# Angewandte Stochastik

Prof. Dr. Evgeny Spodarev | Vorlesungskurs |

4. Thema

# Heutiges Thema

- Quantilplots und Kerndichteschätzung

# Quantilplots (Quantil-Grafiken)

- Nach der ersten beschreibenden Analyse eines Datensatzes $(x_1, \ldots, x_n)$ soll überlegt werden, mit welcher Verteilung diese Stichprobe modelliert werden kann.
- Hier sind die sogenannten *Quantilplots* behilflich.
- Sie zeigen graphisch, wie gut die Daten $(x_1, \ldots, x_n)$ mit dem Verteilungsgesetz $G$ übereinstimmen.
- $G$ ist die Verteilungsfunktion einer hypothetischen Verteilung.

# Quantilplots (Quantil-Grafiken)

- Sei $X$ eine Zufallsvariable mit (unbekannter) Verteilungsfunktion $F_X$.
- Auf Basis der Daten $(X_1, \ldots, X_n)$, $X_i$ unabhängig identisch verteilt und $X_i \stackrel{d}{=} X$ möchte man prüfen, ob $F_X = G$ für eine bekannte Verteilungsfunktion $G$ gilt.
- Methode der *Quantil-Grafiken*: Man vergleicht die entsprechenden Quantil-Funktionen $\hat{F}_n^{-1}$ und $G^{-1}$ von $\hat{F}_n$ und $G$ graphisch.

# Quantilplots (Quantil-Grafiken)

Hierzu

- plotte man $G^{-1}(\frac{k}{n})$ gegen $\hat{F}_n^{-1}(\frac{k}{n}) = X_{(k)}, \quad k = 1, \ldots, n.$
- Falls die Punktwolke

$$\left\{ \left( G^{-1}\left(\frac{k}{n}\right), X_{(k)} \right), \quad k = 1, \ldots, n \right\}$$

näherungsweise auf einer Geraden $y = ax + b$ liegt, so sagt man, dass $F_X(x) \approx G\left(\frac{x-a}{b}\right), \ x \in \mathbb{R}$.

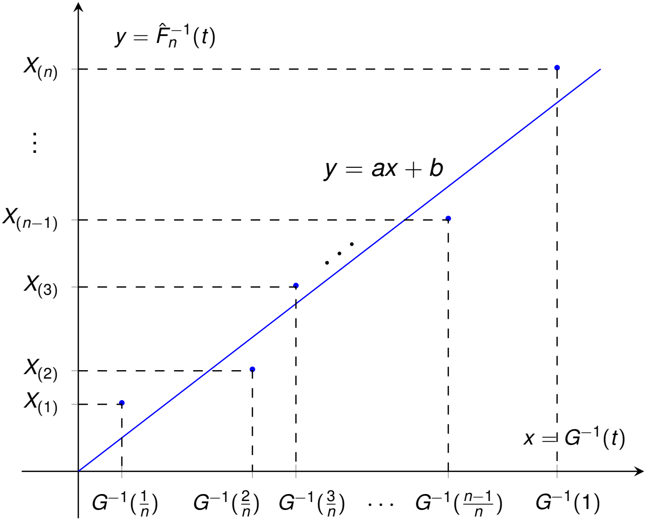



Figure: Quantil-Grafik

# Quantilplots (Quantil-Grafiken)

Diese empirische Vergleichsmethode beruht auf folgenden Überlegungen:

- Man ersetzt die unbekannte Funktion $F_X$ durch die aus den Daten berechenbare Funktion $\hat{F}_n$. Dabei macht man einen Fehler, der allerdings asymptotisch (für $n \to \infty$) klein ist. Dies folgt aus dem Satz von Gliwenko-Cantelli, der besagt, dass

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_X(x) \right| \underset{n \to \infty}{\longrightarrow} 0 \quad \text{f.s.}$$

# Quantilplots (Quantil-Grafiken)

Der Vergleich der entsprechenden Quantil-Funktionen wird durch folgendes Ergebnis bestärkt:

Falls E$X < \infty$, dann gilt

$$\sup_{t\in[0,1]} \left| \int_0^t \left( \hat{F}_n^{-1}(y) - F_X^{-1}(y) \right) dy \right| \xrightarrow[n\to\infty]{\text{f.s.}} 0 \,.$$

$\Rightarrow$ Voraussetzung für Verwendung der Quantil-Grafiken: der Stichprobenumfang $n$ ist ausreichend groß, um $\hat{F}_n^{-1} \approx F_X^{-1}$ zu gewährleisten.

