
Datenanalyse metrischer Daten

Dr. Dirk Hasenclever

IMISE, Leipzig

dirk.hasenclever@imise.uni-leipzig.de

Datensatz vorbereiten

- Unterschiede Rohdaten und abgeleitete Daten
 - Z.B. Rohvariable in Großbuchstaben
- Jede Beobachtung (Zeile) sollte einen eindeutigen Identifier haben.
- Erste Beschreibung der Variablen
 - Skala, Datentyp: Metrisch (numeric, integer)
 - Anteil Fehlwerte: Achtung wenn hoch.

Generelle Empfehlungen

- Immer beides: Visualisierung **und** zahlenmäßige Darstellung
- Suche aussagefähige Parameter auf verständlichen Skalen
- Kein Punktschätzer ohne Konfidenzintervall
- Tests dual zu Konfidenzintervallen

Vier Methoden der Visualisierung metrischer Variabler

- Empirische kumulative Verteilungsfunktion
 - Als primäre Visualisierung empfohlen!
- Histogramm mit Dichteschätzer
- (Boxplots)
- (Stripcharts)

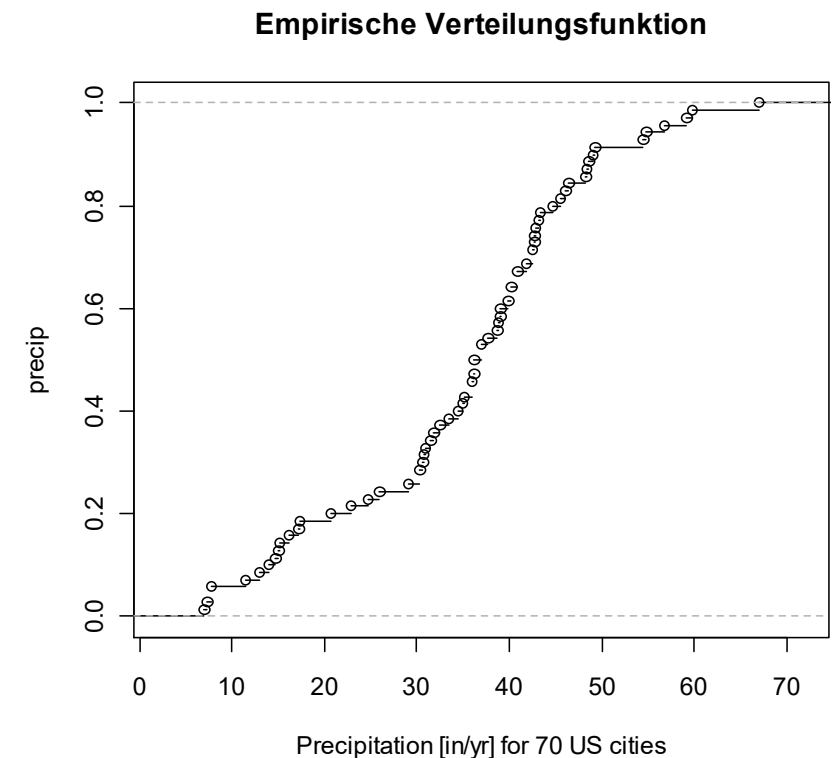
Schätzer für Verteilungsfunktionen

- Histogramm bzw. Dichteschätzer approximieren die theoretische Dichte $f_X(x)$
- Die empirische Verteilungsfunktion

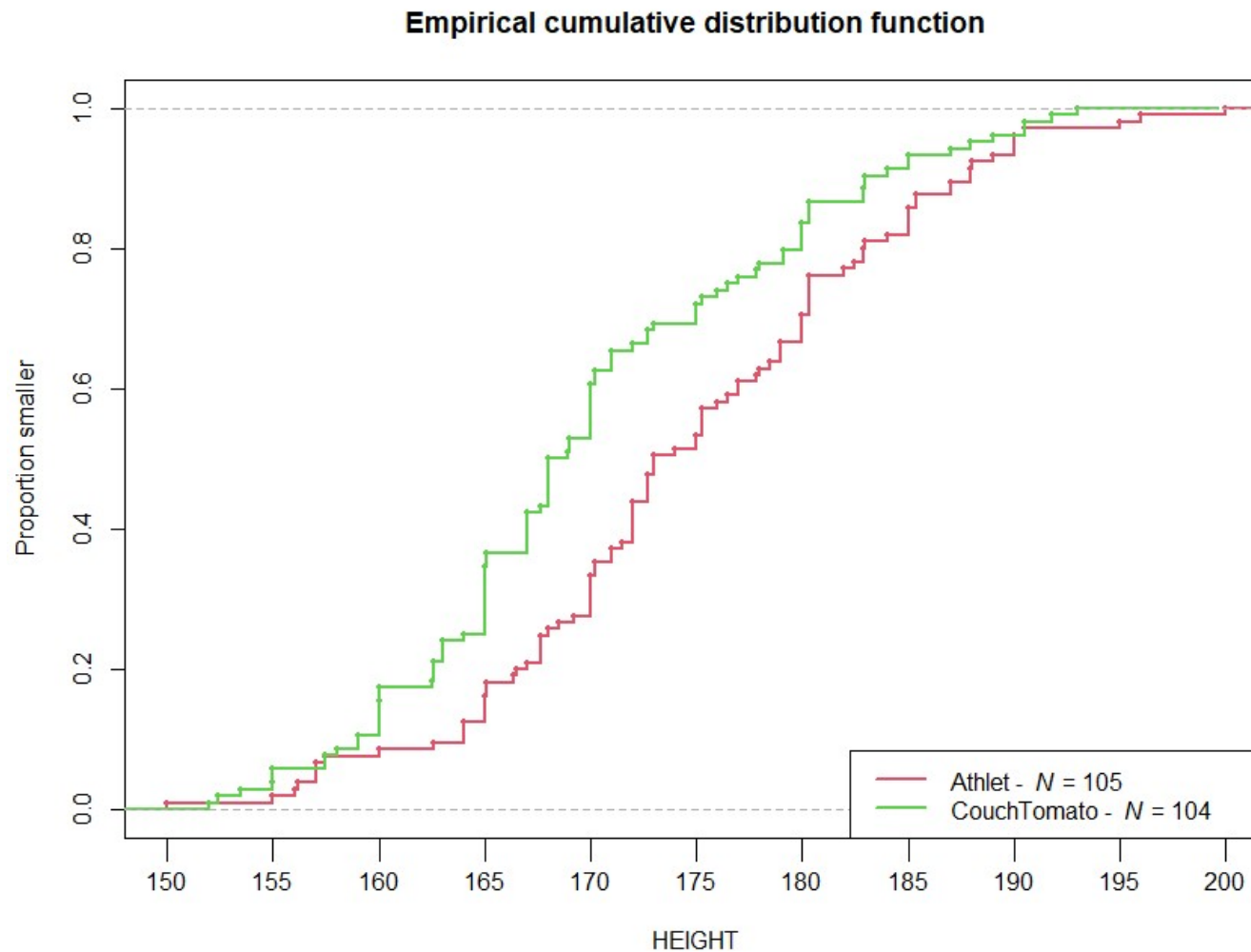
$$\hat{F}_X(x) = \frac{\#\{x_i \leq x\}}{N}$$

Approximiert die theoretische Verteilungsfunktion

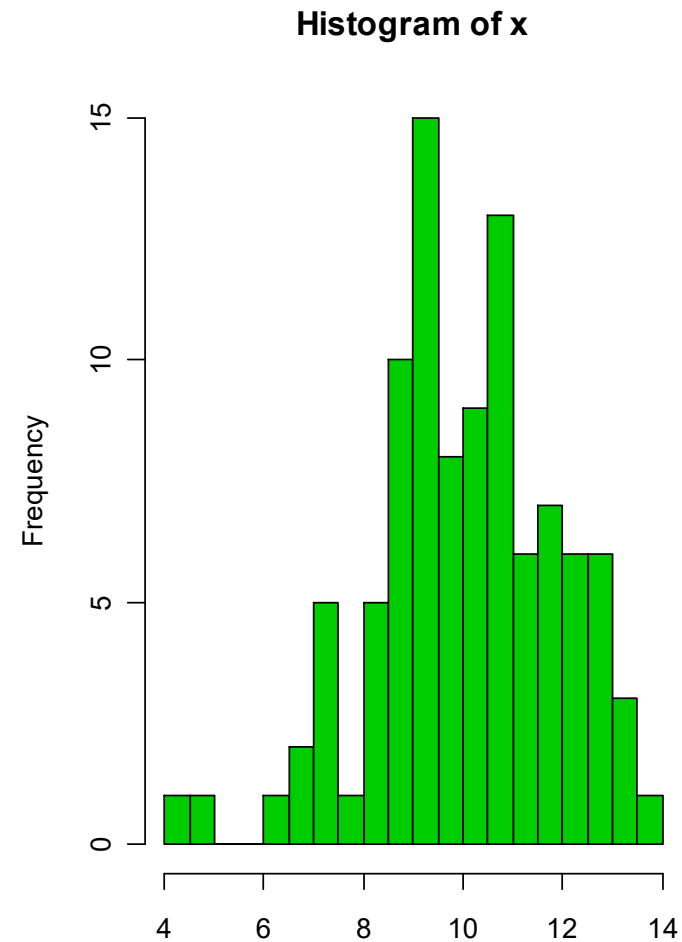
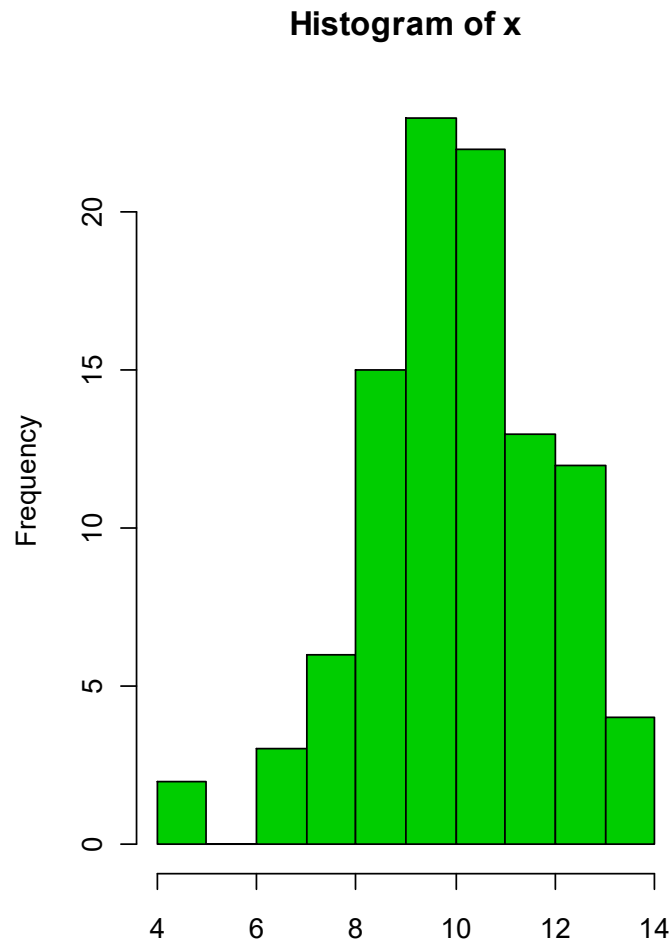
$$F_X(x) = pr(X \leq x)$$



Beispiel mit OurTools::PlotMetricBySplit

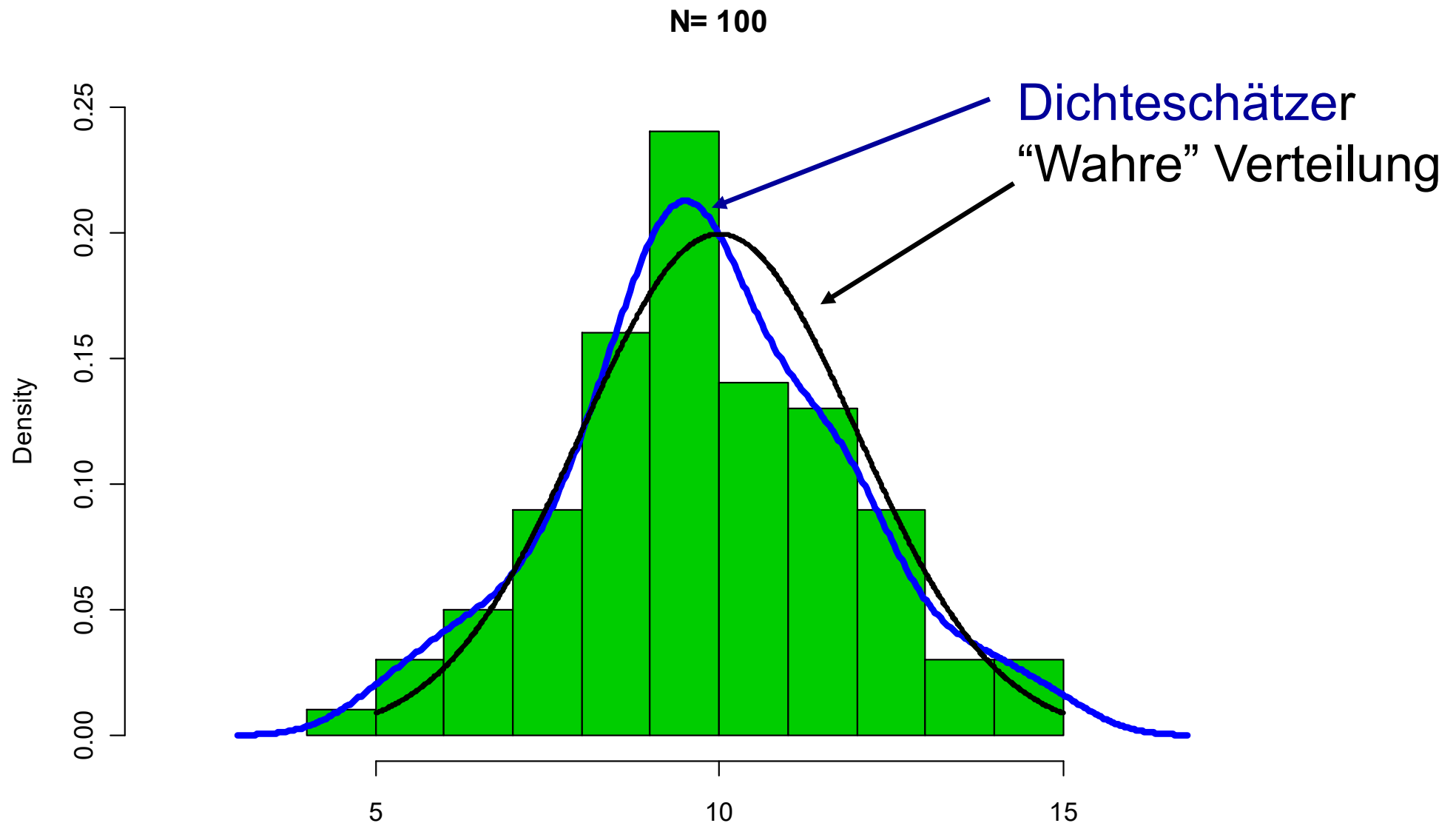


Visualisierung einer stetigen Verteilung: Histogramm

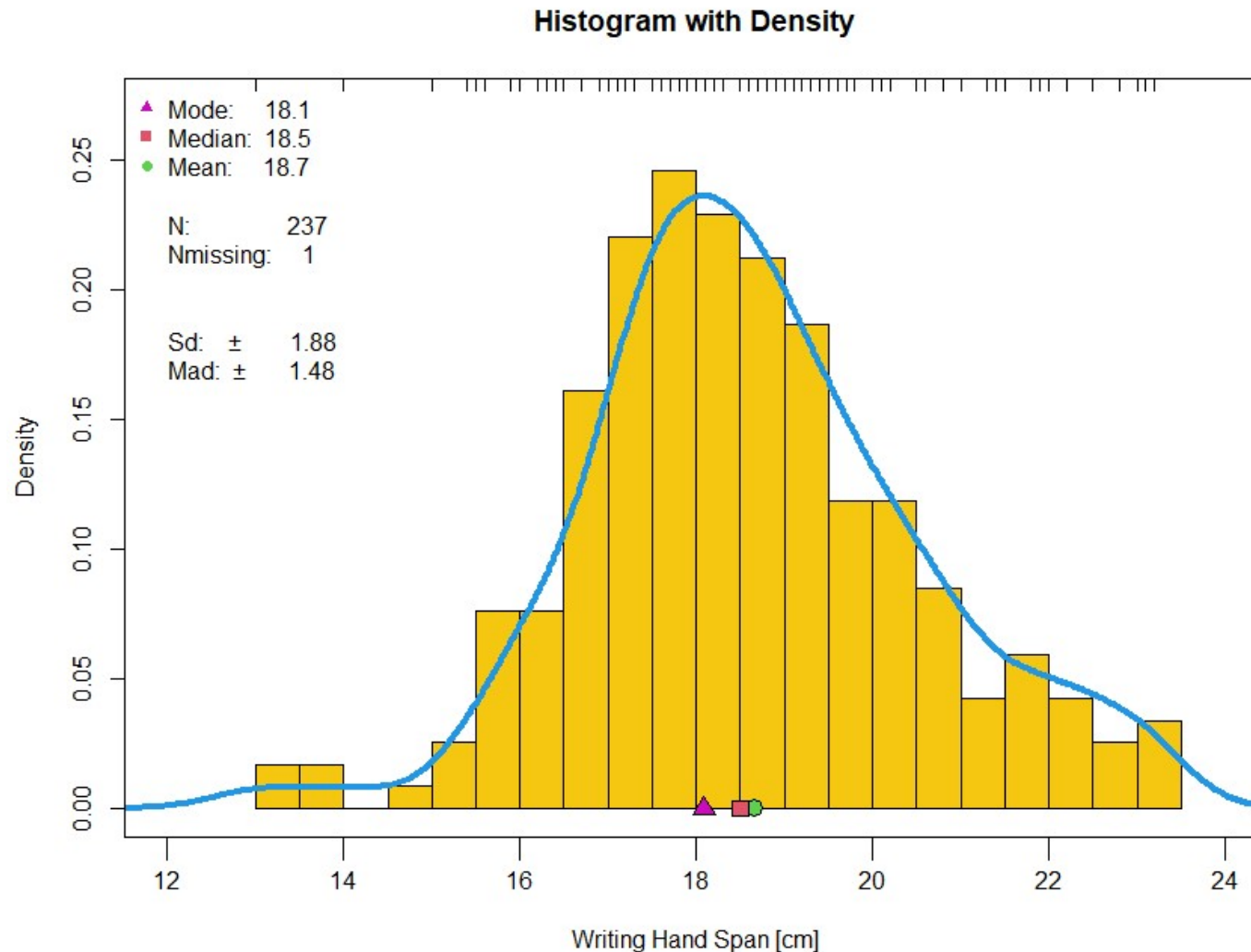


^x Abhängig von Klassenbreite ^x

Visualisierung einer stetigen Verteilung: Dichteschätzer

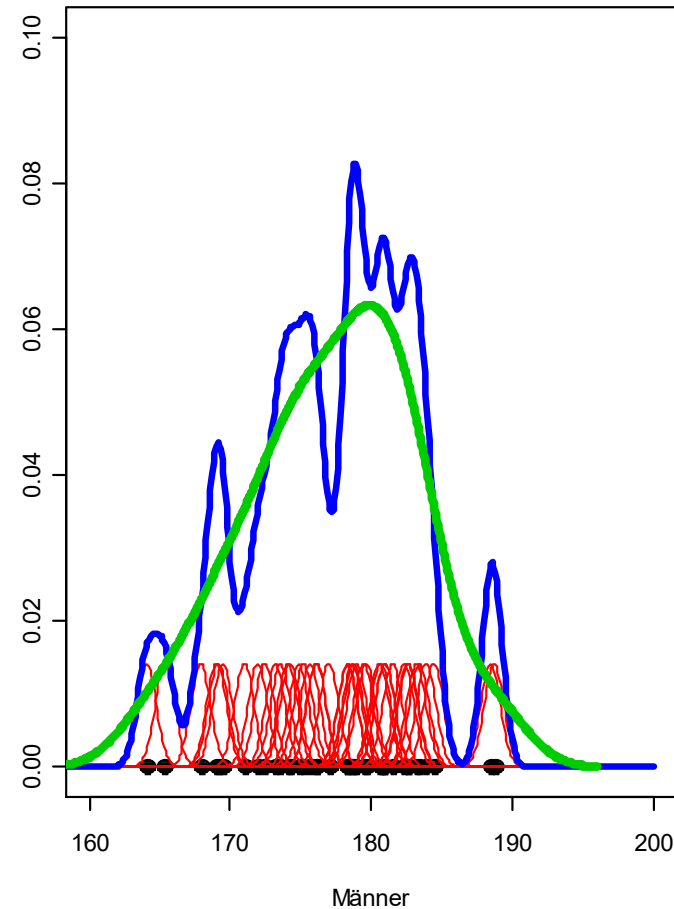


Beispiel: OurTools:: NiceHist

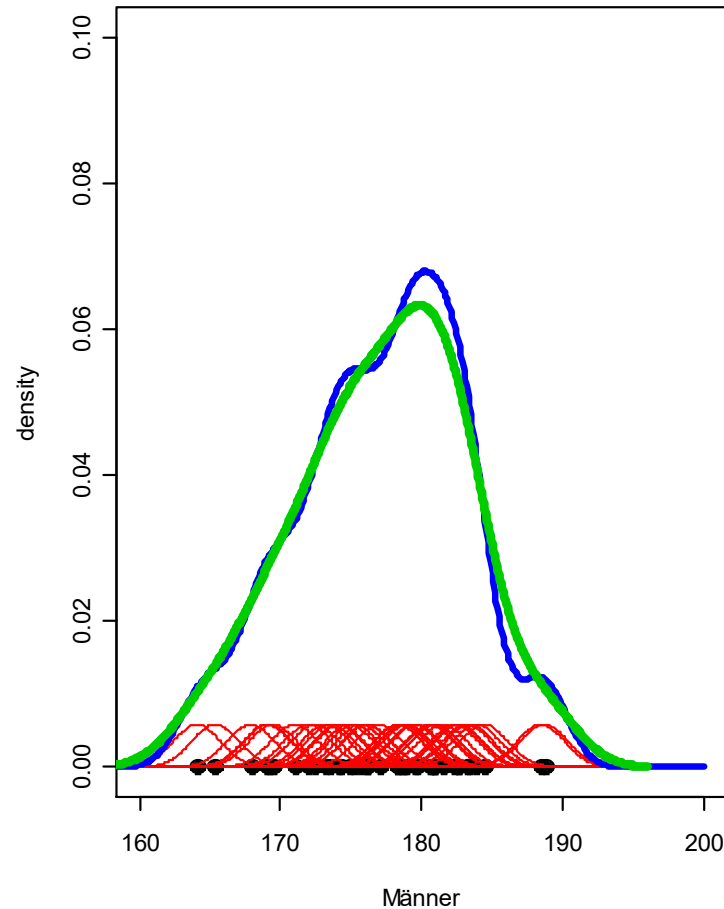


Wie funktioniert ein Dichteschätzer?

