Kapitel 6
Schließende Statistik

Thorsten Dickhaus

Universität Bremen
Institut für Statistik

Mathematik 3: Stochastik
Universität Bremen, Fachbereich 03, SoSe 2025

Einleitung

$Y : \Omega \to \mathcal{Y}$ Zufallsgröße, die das (zufällige) Ergebnis eines Experiments beschreibt.

Statistisches Modell:
$(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta \in \Theta})$

$\mathcal{Y}$ heißt der Stichprobenraum des Experiments.
θ heißt der Parameter des Modells,
Θ heißt der Parameterraum.

Das Tupel $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ heißt ein messbarer Raum.

Der Wert von θ ist unbekannt und unbeobachtbar.
Er soll aus den Daten $Y = y$ inferiert werden.

Illustratives Beispiel

Bernoulli'sches Versuchsschema:

$X = (X_i)_{i=1,\dots,n}$, $X_i \sim$ Bernoulli(p) für alle $1 \leq i \leq n$,
X_i stoch. unabhängige Zufallsvariablen mit Werten in $\{0, 1\}$,
$p \in (0, 1)$ Trefferw'keit, $n \in \mathbb{N}$,

$$Y = \sum_{i=1}^{n} X_i$$

$\mathcal{Y} = \{0, \dots, n\}$, $\mathcal{B}(\mathcal{Y}) = 2^{\mathcal{Y}}$,
$\theta = (n, p) \in \mathbb{N} \times [0, 1] = \Theta$,
$\mathbb{P}_\theta$ ist die Binomialverteilung mit Parametern n und p.

Übersicht

Übersicht

Schätzer und ihre Eigenschaften

$\hat{\theta} : \mathcal{Y} \to \Theta$ heißt eine Schätzvorschrift.

Die Zufallsvariable $\hat{\theta}(Y)$ heißt ein Schätzer,
$\hat{\theta}(y)$ heißt der Schätzwert für θ
basierend auf der Stichprobe (Realisierung) $Y = y$.

Definition: (Eigenschaften eines Schätzers $\hat{\theta}(Y)$)

- $\hat{\theta}(Y)$ *heißt unverzerrt bzw. erwartungstreu, falls* $\forall \theta \in \Theta : \mathbb{E}_\theta[\hat{\theta}(Y)] = \theta$.
- $\hat{\theta}(Y)$ *heißt effizient, falls* $Var_\theta(\hat{\theta}(Y))$ *minimal ist (innerhalb einer gegebenen Klasse von Schätzern).*

Mittlerer quadratischer Schätzfehler, Bias-Varianz-Zerlegung

Sei $\Theta \subseteq \mathbb{R}$. Dann heißt

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$$

mittlerer quadratischer Schätzfehler (mean squared error) von $\hat{\theta}$ bzw. von $\hat{\theta}(Y)$.

Es gilt:

$$\begin{aligned}
\mathrm{MSE}_\theta(\hat{\theta}) &= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}] + \mathbb{E}_\theta[\hat{\theta}] - \theta)^2] \\
&= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2] + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2 \\
&\quad + \underbrace{2(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta)}_{=0}
\end{aligned}$$

Also: $\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2$

Verzerrung bzw. Bias

Der Wert

$$b(\hat{\theta}, \theta) = \mathbb{E}_\theta[\hat{\theta}] - \theta$$

heißt die Verzerrung bzw. der Bias von $\hat{\theta}$ bzw. von $\hat{\theta}(Y)$.

Falls $\hat{\theta}$ unverzerrt ist, so gilt (gleichbedeutenderweise):

$$\forall \theta \in \Theta : b(\hat{\theta}, \theta) = 0$$

Für einen unverzerrten Schätzer eines univariaten Parameters gilt also:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta})$$

Erinnerung: Momente

Definition:

Falls es existiert, so ist das k-te Moment einer (reellwertigen) Zufallsvariable X definiert als

$$m_k(X) = \mathbb{E}[X^k] = \begin{cases} \sum_{x \in \mathcal{X}} x^k f_X(x), & X \text{ diskret verteilt,} \\ \int_{\mathcal{X}} x^k f_X(x) dx, & X \text{ stetig verteilt.} \end{cases}$$

Dabei bezeichnet $\mathcal{X}$ den Träger von X und f_X bezeichnet die Zähl- bzw. Lebesguedichte von X.

Momenten- bzw. plug-in Schätzer

Angenommen, es liegen beobachtbare i.i.d. Zufallsvariablen $Y_1, \ldots, Y_n$ vor, wobei $m_k \equiv m_k(Y_1)$ für ein gegebenes $k \in \mathbb{N}$ existiert.

Dann ist der Momenten- bzw. plug-in Schätzer $\hat{m}_k$ für m_k gegeben durch

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} Y_i^k.$$

(k-tes Moment der induzierten empirischen Verteilung)

Momentenschätzer: Beispiele

Angenommen, Y_1 besitzt ein endliches zweites Moment.

Dann erhalten wir die folgenden Momentenschätzer:

1. $\hat{m}_1 = \bar{Y}_n = \sum_{i=1}^{n} Y_i/n$ (arithmetisches Mittel)
2. $\hat{m}_2 = \sum_{i=1}^{n} Y_i^2/n$
3. $\hat{\sigma}^2 = \hat{m}_2 - \hat{m}_1^2$, denn $\mathrm{Var}(Y_1) = \mathbb{E}[Y_1^2] - \mathbb{E}^2[Y_1]$
   ("plug-in"-Methode, unkorrigierte Stichprobenvarianz)

Verallgemeinerte Momentenmethode (I)

Sei $\Theta \subseteq \mathbb{R}$ der Parameterraum eines statistischen Modells.

Sei $g : \mathbb{R} \to \mathbb{R}$ eine Funktion, so dass das "Moment"

$$m_g(\theta) = \mathbb{E}_\theta \left[g(Y_1) \right]$$

für alle $\theta \in \Theta$ existiert.

