Vorlesung: Statistik I für Studierende der Statistik, Mathematik & Informatik

Fabian Scheipl

Wiederholung & Beispiele: Univariate Deskriptive Statistik

Univariate Deskriptive Statistik

Ziel:

Ehrliche, präzise Beschreibung von (relevanten/interessanten Aspekten von) Daten

- mit möglichst geringem Informationsverlust
- möglichst gut verständlich

Mittel:

1. Grafiken
2. Statistische Kennzahlen

Notation & Begriffe

- **Merkmale** $X, Y, \ldots$: zB Alter, Geschlecht, Nettomiete, etc. . .
- **Untersuchungseinheiten**: Objekte/Personen an denen *Merkmale* gemessen wurden (auch: **Merkmalsträger**), davon n in der **Stichprobe**.
- $x_i, i \in \{1, \ldots, n\}$: **beobachteter** Wert bzw. **Merkmalsausprägung** von X für i-te Beobachtung
- $x_1, \ldots, x_n$ **Rohdaten**, **Urliste**

Häufigkeitsverteilung

Nicht alle **Untersuchungseinheiten** haben *unterschiedliche* **Merkmalsausprägungen**

$\Rightarrow$ Fasse **Urliste** $x_1, \ldots, x_n$ in Häufigkeitstabelle der (geordneten) unterschiedlichen Werte $a_1 < a_2 < \ldots < a_k, \ k \leq n$ die gemessen wurden zusammen:

```
url <- "http://www.statistik.lmu.de/service/datenarchiv/miete/miete03.asc"
mietspiegel <- read.table(file = url, header = TRUE)
mietspiegel$lage <- with(mietspiegel, ordered(wohnbest, labels = c("gut", "beste")))
klein_und_kalt <- subset(mietspiegel, zh0 == 1 & wfl < 60)
# URLISTE:
klein_und_kalt[, "rooms"]
```

```
##  [1] 2 2 2 2 2 2 2 2 2 2 2 2 1 3 2 3 2 2 1 2 3 2 2 2 1 2 2 3 1 2 2 2 2 2 2
## [36] 2 1 2 2 2 1 1 2 2 2 2 2 2 1 2 2 2 3 1 3 2 2 2 2 2 2
```

```
# WERTE / MERKMALSAUSPRÄGUNGEN:
sort(unique(klein_und_kalt[, "rooms"]))
```

```
## [1] 1 2 3
```

Häufigkeiten

- $h_j =$ **absolute Häufigkeit** von a_j: wie oft kommt dieser Wert in der Urliste vor?
- $f_j = h_j/n$ **relative Häufigkeit** von a_j: wie groß ist der Anteil von Untersuchungseinheiten mit genau diesem Wert in der Urliste?

```
cbind(
  absolut            =        table(klein_und_kalt[,"rooms"]),
  relativ            =        table(klein_und_kalt[,"rooms"])/nrow(klein_und_kalt),
  `kumulativ, F(x)` = cumsum(table(klein_und_kalt[,"rooms"])/nrow(klein_und_kalt)))
```

```
##   absolut relativ kumulativ, F(x)
## 1       9   0.148            0.15
## 2      46   0.754            0.90
## 3       6   0.098            1.00
```

- **Absolute Häufigkeitsverteilung:** $h_1, \ldots, h_k$
- **Relative Häufigkeitsverteilung:** $f_1, \ldots, f_k$

Empirische Verteilungsfunktion

Fasse komplette Information in **absoluter/relativer Häufigkeitsverteilung** in einer Funktion zusammen:

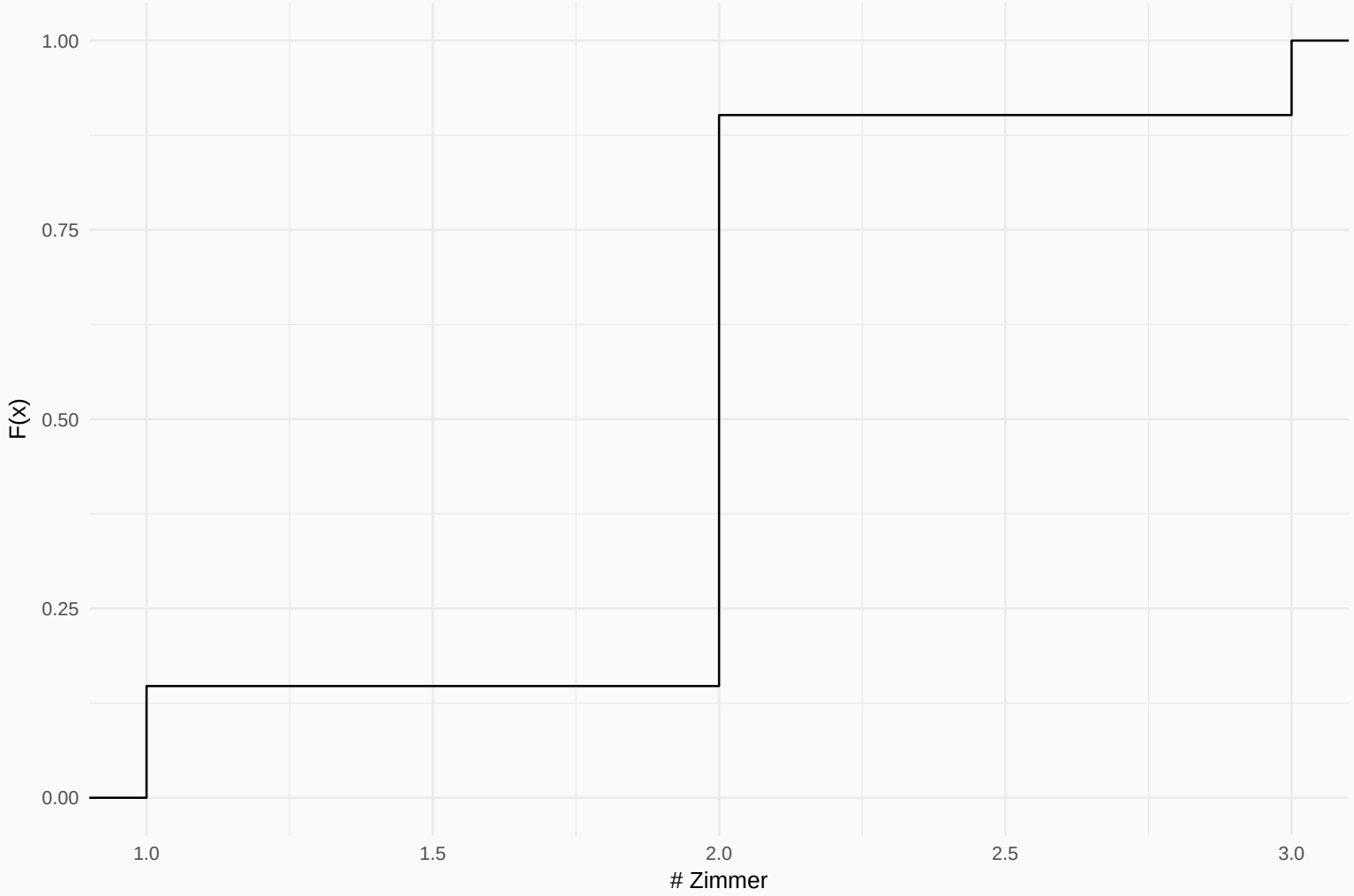
Häufigkeitsfunktion:

$$H(x) := (\text{Anzahl der Werte} <= x) = \sum_{i:a_i \leq x} h_i$$

Verteilungsfunktion:

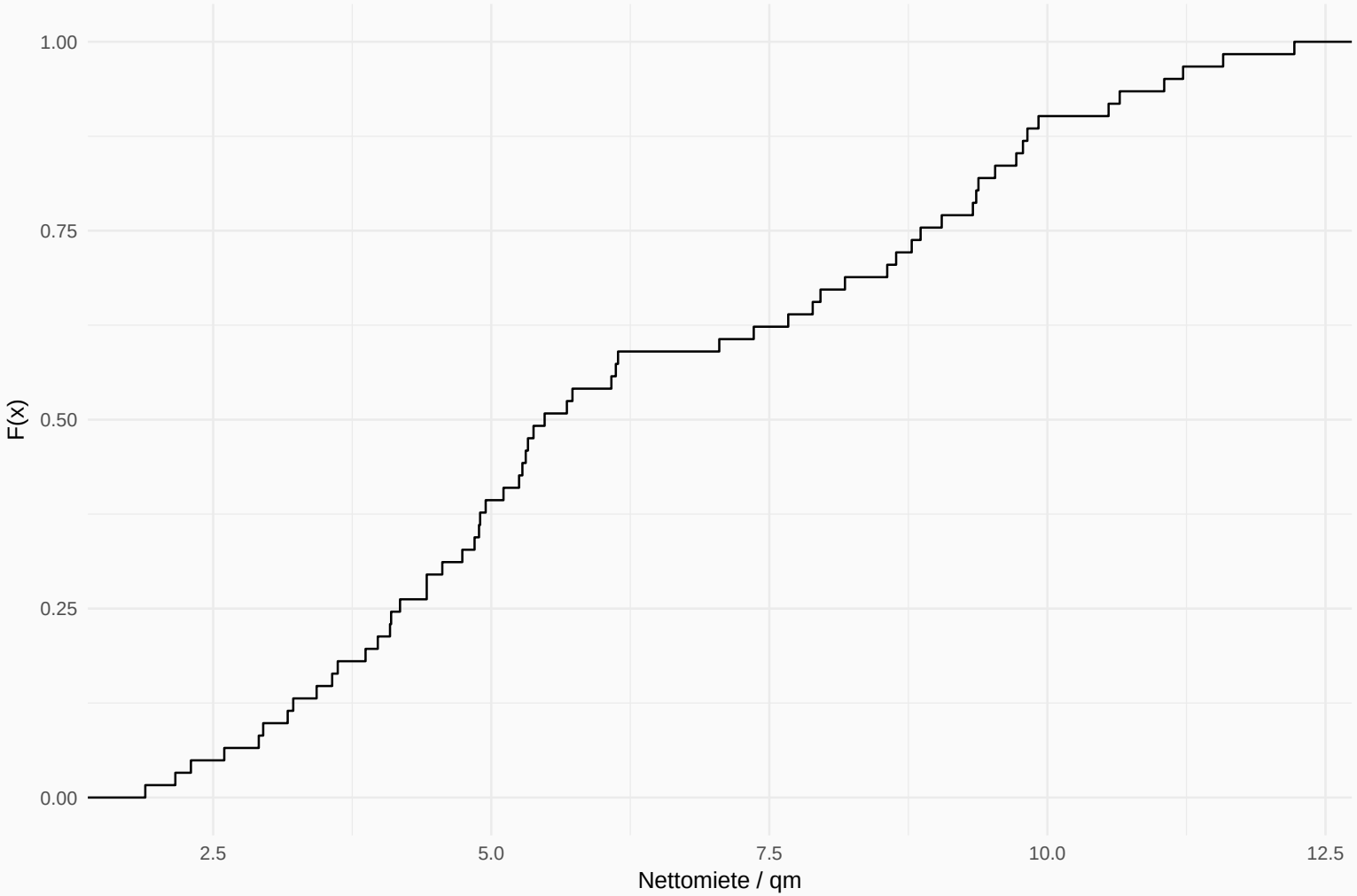
$$F(x) = H(x)/n = \text{Änteil der Werte } \; x_i \; \text{ mit } \; x_i \leq x - \sum_{i:a_i \leq x} f_i$$

Beispiel: F(Zimmerzahl) für Klein & Kalt



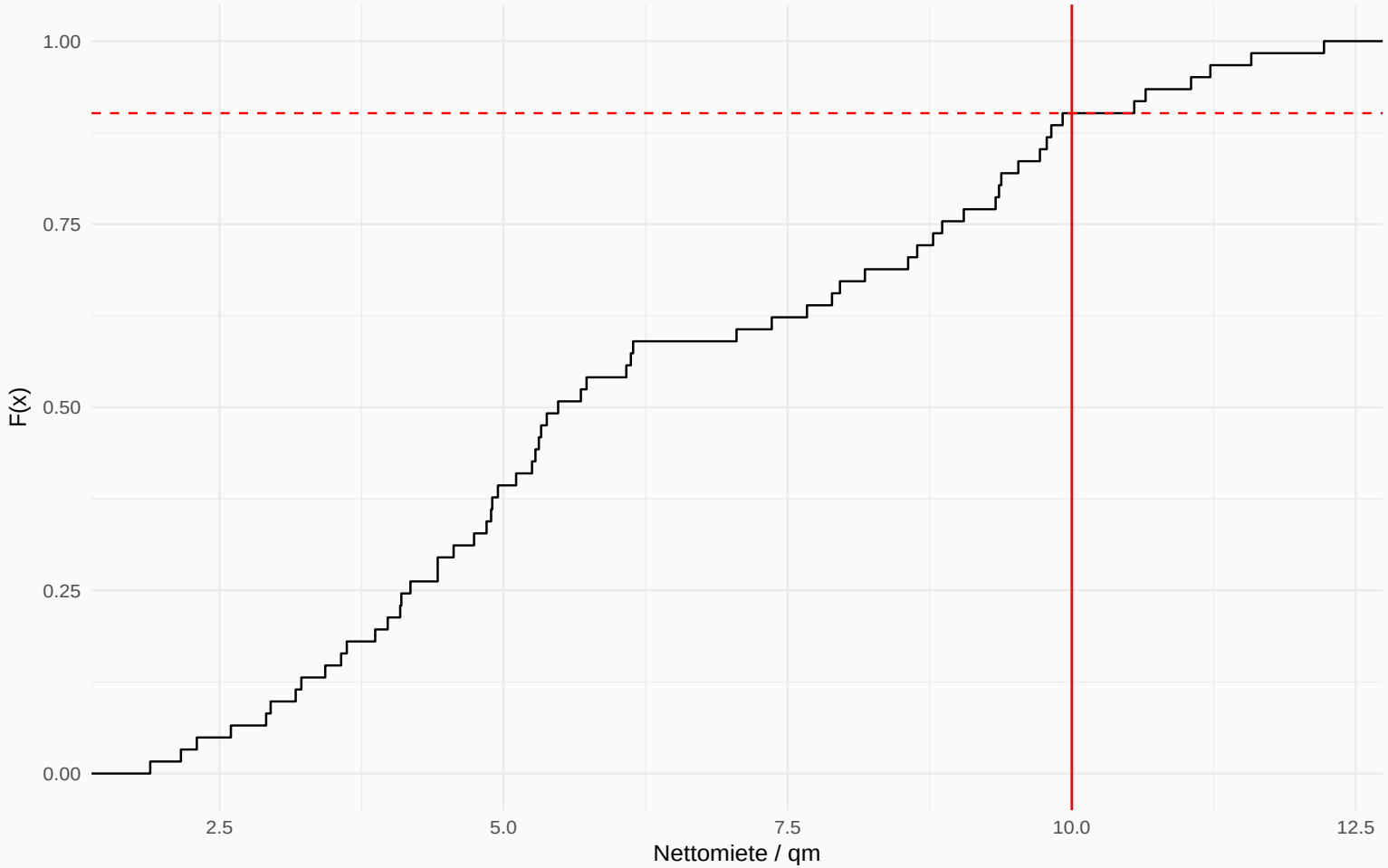
(Vorsicht: Sprungfunktion, oberes Ende der Stufen ist Funktionswert an Sprungstelle!)

