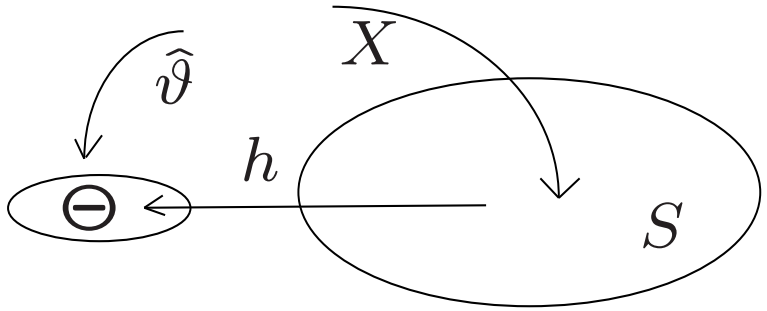
Dazu verarbeitet man X zu einem *Schätzer* $\widehat{\vartheta}$ für ϑ.

Statistisches Modell:

eine Zufallsvariable X, bei deren Verteilung

ein Parameter ϑ frei bleibt:

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \ominus$$

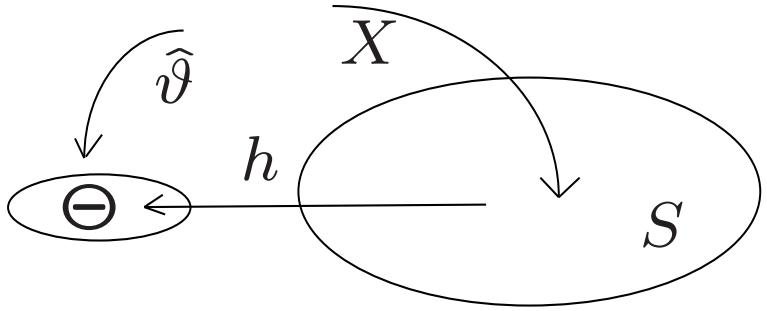


$\ominus$ ... *Menge der Parameter*

S ... *Beobachtungsraum*

$\widehat{\vartheta} := h(X)$ ... *Schätzer* für den Parameter ϑ

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$



Θ ... *Menge der Parameter*

S ... *Beobachtungsraum*

$\widehat{\vartheta} := h(X)$ ... *Schätzer* für den Parameter ϑ

Naheliegende (“naive”) Schätzer:

für p aus einem n-fachen p-Münzwurf $(X_1, \ldots, X_n)$:

$$\widehat{p} = \tfrac{1}{n}(X_1 + \cdots + X_n).$$

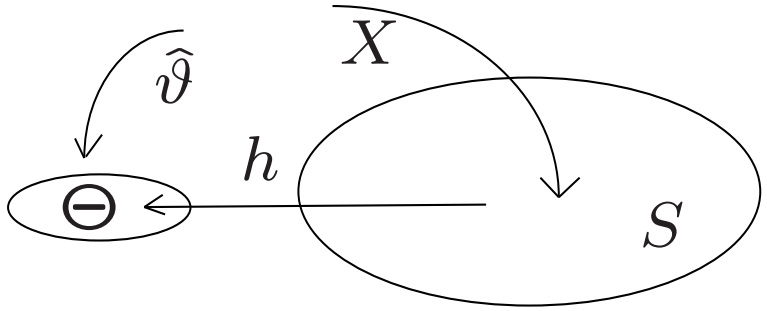
Und allgemeiner:

für μ aus unabhängigen reellwertigen Zufallsvariablen $(X_1, \ldots, X_n)$ mit $\mathbf{E}[X_i] = \mu$:

