Statistik für Informatiker, SS 2018

2. Ideen aus der Statistik

2.3 Schätzprinzipien

Timo Schlüter und Matthias Birkner

http://www.staff.uni-mainz.de/birkner/StatInfo18/

25.6.2018

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Inhalt

Wir haben bisher (nur) den empirischen Mittelwert

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

als Schätzer für den Erwartungswert einer (unbekannten) Verteilung verwendet.

Das ist sehr naheliegend und diese Schätzer haben auch „gute“ Eigenschaften (Erwartungstreue, Konsistenz, asymptotische Normalität), wie wir in Kapitel 2.2 gesehen haben.

Andererseits trifft man auch Situationen, in denen zumindest auf den ersten Blick kein „offensichtlicher“ Schätzer auf der Hand liegt; (spätestens) dann lohnen sich systematischere Ansätze, von denen wir einige in diesem Kapitel betrachten.

Schätzer, allgemein

Wir betrachten eine Zufallsvariable X mit Wertebereich S in einem statistischen Modell $(\Omega, \mathcal{F}, (P_\vartheta)_{\vartheta \in \Theta})$.

Wir verwenden X, um die Beobachtungen zu modellieren (oft schreiben wir $X = (X_1, X_2, \ldots, X_n)$, wenn das Experiment aus n Wiederholungen oder „Bauteilen“ besteht).

Abstrakt gesehen ist ein Schätzer für ϑ eine Funktion $T : S \to \Theta$ (mit der Interpretation, dass wir anhand von Beobachtungen $x \in S$ im Rahmen des Modells „tippen“ würden, dass der Wert von ϑ wohl $T(x)$ ist).

Man nennt diese Situation ein *statistisches Standardmodell*, wenn entweder S diskret ist oder $S \subset \mathbb{R}^n$ gilt und X unter P_ϑ

$$\text{Gewichte } \rho_\vartheta(\cdot) \quad \text{bzw.} \quad \text{Dichte } \rho_\vartheta(\cdot) \qquad \text{für } \vartheta \in \Theta$$

besitzt. Die Funktion

$$\begin{aligned} \rho : \quad S \times \Theta \quad &\to [0, \infty) \\ \cup \qquad\quad \cup \quad &\quad \cup \\ (x, \vartheta) \quad &\mapsto \rho(x, \vartheta) := \rho_\vartheta(x) \end{aligned}$$

heißt die *Likelihood-Funktion* (manchmal auch „Plausibilitäts-Funktion“), für $x \in S$ heißt

$$L_x : \Theta \to [0, \infty), \quad L_x(\vartheta) = \rho(x, \vartheta)$$

die Likelihood-Funktion zum Beobachtungswert x.

ML-Schätzer

Ein Schätzer $T : S \to \Theta$ heißt (ein) *Maximum-Likelihood-Schätzer*, wenn

$$\rho(x, T(x)) = \max_{\vartheta \in \Theta} \rho(x, \vartheta) \quad \forall x \in S$$

(auch kurz ML-Schätzer genannt, engl. MLE = maximum likelihood estimator).

Man schreibt oft auch $\widehat{\vartheta}_{\mathrm{ML}}$ für den ML-Schätzer.

Beispiele. 1. („Rückfangmethode“, engl. “capture-recapture”)

Ein Teich enthalte ϑ Fische (einer gewissen Art, $\vartheta \in \mathbb{N}$ ist der unbekannte Parameter), fange und markiere m, setze wieder aus. Wenn sich die markierten Fische gut verteilt haben, fange erneut n Fische.

Nehmen wir an, wir beobachten unter den erneut gefangenen x markierte Fische.

Beispiele. 1. („Rückfangmethode“, engl. “capture-recapture”)

Formalisierung als statistisches Modell: $S = \{0, 1, \dots, n\}$, $\Theta = \{(m \vee n), (m \vee n) + 1, (m \vee n) + 2, \dots\}$, unter P_ϑ ist die beobachtete Anzahl $X \sim \mathrm{Hyp}_{m,\vartheta-m,n}$ (hypergeometrische Verteilung, s.a. Bsp. 1.15).

Die Likelihood-Funktion ist demnach

$$\rho(x, \vartheta) = \frac{\binom{m}{x}\binom{\vartheta-m}{n-x}}{\binom{\vartheta}{n}},$$

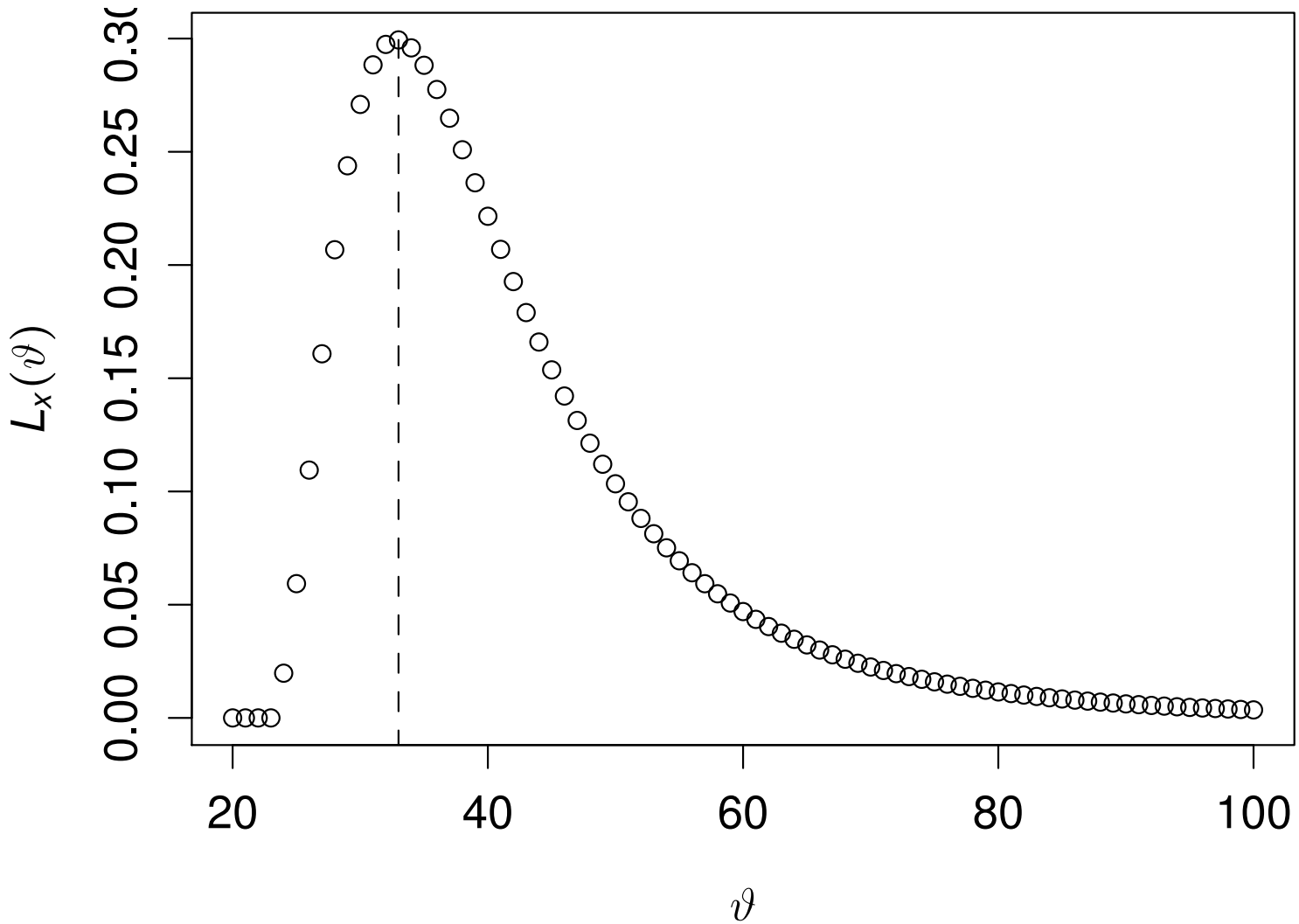
der ML-Schätzer ist

$$\widehat{\vartheta}_{\mathrm{ML}} = \left\lfloor \frac{n}{x} \cdot m \right\rfloor,$$

(mit $\lfloor a \rfloor = \max\{z \in \mathbb{Z} : z \leq a\}$).

Dies ist auch anschaulich plausibel, denn $\dfrac{m}{\widehat{\vartheta}_{\mathrm{ML}}} \approx \dfrac{x}{n}$.

Likelihood-Funktion $L_x(\vartheta)$, hier
$m = 10, n = 20, x = 6$ ($\widehat{\vartheta}_{\mathrm{ML}} = 33$)



$\widehat{\vartheta}_{\mathrm{ML}} = \left\lfloor \frac{n}{x} \cdot m \right\rfloor$: Es ist nämlich

$$\frac{\rho(x,\vartheta)}{\rho(x,\vartheta-1)} = \frac{\binom{\vartheta-m}{n-x}\binom{\vartheta-1}{n}}{\binom{\vartheta}{n}\binom{\vartheta-1-m}{n-x}}$$

$$= \frac{(\vartheta-m)(\vartheta-n)}{\vartheta(\vartheta-m-n+x)} = 1 - \frac{\vartheta x - mn}{\vartheta(\vartheta-m-n+x)} \begin{cases} > 1, & \vartheta < \frac{mn}{x}, \\ = 1, & \vartheta = \frac{mn}{x}, \\ < 1, & \vartheta > \frac{mn}{x} \end{cases}$$

(beachte: stets ist $\vartheta - m \geq n - x$, es gibt im Teich mindestens so viele unmarkierte Fische wie in der Rückfang-Stichprobe).

Bem.: Falls $\frac{mn}{x} \in \mathbb{N}$, so ist der ML-Schätzer hier nicht eindeutig: $\frac{mn}{x} - 1$ und $\frac{mn}{x}$ maximieren die Likelihood.

Beispiele. 2. (Erfolgsw'keit im Binomialmodell per ML schätzen)

Ein zufälliges Experiment mit zwei möglichen Ausgängen (Idealisierung: Werfen einer Münze) werde unabhängig (unter identischen Bedingungen) n-fach wiederholt, wir zählen die Anzahl X der „Erfolge".