# Quantilplots (Quantil-Grafiken)

- Man setzt zusätzlich voraus, dass die Gleichungen

$$y = ax + b\,,$$
$$y = F_X^{-1}(t)\,,$$
$$x = G^{-1}(t)$$

für alle $t$ (und nicht nur näherungsweise für $t = \frac{k}{n}$, $k = 1, \ldots, n$) gelten.

$\Rightarrow G(x) = t = F_X(y) = F_X(ax + b)$ für alle $x$, oder $F_X(y) = G\left(\frac{y-b}{a}\right)$ für alle $y$, weil $x = \frac{y-b}{a}$ ist.

# Quantilplots (Quantil-Grafiken)

- Aus praktischer Sicht ist es besser, Paare $\left(G^{-1}\left(\frac{k}{n+1}\right), X_{(k)}\right)$, $\quad k = 1, \ldots, n$ zu plotten.
- Dadurch wird vermieden, dass $G^{-1}(n/n) = G^{-1}(1) = \infty$ vorkommt, wie es zum Beispiel bei einer Verteilung $G$ der Fall ist, bei der $F(x) < 1$ gilt für alle $x \in \mathbb{R}$.
- Tatsächlich gilt für $k = n$, dass $\frac{n}{n+1} < 1$ und somit $G^{-1}\left(\frac{n}{n+1}\right) < \infty$.

# Beispiel (Exponential-Verteilung, $G(x) = (1 - e^{-\lambda x}) \cdot I(x \geq 0)$)

Es gilt $G^{-1}(y) = -\frac{1}{\lambda} \log(1-y)\,, \quad y \in (0,1)$. So wird man beim Quantil-Plot Paare

$$\left( -\frac{1}{\lambda} \log \left( 1 - \frac{k}{n+1} \right), X_{(k)} \right)\,, \quad k = 1, \ldots, n$$

zeichnen, wobei der Faktor $\frac{1}{\lambda}$ für die Linearität unwesentlich ist und weggelassen werden kann.

# Beispiel (Normalverteilung, $G(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt\,, \quad x \in \mathbb{R}$)

Analytische Berechnung von $\Phi^{-1}$ mit einer geschlossenen Formel nicht möglich.

Aus diesem Grund wird $\Phi^{-1}\left(\frac{k}{n+1}\right)$ numerisch berechnet und in Tabellen oder statistischen Software-Paketen (wie z.B. R) abgelegt.

Um die empirische Verteilung der Daten mit der Normalverteilung zu vergleichen, trägt man Punkte mit Koordinaten

$$\left(\Phi^{-1}\left(\frac{k}{n+1}\right), X_{(k)}\right), \quad k = 1, \ldots, n$$

auf der Ebene auf und prüft, ob sie eine Gerade bilden (vgl. Abb. auf der nächsten Folie).

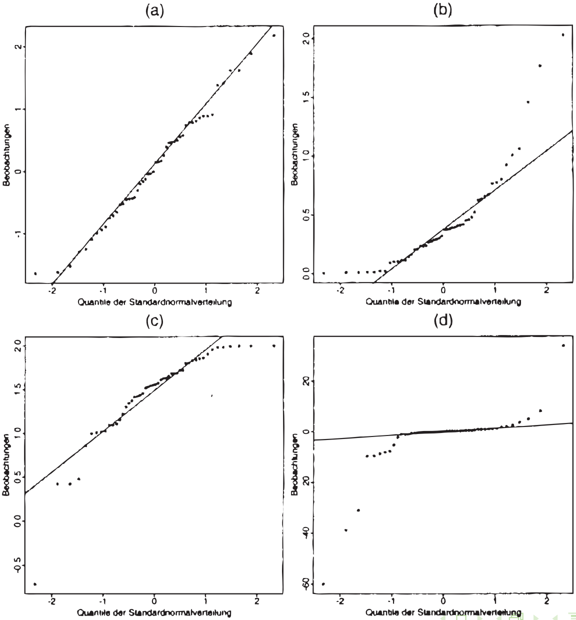

(a)
(b)
(c)
(d)
Beobachtungen
Quantile der Standardnormalverteilung

Figure: QQ-Plot einer Normalverteilung (a), einer linkssteilen Verteilung (b), einer rechtssteilen Verteilung (c) und einer symmetrischen, aber stark gekrümmten Verteilung (d).

- Falls $\bar{x}_n = 0$ und die Verteilung $F_X$ linkssteil ist, so sind die Quantile von $F_X$ kleiner als die von $\Phi$.

  $\Rightarrow$ Der Normal-Quantilplot ist konvex.

- Falls $\bar{x}_n = 0$ und $F_X$ rechtssteil ist, so wird der Normal-Quantilplot konkav sein.

