Kapitel 4

Schätzung

Lernziele

- Was ist ein Schätzer?
- Wie kann man Schätzer bewerten?
- Wie kann man Schätzer konstruieren?

4.1 Motivation

In der Statistik möchte man Kenngrößen der Grundgesamtheit aus Kenngrößen der Stichprobe ermitteln. Da die gezogene Stichprobe jedoch zufällig ist, kann der Wert in der Grundgesamtheit nicht exakt ermittelt werden. Daher verwendet man Verfahren, die diesen Wert so gut wie möglich annähern. Dabei heißt die Vorschrift, mit der aus der Stichprobe ein **Schätzwert** gemacht wird, ein **Schätzer**.

Beispiel 29 Allgemeiner Kontext: *Mit einem ungenauen Messverfahren wird ein Wert (z.B. ein Entstehungsalter oder die Lichtgeschwindigkeit) mehrfach bestimmt. Wir gehen davon aus, dass der Erwartungswert dem tatsächlich zu bestimmenden Wert entspricht. Ziel ist es nun den wahren Erwartungswert aus den Daten zu schätzen.*
Spezielles Beispiel: *1879 wollte Michelson die Lichtgeschwindigkeit bestimmten. Dazu machte er 5 Versuchsreihen. Wir werden die Dritte analysieren:*

```
> help(morley)

morley                package:datasets                R
Documentation

Michaelson-Morley
Speed of Light Data

Description:

     The classical data of Michaelson and Morley on the speed
     of light.  The data consists of five experiments, each
     consisting of 20 consecutive 'runs'.  The response is
     the speed of light measurement, suitably coded.
```

```
> help(morley)
> data(morley)
> morley$Speed <- morley$Speed + 299000
> dotchart(morley$Speed, groups = morley$Expt, main = "Michelsons Speed of Light Data")
```

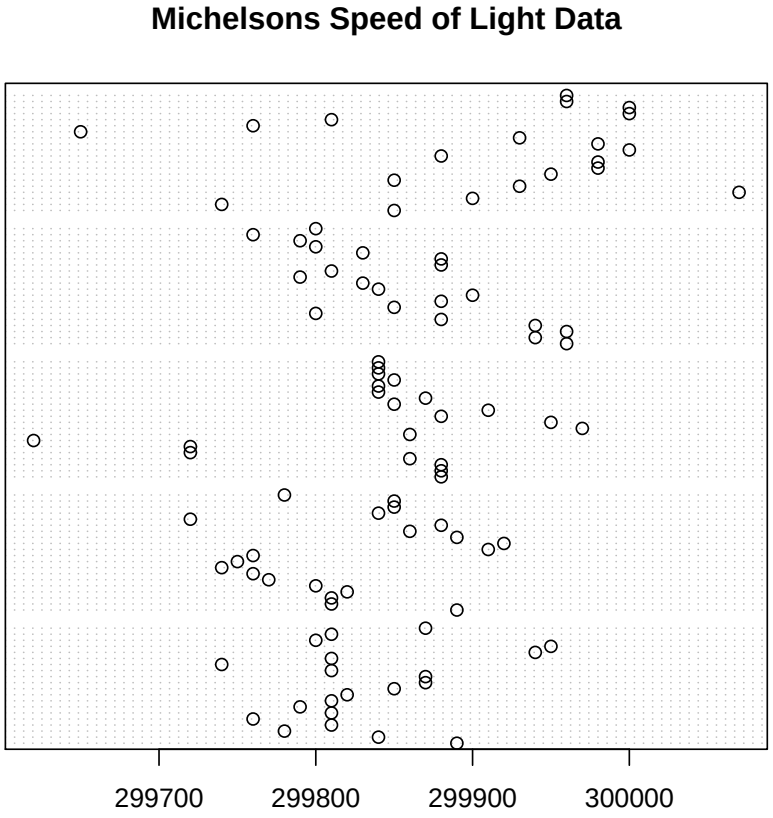

Abbildung 4.1: Ein Punktdiagramm des dritten Experiments von Michelson.

```
Usage:

     morley

Format:

     A data frame contains the following components:

     'Expt' The experiment number, from 1 to 5.

     'Run' The run number within each experiment.

     'Speed' Speed-of-light measurement.

Details:

     The data is here viewed as a randomized block experiment
     with 'experiment' and 'run' as the factors.  'run' may
     also be considered a quantitative variate to account for
     linear (or polynomial) changes in the measurement over
     the course of a single experiment.

Source:

     A. J. Weekes (1986) _A Genstat Primer_. London: Edward
     Arnold.

Examples:

     require(stats) morley£Expt <- factor(morley£Expt)
     morley£Run <- factor(morley£Run) attach(morley) plot(Expt,
     Speed, main = "Speed of Light Data", xlab = "Experiment
     No.")  fm <- aov(Speed ~ Run + Expt, data = morley)
     summary(fm) fm0 <- update(fm, . ~ . - Run) anova(fm0,
     fm) detach(morley)
```

```
> data(morley)
> morley$Speed <- morley$Speed + 299000
> dotchart(morley$Speed, groups = morley$Expt, main = "Michelsons Speed of Light Data")
> dritterVersuch <- morley$Speed[morley$Expt == 3]
> dritterVersuch

 [1] 299880 299880 299880 299860 299720 299720 299620 299860 299970
[10] 299950 299880 299910 299850 299870 299840 299840 299850 299840
[19] 299840 299840
```

Wir haben also nur eine Stichprobe $x_1, \ldots, x_n$, $n = 20$ von Geschwindigkeitsmessungen, die wir als Realisierung von Zufallsgrößen $X_1, \ldots, X_n$ der gleichen Verteilung auffassen. Obwohl jedes Mal die gleiche Größe (Lichtgeschwindigkeit in Luft) gemessen wird, unterscheiden sich die gemessen Werte. Diese Abweichung vom wahren Wert entsteht durch einen zufälligen experimentellen Fehler.

Wenn wir davon ausgehen, dass Michelson seinen Versuch so gestaltet hat, dass der Erwartungswert des gemessenen Wertes genau der wahren Lichtgeschwindigkeit entspricht, entspräche das Problem die Lichtgeschwindigkeit zu bestimmen, dem

Problem den Erwartungswert des Experiments zu bestimmen.
Wir nehmen also als Modell an:

$$E[X_i] = \textit{wahre Lichtgeschwindigkeit}$$

Die wahre Lichtgeschwindigkeit ist also ein Kenngröße oder ein Parameter der Verteilung der als Zufallsvariablen modellierten Beobachtungen.

Wir haben aber nur den Mittelwert der Stichprobe:

```
> mean(dritterVersuch)

[1] 299845
```

Diesen verwenden wir als **Schätzwert** *für die Lichtgeschwindigkeit. Er wird nach einer festen Vorschrift:*

$$\hat{\mu}(X_1, \ldots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

aus den Beobachtungen (Zufallsvariablen) X_i berechnet. $\hat{\mu}(X_1, \ldots, X_n)$ ist eine Abbildung von den Daten, bzw. den sie modellierenden Zufallsgrößen in den Raum der möglichen Werte für den Verteilungsparameter Erwartungswert μ bzw. der Lichtgeschwindigkeit. Eine solche Abbildung heißt, wenn wie messbar ist , ein **Schätzer**. *Ein guter Schätzer sollte möglichst genaue Schätzwerte liefern.*

Wie heute bekannt ist, beträgt die Lichtgeschwindigkeit in Luft [km/s]:

```
> Lichtgeschwindigkeit = 299734.5
```

und dieser Wert entspricht natürlich nicht der wahren Lichtgeschwindigkeit:

```
> mean(dritterVersuch) - Lichtgeschwindigkeit

[1] 110.5
```

Die Differenz aus Schätzer $\hat{\mu}(X_1, \ldots, X_n)$ und wahrem Wert μ:

$$F = \hat{\mu}(X_1, \ldots, X_n) - \mu$$

heißt **Schätzfehler**. *Der Schätzfehler selbst ist wieder eine Zufallsvariable, da ja eine andere Stichprobe einen anderen Schätzfehler produzieren würde:*

```
> fuenfterVersuch <- morley$Speed[morley$Expt == 5]
> fuenfterVersuch

 [1] 299890 299840 299780 299810 299760 299810 299790 299810 299820
[10] 299850 299870 299870 299810 299740 299810 299940 299950 299800
[19] 299810 299870

> mean(fuenfterVersuch) - Lichtgeschwindigkeit

[1] 97

> mean(morley$Speed[morley$Expt == 1])

[1] 299909

> mean(morley$Speed[morley$Expt == 2])
```

```
[1] 299856

> mean(morley$Speed[morley$Expt == 3])

[1] 299845

> mean(morley$Speed[morley$Expt == 4])

[1] 299820.5

> mean(morley$Speed[morley$Expt == 5])

[1] 299831.5
```

Wir können den Schätzfehler selbst nie berechnen, da der wahre Wert ja unbekannt ist. Als Zufallsgröße hat auch der Schätzfehler selbst wieder einen Erwartungswert und eine Varianz.

4.2 Grundbegriffe der Schätztheorie

4.2.1 Parameter der Verteilung

Der zentrale Unterschied zwischen der Statistik und der Wahrscheinlichkeitstheorie besteht in dem Übergang von dem Wahrscheinlichkeitsmodell, in dem die Wahrscheinlichkeitsverteilung P als vollständig bekannt vorausgesetzt wird, zu dem statistischen Modell in dem nur ein von einem unbekannten Parameter θ abhängiges Wahrscheinlichkeitsmodell P_θ vorausgesetzt wird.

Modell 1 (Allgemeines statistisches Modell) *Die Zufallsvariable X ist P_θ verteilt mit unbekanntem $\theta \in \Theta$.*

Alle später einzuführenden statistischen Modell sind Spezialfälle dieses Modells. Typischerweise ist $X = (X_1, \ldots, X_n)$ eine zusammengesetzte Zufallsgröße, die mehrere Beobachtungen umfasst.

Ziel der **Schätztheorie** ist es nun von den Beobachtungen auf gewisse Aspekte von P_θ zurückzuschließen, z.B. auf den wahren Wert von θ oder den Erwartungswert von X unter P_θ.

Interessierende Parameter könnten z.B. sein:

- Kenngrößen der Verteilung, wie z.B. Erwartungswert $E[X]$, Median $F_X^{-1}(0.5)$ oder Varianz $\mathrm{var}(X)$ der Verteilung
- der ganze Parameter θ
- eine beliebige Funktion des Parameter z.B. $g(\theta) = \theta_1$

Normalerweise hat der interessierende Parameter eine typische Bezeichnung (z.B. μ für Erwartungswert, oder θ für den Verteilungsparameter). Im allgemeinsten Fall bezeichnen wir den interessierenden Parameter als g (g steht für "goal") bzw. $g(\theta)$ bzw. $g(P_\theta)$ bzw. $g(F_\theta)$, um klar zu machen, dass der interessierende Parameter eine beliebige Funktion von θ bzw. der tatsächlichen Verteilung ist bzw. der tatsächlichen Verteilungsfunktion F_X ist.

