Programmierung und Deskriptive Statistik

BSc Psychologie WiSe 2022/23

Belinda Fleischmann

(8) Verteilungsfunktionen und Quantile

Empirische Verteilungsfunktionen

Quantile und Boxplots

Übungen und Selbstkontrollfragen

Empirische Verteilungsfunktionen

Quantile und Boxplots

Übungen und Selbstkontrollfragen

Definition (Kumulative absolute und relative Häufigkeitsverteilungen)

$x = (x_1, ..., x_n)$ sei ein Datensatz, $A := \{a_1, ..., a_k\}$ mit $k \leq n$ die im Datensatz vorkommenden verschiedenen Zahlenwerte und h und r die absoluten und relativen Häufigkeitsverteilungen von x, respektive. Dann heißt die Funktion

$$H : A \to \mathbb{N}, a \mapsto H(a) := \sum_{a' \leq a} h(a') \tag{1}$$

die *kumulative absolute Häufigkeitsverteilung* von x und die Funktion

$$R : A \to [0, 1], a \mapsto R(a) := \sum_{a' \leq a} r(a') \tag{2}$$

die *kumulative relative Häufigkeitsverteilung* der Zahlwerte von x.

Bemerkung

- Mit den Definitionen der absoluten und relativen Häufigkeitsverteilungen gilt also

$$H(a) = \text{Anzahl der } x_i \text{ aus } x \text{ mit } x_i \leq a \tag{3}$$

und

$$R(a) = \text{Anzahl der } x_i \text{ aus } x \text{ mit } x_i \leq a \text{ geteilt durch } n. \tag{4}$$

Empirische Verteilungsfunktionen

`cumsum()` erlaubt die Berechnung kumulativer Summen

```
# Einlesen des Beispieldatensatzes und Abbildungsverzeichnisdefinition
fpath    = file.path(pfad_zu_Datenordner, "psychotherapie_datensatz.csv")
D        = read.table(fpath, sep = ",", header = T)
fdir     = file.path(getwd(), "8_Abbildungen")

# Evaluation der absoluten und relativen Häufigkeitsverteilugen von Pre.BDI
x        = D$Pre.BDI                 # Double vector der Pre.BDI Werte
n        = length(x)                 # Anzahl der Datenwerte
H        = as.data.frame(table(x))   # absolute Häufigkeitsverteilung als Dataframe
names(H) = c("a", "h")               # Spaltenbenennung
H$H      = cumsum(H$h)               # kumulative absolute Häufigkeitsverteilung
H$r      = H$h/n                     # relative Häufigkeitsverteilung
H$R      = cumsum(H$r)               # kumulative relative Häufigkeitsverteilung
print(H)
```

```
>     a  h   H    r    R
> 1  14  1   1 0.01 0.01
> 2  15  3   4 0.03 0.04
> 3  16  6  10 0.06 0.10
> 4  17 17  27 0.17 0.27
> 5  18 21  48 0.21 0.48
> 6  19 20  68 0.20 0.68
> 7  20 17  85 0.17 0.85
> 8  21 12  97 0.12 0.97
> 9  22  2  99 0.02 0.99
> 10 23  1 100 0.01 1.00
```

Empirische Verteilungsfunktionen

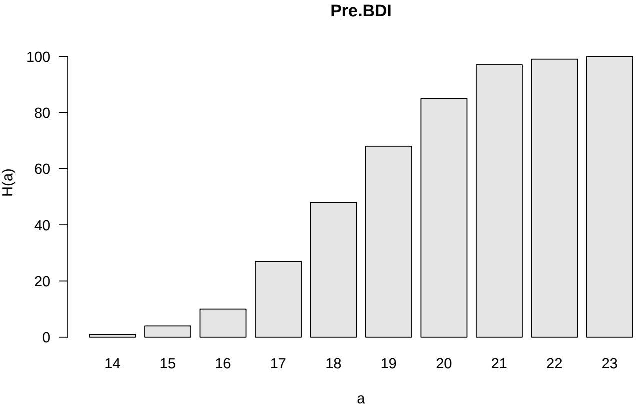
Kumulative absolute Häufigkeitsverteilung der Pre.BDI Werte

```
# Visualisierung der kumulativen absoluten Häufigkeitsverteilung
graphics.off()                    # Abbildungsinitialisierung
dev.new()                         # Abbildungsinitialisierung
Ha           = H$H                # H(a) Werte
names(Ha)    = H$a                # barplot braucht a Werte als names
barplot(                          # Balkendiagramm
Ha,                               # H(a) Werte
col          = "gray90",          # Balkenfarbe
xlab         = "a",               # x Achsenbeschriftung
ylab         = "H(a)",            # y Achsenbeschriftung
ylim         = c(0,110),          # y Achsenlimits
las          = 1,                 # Achsenticklabelorientierung
main         = "Pre.BDI")         # Titel

# PDF Speicherung
dev.copy2pdf(
file         = file.path(fdir, "pds_8_kh.pdf"),
width        = 8,
height       = 5)
```

