

# Programmieren in Anwendungen

Annette Bieniusa

Technische Universität Kaiserslautern

*bieniusa@cs.uni-kl.de*

26.06.2014

# Überblick

Stochastische Modellbildung

Testtheorie

Korrelation

# Stochastische Modellbildung

# Messwerte als Zufallsvariablen

- ▶ Messwerte sind i.d.R. nicht-deterministisch und werden daher als *Zufallsvariablen* modelliert.
- ▶ Das Erheben von Messdaten entspricht dabei einem *Zufallsexperiment*.
- ▶ Beispiel: Körpergröße von Menschen, Bewertungen von Kinofilmen, Dauer von Schwangerschaften

# Warum ist Modellbildung wichtig?

- ▶ Abstraktion, die unwichtige Details ausser Acht lässt
  - ▶ z.B. Abweichungen bei einzelnen Messwerten eines Samples
- ▶ Präzise Beschreibung eines Phänomens anhand weniger Kennwerte: Art der Verteilung und passende Parameter
  - ▶ z.B. Normalverteilung mit Mittelwert und Standardabweichung)
- ▶ Anwendung mathematischer Methoden, um weitere Informationen und Vorhersagen abzuleiten

# Zufallsvariablen und deren Eigenschaften

- ▶ Zufallszahlen werden mit Grossbuchstaben bezeichnet ( $X, Y, Z, \dots$ ), ihre jeweiligen Realisierungen mit entsprechenden Kleinbuchstaben.
- ▶ Eine Zufallsvariable ist *diskret*, falls sie nur abzählbar viele Werte annimmt.
- ▶ Eine *kontinuierliche* Zufallszahlen kann hingegen jeden beliebigen Wert aus einem Intervall in  $\mathbb{R}$  annehmen.

# Reelle Zufallsvariablen

## Definition (Reelle Zufallsvariablen)

Eine reelle Zufallsvariable ist eine Funktion  $X : \Omega \longrightarrow \mathbb{R}$ , die jedem Ereignis  $\omega \in \Omega$  eine reelle Zahl  $X(\omega)$  zuordnet, wobei die Menge aller Ereignisse, deren Realisierung unterhalb eines bestimmten Wertes liegt, ein Ereignis bilden muss.

- ▶ Nicht-reelle Zufallsvariablen können als Zahlenwerte *kodiert* werden.
- ▶ Kontinuierliche Zufallsvariablen werden in der Praxis bisweilen diskretisiert, da nur endlich viele Nachkommastellen erfasst werden können.

# Verteilungsfunktion

## Definition (Verteilungsfunktion)

Sei  $X$  eine Zufallsvariable und  $F^X : \mathbb{R} \rightarrow [0, 1]$  eine monoton steigende, rechtseitig stetige Funktion.  $F^X$  heißt Verteilungsfunktion von  $X$  falls

- ▶  $F^X(x)$  die Wahrscheinlichkeit dafür angibt, dass  $X$  einen Wert kleiner  $x$  realisiert, d.h.  $F^X(x) = P(X \leq x)$
- ▶  $\lim_{x \rightarrow -\infty} F^X(x) = 0$
- ▶  $\lim_{x \rightarrow +\infty} F^X(x) = 1$



# Diskrete Dichtefunktionen

## Definition

Sei  $X$  eine diskrete Zufallsvariable. Die Funktion  $f^X : \mathbb{R} \rightarrow [0, 1]$ , für die gilt, dass

$$F^X(x) = \sum_{x_i \leq x} f^X(x_i),$$

ist die *diskrete Dichte* von  $X$ .

