Inhaltsverzeichnis

Kapitel 5

Hypothesentests und Konfidenzintervalle

"The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge."
(Stephen Hawking)

5.1 Vorbemerkungen

Bisher haben wir uns hauptsächlich mit Schätzfunktionen für Regressionskoeffizienten beschäftigt. Wir haben gesehen, *unter welchen Bedingungen* OLS Schätzfunktionen eine 'größtmögliche' Genauigkeit garantieren.

Eine 'größtmögliche' Genauigkeit ist zwar auf den ersten Blick beruhigend, sagt aber wenig darüber aus, wir groß die Genauigkeit tatsächlich ist, das heißt, wie *vertrauenswürdig* unsere Ergebnisse sind.

Darüber hinaus läuft der wissenschaftliche Erkenntnisprozess häufig anders ab. Am Anfang eines Erkenntnisprozesses steht meist der Wunsch ein beobachtetes Phänomen zu 'verstehen'. Aber was meinen wir mit 'verstehen'? Offensichtlich erklären selbst Experten beobachtete Phänomene höchst unterschiedlich, obwohl sie alle glauben, das Phänomen verstanden zu haben.

Ein bewährter Ansatz "etwas zu verstehen" besteht darin, den Mechanismus, den den wir hinter dem Phänomen vermuteten, in einem – häufig mathematischen – Modell nachzubilden. Wenn dieses Modell ähnliche Ergebnisse produziert wie die in der Realität beobachteten Phänomene, dann sind wir einem reproduzierbaren Erklärungsansatz zumindest deutlich näher gekommen.

Was wir dafür aber benötigen ist eine Methode, mit der wir überprüfen können, inwieweit die Vorhersagen des Modells mit den beobachteten Daten übereinstimmen, und auch, inwieweit die dem Modell zugrunde liegenden Annahmen 'realistisch' sind.

Dies ist nicht ganz einfach, da Modelle immer Vereinfachungen darstellen und nie alle Aspekte des datengenerierenden Prozesses abbilden können. Deshalb werden die Vorhersagen der Modelle kaum jemals exakt mit den beobachteten Daten übereinstimmen. Aber wie groß dürfen die Abweichungen sein?

Und wie sicher dürfen wir sein, dass das Modell tatsächlich eine adäquate Beschreibung der hinter dem datengenerierenden Prozesses liegenden Gesetzmäßigkeiten liefert? Genau darum geht es in diesem Kapitel, nämlich um das Testen.[1]

5.2 Grundlagen: Wie funktionieren Hypothesentests?

Die Anfänge der Hypothesentests gehen auf Pioniere wie Francis Ysidro Edgeworth (1885) und Karl Pearson (1900) zurück, die gegen Ende des 19. Jahrhunderts erstmals begannen die Grundlagen statistischer Tests zu formulieren und systematischer darzustellen.

Die moderne Form von Hypothesentests geht vor allem auf drei Personen zurück, die in den Jahren zwischen 1915 und 1933 die Grundlagen schufen. Auf der einen Seite der große Pionier der modernen Statistik, R.A. Fisher (1925), und auf der anderen Seite Neyman and Pearson (1928*a*,*b*).

Obwohl Fisher und Neyman-Pearson überzeugt waren, dass sich ihre Ansätze grundsätzlich unterscheiden, wird in fast allen modernen Lehrbüchern eine Hybridform dieser beiden Ansätze präsentiert. Für eine Diskussion siehe z.B. Gigerenzer et al. (1990), Lehmann (1993), Spanos (1999, 688ff).

Hier werden wir eher aus didaktischen Gründen als um der historischen Gerechtigkeit willen zuerst kurz den Ansatz von Ronald A. Fisher skizzieren, bevor wir den darauf aufbauenden und heute eher gebräuchlichen Ansatz von Neyman-Pearson präsentieren.

Wir werden allerdings am Schluss sehen, dass alle drei Methoden – p-Werte nach Fisher, Hypothesentests nach Neyman-Pearson als auch Konfidenzintervalle – auf den gleichen Grundlagen beruhen und letztendlich zu den gleichen Schlussfolgerungen führen.

Wir werden die Hypothesentests im folgenden Abschnitt vor allem für Regressionskoeffizienten erläutern, aber die grundlegenden Prinzipien gelten viel allgemeiner, z.B. auch für die F-Tests simultaner Hypothesen oder Spezifikationstests.

5.2.1 Nullhypothesen und Teststatistiken

Bis herauf zum späten 19. Jahrhundert waren die Überlegungen, wie man Informationen aus der Stichprobe mit den theoretischen Vermutungen konfrontieren könnte, eher informeller Natur. Die grundlegende Idee könnte man folgendermaßen skizzieren: wenn wir den wahren – aber unbekannten – Wert eines interessierenden Parameters mit θ bezeichnen[2] und unsere theoretischen Überlegungen einen Wert θ_0

[1] Im Kern interpretieren wir die beobachteten Daten als Ergebnis eines datengenerierenden Prozesses, und unser wissenschaftliches Interesse besteht darin, die dahinter liegenden Gesetzmäßigkeiten in Form von Theorien und Modellen zu erklären. Auf Grundlage dieser theoretischen Arbeit können wir Schlussfolgerungen ziehen, und dabei stellt sich sehr schnell die Frage, inwieweit diese Schlussfolgerungen mit den beobachtbaren Daten *kompatibel* sind.

[2] θ könnte z.B. für den Mittelwert μ, einen Steigungskoeffizienten β_h , eine Varianz σ^2 usw. stehen.

erwarten lassen, so hat die Hypothese die Form

$$\theta = \theta_0$$

Da θ nicht beobachtbar ist können wir dazu keine Aussage machen, aber wenn diese Hypothese wahr ist sollte auch der Unterschied zwischen einem Schätzergebnis $\hat{\theta}$ und der Vermutung θ_0 'möglichst klein' oder ungefähr Null sein

$$|\hat{\theta} - \theta_0| \approx 0$$

Obwohl in dieser Frühzeit noch nicht dargelegt wurde, was konkret unter 'näherungsweise Null' zu verstehen sei, folgen daraus bereits die zwei zentralen Elemente eines Hypothesentests, nämlich

1. eine Vermutung über die Grundgesamtheit $\theta = \theta_0$; und
2. eine Distanzfunktion $|\hat{\theta} - \theta_0|$

Darauf aufbauend entwickelte R. Fisher seinen Ansatz zu Hypothesentests.

Der erste große Beitrag Fishers in diesem Zusammenhang bestand in der Formulierung einer expliziten Nullhypothese

$$\mathrm{H}_0: \quad \theta = \theta_0$$

sowie der Einsicht, welche Implikationen dies für Hypothesentests hat. Man beachte, dass sich die Nullhypothese immer auf die unbeobachtbare Grundgesamtheit – also die Parameter der PRF – bezieht, und nicht auf die Schätzung (Schätzungen sind beobachtbar, da gibt es nichts zu vermuten!).

Die Nullhypothese wird meist als 'Negativhypothese' formuliert, d.h. als Gegenhypothese zur theoretischen Vermutung. Die Nullhypothese beschreibt in in diesem Sinne häufig einen *'worst case'* für die Anfangsvermutung, sodass man wünscht, die *Nullhypothese verwerfen* zu können. Wie wir später sehen werden erlaubt diese Vorgangsweise, die Wahrscheinlichkeit für eine irrtümliche Verwerfung der Nullhypothese kontrolliert klein zu halten.

Prinzipiell können Forscher frei entscheiden wie sie ihre Nullhypothese formulieren, allerdings gibt es – wie wir später zeigen werden – eine Einschränkung technischer Natur: die Nullhypothese muss stets so formuliert werden, dass sie das '=' Zeichen (bzw. bei einseitigen Hypothesen das '$\leq$' oder '$\geq$' Zeichen) enthält (oder in anderen Worten, die H_0 darf *kein* $\neq$, $<$ oder $>$ enthalten).

Die Nullhypothese wird solange als 'wahr' angenommen, bis sie in 'zu starken' Konflikt mit den beobachtbaren Daten – also der Stichprobe – gerät.

Auf den Arbeiten von Gosset (1908) aufbauend erkannte Fisher, wie die Distanz $|\hat{\theta} - \theta_0|$ beurteilt werden kann, nämlich durch die Verwendung einer Teststatistik.

Teststatistiken sind neben Nullhypothesen die zweite unverzichtbare Zutat für Hypothesentests. Teststatistiken sind spezielle Zufallsvariablen, deren theoretische Verteilung *unter Gültigkeit der Nullhypothese* bekannt ist. Ähnlich wie Schätzfunktionen sind Teststatistiken spezielle Funktionen der Stichprobe, d.h. sie ordnen jeder Stichprobe eine reelle Zahl zu. Im Unterschied zu Schätzfunktionen muss die theoretische Stichprobenkennwertverteilung (*'sampling distribution'*) einer Teststatistik unter H_0 aber von vornherein bekannt sein.

Teststatistiken dürfen natürlich keine unbekannten Parameter enthalten, sonst könnte ihr Wert für eine konkrete Stichprobe nicht berechnet werden.

Die Herleitung von Teststatistiken ist Aufgabe der theoretischen Statistik und im allgemeinen kein einfaches Unterfangen. Wie wir später sehen werden sind viele der gebräuchlichen Teststatistiken asymptotischer Natur, d.h. ihre Verteilung ist für kleine Stichproben unbekannt, konvergiert aber mit zunehmender Stichprobengröße gegen eine bekannte theoretische Verteilung.

Wir wollen die prinzipielle Idee von Teststatistiken hier anhand der einfachen t-Statistik für Regressionskoeffizienten erläutern (übrigens eine der eher wenigen Teststatistiken, deren theoretische Verteilung auch für kleine Stichproben bekannt ist).

Zuvor müssen wir aber noch etwas ausholen. Für die bisherigen Ausführungen und Beweise benötigten wir lediglich die vier Gauss-Markov Annahmen. Nun wird erstmals eine theoretische Verteilung eine Rolle spielen. Unter einer *theoretischen Verteilung* wollen wir hier einfach eine Verteilung verstehen, deren Dichte- und Verteilungsfunktion durch eine mathematische Funktion beschrieben werden kann, und die durch bestimmte Parameter (Momente) charakterisiert ist. Konkret wird es um die Normalverteilung und die damit eng verbundenen χ^2-, t- und F-Verteilung gehen. Theoretische Verteilungen werden häufig verwendet, um empirische Verteilungen zu approximieren.

5.2.2 Theoretische Verteilungen

Wir erinnern uns, dass wir für den Beweis der Erwartungstreue, Effizienz und Konsistenz der OLS Schätzfunktion keine explizite Verteilungsannahme benötigten, die vier Gauss-Markov Annahmen reichten. Nur der Einfachheit halber haben wir bisher die Stichprobenkennwertverteilungen als schöne Glockenkurven gezeichnet, aber diese Form war für keines der bisherigen Argumente essentiell.

Aber für die statistische Beurteilung der *'Vetrauenswürdigkeit'* von Schätzergebnissen benötigen wir nun erstmals genauere Information über die Form der entsprechenden *Stichprobenkennwertverteilungen*.

Erinnern wir uns, dass die Stichprobenkennwertverteilung von Schätzfunktionen und Teststatistiken unmittelbar aus dem datengenerierenden Prozess (DGP) und dem *sampling* Prozess folgt.

A priori garantiert uns nichts, dass wir diese unbekannte Stichprobenkennwertverteilung durch eine bekannte theoretische Verteilung darstellen oder zumindest approximieren können. Aber wir können uns wieder fragen, welche Annahmen erforderlich sind, damit wir die unbekannte Stichprobenkennwertverteilung durch eine bekannte theoretische Verteilung approximieren können.

An dieser Stelle kommt erstmals – zusätzlich zu den bereits bekannten Gauss-Markov Annahmen – die Normalverteilungsannahme zum Tragen.

In der Frühzeit der Statistik wurde häufig angenommen, dass die Störterme unabhängig *normalverteilt* sind. Wie wir gleich zeigen werden kann mit Hilfe dieser Normalverteilungsannahme die Stichprobenkennwertverteilung der Koeffizienten unmittelbar hergeleitet werden.

Heute spielt die Annahme normalverteilter Störterme keine so große Rolle mehr, die Rechtfertigung der Sonderstellung der Normalverteilung erfolgt hauptsächlich über zentrale Grenzwertsätze.

Aber unabhängig von ihrer Rechtfertigung hat die Normalverteilung einige einzigartige und für unsere Zwecke sehr angenehme Eigenschaften, die wir vorab ganz kurz reflektieren werden.

Die Rolle der Normalverteilung

Die Normalverteilung hat einige wichtige und sehr angenehme Eigenschaften, die unser Leben im Folgenden wesentlich vereinfachen wird. Eine dieser Eigenschaften ist die *Reproduktivität*, d.h. jede Linearkombination (lineare Funktion) von n normalverteilten Zufallsvariablen X_i ist selbst wieder normalverteilt!

Wenn z.B. $X_i \sim \mathrm{N}(\mu_i, \sigma^2_{X_i})$ mit $i = 1, \ldots, n$, dann ist eine Linearkombination $Y = c_0 + c_1 X_1 + \cdots + c_n X_n$ (für konstante c_i) auch wieder normalverteilt.

Erinnern wir uns, dass der OLS Steigungskoeffizient $\widehat{\beta}_2$ eine lineare Funktion der y_i ist, wobei die Gewichte w_i nur von den erklärenden x Variablen abhängen

$$\widehat{\beta}_2 = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} w_i(\beta_1 + \beta_2 x_i + \varepsilon_i)$$

Für deterministische x_i garantiert uns die Reproduktivität also, dass für $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ auch die Schätzfunktion $\widehat{\beta}_2$ normalverteilt ist.

Wem die Annahme normalverteilter Störterme zu streng oder zu unsinnig ist findet Zuflucht bei zentralen Grenzwertsätzen, die allerdings nur asymptotisch gelten. Grenzwertsätze sind immer anwendbar, wenn in irgendeiner Form über viele i.i.d.-verteilte Zufallsvariablen gemittelt wird (häufig sind weniger strenge Annahmen als i.i.d. erforderlich).

Wenn wir davon ausgehen können, dass die Schätzfunktionen (annähernd) normalverteilt sind kommt uns eine weitere angenehme Eigenschaft der Normalverteilung zugute: jede lineare Transformation einer normalverteilten Zufallsvariable ist selbst wieder normalverteilt.

Diese Eigenschaft erlaubt uns eine *Standardisierung*, d.h. wir können jede beliebige normalverteilte Zufallsvariable $X \sim \mathrm{N}(\mu, \sigma^2_X)$ in eine standardnormalverteilte Zufallsvariable $Z \sim \mathrm{N}(0, 1)$ (d.h. Erwartungswert Null und Varianz Eins) transformieren. Das heißt, für $X \sim \mathrm{N}(\mu, \sigma^2_X)$ folgt

$$Z = \frac{X - \mu}{\sigma_X} \qquad \rightarrow \quad Z \sim \mathrm{N}(0, 1)$$

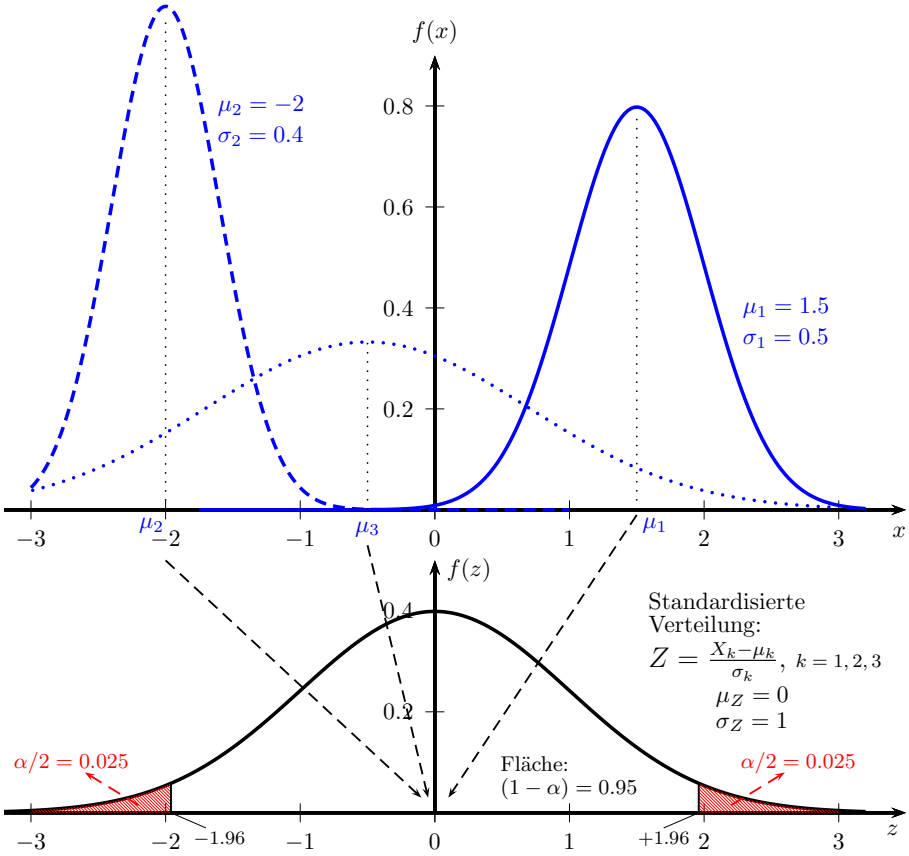


Abbildung 5.1: Normalverteile Zufallsvariablen X_1, X_2, X_3 mit unterschiedlichen Erwartungswerten und Varianzen können *standardisiert* werden zu $Z \sim \mathrm{N}(0, 1)$. Der kritische Wert $\alpha = 0.025$ für die Standardnormalverteilung ist $z_{0.025} = 1.96$.

Diese sogenannte z-Transformation ist in Abbildung 5.1 dargestellt.

Von standardnormalverteilten Zufallsvariablen wissen wir, dass 95% aller Realisationen in das Intervall $[-1.96,\ +1.96]$ fallen werden. Da die Fläche unter jeder Dichtefunktion Eins ist folgt daraus, dass 2.5% der Realisationen links von -1.96 und 2.5% der Realisationen rechts von $+1.96$ erwartet werden, in Summe also 5%.

Diesen Prozentsatz nennen wir *Signifikanzniveau* und er wird meist als α geschrieben.

Jenen Wert, bei dem links und rechts eine Fläche von je $\alpha/2$ abgeschnitten wird, nennen wir den *kritischen Wert*; für die Standardnormalverteilung also $z^{\text{crit}}_{\alpha/2=0.025} = 1.96$; und α wird Signifikanzniveau genannt ($\alpha/2 = 0.025$ da links und rechts 2.5% der Fläche abgeschnitten werden, vgl. Abbildung 5.1). Für andere Signifikanzniveaus und Verteilungen können die kritischen Werte in einer entsprechenden Tabelle nachgeschlagen oder mit Hilfe der Quantilfunktion[3] berechnen.

Wir fassen also zusammen: für jede normalverteilte Zufallsvariable

$$X \sim \mathrm{N}(\mu, \sigma_X^2)$$

gilt folgende Wahrscheinlichkeitsaussage

$$\Pr\left(-1.96 \leq \frac{X-\mu}{\sigma_X} \leq +1.96\right) = 0.95$$

In dieser Form sagt uns diese Wahrscheinlichkeitsaussage, dass wir bei wiederholter Durchführung des Zufallsexperiments erwarten können, dass 95% der Realisationen von $Z = (X-\mu)/\sigma_X$ in das Intervall $[-1.96, +.196]$ fallen.

Dies gilt natürlich auch für unsere OLS Schätzfunktion des Steigungskoeffizienten $\widehat{\beta}_2$. Deshalb gilt für

$$\widehat{\beta}_2 \sim \mathrm{N}(\beta_2, \sigma^2_{\widehat{\beta}_2})$$

folgende Wahrscheinlichkeitsaussage

$$\Pr\left(-1.96 \leq \frac{\widehat{\beta}_2 - \beta_2}{\sigma_{\widehat{\beta}_2}} \leq +1.96\right) = 0.95 \tag{5.1}$$

(wir beschränken uns hier auf das bivariate Modell $y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_+ \hat{\varepsilon}_i$, aber diese Überlegungen gelten analog auch für das multiple Regressionsmodell und für alle Regressionskoeffizienten $\widehat{\beta}_h$).

Allerdings haben wir hier noch ein Problem, im Nenner der standardisierten Zufallsvariable steht der unbeobachtbare Parameter $\sigma_{\widehat{\beta}_2}$, die Standardabweichung von $\widehat{\beta}_2$.

[3] Die Quantilfunktion ist die Umkehrfunktion der Verteilungsfunktion, welche die Fläche *links* von jedem beliebigen z^{crit} angibt. Wenn wir wissen möchten, bei welchem z^{crit} die Fläche der Dichtefunktion *rechts* 0.025 beträgt, müssen wir nur das z^{crit} suchen, bei dem die Fläche *links* 0.975 ist (die Gesamtfläche unter jeder Dichtefunktion ist 1), also in `R`: `qnorm(p = 0.975)` gibt 1.959964.

Ein historisches Ergebnis: Gosset (1908)

Die naheliegende Idee ist natürlich, diese Standardabweichung $\sigma_{\widehat{\beta}_2}$ aus der Stichprobe zu schätzen. Eine Schätzfunktion dafür haben wir bereits im letzten Kapitel hergeleitet; unter den vier Gauss-Markov Annahmen ist

$$\widehat{\text{se}}(\widehat{\beta}_2) := \hat{\sigma}_{\widehat{\beta}_2} = \sqrt{\frac{\frac{\sum_i \hat{\varepsilon}_i^2}{n-2}}{\sum_i (x_i - \bar{x})^2}}$$

Allerdings haben wir hier ein Problem, diese Schätzfunktion ist selbst eine Zufallsvariable (wird sich also von Stichprobe zu Stichprobe unterscheiden), und wenn wir diese Schätzfunktion für die Standardisierung verwenden, erhalten wir das Verhältnis von zwei Zufallsvariablen. Dieses Verhältnis zweier Zufallsvariablen ist aber nicht länger normalverteilt

$$\frac{\widehat{\beta}_2 - \beta_2}{\hat{\sigma}_{\widehat{\beta}_2}} \nsim \text{N}(0,1)$$

(beachten Sie das Dach über dem σ!)

Für dieses Problem hat W.S. Gosset (1908), seinerzeit Chef-Braumeister der Guinness Brauerei, vor mehr als hundert Jahren eine einfache Lösung gefunden. Er konnte zeigen, dass die resultierende Zufallsvariable t-verteilt ist mit $n-2$ Freiheitsgraden

$$\frac{\widehat{\beta}_2 - \beta_2}{\hat{\sigma}_{\widehat{\beta}_2}} \sim \text{t}_{n-2}$$

Die mit den *geschätzten Standardfehlern standardisierten* Schätzfunktionen der OLS Koeffizienten sind unter den obigen Annahmen (Gauss-Markov plus normalverteilte Störterme) also t-verteilt!

Dieses Resultat gilt auch für das multiple Regressionsmodell, wobei in diesem Fall eine t-Verteilung mit $n-k$ Freiheitsgraden folgt.

Intuitiv sollte klar sein, dass durch die Ersetzung eines Parameters durch dessen Schätzfunktion zusätzliche Unsicherheit ins Spiel kommt. Wie Gosset (1908) gezeigt hat, kann diese zusätzliche Unsicherheit exakt durch Verwendung der t-Verteilung mit ihren dickeren Rändern (*'fat tails'*) berücksichtigt werden.

Da Gossets Arbeitgeber die Veröffentlichung dieses Ergebnisses nicht gestattete veröffentlichte er es unter dem Pseudonym "Student", deshalb wird die t-Verteilung manchmal auch Student-Verteilung genannt (vgl. Ziliak, 2008).

Das Resultat von Gosset*

Das bekannte Resultat von Gosset lebt heute zwar in den Standardfehlern aller Regressionsoutputs weiter, spielt aber ansonsten keine sehr große Rolle mehr, da die Rechtfertigung heute weitgehend asymptotischen Überlegungen folgt.

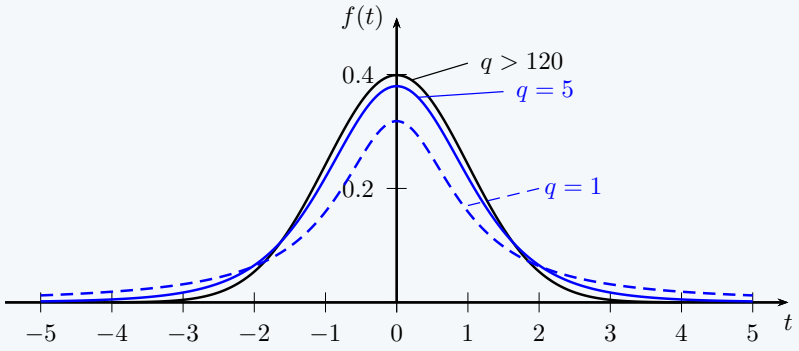
Trotzdem ist es ein wunderschönes Resultat, welches nicht nur die Tragweite theoretischer Überlegungen zeigt, sondern auch für weitergehende Analysen wichtig ist. Deshalb wollen wir es für Interessierte zumindest kurz skizzieren.

Die t–Verteilung

Ein zentrales Ergebnis der theoretischen Statistik ist, dass die Quadratsumme von q unabhängig standardnormalverteilten Zufallsvariablen χ^2 verteilt ist mit q Freiheitsgraden.
Außerdem ist aus der theoretischen Statistik bekannt, dass das Verhältnis einer standardnormalverteilten Zufallsvariable und der Wurzel einer davon unabhängig χ^2-verteilten Zufallsvariable, dividiert durch die Freiheitsgrade, t-verteilt ist, also

$$\frac{\mathrm{N}(0,1)}{\sqrt{\chi_q^2/q}} \sim \mathrm{t}_q$$

Die t-Verteilung ist ähnlich wie die Standardnormalverteilung symmetrisch, hat aber *'fat tails'* ('dicke Ränder'); siehe Abbildung unten. Für große n konvergiert die t-Verteilung gegen die Standardnormalverteilung. De facto macht es ab einer Stichprobengröße $n > 30$ keinen großen Unterschied, ob man in der t-Verteilungstabelle oder in der Standardnormalverteilungstabelle nachschlägt.



Gosset (1908) zeigte, dass für das Regressionsmodell mit normalverteilten Störtermen $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_h x_{ih} + \cdots + \beta_k x_{ik} + \varepsilon_i$ mit $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ gilt

$$\frac{\widehat{\beta}_h - \beta_h}{\hat{\sigma}_{\widehat{\beta}_h}} \sim \mathrm{t}_{n-k}$$

oder in Worten, die mit Hilfe der *geschätzten* Standardfehler $\hat{\sigma}_{\widehat{\beta}_h}$ standardisierten Koeffizienten sind t-verteilt mit $n - k$ Freiheitsgraden.

Um dies zu zeigen erinnern wir uns aus der einführenden Statistik, dass die Quadratsumme von q unabhängig standardnormalverteilten Zufallsvariablen χ^2 verteilt ist mit q Freiheitsgraden, d.h. für $z_i \sim \mathrm{N}(0, 1)$ gilt

$$(z_1^2 + z_2^2 + \cdots + z_q^2) \sim \chi_q^2$$

Ebenso ist bekannt, dass das Verhältnis einer standardnormalverteilten Zufallsvariable und der Wurzel einer davon unabhängig χ^2-verteilten Zufallsvariable, dividiert durch die Freiheitsgrade, t-verteilt ist, also

$$\frac{\mathrm{N}(0,1)}{\sqrt{\chi_q^2/q}} \sim \mathrm{t}_q$$

Um nun zu zeigen, dass $(\widehat{\beta}_h - \beta_h)/\hat{\sigma}_{\widehat{\beta}_h}$ tatsächlich t-verteilt ist, erinnern wir uns, dass die mit Hilfe des 'wahren' Standardfehlers $\sigma_{\widehat{\beta}_h}$ standardisierten Koeffizienten standardnormalverteilt sind

$$\frac{\widehat{\beta}_h - \beta_h}{\sigma_{\widehat{\beta}_h}} \sim \mathrm{N}(0, 1) \tag{5.2}$$

wobei β_h und $\sigma_{\widehat{\beta}_h}$ fixe, aber unbekannte Parameter der Grundgesamtheit sind (also *keine* Zufallsvariablen).

Wenn für alle Störterme gilt $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ (mit $i = 1, \ldots, n$) kann man zeigen, dass die folgende Zufallsvariable χ^2 verteilt ist mit $n - k$ Freiheitsgraden[4]

$$\frac{\sum_i \hat{\varepsilon}_i^2}{\sigma^2} \sim \chi_{n-k}^2$$

wobei $\sum_i \hat{\varepsilon}_i^2$ die Quadratsumme der Residuen und σ^2 die unbeobachtbare Varianz der Störterme ist.

Unter Verwendung der Schätzfunktion für die Varianz der Störterme $\hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2/(n-k)$ können wir dies umschreiben zu (im dritten Term werden Zähler und Nenner durch die feste Zahl $\sum(x_i - \bar{x})^2$ dividiert)

$$\frac{\sum \hat{\varepsilon}_i^2}{\sigma^2} = \frac{(n-k)\hat{\sigma}^2}{\sigma^2} = \frac{\frac{(n-k)\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}}{\frac{\sigma^2}{\sum(x_i-\bar{x})^2}} = \frac{(n-k)\hat{\sigma}^2_{\widehat{\beta}_h}}{\sigma^2_{\widehat{\beta}_h}} \sim \chi_{n-k}^2 \tag{5.3}$$

da $\hat{\sigma}^2_{\widehat{\beta}_h} = \hat{\sigma}^2 / \sum(x_i - \bar{x})^2$ bzw. $\sigma^2_{\widehat{\beta}_h} = \sigma^2 / \sum(x_i - \bar{x})^2$. Man beachte, dass σ^2 die Varianz der Störterme ist, und $\sigma^2_{\widehat{\beta}_h}$ die Varianz des Steigungskoeffizienten $\widehat{\beta}_h$; beide

[4]Der Beweis ist etwas aufwändiger, siehe z.B. Johnston and Dinardo (1996, S. 493).

Varianzen sind unbeobachtbare Parameter, die aus der Stichprobe geschätzt werden müssen.