Empirische Verteilungsfunktionen

Kumulative absolute Häufigkeitsverteilung der Pre.BDI Werte



Empirische Verteilungsfunktionen

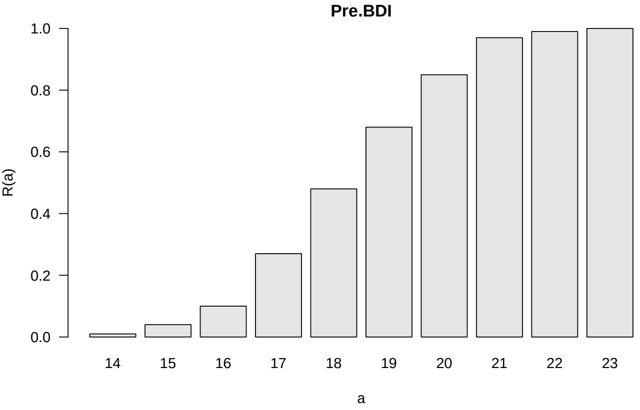
Kumulative relative Häufigkeitsverteilung der Pre.BDI Werte

```
# Visualisierung der kumulativen relativen Häufigkeitsverteilung
graphics.off()                 # Abbildungsinitialisierung
dev.new()                      # Abbildungsinitialisierung
R            = H$R             # R(a) Werte
names(R)     = H$a             # barplot braucht a Werte als names
dev.new()                      # Abbildungsinitialisierung
barplot(                       # Balkendiagramm
R,                             # R(a) Werte
col          = "gray90",       # Balkenfarbe
xlab         = "a",            # x Achsenbeschriftung
ylab         = "R(a)",         # y Achsenbeschriftung
ylim         = c(0,1),         # y Achsenlimits
las          = 1,              # Achsenticklabelorientierung
main         = "Pre.BDI")      # Titel

# PDF Speicherung
dev.copy2pdf(
file         = file.path(fdir, "pds_8_kr.pdf"),
width        = 8,
height       = 5)
```

Empirische Verteilungsfunktionen

Kumulative relative Häufigkeitsverteilung der Pre.BDI Werte



Definition (Empirische Verteilungsfunktion)

$x = (x_1, ..., x_n)$ sei ein Datensatz. Dann heißt die Funktion

$$F : \mathbb{R} \to [0, 1], \xi \mapsto F(\xi) := \frac{\text{Anzahl der } x_i \text{ aus } x \text{ mit } x_i \leq \xi}{n} \tag{5}$$

die empirische Verteilungsfunktion (EVF) von x.

Bemerkungen

- Die empirische Verteilungsfunktion wird auch *empirische kumulative Verteilungsfunktion* genannt.
- Die Definitionsmenge der EVF ist im Gegensatz zu Häufigkeitsverteilungen $\mathbb{R}$ und nicht A
- Die EVF verhält sich zu kumulativen Häufigkeitsverteilungen wie Histogramme zu Häufigkeitsverteilungen.
- Typischerweise sind empirische Verteilungsfunktionen Treppenfunktionen.
- Die (visuelle) Umkehrfunktion der EVF kann zur Bestimmung von Quantilen genutzt werden.

Empirische Verteilungsfunktion der Pre.BDI Werte

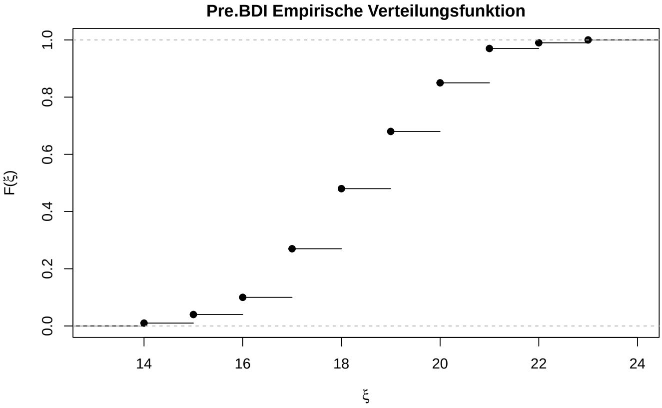
`ecdf()` evaluiert die empirischen Verteilungsfunktion eines Datensatzes

```
graphics.off()                                          # Abbildungsinitialisierung
dev.new()                                               # Abbildungsinitialisierung
x    = D$Pre.BDI                                        # double vector der Pre.BDI Werte
evf  = ecdf(x)                                          # Evaluation der EVF
plot(                                                   # plot() weiß mit ecdf object umzugehen
evf,                                                    # ecdf Objekt
xlab = TeX("$\\xi$"),                                   # x Achsenbeschriftung
ylab = TeX("$F(\\xi)$"),                                # y Achsenbeschriftung
main = "Pre.BDI Empirische Verteilungsfunktion")        # Titel

# PDF Speicherung
dev.copy2pdf(
file        = file.path(fdir, "pds_8_ecdf.pdf"),
width       = 8,
height      = 5)
```

