Zusammenhänge in den Daten

VB

Institut für Veterinär-Epidemiologie und Biometrie, FU Berlin

11/03/2020

Wiederholung

Wiederholung

Zusammenhänge in kategoriellen Daten

Zusammenhänge in kategoriellen Daten

Um Zusammenhänge zwischen kategoriellen Daten festzustellen, können wir keine Mittelwerte oder ähnliche kontinuierliche Maßzahlen betrachten

Zusammenhänge in kategoriellen Daten

Um Zusammenhänge zwischen kategoriellen Daten festzustellen, können wir keine Mittelwerte oder ähnliche kontinuierliche Maßzahlen betrachten

Wir können nur Anzahl oder Häufgkeiten für unterschiedliche Kategorien untersuchen

Zusammenhänge in kategoriellen Daten (1)

- Pearsonsche Chi-Quadrat-Test (χ^2-Test)
- Exakter Test nach Fisher
- Yates-Korrektur
- G-Test
- Maßzahlen der Effektstärke

Kontingenztabelle (Beispiel)

Den 200 Katzen wir das *Tanzen* auf der Schur beigebracht indem sie enweder durch *Futter* oder *Aufmerksamkeit* belohnt werden. Nach einer Woche wird gezählt, wie viele Katzen erfolgreich trainiert werden konnten.

Kontingenztabelle (Beispiel)

Den 200 Katzen wir das *Tanzen* auf der Schur beigebracht indem sie enweder durch *Futter* oder *Aufmerksamkeit* belohnt werden. Nach einer Woche wird gezählt, wie viele Katzen erfolgreich trainiert werden konnten.

		Belohnung		
		Futter	Aufmerksamkeit	Gesamt
Tanzen	Ja	28	48	76
	Nein	10	114	124
	Gesamt	38	162	200

Pearsonsche Chi-Quadrat-Test (χ^2-Test)

Gibt es einen Zusammenhang zwischen der Anzahl der erfolgreich trainierten Katzen und Art der Belohnung?

Pearsonsche Chi-Quadrat-Test (χ^2-Test)

Gibt es einen Zusammenhang zwischen der Anzahl der erfolgreich trainierten Katzen und Art der Belohnung?

Pearsonsche Chi-Quadrat-Test (χ^2-Test)

Vergeleicht die beobachteten Häufigkeiten mit solchen, die man rein zufällig bekommen würde

Pearsonsche Chi-Quadrat-Test (χ^2-Test) (1)

$$\chi^2 = \sum \frac{(\text{observed}_{ij} - \text{model}_{ij})^2}{\text{model}_{ij}}$$

i ist die Reihe und j die Spalte in der Kontingenztabelle

Pearsonsche Chi-Quadrat-Test (χ^2-Test) (1)

$$\chi^2 = \sum \frac{(\text{observed}_{ij} - \text{model}_{ij})^2}{\text{model}_{ij}}$$

i ist die Reihe und j die Spalte in der Kontingenztabelle

$$\text{model}_{ij} = \frac{\text{row total}_i \times \text{column total}_j}{n}$$

für den Katzenfall $\chi^2 = 25.35$. Für das Signifikanzneveau von 0.05 entspricht χ^2-Statistik von 3.4. Da unser Wert höher ist, ist das Ergebnis signifikant.

Dabei $df = (r-1)(c-1)$ mit r - Anzahl der Reihen und c - Anzahl der Spalten.

Theoretische Kontingenztabelle (Beispiel)

Theoretische Kontingenztabelle (Beispiel)

		Belohnung		
		Futter	Aufmerksamkeit	Gesamt
Tanzen	Ja	14.44	61.56	76
	Nein	23.56	100.44	124
	Gesamt	38	162	200

Exakter Test nach Fisher

χ^2-Test basiert auf der Annahme, das die Abschätzung von χ^2 aus der Stichprobe der theoretischen χ^2-Verteilung entspicht.

Das gilt allerding nur für die Anzahl der Meßwerte für jede Kategorie größer als 5.

Für die kleinere Anzahl der Meßwerte pro Kategorie wird *Exakter Test nach Fisher* benutzt.

G-Test (The likelihood ratio)

Alternative zum Pearson-χ^2-Test

$$G = 2 \sum \text{observed}_{ij} \ln \left(\frac{\text{observed}_{ij}}{\text{model}_{ij}} \right)$$

G-Statistik ist auch nach χ^2 verteilt. Für den Beispiel mit den Katzen $G = 24.94$ und daher signifikant.

Yates-Korrektur

$$\chi^2 = \sum \frac{(|\text{observed}_{ij} - \text{model}_{ij}| - 0.5)^2}{\text{model}_{ij}}$$

Für den Katzenbeispiel $\chi^2 = 23.52$ und daher ist χ^2 kleiner geworden.

