Statistik

für Betriebswirtschaft, Internationales Management, Wirtschaftsinformatik und Informatik

Sommersemester 2017

HSA Statistik SS 2017 Sessionlist		
Datum	Statistik für BW/IM/I/Winf	Nr.
15.03.2017	Einführung Statistik	1
22.03.2017	Differentialrechnung, 2-dim Diff.Rechnung	2
29.03.2017	univ. deskr. Stat., Quantile, Plots	3
05.04.2017	Streuung, Konzentrationsmaße	4
12.04.2017	Kontingenztabellen, Mosaikplots, Korrelation	5
19.04.2017	Preisindizes, lineare Regression	6
26.04.2017	Wahrscheinlichkeitsbegriff	7
03.05.2017	Bedingte Wahrscheinlichkeit, Bayes	8
10.05.2017	diskrete Zufallsvariablen	9
17.05.2017	Stetige ZV, Gleichverteilung	10
24.05.2017	Pyramid	
31.05.2017	Normalverteilung, Verteilungsparameter	11
07.06.2017	Schätzfunktionen und Punktschätzer	12
14.06.2017	Konfidenzintervalle	13
21.06.2017	Wiederholung, Besprechung Probeklausur	14
28.06.2017	Prüfungswoche	15

Stefan Etschberger

- Voraussetzung: kardinale Werte $x_1, \ldots, x_n$
- **Beispiel:**

$$\left.\begin{array}{lll} \text{a)} & x_i & 1950 \quad 2000 \quad 2050 \\ \text{b)} & x_i & 0 \quad 0 \quad 6000 \end{array}\right\} \text{je } \bar{x} = 2000$$

- Spannweite: $SP = \max_i x_i - \min_i x_i$

  Im Beispiel:

$$\begin{array}{lll} \text{a)} & SP = 2050 - 1950 & = 100 \\ \text{b)} & SP = 6000 - 0 & = 6000 \end{array}$$

- Mittlere quadratische Abweichung:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}_{\text{Verschiebungssatz}}$$

Streuungsparameter (2)

- **Mittlere quadratische Abweichung** im Beispiel:

  a) $s^2 = \frac{1}{3} \cdot (50^2 + 0^2 + 50^2)$

  $= \frac{1}{3} \cdot (1950^2 + 2000^2 + 2050^2) - 2000^2 = 1666{,}67$

  b) $s^2 = \frac{1}{3} \cdot (2000^2 + 2000^2 + 4000^2)$

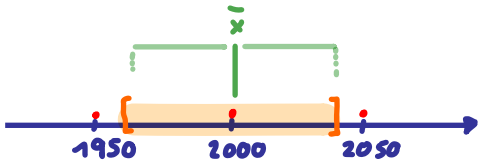
  $= \frac{1}{3} \cdot (0^2 + 0^2 + 6000^2) - 2000^2 = 8000000$

- **Standardabweichung:** $s = \sqrt{s^2}$

  Im Beispiel:

  a) $s = \sqrt{1666{,}67} = 40{,}82$

  b) $s = \sqrt{8000000} = 2828{,}43$



- **Variationskoeffizient:** $V = \frac{s}{\bar{x}}$ (maßstabsunabhängig)

  Im Beispiel:

  a) $V = \frac{40{,}82}{2000} = 0{,}02 \ (\widehat{=} 2\,\%)$

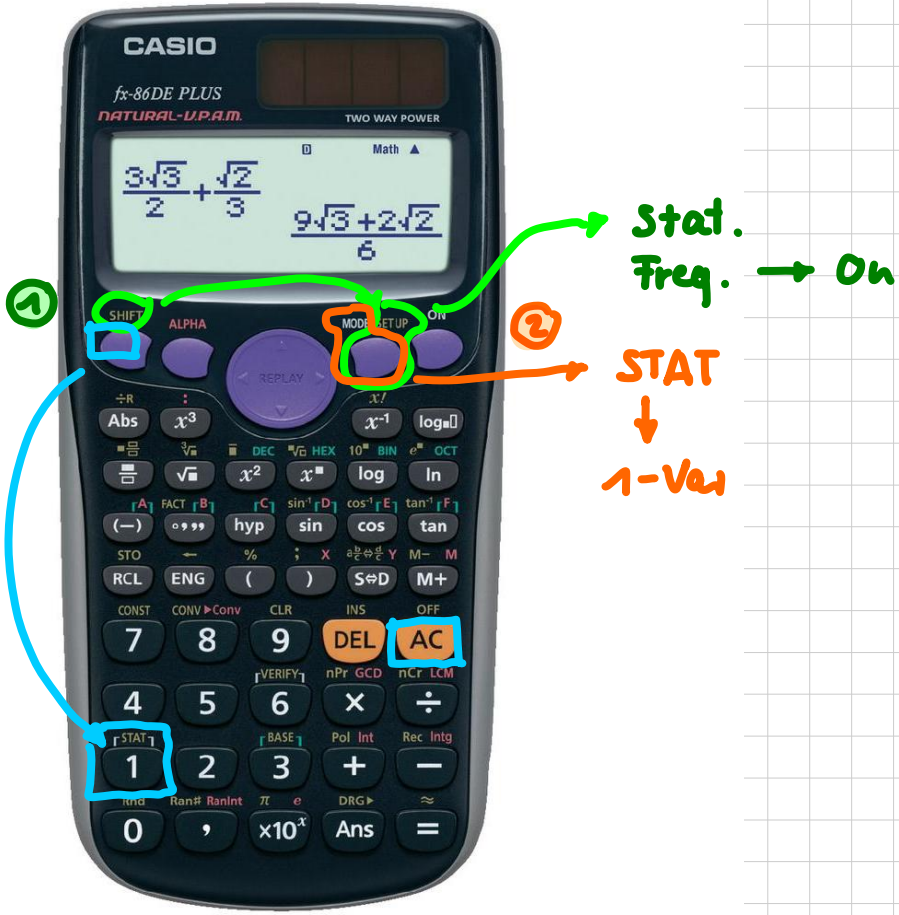
  b) $V = \frac{2828{,}43}{2000} = 1{,}41 \ (\widehat{=} 141\,\%)$

Filiale	s	$\bar{x}$	$V = \frac{s}{\bar{x}}$
1	1 Mio	20 Mio	0.05
2	500000	1 Mio	0.50
3	⋮	.	⋮
⋮			

Univariate Statistik mit dem TR

1. Häufigkeiten einschalten
2. Statistik modus (eindim.)
3. Daten eingeben : z.B.

x
1950
2000
2050



4. Ergebnisse : AC

Shift → STAT → Var

↳ $\bar{x}$ (2000) σ_x (40.82)

Lage und Streuung: Überblick

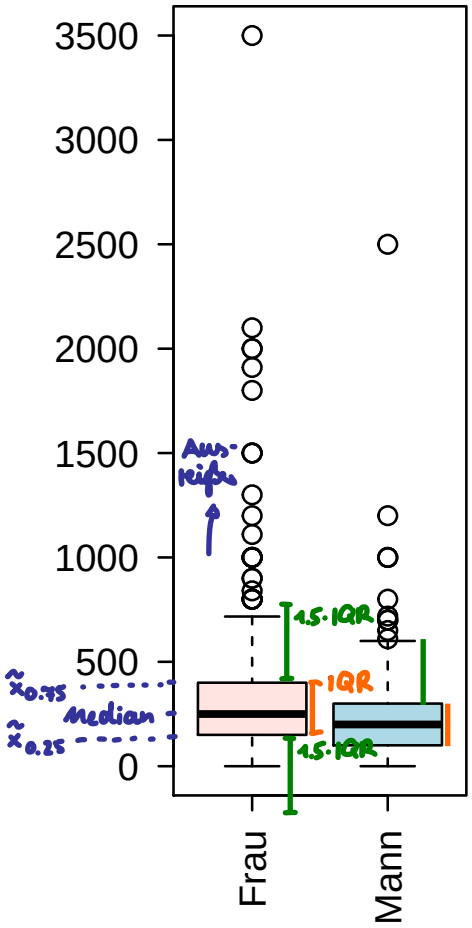

```r
LageStreuung = function(x) {
  x=na.omit(x)  # ignoriere fehlende Werte
  n = length(x) # Anzahl nicht fehlender Werte
  popV = var(x)*(n-1)/n # var() ist nicht mittl. qu. Abweichung
  return(list(mean=mean(x),
              median=median(x),
              Variance=popV,
              StdDev=sqrt(popV),
              VarCoeff=sqrt(popV)/mean(x)))
}
mat1 = sapply(MyData[c("Alter","AlterV","AlterM", # sapply: pro Spalte anwenden
                       "Geschwister", "AnzSchuhe", "AusgSchuhe")],
              LageStreuung)
```

	Alter	AlterV	AlterM	Geschwister	AnzSchuhe	AusgSchuhe
mean	22.00	54.41	51.69	1.51	21.11	278.69
median	21.00	54.00	51.00	1.00	15.00	200.00
Variance	11.12	35.68	25.47	1.21	403.21	74822.42
StdDev	3.33	5.97	5.05	1.10	20.08	273.54
VarCoeff	0.15	0.11	0.10	0.73	0.95	0.98

