Das Zwei-Stichproben-Problem

(X_i, Y_i) zwei Merkmale am i-ten Merkmalsträger (i-ten Objekt) beobachtet.

n Stichprobenpaare $(X_1, Y_1), \ldots, (X_n, Y_n)$ mit entsprechenden
n Realisierungen $(x_1, y_1), \ldots, (x_n, y_n)$.

Zweistichprobenproblem:

Lokationsunterschied in der Verteilung mit Vorzeichen-Test oder Wilcoxon-Test.

Regressionsproblem:

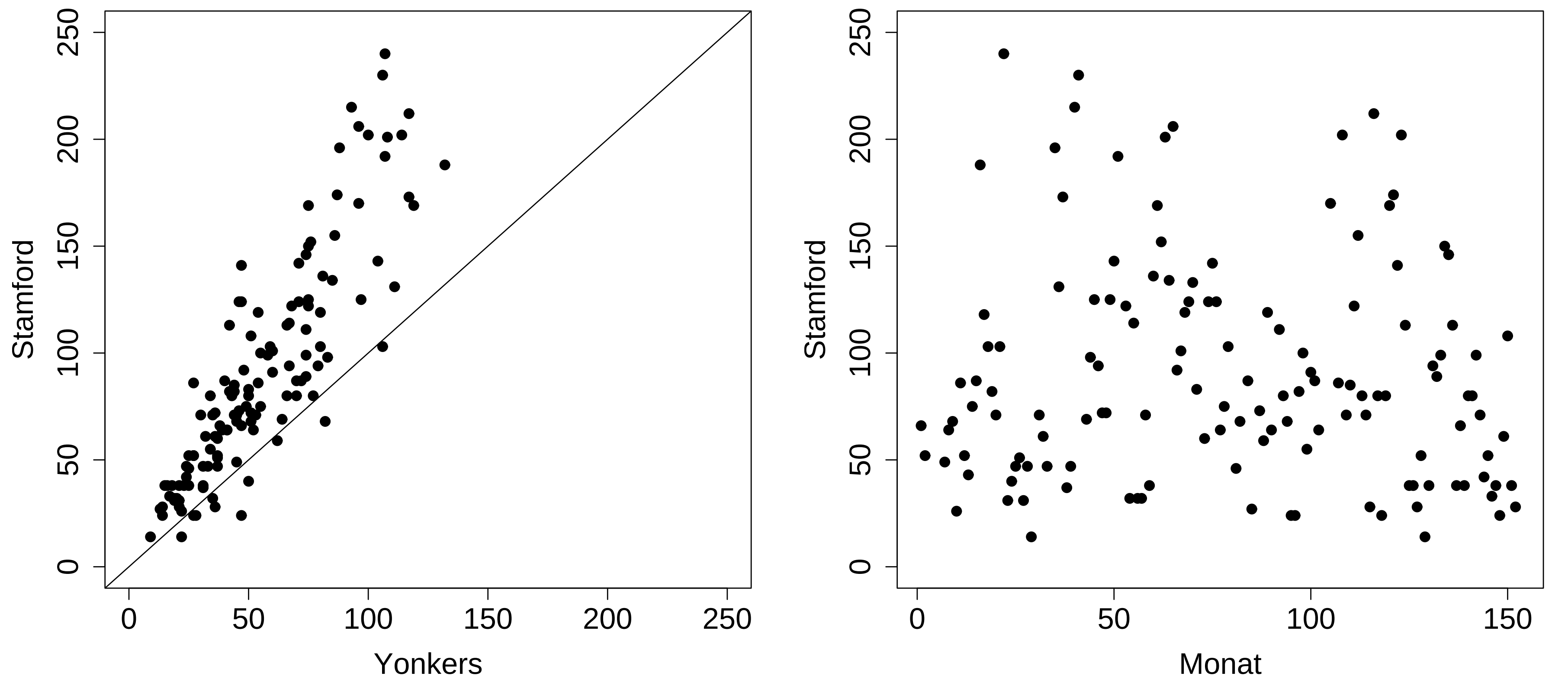
Mit Scatter-Plots funktionalen Zusammenhang erkennen
Glättungsmethoden anwenden, um funktionale Darstellung zu erhalten.

Korrelation und Kontingenz:

Korrelationsmaße, Korrelationstests, Test auf Unabhängigkeit

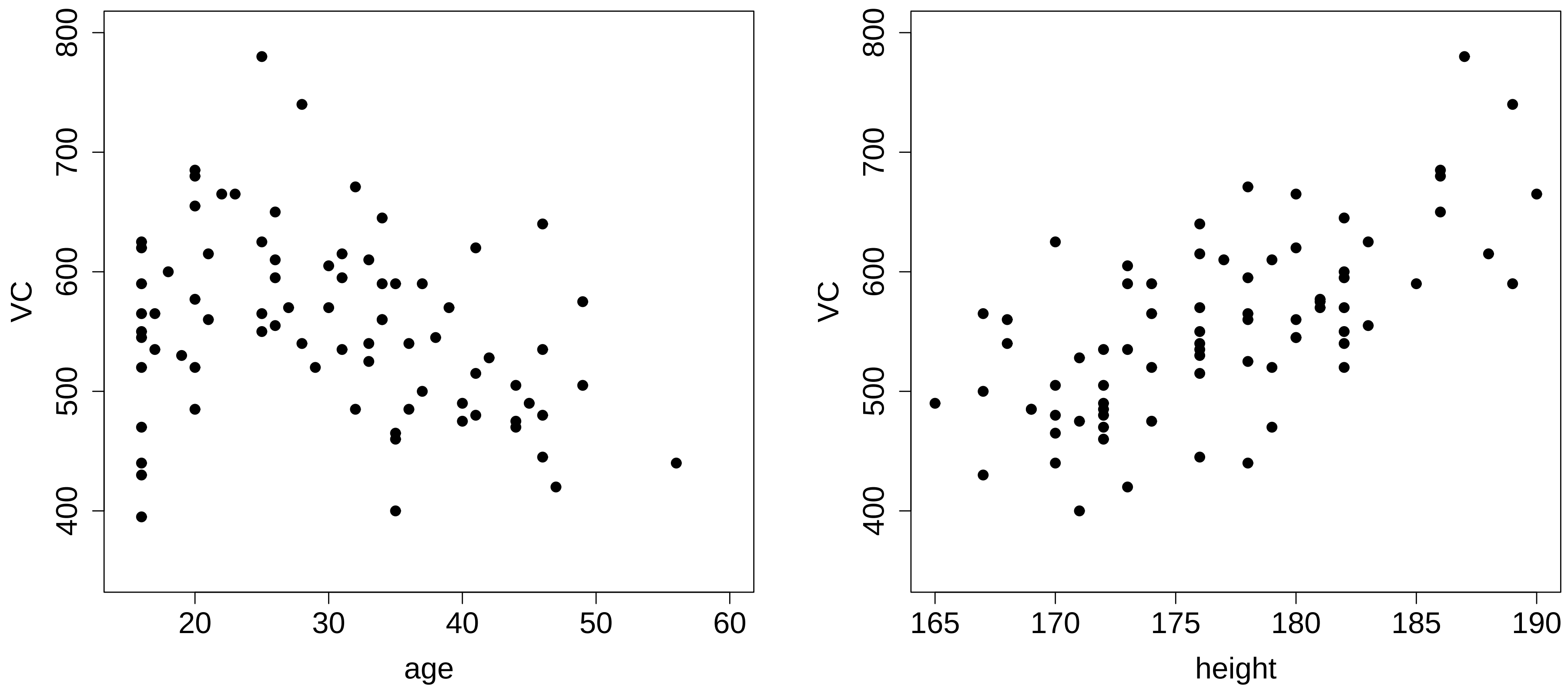
Graphische Verfahren

Bivariate Scatter-Plot Punkte (x_i, y_i).



Stamford vs. Yonkers Ozon-Daten (links) und Stamford Ozon-Daten gegen den Zeitverlauf (rechts).

```
> attach(aimu);  plot(age, VC);  plot(height, VC)
```



Scatter-Plot der `VC`-Daten gegen `age` (links) und `height` (rechts).

Regression

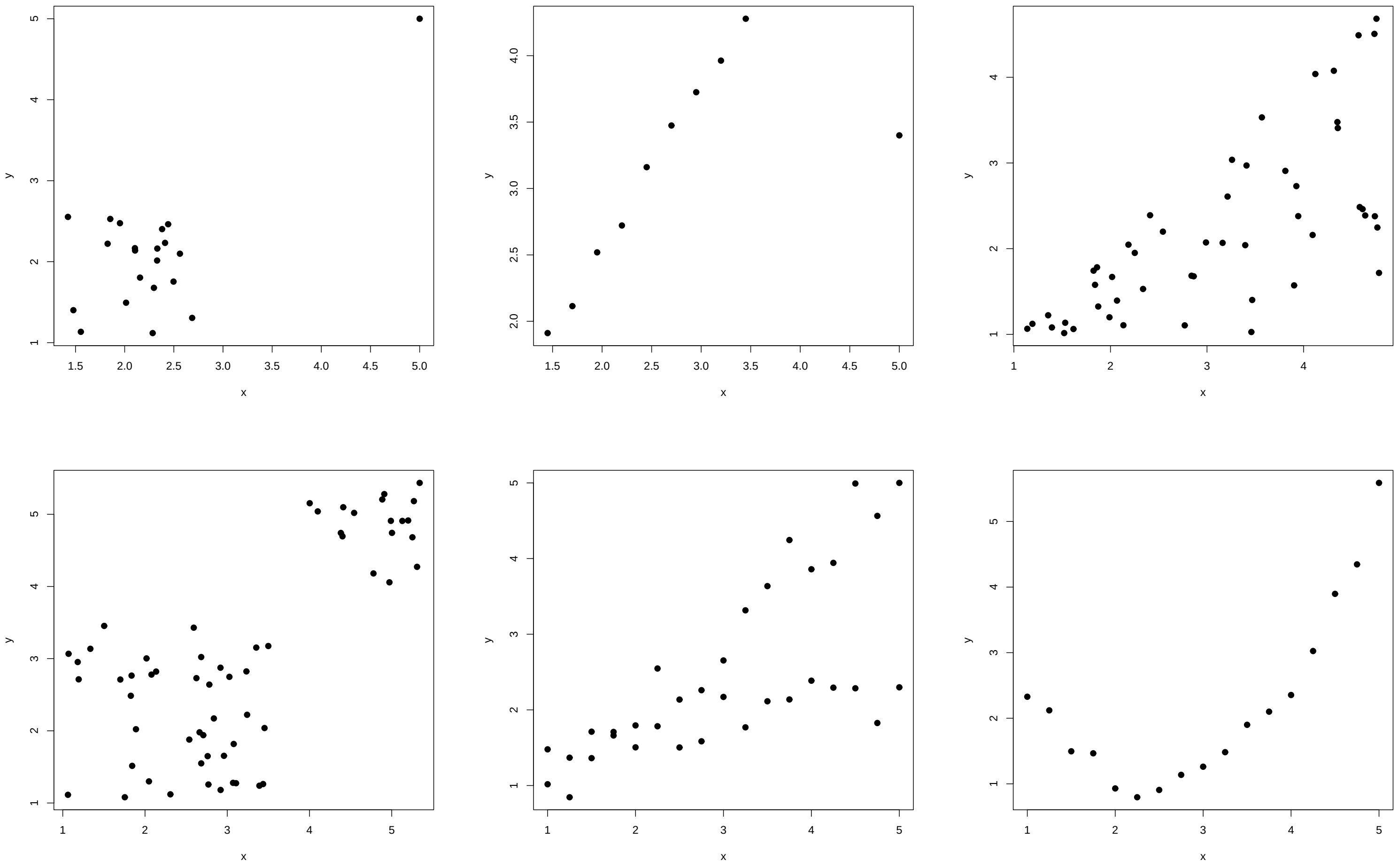
Merkmalspaare durch Scatterplot darstellen, um funktionalen Zusammenhang zu erkennen = **Idee der Regression**.

