Einführung in die induktive Statistik

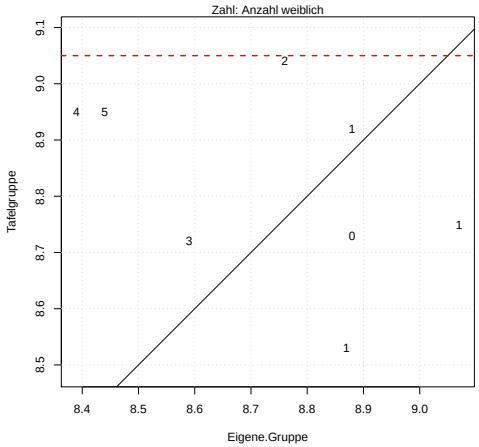
Friedrich Leisch

Institut für Statistik
Ludwig-Maximilians-Universität München

SS 2009

Konfidenzintervalle

Spiel Körpergröße



Inferenzstatistik

Methoden, um mit Informationen aus Stichproben auf Charakteristika der Gesamtpopulation schließen zu können.

Nach Modellierung und Parameterschätzung ist man meist an der Überprüfung konkreter Fragestellungen interessiert (ja/nein-Entscheidungen statt lange Beschreibungen der Daten).

Mögliche Frageformen:

1. Paßt eine Hypothese zu meinem Modell?
2. Widerspricht mein Modell einer gewissen Hypothese?

In der klassischen Inferenzstatistik werden vor allem Fragen der zweiten Form behandelt.

Beispiel: Sonntagsfrage

Vier Wochen vor der österreichischen Nationalratswahl 1999 wurde 499 Haushalten die „Sonntagsfrage“ gestellt: Falls nächsten Sonntag Wahlen wären, welche Partei würden Sie wählen?

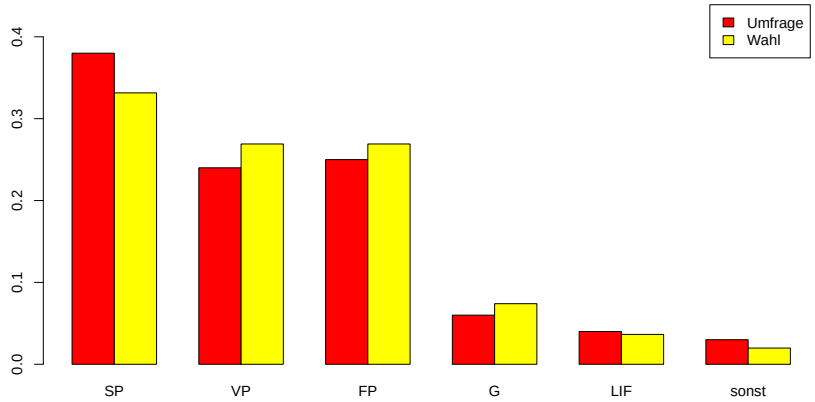
	SPÖ	ÖVP	FPÖ	Grüne	LIF	Sonst
Umfrage	38%	24%	25%	6%	4%	3%
Wahl	33.15%	26.91%	26.91%	7.4%	3.65%	1.98%

Frage 1: War das Ergebnis für die SPÖ überraschend?

Frage 2: Mit welcher Wahrscheinlichkeit mußte das LIF damit rechnen, den Wiedereinzug ins Parlament nicht zu schaffen? Mit welcher Wahrscheinlichkeit die Grünen?

Frage 3: War das Gesamtergebnis überraschend?

Beispiel: Sonntagsfrage



Konfidenzintervall für Mittelwert

Schätzung des Mittelwerts:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Kommt die Stichprobe aus einer Population mit Mittelwert μ und Varianz σ^2, dann gilt für $n \to \infty$

$$\hat{\mu} \sim N(\mu, \sigma^2/n)$$

und mit 95% Wahrscheinlichkeit liegt $\hat{\mu}$ innerhalb des Intervalls

$$\mu - \frac{1.96\sigma}{\sqrt{n}} \leq \hat{\mu} \leq \mu + \frac{1.96\sigma}{\sqrt{n}}$$