Für die Wahrscheinlichkeit eines Ereignisses  $\{X \in A\}$ ,  $A \subseteq \mathbb{R}$  gilt

$$P(X \in A) = \sum_{x_i \in A} f^X(x_i) = \sum_{x_i \in A} P(X = x_i)$$

## Diskrete Dichtefunktionen: Beispiel

- ▶ Die *Binomialverteilung* beschreibt die Anzahl der Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben (Erfolg vs. Misserfolg).
- ▶ Sei  $p \in [0, 1]$  die Erfolgswahrscheinlichkeit,  $n$  die Gesamtanzahl der Versuche und  $k$  die Anzahl der Erfolge.
- ▶ Dichtefunktion

$$P(X = k) = B_{n,p}(k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- ▶ Verteilungsfunktion

$$F^X(x) = P(X \leq x) = \sum_{k=0}^x \binom{n}{k} p^k (1 - p)^{n-k}$$

# Stetige Dichtefunktionen

## Definition

Sei  $X$  eine kontinuierliche Zufallsvariable. Die Funktion  $f^X : \mathbb{R} \rightarrow [0, \infty)$ , für die gilt, dass

$$F^X(x) = \int_{-\infty}^x f^X(t) dt,$$

ist die *stetige Dichte* von  $X$ .

Für die Wahrscheinlichkeit eines Ereignisses  $\{X \in A\}$ ,  $A \subseteq \mathbb{R}$  gilt

$$P(X \in A) = \int_A f^X(t) dt$$

## Beispiel: Normalverteilung

- ▶ Die *Normal-/Gaußverteilung* beschreibt eine Vielzahl von naturwissenschaftlichen Vorgänge und Effekte.
- ▶ Sei  $X$  eine  $N(\mu, \sigma^2)$ -verteilte Zufallsvariable, wobei  $\mu$  den Mittelwert und  $\sigma^2$  die Varianz bezeichnet.
- ▶ Dichtefunktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- ▶ Verteilungsfunktion

$$F^X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

## Beispiel: Exponentialverteilung

- ▶ Die *Exponentialverteilung* modelliert die Größe von Zeitintervallen bis zum Eintreffen eines Ereignisses.
- ▶ Sei  $X$  eine  $Exp(\lambda)$ -verteilte Zufallsvariable, wobei  $\lambda \in \mathbb{R}_{\geq 0}$  die Ereignisrate und  $\frac{1}{\lambda}$  den mittleren Ereignisabstand beschreibt.
- ▶ Dichtefunktion

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- ▶ Verteilungsfunktion

$$F^X(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

# Wahrscheinlichkeitsverteilungen in R

Verteilung	in R	Parameter
Normal-	-norm()	mean, sd
Binomial-	-binom()	size, prob
Exponential-	-exp()	rate
Gleich-	-unif()	min, max
Poisson-	-pois()	lambda

sowie viele weitere.

Präfixe:

- ▶ `r` liefert Zufallszahlen, die der spezifizierten Verteilungsfunktion folgen. Erster Parameter ist hierbei der Umfang der Stichprobe.
- ▶ `d` gibt den Wert der Dichtefunktion, `p` den Wert der Verteilungsfunktion an der jeweiligen Stelle.
- ▶ `q` bestimmt die Quantile für Werte zwischen 0 und 1.

# Zufallszahlen in R

Beispiel: Standardnormalverteilung

- ▶ Ziehen von n Zufallszahlen: `rnorm(n, mean=0, sd=1)`
- ▶ Dichte im Wert x: `dnorm(x, mean=0, sd=1)`  
> `dnorm(c(-1,0,1))`  
0.24197 0.39894 0.24197
- ▶ Verteilungsfunktion im Wert x: `pnorm(x, mean=0, sd=1)`  
> `pnorm(c(-1,0,1))`  
0.15866 0.50000 0.84134
- ▶ Quantil für Wahrscheinlichkeit p: `qnorm(p, mean=0, sd=1)`  
> `qnorm(c(0.25,0.5,0.75))`  
-0.67449 0.00000 0.67449