Empirische Korrelationskoeff: Maß für die **lineare** Abhängigkeit der Y_i von X_i

$$R = \frac{S_{xy}^2}{\sqrt{S_x^2 S_y^2}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

```
> cor(age, VC)
 [1] -0.2914085
> cor(height, VC)
[1] 0.6829789
```

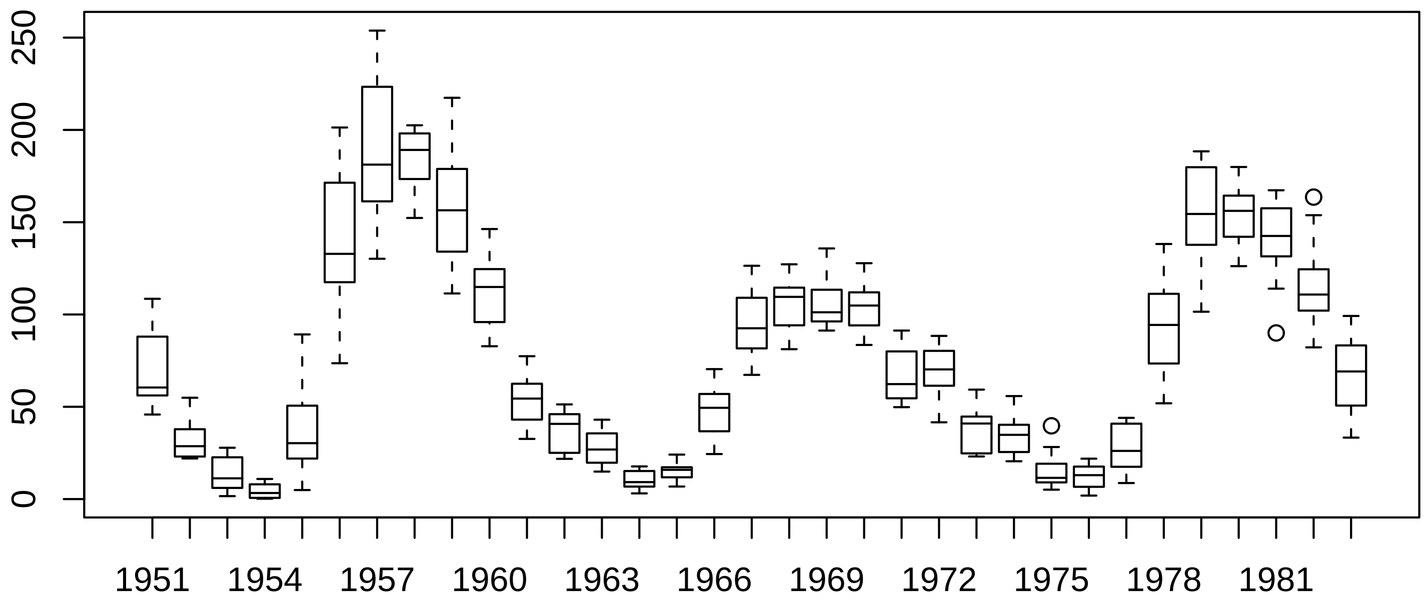
Interpretation von r:



Scatter-Plots von Stichproben mit $r = 0.7$

Analyse der Abhängigkeitsstruktur

(x_i, y_i), $i = 1, \ldots, n$, seien Realisationen von Zufallsvariablen X und Y.

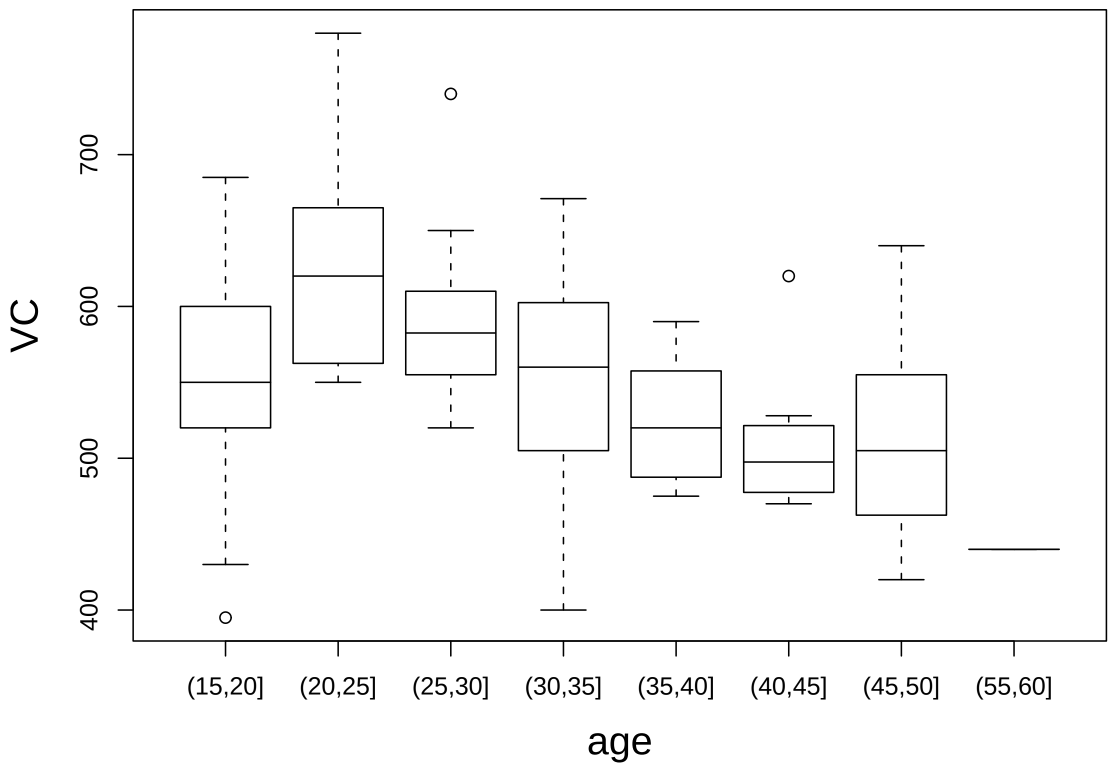


Box-Plot-Serie über die Aktivität der Sonne, 1951-1983.

```
> sunspots <- scan("sunspots.dat")
> year <- gl(235, 12, label=c(1749:1983))    # 12*1749, ..., 12*1983
> plot.it <- (as.numeric(year) >= 203)
> boxplot(sunspots[plot.it] ~ year[plot.it]) # use data from 1951-1983 only
```

Veränderung der empirischen lokalen Verteilung von Y in Abhängigkeit von X

```
> attach(aimu)
> age.class <- cut(age, breaks=seq(15,60,by=5))
> table(age.class)
 (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60]
      21       8      10      16       8       8       7       0       1
> boxplot(VC ~ age.class)
```



Box-Plots der VC-Werte in Abhängigkeit von der Altersgruppe

Lokal gewichtete Regression (LOWESS)

Idee: Mittels **Glättungsverfahren** zu jedem Datenpunkt (x_i, y_i) einen geglätteten Punkt $(x_i, \hat{y}_i)$ berechnen und zwischen x_i und x_{i+1} linear interpolieren

Resultat: geschätzte glatte Regressionskurve $\hat{y} = f(x)$

Zum Teil sehr ausgefeilte Methoden, viele theoretische Resultate, hier nur als numerisch-exploratives Instrument

lowess ist ein Glättungsverfahren. Steht für: 'Locally Weighted Regression Scatter Plot Smoothing'

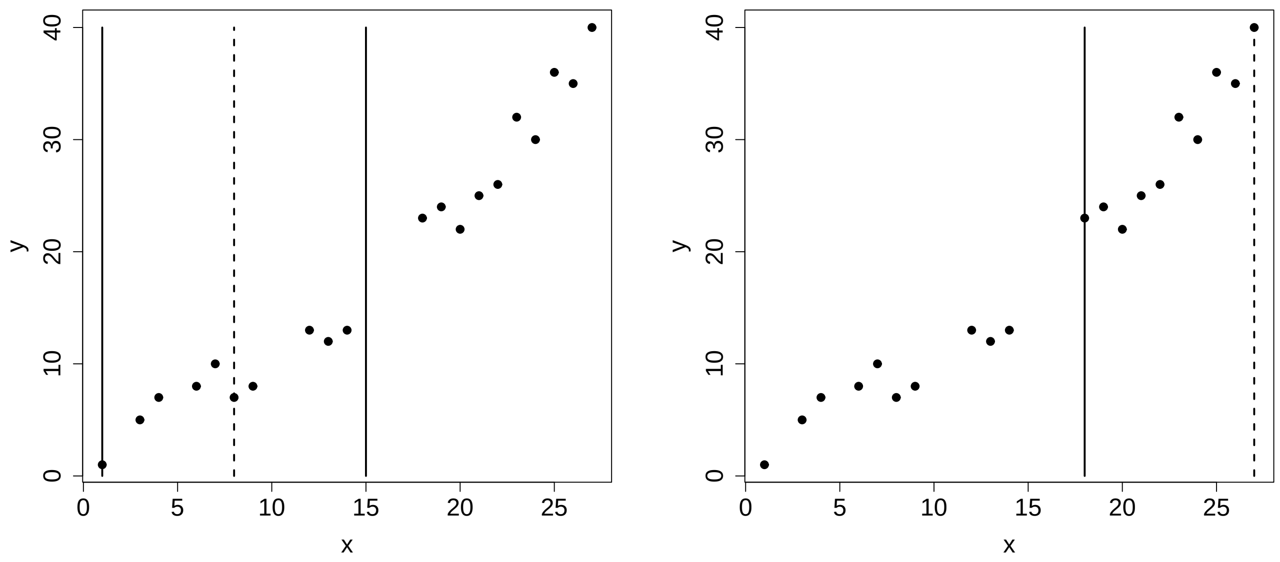
Verfahren besteht aus zwei Teilen:

- Glättung: lokale (in einem Fenster) wird Regression gerechnet.
- Robustifizierung: Da anfällig gegen Ausreißer wird noch 'robustifiziert'

Beschreibung des Verfahrens

Für jeden Datenpunkt (x_i, y_i), $i = 1, \ldots, n$, wird ein geglätteter Punkt $(x_i, \hat{y}_i)$ berechnet. Einzelne Schritte anhand von $n = 20$ künstlichen Punkten erläutert

1. Um (x_i, y_i) ein vertikaler Streifen, sodass die $q = \lceil fn \rceil$ bzgl. der x-Richtung benachbarten (inkl. (x_i, y_i)) Punkte enthalten sind (Parameter $f \in [1/3, 2/3]$)