Da man obendrein zeigen kann, dass die standardnormalverteilte Zufallsvariable (5.2) und die χ^2-verteilte Zufallsvariable (5.3) stochastisch unabhängig sind[5], ist der Quotient dieser beiden Zufallsvariablen eine t-verteilte Zufallsvariable

$$\frac{\mathrm{N}(0,1)}{\sqrt{\frac{\chi^2_{n-k}}{n-k}}} \sim \mathrm{t}_{n-k} \quad \Rightarrow \quad \frac{\frac{\widehat{\beta}_h-\beta_h}{\sigma_{\widehat{\beta}_h}}}{\sqrt{\frac{(n-k)\hat{\sigma}^2_{\widehat{\beta}_h}}{(n-k)\sigma^2_{\widehat{\beta}_h}}}} = \frac{\frac{\widehat{\beta}_h-\beta_h}{\sigma_{\widehat{\beta}_h}}}{\frac{\hat{\sigma}_{\widehat{\beta}_h}}{\sigma_{\widehat{\beta}_h}}} = \frac{\widehat{\beta}_h-\beta_h}{\hat{\sigma}_{\widehat{\beta}_h}} \sim \mathrm{t}_{n-k}$$

Die schöne Überraschung dabei ist, dass sich die unbekannte Populationsvarianz $\sigma^2_{\widehat{\beta}_h}$ herauskürzt.

Damit erhalten wir dieses Resultat von Gosset (1908) für einen Koeffizienten h des multiplen Regressionsmodells mit k Regressoren (inkl. Regressionskonstante; mit $h \in \{1,\ldots,k\}$)

$$\frac{\widehat{\beta}_h-\beta_h}{\hat{\sigma}_{\widehat{\beta}_h}} \sim \mathrm{t}_{n-k}$$

wobei $n-k$ die Freiheitsgrade sind. ■

5.2.3 *p*-Werte nach R.A. Fisher

Auf diesen Grundlagen entwickelte R.A. Fisher seine bekannten p Werte für Hypothesentests.

Von Gosset (1908) wissen wir, dass

$$\frac{\widehat{\beta}_h-\beta_h}{\hat{\sigma}_{\widehat{\beta}_h}} \sim \mathrm{t}_{n-k} \tag{5.4}$$

Dies gilt für den 'wahren' Wert β_h, der leider ein unbeobachtbarer Parameter ist. Aber *wenn* die Nullhypothese $\beta_h = \beta_0$ wahr ist können wir den unbeobachtbaren Parameter β_h durch unsere Vermutung unter der Nullhypothese β_0 ersetzen. Dann muss gelten

$$\hat{t}(\boldsymbol{S}) = \frac{\widehat{\beta}_h-\beta_0}{\hat{\sigma}_{\widehat{\beta}_h}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-k} \tag{5.5}$$

wobei $\boldsymbol{S}$ die Menge aller mögliche Stichproben bezeichnet und $\hat{t}(\boldsymbol{S})$ ausdrücken soll, dass die Teststatistik $\hat{t}(\boldsymbol{S})$ eine Funktion der Stichproben $\boldsymbol{S}$ ist, also eine Zufallsvariable.

Dies gilt nur, wenn die Nullhypothese wahr ist, deshalb schreiben wir $\overset{\mathrm{H}_0}{\sim}$ und lesen dies als "ist unter H_0 verteilt als".

Also, wenn $\mathrm{H}_0 : \beta_2 = \beta_0$ wahr ist und alle erforderlichen Annahmen erfüllt sind, dann ist $\hat{t}(\boldsymbol{S})$ t-verteilt mit $n-k$ Freiheitsgraden.

[5] Intuitiv folgt dies aus der stochastischen Unabhängigkeit des systematischen Teils der Regression und der Residuen.

Es ist wichtig den Unterschied zwischen Gleichung (5.5) und Gleichung (5.4) zu erkennen. Gleichung (5.4) enthält den unbekannten Parameter β_2 und ist deshalb keine Teststatistik.

In Funktion (5.5) wurde der unbekannte Parameter β_2 durch die (bekannte) Vermutung der H_0, d.h. durch β_0, ersetzt, deshalb erfüllt Funktion (5.5) alle Eigenschaften einer Teststatistik: sie ist eine Zufallsvariable, die *jeder möglichen* Stichprobe eine reelle Zahl zugeordnet, sie enthält keine unbekannten Parameter, und die theoretische Verteilung von $\hat{t}$ unter Gültigkeit der H_0 ist bekannt.

Man beachte, dass im Zähler der Teststatistik das Stichproben-Analogon zur Nullhypothese $H_0 : \beta_2 = \beta_0$ in impliziter Form steht, d.h. $\widehat{\beta}_2 - \beta_0 \ (= 0)$, wobei die Schätzfunktion $\widehat{\beta}_2$ eine Zufallsvariable und β_0 ein Skalar ist. Dies ist ein Beispiel für eine Distanzfunktion.

Die prinzipielle Idee von Hypothesentests nach R.A. Fisher können nun anhand von Abbildung 5.2 veranschaulicht werden.

Der datengenerierende Prozess erzeugt n Realisationen von (x_i, y_i) Paaren, und wir können uns vorstellen, dass die entsprechenden Zufallsvariablen 'hinter' diesen Realisationen gewissermaßen alle möglichen Stichproben abbilden, den Stichprobenraum. Damit können wir nicht nur die OLS Schätzfunktion $\widehat{\beta}_2$ berechnen, sondern auch deren z-transformierte (standardisierte) Variante, von der wir wissen, dass sie unter den Gauss-Markov Annahmen und der Annahme unabhängig normalverteilter Störterme ε_i standardnormalverteilt ist

$$\frac{\widehat{\beta}_2 - \beta_2}{\sigma_{\widehat{\beta}_h}} \sim \mathrm{N}(0, 1)$$

Diese 'wahre' Stichprobenkennwertverteilung ist durchgezogen (blau) in Abbildung 5.2 eingezeichnet.

Allerdings können wir diese 'wahre' Stichprobenkennwertverteilung nicht beobachten, wir kennen weder β_2 noch $\sigma_{\widehat{\beta}_2}$, und können deshalb auch nicht sagen, *wo* sie liegt! Aber wir wissen, dass die aus der vorliegenden Stichprobe berechnete empirische Teststatistik t^{emp} eine Realisation aus dieser 'wahren' Stichprobenkennwertverteilung ist.

Eine der tiefen Einsichten R.A. Fishers bestand nun darin, dass wir eine theoretische Teststatistik *unter der Annahme konstruieren können, dass die Nullhypothese gültig* ist. Diese *theoretische Teststatistik*

$$\hat{t} = \frac{\widehat{\beta}_2 - \beta_0}{\hat{\sigma}_{\widehat{\beta}_2}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-k}$$

ist in Abbildung 5.2 (rot) strichliert eingezeichnet. Das empirische Analogon zu dieser theoretischen Stichprobenkennwertverteilung unter H_0 können wir aus der gegebenen Stichprobe berechnen, da wir β_0 kennen und $\hat{\sigma}_{\widehat{\beta}_2}$ aus der Stichprobe schätzen können. Das t-verteilte Resultat ist in Abbildung 5.2 rechts unten eingezeichnet.

Nun können wir die *empirische Teststatistik*

$$t^{\mathrm{emp}} = \frac{b_2 - \beta_0}{s_{\widehat{\beta}_2}}$$

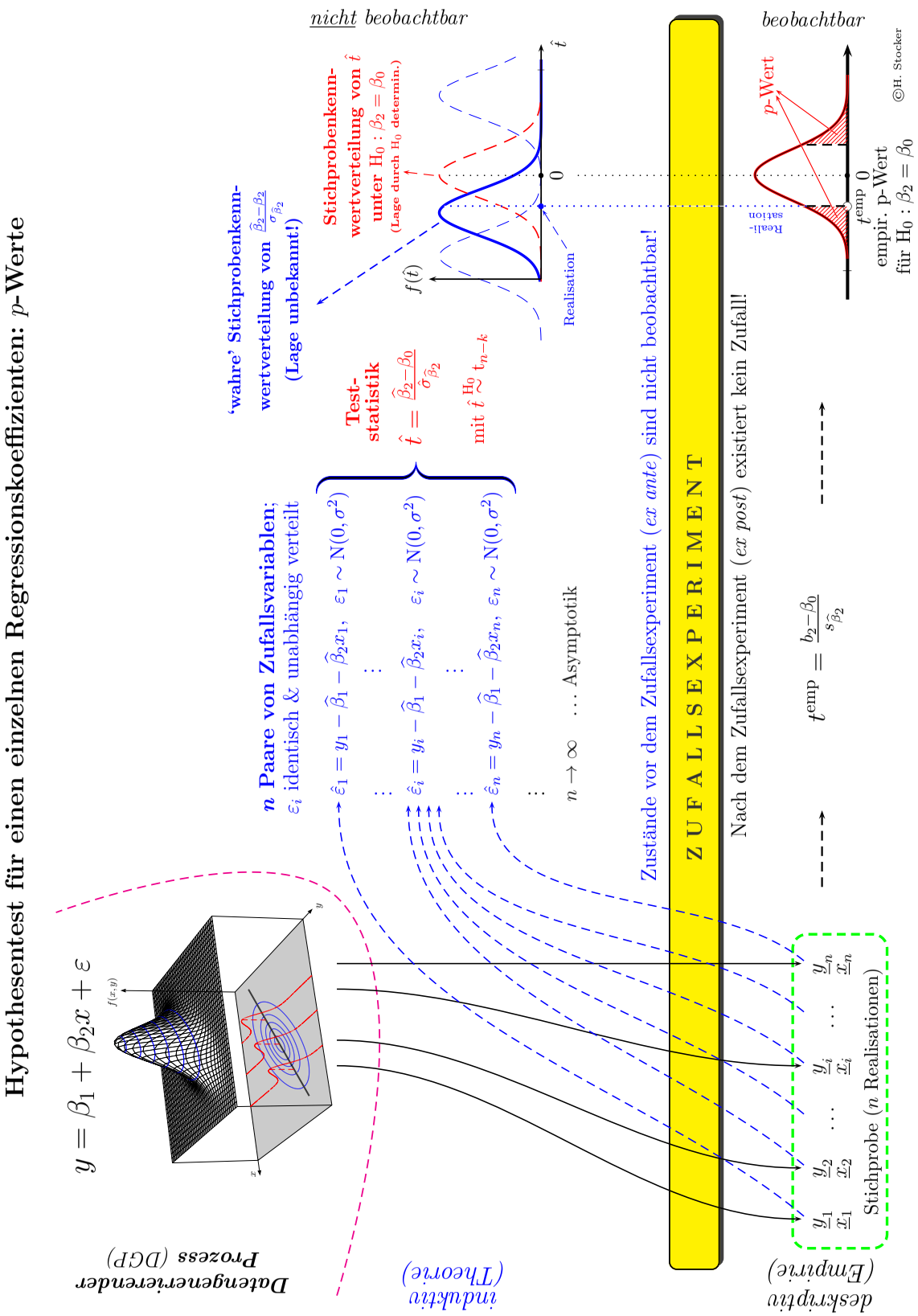


Abbildung 5.2: Hypothesentest eines einzelnen Regressionskoeffizienten nach R.A. Fisher

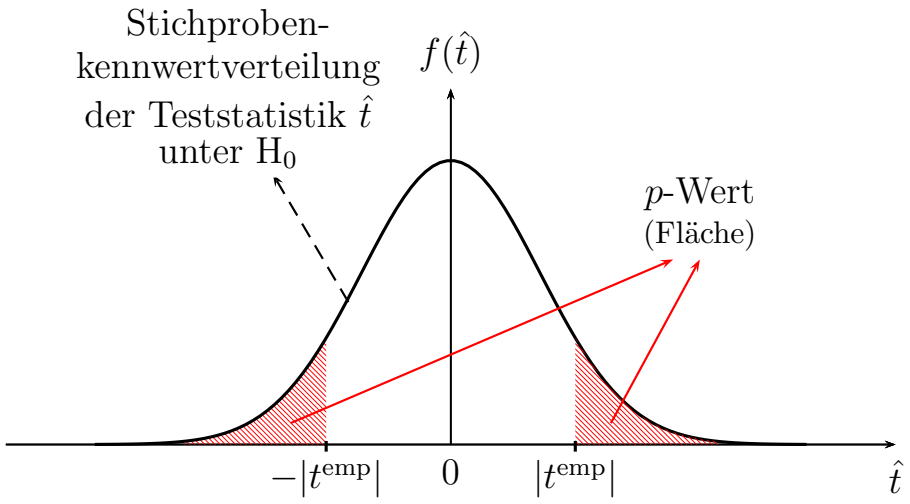


Abbildung 5.3: p-Wert nach R. Fisher (rot schraffierte Fläche) für zweiseitigen Test $H_0: \beta_2 = \beta_0$.

(einer Realisation der theoretischen Teststatistik, also eine einfache reelle Zahl und keine Zufallsvariable) auf der x-Achse der empirischen Stichprobenkennwertverteilung auftragen.

Empirische p-Werte für zweiseitige Nullhypothesen

Für die zweiseitige Nullhypothese $H_0: \beta_2 = \beta_0$ ist der *empirische* p-Wert definiert als die *Fläche unter der Dichtefunktion*, welche rechts von $+|t^{emp}|$ und links von $-|t^{emp}|$ liegt.[6]

Diese Nullhypothese ist *zweiseitig*, weil sowohl sehr weit links als auch sehr weit rechts liegende t^{emp}-Werte als Indizien *gegen* die Nullhypothese interpretiert werden. Konkret ist eine Nullhypothese immer zweiseitig, wenn sie das '=' Zeichen enthält, und sie ist einseitig, wenn sie ein '$\geq$' oder '$\leq$' enthält.

Dies wird in Abbildung 5.3 gezeigt. Da die Gesamtfläche unter einer Dichtefunktion immer Eins ist können p-Werte nur Werte zwischen Null und Eins annehmen.

Diese beiden Flächen definieren den empirischen p-Wert für einen zweiseitigen Test.

Da die t-Verteilung symmetrisch ist reicht es eine der beiden Flächen zu berechnen und diese mit zwei zu multiplizieren.

Exkurs:* In historischen Zeiten war die Berechnung der p-Werte aufwändig, aber heute geben so gut wie alle Statistikprogramme die p-Werte automatisch aus. Computerprogramme berechnen den p-Wert meist mit Hilfe der Verteilungsfunktion

$$F(t) = \int_{-\infty}^{|t^{emp}|} f(x)\, dx$$

Den zweiseitigen p-Wert erhält man aus

$$p^{emp} = 2[1 - F(|t^{emp}|, df)]$$

[6] Wir verwenden hier den Absolutbetrag $|\cdot|$ der Teststatistik, da die Teststatistik auch negativ sein kann. In diesem Fall könnte 'links' und 'rechts' von t^{emp} irreführend sein.

wobei $F(t)$ in diesem Fall die Verteilungsfunktion der t-Verteilung bezeichnet (*nicht* die F-Statistik!), $|t^{\text{emp}}|$ ist der Absolutwert der empirischen Teststatistik, und *df* steht natürlich wieder für die Freiheitsgrade (*'degrees of freedom'*). □

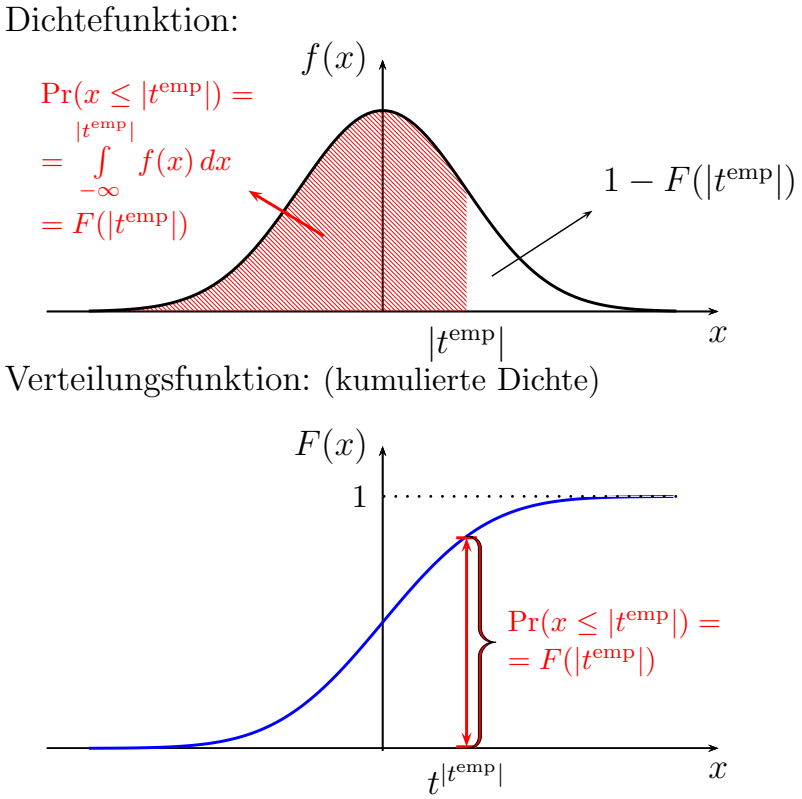


Abbildung 5.4: Dichtefunktion $f(x)$ und Verteilungsfunktion $F(x)$ einer stetigen Zufallsvariable. Die Wahrscheinlichkeit, dass eine stetige Zufallsvariable eine Ausprägung kleiner gleich t^{emp} annimmt, ist gleich der schraffierten *Fläche* unter der Dichtefunktion *links* von t^{emp}, oder gleich der *Länge* der Strecke *über* t^{emp} der Verteilungsfunktion. Der p-Wert für eine zweiseitige Hypothese ist $2[1 - F(t^{\text{emp}})]$.

Kehren wir nun nochmal zurück zu Abbildung 5.2. Falls die Nullhypothese wahr ist, sollte die (blau durchgezogene) 'wahre' Stichprobenkennwertverteilung (fast) genau auf der (rot strichliert eingezeichneten) Stichprobenkennwertverteilung unter der H_0 liegen.

Aber selbst wenn die Nullhypothese exakt richtig ist und alle Annahmen exakt erfüllt sind würden wir aufgrund der Stichprobenfehler (*'repeated sampling'*) in 5% der möglichen Fälle einen p-Wert < 0.05 erwarten. Dies ist zwar ein seltenes Ereignis, aber jederzeit möglich.

Nun stellen Sie sich vor, dass die (blau durchgezogene) 'wahre' Stichprobenkennwertverteilung weit entfernt von der (rot strichliert eingezeichneten) Stichprobenkennwertverteilung unter der H_0 liegt. Die *empirische* t^{emp}-Statistik ist eine Realisation aus der 'wahren' Stichprobenkennwertverteilung, deshalb würden wir in diesem Fall einen (absolut gesehen) großen Wert der *empirische* t^{emp}-Statistik erwarten, der unmittelbar zu einem kleinen p-Wert führt (nahe bei Null).

Deshalb können wir den empirischen p-Wert als Indikator dafür interpretieren, wie gut die Nullhypothese die 'wahre' Stichprobenkennwertverteilung – und damit den datengenerierenden Prozess – beschreibt.

Interpretation empirischer p-Werte

Vorsicht ist wieder bei der Interpretation geboten, jeder *empirische* p–Wert ist eine Realisation, und es macht natürlich überhaupt keinen Sinn sich zu fragen, wie wahrscheinlich ein Ereignis ist, welches bereits stattgefunden hat, es hat stattgefunden!

Aber wir können uns fragen, wie wahrscheinlich es ist, bei einer *hypothetischen neuerlichen Ziehung einer Zufallsstichprobe* (oder sehr häufiger Wiederholungen des Zufallsexperiments) einen mindestens ebenso extremen oder noch extremeren Wert als t^{emp} zu erhalten.

Genau diese Wahrscheinlichkeit gibt uns der *p-Wert* an.

In dieser Interpretation ist der *theoretische* p-Wert eine Zufallsvariable, und der empirische p-Wert p^{emp} eine Realisation dieser Zufallsvariable, den wir aus einer gegebenen Stichprobe berechnen.

Interpretation von p-Werten: Wenn die Nullhypothese wahr ist und alle erforderlichen Annahmen erfüllt sind gibt der p-Wert die Wahrscheinlichkeit dafür an, dass wir bei einer hypothetischen neuerlichen Durchführung des Zufallsexperiments eine empirische Teststatistik erhalten würden, die noch extremer ist als die vorliegende empirische Teststatistik.

Da p-Werte Wahrscheinlichkeiten sind können sie nur Werte zwischen Null und Eins annehmen. Ein sehr kleiner p-Wert nahe bei Null deutet darauf hin, dass entweder die Nullhypothese falsch ist, oder dass ein sehr unwahrscheinliches Ereignis eingetreten ist (oder dass eine zugrunde liegende Annahme verletzt wurde).

Umso kleiner ein p-Wert ist, umso eher werden wir geneigt sein die Nullhypothese zu verwerfen, was wir als empirische Evidenz für unsere Anfangsvermutung interpretieren können.

Etwas salopp können wir den p-Wert auch einfach als Kennzahl dafür interpretieren, wie gut die Nullhypothese die Daten beschreibt. Ein sehr kleiner p-Wert wird als Evidenz gegen die Nullhypothese interpretiert.

p–Werte für einseitige Hypothesentests

Viele Hypothesen sind nicht symmetrisch. Wenn z.B. eine neue Trainingsmethode eingeführt wird, so werden wir *im Durchschnitt* eine Leistungsverbesserung der Teilnehmer erwarten. Ebenso werden wir bei steigenden Einkommen eine Zunahme der Konsumausgaben erwarten, keine Abnahme.

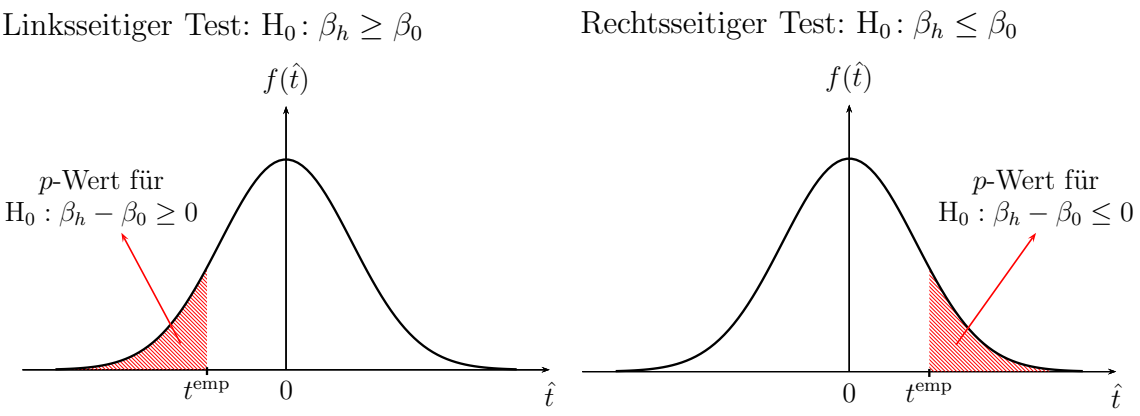


Abbildung 5.5: p-Wert nach R.A. Fisher für einseitige Hypothesentests. Für $H_0: \beta_h - \beta_0 \geq 0$ stellt nur ein *stark negativer* Wert von t^{emp} eine Überraschung dar, für $H_0: \beta_h - \beta_0 \leq 0$ ist erst ein *stark positiver* Wert t^{emp} überraschend.

Solche *einseitigen Hypothesen* können ebenso einfach getestet werden. Auch einseitige Nullhypothesen werden üblicherweise als Negativhypothesen formuliert. Vermuten wir z.B., dass ein Kennwert der Grundgesamtheit β_h *größer* ist als β_0, so ist die Nullhypothese (als Negativhypothese)

$$H_0: \beta_h \leq \beta_0$$

Wenn wir z.B. vermuten, dass die durchschnittliche Leistung nach Einführung einer neuen Trainingsmethode besser ist als die durchschnittliche Leistung vor der Einführung β_0, dann ist die Negativhypothese, dass sich die Leistung nicht verbessert hat oder sogar abgenommen hat.

Auf den ersten Blick scheint es unmöglich diese Hypothese zu testen, da $\beta_h \leq \beta_0$ unendlich viele Fälle umfasst. Etwas nachdenken zeigt allerdings, dass es genügt den Grenzfall $\beta_h = \beta_0$ zu testen, denn wenn man $H_0: \beta_h = \beta_0$ verwerfen kann, können automatisch auch alle extremeren Hypothesen $\beta_h < \beta_0$ verworfen werden.

Deshalb kann für einen einseitigen Test die gleiche Teststatistik wie für zweiseitige Tests verwendet werden, in diesem Fall

$$t^{emp} = \frac{b_h - \beta_0}{s_{\widehat{\beta}_h}}$$

allerdings ist bei einseitigen Tests das *Vorzeichen* der empirischen Teststatistik zu beachten.

Die **Interpretation** des p-Wertes ändert sich nicht: Falls die *Nullhyhpothese* $\beta_h \leq \beta_0$ *wahr ist* und alle erforderlichen Annahmen erfüllt sind würden wir bei *wiederholten Stichprobenziehungen* in $p^{emp} \times 100\%$ der Fälle ein so extremes Ergebnis erwarten wie das beobachtete.

Linksseitige Tests für Nullhypothesen $H_0: \beta_h \geq \beta_0$ funktionieren analog, nur ist in diesem Fall die Fläche *links* von t^{emp} relevant ist, vergleiche linkes Panel von Abbildung 5.5.

Hinweis: Viele finden einseitige Hypothesentests etwas verwirrend, da man das Vorzeichen beachten muss. Etwas einfacher wird es zumindest in diesem einfachen Fall

wenn man beachtet, dass im Zähler dieser Teststatistik das empirische Analogon zur Nullhypothese in impliziter Form steht.

- Für eine *linksseitige* Nullhypothese $H_0: \beta_h - \beta_0 \geq 0$ stellen *positive* Werte von $t^{\text{emp}} = (b_h - \beta_0)/s_{\widehat{\beta}_h}$ keine Überraschung dar, selbst *kleine* negative Werte sind nicht sehr überraschend, erst *große negative* Werte sind überraschend und mit der Nullhypothese schwer vereinbar. Deshalb liegt der Verwerfungsbereich zur Gänze *links* im negativen Bereich.
- Umgekehrt, für eine *rechtsseitige* Nullhypothese $H_0: \beta_h - \beta_0 \leq 0$ sind *negative* oder selbst *kleine positive* Werte von $t^{\text{emp}} = (b_h - \beta_0)/s_{\widehat{\beta}_h}$ wenig überraschend, aber *große positive* Werte sind bei Gültigkeit der Nullhypothese unwahrscheinlich. Der Verwerfungsbereich liegt zur Gänze *rechts* im positiven Bereich.

Was passiert, wenn der p-Wert keine klare Verwerfung der Nullhypothese erlaubt? R.A. Fisher interpretierte große p-Werte ausschließlich als Evidenz dafür, dass die zur Verfügung stehenden Daten keine Grundlage für eine klare Entscheidung liefern. Es sollte klar sein, dass p-Werte nahe bei Eins keinesfalls bedeutet, dass die getestete Nullhypothese wahr sein muss. Genauso wenig 'beweist' ein kleiner p-Wert, dass die Nullhypothese falsch ist. Fisher schreibt:

> "For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning. [...] It would, therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as the are contradicted by the data: but that they are never capable of establishing them as certainly true." (zitiert nach Salsburg, 2002, 107f)

5.2.4 Hypothesentests nach Neyman-Pearson

Auf der Arbeit von R.A. Fisher aufbauend entwickelten Jerzy Neyman (polnischer Mathematiker und Statistiker, 1894 – 1981) und Egon Pearson (1895 – 1980, Sohn von Karl Pearson) die heute verbreitete Form von Hypothesentests.

Alles, was wir bisher über R. Fishers Signfikanztests gehört haben, bleibt auch für Hypothesentests nach Neyman-Pearson gültig, aber bei Neyman-Pearson kommen zwei neue Elemente hinzu

1. Der Nullhypothese wird explizit eine Alternativhypothese H_A gegenüber gestellt, die immer richtig ist, wenn die Nullhypothese falsch ist, d.h. H_0 und H_A schließen sich gegenseitig aus.

   Zur Erinnerung, sowohl Null- als auch Alternativhypothese sind stets Aussagen über unbeobachtbare Parameter der Grundgesamtheit!

2. Es wird *a priori* ein Signifikanzniveau α festgelegt. Im Unterschied zum p–Wert von Fisher, der eine Zufallsvariable ist (bzw. eine Realisation davon), ist das Signifikanzniveau α ein fixer Parameter, der von der Forscherin a priori festgelegt wird.

Diese beiden Erweiterungen erlaubten Neyman-Pearson Hypothesentests als eindeutige *Entscheidungsregel* zu formulieren, die für jede Stichprobe angibt, ob die Nullhypothese verworfen oder beibehalten werden soll. Diese Entscheidung muss natürlich nicht richtig sein, aber wie wir gleich sehen werden erlaubt diese Methodik eine klare Beurteilung der möglichen Fehler.

Wir werden im Folgenden die Testmethodik nach Neyman-Pearson nur für den denkbar einfachsten Fall vorstellen, für den Test eines einzelnern Parameters der Grundgesamtheit, aber diese Testmethodik ist deutlich allgemeiner und kann in verschiedenen Zusammenhängen angewandt werden.

Wir beginnen mit einer Übersicht über die wesentlichen Schritte eines Hypothesentests nach Neyman-Pearson:

Problemanalyse Der erste, wichtigste und meist auch schwierigste Schritt jedes Hypothesentests sollte ein *Nachdenken* über das zu untersuchende Phänomen sein.

Festlegung von Null- und Alternativhypothese Auf Grundlage der theoretischen Überlegungen erfolgt die Formulierung der Nullhypothese H_0 sowie der dazugehörigen Alternativhypothese H_A. In der Alternativhypothese findet sich meist die Anfangsvermutung, bzw. der erhoffte Fall, die Nullhypothese ist die Negativhypothese dazu.

Null- und Alternativhypothese schließen sich gegenseitig aus, es kann nur eine der beiden Hypothesen richtig sein. Getestet wird immer die Nullhypothese, und diese muss so formuliert sein, dass sie das '=' (bzw. bei einseitigen Tests das '$\leq$' oder '$\geq$') enthält.

Wahl der Teststatistik Wir haben bisher erst eine Teststatistik kennen gelernt, nämlich die einfache t-Statistik. Oft stehen aber auch alternative Teststatistiken zur Verfügung, die weniger strenge Annahmen benötigen, dafür aber nur asymptotisch gültig sind. Die Wahl der Teststatistik hängt häufig wesentlich davon ab, welche Annahmen man bereit ist zu akzeptieren.

Festlegung eines Signifikanzniveaus α Im Unterschied zur Testmethodik von Fisher wird hier a priori eine Wahrscheinlichkeit festgelegt, ab der ein Ereignis als hinreichend unwahrscheinlich beurteilt wird, so dass man bei Überschreitung dieses Schwellenwertes die Nullhypothese auf jeden Fall verwerfen wird. Diesen Schwellenwert nennen wir Signifikanzniveau (α).