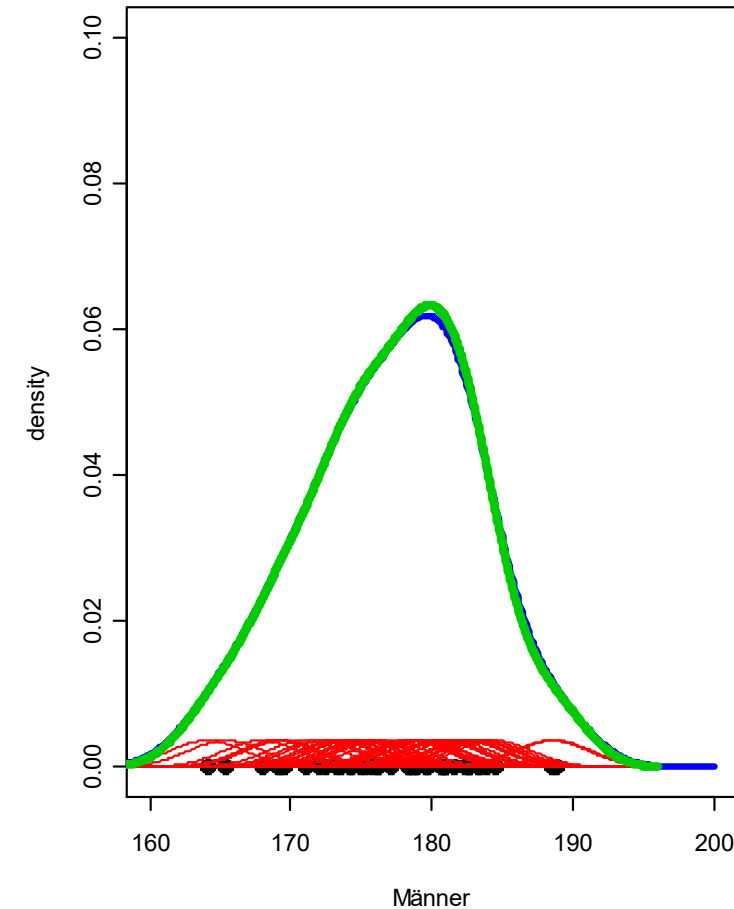
Prinzip Dichteschätzer



Prinzip Dichteschätzer

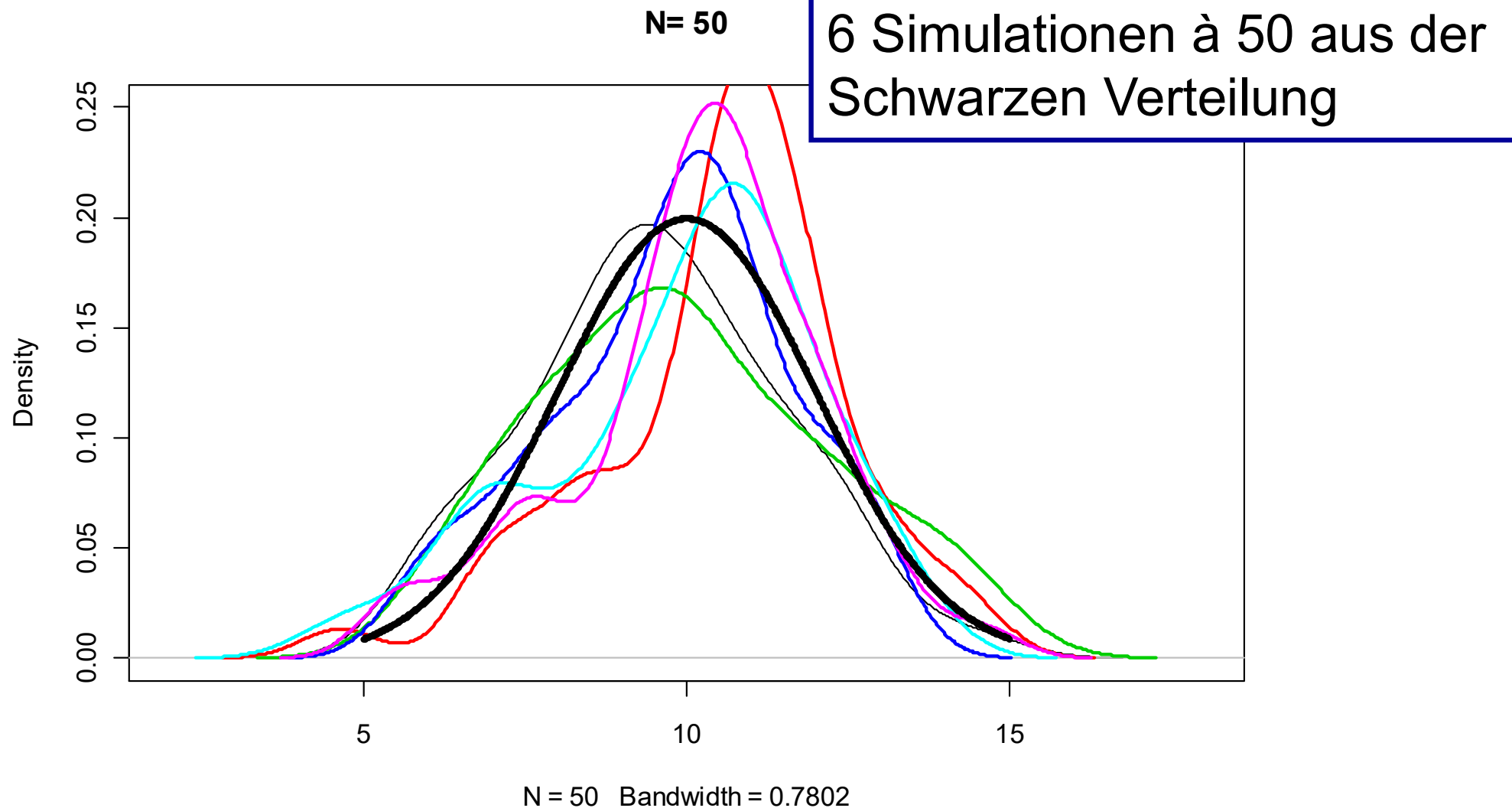


Prinzip Dichteschätzer



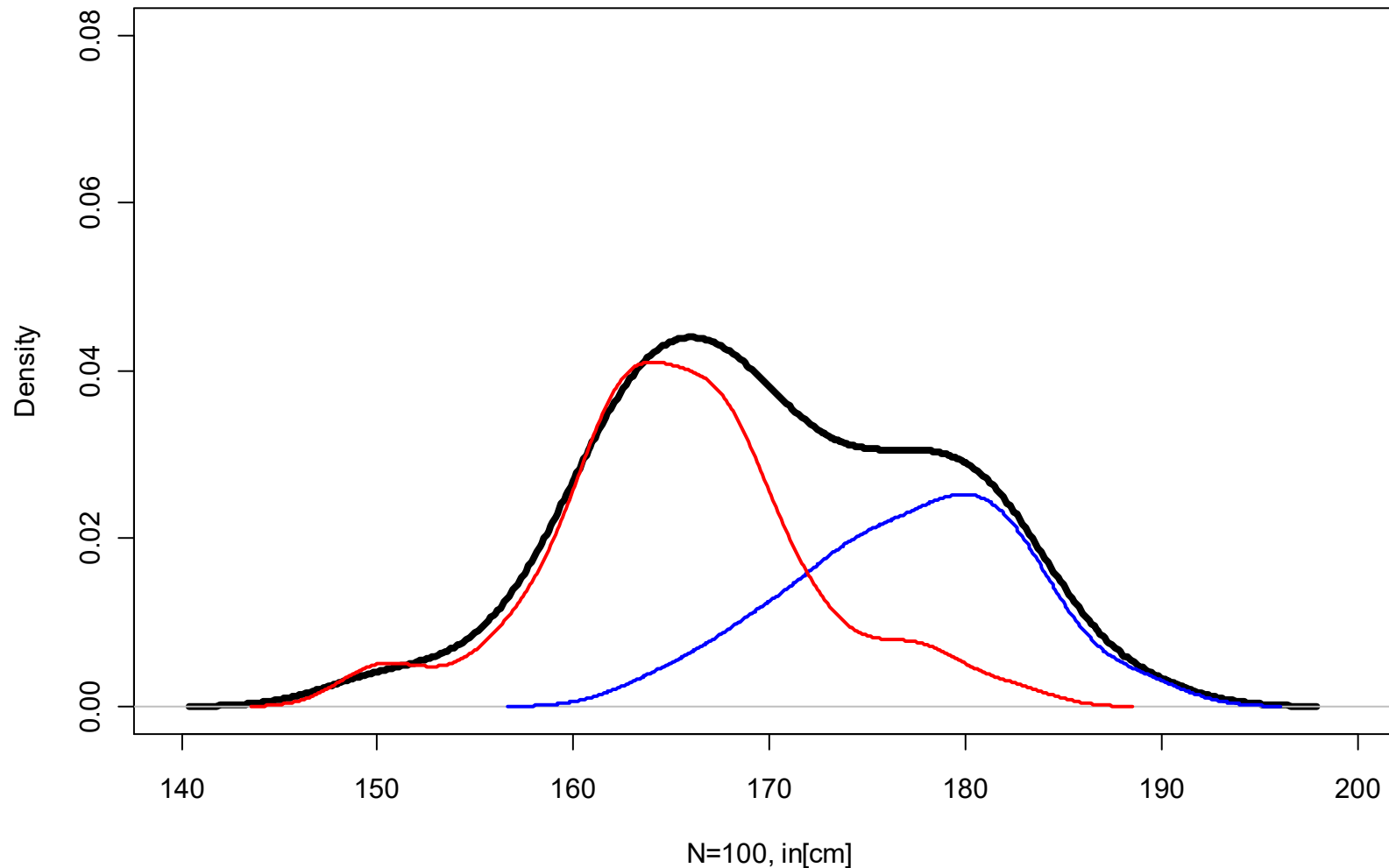
Für verschiedene Breiten der überlagerten “normalverteilte Kerne”

Robuster Blick erforderlich



Verteilungsdichten können Heterogenität verbergen

Körpergrößen-Mischung von Männern und Frauen



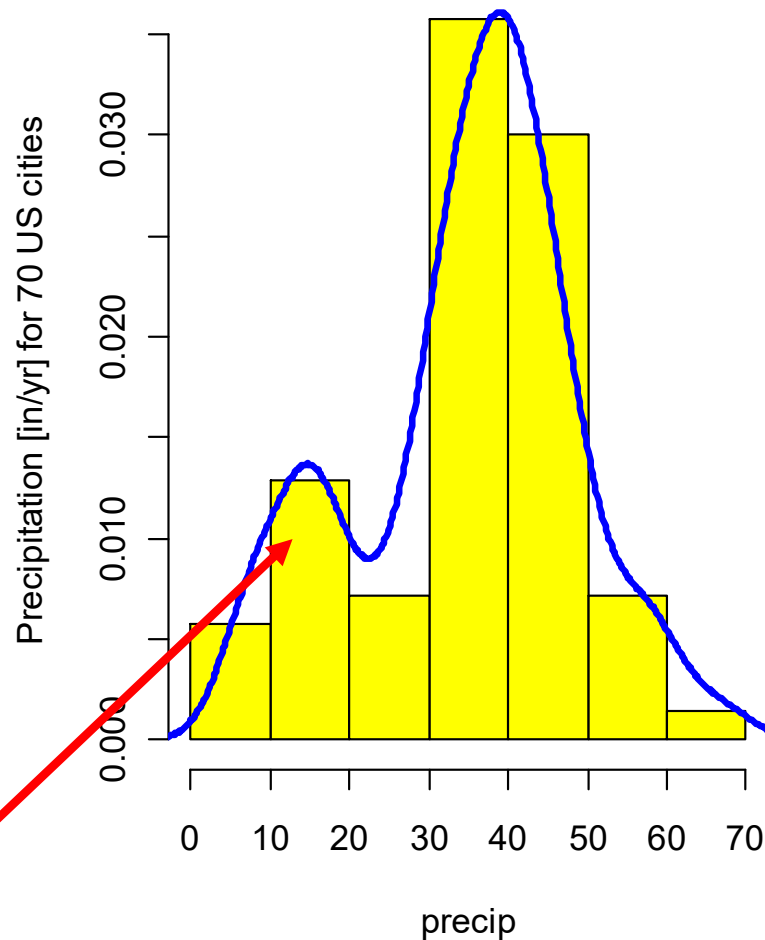
precip Daten zu jährlichen Niederschlagsmengen in US Städten

Phoenix	Reno	Albuquerque	El Paso	Boise	Denver
7.0	7.2	7.8	7.8	11.5	13.0
Los Angeles	Cheyenne	Great Falls	Salt Lake City	Bismark	Sacramento
14.0	14.6	15.0	15.2	16.2	17.2
Spokane	San Francisco	Honolulu	Sioux Falls	Minneapolis/St Paul	Milwaukee
17.4	20.7	22.9	24.7	25.9	29.1
Duluth	Omaha	Wichita	Des Moines	Detroit	Oklahoma City
30.2	30.2	30.6	30.8	31.0	31.4
Sault Ste. Marie	Burlington	Albany	Chicago	Cleveland	Peoria
31.7	32.5	33.4	34.4	35.0	35.1
St Louis	Dallas	Buffalo	Concord	Pittsburg	Kansas City
35.9	35.9	36.1	36.2	36.2	37.0
Columbus	Portland	Indianapolis	Seattle Tacoma	Washington	Cincinnati
37.0	37.6	38.7	38.8	38.9	39.0
Philadelphia	Wilmington	New York	Portland	Charleston	Baltimore
39.9	40.2	40.2	40.8	40.8	41.8
Boston	Raleigh	Richmond	Charlotte	Providence	Louisville
42.5	42.5	42.6	42.7	42.8	43.1
Hartford	Norfolk	Atlantic City	Nashville	Columbia	Houston
43.4	44.7	45.5	46.0	46.4	48.2
Atlanta	Little Rock	Memphis	Jackson	Jacksonville	Juneau
48.3	48.5	49.1	49.2	54.5	54.7
New Orleans	San Juan	Miami	Mobile		
56.8	59.2	59.8	67.0		

Sortiert, N=70

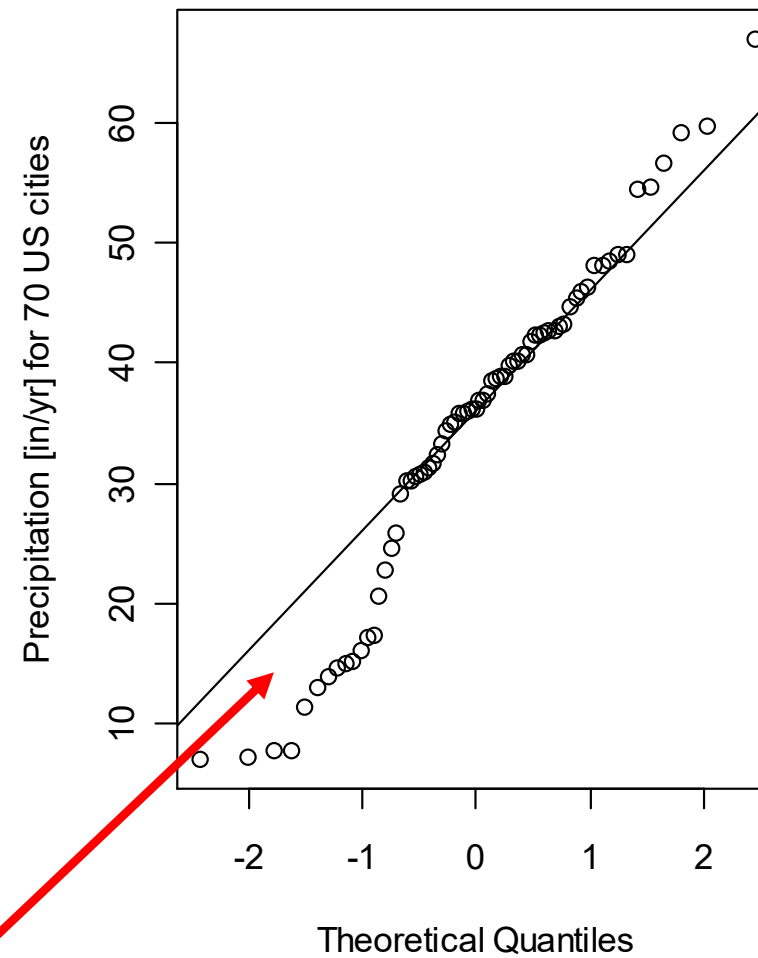
Q-Q Plot zur graphischen Überprüfung, ob grob normalverteilt

Histogram of precip



Wüste

Normal Q-Q Plot



Q-Q Plot Interpretation

- Plote empirische Verteilungsfunktion der Daten gegen entsprechende Quantile der Verteilung, mit der verglichen werden soll.
- Daten sollten bei Übereinstimmung auf einer Geraden liegen.
- Im Beispiel: Hinweis auf Abweichung von Normalverteilung (Symmetrie).
Inhaltlich ist bimodale Dichte plausibel.

