Statistik

für Betriebswirtschaft, Internationales Management, Wirtschaftsinformatik und Informatik

Sommersemester 2016

Veranstaltungen zur Statistik für BW/IM Sommersemester 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Vorlesung Statistik	Etschberger	Mi, 14.00-17.00	B2.14	16.03.2016
Vorlesung Statistik PLUS	Etschberger/Jansen	- Blocktermin -	?	?
Übung Statistik	Etschberger	Mi, 17.00-18.30	A1.10	30.03.2016
Übung Statistik	Jansen	Di, 11.30-13.00	W1.06	22.03.2016
Übung Statistik	Jansen	Di, 14.00-15.30	W2.14	22.03.2016
Übung Statistik	Jansen	Mi, 11.30-13.00	W2.11	30.03.2016
Übung Statistik	Jansen	Do, 14.00-15.30	W2.14	31.03.2016
Übung Statistik	Schneller	Do, 14.00-15.30	W3.03	31.03.2016
Übung Statistik	Schneller	Do, 15.30-17.00	W3.03	31.03.2016
Übung Statistik	Wins	Di, 14.00-15.30	J3.19	22.03.2016
Übung Statistik	Wins	Di, 15.30-17.00	J3.13	22.03.2016
Offener Statistikraum	Etschberger/Tutoren	?	?	?
Veranstaltungen für Teilnehmer der WiMa-Klausur im Juli 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Tutorium Mathematik	Burkart	Do 13.30-15.00	W1.06	07.04.2016
Tutorium Mathematik	Burkart	Do 15.00-16.15	W1.06	07.04.2016
Offener Matheraum	Jansen/Tutoren	?	?	?

HSA Statistik SS 2016 Sessionlist		
Datum	Statistik für IM/BW	Nr.
Mittwoch, 16. März 2016	Einführung, R Installation, Rstudio Einführung, Skalen	1
Mittwoch, 23. März 2016	univ. deskr. Stat., Quantile, Plots	2
Mittwoch, 30. März 2016	Streuung, Konzentrationsmaße	3
Mittwoch, 6. April 2016	Kontingenztabellen, Mosaikplots, Korrelation	4
Mittwoch, 13. April 2016	Preisindizes, lineare Regression	5
Mittwoch, 20. April 2016	Kombinatorik, Wahrscheinlichkeit	6
Mittwoch, 27. April 2016	Wahrscheinlichkeit, diskrete Zufallsvariablen	7
Mittwoch, 4. Mai 2016	Pyramid	
Mittwoch, 11. Mai 2016	Binomial-, Hypergeom.-, Poisson-Verteilung	8
Mittwoch, 18. Mai 2016	Stetige ZV, Gleichverteilung	9
Mittwoch, 25. Mai 2016	Normalverteilung, Verteilungsparameter	10
Mittwoch, 1. Juni 2016	Schätzfunktionen und Punktschätzer	11
Mittwoch, 8. Juni 2016	Konfidenzintervalle	12
Mittwoch, 15. Juni 2016	Tests	13
Mittwoch, 22. Juni 2016	Puffer, WH, Fragen zur Probekl.	14
Mittwoch, 29. Juni 2016	AW Prüfungswoche	

Prof. Dr. Stefan Etschberger
Hochschule Augsburg

Outline

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

Laden	Standardabweichung	$\bar{x}$	$s/\bar{x} = V$
1	500 000	1 Mio	0.5
2	5 Mio	100 Mio	0.05
3	⋮		⋮
⋮			
100			

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

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

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

Lage und Streuung: Überblick

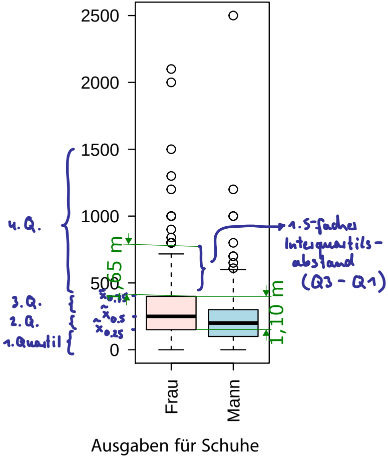

```r
LageStreuung = function(x) {
  x=na.omit(x)  # ignoriere fehlende Werte
  n = length(x) # Anzahl nicht fehlender Werte
  popV = var(x)*(n-1)/n # var() ist nicht mittl. qu. Abweichung
  return(list(mean=mean(x),
              median=median(x),
              Variance=popV,
              StdDev=sqrt(popV),
              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	22.13	54.28	51.64	1.51	21.22	270.45
median	21.00	54.00	51.00	1.00	16.00	200.00
Variance	11.36	35.35	25.74	1.18	415.51	56333.39
StdDev	3.37	5.95	5.07	1.08	20.38	237.35
VarCoeff	0.15	0.11	0.10	0.72	0.96	0.88

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

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



