1.

Absolute Häufigkeit:	h
Wie viele Elemente weisen diesen bestimmten Wert *(= diese bestimmte Ausprägung)* auf? > Anzahl h der Elemente a mit der Ausprägung i	$h_i := h(a_i)$
⚁ ⚃ ⚃ ⚁ ⚅ ⚃ ⚅ ⚃ ⚁ ⚃	$h_2 := h(\text{der Elemente mit der Ausprägung}_{i=2}) = 3$ $h_4 := h(\text{der Elemente mit der Ausprägung}_{i=4}) = 5$ $h_6 := h(\text{der Elemente mit der Ausprägung}_{i=6}) = 2$

⚁	⚃	⚅
3	5	2

2.

Relative Häufigkeit:	f
Wieviel Prozent aller Elemente weisen diesen bestimmten Wert auf? > Anzahl h der Elemente a mit der Ausprägung i im Verhältnis *geteilt durch* zu der Anzahl n aller Werte	$f_i = \frac{h(a_i)}{n}$
⚁ ⚃ ⚃ ⚁ ⚅ ⚃ ⚅ ⚃ ⚁ ⚃	$f_2 = \frac{h(\text{Elemente mit der Ausprägung}_{i=2})}{\text{Anzahl aller Werte}} = \frac{3}{10} = 0{,}3 = 30\,\%$ $f_4 = \frac{h(\text{der El. mit der Ausprägung}_{i=4})}{\text{Anzahl aller Werte}} = \frac{5}{10} = 0{,}5 = 50\,\%$ $f_6 = \frac{h(\text{der El. mit der Ausprägung}_{i=6})}{\text{Anzahl aller Werte}} = \frac{2}{10} = 0{,}2 = 20\,\%$

⚁	⚃	⚅
$\frac{3}{10}$	$\frac{5}{10}$	$\frac{2}{10}$

Absolute kumulierte Häufigkeit:	H
Wie viele aller Elemente weisen *höchstens* diesen Wert auf? > *aufaddierte* Summen Σ der absoluten Häufigkeit h der Elemente a mit der Ausprägung i im geordneten Datensatz von Position $1\, bis\, j$, bis zu $a_i \leq einem\ bestimmten\ Wert\ x$	$H(x) = h(a_1) + \ldots + h(a_j) = \sum_{i: a_i \leq x} h_i$

⚁ ⚁ ⚁ ⚄ ⚄ ⚃ ⚄ ⚃ ⚅ ⚅

⚁	⚃	⚅
3	3+5	3+5+2

$H(des\ Wertes\ 2) = h(der\ Elemente\ mit\ dem\ Wert\ 2) = 3$

$H(des\ Wertes\ 4) = h(der\ El.\ mit\ der\ Ausprägung_{i=2}) + h(der\ El.\ mit\ der\ Ausprägung_{i=4}) = 3 + 5 = 8$

$H(des\ Wertes\ 5)$
$= h(der\ El.\ mit\ der\ Ausprägung_{i=2}) + h(der\ El.\ mit\ der\ Ausprägung_{i=4}) + h(der\ El.\ mit\ der\ Ausprägung_{i=5})$
$= 3 + 5 + 0 = 8$

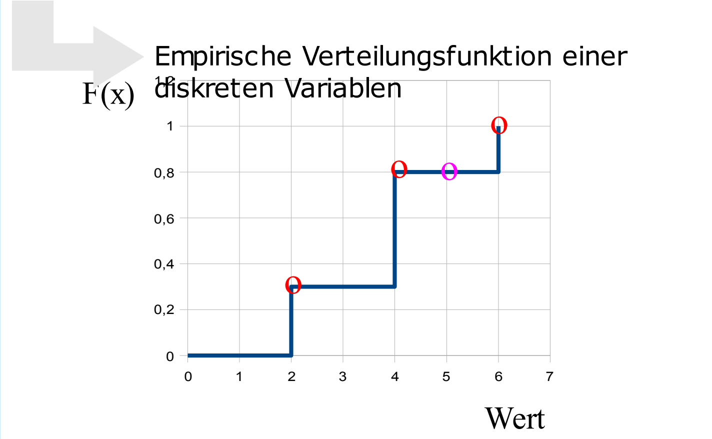
$H(des\ Wertes\ 6)$
$= h(der\ El.\ mit\ der\ Ausprägung_{i=2}) + h(der\ El.\ mit\ der\ Ausprägung_{i=4}) + h(der\ El.\ mit\ der\ Ausprägung_{i=6})$
$= 3 + 5 + 2 = 10$

4.