Tests auf Normalverteilung

```
>shapiro.test(precip)
```

Shapiro-Wilk normality test

data: precip

W = 0.9646, p-value = 0.04493

Kolmogorov-Smirnov Test auf beliebige Verteilungsformen

Eigentlich oft unnötig:

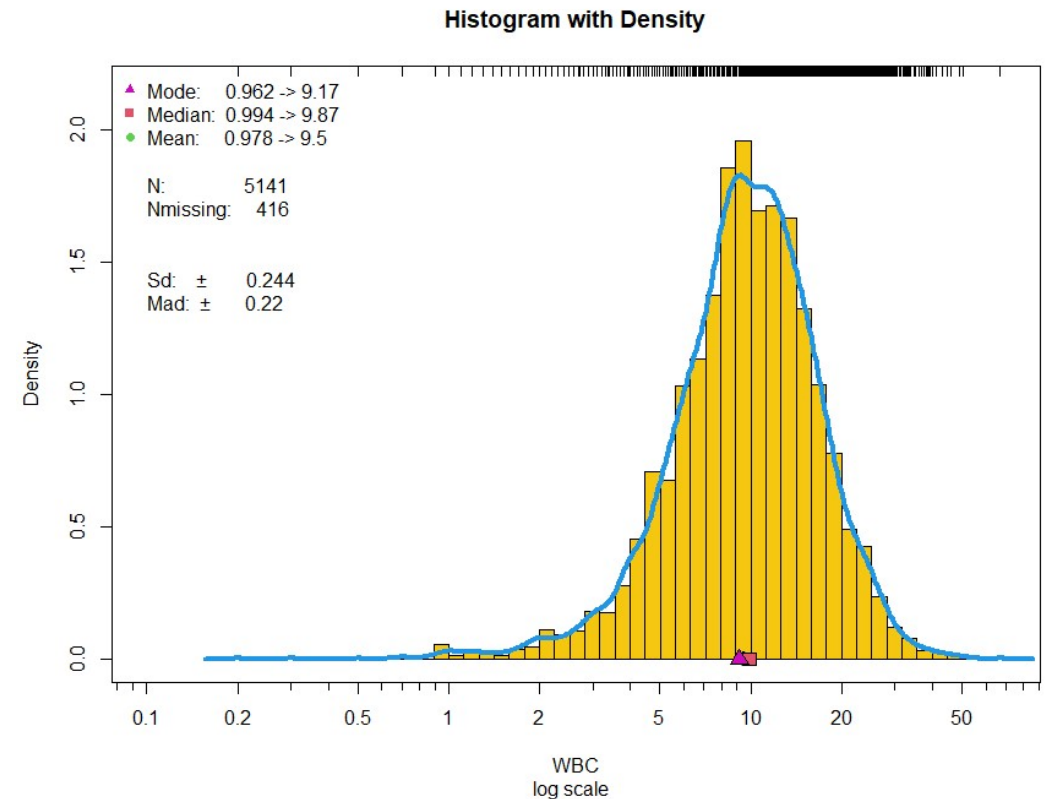
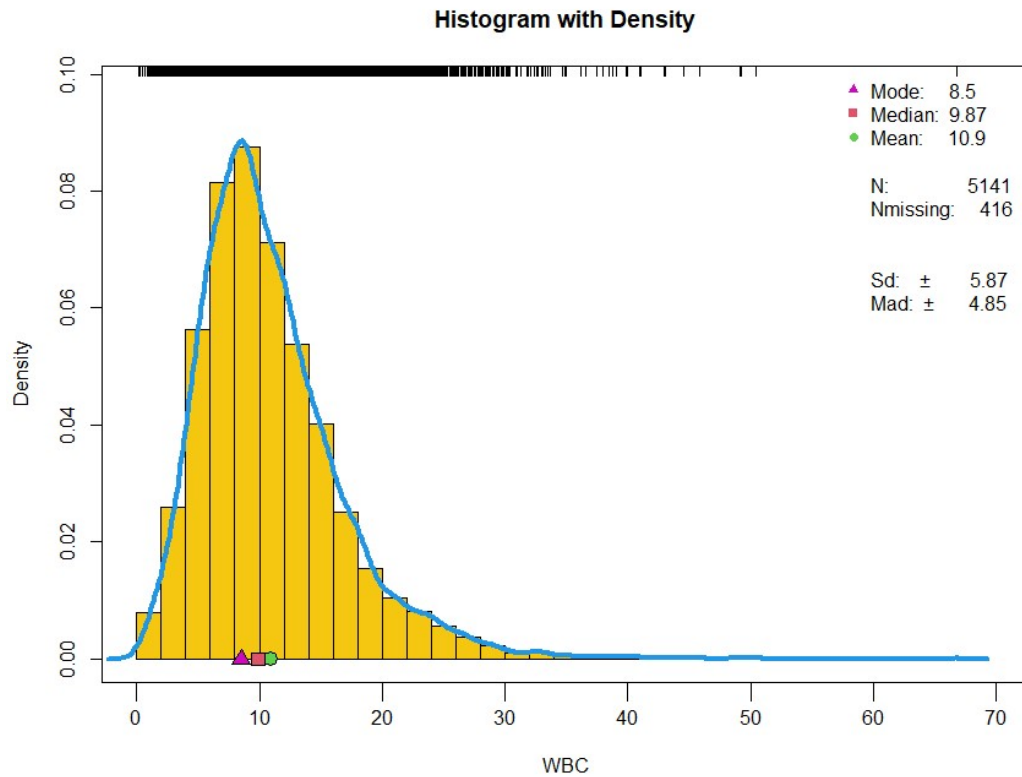
Grafische Diagnose reicht in der Regel

Zentraler Grenzwertsatz -> gewisse Abweichungen tolerabel

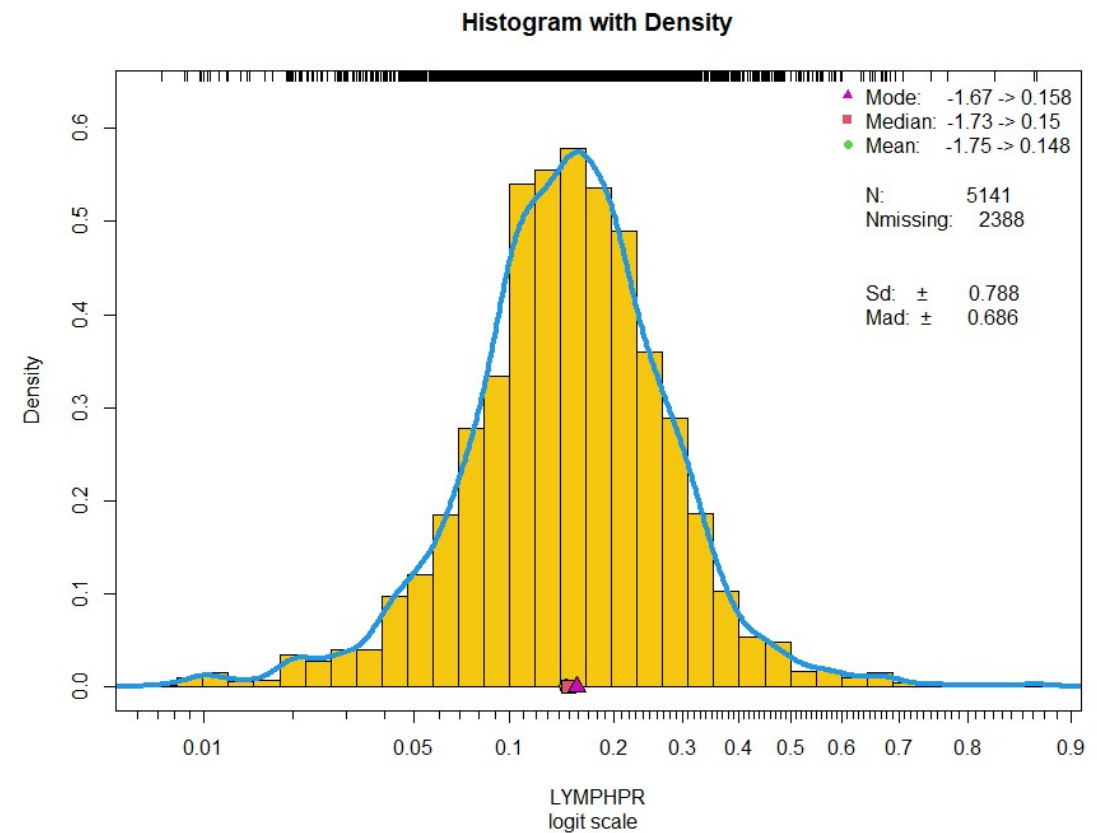
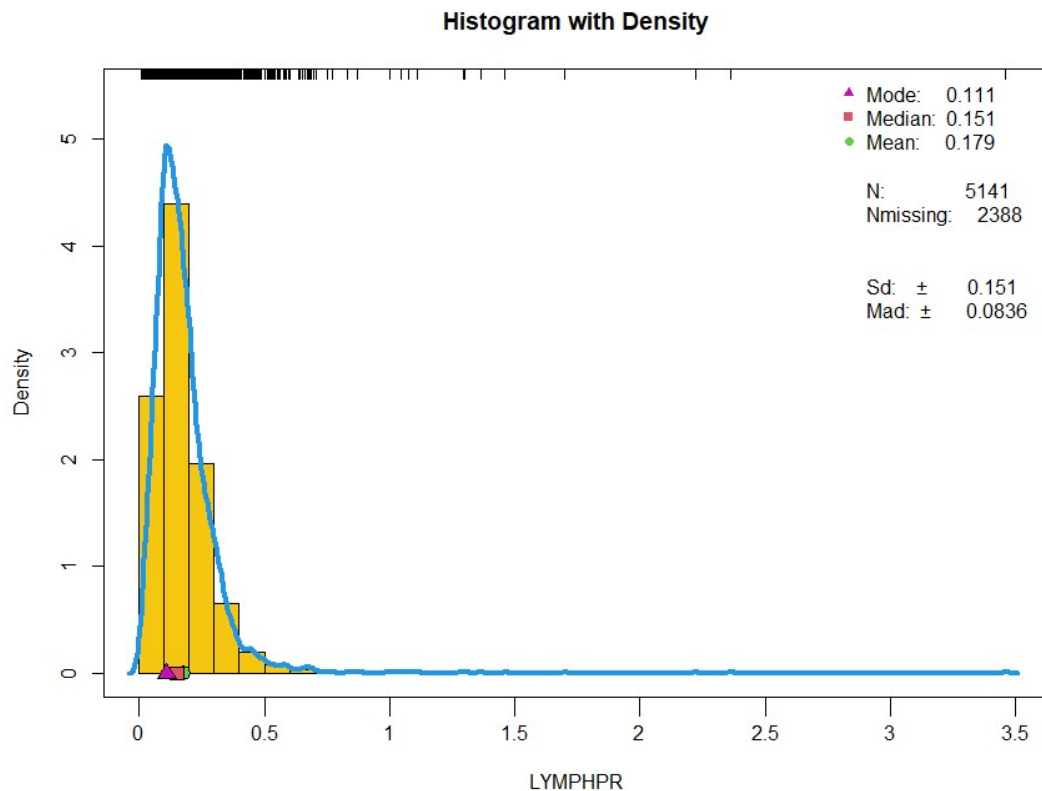
Verwende die „richtige“ Skala

- Alles immer auf der „richtigen“ Skala
 - „Richtig“ = so dass Histogramm grob symmetrisch..
 - Biometrisches Erfahrungswissen
 - LOG10 – z.B. bei Zellzahlen, die stark ansteigen können
 - LOGIT bei Proportions/Prozentzahlen in $[0;1]$
 - ...

Log Skala

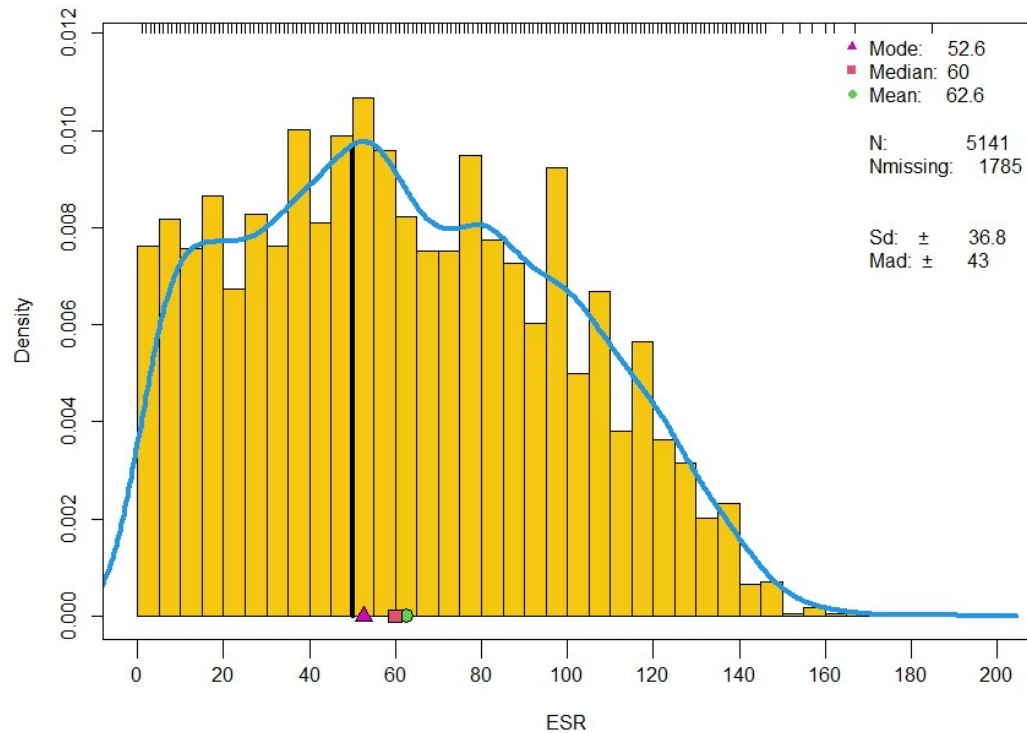


Logit-Skala

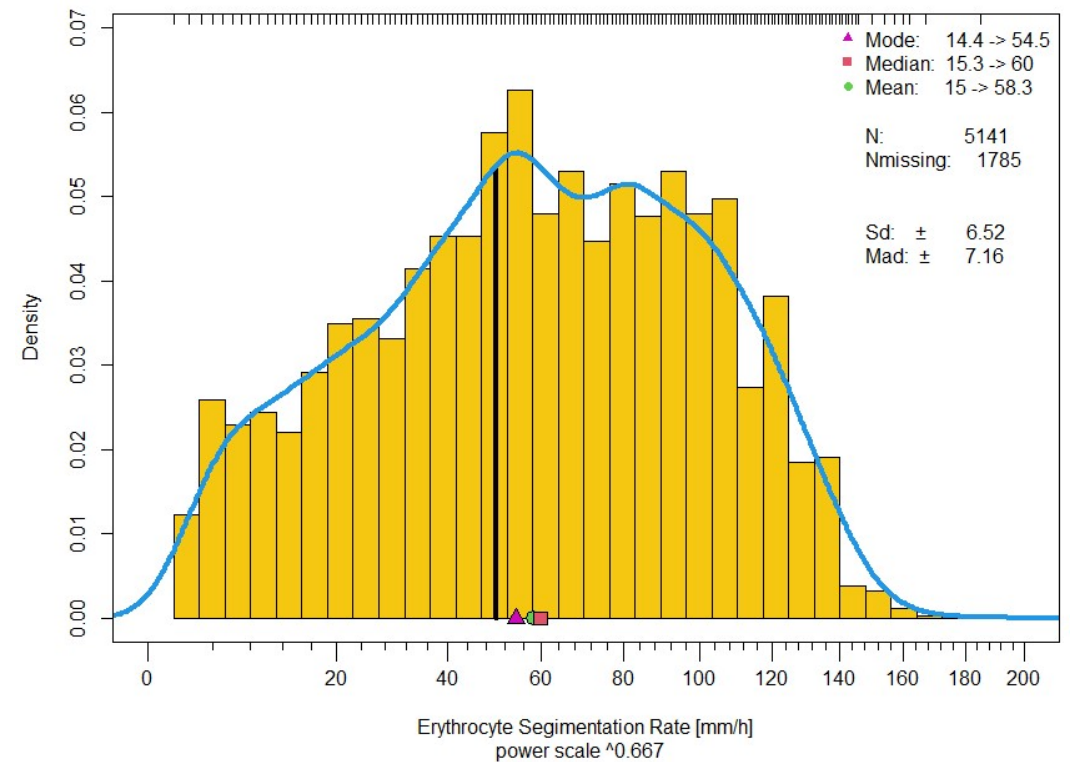


Power Skala (seltener)

Histogram with Density



Histogram with Density

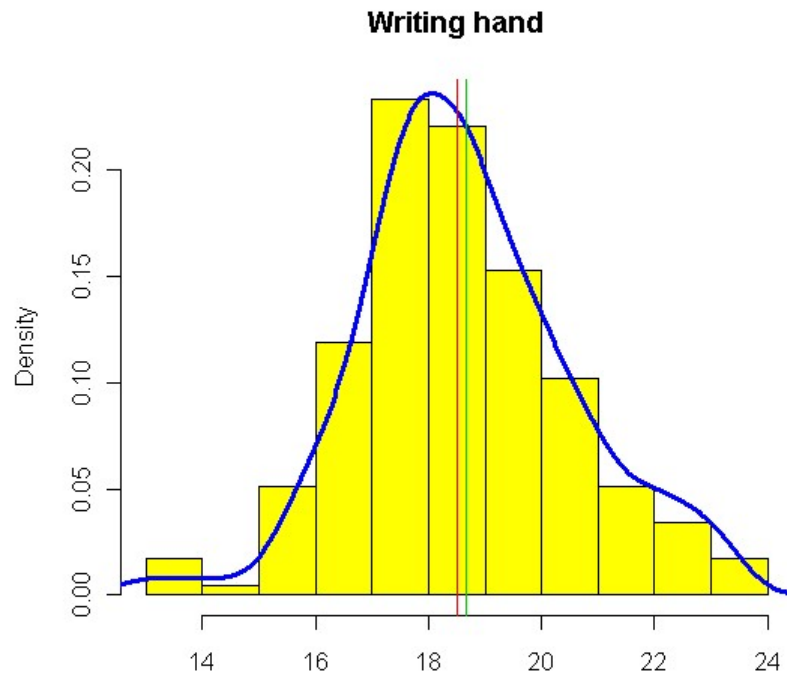


Statt ganze Verteilungen zu schätzen, schätze Kenngrößen für Lokalisation und Skala

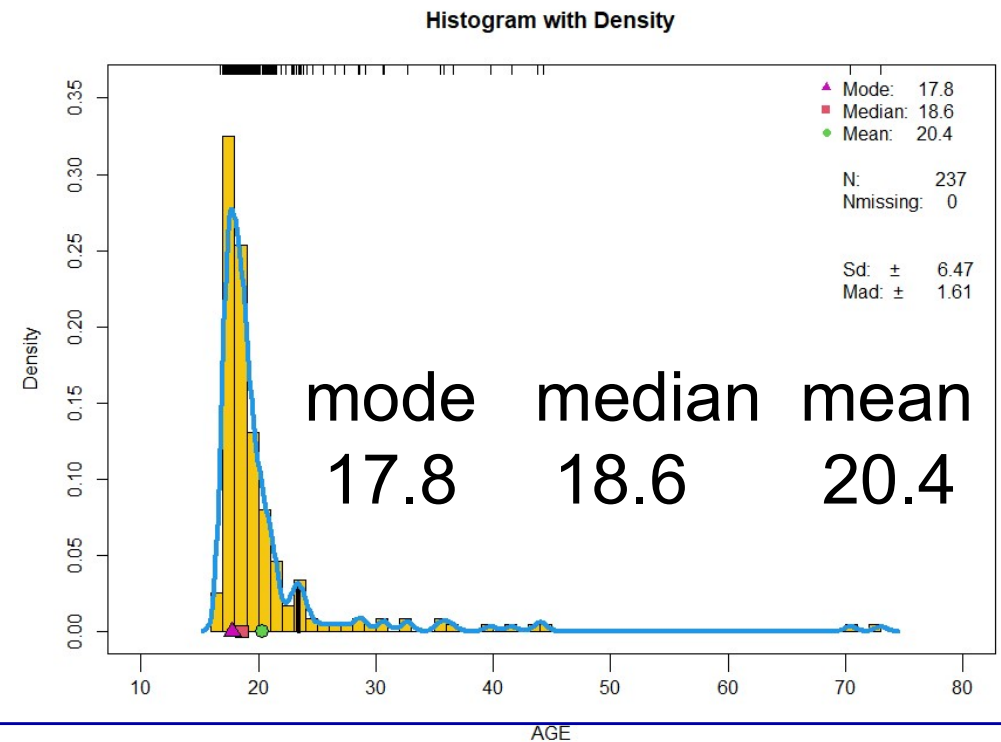
- Um eine Dichte vernünftig zu schätzen,
braucht man meist viele Fälle.
(Daumenregel ~ 100-300 je nach Verteilungstyp)
- Leichter zu schätzen sind
Lokalisation- und Skalenparameter

Lokalisation: Mittelwert , Median und Mode

~ symmetrisch



linkssteil - rechtsschief



Bei rechts-schiefen (= links-steilen) Verteilungen:

Mode < Median < Mean

bei symmetrischen Verteilungen fallen diese drei Lokalisationsparameter zusammen!