χ^2 in R

```
library(gmodels)
food <- c(28,10)
affection <- c(48,114)
catsTable <- cbind(food,affection)
CrossTable(catsTable, fisher = TRUE, chisq = TRUE, expected = TRUE, sresid = TRUE)
```

χ^2 in R (1) I

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  200
##
##
##              |
##              |      food | affection | Row Total |
## -------------|-----------|-----------|-----------|
##         [1,] |        10 |       114 |       124 |
##              |    23.560 |   100.440 |           |
##              |     7.804 |     1.831 |           |
##              |     0.081 |     0.919 |     0.620 |
##              |     0.263 |     0.704 |           |
##              |     0.050 |     0.570 |           |
## -------------|-----------|-----------|-----------|
##         [2,] |        28 |        48 |        76 |
##              |    14.440 |    61.560 |           |
##              |    12.734 |     2.987 |           |
##              |     0.368 |     0.632 |     0.380 |
##              |     0.737 |     0.296 |           |
##              |     0.140 |     0.240 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        38 |       162 |       200 |
##              |     0.190 |     0.810 |           |
## -------------|-----------|-----------|-----------|
```

χ^2 in R (1) II

```
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  25.35569     d.f. =  1     p =  4.767434e-07
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  23.52028     d.f. =  1     p =  1.236041e-06
##
##
## Fisher's Exact Test for Count Data
## ------------------------------------------------------------
## Sample estimate odds ratio:  0.1519927
##
## Alternative hypothesis: true odds ratio is not equal to 1
## p =  1.311709e-06
## 95% confidence interval:  0.06086544 0.352389
##
## Alternative hypothesis: true odds ratio is less than 1
## p =  7.7122e-07
## 95% confidence interval:  0 0.3131634
##
## Alternative hypothesis: true odds ratio is greater than 1
## p =  0.9999999
## 95% confidence interval:  0.07015399 Inf
##
##
##
```

