
Datenanalyse

Binärer Variable

Dr. Dirk Hasenclever
IMISE, Leipzig

dirk.hasenclever@imise.uni-leipzig.de

Empfehlungen

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

Eine Wahrscheinlichkeit

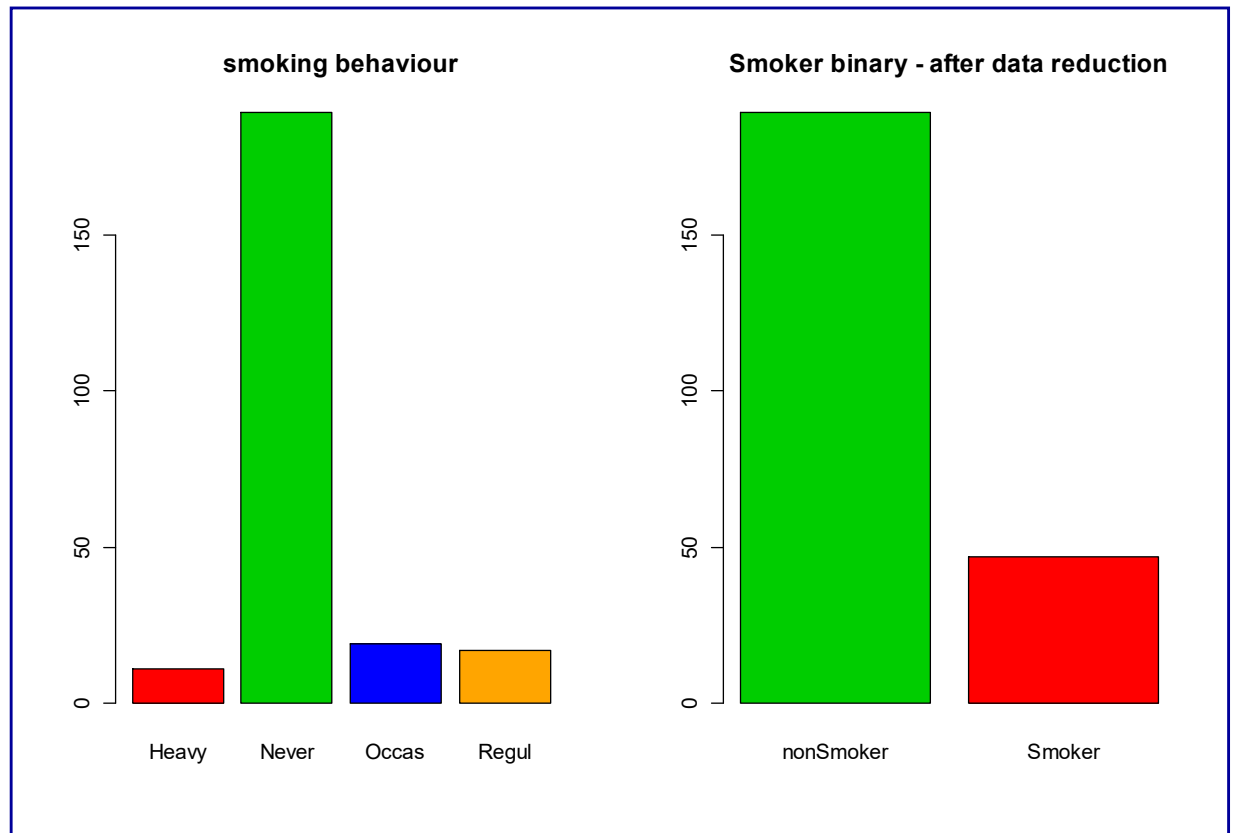
- Schätze eine Wahrscheinlichkeit (ein Stichprobenfall)
 - Sinnvolle, wenn Stichprobe repräsentativ für eine Grundgesamtheit