```
set.seed(4)
x = sample(Umfrage$Groesse, 10)
table(x)

164 166 168 170 171 179 185
  1   2   2   2   1   1   1
```

$n \cdot p = 10 \cdot 0.5 = 5 \in \mathbb{N}$

$\tilde{x}_{0.5} = \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(168+170) = 169 = x_{med}$

Geben Sie an bzw. zeichnen Sie:

$\bar{x} = 170.7$

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

$\tilde{x}_{0.25} = x_{\lceil 0.25 \cdot 10 \rceil} = x_{\lceil 2.5 \rceil} = x_3 = 166$

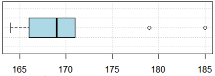
$\tilde{x}_{0.75} = x_8 = 171$

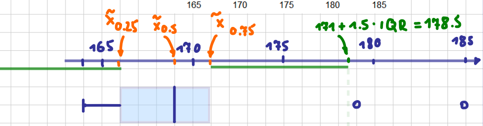
(Bearbeitungszeit: 7 Minuten)

$IQR = Q3 - Q1 = 171 - 166 = 5$

$1.5 \cdot IQR = 7.5$

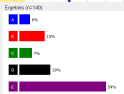
```
# Loesung:
quantile(x, probs=c(0.25, 0.5, 0.75), type=2)
boxplot(x, horizontal = TRUE, col="lightblue")

25% 50% 75%
166 169 171
```





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




```
set.seed(4)
x = sample(Umfrage$Geschwister, 10)
table(x)

 1 1.5   2   7
 4   1   4   1
```

Geben Sie an bzw. zeichnen Sie:

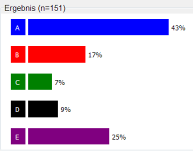
a) Modus, Median, arithmetisches Mittel
b) empirische Quantile zu 25% und 75%
c) Boxplot

(Bearbeitungszeit: 5 Minuten)

```
# Loesung
quantile(x, probs=c(0.25, 0.5, 0.75), type=2)
boxplot(x, horizontal = TRUE, col="lightblue")
```

Umfrage: In der letzten Aufgabe hatte ich

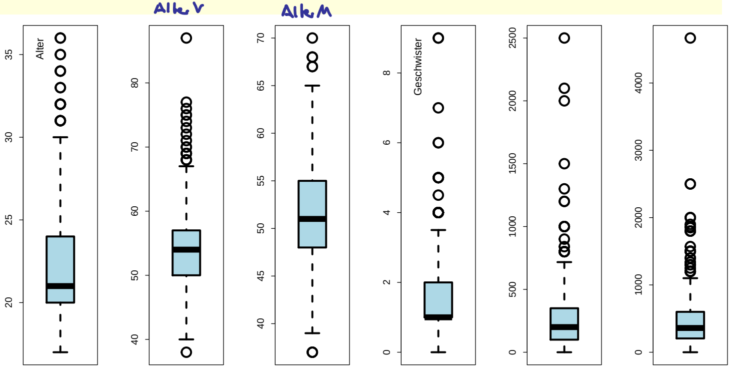
A) Alles richtig
B) Alles bis auf die Zeichnung richtig
C) Einen Fehler in den Zahlen
D) Mehr als einen Fehler in den Zahlen
E) Ich wusste nicht, was zu tun ist oder nicht fertig




```r
summary(MyData)