Streifen um $x_6 = 8$ und $x_{20} = 27$ mit Zentrum (dashed), Rändern (solid)

2. Für alle Punkte im Streifen Nachbarschaftsgewichte definieren, sodass
• (x_i, y_i) größtes Gewicht
• Gewichte im Fenster nehmen mit Abstand zu x_i ab
• Gewichtsfunktion symmetrisch um x_i
• Außerhalb des Streifens Gewicht Null

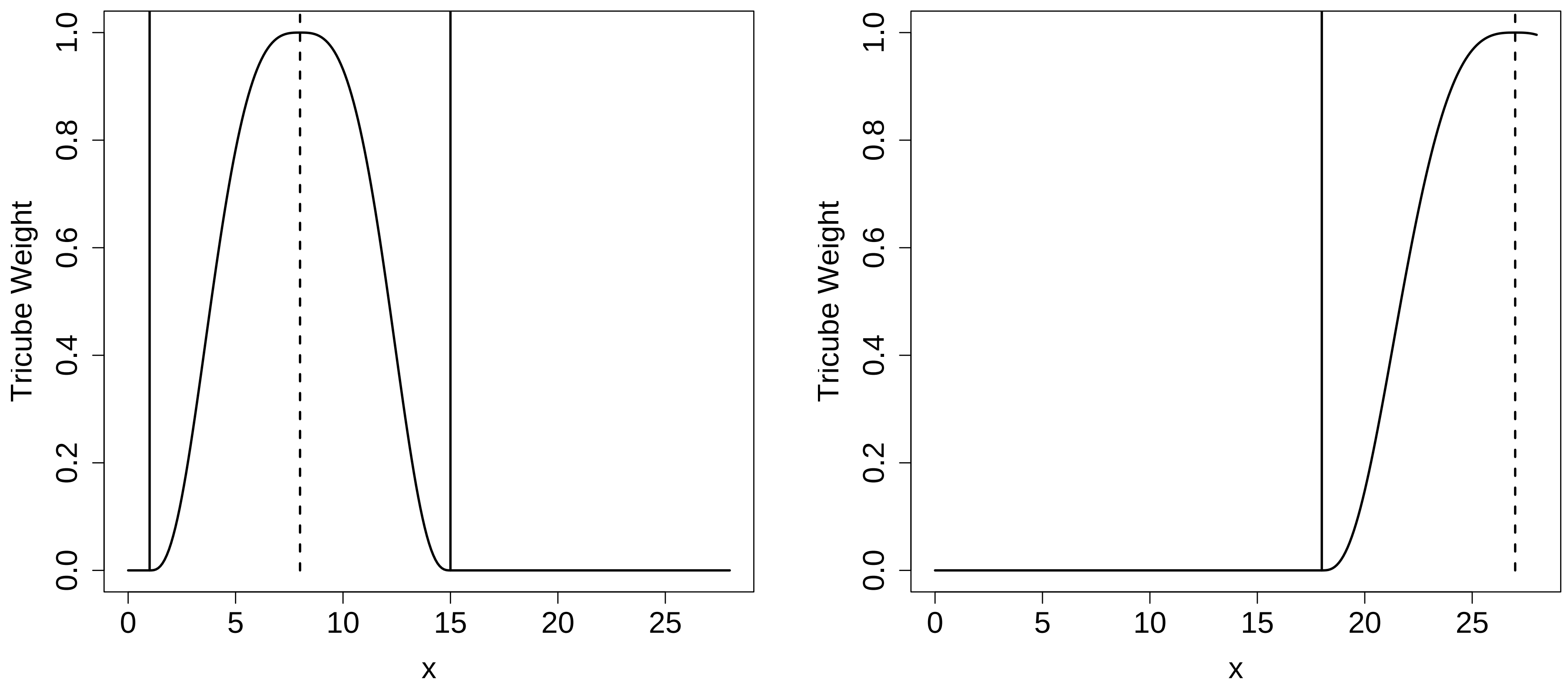
Z.B. mittels Tricube-Funktion

$$T(u) = \begin{cases} (1 - |u|^3)^3 & \text{für} \quad |u| < 1 \\ 0 & \text{sonst} \end{cases}$$

Gewicht definiert als

$$t_i(x_k) = T\left(\frac{x_i - x_k}{d_i}\right),$$

wobei d_i die Entfernung von x_i zum q–nächsten Nachbarn beschreibt.



Nachbarschaftsgewichte um $x_6 = 8$ (links) und $x_{20} = 27$ (rechts)

3. Geglätteter Wert $(x_i, \hat{y}_i)$ mit

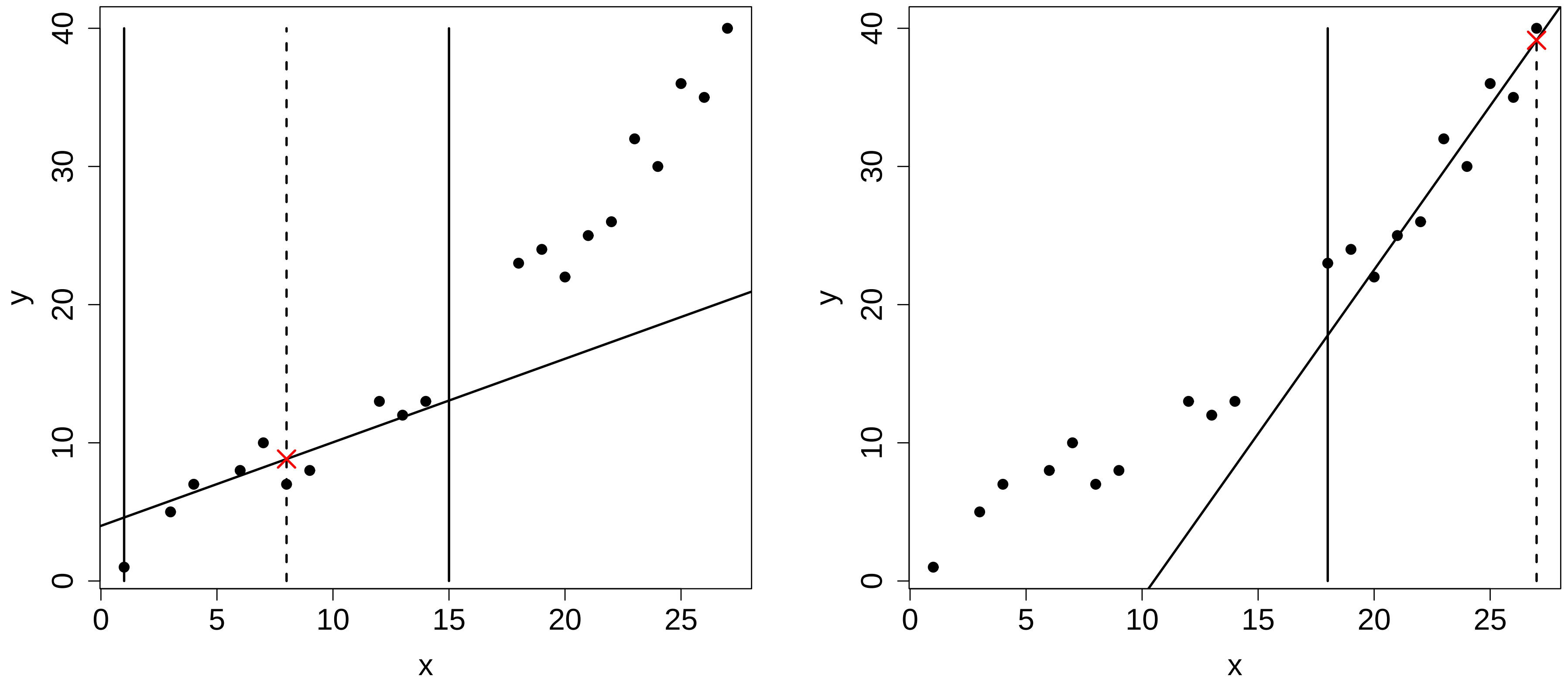
$$\hat{y}_i = \hat{a}_i + \hat{b}_i x_i$$

$(\hat{a}_i, \hat{b}_i)$ nach der Methode der gewichteten Kleinsten Quadrate bestimmen. D.h., $\hat{a}_i$, $\hat{b}_i$ minimieren die gewichtete Fehlerquadratsumme (*Sum of Squared Errors*)

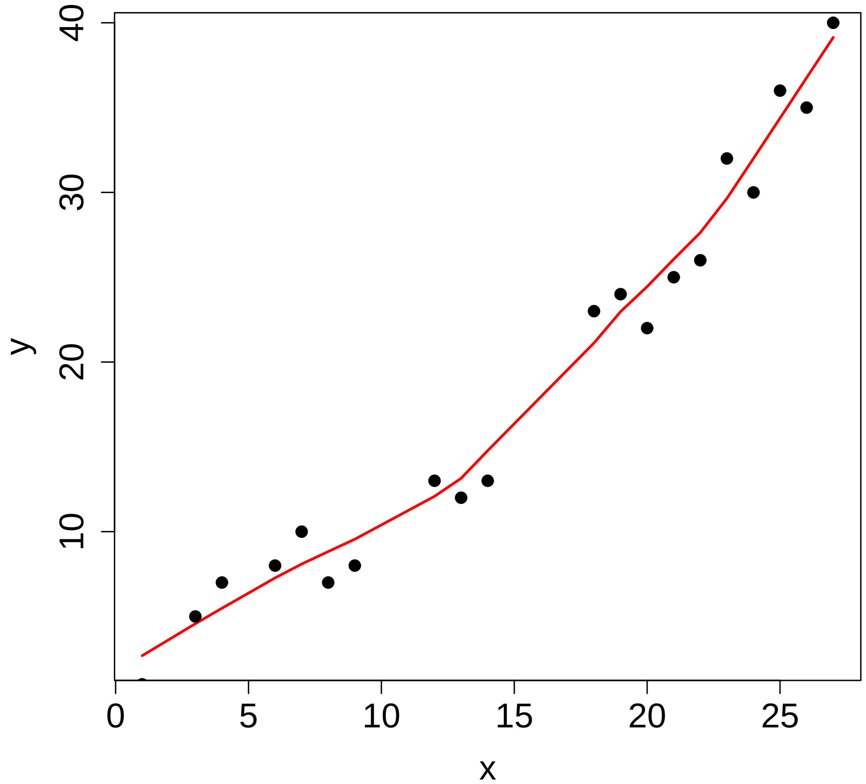
$$SSE_t(a_i, b_i) = \sum_{k=1}^{n} t_i(x_k)(y_k - a_i - b_i x_k)^2 \,.$$