Weil von den Daten auf einen Wert geschlossen werden soll, gehört die Schätztheorie (zusammen mit der Testtheorie und der Entscheidungstheorie) zur **schließenden Statistik**.

4.2.2 Schätzer

Definition 30 (Schätzer) *Eine Statistik $\hat{g}$ (also ein messbare Funktion der Daten), deren Aufgabe es ist als Schätzung für einen Parameter g zu dienen, heißt* **Schätzer** *für g.*

Diese Definition sagt also drei Sachen:

- Ein Schätzer ist eine Funktion der Daten.
- Ein Schätzer ist messbar.
- Ein Schätzer hat eine Aufgabe (aber es ist nicht gesagt, wie gut er diese Aufgabe erfüllt!). Man hofft sozusagen, dass der Wert in der Nähe des wahren g liegen möge.

Außerdem ist zu beachten:

- Typischerweise hat ein Schätzer $\hat{g}$ den gleichen Wertebereich, wie g selbst.
- Schätzer tragen oft die Bezeichnung des zu schätzenden Parameters und tragen zusätzliche einen Hut: z.B. $\hat{g}, \hat{\text{var}}(X), \widehat{F_X^{-1}(0.5)}$
- Der Schätzer $\hat{g} = \hat{g}(X_1, \ldots, X_n) = \hat{g}(X) = \hat{g}$ ist eine Zufallsvariable und $\hat{g} = \hat{g}(x_1, \ldots, x_n)$ seine Realisierung.

Beispiel 31 *Die X_i seinen unabhängig identisch $N(\mu, \sigma^2)$ normalverteilt. Wir interessieren uns für den Wert $g((\mu, \sigma^2)) = \mu = E[X]$. Ein sinnvoller Schätzer $\hat{g}$ ist durch den Mittelwert gegeben:*

$$\hat{g}(X_1, \ldots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Da in diesem Fall der Parameter eine klassische Bezeichnung μ hat, würden wir $\hat{\mu}$ statt $\hat{g}$ schreiben:

$$\hat{\mu}(X_1, \ldots, X_n) = \hat{g}(X_1, \ldots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Bemerkung 32 *Eine Statistik, die den Wert einer unbeobachteten Zufallsvariable (z.B. den von X_{n+1}) vorhersagt (also "schätzt") heißt Prädiktor. Das tritt z.B. in der Zeitreihenanalyse z.B. Vorhersage von Wetter oder Aktienkursen, Geostatistik (Vorhersage von reichen Lagerstätten oder Gefahrenschwerpunkten) und medizinische Statistik (Vorhersage von kritischen Zuständen) auf.*

4.3 Mathematische Schätzerbewertung

Es fällt auf, dass der Begriff des Schätzers selbst sehr blutleer ist. Ideal wäre natürlich, wenn der Schätzer immer oder zumindest mit hoher Wahrscheinlichkeit dem wahren Wert des Parameters entsprechen würde. Leider ist das so in den meisten Situationen gar nicht möglich, so dass man Kompromisslösungen, wie "weicht meistens nur geringfügig ab", "weicht im Mittel so wenig wie möglich ab", "konvergiert gegen den wahren Wert für immer mehr Daten" oder "trifft im Mittel den wahren Wert" verwendet, die dann mathematisch exakt gefasst werden. Dazu hat die mathematische Statistik eine Reihe von Standardeigenschaften für Schätzer entwickelt, die jeder Schätzer hat oder nicht hat.

4.3.1 Verteilung von Schätzern

Der Schätzer $\hat{g}$ ist eine Funktion der Zufallsvariable $X = (X_1, \ldots, X_n)$, welche die Daten modelliert. Da die Daten zufällig sind, sind es damit auch die Schätzer. Bei bekanntem Parametern θ ergibt sich die Verteilung $P^{\hat{g}(X)}$ des Schätzers, mittels des Transformationssatzes aus der Verteilung P_θ^X der Beobachtungen X.

Dabei hängt die Verteilung des Schätzers $\hat{\theta}$ natürlich von dem wahren Parameter θ ab.

4.3.1.1 Simulation der Verteilung eines Schätzers

Für ein gewähltes θ kann der Computer Beobachtungen x mit der Verteilung P_θ^X simulieren.

Beginnen wir mit einem einfachen Standardmodell der parametrischen Statistik:

Modell 2 (Ein-Stichproben-Normalverteilungsmodell mit unbekannter Varianz) *Die X_i, $i = 1, \ldots, n$ seien i.i.d. $N(\mu, \sigma^2)$ normalverteilt mit unbekanntem $\mu \in \mathbb{R}$ und unbekannter Varianz $\sigma^2 \in \mathbb{R}^+$.*

Dieses Modell kann mit dem Befehl:

```
> n = 10
> mu = 3
> sigmaQ = 4
> X = rnorm(n, mu, sqrt(sigmaQ))
> X
```

```
 [1]  2.995445762  4.073423116  2.304251505  3.072828143  2.311007671
 [6]  3.834906096 -0.183286300  4.130858841 -0.125316860  0.006933842
```

simuliert werden. Dazu wurden die Parameter konkret festgelegt.

Wie besprochen (aber noch nicht weiter begründet) liefert der Mittelwert $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ einen Schätzer für den Erwartungswert μ:

```
> mean(X)
```

```
[1] 2.242105
```

Dieser Schätzer liefert also nicht genau 3, obwohl der Datensatz perfekt den Verteilungsgesetzen folgt. Wir simulieren nun diesen Vorgang 100 mal:

```
> replicate(100, mean(rnorm(100, mu, sqrt(sigmaQ))))
```

```
  [1] 2.746309 2.700156 3.232249 2.864645 2.992114 3.187135 3.011086
  [8] 2.654078 3.172835 3.462208 3.211178 3.026951 2.897415 2.798771
 [15] 3.131405 3.278122 2.812690 2.834910 2.797557 3.142536 2.945812
 [22] 2.753959 3.034634 3.080851 3.359784 3.251175 3.153970 3.324136
 [29] 2.774884 3.097528 3.129411 3.407040 2.994403 2.984896 3.136341
 [36] 3.168745 2.824355 3.022520 3.042265 3.366779 2.907438 2.949979
 [43] 3.407356 3.056752 3.136154 3.040016 3.149410 2.865665 3.279350
 [50] 3.298326 2.731152 3.042625 3.201482 2.931755 2.838743 2.817752
 [57] 3.291305 3.233204 3.111327 3.011336 3.199549 3.098188 3.064599
 [64] 3.271385 3.085224 2.954253 3.189905 2.885367 2.919550 2.887635
 [71] 3.333374 2.804082 2.967273 2.984216 2.848184 3.310723 3.125810
 [78] 2.886446 3.023325 3.383897 3.475029 3.103023 2.713692 3.086286
 [85] 2.789608 2.960618 2.546671 3.178360 2.710358 2.883065 3.016367
 [92] 2.766271 2.576913 2.966277 3.026264 3.045630 3.011338 2.908848
 [99] 2.815349 3.025112
```

Offenbar schwankt der Schätzwert leicht um den Erwartungswert $\mu = 3$, ist allerdings nie sonderlich weit von 3 entfernt.

4.3.1.2 Berechnung der Verteilung eines Schätzers

Wir können die Verteilung dieses Schätzers $\hat{\mu}(X) = \bar{X}$ auch genau berechnen. Zunächst ergibt sich die Verteilung von X aus dem Modell als eine multivariate Normalverteilung:

$$X \sim \otimes_{i=1}^{n} N(\mu, \sigma^2) = N\left(\mu \mathbb{1}_n, \sigma^2 \mathsf{Id}_{n\times n}\right)$$

Der Schätzer ist eine Linearkombination dieser Werte:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} 1X_i = \underbrace{\frac{1}{n}\mathbb{1}_n^t}_{\mathbf{c}^t} X = \mathbf{c}^t X$$

und damit nach dem Transformationssatz der Normalverteilung:

$$\hat{\mu} = \mathbf{c}^t X \sim N(\mathbf{c}^t \mu \mathbb{1}_n, \mathbf{c}^t \sigma^2 \mathsf{Id}_{n\times n}\mathbf{c}) = N\left(\mu, \frac{1}{n}\sigma^2\right)$$

Der Schätzer ist also wieder normalverteilt. Dabei entspricht sein Erwartungswert dem zu schätzenden Wert μ und seine Varianz ist deutlich kleiner, als die der X_i und nimmt proportional zu $\frac{1}{n}$ ab. Diese Varianzabnahme proportional zu $\frac{1}{n}$ ist ein typisches Verhalten vieler guter Schätzer und trägt den Namen $\frac{1}{n}$**-Gesetz**.

Für vorgegebene Parameter lässt sich die Verteilung eine Schätzers mit Hilfe eines Transformationssatzes aus der Verteilung der Daten berechnen.

4.3.2 Qualitative Kriterien

Die Schätztheorie hat eine ganze Reihe von Eigenschaften definiert, die ein Schätzer haben kann. Diese Eigenschaften tauchen in der Beschreibung von Schätzern immer wieder auf und werden daher hier kurz eingeführt.

4.3.2.1 Unverzerrtheit als Eigenschaft eines Schätzers

Die Eigenschaft, dass der Erwartungswert des Schätzers dem wahren Parameter entspricht, ist eine populäre Eigenschaft. Sie hat deshalb auch einen eigenen Namen:

Definition 33 *Ein Schätzer $\hat{g}$ in Modell 1 heißt* **unverzerrt** *oder* **erwartungstreu** *(engl.* **unbiased***) für eine reelle Funktion $g(\theta)$ der Parameter, wenn für alle $\theta \in \Theta$ gilt:*

$$E_{P_\theta}[\hat{g}(X)] = g(\theta)$$

Dabei bezeichnet $E_{P_\theta}[\ldots]$ den Erwartungswert unter der Verteilung P_θ.

In unserem Fall war die Funktion g die Abbildung des Parameters (μ, σ^2) auf die erste Komponente μ.

Ist der Schätzer nicht unverzerrt, so spricht man auch von einem **verzerrten** Schätzer. Die Abweichung

$$bias(\theta) := E_{P_\theta}[\hat{g}(X)] - g(\theta)$$

heißt auch **Verzerrung** oder **Bias** von $\hat{g}$.

