Der Mittelwert (arithmetisches Mittel)

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

- bekanntestes Lagemaß
- instabil gegen extreme Werte
- geeignet für intervallskalierte Daten

Mittelwert bei gruppierten Daten

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \\ &= \frac{1}{n}\sum_{j=1}^{k} h_j a_j \end{aligned}$$

h_j : Häufigkeit von a_j

Das geometrische Mittel

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- arithmetisches Mittel auf der log-Skala

$$x_g = exp\left(\frac{1}{n}\sum_{i=1}^{n} log(x_i)\right)$$

- nur geeignet für positive Werte
- geeignet für intervallskalierte Daten

Das harmonische Mittel

$$\bar{x}_H := \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Das harmonische Mittel entspricht dem Mittel durch Transformation

$$t \to \frac{1}{t} \qquad \bar{x}_H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

Das harmonische Mittel

$$\bar{x}_H := \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Das harmonische Mittel entspricht dem Mittel durch Transformation

$$t \to \frac{1}{t} \qquad \bar{x}_H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

Beispiel:

$x_1, \ldots, x_n$ Geschwindigkeiten, mit denen konstante Wegstrecken L zurückgelegt werden

Gesamt-Geschwindigkeit:

$$\frac{L \cdot n}{\frac{L}{x_1} + \ldots + \frac{l}{x_n}} = \bar{x}_H$$

Allgemeine Transformation des Mittelwerts I

Lineare Transformation:

$$\begin{aligned} g(t) &= a + bt \\ y_i &= a + bx_i \quad \Rightarrow \quad \bar{y} = a + b\bar{x} \end{aligned}$$

d.h.

$$\begin{aligned} \overline{a + bx} &= a + b\bar{x} \\ \overline{g(x)} &= g(\bar{x}) \end{aligned}$$

Allgemeine Transformation:

Generell ist $\overline{g(x)} \neq g(\bar{x})$

Allgemeine Transformation des Mittelwerts II

Für konvexe Funktionen g gilt:

$$\begin{aligned} g(\bar{x}) &\leq \overline{g(x)} \\ g\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) &\leq \frac{1}{n}\sum_{i=1}^{n} g(x_i) \qquad \text{(Jensen-Ungleichung)} \end{aligned}$$

g konvex: $\Leftrightarrow$ $\quad g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$
$\forall \lambda \in [0,1], \quad x, y \in D_g$

Beispiel:
$\bar{x}^2 \leq \overline{x^2}$

Vergleich I

Es gilt allgemein für positive x_i

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}$$

Beweis:

a) Die Funktion $g : t \to log(t)$ ist konkav,
da $g''(t) = -\frac{1}{t^2} < 0$

$$\Rightarrow log(\bar{x}) \geq \overline{log(x)}$$

$$\Rightarrow \bar{x} \geq exp\left(\overline{log(x)}\right) = exp\left(\frac{1}{n}\sum_{i=1}^{n} log(x_i)\right)$$

$$= \left(\prod_{i=1}^{n} exp\left(log(x_i)\right)\right)^{\frac{1}{n}} = \bar{x}_G$$

Vergleich II

b) Die Funktion $g_2 : t \to \frac{1}{exp(t)}$ ist konvex,
da $g_2''(t) = \frac{1}{exp(t)} \geq 0$

Daten $log(x_1), \ldots, log(x_n)$

$$g_2\left(\frac{1}{n}\sum_{i=1}^{n} log(x_i)\right) \leq \frac{1}{n}\sum_{i=1}^{n}\left(exp(log(x_i))^{-1}\right)$$

$$\Rightarrow \frac{1}{\sqrt[n]{\prod_{i=1}^{n} x_i}} \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

$$\Rightarrow \underbrace{\sqrt[n]{\prod_{i=1}^{n} x_i}}_{x_G} \geq \underbrace{\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}}_{x_H}$$

Getrimmtes Mittel

Um die Ausreißerempfindlichkeit von $\bar{x}$ abzuschwächen definiert man

$$\bar{x}_\alpha = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} x_{(i)}$$

$x_{(i)}$: geordnete x-Werte
r ist die größte ganze Zahl mit $r \leq n\alpha$

Es wird also der Anteil α der extremsten Werte abgeschnitten.