Skalenparameter

Varianzschätzer

$$\hat{v}ar = \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

Schätzer Standardabweichung

$$sd = \hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Schätzer *median absolute deviation*

$$\hat{m}ad = 1.4826 \cdot \text{median}(|x_i - \text{median}(x_k)|)$$

(Konstante so, dass $\hat{m}ad \sim \hat{\sigma}$ für Normalverteilungen)

mad ist robuster als sd

Writing Hand

sd	mad
----	-----

1.88	1.48	ungefähr gleich (keine Ausreißer)
------	------	-----------------------------------

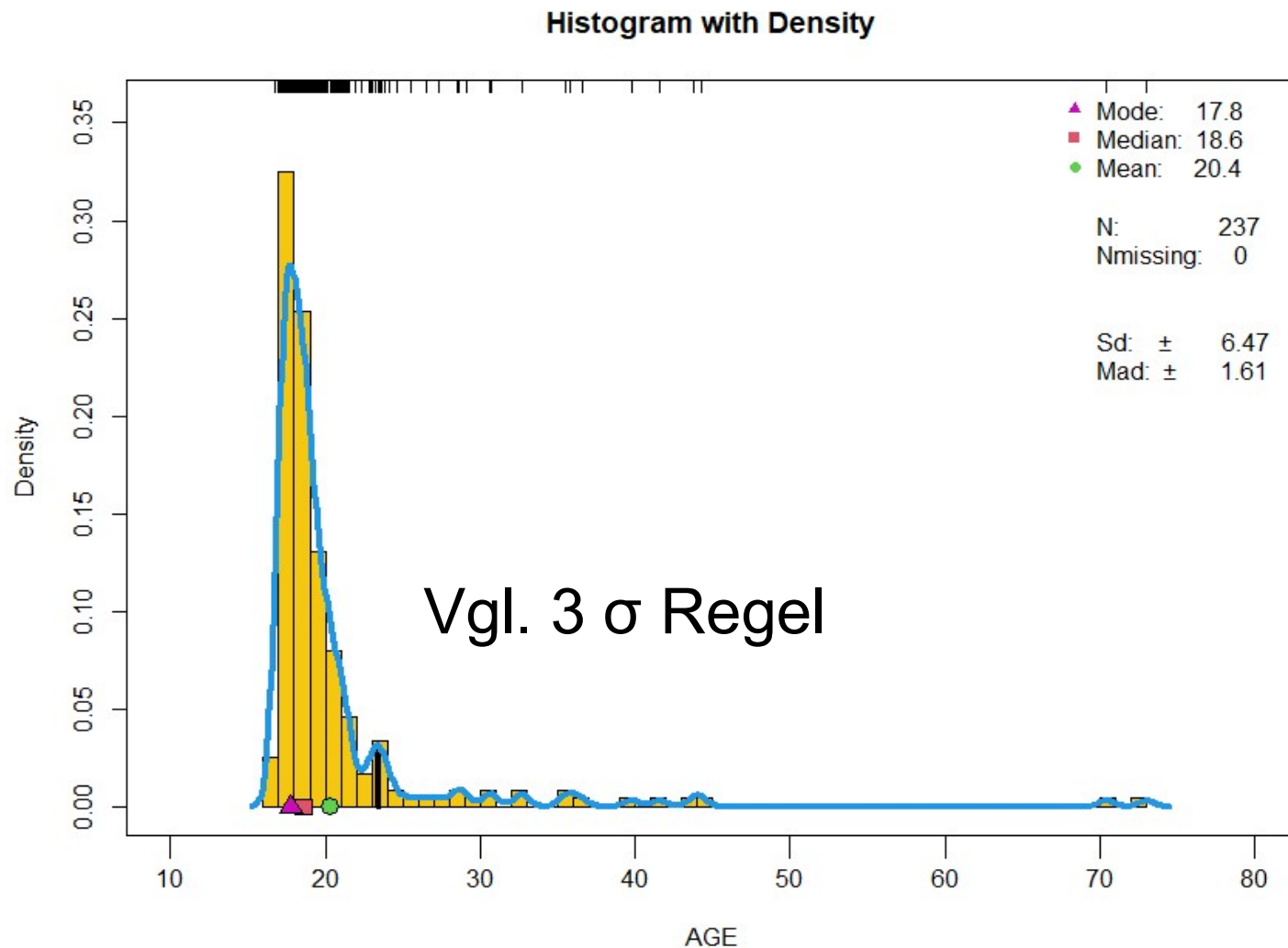
Age

sd	mad
----	-----

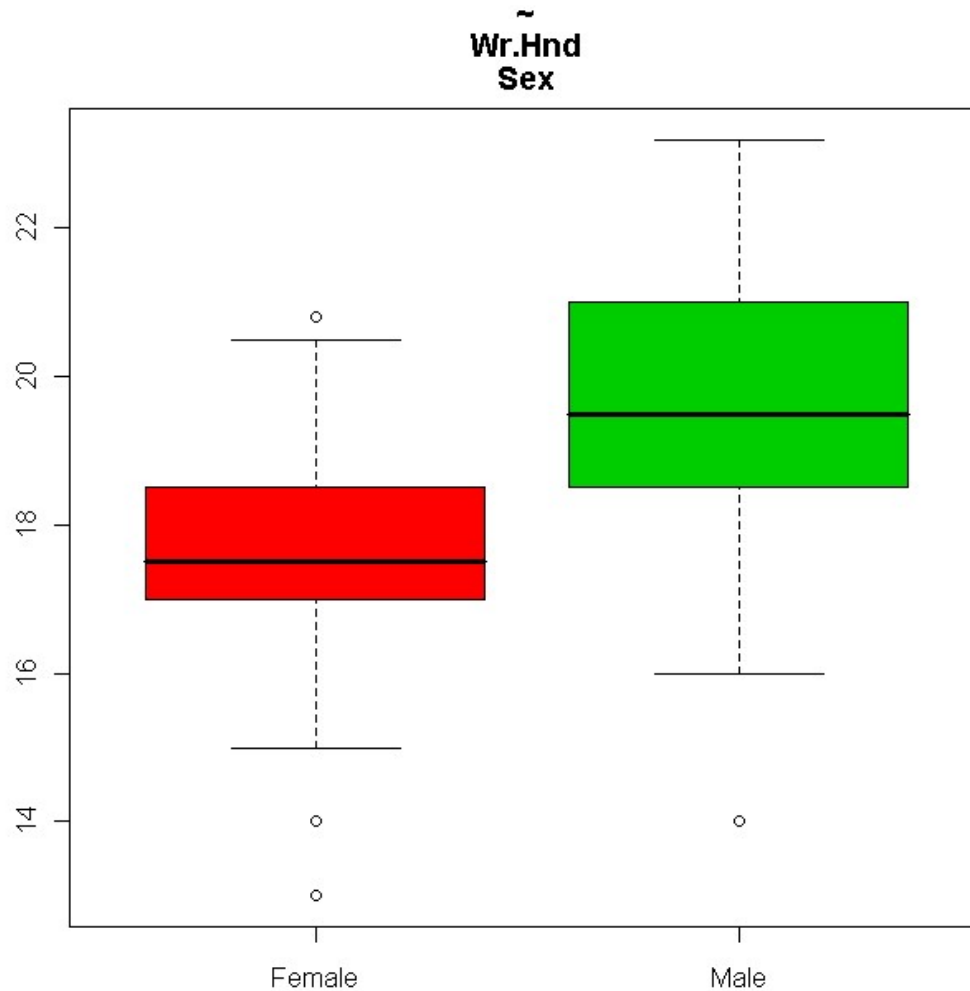
6.47	1.61	stark verschieden wegen der Oldies in der Seminargruppe
------	------	---

mad ist robust gegen Ausreißer

Median ± 3 (oder 5) mad zur Outlier Identifikation



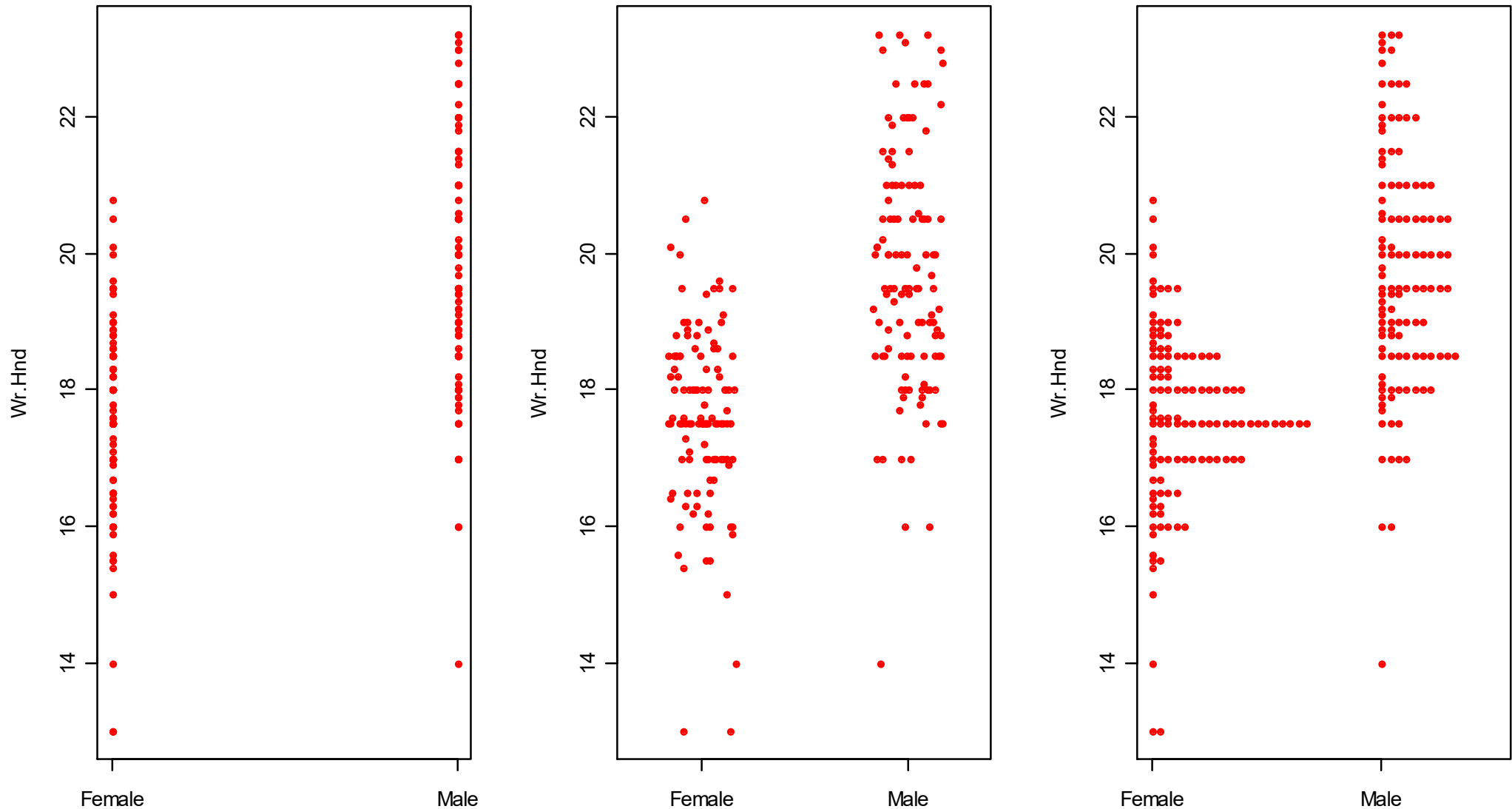
Boxplots zum Visualisieren auf Gruppenunterschieden



	Female	Male
Whisker low	15.00000	16.00000
1.quartil	17.00000	18.50000
median	17.50000	19.50000
3.quartil	18.50000	21.00000
Whisker high	20.50000	23.20000
95%-CI median low	17.28182	19.13482
95%-CI median high	17.71818	19.86518
# outlier	4.00000	1.00000

Kasten zwischen Q1 und Q3.
Median wird als Linie eingezeichnet.
Die Whisker ("Schurrhaare") liegen beim kleinsten/größten Wert innerhalb $[Q1 - 1.5 \cdot (Q3 - Q1), Q3 + 1.5 \cdot (Q3 - Q1)]$
Werte außerhalb werden als Outlier gezeigt

Stripcharts einfach, mit jitter, stacked



Zwei Gruppen t-test

```
> t.test(Wr.Hnd~Sex)
```

Welch Two Sample t-test

```
data: Wr.Hnd by Sex
```

```
t = -10.6187, df = 215.265, p-value < 2.2e-16
```

```
alternative hypothesis: true difference in means is  
not equal to 0
```

```
95 percent confidence interval:
```

```
-2.544482 -1.747753
```

```
sample estimates:
```

mean in group Female	mean in group Male
17.59576	19.74188

Zwei-Stichproben (ungepaarter) t-Test I

$$X_i \sim N(\mu_X, \sigma_X^2) \quad i = 1..n$$

$$Y_k \sim N(\mu_Y, \sigma_Y^2) \quad k = 1..m$$

$$\bar{X}_n \sim N(\mu_X, \frac{\sigma_X^2}{n}) \quad \text{und} \quad \bar{Y}_m \sim N(\mu_Y, \frac{\sigma_Y^2}{m})$$

$$(\bar{X}_n - \bar{Y}_m) \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$$

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N\left(\frac{\mu_X - \mu_Y}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}, 1\right)$$

Zwei-Stichproben (ungepaarter) t-Test II

Unter $H_0 : \mu_X = \mu_Y$ und falls σ_X und σ_Y bekannt sind, gilt :

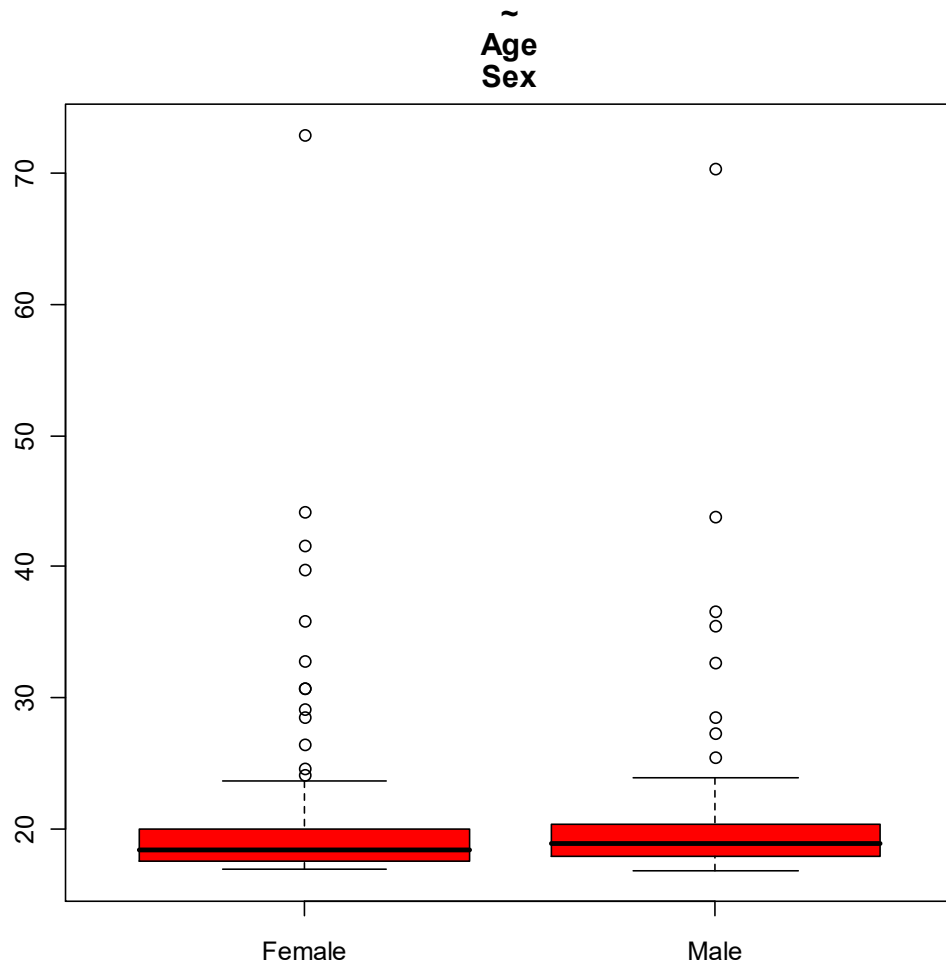
$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

Wenn die Varianzen aus den Daten geschätzt werden müssen, gilt approximativ :

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}}} \sim t_{df=\min(n,m)-1}$$

bzw. eine noch genauere Approximation nach Welch welche gebrochene Freiheitsgrade (df) liefert.

Gruppenvergleich bei stark nicht normalverteilten Daten



Welch Two Sample t-test

```
data: Age by Sex
t = 0.0893, df = 230.208, p-value = 0.929
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
-1.592129 1.743281
sample estimates:
mean in group Female    mean in group Male
      20.40753           20.33196
```

?

Wilcoxon rank sum test with continuity correction

```
data: Age by Sex
W = 5869, p-value = 0.03716
alternative hypothesis: true location shift is not
equal to 0
```

Warum diese Diskrepanz?

- Wegen der Ausreißer sind die Varianzschätzer im t-test stark inflationiert. Daher t-Test extrem konservativ.
- Hier besser: Mann-Whitney U-Test oder (synonym) ungepaarter Wilcoxon-Test.
- Welches Differenzmaß liegt dem MW-U-Test zugrunde?

Proversionswahrscheinlichkeit

X und Y reele Zufallsvariable.

$\text{pr}(X > Y)$ heißt die Proversionswahrscheinlichkeit.

Gegeben zwei Stichproben x_i $i = 1..n$ und y_k $k = 1..m$.

Man kann $\text{pr}(X > Y)$ schätzen, indem man $x_i > y_k$ in allen $n \cdot m$ Paaren (x_i, y_k) auszählt.

Für die Statistik U des Mann Whitney U-Test

bzw. W des Wilcoxon - Test gilt :

$$\hat{\text{pr}}(X > Y) = \frac{W}{n * m}$$

Im Altersbeispiel:

Wenn eine Frau und ein Mann zufällig ausgewählt werden, so wird die Frau in 41% der Fälle älter sein.

Konfidenzintervalle nach

Newcombe RG: Statistics in Medicine 2005

95%CI: [0.353 , 0.496]

Differenz der Pseudomediane

Die Proversionswahrscheinlichkeiten kann man schon bestimmen, sobald man nur ordinal Daten hat.

Hat man eine stetige intervallskalierte Skala, so liegt dem MW-Test als Differenzmaß die Differenz der

Synonym: Pseudomedian,

Median der Walsh- Mittelwerte

oder der Hodge-Lehmann Lokationsschätzer

zugrunde.

$$\text{Pseudomedian:} = \text{median}_{i \leq j} \left(\frac{(x_i + x_j)}{2} \right)$$

Alter nach Geschlecht - Beispiel

```
> wilcox.test(Age~Sex, conf.int = TRUE)
data:  Age by Sex
W = 5869, p-value = 0.03716
alternative hypothesis: true location shift
is not equal to 0
95 percent confidence interval:
 -7.500042e-01 -3.864974e-05
sample estimates:
difference in location  -0.4160449

> pseudomedian(Age[Sex=="Female"])
[1] 18.708
> pseudomedian(Age[Sex=="Male"])
[1] 19.125
```

Ungepaarte versus gepaarte Gruppenvergleiche

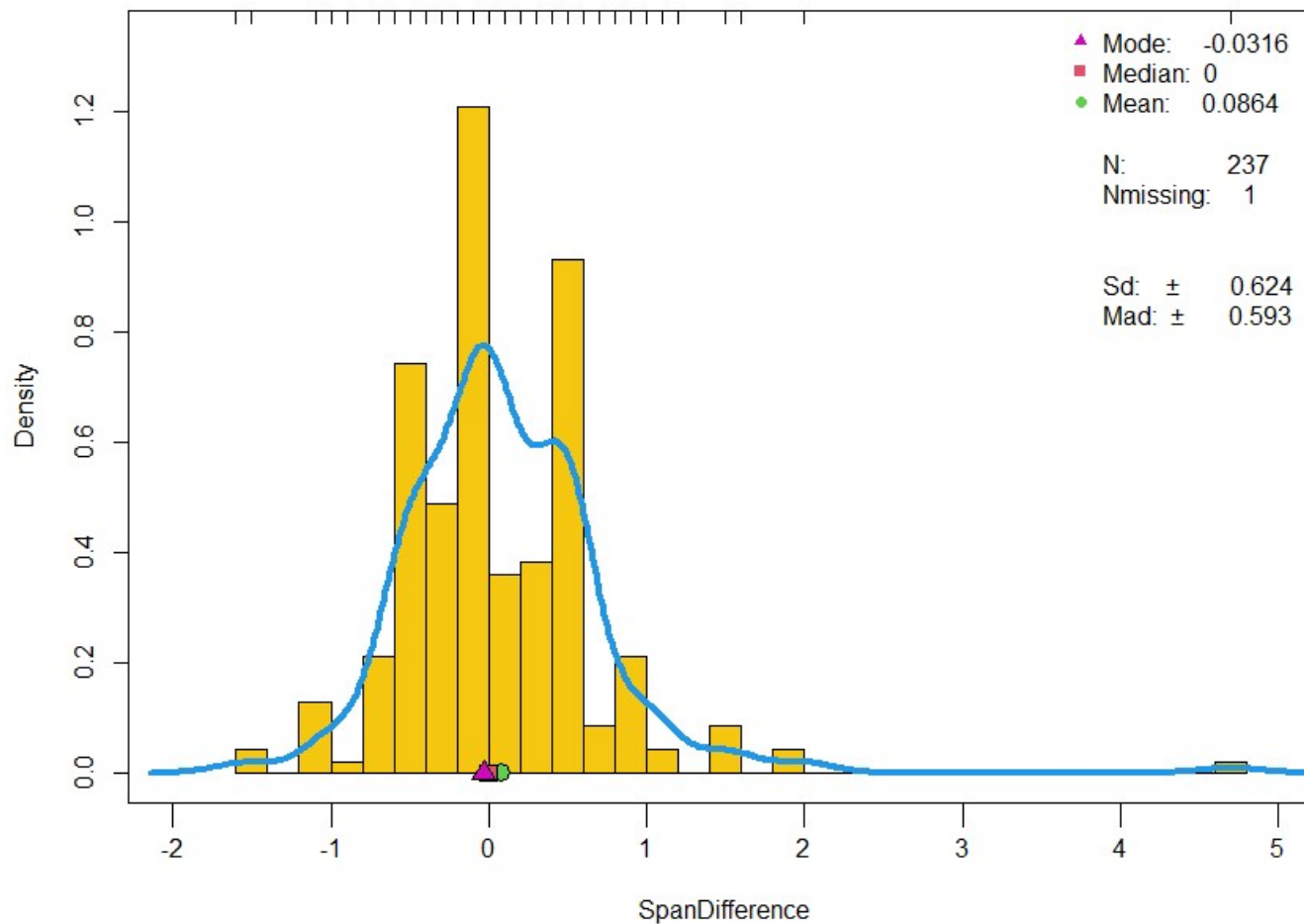
- **Ungepaart (besser: zwei Stichproben-Vergleich)**
 - Alter by sex
- **Gepaart (besser: ein Stichproben-Vergleich) :**
 - Schreibhandspanne versus
Nicht-Schreibhandspanne
 - jeweils derselben Person**
(=Beobachtungseinheit)
 - → **Einstichproben-Test auf Differenzen auf
Lokationsparameter=0**