Smoke				
Heavy	Never	Occas	Regul	
11	189	19	17	

Dichotomisierung

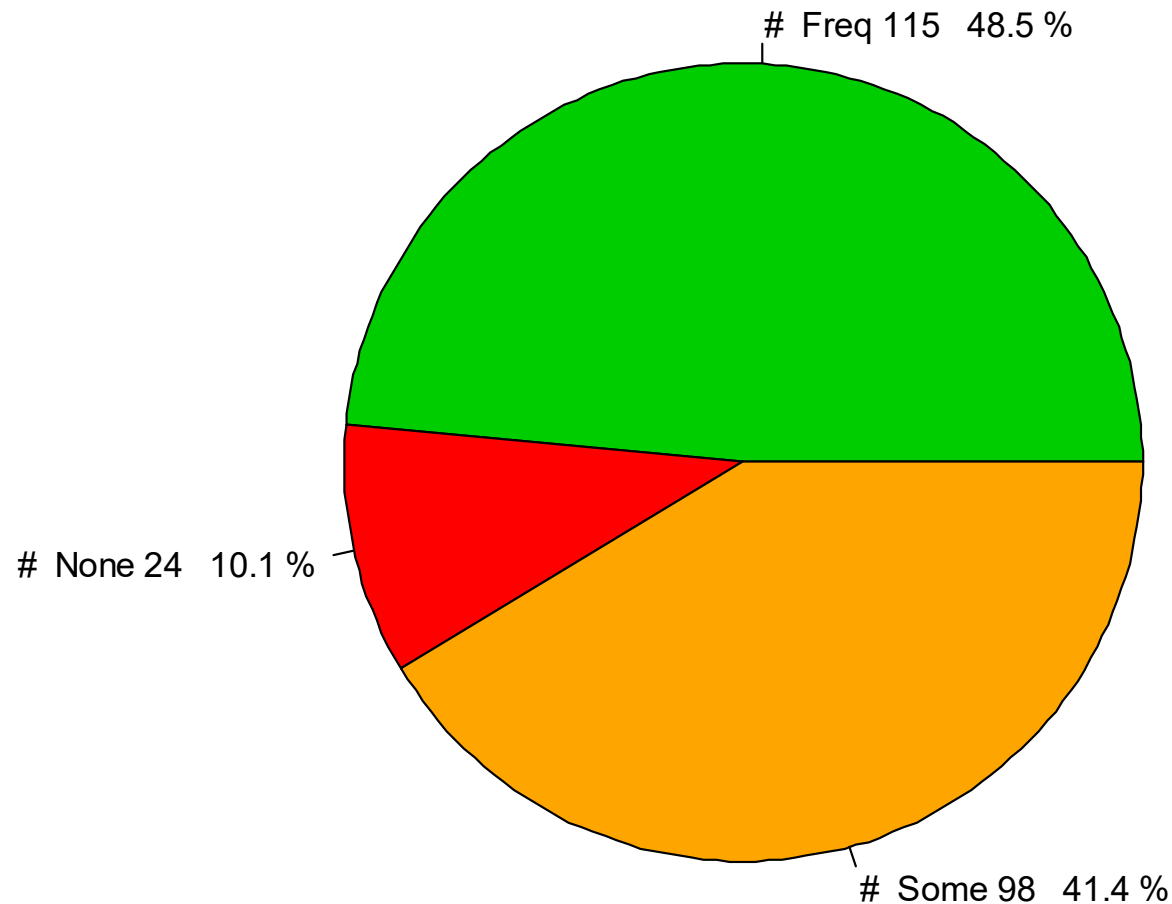


Smoker	
nonSmoker	Smoker
189	47



Pie-Chart alias Kuchendiagramm

exercise behaviour - N=237



Eine Wahrscheinlichkeit II

- Häufigkeitsschätzer mit Konfidenzintervall
- Test gegen vorgegebene Nullhypothese: $H_0: p=0.3$
 - z.B. Bekannte Raucherquote in Normalbevölkerung

```
prop.test(table(Smoker)[2], sum(table(Smoker)), p=0.30)
```

```
1-sample proportions test with continuity  
correction
```

```
data: table(Smoker)[2] out of sum(table(Smoker)), null  
probability 0.3
```

```
X-squared = 10.9542, df = 1, p-value = 0.000934
```

```
alternative hypothesis: true p is not equal to 0.3
```

```
95 percent confidence interval:
```

```
0.1513140 0.2570239
```

```
sample estimates:
```

```
p
```

```
0.1991525 = 47/236
```

Schätzung einer Wahrscheinlichkeit

Nach dem zentralen Grenzwertsatz ist

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

für hinreichend große n .

Das einfachste, gebräuchlichste, aber suboptimale Konfidenzintervallverfahren für p entsteht, wenn man

$$\begin{aligned} \text{CI} &= \{p \mid Z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq Z_{1-\alpha/2}\} \text{ setzt.} \\ &= [\hat{p} \pm Z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}] \end{aligned}$$

Limitationen der Standard Konfidenzintervallverfahren

Dieses Konfidenzintervallverfahren eignet sich für Überschlagsrechnungen und große Fallzahlen.

(Faustregel : $np > 5 \wedge n(1-p) > 5$).

Für kleine Fallzahlen bzw. p nahe 0 oder 1 ist es unbrauchbar.

Für $\hat{p} = 0$ oder 1 hat es Breite 0.

WILSON Konfidenzintervall

Das WILSON Konfidenzintervallverfahren für p entsteht, wenn man

$$CI = \{p \mid Z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq Z_{1-\alpha/2}\} \text{ setzt.}$$

$$= \left[\frac{\hat{p} + \frac{1}{2n} Z_{1-\alpha/2}^2 \pm Z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{Z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n} Z_{1-\alpha/2}^2} \right]$$

WILSON Konfidenzintervall

Das WILSON Konfidenzintervallverfahren für p hat folgende Eigenschaften:

a) Nicht symmetrisch um Häufigkeitsschätzer $\hat{p}$.

b) Mittelpunkt
$$\frac{\hat{p} + \frac{1}{2n} Z^2_{1-\alpha/2}}{1 + \frac{1}{n} Z^2_{1-\alpha/2}}$$

c) Symmetrisch auf logodds Skala

d) Sinnvolle Konfidenzintervalle auch für kleinste Fallzahlen

```
prop.test(0,1)
95 percent confidence
interval:
 0.0000000 0.8966481
sample estimates:
p = 0 =0/1
```