##     Jahrgang        Alter          Groesse      Geschlecht     AlterV          AlterM    
##  Min.   :2014   Min.   :17.00   Min.   :150.0   Frau:389   Min.   :38.00   Min.   :37.00  
##  1st Qu.:2014   1st Qu.:20.00   1st Qu.:166.0   Mann:281   1st Qu.:50.00   1st Qu.:48.00  
##  Median :2015   Median :21.00   Median :172.0              Median :54.00   Median :51.00  
##  Mean   :2015   Mean   :22.13   Mean   :173.1              Mean   :54.28   Mean   :51.64  
##  3rd Qu.:2016   3rd Qu.:24.00   3rd Qu.:180.0              3rd Qu.:57.00   3rd Qu.:55.00  
##  Max.   :2016   Max.   :36.00   Max.   :198.0              Max.   :87.00   Max.   :70.00  
##                                                            NA's   :1       NA's   :1      
##     GroesseV        GroesseM      Geschwister        Farbe        AusgKomm        AnzSchuhe     
##  Min.   :160.0   Min.   : 76.0   Min.   :0.000   blau   : 31   Min.   :   0.0   Min.   :  2.00  
##  1st Qu.:175.0   1st Qu.:162.0   1st Qu.:1.000   gelb   :  5   1st Qu.: 207.5   1st Qu.:  8.00  
##  Median :180.0   Median :165.0   Median :1.000   rot    : 24   Median : 360.0   Median : 16.00  
##  Mean   :179.1   Mean   :166.2   Mean   :1.509   schwarz:333   Mean   : 458.1   Mean   : 21.22  
##  3rd Qu.:183.0   3rd Qu.:170.0   3rd Qu.:2.000   silber : 82   3rd Qu.: 600.0   3rd Qu.: 30.00  
##  Max.   :204.0   Max.   :192.0   Max.   :9.000   weiss  :195   Max.   :4668.0   Max.   :275.00  
##  NA's   :11      NA's   :8                                     NA's   :2                        
##    AusgSchuhe        Essgewohnheiten Raucher      NoteMathe              MatheZufr   Studiengang
##  Min.   :   0.0   carnivor    :420   ja  : 81   Min.   :1.000   unzufrieden   :185   BW  :107   
##  1st Qu.: 100.0   fruktarisch :  1   nein:381   1st Qu.:2.650   geht so       :151   ET  :  1   
##  Median : 200.0   pescetarisch: 26   NA's:208   Median :3.300   zufrieden     :114   IM  : 74   
##  Mean   : 270.5   vegan       :  3              Mean   :3.233   sehr zufrieden: 74   Inf : 48   
##  3rd Qu.: 350.0   vegetarisch : 15              3rd Qu.:4.000   NA's          :146   WI  : 59   
##  Max.   :2500.0   NA's        :205              Max.   :5.000                        NA's:381   
##  NA's   :1                                      NA's   :162                                     
```

Boxplots