„α-getrimmtes Mittel“

Winsorisiertes Mittel (gestutztes Mittel)

Der Anteil α der extremsten Werte wird durch das entsprechende Quantil ersetzt.

Maße für die Streuung

- Spannweite
- Interquartilsabstand
- Standardabweichung und Varianz
- Variationskoeffizient

Die Spannweite (Range)

Definition:

$$q = x_{max} - x_{min}$$

- „Bereich in dem die Daten liegen“
- Wichtig für Datenkontrolle

Der Quartilsabstand

Definition:

$$d_Q = x_{0.75} - x_{0.25}$$

- „Größe des Bereichs in dem die mittlere Hälfte der Daten liegt“
- Bei ordinal skalierten Daten Angabe von $x_{0.75}$ und $x_{0.25}$
- Zentraler 50%-Bereich
- Robust gegen Ausreißer

Standardabweichung und Varianz

Definition

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \textit{Varianz}$$

$$S = \sqrt{S^2} \qquad \textit{Standardabweichung}$$

- Verwende $\widetilde{S}^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ für Vollerhebung
  Division durch $n-1$ eigentlich nur bei Stichproben sinnvoll
- „Mittlere Abweichung vom Mittelwert“
- Intervallskala Voraussetzung
- Empfindlich gegen Ausreißer

Transformationsregel

$y_i = a + bx_i$

$$\begin{aligned} \Rightarrow \widetilde{S}_y^2 &= b^2 \widetilde{S}_x^2 \\ \widetilde{S}_y &= |b| \widetilde{S}_x \quad \text{(Analog für } S_x, S_y\text{)} \end{aligned}$$

Varianz und Standardabweichung sind stabil
mit linearen Transformationen verträglich.

Verschiebungssatz

Für jedes $c \in \mathbb{R}$ gilt:

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - c)^2$$

$$\begin{aligned} c = 0 \Rightarrow \widetilde{S}^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \\ \widetilde{S}^2 &= \overline{x^2} - \bar{x}^2 \end{aligned}$$

Beachte:
Verschiebungssatz für numerische Berechnung mit Computer nicht geeignet.

Streuungszerlegung I

Seien die Daten in r Schichten aufgeteilt:

$$x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n_1+n_2}, \ldots, x_{n_r}$$

Schichtmittelwerte:

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i, \quad \bar{x}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} x_i, \quad \text{usw.}$$

Schichtvarianzen:

$$\widetilde{S}_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2, \quad \widetilde{S}_2^2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2)^2, \quad \text{usw.}$$

Streuungszerlegung II

Dann gilt:

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{j=1}^{r} n_j \bar{x}_j \\ \widetilde{S}^2 &= \frac{1}{n}\sum_{j=1}^{r} n_j \widetilde{S}_j^2 + \frac{1}{n}\sum_{j=1}^{r} n_j (\bar{x}_j - \bar{x})^2 \end{aligned}$$

$$\text{Gesamtstreuung} = \begin{array}{c}\text{Streuung}\\ \text{innerhalb}\\ \text{der Schicht}\end{array} + \begin{array}{c}\text{Streuung}\\ \text{zwischen}\\ \text{den Schichten}\end{array}$$

Variationskoeffizient

Das Verhältnis von Standardabweichung und Mittelwert ist gegeben durch

$$v = \frac{\widetilde{S}}{\bar{x}} \quad \text{mit } \bar{x} > 0$$

Der Variationskoeffizient hat keine Einheit und ist skalenunabhängig.

Er ist eine Maßzahl für die relative Schwankung um den Mittelwert.