Diese Entscheidung wird bereits *vor* Durchführung des eigentlichen Tests getroffen, ab diesem Zeitpunkt wird das Testverfahren gewissermaßen 'auf Autopilot geschaltet', wir müssen nur noch die restlichen Schritte nach Vorschrift ausführen und die resultierende Entscheidung akzeptieren.

Tabelle 5.1: TABELLE: $t_{(q)}$-Verteilung (rechts-seitig), mit q Freiheitsgraden.

q	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
⋮	⋮	⋮	⋮	⋮	⋮
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
⋮	⋮	⋮	⋮	⋮	⋮
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
∞	1.282	1.645	1.960	2.327	2.576

Annahme- und Verwerfungsbereich Mit Hilfe dieses Signifikanzniveaus und der gewählten Teststatistik kann man einen *kritischen Wert* der Teststatistik ermitteln. Dies wird in Abbildung 5.6 (Seite 21) für einen zweiseitigen Test gezeigt. Für einen zweiseitigen Test definieren wir den *Akzeptanzbereich* als jenen Bereich unter der Dichtefunktion der Teststatistik, über dem die Fläche $(1 - \alpha)$ beträgt, und als *Verwerfungsbereich* jenen Bereich an den Rändern, der links und rechts die Flächen $\alpha/2$ 'abschneidet'.

Wir suchen also in Abbildung 5.6 den kritischen Wert $t^{\text{crit}}_{\alpha/2,df}$, sodass die gesamte schraffierte Fläche exakt gleich α ist (bei einem zweiseitigen Test also die Summe der Flächen links von $-|t^{\text{crit}}_{\alpha/2,df}|$ und rechts von $+|t^{\text{crit}}_{\alpha/2,df}|$ gleich α ist).

Diesen kritischen Wert der Verteilung kann man in einer üblichen t–Tabelle nachschlagen, vgl. Tabelle 5.1, oder wenn ein Computer zur Verfügung steht, mit Hilfe der Quantilfunktion berechnen.

Der kritische Wert erlaubt eine Unterteilung des Parameterraums in einen *Annahme-* und *Verwerfungsbereich*, die sich gegenseitig ausschließen Unter einem *Parameterraum* verstehen wir einfach alle möglichen Werte, die ein interessierender Parameter annehmen kann. In vielen Fällen wird dies die Menge der reellen Zahlen $\mathbb{R}$ sein. Der Annahme- und Verwerfungsbereich sind disjunkte Teilmengen des Parameterraums.

Aufgrund des *Stichprobenfehlers* (d.h. des Fehlers, der durch das Zufallselement der Stichprobenziehung zustande kommt) erwarten wir bei wiederholten Stichprobenziehungen (*'repeated sampling'*), dass selbst wenn die Nullhypothese wahr ist $\alpha \times 100$ Prozent der möglichen Realisationen zufällig im *Verwerfungsbereich* fallen werden.

Man beachte, dass bisher die Stichprobe nicht benötigt wurde. Im Idealfall sollte die Stichprobe bisher nicht ausgewertet worden sein.

Empirische Teststatistik Erst jetzt wird die *empirische Teststatistik* t^{emp} berech-

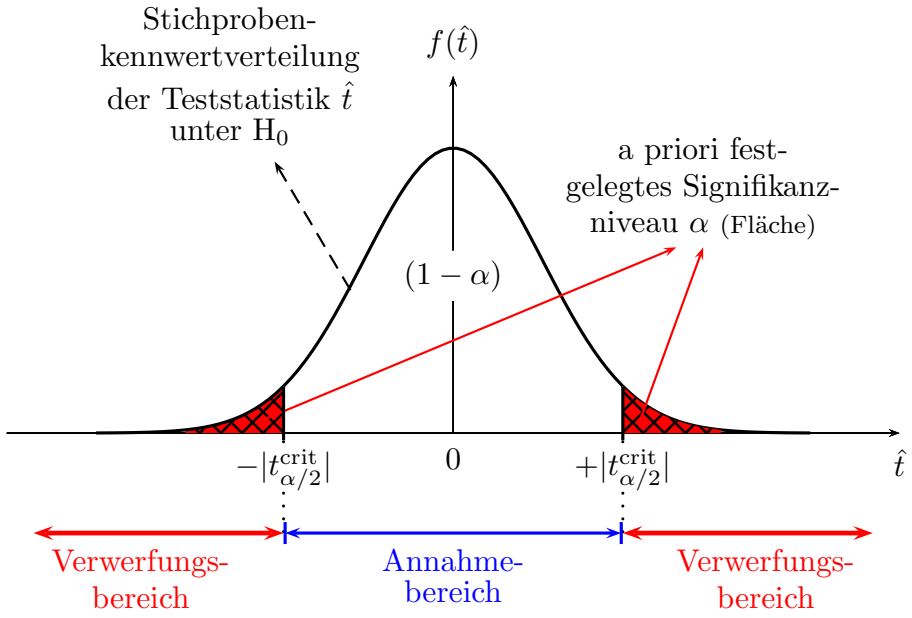


Abbildung 5.6: Stichprobenkennwertverteilung der Teststatistik $t(\boldsymbol{S})$ unter der H_0: $\beta_h = \beta_0$.
Man beachte, dass die der Grafik zugrunde liegende Information ausschließlich aus der statistischen Theorie folgt, nicht aus den Stichprobendaten! Erst in einem nächsten Schritt wird aus der vorliegenden Stichprobe der empirische Wert t^{emp} berechnet und überprüft, ob t^{emp} in den Annahme- oder Verwerfungsbereich fällt.

net, indem die vorliegenden Stichprobendaten, d.h. die *Realisationen* von $\widehat{\beta}_h$ und $\hat{\sigma}_{\widehat{\beta}_h}$ in die empirische Teststatistik

$$t^{emp} = \frac{b_h - \beta_0}{s_{\widehat{\beta}_h}}$$

eingesetzt werden. Das Resultat ist eine Realisation der Teststatistik, also eine konkrete Zahl.

Wenn die Nullhypothese wahr ist würden wir *bei wiederholten Stichprobenziehungen* erwarten, dass $1 - \alpha \times 100$ Prozent der derart berechneten Teststatistiken in den Annahmebereich und $\alpha \times 100$ Prozent in den Verwerfungsbereich fallen.

Entscheidung Die Entscheidungsregel ist nun sehr einfach, wir müssen nur noch schauen, ob die eine beobachtete *empirische Teststatistik* t^{emp} in den Annahme- oder in den Verwerfungsbereich fällt. Liegt die empirische Teststatistik t^{emp} im Annahmebereich kann die Nullhypothese *nicht* verworfen werden, liegt t^{emp} hingegen im Verwerfungsbereich wird die Nullhypothese verworfen.

Da sich Annahme- und Verwerfungsbereich gegenseitig ausschießen führt dies immer zu einer eindeutigen Entscheidung.

Mögliche Fehler Die Entscheidung ist zwar eindeutig, aber sie muss deshalb nicht richtig sein. Konkret können zwei Arten von Fehler auftreten, nämlich

1. Typ I Fehler: Die Nullhypothese ist tatsächlich wahr, aber wir haben zufällig eine extreme Stichprobe gezogen, weshalb wir die nach dieser Entscheidungsregel die Nullhypothese trotzdem verwerfen. Bei wiederholten Stichprobenziehungen sollte dieser Fehler nicht öfter als in $\alpha \times 100$ Prozent der Fälle auftreten.

2. Typ II Fehler: Die Nullhypothese ist tatsächlich falsch, die (erhoffte) Alternativhypothese also richtig, aber die Anwendung der Entscheidungsregel führt dazu, dass die Nullhypothese irrtümlich *nicht* verworfen wird.

Ein Beispiel für einen einzelnen Regressionskoeffizienten

Die Tests einzelner Regressionskoeffizienten verlaufen nach dem üblichen Muster. Wenn wir den Koeffizienten β_h der PRF

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_h x_{ih} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

testen wollen können wir die Teststatistik

$$\hat{t} = \frac{\widehat{\beta}_h - \beta_0}{\hat{\sigma}_{\widehat{\beta}_h}} = \frac{\text{Stochastisches Analogon zur } \mathrm{H}_0 \text{ in impliziter Form}}{\text{Standardfehler vom Zähler}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-k}$$

verwenden ($n-k$ bezeichnet die Freiheitsgrade, k ist die Anzahl der geschätzten Regressionskoeffizienten).

Die am häufigsten getestete Nullhypothese ist, dass der wahre Parameter β_h Null ist (also $\beta_0 = 0$), oder in Worten, dass die Variable x_h tatsächlich *keinen* Einfluss auf $\widehat{y}$ hat. Die zu testende Nullhypothese ist in diesem Fall

$$\mathrm{H}_0\colon \quad \beta_h = 0$$

Diese Nullhypothese liegt auch den von statistischen Programmen ausgegebenen t-Statistiken und p-Werten zugrunde.

Forscher wollen natürlich meist zeigen, dass eine Variable x_h einen Einfluss auf $\widehat{y}$ ausübt, ihre Vermutung (und meist auch Hoffnung) ist also die Alternativhypothese. Nur wenn die Nullhypothese überzeugend *verworfen* werden kann wird dies als empirische Evidenz für die Anfangsvermutung interpretiert.

Da die Standardfehler aus der Stichprobe geschätzt werden müssen ist die theoretische Teststatistik t-verteilt mit $n-k$ Freiheitsgraden. Die theoretische Teststatistik (Zufallsvariable) und empirische Teststatistik (Realisation) ist allgemein

$$\text{theoret.:} \quad \hat{t} = \frac{\widehat{\beta}_h - \beta_0}{\hat{\sigma}_{\widehat{\beta}_h}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-k} \qquad \text{empirisch:} \quad t^{\text{emp}} = \frac{b_h - \beta_0}{s_{\widehat{\beta}_h}}$$

und für die Nullhypothese $\mathrm{H}_0\colon \beta_h = 0$ (d.h. $\beta_0 = 0$) natürlich

$$\text{theoret.:} \quad \hat{t} = \frac{\widehat{\beta}_h}{\hat{\sigma}_{\widehat{\beta}_h}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-k} \qquad \text{empirisch:} \quad t^{\text{emp}} = \frac{b_h}{s_{\widehat{\beta}_h}}$$

d.h., für diese einfache Nullhypothese erhalten wir den empirischen t-Wert, indem wir einfach den Koeffizienten durch dessen Standardfehler dividieren.

Zur Illustration greifen wir auf ein früheres Beispiel mit den Stundenlöhnen (StdL) zurück, die wir auf eine Dummyvariable m für männlich regressieren, d.h. $m = 1$ für Männer und $m = 0$ sonst. Wie wir schon im Kapitel zur deskriptiven Regressionsanalyse gezeigt haben misst in diesem Fall das Interzept den durchschnittlichen Stundenlohn der Referenzkategorie, in diesem Fall Frauen, und der Steigungskoeffizient die Differenz zum durchschnittlichen Stundenlohn von Männern.

Das Regressionsergebnis ist

$$\begin{array}{rcccc} \text{StdL} & = & 12.5 & + & 2.5\ m \\ & & (0.747)^{***} & & (1.057)^{**} \\ & & R^2 = 0.359, & & n = 12 \end{array}$$

(OLS Standardfehler in Klammern)

Wir nehmen an, dass alle Gauss-Markov Annahmen erfüllt sind und $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ (was für eine so kleine Stichprobe eine ziemlich verrückte Annahme ist).

Wie schon erwähnt ist die am häufigsten getestete Nullhypothese in Regressionsmodellen, dass die Parameter β_h der Grundgesamtheit Null sind, also keinen Einfluss auf die abhängige Variable haben.

In diesem Beispiel testen wir

$$\mathrm{H}_0 \colon \beta_2 = 0 \text{ gegen die Alternativhypothese } \mathrm{H}_A : \beta_2 \neq 0$$

Als Teststatistik wählen wir (derzeit in Ermangelung besserer Alternativen) die übliche t Statistik mit $n - 2 = 10$ Freiheitsgraden

$$\hat{t} = \frac{\widehat{\beta}_2 - \beta_0}{\hat{\sigma}_{\widehat{\beta}_h}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-2}$$

mit $\beta_0 = 0$. Als Signifikanzniveau α wählen wir alten Traditionen folgend 0.05.

Auf Grundlage des a priori gewählten Signifikanzniveaus $\alpha = 0.05$ und der Freiheitsgrade können wir die kritischen t-Werte bestimmen.

Da es sich um eine zweiseitige Hypothese handelt suchen wir den Wert $t^{\text{crit}}_{\alpha/2,\text{df}}$, bei welchem links und rechts $\alpha/2 * 100\%$ der Fläche abgeschnitten werden.

Mit Hilfe der kritischen Werte können wir den Annahmebereich $[-t^{\text{crit}}_{\alpha/2}, +t^{\text{crit}}_{\alpha/2}]$ und die Verwerfungsbereiche $[-\infty, -t^{\text{crit}}_{\alpha/2})$ und $(+t^{\text{crit}}_{\alpha/2}, +\infty]$ festlegen.

In unserem Beispiel mit $t^{\text{crit}}_{0.025,10} = 2.228$, siehe Tabelle 5.1 (Seite 20), ist der Annahmebereich $[-2.228;\ +2.228]$, und die Verwerfungsbereiche sind $[-\infty;\ -2.228)$ und $(+2.228;\ +\infty]$.

Bisher benötigten wir, abgesehen von den Freiheitsgraden, keine Informationen aus der Stichprobe.

Erst im nächsten Schritt verwenden wir die vorliegende Stichprobe um die *Schätzungen* b_2 und $s_{\widehat{\beta}_2}$ zu berechnen. Durch Einsetzen dieser Schätzungen und der Vermutung β_0 in die Teststatistik erhalten wir den empirischen t-Wert

$$t^{\text{emp}} = \frac{b_2 - \beta_0}{s_{\widehat{\beta}_2}} = \frac{2.5 - 0}{1.057} = +2.365$$

Im letzten Schritt brauchen wir nur noch zu überprüfen, in welchen Bereich t^{emp} fällt.

- Wenn t^{emp} in den *Verwerfungsbereich* fällt ist entweder ein sehr unwahrscheinliches Ereignis eingetreten, oder die Nullhypothese ist falsch. Was 'hinreichend unwahrscheinlich' bedeutet haben wir bereits bei der Wahl des Signifikanzniveaus entschieden.
- Wenn t^{emp} in den *Annahmebereich* fällt stehen die Stichprobendaten in keinem starken Widerspruch zur Nullhypothese, wir können die Nullhypothese *nicht* verwerfen.

In diesem Beispiel fällt $t^{\text{emp}} = 2.365$ in den Verwerfungsbereich $(+2.228;\ +\infty]$, also werden wir die Nullhypothese auf einem Signifikanzniveau von 5% verwerfen.

Wir finden in der Stichprobe also empirische Evidenz dafür, dass es auch in der Grundgesamtheit Unterschiede im Stundenlohn von Männern und Frauen gibt (zweiseitige Hypothese).

Hinweis: Hätten wir a priori ein Signifikanzniveau von 1% ($\alpha = 0.01$) festgelegt hätten wir den kritischen Wert $t^{\text{crit}}_{0.005,10} = 3.1693$ erhalten. Der Akzeptanzbereich für dieses Signifikanzniveau ist $[-3.1693;\ +3.1693]$, unsere $t^{\text{emp}} = 2.365$ fällt also klar in diesen Akzeptanzbereich und wir hätten die Nullhypothese auf einem Signifikanzniveau von 1% *nicht* verwerfen können.

Faustregel:

> Für Stichproben mit mehr als 30 Beobachtungen ($n > 30$) ist der Koeffizient 'vermutlich' auf dem 5% Signifikanzniveau *signifikant* von Null verschieden, wenn der Absolutbetrag des empirischen t-Wertes größer als 2 ist.

(für $n > 30$ ist die Stichprobenkennwertverteilung annähernd normalverteilt, und $z^{\text{crit}}_{0.025} = 1.96 \approx 2$. Da die empirische Teststatistik $t^{\text{emp}} = b_h / s_{\widehat{\beta}_h}$ folgt obige Faustregel.)

Achtung: Auch Neyman-Pearson Signifikanztests liefern eine *bedingte* Wahrscheinlichkeitsaussage: *gegeben die Nullhypothese ist wahr* und alle erforderlichen Annahmen sind erfüllt, dann können wir bei wiederholter Durchführung des Zufallsexperiments darauf vertrauen, dass die empirische Teststatistik nur in $\alpha \times 100$ Prozent der Fälle in den Verwerfungsbereich fällt.

Weder Fishers p-Werte noch Neyman-Person Signifikanztests liefern per se eine Aussage darüber, ob die Nullhypothese wahr ist oder nicht (!), sondern anders herum, wie groß die Wahrscheinlichkeit dafür ist, dass die empirische Teststatistik in den Verwerfungsbereich fällt, *wenn* die Nullhypothese tatsächlich wahr ist!

5.2.5 t-Test für eine Linearkombination von mehreren Regressionskoeffizienten*

Wir haben bisher nur t-Tests für *einzelne* Regressionskoeffizienten durchgeführt. Die Vorgangsweise lässt sich aber einfach für Tests einer Linearkombination von mehreren Regressionskoeffizienten verallgemeinern.

Beginnen wir mit einer einfachen PRF:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

Wenn wir zum Beispiel testen möchten, ob die Summe von β_2 und β_3 den Wert Eins hat, lauten Null- und Alternativhypothese

$$\mathrm{H}_0 : \ \beta_2 + \beta_3 = 1 \quad \text{gegen} \quad \mathrm{H}_A : \ \beta_2 + \beta_3 \neq 1$$

Dies ist ein sehr spezieller Fall von Null- und Alternativhypothese.

Etwas allgemeiner können zweiseitige Null- und Alternativhypothesen für dieses einfache Modell geschrieben werden als

$$\begin{aligned} \mathrm{H}_0 &: c_1\beta_2 + c_2\beta_3 = \beta_0 \\ \mathrm{H}_A &: c_1\beta_2 + c_2\beta_3 \neq \beta_0 \end{aligned}$$

wobei $c_1,\ c_2$ und $\beta_0 \in \mathbb{R}$ unter der Nullhypothese vermutete Parameter (Konstante) sind.

Für die obige Nullhypothese $\beta_2 + \beta_3 = 1$ sind $c_1 = c_2 = 1$ und $\beta_0 = 1$.

Sollte z.B. getestet werden, ob die Parameter β_2 und β_3 den gleichen Wert haben, würden wir $c_1 = +1,\ c_2 = -1$ und $\beta_0 = 0$ wählen, denn daraus folgt $\mathrm{H}_0 : \ \beta_2 - \beta_3 = 0$, bzw. $\beta_2 = \beta_3$.

Auf diese Weise lassen sich beliebige lineare Nullhypothesen testen, die zwei oder auch mehrere Parameter betreffen. Allerdings können auf diese Weise nur einzelne Hypothesen getestet werden, einen simultanen Test mehrerer Hypothesen werden wir erst in einem späteren Abschnitt kennen lernen.

Das Testprinzip für eine einzelne lineare Hypothese, die mehrere Parameter betrifft, ist einfach: wenn die Nullhypothese wahr ist gilt in der Grundgesamtheit

$$c_1\beta_2 + c_2\beta_3 - \beta_0 = 0$$

Selbst wenn dies in der Grundgesamtheit exakt gilt müssen wir in der Stichprobe damit rechnen, dass aufgrund von Zufallschwankungen dieser Zusammenhang nicht exakt erfüllt ist, also

$$c_1\widehat{\beta}_2 + c_2\widehat{\beta}_3 - \beta_0 = v$$

wobei die Zufallsvariable v 'relativ nahe' bei Null liegen sollte, *wenn die Nullhypothese wahr ist*. Man beachte, dass v eine Linearkombination der Zufallsvariablen $\widehat{\beta}_2$ und $\widehat{\beta}_3$ ist. Wenn $\widehat{\beta}_2$ und $\widehat{\beta}_3$ normalverteilt sind, dann ist auch v normalverteilt. Außerdem ist bei Gültigkeit der Nullhypothese $\mathrm{E}(v) = 0$. Um eine Teststatistik zu erhalten brauchen wir also nur durch den Standardfehler von v zu dividieren. Da v eine Linearkombination der Zufallsvariablen $\widehat{\beta}_2$ und $\widehat{\beta}_3$ ist gestaltet sich dies sehr einfach, wir brauchen nur die Rechenregel für das Rechnen mit Varianzen anzuwenden[7]

$$\widehat{\mathrm{var}}(v) = \widehat{\mathrm{var}}(c_1\widehat{\beta}_2 \pm c_2\widehat{\beta}_3 - \beta_0) = c_1^2\,\widehat{\mathrm{var}}(\widehat{\beta}_2) + c_2^2\,\widehat{\mathrm{var}}(\widehat{\beta}_3) \pm 2c_1c_2\,\widehat{\mathrm{cov}}(\widehat{\beta}_2, \widehat{\beta}_3)$$

[7] $\mathrm{var}(c_0 + c_1x \pm c_2y) = \mathrm{E}[(c_0 + c_1x \pm c_2y) - \mathrm{E}(c_0 + c_1x \pm c_2y)]^2 = c_1^2\,\mathrm{var}(x) + c_2^2\,\mathrm{var}(y) \pm 2c_1c_2\,\mathrm{cov}(x, y)$.

Wir haben bisher zwar nur für das bivariate Modell einen Schätzer für die Kovarianz zwischen Interzept und Steigungskoeffizient, d.h. $\widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_2)$, des bivariaten Modells hergeleitet, aber wir werden im Kapitel zur Matrixschreibweise zeigen, dass auch die Kovarianzen zwischen Steigungskoeffizienten ähnlich einfach berechnet werden können. Die üblichen Programme geben die Varianz-Kovarianzmatrix der Koeffizienten zwar nicht unmittelbar aus, aber man kann einfach mit den entsprechenden Befehlen darauf zugreifen.[8]

Da die Varianz des Nenners aus der Stichprobe berechnet werden muss ist die Teststatistik[9]

$$\frac{v}{\widehat{\text{se}}(v)} = \frac{c_1\widehat{\beta}_2 \pm c_2\widehat{\beta}_3 - \beta_0}{\sqrt{c_1^2\,\widehat{\text{var}}(\widehat{\beta}_2) + c_2^2\,\widehat{\text{var}}(\widehat{\beta}_3) \pm 2c_1c_2\,\widehat{\text{cov}}(\widehat{\beta}_2, \widehat{\beta}_3)}} \overset{\text{H}_0}{\sim} \text{t}_{n-k}$$

unter H_0 t-verteilt mit $n - k$ Freiheitsgraden.

Diese Teststatistik ($\hat{t}$) hat die Form

$$\hat{t} = \frac{\text{Empirisches Analogon zur H}_0 \text{ in impliziter Form}}{\text{Standardfehler vom Zähler}} \overset{\text{H}_0}{\sim} \text{t}_{n-k}$$

Dies ist offensichtlich nur eine kleine Erweiterung der üblichen *Tests für einen Koeffizienten* für Fälle, in denen *eine* Hypothese über eine Linearkombination von zwei (oder mehreren) Koeffizienten getestet werden soll.

Mit Hilfe dieser Teststatistik kann wieder der p-Wert à la Fisher berechnet werden, oder eine Entscheidung nach der Methode Neyman und Pearson gefällt werden.

Beispiel: Angenommen wir erhalten folgende Schätzung einer Cobb-Douglas Produktionsfunktion und möchten testen, ob konstante Skalenerträge vorliegen.[10]

$$\begin{array}{rcccccc} \log(\text{Q}) & = & 2.481 & + & 0.64\,\log(\text{K}) & + & 0.257\,\log(\text{L}) \\ & & (0.129)^{***} & & (0.035)^{***} & & (0.027)^{***} \end{array}$$

$$R^2 = 0.942, \qquad n = 25$$

(OLS Standardfehler in Klammern)

Die dazugehörige Varianz-Kovarianzmatrix der Koeffizienten (*Coefficient Covariance Matrix*) ist

	b_1	b_2	b_3
b_1	0.016543	−0.002869	−0.003187
b_2	−0.002869	0.001206	0.000300
b_3	−0.003187	0.000300	0.000727

[8] In R erhält man die geschätzte Varianz-Kovarianzmatrix der Koeffizienten mit der Funktion `vcov(...)`; in Stata mit dem postestimation Befehl `e(V)`, und in EViews mit `@coefcov`.

[9] Dieser Test ist wie alle t-Tests ein Spezialfall eines Wald Tests, der 1943 von dem Mathematiker Abraham Wald entwickelt wurde.

[10] Eine Cobb-Douglas Funktion hat die Form $Q = AK^{\beta_2}L^{\beta_3}$. Durch Logarithmieren erhält man eine in den Parametern lineare Funktion $\ln(Q) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L)$, mit $\beta_1 = \ln(A)$ die einfach mit OLS geschätzt werden kann.

Bei Cobb-Douglas Funktionen liegen konstante Skalenerträge vor, wenn $\beta_2+\beta_3 = 1$. Die Nullhypothese ist also

$$\mathrm{H}_0: \ \beta_2 + \beta_3 = 1$$

(diese Hypothese ist ein Spezialfall von $\mathrm{H}_0: \ c_1\beta_2 + c_2\beta_3 = \beta_0$ mit $c_1 = c_2 = \beta_0 = 1$)

Die entsprechende t-Statistik ist deshalb

$$\begin{aligned} t^{\text{emp}} &= \frac{b_2 + b_3 - 1}{\sqrt{\widehat{\text{var}}(b_2) + \widehat{\text{var}}(b_3) + 2\,\widehat{\text{cov}}(b_2, b_3)}} \\ &= \frac{0.640 + 0.257 - 1}{\sqrt{0.001206 + 0.000727 + 2 \times 0.000300}} \\ &= \frac{-0.10255}{\sqrt{0.002533}} = -2.037595955 \end{aligned}$$

Der kritische t-Wert für einen zweiseitigen Test mit $n - k = 22$ Freiheitsgraden ist $t^{\text{crit}}_{0.025,22} = 2.0739$, deshalb können wir die Nullhypothese konstanter Skalenerträge (sehr knapp) nicht verwerfen.

Um den p-Wert zu berechnen benötigt man den Wert der Verteilungsfunktion. Wenn wir mit $\Phi(x, q)$ den Wert der Verteilungsfunktion einer t-Verteilung mit q Freiheitsgraden an der Stelle x bezeichnen errechnet sich der p-Wert als $p = 2(1 - \Phi(+2.037595955, 22)) = 0.05379$.[11]

Code zum Beispiel* Solche Tests sind in fast allen Statistik- und Ökonometrieprogrammen fix implementiert, aber man kann sie auch einfach selbst programmieren.

R:

```
rm(list=ls(all=TRUE))
CD <- read.table("https://www.uibk.ac.at/econometrics/data/prodfunkt.csv",
     dec = ".", sep=";", header=TRUE)
eq1 <- lm(log(Q) ~ log(K) + log(L), data = CD)

# Hypothesentest für H_0: beta_2 + beta_3 = 1
library(car)
linearHypothesis(eq1, "log(K) + log(L) = 1")

# oder sehr ausführlich:
tstat <- (coef(eq1)[2] + coef(eq1)[3] - 1) /
     sqrt(vcov(eq1)[2,2] + vcov(eq1)[3,3] + 2*vcov(eq1)[2,3])
pval <- 2*(1-(pt(abs(tstat),22))) ## 0.05379616
```

Stata:

```
insheet using "https://www.uibk.ac.at/econometrics/data/prodfunkt.csv", ///
              delim(";") names clear
generate logQ = log(q)
generate logK = log(k)
generate logL = log(l)
```

[11] In Excel erhalten Sie den p-Wert mit der Funktion '=TVERT(ABS(-2.037595955);22;2)'

```
regress logQ logK logL

# Hypothesentest für H_0: beta_2 + beta_3 = 1
test logK + logL = 1

# oder sehr ausführlich:
matrix V = e(V)   // Var-Cov-Matrix der Koeff.
scalar tstat = (_b[logK] + _b[logL] - 1) / ///
               sqrt( V[1,1] + V[2,2] + 2*V[1,2])
display "t-statistic: " tstat
display "p-value: " 2*ttail(22,abs(tstat))
```

□

Einseitige Signifikanztests nach Neyman-Pearson

Einseitige Tests laufen exakt nach dem gleichen Muster ab, nur in zwei Details unterscheiden sie sich von zweiseitigen Tests. Erstens liegt der Verwerfungsbereich je nach Richtung der Nullhypothese zur Gänze am linken *oder* am rechten Ende der Dichtefunktion, deshalb ist der *kritische Wert* α (statt wie beim zweiseitigen Test $\alpha/2$) zu wählen, und zweitens ist natürlich das Vorzeichen des empirischen und des kritischen Wertes (t^{emp} und $\text{t}_\alpha^{\text{crit}}$) zu beachten. Im übrigen gilt, was schon zu einseitigen Tests nach Fisher gesagt wurde.

Die Begriffe linksseitig bzw. rechtsseitig beziehen sich jeweils auf den *Verwerfungsbereich*, bei linksseitigen Tests liegt dieser zur Gänze im negativen Bereich der Zahlengerade, bei rechtsseitigen Tests im positiven Bereich der Zahlengerade. Dementsprechend liegt die Null immer im Bereich der Nullhypothese.

Abbildung 5.7 soll den Unterschied zwischen ein- und zweiseitigen Hypothesentests verdeutlichen. Die obere Abbildung zeigt den bisherigen zweiseitigen Test, die untere Abbildung einen *rechtsseitigen* Test (d.h. einen Test für $\text{H}_0: \beta_h \leq \beta_0$), und die unterste Abbildung einen linksseitigen Test für $\text{H}_0: \beta_h \geq \beta_0$.

Hinweis: Wenn man unsicher ist wo der Verwerfungsbereich einer einseitigen Nullhypothese liegt ist es manchmal nützlich die Nullhypothese in impliziter Form zu schreiben, d.h. statt $\text{H}_0: \beta_h \leq \beta_0$ schreiben wir $\text{H}_0: \beta_h - \beta_0 \leq 0$. In dieser Form ist einfach zu erkennen, dass negative Werte *nicht* im Widerspruch zur Nullhypothese stehen, und aufgrund des stochastischen Charakters sind auch *kleine* positive Werte keine große Überraschung. Erst Werte die größer sind als der kritische Wert $\text{t}_\alpha^{\text{crit}}$ stehen in starkem Widerspruch zur H_0, deshalb wird der Verwerfungsbereich zur Gänze im positiven Bereich liegen, es handelt sich also um eine rechtsseitige Hypothese.