Odds-Ratio (Chancenverhältnis, relative Chance, Quotenverhältnis)

Maßzahl der Effektstärke im Falle von Assoziation zwischen Variablen (measure of associations)

$$\text{odds}_{\text{dancing after food}} = \frac{\text{had food and danced}}{\text{had food but didn't dance}} = 28/10$$

$$\text{odds}_{\text{dancing after affection}} = \frac{\text{had affection and danced}}{\text{had affection but didn't dance}} = 48/114$$

$$\text{odds ratio} = \frac{\text{odds}_{\text{dancing after food}}}{\text{odds}_{\text{dancing after affection}}} = 6.65$$

Odds-Ratio

odds ratio $= 6.57\,(2.84, 16.43)$ bedeutet, dass die Belohung mit Futter um 6.65 die Chance erhöht, der Katze das Tanzen auf der Schnur beizubringen, verglichen mit der Belohnung durch Zuwendung.

Odds-Ratio

odds ratio $= 6.57\,(2.84, 16.43)$ bedeutet, dass die Belohung mit Futter um 6.65 die Chance erhöht, der Katze das Tanzen auf der Schnur beizubringen, verglichen mit der Belohnung durch Zuwendung.

Generell gilt: - odds ratio > 1 die Chance auf Erfolg ist größer durch die Futter-Belohnung, als durch die Zuwendung

Odds-Ratio

odds ratio $= 6.57\,(2.84, 16.43)$ bedeutet, dass die Belohung mit Futter um 6.65 die Chance erhöht, der Katze das Tanzen auf der Schnur beizubringen, verglichen mit der Belohnung durch Zuwendung.

Generell gilt: - odds ratio > 1 die Chance auf Erfolg ist größer durch die Futter-Belohnung, als durch die Zuwendung

- odds ratio $= 1$ sowohl die Futter-Belohnung als auch durch die Zuwendung hat die gleiche Chance auf Erfolg

Odds-Ratio

odds ratio $= 6.57\,(2.84, 16.43)$ bedeutet, dass die Belohung mit Futter um 6.65 die Chance erhöht, der Katze das Tanzen auf der Schnur beizubringen, verglichen mit der Belohnung durch Zuwendung.