Vergleich von CI-Verfahren

```
> require(epitools)
> # Sehr konservativ Exaktes Verfahren beruhend auf Binomialverteilung
> binom.exact(c(1,3,10), c(10,30,100))
```

	x	n	proportion	lower	upper	conf.level
1	1	10	0.1	0.0025	0.4450	0.95
2	3	30	0.1	0.0211	0.2652	0.95
3	10	100	0.1 0	0.0490	0.1762	0.95

```
> # Bestes Verfahren für die Praxis WILSON Intervalle
> binom.wilson(c(1,3,10), c(10,30,100))
```

	x	n	proportion	lower	upper	conf.level
1	1	10	0.1	0.0178	0.4041	0.95
2	3	30	0.1	0.0345	0.2562	0.95
3	10	100	0.1	0.0552	0.1743	0.95

```
> # Textbook Verfahren (leicht zu erklären, aber nicht besonders gut)
> binom.approx(c(1,3,10), c(10,30,100))
```

	x	n	proportion	lower	upper	conf.level
1	1	10	0.1	-0.0859	0.2859	0.95
2	3	30	0.1	-0.0073	0.2073	0.95
3	10	100	0.1	0.0412	0.1587	0.95

Vier Gestalten einer Vierfeldertafel

> Absolute Häufigkeiten

```
addmargins (table (Sex, Smoker))
```

Sex	Smoker		Sum
	nonSmoker	Smoker	
Female	99	19	118
Male	89	28	117
Sum	188	47	235

> Relative Häufigkeiten

```
addmargins (round (prop. table (table (Sex, Smoker)), 3))
```

Sex	Smoker		Sum
	nonSmoker	Smoker	
Female	0.421	0.081	0.502
Male	0.379	0.119	0.498
Sum	0.800	0.200	1.000

> Zeilenweise bedingte Häufigkeiten

```
addmargins (round (prop. table (table (Sex, Smoker), 1), 3))
```

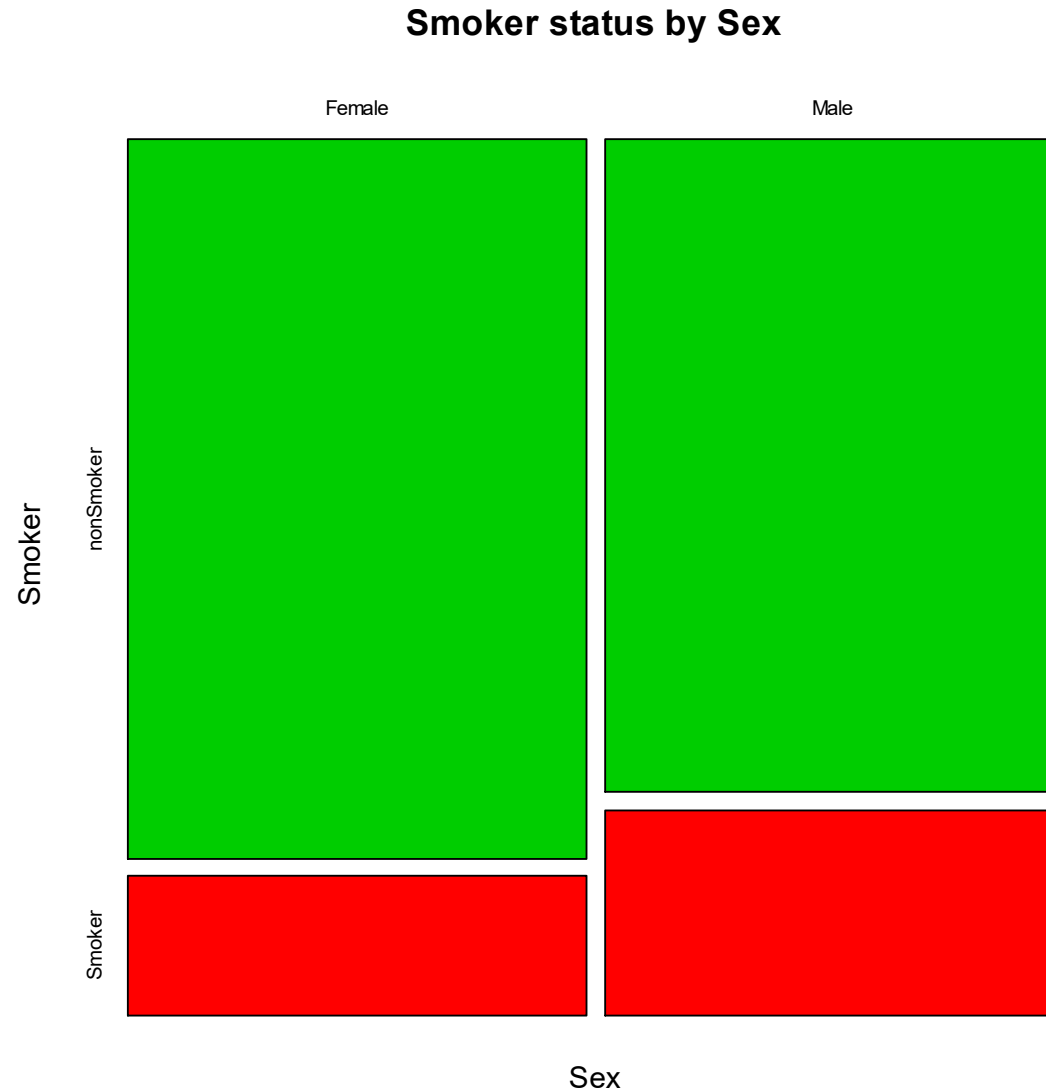
Sex	Smoker		Sum
	nonSmoker	Smoker	
Female	0.839	0.161	1.000
Male	0.761	0.239	1.000
Sum	1.600	0.400	2.000

> Spaltenweise bedingte Häufigkeiten

```
addmargins (round (prop. table (table (Sex, Smoker), 2), 3))
```

Sex	Smoker		Sum
	nonSmoker	Smoker	
Female	0.527	0.404	0.931
Male	0.473	0.596	1.069
Sum	1.000	1.000	2.000

Mosaik - Plot



Rauchen Männer häufiger?