Beispiel 34 *Der Mittelwert ist in Modell 2 ein unverzerrter Schätzer für den Erwartungswert $g(\theta) = E_{P_\theta}[X_1]$:*

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &\quad \textit{identische Verteilung:} \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_1] \\ &= \frac{1}{n} n E[X_1] \\ &= E[X_1] \end{aligned}$$

Früher wurde die Unverzerrtheit eines Schätzers als eines der wichtigsten Qualitätsmerkmale angesehen. Es hat einige interessante Vorteile:

- Es vermittelt einem das Gefühl, das der Schätzer einen nicht, im Mittel, in einer Richtung betrügt.
- Erwartungstreue ist äquivalent damit, dass der Mittelwert über eine steigende Anzahl von unabhängigen Schätzung mit diesem Schätzer nach dem starken Gesetz der großen Zahlen mit Wahrscheinlichkeit 1 schließlich gegen den wahren Parameter g konvergiert. Das ist insbesondere dann interessant, wenn das Verfahren oft angewendet wird.
- Für erwartungstreu Schätzung gibt es eine geschlossene und leistungsfähige Theorie, die anwendbar ist, sobald überhaupt ein erwartungstreuer Schätzer für g existiert. Funktionale $g(P)$ zu denen eine erwartungstreuer Schätzer existiert, heißen **erwartungstreu schätzbar** .

In neuerer Zeit hat das Kriterium der Erwartungstreue an Ansehen verloren, weil:

- Es leistungsfähigere Kriterien und Methoden gibt (z.B. Minimax-Schätzung, Bayes-Schätzung)
- Der beste erwartungstreue Schätzer oft nicht robust gegenüber Ausreißern in den Daten oder Modellfehlern ist. Und selbst bei wenigen Datenfehler drastisch falsche Werte liefern kann.
- Es oft objektiv bessere Schätzer gibt (z.B. Stein-Schätzer).

Dennoch stellt dieses Kriterium immer noch einen wichtigen Grundpfeiler der Statistik dar.

4.3.2.2 Linearität

Linearität: Ein Schätzer heißt linear, wenn er eine lineare Funktion der Daten ist:

$$\hat{g} = \sum_{i=1}^{n} w_i X_i,$$

für Gewichte $w_i \in \mathbb{R}^d$. Der Vorteil linearer Schätzer ist ihre Einfachheit. Ein Beispiel für einen linearen Schätzer ist der Mittelwert. Weitere Beispiele werden wir bei den linearen Modellen kennenlernen.

4.3.2.3 Schwache Konsistenz

Ein Familie von Schätzern $\hat{g}_n(X_1, \ldots, X_n)$, $n = 1, \ldots, \infty$ heißt schwach konsistent für g, wenn:

$$\forall\theta \in \Theta : \forall\epsilon > 0 : \lim_{n\to\infty} P_\theta(\hat{g}(X_1, \ldots, X_n) \notin B(g(\theta), \epsilon)) = 0$$

wobei $B(g(\theta), \epsilon)$ den Ball mit Radius ϵ um $g(\theta)$ bezeichnet. Der Vorteil konsistenter Schätzer ist, dass man wenigstens bei großen Stichprobenumfängen sicher ist, nahe am wahren Wert zu sein. Leider trifft die Konsistenz hier keine Aussage darüber, was "groß" ist. Existiert die Varianz, so ist der Mittelwert nach dem schwachen Gesetz der großen Zahlen ein konsistenter Schätzer für den Erwartungswert. Weiterhin folgt für stetige Funktionen $h(g)$, dass $\hat{h} = h(\hat{g})$ ein schwach konsistenter Schätzer für $h(g)$ ist.

4.3.2.4 Starke Konsistenz

Ein Familie von Schätzern $\hat{g}_n(X_1, \ldots, X_n)$, $n = 1, \ldots, \infty$ heißt stark konsistent für g, wenn:

$$\forall\theta \in \Theta : \forall\epsilon > 0 : P(\{\omega : \lim_{n\to\infty} \hat{g}(X_1, \ldots, X_n) = g(\theta)\}) = 1$$

also wenn mit Wahrscheinlichkeit 1 die Folge der Schätzer, die auf immer mehr Daten basieren, gegen den wahren Wert g konvergiert. Der Vorteil stark konsistenter Schätzer ist, dass man zumindest mit steigendem Stichprobenumfang dem wahren Wert sehr nahe kommt. Der Mittelwert ist nach dem starken Gesetz der großen Zahlen ein stark konsistenter Schätzer für den Erwartungswert. Starke Konsistenz impliziert schwache Konsistenz. Weiterhin folgt für stetige Funktionen $h(g)$, dass $\hat{h} = h(\hat{g})$ ein stark konsistenter Schätzer für $h(g)$ ist.

4.3.2.5 Asymptotische Normalität

Ein Schätzer $\hat{g}$ heißt **asymptotisch normal**, wenn für alle θ um den wahren Parameter die Verteilungsfunktion des zentrierten und mit seiner Standardabweichung normierte Schätzer $(\hat{g} - g)/sd(\hat{g})$ bei steigenden Stichprobenumfang gegen die Verteilungsfunktion einer Standardnormalverteilung konvergiert.

4.3.2.6 Robustheit gegen Ausreißer

Ein Schätzer heißt robust (gegen Ausreißer) mit Bruchpunkt p, wenn

$$\sup_{\vec{y}\in M_p(\vec{x})} |\hat{g}(y_1, \ldots, y_n) - \hat{g}(x_1, \ldots, x_n)| < \infty$$

wobei

$$M_p(\vec{x}) := \{(y_i)_{i=1,\ldots,n} : \text{weniger als } np \text{ viele } y\_i \text{ sind ungleich } x_i\}$$

Der Vorteil robuster Schätzer ist, dass die Schätzung durch eine kleinen Anteil falscher Werte nicht beliebig falsch wird. Der Stichproben Mittelwert ist kein robuster Schätzer, da schon ein einzelne extreme Beobachtung ihn beliebig verändern kann. Hingegen ist der Median der Stichprobe ein robuster Schätzer (für was auch immer), da bis zu unter 50% der Daten beliebig verändert werden können, ohne dass der Median den Bereich der unveränderten Daten verlässt.

4.3.2.7 Robustheit gegen Modellmissspezifikation

Dieser zweite Robustheitsbegriff ist im allgemeinen meist weniger genau gefasst, stellt aber ein zentrales Problem dar, da die genaue Verteilungsform ja so gut wie nie bekannt ist.

Der Mittelwert ist robust gegenüber Modellmissspezifikationen, da er in jedem Modell ein erwartungstreuer Schätzer für den Erwartungswert ist.

4.3.3 Beurteilung mittels Verlustfunktion

4.3.3.1 Die Verlustfunktion

Eine exaktere Methode zur Beurteilung von Schätzern bietet die am leichtesten betriebswirtschaftlich zu motivierende Verlustfunktion:

Definition 35 *Eine $V(\theta, \hat{\theta})$ ist eine Funktion, die den (z.B. finanziellen) Verlust angibt, den man dadurch erleidet, dass man den Wert $\hat{\theta}$ geschätzt hat, anstelle des wahren Wertes θ.*

Beispiel 36 *Wir wollen wissen, ob der Patient Magenkrebs hat oder nicht. Der Parameterbereich ist also $\Theta = \{Krank, Gesund\}$. Je nachdem, wie wir uns entscheiden entstehen Kosten:*

	Diagnose: Krank	*Diagnose: Gesund*
Ist gesund	*Unnötige lebensgefährliche Chemotherapie*	*Alles gut*
Ist Krank	*Patient überlebt durch Chemotherapie*	*stirbt*

Beispiel 37 *Das Gesamtmasse θ an für den Winter benötigten Futtermais soll geschätzt werden. Überschätzt man das um Δ, so zahlt man den $\Delta \cdot p_1$ mehr Mais, als man eigentlich braucht. Dabei ist p_h der billige Preis für Mais in der Erntezeit. Unterschätzt man jedoch die benötigte Menge, so muss die entsprechende Menge im Winter zu den teuren Winterpreisen p_w nachgekauft werden.*
Insgesamt erhält man also die Verlustfunktion:

$$V(\theta, \hat{\theta}) = \begin{cases} (\hat{\theta} - \theta)p_h, wenn\ \hat{\theta} > \theta \\ (\theta - \hat{\theta})p_w, wenn\ \hat{\theta} <= \theta \end{cases}$$

Die folgenden Verlustfunktionen werden oft als Standard benutzt, wenn keine andere Verlustfunktion durch die Anwendung vorgegeben ist:

Definition 38 (Quadratischer und absoluter Verlust) *Die* **quadratische Verlustfunktion** *(**S**quared **E**rror) ist gegeben durch:*

$$SE(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$$

Die **absolute Verlustfunktion** *(**S**quared **E**rror) ist gegeben durch:*

$$AE(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$$

Insbesondere die quadratische Verlustfunktion ist, weil sie mathematisch am leichtesten zu behandeln ist auch die populärste.

4.3.3.2 Erwarteter Verlust

Wird nun ein spezieller Schätzer verwendet, um die θ zu schätzen, so ist der tatsächlich erlittene Verlust natürlich zufällig. Deshalb geht man zu Erwartungswert über:

Definition 39 (Erwarteter Verlust) *Zu einer Verlustfunktion $V(\theta, \hat{\theta})$ und einem Schätzer $\hat{\theta}$ für θ heißt:*

$$EV_{V,\hat{\theta}}(\theta) := E_{P_\theta}[V(\theta, \hat{\theta}(X))]$$

der **e***rwartete* **V***erlust von $\hat{\theta}$.*

Definition 40 (Quadratischer und absoluter Verlust) *Der zur quadratische Verlustfunktion gehörende* **quadratische Verlust** *(***M***ean* **S***quared* **E***rror) ist gegeben durch:*

$$MSE(\theta) = E_{P_\theta}\left[\|\theta - \hat{\theta}(X)\|^2\right]$$

Der zur absoluten Verlustfunktion gehörende **absolute Verlust** *(***M***ean* **A***bsolute* **E***rror) ist gegeben durch:*

$$MAE(\theta) = E_{P_\theta}\left[\|\theta - \hat{\theta}(X)\|^2\right]$$

Der folgende berühmte Satz zeigt, dass sich der quadratische Verlust aus einem Teil für die Verzerrung und einen Teil für die Streuung des Schätzers zusammensetzt:

Satz 41 (Biastheorem)

$$MSE(\theta) = \|bias(\theta)\|^2 + \mathrm{var}_{P_\theta}(\hat{\theta}(X))$$

Dabei bezeichnet der untere Index an der Varianz, bezüglich welcher Verteilung sie zu berechnen ist.

Beweis: Aus der Wahrscheinlichkeitstheorie:

$$\mathrm{var}_{P_\theta}(\hat{\theta}(X)) = \mathrm{var}_{P_\theta}(\hat{\theta}(X) - \theta) = \underbrace{E_{P_\theta}[\|\hat{\theta} - \theta\|^2]}_{MSE} - \|\underbrace{E[\theta - \hat{\theta}]}_{bias}\|^2$$

□

Insbesondere ist also für unverzerrte Schätzer:

$$MSE(\theta) = \mathrm{var}_{P_\theta}(\hat{\theta}(X))$$

Bei unverzerrten Schätzer kann man also den erwarteten Verlust und die Varianz austauschbar benutzen.