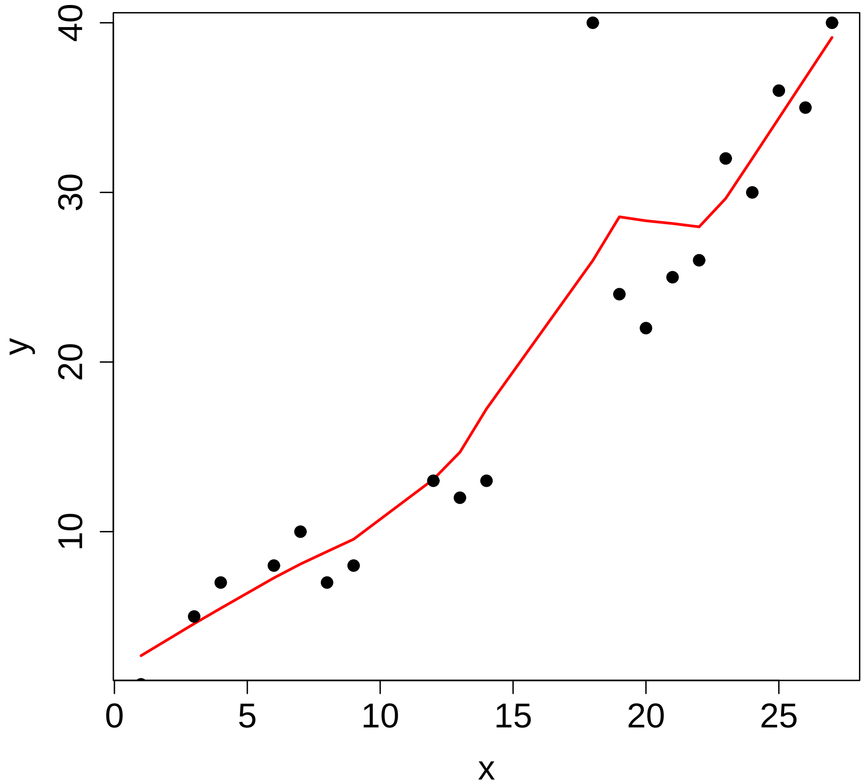
Benötigt werden die Lösungen der beiden Normalgleichungen

$$\begin{aligned} \sum_{k=1}^{n} t_i(x_k)(y_k - \hat{a}_i - \hat{b}_i x_k) &= 0\,, \\ \sum_{k=1}^{n} t_i(x_k)(y_k - \hat{a}_i - \hat{b}_i x_k) x_k &= 0 \end{aligned}$$



Ergebnisse der gewichteten linearen Regression in den Streifen um $x_6 = 8$ (links) und um $x_{20} = 27$ (rechts) mit geglätteten Punkten ($\times$).





`lowess`-Glättung der originalen Punkte (links) und von modif. Daten (rechts)

Verfahren ist anfällig gegenüber Ausreißer in y-Richtung!! (lokale) LS Problem!!

Daher **Robustifizierung:**

Residuen

$$r_i = y_i - \hat{y}_i$$

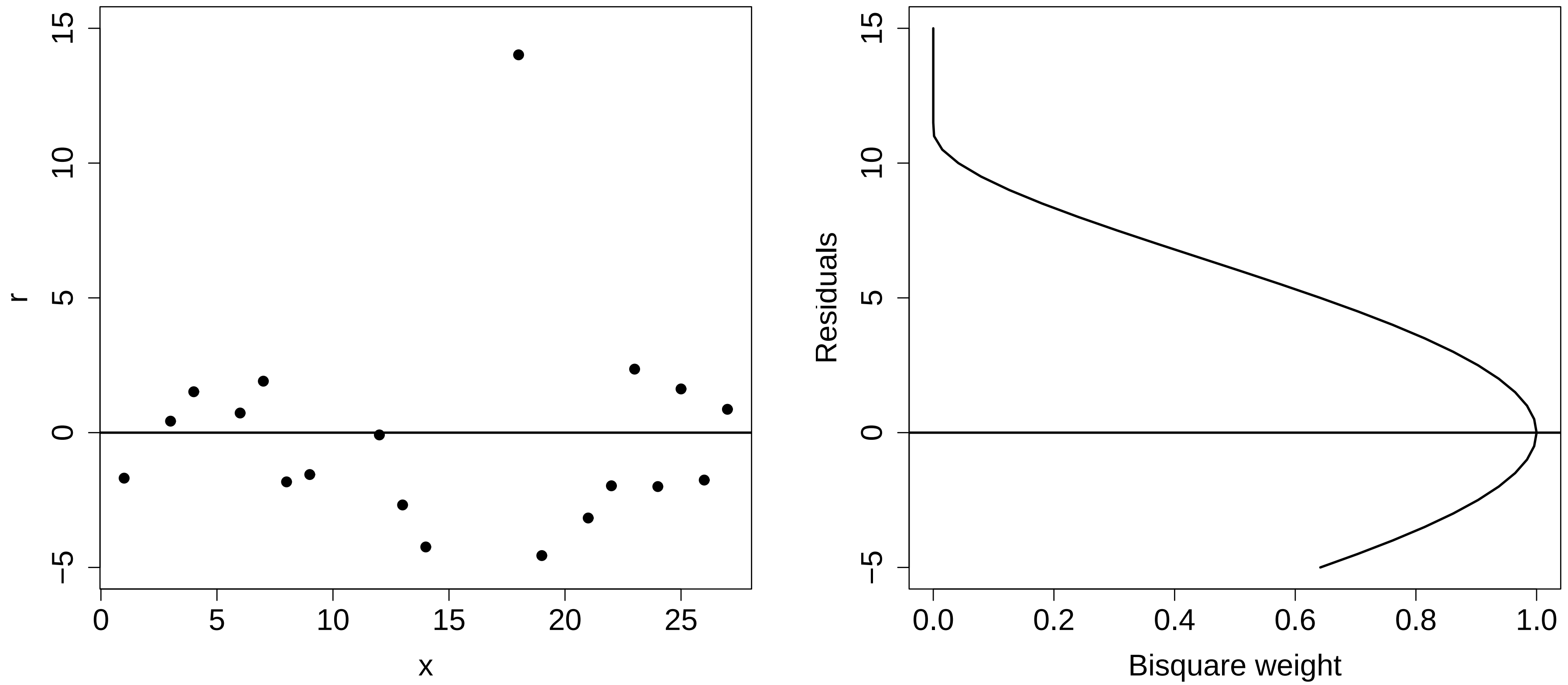
Bisquare-Funktion

$$B(u) = \begin{cases} (1-u^2)^2 & \text{für} \quad |u| < 1, \\ 0 & \text{sonst} \end{cases}$$

Robustheitsgewicht für (x_k, y_k) durch

$$w(x_k) = B\left(\frac{r_k}{6m}\right) \quad \text{mit} \quad m = med(|r|)$$

definiert - großes Residuum gibt kleines Gewicht.



Scatter-Plot der Residuen r_i gegen die x_i (links) und auf diese Residuen angewandte Gewichtung (rechts)

Bemerkung zur Skalierung durch $6m$*:*

Für $R \sim N(0, \sigma^2)$ folgt $|R| \sim H(\sigma^2)$.

Die Halbnormal-Verteilungsfunktion ist $2\Phi\left(|r|/\sigma\right) - 1$. Daher gilt für den Median m dieser Verteilung $2\Phi\left(m/\sigma\right) - 1 = 1/2$. Somit gilt $m/\sigma = z_{3/4}$ und es folgt $m = 0.675\sigma \approx 2/3\sigma$, also $6m \approx 4\sigma$.

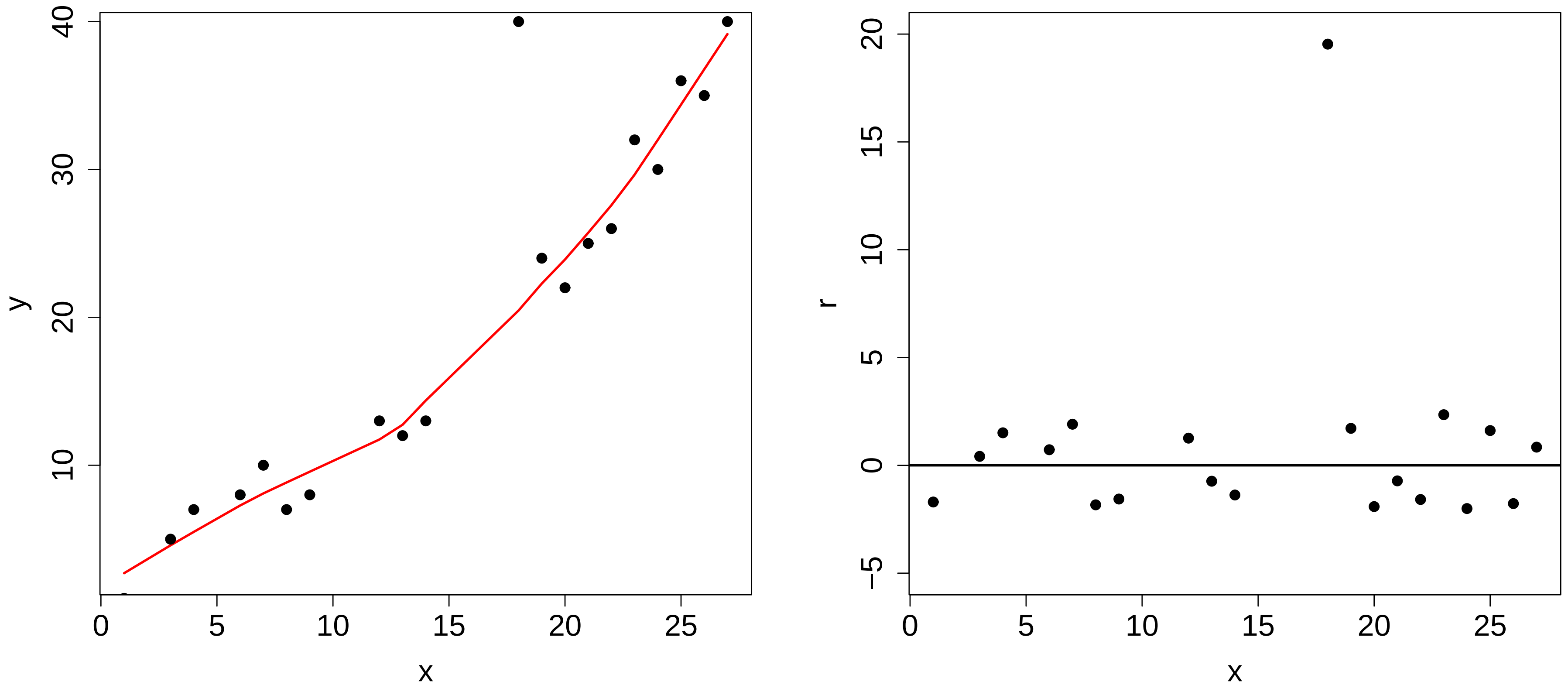
Zurück zu Schritt 2. Verwende Gewichte die sich aus dem Produkt der Nachbarschafts- mit den Robustheitsgewichten zusammensetzen. Bestimme damit Schätzer $\hat{a}_i^*$ und $\hat{b}_i^*$, die die gewichteten Fehlerquadratsummen