Analoges gilt für linksseitige Hypothesen.

	rechtsseitig: $\text{H}_0: \beta_h - \beta_0 \leq 0$	linksseitig: $\text{H}_0: \beta_h - \beta_0 \geq 0$
Akzeptanzbereich:	$[-\infty, +\text{t}_\alpha^{\text{crit}}]$	$[-\text{t}_\alpha^{\text{crit}}, +\infty]$
Verwerfungsbereich:	$(+\text{t}_\alpha^{\text{crit}}, +\infty]$	$[-\infty, -\text{t}_\alpha^{\text{crit}})$

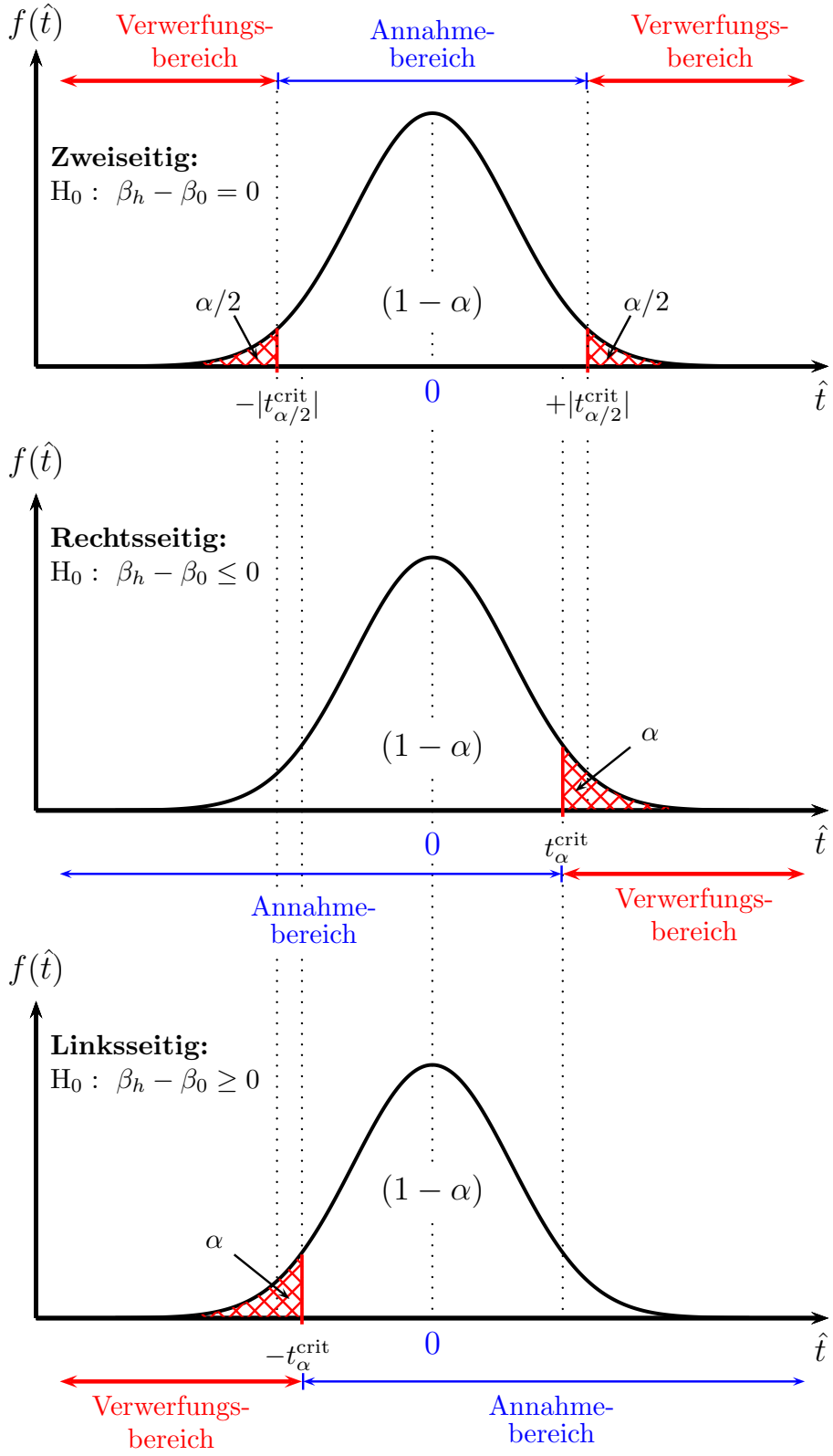


Abbildung 5.7: Zweiseitiger und links- bzw. rechtsseitiger Test mit gleichem Signifikanzniveau α.

Beispiel: Wir haben vorhin mit Hilfe der Regression

$$\begin{array}{rcccc} \text{StdL} & = & 12.5 & + & 2.5\ m \\ & & (0.747)^{***} & & (1.057)^{**} \end{array}$$
$$R^2 = 0.359, \qquad n = 12$$
(OLS Standardfehler in Klammern)

die zweiseitige Hypothese getestet, dass es Lohnunterschiede zwischen Männern und Frauen gibt.

Angenommen wir möchten die Nullhypothese testen, dass die Stundenlöhne von Männern *niedriger* sind als die Stundenlöhne von Frauen, oder in anderen Worten, dass der 'wahre' Koeffizient der Dummyvariable m *nicht positiv* ist (d.h. dass Männer höchstens gleich viel oder weniger verdienen)

$$\mathrm{H}_0\colon \beta_2 <= 0 \text{ gegen die Alternativhypothese } \mathrm{H}_A : \beta_2 > 0$$

Negative Werte des Koeffizienten würden also nicht im Widerspruch zu dieser Nullhypothese stehen $\beta_2 - 0 <= 0$, und selbst leicht positive Werte würden keine große Überraschung darstellen.

Wie schon früher erwähnt brauchen wir nur den Extremfall $\beta_2 = 0$ zu testen, denn wenn wir diese Hypothese verwerfen können, können wir auch alle extremeren Hypothesen $\beta_2 < 0$ verwerfen. Wir können also die gleiche empirische Teststatistik wie für den zweiseitigen Test verwenden.

Als Teststatistik wählen wir wieder die übliche t Statistik mit $n - 2 = 10$ Freiheitsgraden

$$\hat{t} = \frac{\widehat{\beta}_2 - \beta_0}{\hat{\sigma}_{\widehat{\beta}_h}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-2}$$

mit $\beta_0 = 0$, und als Signifikanzniveau α wählen wir wieder 0.05.

Wir nehmen wieder an, dass alle Gauss-Markov Annahmen erfüllt sind und $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$.

Auf Grundlage des a priori gewählten Signifikanzniveaus $\alpha = 0.05$ und der Freiheitsgrade $df = 10$ können wir aus Tabelle 5.1 den kritischen t-Wert $t^{\mathrm{crit}_{0.05,10}} = 1.812$ entnehmen (da es sich um einen einseitigen Test handelt müssen wir bei $\alpha = 0.05$ nachschlagen).

Dies definiert den Annahme- und Verwerfungsbereich. Es handelt sich um einen rechtsseitigen Test, da negative Werte in keinem Widerspruch zur Nullhypothese stehen

Akzeptanzbereich:	$(-\infty,\ +t^{\mathrm{crit}}_{\alpha,df}]$	$\rightarrow$	$(+\infty,\ +1.812]$
Verwerfungsbereich:	$(+t^{\mathrm{crit}}_{\alpha,df},\ +\infty]$	$\rightarrow$	$(+1.812,\ +\infty]$

Erst jetzt benötigten wir die Informationen aus der Stichprobe, d.h., die empirische Teststatistik

$$t^{\mathrm{emp}} = \frac{b_2 - \beta_0}{s_{\widehat{\beta}_2}} = \frac{2.5 - 0}{1.057} = +2.365$$

Dieser empirische Wert fällt klar in den Verwerfungsbereich, d.h. wenn die Nullhypothese $\beta_2 <= 0$ wahr ist und alle erforderlichen Annahmen erfüllt sind, würden

wir in weniger als 5% der möglichen Stichprobenziehungen ein so extremes Resultat erwarten.

Aus diesem Grund verwerfen wir die Nullhypothese und schließen, dass es empirische Evidenz dafür gibt, dass die Stundenlöhne von Männern auch in der Grundgesamtheit höher sind als die von Frauen.

Durch die Einführung einer expliziten Alternativhypothese und eines a priori festgelegten Signifikanzniveaus α kann eine eindeutige Entscheidungsregel gefunden werden, ob die Nullhypothese verworfen oder nicht verworfen werden soll. Diese Entscheidungsregel ist zwar eindeutig, aber das bedeutet nicht, dass die Entscheidung auch richtig sein muss.

5.2.6 Typ I & Typ II Fehler

Aufgrund des stochastischen Charakters der Teststatistik können zwei Arten von Fehlentscheidungen passieren. Zum einen kann eine tatsächlich richtige Nullhypothese irrtümlich verworfen werden (Typ I Fehler, *Verwerfungsfehler*), oder eine tatsächlich falsche Nullhypothese kann irrtümlich *nicht* verworfen werden (Typ II Fehler, *Nicht-Verwerfungsfehler*). Dies wird in Tabelle 5.2 dargestellt.

Tabelle 5.2: Typ I und Typ II Fehler

Entscheidung auf Grundlage eines statistischen Tests:	**Wahrer Sachverhalt: H_0 ist wahr**	**Wahrer Sachverhalt: H_0 ist falsch**
Nullhypothese H_0 wird *nicht* verworfen	korrekte Entscheidung $(1-\alpha)$	*Typ II Fehler*
Nullhypothese H_0 wird verworfen	*Typ I Fehler*	korrekte Entscheidung ($1-\beta$: "*Power*")

Den Typ I Fehler haben wir bereits kennengelernt: selbst wenn die Nullhypothese die Grundgesamtheit exakt beschreibt und alle Annahmen erfüllt sind, müssen wir aufgrund der Stichprobenfehler in 5% der Fälle damit rechnen, einen p-Wert < 0.05 zu erhalten.

Der wesentliche Grund, warum Neyman-Pearson ein fest vorgegebenes Signifikanzniveau α einführten, besteht darin, dass erst dies die Bestimmung eines Typ II Fehlers ermöglicht (und damit, wie wir später sehen werden, die *Power* einer Teststatistik).

Bei einem *Typ II Fehler* wird eine tatsächlich falsche Nullhypothese *nicht* verworfen (*Nicht-Verwerfungsfehler*).

Wenn wir z.B. vermuten, dass eine Variable x_h die abhängige Variable $\widehat{y}$ beeinflusst, dann ist die Nullhypothese (= Gegenhypothese) für den Regressionskoeffizienten $H_0: \beta_h - \beta_0 = 0$.

Aufgrund des Stichprobenfehlers kann es allerdings passieren, dass die zufällige Realisation der Teststatistik *keine* Verwerfung der Nullhypothese erlaubt, obwohl die

Variable x_h in der Grundgesamtheit sehr wohl einen Einfluss auf $\widehat{y}$ ausübt. In diesem Fall machen wir einen Typ II Fehler.

In diesem Sinne bedeutet ein Typ II Fehler immer eine *"verpasste Entdeckung"*, ein tatsächlich existierender Zusammenhang oder Unterschied wird nicht erkannt.

Man beachte, dass die Entscheidung immer auf Grundlage der H_0 erfolgt, die üblicherweise als Negativhypothese zu unserer Anfangsvermutung formuliert wird. Dies impliziert eine Asymmetrie zwischen Null- und Alternativhypothese. Um diese Asymmetrie zu rechtfertigen werden Hypothesentests nach Neyman-Pearson häufig mit der Unschuldsvermutung bei einem Gerichtsprozess verglichen. In einer Welt ohne Unsicherheit sollten Unschuldige frei gesprochen und Schuldige verurteilt werden. In einer Welt mit Unsicherheit kann es bei Indizienprozessen allerdings passieren, dass – analog zu einem Typ I Fehler – ein Unschuldiger verurteilt wird, oder – analog zu einem Typ II Fehler – ein Schuldiger freigesprochen wird; vergleiche Tabelle 5.3.

Tabelle 5.3: Vergleich Typ I und Typ II Fehler mit einem Gerichtsurteil

	Angeklagter ist unschuldig	**Angeklagter ist schuldig**
Gericht fällt Entscheidung "unschuldig"	richtige Entscheidung	Schuldiger wird freigesprochen
Gericht fällt Entscheidung "schuldig"	Unschuldiger wird verurteilt	richtige Entscheidung

Wenn alle erforderlichen Annahmen erfüllt sind wird die Wahrscheinlichkeit einen Typ I Fehler zu machen unmittelbar durch das Signifikanzniveau α bestimmt.

$$\begin{aligned}\Pr\left[\text{Typ I Fehler}\right] &= \Pr\left[\text{H}_0 \text{ ablehnen}|\text{H}_0 \text{ wahr}\right] \\ &= \alpha\end{aligned}$$

Man beachte, dass es sich dabei um eine *bedingte* Wahrscheinlichkeit handelt: α gibt die Wahrscheinlichkeit dafür an, dass die Teststatistik in den Verwerfungsbereich fallen wird, *gegeben dass die Nullhypothese wahr ist.*

Wann immer die potentiellen Kosten eines Typ I Fehlers sehr hoch sind sollte ein entsprechend kleines Signifikanzniveau gewählt werden (z.B. $\alpha = 0.01$).

Es zeigt sich aber, dass dies auch Kosten hat, denn die Wahl eines kleinen α erhöht automatisch die Wahrscheinlichkeit eines Typ II Fehlers.

Ein Typ II Fehler (manchmal auch β-Fehler genannt) ist ein *'Nicht-Verwerfungs Fehler'*, eine tatsächlich falsche Nullhypothese wird nicht abgelehnt

$$\Pr\left[\text{Typ II Fehler}\right] = \Pr\left[\text{H}_0 \text{ nicht ablehnen}|\text{H}_0 \text{ falsch}\right]$$

Dies wird anhand des Vergleichs mit einem Indizienprozess unmittelbar klar. Wenn man unbedingt vermeiden möchte einen Unschuldigen zu verurteilen (Typ I Fehler)

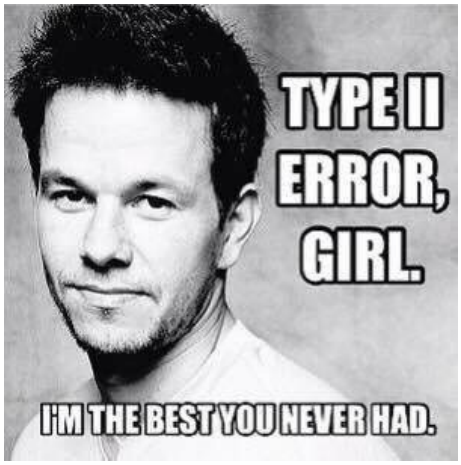


Abbildung 5.8: Typ II Fehler; Quelle: die (Un-)Tiefen des Internets.

und deshalb sehr strenge Anforderungen an die Beweise stellt (ein kleines α wählt), wird dies automatisch dazu führen, dass mehr Schuldige irrtümlich freigesprochen werden (Typ II Fehler)!

In der Medizin werden Typ I Fehler häufig *'false positive'* genannt, und Typ II Fehler *'false negative'*. Wenn mit Hilfe eines Tests festgestellt werden soll, ob jemand an einer Krankheit leidet oder nicht, bedeutet ein 'positiv', dass die Person an der Krankheit leidet, ein *'false positive'* bedeutet deshalb, dass der Test die Krankheit anzeigt, obwohl die Person gesund ist. Dies ist natürlich ein Typ I Fehler.

Wenn der Test hingegen anzeigt, dass die Person gesund ist (Ergebnis 'negativ'), obwohl die Person tatsächlich krank ist (*'false negative'*), dann passiert natürlich ein Typ II Fehler.

Beispiel Stellen Sie sich vor, es gibt 1000 zu testende Hypothesen, und 100 davon sind tatsächlich war. Aufgrund des Typ I Fehlers würden Sie bei einem $\alpha = 0.05$ ca. 45 (= 5% von 900) tatsächlich falsche Hypothesen für richtig halten, und damit einen Typ I Fehler machen.

Aufgrund des Typ II Fehlers würden Sie einige der tatsächlich richtigen Hypothesen nicht erkennen, d.h. für falsch halten. Wie groß dieser Anteil ist hängt von mehreren Faktoren ab, u.a. – wie wir später sehen werden – vom Siginifikanzniveau α. □

Da diese Überlegungen sehr allgemein gelten werden wir im Folgenden für den interessierenden Parameter wieder das Symbol θ (*theta*) verwenden, wobei θ wieder für einen Koeffizienten β_h, Mittelwert μ, Varianz σ^2 etc. stehen kann.

Um diese zwei Arten von Fehlern grafisch darstellen zu können wählen wir eine sehr einfache Null- und Alternativhypothese

$$\mathrm{H}_0 : \theta = \theta_0 \quad \text{gegen} \quad \mathrm{H}_A : \theta = \theta_A$$

wobei θ_0 und θ_A fixe Zahlen sind, d.h. sowohl Null- als auch Alternativhypothese umfassen nur einen einzelnen Punkt. Abbildung 5.9 zeigt die Stichprobenkennwertverteilungen unter H_0 und H_A.

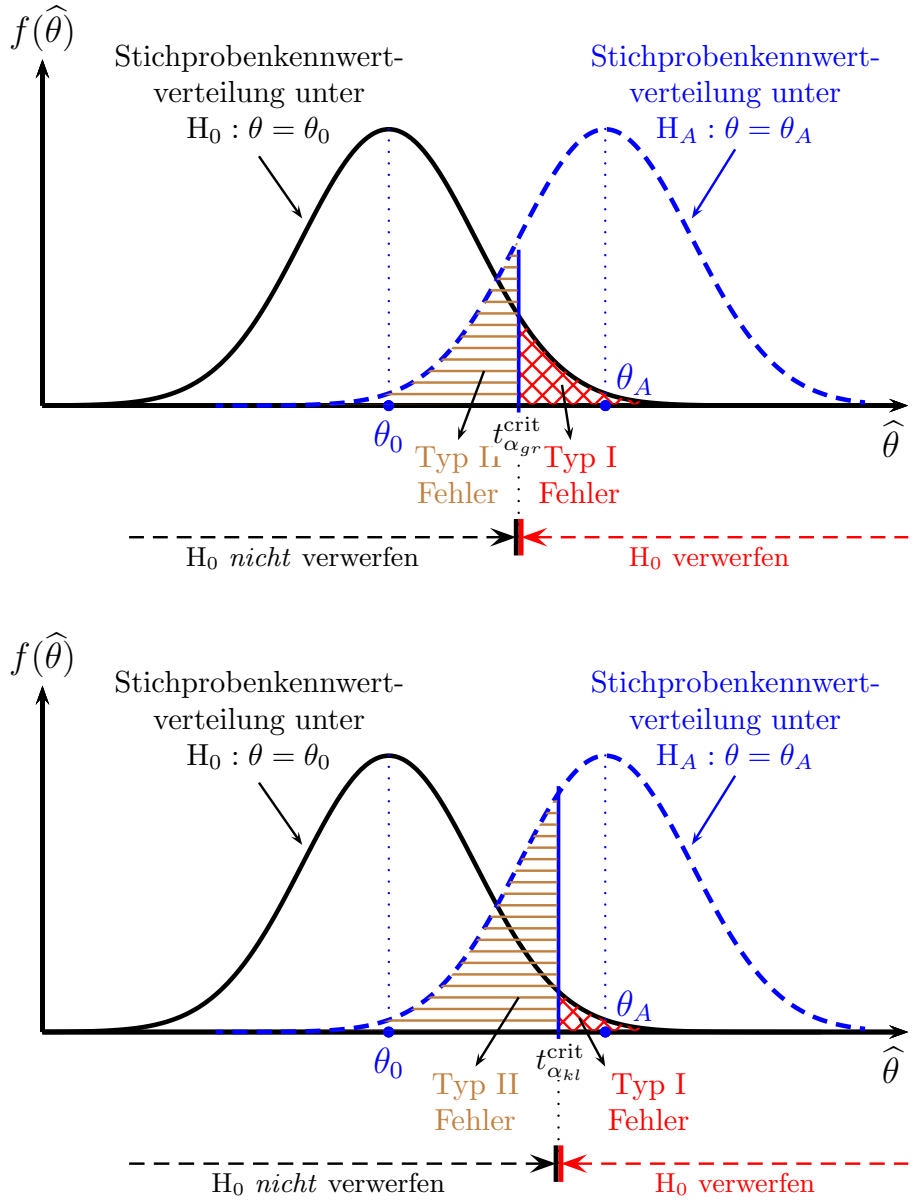


Abbildung 5.9: Typ I und Typ II Fehler: Durch die Wahl eines höheren Signifikanzniveaus (d.h. kleineren α) sinkt die Wahrscheinlichkeit eines Typ I Fehlers, aber dadurch steigt die Wahrscheinlichkeit eines Typ II Fehlers.[Folien: local]

Die linke durchgezogene Verteilung in Abbildung 5.9 ist die Stichprobenkennwertverteilung von $\widehat{\theta}$ unter Gültigkeit der Nullhypothese. Wenn die Nullhypothese tatsächlich wahr ist fallen bei oftmaliger Wiederholung $\alpha \times 100$ Prozent der geschätzten $\widehat{\theta}$ in den Bereich rechts von t^{crit} (einseitiger Test), und die schraffierte Fläche darüber gibt die Wahrscheinlichkeit für einen Typ I Fehler an.

Ein Typ II Fehler passiert hingegen, wenn eine tatsächlich falsche Nullhypothese *nicht* verworfen wird. Wenn die Nullhypothese falsch ist muss die Alternativhypothese wahr sein, in diesem sehr einfachen Fall, in dem die Alternativhypothese nur einen Punkt umfasst, also $\theta = \theta_A$.

Die rechte strichlierte Verteilung in Abbildung 5.9 zeigt die Stichprobenkennwertverteilung von $\widehat{\theta}$ wenn die Alternativhypothese richtig ist. Die Wahrscheinlichkeit eines Typ II Fehlers, also dass die Alternativhypothese richtig und die Nullhypothese trotzdem *nicht* abgelehnt wird, entspricht der horizontal schraffierten Fläche in Abbildung 5.9 (die Fläche unter der strichliert gezeichneten Verteilung links von t^{crit}).

Wie man aus dem Vergleich der oberen und unteren Abbildung erkennen kann führt die Wahl eines kleineren α zwar zu einer Abnahme der Wahrscheinlichkeit eines Typ I Fehlers, aber gleichzeitig zur Zunahme der Wahrscheinlichkeit eines Typ II Fehlers!

Tatsächlich wissen wir nicht, ob die Null- oder die Alternativhypothese richtig ist, deshalb können wir auch nicht wissen, welche der beiden Stichprobenkennwertverteilungen die 'wahre' ist. Aber wir wissen, dass nur entweder die Nullhypothese *oder* die Alternativhypothese richtig sein kann.

Wenn die Stichprobenkennwertverteilung unter H_0 wahr ist, machen wir bei einem einseitigen Test in $\alpha \times 100$ Prozent der Fälle einen Typ I Fehler. Wenn aber die Alternativhypothese wahr ist machen wir einen Typ II Fehler, dessen Wahrscheinlichkeit in Abbildung 5.9 durch die Fläche des horizontal schraffierten Bereichs (links von t^{crit}) gegeben ist. Man beachte, dass der Wert des Typ II Fehlers auch vom wahren θ abhängt.

Daraus lassen sich zwei wichtige Schlussfolgerungen ziehen:

1. Die Wahrscheinlichkeit von Typ I und Typ II Fehler sind invers verknüpft, d.h. die Wahl eines hohen Signifikanzniveaus (kleinen α) reduziert zwar die Wahrscheinlichkeit eines Typ I Fehlers, aber erhöht gleichzeitig die Wahrscheinlichkeit eines Typ II Fehlers (vergleiche obere und untere Grafik in Abbildung 5.9).

2. Die Wahrscheinlichkeit eines Typ II Fehlers ist ceteris paribus umso größer, je näher θ_A bei θ_0 liegt.

Auch wenn es nicht möglich ist beide Arten von Fehlern gleichzeitig zu kontrollieren, so kann man doch zeigen, dass diese Teststrategie zumindest 'bestmöglich' in dem Sinne ist, dass sie unter bestimmten Annahmen für eine gegebene Wahrscheinlichkeit eines Typ I Fehlers die Wahrscheinlichkeit eines Typ II Fehlers minimiert.

5.2.7 Trennschärfe (*Power*) eines Tests

Erinnern wir uns, Neyman-Pearson entwickelten ihre Methode u.a. um ein Kriterium zu finden, das eine Beurteilung der Güte von Testfunktionen ermöglichen sollte. Ideal wäre natürlich ein Test, der falsche Nullhypothesen mit Wahrscheinlichkeit Eins verwirft und korrekte Nullhyhpothesen mit Wahrscheinlichkeit Eins nicht verwirft, aber eine solche Teststatistik existiert für reale Situationen natürlich nicht.

Da die Wahrscheinlichkeiten für Typ I und Typ II Fehler invers verknüpft sind ist es hoffnungslos einen Test zu suchen, der beide Arten von Fehler minimiert. Deshalb entschieden sich Neyman-Pearson die Wahrscheinlichkeit für einen Typ I Fehler – das Signifikanzniveau α – vorzugeben und Tests zu suchen, die für ein vorgegebenes α die Wahrscheinlichkeit eines Typ II Fehlers minimieren. Das impliziert indirekt, dass Null- und Alternativhypothese asymmetrisch behandelt werden, ähnlich wie bei der Unschuldsvermutung vor Gericht. Diese Überlegungen führen zur Power eines Tests.

Die *Power eines Tests* gibt die Wahrscheinlichkeit dafür an, dass eine tatsächlich falsche Nullhypothese auch verworfen wird.

Da der Typ II Fehler die Wahrscheinlichkeit angibt, mit der eine falsche Nullhypothese *nicht* verworfen wird, ist die Power einfach die Gegenwahrscheinlichkeit zum Typ II Fehler.

$$\begin{aligned} \text{Power} &= \Pr\left[\text{H}_0 \text{ verwerfen } |\text{H}_0 \text{ ist falsch}\right] \\ &= 1 - \Pr\left[\text{H}_0 \text{ akzeptieren } |\text{H}_0 \text{ ist falsch}\right] \\ &= 1 - \Pr\left[\text{Typ II Fehler}\right] \end{aligned}$$

Dies wird in Abbildung 5.10 gezeigt.

Da die Power eine Wahrscheinlichkeit ist und als Fläche unter einer Dichtefunktion gemessen wird kann sie nur Werte zwischen Null und Eins annehmen. Natürlich sollte der Wert der '*Power*', d.h. die Wahrscheinlichkeit mit der eine falsche Nullhypothese tatsächlich verworfen wird, möglichst nahe bei Eins liegen.

Im vorhergehenden Beispiel hatten wir eine sehr einfache Null- und Alternativhypothese, nämlich $\text{H}_0 \colon\ \theta = \theta_0$ und $\text{H}_A :\ \theta = \theta_A$. Nun wollen wir eine etwas realistischere Alternativhypothese untersuchen, nämlich

$$\begin{aligned} \text{H}_0 \colon &\qquad \theta = \theta_0 \\ \text{H}_A :&\qquad \theta \neq \theta_0 \end{aligned}$$

Dies ändert nichts für den Typ I Fehler (Verwerfungsfehler), dieser wird nach wie vor durch das gewählte Signifikanzniveau angegeben.

Aber in diesem Fall kann die Wahrscheinlichkeit für einen Typ II Fehler (falsche H_0 akzeptieren) nicht mehr durch eine einfache Zahl angegeben werden, sondern hängt von θ ab. Wenn θ sehr nahe bei θ_0 liegt wird ceteris paribus die Wahrscheinlichkeit für einen Typ II Fehler höher sein, und also die Power des Tests niedriger sein.

Dies wird in Abbildung 5.11 gezeigt, die 'Power' ist ceteris paribus umso größer, je weiter der 'wahre' Wert θ von θ_0 entfernt liegt, sie hängt also von θ ab. Wenn man die Power als Funktion aller möglichen θ darstellt erhält man die *Teststärkefunktion* (*'power function'*), die im untersten Panel von Abbildung 5.11 dargestellt ist.

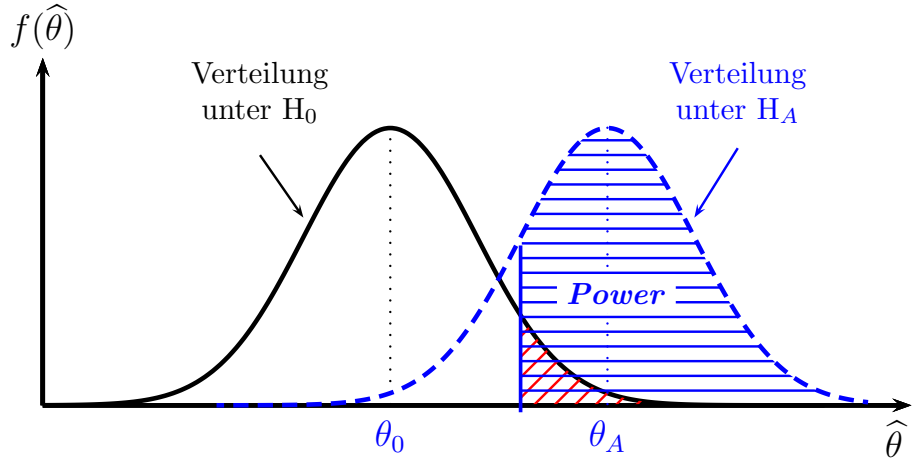


Abbildung 5.10: Die 'Power' oder Trennschärfe eines Tests ist *Eins minus Wahrscheinlichkeit eines Typ II Fehlers*, also die Wahrscheinlichkeit, mit der eine falsche Nullhypothese tatsächlich verworfen wird.

Effizientere Schätzfunktionen erlauben trennschärfere Tests

Die 'Power' eines Tests nimmt mit der Stichprobengröße zu. Abbildung 5.12 zeigt in der oberen Grafik die Power für eine kleine Stichprobe (kleines n), und in der unteren Grafik für eine große Stichprobe. Die zugrunde liegende Null- und Alternativhypothese ist in beiden Grafiken gleich. Offensichtlich ist die 'Power' bei der großen Stichprobe deutlich größer!

Man beachte, dass die Power aus zwei Gründen zunimmt: erstens ist die Varianz der unbeobachtbaren wahren Verteilung bei einer größeren Stichprobe kleiner, und zweitens liegt der kritische Wert der Verteilung $t_{\alpha_g}^{\text{crit}}$ näher beim θ_0.

Abbildung 5.13 zeigt zwei typische Teststärkefunktionen, eine mit kleinerer 'Power' (strichlierte Linie) und eine mit größerer 'Power'.

Frage: Wie sieht die Power-Funktion für einseitige (links- bzw. rechtsseitige) Tests aus?