Beispiel 42 *Für Modell 2 haben wird die Verteilung des Mittelwertes $N(\mu, \frac{1}{n}\sigma^2)$ bereits ausgerechnet. Die Varianz des Schätzers $\bar{X}$ beträgt also $\frac{1}{n}\sigma^2$ und damit, da der Schätzer nach Beispiel 34 unverzerrt ist, auch der MSE:*

$$MSE(\theta) = \frac{1}{n}\sigma^2$$

Der MSE ist in diesem Modell also konstant, aber zunächst unbekannt.

Oft wird anstelle des MSE seine Wurzel der Root Mean squared error $RMSE = \sqrt{MSE}$, auch **Standardschätzfehler** genannt, angegeben. Die Interpretation des $RMSE$ entspricht in etwa der Standardabweichung.

4.3.3.3 Verlustbasierte Gütekriterien

Definition 43 *Ein Schätzer $\hat{g}_1$ für $g(\theta)$ heißt besser, als ein Schätzer $\hat{g}_2$ bezüglich des Verlustes V, wenn für alle $\theta \in \Theta$ gilt:*

$$EV_{V,\hat{g}_1} \leq EV_{V,\hat{g}_2}(\theta)$$

und für mindestens ein θ Ungleichheit gilt.

Beispiel 44 *Nehmen wir einmal an wir wissen $n = 5$, dass $\sigma^2 = 1$ und dass $|\mu| < 1$. Jetzt betrachten wir statt des Mittelwerts $\hat{\mu}(X) = \bar{X}$ einen alternativen Schätzer $\hat{\mu}'(X) = 0.9\bar{X}$, dann gilt:*

$$EV_{V,\hat{\mu}} = \frac{1}{n} = 0.2$$

und

$$EV_{V,\hat{\mu}'} = E[\|0.9\bar{X} - X\|^2] = \underbrace{bias_{\hat{\mu}_2}((\mu,\sigma))}_{-0.1\mu}{}^2 + \mathrm{var}_{N(\mu,\sigma^2)}(\hat{\mu}_2) = 0.01\mu^2 + 0.9^2\frac{1}{n} \leq 0.172$$

In dieser Situation ist also $\hat{\mu}'$ ein gleichmäßig besserer Schätzer für den Erwartungswert $\mu = E[X_1]$ als der Mittelwert der Daten.

Das Gefühl, dass man doch immer einen möglichst guten Schätzer verwenden möchte, führte zu der Definition, dass jeder Schätzer der noch verbessert werden kann, ausgeschlossen sein sollte.

Definition 45 (Zulässigkeit) *Ein Schätzer heißt* **zulässig** *wenn es keine besseren Schätzer gibt.*

Leider ist es, insbesondere im Falle mehrdimensionaler Parameter, aber sehr schwierig zulässige Schätzer zu konstruieren, so dass fast alle praktisch verwendeten Schätzer zumindest oft nach dieser Definition unzulässig sind. Die Zulässigkeit wird daher als Kriterium praktisch nicht verwendet.

Definition 46 (Minimax-Schätzer) *Ein Schätzer heißt Minimax-Schätzer, wenn es keinen Schätzer gibt, der für diesen Parameter im gleichen Modell einen kleineren maximalen Verlust*

$$\sup_{\theta\in\Theta} EV_{V,\hat{\theta}}(\theta)$$

hat.

Minimax-Schätzer sind häufig schwierig direkt zu konstruieren, da man dazu eine Optimierungsaufgabe lösen muss, bei der alle messbaren Funktionen zur Wahl stehen. Oft werden Minimax-Schätzer gefunden, indem ein anderes Kriterium zur Konstruktion des Schätzers verwendet wird und seine Minimax-Eigenschaft später nachgewiesen wird.

4.3.3.4 Asymptotische verlustbasierte Gütekriterien

Ähnlich der Konsistenz kann man auch quantitative Gütekriterien asymptotisch beschreiben.

Definition 47 (Asymptotische relative Effizienz) *Eine Familie $\hat{g}_n(X_1,\ldots,X_n)$ von Schätzern für g heißt* **asymptotisch relativ effizient** *bezüglich einer anderen Familien $\hat{g}'_n(X_1,\ldots,X_n)$, wenn*

$$\forall\theta \in \Theta : \limsup_{n\to\infty} \frac{EV_{\hat{g}_n}}{EV_{\hat{g}'_n}} \leq 1$$

Definition 48 (Asymptotische Effizienz) *Eine Familie $\hat{g}_n(X_1, \ldots, X_n)$ von Schätzern für g heißt* **asymptotisch relativ** *bezüglich einer anderen Familien $\hat{g}'_n(X_1, \ldots, X_n)$, wenn sie bezüglich aller denkbarer Familien von Schätzern asymptotisch relativ effizient ist.*

4.4 Konstruktion von Schätzern

Das Problem der Schätzung tritt in der Statistik in immer neuen Konstellationen auf. Deshalb hat die mathematische Statistik eine Reihe von Prinzipien entwickelt, die es einem erlauben für große Klassen von verschiedenen Situationen mehr oder weniger leistungsfähige Schätzer zu erzeugen. Die nach den verschiedenen Prinzipien erzeugten Schätzer haben oft typische Eigenschaften, die man für manche Klassen von Situationen zeigen kann. Die Schätzprinzipien bauen z.T. aufeinander auf und werden daher vom einfachen zum komplizierteren besprochen.

4.4.1 Mittelwert

Das einfachste Schätzprinzip ist das des Mittelwerts. Es funktioniert, wenn aus einer unabhängig identisch verteilten Stichprobe der Erwartungswert $E[T(X)]$ einer beliebigen transformierten Größe $T(X)$ geschätzt werden soll.

- **Name:** Mittelwertsschätzer $\bar{T}$
- **Anwendungsbereich**: i.i.d. Stichprobe $X_i,\ i = 1, \ldots, n$, $g = E[T(X)]$
- **Schätzer**: $\hat{g} = \frac{1}{n} \sum_{i=1}^{n} T(X_i)$
- **Eigenschaften:**
  - Erwartungstreu wg. $E[\hat{g}] = \frac{1}{n} \sum_{i=1}^{n} E[T(X_i)] = E[T(X)]$
  - Stark konsistent wg. starkem Gesetz der großen Zahlen
- **Übliche Voraussetzungen:** Existenz und Konstanz des Erwartungswertes, stochastische Unabhängigkeit für Konsistenz.

Beispiel 49 Mittelwertschätzer

- *Schätzung des Erwartungswertes von X bei beliebigen Verteilungsmodellen.*

```
> mean(morley$Speed)

[1] 299852.4
```

- *Schätzung der Wahrscheinlichkeit, mit der ein Ereignis $X \in A$ auftritt: Setzte dazu $T(X) = 1_A(X)$*

```
> mean(morley$Speed > 3e+05)

[1] 0.01
```

Das Mittelwertsprinzip kann man also insbesondere auch anwenden, um Wahrscheinlichkeiten zu schätzen.

4.4.2 Empirische Verteilungsfunktion

Jeder Punkt der Verteilungsfunktion stellt eine Wahrscheinlichkeit dar und die Verteilungsfunktion beschreibt die Verteilung vollständig. Wir wenden also das Prinzip an, um die "ganze Verteilung" zu schätzen:

- **Name:** empirische Verteilungsfunktion $\hat{F}(x)$
- **Anwendungsbereich**: reelle i.i.d. Stichprobe $X_i,\ i = 1, \ldots, n,\ g = F_X$
- **Schätzer**: $\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x]}(X_i)$
- **Eigenschaften:**
  - Erwartungstreu wg. $E[\hat{F}(X)] = F_X(x)$
  - Stark konsistent bezüglich der Supremumsnorm
    (Beweisskizze: Man zeigt Konvergenz mit Wahrscheinlichkeit 1 auf einer Dichten aber abzählbaren Menge von Punkten x mittels des vorherigen Resultats und schätzt die Differenz an Zwischenstellen mittels der Isotonie gegenüber der Differenz an den betrachteten Stellen ab.) Die empirische Verteilungsfunktion konvergiert also mit Wahrscheinlichkeit 1 punktweise gleichmäßig gegen die wahre Verteilungsfunktion.
- **Übliche Voraussetzungen:** Stochastische Unabhängigkeit für Konsistenz.

4.4.3 V-Schätzer

Wenn nun die empirische Verteilungsfunktion stark konsistent schätzbar ist, so natürlich auch jedes ihrer stetigen Funktionale $g(F_X)$. Diese Idee verwenden die sogenanntes V-Schätzer (V-steht hier für Verteilungs-...)

- **Name:** V-Schätzer
- **Anwendungsbereich**: reelle i.i.d. Stichprobe $X_i,\ i = 1, \ldots, n,\ g = g(F_X)$ stetig
- **Schätzer**: $\hat{g} = g(\hat{F})$
- **Eigenschaften:**
  - Stark konsistent
    (Beweisskizze: Ist g ein stetiges Funktional, so konvergiert $g(\hat{F})$ mit $\hat{F}$.)
- **Übliche Voraussetzungen:** g ist eine im Sinne der Supremumsnorm stetiges Funktional auf F_X .