Beispiel: F(Quadratmetermiete) für Klein & Kalt



Beispiel: F(Quadratmetermiete) für Klein & Kalt

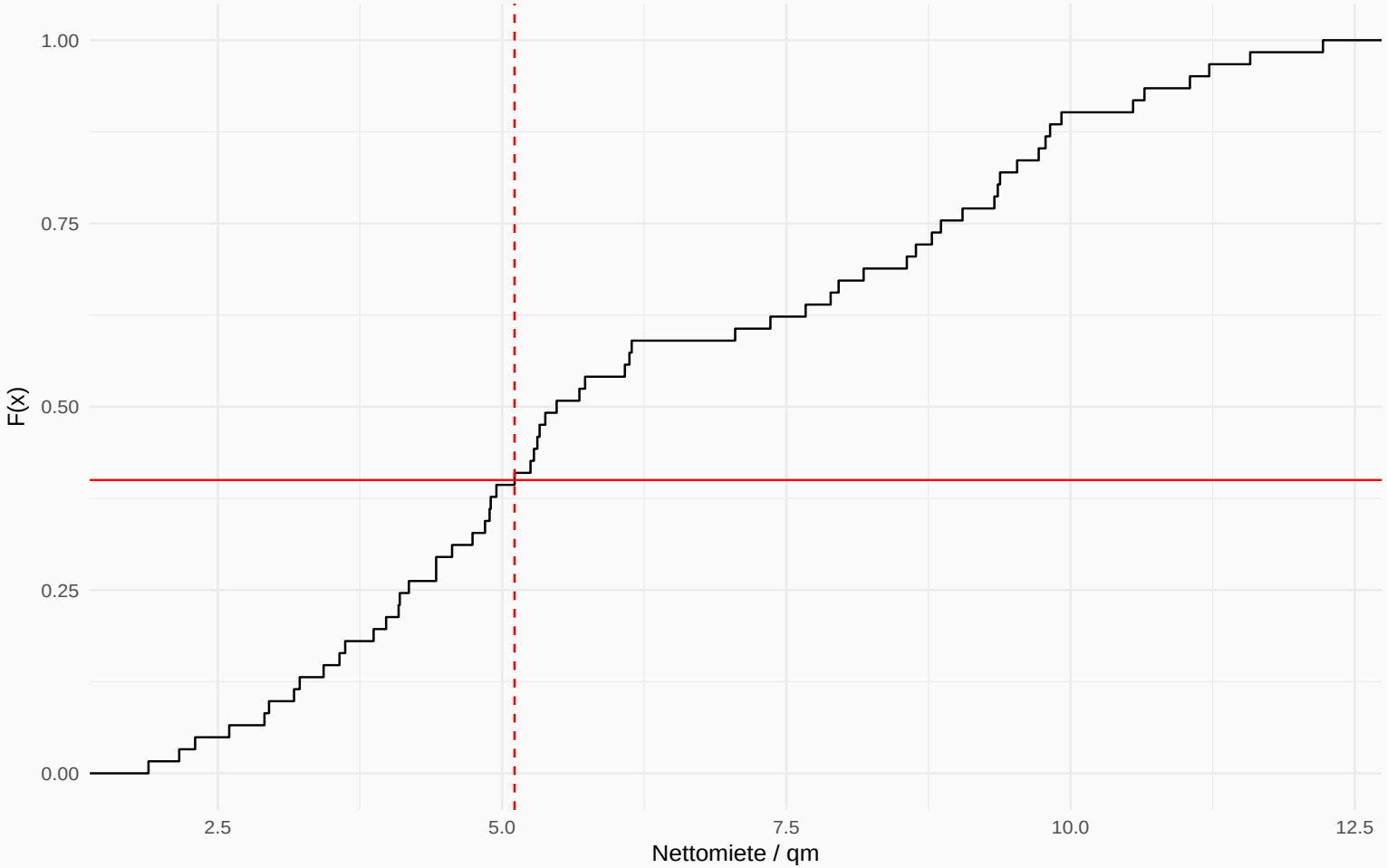
Wie groß ist der Anteil von Wohnungen mit Quadratmetermiete < 10 €?



$\Rightarrow \approx 90\%$

Beispiel: F(Quadratmetermiete) für Klein & Kalt

Wie teuer sind die günstigsten 30% der Wohnungen höchstens? (also: Was ist das 30%-Perzentil oder 0.3-Quantil der Nettoquadratmetermiete?)



$\Rightarrow \approx 5€$

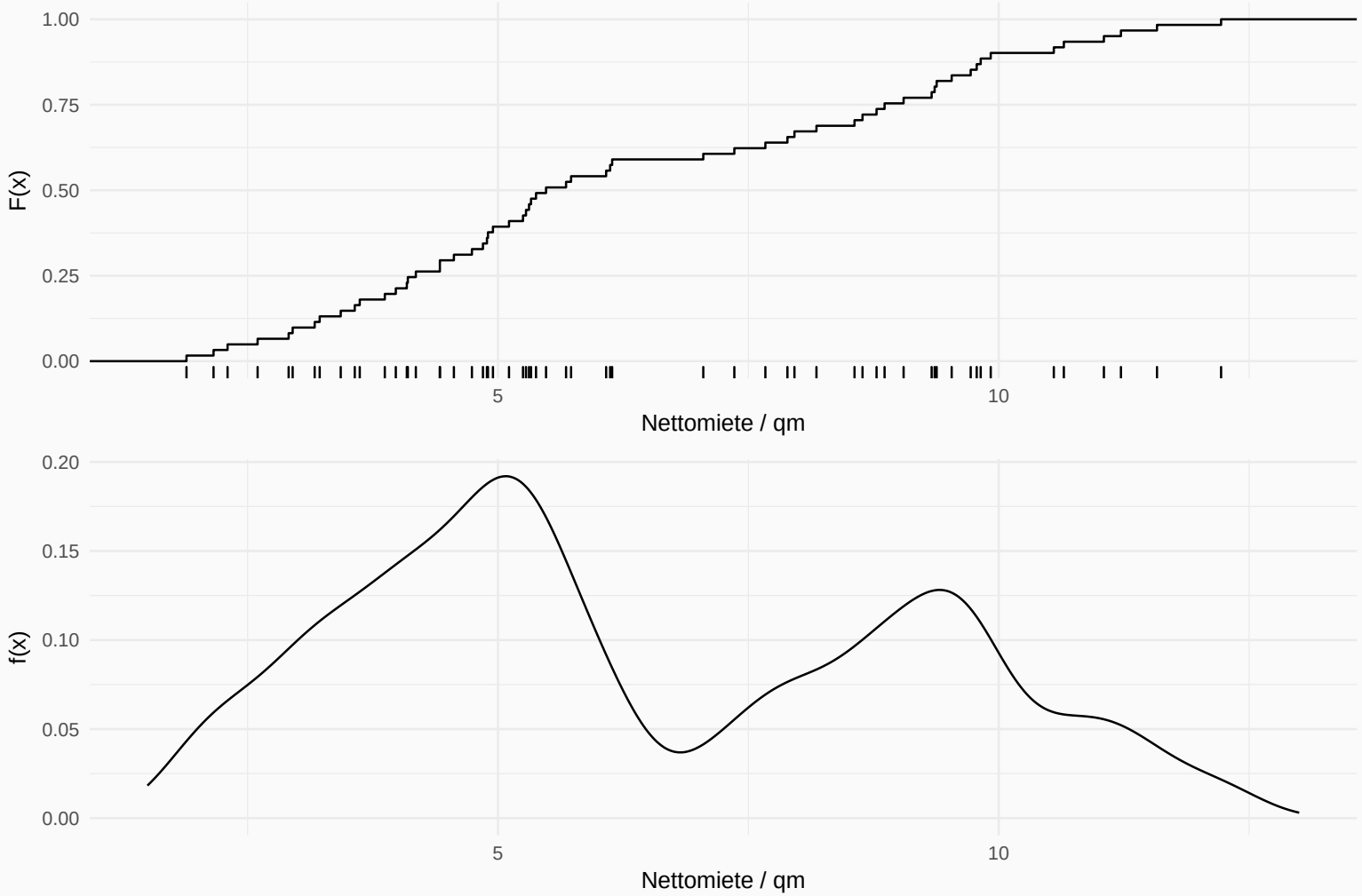
Dichtefunktion

Dichtefunktion $f(x)$ so dass Fläche unter der Dichte über ein Intervall (in etwa) der relativen Häufigkeit von Werten in diesem Intervall entspricht, d.h.