Ferner angenommen, die so definierte Funktion $m_g : \Theta \to \mathbb{R}$ ist stetig und umkehrbar. Dann gilt:

$$\theta = m_g^{-1} \left(\mathbb{E}_\theta \left[g(Y_1) \right] \right)$$

Verallgemeinerte Momentenmethode (II)

Auf der Basis der vorangegangenen Überlegungen liefert die plug in-Methode den Schätzer

$$\hat{\theta} = m_g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(Y_i)\right).$$

Häufig wird $g(y) = y$ oder $g(y) = y^2$ gewählt.

Verallgemeinerte Momentenmethode: Beispiel

Angenommen, Y_1 ist exponentialverteilt mit unbekanntem Intensitätsparameter $\theta > 0$.

Dann gilt gemäß Kapitel 4:

$$\mathbb{E}_\theta[Y_1] = \frac{1}{\theta}$$

Wählen wir nun die durch $g(y) = y$ gegebene Momentenfunktion g, so ergibt sich als plug in- Schätzer

$$\hat{\theta} = \frac{1}{\bar{Y}_n} = \frac{n}{\sum_{i=1}^n Y_i}.$$

Momentenschätzer sind nicht immer unverzerrt!

Die unkorrigierte Stichprobenvarianz $\hat{\sigma}^2$ ist Momentenschätzer für $\sigma^2 = \text{Var}(Y_1)$ im i.i.d.-Modell.

Allerdings ist

$$\mathbb{E}_{\sigma^2}\left[\hat{\sigma}^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Somit ist $\hat{\sigma}^2$ verzerrt. In der Praxis wird daher die korrigierte Stichprobenvarianz

$$\tilde{\sigma}^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$

oftmals bevorzugt, da sie σ^2 erwartungstreu schätzt.

Definition: (Likelihoodfunktion)

Gegeben: Statistisches Modell $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$

Angenommen, Zufallsvariablen $(Y_i)_{i=1,\ldots,n}$ *mit Realisierungen* $(y_i)_{i=1,\ldots,n}$ *liegen vor.*

Dann definieren wir

$$Z(\mathbf{y}, \theta) = \prod_{i=1}^{n} p(y_i, \theta), \quad \mathbf{y} = (y_1, \ldots, y_n).$$

$p(\cdot, \theta)$*: Zähl- oder Lebesguedichte von* Y_1 *unter* $\theta \in \Theta$*.*

$Z(\mathbf{y}, \theta)$ ist (eine Approximation für) die W'keit, die Stichprobe $\mathbf{y} = (y_1, \ldots, y_n)$ unter $\theta \in \Theta$ zu beobachten.

Maximum-Likelihood-Schätzer (MLE)

Der MLE für θ ist gegeben durch

$$\hat{\theta}(\mathbf{y}) = \arg\max_{\theta \in \Theta} Z(\mathbf{y}, \theta).$$

In vielen Modellen ist es (analytisch und numerisch) einfacher, die äquivalente Formulierung

$$\hat{\theta}(\mathbf{y}) = \arg\max_{\theta \in \Theta} L(\mathbf{y}, \theta),$$

zu verwenden, mit $L(\mathbf{y}, \theta) = \log\left(Z(\mathbf{y}, \theta)\right) = \sum_{i=1}^{n} \log(p(y_i, \theta))$.

Oftmals ist garantiert, dass $L(\mathbf{y}, \cdot)$ strikt konkav ist, so dass der MLE dann eindeutig bestimmt ist.

MLE: Beispiel (I)

Wir betrachten das Binomialmodell mit fest vorgegebenem Stichprobenumfang $n = 10$ und unbekannter Trefferw'keit p.

Angenommen, es werden genau neun "Treffer" beobachtet, also: $y = 9$.

$$\begin{aligned} Z(y = 9, p) &= p^9(1-p) \\ L(y, p) &= 9\ln(p) + \ln(1-p) \\ \frac{dL(y,p)}{dp} &= \frac{9}{p} - \frac{1}{1-p} \end{aligned}$$

MLE: Beispiel (II)

Wir bestimmen die Nullstelle der Ableitung:

$$\begin{aligned} \frac{dL(y,p)}{dp} &= 0 \\ \iff \frac{9}{p} &= \frac{1}{1-p} \\ \iff 9 - 9 \cdot p &= p \\ \iff 9 &= 10 \cdot p \end{aligned}$$

Somit erhalten wir den ML-Schätzwert $\hat{p}(y) = 9/10$.

Kleinste Quadrate (KQ)-Methode

Modellannahme:

Jeder Datenpunkt y_i ist additiv zusammengesetzt aus einer Funktion $f_i(\theta)$ des Parameters von Interesse plus einer zufälligen Störung (Rauschen).

Definition:

Die Fehlerquadratsumme ist dann gegeben durch

$$S(\theta) = \sum_{i=1}^{n} (y_i - f_i(\theta))^2.$$

KQ-Methode: Minimiere $S(\theta)$ bzgl. θ !

KQ-Methode: Beispiel (I)

Im Falle der einfachen linearen Regression ist

$$f_i(\theta) = f_i(\alpha, \beta) = \alpha + \beta x_i$$

für die Ausprägung $x_i \in \mathbb{R}$ eines Regressors für Beobachtungseinheit i.

Wir beobachten also Merkmalspaare $(x_1, y_1), \ldots, (x_n, y_n)$.

Es ergibt sich:

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

KQ-Methode: Beispiel (II)

$$\begin{aligned}
\frac{\partial S(\alpha,\beta)}{\partial \alpha} &= -2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i),\\
\frac{\partial S(\alpha,\beta)}{\partial \beta} &= -2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)x_i,\\
\\
\hat{\alpha} &= \bar{y}_n-\hat{\beta}\bar{x}_n,\\
\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}=\frac{s_{xy}}{s_x^2}.
\end{aligned}$$

Übersicht

Konfidenzschätzung

Sei ein statistisches Modell $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$ mit $\Theta \subseteq \mathbb{R}$ gegeben.