Relative kumulierte Häufigkeit:	F
Wieviel Prozent aller Elemente weisen *höchstens* diesen Wert auf?	
> *aufaddierte* Summen Σ der relativen Häufigkeit f der Elemente a mit der Ausprägung i im geordneten Datensatz von Position $1\,bis\,j$, bis zu $a_i \leq einem\ bestimmten\ Wert\ x$	$F(x)=f(a_1)+...+f(a_j)=\sum_{i:a_i\leq x} f_i$
> absolute kumulierte Häufigkeit H des Wertes x im Verhältnis zu $geteilt\ durch$ der Anzahl n aller Werte	$F(x)=\frac{H(x)}{n}$

⚁ ⚁ ⚁ ⚃ ⚃ ⚃ ⚃ ⚃ ⚅ ⚅

⚁	⚃	⚅
$\frac{3}{10}$	$\frac{3+5}{10}$	$\frac{3+4+2}{10}$



nach: $F(x)=f(a_1)+...+f(a_j)=\sum_{i:a_i\leq x} f_i$

$F(des\ Wertes\ 1)=f(der\ Elemente\ mit\ dem\ Wert\ 1)=0{,}2=20\,\%$

$F(des\ Wertes\ 4)=f(der\ El.\ m.d.Auspr._{i=2})+f(der\ El.\ m.d.Auspr._{i=4})=0{,}3+0{,}5=0{,}8=80\,\%$

$F(des\ Wertes\ 5)=f(der\ El.\ m.d.Auspr._{i=2})+f(der\ El.\ m.d.Auspr._{i=4})+f(der\ El.\ m.d.Auspr._{i=5})$
$=0{,}3+0{,}5+0=0{,}8=80\,\%$

$F(des\ Wertes\ 6)=f(der\ El.\ m.d.Auspr._{i=2})+f(der\ El.\ m.d.Auspr._{i=4})+f(der\ El.\ m.d.Auspr._{i=6})$
$=0{,}3+0{,}5+0{,}2=1=100\,\%$

nach: $F(x)=\frac{H(x)}{n}$

$$F(des\ Wertes\ 2)=\frac{H(des\ Wertes\ 2)}{Anzahl\ aller\ Werte}=\frac{3}{10}=0{,}3=30\,\%$$

$$F(des\ Wertes\ 4)=\frac{H(des\ Wertes\ 4)}{Anzahl\ aller\ Werte}=\frac{8}{10}=0{,}8=80\,\%$$

$$F(des\ Wertes\ 6)=\frac{H(des\ Wertes\ 6)}{Anzahl\ aller\ Werte}=\frac{10}{10}=1=100\,\%$$

5.

Modus / Modalwert:	x_{mod}
Welcher Wert tritt am häufigsten auf?	
⚂ ⚃ ⚃ ⚂ ⚅ ⚃ ⚅ ⚃ ⚂ ⚃ ⚂: 3, ⚃: 5, ⚅: 2	$x_{mod}=4$

⚂	⚃	⚅
3	5	2

6.

Median:	$\tilde{x}$ oder x_{med}
Welcher Wert steht in der Mitte *des geordneten Datensatzes*? > *wenn n ungerade:* Wert x an der mittleren Position i des geordneten Datensatzes (mit dem Umfang n), für die gilt: $i=\frac{n+1}{2}$ > *wenn n gerade:* Mitte der Werte x an den *beiden* mittleren Positionen i des geordneten Datensatzes (mit dem Umfang n), für die gilt: $i=\frac{n}{2}$ bzw. $i=\frac{n}{2}+1$	$\tilde{x}=x_{(\frac{n+1}{2})}$ *falls n ungerade* $\tilde{x}=\frac{1}{2}\cdot(x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)})$ *falls n gerade* und metrisch skaliert!
(Tabelle unten)	n = 10 -> gerade: $\tilde{x}=\frac{1}{2}\cdot(\mathit{Wert\ an\ der\ Position}_{(\frac{10}{2})}+\mathit{Wert\ an\ der\ Position}_{(\frac{10}{2}+1)})$ $=\frac{1}{2}\cdot(\mathit{Wert\ an\ der\ Position}_5+\mathit{Wert\ an\ der\ Position}_6)$ $=\frac{1}{2}\cdot(4+4)=\frac{8}{2}=4$

1	2	3	4	5	6	7	8	9	10
⚂	⚂	⚂	⚃	⚃	⚃	⚃	⚃	⚅	⚅

7.