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

- Da $F(b) := \#\{x_i | x_i \leq b\}$ gilt also $F(b) \approx \int_{-\infty}^{b} f(x)dx$
- Also: Dichte ist so etwas wie die Ableitung der Verteilungsfunktion
  - Dort wo starker Anstieg der Verteilungsfunktion (große Ableitung) liegen viele Beobachtungen
    $\Rightarrow$ hohe "Dichte" an Beobachtungen
  - keine exakte Übereinstimmung da "verschmiert" durch Glättung mit Kernfunktionen

Dichtefunktion



Statistische Kennzahlen:

Lage: Wo liegen die (typischen/häufigsten/...) Werte des Merkmals?
Streuung: Wie stark / über welchen Bereich streuen die Werte des Merkmals? Wie groß ist die Schwankung?

Form der Verteilung: **Modalität**, **Schiefe**, **Konzentration**

Lagemaße:

- Welche sinnvoll sind hängt vom Skalenniveau ab
- Modus: häufigster Wert
- Median: x_{med}: $F(x_{\text{med}}) = 0.5$ “Mitte” der Daten
- arithmetisches Mittel (evtl. auf transformierter Skala): Durchschnitt der Daten

Lagemaß: Modus

Definition: **Häufigster Wert**

Eigenschaften:

- oft nicht eindeutig
- nur bei gruppierten Daten oder bei Merkmalen mit wenigen Ausprägungen sinnvoll
- stabil bei allen *eindeutigen* Transformationen
- geeignet für alle Skalenniveaus

Lagemaß: Median

Definition: Der **Median** ($\tilde{x}_{\text{med}}$) ist der Wert für den gilt

- mindestens 50% der Daten sind kleiner oder gleich $\tilde{x}_{\text{med}}$,
- mindestens 50% der Daten sind größer oder gleich $\tilde{x}_{\text{med}}$.

Eigenschaften des Medians

- anschaulich
- stabil gegenüber monotonen Transformationen
- geeignet für mindestens ordinale Daten
- stabil gegenüber Ausreißern (da nur Rangfolge der Daten berücksichtigt)

Lagemaß: Quantil

Verallgemeinerung des Medians: p-**Quantil** ist der Wert $\tilde{x}_p$ für den gilt

- mindestens Anteil p der Daten sind kleiner oder gleich $\tilde{x}_p$,
- mindestens Anteil $1 - p$ der Daten sind größer oder gleich $\tilde{x}_p$.

Also:

$$F(\tilde{x}_p) = p$$

Lagemaß: Mittelwert (arithmetisches Mittel)

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

- bekanntestes Lagemaß
- instabil gegen extreme Werte
- geeignet für mindestens intervallskalierte Daten
- “Durchschnitt”

Lagemaß: Mittelwert-Varianten:

- **getrimmtes** oder **winsorisiertes** arithmetisches Mittel: robust gegen Ausreißer
- **geometrisches** Mittel $\bar{x}_G$:
  - exponierter Mittelwert der logarithmierten Daten (also $x_i > 0$!)
  - für durchschnittliche Änderungsraten u.ä.
  -

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

Bsp: Geometrisches Mittel:

```
# zB UMSATZENTWICKLUNG:
(umsatz <- c("2010" = 20,  "2011" = 22.2, "2012" = 17.5, "2013" = 24.5, "2014" = 30))
## 2010 2011 2012 2013 2014
## 20.0 22.2 17.5 24.5 30.0
(wachstumsraten <- umsatz[2:5]/umsatz[1:4]) # umsatz2011/umsatz2010, u.2012/u.2011, etc...
##  2011  2012  2013  2014
## 1.110 0.788 1.400 1.224

# Gesamtes Wachstum: 50% (2010:20 -> 2014:30)
# GEOMETRISCHES MITTEL der Wachstumsraten = mittlere Veränderung pro Jahr:
(mittlere_rate <- exp(mean(log(wachstumsraten))))
## [1] 1.11
# --> mittleres Umsatzwachstum in diesem Zeitraum ist 11% pro Jahr

# Also:
unname(umsatz["2010"] * mittlere_rate ^ 4)
## [1] 30
# == umsatz["2014"]

# Aber eben nicht wenn man naiv arithmetisches Mittel der Wachstumsraten nimmt:
(mittlere_rate_naiv <- mean(wachstumsraten))
## [1] 1.13
unname(umsatz["2010"] * mittlere_rate_naiv ^ 4)
## [1] 32.7
# --> nicht das selbe wie umsatz["2014"]!
```

Arithmetisches Mittel transformierter Daten:

Lineare Transformation:

$$y = g(x) = a + bx \;\Rightarrow\; \bar{y} = a + b\bar{x}$$

- Sonst im allgemeinen $\overline{g(x)} \neq g(\bar{x})$
- Für **konvexes** g: $\overline{g(x)} \geq g(\bar{x})$ (zB: $(\bar{x})^2 < \overline{(x^2)}$)

Streuungsmaße

- Spannweite: Differenz zwischen Maximum und Minimum
- Interquartilsabstand: Differenz zwischen 75% und 25% Quantilen
- Standardabweichung und Varianz: Mittelwert der (quadrierten) Abweichung vom Mittelwert
- Variationskoeffizient: Schwankung *relativ* zum Mittelwert

Die Spannweite (Range)

Definition:

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

- Bereich in dem die Daten liegen
- Wichtig für Datenkontrolle: Plausibilität, Eingabe-/Codierungsfehler, etc.

Interquartilsabstand

Definition:

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

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

Standardabweichung und Varianz

Definition:

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

$$\textbf{Standardabweichung } S_x := \sqrt{S_x^2}$$

- S_x = "Mittlere Abweichung vom Mittelwert"
- Mindestens Intervallskala
- empfindlich gegen Ausreißer

Streuungszerlegung I

Seien die Daten in r Schichten aufgeteilt:

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

Schichtmittelwerte:

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

Schichtvarianzen:

$$\widetilde{S}_{x1}^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2,$$

$$\widetilde{S}_{x2}^2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2)^2 \; \text{usw.}$$