Hinweis: Die Diskussion um die Power von Testfunktionen war übrigens auch ein Hauptgrund für das Zerwürfnis zwischen R.A. Fisher und Neyman-Pearson. Während der p-Wert nach Fisher zwar die Beurteilung der Wahrscheinlichkeit für einen Typ I Fehler erlaubt, kann die Wahrscheinlichkeit eines Typ II Fehlers damit nicht angegeben werden.

Neyman-Pearson suchten nach einer Möglichkeit, ähnlich wie für Schätzfunktionen Eigenschaften von *Teststatistiken* anzugeben, und stießen dabei auf die Power. Um die Power zu definieren muss allerdings ein vorgegebenes Signifikanzniveau α angenommen werden. Tatsächlich konnten sie mit Hilfe der Power 'Gleichmäßig beste Tests' (*'Uniformly Most Powerful (UMP) Tests'*) definieren, die – vereinfacht ausgedrückt – die Wahrscheinlichkeit eines Typ II Fehlers für einen beschränkten Bereich eines Typ I Fehlers minimieren. Es zeigte sich allerdings, dass solche UMP Tests nicht immer existieren; in der Ökonometrie spielt dieses Kriterium im allgemeinen keine große Rolle.

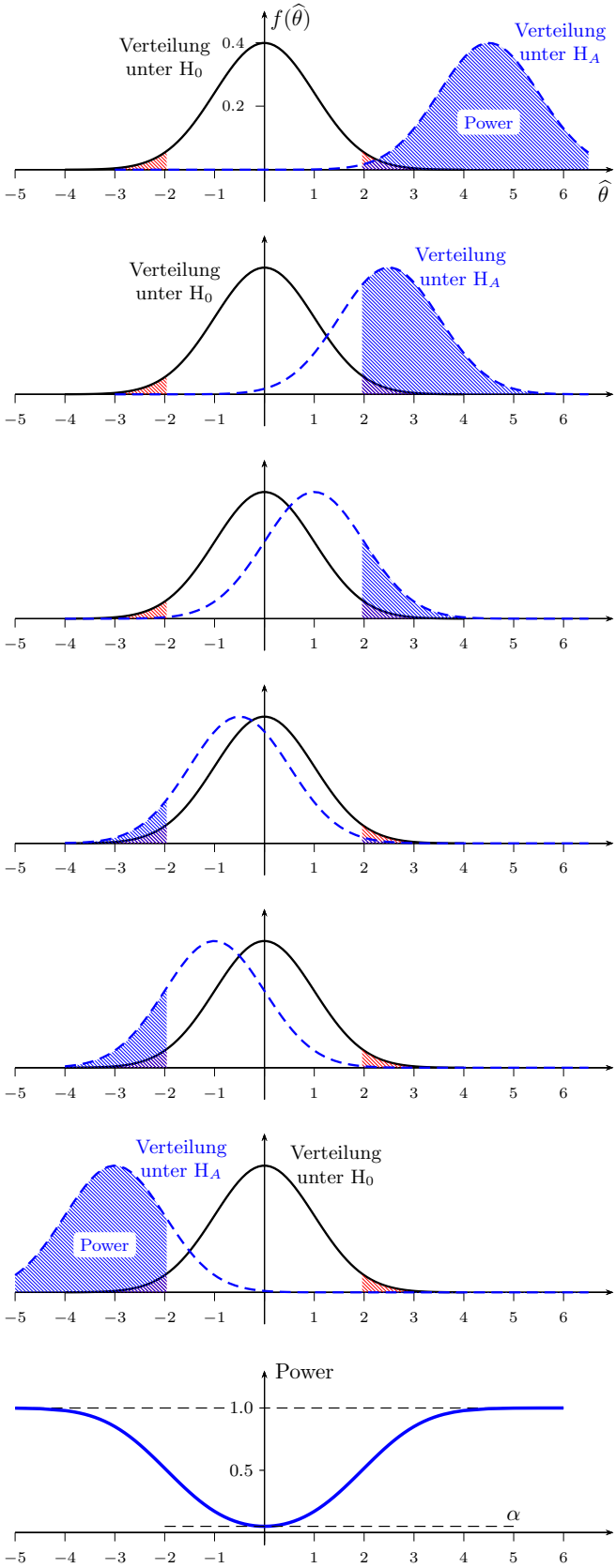


Abbildung 5.11: Powerfunktion eines zweiseitigen Tests. Die Power ist eine Funktion vom wahren Wert θ.

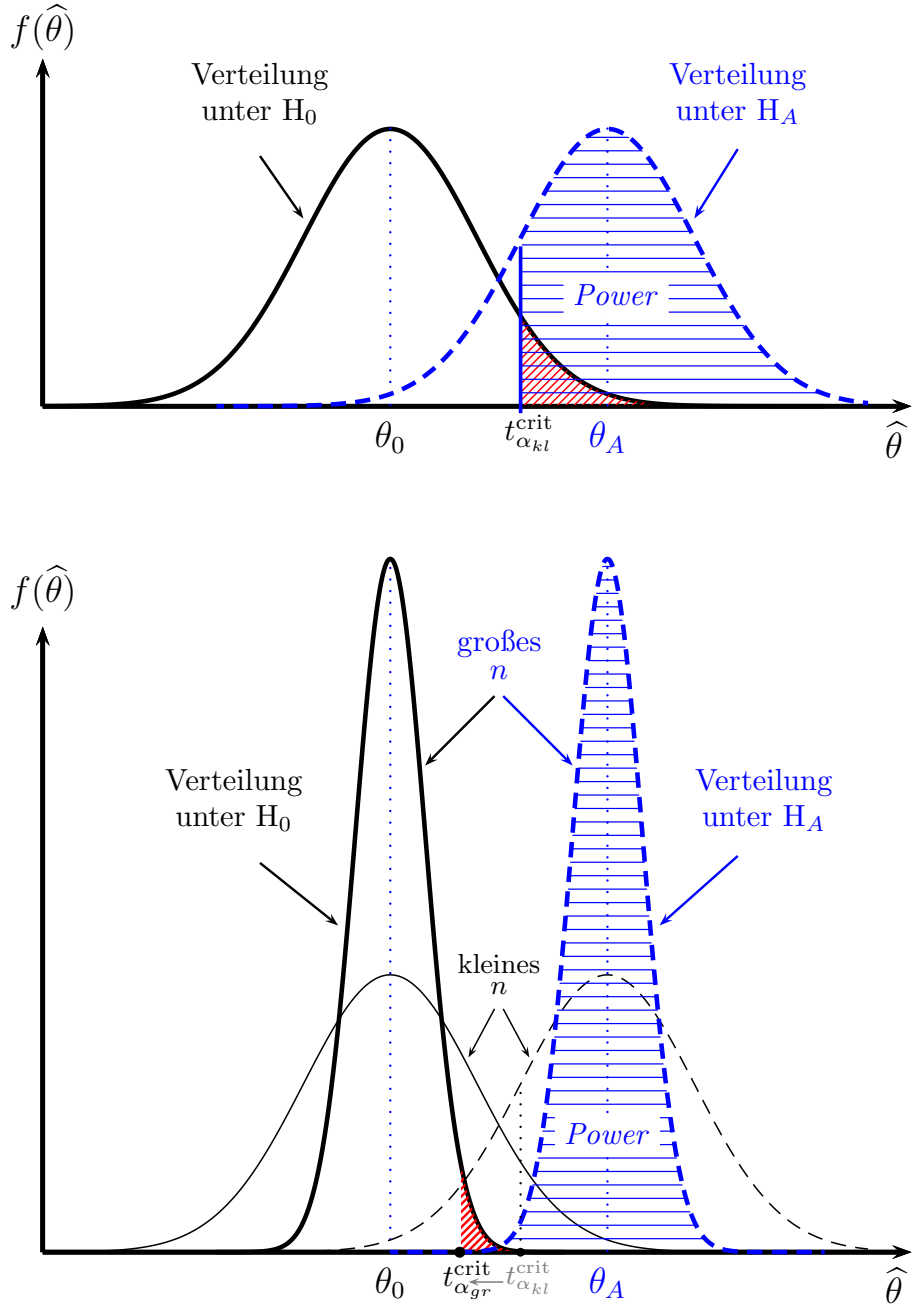


Abbildung 5.12: Die 'Power' eines Tests nimmt ceteris paribus mit der Stichprobengröße n zu.

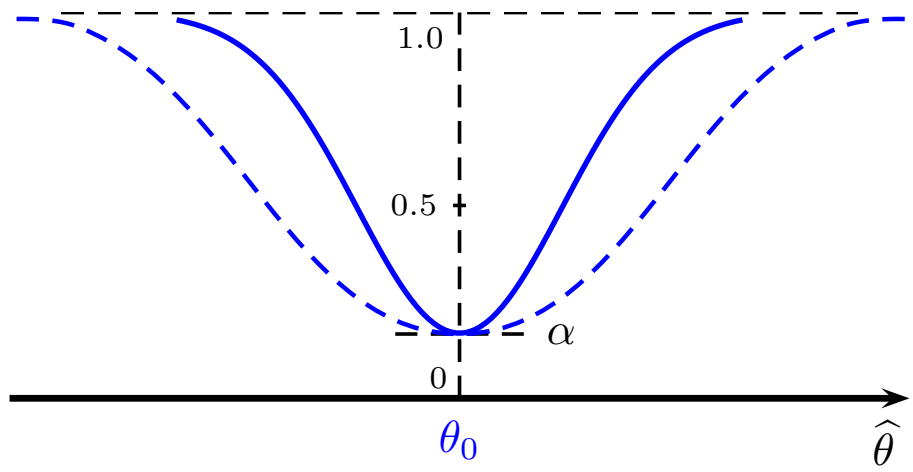


Abbildung 5.13: Teststärkefunktionen ('Power Functions') für zwei Tests (zweiseitig). Die durchgezogene Teststärkefunktion hat eine größere 'Power' als die strichlierte.

Hingegen wurden Neyman-Pearson Tests mit einem vorgegebenen Signifikanzniveau α als Entscheidungsregel zum weitgehend akzeptierten Standard.

Fisher konnte sich bis zu seinem Lebensende nicht damit anfreunden, und sah zwischen den beiden Methoden unüberbrückbare Gegensätze. Er schreibt

> "Moreover, it is a fallacy, so well known as to be a standard example, to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened. In an acceptance procedure, on the other hand, acceptance is irreversible, whether the evidence for it was strong or weak. It is the result of applying mechanically rules laid down in advance; no thought is given to the particular case, and the tester's state of mind, or his capacity for learning, is inoperative. By contrast, the conclusions drawn by a scientific worker from a test of significance are provisional, and involve an intelligent attempt to understand the experimental situation." (Fisher, 1955, 73f)

Historisch gesehen hat sich die Vorgangsweise von Neyman-Pearson weitgehend durchgesetzt, und Fishers p-Wert wird heute vielfach für Tests im Sinne von Neyman-Pearson verwendet, indem er mit dem vorgegebenen Signifikanzniveau α verglichen wird.

Aber es gibt auch kritische Stimmen, Spanos (1999, 720ff) argumentiert zum Beispiel, dass sich die Methode von Neyman-Pearson v.a. für klar strukturierte Probleme eignet, während die Methode von Fisher besser geeignet sei für offenere und breitere Fragestellungen, wie z.B. Spezifikationstests. □

5.3 Konfidenzintervalle

Als Jerzy Neyman seine Idee mit den Konfidenzintervallen 1934 erstmals der Royal Statistical Society vorstellte stieß er nicht auf ungeteilte Begeisterung. Nach dem üblichen Dank brachte der *Chair* der Sitzung seine Skepsis zum Ausdruck:

> "I am not certain whether to ask for an explanation or to cast a doubt. [...] I am referring to Dr. Neyman's confidence limits. I am not at all sure that the 'confidence' is not a 'confidence trick'. [...] The statement of the theory is not convincing, and until I am convinced I am doubtful of its validity." (Salsburg, 2002, 121)

Die Zweifel des *Chairs* sind natürlich längst ausgeräumt, aber vielen, die erstmals mit Konfidenzintervallen konfrontiert werden, geht es ähnlich.

Dabei beruhen Konfidenzintervalle auf den gleichen Grundlagen wie Hypothesentests, und wir werden später sehen, dass es sogar eine Dualität zwischen Konfidenzintervallen und Hypothesentests gibt.

Allerdings verfolgen Konfidenzintervalle in der Regel einen etwas unterschiedlichen Zweck. Während Hypothesentests in erster Linie zur Überprüfung dienen, inwieweit eine vorgegebene Vermutung mit den beobachteten Daten kompatibel ist, dienen Konfidenzintervalle in erster Linie dazu, einen Eindruck von der Messgenauigkeit einer Schätzung zu vermitteln.

OLS Schätzfunktionen, mit denen wir uns bisher beschäftigt haben, sind Beispiele für sogenannte *Punktschätzer*, sie ordnen jeder möglichen Stichprobe einen Punkt auf der Zahlengerade zu.

Wenn die Gauss-Markov Annahmen erfüllt sind kann man zeigen, dass diese OLS Schätzer *effizient* sind, d.h. sie bieten in der Klasse der linearen erwartungstreuen Schätzfunktionen die 'größtmögliche Genauigkeit' (sie sind 'BLUE').

Eine solche 'größtmögliche Genauigkeit' ist zwar beruhigend, aber 'größtmöglich' sagt wenig darüber aus, wie präzise die Schätzung tatsächlich ist. Wenn wir auf Grundlage unserer Schätzungen Entscheidungen treffen müssen wollen wir vermutlich wissen, inwieweit wir den Schätzungen 'vertrauen' können, bzw. *wie* genau sie sind.

Die bisher verwendeten Punktschätzer, wie z.B. $\widehat{\beta}_h$ oder $\hat{\sigma}^2$, können uns diese Information nicht liefern, sie geben keine Hinweise darüber, wie 'nahe' sie beim wahren Wert liegen, bzw. wie präzise sie sind, und täuschen damit möglicherweise eine Scheingenauigkeit vor.

In diesem Abschnitt zeigen wir, dass man mit Hilfe der Punktschätzer und deren Standardfehler ein Intervall berechnen kann, das zusätzlich den Stichprobenfehler (*'sampling error'*) berücksichtigt, und uns damit etwas über die 'Vertrauenswürdigkeit' einer Schätzung verrät.

5.3.1 Das Grundprinzip von Konfidenzintervallen

Beginnen wir mit einer einfachen normalverteilten Zufallsvariable

$$X \sim \mathrm{N}(\mu, \sigma_X^2)$$

Für die standardisierte (d.h. z transformierte) Zufallsvariable gilt folgende Wahrscheinlichkeitsaussage

$$\Pr\left(-1.96 \leq \frac{X-\mu}{\sigma_X} \leq +1.96\right) = 0.95 \tag{5.6}$$

Dies sagt uns, dass wir bei wiederholter Durchführung des Zufallsexperiments erwarten können, dass 95% der Realisationen von $Z = (X - \mu)/\sigma_X$ in das Intervall $[-1.96, +.196]$ fallen werden.

Wir können diese Wahrscheinlichkeitsaussage einfach umschreiben zu

$$\Pr\left[\mu - 1.96\,\sigma_X \leq X \leq \mu + 1.96\,\sigma_X\right] = 0.95$$

Man beachte, dass hier die Intervallgrenzen $\mu \pm 1.96\,\sigma_X$ feste Zahlen sind, da μ und σ_X Parameter der Grundgesamtheit sind.

Am einfachsten versteht man ein Konfidenzintervall, indem wir Gleichung (5.6) so umschreiben, dass der unbekannte (Lage-) Parameter μ im Zentrum und die Zufallsvariable X in der Unter- und Obergrenze steht, nämlich[12]

$$\Pr\left[X - 1.96\,\sigma_X \leq \mu \leq X + 1.96\,\sigma_X\right] = 0.95$$

Da nun die die beiden Intervallgrenzen Zufallsvariablen sind ändert sich die Wahrscheinlichkeitsaussage: bei unendlich vielen Wiederholungen des zugrunde liegenden Zufallsexperiments können wir damit rechnen, dass in 95% der Fälle das 'wahre' μ zwischen den beiden stochastischen Intervallgrenzen liegt.

Oder in anderen Worten, wenn die Annahme $X \sim \mathrm{N}(\mu, \sigma_X^2)$ erfüllt ist wird das stochastische Intervall in 95% der Wiederholungen das 'wahre' μ überdecken. Dies gilt für eine einzelne Zufallsvariable X.

Wir interessieren uns aber in den allermeisten Fällen nicht für einzelne Zufallsvariablen, sondern für *Schätzfunktionen*. Wenn diese, wie OLS Schätzer, die gewichtete Summe von Zufallsvariablen sind, können diese Ideen leicht übertragen werden.

5.3.2 Konfidenzintervalle für einzelne Regressionskoeffizienten

Die Berechnung von Konfidenzintervallen für Regressionskoeffizienten funktioniert völlig analog.

Sei die PRF (*population regression function*)

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_h x_{ih} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

mit $i = 1, \ldots, n$, $h \in \{1, \ldots, k\}$, und $\widehat{\beta}_h$ eine erwartungstreue Schätzfunktion für den Parameter β_h, dann gilt für den standardisierten Koeffizienten

$$\frac{\widehat{\beta}_h - \beta_h}{\hat{\sigma}_{\widehat{\beta}_h}} \sim \mathrm{t}_{n-k}$$

[12] Ausgehend von Gleichung (5.6)

$$\begin{aligned}
\Pr\left(-1.96\sigma_X \leq X - \mu \leq +1.96\sigma_X\right) &= 0.95 \\
\Pr\left(-X - 1.96\sigma_X \leq -\mu \leq -X + 1.96\sigma_X\right) &= 0.95 \qquad / \times (-1) \\
\Pr\left(X + 1.96\sigma_X \geq \mu \geq X - 1.96\sigma_X\right) &= 0.95 \\
\Pr\left(X - 1.96\sigma_X \leq \mu \leq X + 1.96\sigma_X\right) &= 0.95
\end{aligned}$$

mit

$$\Pr\left[-t^{\text{crit}}_{\alpha/2,df} \leq \frac{\widehat{\beta}_h - \beta_h}{\hat{\sigma}_{\widehat{\beta}_h}} \leq +t^{\text{crit}}_{\alpha/2,df}\right] = 0.95$$

Man beachte, dass die mit dem *geschätzten* Standardfehler $\hat{\sigma}_{\widehat{\beta}_h}$ standardisierten Koeffizienten t-verteilt sind, siehe Gosset (1908).

Wir können diese Wahrscheinlichkeitsaussage wieder umformen zu

$$\Pr\left[\widehat{\beta}_h - t^{\text{crit}}_{\alpha/2,df} \cdot \hat{\sigma}_{\widehat{\beta}_h} \;\leq\; \beta_h \;\leq\; \widehat{\beta}_h + t^{\text{crit}}_{\alpha/2,df} \cdot \hat{\sigma}_{\widehat{\beta}_h}\right] = 1 - \alpha \tag{5.7}$$

Man beachte, dass hier der unbekannte Parameter β_h (eine einfache Zahl) im Zentrum steht, und die Zufallsvariablen $\widehat{\beta}_h$ und $\hat{\sigma}_{\widehat{\beta}_h}$ links und rechts vom Ungleichheitszeichen stehen.

Diese beiden *stochastischen 'Grenzen'* bilden ein $(1-\alpha) \times 100$ Prozent *Konfidenzintervall* für den Regressionskoeffizienten $\widehat{\beta}_h$.

$$\left[\widehat{\beta}_h - t^{\text{crit}}_{\alpha/2,df}\, \hat{\sigma}_{\widehat{\beta}_h},\quad \widehat{\beta}_h + t^{\text{crit}}_{\alpha/2,df}\, \hat{\sigma}_{\widehat{\beta}_h}\right]$$

Dies ist wieder ein *theoretisches* Konfidenzintervall, welches allgemein für die Zufallsvariablen $\widehat{\beta}_h$ und $\hat{\sigma}_{\widehat{\beta}_h}$ definiert ist.

Das entsprechende *empirische Konfidenzintervall* erhalten wir wieder, indem wir die Realisationen der Stichprobe b_h und $s_{\widehat{\beta}_h}$ einsetzen

$$\left[b_h - t^{\text{crit}}_{\alpha/2,df}\, s_{\widehat{\beta}_h},\quad b_h + t^{\text{crit}}_{\alpha/2,df}\, s_{\widehat{\beta}_h}\right]$$

oder kürzer

$$b_h \pm t^{\text{crit}}_{\alpha/2,df}\, s_{\widehat{\beta}_h}$$

wobei $s_{\widehat{\beta}_h}$ der geschätzte Standardfehler des Koeffizienten h ist.

Solche Konfidenzintervalle können natürlich nicht nur für Regressionskoeffizienten berechnet werden, sondern für beliebige Schätzfunktionen, z.B. für Mittelwerte μ, oder – wie wir später zeigen werden – auch für die Standardfehler von Regressionskoeffizienten.

5.3.3 Interpretation von Konfidenzintervallen

Während die Berechnung von Konfidenzintervallen einfach ist, hatte mit deren Interpretation nicht nur der eingangs erwähnte *Chair* Probleme, sondern auch unzählige Studierende seither.

Dabei ist es wieder relativ einfach, wenn wir uns die theoretische Stichprobenkennwertverteilung und deren Realisation vor das geistige Auge rufen, vgl. Abbildung 5.14.

Wir beobachten eine einzelne Realisation aus der Stichprobenkennwertverteilung, und darum herum basteln wir ein *empirisches* Konfidenzintervall.

Da sowohl die Realisation b_h als auch der unbekannte Parameter β_h einfache Zahlen sind können wir darüber natürlich keine Zufallsaussage treffen.

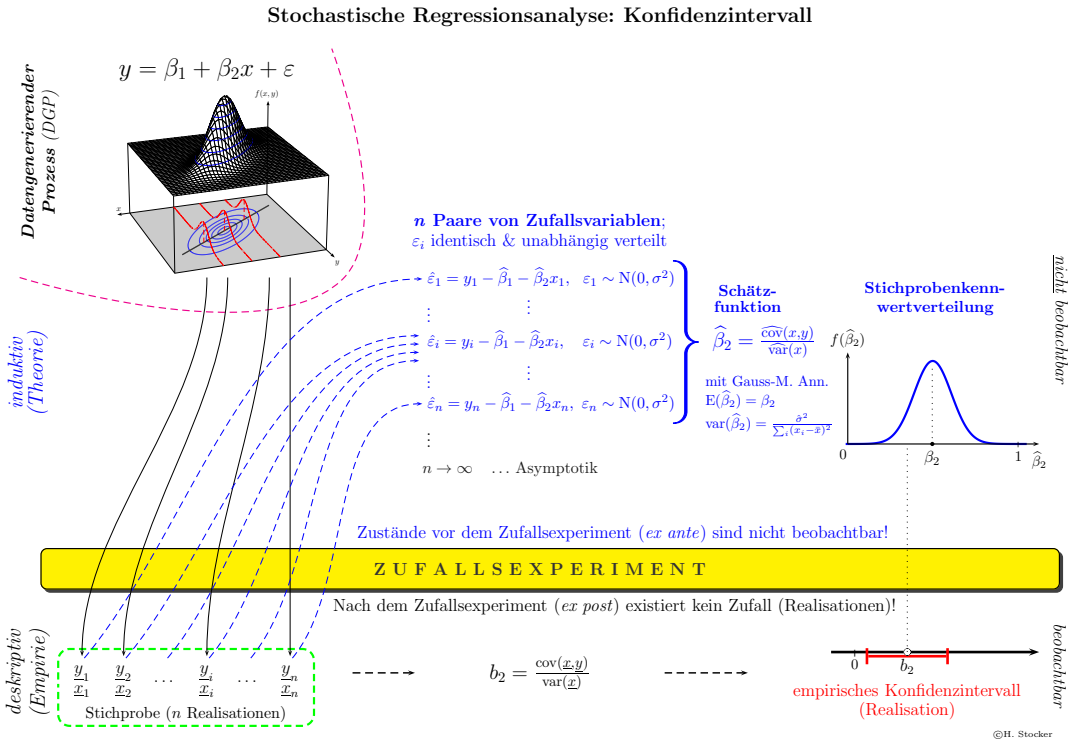


Abbildung 5.14: Konfidenzintervall für einen einzelnen Regressionskoeffizienten

Die häufig gehörte (und falsche!) Aussage: "ein empirisches Konfidenzintervall gibt den Bereich an, in dem der wahre Parameter mit 95% Wahrscheinlichkeit liegt" ist (in einer frequentistischen Denkweise) ebenso unsinnig wie die Behauptung, "die Zahl 10 liegt mit einer Wahrscheinlichkeit von 95% zwischen den Zahlen 9 und 11". Der Parameter liegt in dem empirischen Intervall oder er liegt nicht darin, da existiert keine Wahrscheinlichkeit.

Für die *stochastische Interpretation* müssen wir auf die zugrunde liegenden Zufallsvariablen Bezug nehmen, d.h. auf das *theoretische Konfidenzintervall*, welches durch die Wahrscheinlichkeitsaussage (5.7) definiert ist.

Dies kann man sich mit Hilfe der Vorstellung eines (*'repeated sampling'*) veranschaulichen: Wenn wir sehr viele Stichproben ziehen würden und für jede dieser Stichproben ein empirisches Konfidenzintervall berechnen würden, so könnten wir darauf vertrauen, dass z.B. 95% dieser Konfidenzinteravlle den wahren Parameter β_h überdecken würden.

Diese Idee ist in Abbildung 5.15 dargestellt, die oben die theoretische Stichprobenkennwertverteilung zeigt, und unten sieben Realisationen von empirischen Konfidenzintervallen, die auf sieben unterschiedlichen Stichproben beruhen.

Tatsächlich haben wir keine Möglichkeit zu erkennen, ob das aus unserer Stichprobe berechnete empirische Konfidenzintervall den wahren Parameter β_h überdeckt oder nicht.

Hätten wir zufällig die Stichprobe gezogen, die in Abbildung 5.15 zu Konfidenzintervall 5 führt, hätten wir Pech gehabt, denn dieses Konfidenzintervall überdeckt den wahren Parameter nicht, aber wir hätten keine Möglichkeit dies zu erkennen!

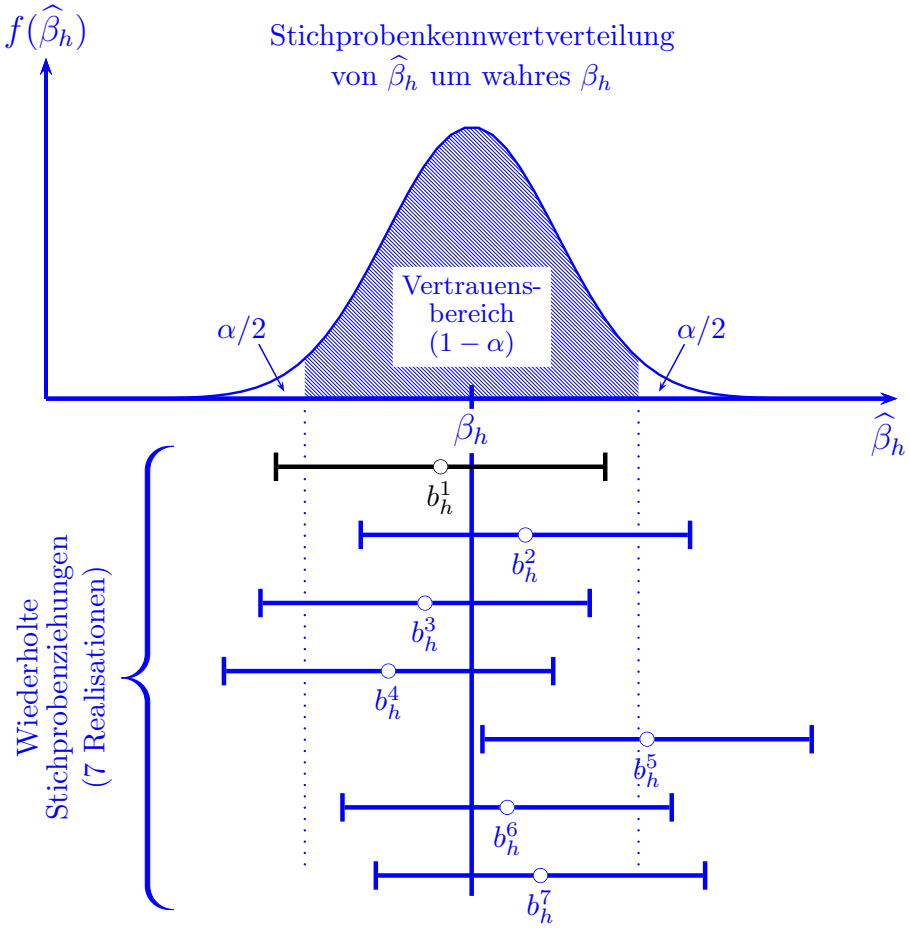


Abbildung 5.15: Parameter β_h und Konfidenzintervalle für 7 verschiedene Stichproben. In der praktischen Arbeit beobachten wir nur eine einzige Stichprobe und deshalb nur eine einzige Realisation eines empirischen Konfidenzintervalls. Die Theorie der Stichprobenkennwertverteilungen sagt uns, dass bei wiederholten Stichprobenziehungen $(1-\alpha)100\%$ aller hypothetischen Konfidenzintervalle das wahre β_h überdecken.
Obwohl wir nie sicher sein können, ob unser beobachtetes Konfidenzintervall den wahren Wert β_h überdeckt, gibt uns die *Breite des Konfidenzintervalls* wichtige Information über die Genauigkeit der Schätzung! *Quelle:* nach Bleymüller (2012).

Wenn aber alle erforderlichen Annahmen erfüllt sind können wir der Wahrscheinlichkeitsaussage mit den stochastischen Intervallgrenzen

$$\Pr\left[\widehat{\mu} - t^{\text{crit}}_{\alpha/2,df} \cdot \hat{\sigma}_{\widehat{\mu}} \;\leq\; \mu \;\leq\; \widehat{\mu} + t^{\text{crit}}_{\alpha/2,df} \cdot \hat{\sigma}_{\widehat{\mu}}\right] = 1 - \alpha$$

vertrauen und im Sinne des *repeated sampling* interpretieren (für $\alpha = 0.05$):

> Wenn wir – hypothetisch – unendlich viele Stichproben ziehen würden, und für jede dieser Stichproben ein 95% Konfidenzintervall berechnen würden, so könnten wir darauf vertrauen, dass 95% dieser Konfidenzintervalle den wahren Parameter β_h überdecken werden.[13]

Wir können zwar nicht wissen, ob unser berechnetes empirisches Konfidenzintervall den wahren Parameter β_h überdeckt, aber wir können darauf vertrauen, dass 95% (oder allgemeiner $(1-\alpha) \times 100\%$) aller hypothetisch möglichen Konfidenzintervalle den wahren Wert überdecken werden!