Konfidenzintervall für Mittelwert

Mit $c = 1.96\sigma/\sqrt{n}$ haben wir 2 Ungleichungen

$$\begin{aligned} \mu - c &\leq \hat{\mu} \\ \hat{\mu} &\leq \mu + c \end{aligned}$$

und wollen eigentlich Aussagen über μ treffen. Bringen wir die Konstante jeweils auf die andere Seite so erhalten wir

$$\begin{aligned} \mu &\leq \hat{\mu} + c \\ \hat{\mu} - c &\leq \mu \end{aligned}$$

oder kürzer in einer Zeile

$$\hat{\mu} - c \leq \mu \leq \hat{\mu} + c$$

Konfidenzintervall für Mittelwert

Setzen wir für c wieder $1.96\sigma/\sqrt{n}$ ein, so erhalten wir, daß mit 95% Wahrscheinlichkeit

$$\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}$$

gilt, falls die Stichprobe wirklich aus einer Population mit Mittelwert μ und Varianz σ^2 stammt.

Die obere und untere Intervallgrenze hängen aber nicht von μ ab $\rightarrow$ die Aussage trifft für alle μ aus diesem Intervall gleichermaßen zu.

Wir schließen also: Mit 95% Wahrscheinlichkeit liegt der wahre Mittelwert irgendwo in dem gegebenen Intervall.

Beispiel: Sonntagsfrage

Beginnen wir mit dem Ergebnis der SPÖ (Frage 1): Die Stimmen für jede einzelne Partei können wir als binomialverteilt ansehen (jeweils gegen den Rest).

Wir haben also einen geschätzten Erwartungswert von $\hat{p}_s = 0.38$ und eine zugehörige Varianz von

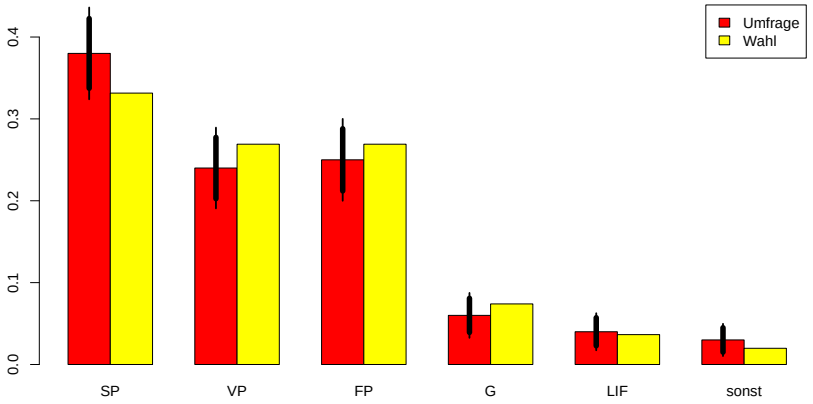
$$\hat{\sigma}_S^2 = \hat{p}_S(1 - \hat{p}_S) = 0.2356$$

somit erhalten wir

$$c = 1.96\sqrt{\frac{0.2356}{499}} = 0.04259$$

und es war zu erwarten, daß das Ergebnis der SPÖ mit 95% Wahrscheinlichkeit im Intervall [33.7, 42.3] liegen würde. Das 99% Intervall ist [32.4, 43.6].

Beispiel: Sonntagsfrage



Beispiel: Sonntagsfrage

Frage 2 betreffend den Verbleib der Grünen und des LIF können wir auch so formulieren: Wie groß ist die Wahrscheinlichkeit, daß $p_G \geq 0.04$ bzw. $p_L \geq 0.04$.

$$\mathbb{P}\{p \geq 0.04\} = \mathbb{P}\{\hat{p} - p \leq \hat{p} - 0.04\}$$

Es gilt

$$\begin{aligned}
\hat{p} &\sim N(p, \sigma^2/n) \\
\hat{p} - p &\sim N(0, \sigma^2/n) \\
\frac{(\hat{p} - p)\sqrt{n}}{\sigma} &\sim N(0, 1) \\
\frac{(\hat{p} - p)\sqrt{n}}{\sqrt{\hat{p}(1 - \hat{p})}} &\sim N(0, 1)
\end{aligned}$$

Beispiel: Sonntagsfrage

$$\mathbb{P}\{p \geq 0.04\} = \mathbb{P}\left\{\frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} \leq \frac{(\hat{p}-0.04)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}\right\}$$