Lage und Streuung als Grafik: Boxplot

- Graphische Darstellung von Lage und Streuung
- Box: Oberer/Unterer Rand: 3. bzw. 1. Quartil ($\tilde{x}_{0,75}$ bzw. $\tilde{x}_{0,25}$),
- Linie in Mitte: Median
- Whiskers: Länge: Max./Min Wert, aber beschränkt durch das 1,5-fache des Quartilsabstands (falls größter/kleinster Wert größeren/kleineren Abstand von Box: Länge Whiskers durch größten/kleinsten Wert innerhalb dieser Schranken)
- Ausreißer: Alle Objekte außerhalb der Whisker-Grenzen

```
boxplot(AusgSchuhe ~ Geschlecht,
        col=c("mistyrose", "lightblue"),
        data=MyData, main="", las=2)
```



Ausgaben für Schuhe

Aufgabe 1

Stichprobe aus Umfrage: Größen

198 173 172 172 187 175 162 169

Geben Sie an bzw. zeichnen Sie:

a) Modus, Median, arithmetisches Mittel
b) empirische Quantile zu 25% und 75%
c) Boxplot

(Bearbeitungszeit: 8 Minuten)

a) Modus: 172, $\bar{x} = 176$

Median: Sortierte Urliste

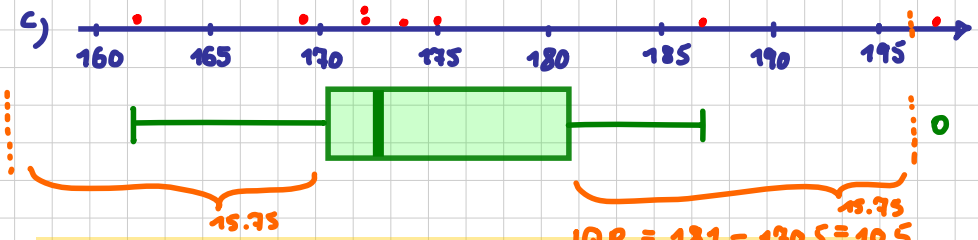
1	2	3	4	5	6	7	8
162	169	172	172	173	175	187	198

$\tilde{x}_{0.50} = \frac{1}{2}(x_4 + x_5) = \frac{1}{2}(172 + 173) = 172.5$

$8 \cdot 0.5 = 4 \in \mathbb{N}$

b) $\tilde{x}_{0.25} = \frac{1}{2}(x_2 + x_3) = 170.5$, $\tilde{x}_{0.75} = \frac{1}{2}(x_6 + x_7) = 181$

c)



$IQR = 181 - 170.5 = 10.5$

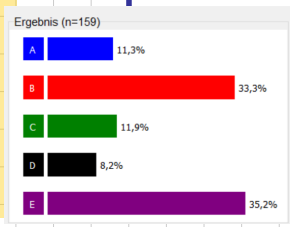
$1.5 \cdot IQR = 15.75$

$170.5 - 15.75 = 154.75$

$181 + 15.75 = 196.75$

Umfrage: In der letzten Aufgabe hatte ich

A) Alles richtig
B) Alles bis auf die Zeichnung richtig
C) Einen Fehler in den Zahlen
D) Mehr als einen Fehler in den Zahlen
E) Ich wusste nicht, was zu tun ist oder bin nicht fertig geworden



Aufgabe 2

Stichprobe aus Umfrage: Anzahl Geschwister

0	1	2	3	4	6
1	14	7	1	1	1

Geben Sie an bzw. zeichnen Sie:

a) Modus, Median, arithmetisches Mittel
b) empirische Quantile zu 25% und 75%
c) Boxplot