$$\widehat{\mu} = \tfrac{1}{n}(X_1 + \cdots + X_n).$$

2. Das Maximum-Likelihood-Prinzip

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$

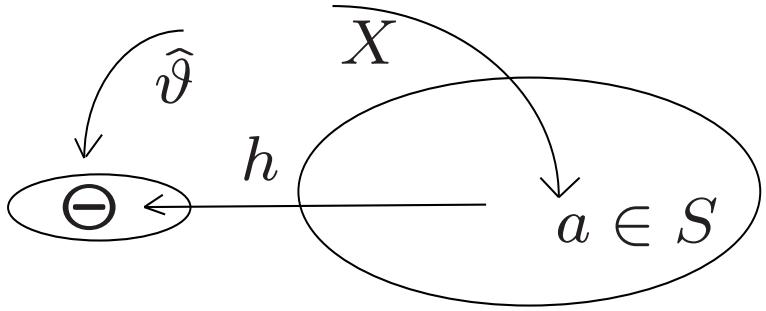


Ein tragfähiges Prinzip zur Wahl der Abbidung h:

Sei $h(a)$ dasjenige ϑ, für das die Wahrscheinlichkeit,

den Ausgang a zu erhalten, maximal wird.

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$



Ein tragfähiges Prinzip zur Wahl der Abbidung h:

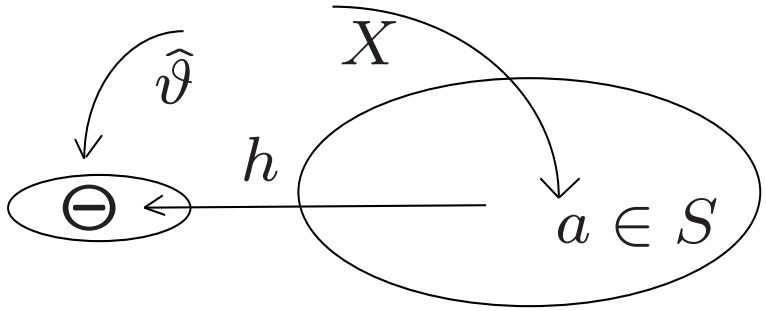
Für diskretes X: Wähle $h(a)$ so, dass

$$\mathbf{P}_{h(a)}(X = a) = \max_{\vartheta \in \Theta} \mathbf{P}_\vartheta(X = a).$$

Die Zufallsvariable $\widehat{\vartheta} := h(X)$ nennt man dann

Maximum-Likelikood-Schätzer für ϑ auf der Basis von X:

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$



Für eine Familie mit Dichten, $\rho_\vartheta(da) = f_\vartheta(a)\, da, \quad \vartheta \in \Theta$,

geht man analog vor: Wähle $h(a)$ so, dass

$$f_{h(a)}(a) = \max_{\vartheta \in \Theta} f_\vartheta(a).$$

Die Zufallsvariable $\widehat{\vartheta} := h(X)$ nennt man dann

Maximum-Likelikood-Schätzer für ϑ auf der Basis von X.

3. Beispiel: Münzwurf.

$(X_1, \ldots, X_n)$ sei n-facher p-Münzwurf mit unbekanntem p,

$K_n := X_1 + \cdots + X_n$ die Anzahl der Erfolge.

Beobachtet wird die Realisierung $(a_1, \ldots, a_n)$

mit $k = a_1 + \ldots + a_n$.

Behauptung: Unter allen p ist k/n derjenige Parameter,

für den $\mathbf{P}_p(X_1 = a_1, \ldots, X_n = a_n) = p^k(1-p)^{n-k}$

maximal ist.

Wir betrachten zuerst die Fälle $k = n$ und $k = 0$.

Hierfür stimmt die Behauptung wegen

$$\mathbf{P}_1(K_n = n) = \mathbf{P}_0(K_n = 0) = 1.$$

Behauptung: Unter allen p ist k/n derjenige Parameter, für den $\mathbf{P}_p(X_1 = a_1, \ldots, X_n = a_n) = p^k(1-p)^{n-k}$ maximal ist.

Denn: Der Logarithmus der rechten Seite ist

$$k \ln p + (n-k) \ln(1-p).$$

Die Fälle $k \in \{0, 1\}$ hatten wir schon betrachtet.

Für $0 < k < n$ hat dieser Ausdruck sein Maximum in $p = k/n$, wie man durch Differenzieren feststellt.