Empirische Verteilungsfunktionen

Empirische Verteilungsfunktion der Pre.BDI Werte



Empirische Verteilungsfunktionen

Quantile und Boxplots

Übungen und Selbstkontrollfragen

Definition (p-Quantil)

$x = (x_1, ..., x_n)$ sei ein Datensatz und

$$x_s = \left(x_{(1)}, x_{(2)}, ..., x_{(n)}\right) \text{ mit } \min_{1\leq i\leq n} x_i = x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} = \max_{1\leq i\leq n} x_i \tag{6}$$

der zugehörige aufsteigend sortierte Datensatz. Weiterhin bezeichne $\lfloor\cdot\rfloor$ die Abrundungsfunktion.Dann heißt für ein $p \in [0, 1]$ die Zahl

$$x_p := \begin{cases} x_{(\lfloor np+1\rfloor)} & \text{falls } np \neq \mathbb{N} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right) & \text{falls } np \in \mathbb{N} \end{cases} \tag{7}$$

das p-*Quantil* von x.

Bemerkungen

- Mindestens $p \cdot 100\%$ aller Werte in x sind kleiner oder gleich x_p.
- Mindestens $(1 - p) \cdot 100\%$ aller Werte in x sind größer als x_p .
- Das p-Quantil teilt den geordneten Datensatz im Verhältnis p zu $(1 - p)$ auf.
- $x_{0.25}, x_{0.50}, x_{0.75}$ heißen *unteres Quartil*, *Median*, und *oberes Quartil*, respektive.
- $x_{j\cdot 0.10}$ für $j = 1, ..., 9$ heißen *Dezile*,
- $x_{j\cdot 0.01}$ für $j = 1, ..., 99$ heißen *Percentile*.

Quantile und Boxplots

Beispiel p-Quantil (Henze (2018), Kapitel 5)

Datensatz und sortierter Datensatz

i	1	2	3	4	5	6	7	8	9	10
x_i	8.5	1.5	75	4.5	6.0	3.0	3.0	2.5	6.0	9.0
$x_{(i)}$	1.5	2.5	3.0	3.0	4.5	6.0	6.0	8.5	9.0	75

0.25-Quantil

Es ist $n = 10$ und es sei $p := 0.25$. Dann gilt $np = 10 \cdot 0.25 = 2.5 \notin \mathbb{N}$. Also folgt

$$x_{0.25} = x_{(\lfloor 2.5+1 \rfloor)} = x_{(3)} = 3.0 \tag{8}$$

0.80-Quantil

Es ist $n = 10$ und es sei $p := 0.80$. Dann gilt $np = 10 \cdot 0.80 = 8 \in \mathbb{N}$. Also folgt

$$x_{0.80} = \frac{1}{2}\left(x_{(8)} + x_{(8+1)}\right) = \frac{1}{2}\left(x_{(8)} + x_{(9)}\right) = \frac{8.5 + 9.0}{2} = 8.75. \tag{9}$$

Quantile und Boxplots

Beispiel p-Quantil (Henze (2018), Kapitel 5) (fortgeführt)

"Manuelle" Quantilbestimmung anhand obiger Definition

```
x   = c(8.5,1.5,12,4.5,6.0,3.0,3.0,2.5,6.0,9.0)   # Beispieldaten
n   = length(x)                                    # Anzahl Datenwerte
x_s = sort(x)                                      # sortierter Datensatz
p   = 0.25                                         # np \notin \mathbb{N}
x_p = x_s[floor(n*p + 1)]                          # 0.25 Quantil
print(x_p)                                         # Ausgabe
```

```
> [1] 3
```

```
p   = 0.80                                         # np \in \mathbb{N}
x_p = (1/2)*(x_s[n*p] + x_s[n*p + 1])              # 0.80 Quantil
print(x_p)                                         # Ausgabe
```

```
> [1] 8.75
```

`quantile()` wertet Quantile anhand der Quantildefinition `type` aus.