Beispiel 50 V-Schätzer

- *Der Odd bei einer Bernouliverteilung B(p): ist gegeben durch $g(p) = \frac{p}{1-p}$, also durch das Verhältnis der beiden Wahrscheinlichkeiten: Wieviel wahrscheinlicher ist es zu gewinnen als zu verlieren. Offenbar lässt sich g durch*

$$g(F) = \frac{1 - F(0)}{F(0)}$$

ausdrücken und daher ist mit $\hat{F}_X(0) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,0]} X_i = n - N$ bei N Erfolgen auch

$$\hat{g} = \frac{1 - \hat{F}(0)}{\hat{F}(0)} = \frac{N}{n - N}$$

ein stark konsistenter Schätzer für den Odd g.

```
> plot(ecdf(morley$Speed))
```

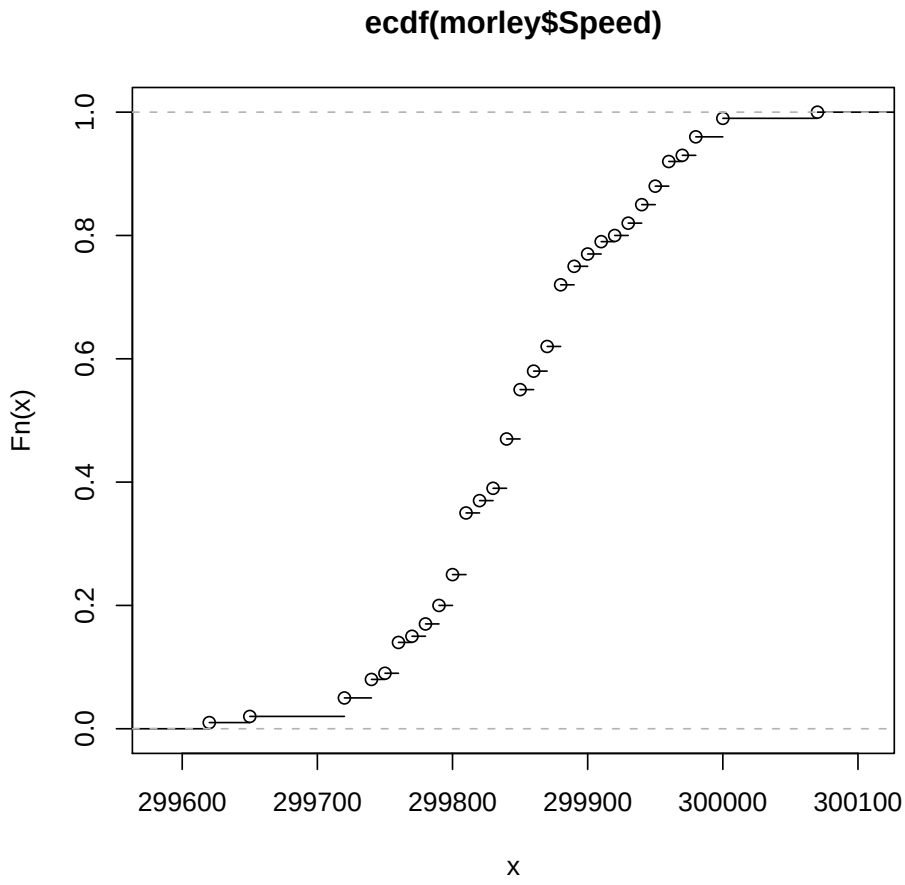

Abbildung 4.2: Empirische Verteilungsfunktion

- *$g = \mathrm{var}(X)$ lässt sich als: $g(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$ schreiben, wobei das Lebesgue–Stiltjesintegral $\int T(x)dF(x)$ das Integral $\int x dP_F(x)$ über die zur Funktion F gehörige Verteilung P_F bezeichnet.*

  Der V-Schätzer ergibt sich also als:

$$\begin{aligned}
\hat{g} &= \int x^2 dP_{\hat{F}}(x) - \left(\int x dP_{\hat{F}}(x)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2
\end{aligned}$$

4.4.4 Vorbereitung auf den U-Schätzer

Stellen wir einmal fest, ob dieser Varianz-Schätzer erwartungstreu ist. Dazu verwenden wir die Substitution $Y_i = X_i - E[X]$, was $E[Y] = 0$ und $E[Y^2] = \mathrm{var}(X)$ impliziert:

$$\begin{aligned}
E[\hat{g}] &= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \bar{X})^2] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[(Y_i - \bar{Y})^2] \\
&= \frac{1}{n}\sum_{i=1}^{n} (E[Y_i^2] - 2E[Y_i\bar{Y}] + E[\bar{Y}^2]) \\
&= \frac{1}{n}\sum_{i=1}^{n} (\mathrm{var}(X) - 2\frac{1}{n}E[Y_iY_i] - 0 + \frac{1}{n}\mathrm{var}(X)) \\
&= \frac{1}{n}\sum_{i=1}^{n} (\mathrm{var}(X) - 2\frac{1}{n}\mathrm{var}(X) + \frac{1}{n}\mathrm{var}(X)) \\
&= \frac{1}{n} n(1 - \frac{1}{n})\mathrm{var}(X) \\
&= \frac{n-1}{n}\mathrm{var}(X)
\end{aligned}$$

Es wird als wesentliches Problem der V-Schätzer angesehen, dass sie in solchen einfachen Situationen nicht erwartungstreu sind.

Das Problem entsteht offenbar durch die Terme in denen die gleichen X_i aus verschiedenen Integralen aufeinandertreffen.

4.4.5 U-Schätzer

Hier kommen die U-Schätzer (U für Unverzerrt) ins Spiel, die sich zum Schätzen kombinierter Integrale über die Verteilungsfunktion eignen. U-Schätzer vermeiden dieses Aufeinandertreffen explizit und lassen die entsprechenden Terme einfach weg:

- **Name:** U-Schätzer
- **Anwendungsbereich**: reelle i.i.d. Stichprobe $X_i,\ i = 1, \ldots, n$,
  $g(F) = \int \ldots \int T(x_1, \ldots, x_d) dF(x_1) \ldots dF(x_d)$

- **Schätzer**: Der Schätzer entspricht dem Mittelwert über T für alle Kombinationen verschiedener X_i:

$$\hat{g} = \frac{1}{N} \sum_{i_1 \neq \ldots, \neq i_n} T(X_{i_1}, \ldots, X_{i_n})$$

  mit N =Anzahl der Summanden.

- **Eigenschaften:**
  - Stark konsistent
    (Beweisskizze: Der Unterschied zum entsprechenden V-Schätzer konvergiert gegen 0)
  - Erwartungstreu
    (Beweisskizze: Offenbar ist der Erwartungswert jedes Summanden der Summe gleich $g(F)$)

- **Übliche Voraussetzungen:** wie angegeben.

Betrachten wir dazu nochmal das Varianzbeispiel:

$$\begin{aligned}
\text{var}(X) &= g(F) = \int (x - \int y dF(y))^2 dF(x) \\
&= \int x^2 - 2x \int y dF(y) + (\int y dF(y))^2 dF(x) \\
&= \int \int x^2 - 2xy dF(y) dF(x) + (\int y dF(y))^2 \\
&= \int \int x^2 - 2xy dF(y) dF(x) + \int y dF(y) \int x dF(x) \\
&= \int \int x^2 - xy dF(y) dF(x)
\end{aligned}$$

Also erhalten wir den U-Schätzer:

$$\begin{aligned}
\hat{var} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} x_i^2 - x_i y_i \\
&= \frac{1}{n(n-1)} \left((n-1) \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_i - \sum_{i=1^n} x_i^2 \right) \right) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \frac{n}{n-1} \bar{x}^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}$$

Die Umformung dient der Elimination der Doppelsumme und der direkteren Interpretation als mittlere quadratische Abweichung vom Mittelwert.

Dieser Schätzer für die Varianz wurde erstmals von C.F. Gauß angegeben und vertreten, da dieser Schätzer im Gegensatz zu dem vorher verwendeten V-Schätzer oben erwartungstreu ist.

Beispiel 51 U-Schätzer

- *nichtzentrierte Momente:* $\mu_d \to T(x_1) = x_1^d$
- *zentrierte Momente:* $\mu_d \to T(x_1, \ldots, x_d) = x_1^d - x_1 \ldots x_d$
- *standardisierte Momente kann man mit U-Schätzern nicht schätzen. Dazu verwendet man meist Momentenschätzer.*

4.4.6 Überleitung: Schätzung in Modellen

Die bisher besprochenen Schätzprinzipien eignen sich sehr gut, wenn der Schätzer als Funktional der Verteilungsfunktion leicht angegeben werden kann. Oft hat man jedoch ein parametrisches statistisches Modell und möchte die Parameter der Verteilung schätzen, die sich oft nur sehr umständlich anderesherum wieder als Funktional der Verteilung schreiben lassen. Das typische Modell für diesen nächsten Abschnitt ist also:

$$X_i \sim P_\theta$$

und $g(\theta) = \theta$. Dabei ist θ des öfteren auch mehrdimensional.

4.4.7 Momentenschätzer

Mit dem U-Schätzer lassen sich also die Momente der Verteilung leicht schätzen. Der Momentenschätzer versucht das nun für die Schätzung des Verteilungsparameters auszunutzen. Dazu werden zunächst einige ausgewählte Momente z.B. Erwartungswert und Varianz mittels U-Schätzer geschätzt und dann derjenige Verteilungsparameter angegeben, der zu diesen Momenten korrespondiert. Ist diese Zuordnung beiderseitig stetig, so ist der resultierende Schätzer wieder stark konsistent.

- **Name:** Momenten-Schätzer
- **Anwendungsbereich**: reelle i.i.d. Stichprobe $X_i \sim P_\theta,\ i = 1, \ldots, n$, gesucht $g(\theta)$ stetige Funktion von θ, $\theta \in \mathbb{R}^p$
- **Nutzerentscheidung**: Welche Momente $m_1, \ldots, m_p$ sollen verwendet werden.
- **Schätzer**: Bestimme zunächst die U-Schätzer $\hat{m}_1, \ldots, \hat{m}_p$ von $m_1, \ldots, m_p$ und berechne $m_1(\theta), \ldots, m_p(\theta)$ als Funktion vom Verteilungsparameter. Dann wähle $\hat{\theta}$, so dass gilt:
  $$\forall j : m_j(\hat{\theta}) = \hat{m}_j$$
  und setze $\hat{g} = g(\hat{\theta})$.
- **Eigenschaften:**
  - Stark konsistent falls $\hat{\theta}$ eine stetige Funktion der $\hat{m}_j$ ist.
    (Beweisskizze: Wenn die $\hat{m}_j$ konvergieren, so wegen der Stetigkeit auch die $\hat{\theta}$, und damit wegen der vorausgesetzten Stetigkeit auch $g(\hat{\theta})$)
- **Übliche Voraussetzungen:** wie angegeben.

Beispiel 52 Momentenschätzer

- *Eine Schätzung für die Standardabweichung ergibt sich als Wurzel der Schätzung für die Varianz.*
  $$\hat{sd}(X) = \sqrt{\hat{var}(X)}$$

- *Eine nette Übungsaufgabe ist der Schätzer für den λ-Parameter der Exponentialverteilung.*

Momentenschätzer sind im allgemeinen nicht sehr hoch angesehen, da sie in keiner Weise garantieren, dass die Schätzung effizient oder robust ist. Die Effizienz wird insbesondere durch die Nutzerentscheidung, welche Momente verwendet werden sollen beeinflusst.

4.4.8 ML-Schätzer

Ein sehr beliebtes Arbeitspferd ist der **Maximum–Likelihood–Schätzer** oder **ML-Schätzer**. Im Unterschied zu den bisher vorgestellten Schätzern funktioniert er theoretisch auch ohne die Annahme der identischen Verteilung und theoretisch auch für beliebige Parameterräume.