Mittlere absolute Abweichung (MAD)

Die mittlere absolute Abweichung ist definiert als

$$MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

$$MedAD := median(|x_i - x_{med}|)$$

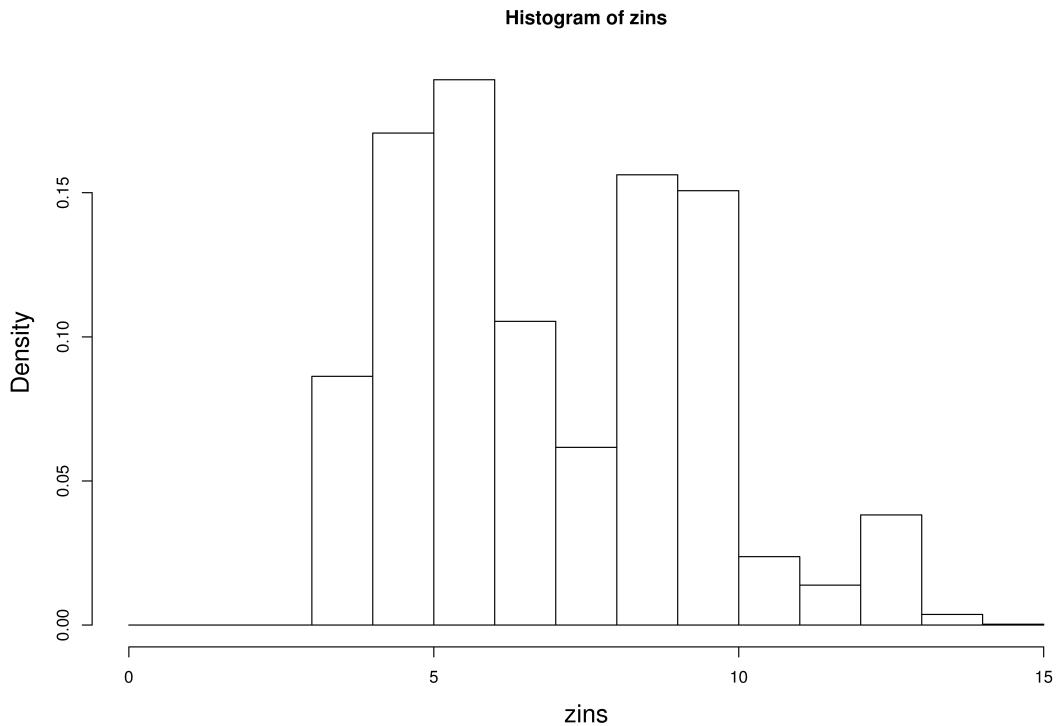
Wegen der Jensen-Ungleichung gilt: $MAD \leq \widetilde{S}$

MAD/ *MedAD* :

- nicht so „schöne“ theoretische Eigenschaften
- klarer interpretierbar als $\widetilde{S}$
- weniger Ausreißer-empfindlich