Definition:

Ein Konfidenzintervall $I = [\hat{\theta}_{\text{unten}}, \hat{\theta}_{\text{oben}}]$ *besteht aus zwei Abbildungen* $\hat{\theta}_{\text{unten / oben}} : \mathcal{Y} \to \Theta$*, so dass*

$$\forall\theta \in \Theta : \mathbb{P}_\theta(\hat{\theta}_{\text{unten}} \leq \theta \leq \hat{\theta}_{\text{oben}}) \geq 1 - \alpha$$

für ein gegebenes Konfidenzniveau $(1 - \alpha) \in (0, 1)$*.*

Idealerweise sollten die Funktionen $\hat{\theta}_{\text{unten}}$ und $\hat{\theta}_{\text{oben}}$ so gewählt werden, dass die Länge von I minimal ist.
(Informativität der Datenanalyse!)

Konfidenzintervalle unter (Gauß'scher) Normalverteilung

Sei $\hat{\theta}(Y)$ ein unverzerrter und normalverteilter Punktschätzer für θ mit endlicher Varianz σ^2.

Dann wird die (mittlere bzw. erwartete) Länge von I minimal, falls I symmetrisch um $\hat{\theta}$ liegt.

Wir rechnen:

$$\begin{aligned} \mathbb{P}_\theta(\hat{\theta} - c \leq \theta \leq \hat{\theta} + c) &= 1 - \alpha \\ \iff c &= \sigma\Phi^{-1}(1 - \alpha/2), \end{aligned}$$

wobei Φ die Vtfkt. der Standardnormalverteilung bezeichnet.

Für $\alpha = 5\%$ erhalten wir $c \approx 1.96\sigma \approx 2\sigma$. Deswegen wird diese Methode auch 2σ-Regel genannt.

Höhere Dimensionen (von Θ)

Falls Θ hoch-dimensional ist, also $\Theta \subseteq \mathbb{R}^k, k >> 1$,
so sind klassische Schätzmethoden oftmals suboptimal.

Einige moderne Schätzmethoden für diesen Fall sind:

- Regularisierte Schätzer
  (mit einem Strafterm für Komplexität)
- Schrumpfungs- bzw. Shrinkage-Schätzer
  (James-Stein, etc.)
- Statistische Lernverfahren (ERM, SRM, ...)

Übersicht

Definition: (Statistischer Test)

Y: Zufallsgröße mit Werten in $\mathcal{Y}$,
$(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), (\mathbb{P}_\theta)_{\theta\in\Theta})$ statistisches Modell

Aufgabe: Teste $H_0 : \theta \in \Theta_0$ gegen $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$
auf der Basis der Daten $Y = y$.

(Nicht-randomisierter) statistischer Test:

$\varphi : \mathcal{Y} \to \{0, 1\}$ mit der Konvention

$$\varphi(y) = 1 \iff: \quad H_0 \text{ wird zu Gunsten von } H_1 \text{ verworfen,}$$
$$\varphi(y) = 0 \iff: \quad H_0 \text{ wird beibehalten (also nicht verworfen).}$$

Randomisierter Test:

$\varphi : \mathcal{Y} \to [0, 1]$ liefert eine Ablehnwahrscheinlichkeit.

Entscheidungsstruktur, Bezeichnungen

$\varphi(y)$	H_0 wahr	H_0 falsch
1	Typ I-Fehler	Güte
0	Niveau	Typ II-Fehler

H_0 einfach (einelementig)	$\lvert\Theta_0\rvert = 1$
H_0 zusammengesetzt	$\lvert\Theta_0\rvert > 1$ (z. B. ein Intervall auf $\mathbb{R}$)

φ einseitig	$\Theta_1 \subseteq \mathbb{R}$ auf einer Seite beschränkt
φ zweiseitig	$\Theta_1 \subseteq \mathbb{R}$ beidseitig unbeschränkt

Teststatistik, Niveau α-Test, Gütefunktion

Eine Teststatistik ist eine Abbildung $T : \mathcal{Y} \to \mathbb{R}$.

Für vorgegebenes $\alpha \in [0, 1]$ sei eine Ablehnregion $\Gamma_\alpha \subset \mathbb{R}$ so, dass

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(T \in \Gamma_\alpha \right) \leq \alpha.$$

Dann definiert $\varphi(y) = \mathbf{1}_{\Gamma_\alpha}(T(y))$ einen Test zum Signifikanzniveau α.

Gütefunktion: $\beta_\varphi(\theta) = \mathbb{P}_\theta(T \in \Gamma_\alpha)$ für $\theta \in \Theta_1$

Optimale Signifikanztests

Definition:

Ein Test φ heißt gleichmäßig bester Niveau α-Test, falls für jeden anderen Niveau α-Test $\tilde{\varphi}$ gilt:

$$\forall \theta_1 \in \Theta_1 : \beta_{\tilde{\varphi}}(\theta_1) \leq \beta_{\varphi}(\theta_1)$$

Leider existieren gleichmäßig beste Niveau α-Tests nur in eingeschränkten Modellklassen bzw. nur für spezielle Typen von Testproblemen.

Lemma: (**Neyman-Pearson-Test**)

Betrachte zwei einfache Hypothesen $H_0 : \theta = \theta_0$ *und* $H_1 : \theta = \theta_1$.

Dann ist der Neyman-Pearson-Test, gegeben durch

$$\varphi(y) = \begin{cases} 0\,, & \text{falls } q(y) < c_\alpha \\ \gamma\,, & \text{falls } q(y) = c_\alpha, \\ 1\,, & \text{falls } q(y) > c_\alpha, \end{cases} \qquad \text{wobei } \; q(y) = \frac{Z(y,\theta_1)}{Z(y,\theta_0)}$$

den Likelihood-Quotienten bezeichnet, der beste Niveau α*-Test.*

Dabei werden $c_\alpha \in \mathbb{R}$ *und* $\gamma \in [0,1]$ *so gewählt, dass* $\mathbb{P}_{\theta_0}(\varphi = 1) = \alpha$ *gilt.*

Kritische Werte

Der Wert c_α heißt kritischer Wert des Tests.