## Plotten von Verteilungen in R

Beispiel: Standardnormalverteilung

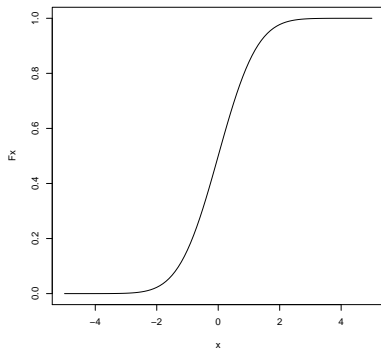
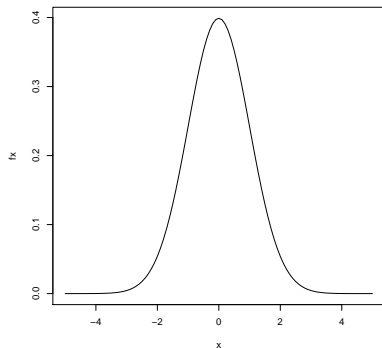
```
x <- seq(from = -10, to = 10, by = 0.1)
```

```
fx <- dnorm(x)
```

```
plot(x,fx,type="l")
```

```
Fx <- pnorm(x)
```

```
plot(x,Fx,type="l")
```





# Kenngößen von Zufallsvariablen

## Definition (Erwartungswert)

Sei  $X$  eine Zufallsvariable und  $f^X$  ihre Dichtefunktion. Der *Erwartungswert*  $E[X]$  von  $X$  ist definiert als

- ▶  $E[X] = \sum_{x_i} x_i f^X(x_i)$  für diskrete Zufallsvariablen
- ▶  $E[X] = \int_{-\infty}^{+\infty} x f^X(x) dx$  für diskrete Zufallsvariablen

## Definition (Varianz)

Die *Varianz* von  $X$  ist definiert durch

$$\sigma^2 = \text{Var}[X] = E[(X - E[X])^2]$$

# Punktschätzer

- ▶ Ein *Punktschätzer* schätzt eine relevante Kenngröße einer Verteilung durch die Angabe eines einzelnen Wertes.
- ▶ Beispiel: Ein Schätzer für den Erwartungswert einer Zufallsvariablen, basierend auf einer Stichprobe  $x_1, \dots, x_n$ , ist das arithmetische Mittel  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  (in R: `mean()` ).
- ▶ Beispiel: Ein Schätzer für die Varianz einer Zufallsvariablen ist  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  (in R: `var()` ).

# Intervallschätzer

- ▶ Punktschätzer enthalten keine Information über ihre Genauigkeit.
- ▶ Sei  $I_\alpha \subset \mathbb{R}$  ein Intervall, für das gilt:

$$P(\lambda \in I_\alpha) = 1 - \alpha$$

wobei  $\lambda$  die zu schätzende Kenngröße bezeichnet. Dann heißt  $I_\alpha$  ein Konfidenzintervall zum Niveau  $1 - \alpha$  für  $\lambda$ .

- ▶ D.h. Mit Wahrscheinlichkeit  $1 - \alpha$  liegt der Parameter  $\lambda$  im geschätzten Intervall  $I_\alpha$ .

## Intervallschätzer: Beispiel

- ▶ Für  $N(\mu, \sigma^2)$ -verteilte Zufallsvariablen ist das Konfidenzintervall zum Niveau  $1 - \alpha$  für den Erwartungswert  $\mu$  gegeben durch:

$$I_\alpha = \left[ \bar{x} - t_{1-\frac{\alpha}{2}, n-1} \frac{S_n}{\sqrt{n}}; \bar{x} + t_{1-\frac{\alpha}{2}, n-1} \frac{S_n}{\sqrt{n}} \right]$$

wobei  $t_{1-\frac{\alpha}{2}, n-1}$  das  $1 - \frac{\alpha}{2}$ -Quantil der Studentischen  $t$ -Verteilung mit  $n - 1$  Freiheitsgraden bezeichnet.

- ▶ Berechnung in R für  $\alpha = 0.05$ :
  - > `konf.level <- qt(0.975, length(x) - 1)`  
    \* `(sd(x)) / sqrt (length(x))`
  - > `lower <- mean(x) - konf.level`
  - > `upper <- mean(x) + konf.level`

# Testtheorie

# Hypothesen

- ▶ Bei einem statistischen Testproblem stellt man eine Behauptung auf (*Nullhypothese*,  $H_0$ ), deren Gültigkeit mit Hilfe statistischer Verfahren überprüft werden soll.
- ▶ Die entgegengesetzte Aussage bezeichnet man als *Alternativhypothese*  $H_1$ .
- ▶ Nur die Ablehnung von  $H_0$  stellt eine statistisch verlässliche Entscheidung dar, nicht aber die Annahme der Nullhypothese!