Auf den ersten Moment wirkt diese Aussage wenig hilfreich, wir haben nicht einmal zwei Stichproben, geschweige denn unendlich viele.

Was sollen wir also mit diesem Konfidenzintervall anfangen? Viele Anfänger sind verunsichert und frustriert, dass die Realisation des Konfidenzintervalls so wenig über den wahren Parameter verrät, dass sie darüber den wesentlichen Punkt übersehen, *die Breite des Konfidenzintervalls*!

Durch die Anwendung dieser Methode wird die Breite des Konfidenzintervalls derart bestimmt, dass bei wiederholten Stichprobenziehungen genau $(1-\alpha) \times 100\%$ aller hypothetisch möglichen Konfidenzintervalle den wahren Parameter β_h überdecken werden. Damit gibt uns die Breite des Konfidenzintervalls unmittelbar Auskunft über die Vertrauenswürdigkeit der Schätzung. Umso enger ein Konfidenzintervall ist, umso mehr werden wir der Schätzung vertrauen.

Aber benötigen wir dafür wirklich ein so aufwändiges Verfahren? Ja! Ein großer Vorteil von Konfidenzintervallen besteht darin, dass sie nach einer festen und nachvollziehbaren Regel berechnet werden, und dass alle Forscherinnen, die sich an dieses Regelwerk halten, zu vergleichbaren Ergebnissen kommen. Sie gehören deshalb längst zum unverzichtbaren Instrumentarium aller empirisch arbeitenden Wissenschaftler.

Wovon hängt die Breite des Konfidenzintervalls ab?

Konfidenzintervalle erlauben die Beurteilung der Genauigkeit (Präzision) einer Schätzung. Sie werden nach einer transparenten und nachvollziehbaren Regel berechnet, und ermöglichen dadurch auch Vergleiche von Schätzungen. Ceteris paribus wird man in der Regel eine Schätzung mit einem engeren Konfidenzintervall bevorzugen.

[13] Alternativ könnten wir auch sagen, dass das aus der *nächsten* – noch nicht gezogenen – Stichprobe berechnete 95% Konfidenzintervall mit 95% Wahrscheinlichkeit den wahren Parameter β_h überdecken wird, aber diese Interpretation ist in diesem Zusammenhang weniger hilfreich.

Um zu erkennen, wovon die Breite des Konfidenzintervalls abhängt, müssen wir uns nur an die Determinanten des empirischen Konfidenzintervalls erinnern. Für den Steigungskoeffizienten einer Regression ist das *empirische Konfidenzintervall*

$$\left[b_h - t^{\text{crit}}_{\alpha/2,df}\, s_{\widehat{\beta}_h},\quad b_h + t^{\text{crit}}_{\alpha/2,df}\, s_{\widehat{\beta}_h}\right] \qquad \text{(bzw. kürzer } b_h \pm t^{\text{crit}}_{\alpha/2,df}\, s_{\widehat{\beta}_h}\text{)}$$

wobei $s_{\widehat{\beta}_h}$ der geschätzte Standardfehler des Koeffizienten h ist.
Für eine bivariate Regression ist

$$b_2 = \frac{\text{cov}(x,y)}{\text{var}(x)}, \quad s_{\widehat{\beta}_2} = \sqrt{\frac{\frac{\sum_i e_i^2}{n-2}}{\sum_i (x_i - \bar{x})^2}}$$

Daraus erkennen wir, dass die Breite des Konfidenzintervalls neben dem frei gewählten Konfidenzniveau $1-\alpha$ (über $\text{t}^{\text{crit}}_{\alpha/2,df}$) von allen Faktoren abhängt, die den Standardfehler der Koeffizienten $\hat{\sigma}_{\widehat{\beta}_h}$ (bzw. deren Realisation $s_{\widehat{\beta}_2}$)beeinflussen
Das Konfidenzintervall ist ceteris paribus umso *enger*, ...

- je größer die Stichprobe ist, da dies zu einem kleineren Standardfehler führt. Darüber hinaus führt dies über die Freiheitsgrade auch zu einem kleineren kritischen Wert $t^{\text{crit}}_{\alpha/2,df}$.
- umso kleiner die Varianz der Störterme σ^2 ist,
- umso größer die Streuung der x (d.h. $\sum_i (x_i - \bar{x})^2$) ist,
- umso weniger die Regressoren untereinander korreliert sind,
- je kleiner das frei gewählte Konfidenzniveau $1-\alpha$ ist (d.h. ein 90% Konfidenzintervall ist ceteris paribus enger als ein 99% Konfidenzintervall[14]). Dieser Effekt wirkt sich über den kritischen Wert $t^{\text{crit}}_{\alpha/2,df}$ auf die Breite aus.

Konfidenzintervalle können natürlich nicht nur für Mittelwerte berechnet werden, sondern auch für Differenzen zwischen Mittelwerten, für Anteile, und natürlich auch für Regressionskoeffizienten.

Beispiel Im Kapitel zur deskriptiven Regressionsanalyse hatten wir ein Beispiel mit den Stundenlöhnen (StdL) und dem Geschlecht ($m = 1$ für Männer und Null sonst) von 12 Personen. Nun interpretieren wir diese Zahlen als Stichprobe aus einer unbekannten Grundgesamtheit. Das Ergebnis war

$$\begin{array}{rcccc} \text{StdL} & = & 12.5 & + & 2.5\ \text{m} \\ & & (0.747)^{***} & & (1.057)^{**} \end{array}$$

$$R^2 = 0.359, \quad s = 1.83, \quad n = 12$$

(OLS Standardfehler in Klammern)

[14] Konfidenzintervalle, die nur in 90% der Fälle den wahren Wert enthalten müssen, können ceteris paribus schmaler sein als Konfidenzintervalle, die in 99% der Fälle den wahren Wert enthalten sollen.

Wir erinnern uns, dass bei einer Regression auf eine Dummy Variable das Interzept den Wert von y der Referenzkategorie angibt, und der Koeffizient der Dummy Variable den Unterschied zur Referenzkategorie misst. In diesem Beispiel ist also der durchschnittliche Stundenlohn von Frauen 12.5 Euro, und Männer verdienen um 2.5 Euro *mehr*.

Ein empirisches 95% Konfidenzintervall für die Lohndifferenz erhalten wir mit dem kritischen Wert der t-Verteilung für $\alpha/2 = 0.025$ und $n - k = 12 - 2 = 10$ Freiheitsgraden $\mathrm{t}^{\mathrm{crit}}_{0.025,10} = 2.228$ (siehe Tabelle 5.1, Seite 20)

$$2.5 \pm 2.228 \cdot 1.057 \qquad \text{oder} \qquad [0.145,\ 4.855]$$

Mit R können können sie mit der Funktion `confint()` berechnet werden[15]

```
d <- read.csv2(url("http://www.uibk.ac.at/econometrics/data/stdl_bsp1.csv"))
eq <- lm(StdL ~ m, data = d)
confint(eq)

#                  2.5 %    97.5 %
# (Intercept) 10.8350966 14.164903
# m            0.1454711  4.854529
```

Beispiel 2: In einem früheren Beispiel haben wir den Preis von Gebrauchtautos auf Alter und Kilometerstand regressiert.

$$\begin{array}{rcccccc} \widehat{\text{Preis}} & = & 22649.88 & - & 1896.26\ \text{Alter} & - & 0.031\ \text{km} \\ & & (411.87) & & (235.215) & & (0.008) \end{array}$$

$$R^2 = 0.907, \qquad n = 40$$

(Standardfehler in Klammern)

Die Koeffizienten weisen das erwartete Vorzeichen auf und haben die übliche Interpretation, z.B. sinkt der erwartete (gefittete) Preis bei konstanter km-Zahl jedes Jahr um 1896.26 Euro.

Dies ist eine Punktschätzung, die uns nichts über die Genauigkeit der Schätzung verrät. Um ein 95% Konfidenzintervall zu berechnen benötigen wir den kritischen t-Wert, und um diesen nachschlagen zu können die Freiheitsgrade.

Wir haben insgesamt 40 Beobachtungen ($n = 40$) und drei geschätzte Koeffizienten (b_1, b_2 und b_3), also $n-k = 40-3 = 37$ Freiheitsgrade. Tabelle 5.1 (Seite 20) gibt uns nur die die kritischen Werte für 30 und 40 Freiheitsgrade, der kritische Wert für 40 Freiheitsgrade wäre z.B. $t^{\mathrm{crit}}_{0.025,40} = 2.021$. Wem dies zu ungenau ist kann mit Hilfe der Quantilfunktion eines geeigneten Programmes den genauen Wert $t^{\mathrm{crit}}_{0.025,37} = 2.0262$ berechnen.[16]

Die Realisation des 95% Konfidenzintervalls für das Alter ist also

$$[-1896.26 - 2.0262 * 235.215;\ -1896.26 + 2.0262 * 235.215]$$

[15] Stata gibt die Konfidenzintervalle standardmäßig aus.

[16] Die Funktionen sind für R: `qt(p = 0.975, df = 37)` und für Stata: `invttail(37,0.025)` (siehe `help density_functions`), und für EViews: `@qtdist(0.975,37)`.

bzw. $[-2372.854;\ -1419.674]$.

Wenn wir ein 99% Konfidenzintervall berechnen wollen benötigen wir nur den entsprechenden kritische Wert $t^{\text{crit}}_{0.005,37} = 2.715409$. Dieser ist größer als für ein 95% Konfidenzintervall, deshalb ist das 99% Konfidenzintervall breiter $[-2534.968;\ -1257.56]$.

In der Praxis werden wir die Konfidenzintervalle natürlich von einem entsprechenden Programm berechnen lassen, z.B. von R:

```
df <- read.csv2(url("http://www.hsto.info/econometrics/data/auto40.csv"))
eq1 <- lm(Preis ~ Alter + km, data = df)

confint(eq1, level = 0.99)
#                    0.5 %        99.5 %
# (Intercept)  2.153149e+04  2.376828e+04
# Alter       -2.534968e+03 -1.257560e+03
# km          -5.236829e-02 -9.652361e-03
```

(man beachte, dass `e+04` eine Kurzschreibweise für 10^4 ist, d.h. `2.1531e+04`= $2.153149 \times 10^4 = 2.153149 \times 10000 = 21531.49$)

Konfidenzellipsen für zwei Regressionskoeffizienten

Die Idee von Konfidenzintervallen kann auch auf zwei Regressionsoeffizienten verallgemeinert werden, Abbildung 5.16 zeigt eine solche für die beiden Koeffizienten von Alter und km des vorhergehenden Beispiels. Die Interpretation ist völlig analog zu früher, bei wiederholten Stichprobenziehungen würden wir erwarten, dass $(1-\alpha)*100$ Prozent der daraus berechneten Konfidenzellipsen das wahre Koeffizientenpaar (β_2, β_3) überdecken würden.

5.3.4 Ein Konfidenzintervall für den *Standardfehler* eines Koeffizienten*

Natürlich sind auch die Schätzfunktionen für die Standardfehler der Koeffizienten Zufallsvariablen und haben als solche eine Stichprobenkennwertverteilung.

Wir haben in Gleichung (5.3) gesehen, dass

$$\frac{(n-k)\hat{\sigma}^2_{\widehat{\beta}_h}}{\sigma^2_{\widehat{\beta}_h}} \sim \chi^2_{n-k}$$

Dies können wir nützen, um ein 95% Konfidenzintervall für den *Standardfehler* von $\widehat{\beta}_2$, d.h. $\sigma^2_{\widehat{\beta}_2}$ zu konstruieren.

Die entsprechende Wahrscheinlichkeitsaussage ist

$$\Pr\left[\chi^{2\ \text{crit,lower}}_{\alpha/2,n-k} \ \leq\ \frac{(n-k)\hat{\sigma}^2_{\widehat{\beta}_2}}{\sigma^2_{\widehat{\beta}_2}}\ \leq\ \chi^{2\ \text{crit,upper}}_{\alpha/2,n-k}\right] = 1-\alpha$$

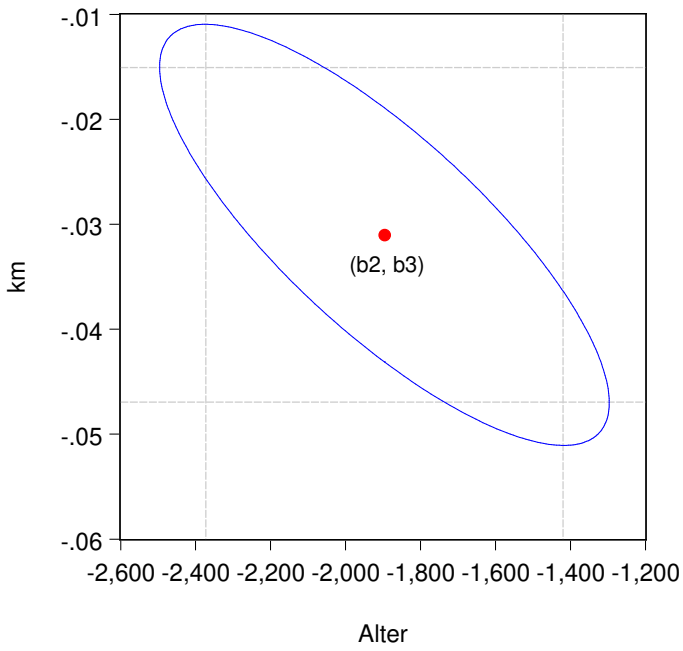


Abbildung 5.16: Empirische Konfidenzellipse für die zwei Koeffizienten von Alter und km des vorhergehenden Beispiels, Grau strichliert sind die individuellen Konfidenzintervalle der Koeffizienten eingezeichnet (erstellt mit EViews).

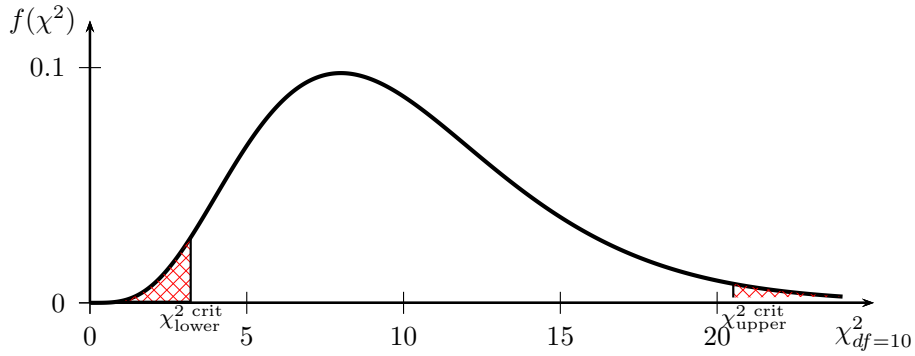


Abbildung 5.17: χ^2 Verteilung mit 10 Freiheitsgraden und den kritischen Werten für $\alpha/2 = 0.025$. Der untere kritische Wert ist $\chi^{2\text{ crit,lower}}_{0.025,10} = 3.25$, der obere kritische Wert ist $\chi^{2\text{ crit,upper}}_{0.975,10} = 20.48$.

Eine χ^2-Verteilung mit den kritischen Werten für $\alpha = 0.05$ und 10 Freiheitsgrade finden Sie in Abbildung 5.17. Da es sich um keine symmetrische Verteilung handelt unterscheiden sich unterer und oberer kritischer Wert (diese können z.B. in R mit der Quantilfunktion `qchisq(p, df)` ermittelt werden).

Dies können wir wieder umformen, sodass die stochastischen Werte in den Intervallgrenzen und der Parameter $\sigma^2_{\widehat{\beta}_2}$ im Zentrum steht

$$\Pr\left[\hat{\sigma}_{\widehat{\beta}_2}\sqrt{\frac{n-k}{\chi^{2\ \text{crit,upper}}_{\alpha/2,n-k}}} \;\leq\; \sigma_{\widehat{\beta}_2} \;\leq\; \hat{\sigma}_{\widehat{\beta}_2}\sqrt{\frac{n-k}{\chi^{2\ \text{crit,lower}}_{\alpha/2,n-k}}}\right] = 1-\alpha$$

Für das frühere Beispiel mit den Stundenlöhnen erhalten wir für $n-2 = 10$ Freiheitsgrade die kritischen Werte $\chi^{2\ \text{crit,lower}}_{0.025,10} = 3.25$ und $\chi^{2\ \text{crit,upper}}_{0.975,10} = 20.48$. Der Standardfehler des Koeffizienten von m ist $s_{\widehat{\beta}_2} = 1.057$.

Wenn ich mich nicht verrechnet habe ist das empirische Konfidenzintervall des Standardfehlers [0.738, 1.854].

5.3.5 Dualität zwischen Konfidenzintervallen und Hypothesentests

Konfidenzintervalle, Fishers p-Wert und die Entscheidungsregel nach Neyman-Pearson beruhen letztendlich alle auf der gleichen Information, deshalb ist es wenig verwunderlich, dass sie alle zu gleichen Schlussfolgerungen führen und in einem gewissen Sinne auch austauschbar sind.

Unter ziemlich allgemeinen Bedingungen können Konfidenzintervalle auch als die Menge aller Punkte interpretiert werden, die keine Ablehnung der entsprechenden Nullhypothese erlauben.

Wenn wir für eine Regression z.B. die Nullhyothese $\mathrm{H}_0\colon \beta_h = 0$ testen möchten, und das entsprechende empirische Konfidenzintervall den Nullpunkt überdeckt, dann folgt daraus, dass wir obige Nullhypothese *nicht* ablehnen können.

Etwas allgemeiner ist ein Konfidenzintervall einfach ein Intervall auf der Linie der reellen Zahlen, welches alle reellen Zahlen θ_0 enthält, für die die Nullhypothese $\theta = \theta_0$ durch einen entsprechenden Hypothesentest *nicht* abgelehnt werden kann (vgl. Davidson and MacKinnon, 2003, 177). Dies gilt analog für Konfidenzellipsen.

Allerdings sollte dies nicht darüber hinweg täuschen, dass diese Methoden für unterschiedliche Zielsetzungen entwickelt wurden. Konfidenzintervalle sollten uns in erster Linie Information über die Genauigkeit von Schätzern geben, während Hypothesentests uns erlauben zu überprüfen, inwieweit eine a priori vermutete Hypothese mit den beobachteten Stichprobendaten kompatibel ist.

5.4 Simultane Tests mehrerer linearer Hypothesen

Bisher haben wir nur einzelne Hypothesen mit Hilfe einer t-Statistik getestet. Nun werden wir zeigen, dass mit Hilfe von F-Statistiken auch mehrere lineare Hypothesen *simultan* getestet werden können.

Die prinzipielle Vorgangsweise bleibt völlig gleich, wir können auch hier wieder Konfidenzintervalle (bzw. Konfidenzellipsen) und p-Werte berechnen, oder Tests nach Neyman-Pearson durchführen. Auch bleibt alles, was über Typ I & II Fehler, die Power sowie zur Kritik an Hypothesentests gesagt wurde, uneingeschränkt gültig.

Wir beginnen mit der vermutlich bekanntesten dieser F-Statistiken, mit der F-total Statistik, die sich in so gut wie jedem Regressionsoutput findet.

5.4.1 ANOVA-Tafel und die F-total Statistik

Wir erinnern uns, dass wir für die Herleitung des Bestimmtheitsmaßes R^2 folgende Streuungs-Zerlegung vornahmen

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{\sum \hat{\varepsilon}_i^2}_{\text{SSR}}$$

wobei TSS für 'Total Sum of Squares', ESS für 'Explained Sum of Squares' und SSR für 'Sum of Squared Residuals'[17] steht.

Dies führt zu folgender **ANOVA-Tafel** (*'ANalysis Of VAriance Table'*)

	Sum of Squares	df	Mean Square
Regression	ESS	$k-1$	$\text{ESS}/(k-1)$
Residuen	SSR	$n-k$	$\text{SSR}/(n-k)$
Total	TSS	$n-1$	

wobei df für '*degrees of freedom*' (Freiheitsgrade) steht.

Mit Hilfe der ANOVA-Tafel, die von den vielen Statistik-Programmen ausgegeben wird, kann man eine F-Statistik für die Nullhypothese

$$\text{H}_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

konstruieren; das heißt für die Nullhypothese, dass alle Koeffizienten *mit Ausnahme des Interzepts* β_1 simultan gleich Null sind, oder in anderen Worten, dass alle $k-1$ Steigungskoeffizienten simultan gleich Null sind. Eine andere Möglichkeit diese Nullhypothese zu interpretieren ist

$$\text{H}_0 : \quad \text{E}(y|x_2, \ldots, x_k) = \text{E}(y)$$

wenn der bedingte Erwartungswert von y gleich dem unbedingten Erwartungswert ist leisten die erklärenden Variablen auch gemeinsam keinen Erklärungsbeitrag. In diesem Falle erklärt eine Regression auf die Regressionskonstante $y = \beta_1 + \varepsilon$ die Daten gleich gut wie die 'lange' Regression $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$.

Die Alternativhypothese ist

$$\text{H}_A : \text{ mindestens einer der Steigungskoeffizienten ist ungleich Null}$$

[17] Manchmal wird dafür auch RSS für *'Residual Sum Squared'* geschrieben.

Wenn diese Nullhypothese wahr ist, sollten alle x Variablen auch gemeinsam keinen Erklärungsbeitrag für y leisten.

Die Nullhypothese, dass alle Regressoren auch gemeinsam keinen Erklärungsbeitrag leisten können, kann mit der folgenden Teststatistik getestet werden

$$\hat{F}\text{-total Statistik} = \frac{\frac{\sum_i (\hat{y}_i - \bar{y})^2}{k-1}}{\frac{\sum_i \hat{\varepsilon}_i^2}{n-k}} = \frac{\text{ESS}/(k-1)}{\text{SSR}/(n-k)} \overset{\text{H}_0}{\sim} \text{F}_{(k-1,n-k)}$$

Diese Teststatistik ist F-verteilt mit $k-1$ Zähler- und $n-k$ Nennerfreiheitsgraden. Dabei ist $k-1$ die Anzahl der zu testenden Koeffizienten (d.h. alle Koeffizienten außer Interzept) und $n-k$ sind die Residuen-Freiheitsgrade.

*Hinweis:** Wenn die Störterme der Grundgesamtheit normalverteilt sind ist bei Zutreffen der Nullhypothese die Quadratsumme $\sum_i (\hat{y}_i - \bar{y})^2/\sigma^2$ bekanntlich χ^2-verteilt mit $k-1$ Freiheitsgraden, und $\sum_i \hat{\varepsilon}_i^2/\sigma^2$ unabhängig davon χ^2-verteilt mit $n-k$ Freiheitsgraden. Das Verhältnis zweier unabhängig χ^2 verteilten Zufallsvariablen, die beide durch die entsprechenden Freiheitsgrade dividiert wurden, ist F-verteilt (vgl. Statistischen Appendix). □

Wenn die durch alle erklärenden x Variablen gemeinsam erklärte Streuung ESS sehr klein ist im Verhältnis zur unerklärten Streuung SSR würde man einen sehr kleinen Wert der empirischen Statistik F^{emp} erwarten. Wenn umgekehrt ESS groß ist im Verhältnis zu SSR leisten alle x Variablen gemeinsam offensichtlich einen großen Erklärungsbeitrag.

Die Nullhypothese $\beta_2 = \beta_3 = \cdots = \beta_k = 0$ wird deshalb verworfen, wenn der empirische Wert dieser F-Statistik *größer* ist als der kritische Wert F^{crit} der F-Statistik, vgl. Abbildung 5.18. Fällt der empirische F^{emp}-Wert in den Verwerfungsbereich wird H_0 verworfen, anderenfalls akzeptiert.

Der zur F-total Statistik gehörende p-Wert ist wieder die Fläche unter der Verteilung rechts vom berechneten F^{emp}-Wert, und wird standardmäßig von allen statistischen Programmpaketen ausgewiesen.

Beispiel: STATA liefert z.B. für die Schätzung der Produktionsfunktion auf Seite 26 den in Tabelle 5.4 wiedergegebenen Regressionsoutput; im Kopf des Outputs wird die komplette ANOVA-Tafel wiedergegeben.

Dabei ist `Model SS` die ESS ('Explained Sum of Squares'), die `Residual SS` die SSR ('Sum of Squared Residuals') und `Total SS` die TSS ('Total Sum of Squares').

Für Produktionsfunktions-Beispiel folgt (siehe Tabelle 5.4)

$$F^{\text{emp}}\text{-total Statistik} = \frac{3.01454182/(3-1)}{0.187112359/(25-3)} = 177.2195$$

Andere Programme, wie z.B. R, geben die ESS und TSS nicht automatisch aus, sondern nur die F-Statistik mit dem dazugehörigen p-Wert. □

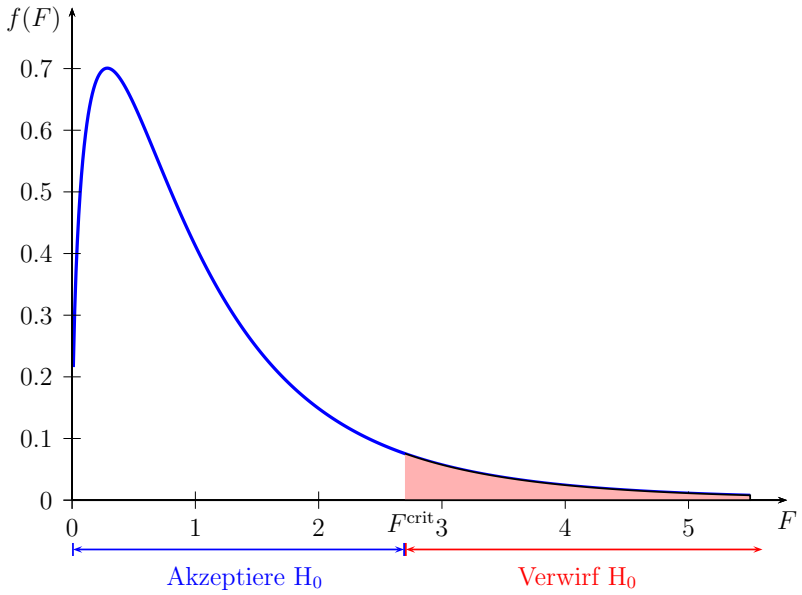


Abbildung 5.18: F-Test nach Neyman-Pearson (mit 3 Zähler- und 12 Nennerfreiheitsgraden

Tabelle 5.4: STATA Output

```
   Source |       SS       df       MS              Number of obs =      25
----------+------------------------------           F(  2,    22) =  177.22
    Model |  3.01454182     2  1.50727091           Prob > F      =  0.0000
 Residual |  .187112359    22  .008505107           R-squared     =  0.9416
----------+------------------------------           Adj R-squared =  0.9362
    Total |  3.20165418    24  .133402257           Root MSE      =  .09222

------------------------------------------------------------------------------
    log_Q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
    const |   2.481079   .1286189     19.29   0.000      2.21434    2.747819
    log_K |   .6401108   .0347308     18.43   0.000     .5680835    .7121381
    log_L |   .2573354   .0269591      9.55   0.000     .2014256    .3132451
------------------------------------------------------------------------------
```

Exkurs* Diese F-total Statistik kann alternativ auch mit Hilfe des Bestimmtheitsmaßes R^2 berechnet werden, denn unter Berücksichtigung von TSS = ESS + SSR und $R^2 = \text{ESS}/\text{TSS}$

$$\begin{aligned} F\text{-total Stat} &= \frac{\text{ESS}/(k-1)}{\text{SSR}/(n-k)} \\ &= \frac{n-k}{k-1}\,\frac{\text{ESS}}{\text{SSR}} \\ &= \frac{n-k}{k-1}\,\frac{\text{ESS}}{\text{TSS}-\text{ESS}} \\ &= \frac{n-k}{k-1}\left[\frac{\text{ESS}/\text{TSS}}{1-(\text{ESS}/\text{TSS})}\right] \\ &= \frac{n-k}{k-1}\left[\frac{R^2}{1-R^2}\right] \end{aligned}$$

Also

$$F\text{-total Statistik} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

□

Die Alternativhypothese, dass *zumindest für einen* der Regressoren der wahre Wert des Regressionskoeffizienten von Null verschieden ist, wird akzeptiert, wenn die Nullhypothese $\text{H}_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$ (d.h. dass alle Koeffizienten mit Ausnahme des Interzepts β_1 gleich Null sind) verworfen wird.

Man beachte, dass es sich dabei um einen simultanen Test von insgesamt $k-1$ Hypothesen handelt, wobei $k-1$ die Anzahl der Steigungskoeffizienten ist.

Dies ist also ein Test, ob alle Regressoren *gemeinsam* einen Beitrag zur 'Erklärung' von y leisten. Wenn der p-Wert dieser F-Statistik größer ist als das a priori festgelegte Signifikanzniveau sollte die Schätzung verworfen werden.

Dieser F-total Test wird häufig auch einfach F-Test genannt, zum Beispiel im Regressionsoutput fast jeder Statistiksoftware, was aber etwas verwirrend ist, da jede F-verteilte Teststatistik zu einem F-Test führt, und davon gibt es viele.

Achtung: Man beachte, dass es nicht reicht zu überprüfen, ob alle Koeffizienten individuell signifikant von Null verschieden sind. Es ist sehr gut möglich und passiert auch häufig, dass kein einziger Koeffizient signifikant von Null verschieden ist, obwohl alle Koeffizienten gemeinsam hoch signifikant von Null verschieden sind! Wie wir später sehen werden tritt dieser Fall bei *Multikollinearität* (d.h. wenn die erklärenden Variablen untereinander hoch korreliert sind) relativ häufig auf.

Konkret, die mit der F-Statistik getestete gemeinsame Nullhypothese $\text{H}_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$ darf *nicht* durch eine Reihe individueller t-Tests ersetzt werden!

Der Grund dafür liegt darin, dass die F-Statistik die mögliche Korrelation zwischen den OLS-Schätzern $\widehat{\beta}_2, \widehat{\beta}_3, \ldots, \widehat{\beta}_k$ berücksichtigt, während diese bei individuellen t-Tests unberücksichtigt bleibt.

Dies wird in Abbildung 5.19 veranschaulicht, die zwei bivariate Normalverteilungen zeigt, links ohne und rechts mit einer Korrelation zwischen den Zufallsvariablen.