Falls der Likelihood-Quotient monoton in einer Statistik T ist, so können wir äquivalenterweise kritische Werte für $T(y)$ angeben.

Ist nämlich z. B. $q(y) = f(T(y))$ für strikt isotones f, so ist

$$q(y) > c_\alpha \iff T(y) > f^{-1}(c_\alpha).$$

Somit ist dann $f^{-1}(c_\alpha)$ der kritische Wert für $T(y)$.

(Allgemeiner) Likelihood-Quotienten-Test

Für ein allgemeines Testproblem $H_0 : \theta \in \Theta_0$ gegen $H_1 : \theta \in \Theta_1$ kann der Likelihood-Quotienten (LQ)-Test verwendet werden, der als Teststatistik den Likelihood-Quotienten Λ verwendet.

Dieser ist gegeben durch

$$\Lambda(y) = \frac{\sup_{\theta \in \Theta} Z(y, \theta)}{\sup_{\theta \in \Theta_0} Z(y, \theta)} \in [1, \infty].$$

Große Werte von $\Lambda(y)$ sprechen für H_1.

p-Wert

Der zu einem gegebenen Test φ gehörige p-Wert bezeichnet das kleinste Signifikanzniveau $\alpha_{\min}(y)$, zu dem H_0 gerade noch abgelehnt werden kann, wenn $Y = y$ beobachtet wurde. Man spricht daher auch vom beobachteten Signifikanzniveau von φ gegeben $Y = y$.

Statistik-Software-Systeme geben typischerweise p-Werte statt expliziter Testentscheidungen aus.

Die Nullhypothese H_0 kann bei beobachteter Stichprobe $Y = y$ zu Gunsten von H_1 zum Niveau α verworfen werden, falls der entsprechende p-Wert kleiner ist als α.

Beispiel: Binomialtest

Angenommen, wir möchten prüfen, ob eine gewisse Münze "fair" ist oder nicht: $H_0 : p = p_0 = 0.5$ versus $H_1 : p \neq p_0$

```
> binom.test(15,20,p=0.5,alternative="two.sided")

        Exact binomial test

number of successes = 15, number of trials = 20, p-value = 0.04139
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval: 0.5089541 0.9134285
sample estimates: probability of success = 0.75
```

Dualität von Tests und Konfidenzbereichen:

Wir verwerfen $H_0 : p = p_0$ zum Niveau α, falls das $(1 - \alpha)$-Konfidenzintervall für p den Wert p_0 **nicht überdeckt.**

Einstichproben-t-Test

Seien $Y_1, \ldots, Y_n$ i. i. d. mit $Y_1 \sim \mathcal{N}(\mu, \sigma^2)$, für unbekannte Parameterwerte μ und σ^2.
$H_0 : \mu = \mu_0$ (oder auch einseitige Nullhypothese)

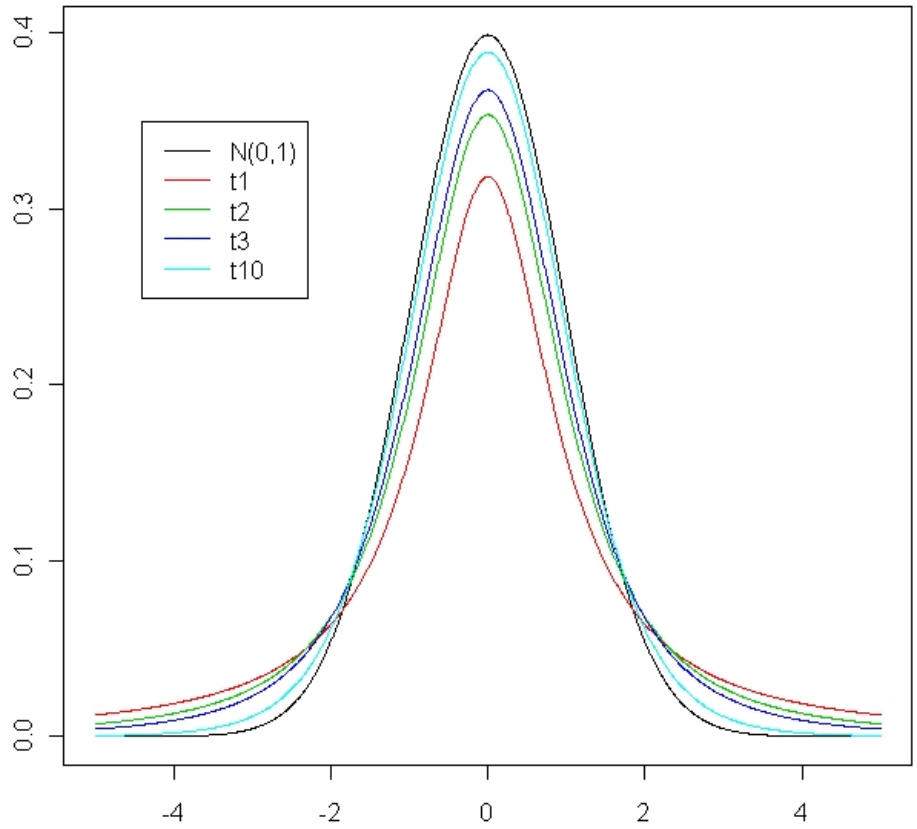
Teststatistik:

$$T(Y) = \frac{\sqrt{n(n-1)}(\bar{Y} - \mu_0)}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

es kann gezeigt werden, dass $|T(y)|$ (bzw. $T(y)$) eine monotone Transformation des Likelihood-Quotienten $\Lambda(y)$ ist.

Ferner gilt $T(Y) \underset{H_0}{\sim} t_{n-1}$ unabhängig von μ_0 und σ^2
(Student, 1908).