## Beispiel: Zweiseitige Hypothesen

Besitzen Männer und Frauen in Deutschland unterschiedliche Intelligenzquotienten?

- ▶ Sei  $IQ_m$  der Intelligenzquotient deutscher Männer,  $IQ_w$  der Intelligenzquotient deutscher Frauen.
- ▶ Formalisiertes Testproblem:

$$H_0 : IQ_m = IQ_w \quad \text{vs.} \quad H_1 : IQ_m \neq IQ_w$$

- ▶ Da die Alternativhypothese aus zwei Möglichkeiten besteht (IQ der Frauen ist größer oder kleiner), spricht man von einer *zweiseitigen* Hypothese.

# Einseitige Hypothesen

Ist ein neues Medikament besser als ein bereits zugelassenes Medikament?

- ▶ Sei  $q$  die Heilungswahrscheinlichkeit für das bereits zugelassene Medikament,  $p$  die Heilungswahrscheinlichkeit für das neue Medikament.
- ▶ Das Testproblem lautet dann:

$$H_0 : p \leq q \quad \text{vs.} \quad H_1 : p > q$$

- ▶ Hierbei handelt es sich um eine *einseitige* Hypothese.



## Fehler 1. und 2. Art

- ▶ Bei einem *Fehler 1. Art* wird die Nullhypothese irrtümlicherweise verworfen.
- ▶ Bei einem *Fehler 2. Art* wird die Nullhypothese irrtümlicherweise beibehalten.
- ▶ Die Wahrscheinlichkeit für einen Fehler 1. Art bezeichnet man als *Signifikanzniveau*  $\alpha$  des Tests.

## p-Wert

- ▶ Zur Überprüfung einer Nullhypothese berechnet man mit Hilfe der Stichprobe eine *Prüfgröße / Teststatistik*  $T(x_1, \dots, x_n)$ , an Hand derer eine Entscheidung getroffen wird.
- ▶ Der *p-Wert* ist die Wahrscheinlichkeit dafür, dass man unter der Nullhypothese  $H_0$  die ermittelte Teststatistik beobachtet.
- ▶ Je kleiner der p-Wert, desto unwahrscheinlicher ist die Gültigkeit von  $H_0$ .
- ▶ Häufig wird die Nullhypothese  $H_0$  abgelehnt bei einem p-Wert  $\leq 0.05$  (bzw. 5%).

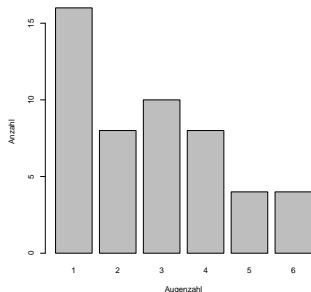
# Testentscheidungen

1. Die Hypothese, deren Gültigkeit man zeigen will, muss als Alternativhypothese  $H_1$  formuliert werden.
2. Je nach Testergebnis:
  - ▶ Ist  $p \leq 0.05$ , wird die Nullhypothese  $H_0$  verworfen, und man entscheidet sich für  $H_1$  mit Irrtumswahrscheinlichkeit von 5%.
  - ▶ Ist  $p > 0.05$ , ist keine Testentscheidung möglich!

## $\chi^2$ -Verteilungstest

- ▶ Der  $\chi^2$ -Verteilungstest testet, ob Daten einer bestimmten Wahrscheinlichkeitsverteilung folgen.
- ▶ Voraussetzungen: Sei  $x_1, \dots, x_n$  eine Stichprobe bestehend aus Realisationen von unabhängigen und identisch verteilten Zufallsvariablen, deren Wertebereich aus  $J$  Kategorien besteht. Für jede dieser Kategorien muss gelten, dass die erwartete Häufigkeit in jeder Kategorie mindestens 5 beträgt:  $E_j \geq 5$  für  $j = 1, \dots, J$ .
- ▶ Der  $\chi^2$ -Verteilungstest überprüft, ob die einzelnen Kategorien in einer vorgegebenen Häufigkeit auftreten.