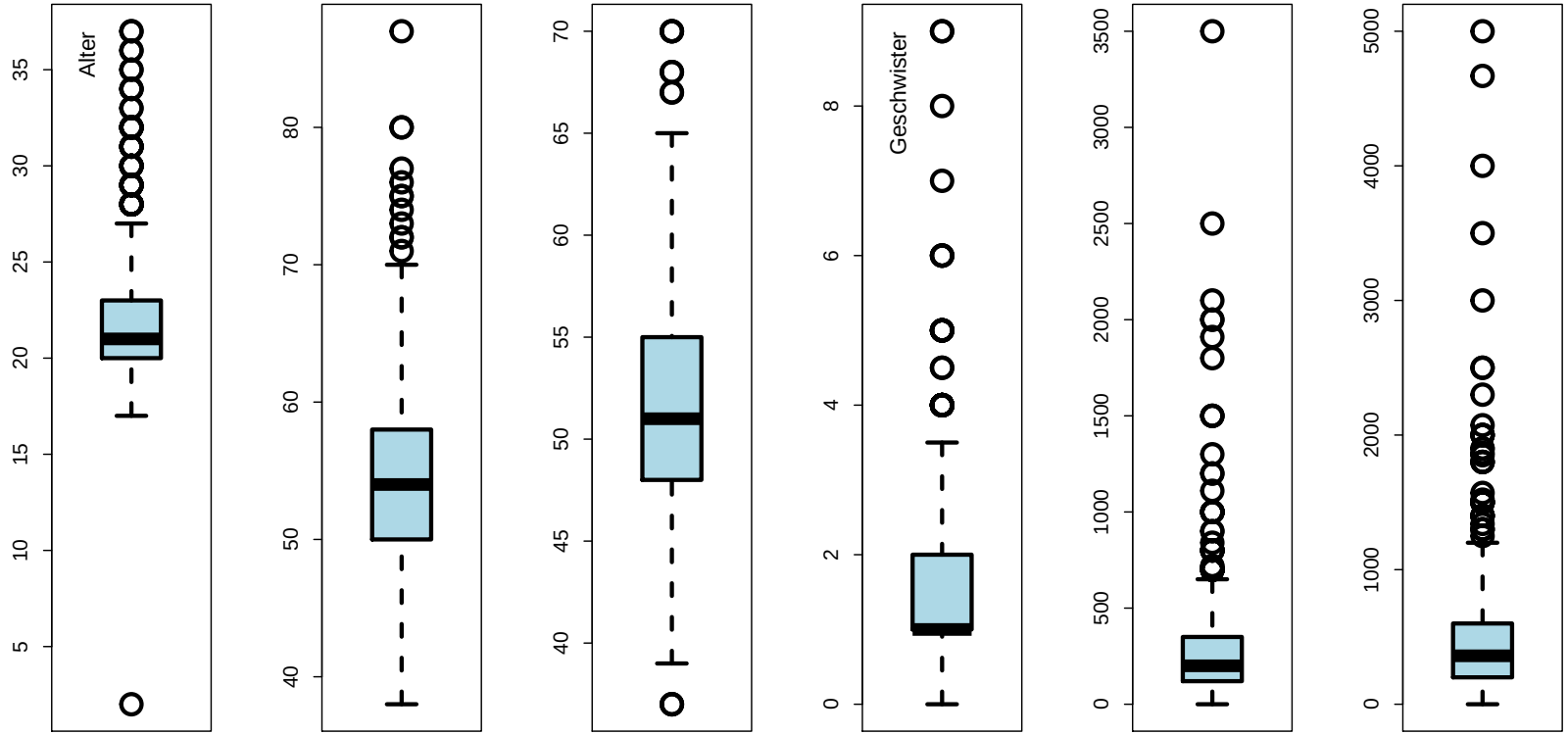
(Bearbeitungszeit: 6 Minuten)


```
summary(MyData)

##     Jahrgang       X              Alter       Groesse      Geschlecht     AlterV
##  Min.   :2014   Mode:logical   Min.   : 2   Min.   :150.0   Frau:543   Min.   :38.00
##  1st Qu.:2015   NA's:939       1st Qu.:20   1st Qu.:167.0   Mann:396   1st Qu.:50.00
##  Median :2016                  Median :21   Median :173.0              Median :54.00
##  Mean   :2016                  Mean   :22   Mean   :173.5              Mean   :54.41
##  3rd Qu.:2016                  3rd Qu.:23   3rd Qu.:180.0              3rd Qu.:58.00
##  Max.   :2017                  Max.   :37   Max.   :198.0              Max.   :87.00
##  NA's   :79                                                            NA's   :1
##      AlterM         GroesseV        GroesseM      Geschwister        Farbe        AusgKomm
##  Min.   :37.00   Min.   :157.0   Min.   : 76.0   Min.   :0.000   blau   : 42   Min.   :   0.0
##  1st Qu.:48.00   1st Qu.:175.0   1st Qu.:162.2   1st Qu.:1.000   gelb   : 10   1st Qu.: 200.0
##  Median :51.00   Median :180.0   Median :167.0   Median :1.000   rot    : 29   Median : 360.0
##  Mean   :51.69   Mean   :179.3   Mean   :166.5   Mean   :1.511   schwarz:475   Mean   : 464.2
##  3rd Qu.:55.00   3rd Qu.:183.0   3rd Qu.:170.0   3rd Qu.:2.000   silber :119   3rd Qu.: 600.0
##  Max.   :70.00   Max.   :204.0   Max.   :192.0   Max.   :9.000   weiss  :261   Max.   :5000.0
##  NA's   :1       NA's   :17      NA's   :13                      weiß   :  3   NA's   :2
##    AnzSchuhe        AusgSchuhe         Essgewohnheiten Raucher      NoteMathe
##  Min.   :  1.00   Min.   :   0.0   carnivor    :665    ja  :145   Min.   :1.000
##  1st Qu.:  8.00   1st Qu.: 120.0   fruktarisch :  3    nein:586   1st Qu.:2.300
##  Median : 15.00   Median : 200.0   pescetarisch: 36    NA's:208   Median :3.300
##  Mean   : 21.11   Mean   : 278.7   vegan       :  4               Mean   :3.257
##  3rd Qu.: 30.00   3rd Qu.: 350.0   vegetarisch : 26               3rd Qu.:4.000
##  Max.   :275.00   Max.   :3500.0   NA's        :205               Max.   :5.000
##                   NA's   :1                                       NA's   :227
##          MatheZufr   Studiengang
##  unzufrieden   :258   BW  :217
##  geht so       :193   ET  :  1
##  zufrieden     :159   IM  :153
##  sehr zufrieden:118   Inf : 57
##  NA's          :211   WI  :129
##                       NA's:382
##
```