Dichten von Studentischen t-Verteilungen



Kritische Bereiche

Ablehnbereiche Γ_α der Niveau α-t-Tests
für ein- bzw. zweiseitige Hypothesen:

H_0	H_1	Γ_α
$\mu \leq \mu_0$	$\mu > \mu_0$	$\{y \in \mathcal{Y} : T(y) > t_{n-1;1-\alpha}\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\{y \in \mathcal{Y} : T(y) < -t_{n-1;1-\alpha}\}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\{y \in \mathcal{Y} : \lvert T(y)\rvert > t_{n-1;1-\alpha/2}\}$

Dabei bezeichnet $t_{n-1;\gamma}$ das γ-Quantil der Studentischen t-Verteilung mit $n-1$ Freiheitsgraden.

Einstichproben-t-Test in R

Temperatur-Datensatz `temp` sei gegeben.
Nullhypothese: Erwartete Temperatur ist $\leq 37°C$

```
> temp<-c(36.8,37.2,37.5,36.9,37.0,37.4,37.9,38.0)
> t.test(temp,alternative="greater",mu=37)

        One Sample t-test

data:  temp
t = 2.1355, df = 7, p-value = 0.03505
alternative hypothesis: true mean is greater than 37
95 percent confidence interval:
 37.03807      Inf
sample estimates:
mean of x
  37.3375
```

Fallzahlabschätzung (I)

Beispielhaft:

Einseitiges Hypothesenpaar $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$
Analysiere Trennschärfe (Güte) als Funktion in n.

Eine Differenz $\delta = \mu - \mu_0$ soll mit Wahrscheinlichkeit $(1 - \beta) \in (0, 1)$ aufgedeckt werden können, d. h., unter $\mu_1 = \mu_0 + \delta$ soll $\beta_\varphi(\mu_1) = 1 - \beta$ gelten.

Fallzahlabschätzung (II)

Sei $\nu := n - 1$. Wir rechnen:

$$\begin{aligned} 1-\beta &\stackrel{!}{=} \mathbb{P}_{\mu_1}\left(\sqrt{n}\frac{(\bar{X}-\mu_0)}{s} > t_{\nu;1-\alpha}\right) \\ &= \mathbb{P}_{\mu_1}\left(\sqrt{n}\frac{(\bar{X}-\mu_1)}{s} > t_{\nu;1-\alpha} - \sqrt{n}\frac{\delta}{s}\right), \text{ also} \end{aligned}$$

$$t_{\nu;1-\alpha} - \sqrt{n}\delta/s \stackrel{!}{=} t_{\nu;\beta} \Longleftrightarrow n = (s/\delta)^2 \cdot (t_{\nu;\alpha} + t_{\nu;\beta})^2.$$

Problem: Man benötigt eine Vorschätzung von s !

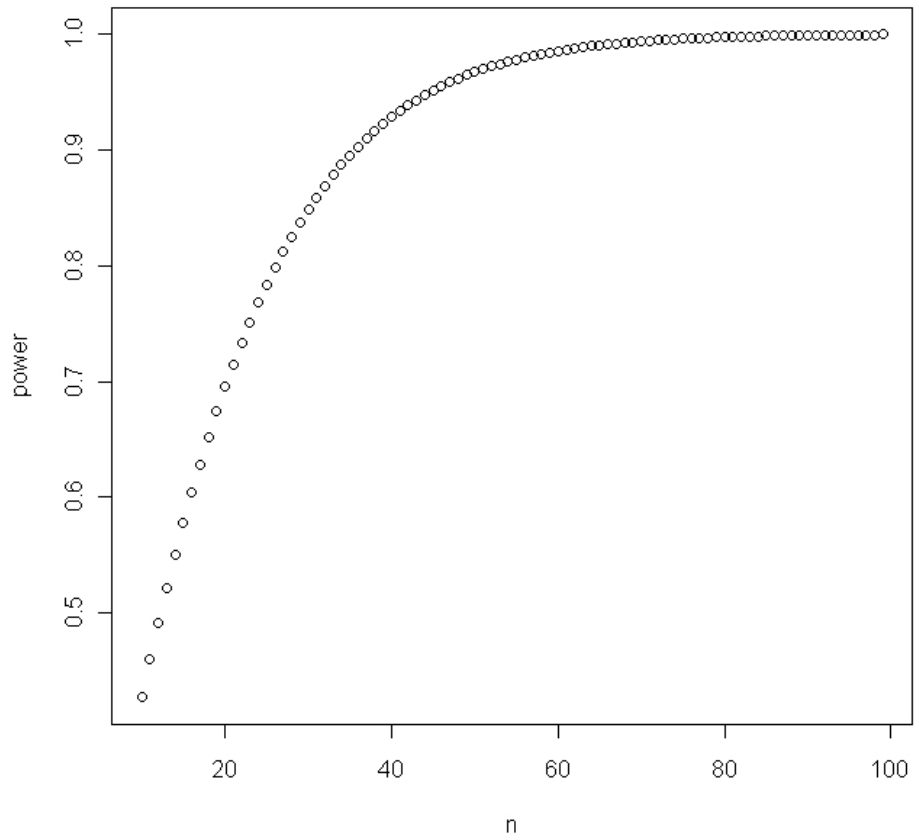
Powerberechnung in R

```
> d<-15;s<-30;r<-d/s;alpha<-0.05;beta<-0.2
> n1<-ceiling((qnorm(1-alpha)+qnorm(1-beta))^2/(r^2));n1
[1] 25
> n2<-ceiling((qt(1-alpha,n1-1)+qt(1-beta,n1-1))^2/(r^2));n2
[1] 27
> n3<-ceiling((qt(1-alpha,n2-1)+qt(1-beta,n2-1))^2/(r^2));n3
[1] 27
> power.t.test(delta=15,sd=30,sig.level=0.05,power=0.80,
+ type="one.sample",alternative="one.sided")

     One-sample t test power calculation

              n = 26.13751
          delta = 15
             sd = 30
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

Grafik: Fallzahlabschätzung (beispielhaft)



χ^2–Anpassungstest

Beobachtbar seien reellwertige i.i.d. Zufallsvariablen mit unbekannter Verteilungsfunktion F von Y_1.

Zu testen sei $H_0 : F = F_0$ gegen $H_1 : F \neq F_0$.

Beobachtungen werden aufgeteilt in m Klassen mit beobachteten Häufigkeiten n_j, $j = 1, \ldots, m$.

Wir vergleichen diese mit den unter F_0 erwarteten Häufigkeiten $n_{j,0}$, $j = 1, \ldots, m$.

Teststatistik des χ^2–Anpassungstests:

$$T(n_1, \ldots, n_m) = \sum_{j=1}^{m} \frac{(n_j - n_{j,0})^2}{n_{j,0}}$$

Unter H_0 ist diese Teststatistik T für große Stichprobenumfänge approximativ Chi-Quadrat-verteilt mit $m - 1$ Freiheitsgraden.

Ablehnbereich des zugehörigen χ^2-Signifikanztests:

$$\Gamma_\alpha = (\chi^2_{m-1;1-\alpha}, \infty),$$

wobei $\chi^2_{m-1;1-\alpha}$ das $(1 - \alpha)$-Quantil der Chi-Quadrat-Verteilung mit $m - 1$ Freiheitsgraden bezeichnet.

Anwendungsbeispiel: χ^2–Test

Problem: Teste nach 60 Würfen, ob ein Würfel fair ist.