Der Maximum-Likelihood-Schätzer basiert auf der Idee denjenigen Parameter als Schätzung anzugeben, unter dem die Beobachtung am wahrscheinlichsten ist:

$$\hat{g}(x) = g\left(\operatorname*{argmax}_{\theta\in\Theta} P_\theta(\{\omega : X(\omega) = x\})\right)$$

Beispiel 53 *Es soll die Erfolgswahrscheinlichkeit, bei einem binomialen Experiment bestimmt werden: Es wurden n unabhängige Würfe durchgeführt von denen x ein Erfolg waren. Das Modell ist also $X \sim P_p = Bi(n,p)$ und gesucht ist ein Schätzer für p bei bekanntem n. Gemäß der Binomialverteilung gilt:*

$$L(p) := P_p(\{\omega : X(\omega) = x\}) = \binom{n}{x} p^x (1-p)^{n-x}$$

Die Wahrscheinlichkeit, betrachtet als Funktion des Parameters, wird dabei als Likelihood (im Gegensatz zu Probability) bezeichnet und meist mit L abgekürzt. Der Maximum-Likelihood-Schätzer für p ist damit gegeben durch:

$$\hat{p} = \operatorname*{argmax}_{p\in[0,1]} L(p)$$

Dieses Maximum kann durch Kurvendiskussion und Differenzieren leicht bestimmt werden:

$$\begin{aligned}
\frac{\partial}{\partial p} L(p) &= \binom{n}{x} \frac{d}{dp} \exp\left(x \ln p + (n-x)\ln(1-p)\right) \\
&= \binom{n}{x} \exp\left(x \ln p + (n-x)\ln(1-p)\right) \left(\frac{x}{p} - \frac{n-x}{1-p}\right) \\
&= \dots
\end{aligned}$$

Oh, das ist wohl doch nicht so leicht. Aus tiefliegenden informationstheoretischen Gründen hilft uns an dieser Stelle immer (ja immer) ein Trick: Wenn man das Maximum einer Funktion sucht, dann kann man sie auch isoton transformieren. Und es wird an dieser Stelle immer einfacher, wenn man den Logarithmus bildet. Wir minimieren also statt $L(p)$ immer seinen Logarithmus:

$$l(p) := \ln L(p)$$

$l(p)$ heißt die Loglikelihood.

$$\begin{aligned}
\frac{\partial}{\partial p} l(p) &= \frac{d}{dp} \ln \binom{n}{x} + \frac{d}{dp}(x \ln p) + \frac{d}{dp}((n-x)\ln(1-p)) \\
&= 0 + \frac{x}{p} - \frac{n-x}{1-p}
\end{aligned}$$

Um das Maximum zu finden setzten wir das gleich 0:

$$\begin{aligned} 0 &= \frac{\partial}{\partial p} l(p) \\ 0 &= \frac{x}{p} - \frac{n-x}{1-p} \\ 0 &= x(1-p) - (n-x)p \\ 0 &= x + np \\ p &= \frac{x}{n} \end{aligned}$$

Ein eventuelles Maximum liegt also bei $p = \frac{x}{n}$. Überprüfen wir mittels der zweiten Ableitung, ob das ein Maximum ist:

$$\frac{\partial^2}{\partial p^2} l(p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} < 0$$

- *Damit ist die Funktion über den ganzen Bereich negativ gekrümmt und somit auch an der Stelle $\frac{x}{n}$. Es handelt sich also um ein Maximum.*
- *Wegen der durchgehend negativen Krümmung können die Werte auch am Rande des Definitionsbereichs nicht größer sein. Alternativ kann man natürlich nachrechnen.*

$$l(0) = \ln \binom{n}{x} + x(-\infty) + 0 = -\infty < l(\frac{x}{n})$$

$$l(1) = \ln \binom{n}{x} + x0 + (n-x)(-\infty) = -\infty < l(\frac{x}{n})$$

- *Wir haben also das eindeutige Maximum bei $p = \frac{x}{n}$ gefunden.*

Der Maximum Likelihood Schätzer wird also zu

$$\hat{p} := \frac{x}{n}$$

bestimmt. Das ist kein sonderlich überraschendes Ergebnis, da das z.B. dem Wert des Momentenschätzers zum Mittelwert entspricht.

Da in stetigen Situationen die Beobachtung stets die Wahrscheinlichkeit 0 hat, verwendet man in diesem Fall anstelle der Wahrscheinlichkeit der Beobachtung ihre Dichte:

$$L(\theta) := \frac{dP_\theta(x)}{d\mu(x)}$$

bezüglich irgend eines dominierenden Maßes μ, $\forall\theta \in \Theta : P_\theta \ll \mu$. Die Dichte als Funktion des Parameters heißt dann wieder Likelihood. Wieder geht man zum Logarithmus über und definiert eine log-Likelihood:

$$l(\theta) := \ln \frac{dP_\theta(x)}{d\mu(x)}$$

wobei jeweils x den tatsächlich beobachteten Wert darstellt.

- **Name:** ML-Schätzer
- **Anwendungsbereich**: parametrisches statistisches Modell $X \sim P_\theta, \theta \in \Theta$ (d.h. endlich dimensionales oder endliches Θ). Gesucht wird ein Parameter $g(\theta)$.

- **Schätzer**: $\hat{g}(x) := g\left(\underset{\theta\in\Theta}{\operatorname{argmax}} \frac{dP_\theta(x)}{\mu(x)}\right) = g\left(\underset{\theta\in\Theta}{\operatorname{argmax}} \ln \frac{dP_\theta(x)}{\mu(x)}\right)$
  für ein beliebiges dominierendes Maß μ.

- **Eigenschaften:**

  - Stark konsistent ungefähr falls
    - $X = (X_1, \ldots, X_n)$, i.i.d. Zufallsfolge
    - $\dim \Theta = d < \infty$ unabhängig von n.
    - alles hinreichend regulär (z.B. zweimal stetig differenzierbar) ist.

    (Beweisskizze: Im Spezialfall lässt sich der Beweis analog zum Beweis der Konsistenz der Momentenmethode führen.)
  - Oft asymptotisch effizient (näheres in der mathematischen Statistik): d.h. für große Stichprobenumfänge lohnt sich das Weitersuchen nach einem besseren Schätzer nicht. (Beweisskizze: Der Beweis erfolgt durch Abschätzung gegen die sogenannte Cramer-Rao-Schranke.)
  - Der Wert ist unabhängig von μ und der Parametrisierung der Verteilung.

- **Übliche Voraussetzungen:** ein korrektes endlich dimensionales Modell.

Haben wir eine i.i.d. Stichprobe $X_i,\ i = 1, \ldots, n$, so ist das Vorgehen:

- Sei $f_\theta(x)$ die Dichtefunktion von P_θ

- Die Unabhängigkeit impliziert:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i)$$

- Übergang zur Loglikelihood:

$$l_i(\theta) = \ln f_\theta(X_i)$$

$$l(\theta) = \ln \prod_{i=1}^{n} f_\theta(X_i) = \sum_{i=1}^{n} l_i(\theta)$$

- Ableiten:

$$s(\theta) = \sum_{i=1}^{n} l_i'(\theta)$$

  $s(\theta)$ heißt auch die **Scorefunktion**.

- Nullsetzen:

$$0 = \sum_{i=1}^{n} l_i'(\theta)$$

  und nach θ auflösen. Diese Gleichung heißt Schätzgleichung. Die Lösung $\hat{\theta}$ bildet den Maximum-Likelihood-Schätzer von θ.

- Den Maximum-Likelihood-Schätzer für g erhält man indem man den Maximum-Likelihood-Schätzer für θ einsetzt: $\hat{g} = g(\hat{\theta})$

4.4.9 M-Schätzer

Ein wesentlicher Vorwurf an die Maximum-Likelihood-Schätzer ist, dass sie oft nicht robust gegen Ausreißer oder Modellmissspezifikationen in den Bereichen geringer Dichte sind. Um dieses Problem zu reparieren, verwendet die robuste Statistik oft eine abgewandelte Version des ML-Schätzers in dem nicht die Loglikelihood, sondern eine andere Funktion $S(\theta)$ maximiert wird. Daher spricht man von M-Schätzern, wie Maximierungsschätzern. Diese Funktion besteht aus einer Summe über Schätzfunktionen $\Phi(x, \theta)$, die in einem geeigneten Sinne die Möglichkeit x zu beobachten, wenn θ der wahre Parameter ist, bewerten.

$$S(\theta) = \sum_{i=1}^{n} \Phi(x_i, \theta)$$

Da man das Maximum sucht, kann man zur Ableitung übergehen:

$$S'(\theta) = \sum_{i=1}^{n} \phi(x_i, \theta)$$

mit $\phi(x_i, \theta) = \Phi'(x_i, \theta)$ und den Schätzer als Lösung der Schätzgleichung:

$$\sum_{i=1}^{n} \phi(x_i, \hat{\theta}) = 0$$

bestimmen. Damit das eine sinnvolle Schätzung ist, sollten Φ bzw. ϕ einige Eigenschaften haben:

- Damit das Maximum eindeutig ist, sollte Φ für festes x (am besten strikt) konvex (von oben) sein.
- Der Erwartungswert von Φ unter θ sollte ein Maximum in θ haben:

$$E_{P_\theta}[\Phi'(X_i, \theta)] = 0 \tag{4.1}$$

Zusammenfassend lautet das Verfahren also:

- **Name:** M-Schätzer
- **Anwendungsbereich**: parametrisches statistisches Modell $X_i \sim P_\theta, \theta \in \Theta$ (d.h. endlich dimensionales Θ). Gesucht wird ein Parameter $g(\theta)$.
- **Schätzer**: $\hat{g}(x) := g\left(\underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \Phi(x_i, \theta)\right)$
  für ein strikt konvexes Φ, welches Gleichung 4.1 erfüllen.
- **Eigenschaften:** Meist stark konsistent (Beweisskizze: Beweis ist kompliziert und benutzt $\frac{1}{n}\sum_{i=1}^{n} \Phi'(x_i, \theta)$ konvergiert P-f.s. gegen 0)

4.4.10 Spezielle Schätzprinzipien

4.4.10.1 Kleinste Quadrate

Kleinste Quadrate Schätzer werden angewendet, wenn eine Modell der Form

$$X_i = F_i(\theta) + \epsilon_i$$

mit einer bekannten Funktion $F_i(\theta)$ und unbekannten zufälligen ϵ_i mit Erwartungswert 0 und gemeinsamer Varianz σ^2 gegen ist. Das Verfahren lautet:

$$\hat{g}(\theta) = g\left(\operatorname*{argmin}_{\theta} \|X_i - F_i(\theta)\|^2\right)$$

Das Verfahren ergibt als sich Maximum-Likelihood-Schätzer, wenn die ϵ_i i.i.d. $N(0, \sigma^2)$ verteilt sind. Aber auch für andere Verteilungen funktioniert das Verfahren unter den obigen Voraussetzungen.