Vergleich zweier Häufigkeiten

- **Frage: Ist die Wahrscheinlichkeit zu rauchen abhängig vom Geschlecht?**
- **Voraussetzung:**
 - Auswahl der Männer bzw. der Frauen gleichartig
(zwei Stichproben - nicht differentiell verzerrt) UND
 - repräsentativ für jeweilige Grundgesamtheit
(absolute Raten machen Sinn)
- **Schätze: Differenz der Rauchwahrscheinlichkeiten.**

> Zeilenweise bedingte Häufigkeiten

```
addmargins(round(prop.table(table  
(Sex, Smoker), 1), 3))
```

Sex	Smoker		Sum
	nonSmoker	Smoker	
Female	0.839	0.161	1.000
Male	0.761	0.239	1.000
Sum	1.600	0.400	2.000

```
> tt<-table(Sex, Smoker)
```

```
> prop.test(tt[,2], apply(tt, 1, sum))
```

```
2-sample test for equality of  
proportions with continuity correction  
X-squared = 1.7883, df = 1, p-value = 0.1811  
alternative hypothesis: two.sided  
95 percent confidence interval:  
-0.189 0.0329  
sample estimates:  
prop 1 prop 2  
0.161 0.239
```

Vergleich zweier Häufigkeiten

- **Frage:** Ist die Rate der Männer bei Rauchern höher als bei Nichtrauchern?
- **Voraussetzung:**
 - Auswahl der Männer bzw. der Frauen gleichartig
(zwei Stichproben - nicht differentiell verzerrt) UND
 - repräsentativ für jeweilige Grundgesamtheit
(absolute Raten machen Sinn)
- **Schätze:** Differenz der Wahrscheinlichkeit ein Mann zu sein, in den Vergleichsgruppen.

```
> Spaltenweise relative Häufigkeiten  
addmargins(round(prop.table(table(Sex,Smoker),2),3))
```

Sex	Smoker		Sum
	nonSmoker	Smoker	
Female	0.527	0.404	0.931
Male	0.473	0.596	1.069
Sum	1.000	1.000	2.000

```
> tt<-table(Sex,Smoker)  
> prop.test(tt[,2],apply(tt,2,sum))
```

2-sample test for equality of proportions
with continuity correction

X-squared = 1.7883, df = 1, p-value = 0.1811

alternative hypothesis: two.sided

95 percent confidence interval:

-0.293 0.048

sample estimates:

prop 1	prop 2
0.473	0.595

Chi2-Test

Der **Chi-Quadrat-Vierfeldertest** (engl. *Cross tab* oder *Fourfold Test*) ist ein **statistischer Test**. Er dient dazu, zu prüfen, ob zwei dichotome Merkmale **stochastisch unabhängig** voneinander sind bzw. ob die Verteilung eines dichotomen Merkmals in zwei Gruppen identisch ist.^[5]

Vorgehensweise [Bearbeiten]

Der Vierfeldertest beruht auf einer (2x2)-**Kontingenztafel**, die die (bivariate) **Häufigkeitsverteilung** zweier **Merkmale** visualisiert:

	Merkmal X		
Merkmal Y	Ausprägung 1	Ausprägung 2	Zeilensumme
Ausprägung 1	a	b	a+b
Ausprägung 2	c	d	c+d
Spaltensumme	a+c	b+d	n = a+b+c+d

Laut einer **Faustformel** muss der **Erwartungswert** aller vier Felder mindestens 5 betragen. Der Erwartungswert wird dabei berechnet aus Zeilensumme*Spaltensumme/Gesamtzahl. Bei einem Erwartungswert kleiner 5 empfehlen Statistiker den **Exakten Fisher-Test**.

Teststatistik [Bearbeiten]

Um die **Nullhypothese** zu prüfen, dass beide Merkmale **stochastisch unabhängig** sind, wird zunächst folgende **Prüfgröße** für einen zweiseitigen Test berechnet:

$$\widehat{\chi^2} = \frac{n \cdot (a \cdot d - c \cdot b)^2}{(a + c) \cdot (b + d) \cdot (a + b) \cdot (c + d)} \overset{a}{\sim} \chi_1^2.$$

Die Prüfgröße ist näherungsweise **Chi-Quadrat-verteilt** mit einem **Freiheitsgrad**. Sie sollte nur dann verwendet werden, wenn in jeder der beiden Stichproben mindestens sechs Merkmalsträger (Beobachtungen) enthalten sind.