Boxplots

```
for(attribute in c("Alter", "AlterV", "AlterM", "Geschwister",
                   "AusgSchuhe", "AusgKomm")) {
  data=MyData[, attribute]
  boxplot(data, # all rows, column of attribute
          col="lightblue",     # fill color
          lwd=3,               # line width
          cex=2,                # character size
          oma=c(1,1,2,1)
          )
  text(0.7,max(data), attribute, srt=90, adj=1)
}
```



- Gegeben: kardinale Werte $0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_n$
- **Achtung!** Die Werte müssen aufsteigend sortiert werden!
- Lorenzkurve:

> Wieviel Prozent der Merkmalssumme entfällt auf die x Prozent kleinsten Merkmalsträger?

- **Beispiel:** Die 90 % ärmsten besitzen 20 % des Gesamtvermögens.
- Streckenzug: $(0,0),\ (u_1,v_1),\ \ldots,\ (u_n,v_n) = (1,1)$ mit

$$v_k = \text{Anteil der } k \text{ kleinsten MM-Träger an der MM-Summe} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{n} x_i}$$

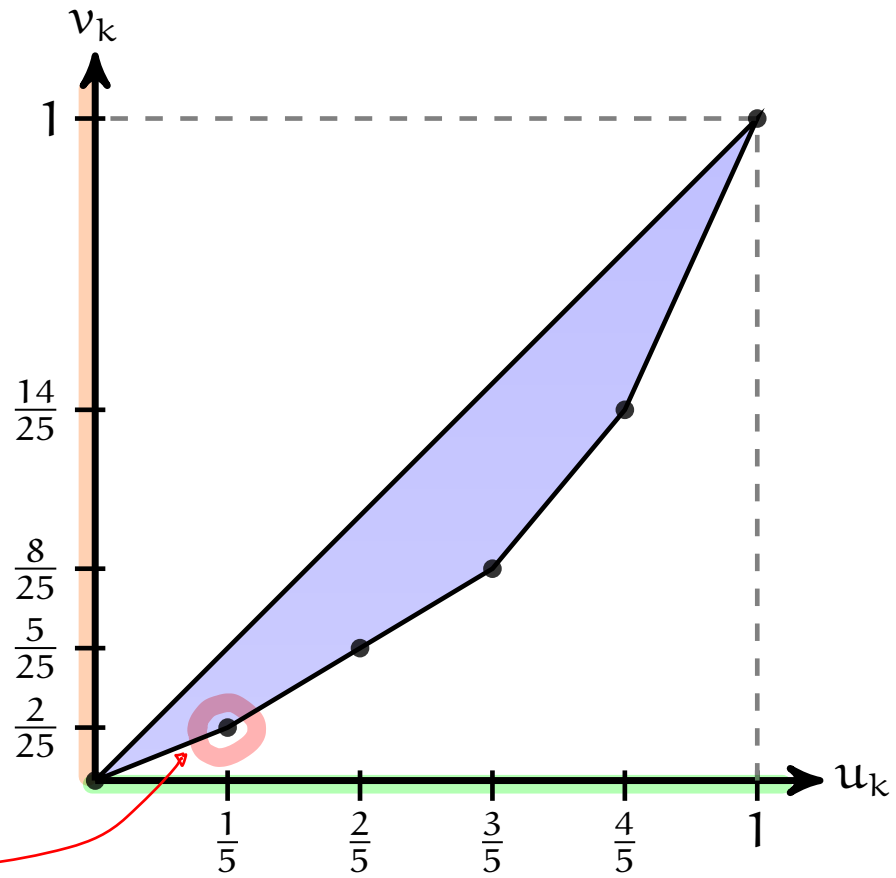
$$u_k = \text{Anteil der } k \text{ kleinsten an der Gesamtzahl der MM-Träger} = \frac{k}{n}$$