4.4.10.2 Kernschätzer

Zur Schätzung von Dichten kann kein geeigneter Mittelwert herangezogen werden, da die Dichte ja dem Grenzübergang, des Anteils der Wahrscheinlichkeit in einem infinitisimal kleinen Bereich entspricht. Zur Schätzung von Dichten verwendet man daher einen gewichteten Mittelwert über einen kleine Bereich:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k_b(X_i - x)1$$

wobei $k(X_i - x)$ einer Gewichtungsfunktion entspricht, die sagt wie wichtig ein Punkt in einer gegeben Entfernung ist. Dabei gilt jeweils:

$$\int k_b(x)dx = 1$$

Die Funktion k_b heißt Kernfunktion oder einfach Kern. Dabei ist es wichtig die Breite des Kerns an das Problem anzupassen. Die Kernfunktionen haben daher einen Breitenparameter b mit dem sie aus einem Vorläuferkern k ermittelt werden:

$$k_b(x) := \frac{1}{b}k(\frac{x}{b})$$

Wählt man b zu klein, so wird der Mittelwert über einen zu kleinen Bereich gebildet und man erhält eine zu große Varianz der geschätzten Dichte. Wird b zu groß gewählt, so werden Bereiche anderer Dichte zu stark in die Mittelwertbildung mit einbezogen und man erhält einen stark verzerrten Schätzer.

4.5 Bayes-Schätzer

4.5.1 Definition des Bayes-Schätzers

Ein anderer Ansatz für die Konstruktion von Schätzern bietet die Formel von Bayes. Dazu wird angenommen, dass die Parameter θ selbst zufällig sind und einer Verteilung P_{prior} auf einem Messraum $(\Theta, \mathfrak{T})$ folgen. Der Parameter wird also selbst zu einer Zufallsvariable Θ mit einer Verteilung $P^{\Theta} = P_{\text{prior}}$. Diese Verteilung heißt auch a-priori-Verteilung, weil sie vor der Datenerhebung gegen ist. Sie modelliert das Vorauswissen oder die Vorausvermutung des Statistikers und ist somit eine subjektive Wahrscheinlichkeit und entspricht der vermuteten marginalen Verteilung P^{Θ} des Parameters.

Das Modell P_θ wird als bedingte Verteilung gegeben θ aufgefasst:

$$P^X(A|\Theta = \theta) = P_\theta(A)$$

Mit der Formel von Bayes lässt sich nun die bedingte Verteilung des Parameters gegeben die Daten berechnen:

$$P_{\text{post}}(A) := P^{\Theta}(A|X = x)$$

```
> opar <- par(mfrow = c(2, 2))
> plot(density(morley$Speed, width = 10), main = "width=10")
> points(morley$Speed, runif(nrow(morley), 0, 1e-04), pch = 1)
> plot(density(morley$Speed, width = 20), main = "width=20")
> points(morley$Speed, runif(nrow(morley), 0, 1e-04), pch = 1)
> plot(density(morley$Speed, width = 40), main = "width=40")
> points(morley$Speed, runif(nrow(morley), 0, 1e-04), pch = 1)
> plot(density(morley$Speed, width = 80), main = "width=80")
> points(morley$Speed, runif(nrow(morley), 0, 1e-04), pch = 1)
> par(opar)
```

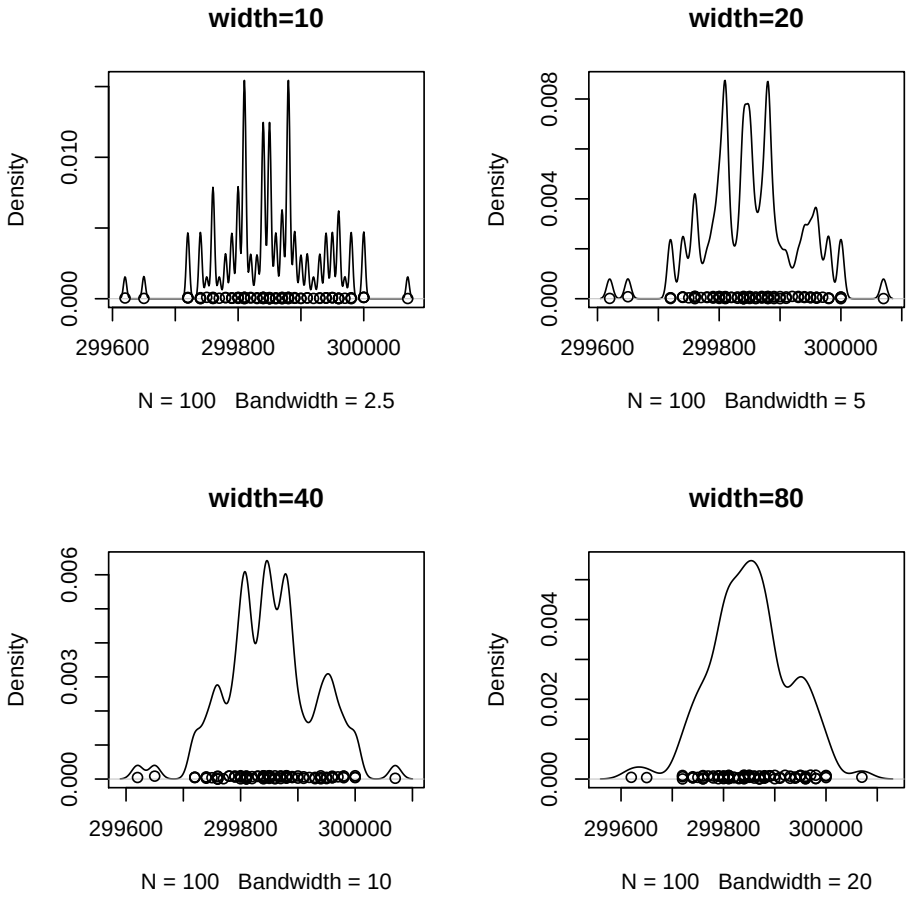

Abbildung 4.3: Kerndichteschätzer mit unterschiedlichen Breitenparametern

```
> plot(density(morley$Speed))
> points(morley$Speed, runif(nrow(morley), 0, 1e-04), pch = 1)
```

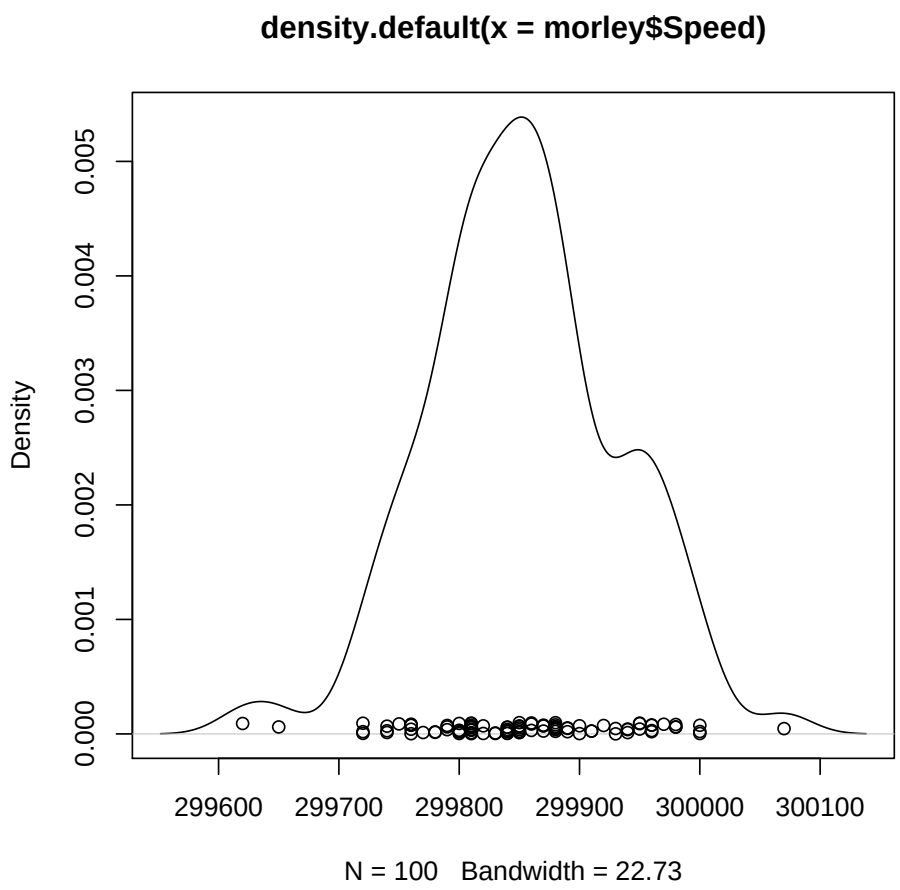

Abbildung 4.4: Kerndichteschätzer

```
> opar <- par(mfrow = c(2, 2))
> x <- c(0)
> plot(density(0, kernel = "gaussian", width = 1), main = "Gausskern")
> plot(density(0, kernel = "epanechnikov", width = 1),
+     main = "Epanechnikov Kern")
> plot(density(0, kernel = "rectangular", width = 1), main = "Rechteck Kern")
> plot(density(0, kernel = "cosine", width = 1), main = "Cosinus Kern")
> par(opar)
```

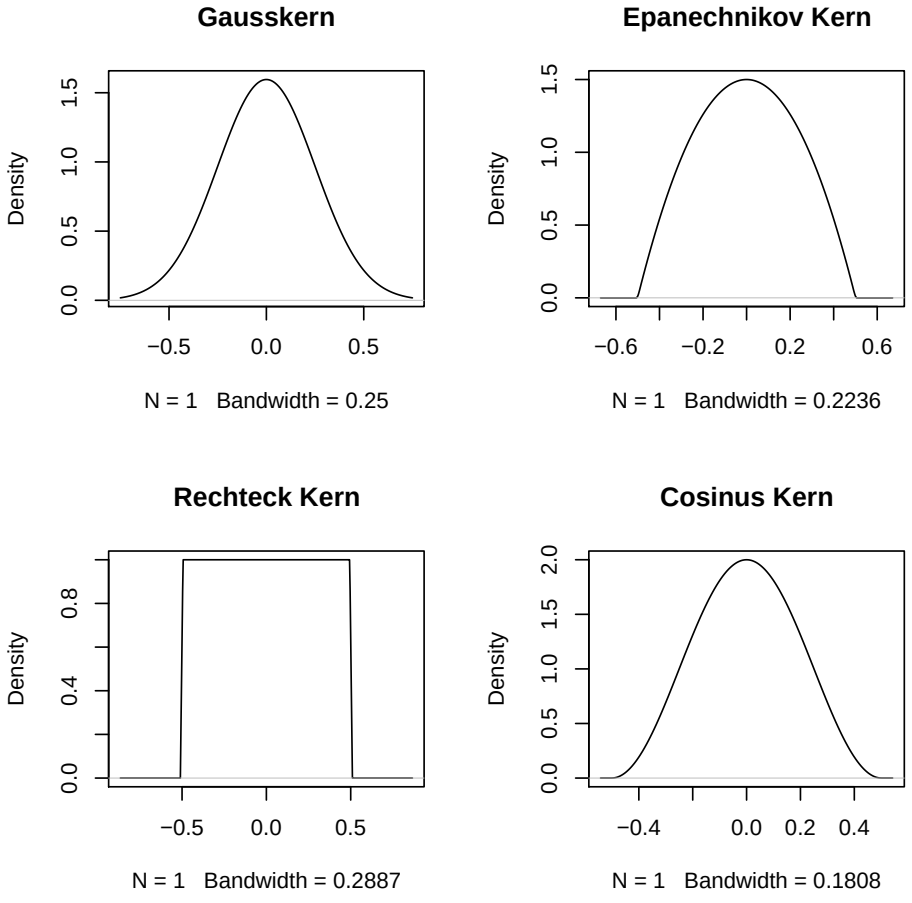

Abbildung 4.5: Typische Kerne

Zur Anwendung der Formel von Bayes, die ja in Dichten formuliert ist, müssen dominierende Maße $\mu \gg P_\theta$ und $\nu \gg P^\Theta$ benutzt werden. Damit ergibt sich