$$SSE_{tw}(a_i, b_i) = \sum_{k=1}^{n} w(x_k) t_i(x_k) (y_k - a_i - b_i x_k)^2$$

minimieren. Dadurch resultieren die geglätteten Werte

$$\hat{y}_i^* = \hat{a}_i^* + \hat{b}_i^* x_i$$

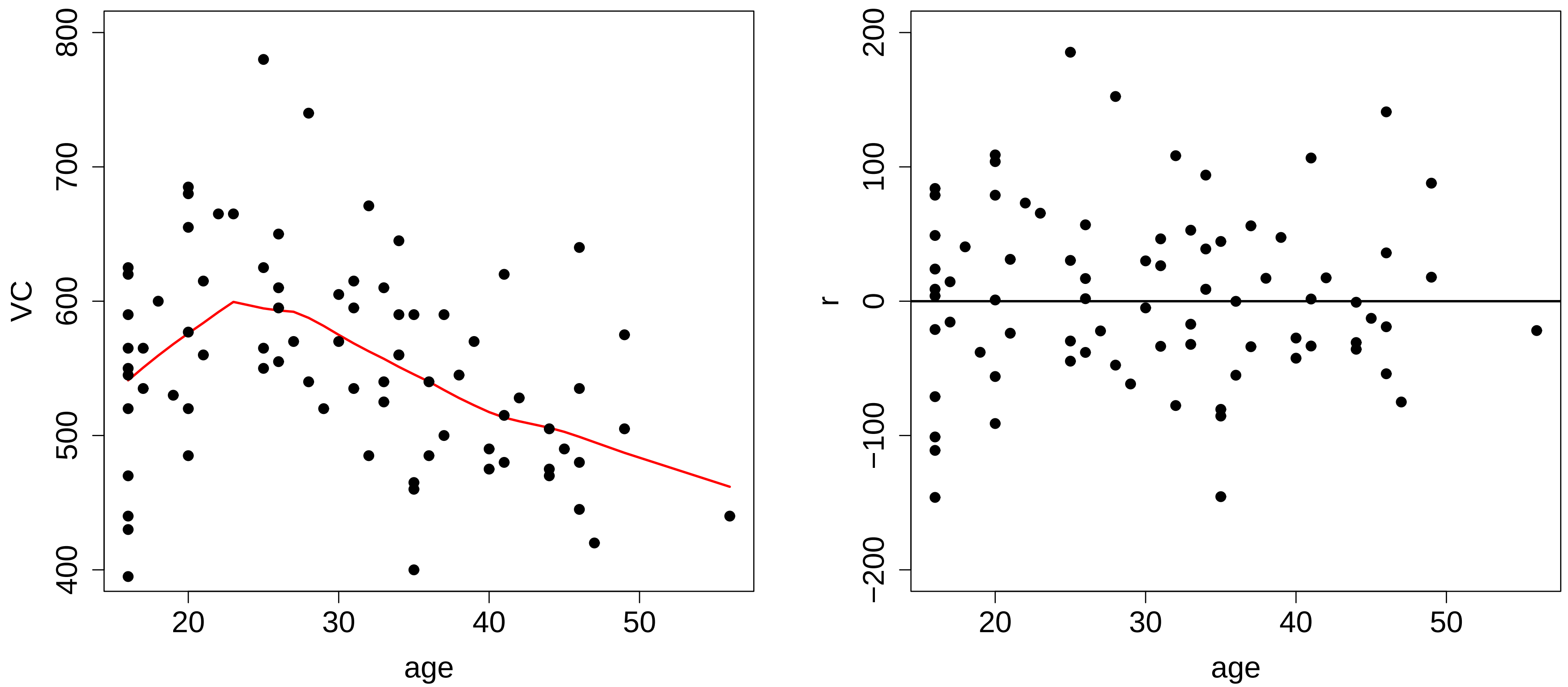
Diese Robustifizierung sollte zweimal durchgeführt werden.



Werte nach 1. Robustifizierung (links) mit Scatter-Plot der Residuen r_i^* (rechts).

```
> o <- order(age);  o.age <- age[o];  o.VC  <- VC[o]
> lowess.fit <- lowess(o.VC ~ o.age, f=1/2, iter=2)
> plot(lowess.fit);  points(age, VC)

> yhat <- lowess.fit$y;  r <- o.VC - yhat
> plot(o.age, r);  abline(h=0)
```



Ergebnis von `lowess` für `VC` in Abhängigkeit von `age` mit $f = 1/2$ und zweimaliger Robustifizierung (links) sowie entsprechende Residuen (rechts).

Lokationstests

Stichprobensituation: Zweidimens. Stichprobenvariablen $(X_1, Y_1), \ldots, (X_n, Y_n)$. Die Paare (X_i, Y_i) und (X_j, Y_j), mit $i \neq j$, sind stochastisch unabhängig.

Fragestellung: Sei $D_i = Y_i - X_i \overset{iid}{\sim} F$ mit $\mathsf{E}(D_i) = \mathsf{E}(Y_i) - \mathsf{E}(X_i) = \theta$. **Behandlungseffekt** mittels $H_0 : \theta = 0$ gegen $H_1 : \theta \neq 0$ zu falsifizieren.

Antwort: $\overline{D} = \overline{Y} - \overline{X}$ ist unverzerrt für θ. Untersuchung dieser Größe unter H_0.

Parametrischer Test bei Normalverteilung

Annahmen: $X_i \stackrel{iid}{\sim} N(\mu_X, \sigma_X^2)$, $Y_i \stackrel{iid}{\sim} N(\mu_Y, \sigma_Y^2)$, und $\text{cov}(X_i, Y_i) \neq 0$

$$D_i = Y_i - X_i \stackrel{iid}{\sim} N\left(\mu_Y - \mu_X, \sigma_X^2 + \sigma_Y^2 - 2\text{cov}(X_i, Y_i)\right)$$

und

$$\overline{D} = \overline{Y} - \overline{X} \sim N(\mu_D, \sigma_D^2/n)\,.$$

Testbare Hypothesen:

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$T < c_3$ oder $T > c_4$	$c_3 = t_{\alpha/2}$ $c_4 = t_{1-\alpha/2}$
$\mu_X \leq \mu_Y$	$\mu_X > \mu_Y$	$T < c_1$	$c_1 = t_\alpha$
$\mu_X \geq \mu_Y$	$\mu_X < \mu_Y$	$T > c_2$	$c_2 = t_{1-\alpha}$

$$T = \frac{\overline{D}}{S_D}\sqrt{n}\,, \qquad \overline{D} = \overline{Y} - \overline{X} \qquad \text{und} \quad S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2}\,.$$

Unter H_0 gilt offensichtlich $T \sim t_{n-1}$.

Beispiel: An $n = 10$ PKW's wird die Leistung 2er Kraftstoffe A und B getestet. Dabei ergaben sich die Fahrleistungen in km:

PKW	1	2	3	4	5	6	7	8	9	10
A	89	110	105	101	90	92	104	100	101	98
B	95	109	111	110	91	95	106	99	104	101
d_i	6	-1	6	9	1	3	2	-1	3	3

```
> A <- c(89, 110, 105, 101, 90, 92, 104, 100, 101,  98)
> B <- c(95, 109, 111, 110, 91, 95, 106,  99, 104, 101)
> t.test(A, B, paired = TRUE)
        Paired t-test
 data:  A and B
 t = -3.0846, df = 9, p-value = 0.01304
 alternative hypothesis: true difference in means is not equal to 0
 95 percent confidence interval:
  -5.3734398 -0.8265602
 sample estimates:  mean of the differences
                    -3.1
```

Wir vermuten, dass B besser als A ist und testen $H_0 : A$ ist besser als B.

```
> t.test(A, B, alt="less", paired = TRUE)

        Paired t-test

data:  A and B
t = -3.0846, df = 9, p-value = 0.006521
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.257744
sample estimates:
mean of the differences
                  -3.1
```

Dieser einseitige p-Wert entspricht dem halben p-Wert des zweiseitigen Test.

Eigentlich vermuten wir, dass Treibstoff B im Mittel um mindestens 2 Liter besser als A ist. Die entsprechende Nullhypothese kann jetzt nicht mehr verworfen werden.

```
> t.test(A, B, alt="less", paired = TRUE, mu=-2)

        Paired t-test

 data:  A and B
 t = -1.0945, df = 9, p-value = 0.1511
 alternative hypothesis: true difference in means is less than -2
```

Vorzeichentest

Teststatistik: Anzahl der Differenzen $Y_i - X_i$ mit positiven Vorzeichen

Annahmen: Differenzen $D_i = Y_i - X_i$ sind iid Variablen mit $P(X_i = Y_i) = 0$.

Hypothesen:
- Test A: $H_0 : P(X < Y) = P(X > Y), \quad H_1 : P(X < Y) \neq P(X > Y)$,
- Test B: $H_0 : P(X < Y) \leq P(X > Y), \quad H_1 : P(X < Y) > P(X > Y)$,
- Test C: $H_0 : P(X < Y) \geq P(X > Y), \quad H_1 : P(X < Y) < P(X > Y)$.

Teststatistik

$$T = \sum_{i=1}^{n} Z_i \quad \text{mit} \quad Z_i = \begin{cases} 1 & \text{falls} \quad X_i < Y_i\,, \\ 0 & \text{falls} \quad X_i > Y_i\,, \end{cases}$$

unter H_0 gilt $T \sim \text{Binomial}(n, 1/2)$.

Entscheidungsregel: Lehne H_0 ab, falls

- Test A: $t \leq t_{\alpha/2}$ oder $t \geq n - t_{\alpha/2}$,
- Test B: $t \geq n - t_{\alpha}$,
- Test C: $t \leq t_{\alpha}$,

wobei t_α das α-Quantil der Binomialverteiltung ist.

Beispiel: Treibstoff A und B wird auf unterschiedliche Fahrleistung getestet.

PKW	1	2	3	4	5	6	7	8	9	10
A	89	110	105	101	90	92	104	100	101	98
B	95	109	111	110	91	95	106	99	104	101
d_i	6	-1	6	9	1	3	2	-1	3	3
z_i	1	0	1	1	1	1	1	0	1	1

$t = 8$ beobachtet. Für $\alpha = 0.055$ ist $t_\alpha = 2$, damit $t_{1-\alpha} = n - t_\alpha = 10 - 2 = 8$.
H_0: 'Treibstoff A ist im Mittel besser als B' wird gerade noch abgelehnt (Test B).

```
> binom.test(sum(A>B), length(A), p=1/2, alt="two.sided")
        Exact binomial test
 data:  sum(A > B) and length(A)
 number of successes = 2, number of trials = 10, p-value = 0.1094
 alternative hypothesis: true probability of success is not equal to 0.5
 95 percent confidence interval:
  0.02521073 0.55609546
 sample estimates:
 probability of success
                  0.2
> binom.test(sum(A>B), length(A), p=1/2, alt="less")
        Exact binomial test
 data:  sum(A > B) and length(A)
 number of successes = 2, number of trials = 10, p-value = 0.05469
 alternative hypothesis: true probability of success is less than 0.5
 95 percent confidence interval:
  0.0000000 0.5069013
 probability of success
                  0.2
```

Vorzeichentest als Test auf Quantile

Bei kardinalem Meßniveau auch zur Prüfung der Hypothese

$$H_0' : \text{'Der Median von } Y - X \text{ ist } m_0\text{'}$$

verwendbar.

Die Teststatistik ist dann

$$T' = \sum_{i=1}^{n} Z_i' \quad \text{mit} \quad Z_i' = \begin{cases} 1 & \text{falls} \quad m_0 < Y_i - X_i, \\ 0 & \text{falls} \quad m_0 > Y_i - X_i. \end{cases}$$

Wilcoxon-Test

Entspricht dem Wilcoxon-Vorzeichen-Rang-Test für den Median beim Einstichproben-Problem mit $D_i = Y_i - X_i \overset{iid}{\sim} F$, F stetig und symmetrisch um den Median der Differenzen m_D verteilt.

Hypothesen:
- Test A: $H_0 : m_D = 0; \quad H_1 : m_D \neq 0$,
- Test B: $H_0 : m_D = 0; \quad H_1 : m_D > 0$,
- Test C: $H_0 : m_D = 0; \quad H_1 : m_D < 0$.