$S = \{0, 1, \ldots, n\}$, unter P_ϑ, $\theta \in \Theta = [0, 1]$ ist $X \sim \mathrm{Bin}_{n,\vartheta}$, also

$$\rho(x, \vartheta) = L_x(\vartheta) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x}$$

Beispiele. 2. $\rho(x,\vartheta) = L_x(\vartheta) = \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}$

$$\frac{d}{d\vartheta}\log L_x(\vartheta) = \frac{d}{d\vartheta}\Big(\log\binom{n}{x} + x\log\vartheta + (n-x)\log(1-\vartheta)\Big)$$
$$= \frac{x}{\vartheta} - \frac{n-x}{1-\vartheta} = 0 \iff \vartheta = \frac{x}{n}$$

d.h. hier ist $\widehat{\vartheta}_{\mathrm{ML}} = \dfrac{x}{n}$.

(Es ist $\frac{d}{d\vartheta}\log L_x(\vartheta) > 0$ für $\vartheta < x/n$ und $\frac{d}{d\vartheta}\log L_x(\vartheta) < 0$ für $\vartheta > x/n$, d.h. es handelt sich tatsächlich um ein Maximum; Inspektion zeigt, dass auch in den Randfällen $x = 0$ und $x = n$ $\widehat{\vartheta}_{\mathrm{ML}} = \frac{x}{n}$ gilt.)

$\widehat{\vartheta}_{\mathrm{ML}}$ ist hier auch ein sehr „naheliegender“ Schätzer: Wir schätzen die (unbekannte) Erfolgswahrscheinlichkeit durch die relative Anzahl der beobachteten Erfolge.

Beispiele. 3. (Normales Modell mit bekannter Varianz)

n Beobachtungen seien u.i.v. $\sim \mathcal{N}_{\vartheta,\sigma^2}$, $\sigma^2 > 0$ sei bekannt und $\vartheta \in \Theta = \mathbb{R}$ soll geschätzt werden.

Mit $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ ist

$$\rho(x,\vartheta) = L_x(\vartheta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x_i - \vartheta)^2}{2\sigma^2} \Big)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \vartheta)^2 \Big)$$

d.h.

$$L_x(\vartheta) \stackrel{!}{=} \max \iff \sum_{i=1}^{n} (x_i - \vartheta)^2 \stackrel{!}{=} \min .$$

Beispiele. 3. (Normales Modell mit bekannter Varianz)

$$L_x(\vartheta) \stackrel{!}{=} \max \iff \sum_{i=1}^{n}(x_i - \vartheta)^2 \stackrel{!}{=} \min .$$

Mit $\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ ist

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \vartheta)^2 = \left(\overline{x} - \vartheta\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2,$$

d.h. es ist $\widehat{\vartheta}_{\mathrm{ML}} = \overline{x}$, das empirische Mittel der Beobachtungen.

(Für die Gleichung beachte $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2$.)

Beispiele. 3. (Normales Modell mit bekannter Varianz)
Die ML-Schätzer $\widehat{\vartheta}_{\mathrm{ML}} = \overline{x}$ stimmen hier mit dem naheliegenden Lageschätzer aus Kap. 2.2 überein, sie sind insbesondere konsistent und (hier sogar exakt) normalverteilt.

(Dies gilt asymptotisch für $n \to \infty$ recht allgemein für ML-Schätzer.)

Beispiele. 4. (Normales Modell, unbekannter Erwartungswert und unbekannte Varianz)

n Beobachtungen seien u.i.v. $\sim \mathcal{N}_{\mu,v}$ mit unbekannten $\mu \in \mathbb{R}$ und $v \in (0, \infty)$.

(Formalisierung: $S = \mathbb{R}^n$, $\Theta = \{\vartheta = (\mu, v) : \mu \in \mathbb{R}, v > 0\}$, unter $P_{(\mu,v)}$ ist $X = (X_1, \ldots, X_n) \sim \mathcal{N}_{\mu,v}^{\otimes n}$)

Wie in 3. ist

$$\log L_x\big((\mu, v)\big) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log v - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2,$$

nach obigem ist $\widehat{\mu}_{\mathrm{ML}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ Maximierer bezüglich μ (für jeden Wert von v).

Beispiele. 4. (Normales Modell, unbekannter Erwartungswert und unbekannte Varianz)

$$\log L_x\big((\mu, v)\big) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log v - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2,$$

somit ist

$$\frac{\partial}{\partial v}\log L_x\big((\mu, v)\big)\Big|_{\mu=\widehat{\mu}_{\mathrm{ML}}} = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{\mathrm{ML}})^2$$

also

$$\frac{\partial}{\partial v}\log L_x\big((\widehat{\mu}_{\mathrm{ML}}, v)\big) = 0 \iff v = \widehat{v}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{\mathrm{ML}})^2.$$

(Und man prüft: $\log L_x\big((\widehat{\mu}_{\mathrm{ML}}, v)\big)$ ist wachsend für $v < \widehat{v}_{\mathrm{ML}}$, fallend für $v > \widehat{v}_{\mathrm{ML}}$.)

Beachte: Der ML-Schätzer für die unbekannte Varianz ist hier die (unkorrigierte) Stichprobenvarianz, also ist er nicht erwartungstreu (vgl. Kap. 2.2; allerdings ist er für halbwegs großes n auch nicht weit weg von Erwartungstreue).

Beispiele. 5. Beobachtungen seien u.i.v. uniform auf $\{1, 2, \ldots, \vartheta\}$ (mit einem unbekannten $\vartheta \in \mathbb{N}$).

Mögliche Interpretation: In einer Stadt gibt es ϑ Taxis, die mit $1, 2, \ldots, \vartheta$ durchnummeriert sind. Wir beobachten die Nummern von n zufällig gewählten Taxis und schätzen, wieviele Taxis es insgesamt gibt.

Formalisierung: $X = (X_1, \ldots, X_n)$ mit Werten in $S = \mathbb{N}^n$, $\rho\big((x_1, \ldots, x_n), \vartheta\big) = \vartheta^{-n} \mathbf{1}_{x_1, \ldots, x_n \le \vartheta}$ für $\vartheta \in \Theta = \mathbb{N}$.

Es ist

$$\widehat{\vartheta}_{\mathrm{ML}} = \max\{x_1, x_2, \ldots, x_n\},$$

denn

$$L_{(x_1, \ldots, x_n)}(\vartheta) = \begin{cases} \dfrac{1}{\vartheta^n}, & \text{falls } \vartheta \ge x_1, x_2, \ldots, x_n, \\ 0, & \text{sonst.} \end{cases}$$

Beispiele. 6. n Beobachtungen seien u.i.v. $\sim \mathrm{Poi}_\vartheta$ (mit einem unbekannten $\vartheta \in \Theta := (0, \infty)$), d.h. $\rho\big((x_1, \ldots, x_n), \vartheta\big) = \prod_{i=1}^n e^{-\vartheta} \frac{\vartheta^{x_i}}{x_i!}$

Für $x = (x_1, \ldots, x_n)$ ist

$$L_x(\vartheta) = \frac{\exp\big(-n\vartheta + (x_1 + \cdots + x_n) \log \vartheta \big)}{x_1! x_2! \cdots x_n!}$$

also

$$\frac{\partial}{\partial \vartheta} \log L_x(\vartheta) = -n + \frac{x_1 + \cdots + x_n}{\vartheta} \stackrel{!}{=} 0 \quad \Longleftrightarrow \quad \vartheta = \widehat{\vartheta}_{\mathrm{ML}}$$

mit $\widehat{\vartheta}_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^n x_i$, d.h. der ML-Schätzer ist hier wiederum der empirische Mittelwert.

(Wegen $\mathbb{E}[X] = \vartheta$ für $X \sim \mathrm{Poi}_\vartheta$ ist es auch zugleich der sogenannte „Momentenschätzer“.)

Betrachen wir Beispiel 1.21 nochmals in diesem Licht:

Ladislaus von Bortkewitsch, *Das Gesetz der kleinen Zahlen*, Teubner, 1898, § 12, 4. („Die durch Schlag eines Pferdes im preußischen Heere getöteten“) berichtet für 20 Jahre (1875–1894) und 10 Armeekops der preußischen Kavallerie, also insgesamt $20 \cdot 10 = 200$ „Korpsjahre“ berichtet, in wievielen davon sich x Todesfälle durch Schlag eines Pferds ereigneten (Tabelle b) auf S. 25):

Ergebnis x	Anz. „Korpsjahre“	$200 \times \mathrm{Poi}_{0,61}(x)$
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
≥ 5	0	0,08

Ergebnis x	Anz. „Korpsjahre“	$200 \times \mathrm{Poi}_{0,61}(x)$
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
≥5	0	0,08

Sei X_i = Anzahl der im i-ten Korpsjahr Getöteten $\sim \mathrm{Poi}_\vartheta$, u.a. für verschiedene i.

Die Daten geben uns zwar nicht die Beobachtungen x_i selbst, enthalten aber genug Informationen, um

$$\widehat{\vartheta}_{\mathrm{ML}} = \frac{1}{200}\sum_{i=1}^{200} x_i$$
$$= \frac{109}{200}\cdot 0 + \frac{65}{200}\cdot 1 + \frac{22}{200}\cdot 2 + \frac{3}{200}\cdot 3 + \frac{1}{200}\cdot 4 + 0 = 0{,}61$$

zu berechnen.

Bericht („gute“ Eigenschaften der ML-Schätzer)**.**

Betrachte ein Produktmodell, d.h. $X = (X_1, \ldots, X_n)$ mit Werten in $S = S_1^n$ und

$$\rho\big((x_1, \ldots, x_n), \vartheta\big) = \prod_{i=1}^{n} \rho^{(1)}(x_i, \vartheta)$$

für eine Likelihood-Funktion $\rho^{(1)}$ auf $S_1 \times \Theta$ (unter P_ϑ sind die Beobachtungen $X_1, \ldots, X_n$ u.i.v. mit Gewichten / Dichte $\rho^{(1)}(\cdot, \vartheta)$).