```r
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)
}
```



Konzentrationsmaße

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

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

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

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

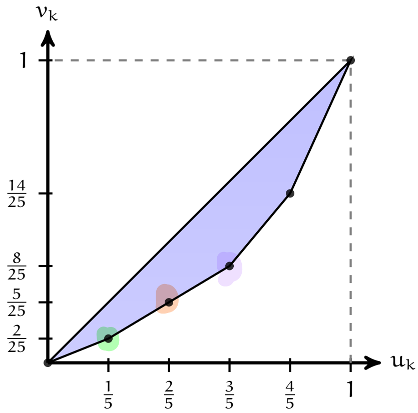
$$u_k = \text{Anteil der } k \text{ kleinsten an der Gesamtzahl der MM-Träger} = \frac{k}{n}$$

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

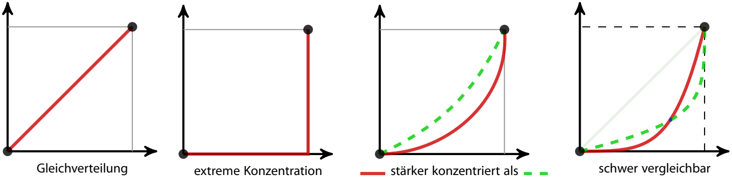


Knickstellen:

- Bei i-tem Merkmalsträger $\iff x_{i+1} > x_i$
- 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

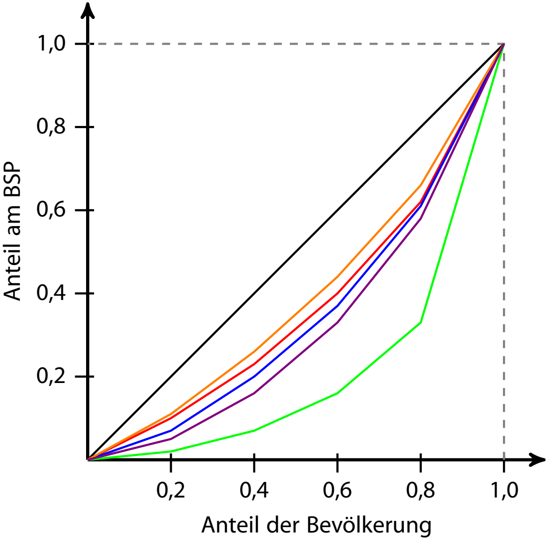
Vergleich von Lorenzkurven:



Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP

Bangladesch
Brasilien
Deutschland
Ungarn
USA

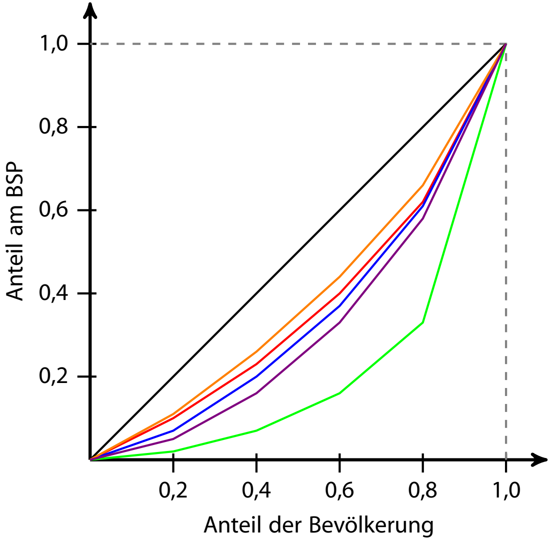
(Stand 2000)



Lorenzkurve: Beispiel Bevölkerungsanteil gegen BSP

Bangladesch
Brasilien
Deutschland
Ungarn
USA

(Stand 2000)



Gini-Koeffizient

- Numerisches Maß der Konzentration: Gini-Koeffizient G

$$G = \frac{\text{Fläche zwischen } 45^\circ\text{-Linie und } L}{\text{Fläche unter } 45^\circ\text{-Linie}} =$$

$= \cdot 2$

$\frac{1}{2}$

- Aus den Daten:

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

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

➡ Normierter Gini-Koeffizient:

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

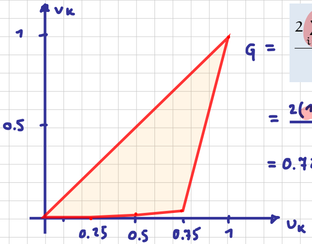
Beispiel (Lorenzk. und Gini-koeffiz.)

Urliste : Bargeld in Hosentasche

0, 96, 3, 1

Sortieren:

					Σ
	0	1	3	96	100
p_k	0	0.01	0.03	0.96	1
v_k	0	0.01	0.04	1	
u_k	1/4	2/4	3/4	1	

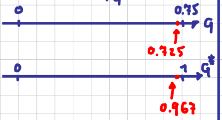


$$G = \frac{2\sum_{i=1}^{n} i\,p_i - (n+1)}{n}$$

$$= \frac{2(1\cdot 0 + 2\cdot 0.01 + 3\cdot 0.03 + 4\cdot 0.96) - (4+1)}{4}$$

$$= 0.725$$

normiert : $G^* = \frac{0.725}{3/4} \approx 0.967$



Jetzt : Urliste : 0, 0, 0, 100

$$\Rightarrow \quad G = \frac{2\cdot(\ 4\cdot 1.0\ ) - 5}{4} = \frac{3}{4}$$

Urliste : $\underbrace{0, 0, 0, \dots, 0}_{n-1}$, 100

$$G = \frac{2(n\cdot 1) - (n+1)}{n} = \frac{n-1}{n} = 1 - \frac{1}{n}$$

Gini-Koeffizient: Beispiel

Beispiel:

i	1	2	3	4	$\sum$
x_i	1	2	2	15	20
p_i	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{15}{20}$	1

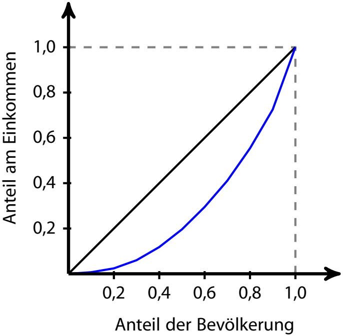
$$G = \frac{2 \cdot \left(1 \cdot \frac{1}{20} + 2 \cdot \frac{2}{20} + 3 \cdot \frac{2}{20} + 4 \cdot \frac{15}{20}\right) - (4+1)}{4} = 0{,}525$$

Mit $G_{\text{max}} = \frac{4-1}{4} = 0{,}75$ folgt

$$G_* = \frac{4}{4-1} \cdot 0{,}525 = 0{,}7$$

Armutsbericht der Bundesregierung 2008

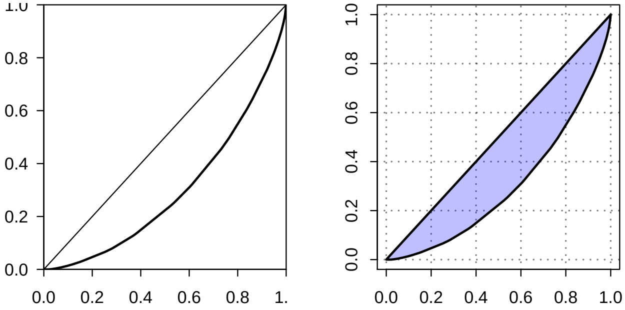
- Verteilung der Bruttoeinkommen in Preisen von 2000
- aus unselbständiger Arbeit der Arbeitnehmer/-innen insgesamt



	2002	2003	2004	2005
Arithmetisches Mittel	24.873	24.563	23.987	23.648
Median	21.857	21.531	20.438	20.089
Gini-Koeffizient	0,433	0,441	0,448	0,453

Lorenzkurve mit R