Grüne:

$$\frac{(\hat{p}-0.04)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} = \frac{(0.06-0.04)\sqrt{499}}{\sqrt{0.06(1-0.06)}} \approx 1.881225$$

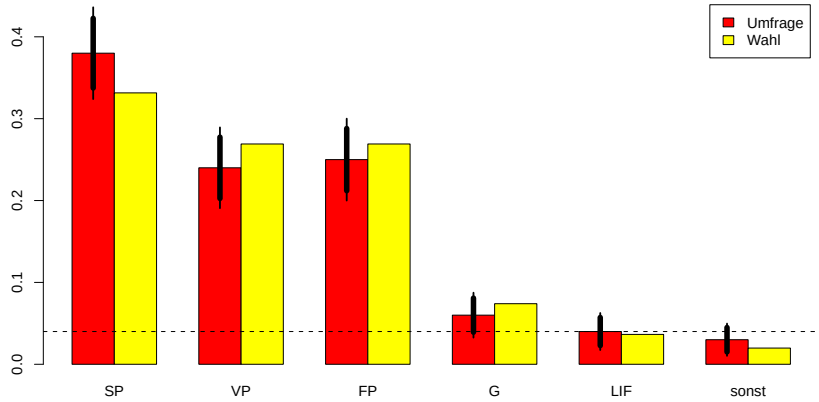
$$\Phi(1.881225) \approx 97\%$$

LIF:

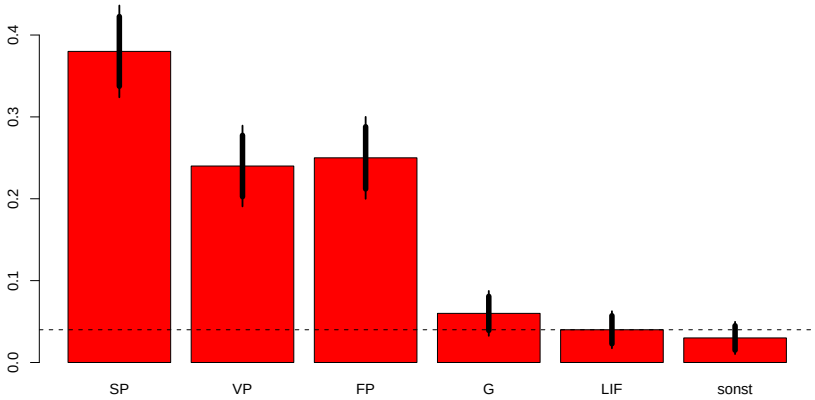
$$\frac{(\hat{p}-0.04)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} = \frac{(0.04-0.04)\sqrt{499}}{\sqrt{0.06(1-0.06)}} = 0$$

$$\Phi(0) = 50\%$$

Beispiel: Sonntagsfrage



Beispiel: Sonntagsfrage



Beispiel: Sonntagsfrage

Die Beantwortung der Frage 3, ob das Gesamtergebnis überraschend war, ist nicht ganz so einfach:

- Dürfen wir einfach 95% Konfidenzintervalle für alle Parteien bilden und das Gesamtergebnis als „nicht überraschend“ klassifizieren, wenn alle innerhalb der Konfidenzintervalle liegen?
- Was passiert, wenn 20 Parteien antreten?
- Sollen wir dann einen „Ausreißer“ akzeptieren? Oder vielleicht gar 2?
- Sind die Ergebnisse der Parteien voneinander unabhängig?

→ darauf kommen wir später nochmal zurück.

$(1-\alpha)$ Konfidenzintervall

Das von den Schätzstatistiken

$$G_u = g_u(X_1, \ldots, X_n) \quad \leq \quad G_o = g_o(X_1, \ldots, X_n)$$

definierte Intervall $[G_u, G_o]$ heißt $(1-\alpha)$ Konfidenzintervall für θ, falls für jede vorgegebener Irrtumswahrscheinlichkeit $\alpha \in [0, 1]$

$$\mathbb{P}\{G_u \leq \theta \leq G_o\} = 1 - \alpha$$

gilt.

Achtung: G_u und G_o sind Statistiken der Stichprobe, und damit Zufallsvariablen.