Es gibt mindestens neun verschiedene Quantildefinitionen (cf. Hyndman and Fan (1996))

```
x_p = quantile(x, 0.80, type = 1)                  # 0.80 Quantil, Definition 1
print(x_p)
```

```
> 80%
> 8.5
```

```
x_p = quantile(x, 0.80, type = 2)                  # 0.80 Quantil, Definition 2
print(x_p)
```

```
>  80%
> 8.75
```

Beispiel p-Quantil (Henze (2018), Kapitel 5) (fortgeführt)

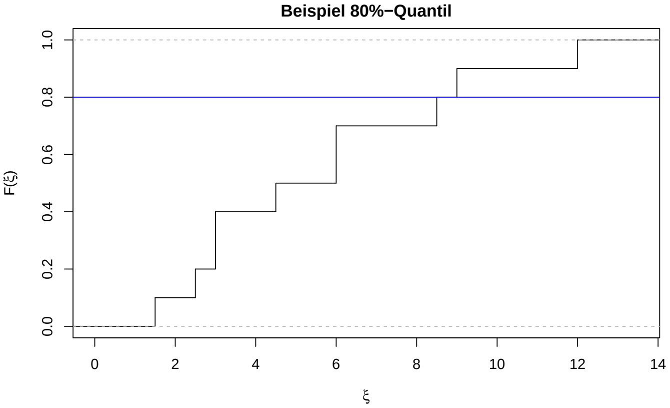
Kombination von `ecdf()`und `abline()` erlaubt prinzipiell die visuelle Bestimmung von Quantilen

```
graphics.off()                                   # Abbildungsinitialisierung
dev.new()                                        # Abbildungsinitialisierung
evf          = ecdf(x)                           # Evaluation der EVF
plot(                                            # plot() weiß mit ecdf object umzugehen
evf,                                             # ecdf Objekt
xlab         = TeX("$\\xi$"),                    # x Achsenbeschriftung
ylab         = TeX("$F(\\xi)$"),                 # y Achsenbeschriftung
verticals    = TRUE,                             # vertikale Linien
do.points    = FALSE,                            # keine Punkte
main         = "Beispiel 80%-Quantil")           # Titel
abline(                                          # horizontale Linie
h            = 0.80,                             # y Ordinate der Linie
col          = "blue")                           # blau

# PDF Speicherung
dev.copy2pdf(
file         = file.path(fdir, "pds_8_ecdf_abline_x.pdf"),
width        = 8,
height       = 5)
```

Beispiel p-Quantil (Henze (2018), Kapitel 5) (fortgeführt)

Kombination von `ecdf()`und `abline()` erlaubt prinzipiell die visuelle Bestimmung von Quantilen

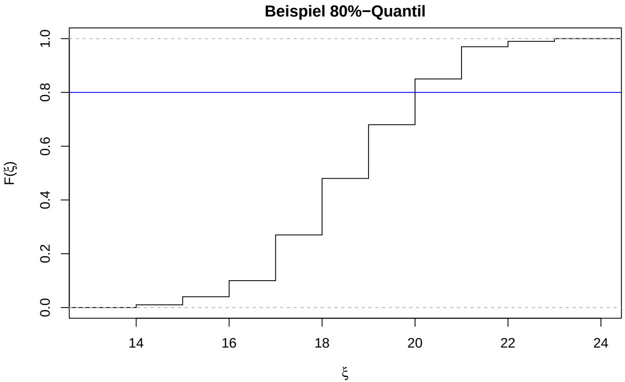


80% Quantil der Pre-BDI Daten

```
graphics.off()                              # Abbildungsinitialisierung
dev.new()                                   # Abbildungsinitialisierung
x            = D$Pre.BDI                    # Double vector der Pre.BDI Werte
evf          = ecdf(x)                      # Evaluation der EVF
plot(                                       # plot() weiß mit ecdf object umzugehen
evf,                                        # ecdf Objekt
xlab         = TeX("$\\xi$"),               # x Achsenbeschriftung
ylab         = TeX("$F(\\xi)$"),            # y Achsenbeschriftung
verticals    = TRUE,                        # vertikale Linien
do.points    = FALSE,                       # keine Punkte
main         = "Beispiel 80%-Quantil")      # Titel
abline(                                     # horizontale Linie
h            = 0.80,                        # y Ordinate der Linie
col          = "blue")                      # blau

# PDF Speicherung
dev.copy2pdf(
file         = file.path(fdir, "pds_8_ecdf_abline_prebdi.pdf"),
width        = 8,
height       = 5)
```

80% Quantil der Pre-BDI Daten




```
x_p = quantile(D$Pre.BDI, 0.80, type = 2) # 0.80 Quantil, Definition 2
print(x_p)
```

```
> 80%
>  20
```

Quantile und Boxplots

Boxplots

Ein Boxplot visualisiert eine Quantil-basierte Zusammenfassung eines Datensatzes.