$P^\Theta(\cdot|X = x)$ nach der Formel von Bayes gemäß:

$$\begin{aligned} P_{\text{post}}(A) &:= P^\Theta(A|X = x) \\ &= \int_A \frac{\frac{dP_\theta(x)}{d\mu(x)} \frac{dP^\Theta(\theta)}{d\nu(\theta)}}{\int_\Omega \frac{dP_{\theta'}(x)}{d\mu(x)} \frac{dP^\Theta(\theta')}{d\nu(\theta')} d\nu(\theta')} d\nu(\theta) \\ &= \frac{\int_A \frac{dP_\theta(x)}{d\mu(x)} dP^\Theta(\theta)}{\int \frac{dP_{\theta'}(x)}{d\mu(x)} dP^\Theta(\theta')} \end{aligned}$$

oder formuliert mit Hilfe der Dichten:

$$\begin{aligned} f(\theta) &:= \frac{dP^\theta(\theta)}{d\nu(\theta)} = \frac{dP_{\text{prior}}}{d\nu} \\ f_\theta(x) &:= \frac{dP^X(x|\theta = \theta)}{d\mu(x)} = \frac{dP_\theta}{d\mu} \\ f_x(\theta) &:= \frac{dP^\Theta(\theta|X = x)}{d\nu(\theta)} = \frac{dP_{\text{post}}}{d\nu} \end{aligned}$$

als

$$f_x(\theta) = \frac{f_\theta(x) f(\theta)}{\int f_{\theta'}(x) f(\theta') d\nu(\theta')}$$

Als Schätzwert kann dann der Wert gewählt werden, der den geringsten erwarteten Verlust hat.

$$\hat{g} = \operatorname*{argmin}_w E_{P_{\text{post}}}[v(g(\theta), w)]$$

Der so konstruierte Schätzer heißt **Bayes-Schätzer** von g bezüglich der Verlustfunktion v und der a-priori-Verteilung P_{prior}. Der Bayesschätzer minimiert offenbar den unter der a-priori-Verteilung P_{prior} erwarteten Verlust.

$$\begin{aligned} BV_{\hat{g}} &:= E_{P_{\text{prior}}}[EV(g(\theta), \hat{g})] \\ E_{P_{\text{prior}}}[E_{P_\theta}[v(g(\theta), \hat{g}(x))]] & \\ &= E_{P_\theta}\left[E_{P_{\text{prior}}}[v(g(\theta), \hat{g}(x))]\right] \end{aligned}$$

da der innere Teil für jedes θ minimal ist. Dieser unter der a-priori-Verteilung erwartete Verlust BV heißt auch Bayes-Verlust.

4.5.2 Bayes-Schätzer bei quadratischem Verlust

Es soll noch darauf hingewiesen werden, dass der quadratische Verlust $v(x, y) = \|x - y\|^2$ durch den bedingten Erwartungswert

$$\hat{g}(x) = E_{P_{\text{post}}}[\Theta] = E[\Theta|X = x]$$

minimiert wird. Der einfachste Bayes Schätzer entspricht also dem Erwartungswert der a-posteriori-Verteilung.

4.5.3 Diskussion des Bayes-Schätzers

Der Bayes-Schätzer basiert als einziges Schätzprinzip auf einem vollständigen und zwingenden mathematischen Modell des Schätzproblems und garantiert als einziges Schätzprinzip Optimalitätseigenschaften schon bei kleinen Stichprobenumfängen. Es gibt eine weitentwickelte mathematische Theorie zur Bayes-Schätzung.

Dennoch gibt es eine Reihe von Einwänden und praktischen Problemen, die dazu geführt haben, dass bayessche Verfahren sich nicht allgemein durchgesetzt haben:

4.5.3.1 Wahl der a-priori-Verteilung

Die Wahl der a-priori-Verteilung hat eine subjektive Komponente. Zum Einen ergibt sich daraus das praktische Problem, dass verschiedene Menschen verschiedene a-priori-Verteilungen wählen wollen und daher verschiedene Bayes-Schätzer brauchen, so dass man keine einheitlichen auf ewig, feststehende Schätzer "lernen" kann. Zum anderen wird der Schätzer dadurch als subjektiv angreifbar. Das ist besonders dann problematisch, wenn verschiedene Interessengruppen verschiedene Vorstellungen entwickeln, wie die Welt zu sein hat. Es ist daher oft leichter einen einfach konstruierten Schätzer politisch zu vertreten, als einen Bayes-Schätzer, der im Geruch der Subjektivität steht.

Zur Lösung des ersten Problems mögen die sogenannten **konjugierten a-priori-Verteilungsklassen** beitragen, bei denen auch die a-priori-Verteilung Parameter hat, die vom Benutzer nach seinen Vorstellungen gesetzt werden können. Die Berechnung des Bayes-Schätzers erfolgt dann für die gesamte Verteilungsklasse und dass Ergebnis ist vom gewählten Parameter abhängig. An diesen Klassen kann man auch erkennen, dass sich die a-posteriori-Verteilungen selbst bei radikalen Unterschiede in der a-priori-Verteilungen kaum unterscheiden, wenn nur genügend vielen Daten vorliegen.

Zur Lösung des zweiten Problems wurden die sogenannten **objektiven a-priorie-Verteilungen** entwickelt, welche die a-priori-Verteilung nach nichtsubjektiven informationstheoretischen Methoden bestimmen. Einerseits widerspricht das allerdings gerade dem bayesschen Ansatz, der ja versucht das Vorauswissen angemessen zu verwenden und andererseits ergeben sich mit diesen a-priori-Verteilungen oft Schätzer, die auch mit den nicht bayesschen Methoden ermittelt werden, eben weil die vorhandene Information nicht genutzt wird, so dass sich in diesem Fall die Anwendung der aufwendigeren Bayes-Methoden oft nicht lohnt. Andererseits lassen sich gerade in komplizierteren Schätzproblem gerade auf diesem Weg die klassischen Schätzer überhaupt erst konstruieren, so dass es letztlich nur noch eine Frage ist, mit welcher Methode man einen Schätzer legitimiert.

4.5.3.2 Praktische Durchführung

Ein nicht unwesentliches Problem stellte lange Zeit die praktische Berechnung von Bayes-Schätzern für kompliziertere Modelle dar. Das praktische Problem steckt jeweils in der Berechnung des typischerweise hochdimensionalen Integrals im Nenner. Dadurch wurde die Bayes-Statistik lange Zeit in die Ecke der theoretischen Spielerei gedrängt und erwies sich als für viele praktische Probleme ungeeignet. Die anderen Schätzer konnten leichter konstruiert werden und mussten sich ohne Konkurrenz kaum legitimieren.

Durch die moderne Computertechnik und neue Algorithmen, welche diese Integrale durch Simulation von Zufallsexperimenten bestimmen, haben sich in den letzten Jahren die Verhältnisse fast umgekehrt. Während die anderen Konstruktionsprinzipien an ihre technischen Grenzen gestoßen sind, lassen sich Bayes-Schätzer mit massivem Computereinsatz praktisch ohne theoretischen Aufwand auch für sehr komplizierte Modelle berechnen und durch die einfache Theorie legitimieren, so dass ein eher geringer theoretischer Aufwand nötig ist. Diese Algorithmen, welche stochastische Simulation zu Berechnungen einsetzten, heißen **Monte-Carlo-Algorithmen**. Man sollte aber beachten, dass die Konvergenz dieser Algorithmen oft schwer zu überprüfen ist, so dass auch im Bereich der Bayes-Schätzer von einem naiven Einsatz ohne hinreichende Sachkenntnis abzuraten ist.

4.5.3.3 Genau formulierte Modelle

Auch das genau formulierte Modell macht der Bayes-Statistik in verschiedener Weise zu schaffen. Zum Einem erlaubt die exakte Formulierung des Modells und semikonstruktive Angabe des Schätzers als argmin für eine beliebige Verlustfunktion theoretisch die Konstruktion des besten Schätzers für die jeweilige Situation. Dadurch das man die Aufgabe aber so klar formulieren muss, ergibt sich auch ein Anspruch die Aufgabe richtig zu formulieren. Praktisch kann man die sich ergebenden Rechenaufgaben dann nicht unbedingt lösen. Dieses Problem scheint aber in den anderen Ansätzen gar nicht lösbar und wird daher dort einfach ignoriert.

Zu anderen entsteht, wie bei der Maximum-Likelihood-Schätzung das Problem, dass der mathematische Apparat garantiert, dass die Daten gemäß dem Modell optimal ausgenutzt werden. Sollten sich aber Fehler in den Daten befinden, oder das Modell in Details falsch sein, so kann das starke Auswirkungen auf den Schätzwert haben. Bayes-Schätzer sind also oft nicht robust gegen Ausreißer und Modellfehler. Dazu kommt, das immer ein parametrisches Modell formuliert werden muss, da sonst meist kein dominierendes Maß existiert. Der bayesschen Statistik steht also der Ausweg zur Robustifizierung nicht in der Weise offen, wie den anderen statistischen Ansätzen.

Eine robuste und nichtparametrische Bayes-Statistik steckt allerdings noch in Kinderschuhen und ist in Praxis noch nicht angekommen.

4.5.4 Bayes-Schätzer als Schätzverfahren

- **Name:** Bayes-Schätzer
- **Anwendungsbereich**: parametrisches statistisches Modell $X_i \sim P_\theta, \theta \in \Theta$ (d.h. endlich dimensionales Θ). Gesucht wird ein Parameter $g(\theta)$.
- **Schätzer**: $\hat{g}(x) := \underset{g}{\operatorname{argmin}} E_{P_{\text{post}}}\left[v(g(\Theta), g)\right]$
- **Eigenschaften:** Minimales Bayes-Risiko, meistens stark konsistent,...