Anmerkungen zur Vorlesung Statistik vom 11.10.2016

Inhalt:

univariate deskriptive Statistik:

- Häufigkeiten und Darstellungsmöglichkeiten (Plots)
- Lageparameter: Modus, arithmetisches Mittel, Median, Quantile
- empirische Verteilungsfunktion

Aufgabensammlung Statistik:

04.10.2016
Hausaufgabe
A1 - A7

11.10.2016
Hausaufgabe
A8 - A13
A10: ohne
Boxplot

http://bit.ly/2dXMJY6

Notation:

X (Großbuchstaben) : Merkmalsvariable (später auch Zufallsvariable)
(typisch auch: Y, Z)

x : Realisationen einer Merkmalsvariable

i $(i = 1, \dots, n)$: Index des Realisationen, steht für einen bestimmten Merkmalsträger

n : Anzahl der Merkmalsträger

$(x_1, x_2, \dots, x_n)$: Urliste

weil Urlisten oft unübersichtlich → Häufigkeits-verteilungen

a_j : Merkmalsausprägung des Merk var
Index j $(j = 1, \dots, J)$

$h(a_j)$: absolute Häufigkeiten (# des Auftretens)

$H(a_j) = \sum_{k=1}^{j} h(a_k)$ kummulierte absolute Häufigkeiten

$f(a_j) = \frac{h(a_j)}{n}$: relative Häufigkeit

$F(a_j) = \sum_{k=1}^{j} f(a_k)$ kummulierte relative Häufigkeit

Anmerkung: $H(a_j)$ und $F(a_j)$ (≙ Häufigkeitsverteilungen) sind nur sinnvoll, falls Merkmal X geordnet werden kann.

Falls X metrisch ↝ empirische Verteilungsfunktion
(≙ Erweiterung der kummulierten relativen Häufigkeit auf alle Zwischenwerte)

↓ d.h.

wir betrachten nicht mehr nur $F(a_j)$, $j = 1, \dots, J$ sondern
$F(x)$ $\forall\, x \in \mathbb{R}$

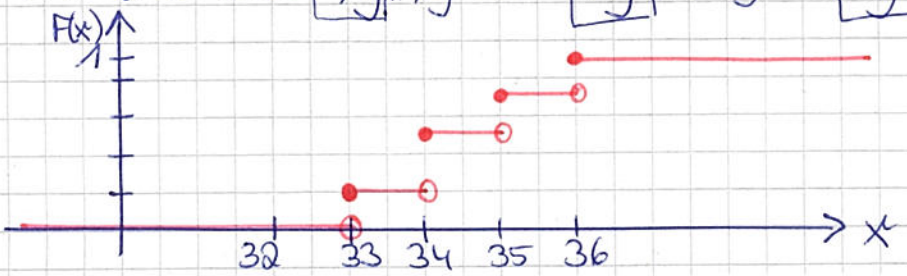
4.1 sortierte „Urliste": 33 33 34 34 34 35 35 36 36

j	1	2	3	4	
a_j	33	34	35	36	
$h(a_j)$	2	3	2	2	$\Sigma = 9 = n$
$f(a_j)$	2/9	3/9	2/9	2/9	
$F(a_j)$	2/9	5/9	7/9	9/9 = 1	

↓ empirische Vtlgsfkt. $F(x)$

-10 ... -5 ... 0 ... 32 32,9 33 33,1 ... 34 ... 34,5 ... 35 35,001 ... 36 36,001 ... 100

$F(x)$: 0 ... 0 ... 0 ... 0 ... 0 | 2/9 | 2/9 ... | 5/9 | 5/9 ... | 7/9 | 7/9 ... | 1 | 1 ... 1



F4.1 Wesentliche Eigenschaften der empirischen Vtlgsfkt.:

1) $F(x) = 0$ für $x < a_1$ (≙ kleinste Merkmalsausprägung)
2) $F(x) = 1$ für $x \geq a_j$ (≙ größte -"-)
3) „Treppenfunktion" mit Sprungstellen an den Merkmalsausprägungen a_j (→ # Sprungstellen = 7)
4) mit Sprunghöhe $f(a_j)$

⚠ Je mehr Sprungstellen, desto glatter wird die Funktion.