$\rho = 0$: $\rho = -0.7$:

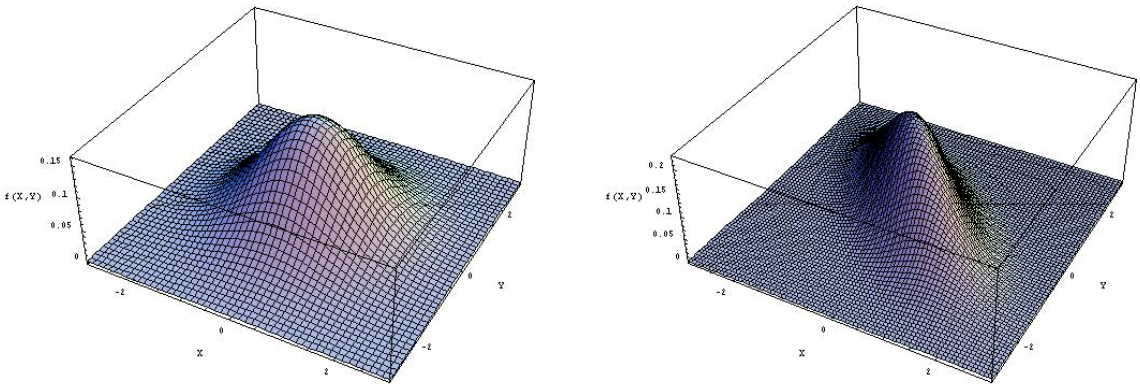


Abbildung 5.19: Bivariate Normalverteilungen ohne und mit Korrelation zwischen den Zufallsvariablen

Wenn die Variablen unkorreliert sind (linke Grafik) wird die gemeinsame Signifikanz durch einen *Signifikanzkreis* dargestellt, bei Korrelation zwischen den Variablen (rechte Grafik) durch eine Konfidenzellipse, die umso schmaler wird, je höher die Korrelation ist.

Wir werden später sehen, dass viele der üblichen Teststatistiken für den simultanen Test mehrerer Hypothesen unter den Standardannahmen F-verteilt sind. Sollten z.B. im Modell

$$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \hat{\varepsilon}_i$$

die beiden Koeffizienten $\widehat{\beta}_2$ und $\widehat{\beta}_3$ simultan getestet werden, so kann man die folgende Teststatistik verwenden:

$$\hat{F} = \frac{1}{2\hat{\sigma}^2}\left[S_{22}(\widehat{\beta}_2 - \beta_2)^2 + 2S_{23}(\widehat{\beta}_2 - \beta_2)(\widehat{\beta}_3 - \beta_3) + S_{33}(\widehat{\beta}_3 - \beta_3)^2\right] \overset{\mathrm{H}_0}{\sim} \mathrm{F}_{2,n-3}$$

mit

$$\begin{aligned} S_{22} &= \sum_i (x_{i2} - \bar{x}_{\cdot 2})^2 \\ S_{23} &= \sum_i (x_{i2} - \bar{x}_{\cdot 2})(x_{i3} - \bar{x}_{\cdot 3}) \\ S_{33} &= \sum_i (x_{i3} - \bar{x}_{\cdot 3})^2 \end{aligned}$$

Diese Teststatistik ist unter H_0 F-verteilt mit 2 Zähler- und $n-3$ Nennerfreiheitsgraden. Diese F-Statistik kann für die Konstruktion einer *Konfidenzregion* verwendet werden.

Wenn wir mit F^{crit} den kritischen F-Wert mit 2 Zähler- und $n-3$ Nennerfreiheitsgraden bezeichnen ist diese Konfidenzregion

$$\left[S_{22}(\widehat{\beta}_2 - \beta_2)^2 + 2S_{23}(\widehat{\beta}_2 - \beta_2)(\widehat{\beta}_3 - \beta_3) + S_{33}(\widehat{\beta}_3 - \beta_3)^2\right] \leq F^{\mathrm{crit}}(2\hat{\sigma}^2)$$

Dies definiert eine Ellipsengleichung, das heißt, bei dem simultanen Test zweier Koeffizienten erhält man anstelle eines Konfidenzintervalls eine *Konfidenzellipse*.

Sind die Koeffizienten $\widehat{\beta}_2$ und $\widehat{\beta}_3$ unkorreliert erhält man einen Kreis, und umso höher die Korrelation zwischen $\widehat{\beta}_2$ und $\widehat{\beta}_3$ ist, umso schmaler ist die Ellipse.

Werden mehr als zwei Hypothesen gleichzeitig getestet erhält man ein höherdimensionales Ellipsoid.

Beispiel: Das folgende Beispiel zeigt einen Fall, bei dem die F-Statistik die gemeinsame Nullhypothese $\beta_2 = \beta_3 = 0$ ablehnt, die individuellen t-Tests einzeln aber weder die Ablehnung von $\beta_2 = 0$ noch von $\beta_3 = 0$ erlauben (auf dem 5% Niveau).

$$\begin{array}{rcccccc} y & = & 7.888 & + & 0.2\ x_2 & + & 0.634\ x_3 \\ & & (5.028) & & (0.28) & & (0.344)^* \end{array}$$

F-Statistik (df: 2, 47): 34.95, p-Wert: 5.023e-10

$R^2 = 0.598, \qquad n = 50$

Die F-Statistik für die Nullhypothese $H_0: \beta_2 = 0$ *und* $\beta_3 = 0$ ist in diesem Beispiel 34.95 und kann damit mit einer Wahrscheinlichkeit kleiner als 0.0001 verworfen werden. In diesem Beispiel sind die Regressoren x_2 und x_3 hoch korreliert, d.h. es liegt Multikollinearität vor (der Korrelationskoeffizient zwischen x_2 und x_3 ist 0.95!), was auch zu einer Korrelation zwischen den geschätzten Koeffizienten führt.

Die geschätzte Varianz-Kovarianzmatrix der Koeffizienten ist

	b_1	b_2	b_3
b_1	25.27815	0.308261	−0.808990
b_2	0.308261	0.078356	−0.091832
b_3	−0.808990	−0.091832	0.118481

deshalb ist die Korrelation $\text{corr}(\widehat{\beta}_2, \widehat{\beta}_3) = -0.0918/\sqrt{0.0784 * 0.1185} = -0.953$.

Aufgrund dieser Korrelation ist die Konfidenzellipse in Abbildung 5.20 ziemlich schmal. Die Konfidenzintervalle für $\widehat{\beta}_2$ und $\widehat{\beta}_3$ einzeln sind in Abbildung 5.20 strichliert eingezeichnet.

Zur Interpretation von Konfidenzellipsen gilt analoges wie für Konfidenzintervalle, wenn sehr viele Stichproben gezogen würden könnten wir damit rechnen, dass $(1 - \alpha) * 100\%$ der resultierenden Konfidenzregionen die wahren Werte β_2 und β_3 enthalten würden.

Erinnern wir uns, dass es eine enge Beziehung zwischen Konfidenzintervallen und Hypothesentests gibt. Wenn wir den unter der Nullhypothese vermuteten Wert des Parameters h mit $\beta'_{h,0}$ bezeichnen (für $h = 1, \ldots, k$), kann die Nullhypothese geschrieben werden als $H_0: \beta_h = \beta'_{h,0}$. Wenn der unter H_0 vermutete Wert $\beta'_{h,0}$ im Konfidenzintervall liegt kann die Nullhypothese *nicht* verworfen werden, wenn der Wert $\beta'_{h,0}$ hingegen außerhalb des Konfidenzintervalls liegt darf die Nullhypothese verworfen werden. Analoges gilt auch für die Konfidenzellipse.

In Abbildung 5.20 ist ersichtlich, dass in diesem Beispiel weder die Nullhypothese $H_0: \beta_2 = 0$ noch die Nullhypothese $H_0: \beta_3 = 0$ einzeln abgelehnt werden kann,

der Nullpunkt liegt innerhalb der beiden individuellen Konfidenzintervalle (als grau strichlierte Linien eingezeichnet).

Hingegen kann die gemeinsame Nullhypothese, dass beide Steigungskoeffizienten simultan gleich Null sind $H_0 : \beta_2 = 0$ *und* $\beta_3 = 0$, abgelehnt werden, der Nullpunkt liegt außerhalb der Konfidenzellipse!

Man sieht in Abbildung 5.20, dass es sehr viele Punkte wie z.B. (β_2', β_3') gibt, die ähnlich wie der Nullpunkt sowohl im Konfidenzintervall von $\widehat{\beta}_2$ als auch im Konfidenzintervall von $\widehat{\beta}_3$ liegen, aber *nicht* in der Konfidenzellipse für $\widehat{\beta}_2$ *und* $\widehat{\beta}_3$.

Andererseits sind auch Fälle möglich, wie z.B. Punkt (β_2'', β_3'') in Abbildung 5.20, die außerhalb der beiden individuellen Konfidenzintervalle für $\widehat{\beta}_2$ und $\widehat{\beta}_3$ liegen, aber innerhalb der Konfidenzellipse. Das bedeutet, dass beide Koeffizienten individuell signifikant von β_2'' und β_3'' verschieden sind (d.h. es kann sowohl $H_0 : \beta_2 = \beta_2''$ als auch $H_0 : \beta_3 = \beta_3''$ individuell verworfen werden), aber dass die gemeinsame Nullhypothese $H_0 : \beta_2 = \beta_2''$ und $\beta_3 = \beta_3''$ *nicht* verworfen werden kann.

Im Beispiel von Abbildung 5.20 ist der empirische t-Wert für die $H_0 : \beta_2 = 0.85$ gleich $t^{\text{emp}} = 2.32$ (mit $p = 0.025$), für $H_0 : \beta_3 = -0.1$ ist $t^{\text{emp}} = 2.132$ ($p = 0.038$), beide Nullhypothesen können also auf einem Signifikanzniveau von 5% verworfen werden. Aber der empirische F-Wert der gemeinsamen Nullhypothese $H_0 : \beta_2 = 0.85$ und $\beta_3 = -0.1$ ist $F^{\text{emp}} = 2.74$ mit $p = 0.075$, die gemeinsame Nullhypothese kann also auf einem Signifikanzniveau von 5% *nicht* verworfen werden!

Die t-Statistiken der einzelnen Koeffizienten können also signifikant sein, obwohl die F-Statistik für die gemeinsame Nullhypothese nicht signifikant ist. Solche Fälle sind allerdings selten und treten meist nur bei hoher Multikollinearität auf.

Als nächstes werden wir eine allgemeinere Möglichkeiten kennen lernen eine F-Statistik zu berechnen, die den Test komplexerer Nullhypothesen erlaubt.

5.4.2 Simultane Tests für mehrere lineare Restriktionen

In diesem Abschnitt werden wir zeigen, wie mehrere Hypothesen *simultan* getestet werden können. Diese Tests beruhen auf der grundlegende Idee, dass Nullhypothesen immer als *Restriktion* auf ein allgemeineres Modell aufgefasst werden können, d.h. als Beschränkung.

Um dies zu erläutern kehren wir nochmals zur F-total Statistik zurück. Angenommen die PRF (*'population regression function'*) sei

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

und wir wollen testen, ob die Regressoren x_2 und x_3 einzeln *und/oder gemeinsam* einen Einfluss auf y ausüben.

Die entsprechende Nullhypothesen sind also

$$H_0 : \quad \beta_2 = 0 \quad \text{und} \quad \beta_3 = 0$$

Wenn – und nur wenn – diese H_0 wahr ist beschreibt die *restringierte* PRF

$$y_i = \beta_1 + \varepsilon_i \quad (= \beta_1 + 0 \cdot x_{i2} + 0 \cdot x_{i3} + \varepsilon_i)$$

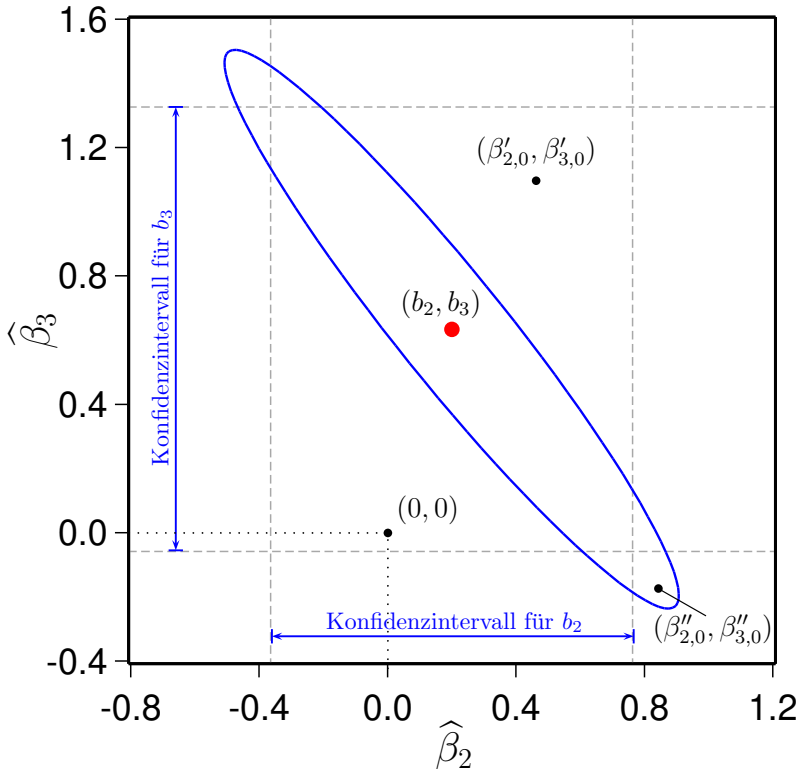


Abbildung 5.20: 95%-Konfidenzellipse für beide Koeffizienten und Konfidenzintervalle für die beiden einzelnen Koeffizienten für die Regression aus dem Beispiel auf Seite 57. Die t-Statistiken der einzelnen Koeffizienten zeigen, dass die Koeffizienten einzeln nicht signifikant von Null verschieden sind (für $\alpha = 0.05$), der Nullpunkt $(0, 0)$ liegt innerhalb der individuellen Konfidenzintervalle. Hingegen zeigt die F-total-Statistik, dass die Koeffizienten *gemeinsam* hochsignifikant von Null verschieden sind, der Nullpunkt liegt außerhalb der Konfidenzellipse!
Zum Beispiel können weder die $H_0 : \beta_2 = 0$ noch die $H_0 : \beta_3 = 0$ einzeln verworfen werden, aber die simultane $H_0 : \beta_2 = 0$ und $\beta_3 = 0$ wird abgelehnt (der Punkt $(0, 0)$ liegt wie viele weitere Punkte (z.B. $(\beta'_{2,0}, \beta'_{3,0})$) innerhalb der einzelnen Konfidenzintervalle, aber außerhalb der Konfidenzellipse.
Es ist auch möglich, dass die Nullhypothesen einzeln abgelehnt werden können, aber die simultane H_0 nicht abgelehnt werden kann, z.B. werden die $H_0 : \beta_2 = \beta''_{2,0}$ und die $H_0 : \beta_3 = \beta''_{3,0}$ beide einzeln verworfen, aber die simultane $H_0 : \beta_2 = \beta''_{2,0}$ und $\beta_3 = \beta''_{3,0}$ kann *nicht* verworfen werden!

die Daten gleich gut. Dies ist ein *Spezialfall* der allgemeinen PRF und ist nur korrekt, wenn die Nullhypothese wahr ist. In diesem Sinne können wir die Nullhypothesen als eine *Restriktion* auf ein allgemeineres Modell auffassen.

Die PRF können wir natürlich nicht beobachten, aber wir können die Gleichung mit und ohne Restriktion aus den Stichprobendaten schätzen. Wenn die Nullhypothesen wahr sind sollten sich die Residuen der beiden Schätzungen nicht wesentlich unterscheiden, weshalb die *Quadratsumme* der Residuen beider Modelle zumindest ähnlich sein sollte.

Wenn die Nullhypothese aber falsch ist würden wir erwarten, dass die Quadratsumme der Residuen des *restringierten* Modells (d.h. des Modells unter Berücksichtigung der Nullhypothese) deutlich *größer* ist.[18] Deshalb können die Quadratsummen der Residuen für einen Test verwenden. Tatsächlich kann man zeigen, dass aus diesen Quadratsummen der Residuen eine einfache Teststatistik konstruiert werden kann, die einen Test der Nullhypothese(n) erlaubt.

Wenn wir mit SSR_u (*Sum of Squared Residuals*) die Quadratsumme der Residuen *ohne Restriktionen* (der Subindex u steht für *'unrestricted'*) und mit SSR_r die Quadratsumme der Residuen des $r$estringierten Modells bezeichnen, so kann man zeigen, dass

$$\hat{F} = \frac{(\sum_i \hat{\varepsilon}_{r,i}^2 - \sum_i \hat{\varepsilon}_{u,i}^2)/q}{(\sum_i \hat{\varepsilon}_{u,i}^2)/(n-k)} := \frac{(\text{SSR}_r - \text{SSR}_u)/q}{\text{SSR}_u/(n-k)} \overset{\text{H}_0}{\sim} \text{F}_{q,n-k} \tag{5.8}$$

unter H_0 F-verteilt ist mit q Zähler- und $n-k$ Nennerfreiheitsgraden, wobei die Zählerfreiheitsgrade q der Anzahl der Restriktionen entspricht.[19]

Die Quadratsumme der Residuen des restringierten Modells ist $\sum_i \hat{\varepsilon}_{r,i}^2$ $(= \text{SSR}_r)$, und die des nicht restringierten Modells ist $\sum_i \hat{\varepsilon}_{u,i}^2$ $(= \text{SSR}_u)$.

Die Herleitung dieser Teststatistik ist etwas aufwändiger und wird im Kapitel zur Matrixnotation des OLS Modells kurz skizziert.

Dieser Test, bei dem ein Modell *ohne Restriktion(en)* mit einem Modell *unter Berücksichtigung einer oder mehrerer Restriktionen* verglichen wird, ist eine spezielle Form eines *Wald Tests*, benannt nach dem Mathematiker Abraham Wald (1902–1950, `https://de.wikipedia.org/wiki/Abraham_Wald`).

Ein Test zwischen zwei Modellen, bei denen es möglich ist eines der beiden Modelle durch geeignete Parameterrestriktionen in das andere Modell überzuführen, wird im Englischen als '*nested test*' (geschachtelte Hypothesen) bezeichnet. In diesem Kapitel werden wir uns ausschließlich auf solche '*nested tests*' beschränken.

[18] Da die Nullhypothese eine Restriktion darstellt kann die Anpassung des restringierten Modells nie besser sein als die des nicht restringierten Modells, deshalb kann die Quadratsumme der Residuen des restringierten Modells nie kleiner sein als die des nicht restringierten Modells.

[19] Wenn nur eine einzelne Hypothese getestet wird kann man zeigen, dass diese F-Statistik das Quadrat der entsprechenden t-Statistik ist.

Beispiel: Kehren wir nochmals zurück zum Beispiel mit der Produktionsfunktion (Seite 53). Die dort berechnete F-total Statistik erhalten wir natürlich auch, wenn wir die Quadratsummen der Residuen des restringierten und nicht restringierten Modells mit Hilfe der obigen Teststatistik vergleichen. Die nicht restringierte PRF ist

$$\log(Q)_i = \beta_1 + \beta_2 \log(K)_i + \beta_3 \log(L)_i + \varepsilon_i$$

und die Nullhypothese der F-total Statistik, dass alle Steigungskoeffizienten Null sind

$$\mathrm{H}_0: \quad \beta_2 = 0 \quad \text{und} \quad \beta_3 = 0$$

umfasst 2 Restriktionen ($q = 2$). Das restringierte Modell ist deshalb $\log(Q)_i = \beta_1 + \varepsilon_i^*$.

In R erhalten wir die Quadratsumme der Residuen einer Regression `equ` mit `deviance(equ)`. Mit dem folgenden R-Code können wir die H_0 testen:

```
pf <- read.csv(url("https://www.uibk.ac.at/econometrics/data/prodfunkt.csv"))

eq_u <- lm(log(Q) ~ log(K) + log(L), data = pf)
eq_r <- lm(log(Q) ~ 1, data = pf)

SSR_u <- deviance(eq_u)
SSR_r <- deviance(eq_r)

F_emp <- ((SSR_r - SSR_u)/2) / (SSR_u/(nobs(eq_u) - 3))
## 177.2195
p_val <- 1 - pf(F_emp, 2, nobs(eq_u) - 3)
## 2.708944e-14
```

Dies dient nur zur Demonstration, natürlich wird man den Test einfacher und sicherer mit dem R-Befehl anova() durchführen

```
anova(eq_r, eq_u)
## Analysis of Variance Table
##
## Model 1: log(Q) ~ 1
## Model 2: log(Q) ~ log(K) + log(L)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     24 3.2017
## 2     22 0.1871  2    3.0145 177.22 2.716e-14 ***
```

Die Schätzungen finden sich in Tabelle 5.5, Spalte (1) zeigt das nicht restringierte Modell, und Spalte (2) das restringierte Modell für die H_0: $\beta_2 = 0$ und $\beta_3 = 0$.

Zur Veranschaulichung ermitteln wir alternativ auch daraus die empirische F-Statistik

$$F^{\text{emp}} = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_u)/q}{\mathrm{SSR}_u/(n-k)} = \frac{(3.202 - 0.187)/2}{0.187/(25-3)} = 177.220$$

Tabelle 5.5: Spalte (1): *'unrestricted'*; Spalte (2): *'restricted'*, für F-total Test; Spalte (3): *'restricted'*, für Test auch konstante Skalenerträge.

	Dependent variable:		
	log(Q)		log(Q) - log(L)
	(1)	(2)	(3)
Constant	2.481***	4.376***	2.236***
	(0.129)	(0.073)	(0.048)
log(K)	0.640***		
	(0.035)		
log(L)	0.257***		
	(0.027)		
log(K) - log(L)			0.701***
			(0.019)
Observations	25	25	25
R^2	0.942	0.000	0.984
Sum of Squared Resid.	0.187	3.202	0.222
F Statistic	177.220***		1,392.158***
Note:			*p<0.1; **p<0.05; ***p<0.01

Der kritische Wert ist $F^{\text{crit}}_{0.05,2,22} = 3.443$, die empirische F-Statistik fällt also klar in den Verwerfungsbereich (mit R erhalten Sie diesen kritischen Wert mit `qf(0.95, 2, 22)`).
Dieser Test ist sehr allgemein und kann für den Test beliebiger linearer Restriktionen auf Koeffizienten verwendet werden (solange $q < n$).

Beispiel: konstante Skalenerträge Wir haben diese Produktionsfunktion bereits früher (Seite 26) mit Hilfe einer t-Statistik auf konstante Skalenerträge getestet, d.h.

$$\mathrm{H}_0 \colon \beta_2 + \beta_3 = 1$$

Dies ist *eine* Restriktion ($q = 1$), falls Sie bezüglich der Anzahl der Restriktionen unsicher sind zählen Sie einfach die '=' Zeichen.
Das restringierte Modell erhalten wir durch Substitution der (umgeschriebenen) $\mathrm{H}_0 \colon \beta_3 = 1 - \beta_2$

$$\log(Q)_i = \beta_1 + \beta_2 \log(K)_i + (1 - \beta_2) \log(L)_i + \varepsilon_i^*$$

und kann einfach in der folgenden Form geschätzt werden

$$\log(Q)_i - \log(L)_i = \beta_1 + \beta_2(\log(K)_i - \log(L)_i) + \varepsilon_i^*$$

Die Schätzungen dieses restringierten Modells finden sich in Spalte (3) von Tabelle 5.5, die empirische Teststatistik ist

$$F^{\text{emp}} = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_u)/q}{\mathrm{SSR}_u/(n-k)} = \frac{(0.222 - 0.187)/1}{0.187/(25-3)} = 4.117647$$

Der kritische Wert ist $F^{\text{crit}}_{0.05,1,22} = 4.301$, die empirische F-Statistik fällt also in den Annahmebereich, wir können die Nullhypothese konstanter Skalenerträge auf einem Signifikanzniveau von 5% *nicht* verwerfen.

In R können Sie in diesem Fall nicht den `anova()` Befehl verwenden, da dieser für beide Modelle die identische abhängige Variable benötigt (hier haben wir $\log(Q)$ und im anderen Fall $\log(Q) - \log(L)$).
Aber wir können das R package `car` (und das schon vorher geschätzte nicht restringierte Modell `eq_u` verwenden

```
library(car)
linearHypothesis(eq_u, c("log(K) + log(L) = 1"))

## Linear hypothesis test
## Hypothesis:
## log(K)  + log(L) = 1

## Model 1: restricted model
## Model 2: log(Q) ~ log(K) + log(L)

##  Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     23 0.22242
## 2     22 0.18711  1  0.035311 4.1517 0.0538 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

oder alternativ und etwas umständlicher