Bericht („gute“ Eigenschaften der ML-Schätzer)**.**

Sei $\widehat{\vartheta}_{\mathrm{ML},n} = \widehat{\vartheta}_{\mathrm{ML},n}(X_1, \ldots, X_n)$ der ML-Schätzer basierend auf n unabhhängigen Beobachtungen, dann gilt (unter gewissen Bedingungen an $\rho^{(1)}(\cdot,\cdot)$) :

- Die ML-Schätzer sind konsistent, d.h. $P_\vartheta\big(|\widehat{\vartheta}_{\mathrm{ML},n} - \vartheta| \geq \varepsilon\big) \xrightarrow[n\to\infty]{} 0$ für jedes $\varepsilon > 0$ und jedes $\vartheta \in \Theta$.
- Die ML-Schätzer sind asymptotisch normal mit (asymptotisch optimaler) Varianz $1/(nI(\vartheta))$, d.h. $\sqrt{n}\big(\widehat{\vartheta}_{\mathrm{ML},n} - \vartheta\big) \xrightarrow[n\to\infty]{d} \mathcal{N}_{0,1/I(\vartheta)}$ unter P_ϑ für jedes $\vartheta \in \Theta$, wobei $I(\vartheta) := \mathrm{Var}_\vartheta\Big[\dfrac{d}{d\vartheta} \log \rho(X,\vartheta)\Big]$ die sogenannte Fisher-Information ist.

(Bem.: Im Fall des ML-Schätzers für den Erwartungswert im normalen Modell haben wir dies schon gesehen;
die Botschaft hier ist: es gilt ziemlich allgemein.)

Wir folgten bisher (und werden dies auch in den späteren Kapiteln wieder tun) dem klassischen, sogenannten frequentistischen Ansatz der Statistik:

Wir fassen eine Menge Θ von „Parametern“ ins Auge, für $\Theta \ni \vartheta$ beschreibt (in einem statistischen Modell) P_ϑ die Verteilung der Beobachungen, wenn dieses ϑ der tatsächlich zutreffende (sozusagen der „wahre“) Parameter ist. In der konkreten Anwendungssituation kennen wir dieses „wahre“ ϑ natürlich i.A. nicht, wir fassen es als zwar unbekannte, aber prinzipiell feste Größe auf. Wahrscheinlichkeitsaussagen beziehen sich *nicht* auf ϑ, sondern auf zufällige Beobachtungen unter P_ϑ.

Ansatz der Bayes-Statistik

Dies ist anders in der *Bayes-Statistik*: Man wählt eine Wahrscheinlichkeitsverteilung α auf Θ, die *a priori*-Verteilung (auch *Vorbewertung*) und stellt sich vor, dass die Daten einem zweistufigen Experiment entstammen:

- Zunächst wird der Parameter ϑ gemäß der a priori-Verteilung α erzeugt (insbesondere ist ϑ jetzt selbst eine Zufallsvariable),
- dann werden die Beobachtungen X zufällig erzeugt mit einer Verteilung, die vom gewählten ϑ abhängt.

Insbesondere besitzt in dieser Formulierung das Paar (X, ϑ) eine *gemeinsame Verteilung*.

Es gelte: Die a priori-Verteilung auf Θ hat Dichte bzw. Gewichte $\alpha(\vartheta)$

(je nachdem, ob ϑ kontinuierlich oder diskret verteilt ist; wir betrachten im Folgenden nur den Fall, dass $\Theta \subset \mathbb{R}$ ein Intervall ist und ϑ eine Dichte besitzt)

Interpretation: Ohne Kenntnis der Beobachtungen nehmen wir an, dass ϑ die a priori-Verteilung besitzt (z.B. aus „Erfahrung“ oder aus „Expertenwissen“).

Wir interpretieren die Likelihood-Funktion $\rho : S \times \Theta \to [0, \infty)$ als

$$\rho(x, p) = P(X = x \mid \vartheta = p)$$

(bzw. $P(X \in dx \mid \vartheta = p) = \rho(x, p)\, dx$ falls X eine Dichte besitzt)

Mit Formel von der totalen Wahrscheinlichkeit (Satz 1.41, 1.) ist

$$P(X = x) = \int_\Theta \rho(x,t)\,\alpha(t)\,dt$$

(bzw. $P(X \in dx) = \int_\Theta \rho(x,t)\,\alpha(t)\,dt\,dx$, d.h. $P(X \leq x) = \int_\Theta \int_{-\infty}^{x} \rho(y,t)\,dy\,\alpha(t)\,dt$, wenn X eine Dichte besitzt),

mit der Formel von Bayes (Satz 1.41, 2.) ist

$$P(\vartheta \in dp \mid X = x) = \frac{\rho(x,p)\alpha(p)\,dp}{\int_\Theta \alpha(\vartheta')\rho(x,\vartheta')\,d\vartheta'}$$

Die *a posteriori-Dichte* (bzw. a posteriori-Gewicht, wenn ϑ diskret) bei Beobachtung x,

$$\pi_x(\vartheta) = \frac{\alpha(\vartheta)\rho(x,\vartheta)}{\int_\Theta \alpha(\vartheta')\rho(x,\vartheta')\,d\vartheta'},$$

ist die Dichte von ϑ, bedingt auf Beobachtung $X = x$, d.h.

$$P(\vartheta \le u \mid X = x) = \int_{\Theta\cap(-\infty,u]} \pi_x(p)\,dp$$

Der *Bayes-Schätzer* (zur a priori-Verteilung α) ist

$$\widehat{\vartheta}_{\mathrm{B}} = \widehat{\vartheta}_{\mathrm{B}}(x) := \mathbb{E}_{\pi_x}[\vartheta] = \int_\Theta \vartheta\,\pi_x(\vartheta)\,d\vartheta$$

(d.h. der Erwartungswert von ϑ bedingt auf $X = x$).

(Wir betrachten hier nur den Fall, dass Θ ein Intervall ist.)

Definition. Für einen Schätzer $Y = Y(X)$ (für ϑ) ist

$$F_\alpha(Y) := \int_\Theta \mathbb{E}\big[(Y-\vartheta)^2 \,\big|\, \vartheta = p\big]\alpha(p)\,dp$$

der erwartete quadratische Fehler (zur Vorbewertung α).

Der Bayes-Schätzer minimiert den erwarteten quadratischen Fehler (zur Vorbewertung α) :

Satz. Stets gilt $F_\alpha(Y) \geq F_\alpha(\widehat{\vartheta}_{\mathrm{B}}(X))$.

$F_\alpha(Y) \geq F_\alpha(\widehat{\vartheta}_{\mathrm{B}}(X))$.

Beweis.

$$
\begin{aligned}
& F_\alpha(Y(X)) - F_\alpha(\widehat{\vartheta}_{\mathrm{B}}(X)) \\
&= \int_\Theta \mathbb{E}\Big[(Y(X) - \vartheta)^2 - (\widehat{\vartheta}_{\mathrm{B}}(X) - \vartheta)^2 \,\Big|\, \vartheta = p\Big]\alpha(p)\,dp \\
&= \int_\Theta \mathbb{E}\Big[Y(X)^2 - 2Y(X)\vartheta - \widehat{\vartheta}_{\mathrm{B}}(X)^2 + 2\vartheta\widehat{\vartheta}_{\mathrm{B}}(X) \,\Big|\, \vartheta = p\Big]\alpha(p)\,dp \\
&= \mathbb{E}\Big[Y(X)^2 - 2Y(X)\vartheta - \widehat{\vartheta}_{\mathrm{B}}(X)^2 + 2\vartheta\widehat{\vartheta}_{\mathrm{B}}(X)\Big] \\
&= \mathbb{E}\Big[Y(X)^2 - \widehat{\vartheta}_{\mathrm{B}}(X)^2\Big] - 2\mathbb{E}\Big[Y(X)\vartheta\Big] + 2\mathbb{E}\Big[\vartheta\widehat{\vartheta}_{\mathrm{B}}(X)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&F_\alpha(Y(X)) - F_\alpha(\widehat{\vartheta}_{\mathrm{B}}(X)) \\
&= \mathbb{E}\big[Y(X)^2 - \widehat{\vartheta}_{\mathrm{B}}(X)^2\big] - 2\mathbb{E}\big[Y(X)\vartheta\big] + 2\mathbb{E}\big[\vartheta\widehat{\vartheta}_{\mathrm{B}}(X)\big]
\end{aligned}
$$

Weiter ist

$$
\begin{aligned}
&\mathbb{E}\big[\vartheta\widehat{\vartheta}_{\mathrm{B}}(X)\big] \\
&= \sum_{x\in S} \mathbb{E}\big[\vartheta\widehat{\vartheta}_{\mathrm{B}}(X) I_{\{X=x\}}\big] = \sum_{x\in S} P(X=x)\widehat{\vartheta}_{\mathrm{B}}(x)\mathbb{E}[\vartheta \mid X=x] \\
&= \sum_{x\in S} P(X=x)\widehat{\vartheta}_{\mathrm{B}}(x)\mathbb{E}_{\pi_x}[\vartheta] = \sum_{x\in S} P(X=x)\big(\widehat{\vartheta}_{\mathrm{B}}(x)\big)^2 \\
&= \mathbb{E}\big[\big(\widehat{\vartheta}_{\mathrm{B}}(X)\big)^2\big]
\end{aligned}
$$

und analog

$$
\mathbb{E}\big[\vartheta Y(X)\big] = \mathbb{E}\big[Y(X)\widehat{\vartheta}_{\mathrm{B}}(X)\big].
$$

Insgesamt:

$$
\begin{aligned}
& F_\alpha(Y(X)) - F_\alpha(\widehat{\vartheta}_{\mathrm{B}}(X)) \\
&= \mathbb{E}\big[\, Y(X)^2 - \widehat{\vartheta}_{\mathrm{B}}(X)^2 - 2Y(X)\widehat{\vartheta}_{\mathrm{B}}(X) + 2\widehat{\vartheta}_{\mathrm{B}}(X)^2 \big] \\
&= \mathbb{E}\big[\big(Y(X) - \widehat{\vartheta}_{\mathrm{B}}(X)\big)^2\big] \geq 0.
\end{aligned}
$$

□

Beispiel (Münzwurf mit zufälliger Erfolgswahrscheinlichkeit)**.** ϑ wird gemäß einer Verteilung α auf $\Theta := [0,1]$ „ausgewürfelt“, dann:

$n \in \mathbb{N}$, gegeben $\vartheta = u \in [0,1]$ seinen $X_1, X_2, \ldots, X_n$ unabhängig und jeweils $\sim \mathrm{Ber}_u$
(d.h. $P(X_i = 1 \mid \vartheta = u) = u = 1 - P(X_i = 0 \mid \vartheta = u)$).