Generell gilt: - odds ratio > 1 die Chance auf Erfolg ist größer durch die Futter-Belohnung, als durch die Zuwendung

- odds ratio $= 1$ sowohl die Futter-Belohnung als auch durch die Zuwendung hat die gleiche Chance auf Erfolg
- odds ratio < 1 die Chance auf Erfolg ist kleiner durch die Futter-Belohnung, als durch die Zuwendung

Kontingenzkoeffizient

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

n - Stichprobenumfang

Kontingenzkoeffizient

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

n - Stichprobenumfang

Der Nachteil von C ist das der immer kleiner als 1 ist und hängt von der Form der Kontingenztabelle. Daher betrachten man den *korrigierten Kontingenzkoeffizient*

$$C = \frac{C}{C_{\max}} = \sqrt{\frac{\chi^2 \cdot m}{(\chi^2 + n)(m-1)}}$$

m - $\min(r, c)$

Mosaic plots I

Datensatz über Zullasungen in UC Berkeley für 6 größte Fachbereiche im Jahr 1973

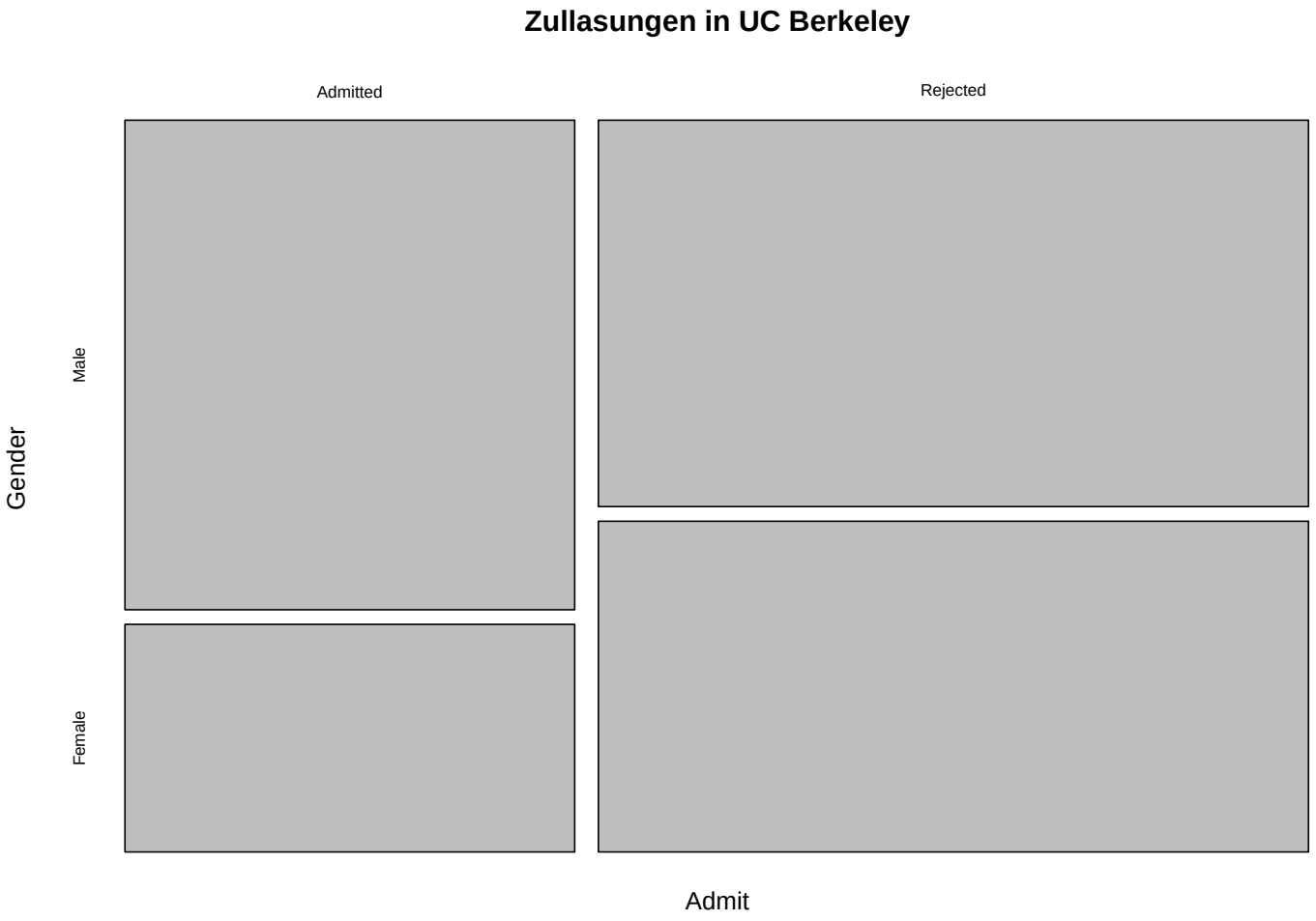
```
## , , Dept = A
##
##           Gender
## Admit      Male Female
##   Admitted  512     89
##   Rejected  313     19
##
## , , Dept = B
##
##           Gender
## Admit      Male Female
##   Admitted  353     17
##   Rejected  207      8
##
## , , Dept = C
##
##           Gender
## Admit      Male Female
##   Admitted  120    202
##   Rejected  205    391
##
## , , Dept = D
##
##           Gender
## Admit      Male Female
##   Admitted  138    131
##   Rejected  279    244
##
## , , Dept = E
##
##           Gender
## Admit      Male Female
```

Mosaic plots II

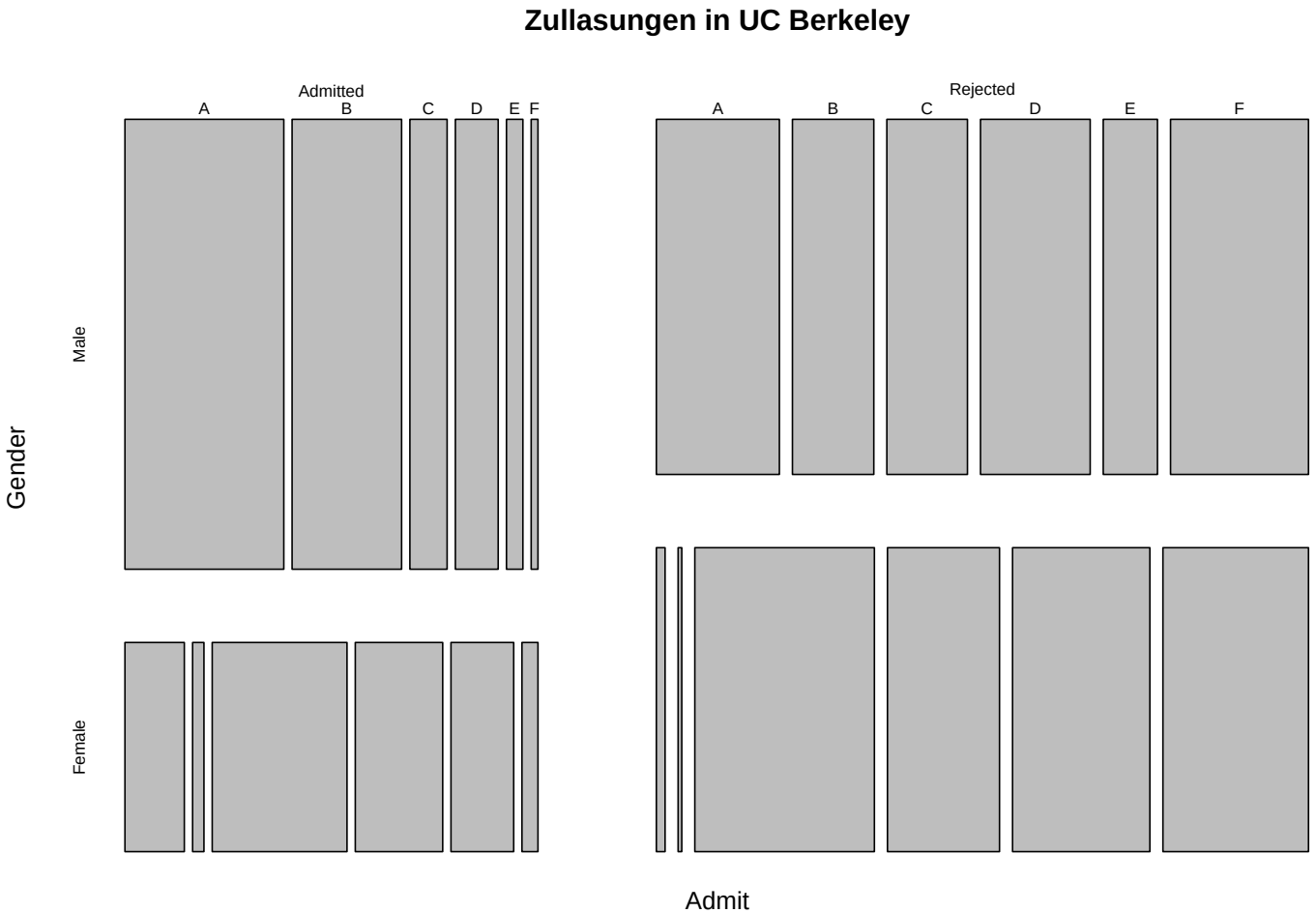
```
##   Admitted   53     94
##   Rejected  138    299
##
## , , Dept = F
##
##           Gender
## Admit      Male Female
##   Admitted   22     24
##   Rejected  351    317
```

	Male	Female
Admitted	1198	557