Teststatistik:

$$W^+ = \sum_{i=1}^{n} Z_i R(|D_i|) \quad \text{mit} \quad Z_i = \begin{cases} 1 & \text{für} \quad D_i > 0, \\ 0 & \text{für} \quad D_i < 0, \end{cases}$$

wobei $R(|D_i|)$ den Rang von $|D_i|$ beschreibt.

Entscheidungsregel:
- Test A: $w^+ \leq w_{\alpha/2}$ oder $w^+ \geq w_{1-\alpha/2}$,
- Test B: $w^+ \geq w_{1-\alpha}$,
- Test C: $w^+ \leq w_{\alpha}$.

Beispiel: Kraftstoffarten A und B.

PKW	1	2	3	4	5	6	7	8	9	10
A	89	110	105	101	90	92	104	100	101	98
B	95	109	111	110	91	95	106	99	104	101
d_i	6	-1	6	9	1	3	2	-1	3	3
$r(\|d_i\|)$	8.5	2	8.5	10	2	6	4	2	6	6

Ränge der $|d_i|$: wegen der Bindungen der Realisationen 1, 3 und 6 wurden Durchschnittsränge verwendet.

$w^+ = 51$ beobachtet, für $\alpha = 0.05$ ist $w_{\alpha/2} = w_{0.025} = 8$ und $w_{1-\alpha/2} = w_{0.975} = n(n+1)/2 - w_{0.025} = 47$, d.h. für Test A wird H_0 wegen $51 > 47$ abgelehnt.

```
> wilcox.test(A, B, paired = TRUE, alt="less")

        Wilcoxon signed rank test with continuity correction

data:  A and B
V = 4, p-value = 0.009182
alternative hypothesis: true mu is less than 0
```

Beachte!!

Hypothese $H_0 : m_D = 0$ ist nicht zur Hypothese $H_0 : m_X = m_Y$ äquivalent.

Falls X bzw. Y symmetrisch um ihre Mediane m_X bzw. m_Y verteilt sind, dann sind folgende Aussagen äquivalent.

- $m_D = 0$,
- $m_X = m_Y$,
- $\mathsf{E}(X) = \mathsf{E}(Y)$.

Korrelation und Unabhängigkeit

'Korrelation' und 'Kontingenz' beschreiben Zusammenhang von Merkmalen.

Kontingenz bei nominalen, Korrelation bei zumindest ordinal skalierten Daten.

Korrelationskoeffizient (bei zumindest ordinal skalierten Zufallsvariablen)

$$\rho(X,Y) = \frac{\mathsf{cov}(X,Y)}{\sqrt{\mathsf{var}(X)\mathsf{var}(Y)}}$$

beschreibt den Grad der linearen Abhängigkeit der Merkmale X und Y.

Für kategoriale Daten, z.B. Vierfeldertafel,

	W	M
B	p_{11}	p_{12}
$\tilde{B}$	p_{21}	p_{22}
	$p_{\bullet 1}$	$p_{\bullet 2}$

odds(B,W) $=$ Chance Brillenträger zu sein falls weiblich

odds(B,M) $=$ Chance Brillenträger zu sein falls männlich

Falls Chancen für W und M gleich, so ist B unabhängig vom Geschlecht

Odds Ratio

$$\theta = \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)}.$$

Falls X, Y unabhängig gilt: $P(X,Y) = P(X)P(Y)$ und somit $\theta = 1$.

Eigenschaften von ρ

1. $-1 \leq \rho \leq +1$

2. $|\rho| = 1 \Leftrightarrow P(Y = aX + b) = 1$, wobei $a \neq 0$ und b Konstanten sind

3. X, Y stochastisch unabhängig $\Rightarrow \rho = 0$

4. Sind $(X, Y) \sim N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ und ist $\rho = 0$, so gilt: X und Y sind stochastisch unabhängig (Umkehrung von 3. nur bei Normalverteilung).

Bivariate Normalverteilung

Definition: Die Dichte der bivariaten Normalverteilung $N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ ist

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right).$$

Testprobleme:

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\rho = 0$	$\rho \neq 0$	$T < c_3$ oder $T > c_4$	$c_3 = t_{\alpha/2}$ $c_4 = t_{1-\alpha/2}$
$\rho = 0$	$\rho > 0$	$T > c_1$	$c_1 = t_{1-\alpha}$
$\rho = 0$	$\rho < 0$	$T < c_2$	$c_2 = t_\alpha$

Teststatistik (unter H_0):

$$T = R\sqrt{\frac{n-2}{1-R^2}} \sim t_{n-2}$$

mit

$$R = \frac{S_{XY}^2}{\sqrt{S_X^2 S_Y^2}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Falls (X_i, Y_i) unabhängig aus bivariater Normalverteilung, so ist R der Maximum-Likelihood-Schätzer von ρ.

Konfidenzintervalle

Konfidenzintervalls für ρ: Verteilung von R (unter Alternative) nötig.

Falls $\rho = 0$: $T \sim t_{n-2}$. Für $\rho \neq 0$: R transformieren;

Für die Fisher Z-Transformierte gilt asymptotisch

$$Z = \frac{1}{2} \log \frac{1+R}{1-R} \overset{as}{\sim} N\left(\frac{1}{2} \log \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right).$$

Mit $\mu = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$ bzw. $\rho = \frac{e^{\mu}-e^{-\mu}}{e^{\mu}+e^{-\mu}} = \tanh(\mu)$.

$$U = \sqrt{n-3}(Z-\mu) \overset{as}{\sim} N(0,1).$$

Damit

$$P\left(z_{\alpha/2} \le U \le z_{1-\alpha/2}\right) = P\left(-\frac{z_{1-\alpha/2}}{\sqrt{n-3}} \le Z - \mu \le \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right) = 1 - \alpha.$$

Setze $a = \frac{z_{1-\alpha/2}}{\sqrt{n-3}}$, wegen der strengen Monotonie von $\tanh$ ergibt sich

$$P\left(\tanh(Z-a) \le \tanh(\mu) = \rho \le \tanh(Z+a)\right) = 1 - \alpha.$$

Beispiel: $n = 9$ Arbeiter bewerben sich für eine freie Stelle.

2 Kommissionen A und B testen die Bewerber und vergeben Punkte.

Wie groß ist das Maß der Übereinstimmung im Urteil der beiden Kommissionen? (Annahme einer bivariaten Normalverteilung)

Bewerber i	1	2	3	4	5	6	7	8	9
$x_i \ldots$ Punkte von A	75	62	87	76	73	66	81	74	77
$y_i \ldots$ Punkte von B	82	69	89	84	80	68	79	70	74

$\overline{x} = 74.56$, $s_X^2 = 54.78$, $\overline{y} = 77.22$, $s_Y^2 = 54.19$ und $s_{XY}^2 = 42.99$. $r = 0.789$.
$t_{7,1-0.05} = 1.89$. $t = 0.789\sqrt{\frac{7}{1-0.789^2}} = 3.40 > 1.89$, also H_0 verwerfen.

Zweiseitiges 95% Konfidenzintervall für ρ: $z_{0.975} = 1.96$ somit $a = \frac{1.96}{\sqrt{6}} = 0.80$.
Transformierte $z = \frac{1}{2}\log\frac{1.789}{0.211} = 1.069$. Somit $\tanh(z-a) = \tanh(0.269) = 0.26$ bzw. $\tanh(z+a) = \tanh(1.869) = 0.95$, also $KIV(\rho) = (0.26, 0.95)$.

```
> A <- c(75,62,87,76,73,66,81,74,77)
> B <- c(82,69,89,84,80,68,79,70,74)
> cor.test(A, B)

        Pearson's product-moment correlation

data:  A and B
t = 3.3971, df = 7, p-value = 0.01149
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2622175 0.9534845
sample estimates:
      cor
0.7889483
```

```
> cor.test(A, B, alt="greater")

        Pearson's product-moment correlation

data:  A and B
t = 3.3971, df = 7, p-value = 0.005744
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
0.3774919 1.0000000
sample estimates:
      cor
0.7889483
```

Rangkorrelationskoeffizient von Spearman (1904)

Daten zumindest ordinal skaliert:

Bilde mit $r_1, \ldots, r_n$, den Rängen der x_i, und mit $s_1, \ldots, s_n$, jenen der y_i, den Korrelationskoeffizient nach Pearson, d.h.

$$r_S = \frac{\sum_{i=1}^{n} (r_i - \overline{r})(s_i - \overline{s})}{\sqrt{\sum_{i=1}^{n} (r_i - \overline{r})^2 \sum_{i=1}^{n} (s_i - \overline{s})^2}} = \frac{s_{rs}^2}{\sqrt{s_r^2 s_s^2}} .$$

Vereinfachte Schreibweise von r_S möglich.

Mit

$$\overline{r} = \frac{1}{n}\sum_{i=1}^{n} r_i = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2} = \overline{s}$$

folgt

$$\sum_{i=1}^{n}(r_i - \overline{r})^2 = \sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2 = \frac{(n-1)n(n+1)}{12} = \sum_{i=1}^{n}(s_i - \overline{s})^2 \,.$$

Betrachte $d_i = r_i - s_i$, bzw. besser

$$d_i = \left(r_i - \frac{n+1}{2}\right) - \left(s_i - \frac{n+1}{2}\right) ,$$

so folgt

$$\begin{aligned}
\sum_{i=1}^{n} d_i^2 &= \sum_{i=1}^{n}\left(r_i - \frac{n+1}{2}\right)^2 + \sum_{i=1}^{n}\left(s_i - \frac{n+1}{2}\right)^2 - 2\sum_{i=1}^{n}\left(r_i - \frac{n+1}{2}\right)\left(s_i - \frac{n+1}{2}\right) \\
&= 2\frac{(n-1)n(n+1)}{12} - 2\frac{(n-1)n(n+1)}{12} r_S \\
&= \frac{(n-1)n(n+1)}{6}(1 - r_S)\,.
\end{aligned}$$

Daraus resultiert die einfache Darstellung

$$r_S = 1 - \frac{6\sum\limits_{i=1}^{n} d_i^2}{(n-1)n(n+1)}\,.$$

Eigenschaften von r_S:

1. $-1 \leq r_S \leq +1$,

2. $r_S = +1 \Leftrightarrow r_i = s_i \quad \forall i = 1, \ldots, n$,

3. $r_S = -1 \Leftrightarrow r_i = n + 1 - s_i \quad \forall i = 1, \ldots, n$,

4. r_S ist invariant bzgl. monotoner Trafos der Daten.

Interpretation von r_S:

1. r_S nahe $+1$: Hinweis auf eine starke positive Korrelation, d.h. hat x_i einen hohen (niedrigen) Rang, hat auch y_i hohen (niedrigen) Rang.

2. r_S nahe -1: Hinweis auf eine starke negative Korrelation, d.h. hat x_i einen hohen (niedrigen) Rang, so hat y_i einen niedrigen (hohen) Rang.