```
> AZ<-c(1,2,3,4,5,6)
> H<-c(7,12,9,15,7,10)
> chisq.test(H)

        Chi-squared test for given probabilities

data:  H
X-squared = 4.8, df = 5, p-value = 0.4408
```

χ^2–Unabhängigkeitstest

Angenommen, wir beobachten n Tupel $(x_1, y_1), \ldots, (x_n, y_n)$ zweier diskreter Merkmale.

Mögliche Werte für Merkmal X:

$a_1, a_2, \ldots, a_k$

Mögliche Werte für Merkmal Y:

$b_1, b_2, \ldots, b_\ell$

Wir möchten prüfen, ob X stochastisch unabhängig von Y ist oder nicht.

Kontingenztafel

Die Kontingenztafel (auch: $(k \times \ell)$-Feldertafel) ist die Matrix der absoluten Häufigkeiten aller Kombinationen (a_i, b_j), $1 \leq i \leq k$, $1 \leq j \leq \ell$, in der Stichprobe vom Umfang n.

	b_1	b_2	$\ldots$	b_ℓ	$\sum$
a_1	n_{11}	n_{12}	$\ldots$	$n_{1\ell}$	$n_{1.}$
a_2	n_{21}	n_{22}	$\ldots$	$n_{2\ell}$	$n_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
a_k	n_{k1}	n_{k2}	$\ldots$	$n_{k\ell}$	$n_{k.}$
$\sum$	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.\ell}$	n

Teststatistik des χ^2–Unabhängigkeitstests

$$Q\left((n_{ij})\right) = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(n_{ij} - e_{ij})^2}{e_{ij}},$$

wobei $e_{ij} = n_{i.} n_{.j} / n$ die unter H_0 erwartete absolute Häufigkeit von Merkmalskombination (a_i, b_j) in der Stichprobe vom Umfang n bezeichnet, gegeben die Randhäufigkeiten $\mathbf{n} = (n_{1.}, n_{2.}, \ldots, n_{k.}, n_{.1}, n_{.2}, \ldots, n_{.\ell}) \in \mathbb{N}^{k+\ell}$.

Testentscheidung des χ^2–Unabhängigkeitstests

Unter H_0 ist diese Teststatistik Q für große Stichprobenumfänge approximativ Chi-Quadrat-verteilt mit $(k-1) \times (\ell - 1)$ Freiheitsgraden.

Ablehnbereich des zugehörigen χ^2-Signifikanztests:

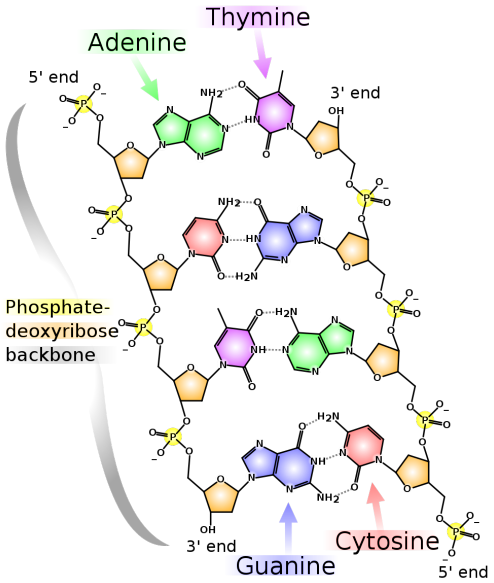
$$\Gamma_\alpha = (\chi^2_{(k-1)\times(\ell-1);1-\alpha}, \infty),$$

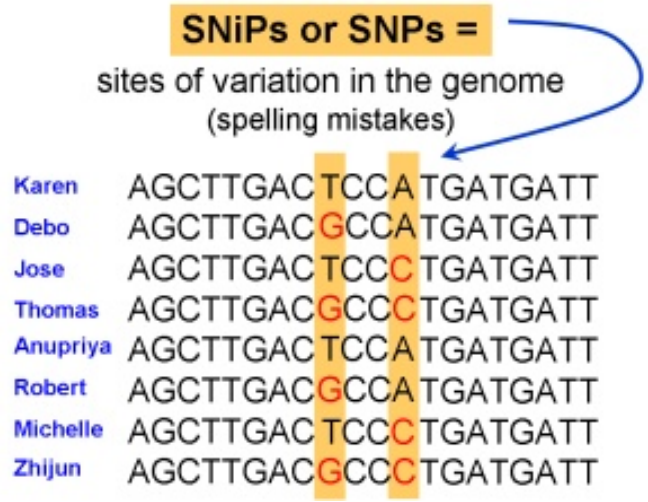
wobei $\chi^2_{(k-1)\times(\ell-1);1-\alpha}$ das $(1-\alpha)$-Quantil der Chi-Quadrat-Verteilung mit $(k-1) \times (\ell - 1)$ Freiheitsgraden bezeichnet.

Übersicht

Genetische Assoziationsstudien

Fragestellung: Welche Genorte sind mit gewissen Phänotypen (typischerweise: Krankheiten) assoziiert?