Ausschalten von Störgrößen durch gepaarte Designs

- Frage: Ist die Schreibhand im Allgemeinen durch ihren stärkeren Gebrauch größer als die Nichtschreibhand ?
- **Dummes Experiment:**
 - Teile Studenten zufällig in zwei Gruppen,
 - messe in einer Gruppe die Schreibhandspanne,
 - in der anderen die Nicht-Schreibhandspanne
 - Zwei Stichproben t-Test
- Inter-individuelle **Störgröße:** Allgemeine Körpergröße
- **Besser:** Inter-individuelle Varianz wird ausgeschaltet, wenn man an jeder Beobachtungseinheit beide Messungen durchführt und **intra-individuelle Differenzen** betrachtet.
 - → Ein Stichproben t-Test

Histogramm der Differenzen: Schreibhandspanne - Nicht-Schreibhandspanne

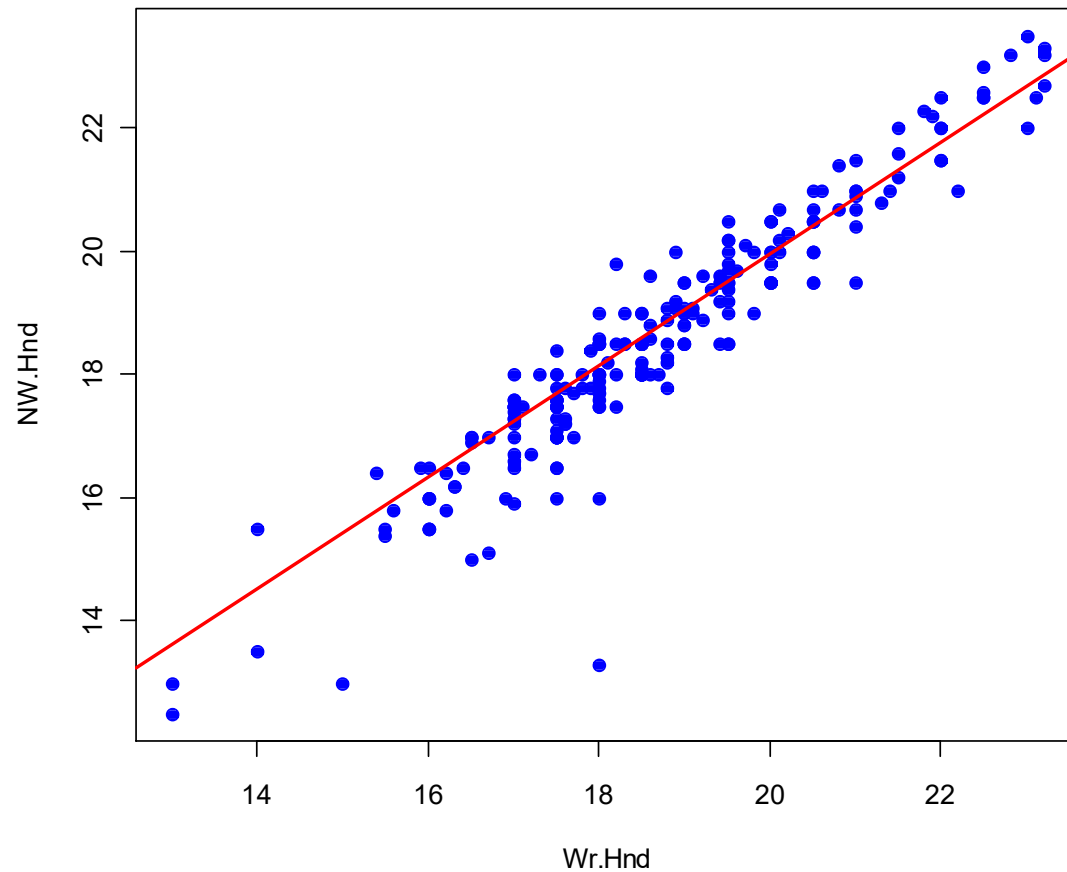
Histogramm with Density



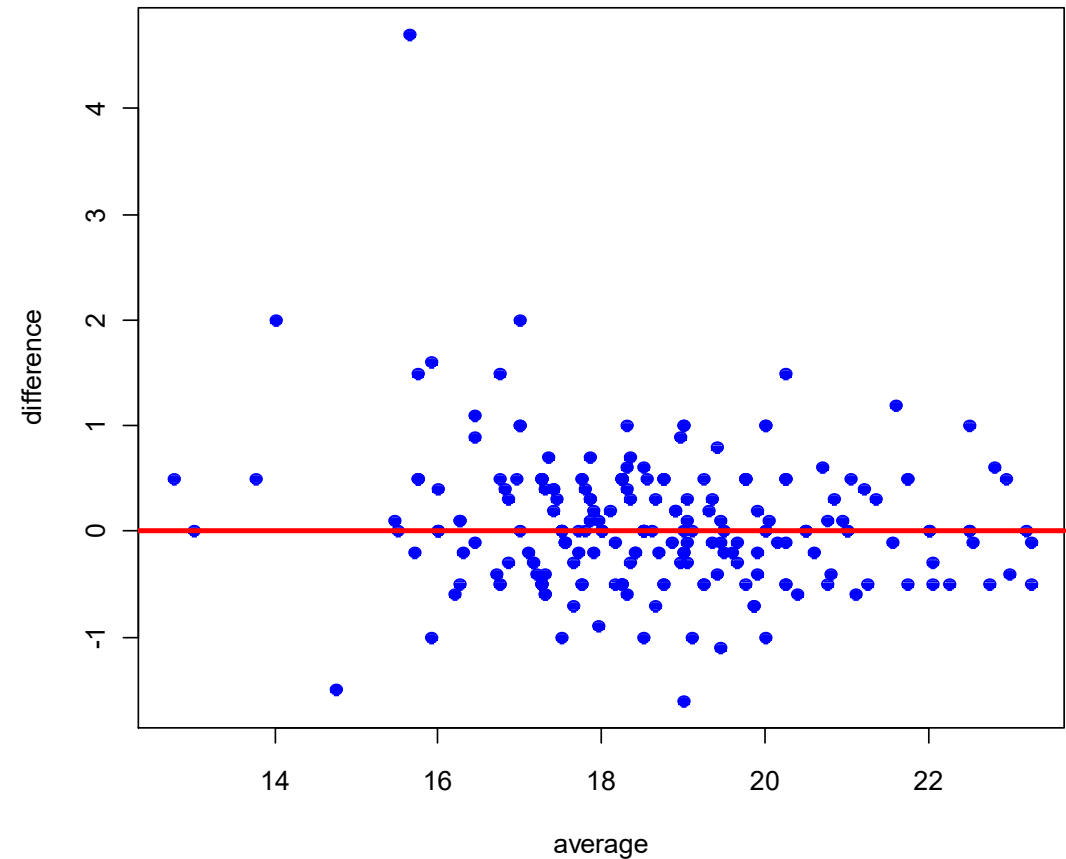
```
> pseudomedian(Wr.Hnd)
[1] 18.6
> pseudomedian(NW.Hnd)
[1] 18.5
```

Scatterplot und Bland Altman Plot

scatterplot



Bland Altman Plot



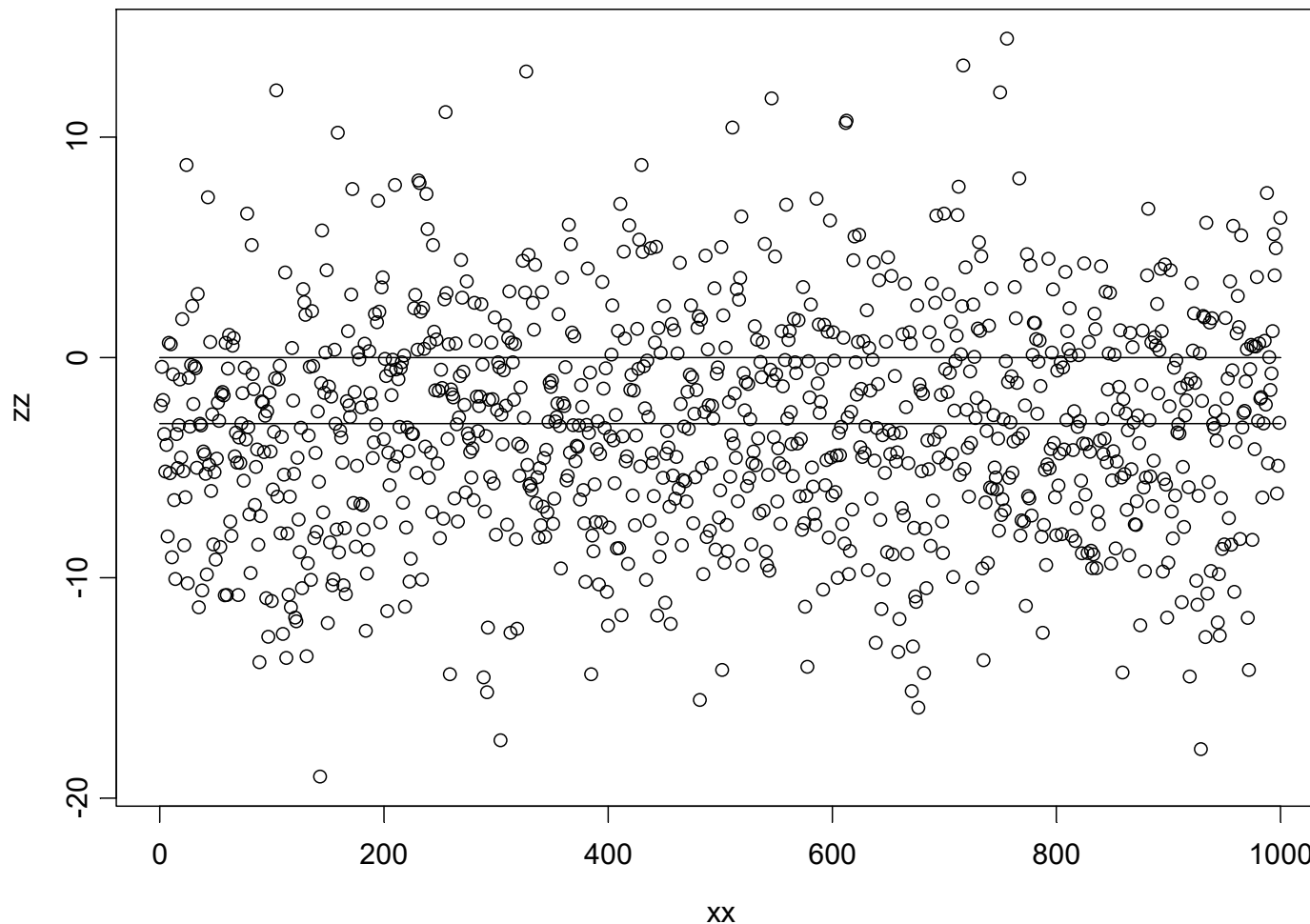
Bland Altman Plot

- Technik zum Vergleich von Messgrößen, die ungefähr **dasselbe** messen
- **Standardvisualisierung zum Vergleich von Messmethoden**
- **CAVE: Lineare Regression untersucht nur linearen Zusammenhang, nicht Übereinstimmung !**
- **BA-Plot macht sichtbar, ob Messfehler konstant additiv oder vom zu messenden Wert anhängen.**

Bland Altman Plot – Messfehlertypik

$y = x + \varepsilon; \varepsilon \sim \text{Normal}(\text{bias}, \sigma^2)$

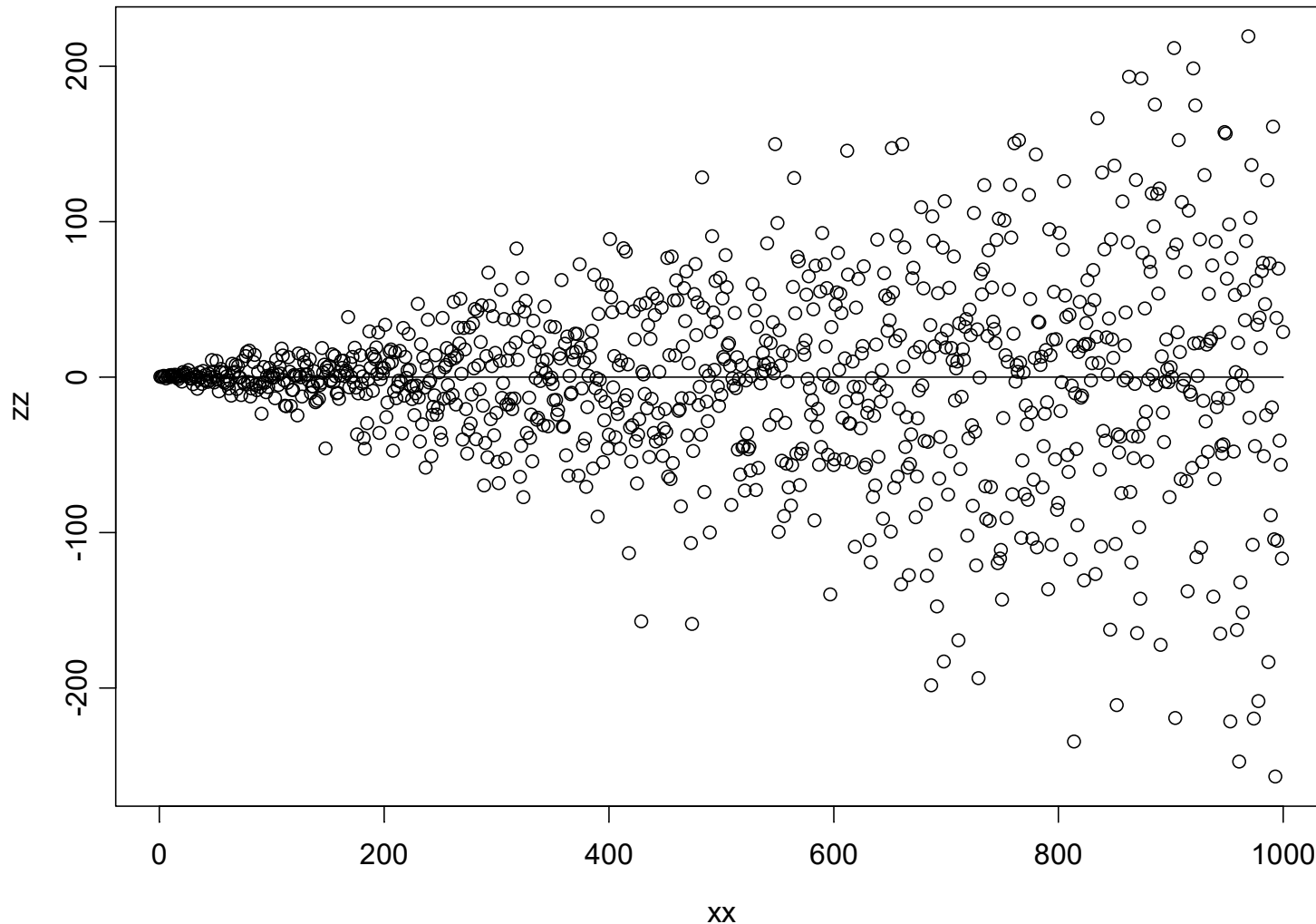
Konstanter Bias; Konstanter additiver Fehler



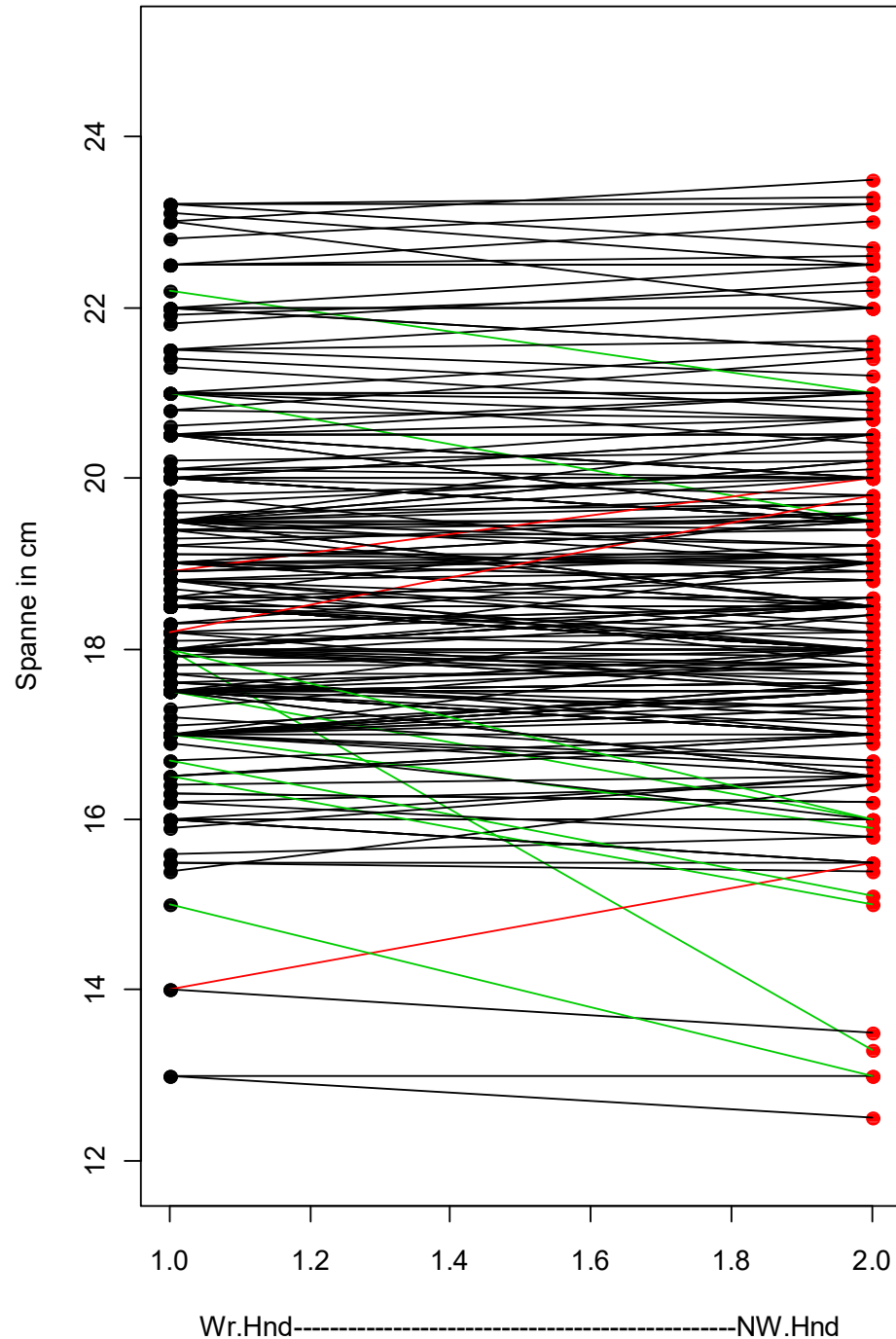
Bland Altman Plot – Messfehlertypik

$$y = \exp(\varepsilon')x; \varepsilon' \sim \text{Normal}(0, \sigma'^2)$$

Multiplikativer Fehler unabhängig vom zu messenden Wert



Paar Plot



Paarplot

Messungen an einer Beobachtungseinheit werden verbunden.

Rot : $Wr.Hnd - NW.Hnd < -1$

Grün: $Wr.Hnd - NW.Hnd > +1$

Ein-Stichproben (gepaarter) t-Test I

Gegeben Paare von Messungen an einer Beobachtungseinheit

$$(X_i, Y_i), \quad i = 1..n$$

Bilde die Differenzen $D_i := (X_i - Y_i)$

Ang.: Approximativ normalverteilt: $D_i \sim N(\mu_D, \sigma_D^2) \quad i = 1..n$

Dann gilt: $\bar{D}_n \sim N(\mu_D, \frac{\sigma_D^2}{n})$

Also:
$$\frac{\bar{D}_n}{\sqrt{\frac{\sigma_D^2}{n}}} = \frac{\bar{D}_n}{\sigma_D} \cdot \sqrt{n} \sim N\left(\frac{\mu_D}{\sigma_D} \cdot \sqrt{n}, 1\right)$$

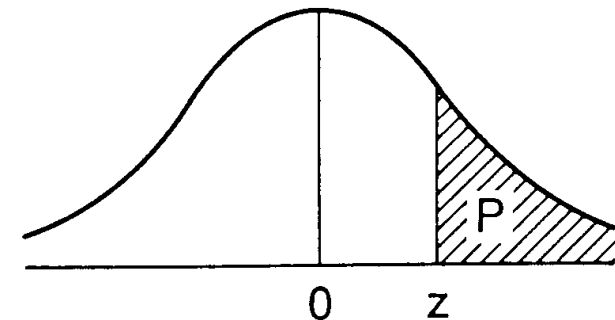
Ein-Stichproben (gepaarter) t-Test II

Unter der Nullhypothese $H_0: \mu_D = 0$

$$\frac{\bar{D}_n}{\sigma_D} \cdot \sqrt{n} \sim N(0, 1) \text{ standardnormalverteilt.}$$

Falls σ_D aus den Daten geschätzt werden muss :

$$\frac{\bar{D}_n}{\hat{\sigma}_D} \cdot \sqrt{n} \sim t_{df=n-1} \text{ t - verteilt}$$



Ein-Stichproben t-Test

```
t.test(Wr.Hnd,NW.Hnd,paired=TRUE)
```

Paired t-test

data: Wr.Hnd and NW.Hnd

t = 2.1268, df = 235, p-value = 0.03448

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:

0.006367389 0.166513967

mean of the differences

0.08644068

Ein-Stichproben Wilcoxon signed rank test

```
> wilcox.test(Wr.Hnd,NW.Hnd,paired=TRUE,conf.int = TRUE)
```

Wilcoxon signed rank test with continuity correction

data: Wr.Hnd and NW.Hnd

V = 10919, p-value = 0.08225

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

-1.402869e-05 1.500411e-01

sample estimates:

(pseudo)median 0.05005861

MERKE:

Wenn Daten ~normalverteilt hat wilcox.test nur

~95% der Power des t-Tests

Hier im Beispiel gerade auf der Kippe!