Mittelwert: $\overline{x}$

arithmetisches Mittel
(bei metrisch skalierten Daten)

Werte:

$$\frac{\textit{Summer aller Werte}}{\textit{Anzahl der Werte}}$$

$$\overline{x} := \frac{1}{n} \cdot (x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Absolute Häufigkeiten:

$$\frac{\textit{Wert} \cdot \textit{wie oft} + \textit{Wert} \cdot \textit{wie oft} + \ldots}{\textit{Anzahl der Werte}}$$

$$\overline{x} := \frac{1}{n} \cdot (a_1 \cdot h_1 + \ldots + a_k \cdot h_k) = \frac{1}{n} \cdot \sum_{i=1}^{n} a_j \cdot h_j$$

Relative Häufigkeiten:

$$\begin{gathered} \textit{Wert} \cdot \textit{wie oft}\,(\textit{relativ}) \\ + \\ \textit{Wert} \cdot \textit{wie oft}\,(\textit{relativ}) \\ + \\ \ldots \end{gathered}$$

$$\overline{x} := a_1 \cdot f_1 + \ldots + a_k \cdot f_k = \sum_{j=1}^{k} a_j \cdot f_j$$

⚁ ⚃ ⚃ ⚁ ⚅ ⚃ ⚅ ⚃ ⚁ ⚃

⚁	⚃	⚅
3	5	2

$$\overline{x} := \frac{1}{10} \cdot (2+4+4+2+6+4+6+4+2+4) = 3{,}8$$

$$=> \quad \frac{1}{10} \cdot [2 \cdot 3 + 4 \cdot 5 + 6 \cdot 2]$$

$$=> \quad 2 \cdot 0{,}3 + 4 \cdot 0{,}5 + 6 \cdot 0{,}2$$

Empirische Varianz / Stichprobenvarianz:

s^2 oder $\tilde{s}^2$

Wie stark streuen die Daten um den Mittelwert?

-> Mittel der *quadrierten Abweichungen* $\boxed{(x_i-\bar{x})^2}$

Werte:

$$\frac{(Wert-Mittelwert)^2+(Wert-Mittelwert)^2+\ldots}{Anzahl\ der\ Werte}$$

$$s^2=\frac{1}{n}\cdot[(x_1-\bar{x})^2+\ldots+(x_n-\bar{x})^2]$$
$$=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

quadrierte Werte / Zerlegungsformel:

$$\frac{(Wert^2\text{-}Mittelwert^2)+(Wert^2\text{-}Mittelwert^2)+\ldots}{Anzahl\ der\ Werte}$$

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i^2-\bar{x}^2)=\overline{x^2}-\bar{x}^2$$

Häufigkeiten:

$$(Wert-Mittelwert)^2\cdot wie\,oft\,(relativ)$$
$$+$$
$$(Wert-Mittelwert)^2\cdot wie\,oft\,(relativ)$$
$$+$$
$$\ldots$$

$$s^2=(a_1-\bar{x})^2\cdot f_1+\ldots+(a_k-\bar{x})^2\cdot f_k$$
$$=\sum_{j=1}^{k}(a_j-\bar{x})^2\cdot f_j$$

⚁ ⚃ ⚃ ⚁ ⚅ ⚃ ⚅ ⚃ ⚁ ⚃

⚁	⚃	⚅
3	5	2

$\bar{x}=3{,}8$

$$s^2=\frac{1}{10}[(2-3{,}8)^2+(4-3{,}8)^2+(4-3{,}8)^2+(2-3{,}8)^2+\ldots]$$
$$=1{,}96$$

=> $s^2=\frac{1}{10}[(2^2-3{,}8^2)+(4^2-3{,}8^2)+(4^2-3{,}8^2)+\ldots]$

=> $s^2=(2-3{,}8)^2\cdot 0{,}3+(4-3{,}8)^2\cdot 0{,}5+(6-3{,}8)^2\cdot 0{,}2$

Korrigierte Varianz:

(beim Schätzen und Testen bevorzugt)

$$s^{*2}:=\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2=\boxed{\frac{n}{n-1}\cdot s^2}$$

$$s^{*2}=\frac{10}{10-1}\cdot 1{,}96\approx 2{,}18$$

Standardabweichung

Wie weit weichen die einzelnen Werte *im Durchschnitt* vom Mittelwert ab?

$$s=\sqrt{s^2}$$

$$s=\sqrt{1{,}96}=1{,}4$$

9.