Also ist

$$\frac{1}{n}(X_1 + \cdots + X_n)$$

der Maximum-Likelihood-Schätzer für p.

Für $k = n$ (kein Misserfolg in n Versuchen) ergibt sich 1 als Maximum-Likelikood-Schätzung von p.

Das ist möglicherweise zu optimistisch.

Eine Alternative bietet der sogenannte *Bayes-Schätzer* (vgl Buch S. 127).

Hier denkt man an ein zweistufiges Experiment:

1. eine auf $[0, 1]$ uniform verteilte Zufallsvariable U
2. gegeben $\{U = u\}$ einen Münzwurf mit Erfolgswahrscheinlichkeit U.

$$\tilde{p} := \mathbf{E}[U|K] \ .$$

Erinnerung an Vorlesung 13a:

$Z_1, Z_2, \ldots$ sei ein Münzwurf mit uniform auf $[0, 1]$ verteiltem zufälligem Erfolgsparameter U,

K_n sei die Anzahl der Erfolge in den ersten n Versuchen.

Dann ist

$$\mathbf{E}[U \mid K_n = k] = \frac{k+1}{n+2}.$$

Man nennt dies auch den

Bayes-Schätzer für die Erfolgswahrscheinlichkeit

(bei a priori uniform verteilter Erfolgswahrscheinlichkeit).

Eine Beziehhung zur Pólya-Urne:

In Vorlesung 13a haben wir gesehen (vgl. Buch S. 113/114):

Ein Münzwurf $(Z_1, Z_2, \ldots)$

mit uniform verteilter Erfolgswahrscheinlichkeit U

ist so verteilt wie die Folge der Zuwächse in Richtung Osten

in einer Nordost-Wanderung à la Pólya.

Also:

$$\mathbf{E}[U|K_n = k] = \mathbf{P}[Z_{n+1} = 1|K_n = k] = \frac{k+1}{n+2}.$$

4. Beispiel:

Unabhängige, identisch normalverteilte Zufallsvariable.

$X_1, \ldots, X_n$ seien unabhängig und

N(μ, σ^2)-verteilt, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$.

Behauptung:

Der ML-Schätzer für $\vartheta = (\mu, \sigma)$ ist dann $(\widehat{\mu}, \widehat{\sigma})$

mit

$$\widehat{\mu} := \tfrac{1}{n}(X_1 + \cdots + X_n),$$

$$\widehat{\sigma}^2 := \tfrac{1}{n}\big((X_1 - \widehat{\mu})^2 + \cdots + (X_n - \widehat{\mu})^2\big) .$$

Denn: Die gemeinsame Dichtefunktion ist

$$\varphi_{\mu,\sigma^2}(a_1) \cdots \varphi_{\mu,\sigma^2}(a_n)$$

mit $\varphi_{\mu,\sigma^2}(x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}$, $x \in \mathbb{R}$.

Damit ist die Aufgabe: Finde für gegebenes $(a_1, \ldots, a_n)$

das maximierende (μ, σ^2).

Im ersten Schritt betrachten wir die Abbildung

$\mu \mapsto \ln \varphi_{\mu,\sigma^2}(a_1) + \cdots + \ln \varphi_{\mu,\sigma^2}(a_n)$ für festes σ^2.

Diese wird maximiert bei $m = \frac{1}{n}(a_1 + \cdots + a_n)$ (warum?)

Denn: Die gemeinsame Dichtefunktion ist

$$\varphi_{\mu,\sigma^2}(a_1) \cdots \varphi_{\mu,\sigma^2}(a_n)$$

mit $\varphi_{\mu,\sigma^2}(x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}$, $x \in \mathbb{R}$.

Damit ist die Aufgabe: Finde für gegebenes $(a_1, \ldots, a_n)$
das maximierende (μ, σ^2).