Uni- und multimodale Verteilungen

unimodal = eingipflig, multimodal = mehrgipflig

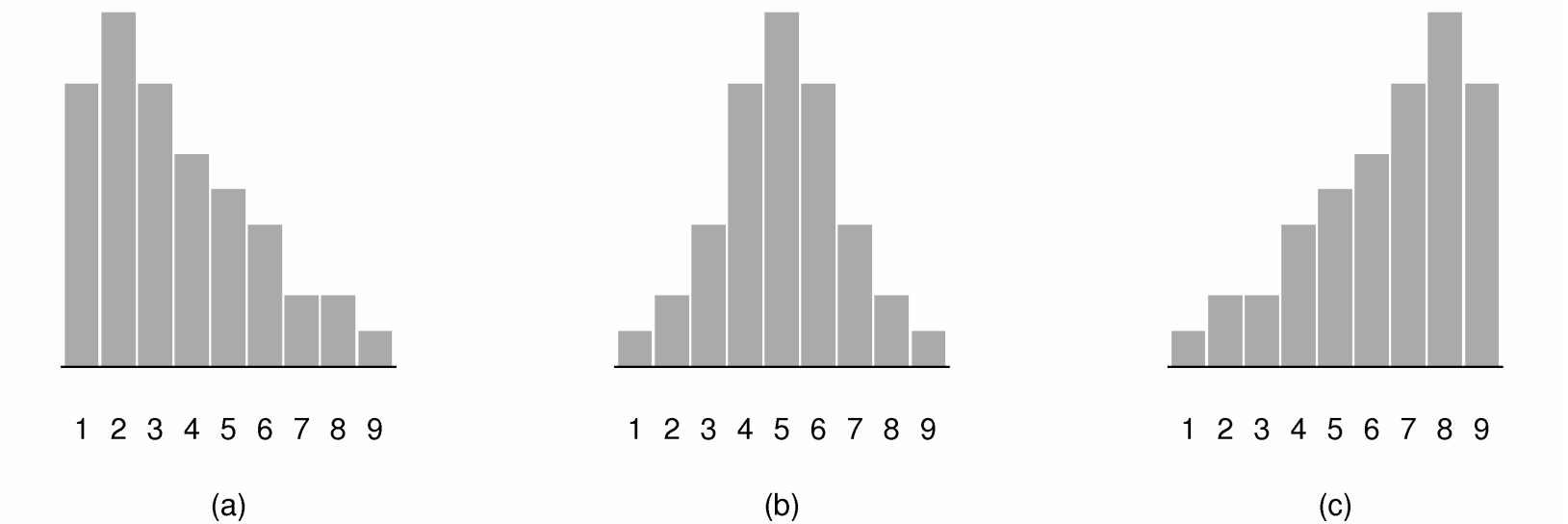


Das Histogramm der Zinssätze zeigt eine bimodale Verteilung.

Symmetrie und Schiefe I

symmetrisch	$\Leftrightarrow$	Rechte und linke Häfte der Verteilung sind annähernd zueinander spiegelbildlich
linkssteil (rechtsschief)	$\Leftrightarrow$	Verteilung fällt nach links deutlich steiler und nach rechts langsamer ab
rechtssteil (linksschief)	$\Leftrightarrow$	Verteilung fällt nach rechts deutlich steiler und nach links langsamer ab

Symmetrie und Schiefe II



Eine linkssteile (a), symmetrische (b) und rechtssteile Verteilung (c)

Lageregeln

Symmetrische und unimodale Verteilung:

$\bar{x} \approx x_{med} \approx x_{mod}$

Linkssteile Verteilung:

$\bar{x} > x_{med} > x_{mod}$

Rechtssteile Verteilung:

$\bar{x} < x_{med} < x_{mod}$

Bei gruppierten Daten: Auch für Histogramme gültig

Beachte:
Form der Verteilung bleibt bei linearen Transformationen gleich.
Änderung bei nichtlinearen Transformationen.

Maßzahlen für die Schiefe I

Quantilskoeffizient:

$$g_p = \frac{(x_{1-p} - x_{med}) - (x_{med} - x_p)}{x_{1-p} - x_p}$$

$p = 0.25$ Quartilskoeffizient

Werte des Quantilskoeffizienten:

$g_p = 0$ für symmetrische Verteilungen
$g_p > 0$ für linkssteile Verteilungen
$g_p < 0$ für rechtssteile Verteilungen

Maßzahlen für die Schiefe II

Momentenkoeffizient der Schiefe

$$g_m = \frac{m_3}{\widetilde{s}^3} \quad \text{mit} \quad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Werte des Momentenkoeffizienten

$g_m = 0$ für symmetrische Verteilungen
$g_m > 0$ für linkssteile Verteilungen
$g_m < 0$ für rechtssteile Verteilungen

Dichtefunktion I

Histogramm:

Anteil = Fläche unter der Kurve

Histogramm ist stückweise konstante Funktion

Probleme: Abhängigkeit von der Wahl der Klassengrenzen

Ersetze Histogramm durch glatte Funktion f

Dichtefunktion II

Dichte

Eine positive stetige Funktion heißt Dichte(-funktion), wenn $f(x) \geq 0$ und

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Die Flächen unter der Dichte sollen den approximativen Häufigkeiten entsprechen, d.h.