Streuungszerlegung II

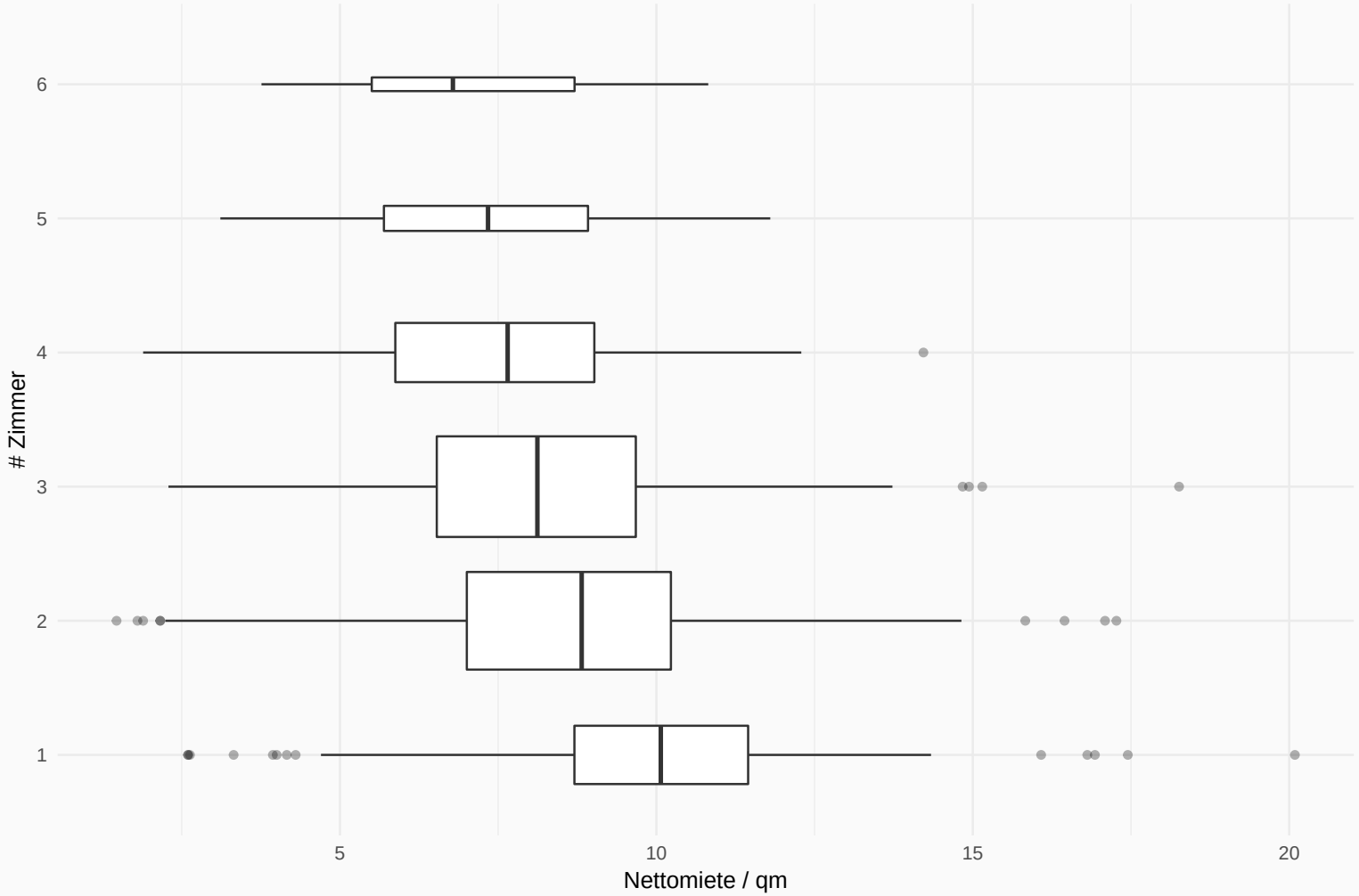
Dann gilt:

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

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

Streuungszerlegung: Beispiel 1

Quadratmetermiete - Zimmerzahl:



Streuungszerlegung: Beispiel 1

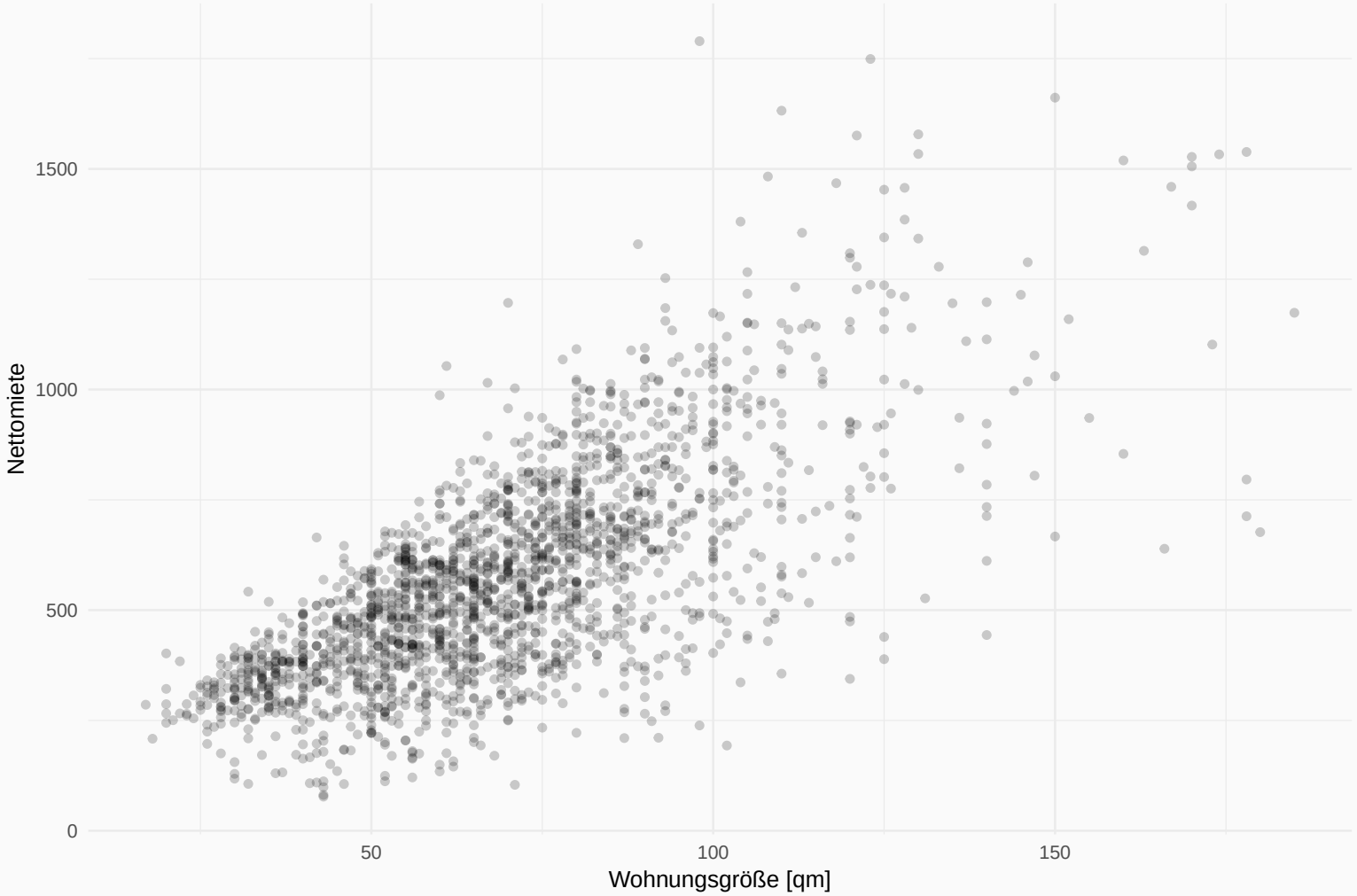
```
# SCHICHTGRÖSSEN (n_j):
(group_sizes <- with(mietspiegel, tapply(nmqm, rooms, length)))
##   1   2   3   4   5   6
## 255 715 759 263  47  14
n <- sum(group_sizes)

# SCHICHTMITTELWERTE (xquer_j):
(grouped_means <- with(mietspiegel, tapply(nmqm, rooms, mean)))
##    1    2    3    4    5    6
## 10.0  8.6  8.0  7.5  7.3  7.1

# SCHICHTVARIANZEN (S^2_xj):
(grouped_variance <- with(mietspiegel, tapply(nmqm, rooms, var)))
##   1   2   3   4   5   6
## 5.7 6.0 5.5 4.6 5.9 5.6
```

Streuungszerlegung: Beispiel 1

```r
# GESAMTMITTELWERT & -VARIANZ:
(total <- with(mietspiegel, c(mean = mean(nmqm), var = var(nmqm))))
## mean  var
##  8.4  6.1

# STREUUNG INNERHALB DER SCHICHTEN = MITTELWERT DER SCHICHTVARIANZEN:
(within <- sum(grouped_variance * group_sizes)/n)
## [1] 5.6

# STREUUNG ZWISCHEN DEN SCHICHTEN = (gewichtete) VARIANZ DER
# SCHICHTMITTELWERTE:
(between <- sum((grouped_means - total["mean"])^2 * group_sizes)/n)
## [1] 0.52

# also total['var'] = between + within
between/total["var"]
##   var
## 0.085
# --> nur etwa 8.5% der Gesamtstreuung ZWISCHEN den Schichten, Rest
# INNERHALB...
```