Quantile, p-Quantil

x_p

Wert *an der Stelle des geordneten Datensatzes*, an der eine bestimmte %zahl aller Elemente < *oder* = dem Wert an dieser Stelle sind

falls Stelle $i = n \cdot p = ganze\ Zahl$

$$\frac{Wert\ an\ der\ Stelle\ (i) + Wert\ an\ der\ Stelle\ (i+1)}{2}$$

im geordneten Datensatz

falls np ganzzahlig

$$x_p = \frac{1}{2} \cdot \left(x_{(n \cdot p)} + x_{([n \cdot p]+1)}\right)$$

falls Stelle $i = n \cdot p = Kommazahl$

$$Wert\ an\ der\ Stelle\ (i+1)$$

im geordneten Datensatz

falls np nicht ganzzahlig:

$$x_p = x_{([n \cdot p]+1)}$$

1	2	3	4	5	6	7	8	9	10
⚁	⚁	⚁	⚃	⚃	⚃	⚃	⚃	⚅	⚅

p=0,25 **unteres Quartil**

$$x_{0,25} = x_{([10 \cdot 0{,}25]+1)} = x_{3,5} = 2$$

p=0,75 **oberes Quartil**

$$x_{0,75} = x_{([10 \cdot 0{,}75]+1)} = x_{8,5} = 4$$

p=0,5 Median

$$x_p = \frac{1}{2} \cdot \left(x_{(10 \cdot 0{,}5)} + x_{([10 \cdot 0{,}5]+1)}\right) = \frac{1}{2} \cdot (x_5 + x_6)$$

$$= \frac{4+4}{2} = 4$$

p=0,1 **Dezil D1**

$$x_p = \frac{1}{2} \cdot \left(x_{(10 \cdot 0{,}1)} + x_{([10 \cdot 0{,}1]+1)}\right) = \frac{1}{2} \cdot (x_1 + x_2)$$

$$= \frac{2+2}{2} = 2$$

p=0,9 **Dezil D9**

$$x_p = \frac{1}{2} \cdot \left(x_{(10 \cdot 0{,}9)} + x_{([10 \cdot 0{,}9]+1)}\right) = \frac{1}{2} \cdot (x_9 + x_{10})$$