Rejected	1493	1278

Simpson-Paradoxon (1)



Simpson-Paradoxon (2)



Simpson-Paradoxon (3)

Es scheint, dass die Bewertung verschiedener Gruppen unterschiedlich ausfällt, je nachdem ob man die Ergebnisse der Gruppen kombiniert oder nicht. Dieses Phänomen tritt oft bei statistischen Auswertungen in den Sozialwissenschaften und in der Medizin auf [*Wikipedia*]

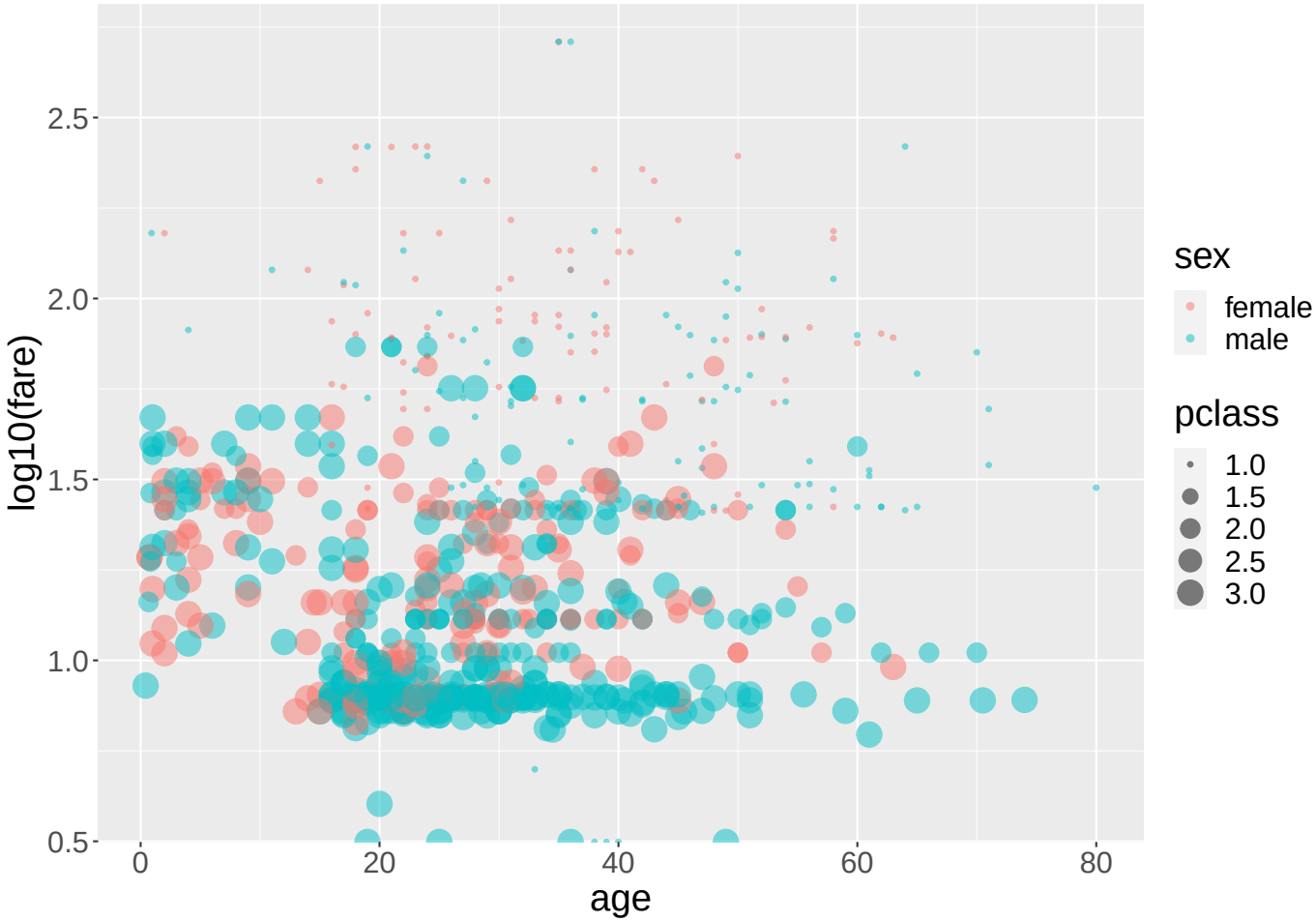
Simpson-Paradoxon (3)

Es scheint, dass die Bewertung verschiedener Gruppen unterschiedlich ausfällt, je nachdem ob man die Ergebnisse der Gruppen kombiniert oder nicht. Dieses Phänomen tritt oft bei statistischen Auswertungen in den Sozialwissenschaften und in der Medizin auf [*Wikipedia*]

Liegen je nach Beurteilungsweise deutlich unterschiedliche Ergebnisse vor, kann dies auf nicht erfasste Einflussfaktoren zurückgeführt werden. Wollen Auswertende mögliche Fehlschlüsse vermeiden, müssen sie diese Einflussfaktoren finden, soweit sie vorhanden sind. Das Vorliegen eines Simpson-Paradoxons kann hier als Indikator dienen. [*Wikipedia*]

Zusammenhänge in kontinuierlichen Daten

Bivariate Daten



Zusammenhänge in kontinuierlichen Daten

- Pearson-Korrelationskoeffizient
- Spearman-Rangkorrelationskoeffizient

Pearson-Korrelationskoeffizient

Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Covariance

$$\mathrm{cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

Pearson-Korrelationskoeffizient

$$r = \frac{\operatorname{cov}(x, y)}{s_x s_y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1) s_x s_y}$$

Korregierter z-score

$$z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

Sein Standardfehler

$$SE_{z_r} = \frac{1}{\sqrt{n-3}}$$

Konfidenzintervalle

$$z_r \pm (1.96 \times SE_{z_r}), \text{ Transformation: } r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}$$

Korrelationskoeffizient (Beispiel)