Somit: Für $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ ist

$$\rho(x, \vartheta) = \vartheta^{\#\{i \le n : x_i = 1\}} (1 - \vartheta)^{\#\{i \le n : x_i = 0\}}.$$

Beispiel (Münzwurf mit zufälliger Erfolgswahrscheinlichkeit)**.**

Eine Situation, in der dieses Modell sinnvoll ist, könnte folgende sein: Nehmen wir an, ein Versicherungsnehmer hat jedes Jahr mit einer gewissen (zu ihm „gehörigen“) Wahrscheinlichkeit ϑ einen Schadensfall (unabhängig über die Jahre), und $\alpha(\vartheta)\,d\vartheta$ beschreibt die Verteilung der Schadenswahrscheinlichkeiten aller Kunden dieser Versicherung (diese Verteilung sei der Versicherung aus Erfahrungswerten bekannt).

Mit $\rho(x,\vartheta) = \mathrm{Bin}_{n,\vartheta}(x)$ ist dann die Wahrscheinlichkeit, dass ein „typischer Kunde“ in n Jahren k Schadensfälle verursacht $\int_0^1 \alpha(\vartheta)\rho(k,\vartheta)\,d\vartheta$, und $\pi_k(\vartheta)$ ist die Verteilung der Schadenswahrscheinlichkeit pro Jahr eines Kunden, bedingt darauf, dass er in den letzten n Jahren k Schäden hatte. Diese Information kann die Versicherung beispielsweise für Vertragsverlängerung, Tarifanpassung, etc. benutzen.

Wir betrachten hier (nur) den Fall $\alpha = \mathrm{unif}_{[0,1]}$.

Gehe über zu $Y = X_1 + \cdots + X_n$, dann ist gegeben $\vartheta = u$, $Y \sim \mathrm{Bin}_{n,u}$ und für $k \in \{0, 1, \ldots, n\}$

$$P(Y = k) = \int_0^1 \binom{n}{k} u^k (1-u)^{n-k}\, du$$
$$= \frac{n!}{k!(n-k)!} \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}$$

(d.h. Y ist uniform auf $\{0, 1, \ldots, n\}$).

(Für obiges Integral brauchen wir eine kleine Nebenrechnung, siehe folgende Folie.)

Definition und Lemma. Für $a, b \in (0, \infty)$ ist die Beta-Funktion gegeben durch

$$B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

wobei $\Gamma(a) = \int_0^\infty y^{a-1}e^{-y}\,dy$ die Gamma-Funktion ist

(beachte: $\Gamma(a+1) = a\Gamma(a)$, speziell für $a \in \mathbb{N}$ ist $\Gamma(a) = (a-1)!$, wie man mit partieller Integration nachrechnen kann).

Für $a, b \in \mathbb{N}$ kann man explizit rechnen:

Es ist dann $B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$, denn

$B(a,1) = \int_0^1 u^{a-1}\,du = \left[\frac{1}{a}u^a\right]_{u=0}^{u=1} = \frac{1}{a}$ und für $b \in \{2,3,\dots\}$ ist (mit partieller Integration)

$$\int_0^1 u^{a-1}(1-u)^{b-1}\,du$$

$$= \left[\tfrac{1}{a}u^a(1-u)^{b-1}\right]_{u=0}^{u=1} - \int_0^1 \tfrac{1}{a}u^a \cdot (b-1)(1-u)^{b-2}(-1)\,du$$

$$= \frac{b-1}{a}\int_0^1 u^a(1-u)^{b-2}\,du$$

also

$$B(a,b) = \frac{b-1}{a}B(a+1,b-1) = \frac{(b-1)\cdot(b-2)\cdots 2\cdot 1\cdot B(a+b-1,1)}{a\cdot(a+1)\cdots(a+b-3)\cdots(a+b-2)}$$

$$= \frac{(b-1)\cdot(b-2)\cdots 2\cdot 1}{a\cdot(a+1)\cdots(a+b-2)\cdots(a+b-1)} = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

Definition und Lemma (Beta-Verteilungen)**.** $r, s > 0$. Eine ZV V mit Werten in $[0,1]$ ist *Beta-verteilt* mit Parametern $r, s > 0$, in Formeln $V \sim \beta_{r,s}$, wenn V die Dichte

$$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} v^{r-1}(1-v)^{s-1}\mathbf{1}_{(0,1)}(v)$$

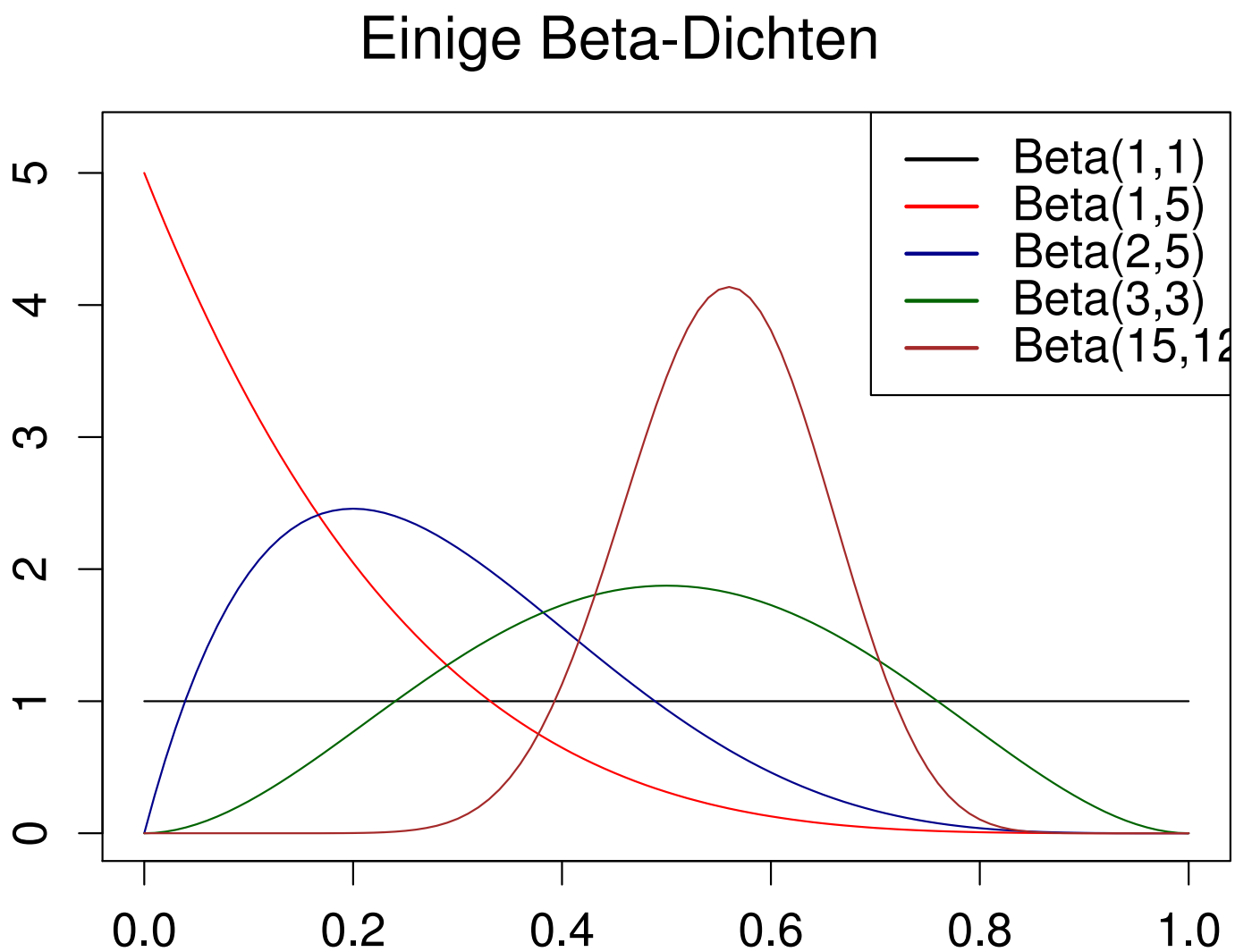
besitzt. Es gilt dann

$$\mathbb{E}[V] = \frac{r}{r+s}.$$

Denn

$$\mathbb{E}[V] = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \int_0^1 v \cdot v^{r-1}(1-v)^{s-1}\, dv$$

$$= \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \frac{\Gamma(r+1)\Gamma(s)}{\Gamma(r+s+1)} = \frac{\Gamma(r+s)}{\Gamma(r+s+1)} \frac{\Gamma(r+1)}{\Gamma(r)} = \frac{r}{r+s}$$


Einige Beta-Dichten
Beta(1,1)
Beta(1,5)
Beta(2,5)
Beta(3,3)
Beta(15,12
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0