Die Teststatistik ist für alle drei Auffassungen der Vierfeldertafel sinnvoll und gleich.

wikipedia

Verschiedene Differenzmaße für Wahrscheinlichkeiten

- Differenz innerhalb $[0; 1]$
 - $(p_1 - p_2)$

- Differenz auf der logit alias logodds Skala: log Oddsratio
 - $(\log(p_1/(1-p_1)) - \log(p_2/(1-p_2))) =$
 - $\log(p_1 / (1-p_1) / (p_2 / (1-p_2))) =$
 - $\log(p_1 * (1-p_2) / (1-p_1) / p_2) =$
 - $\log(\text{oddsratio})$

- Differenz auf log Skala: log Relative Risk
 - $(\log(p_1) - \log(p_2)) = \log(p_1/p_2)$

Beispiel: Differenzmaße im Vergleich

- Komplikationsraten: $p_1 = 0.01$; $p_2 = 0.02$
- Differenz: $p_1 - p_2 = -0.01$
- Log oddsratio: $\text{logit}(p_1) - \text{logit}(p_2) = -0.703$
- Odds ratio: $p_1 * (1 - p_2) / (1 - p_1) / p_2 = 0.495$
- log risk ratio: $\log(p_1) - \log(p_2) = -0.693$
- Risk Ratio: $p_1 / p_2 = 0.5$
- 50% Reduktion der Komplikationsrate, d.h. um 1%....

CompareSuccessRates

➤ `CompareSuccessRatesBySplit(DAT, FAC="Smoker", SUCCESS = 2, SPLIT="SEX", DIGITS = 5, CONF.LEVEL = 0.95)`

	Measure	Diff	Lower95	Upper95	p_test	Method
1	Prob Diff	0.07830	-0.03207	0.18867	0.181	prop.test
2	Odds Ratio	0.61003	0.31491	1.16068	0.136	glm link=logit
3	Relative Risk	0.90667	0.79262	1.03045	0.136	glm link=log

Zusammenhang in Vierfeldertafeln

- Frage: Sind Rauchen und Geschlecht assoziiert? (Besteht Zusammenhang?)
- Voraussetzung:
 - eine Stichprobe (Selektion unangesehen der betrachteten Faktoren) ODER
 - zwei Stichproben nicht in vergleichbarere Weise Repräsentativ (z.B. Fall Kontroll Studie)
 - Rede über absolute Raten machen keinen Sinn
- Schätze: logoddsratio mit Konfidenzintervall.

$$\text{oddsratio} := \frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}}$$

$$\text{schätzer(oddsratio)} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

Für nicht zu kleine N gilt asymptotisch :

$$\log\left(\frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}\right) \sim N\left(\log\left(\frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}}\right), \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}\right)$$

OddsRatio ändert sich nicht, wenn man Zeilen oder Spalten mit positivem Faktor multipliziert.

Beispiel: Schätzung der OddsRatio

```
>require(epitools)
```

```
>oddsratio(table(Sex,Smoker))
```

```
$data
```

Sex	Smoker		Total
	nonSmoker	Smoker	
Female	99	19	118
Male	89	28	117
Total	188	47	235

```
$measure
```

odds ratio with 95% C.I.

estimate	lower	upper
1.63	0.85	3.17

```
$p.value
```

two-sided (Verschiedene Testvarianten)

```
midp.exact
```

fisher.exact chi.square

0.1382

0.1451

0.1335

Assoziationsmaße mit werten in [-1 ; 1]

- $\text{YuleQ} = (\text{OddsRatio} - 1) / (\text{OddsRatio} + 1)$
- $\text{YuleY} = (\sqrt{\text{OddsRatio}} - 1) / (\sqrt{\text{OddsRatio}} + 1)$
- $\text{RR} = \text{Pearson correlation on } 0/1$
 - Only sensitive against diagonal form - not L shape!!