$$= \frac{6+6}{2} = 6$$

Quartilsabstand / Interquartilsabstand

Abstand zwischen oberem und unterem Quartil

$$Q := x_{0,75} - x_{0,25}$$

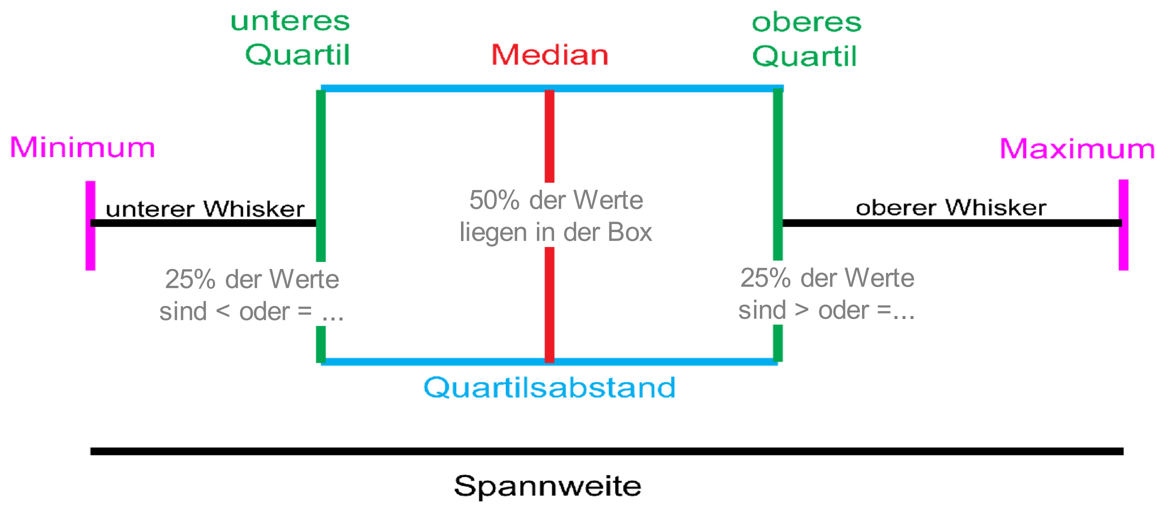
$$Q := x_{8,5} - x_{3,5} = 4 - 2 = 2$$

10.

Spannweite / Range	R
Abstand zwischen kleinstem und größtem Wert *größter (letzter) Wert - kleinster (erster) Wert* im geordneten Datensatz	$R := x_{(n)} - x_{(1)}$
1 2 3 4 5 6 7 8 9 10 ⚁ ⚁ ⚁ ⚃ ⚃ ⚃ ⚃ ⚃ ⚅ ⚅	$R := 6 - 2 = 4$

11.

BOXPLOTS



Kennwert	Beschreibung	Lage im Boxplot
Minimum	Kleinster Datenwert des Datensatzes	Ende eines Whiskers oder entferntester Ausreißer
Unteres Quartil	Die kleinsten 25% der Datenwerte sind kleiner oder gleich diesem Kennwert	Beginn der Box
Zentralwert oder Median	Die kleinsten 50% der Datenwerte sind kleiner oder gleich diesem Kennwert	Senkrechter Strich innerhalb der Box nicht zwangsläufig in der Mitte der Box!
Oberes Quartil	Die kleinsten 75% der Datenwerte sind kleiner oder gleich diesem Kennwert	Ende der Box
Maximum	Größter Datenwert des Datensatzes	Ende eines Whiskers oder entferntester Ausreißer
Spannweite	Gesamter Wertebereich des Datensatzes	Länge des gesamten Boxplots (inklusive Ausreißer)
Quartilabstand	Wertebereich in dem sich die mittleren 50% der Daten befinden	Ausdehnung der Box

Quelle: http://de.wikipedia.org/wiki/Boxplot#Zusammenfassung_der_Kennwerte

12.

LORENZKURVE

u_i v_i

Wieviel Prozent der Summe aller Werte verteilen sich auf wieviel Prozent der Merkmalsträger?

-> *Relative* Konzentration
im geordneten Datensatz

xAchse: $\frac{Anzahl\ i\ der\ Merkmalsträger}{Gesamtanzahl\ n\ der\ Merkmalsträger}$

$u_i := \frac{i}{n}$

yAchse: $\frac{Merkmalssumme\ bis\ zum\ Wert_i(p_i)}{Gesamtsumme\ der\ Merkmale\ p_n}$

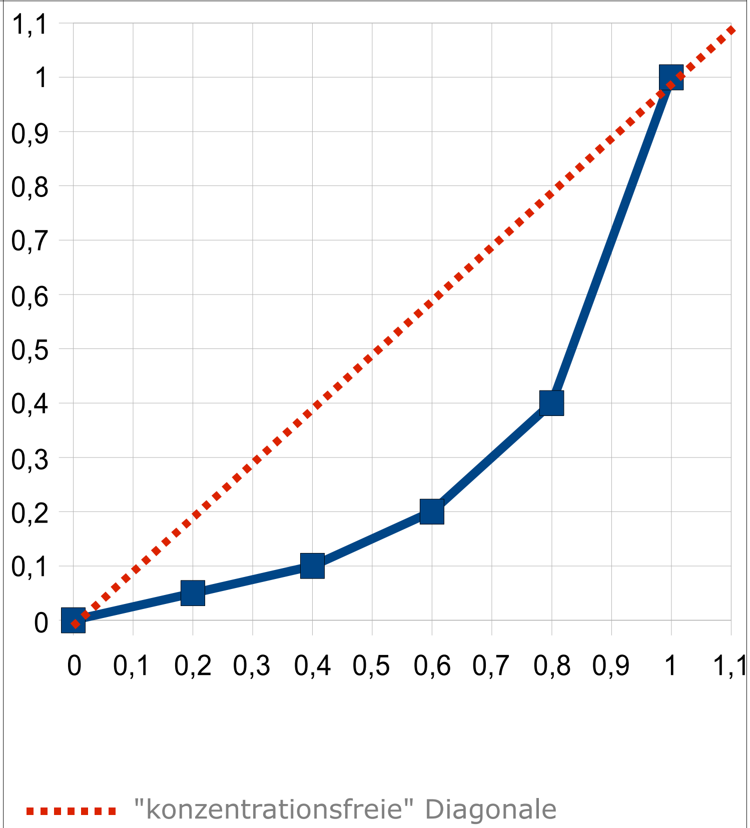
$v_i := \frac{p_i}{p_n}$ $\quad$ $p_i := x_{(1)} + x_{(2)} + \ldots + x_{(i)}$