Lorenzkurve: Beispiel

Markt mit fünf Unternehmen; Umsätze: 6, 3, 11, 2, 3 (Mio. €)

$\Rightarrow n = 5,\ \sum_{k=1}^{5} x_k = 25$

k	1	2	3	4	5
x_k	2	3	3	6	11
p_k	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{3}{25}$	$\frac{6}{25}$	$\frac{11}{25}$
v_k	$\frac{2}{25}$	$\frac{5}{25}$	$\frac{8}{25}$	$\frac{14}{25}$	1
u_k	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1

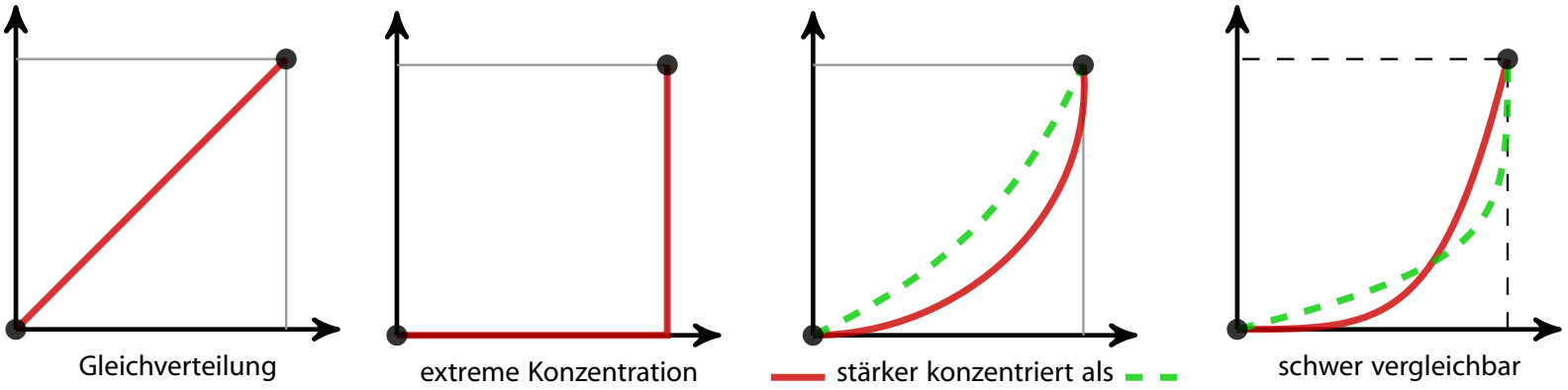


Knickstellen:

- Bei i-tem Merkmalsträger $\Longleftrightarrow$ $x_{i+1} > x_i$
- Empirische Verteilungsfunktion liefert Knickstellen:

a_j	2	3	6	11
$h(a_j)$	1	2	1	1
$f(a_j)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
$F(a_j)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1