Realisiertes Konfidenzintervall:
$[g_u, g_o]$; $g_u = g_u(x_1, \ldots, x_n)$, $g_o = g_o(x_1, \ldots, x_n)$

Symmetrisches KI

Von einem symmetrischen Konfidenzintervall spricht man, falls θ mit gleicher Wahrscheinlichkeit links oder rechts außerhalb des Intervalles liegt:

$$\mathbb{P}\{\theta < G_u\} = \mathbb{P}\{\theta > G_o\} = \frac{\alpha}{2}$$

Falls die Verteilung von θ bekannt ist, liefern die Quantile zu den Werten $\alpha/2$ und $1 - \alpha/2$ die Intervallgrenzen.

Einseitige KI

Falls nur eine untere oder obere Schranke für θ von Interesse ist, wird G_o oder G_u auf ∞ gesetzt:

$$\mathbb{P}\{\theta \leq G_o\} = \mathbb{P}\{\infty \leq \theta \leq G_o\} = 1 - \alpha$$

$$\mathbb{P}\{G_u \leq \theta\} = \mathbb{P}\{G_u \leq \theta \leq \infty\} = 1 - \alpha$$

Bei bekannter Verteilung von θ liefern die Quantile zu den Werten $1 - \alpha$ bzw. α die Schranken.

KI für Mittelwert

Gegeben sei eine normalverteilte Stichprobe mit bekannter Varianz σ^2. Dann ist

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

standardnormalverteilt, die Schranken des Konfidenzintervalls sind

$$\begin{aligned} G_u &= \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \\ G_o &= \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

wobei z_α das α-Quantil der $N(0, 1)$ ist.

KI für Mittelwert

Gegeben sei eine normalverteilte Stichprobe mit unbekannter Varianz σ^2. Dann ist

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

t-verteilt mit $n-1$ Freiheitsgraden, die Schranken des Konfidenzintervalls sind

$$\begin{aligned} G_u &= \bar{X} - t_{1-\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} \\ G_o &= \bar{X} + t_{1-\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} \end{aligned}$$

wobei $t_\alpha(n-1)$ das α-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden ist.

KI für Mittelwert

Gegeben sei eine beliebig verteilte Stichprobe mit bekannter Varianz σ^2. Dann ist

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

approximativ standardnormalverteilt, die Schranken des entsprechenden approximativen Konfidenzintervalls sind

$$\begin{aligned} G_u &= \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \\ G_o &= \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

wobei z_α das α-Quantil der $N(0,1)$ ist. Bei unbekannter Varianz wird diese wieder durch $\hat{\sigma}^2$ ersetzt, das KI sollte dann aber erst für größere n verwendet werden (und damit in jedem Fall die Normalverteilungsquantile).

KI für Varianz

Gegeben sei eine normalverteilte Stichprobe mit unbekannter Varianz σ^2. Dann ist

$$q = \frac{n-1}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-1}$$

χ^2-verteilt mit $n-1$ Freiheitsgraden, die Schranken des Konfidenzintervalls für die Varianz sind

$$\begin{aligned} G_u &= \frac{n-1}{q_{1-\alpha/2}}\hat{\sigma}^2 \\ G_o &= \frac{n-1}{q_{\alpha/2}}\hat{\sigma}^2 \end{aligned}$$

wobei q_α das α-Quantil der χ^2-Verteilung mit $n-1$ Freiheitsgraden ist.

KI für Anteilswert

Gegeben sein eine dichotome Stichprobe mit den Ausprägungen 0 und 1 und $\mathbb{P}(X=1) = \pi$. Dann ist

$$\sum_{i=1}^{n} X_i \sim B(n,\pi) \qquad \frac{\bar{X} - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1)$$

Die Schranken des Konfidenzintervalls für π sind

$$\begin{aligned} G_u &= \hat{\pi} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \\ G_o &= \hat{\pi} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \end{aligned}$$

Breite von Konfidenzintervallen

Beispiel Konfidenzintervall für μ:

$$G_u = \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$G_o = \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