Streuungszerlegung: Beispiel 2

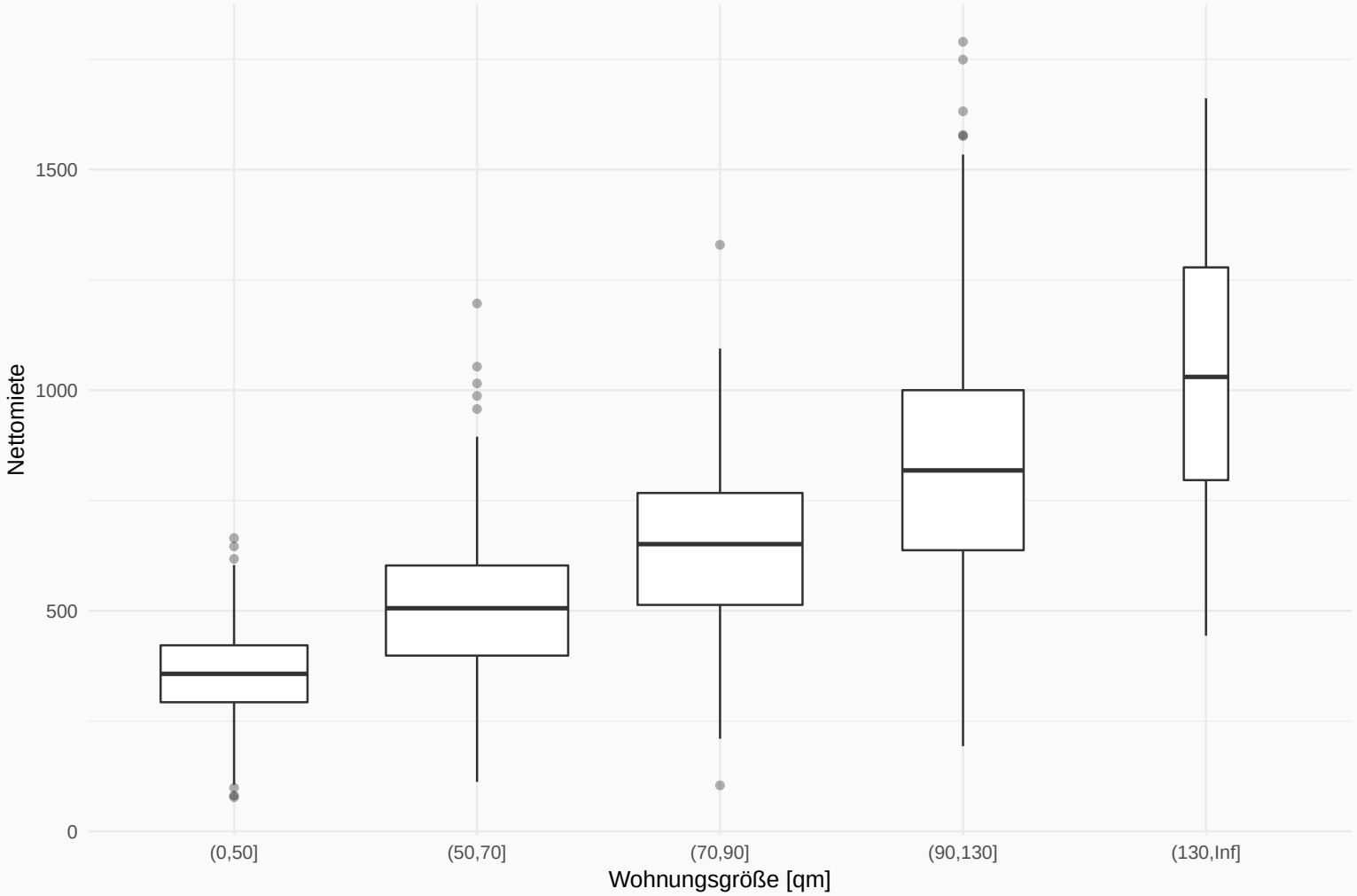
Nettomiete - Wohnungsgröße:



Streuungszerlegung: Beispiel 2

Nettomiete - Wohnungsgröße (gruppiert):

```
mietspiegel$groesse <- cut(mietspiegel$wfl, c(0, 50, 70, 90, 130, Inf))
```



Streuungszerlegung: Beispiel 2

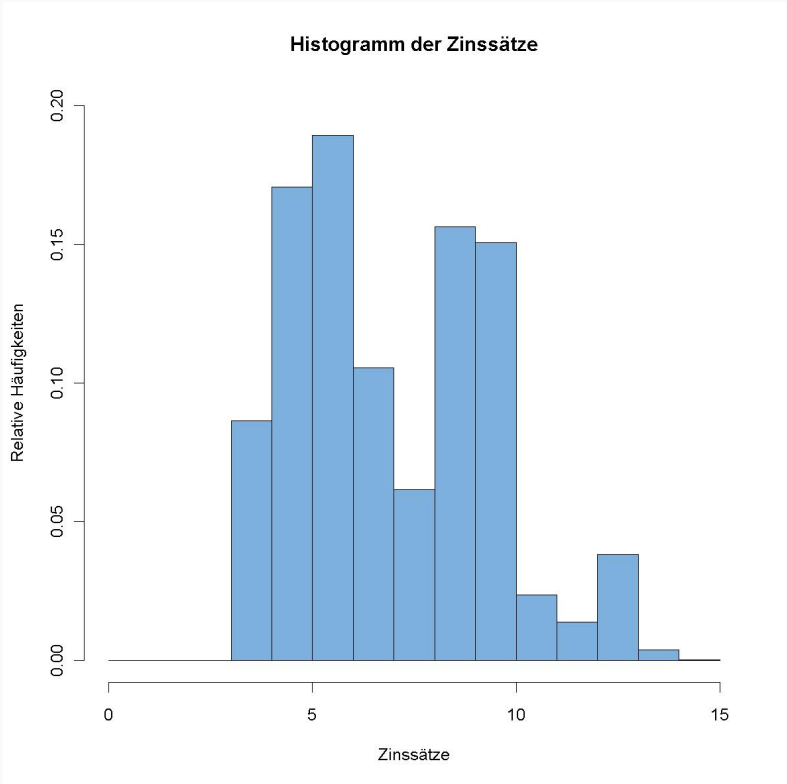
```
# SCHICHTGRÖSSEN (n_j):
(group_sizes <- with(mietspiegel, tapply(nm, groesse, length)))
##    (0,50]   (50,70]   (70,90]  (90,130] (130,Inf]
##       449       691       566       306        41
n <- sum(group_sizes)

# SCHICHTMITTELWERTE (xquer_j):
(grouped_means <- with(mietspiegel, tapply(nm, groesse, mean)))
##    (0,50]   (50,70]   (70,90]  (90,130] (130,Inf]
##       361       501       645       831      1046

# SCHICHTVARIANZEN (S^2_xj):
(grouped_variance <- with(mietspiegel, tapply(nm, groesse, var)))
##    (0,50]   (50,70]   (70,90]  (90,130] (130,Inf]
##     10768     23045     32823     78747    101714
```

Streuungszerlegung: Beispiel 2

```
# GESAMTMITTELWERT & -VARIANZ:
(total <-  with(mietspiegel, c(mean = mean(nm), var = var(nm))))
##  mean   var
##   570 60238

# STREUUNG INNERHALB DER SCHICHTEN  = MITTELWERT DER SCHICHTVARIANZEN:
(within <- sum(grouped_variance * group_sizes) / n)
## [1] 32929

# STREUUNG ZWISCHEN DEN SCHICHTEN  = (gewichtete) VARIANZ DER SCHICHTMITTELWERTE:
(between <- sum((grouped_means - total["mean"])^2 * group_sizes) / n)
## [1] 27400

# also total["var"] = between + within
# (hier nicht ganz exakt wegen Rundungsfehler...)
between / total["var"]
##  var
## 0.45
# --> hier etwa 45% der Gesamtstreuung ZWISCHEN den Schichten, 55% INNERHALB.
```