Kurve: monoton steigend, konvex
=**hängt immer nach unten durch**

(je mehr, desto größer die Konzentration)

5 20 10 5 60

Wert x_i	Summe bis p_i	v_i	Merkmals-träger i	u_i
5	5	5/100= 0,05	1	1/5= 0,2
5	5+5= 10	10/100= 0,1	2	2/5= 0,4
10	10+10= 20	20/100= 0,2	3	0,6
20	20+20= 40	40/100= 0,4	4	0,8
60	40+60= 100	100/100= 1	5	1



13.

Gini-Koeffizient

G

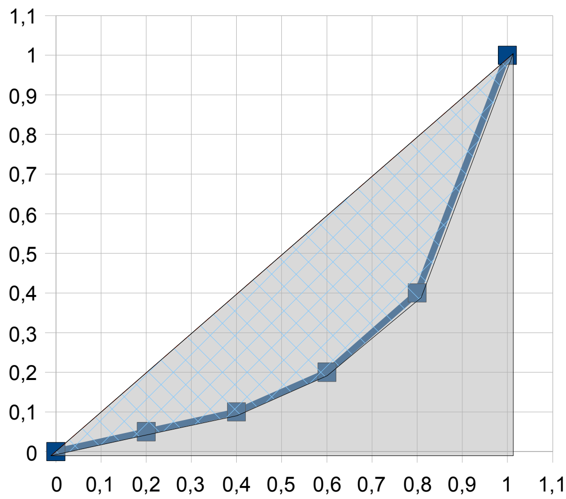
Konzentrationsmaß zur Lorenzkurve

-> *Relative* Konzentration

$$\frac{Fläche\ zwischen\ Diagonale\ und\ Lorenzkurve}{Fläche\ zwischen\ Diagonale\ und\ xAchse\,(u_i)}$$

$$\frac{2 \cdot gewichtete\ Merkmalssumme\ \ q_n}{Anzahl\ \ n\ der\ Werte\ \cdot\ Gesamtsumme\ der\ Werte\ p_n} - \frac{Anzahl+1}{Anzahl}$$

$$G := \frac{2 \cdot q_n}{n \cdot p_n} - \frac{n+1}{n}$$



mit

$$q_n := \sum_{i=1}^{n} i \cdot x_i = 1 \cdot x_{(1)} + 2 \cdot x_{(2)} + \ldots + i \cdot x_{(i)}$$

im geordneten Datensatz

$$G := \frac{2 \cdot \sum_{i=1}^{n} i \cdot x_i}{n \cdot \sum_{i=1}^{n=1} x_i} - \frac{n+1}{n}$$

5 20 10 5 60

i	Wert x_i	$i \cdot x_i$
1	5	1x5= 5
2	5	2x5= 10
3	10	3x10=30
4	20	4x20=80
5	60	5x60=300
n = 5	p_n = 100	q_n = 425

$$q_n := 1 \cdot 5 + 2 \cdot 5 + \ldots + 5 \cdot 60 = 425$$

$$G := \frac{2 \cdot 425}{5 \cdot 100} - \frac{5+1}{5} = 1{,}7 - 1{,}2 = 0{,}5$$

normierter Gini-Koeffizient

$$G^* = \frac{G}{G_{max}} = \frac{n}{n-1} \cdot G$$

mit

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

unabhängig von Länge n der Urliste

$$G^* = \frac{0{,}5}{\frac{5-1}{5}} = \frac{4}{5} \cdot 0{,}5 = 0{,}625$$

14.