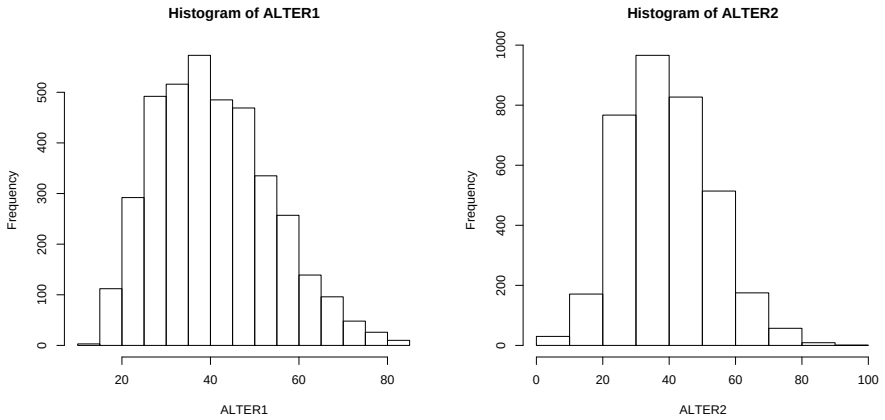
Breite b:

$$b = 2 \cdot z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

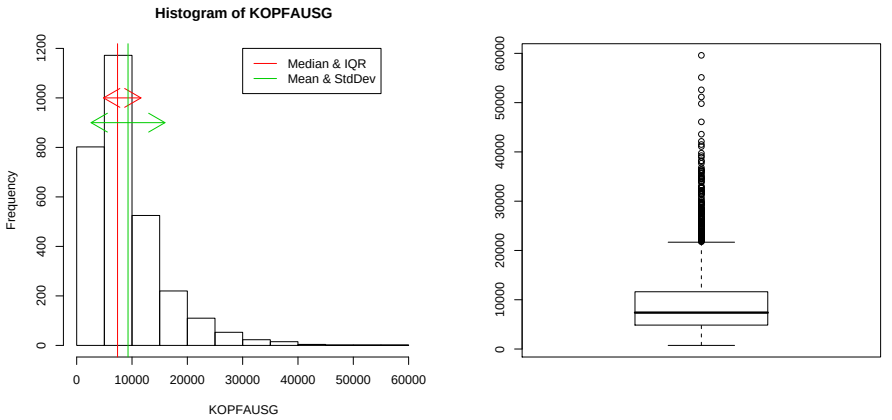
- $1-\alpha$ größer (kleiner) $\Rightarrow$ $z_{1-\frac{\alpha}{2}}$ größer (kleiner) $\Rightarrow$ KI breiter (schmaler)
- n größer (kleiner) $\Rightarrow$ KI schmaler (breiter)
- $n \to nc \Rightarrow$ Breite verändert sich um Faktor $\sqrt{c}$

Überprüfung von Verteilungsannahmen

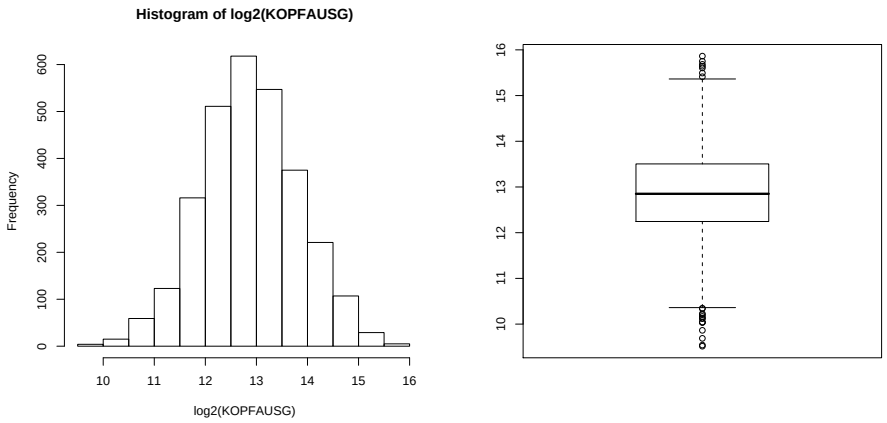
Beispiel GBÖ: Ausgaben



Beispiel GBÖ: Ausgaben



Beispiel GBÖ: log2(Ausgaben)



Beispiel GBÖ: Welche Verteilung?