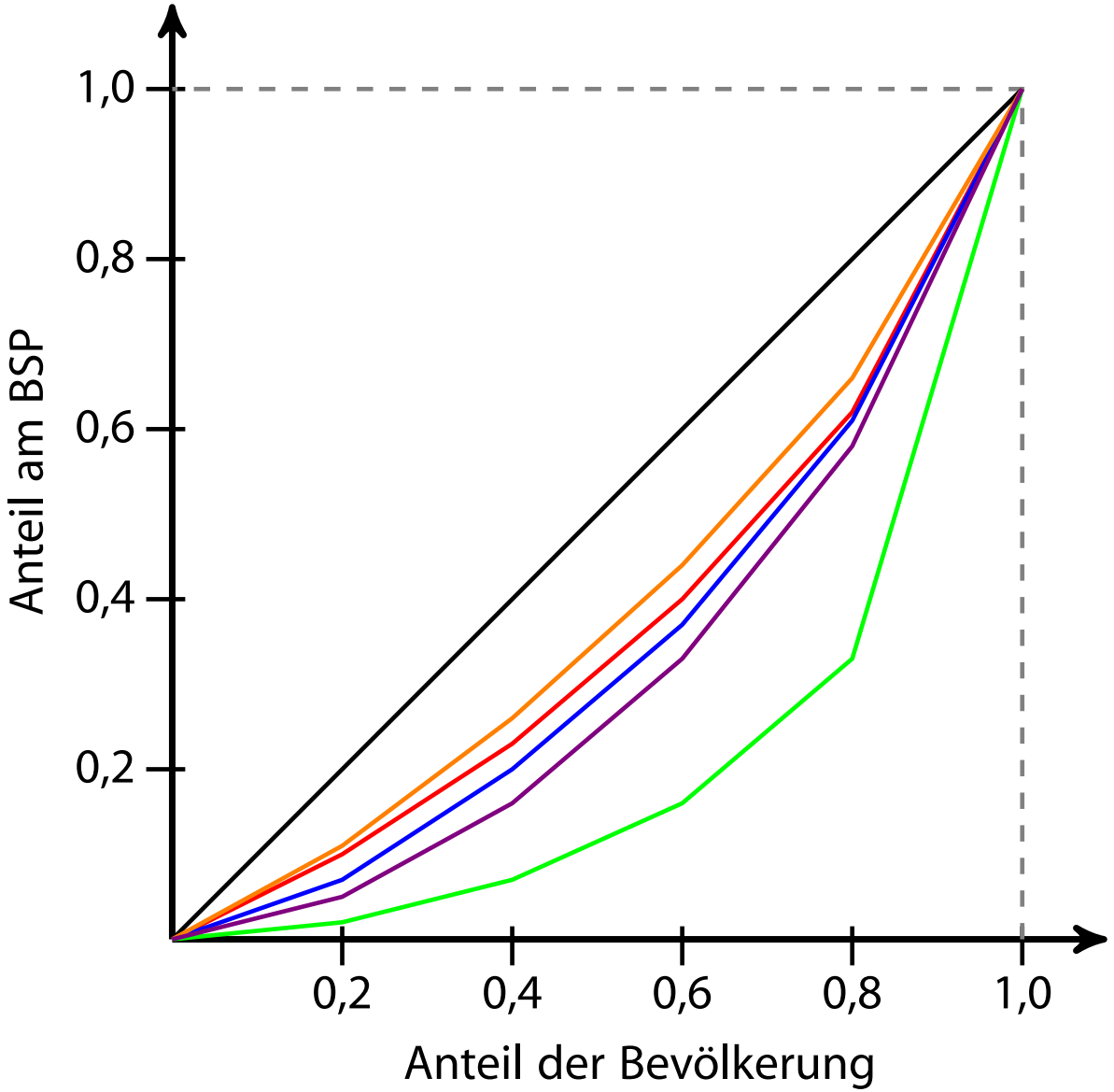
Vergleich von Lorenzkurven:



Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP

Bangladesch
Brasilien
Deutschland
Ungarn
USA

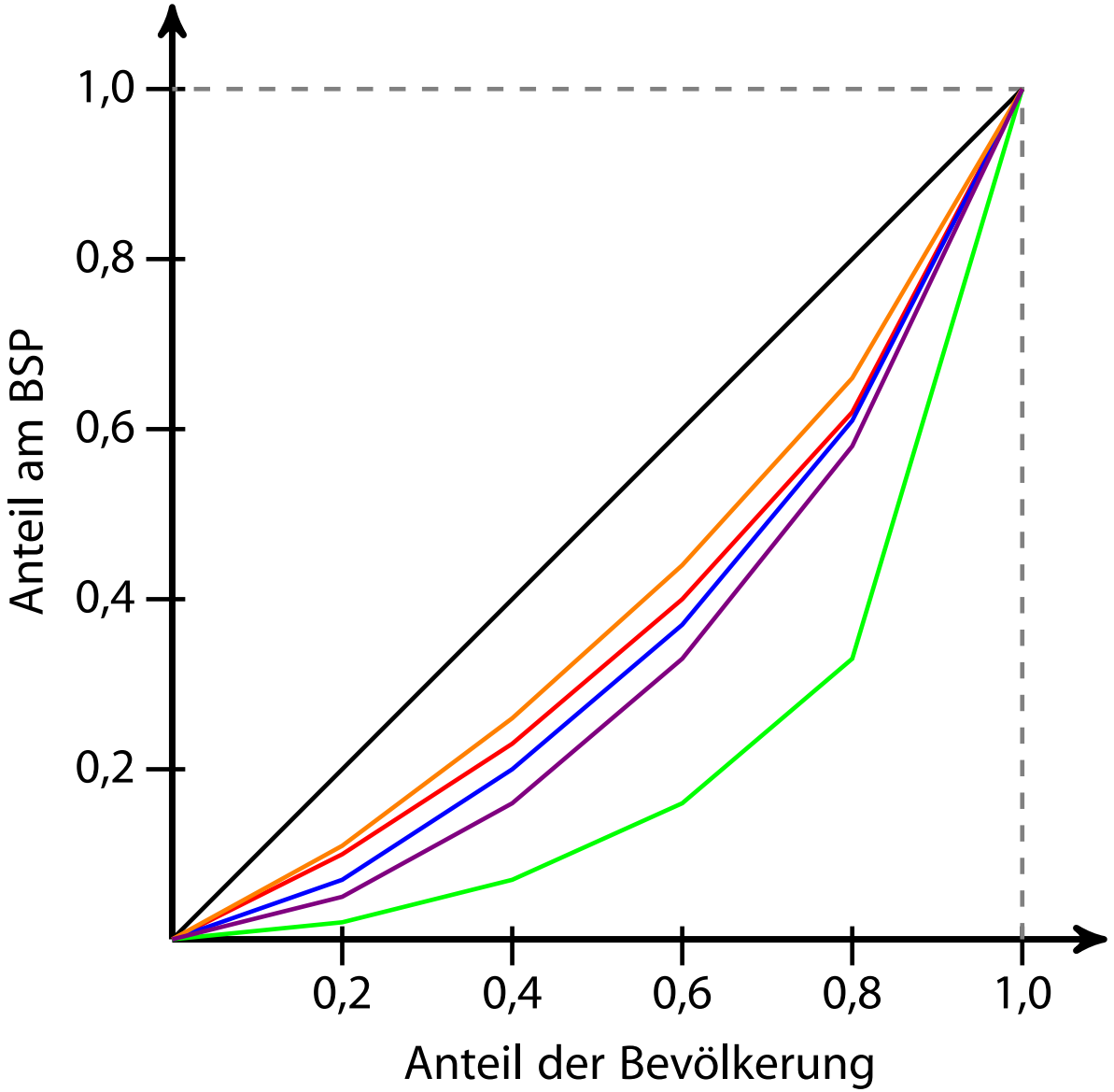
(Stand 2000)



Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP

Bangladesch
Brasilien
Deutschland
Ungarn
USA

(Stand 2000)



Gini-Koeffizient

- Numerisches Maß der Konzentration: Gini-Koeffizient G

$$G = \frac{\text{Fläche zwischen } 45°\text{-Linie und L}}{\text{Fläche unter } 45°\text{-Linie}} = \frac{\text{△}}{\text{△}} = \text{△} \cdot 2$$

- Aus den Daten:

$$G = \frac{2\sum_{i=1}^{n} i\,p_i - (n+1)}{n} \quad \text{wobei} \quad p_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

Bsp: $n=3$, x_i: 1, 2, 7 ($\sum x_i = 10$), p_i: 0.1, 0.2, 0.7

$G = \frac{1}{3}[2(1\cdot 0.1 + 2\cdot 0.2 + 3\cdot 0.7) - (3+1)]$

$= \frac{1}{3}[2\cdot 2.6 - 4] = 0.4$

- Problem: $G_{max} = \frac{n-1}{n}$

- Normierter Gini-Koeffizient:

$$G_* = \frac{n}{n-1} \cdot G \in [0;1]$$

Weitere Konzentrationsmaße

- Konzentrationskoeffizient:

$$CR_g = \text{Anteil, der auf die } g \text{ größten entfällt} = \sum_{i=n-g+1}^{n} p_i = 1 - v_{n-g}$$

- Herfindahl-Index:

$$H = \sum_{i=1}^{n} p_i^2 \qquad (\in [\tfrac{1}{n}; 1])$$

Es gilt: $H = \frac{1}{n}(V^2 + 1)$ bzw. $V = \sqrt{n \cdot H - 1}$

- Exponentialindex:

$$E = \prod_{i=1}^{n} p_i^{p_i} \qquad (\in [\tfrac{1}{n}; 1]) \qquad \text{wobei} \qquad 0^0 = 1$$

- Im Beispiel mit $x = (1, 2, 2, 15)$:

$$CR_2 = \tfrac{17}{20} = 0{,}85$$

$$H = \left(\frac{1}{20}\right)^2 + \cdots + \left(\frac{15}{20}\right)^2 = 0{,}59$$

$$E = \left(\frac{1}{20}\right)^{\frac{1}{20}} \cdots \left(\frac{15}{20}\right)^{\frac{15}{20}} = 0{,}44$$