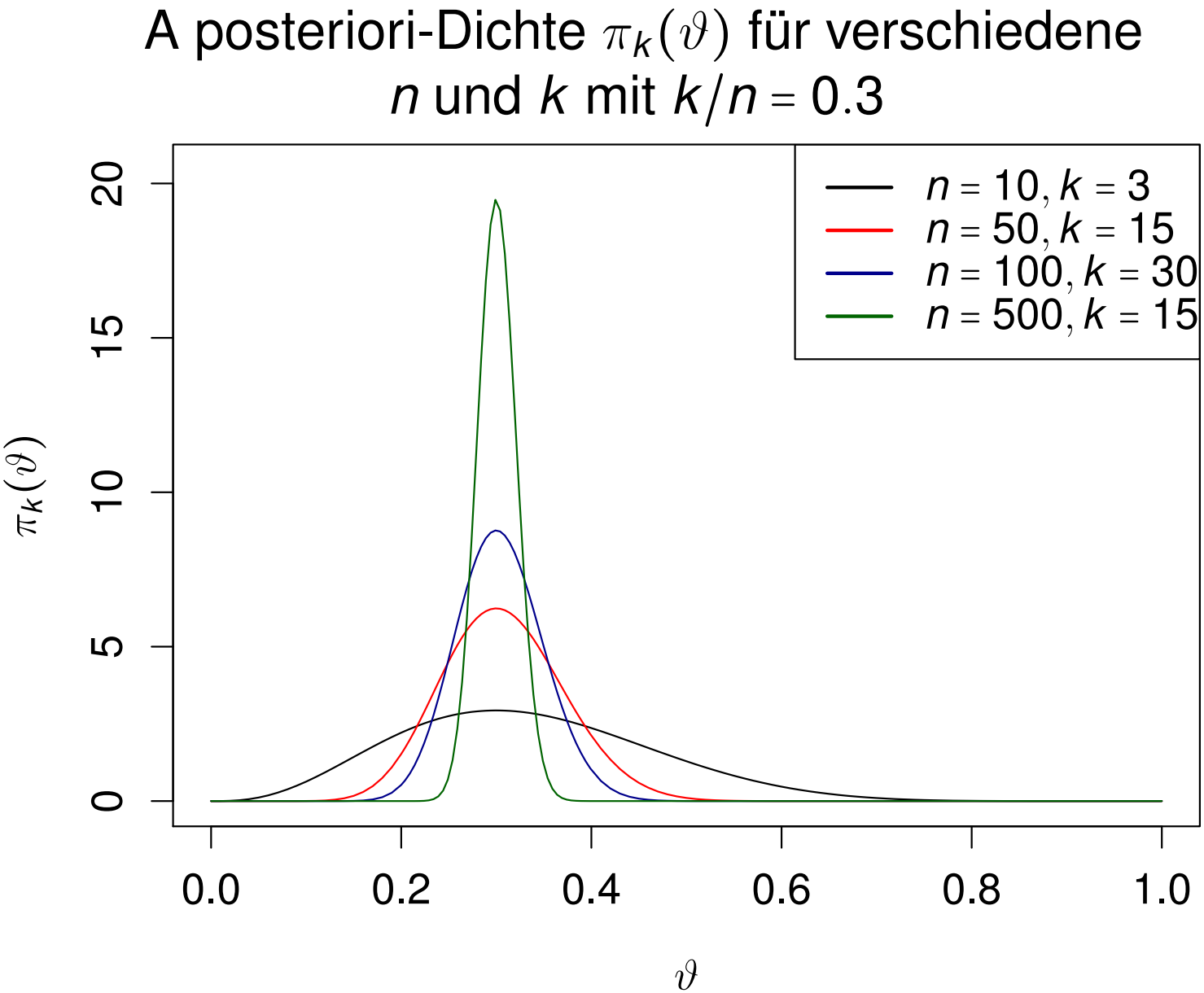
Zurück zum **Beispiel** „Münzwürfe mit zufälliger Erfolgsw'keit“:

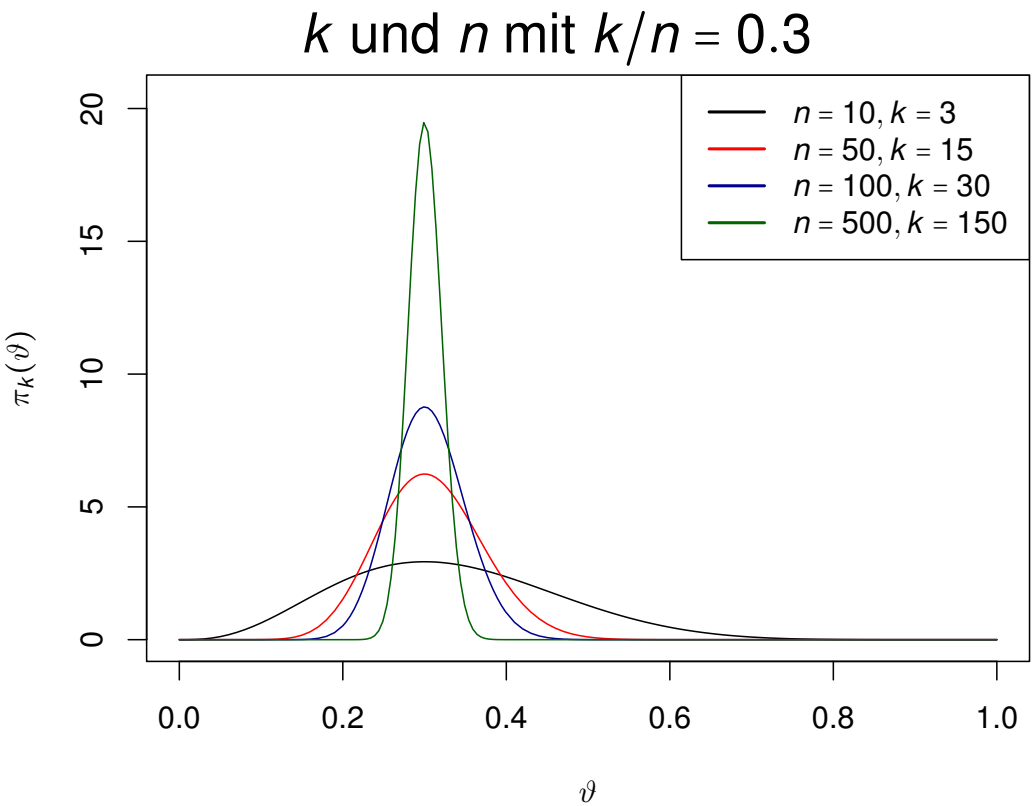
Die a posteriori-Verteilung ist $\mathcal{L}(\vartheta \mid Y = k) = \beta_{k+1,n-k+1}$:

$$\begin{aligned} P(\vartheta \in dp \mid Y = k) &= \frac{P(\vartheta \in dp, Y = k)}{P(Y = k)} \\ &= \frac{1}{1/(n+1)} \binom{n}{k} p^k (1-p)^{n-k}\, dp \\ &= \frac{(n+1)!}{k!(n-k)!} p^k (1-p)^{n-k}\, dp \end{aligned}$$

Demnach (mit obigem Lemma zur Beta-Verteilung) ist

$$\widehat{\vartheta}_{\mathrm{B}} = \widehat{\vartheta}_{\mathrm{B}}(Y) = \frac{Y+1}{n+2}$$


A posteriori-Dichte $\pi_k(\vartheta)$ für verschiedene
n und k mit $k/n = 0.3$
$n = 10, k = 3$
$n = 50, k = 15$
$n = 100, k = 30$
$n = 500, k = 15$
$\pi_k(\vartheta)$
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
ϑ



Wir sehen, dass für großes n die a posteriori-Verteilung recht eng um $\frac{Y+1}{n+2} \approx \frac{Y}{n}$ konzentriert ist.

Zudem: die frequentistische und die Bayes'sche „Antwort" stimmen für große n „nahezu" überein.

Bemerkung. Laplace[1] antwortete auf die die von ihm (vielleicht mit einem Augenzwinkern) gestellte Frage:

„Angenommen die Sonne ist bis heute n-mal aufgegangen. Mit welcher Wahrscheinlichkeit geht sie morgen auf?“

$$\frac{n+1}{n+2}$$

Dies passt zur Antwort des Bayes-Schätzers in obigem Beispiel.

[1] Pierre-Simon Laplace, 1749–1827; zitiert nach Kersting & Wakolbinger, *Elementare Stochastik*, 2. Aufl., Birkhäuser 2010, S. 127

Der sogenannte kleinste-Quadrate-Ansatz ist ebenfalls ein recht allgemeines Prinzip zur Konstruktion von Schätzern, wir betrachten ihn hier am Beispiel des linearen Regressionsmodells:

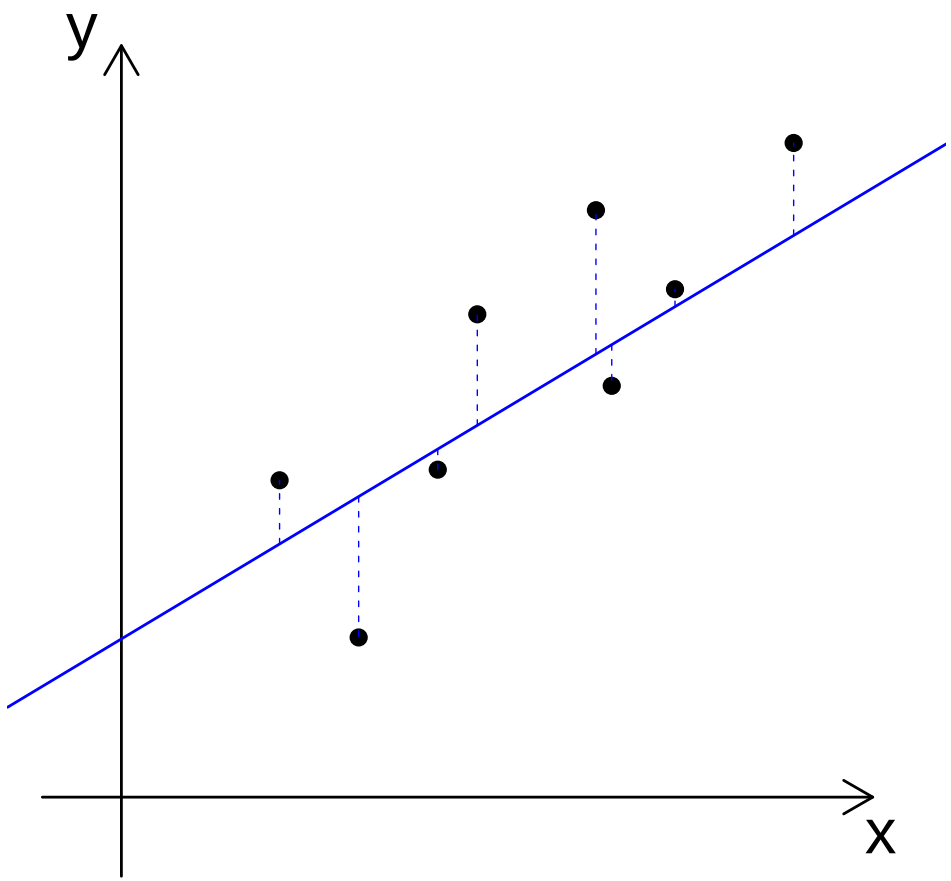
Nehmen wir an, die Beobachtungen bestehen aus n Messwertpaaren (x_i, y_i), $i = 1, \ldots, n$ (Werte in $\mathbb{R}^2$) und wir vermuten aus theoretischen Gründen einen zumindest „ungefähren“ (affin-)linearen Zusammenhang, d.h. bei „perfekter“ Messung und „perfektem“ Zusammenhang gälte

$$y_i = \beta_0 + \beta_1 x_i$$

für gewisse (uns unbekannte) Zahlen β_0 und β_1.