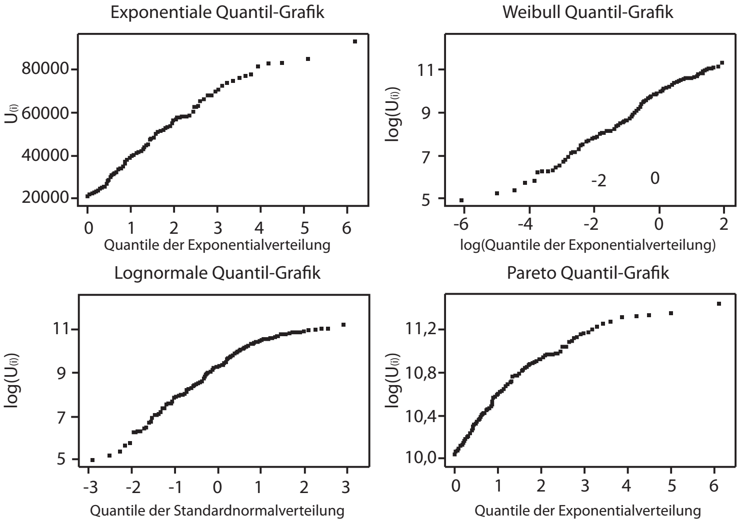



Figure: Ordnungsstatistiken einer Stichprobe von Schadenhöhen der Industrie-Unfälle in Belgien im Jahr 1992

# Beispiel (Haftpflichtversicherung (Belgien, 1992))

- In obiger Abbildung sind Ordnungsstatistiken der Stichprobe von $n = 227$ Schadenhöhen der Industrie-Unfälle in Belgien im Jahr 1992 (Haftpflichtversicherung) gegen Quantile von Exponential-, Pareto-, Standardnormal- und Weibull-Verteilungen geplottet.
- Im Bereich von Kleinschäden zeigen die Exponential- und Pareto-Verteilungen eine gute Übereinstimmung mit den Daten.
- Die Verteilung von mittelgroßen Schäden kann am besten durch die Lognormal- und Weibull-Verteilungen modelliert werden.
- Für Großschäden erweist sich die Weibull-Verteilung als geeignet.

# Beispiel (Rendite der BMW-Aktie)

In der folgenden Abbildung ist der Quantilplot für Renditen der BMW-Aktie beispielhaft zu sehen.

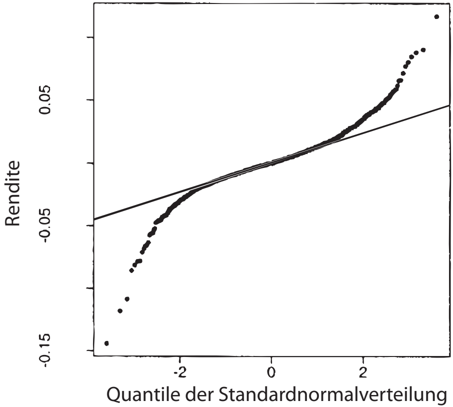



Figure: Quantilplot der Rendite der BMW-Aktie

# Kerndichteschätzung

- Sei eine Stichprobe $(x_1, \ldots, x_n)$ von unabhängigen Realisierungen eines absolut stetig verteilten Merkmals $X$ mit Dichte $f_X$ gegeben.
- Mit Hilfe der Histogramme lässt sich $f_X$ graphisch durch eine Treppenfunktion $\hat{f}_X$ darstellen.
- Dabei gibt es zwei entscheidende Nachteile der Histogrammdarstellung:
  1. Willkür in der Wahl der Klasseneinteilung $[c_{i-1}, c_i]$,
  2. Eine (möglicherweise) stetige Funktion $f_X$ wird durch eine Treppenfunktion $\hat{f}_X$ ersetzt.
- Auf den folgenden Folien werden wir versuchen, diese Nachteile beseitigen, indem wir eine Klasse von Kerndichtenschätzern einführen, die (je nach Wahl des Kerns) auch zu stetigen Schätzern $\hat{f}_X$ führen.

# Definition

Der Kern $K(x)$ wird definiert als eine nicht-negative messbare Funktion auf $\mathbb{R}$ mit der Eigenschaft $\int_{\mathbb{R}} K(x)\,dx = 1$.

# Definition

Der *Kerndichteschätzer* der Dichte $f_X$ aus den Daten $(x_1, \ldots, x_n)$ mit Kernfunktion $K(x)$ ist gegeben durch

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R},$$

wobei $h > 0$ die sogenannte *Bandbreite* ist.