$$\int_{a}^{b} f(x)dx \approx \frac{1}{n}\#\{x_i | a < x_i \leq b\}$$

Dichte und Verteilung

$$F(x_0) = \frac{1}{n}\#\{x_i | x_i \leq x_0\} \approx \int_{-\infty}^{x_0} f(x)dx$$

$$\hat{F}(x_0) = \int_{-\infty}^{x_0} \hat{f}(x)dx$$

Quantile

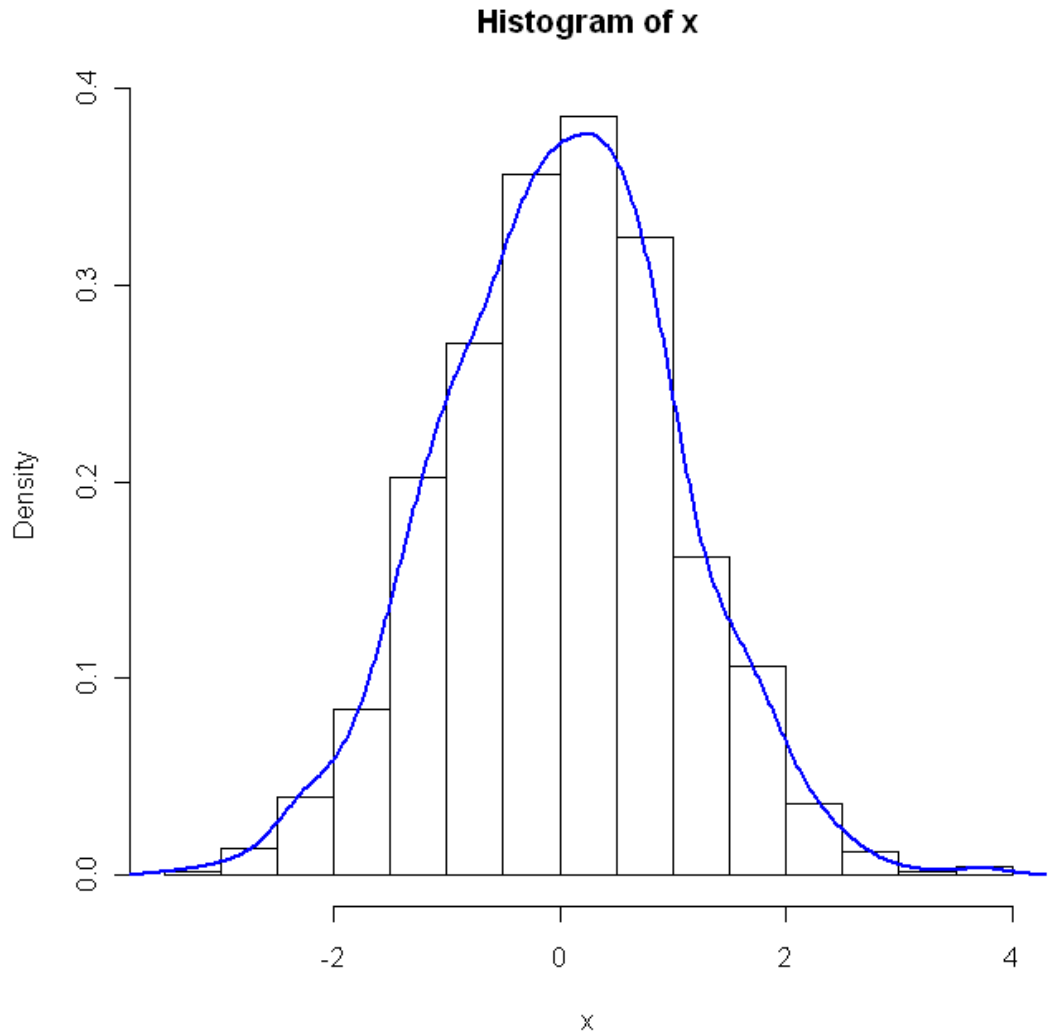
$0 < p < 1$

x_p ist der Wert auf der x-Achse, für den gilt:

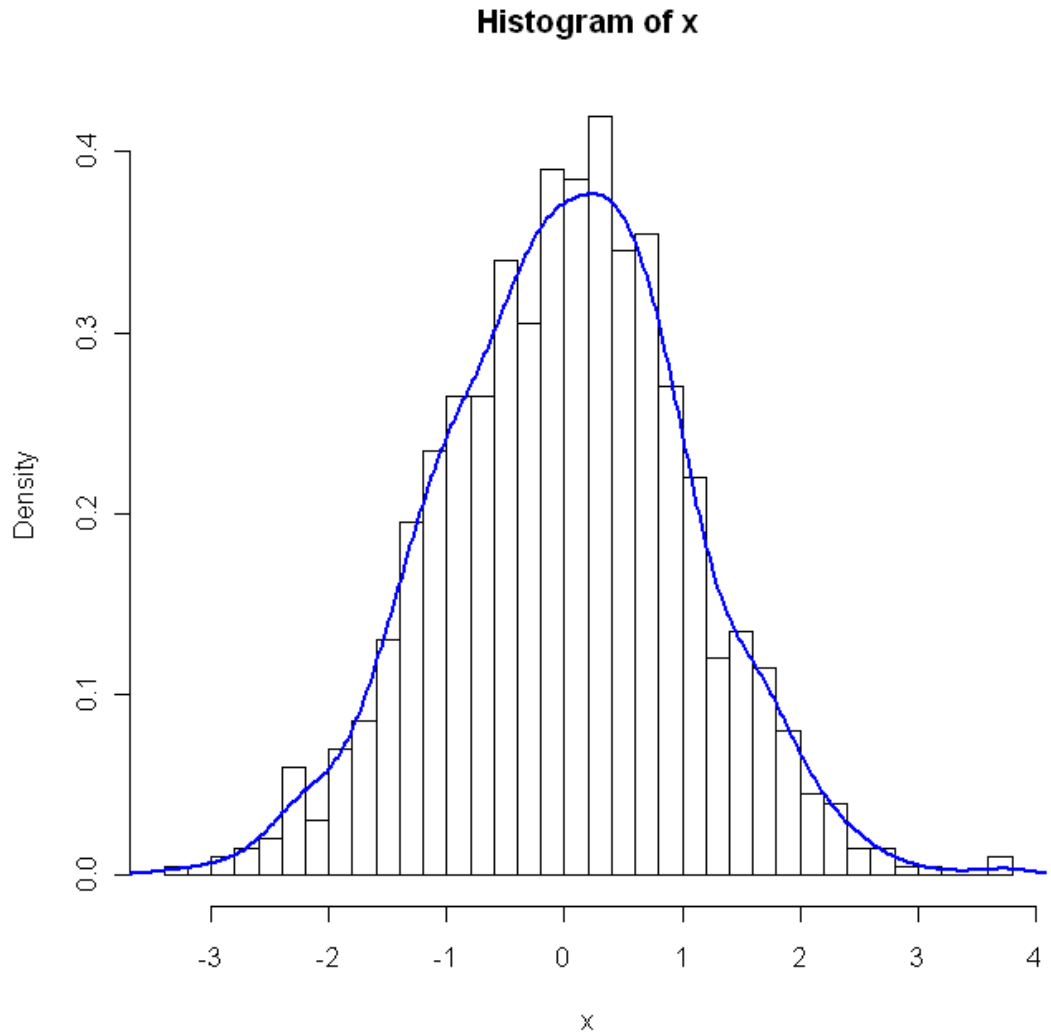
$$\int_{-\infty}^{x_p} f(x)dx = p$$

Der Median teilt die Fläche in 2 gleich große Teile.