Typische Dimensionalität: $500.000 - 2.5$ Mio.
Genorte gleichzeitig analysieren!

Kontingenztafeln in genetischen Fall-Kontroll-Assoziationsstudien

An einem gegebenen Genort lässt sich das Datenmaterial wie folgt zusammenfassen.

SNP-Ausprägung	A_1A_1	A_1A_2	A_2A_2	Σ
krank	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1.}$
gesund	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Im Falle allelischer Tests:

Allel	A_1	A_2	Σ
krank	$n_{1,1}$	$n_{1,2}$	$n_{1.}$
gesund	$n_{2,1}$	$n_{2,2}$	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	n

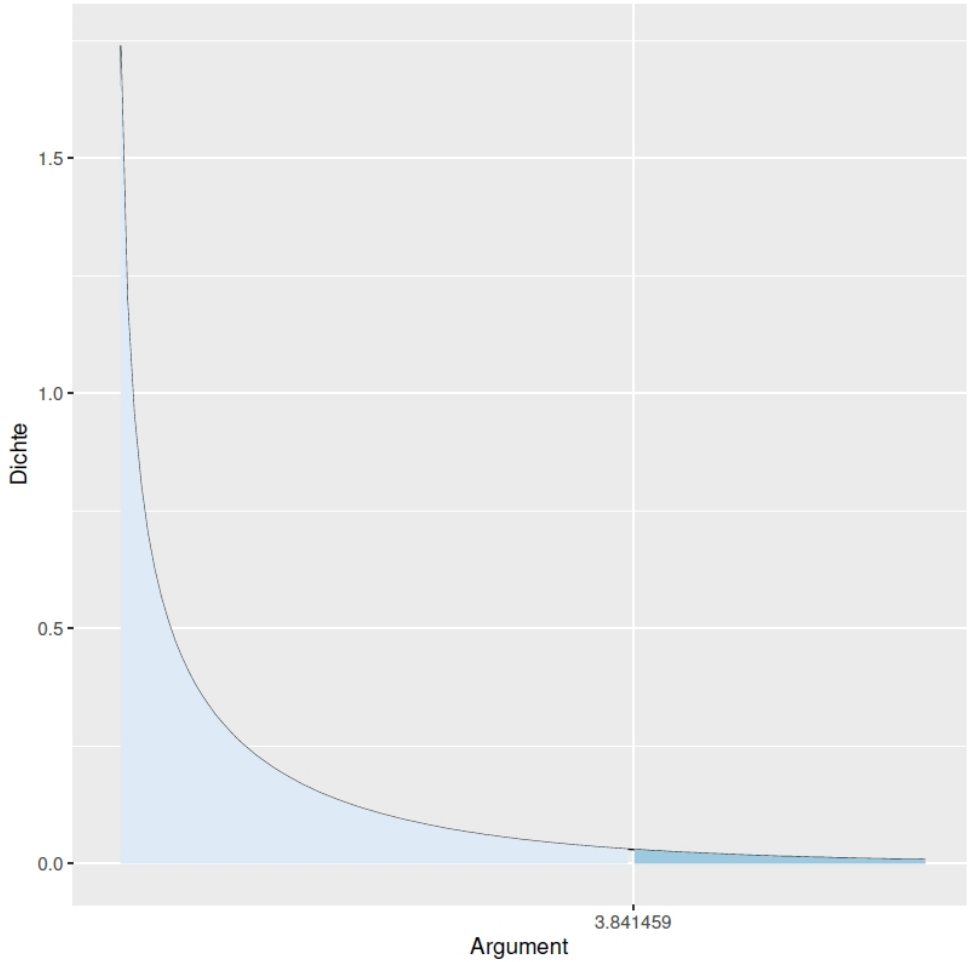
Nullhypothese und Alternativhypothese für einen gegebenen Genort

In unserem Fall ist eine adäquate Formulierung z. B.:

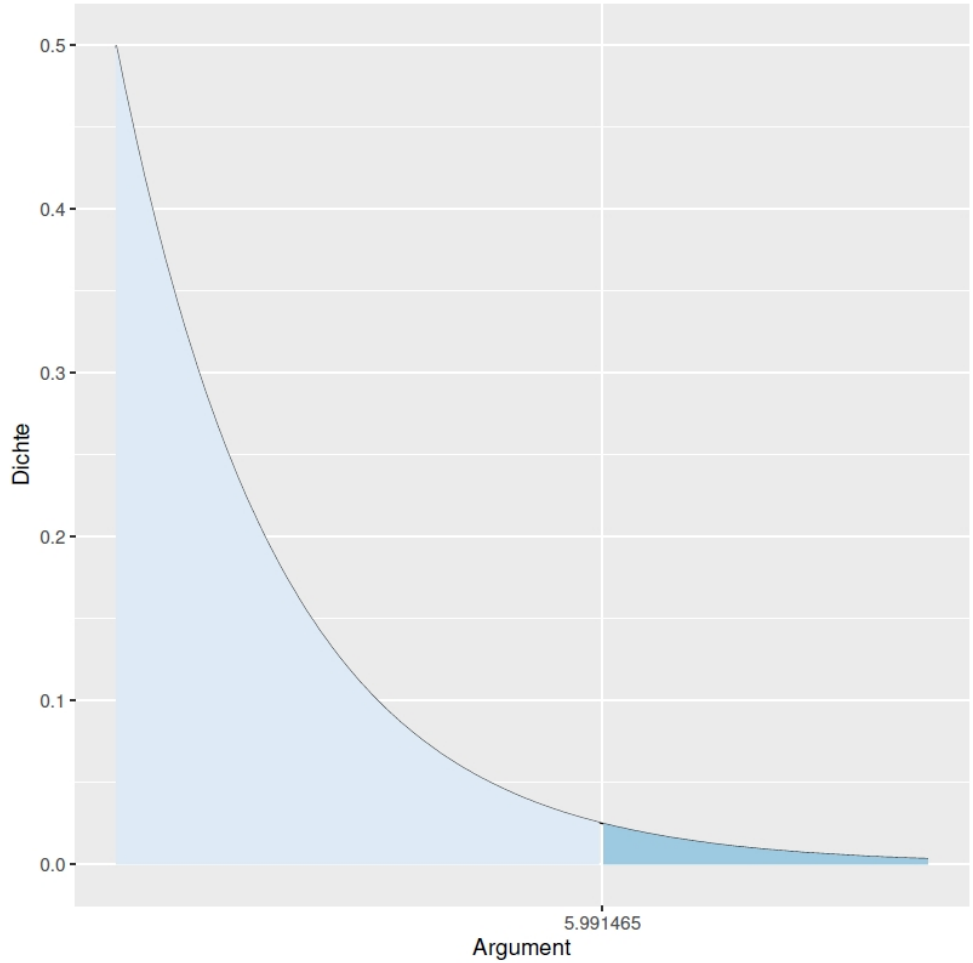
H_0: Am betrachteten Genort gibt es **keine** Assoziation zwischen Genotyp und Phänotyp,

H_1: Am betrachteten Genort gibt es **eine** Assoziation zwischen Genotyp und Phänotyp.

Chi-Quadrat-Verteilung mit einem Freiheitsgrad



Chi-Quadrat-Verteilung mit zwei Freiheitsgraden



Zahlenbeispiel (I)

Angenommen, wir haben an einem bestimmten Genort die folgende (2×2)-Kontingentafel beobachtet.

Allel	A_1	A_2	Σ
krank	32	10	42
gesund	40	20	60
Σ	72	30	102

Wir erhalten

$$e_{11} = \frac{42 \cdot 72}{102} \approx 29{,}65, \quad e_{12} = \frac{42 \cdot 30}{102} \approx 12{,}35$$

$$e_{21} = \frac{60 \cdot 72}{102} \approx 42{,}35, \quad e_{22} = \frac{60 \cdot 30}{102} \approx 17{,}65$$

Zahlenbeispiel (I)

Angenommen, wir haben an einem bestimmten Genort die folgende (2×2)-Kontingentafel beobachtet.

Allel	A_1	A_2	Σ
krank	32	10	42
gesund	40	20	60
Σ	72	30	102

Wir erhalten

$$e_{11} = \frac{42 \cdot 72}{102} \approx 29{,}65, \qquad e_{12} = \frac{42 \cdot 30}{102} \approx 12{,}35$$

$$e_{21} = \frac{60 \cdot 72}{102} \approx 42{,}35, \qquad e_{22} = \frac{60 \cdot 30}{102} \approx 17{,}65$$

Zahlenbeispiel (II)

Daraus ergibt sich

$$Q\left((n_{ij})\right) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx 1{,}0794.$$

Wählen wir als Signifikanzniveau $\alpha = 5\%$, so ist der kritische Wert c_α des Tests gegeben als das 95%-Quantil der Chi-Quadrat-Verteilung mit einem Freiheitsgrad.