Im zweiten Schritt differenzieren wir die Abbildung

$$\sigma \mapsto \ln \varphi_{m,\sigma^2}(a_1) + \cdots + \ln \varphi_{m,\sigma^2}(a_n)$$

und bekommen die Maximalstelle

$$\tfrac{1}{n}\big((a_1 - m)^2 + \cdots + (a_n - m)^2\big). \quad \square$$

Aus Übungsaufgabe 29 wissen wir:

$$\mathbf{E}_{(\mu,\sigma^2)}\left[\widehat{\sigma}^2\right] = \tfrac{n-1}{n}\sigma^2.$$

Der Schätzer $\widehat{\sigma}^2$ ist also nicht "erwartungstreu",

wohl aber seine Modifikation

$$s^2 := \tfrac{n}{n-1}\widehat{\sigma}^2 = \tfrac{1}{n-1}\big((X_1 - \widehat{\mu})^2 + \cdots + (X_n - \widehat{\mu})^2\big).$$

Und es gilt sogar (siehe Buch S. 138) der Satz von Gosset-Fisher:
$\widehat{\mu}$ und s^2 sind unabhängig, und $\frac{s^2}{\sigma^2}$ ist so verteilt wie die Summe aus $n-1$ Quadraten von unabhängigen standard-normalverteilten Z_i.

5. Beispiel:

Einfache lineare Regression.

$x_1, \ldots, x_n$ seien feste reelle Zahlen,

$$Y_i = \beta_0 + \beta_1 x_i + \sigma Z_i, \quad i = 1, \ldots, n,$$

mit $Z_1, \ldots, Z_n$ unabhängig, N$(0, 1)$-verteilt.

Die Dichtefunktion von $Y = (Y_1, \ldots, Y_n)$ hat

am Ausgang $a = (a_1, \ldots, a_n)$ den Wert

$$(*) \quad \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{|a-\mu|^2}{2\sigma^2}\Big),$$

mit $\mu_i := \beta_0 + \beta_1 x_i, \quad \mu := (\mu_1, \ldots, \mu_n)$.

Wieder maximieren wir zuerst bei festgehaltenem σ.

Damit ergeben sich die uns aus Vorlesung 8b (letzte Folie)

wohlbekannten Koeffizienten der Regressionsgeraden:

$$b_1 = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad b_0 = \bar{a} - b_1\bar{x}\ .$$

Betrachten wir jetzt $(*)$ als Funktion von σ,

mit $\widehat{\mu}_i := b_0 + b_1 x_i$ statt μ_i,

so finden wir durch Logarithmieren und Differenzieren

dessen Maximalstelle bei

$$\frac{1}{n}\sum_{i=1}^{n}(a_i - \widehat{\mu}_i)^2.$$

Der ML-Schätzer für $(\beta_0, \beta_1, \sigma^2)$ ergibt sich durch Einsetzen von Y anstelle von a:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \widehat{\beta}_0 = \bar{Y} - \bar{\beta}_1 \bar{x} \ ,$$

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2.$$

Ein erwartungstreuer Schätzer mit schönen Verteilungseigenschaften (siehe Buch S. 138) ist

$$s^2 := \frac{n}{n-2} \widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right)^2.$$