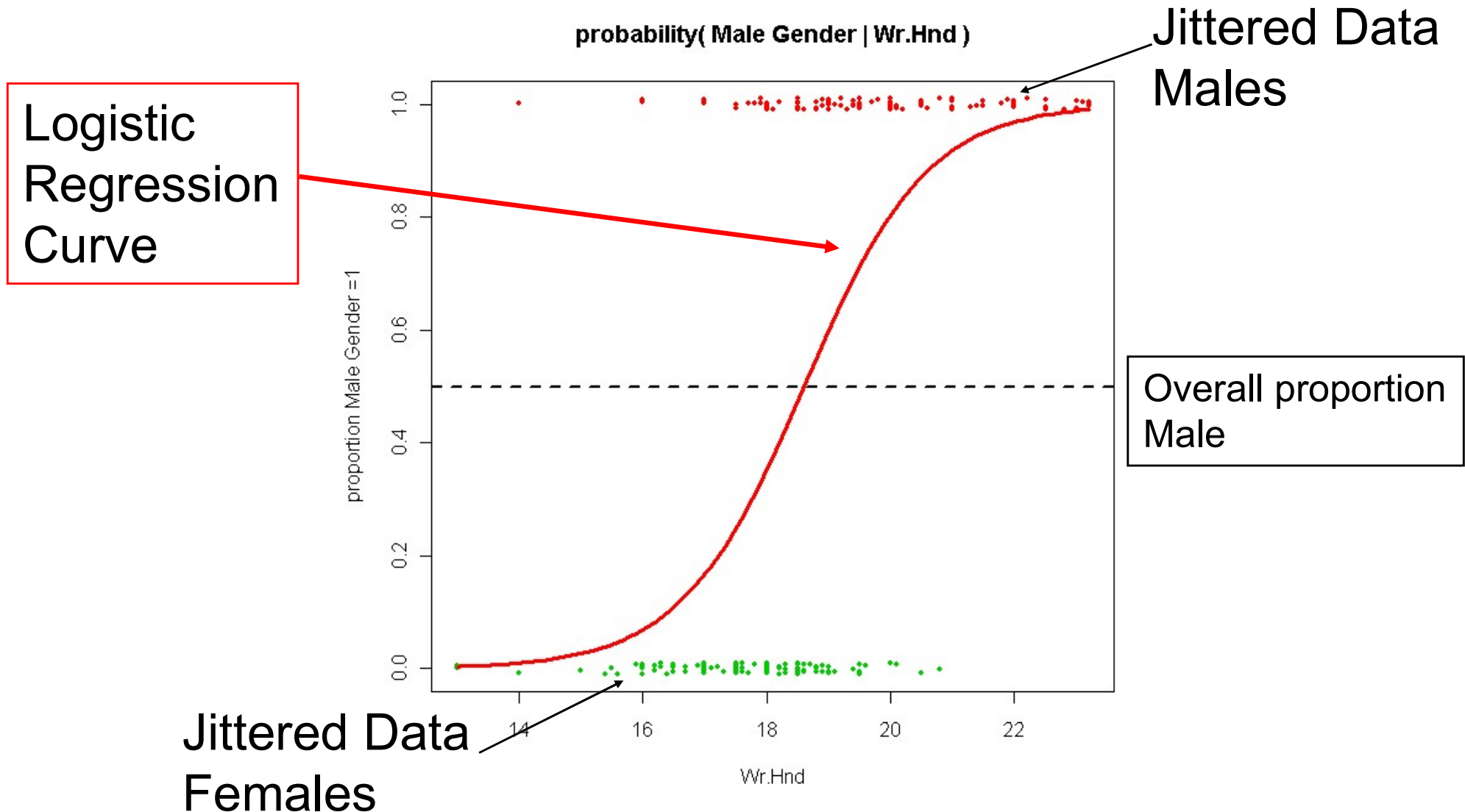
Binary Association

```
> BinaryAssociation( (DAT) )
```

```
[1] "Non-binary variables discarded: Id, WR.HND, NW.HND,  
SpanDifference, SpanMean, FOLD, PULSE, CLAP, EXER, SMOKE,  
HEIGHT, AGE"
```

	YuleQ:	SEX	W.HND	Athlet	Smoker	M.I	Oldies
1	SEX	1.000	-0.185	-0.267	-0.242	-0.059	-0.171
2	W.HND	-0.185	1.000	-0.205	0.232	0.126	1.000
3	Athlet	-0.267	-0.205	1.000	0.267	0.123	0.025
4	Smoker	-0.242	0.232	0.267	1.000	0.068	-0.196
5	M.I	-0.059	0.126	0.123	0.068	1.000	-0.272
6	Oldies	-0.171	1.000	0.025	-0.196	-0.272	1.000

Ein Stichproben Zusammenhang Binäre Größe und stetige Variable



Regression bei binären Daten

Y binär.

Logistische Regression:

$$pr(Y = 1 | z) = \frac{\exp(\alpha + \beta z)}{1 + \exp(\alpha + \beta z)}$$

z ist Kovariate

α , β zu schätzende
Parameter

Linkfunktion zwingt linearen Prädiktor $\alpha + \beta z$
auf Werte zwischen 0 und 1.