```r
require(ineq) # inequality Paket
Lorenz = Lc(na.exclude(MyData$AusgSchuhe))
   plot(Lorenz, xlab="", ylab="", main="") # Standard plot
   plot(c(0,1), c(0,1), type="n", # bisschen netter
       panel.first=grid(lwd=1.5, col=rgb(0,0,0,1/2)),
       xlab="", main="", ylab="")
   polygon(Lorenz$p, Lorenz$L, density=-1, col=rgb(0,0,1,1/4), lwd=2)
```




```r
Gini(na.exclude(AusgSchuhe)) # Gini-Koeffizient

## [1] 0.4069336
```

i	1	2	3	4	5
p_i	p_1	p_2	p_3	p_4	p_5

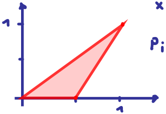
$CR_2 = \sum_{i=5-2+1}^{5} p_i = \sum_{i=4}^{5} p_i = p_4 + p_5$

- **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:**

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



$x = 0\text{ Mio}, 1\text{ Mio}$

$p_i = 0, 1$

$H = \sum p_i^2 = 0^2 + 1^2 = 1$

$E = \prod p_i^{p_i} = 0^0 \cdot 1^1 = 1$

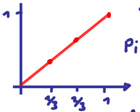
Es gilt: $H = \frac{1}{n}(V^2 + 1)$ bzw. $V = \sqrt{n \cdot H - 1}$

- **Exponentialindex:**

$0^0 \cdot 0.01^{0.01} \cdot 0.03^{0.03} \cdot 0.96^{0.96}$

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

Urlisk: $x = 2\text{ Mio}, 2\text{ Mio}, 2\text{ Mio}$



$p_i = 1/3, 1/3, 1/3$

$H = \sum p_i^2 = (1/3)^2 + (1/3)^2 + (1/3)^2 = 3 \cdot 1/9 = 1/3$

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

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

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

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

Beispiel: $x = \underbrace{1\text{ Mio}, 1\text{ Mio}, \ldots, 1\text{ Mio}}_{n\text{-mal}}$

$p_i = 1/n, 1/n, \ldots, 1/n$

$\Rightarrow H = (1/n)^2 + \cdots + (1/n)^2 = n \cdot (1/n)^2 = 1/n$