```
#fig.show = 'hide'
#results='hide'
library(ggplot2)
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/4_Statistische_Tests_and_Regression/Data/Data\ files/Album\ Sales\ 1.dat', s
cor(df)
```

```
##           adverts     sales
## adverts 1.0000000 0.5784877
## sales   0.5784877 1.0000000
```

```
cor.test(df$adverts,df$sales)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$adverts and df$sales
## t = 9.9793, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4781207 0.6639409
## sample estimates:
##       cor
## 0.5784877
```

Spearman-Korrelationskoeffizient (Beispiel)

```r
cor(df, method = "spearman")
```

```
##           adverts     sales
## adverts 1.0000000 0.5541557
## sales   0.5541557 1.0000000
```

```r
cor.test(df$adverts, df$sales, alternative = "greater", method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$adverts and df$sales
## S = 594444, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.5541557
```

Kendall-tau (Beispiel)

```
cor(df, method = "kendall")
```

```
##            adverts     sales
## adverts 1.0000000 0.3985301
## sales   0.3985301 1.0000000
```

```
cor.test(df$adverts, df$sales, alternative = "greater", method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$adverts and df$sales
## z = 8.2362, p-value < 2.2e-16
## alternative hypothesis: true tau is greater than 0
## sample estimates:
##       tau
## 0.3985301
```

Wahrscheinlichkeitsverteilungen

Normalverteilung oder Gauß-Verteilung