- Wie gut passen eine Normalverteilung zum Alter bzw. eine Log-Normalverteilung zu den Augaben?
- Optisch nach Histogramm scheinbar recht gut, aber was ist gut?
- Stichprobe extrem groß, daher sollte es hier eigentlich sehr leicht sein.

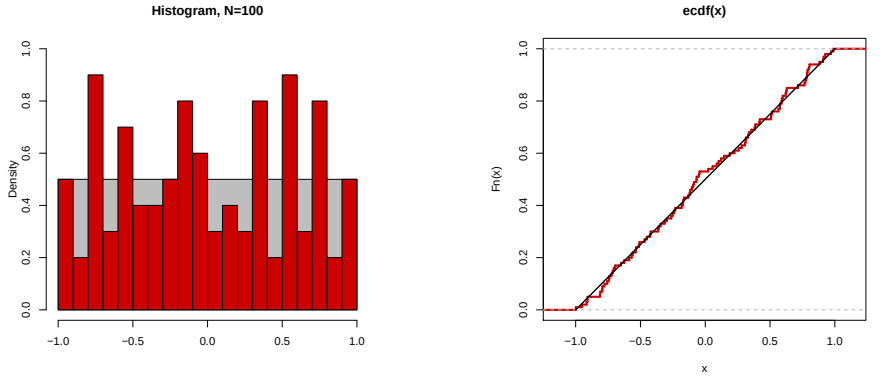
Empirische Dichte und Verteilung

- Als einfache Visualisierung der Wahrscheinlichkeiten einer diskreten Verteilung bzw. Dichte einer stetigen Verteilung verwenden wir Balkendiagramme und Histogramme.
- Verteilung F und empirische Verteilung $\widehat{F}_n$:

$$\begin{aligned} F(x) &= \mathbb{P}\{X \leq x\} \\ \widehat{F}_n(x) &= \frac{\{\#i : x_i \leq x\}}{n} = \frac{\text{günstige}}{\text{mögliche}} \end{aligned}$$

  ist unverzerrter Schätzer für die unbekannte wahre Verteilung
- Echte Dichteschätzung sehr komplexes Thema, es gibt Bücher die sich ausschließlich mit diesem Problem befassen, naive Schätzer wie Histogramm sind meist nicht glatt genug (in Abhängigkeit von Breite der Klassen).

Empirische Dichte und Verteilung



→ Verteilungsschätzer viel „glatter“ (für echten Dichteschätzer muß Histogramm geglättet werden).

Empirische Dichte und Verteilung

Wenn Verteilungsschätzer viel einfacher sind, warum beschäftigt man sich dann überhaupt mit dem Problem der Dichteschätzung?

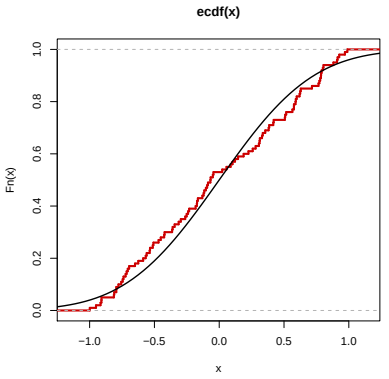
- Dichte ist für Menschen viel intuitiver zu lesen.

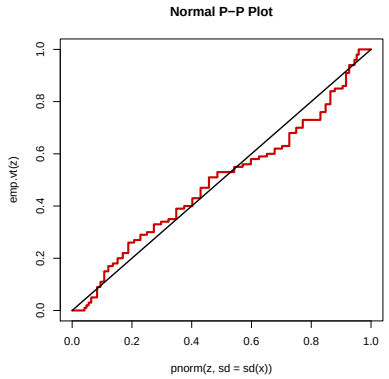
Warum machen Verteilungsschätzer dennoch Sinn?

- Nützlich zum Vergleich von Verteilungen.