Uni- und multimodale Verteilungen

unimodal = eingipflig, **multimodal** = mehrgipflig

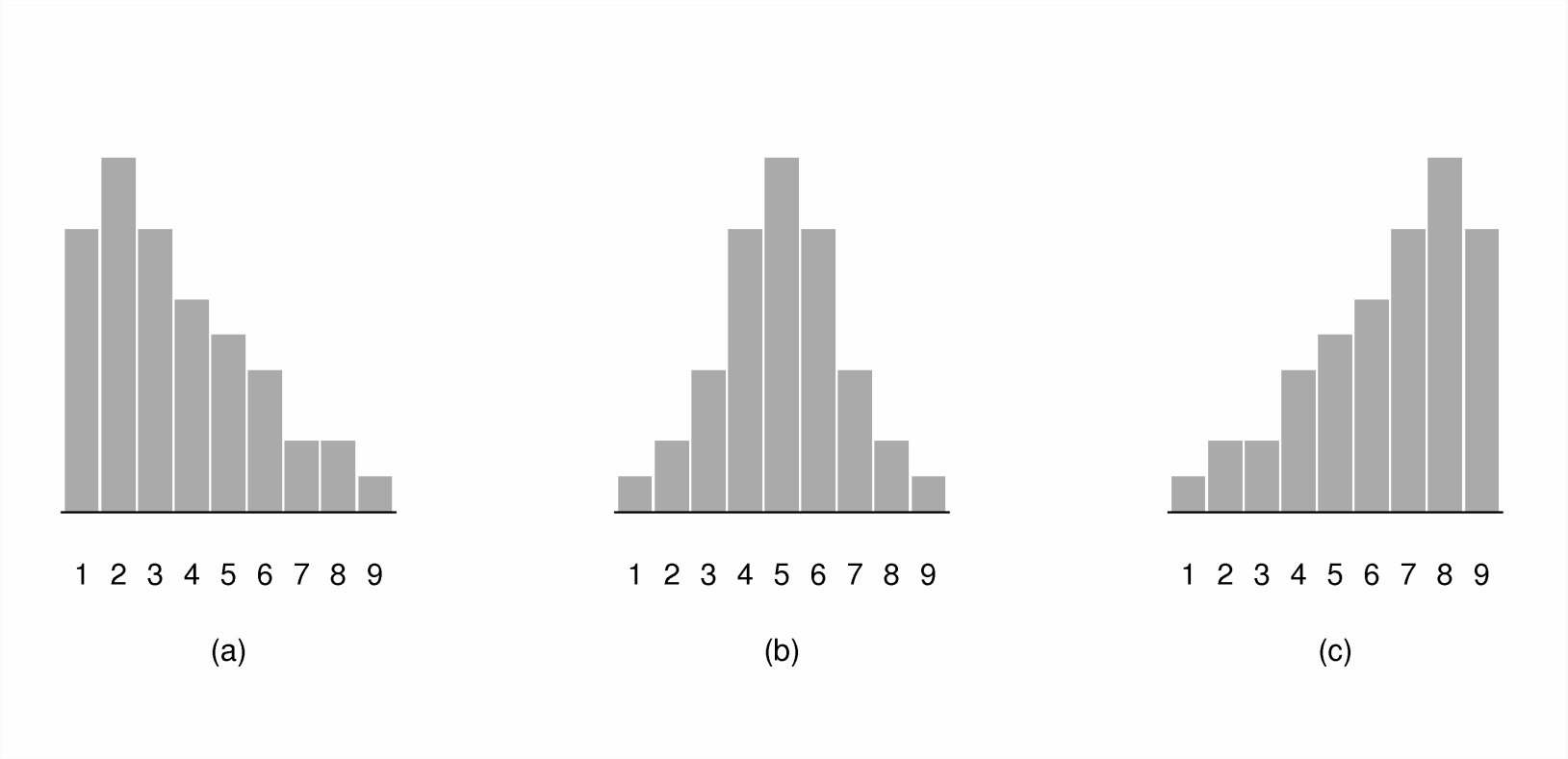


Das Histogramm der Zinssätze zeigt eine bimodale Verteilung.

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

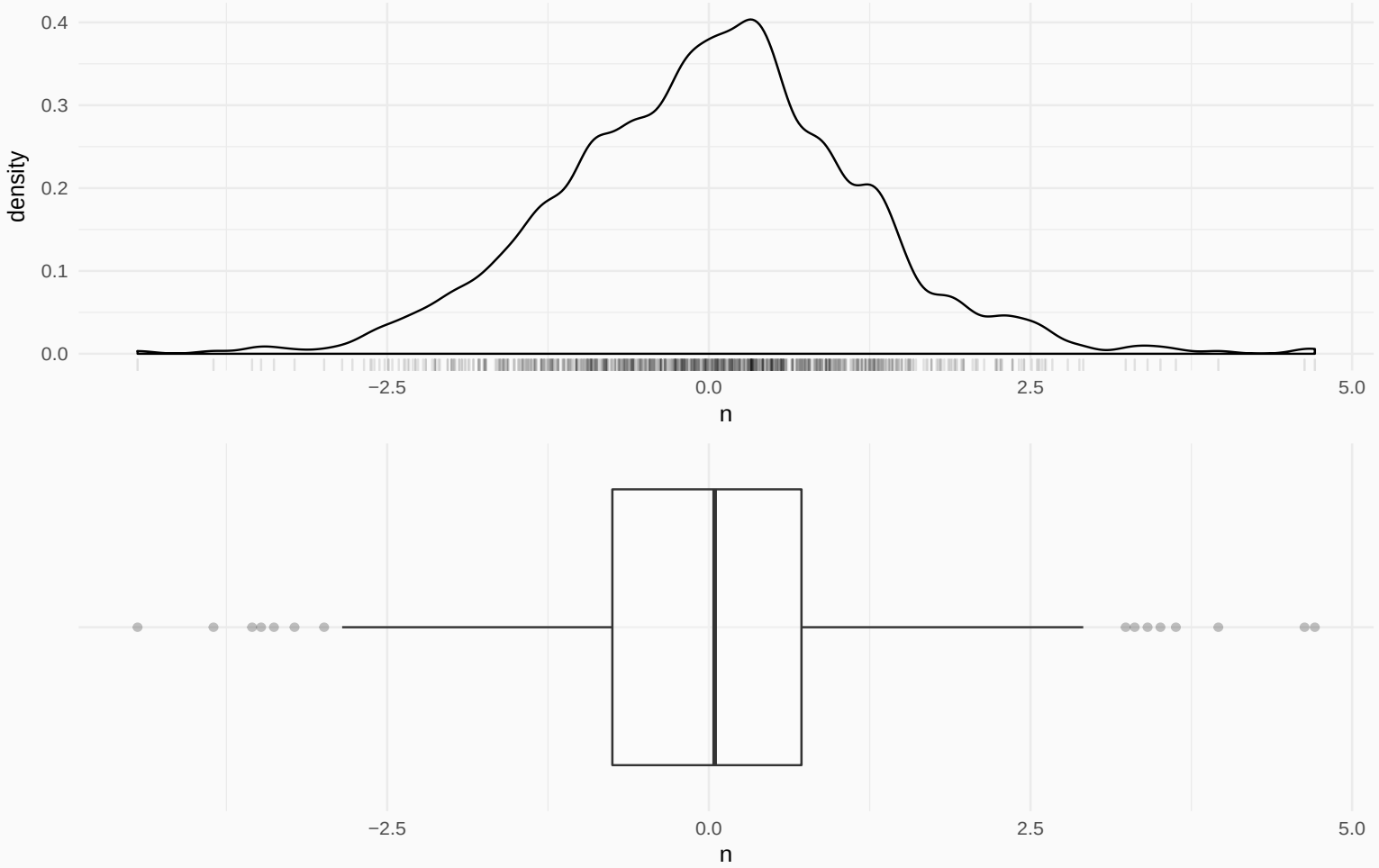


Eine linkssteile (a), symmetrische (b) und rechtssteile Verteilung (c)

Lageregeln

- Symmetrische und unimodale Verteilung:
  $\bar{x} \approx x_{med} \approx x_{mod}$
- Linkssteile Verteilung: $\bar{x} > x_{med} > x_{mod}$
- Rechtssteile Verteilung: $\bar{x} < x_{med} < x_{mod}$