(Ein „Lehrbuchbeispiel“: y_i ist die Länge einer Stahlfeder bei Zugbelastung mit Gewicht x_i innerhalb des Gültigkeitsbereich des Hooke’schen Gesetzes.)

Aufgrund beispielsweise von Messungenauigkeiten (oder womöglich auch weil der lineare Zusammenhang in Wirklichkeit nur approximativ gilt) werden die realen Datenpunkte typischerweise nicht auf einer Geraden liegen.



Formulierung als statistisches Modell:

$x_1, \ldots, x_n$ seien feste (bekannte) Werte (x ist die „erklärende Variable“), für $\vartheta = (\beta_0, \beta_1) \in \Theta = \mathbb{R}^2$ sei unter P_ϑ

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, \ldots, n$$

mit ε_i u.i.v. mit $\mathbb{E}[\varepsilon_i] = 0, \; \mathrm{Var}[\varepsilon_i] = \sigma^2$

und wir fassen die beobachteten y_i-Werte als Realisierungen der Y_i auf (y ist die „abhängige Variable“ oder „Zielgröße“).

Ein naheliegender Ansatz, $\vartheta = (\beta_0, \beta_1)$ zu schätzen, ist der *kleinste-Quadrate-Schätzer*: Finde $\widehat{\beta}_0, \widehat{\beta}_1$ so, dass

$$\sum_{i=1}^{n} \left(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right)^2 = \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Die Lösung kennen wir schon (vgl. Beob. 1.84), die wir hier gewissermaßen nur „statistisch aussprechen“): Mit

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\sigma_x^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad \sigma_y^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$\mathrm{cov}_{x,y} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

($\mathrm{cov}_{x,y}$ ist die „empirische Kovarianz“ der x- und der y-Werte) ist

$$\widehat{\beta}_1 = \frac{\mathrm{cov}_{x,y}}{\sigma_x^2}, \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

Die Gerade $x \mapsto \widehat{\beta}_0 + \widehat{\beta}_1 x$ heißt auch die *Ausgleichsgerade*, der Wert $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ der anhand von x_i „vorhergesagte Wert“ oder „Ausgleichswert“.

Man nennt weiter

$$r_i := y_i - \widehat{y}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$$

das *Residuum* zum i-ten Beobachtungswert (der „Rest“ der Abweichung, die von dem Modell (nur) durch den „Rauschterm“ erklärt wird).

$$\kappa_{x,y} := \frac{\mathrm{cov}_{x,y}}{\sigma_x \sigma_y}$$

ist der (empirische) Korrelationskoeffizient, auch *Pearsons Korrelationskoeffizient*[2].

[2] nach Karl Pearson, 1858–1936

$$\kappa_{x,y} := \frac{\mathrm{cov}_{x,y}}{\sigma_x \sigma_y}$$

(stets ist $-1 \leq \kappa_{x,y} \leq 1$, nach Cauchy-Schwarz-Ungleichung).

$R = \kappa_{x,y}^2$ nennt man auch das *Bestimmtheitsmaß*.

Je näher R an 1 liegt, um so besser passt die lineare Approximation der y-Werte durch die x-Werte.

Das sieht man auch gut an der alternativen Formel

$$R = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^n r_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

Zu den Formeln: Betrachte eine ZV $(\widetilde{X}, \widetilde{Y})$ mit Werten in $\mathbb{R}^2$, deren Verteilung die empirische Verteilung der Datenpunkte $\frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i)}$ ist, so ist

$$\begin{aligned}
&\mathbb{E}[\widetilde{X}] = \overline{x}, \quad \mathbb{E}[\widetilde{Y}] = \overline{y},\\
&\mathrm{Var}[\widetilde{X}] = \sigma_x^2, \quad \mathrm{Var}[\widetilde{Y}] = \sigma_y^2,\\
&\mathrm{Cov}[\widetilde{X}, \widetilde{Y}] = \mathrm{cov}_{x,y}, \quad \kappa_{\widetilde{X},\widetilde{Y}} = \kappa_{x,y}
\end{aligned}$$

und die Behauptung folgt wörtlich aus Beob. 1.84, dort hatten wir gerechnet:

$$\min_{\beta_0,\beta_1\in\mathbb{R}} \mathbb{E}\big[(\widetilde{Y} - \beta_1\widetilde{X} - \beta_0)^2\big] = (1 - \kappa^2_{\widetilde{X},\widetilde{Y}}) \min_{\beta_0\in\mathbb{R}} \mathbb{E}\big[(\widetilde{Y} - \beta_0)^2\big]$$

$$\Big(= (1 - \kappa^2_{\widetilde{X},\widetilde{Y}})\mathrm{Var}[\widetilde{Y}]\Big)$$

$$\min_{\beta_0,\beta_1\in\mathbb{R}} \mathbb{E}\big[(\widetilde{Y}-\beta_1\widetilde{X}-\beta_0)^2\big] = (1-\kappa^2_{\widetilde{X},\widetilde{Y}}) \min_{\beta_0\in\mathbb{R}} \mathbb{E}\big[(\widetilde{Y}-\beta_0)^2\big]$$

denn der Ausdruck auf der linken Seite ist

$$\begin{aligned}
&\mathrm{Var}[\widetilde{Y}-\beta_1\widetilde{X}-\beta_0] + \big(\mathbb{E}[\widetilde{Y}]-\beta_1\mathbb{E}[\widetilde{X}]-\beta_0\big)^2 \\
&= \mathrm{Var}[\widetilde{Y}] - 2\beta_1\mathrm{Cov}[\widetilde{X},\widetilde{Y}] + \beta_1^2\mathrm{Var}[\widetilde{X}] + \big(\mathbb{E}[\widetilde{Y}]-\beta_1\mathbb{E}[\widetilde{X}]-\beta_0\big)^2 \\
&= \sigma^2_{\widetilde{Y}} - 2\beta_1\sigma_{\widetilde{X}}\sigma_{\widetilde{Y}}\kappa_{\widetilde{X},\widetilde{Y}} + \beta_1^2\sigma^2_{\widetilde{X}} + \big(\mathbb{E}[\widetilde{Y}]-\beta_1\mathbb{E}[\widetilde{X}]-\beta_0\big)^2 \\
&= \sigma^2_{\widetilde{Y}}\big(1-\kappa^2_{\widetilde{X},\widetilde{Y}}\big) + \sigma^2_{\widetilde{X}}\Big(\beta_1 - \frac{\sigma_{\widetilde{Y}}}{\sigma_{\widetilde{X}}}\kappa_{\widetilde{X},\widetilde{Y}}\Big)^2 + \big(\mathbb{E}[\widetilde{Y}]-\beta_1\mathbb{E}[\widetilde{X}]-\beta_0\big)^2,
\end{aligned}$$

was offensichtlich minimal wird für die Wahl

$$\beta_1 = \beta_1^* := \frac{\sigma_{\widetilde{Y}}}{\sigma_{\widetilde{X}}}\kappa_{\widetilde{X},\widetilde{Y}}, \qquad \beta_0 = \beta_0^* := \mathbb{E}[\widetilde{Y}] - \beta_1^*\mathbb{E}[\widetilde{X}]$$

und dann den Wert $(1-\kappa^2_{\widetilde{X},\widetilde{Y}})\sigma^2_{\widetilde{Y}}$ hat.

$$\min_{\beta_0,\beta_1\in\mathbb{R}} \mathbb{E}\big[(\widetilde{Y}-\beta_1\widetilde{X}-\beta_0)^2\big] = (1-\kappa^2_{\widetilde{X},\widetilde{Y}}) \min_{\beta_0\in\mathbb{R}} \mathbb{E}\big[(\widetilde{Y}-\beta_0)^2\big]$$

$$\Big(= (1-\kappa^2_{\widetilde{X},\widetilde{Y}})\mathrm{Var}[\widetilde{Y}]\Big)$$

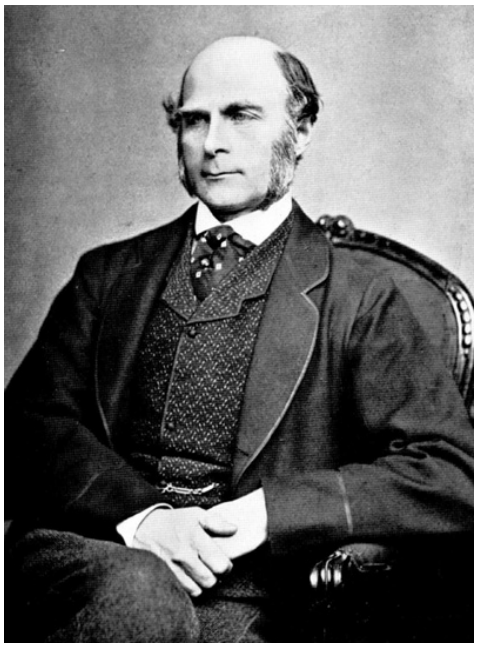
Für den Zusatz beachte analog:

$$\mathbb{E}\big[(\widetilde{Y}-\beta_0)^2\big] = \mathbb{E}[\widetilde{Y}^2] - 2\beta\mathbb{E}[\widetilde{Y}] + \beta^2 = \mathrm{Var}[\widetilde{Y}] + \big(\beta - \mathbb{E}[\widetilde{Y}]\big)^2$$

ist minimal für die Wahl $\beta = \mathbb{E}[\widetilde{Y}]$.