3. Ist r_S nahe 0, besteht kein Zusammenhang (Unkorreliertheit).

Beispiel: Bewerbungen

Bewerber i	1	2	3	4	5	6	7	8	9	
Pkte von A	75	62	87	76	73	66	81	74	77	$\overline{x} = 74.6, s_X^2 = 54.8$
Pkte von B	82	69	89	84	80	68	79	70	74	$\overline{y} = 77.2, s_Y^2 = 54.2$
Rang r_i von A	5	1	9	6	3	2	8	4	7	
Rang s_i von B	7	2	9	8	6	1	5	3	4	
d_i^2	4	1	0	4	9	1	9	1	9	$\sum d_i^2 = 38$

Jetzt gilt $r_S = 1 - \frac{228}{720} = 0.683.$

Bewertung von r_S

1. r_S schon bei ordinalem Meßniveau anwendbar

2. r_S gut interpretierbar falls er in der Nähe von -1, 0 oder $+1$ liegt. Sonst schlecht zu interpretieren.

3. r_S ist nicht als Schätzung für ρ geeignet.

Simulationsstudie: Untersuche r_S als Schätzer für ρ. Unterstelle den Daten Normalverteilung.

```
> S <- 1000
> erg <- rep(0, S)
> for(s in seq(1, S)) {
    data <- rmvnorm(20, mean = c(0, 3), sd = c(1, 2), rho = 0.8)
    x <- data[, 1]
    y <- data[, 2]
    erg[s] <- cor.test(x, y, method = "spearman")[[4]]
}
> mysummary(erg)
  Mean   StDev  Median     Iqr   Size
 0.745   0.121   0.770   0.150   1000
```

Test auf Unabhängigkeit

r_S auch für Test auf Unabhängigkeit verwendbar

- Test A: $H_0 : X$ und Y sind unabhängig, $H_1 : X$ und Y sind korreliert,
- Test B: $H_0 : X$ und Y sind unabhängig, $H_1 : X$ und Y sind positiv korreliert,
- Test C: $H_0 : X$ und Y sind unabhängig, $H_1 : X$ und Y sind negativ korreliert.

Teststatistik: Hotelling-Pabst-Statistik

$$D = \sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (R_i - S_i)^2 = \frac{(n-1)n(n+1)}{6}(1 - R_S)$$

(ist mit R_S linear verknüpft)

Durch Umnumerieren der X_i $(r_i = i)$ ist

$$\begin{aligned} D &= \sum_{i=1}^{n}(i-S_i)^2 = \sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n} iS_i + \sum_{i=1}^{n} S_i^2 \\ &= \frac{n(n+1)(2n+1)}{3} - 2\sum_{i=1}^{n} iS_i \,. \end{aligned}$$

Verteilung von D, und damit von R_S, hängt nur von $\sum iS_i$ ab.

Unter Annahme der Unabhängigkeit nimmt $(S_1, \ldots, S_n)$ die $n!$ Permutationen von $(1, \ldots, n)$ mit gleicher Wahrscheinlichkeit an.

Exakte Berechnung

Verteilung kann wieder explizit durch Abzählen berechnet werden.

(s_1, s_2, s_3)	$\sum is_i$	d	r_S
(1,2,3)	14	0	1
(1,3,2)	13	2	1/2
(2,1,3)	13	2	1/2
(2,3,1)	11	6	-1/2
(3,1,2)	11	6	-1/2
(3,2,1)	10	8	-1

R_S hat die folgende Wahrscheinlichkeitsfunktion:

r_S	-1	$-1/2$	$1/2$	1
$P(R_S = r_S)$	$1/6$	$2/6$	$2/6$	$1/6$

Beachte: kleine (große) Werte von $D \quad \Rightarrow \quad$ positive (negative) Korrelation.

Weiters ist für $i = 1, \ldots, n$ wegen $P(S_i = j) = 1/n$ (diskret gleichverteilt)

$$\begin{aligned}
\mathsf{E}(S_i) &= \frac{n+1}{2}, \qquad \mathsf{var}(S_i) = \frac{n^2-1}{12}, \\
\mathsf{cov}(S_i, S_j) &= \mathsf{E}(S_i S_j) - \mathsf{E}(S_i)\mathsf{E}(S_j) = -\frac{n+1}{12} \quad \forall i \neq j\,.
\end{aligned}$$

Daraus folgt sofort

$$\mathsf{E}\left(\sum_{i=1}^{n} i S_i\right) = \sum_{i=1}^{n} i \mathsf{E}(S_i) = \frac{n(n+1)^2}{4},$$

$$\mathsf{var}\left(\sum_{i=1}^{n} i S_i\right) = \sum_{i=1}^{n} i^2 \mathsf{var}(S_i) + \sum_{i \neq j} ij \mathsf{cov}(S_i, S_j) = \frac{(n-1)n^2(n+1)^2}{144}$$

und schließlich

$$\mathsf{E}(D) = \frac{(n-1)n(n+1)}{6}, \qquad \mathsf{var}(D) = \frac{(n-1)n^2(n+1)^2}{36},$$

sowie

$$\mathsf{E}(R_S) = 0, \quad \mathsf{var}(R_S) = \frac{1}{n-1}.$$

Testentscheidungen

Für $n \leq 11$ wird H_0 abgelehnt, falls

- Test A: $d \leq d_{\alpha/2}$ oder $d \geq d_{1-\alpha/2}$,
- Test B: $d \leq d_{\alpha}$,
- Test C: $d \geq d_{1-\alpha}$.

Quantile d_α entnehme Tabelle M (Beachte: d klein, r_S groß!)

Approximationen:

• Für $12 \leq n \leq 20$ approximiere $T = R_S\sqrt{\frac{n-2}{1-R_S^2}}$ durch t_{n-2}-Verteilung. (Beachte: großes t, großes r_S).

• Für $n > 20$ verwende Approximation $Z = \frac{R_S - E(R_S)}{\sqrt{Var(R_S)}} = R_S\sqrt{n-1} \overset{as}{\sim} N(0,1)$. (Beachte: großes z, großes r_S).

Beispiel:

Vergleich der Testergebnis-Reihen: $d = \sum d_i^2 = 38$ beobachtet. Für $\alpha = 0.05$ liefert die Tabelle M für Test B: $d = 38 < d_{0.05} = 48$, was zur Ablehnung der Unabhängigkeitshypothese führt.

```
> cor.test(A, B, alt="greater", method="spearman")

         Spearman's rank correlation rho

 data:  A and B
 S = 38, p-value = 0.02516
 alternative hypothesis: true rho is greater than 0
 sample estimates:
       rho
 0.6833333
```

Kendall's τ

Abhängigkeit durch Anzahl *konkordanter* und *diskordanter* Paare beschreiben.

Definition: Das Paar $[(x_i, y_i), (x_j, y_j)]$ heißt konkordant (übereinstimmend), falls

(a) $x_i < x_j \Rightarrow y_i < y_j$, oder

(b) $x_i > x_j \Rightarrow y_i > y_j$

gilt. Andernfalls heißen die Paare diskordant.

Es gibt $\binom{n}{2}$ Paare $[(x_i, y_i), (x_j, y_j)]$ mit $i < j$.

n_k Anzahl konkordanter Paare, n_d Anzahl diskordanter Paare, mit $n_k + n_d = \binom{n}{2}$.

Sinnvolles Maß für die Korrelation zwischen X und Y ist

$$\tau = \frac{n_k - n_d}{\binom{n}{2}}.$$

Eigenschaften: Bereich: $-1 \leq \tau \leq +1$

1. $\tau = +1 \Leftrightarrow n_k = \binom{n}{2} \Leftrightarrow$ perfekte positive Korrelation,

2. $\tau = 0 \Leftrightarrow n_k = n_d \Leftrightarrow$ keine Korrelation,

3. $\tau = -1 \Leftrightarrow n_d = \binom{n}{2} \Leftrightarrow$ perfekte negative Korrelation.

τ als Statistik für den Test auf Unabhängigkeit der Variablen X und Y verwendbar.

Für den Test einfacher ist Kendall's S

$$S = N_k - N_d$$

Quantile s_d können aus der Tabelle N entnommen werden.

Normalverteilungsapproximation: Schon für $n \geq 8$ gilt approximativ

$$\tau \overset{as}{\sim} N\left(0, \frac{2(2n+5)}{9n(n-1)}\right).$$

Beispiel: Testergebnisse liefern folgende konkordante bzw. diskordante Paare:

Konkordante Beurteilungen	[1,2],	[1,3],	[1,4],	[1,5],	[1,6],	[1,8],	[2,3],	[2,4],	[2,5]
	[2,7],	[2,8],	[2,9],	[3,4],	[3,5],	[3,6],	[3,7],	[3,8],	[3,9]
	[4,5],	[4,6],	[4,8],	[5,6],	[6,7],	[6,8],	[6,9],	[7,8],	[7,9]
	[8,9]								
Diskordante Beurteilungen	[1,7],	[1,9],	[2,6],	[4,7],	[4,9],	[5,7],	[5,8],	[5,9]	

$n_k = 28$, $n_d = 8$, $\binom{n}{2} = 36$, $\tau = \frac{28-8}{36} = 0.556$. Kendall's S ist $s = 20$. Tabelle N: $n = 9$, $s = 20$ liefert $\alpha = 0.025 < 0.05$, auch hier H_0 verwerfen.

```
> cor.test(A, B, alt="greater", method="kendall")

        Kendall's rank correlation tau

data:  A and B
T = 28, p-value = 0.02231
alternative hypothesis: true tau is greater than 0
sample estimates:
      tau
0.5555556
```

Kontingenztafeln: χ^2-Test auf Unabhängigkeit

Test der Hypothese der Unabhängigkeit kategorialer Merkmale.

X und Y sind Faktoren mit k bzw. ℓ Stufen (mit A und B bezeichnet).

D.h., X und Y können nur in $\{a_1, \ldots, a_k\}$ und $\{b_1, \ldots, b_\ell\}$ realisieren.

Stichprobe: $(X_i, Y_i) \overset{iid}{\sim} F$, $i = 1, \ldots, n$.

Information der Stichprobe (Häufigkeiten) darstellbar in einer Kontingenztatabelle.

Testproblem:

- H_0 : A und B sind unabhängig,
- H_1 : A und B sind abhängig

Sei

$$P(A=a_i, B=b_j)=\pi_{ij}\,, \qquad P(A=a_i)=\pi_{i+}\,, \quad P(B=b_j)=\pi_{+j}\,,$$

so gilt für stochastisch unabhängige Variablen A und B

$$\pi_{ij}=\pi_{i+}\pi_{+j}\,.$$

Kontingenztafel:

	b_1	$\ldots$	b_ℓ	$N_{i+}=\sum_{j=1}^{\ell} N_{ij}$
a_1	$N_{11}\,,\ n\pi_{11}$	$\ldots$	$N_{1\ell}\,,\ n\pi_{1\ell}$	N_{1+}
a_2	$N_{21}\,,\ n\pi_{21}$	$\ldots$	$N_{2\ell}\,,\ n\pi_{2\ell}$	N_{2+}
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
a_k	$N_{k1}\,,\ n\pi_{k1}$	$\ldots$	$N_{k\ell}\,,\ n\pi_{k\ell}$	N_{k+}
$N_{+j}=\sum_{i=1}^{k} N_{ij}$	N_{+1}	$\ldots$	$N_{+\ell}$	n

Statistik

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(N_{ij} - n\hat{\pi}_{ij})^2}{n\hat{\pi}_{ij}} \, .$$

Hierbei bezeichnet:

- N_{ij} die beobachtete Häufigkeit in (A_i, B_j),
- $n\hat{\pi}_{ij}$ die unter H_0 ML-geschätzte erwartete Häufigkeit in Zelle (A_i, B_j)

$$\mathsf{E}_0(N_{ij}) = n\pi_{ij} = n\pi_{i+}\pi_{+j}$$

MLE bei Multinomialproblem: relative Häufigkeiten, d.h.

$$\widehat{\mathsf{E}_0}(N_{ij}) = n\hat{\pi}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\frac{N_{i+}}{n}\frac{N_{+j}}{n}$$

- N_{i+}, N_{+j} die Randhäufigkeiten, n die Gesamthäufigkeit.