Maximum Likelihood Schätzung

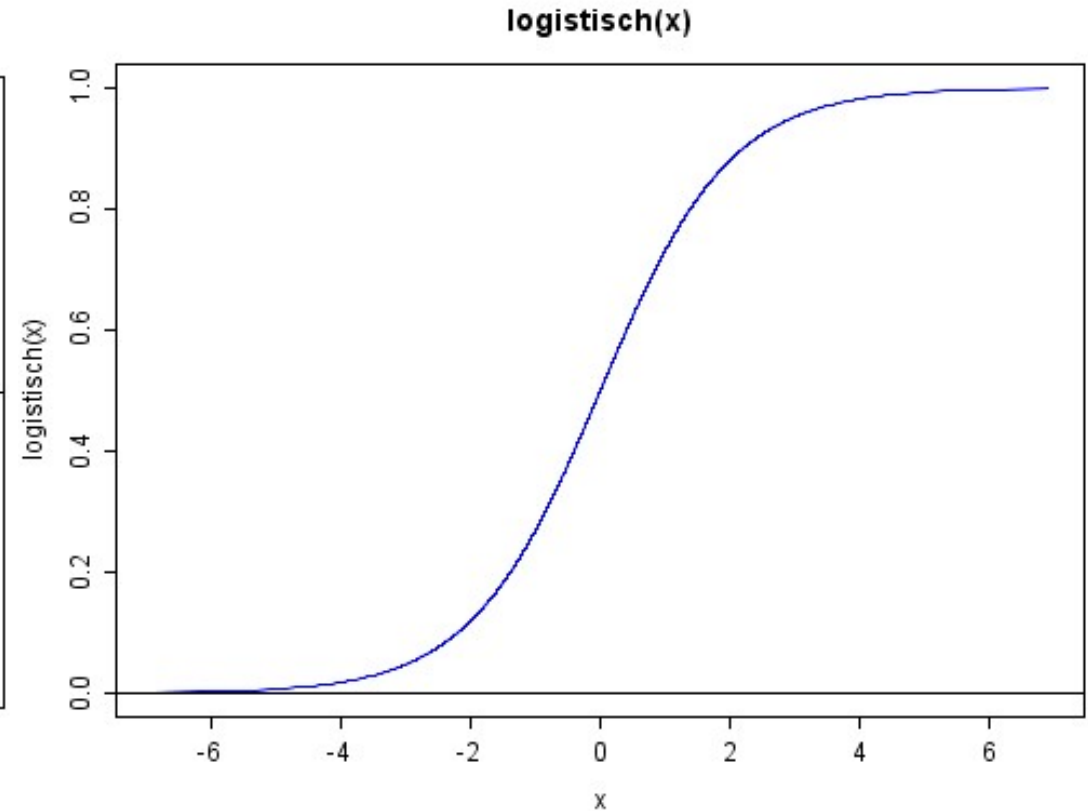
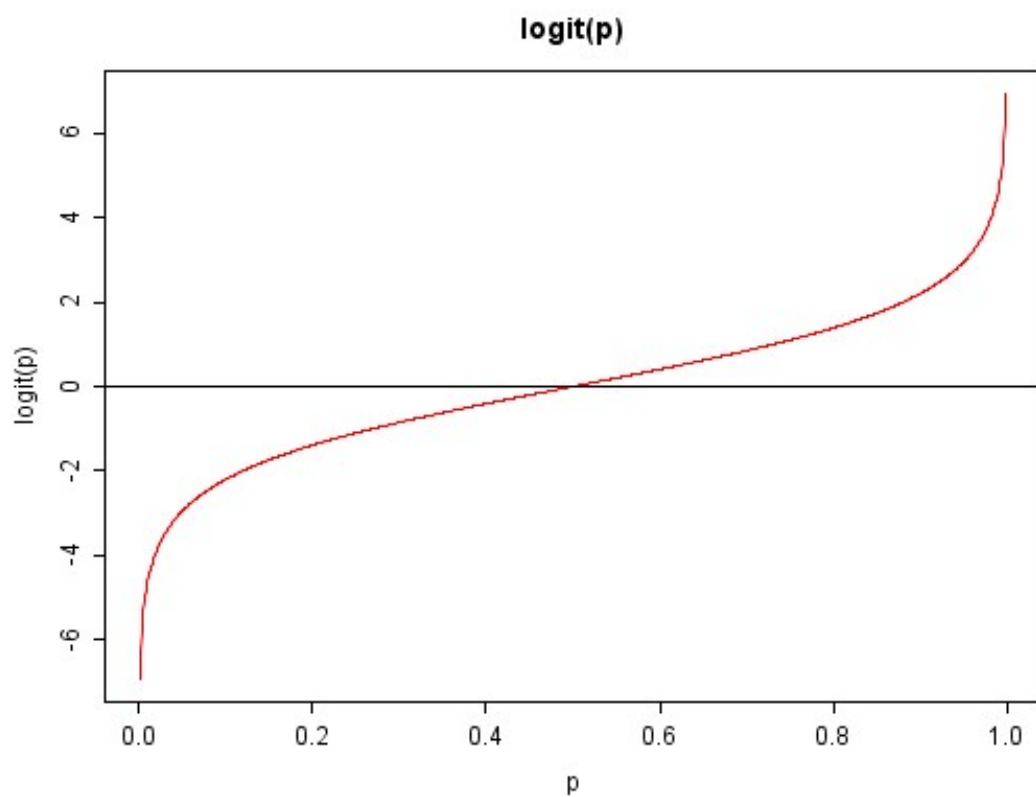
- Mit dieser Formel, kann man für jede Beobachtung (Y_i, z_i) ihre Wahrscheinlichkeit gegeben α und β berechnen.

- $L(\alpha, \beta) = \text{Produkt über } i \text{ pr}((Y_i, z_i) | \alpha, \beta)$

ist die Likelihoodfunktion, die die Wahrscheinlichkeit der Daten geben α, β beschreibt.

Die besten α, β , sind diejenigen Werte, an welchen die Likelihoodfunktion ihr Maximum annimmt.