Interpretation des Beispiels: Quantitativ in der Größenordnung des Messfehlers!
t- Test möglicherweise verzerrt durch Ausreißer. -> Kein relevanter Unterschied.

Wilcoxon Test für Paardifferenzen

„Wilcoxon-matched-pairs-signed-rank-Test“

1. Messung: $x_{A1}, x_{A2}, \dots, x_{An}$

2. Messung: $x_{B1}, x_{B2}, \dots, x_{Bn}$

Berechnung von $d_i = x_{Ai} - x_{Bi}$, $i = 1, 2, \dots, n$.

H_0 : Verteilung der Differenzen symmetrisch um 0

H_1 : Verteilung der Differenzen nicht symmetrisch um 0

Vorgehensweise: (Für kleine Fallzahlen)

- Bildung der Rangwerte r_i der absoluten Differenzen $|d_i|$, $i = 1, 2, \dots, n$, sowie Rangsummen r_+ (r_-) der Rangwerte mit positiven (negativen) Differenzen.
- Ablehnung von H_0 , falls der kleinere der beiden Werte den kritischen Wert des Wilcoxon-Tests nicht überschreitet. ⁴

Wilcoxon Test für Paardifferenzen

Beispiel

Angina-Pectoris-Anfälle unter Pronethalol bzw. Placebo im „Zwei-Perioden“ Design mit zufälliger Reihenfolge (Trampisch, Windeler, 1997); $n = 12$.

Anfälle					
unter Placebo	unter Pronethalol	d_i	Vorzeichen	$ d_i $	Rangwerte
71	29	42	+	42	12
323	348	-25	-	25	11
8	1	7	+	7	6
14	7	7	+	7	6
23	16	7	+	7	6
34	25	9	+	9	8
79	65	14	+	14	9
60	41	19	+	19	10
2	0	2	+	2	1.5
3	0	3	+	3	3
17	15	2	+	2	1.5
7	2	5	+	5	4

Beachte: Bei Bindungen („ties“) werden Rangwerte gemittelt

$$r_+ = 67 ; r_- = 11$$

$$\text{Kontrolle: } r_+ + r_- = \frac{n \cdot (n+1)}{2} = \frac{12 \cdot 13}{2} = 78$$

5

Wassmer Internet

Wilcoxon Test für Paardifferenzen

Kritische Werte für den Wilcoxon-Test

	α (zweiseitig); 2α (einseitig)			
n	0,10	0,05	0,02	0,01
5	0	–	–	–
6	2	0	–	–
7	3	2	0	–
8	5	3	2	0
9	8	5	3	2
10	10	8	5	3
11	13	10	7	5
12	17	13	10	7
13	21	17	13	10
14	25	21	16	13
15	30	25	20	16
20	60	52	43	38
25	100	89	77	68

Kleinere Rangsumme $r_- = 11 \leq 13$

Deshalb kann H_0 zum Niveau $\alpha = 5\%$ (zweiseitig) abgelehnt werden

Wilcoxon Test für Paardifferenzen

asymptotische Version

- H0: **Verteilung der Differenzen symmetrisch um 0**

- Testgröße:

$$W = r_+ - r_- = \sum_{j=1}^n \text{Rang}(x_j) * \text{sign}(x_j)$$

$$E_0(W) = 0$$

$$\text{var}_0(W) = \text{var}_0\left(\sum_{j=1}^n \text{Rang}(x_j) * \text{sign}(x_j)\right) =$$

$$= \sum_{j=1}^n \text{var}_0(\text{Rang}(x_j) * \text{sign}(x_j)) =$$

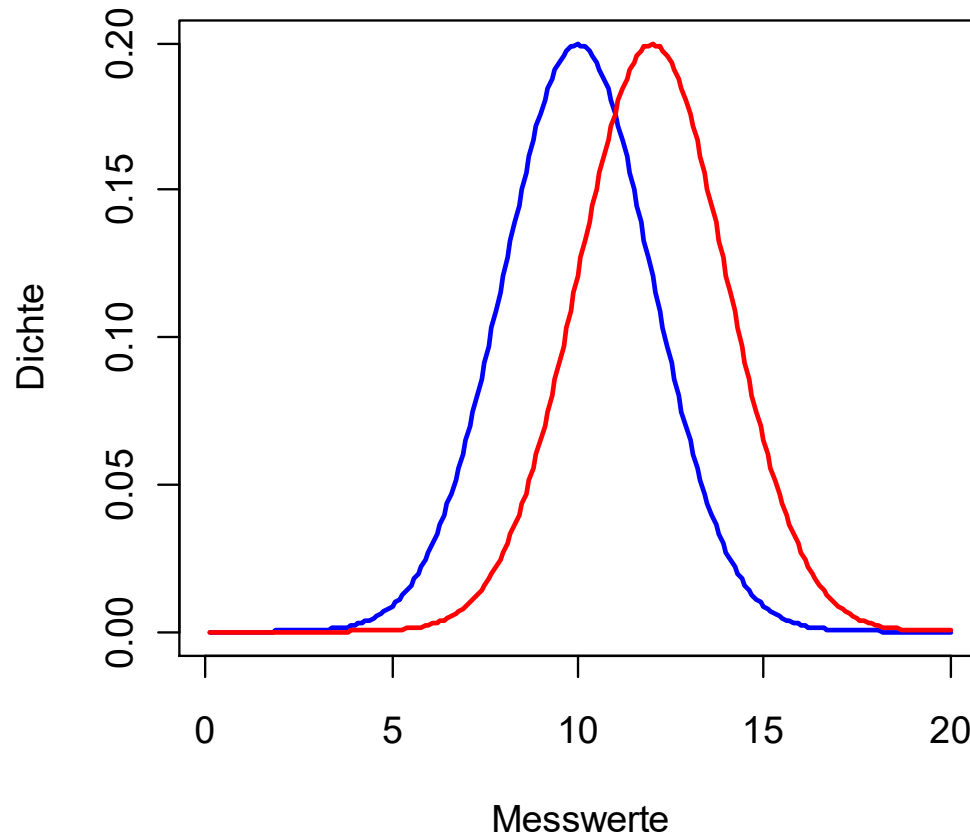
$$= \sum_{j=1}^n \text{Rang}(x_j)^2 * \text{var}_0(\text{sign}(x_j)) =$$

$$= \sum_{j=1}^n j^2 * 1 = \frac{n * (n + 1)(2n + 1)}{6}$$

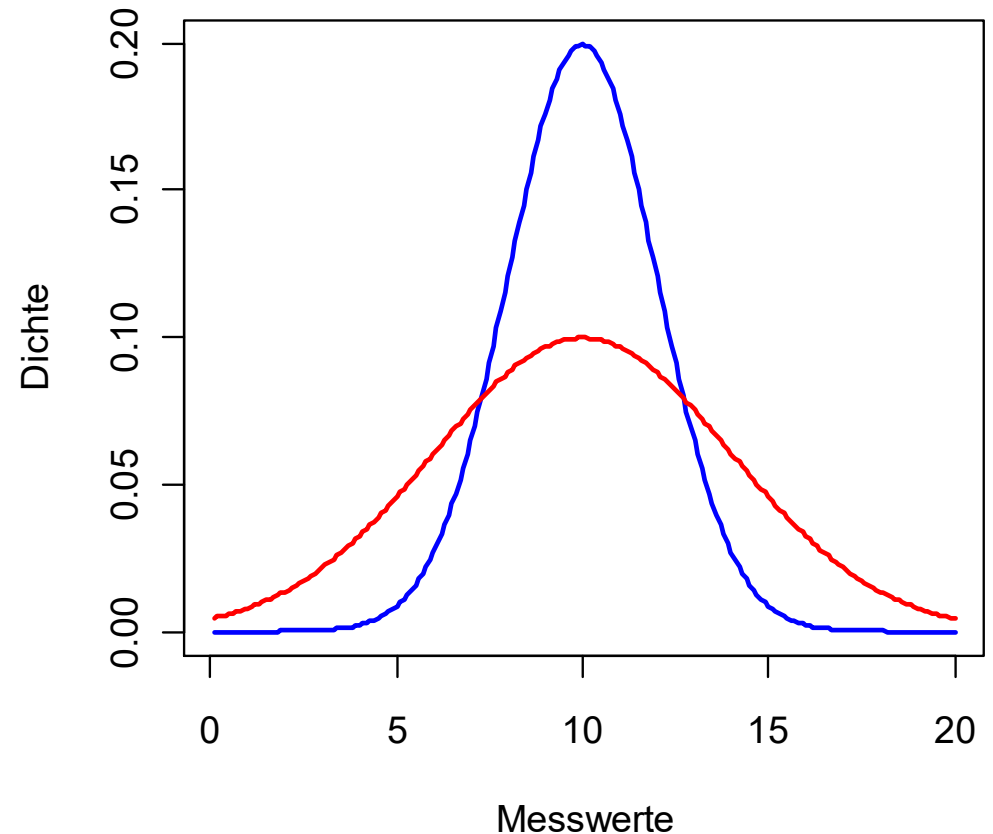
$$W \sim N(0, \text{var}_0(W)) \quad \text{für große } n$$

Unterschied in Varianzen

Unterschied im Erwartungswert



Unterschied in der Varianz



Unterschied in Varianzen

- Unterschiede in Varianzen kann man durch das Varianzverhältnis quantifizieren.
- F-test testet die Nullhypothese Varianzverhältnis=1

```
> var(Wr.Hnd, na.rm=TRUE)
[1] 3.530571
> var(NW.Hnd, na.rm=TRUE)
[1] 3.869356
>
> var.test(Wr.Hnd, NW.Hnd)
      F test to compare two variances
data:  Wr.Hnd and NW.Hnd
F = 0.9124, num df = 235, denom df = 235, p-value = 0.483
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7061286 1.1790402
sample estimates:
ratio of variances
      0.912444
```

Warum ist das jetzt Blödsinn?

```
> var.test(Wr.Hnd, Age)
```

F test to compare two variances

data: Wr.Hnd and Age

F = 0.0842, num df = 235, denom df = 236, p-value < 2.2e-16

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.06520278 0.10881125

sample estimates:

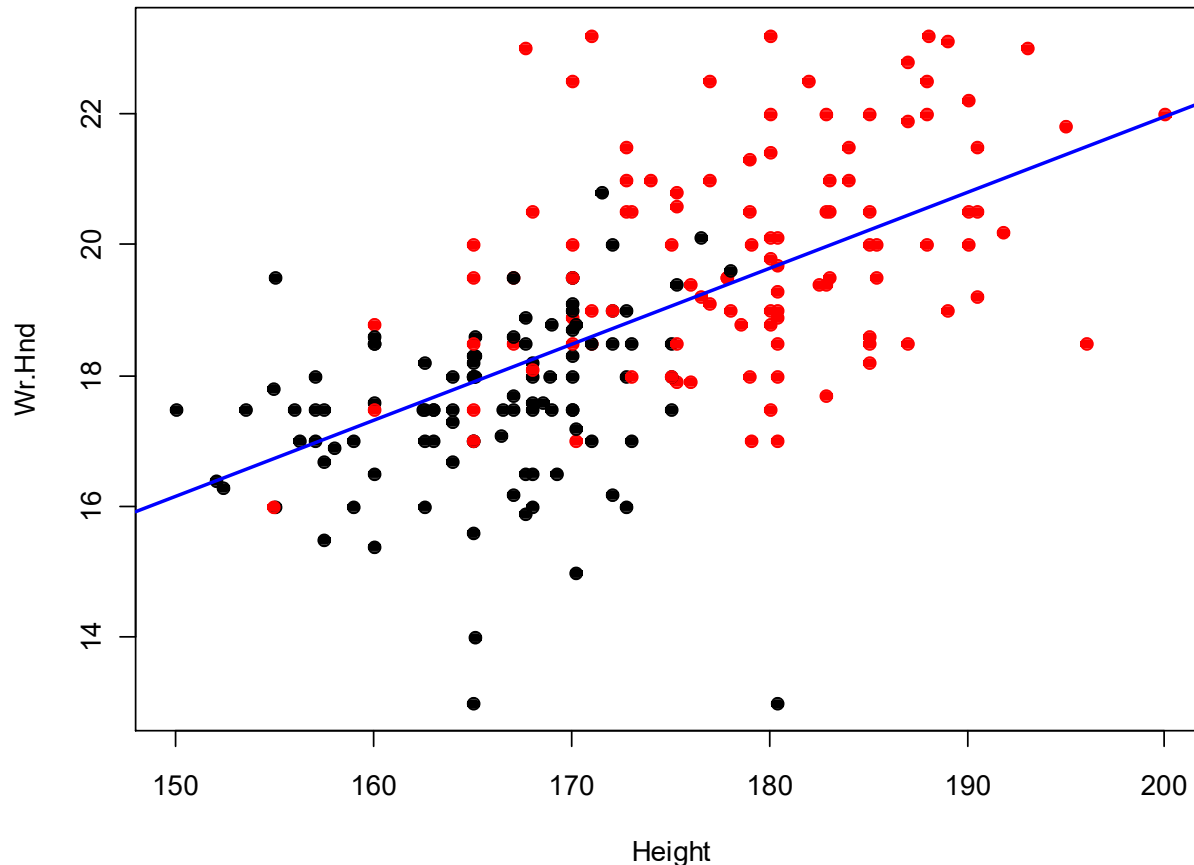
ratio of variances

0.08422764

(Hinweis: F-Test anfällig gegen Ausreißer
ansari.test() Vgl. Sachs p431)

Visualisierung: Eingefärbtes Streudiagramm

Wr.Hnd ~ Height coloured by Sex



Analysiere
Linearen Zusammenhang

**Pearson'scher
Korrelationskoeffizient**

```
>cor (Height, Wr.Hnd, use="pairwise.complete.obs")  
[1] 0.600991
```

Kovarianz

$$\text{cov}(X, Y) := E((X - E(X)) \cdot (Y - E(Y)))$$

Eigenschaften:

- a) $\text{cov}(X, Y) = \text{cov}(Y, X)$ (Symmetrie)
- b) $\text{cov}(a + X, Y) = \text{cov}(X, Y)$ (translationsinvariant)
- c) $\text{cov}(\alpha X + \beta Z, Y) = \alpha \cdot \text{cov}(X, Y) + \beta \cdot \text{cov}(Z, Y)$ (linear)

Folgen direkt aus Eigenschaften des Erwartungswerts.

Unabhängigkeit und Kovarianz

Notiz: Sind X und Y unabhängig, so gilt:

$$\text{cov}(X, Y) = 0$$

Beweis: $E((X - E(X)) \cdot (Y - E(Y))) =$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X)) \cdot (y - E(Y)) \cdot f_{X,Y}(x, y) dx dy =$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X)) \cdot (y - E(Y)) \cdot f_X(x) \cdot f_Y(y) dx dy =$$

$$\int_{-\infty}^{\infty} (y - E(Y)) \cdot f_Y(y) \left(\int_{-\infty}^{\infty} (x - E(X)) \cdot f_X(x) dx \right) dy = 0$$

Vorsicht: Umkehrung gilt nicht!

Unabhängigkeit und Varianz

$$\begin{aligned}\text{var}(X+Y) &= \text{cov}(X+Y, X+Y) \\ &= \text{cov}(X, X) + \text{cov}(Y, Y) + 2 \cdot \text{cov}(X, Y) \\ &= \text{var}(X) + \text{var}(Y) + 2 \cdot \text{cov}(X, Y)\end{aligned}$$

Notiz: Sind X und Y unabhängig, so gilt:

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$$

Korrelation

Die Kovarianz ändert sich bei Skalentransformationen

$$X \rightarrow \alpha X \text{ oder } Y \rightarrow \beta Y$$

Daher definiert man die Korrelation

$$\text{cor}(X, Y) := \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)} \cdot \sqrt{\text{var}(Y)}}$$

Eigenschaften:

a) $-1 \leq \text{cor}(X, Y) \leq +1$ (Beweis mit Cauchy-Schwarz Ungleichung)

b) Die Korrelation ist invariant unter affinen Transformationen

$$X \rightarrow \tau + \alpha X \text{ oder } Y \rightarrow \tau + \beta Y.$$

c) Wenn $Y = a + bX$: $\text{cor}(X, Y) = \frac{b \cdot \text{cov}(X, X)}{\sqrt{\text{var}(X)} \cdot |b| \cdot \sqrt{\text{var}(X)}} = \text{sign}(b)$

Schätzer für Kovarianz und Korrelation

Gegeben Paare $(x_i, y_i), i = 1..n$

$$\hat{c}ov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n) \cdot (y_i - \bar{y}_n)$$

$$\hat{c}or(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x}_n) \cdot (y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y}_n)^2}}$$

Test und Konfidenzintervall für Korrelationskoeffizienten

```
> cor.test(Height,Wr.Hnd,use="pairwise.complete.obs")
```