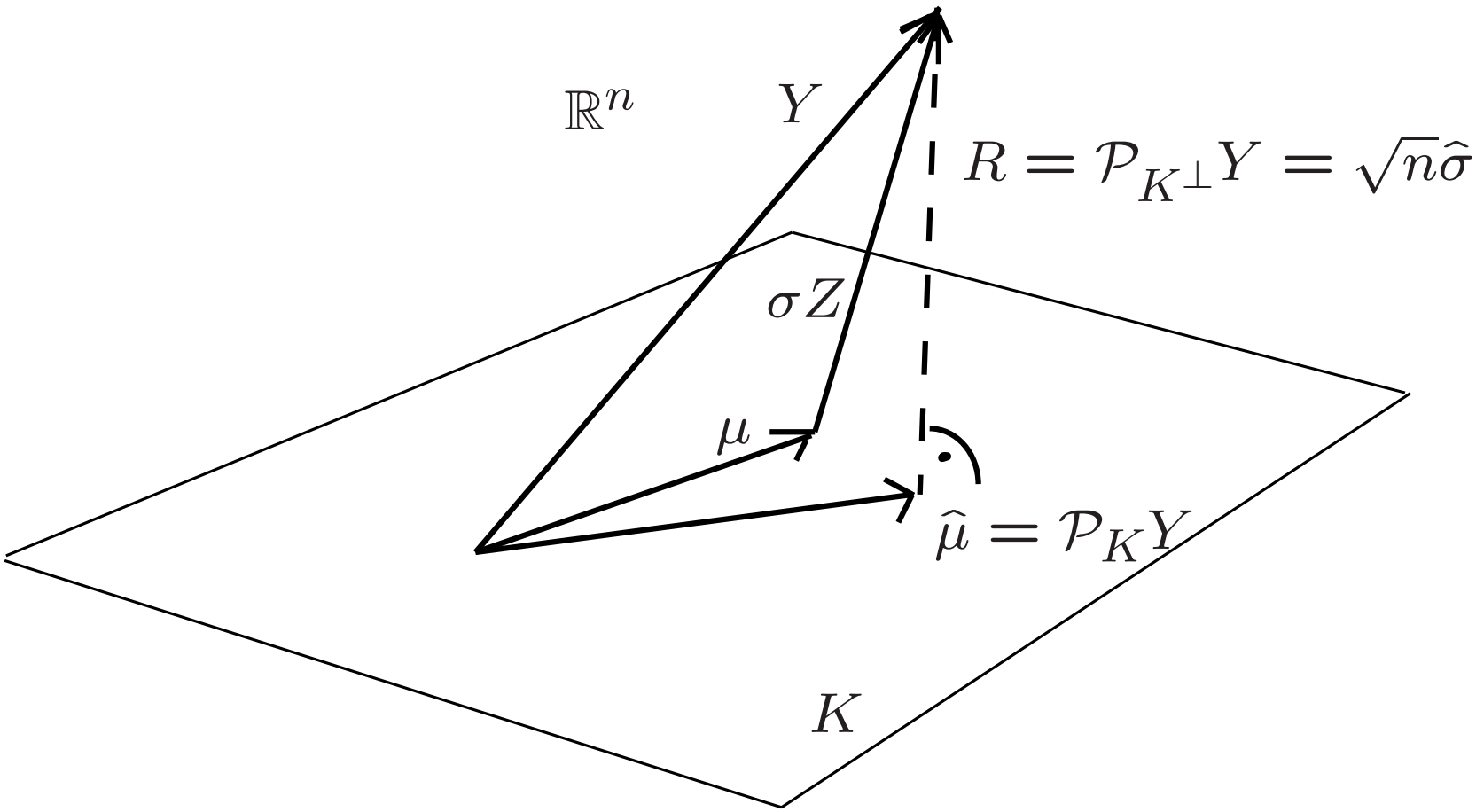
Für alle, die Geometrie mögen:

Der n-dimensionale Datenvektor a soll dargestellt werden durch einen 2-dimensionalen "systematischen Beitrag" plus ein (möglichst kleines) Residuum ("Rest"):

$$a = b_0\mathbf{1} + b_1 x + r, \quad \text{mit } |r|^2 = \min!.$$

Dabei ist $\mathbf{1} = (1, \ldots, 1)$ und $x := (x_1, \ldots, x_n)$.

Im folgenden Bild (vgl Buch S. 136) ist K der von $\mathbf{1}$ und x aufgespannte 2-dimensionale Teilraum des Datenraumes $\mathbb{R}^n$.


$\mathbb{R}^n$
Y
$R = \mathcal{P}_{K^\perp} Y = \sqrt{n}\hat{\sigma}$
σZ
μ
$\hat{\mu} = \mathcal{P}_K Y$
K

6. Beispiel:

Uniforme Verteilung

$X_1, \ldots, X_n$ seien unabhängig und uniform verteilt auf $[0, \vartheta]$.

Was ist der ML-Schätzer für ϑ?