Beispiel ecdf für Variable „Alter" mit R

FL 4.2

? Für welchen Wert von x erreicht $F(x)$ einen bestimmten Wert p ($p \in [0;1]$)
↳ Wahrscheinlichkeit

Beispiel (Fortsetzung):

„Sortierte" Urliste: 33 33 34 34 34 35 35 36 36
$x_{[1]}$ $x_{[2]}$ $x_{[3]}$ $x_{[4]}$ $x_{[9]}$

also: $n = 9$

Ges.: $\tilde{x}_{0,4}$ und $\tilde{x}_{2/9}$

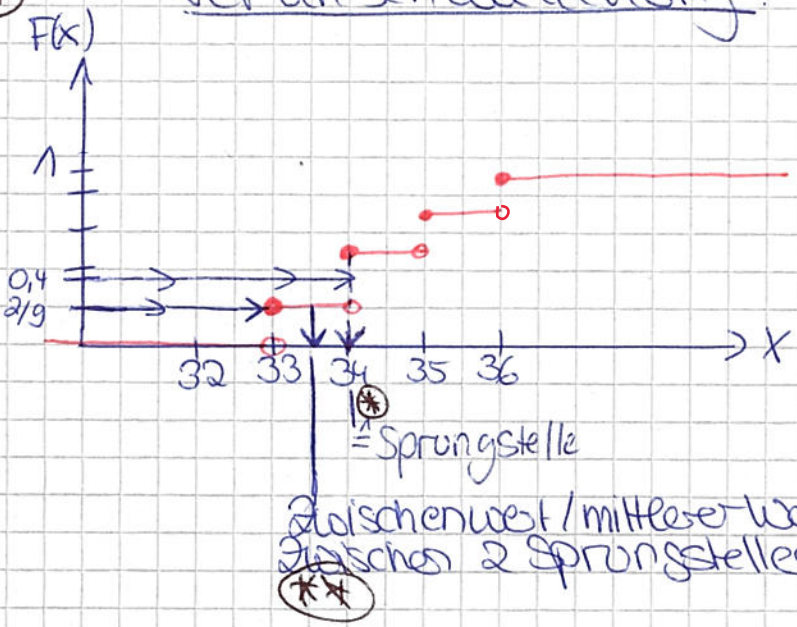
1) $\underbrace{0,4}_{p} \cdot \underbrace{9}_{n} = 3,6 \notin \mathbb{N}_0 \Rightarrow \tilde{x}_{0,4} = x_{\lceil n \cdot p \rceil} = x_{\lceil 3,6 \rceil} =$ Aufrundungsfunktion
$x_{[4]} = 34$
4. Wert der sort. Urliste

2) $\underbrace{2/9}_{p} \cdot \underbrace{9}_{n} = 2 \in \mathbb{N}_0 \Rightarrow \tilde{x}_{2/9} = \frac{1}{2}(x_{[n \cdot p]} + x_{[n \cdot p+1]}) =$
$= \frac{1}{2}(x_{[2]} + x_{[3]}) = \frac{1}{2} \cdot (33+34)$
$= 33,5$

(2) R-Umsetzung!

(4)

(1) Veranschaulichung:



(*) Quantieswert zu $p = 0{,}4$, also $\tilde{x}_{0.4}$

(**) Quantieswert zu $p = 2/9$, also $\tilde{x}_{2/9}$

F43-46

Visualisierungsmöglichkeiten mit R

43 • **Balkendiagramm** (Säulen): für absolute Häufigkeiten
für relative Häufigkeiten
Attribut: ylim = c(,)

44 • **Kreisdiagramm**: w_j = Winkel für Merkmalsausprägung a_j
$w_j = 360° \cdot f(a_j)$
Funktion: pie()

45 • **Balkendiagramme, Klassen getrennt oder gestapelt:** ...

346 • klassierte Daten → **Histogramme**
Besondere Eigenschaft: Fläche ist proport. zur Häufigk.

c ist frei wählbar,
z.B. c = 1 oder c = 1/12

Möglichkeit Mehrere und Mehr gegenüberstellen / nächster Schritt: Dichtefunktion std. Normv

F45dd

Nun: Lage- und Streuungsparameter & visuelle Darst.
- Modus
- Median
- arithm. Mittel
- Spannweite SP
- MSE (mittlere quad. Abw) MQA s^2
- Standardabweichung s
- Variationskoeffizient v
- Boxplot

F50

Modus: = Modalwert x_{Mod} = die häufigste Ausprägung a_j

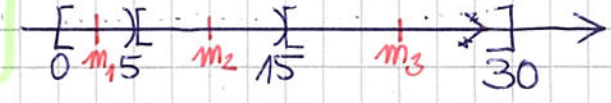
Median: = der mittlere Wert

$x_{Med} = \tilde{x}_{0.5}$

arithmetisches Mittel: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{k} a_j \cdot h(a_j) = \sum_{j=1}^{k} a_j \cdot f(a_j)$

bei klassierten Daten.
(≙ Schätzwert!) weil:

Beispiel von Folie 24:



→ Überschätzung: $\bar{x}$ ist zu hoch
→ Unterschätzung: $\bar{x}$ ist zu niedrig

F52

③ Frau = x_{Mod} mit größtes Anzahl

① na.exclude(.) falls Merkmalsvariable fehlende Werte enthält

② für metrische Merkmale
summary(.) statt separat median(.)/mean möglich

④ summary(.) für Faktoren ≙ Häufigkeitstabelle