## Beispiel: Gezinkter Würfel



- ▶ Gegeben eine Messreihe mit 50 Ergebnissen von Würfelswürfen mit einem Würfel
- ▶ Vermutung: Würfel ist gezinkt!
- ▶ Nullhypothese  $H_0$ : Die erwarteten Häufigkeiten sind für alle Augen gleich.

## Beispiel: Gezinkter Würfel

- ▶ Absolute Häufigkeiten

```
> table(wuerfel$augenzahl)
 1  2  3  4  5  6
16  8 10  8  4  4
```

- ▶ Beobachtete relative Häufigkeiten

```
> prop.table(table(wuerfel$augenzahl))
 1    2    3    4    5    6
0.32 0.16 0.20 0.16 0.08 0.08
```

- ▶ Erwartete relative Häufigkeiten:  $p = \text{rep}(1/6, 6)$

```
 1    2    3    4    5    6
1/6  1/6  1/6  1/6  1/6  1/6
```

## Beispiel: Gezinkter Würfel

```
> chisq.test(table(wuerfel$augenzahl),p = rep(1/6,6))
```

Chi-squared test for given probabilities

```
data: table(wuerfel$augenzahl)
```

```
X-squared = 11.92, df = 5, p-value = 0.0359
```

- ▶ Da der p-Wert  $< 0.05$ , kann  $H_0$  auf dem 5%-Signifikanzniveau verworfen werden.

# Binomialtest

- ▶ Voraussetzungen: Sei  $x_1, \dots, x_n$  eine Stichprobe bestehend aus Realisationen von unabhängigen Wiederholungen eines Zufallsexperiments, deren Wertebereich aus zwei Kategorien besteht.
- ▶ Der Binomialtest überprüft, ob die beiden Kategorien in einer vorgegebenen Häufigkeit auftreten.
- ▶ Nullhypothese  $H_0$ : Die erwartete Häufigkeit  $q$  für die erste Kategorie beträgt (maximal)  $q_0$ .

$$H_0 : q = q_0 \quad \text{bzw.} \quad H_0 : q \leq q_0$$



## Beispiel: Füllmengen in Verpackungen

- ▶ Ein Hersteller von Gummibärchen garantiert, dass höchstens 2.5% seiner Verpackungen von der gekennzeichneten Füllmenge abweichen.
- ▶ Ein Verbrauchermagazin will diese Behauptung überprüfen und misst nach. Bei 19 von 540 Verpackungen weicht das Gewicht um mehr als die zulässige Schwankung vom vorgesehenen Gewicht ab.
- ▶ In der Stichprobe verstoßen 3.5% der Verpackungen gegen die Spezifikation, aber ist diese Abweichung in der Stichprobe signifikant?

## Beispiel: Füllmengen in Verpackungen

```
> binom.test(19,540,0.025,alternative="greater")
Exact binomial test

data: 19 and 540
number of successes = 19, number of trials = 540,
  p-value = 0.08892
alternative hypothesis: true probability of success
  is greater than 0.025
95 percent confidence interval:
 0.02316077 1.00000000
sample estimates:
probability of success
 0.03518519
```

- ▶ `alternative="greater"` testet die Hypothese, dass der Anteil der abweichenden Stichproben größer als 0.025 ist.
- ▶  $H_0$  kann auf dem 5%-Signifikanzniveau nicht verworfen werden.
- ▶ Dem Hersteller kann auf Grund dieser Stichprobe kein Betrug vorgeworfen werden.