Pearson's product-moment correlation

data: Height and Wr.Hnd

t = 10.7923, df = 206, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5063486 0.6813271

sample estimates:

cor

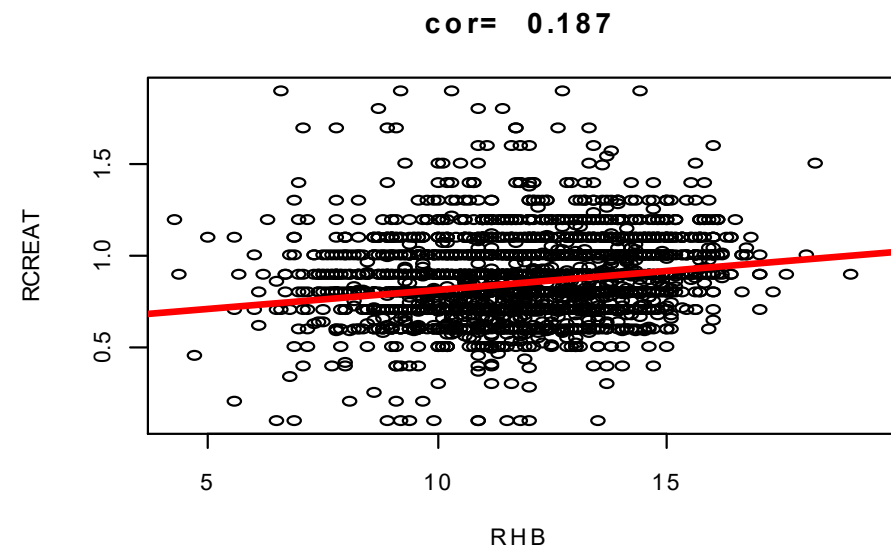
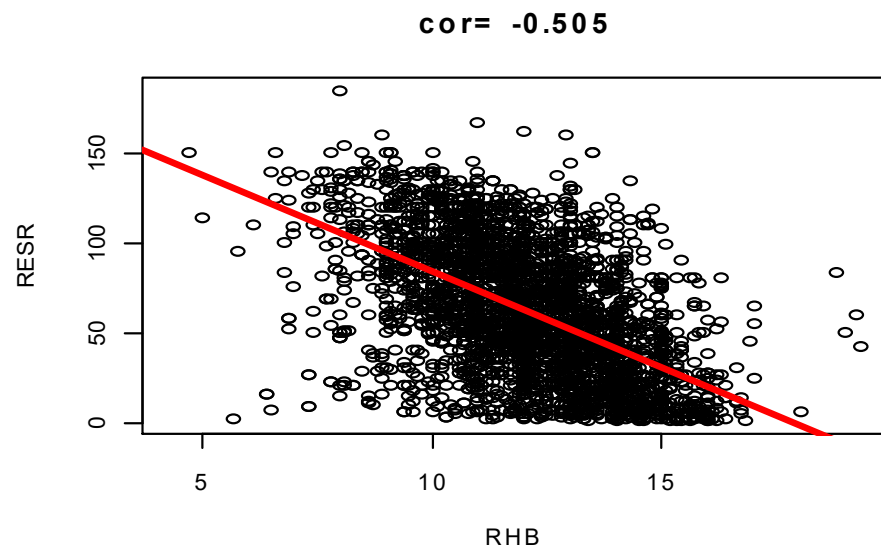
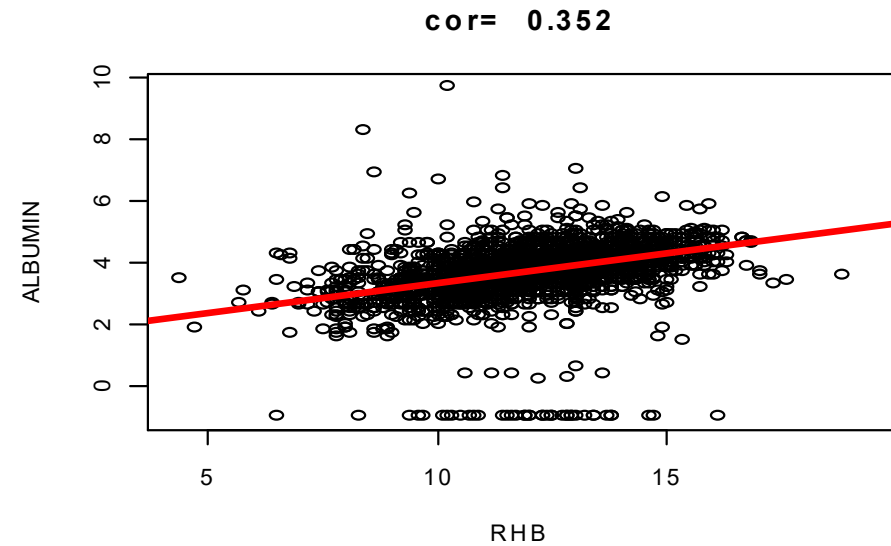
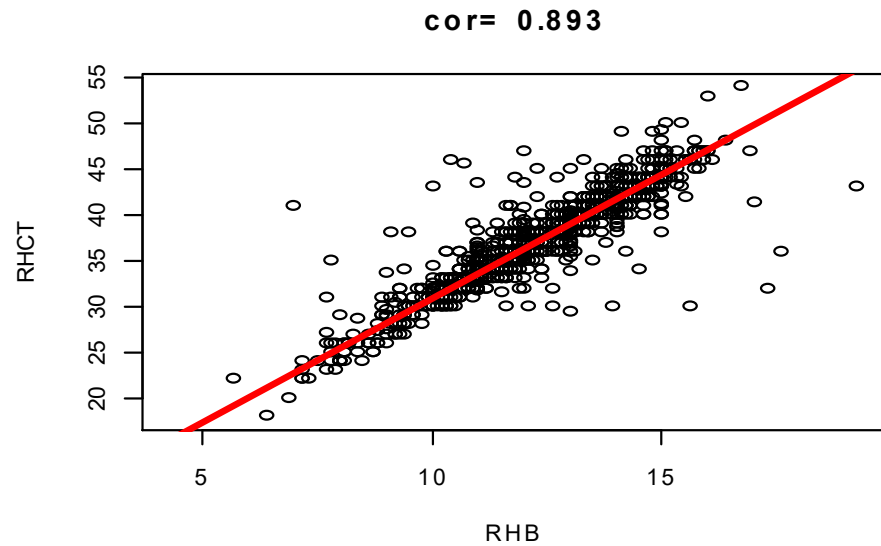
0.600991

Korrelationsmatrix

```
>  
cor(cbind(RHB,RHCT,ALBUMIN,RESR,RCREAT),use="pairwise.complete.  
obs" )
```

	RHB	RHCT	ALBUMIN	RESR	RCREAT
RHB	1.0000000	0.8927524	0.3524108	-0.5048772	0.1869212
RHCT	0.8927524	1.0000000	0.4498254	-0.5002600	0.1466265
ALBUMIN	0.3524108	0.4498254	1.0000000	-0.2871227	0.1279596
RESR	-0.5048772	-0.5002600	-0.2871227	1.0000000	-0.1133845
RCREAT	0.1869212	0.1466265	0.1279596	-0.1133845	1.0000000

Beispiele: Korrelationen zwischen Blutwerten



Korrelationskoeffizient nach Pearson

■ Grobe Orientierung zur Beurteilung der Stärke eines Zusammenhangs:

$|r| < 0,3$ niedrig

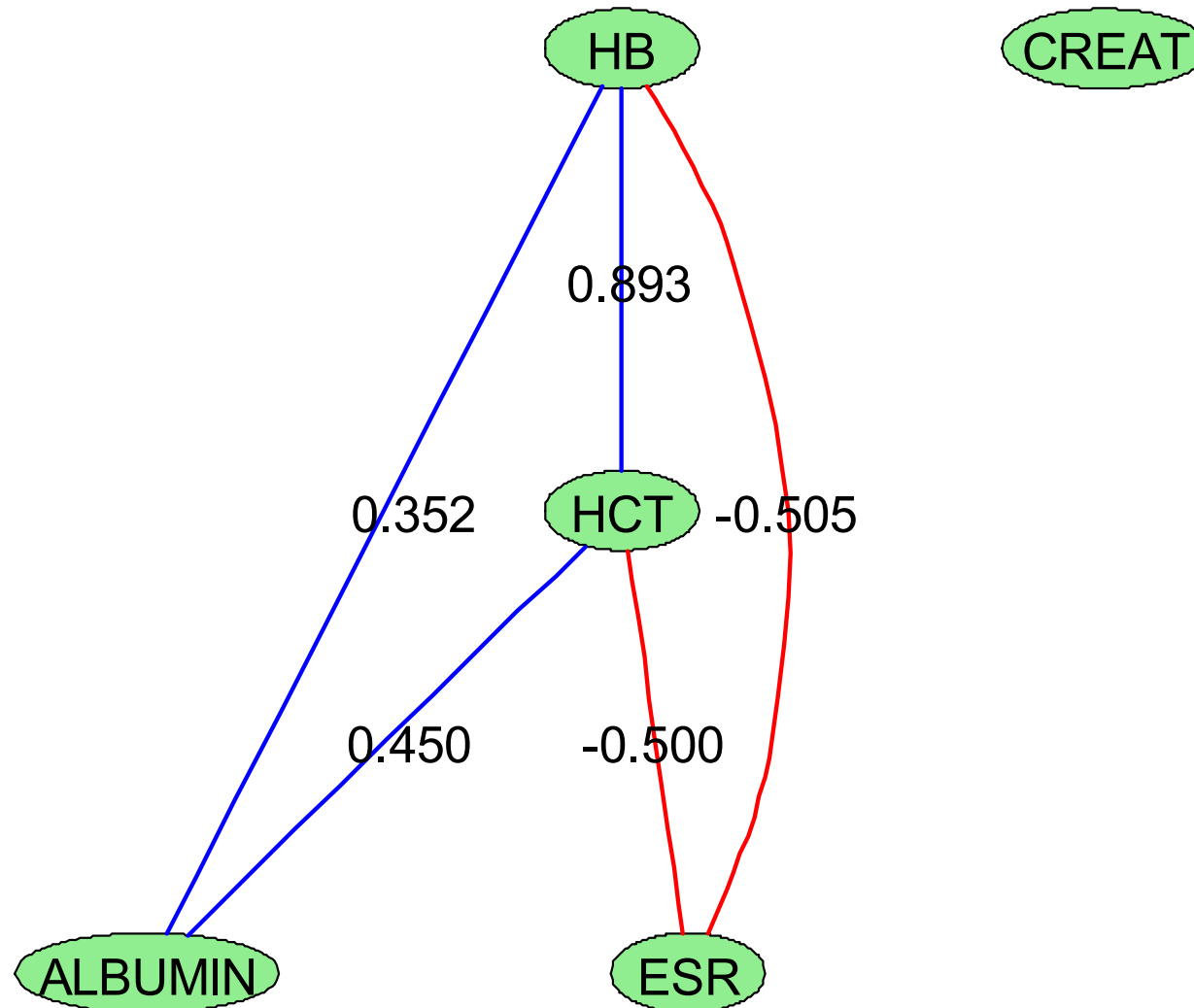
$0,3 \leq |r| < 0,6$ mittel

$0,6 \leq |r| < 0,8$ stark

$0,8 \leq |r| < 1,0$ sehr stark

$|r| = 1,0$ funktionaler Zusammenhang

Graph als Visualisierung von Zusammenhängen



Fehlschlüsse möglich durch

- Formale Korrelation
- durch Ausreißer bedingte Korrelation
- durch Inhomogenitäten bedingte Korrelation
- Gemeinschaftskorrelation

Formale Korrelation

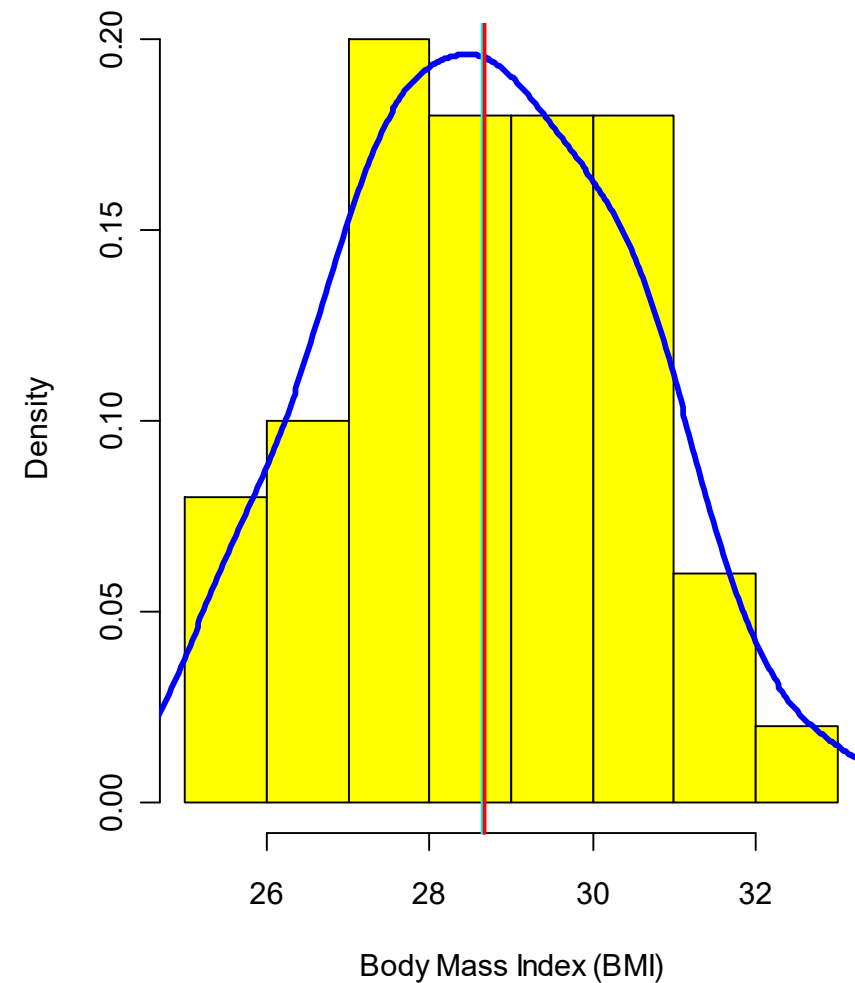
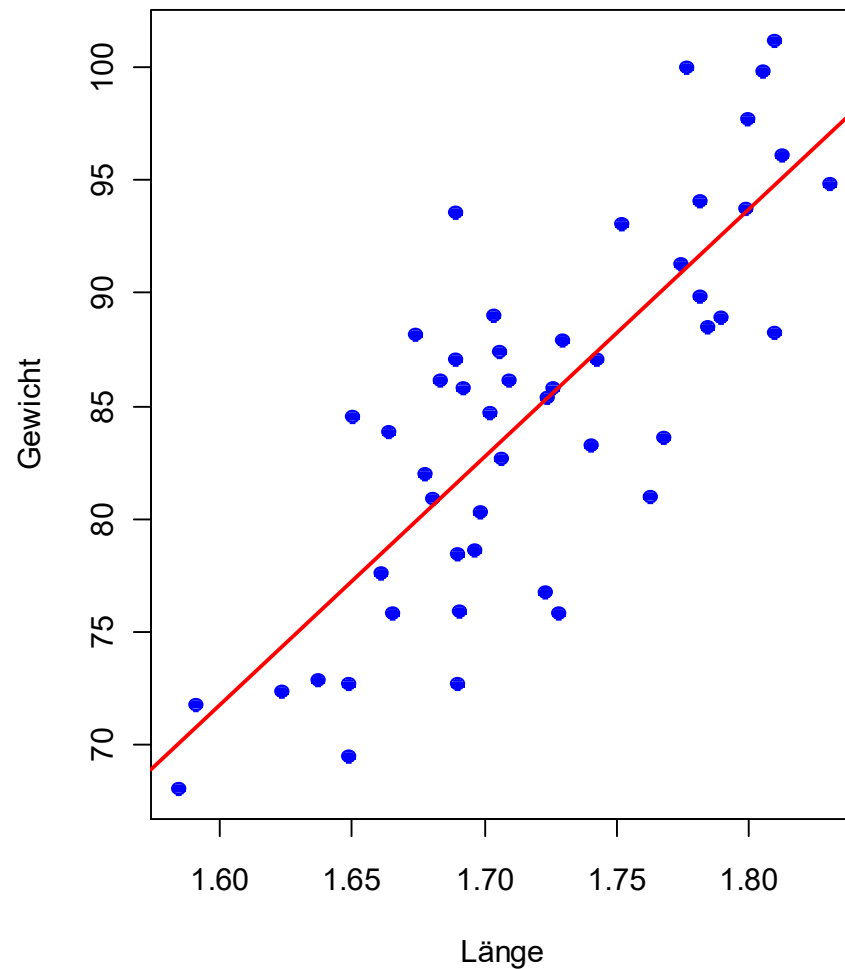
Beispiel: Differential-Blutbild

i	Myeloische Blutzellen x %	Lymphatische Blutzellen y %
1	73	27
2	85	15
3	69	31
4	94	6
5	70	30
6	52	48
7	73	27
22	64	36
23	88	12
24	65	35
25	83	17
26	80	20
27	57	43
28	74	26
29	53	47
30	86	14

$$r = -1$$

**Formale Korrelation durch
Differenzbildung bei Prozentzahlen**

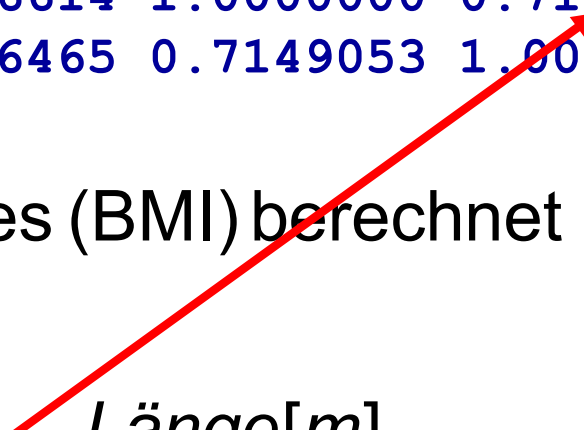
Korrelation durch rechnerische Abhängigkeit



Korrelation durch rechnerische Abhängigkeit

```
> cor(cbind(Länge, Gewicht, BMI))
```

	Länge	Gewicht	BMI
Länge	1.0000000	0.7926614	0.1426465
Gewicht	0.7926614	1.0000000	0.7149053
BMI	0.1426465	0.7149053	1.0000000

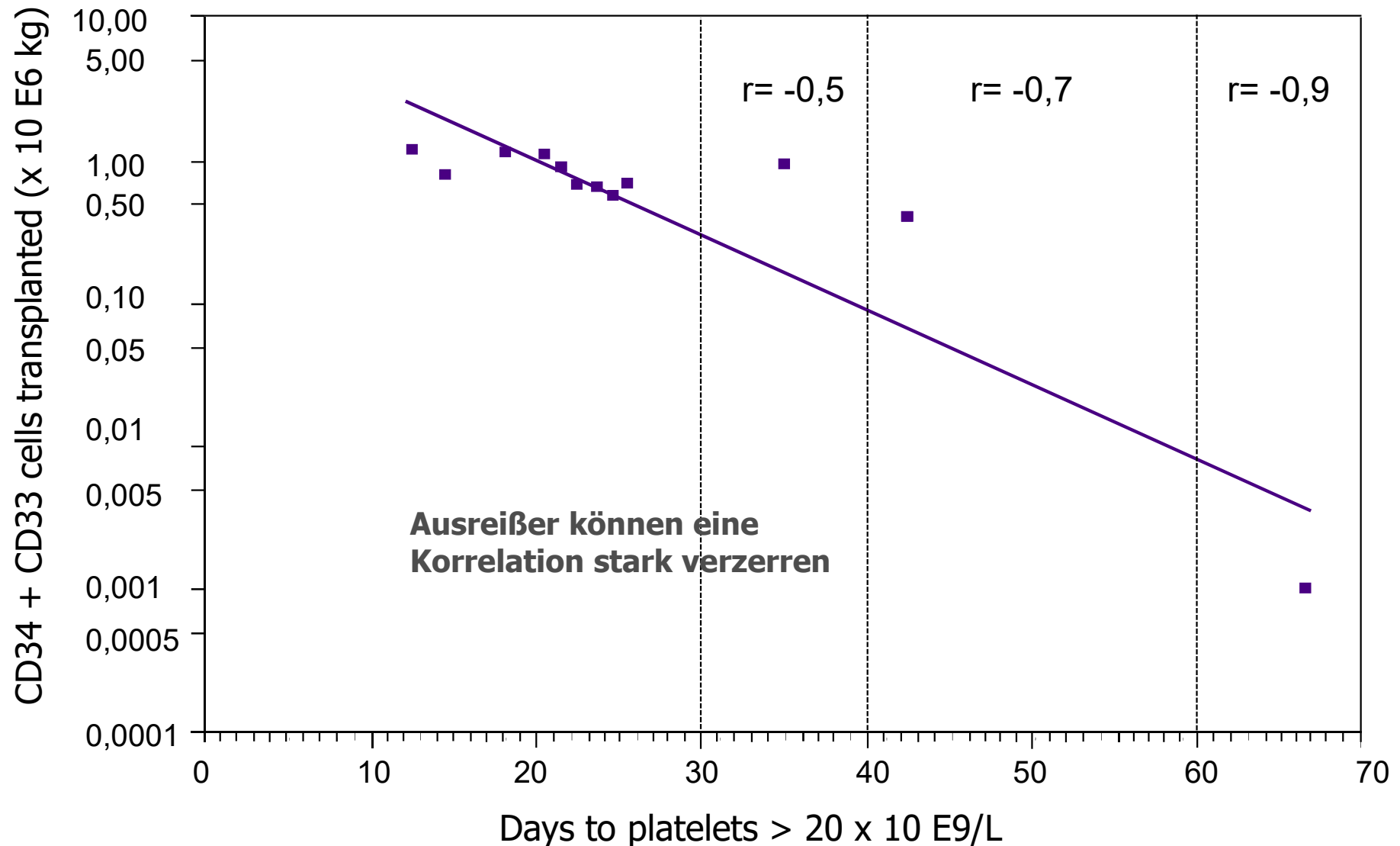


Body mass indes (BMI) berechnet nach :

$$BMI := \frac{Länge[m]}{Gewicht[kg]^2}$$

Correlation between platelet recovery and number of transplanted cells

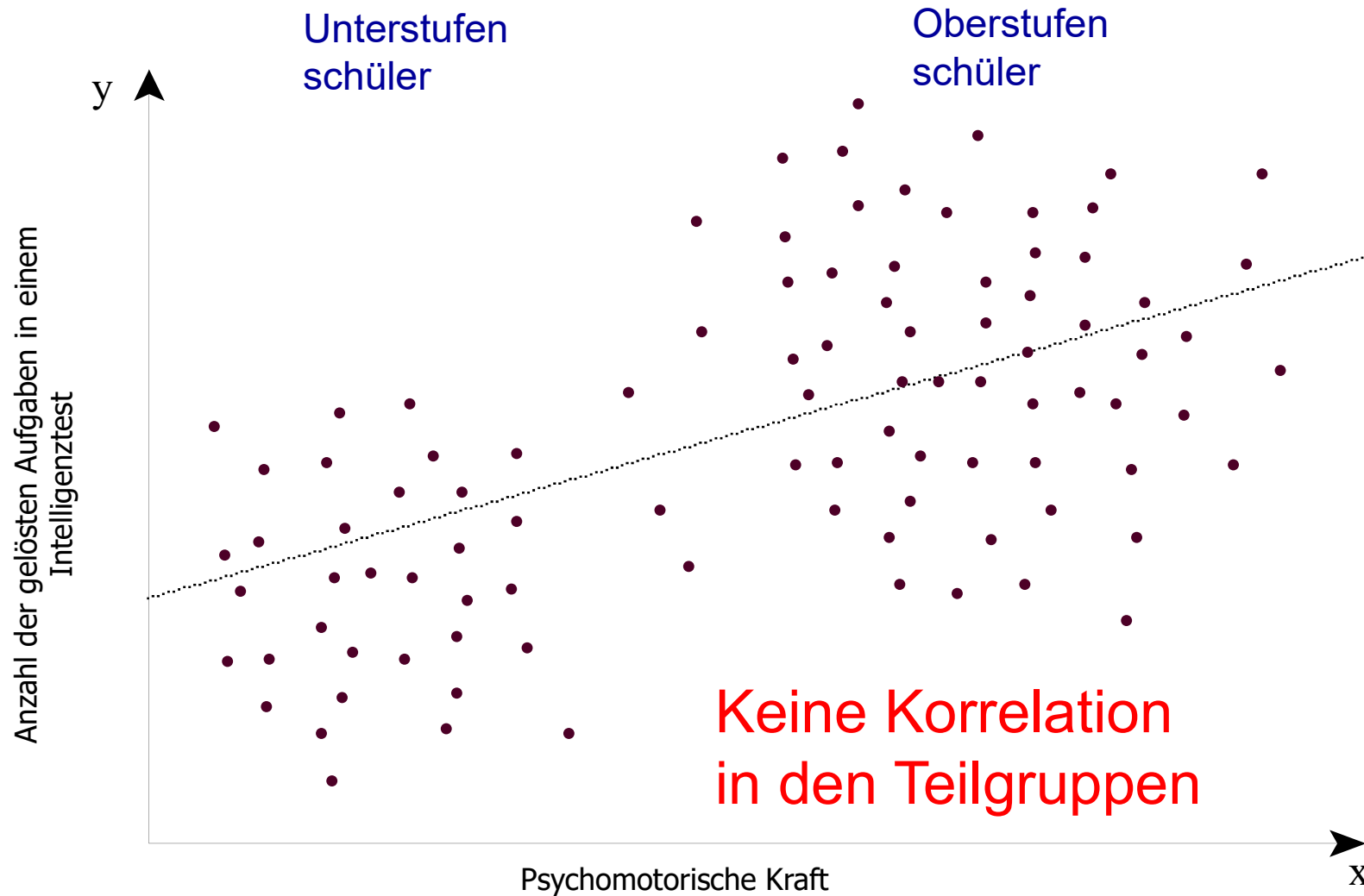
After peripheral blood stem cell transplantation