Typischerweise werden $\min x, x_{0.25}, x_{0.50}, x_{0.75}, \max x$ visualisiert.

- $\min x$ und $\max x$ werden oft als "Whiskerendpunkte" dargestellt.
- $x_{0.25}$ und $x_{0.75}$ sind untere und obere Grenze der zentralen grauen Box.
- $x_{0.50}$ wird als Strich in der zentralen grauen Box abgebildet.

$d_Q := x_{0.75} - x_{0.25}$ heißt *Interquartilsabstand* und dient als Verteilungsbreitenmaß

`summary()` liefert wesentliche Kennzahlen

```
# Sechswertezusammenfassung
summary(D$Pre.BDI)
```

```
>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
>    14.0    17.0    19.0    18.6    20.0    23.0 
```

Boxplots

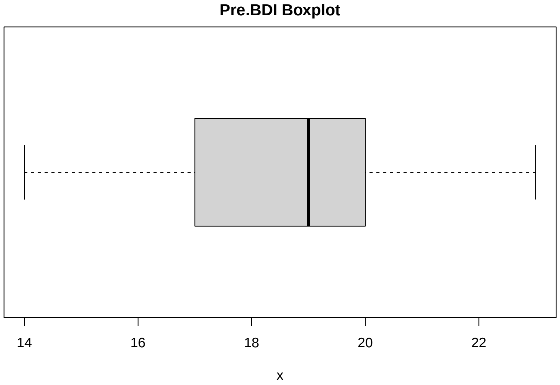
`boxplot()` erstellt einen Boxplot,

```
# Boxplot
graphics.off()                          # Abbildungsinitialisierung
dev.new()                               # Abbildungsinitialisierung
boxplot(                                # Boxplot
D$Pre.BDI,                              # Datensatz
horizontal  = T,                        # horizontale Darstellung
range       = 0,                        # Whiskers bis zu min x und max x
xlab        = "x",                      # x Achsenbeschriftung
main        = "Pre.BDI Boxplot")        # Titel

# PDF Speicherung
dev.copy2pdf(
file        = file.path(fdir, "pds_8_boxplot_prebdi.pdf"),
width       = 8,
height      = 5)
```

Quantile und Boxplots

Boxplots



Es gibt viele Boxplotvariationen (cf. McGill, Tukey, and Larsen (1978)), eine genaue Erläuterung ist immer nötig!

Empirische Verteilungsfunktionen

Quantile und Boxplots

Übungen und Selbstkontrollfragen

Übungen und Selbstkontrollfragen

1. Definieren Sie die Begriffe der kumulativen absoluten und relativen Häufigkeitsverteilungen.
2. Erzeugen und visualisieren Sie die kumulativen absoluten und relativen Häufigkeitsverteilungen der Post.BDI Daten des Beispieldatensatzes *psychotherapie_datensatz.csv*.
3. Definieren Sie den Begriff der empirischen Verteilungsfunktion.
4. Erzeugen und visualisieren Sie die empirische Verteilungsfunktion der Post.BDI Daten.
5. Erläutern Sie den Begriff des sortierten Datensatzes.
6. Definieren Sie den Begriff des p-Quantils.
7. Berechnen Sie das obere Quartil des Beispieldatensatzes auf Folie 16.
8. Definieren Sie die Begriffe unteres Quartil, Median und oberes Quartil mithilfe des p-Quantils.
9. Berechnen Sie das untere Quartil, den Median und das obere Quartil der Post.BDI Daten. Vergleiche Sie ihre Ergebnisse mit der Ausgabe der summary() Funktion.
10. Erstellen Sie einen Boxplot der Post.BDI Daten.

References

Henze, Norbert. 2018. *Stochastik für Einsteiger*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-22044-0.

Hyndman, Rob J., and Yanan Fan. 1996. "Sample Quantiles in Statistical Packages." *The American Statistician* 50 (4): 361. https://doi.org/10.2307/2684934.

McGill, Robert, John W. Tukey, and Wayne A. Larsen. 1978. "Variations of Box Plots." *The American Statistician* 32 (1): 12. https://doi.org/10.2307/2683468.