Herfindahl-Index

H

Wieviel Merkmalsträger haben wieviel Prozent der Gesamtmerkmalssumme?

-> *Absolute* Konzentration

$$\left(\frac{einzelner\ Wert}{Gesamtsumme\ aller\ Werte}\right)^2 + \left(\frac{einzelner\ Wert}{Gesamtsumme\ aller\ Werte}\right)^2 + \ldots$$

=>

$$\frac{1}{Gesamtsumme\ aller\ Werte^2} \cdot einzelner\ Wert^2 + einzelner\ Wert^2 + \ldots$$

$$H := \sum_{i=1}^{n} \left(\frac{x_i}{p_n}\right)^2$$

=>

$$H := \frac{1}{p_n^2} \cdot \sum_{i=1}^{n} x_i^2$$

5   20   10   5   60

Wert x_i	$(x_i)^2$
5	25
5	25
10	100
20	400
60	3600
p_n = 100	
p_n^2 = 10 000	

$$H = \frac{1}{100^2} \cdot 5^2 + 5^2 + 10^2 + 20^2 + 60^2$$

$$= \frac{1}{10\,000} \cdot 25 + 25 + 100 + 400 + 3600 = \frac{4\,150}{10\,000} = 0{,}415$$

zu interpretieren im Vergleich zu

$$H_{min} := \frac{1}{n}$$

Je mehr Merkmalsträger mit gleichem Anteil an der Merkmalssumme beteiligt sind, desto näher liegt H an H_{min} .

$$H_{min} = \frac{1}{5} = 0{,}2$$

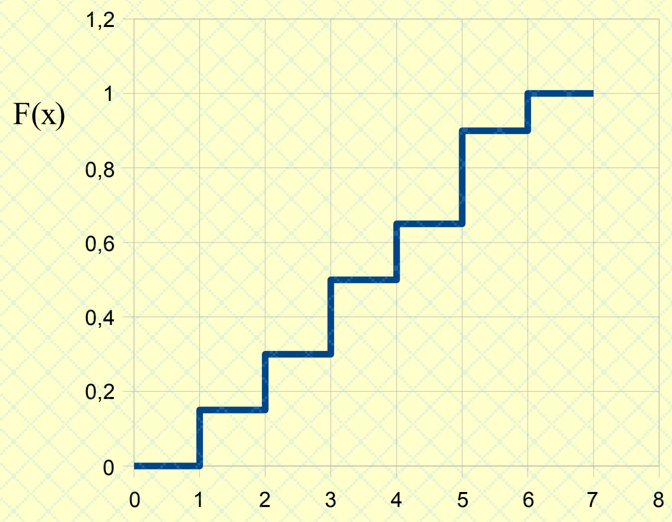
$$H = 0{,}415$$

BEISPIEL

Urliste:

⚁⚃⚂⚀⚅⚄⚁⚂⚄⚄⚀⚀⚃⚁⚂⚂⚄⚅⚄⚃

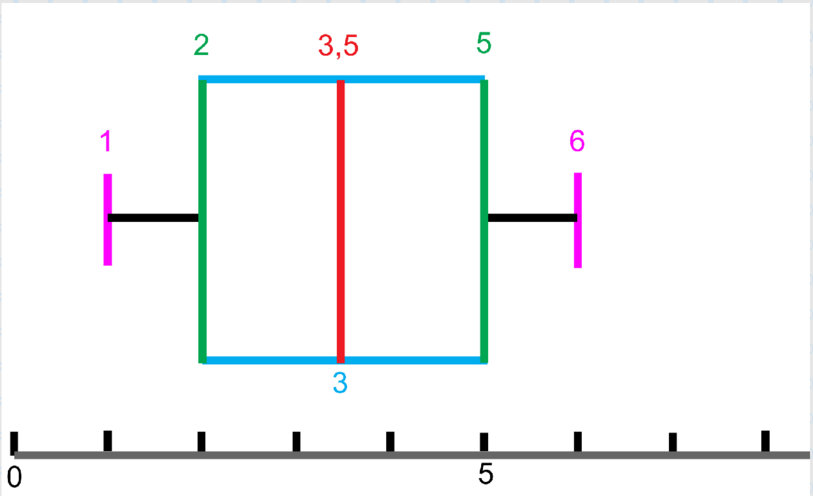
geordneter Datensatz	Wert	h	f	H	F
1	⚀	3	0,15	3	0,15
2	⚀				
3	⚀				
4	⚁	3	0,15	6	0,3
5	⚁				
6	⚁				
7	⚂	4	0,2	10	0,5
8	⚂				
9	⚂				
10	⚂				
11	⚃	3	0,15	13	0,65
12	⚃				
13	⚃				
14	⚄	5	0,25	18	0,9
15	⚄				
16	⚄				
17	⚄				
18	⚄				
19	⚅	2	0,1	20	1
20	⚅				