Nachschauen in einer Tabelle oder Benutzung von Computer-Software liefert $c_\alpha \approx 3{,}841$.

Also wird H_0 zum 5%-Niveau auf der Basis der Daten nicht abgelehnt.

Zahlenbeispiel (II)

Daraus ergibt sich

$$Q\left((n_{ij})\right) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx 1{,}0794.$$

Wählen wir als Signifikanzniveau $\alpha = 5\%$, so ist der kritische Wert c_α des Tests gegeben als das 95%-Quantil der Chi-Quadrat-Verteilung mit einem Freiheitsgrad.

Nachschauen in einer Tabelle oder Benutzung von Computer-Software liefert $c_\alpha \approx 3{,}841$.

Also wird H_0 zum 5%-Niveau auf der Basis der Daten nicht abgelehnt.

Zahlenbeispiel (II)

Daraus ergibt sich

$$Q\left((n_{ij})\right) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx 1{,}0794.$$

Wählen wir als Signifikanzniveau $\alpha = 5\%$, so ist der kritische Wert c_α des Tests gegeben als das 95%-Quantil der Chi-Quadrat-Verteilung mit einem Freiheitsgrad.

Nachschauen in einer Tabelle oder Benutzung von Computer-Software liefert $c_\alpha \approx 3{,}841$.

Also wird H_0 zum 5%-Niveau auf der Basis der Daten nicht abgelehnt.

Zahlenbeispiel (II)

Daraus ergibt sich

$$Q\left((n_{ij})\right) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx 1{,}0794.$$

Wählen wir als Signifikanzniveau $\alpha = 5\%$, so ist der kritische Wert c_α des Tests gegeben als das 95%-Quantil der Chi-Quadrat-Verteilung mit einem Freiheitsgrad.

Nachschauen in einer Tabelle oder Benutzung von Computer-Software liefert $c_\alpha \approx 3{,}841$.

Also wird H_0 zum 5%-Niveau auf der Basis der Daten nicht abgelehnt.

Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 4 2012 *Article* 12

How to analyze many contingency tables simultaneously in genetic association studies

Thorsten Dickhaus, *Humboldt-University, Berlin*
Klaus Straßburger, *German Diabetes Center, Düsseldorf*
Daniel Schunk, *Johannes Gutenberg-Universität Mainz and University of Zurich*
Carlos Morcillo-Suarez, *Universitat Pompeu Fabra, Barcelona*
Thomas Illig, *Helmholtz Zentrum München*
Arcadi Navarro, *ICREA and Universitat Pompeu Fabra, Barcelona*