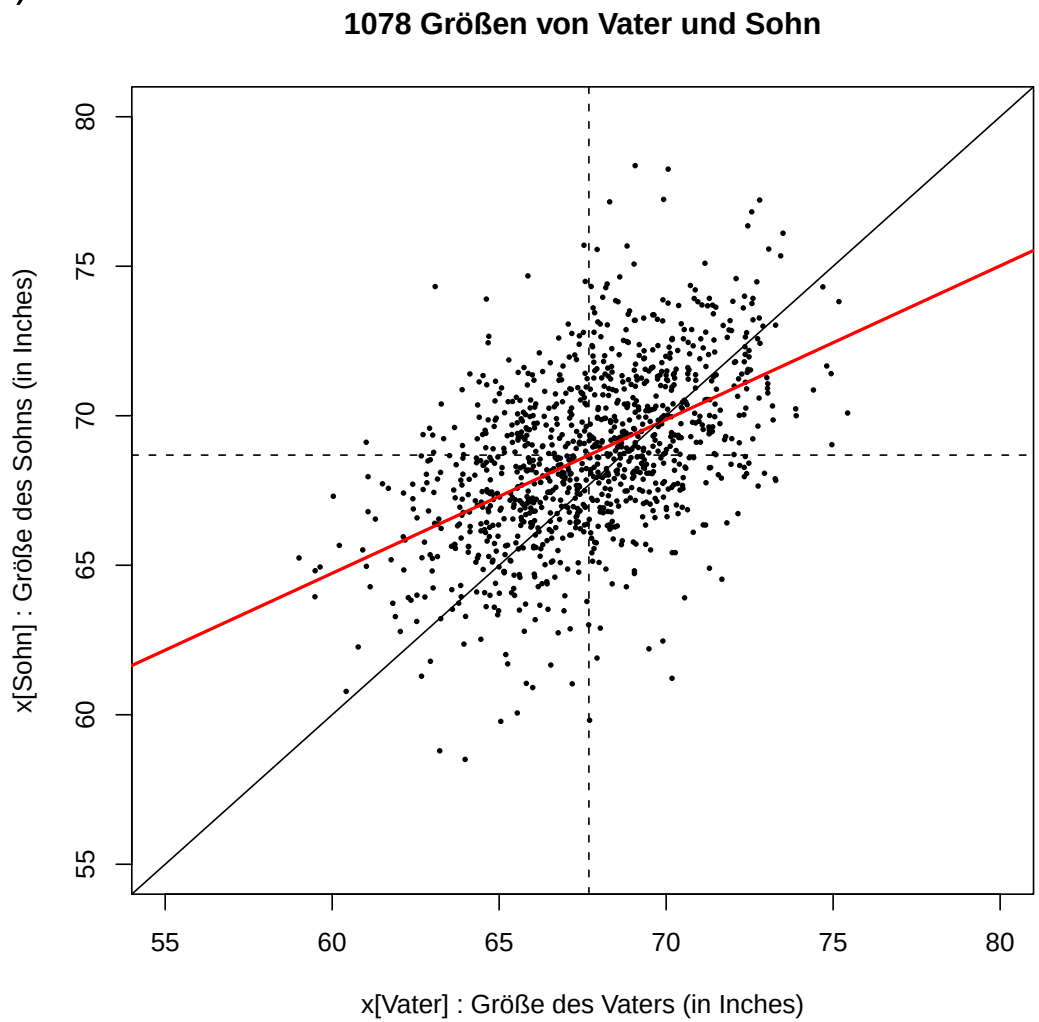
Übrigens: Wenn man zusätzlich annimmt, dass die ε_i u.i.v. $\sim \mathcal{N}_{0,\sigma^2}$ sind, so ist der kleinste-Quadrate-Schätzer hier auch zugleich der Maximum-Likelihood-Schätzer (mit einer Rechnung analog zum Beispiel für den Erwartungswert-ML-Schätzer).

Woher kommt der Name „Regression“ (nach lat. regressio, Zurückkommen)?



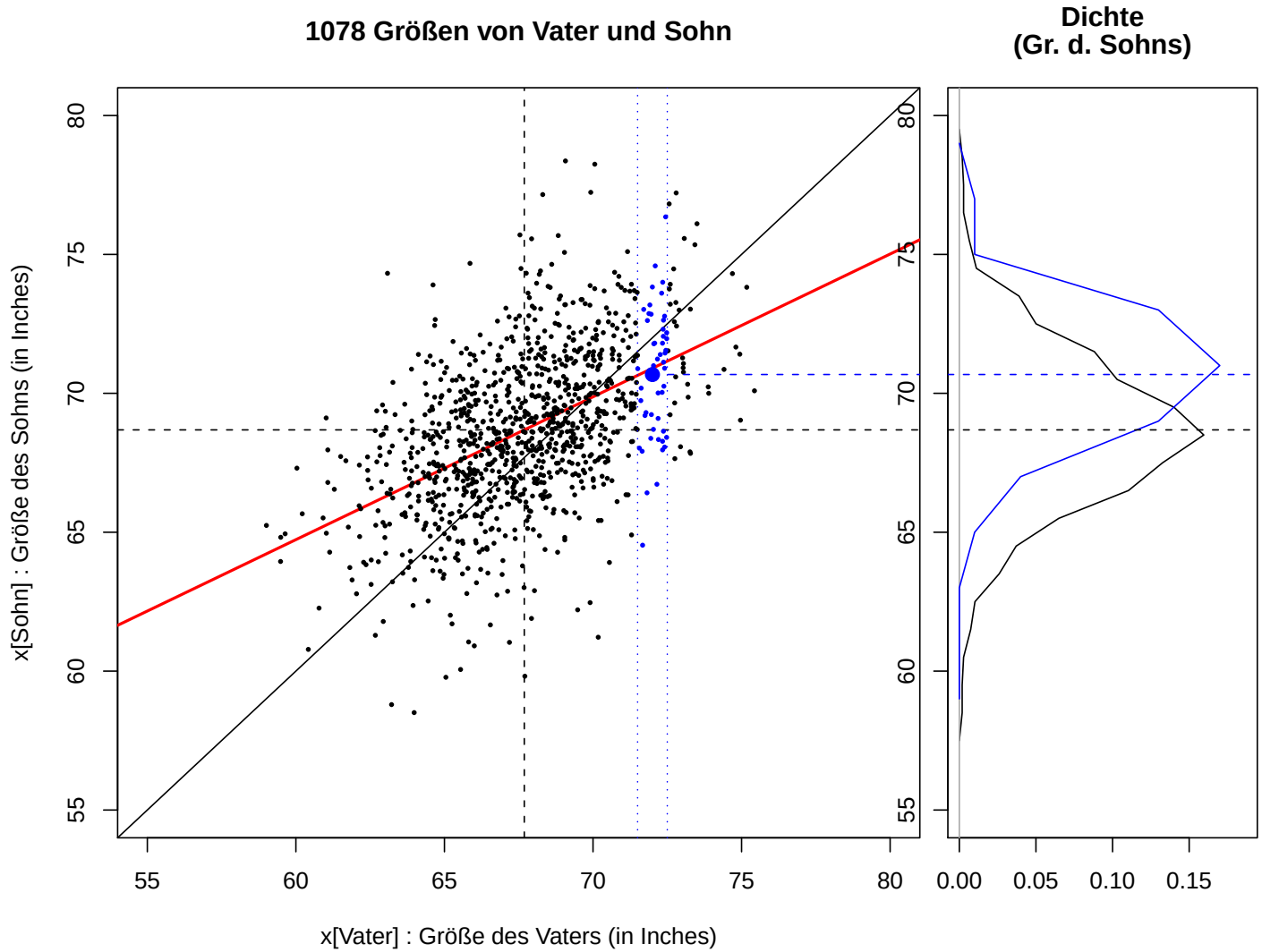
Francis Galton (1822–1911, engl. Wissenschaftler) hat angesichts biometrischer Beobachtungen den Ausdruck “regression towards the mean” geprägt.

Ein (relativ berühmter) Datensatz von Karl Pearson (1858-1936)

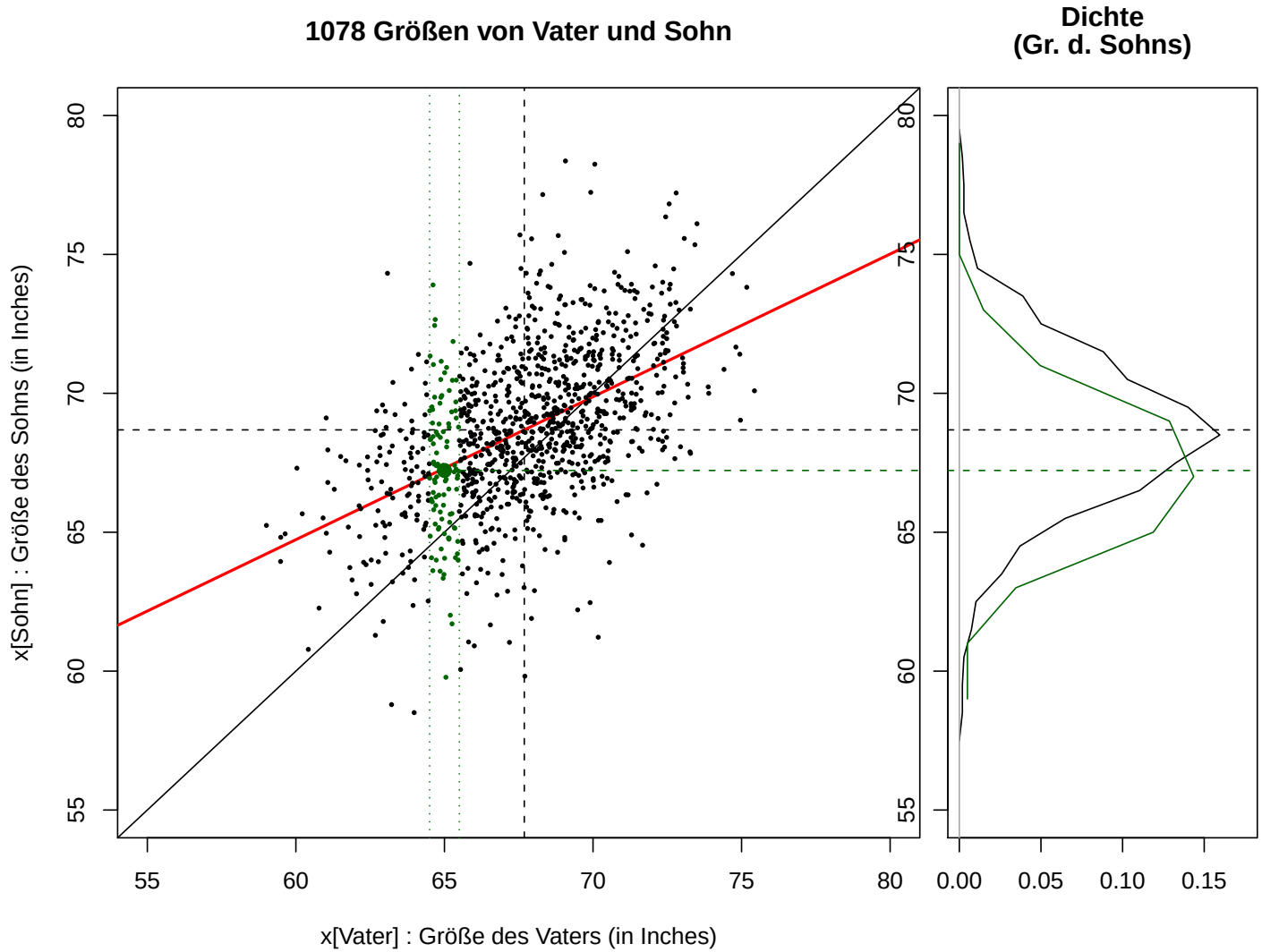


$\overline{x}_{\text{Vater}} = 67.7$, $\overline{x}_{\text{Sohn}} = 68.7$, $\sigma^2_{\text{Vater}} = 7.52$ ($\sigma_{\text{Vater}} = 2.74$, $\sigma_{\text{Sohn}} = 2.81$), $\text{cov}(x_{\text{Vater}}, x_{\text{Sohn}}) = 3.87$