Empirische Dichte und Verteilung

Vergleich Gleichverteilung mit Normalverteilung:





Quantile

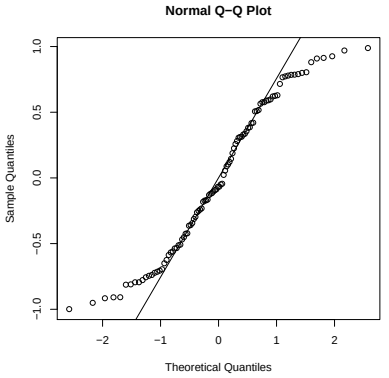
Eine duale Sichtweise: statt der Verteilungsfunktion $F(x)$ verwenden wir die Quantilsfunktion $F^{-1}(\alpha) \rightarrow$ QQ (Quantil-Quantil) Diagramme.

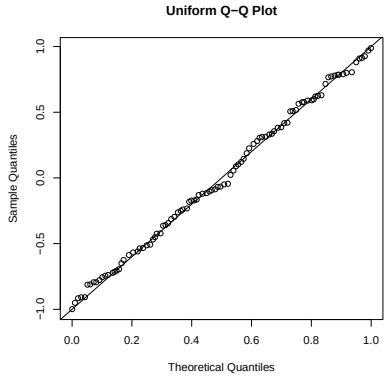
Verwendung sowohl zum Vergleich zweier Stichproben als auch zum Vergleich einer Stichprobe mit den theoretischen Quantilen einer Verteilung. Ersetzt die antiquierte Methode der „Wahrscheinlichkeitspapiere“.

Vorteile:

- F^{-1} „lebt“ für alle Verteilungen auf dem Intervall $[0, 1]$.
- Für mehrere Familien von Verteilungen, inkl. Gleichverteilung und Normalverteilung gilt, daß QQ Diagramme gerade Linien ergeben, auch wenn man falsche Parameter wählt.

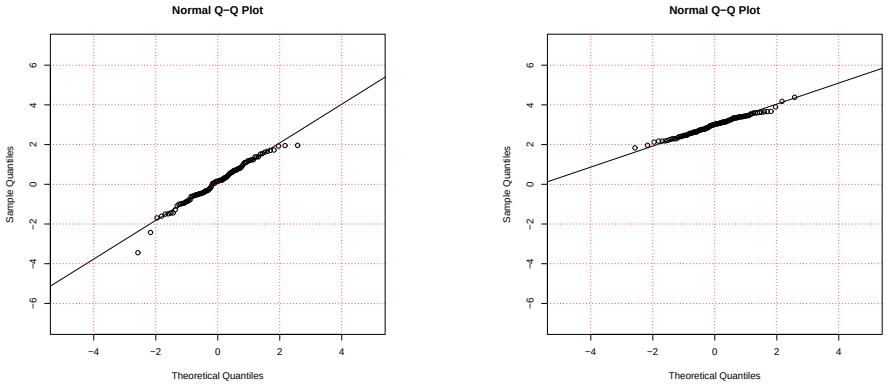
Quantile





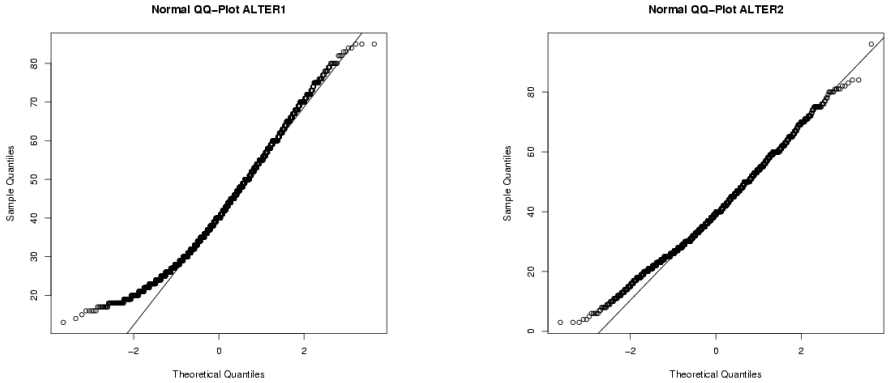
Quantile

Vergleich der QQ Diagramme einer N(0,1) mit einer N(3,0.25):



Gerade wird durch 1. und 3. Quartil der Stichprobe gelegt. Achsenabschnitt entspricht Mittelwert, Steigung der Standardabweichung.

QQ-Plot: ALTER1 und ALTER2



Ab ca. 30 Jahren greift die Normalapproximation recht gut.

QQ-Plot: KOPFAUSG