## Mittelwertsvergleich mit t-Test

- ▶ Der t-Test überprüft, ob sich die Mittelwerte zweier Gruppen normalverteilter Zufallsvariablen voneinander unterscheiden.
- ▶ In R: `t.test(x, y, alternative, paired, var.equal)`
- ▶ `x,y`: zu vergleichende Daten
- ▶ `alternative=c("two.sided", "less", "greater")`: Varianten für die Alternativhypothese
- ▶ `var.equal = TRUE`: Gibt an, ob Varianzgleichheit bei den Populationen vorliegt
- ▶ `paired`: Gibt an, ob `x` und `y` als gepaarte Stichprobe anzusehen sind

## Beispiel: Nettokaltmieten

- ▶ Unterscheiden sich die Nettokaltmieten pro  $m^2$  bei Ein- und Zweizimmerwohnungen?
- ▶ Gegeben sei folgende Stichprobe:

EZ	8.70	11.28	13.24	8.37	12.16	11.04	10.47	11.16	4.28	19.54
ZZ	3.36	18.35	5.19	8.35	13.10	15.65	4.29	11.36	9.09	

```
t.test(EZ,ZZ, var.equal = FALSE, paired = FALSE)
Welch Two Sample t-test
data: EZ and ZZ
t = 0.5471, df = 14.788, p-value = 0.5925
alternative hypothesis: true difference in means is
not equal to 0
```

## Beispiel: Schwangerschaftsdauer

- ▶ Kommen Erstgeborene nach längerer Schwangerschaftsdauer zur Welt als ihre Geschwister?
- ▶ Aus dem Datensatz der NSFG (siehe frühere Vorlesung):

```
> t.test(firstborns$duration, laterborns$duration,  
         var.equal = FALSE, paired = FALSE, alternative="g")
```

Welch Two Sample t-test

```
data: firstborns$duration and laterborns$duration  
t = 1.3771, df = 8981.579, p-value = 0.08426  
alternative hypothesis: true difference in means is greater than 0  
95 percent confidence interval:  
 -0.01518543      Inf  
sample estimates:  
mean of x mean of y  
 38.60095  38.52291
```

# Test auf Normal-Verteilung bei metrischen Daten

- ▶ Bei vielen Testverfahren wird vorausgesetzt, dass die zu untersuchende Zufallsvariable normalverteilt ist.
- ▶ Der Shapiro-Wilk-Test (`shapiro.test()`) wird häufig als Vortest angewandt. Er bewertet die folgende Nullhypothese:  
 $H_0$  : Die Zufallsvariable ist  $N(\mu, \sigma^2)$ -verteilt, wobei  $\mu \in \mathbb{R}, \sigma^2 > 0$  beliebig sind.
- ▶ Voraussetzungen:
  - ▶ Beobachtungen müssen unabhängig voneinander sein
  - ▶ Stichprobengröße  $3 < n < 5000$
- ▶ Über den Test kann nur entschieden werden, ob die Daten *nicht* normalverteilt sind!
- ▶ Es ist hilfreich darüber hinaus grafische Hilfsmittel zu verwenden, um die Verteilung der Daten zu approximieren.
- ▶ Typische Visualisierungen: Histogramme, Box-Plots, Q-Q-Diagramme

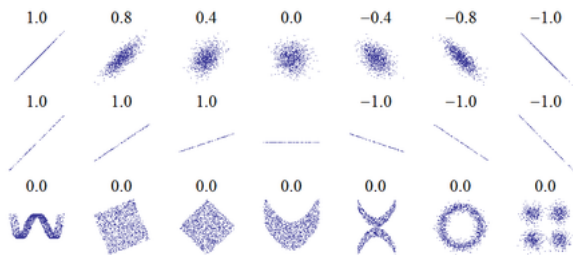
# Korrelation

# Korrelation

- ▶ Eine Korrelation beschreibt eine Beziehung zwischen zwei oder mehreren Merkmalen, Ereignissen oder Größen.
- ▶ Der Korrelationskoeffizient ist ein dimensionsloses Maß für den Grad des *linearen Zusammenhangs* zwischen zwei (mindestens intervallskalierbaren) Merkmalen.

Pearson'scher Korrelationskoeffizient:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)} \quad \text{mit} \quad \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$





## Korrelationstest auf metrischen Daten

- ▶ Mit der Funktion `cor.test(data1,data2)` kann man verschiedene Korrelationskoeffizienten und deren Signifikanz berechnen.
- ▶ Standardmässig wird der Pearson'sche Korrelationskoeffizient verwendet.
- ▶ Beispiel: Korrelation zwischen Sonneneinstrahlung und Temperatur bei Luftqualitätsmessungen

```
> cor.test(airquality$Solar, airquality$Temp)
```

```
      Pearson's product-moment correlation  
data:  airquality$Solar and airquality$Temp  
t = 3.4437, df = 144, p-value = 0.0007518  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 0.1187113 0.4194913  
sample estimates:  
      cor  
0.2758403
```

## Korrelation für kategorielle Daten

- ▶ Eine *Kontingenztafel* listet die Anzahl der Beobachtungen in den Schnittmengen der jeweiligen Kategorien.

