1.5 Ungleichung von Chebyshev

Satz 1.2: (Ungleichung von Chebyshev)

Sei $\{x_1, x_2, \dots, x_n\}$ eine Stichprobe mit Mittelwert $\bar{x}$ und Stichprobenstandardabweichung $s > 0$.
Weiterhin sei

$$S_k = \{ i,\ 1 \le i \le n :\ |x_i - \bar{x}| < k \cdot s \}$$

und $N(S_k)$ die Anzahl der Elemente in S_k, d.h. die Anzahl der x_i, die in $]\bar{x} - k \cdot s ; \bar{x} + k \cdot s[$ liegen.
Dann gilt für alle $k \ge 1$:

$$\boxed{\frac{N(S_k)}{n} > 1 - \frac{1}{k^2}}$$

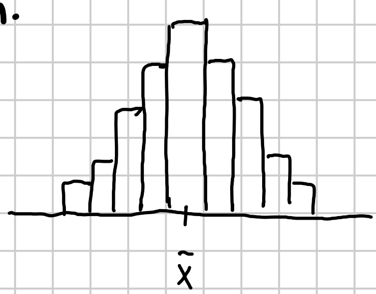
Bemerkungen:

(1) Die Ungleichung von Chebyshev macht eine allgemeine (und damit nicht sehr genaue Aussage) über die Lage der Daten um den Mittelwert:

$k = 2$: Mehr als 75% der Daten liegen in dem $2 \cdot s$-Bereich um $\bar{x}$.
$]\bar{x} - 2s ; \bar{x} + 2s[$

$k = 3$: Mehr als 89% der Daten liegen in dem $3s$-Bereich um $\bar{x}$.

(2) Sind die Daten näherungsweise normalverteilt, d.h.



dann lässt sich die Aussage präzisieren
- ca. 68% liegen im s-Bereich um $\bar{x}$
- ca. 95% —"— $2s$-Bereich —"—
- ca. 99% —"— $3s$-Bereich —"—

1.6 Korrelation (multivariate Daten)

Neben der statistischen Analyse einzelner Merkmale multivariater Daten, interessiert man sich auch für Beziehungen zwischen den Merkmalen, also zwischen Wertepaaren (x_i, y_i) $(1 \leq i \leq n)$ eines Datenframes.

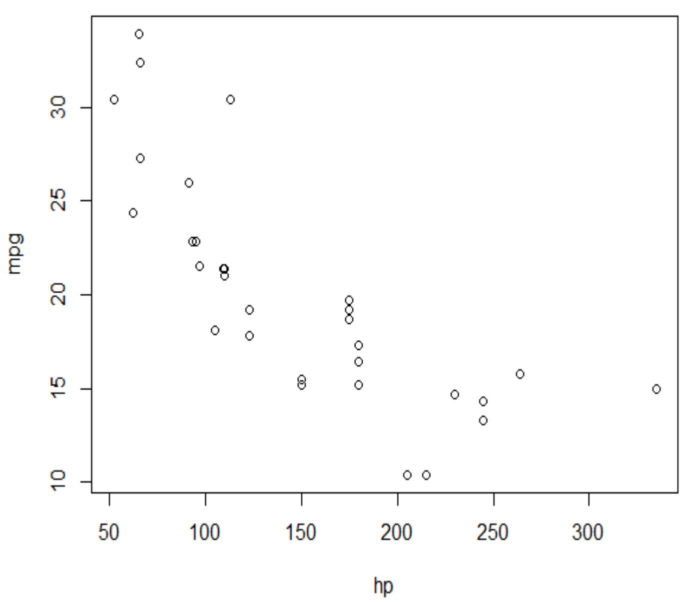
Grafische Veranschaulichung durch ein Streudiagramm im xy-Koord.system.
R-Funktion: plot (x,y)

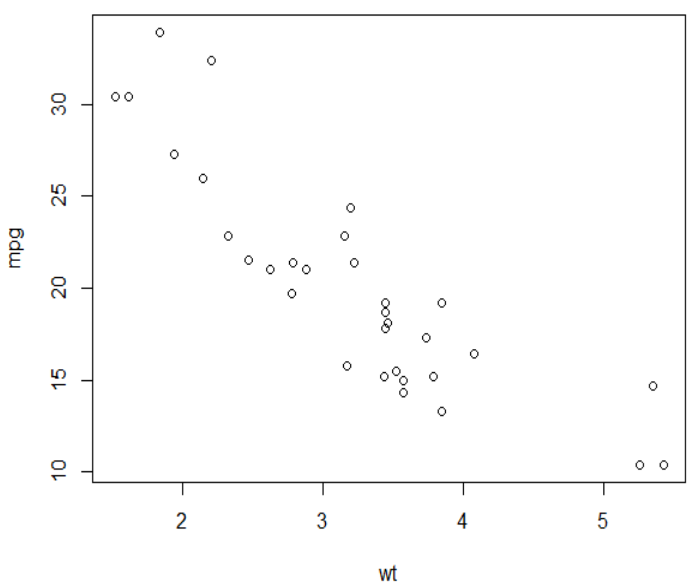
Die Kennzahl, die angibt, ob ein näherungsweiser linearer Zusammenhang zwischen den Merkmalen x und y besteht heißt Korrelationskoeffizient.

__Beispiel:__ Streudiagramme und Korrelationskoeff r

(mtcars$hp, mtcars$mpg)
$r \approx -0.776$

(mtcars$wt, mtcars$mpg)
$r \approx -0.868$





Definition 1.4:

Seien s_x und s_y die Stichprobenstandardabweichungen der x- bzw. y-Werte. Dann ist der Stichprobenkorrelationskoeffizient r der Datenpaare (x_i, y_i), $i=1,\ldots,n$ wie folgt definiert:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x \cdot s_y}$$

Die Datenpaare (x_i, y_i) heißen positiv korreliert, wenn $r > 0$,
bzw. negativ korreliert, wenn $r < 0$.

Bemerkungen:

(1) R-Funktion: cor(x,y)

(2) $s_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ heißt empirische Kovarianz,

d.h. $r = \frac{s_{xy}}{s_x \cdot s_y}$ (ist dimensionslos)

(3) Berechnung mit Hilfe des Verschiebungssatzes

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}\right)\left(\sum y_i^2 - n\bar{y}\right)}}$$

Satz 1.3:

Der Stichprobenkorrelationskoeffizient r hat folgende Eigenschaften:

(1) $-1 \leq r \leq 1$

(2) Falls zwischen x_i und y_i ein linearer Zusammenhang

$y_i = a x_i + b$ mit positiver Steigung a

besteht, dann gilt: $r = 1$.

(3) Falls zwischen x_i und y_i ein linearer Zusammenhang

$y_i = a x_i + b$ mit negativer Steigung a

besteht, dann gilt: $r = -1$.

Bemerkung:

(1) Negative Korrelation, d.h. $r < 0$, bedeutet:
Positive $r > 0$

Je kleiner der Wert des Merkmals x, desto größer der Wert des anderen Merkmals y und umgekehrt.
kleiner

(2) Der Korrelationskoeff. liefert nur eine Aussage, ob näherungsweise ein linearer Zusammenhang besteht.
Andere Zusammenhänge lassen sich nur am Streudiagramm erkennen.

Regressionsgerade im Fall eines näherungsweisen linearen Zusammenhangs:

$$y = a x + b$$

Es gilt:

Steigung

$$a = r \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

y-Abschnitt

$$b = \bar{y} - a \bar{x}$$