Zur Verteilung von X^2: Multinomiale Zufallsmatrix

$$N = \begin{pmatrix} N_{11} & \dots & N_{1\ell} \\ \vdots & & \vdots \\ N_{k1} & \dots & N_{k\ell} \end{pmatrix}.$$

Unter H_0 resultiert als Likelihood für N

$$P(N_{11} = n_{11}, \dots, N_{k\ell} = n_{k\ell} | \pi_{11}, \dots, \pi_{k\ell}) = \frac{n!}{n_{11}! \cdot \ldots \cdot n_{k\ell}!} \, \pi_{11}^{n_{11}} \cdot \ldots \cdot \pi_{k\ell}^{n_{k\ell}}$$

$$\overset{H_0}{=} \text{const} \cdot (\pi_{1+}\pi_{+1})^{n_{11}} \cdot \ldots \cdot (\pi_{k+}\pi_{+\ell})^{n_{k\ell}}.$$

Unter H_0 die marginalen Parameter $\pi_{1+}, \dots, \pi_{k+}$, und $\pi_{+\ell}, \dots, \pi_{+\ell}$ schätzen.

Somit liegt ein χ^2-Test auf Anpassung mit ML-geschätzten Parametern vor, also

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(N_{ij} - n\hat{\pi}_{i+}\hat{\pi}_{+j})^2}{n\hat{\pi}_{i+}\hat{\pi}_{+j}} \overset{as}{\sim} \chi^2_\nu \,,$$

wobei

$$\nu = k\ell - 1 - (k - 1 + \ell - 1) = (k-1)(\ell-1)$$

die Freiheitsgrade beschreibt. H_0 wird verworfen, falls $X^2 \geq \chi^2_{1-\alpha;(k-1)(\ell-1)}$.

Die Güte der Approximation ist gut, falls

1. nach Cochran (1954)

- kein $n\hat{\pi}_{ij} < 1$ und
- für maximal 20% der Felder gilt: $n\hat{\pi}_{ij} < 5$;

2. nach Conover (1971)

- fast alle $n\hat{\pi}_{ij}$ von derselben Größenordnung sind,
- alle $n\hat{\pi}_{ij} > 1$,
- die Anzahl der Klassen groß ist.

Beispiel Rezidiveintritt beim Zervixkarzinom: Abhängigkeit von der Anzahl der befallenen Lymphknotenstationen (LKS)?

	LKS	0	1	2	> 2	n_{i+}
$Rezidiv$	ja	43(77.02)	29(20.43)	23(13.75)	28(11.79)	123
	$nein$	153(118.98)	23(31.57)	12(21.25)	2(18.21)	190
	n_{+j}	196	52	35	30	313

$x^2 = (43 - 77.02)^2/77.02 + \cdots + (2 - 18.21)^2/18.21 = 77.63$

$\chi^2_{\nu;0.95}$ $\nu = (4-1)(2-1) = 3$

Beispiel (Agresti): Hängt religiöse Einstellung mit Ausbildungsgrad zusammen? Dazu wurden $n = 2726$ zufällig ausgewählte Person befragt.

	Religiöser Glaube			
Ausbildung	Fundamentalist	gemäßigt	liberal	Total
Weniger als high school	178	138	108	424
High school oder junior college	570	648	442	1660
Bachelor oder Graduate	138	252	252	642
Total	886	1038	802	2726

```
> n <- matrix(c(178,138,108,570,648,442,138,252,252), 3, 3, byrow=TRUE)
> n
     [,1] [,2] [,3]
[1,]  178  138  108
[2,]  570  648  442
[3,]  138  252  252
```

```
> chisq.test(n)

        Pearson's Chi-squared test

data:  n
X-squared = 69.1568, df = 4, p-value = 3.42e-14
```

Extrem starkes Anzeichen, dass hier Assoziation (Abhängigkeit) vorliegt.

Geschätzte Erwartungswerte $\hat{\mathsf{E}}(N_{ij})$

```
> chisq.test(n)$expected
         [,1]     [,2]     [,3]
[1,] 137.8078 161.4497 124.7425
[2,] 539.5304 632.0910 488.3786
[3,] 208.6618 244.4593 188.8789
```

und Residuen $(N_{ij} - \hat{\mathsf{E}}(N_{ij}))/\sqrt{\hat{\mathsf{E}}(N_{ij})}$

```
> chisq.test(n)$residuals
          [,1]       [,2]      [,3]
 [1,]  3.423775 -1.8455228 -1.499038
 [2,]  1.311771  0.6327815 -2.098646
 [3,] -4.891737  0.4822914  4.592852
```

Zelle (3,1) Bachelor/Graduate — Fundamentalist macht den größten Beitrag zur Statistik aus. Unter H_0 wird für diese Zelle eine viel größere Anzahl erwartet ($\hat{\mathsf{E}}(N_{ij}) = 208.66$) als dies beobachtet werden konnte ($n_{31} = 138$).

Interessanterweise waren in dieser Gruppe (Zeile) viel mehr liberale als erwartet. Ohne Bachelors/Graduierte erhält man

```
> n <- matrix(c(178,138,108,570,648,442), 2, 3, byrow=TRUE); chisq.test(n)

        Pearson's Chi-squared test

 data:  n
 X-squared = 9.439, df = 2, p-value = 0.00892
```

```
> chisq.test(n)$residuals
          [,1]       [,2]       [,3]
[1,]  2.092665 -1.7330336 -0.3686978
[2,] -1.057617  0.8758623  0.1863372
```

Dies ist starker Hinweis, dass für diese Personen der Ausbildungsgrad nicht unabhängig vom Fundamentalismus zu sein scheint. Die Residuen sind jetzt auch viel gleichmässiger als zuvor.

Exakter Test von Fisher

2×2 Kontingenztabelle (Vierfeldertafel): für $n \leq 40$ Fisher-Test verwendbar.

Bereits gezeigt, dass die gemeinsame Verteilung von $N = (N_{11}, N_{12}, N_{21}, N_{22})$ unter H_0 (A, B sind unabhängig) geschrieben werden kann als

$$P(N_{11} = n_{11}, \cdots, N_{22} = n_{22}) = \frac{n!}{n_{11}! \cdot \ldots \cdot n_{22}!} (\pi_{1+}\pi_{+1})^{n_{11}} \cdot \ldots \cdot (\pi_{2+}\pi_{+2})^{n_{22}} .$$

Also hängt die Verteilung von N nur von den marginalen Häufigkeiten $(\pi_{1+}, \pi_{2+}, \pi_{+1}, \pi_{+2})$ ab, die somit eine suffiziente Statistik sind.

Somit ist die konditionale Verteilung von N gegeben die suffiziente Statistik von den Parametern unabhängig.

Um dies zu erreichen, müssen wir also die beiden Ränder festhalten.

Prinzip: Alle möglichen Tafeln mit gleichen Rändern wie Stichprobentafel

	B_1	B_2	n_{i+}
A_1	a	b	$a+b$
A_2	c	d	$c+d$
n_{+j}	$a+c$	$b+d$	n

Andere mögliche Tafeln mit diesen Rändern für $0 \leq x \leq \min(a+b, a+c)$

	B_1	B_2	n_{i+}
A_1	x	$a+b-x$	$a+b$
A_2	$a+c-x$	$d-a+x$	$c+d$
n_{+j}	$a+c$	$b+d$	n

Auftrittswahrscheinlichkeit der Stichprobentafel unter H_0 (A, B unabhängig) mittels **hypergeometrischem Modell**.

Urne mit n Kugeln: davon $a+c$ rote, $b+d$ blaue.

Ziehe $a+b$ Kugeln ohne Zurücklegen. Sei X die Anzahl der gezogenen roten Kugeln.

$$P(X=x) = \frac{\binom{a+c}{x}\binom{b+d}{a+b-x}}{\binom{n}{a+b}}, \qquad \text{mit} \quad \mathsf{E}(X) = \frac{(a+b)(a+c)}{n}.$$

Teststatistik: X

- H_0: A und B sind unabhängig, H_1: A und B sind abhängig
- $H_0^* : \forall i,j = 1,2 : p_{ij} = p_{i.}p_{.j}; \quad \exists (i,j) : H_1^* : p_{ij} \neq p_{i.}p_{.j}$

Testentscheidung: H_0 (bzw. H_0^*) ablehnen, falls $x \leq c_{\alpha/2}$ oder $x \geq c_{1-\alpha/2}$ ist. c_α ist α-Quantil von $Hyp(n, a+c, a+b)$.

Beispiel: Gegeben sei eine Stichprobentafel mit $a = 2$, $b = 8$, $c = 3$ und $d = 7$. Für die Wahrscheinlichkeitsverteilung von X folgt nun:

$$\begin{aligned}
P(X = 0) &= \frac{15!\ 10!}{5!\ 20!} = 0.0163 \\
P(X = 1) &= 0.0163\frac{50}{6} = 0.1354 \\
P(X = 2) &= 0.1354\frac{36}{14} = 0.3482 \\
P(X = 3) &= 0.3482\frac{24}{24} = 0.3482 \\
P(X = 4) &= 0.3482\frac{14}{36} = 0.1354 \\
P(X = 5) &= 0.1354\frac{6}{50} = 0.0163
\end{aligned}$$

Für $\alpha = 0.0326$ ist $c_{\alpha/2} = 0$, $c_{1-\alpha/2} = 5$. Da $x = 2$ wird H_0 nicht abgelehnt.

Beispiel: Fisher's Tea Drinker (Fisher, 1935, Agresti, p.92) Eine Kollegin Fisher's behauptete, unterscheiden zu können, ob zuerst Tee oder Milch in die Tasse gegeben wurde. Sie testete acht Tassen, in die jeweils viermal Milch und Tee zuerst gegeben wurde, und kam dabei auf folgendes Ergebnis:

	Vermutung		
Zuerst in der Tasse	Milch	Tee	Total
Milch	3	1	4
Tee	1	3	4
Total	4	4	8

```
> n <- matrix(c(3, 1, 1, 3), 2, 2, byrow=TRUE)
> fisher.test(n)

        Fisher's Exact Test for Count Data

 data:  n
 p-value = 0.4857
```

```
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.2117329 621.9337505
sample estimates:
odds ratio
 6.408309

> chisq.test(n)

       Pearson's Chi-squared test with Yates' continuity correction

data:  n
X-squared = 0.5, df = 1, p-value = 0.4795
```

Obwohl der Ausgang des Experiments beeindruckte, war seine Meinung über seine Kollegin somit bekräftigt. Man bemerke, dass dafür nur bei perfekter Übereinstimmung ($x = 4$) die Hypothese H_0 mit einem p-Wert von 0.029 verworfen werden kann. Andererseits sollte ihn jedoch die Kollegin laut Fisher's Tochter doch von ihrer Fähigkeit überzeugt haben.