$x_{min}=1$ $R:=5$ $x_{max}=6$

$p_{0,25}=x_5=2$ $Q:=3$ $p_{0,75}=x_{15}=5$

$$p_{0,5}=\frac{x_{10}+x_{11}}{2}=3{,}5$$

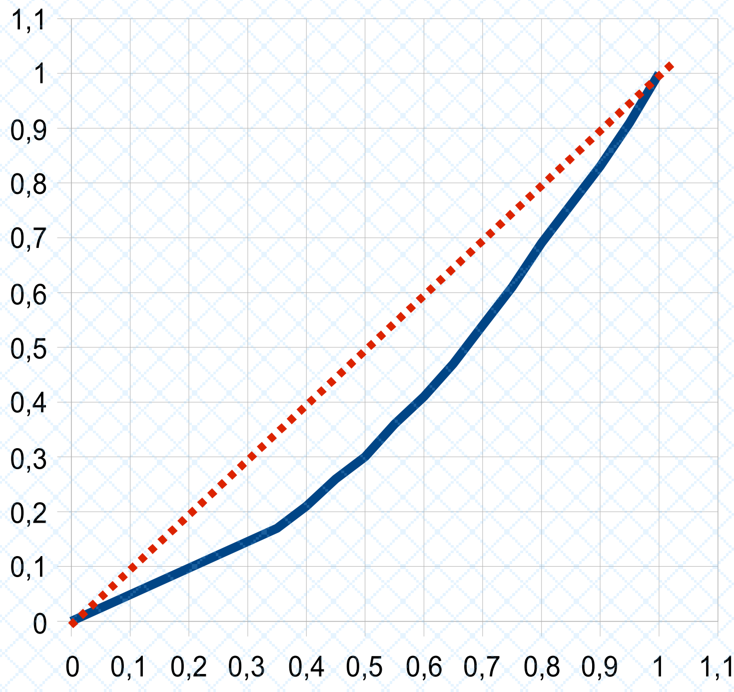


Lorenzkurve und Gini-Koeffizient

Wert x_i	Summe bis p_i	v_i	Merkmals-träger i	u_i	$i \cdot x_i$
1	1	1/70=0,01	1	1/20=0,05	1x1= 1
1	1+1=2	2/70=0,03	2	0,1	2x1= 1
1	2+1=3	3/70=0,04	3	0,15	3x1= 1
2	3+2=5	5/70=0,07	4	0,2	4x2= 8
2	5+2=7	7/70=0,1	5	0,25	5x2= 10
2	7+2=9	9/70=0,13	6	0,3	6x2= 12
3	12	0,17	7	0,35	21
3	15	0,21	8	0,4	24
3	18	0,26	9	0,45	27
3	21	0,3	10	0,5	30
4	25	0,36	11	0,55	44
4	29	0,41	12	0,6	48
4	33	0,47	13	0,65	52
5	38	0,54	14	0,7	70
5	43	0,61	15	0,75	75
5	48	0,69	16	0,8	80
5	53	0,76	17	0,85	85
5	58	0,83	18	0,9	90
6	64	0,91	19	0,95	114
6	$p_n=$ **70**	1	$n=$ **20**	1	120
					$q_n=$ **880**

$$G := \frac{2 \cdot 880}{20 \cdot 70} - \frac{20+1}{20} \approx 0{,}21$$

$$G^* = \frac{0{,}21}{\frac{20-1}{20}} = \frac{20}{19} \cdot 0{,}21 \approx 0{,}22$$



Herfindahl

$$H = \frac{1}{70^2} \cdot 1+1+1+4+4+4+9+9+9+9+16+16+16+25+25+25+25+25+36+36 = \frac{296}{4900} \approx 0{,}06$$

$$H_{min} = \frac{1}{70} \approx 0{,}01$$