Logit Scala I



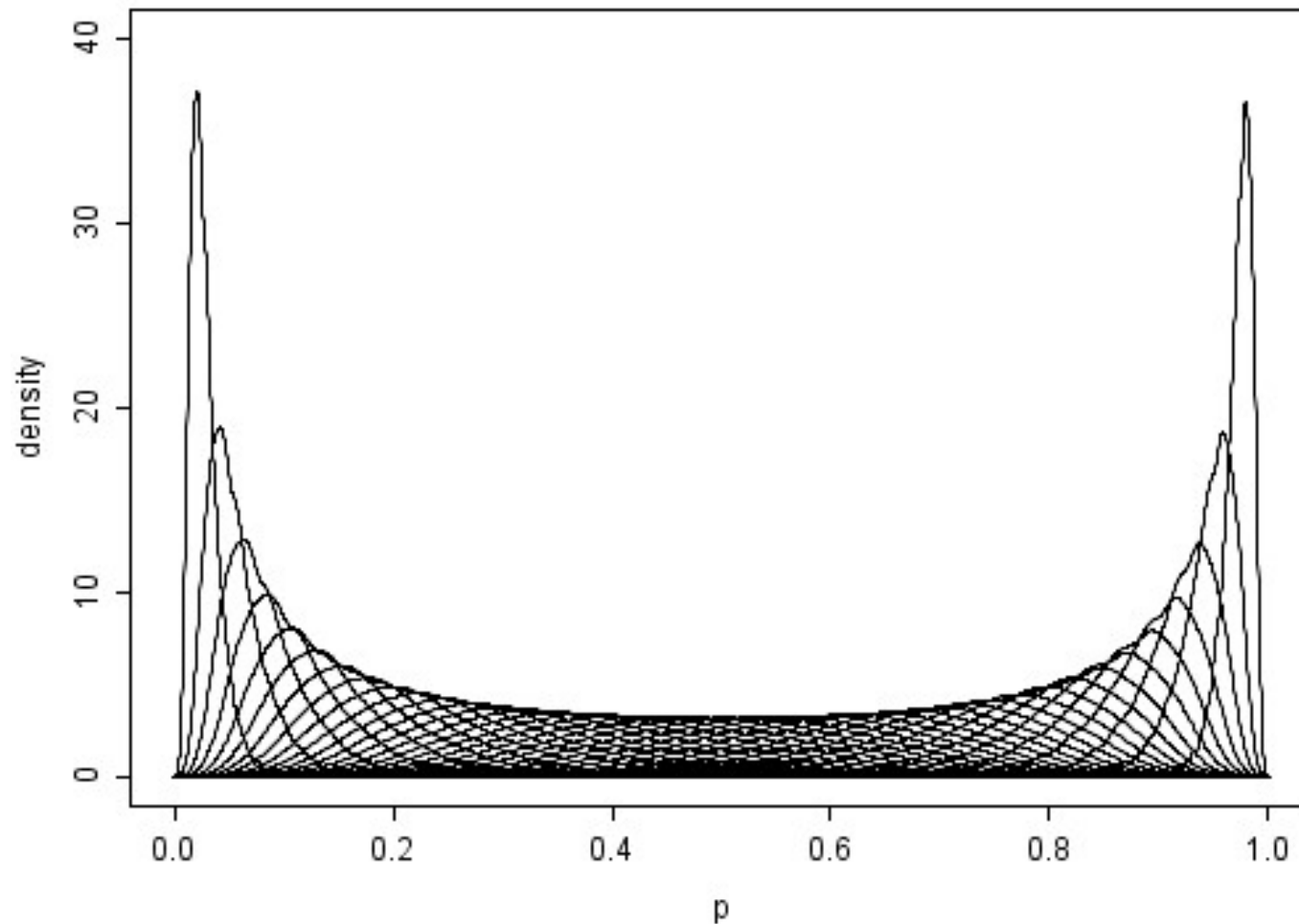
Logistische Linkfunktion:

$$g^{-1}(p) = \ln\left(\frac{p}{1-p}\right) \text{ (Logodds);}$$

$$g(x) = \frac{\exp(x)}{1 + \exp(x)};$$

Logit Scale II

Illustration logit scale



Random noise
on logit scale

Beispiel logistische Regression I

- Klinisches Problem
 - Hodgkin's Lymphom
 - Klinisch nur eine Lymphknotenregion oberhalb des Zwerchfells befallen
 - Diagnostische Bauchöffnung (Laparotomie) zur Abklärung eines möglichen abdominalen Befalls
 - Daten: befallen= j/n + diverse Kovariaten: Alter, Geschlecht, Mediastinum befallen= j/n , Histologie, extranodaler Befall= j/n

Beispiel logistische Regression II

Table V-2. Prognostic factors of positive laparotomy findings in CS IA supradiaphragmatic patients (N=1,169)

Patients with positive laparotomy (PS III-IV): n = 287 (24.6 %)

STEPWISE LOGISTIC REGRESSION

Parameters included in the model (number of dummy variables) = sex, age (4), mediastinal involvement (2), histological subtype (2), extra nodal localization.

Step	Variable (codes)	log. likelihood	Coeff/s.e.	p value
0		- 649.84		
1	SEX (M=-1, F=1)	- 635.03	- 0.396/0.084	< 0.001
2	HISTOLOGY (LP-NS=-1, MC-LD-uncl.=1)	- 625.90	+ 0.295/0.070	< 0.001
3	AGE (<50=-1, 50+=1)	- 622.50	+ 0.243/0.091	0.009
	CONSTANT		- 1.079/0.102	

Henry-Amar et al. (IDHD) 1990

Beispiel logistische Regression III

Pattern	number of patients	observed proportion of LAP +	predicted probability of LAP +
Male, 50+ yrs, MC-LD	50	0.38	0.46
Female, 50+ yrs, MC-LD	22	0.36	0.28
Male, <50 yrs, MC-LD	276	0.36	0.35
Male, 50+ yrs, LP-NS	66	0.32	0.32
Female, 50+ yrs, LP-NS	35	0.26	0.18
Male, <50 yrs, LP-NS	402	0.23	0.23
Female, <50 yrs, MC-LD	76	0.17	0.19
Female, <50 yrs, LP-NS	236	0.11	0.12

Henry-Amar et al. (IDHD) 1990