# Beispiele für Kerne

1. *Rechteckskern:*

$$K(x) = \frac{1}{2} \cdot I(x \in [-1,1))\,.$$

Dabei ist

$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \begin{cases} \frac{1}{(2h)}\,, & x_i - h \le x < x_i + h\,, \\ 0\,, & \text{sonst,} \end{cases}$$

und somit

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{k} K\left(\frac{x - x_i}{h}\right) = \frac{\#\{x_i \in [x - h, x + h)\}}{2nh}\,,$$

das auch *gleitendes Histogramm* genannt wird. Dieser Dichteschätzer ist (noch) nicht stetig, was durch die (besonders einfache rechteckige unstetige) Form des Kerns erklärt wird.

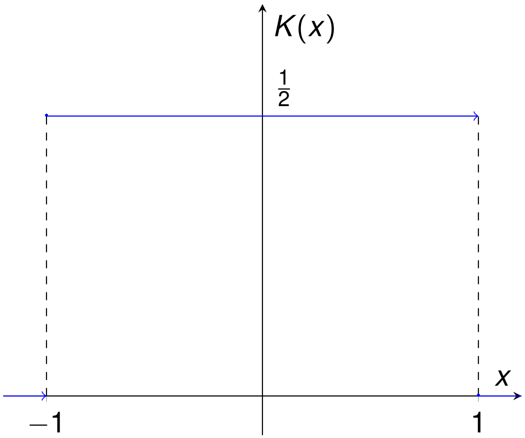
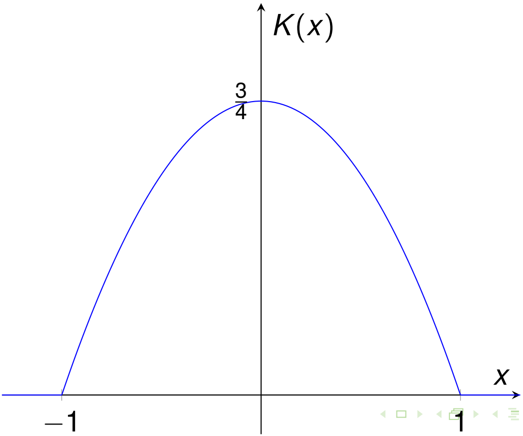



Figure: Rechteckkern

# Beispiele für Kerne

2. *Epanechnikov-Kern:*

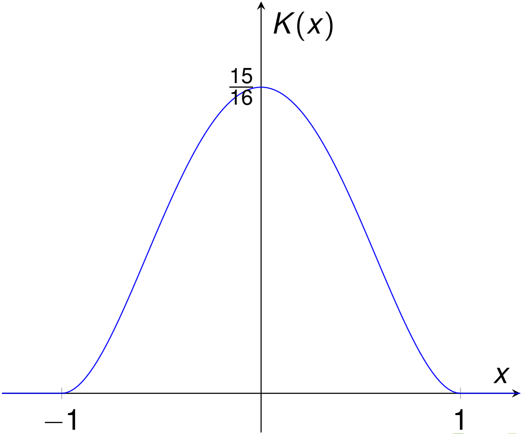
$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2)\,, & x \in [-1, 1) \\ 0\,, & \text{sonst.} \end{cases}$$



# Beispiele für Kerne

3. *Bisquare-Kern:*

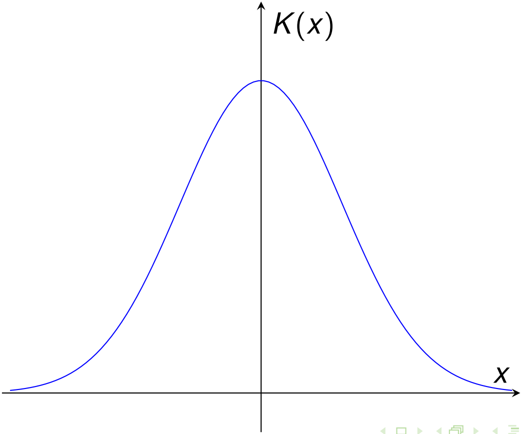
$$K(x) = \frac{15}{16}\left((1-x^2)^2 \cdot I(x \in [-1,1))\right) .$$



# Beispiele für Kerne

4. *Gauss-Kern:*

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$



- Dabei ist die Wahl der Bandbreite $h$ entscheidend für die Qualität der Schätzung.
- Je größer $h > 0$, desto glatter wird $\hat{f}_X$ sein und desto mehr "Details" werden "herausgemittelt".
- Für kleinere $h$ wird $\hat{f}_X$ rauer.
- Dabei können aber auch Details auftreten, die rein stochastischer Natur sind und keine Gesetzmäßigkeiten zeigen.
- Mit der adäquaten Wahl von $h$ beschäftigen sich viele wissenschaftliche Arbeiten, die empirische Faustregeln, aber auch kompliziertere Optimierungsmethoden dafür vorschlagen.