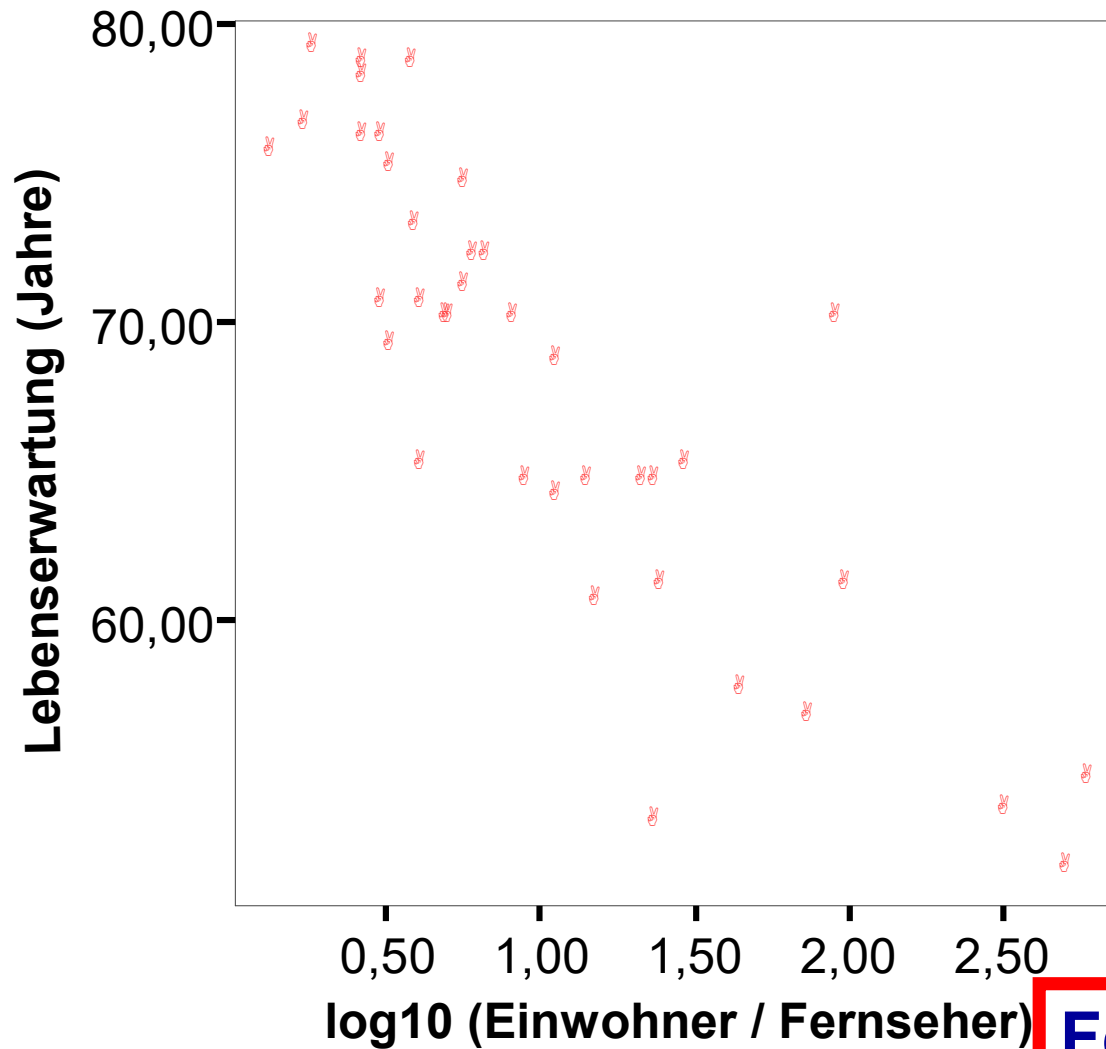
Korrelation durch Inhomogenitäten

- kann entstehen, wenn **heterogene Stichproben** zu einer Untersuchungspopulation zusammengefasst werden
- Korrelation innerhalb der einzelnen Stichproben kann anders sein als Korrelation innerhalb der gesamten Untersuchungspopulation
- Beispiel: Zufallsstichprobe von Schülern eines Gymnasiums
Punktwerte in einem Intelligenztest korrelieren mit psychomotorischer Kraft

Beispiel: Inhomogenitäten



Schein- oder Gemeinschaftskorrelation



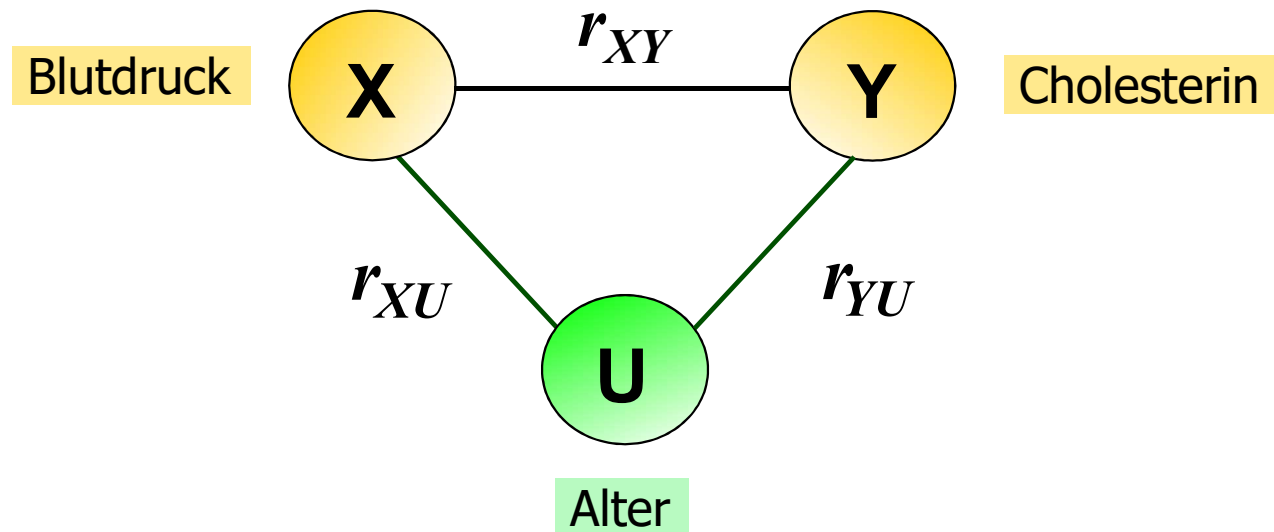
LAND	Lebenserwartung	Einwohner/TV
Argentinien	71	4
Bangladesh	54	315
Brasilien	65	4
Kanada	77	2
China	70	8
Kolumbien	71	6
Ägypten	61	15
Äthiopien	52	503
Frankreich	78	3
Deutschland	76	3
Indien	58	44
Indonesien	61	24
Iran	65	23
Italien	79	4
Japan	79	2
Kenya	61	96
Nord-Korea	70	90
Süd-Korea	70	5
Mexiko	72	7

...

Fernsehen verlängert das Leben!

Partielle Korrelation: Herausrechnen einer dritten Einflussgröße (Störgröße, confounding factor)

- Korrelation zwischen zwei Merkmalen unter Berücksichtigung (d.h. Adjustierung) eines dritten korrelierten Merkmals

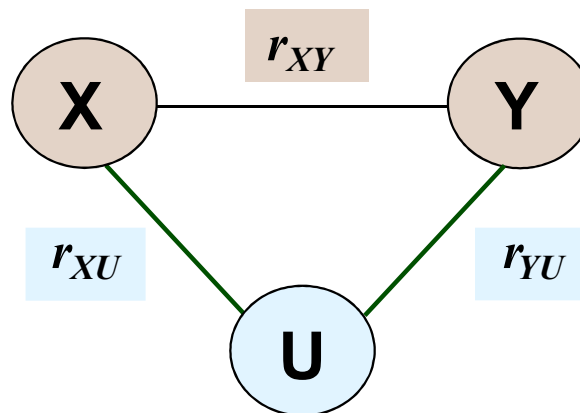


Partieller Korrelationskoeffizient

Statt $\text{cor}(X, Y)$ wird $\text{cor}(\text{residua}(X|U), \text{residua}(Y|U))$ berechnet.

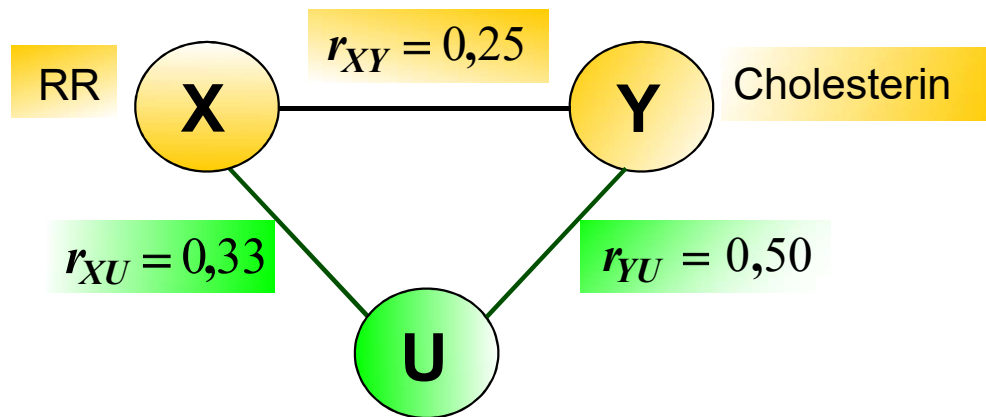
$$r_{XY/U} = \frac{r_{XY} - r_{XU} \cdot r_{YU}}{\sqrt{(1 - r_{XU})^2 \cdot (1 - r_{YU})^2}}$$

↑
AdjustiERG.

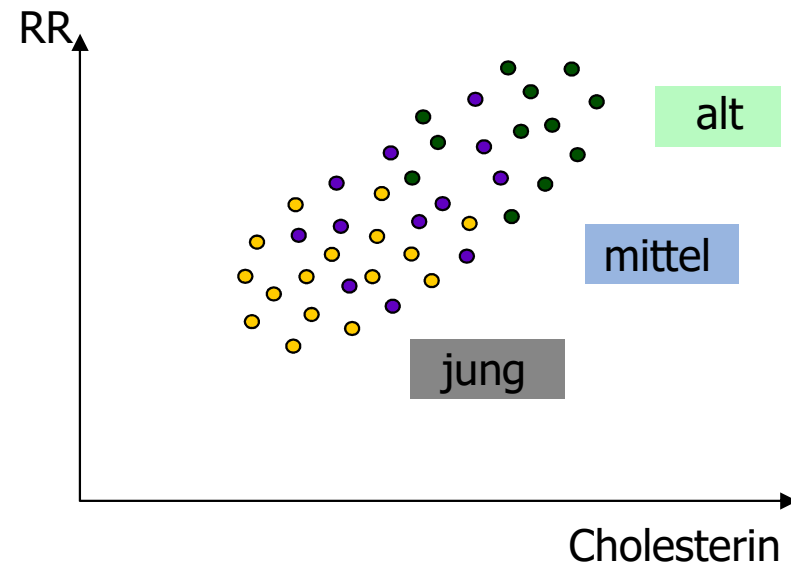


Partielle Korrelation

$n = 142 \text{ Frauen}$



Skizze der Zusammenhänge

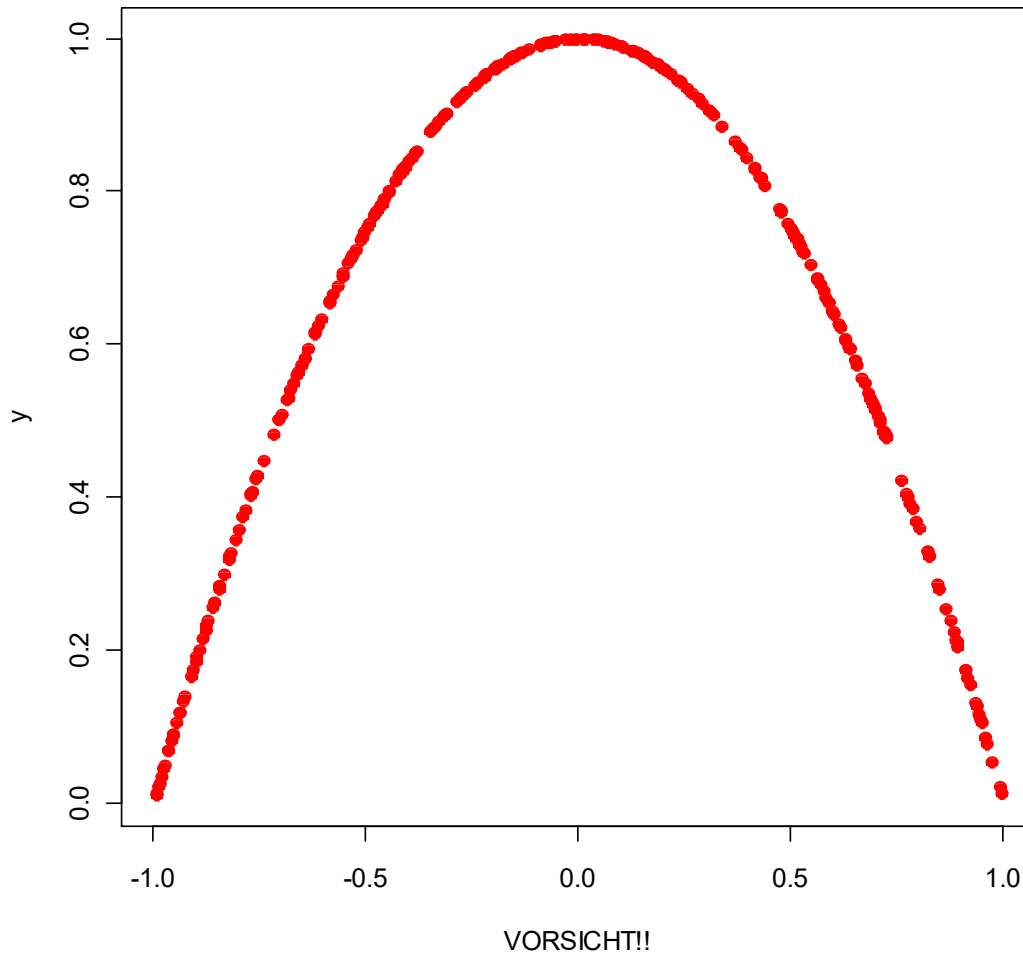


$$r_{XY/U} = \frac{0,25 - 0,33 \cdot 0,50}{\sqrt{(1 - 0,33)^2 \cdot (1 - 0,50)^2}} = \underline{\underline{0,1}}$$

(aus Hartung S. 562)

U-förmiger Zusammenhang

Keine lineare oder monotone Korrelation, aber...



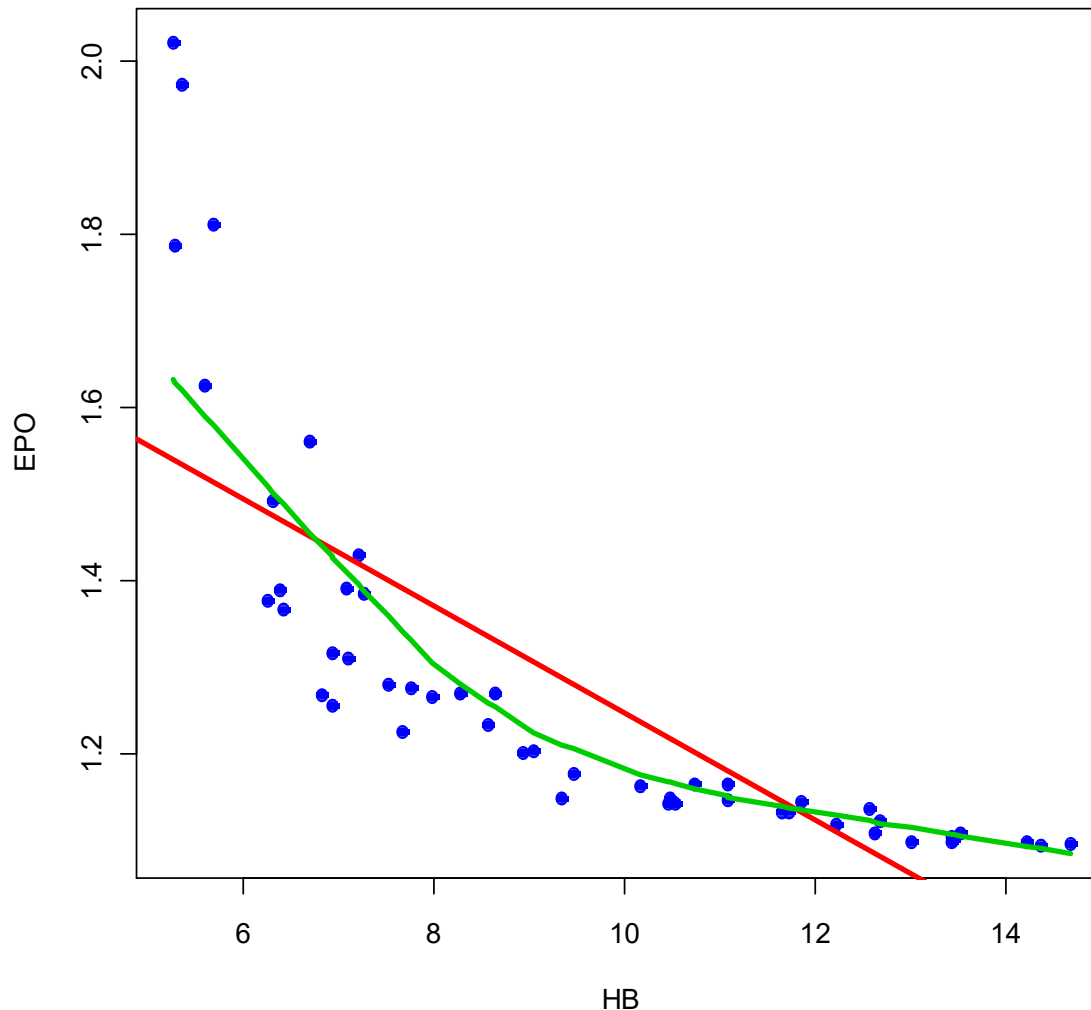
cor = 0.084
??????????

Beispiel für

Starke Abhängigkeit
aber $\text{cor} \approx 0$

Nicht linearer, aber monotoner Zusammenhang

Endogenes EPO bei Patienten mit Blutverlust



```
> cor(HB, EPO)
[1] -0.7837796
```

Korrelationskoeffizienten,
die monotonen Zusammenhang
messen:

```
> cor(HB, EPO, method="kendall")
[1] -0.8595918
```

```
> cor(HB, EPO, method="spearman")
[1] -0.9680192
```

Lowess glätter

Monotoner Zusammenhang

Kendall'scher Korrelationskoeffizient beruht auf bedingter Wahrscheinlichkeit

$$pr(Y_i > Y_k | X_i > X_k)$$

Spearman'scher Korrelationskoeffizient ersetzt X durch $\text{rank}(X)$ und Y durch $\text{rank}(Y)$ und berechnet damit den Pearson'schen Korrelationskoeffizient. (Robustifizierung)

Beide sind invariant unter monotonen Transformationen der Daten.