Die Dichtefunktion von $(X_1, \ldots, X_n)$ ist

$$f_\vartheta(a_1, \ldots, a_n) = \frac{1}{\vartheta^n}, \qquad 0 \leq \min(a_i) \leq \max(a_i) \leq \vartheta.$$

Für festes $(a_1, \ldots, a_n)$ wird sie maximiert bei $\vartheta = \max(a_i)$.

Also ist

$$\widehat{\vartheta} = \max\{X_i : i = 1, \ldots, n\}.$$

Unter $\mathbf{P}_\vartheta$ ist $\max\{X_i : i = 1, \ldots, n\}$ so verteilt wie

$$\vartheta \max\{U_i : i = 1, \ldots, n\},$$

mit $U_1, \ldots, U_n$ unabhängig und Unif([0, 1])-verteilt.

Also ist (vgl. Vorlesung 13a, Folie 24)

$$\mathbf{E}_\vartheta[\widehat{\vartheta}] = \mathbf{E}_\vartheta[\max\{X_i : i = 1, \ldots, n\}] = \tfrac{n}{n+1}\vartheta.$$

7. Beispiel:
Zweiseitige Exponentialverteilung.*

*nachgetragen in der Vorlesung am 05.01.2019

Für $\vartheta \in \mathbb{R}$ seien $X_1, \ldots, X_n$

unabhängig und identisch verteilt mit Dichte

$$g_\vartheta(x) := \tfrac{1}{2} e^{-|x-\vartheta|},\ x \in \mathbb{R}.$$

Was ist der ML-Schätzer für ϑ?

$$f_\vartheta(a_1,\ldots,a_n) := \left(\tfrac{1}{2}\right)^n \prod_{i=1}^{n} e^{-|a_i-\vartheta|},\ a_i \in \mathbb{R}$$

Betrachten wir erst einmal den Fall $n = 2$,

und für festes $a_1, a_2 \in \mathbb{R}$ die Funktion

$$\ell : \vartheta \mapsto |a_1 - \vartheta| + |a_2 - \vartheta|.$$

Angenommen, $a_1 < a_2$.

Für $\vartheta < a_1$ hat ℓ die Steigung -2,

für $\vartheta > a_2$ hat ℓ die Steigung $+2$,

für $\vartheta \in (a_1, a_2)$ hat ℓ die Steigung 0.

Also ist jedes $\vartheta \in [a_1, a_2]$ Minimalstelle von ℓ.

$$f_\vartheta(a_1,\ldots,a_n) := \left(\tfrac{1}{2}\right)^n \prod_{i=1}^{n} e^{-|a_i-\vartheta|},\ a_i \in \mathbb{R}$$

Betrachten wir jetzt den Fall $n = 3$,

und für festes $a_1, a_2, a_3 \in \mathbb{R}$ die Funktion

$$\ell : \vartheta \mapsto |a_1 - \vartheta| + |a_2 - \vartheta| + |a_3 - \vartheta|.$$

Angenommen, $a_1 < a_2 < a_3$.

Für $\vartheta < a_2$ hat ℓ negative Steigung,

für $\vartheta > a_2$ hat ℓ positive Steigung.

Also ist a_2 die einzige Minimalstelle von ℓ.

Was 2 und 3 recht ist, soll n billig sein.

Definition. Seien $a_1, \ldots, a_n \in \mathbb{R}$.

Eine Zahl heißt **Median** von $a_1, \ldots, a_n$,

wenn

ebenso viele der a_i links wie rechts von ihr liegen.

Für ungerades n führt die Definition auf einen einzigen Wert,

für gerades n führt sie auf ein Intervall.

Fazit:

Im Besipiel der zweisitigen Exponentialverteilung ist der ML-Schätzer für den Lageparameter (das “Zentrum”) ϑ von der Form

$$h(X_1, \ldots, X_n) := \text{ Median von } X_1, \ldots, X_n$$

(und nicht, wie man auf die Schnelle vielleicht vermuten würde, der arithmetische Mittelwert der X_i).