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \mathcal{N}(\mu, \sigma)$$

μ — Mittelwert

σ — Standardabweichung

Normalverteilung oder Gauß-Verteilung (1)

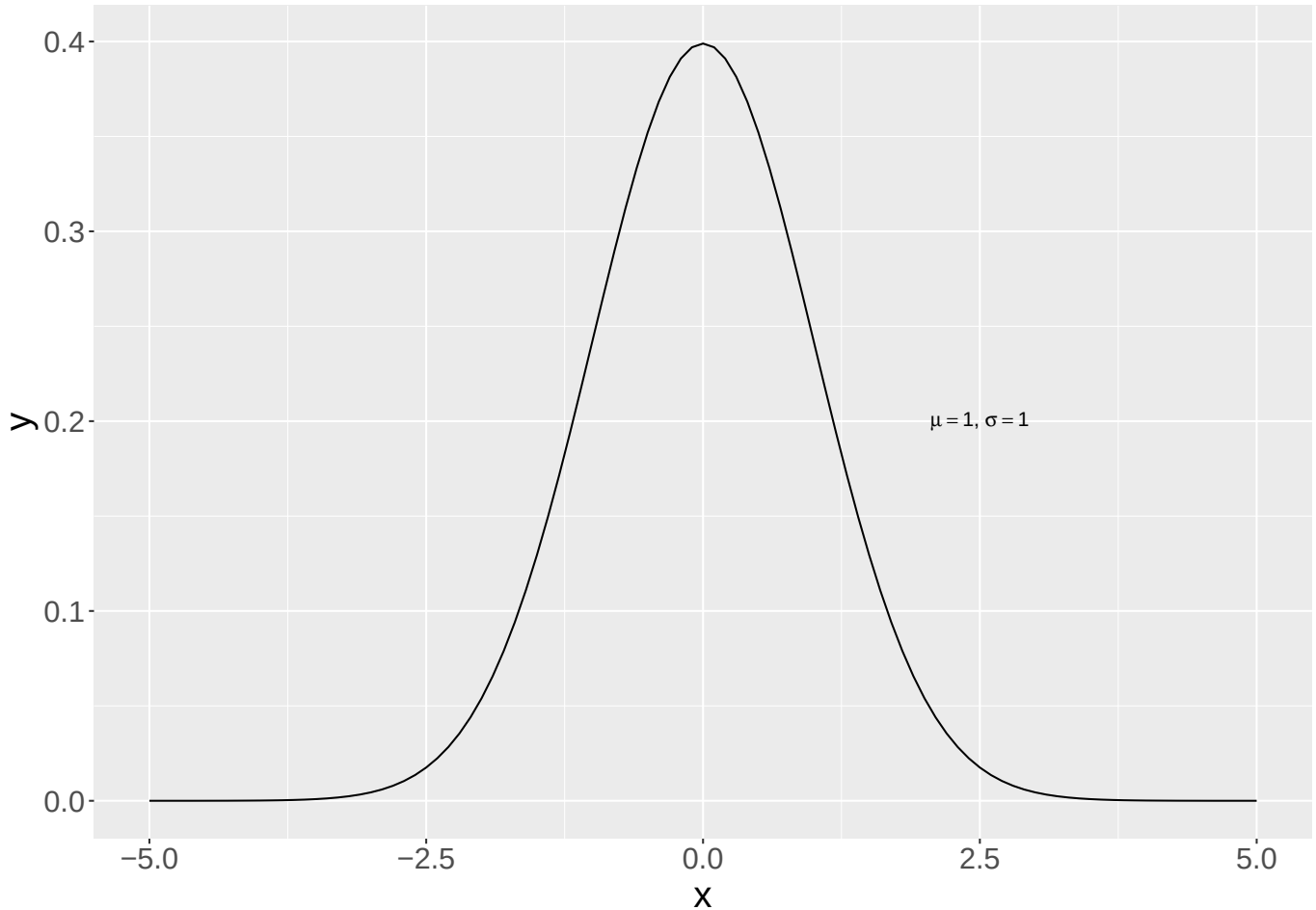


Figure 1: Normalverteilung

Binomiale Verteilung

$$p(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

n - Anzahl der Versuche

p - Erfolgs- oder Trefferwahrscheinlichkeit

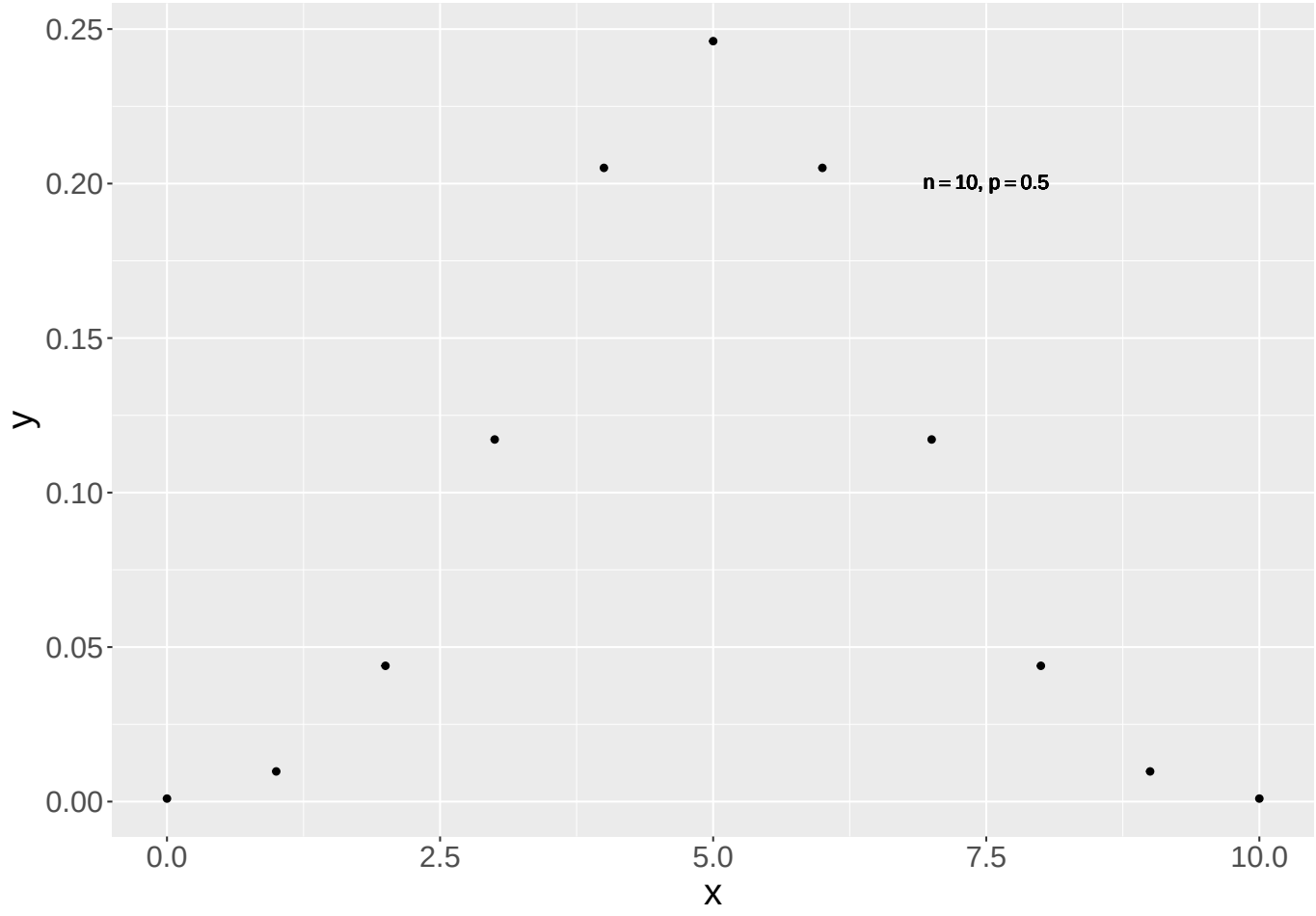


Figure 2: Binomiale Verteilung

Poisson-Verteilung

$$P_\lambda(k) = \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda}$$

λ Erwartungswert und gleichzeitig die Varianz der Poisson-Verteilung

Poisson-Verteilung (1)

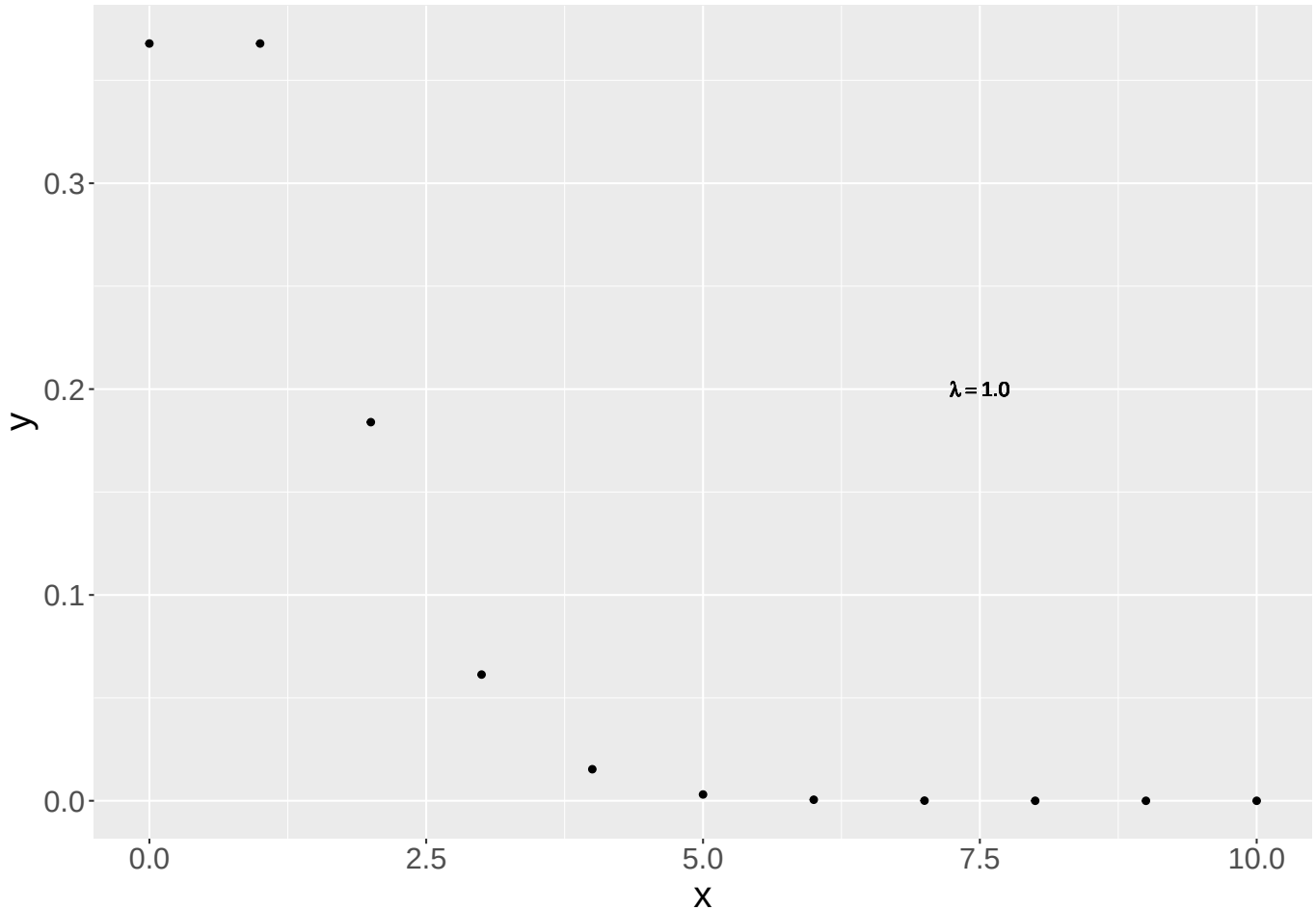


Figure 3: Poisson-Verteilung