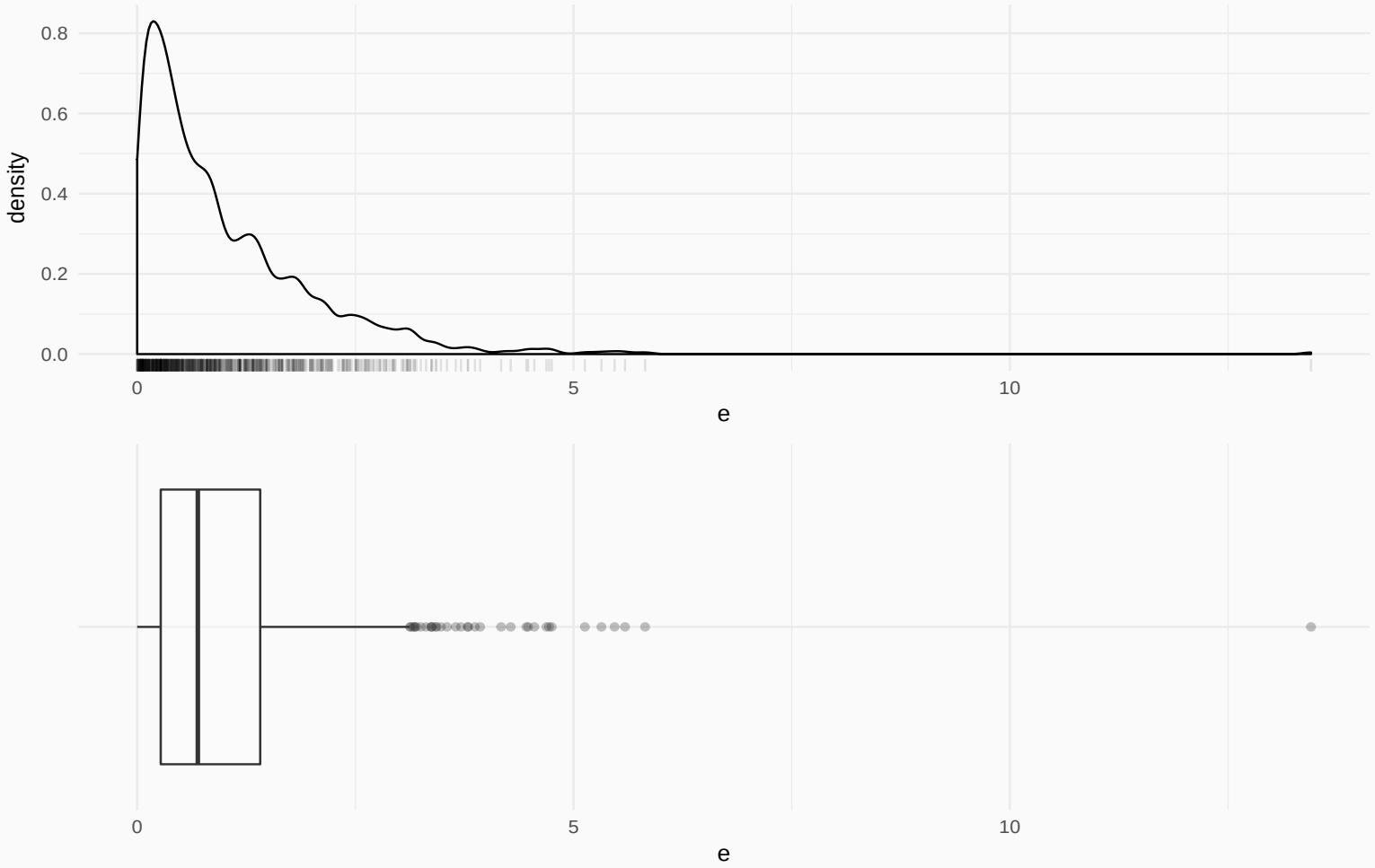
Lageregeln: symmetrisch



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

```
##       mean median.50%       mode
##     0.0029     0.0450     0.3300
```

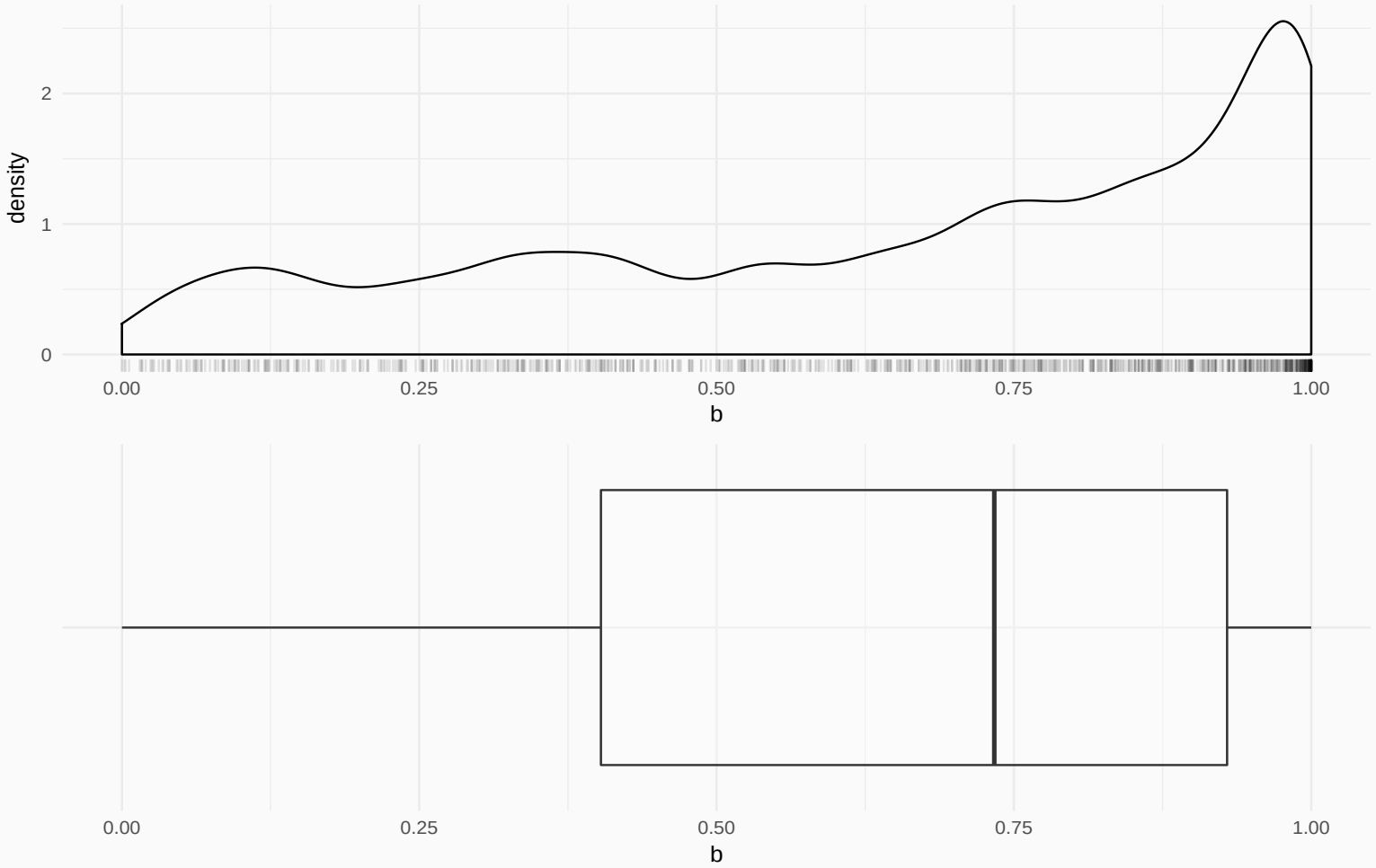
Lageregeln: linkssteil/rechtsschief



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

```
##       mean median.50%       mode
##       0.99       0.69       0.27
```

Lageregeln: linksschief/rechtssteil



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

```
##       mean median.50%       mode
##       0.89       0.92       1.00
```