(Korrelationskoeffizient $\kappa = \text{cov}(x_{\text{Vater}}, x_{\text{Sohn}})/(\sigma_{\text{Vater}}\,\sigma_{\text{Sohn}})) = 0.50$)

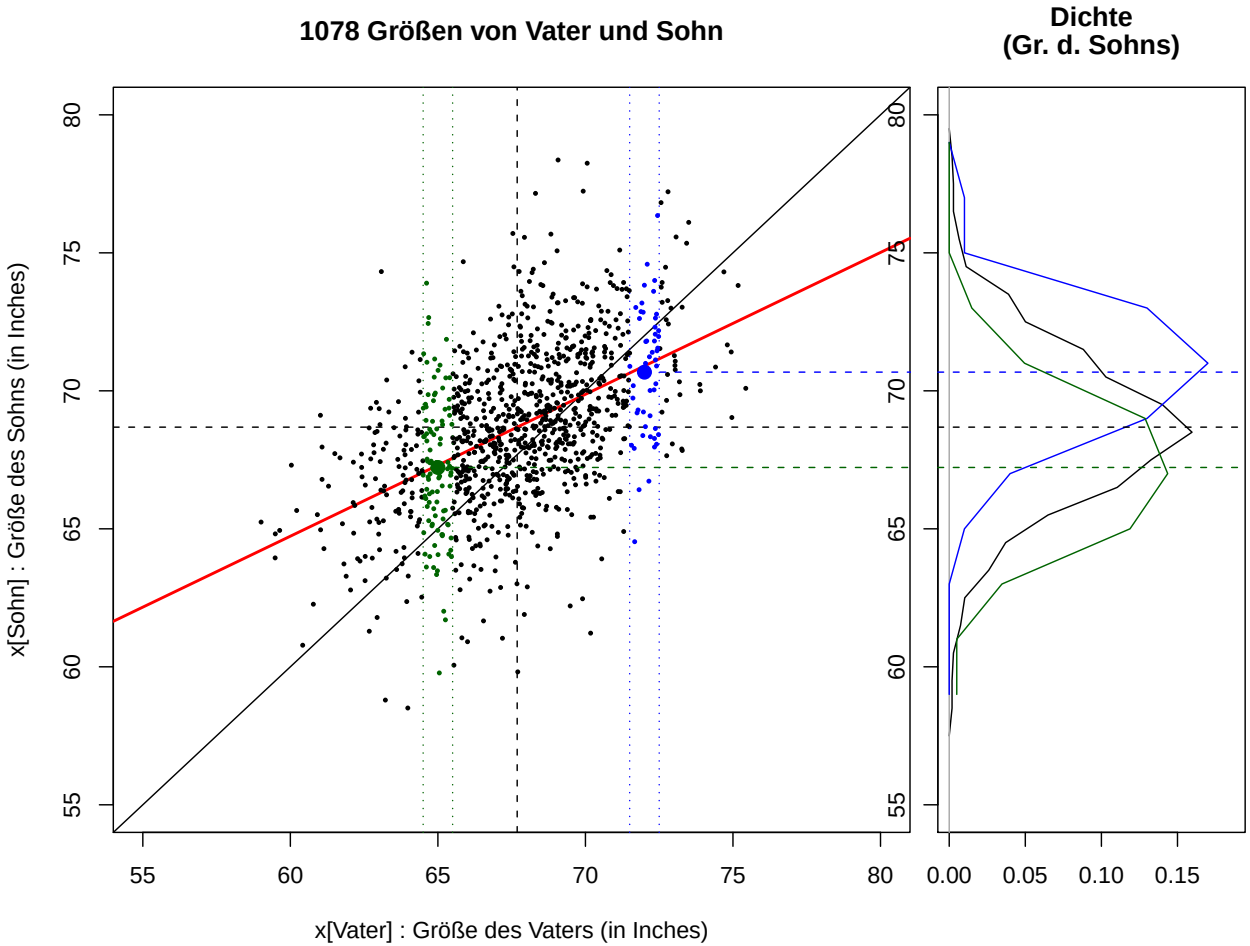


Betrachten wir die Söhne von überdurchschnittlich großen Vätern (z.B. Väter, die ca. 72 Inches groß sind):
Diese Söhne sind überdurchschnittlich groß (im Vergleich zu allen Söhnen), aber im Mittel kleiner als ihr Vater.



Betrachten wir andererseits die Söhne von unterdurchschnittlich großen Vätern (z.B. Väter, die ca. 65 Inches groß sind):
Diese Söhne sind unterdurchschnittlich groß (im Vergleich zu allen Söhnen), aber im Mittel größer als ihr Vater.

“Regression towards the mean”

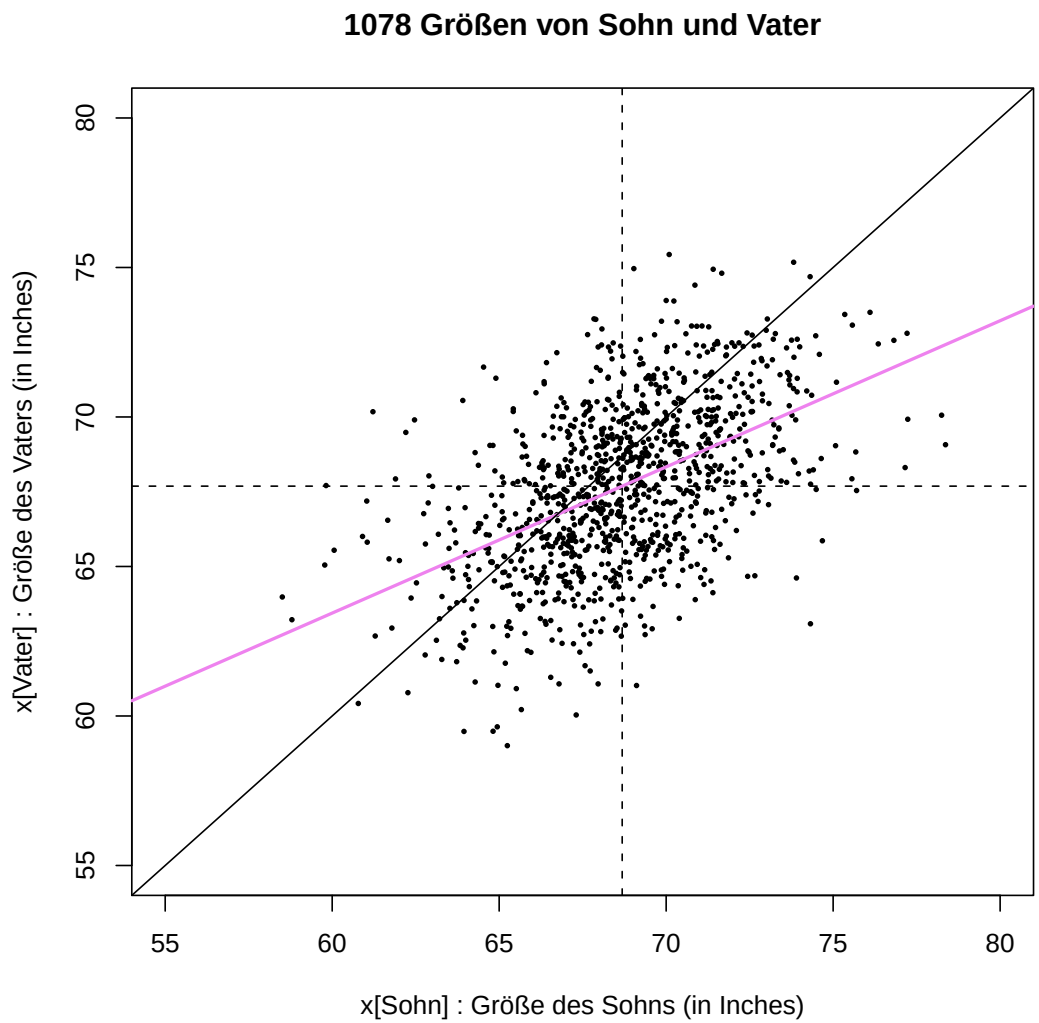


Wir sehen: Söhne überdurchschnittlich großer Väter sind im Mittel kleiner als ihr Vater (aber immer noch größer als der Populationsdurchschnitt), für Söhne unterdurchschnittlich großer Väter ist es umgekehrt: „Rückkehr zum Mittelwert“.

Bemerkung: Das beobachtete Phänomen der „Rückkehr zum Mittelwert“ bedeutet nicht notwendigerweise einen tieferen kausalen Zusammenhang, es tritt stets im Zusammenhang mit natürlicher Variabilität auf (technisch gesehen stets, wenn für den Korrelationkoeffizient κ gilt $|\kappa| < 1$).

Bestimmen wir (spaßeshalber) im Größen-Beispiel die Regressionsgerade für die Größe des Vaters als Funktion der Größe des Sohns:

Wir hatten $\overline{x}_{\text{Vater}} = 67.7$, $\overline{x}_{\text{Sohn}} = 68.7$,
$\text{cov}(x_{\text{Vater}}, x_{\text{Sohn}}) = 3.87$, $\sigma^2_{\text{Sohn}} = 7.92$
und finden die Regressionsgerade
$x_{\text{Vater}} = 34.1 + 0.489 x_{\text{Sohn}}$



Regressionsgerade: $x_{\text{Vater}} = 34.1 + 0.489 x_{\text{Sohn}}$