	1	2	...	J	Summe
1	$O_{11}$	$O_{12}$	...	$O_{1J}$	$O_{1\bullet}$
2	$O_{21}$	$O_{22}$	...	$O_{2J}$	$O_{2\bullet}$
...					
I	$O_{I1}$	$O_{I2}$	...	$O_{IJ}$	$O_{I\bullet}$
Summe	$O_{\bullet 1}$	$O_{\bullet 2}$	...	$O_{\bullet J}$	$n$

## Beispiel: Haar- und Augenfarbe

Kontingenztafel basierend auf einer Befragung von 592 Personen

	blau	braun	gruen	nuss	gesamt
blond	94	7	16	10	127
braun	84	119	29	54	286
rot	17	26	14	14	71
schwarz	20	68	5	15	108
gesamt	215	220	64	93	592

## $\chi^2$ - Unabhängigkeitstest

- ▶ Voraussetzungen: Von einem Paar  $(X, Y)$  von Zufallsvariablen liegt eine Stichprobe  $(x_1, y_1), \dots, (x_n, y_n)$  von unabhängigen Wiederholungen vor. Die erwarteten Häufigkeiten  $E_{ij}$  müssen mindestens den Wert 5 betragen, wobei diese wie folgt aus der Kontingenztabelle ermittelt werden können:

$$E_{ij} = \frac{O_{i\bullet} \cdot O_{\bullet j}}{n}$$

## Beispiel: Haar- und Augenfarbe

```
> chisq.test(data$haar, data$auge)
```

Pearson's Chi-squared test

data: ha\$haar and ha\$auge

X-squared = 138.2898, df = 9, p-value < 2.2e-16

- ▶ Nullhypothese  $H_0$ : Haar- und Augenfarbe sind unabhängig voneinander.
- ▶ Da der p-Wert  $< 0.05$ , kann die Nullhypothese verworfen werden.

# Korrelation und Kausalzusammenhang

## Beispiele

- ▶ Je mehr Eiscreme in einem Monat verkauft wird, desto höher ist auch die Rate von Ertrunkenen am Meer und Badeseen. Daher impliziert der Genuss von Eiscreme eine erhöhte Gefahr von Badeunfällen.
- ▶ Je mehr Feuerwehrleute ein Feuer bekämpfen, desto größer ist das Feuer. Der verstärkte Einsatz von Löschkräften führt also zu größeren Brandschäden.
- ▶ Aus der Korrelation zweier Ereignissen lässt sich nicht ableiten, dass eines der Ereignisse das andere bedingt.
- ▶ Beide könnten auch durch eine dritte Größe bedingt sein oder inhaltlich unabhängig sein (*Scheinkorrelation*, Beispiel: Zusammenhang zwischen Geburtenzahlen und Vorkommen von Störchen).
- ▶ Lesehinweis: [http://en.wikipedia.org/wiki/Correlation\\_does\\_not\\_imply\\_causation](http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation)

## Ausblick: Regressionsanalyse

- ▶ Mathematische Modelle zur genaueren Charakterisierung des Zusammenhangs zweier Faktoren
- ▶ Einfaches lineares Regressionsmodell:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

- ▶ Residuum  $\epsilon$  umfasst die “Zufallskomponente” in der Beobachtung
- ▶ Dabei wird angenommen, dass  $\epsilon$  normalverteilt unter  $N(0, \sigma^2)$  ist.
- ▶ Parameter  $\beta_0$  und  $\beta_1$  werden mit Hilfe der Stichprobe geschätzt, indem beispielsweise die Residuenquadratsumme minimiert wird.
- ▶ In R: `lm(data1 ~ data2)`