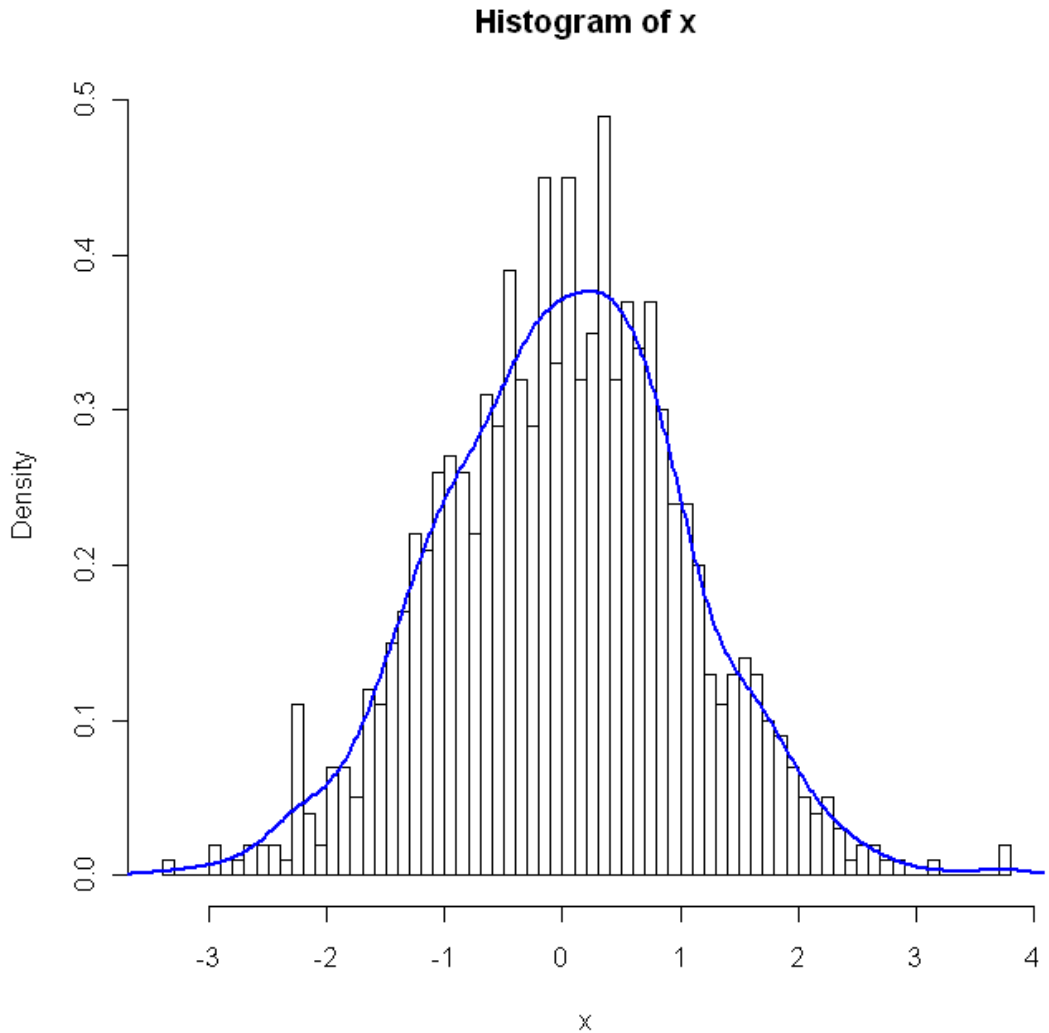
Beispiele Histogramm und Dichte



Beispiele Histogramm und Dichte



Beispiele Histogramm und Dichte



Berechnung von Dichte-Kurven

$$\hat{f}(x) = \frac{\frac{1}{n}\#\{x_i | x_i \in [x-h, x+h)\}}{2h}$$

$\Rightarrow$ „Gleitendes Histogramm“

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

mit $K(u) = \begin{cases} \frac{1}{2} & \text{für } -1 \leq u < 1 \\ 0 & \text{sonst} \end{cases}$

K : Kernfunktion

Kern-Dichteschätzer

$K(u)$ sei Kernfunktion, d.h.

$K(u) \geq 0$ und $\int_{-\infty}^{\infty} K(u)du = 1$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

heißt Kern-Dichteschätzer

Kerne:

Epanechnikov-Kern	$K(u) = \frac{3}{4}(1 - u^2)$	für $-1 \leq u < 1$, 0 sonst.
Bisquare-Kern	$K(u) = \frac{15}{16}(1 - u^2)^2$	für $-1 \leq u < 1$, 0 sonst.
Gauß-Kern	$K(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}u^2\right)$	für $u \in \mathbb{R}$

Bemerkungen zur Dichteschätzung

- Abhängigkeit von der Bandweite $h \rightarrow$ Verfahren zur Bestimmung von h aus den Daten
- Abhängigkeit von der Wahl des Kerns eher unbedeutend
- Kerndichteschätzungen sind insbesondere bei größeren Datenmengen Histogrammen vorzuziehen