```
eq_rc <- lm(I(log(Q) - log(L)) ~ I(log(K) - log(L)), data = pf)
F_Stat <- ((deviance(eq_rc) - deviance(eq_u))/1) /
           (deviance(eq_u)/(25-3))  ## 4.151729
p_val <- 1 - pf(F_Stat, 1, 22)      ## 0.053796
```

Hinweis: Da wir hier nur eine Nullhyhpothese getestet haben konnten wir dies bereits in einem früheren Abschnitt (Seite 26) mit Hilfe einer t-Statistik testen. Wie Sie einfach überprüfen können stimmen die p-Werte exakt überein. Darüber hinaus kann man zeigen, dass im Fall einer einzigen Hypothese die F-Statistik exakt das Quadrat der t-Statistik ist.

Achtung: bei fehlenden Werten ist darauf zu achten, dass restringiertes und nicht-restringiertes Modell auf der gleichen Beobachtungszahl beruhen. Wenn einzelne Beobachtungen einer x Variable fehlen, die im restringierten Modell nicht vorkommt, kann es passieren, dass restringiertes und nicht-restringiertes Modell auf einer anderen Datengrundlage beruhen und die Statistik deshalb fehlerhaft berechnet wird.

Einige Beispiele für lineare Restriktionen

Im Prinzip werden für geschachtelte (*'nested'*) Tests immer zwei Gleichungen geschätzt, eine Gleichung ohne Restriktion(en), das *nicht-restringierte* Modell (*'unrestricted'*), und eine Gleichung *mit* Restriktion(en) – das *restringierte* Modell – das man durch Einsetzen der Nullhypothese in das nicht restringierte Modell erhält, und diese beiden Modelle anhand der F-Statistik verglichen.

Damit können alle linearen Restriktionen getestet werden, hier zeigen wir nur einige wenige Beispiele. Im Folgenden bezeichnet q wieder die Anzahl der Restriktionen (= Zählerfreiheitsgrade). Man beachte, dass alle Hypothesen mit nur einer Restriktion alternativ auch mit einer t-Statistik getestet werden könnten. Da der F-Test aber allgemeiner ist verwenden fast alle Programme auch für den Test einer einzelnen Hypothese die F-Statistik.

- PRF: $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

  Restriktion: $\boxed{\mathrm{H}_0 : \ \beta_3 = 1}$ $(q = 1)$

  SRF *ohne* Restriktion:
  $$y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \hat{\varepsilon}$$

  SRF *mit* Restriktion:
  $$y - x_3 = \widehat{\beta}_1^* + \widehat{\beta}_2^* x_2 + \hat{\varepsilon}^*$$

- PRF: $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

  Restriktion: $\boxed{\mathrm{H}_0 : \ \beta_2 = \beta_3}$ $(q = 1)$

  SRF *ohne* Restriktion:
  $$y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \hat{\varepsilon}$$

  SRF *mit* Restriktion:
  $$y = \widehat{\beta}_1^* + \widehat{\beta}_2^* (x_2 + x_3) + \hat{\varepsilon}^*$$

Wenn die Nullhypothese $\beta_2 = \beta_3$ wahr ist würden wir für die Schätzungen des restringierten und nicht-restringierten Modells zumindest 'sehr ähnliche' Ergebnisse erwarten.

- PRF: $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

  Restriktionen: $\boxed{\mathrm{H}_0: \ \beta_2 = 0 \ \ \& \ \ \beta_3 = 0 \ \ \& \ \ \beta_4 = 0}$ $(q = 3)$

  SRF *ohne* Restriktion:
  $$y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4 + \hat{\varepsilon}$$

  SRF *mit* Restriktion:
  $$y = \widehat{\beta}_1^* + \hat{\varepsilon}^*$$

  Das ist genau die Hypothese, die auch mit der F-total Statistik getestet wird.

- PRF: $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

  Restriktionen: $\boxed{\mathrm{H}_0: \ \beta_2 = \beta_3 = \beta_4}$ $(q = 2)$

  SRF *ohne* Restriktion:
  $$y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4 + \hat{\varepsilon}$$

  SRF *mit* Restriktion:
  $$y = \widehat{\beta}_1^* + \widehat{\beta}_2^* (x_2 + x_3 + x_4) + \hat{\varepsilon}^*$$

- PRF: $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

  Restriktion: $\boxed{\mathrm{H}_0: \ \beta_2 + \beta_3 = 1}$ $(q = 1)$

  SRF *ohne* Restriktion:
  $$y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \hat{\varepsilon}$$

  SRF *mit* Restriktion:
  $$y = \widehat{\beta}_1^* + \widehat{\beta}_2^* x_2 + (1 - \widehat{\beta}_2^*) x_3 + \hat{\varepsilon}^*$$
  diese Gleichung kann in der folgenden Form geschätzt werden:
  $$y - x_3 = \widehat{\beta}_1^* + \widehat{\beta}_2^* (x_2 - x_3) + \hat{\varepsilon}^*$$

  Man beachte, dass in diesem Fall zwei neue Variablen $y - x_3$ und $x_2 - x_3$ angelegt werden müssen.

- PRF: $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

  Restriktionen: $\boxed{\mathrm{H}_0: \ \beta_2 = 0.5 \times \beta_3 \ \ \& \ \ \beta_4 = -1}$ $(q = 2)$

  SRF *ohne* Restriktion:
  $$y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4 + \hat{\varepsilon}$$

  SRF *mit* Restriktionen:
  $$y + x_4 = \widehat{\beta}_1^* + \widehat{\beta}_3^* [0.5 x_2 + x_3] + \hat{\varepsilon}^*$$

Wenn die entsprechenden Nullhypothesen die Daten sehr gut beschreiben würden wir für die Schätzungen des restringierten und nicht-restringierten Modells sehr ähnliche Ergebnisse erwarten, insbesondere sollten sich auch die Quadratsummen der Residuen des nicht-restringierten und des restringierten Modells nicht stark unterscheiden. Deshalb können diese Hypothesen einfach mit obiger F-Statistik getestet werden.

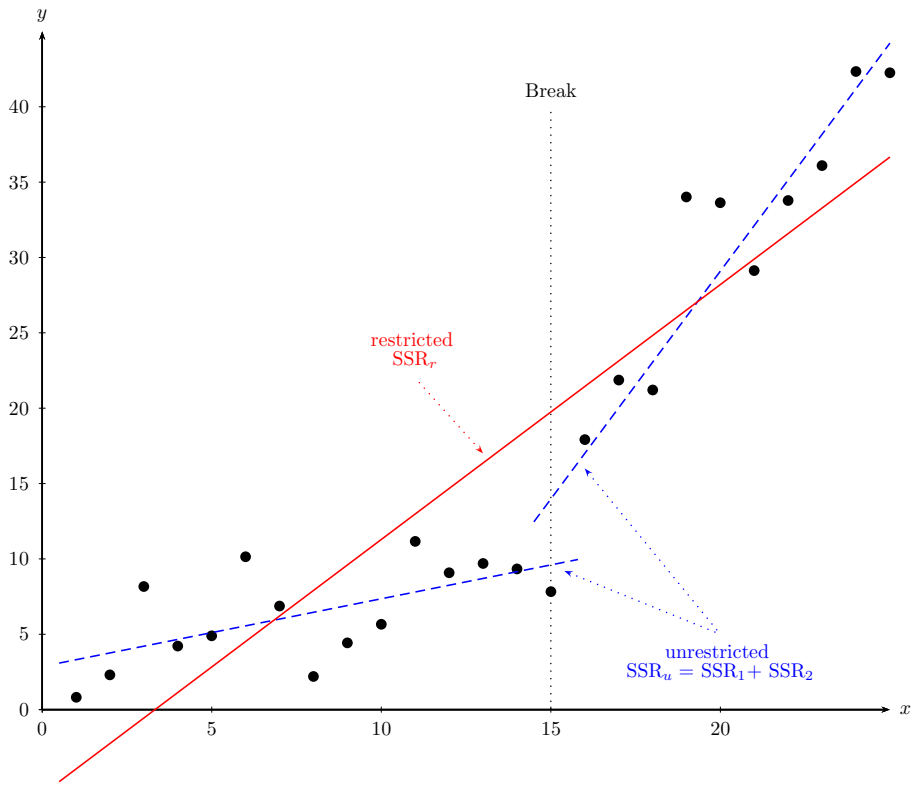


Abbildung 5.21: Chow-Test auf Strukturbruch, das *nicht*-restringierte Modell besteht aus den beiden blau strichliert eingezeichneten Einzelgleichungen für die getrennten Stichproben, das restringierte Modell aus der rot solid eingezeichneten Regression über alle Beobachtungen.

5.4.3 Ein Test auf Strukturbrüche (*Chow Test*)

Ein häufiges empirisches Problem ist die Überprüfung der Wirksamkeit von Maßnahmen, oder die Erkennung systematischer Unterschiede zwischen zwei (oder mehreren) Gruppen. Solche Hypothesen können mit einem Chow Test einfach überprüft werden. Bei einem Chow Test wird überprüft, inwieweit sich die Koeffizienten zweier Regressionsgleichungen systematisch unterscheiden. Deshalb ist er ein Spezialfall eines *Strukturbruchtests*.

Am einfachsten kann die grundlegende Idee anhand von Abbildung 5.21 erläutert werden. Stellen Sie sich vor, auf der x-Achse sei die Zeit aufgetragen, und in der Periode 15 wäre eine Maßnahme gesetzt worden (z.B. ein Förderprogramm oder eine Steuererhöhung).

Falls die Maßnahme wirkungslos war würden wir erwarten, dass die (rot) durchgezogene Regressionslinie die Daten gleich gut erklärt wie zwei Einzelregressionen, von denen eine über die Beobachtungen *vor* Setzung der Maßnahme und die andere für die Beobachtungen *nach* Setzung der Maßnahme geschätzt wurde. Hatte die Maßnahme hingegen reale Auswirkungen würden wir erhebliche Unterschiede erwarten. Wir benötigen nur noch ein Testverfahren, um solche Unterschiede in realen Daten zu erkennen.

Der Ökonometriker G. Chow (`https://de.wikipedia.org/wiki/Gregory_Chow`) erkannte, dass ein solcher Test als relativ einfache Erweiterung des früheren F-Tests für genestete Modelle (*'nested models'*) durchgeführt werden kann.

Er konnte zeigen – und das ist der zentrale Punkt –, dass die *Summe* der beiden Quadratsummen der Residuen der Einzelgleichungen die Quadratsumme der Residuen des nicht restringierten Modells bildet

$$\mathrm{SSR}_u = \mathrm{SSR}_1 + \mathrm{SSR}_2$$

Etwas allgemeiner für das multiple Regressionsmodell: wenn wir vermuten, dass in der Grundgesamtheit zwischen zwei Gruppen (oder Perioden) Unterschiede bestehen, schätzen wir für beide Gruppen getrennte Regressionen (die Gruppen sind mit einem hochgestellten Index 1 bzw. 2 gekennzeichnet)

$$\begin{aligned} y_i^1 &= \widehat{\beta}_1^1 + \widehat{\beta}_2^1 x_{i2}^1 + \cdots + \widehat{\beta}_k^1 x_{ik}^1 + \hat{\varepsilon}_i^1 \qquad \Rightarrow \quad \mathrm{SSR}_1 \\ &\quad \text{mit } i = 1, \ldots, n_1 \\ y_j^2 &= \widehat{\beta}_1^2 + \widehat{\beta}_2^2 x_{j2}^2 + \cdots + \widehat{\beta}_k^2 x_{jk}^2 + \hat{\varepsilon}_j^2 \qquad \Rightarrow \quad \mathrm{SSR}_2 \\ &\quad \text{mit } j = n_1 + 1, \ldots, n \end{aligned}$$

mit $i = 1, 2, \ldots, n_1$, und $j = n_1 + 1, n_1 + 2, \ldots, n$.

Die Nullhypothese ist, dass die folgenden Bedingungen simultan erfüllt sind

$$\mathrm{H}_0: \quad \beta_1^1 = \beta_1^2, \quad \beta_2^1 = \beta_2^2, \quad \ldots, \quad \beta_k^1 = \beta_k^2$$

d.h., diese Nullhypothese umfasst insgesamt k Restriktionen (alle k Koeffizienten der beiden Gleichungen sind gleich groß), weshalb wir k Zählerfreiheitsgrade der F-Statistik benötigen.

Für das nicht restringierte Modell müssen zwei Regressionen mit je k Variablen geschätzt werden, also insgesamt $2k$ Koeffizienten, deshalb ist die Anzahl der Nennerfreiheitsgrade des Chow Tests $n - 2k$.[20]

Deshalb ist nach Chow (1960) die folgende Teststatistik unter dieser Nullhypothese F-verteilt mit k Zähler- und $(n - 2k)$ Nennerfreiheitsgraden:

$$\boxed{\text{Chow Test:} \qquad \hat{F} = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_u)/k}{\mathrm{SSR}_u/(n-2k)} \overset{\mathrm{H}_0}{\sim} \mathrm{F}_{k,n-2k}}$$

mit $\mathrm{SSR}_u = \mathrm{SSR}_1 + \mathrm{SSR}_2$, wobei n wieder die Gesamtzahl der Beobachtungen und k die Anzahl der geschätzten Koeffizienten ist.

Die Alternativhypothese ist, dass *mindestens eine* der Restriktionen *nicht* erfüllt ist.

Annahmen des Chow Tests: der Chow Test liefert nur gültige Resultate, wenn die Störterme identisch und unabhängig verteilt sind, und – zumindest in kleinen Stichproben – wenn die Störterme normalverteilt sind.[21]

[20] Etwas ausführlicher, n_1 und n_2 die beiden Stichprobengrößen, und $n_1 + n_2 = n$. Die Freiheitsgrade sind deshalb $(n_1 - k) + (n_2 - k) = n_1 + n_2 - 2k = n - 2k$.

[21] Es gibt auch robuste Versionen dieses Tests. Das R package `lmtest` hat z.B. eine Funktion `waldtest()` mit der Option `vcov`, an die eine robuste Varianz-Kovarianzmatrix übergeben werden kann.

Tabelle 5.6: Chow Strukturbruchtest, Beispiel (vgl. Abbildung 5.21)

	Dependent variable: y		
	unrestr. 1	unrestr. 2	restricted
	(1)	(2)	(3)
Constant	2.863*	−22.398**	−5.627**
	(1.424)	(7.804)	(2.377)
x	0.449**	2.616***	1.692***
	(0.157)	(0.377)	(0.160)
Observations	15	10	25
R^2	0.387	0.857	0.830
Sum of Squared Resid.	89.317	93.815	764.471
F Statistic	8.224**	48.134***	111.920***
Note:			*p<0.1; **p<0.05; ***p<0.01

Beispiel: Tabelle 5.6 zeigt die Regressionen hinter Abbildung 5.21. Anhand der Quadratsumme der Residuen (*Sum od Squared Resid.*) können wir die $H_0 : \beta_1^1 = \beta_1^2$ und $\beta_2^1 = \beta_2^2$ (mit $q = 2 = k$) testen

$$F^{\text{emp}} = \frac{(\text{SSR}_r - \text{SSR}_u)/k}{\text{SSR}_u/(n-2k)} = \frac{(764.471 - (89.317 + 93.815))/2}{(89.317 + 93.815)/(25-4)} = 33.331$$

Der kritische Wert ist $F^{\text{crit}}_{0.05,2,21} = 3.4668$, deshalb können wir die Nullhypothese überzeugend verwerfen:

Interpretation: Wenn tatsächlich kein Unterschied zwischen den Koeffizienten der beiden Gleichungen besteht (d.h. wenn die Nullhypothese wahr ist) und wenn alle Annahmen erfüllt sind, dann würden wir bei wiederholter Durchführung des Zufallsexperiments in weit weniger als 5% der Fälle einen so extremen Wert der empirischen Teststatistik erwarten.

Das können wir auch einfach mit R berechnen

```
## Chow Test: Beispiel
dat <- read.csv(url("https://www.uibk.ac.at/econometrics/data/chow_bsp1.csv"))
eq_1 <- lm(y ~ x, subset = x <= 15, data = dat)
eq_2 <- lm(y ~ x, subset = x > 15,  data = dat)

SSR_u <- deviance(eq_1) + deviance(eq_2)  ## unrestricted
SRR_r <- deviance(lm(y ~ x, data = dat))  ## restricted

F_stat <- ((SRR_r - SSR_u)/2) / (SSR_u/(25-4))
## 33.33141
p_val  <- 1 - pf(F_stat, 2, 21)
## 3.046077e-07
```

Einfacher und schneller geht es mit dem R package `strucchange` (Zeileis et al., 2002)

```
library(strucchange)
sctest(y ~ x, data = dat, type = "Chow", point = 15)

##  Chow test
## data:  y ~ x
## F = 33.331, p-value = 3.046e-07
```

□

Hinweis: Wir hätten diesen Test auch mit Hilfe einer Dummy Variable durchführen können. Dazu definieren wir eine geeignete Dummyvariable d und schätzen das Interaktionsmodell

$$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i + \widehat{\beta}_3 d_i + \widehat{\beta}_4 x_i \times d_i + \hat{\varepsilon}_i$$

Die Nullhypothese ist $\mathrm{H}_0 \colon \beta_3 = 0$ und $\beta_4 = 0$.

Ein einfacher F-Test dieser Hypothese führt zum exakt gleichen Ergebnis wie der Chow Test! Dies kann einfach mit Hilfe von R demonstriert werden

```
# Interaktionsmodell mit Dummyvariable
dat$d <- ifelse(dat$x <= 15, 0, 1)
eq_d <- lm(y ~ x*d, data = dat)
eq_r <- lm(y ~ x, data = dat)

anova(eq_r, eq_d)
## Analysis of Variance Table
## Model 1: y ~ x
## Model 2: y ~ x * d
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     23 764.47
## 2     21 183.13  2    581.34 33.331 3.046e-07 ***
```

Da der Chow Test und ein entsprechender F-Test eines Interaktionsmodells mit Dummy Variablen zu numerisch exakt gleichen Ergebnissen führt ist es eine Frage der Bequemlichkeit, für welches Verfahren man sich entscheidet. Das Dummyvariablen Interaktionsmodell bietet mehr Flexibilität, kann aber in umfangreicheren Modellen schnell unübersichtlich werden. □

Beispiel: Lohngleichung Um ein etwas umfangreicheres Modell zu schätzen greifen wir auf die EU-Silc Daten der Statistik Austria zurück um eine einfache Lohngleichung zu schätzen. Dabei ist HBA (für Höchster BildungsAbschluss) eine kategoriale Variable (`factor`)

$$\widehat{\log(\mathrm{StdL})}_i = \widehat{\beta}_1 + \sum_{j=2}^{6} \widehat{\beta}_j \mathrm{HBA}_{ij} + \widehat{\beta}_7 \mathrm{Erf}_i + \widehat{\beta}_8 \mathrm{Erf}_i^2$$

Wenn wir uns z.B. für Lohnunterschiede zwischen Männern und Frauen interessieren könnten wir einfach eine Dummyvariable `weibl` hinzufügen. Dies würde aber implizieren, dass sich das Geschlecht nur auf das Level (d.h. Interzept) auswirkt, aber dass

Bildung (HBA) und Erfahrung bei Männern und Frauen die gleichen Auswirkungen haben.

Wenn wir testen möchten, ob es generelle Unterschiede zwischen den Geschlechtern gibt, können wir getrennte Regressionen für Männer und Frauen rechnen, und mit Hilfe eines Chow Tests überprüfen, ob diese in einem statistischen Sinne signifikant sind; die Schätzungen finden Sie in Tabelle 5.7.

Der ausführliche R Code für den Chow Test ist[22]

```
## Stundenlöhne: EU-Silc 2018, Statistik Austria
rm(list = ls())
s <- read.csv2("https://www.hsto.info/econometrics/data/silc2018.csv",
               stringsAsFactors = TRUE)
# cat(shQuote(levels(s$HBA)), sep = ", ")
s$HBA <- factor(s$HBA, levels = c(   ## levels ordnen
  "Pflichtschule",
  "Lehre mit Berufsschule",
  "Fach- oder Handelsschule",
  "Matura",
  "Abschluss an einer Universitaet, (Fach-)Hochschule",
  "Anderer Abschluss nach der Matura")
)

# Chow Test
eq_r <- lm(log(StdL) ~ HBA + Erf + I(Erf^2), data = s)
eq_m <- lm(log(StdL) ~ HBA + Erf + I(Erf^2), data = s, subset = weibl == 0)
eq_w <- lm(log(StdL) ~ HBA + Erf + I(Erf^2), data = s, subset = weibl == 1)

SSR_u <- deviance(eq_m) + deviance(eq_w)
SSR_r <- deviance(eq_r)
k <- length(coef(eq_r))  ## 8
n <- nobs(eq_r)          ## 3298

F_stat <- ((SSR_r - SSR_u)/k) / (SSR_u/(n - 2*k))
## 22.62028
p_val  <- 1 - pf(F_stat, k, n-2*k)
## 0
```

Die Schätzungen der Lohngleichungen in Tabelle 5.7 zeigen, dass alle Koeffizienten das erwartete Vorzeichen aufweisen, und dass alle Koeffizienten hoch signifikant von Null verschieden sind.

Die empirische F-Statistik des Chow Tests $F\text{emp} = 22.62028$ (mit p-Wert $= 0.000\ldots$) sagt uns, dass *wenn* in der Grundgesamtheit die Koeffizienten von Männern und Frauen gleich sind (also die Nullhypothese richtig ist), und *wenn* alle erforderlichen Annahmen erfüllt sind (u.a. die Gauss-Markov Annahmen), dann würden wir bei wiederholten Stichprobenziehungen nur in einer verschwindend geringen Anzahl von Fällen einen ähnlich extremen Wert der Teststatistik erwarten.

Achtung: Man beachte, dass die Annahmen u.a. ein richtig spezifiziertes Modell voraussetzen. In diesem Fall können wir davon ausgehen, dass zentrale Variablen

[22] In fast allen Fällen ist es natürlich klüger sich auf Routinen in geeigneten packages zu verlassen.

Tabelle 5.7: Lohngleichung (EU-Silc 2018, Statistik Austria)

	Dependent variable: log(StdL)		
	Männer	Frauen	Alle
	(1)	(2)	(3)
Constant	2.083***	2.011***	2.041***
	(0.048)	(0.043)	(0.033)
Lehre mit Berufsschule	0.140***	0.134***	0.164***
	(0.038)	(0.035)	(0.026)
Fach- oder Handelsschule	0.275***	0.307***	0.276***
	(0.048)	(0.037)	(0.030)
Matura	0.412***	0.349***	0.386***
	(0.042)	(0.037)	(0.028)
Universitaet, FH	0.644***	0.613***	0.640***
	(0.040)	(0.036)	(0.027)
Anderer Abschl. nach Matura	0.548***	0.587***	0.556***
	(0.076)	(0.054)	(0.045)
Erf	0.030***	0.023***	0.025***
	(0.003)	(0.003)	(0.002)
Erf^2	−0.0004***	−0.0003***	−0.0003***
	(0.0001)	(0.0001)	(0.00005)
Observations	1,579	1,719	3,298
R^2	0.306	0.268	0.273
Sum of Squared Resid.	201.076	226.966	451.643
F Statistic	98.787***	89.699***	176.165***

Note: *p<0.1; **p<0.05; ***p<0.01
Referenzkategorie: Pflichtschule

wie z.B. Charakterzüge, Intelligenz und ähnliches nicht berücksichtigt wurden, die sowohl mit dem Stundenlohn als auch dem Bildungsniveau korreliert sind. Deshalb sind die Koeffizienten verzerrt und *alle Teststatistiken ungültig*!

Hinweis: In diesem Fall ist ein Test mit Interaktionen und `anova()` deutlich einfacher und schneller, und obendrein flexibler

```
eq_r <- lm(log(StdL) ~  HBA + Erf + I(Erf^2), data = s)
eq_u <- lm(log(StdL) ~ (HBA + Erf + I(Erf^2) * weibl), data = s)
anova(eq_r, eq_u)
```

```
## Analysis of Variance Table
## Model 1: log(StdL) ~ HBA + Erf + I(Erf^2)
## Model 2: log(StdL) ~ weibl * (HBA + Erf + I(Erf^2))
##  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1   3290 451.64
## 2   3282 428.04  8    23.601 22.62 < 2.2e-16 ***
```

Damit könnten wir z.B. auch einfach testen, ob sich die Berufserfahrung *bei gleicher Bildung* und bei Berücksichtigung eines Level-Effekts (d.h. unterschiedliches Interzept) bei Männern und Frauen unterschiedlich auswirkt

```
eq_erf_r  <- lm(log(StdL) ~ HBA + Erf + I(Erf^2) + weibl, data = s)
eq_erf_u <- lm(log(StdL) ~ HBA + weibl * (Erf + I(Erf^2)), data = s)
anova(eq_erf_r, eq_erf_u)
```

```
## Analysis of Variance Table
## Model 1: log(StdL) ~ HBA + Erf + I(Erf^2) + weibl
## Model 2: log(StdL) ~ HBA + weibl * (Erf + I(Erf^2))
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   3289 429.03
## 2   3287 428.69  2   0.34036 1.3049 0.2714
```

Gegeben allen früheren Vorbehalten schließen wir daraus, dass sich bei gegebenem Bildungsniveau und wenn wir Unterschiede im Interzept zulassen der Unterschied beim Effekt der Berufserfahrung zwischen Männern und Frauen nicht statistisch signifikant von Null verschieden ist. □

Beim klassischen Chow Strukturbruchtest wird die Nullhypothese getestet, ob zwei Teilstichproben vom gleichen datengenerierenden Prozess erzeugt wurden. Bei Querschnittsdaten beruhen die Teilstichproben oft auf klar unterscheidbaren Gruppen (z.B. Männer & Frauen, Länder, ...), aber bei Zeitreihen ist oft weniger klar, wann ein Strukturbruch stattgefunden haben könnte.

Der klassische Chow Strukturbruchtest ist aber nur anwendbar, wenn der Zeitpunkt des Strukturbruchs a priori bekannt ist.[23] Der Quandt-Andrews Test ermöglicht einen Strukturbruchtest auch in Fällen, in denen der Zeitpunkt eines potentiellen Strukturbruchs nicht bekannt ist, und erlaubt darüber hinaus häufig zugleich Schätzung des Zeitpunktes eines möglichen Strukturbruchs.

[23] Auch bei Querschnittsdaten stellt sich dieses Problem manchmal, wenn die Daten nach einer Variable sortiert wurden und ein Schwellenwert vermutet wird, ab dem sich der Zusammenhang ändert.

5.4.4 Quandt-Andrews Test auf Strukturbrüche bei unbekannten Bruchzeitpunkten

Die Grundidee des Quandt-Andrews Tests (auch bekannt als *'Quandt-Likelihood-Ratio-Statistik'* (QLR) oder auch *'sup-Wald-Statistik'*) ist einfach. Angenommen wir vermuten, dass ein Strukturbruch zwischen zwei Zeitpunkte t_1 und t_2 stattgefunden hat; dann wird einfach für jedes Datum zwischen diesen Zeitpunkten ein Chow Strukturbruchtest durchgeführt und die entsprechende F-Statistik berechnet.

Die Quandt-Andrews Statistik (QLR) ist einfach der größte Wert all dieser F Statistiken

$$\text{QLR} = \max_{t_1<t<t_2} \{F(t_0), \ldots, F(t_1)\}$$

Die Nullhypothese besagt, dass alle Parameter des Regressionsmodells über die Zeit konstant sind.

Die Zeitpunkte t_1 und t_2 sollten nicht zu nahe am Anfang oder Ende der Zeitreihe liegen, da sonst die Eigenschaften dieses asymptotischen Tests sehr schlecht werden. Deshalb wird t_1 und t_2 häufig so gewählt, dass die ersten und letzten 7.5% der Beobachtungen, die vor t_1 bzw. nach t_2 liegen, also 15% aller Beobachtungen, ausgeschlossen werden. Diese 15% werden als *'trimming value'* bezeichnet.

Allerdings ist diese Quandt-Andrews Statistik (QLR) *nicht* mehr F verteilt, da die größte einer Reihe von F Statistiken ausgewählt wurde. Die Verteilung hängt unter anderem von der Anzahl der getesteten Restriktionen, von den 'trimming Werten' t_1/T und t_2/T, usw. ab.

Andrews (1993) konnte die Verteilung dieser Statistik bestimmen und Hansen (1997) ungefähre asymptotische p Werte dazu ermitteln. Deshalb ist diese Statistik sehr einfach anzuwenden. Darüber hinaus hat sie einige nützliche Eigenschaften

- Ähnlich wie der Chow Strukturbruchtest kann sie für einen Test aller Regressionskoeffizienten oder für den Test einer Teilmenge der Koeffizienten verwendet werden.
- Der Test kann auch Hinweise auf mehrere Strukturbrüche geben. Diese sind einfach zu erkennen, indem man die Werte der Teststatistiken für die einzelnen Perioden betrachtet.
- Falls es einen offensichtlichen Strukturbruch gibt kann man den Zeitpunkt mit dem maximalen Wert der F Statistik als Schätzer für den Zeitpunkt des Strukturbruchs verwenden.

Der Quandt-Andrews Test ist in Stata mit dem Postestimation Befehl `estat sbsingle` verfügbar.

In `R` ist dieser und eine große Zahl weiterer Strukturbruchtests im Packet `strucchange` (Zeileis et al., 2002) Zeieis verfügbar.

5.4.5 Ein allgemeiner Spezifkationstest: Ramsey's RESET Test

RESET ist die Abkürzung für *Regression Specification Error Test* und wurde von Ramsey (1969) vorgeschlagen. Es ist ein Test für Spezifikationsfehler sehr allgemeiner Art. Heute wird der RESET Test – wenn überhaupt – vor allem als Test auf die korrekte Funktionsform verwendet.

Ist das wahre Modell nicht-linear und wir schätzen irrtümlich eine lineare Regressionsgleichung, so wird diese nur in kleinen Bereichen die Daten adäquat abbilden. Eine Einbeziehung von Potenzen der *gefitteten* Werte von y würde die Qualität der Schätzung vermutlich verbessern. Auf dieser Grundidee beruht der RESET Test.

Die Durchführung des Tests ist (in der einfachsten Form) simpel:

1. Schätze das Modell (z.B. $y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \hat{\varepsilon}$) mittels OLS und berechne die gefitteten Werte von y, d.h. $\hat{y}$.
2. Schätze die ursprüngliche Gleichung und inkludiere zusätzlich (nicht-lineare) Transformationen von $\hat{y}$. Üblicherweise werden Potenzen von $\hat{y}$ verwendet, also $y = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 \hat{y}^2 + \widehat{\beta}_4 \hat{y}^3 + \hat{\varepsilon}$
3. Teste mittels Wald oder LR-Test, ob die Koeffizienten der Transformationen von $\hat{y}$ gemeinsam signifikant von Null verschieden sind (z.B. $H_0: \quad \beta_3 = \beta_4 = 0$).

Ein Vorteil des RESET-Tests besteht darin, dass das Alternativmodell nicht explizit spezifiziert werden muss. Andererseits ist der Test nicht konstruktiv und gibt keine Hinweise auf die 'richtige' Spezifikation (vgl. (Johnston and Dinardo, 1996, 121), (Wooldridge, 2005, 308f)).

In R sind mehrere Versionen dieses Tests mit dem Befehl `resettest` verfügbar, in Stata ist er als `ovtest` implementiert (etwas irreführend als *omitted variable test* benannt).

5.5 Wie vertrauenswürdig sind publizierte Hypothesentests?

Von Ehen sagt man, dass sie im Himmel geschlossen, aber auf Erden ausgetragen werden. Ähnlich verhält es sich mit Hypothesentests, die Theorie wurde ohne Zweifel von Genies entwickelt, aber ihre Anwendung auf Erden ist nicht immer über jeden Zweifel erhaben.

Es gibt eine ganze Reihe von Gründen, warum man bei Hypothesentests, deren genaues Zustandekommen man nicht kennt (was bei praktisch allen publizierten Studien der Fall ist) skeptisch bleiben sollte.

5.5.1 Fehlspezifikationen und nicht identifizierte Modelle

Wir haben schon früher betont, dass der 'wahre' datengenerierende Prozess (DGP) nicht beobachtbar ist. Um diesen schätzen zu können müssen wir aber Annahmen darüber treffen, wie er aussieht, wir müssen eine Spezifikation wählen. Wenn wir dabei wichtige Variablen übersehen, eine falsche Funktionsform unterstellen, oder *'feed-back'* Mechanismen übersehen, führt dies zu einer Fehlspezifikation und die resultierenden Schätzer sind weder erwartungstreu noch konsistent.

Außerdem haben wir bisher immer völlig selbstverständlich echte Zufallsstichproben unterstellt. Tatsächlich können solche fast nur in Lehr- und Märchenbüchern beobachtet werden. In vielen Fällen ist es fast ein Ding der Unmöglichkeit eine echte Zufallsstichprobe zu ziehen und – bei Befragungen – immer ehrliche Antworten zu erhalten. Verzerrte Stichproben führen zu verzerrten Ergebnissen, und daran kann kein Hypothesentest etwas ändern!

Selbstverständlich sind auch die Teststatistiken solcher fehlspezifizierter Modelle völlig wertlos.

Ein verwandtes Problem sind nicht identifizierte Modelle. Zur Erinnerung, bei nicht identifizierten Modellen reicht selbst die Kenntnis der 'wahren' gemeinsamen Verteilung der relevanten Variablen nicht aus um die interessierenden Parameter konsistent schätzen zu können. Schätzungen sowie Hypothesentests solcher nicht identifizierter Modelle können nicht interpretiert werden und sind ebenfalls wertlos.

Statistische Tests setzen korrekt spezifizierte und identifizierbare Modelle voraus. Sind die Modelle falsch spezifiziert sind die Tests nicht interpretierbar. Um Hinweise für eine möglichst korrekte Spezifikation zu finden ist theoretisches Nachdenken unumgänglich.

Ohne theoretischer Vorarbeit kann auch leicht passieren, was ein berühmter Statistiker einmal einen Typ III Fehler nannte.

> "In 1948, Frederick Mosteller (1916-2006) argued that a 'third kind of error' was required to describe circumstances he had observed, namely:
>
> - Type I error: 'rejecting the null hypothesis when it is true'.
> - Type II error: 'accepting the null hypothesis when it is false'.

- Type III error: 'correctly rejecting the null hypothesis for the wrong reason'."[24,25]

Ein nettes Beispiel für einen solchen 'Type III error' liefert die Medizin[26]

> "Selbstversuche haben in der Medizin Tradition. An die Grenze der menschlichen Belastbarkeit ging dabei der angehende Arzt Stubbins Ffirth am Anfang des 19. Jahrhunderts. Er war überzeugt, dass Malaria nicht ansteckend ist, sondern auf übermäßige Hitze, Essen und Lärm zurückzuführen sei.
>
> Um seine These zu erhärten, setzte er sich selbst der Krankheit aus. Zuerst brachte er nur kleine Mengen von frischem Erbrochenem in sich selbst zugefügte kleine Kratzer ein. Danach tropfte er kleine Mengen in seine Augen. Am Ende der Testreihe aß er die frischen Exkremente eines Kranken. Wie durch ein Wunder blieb er tatsächlich gesund. Er sah seine Behauptung somit belegt."

Der wackere Stubbins Ffirth konnte damals nicht wissen, dass die Malaria durch den Biss der weiblichen Stechmücke Anopheles übertragen wird, sein heldenmütiger Verzehr frischer Exkremente lieferte offensichtlich keinen Beweis für die Richtigkeit seiner Hypothese, dass Hitze, Lärm und schlechtes Essen die Ursache für die Malaria sei.

Im ökonometrischen Sinne ist ihm ein Spezifikationsfehler unterlaufen. Selbst wenn er entsprechende Beobachtungsdaten gesammelt und einen Hypothesentest durchgeführt hätte, hätte ihm dies wenig geholfen. Er hätte genauso festgestellt, dass Hitze die Wahrscheinlichkeit an Malaria zu erkranken statistisch signifikant erhöht, weil er nicht berücksichtigte, dass die Anopheles Mücke vorwiegend in heißen Gegenden vorkommt (*'omitted variable bias'*).

5.5.2 Data- and Estimator Mining (*p-hacking*)

Die Theorie der Hypothesentests beruht darauf, dass eine a priori festgelegte Nullhypothese *einmalig* mit den Daten konfrontiert wird.

Da die Grundgesamtheit (bzw. der 'wahre' datengenerierende Prozess) nicht beobachtbar ist, ist die Versuchung groß, verschiedene Spezifikationen 'zu probieren'. Wenn wir bei einem Signifikanzniveau von 5% hundert Spezifikationen 'probieren' müssen wir damit rechnen, in fünf Fällen die Nullhypothesen irrtümlich zu verwerfen (Typ I Fehler). Es sollte klar sein, dass derart zustande gekommene Ergebnisse wertlos sind (vgl. Abb. 5.24).

Ein verwandtes Problem entsteht, wenn man verschiedene Schätz- und Testverfahren probiert und anschließend die statistisch signifikanten Ergebnisse selektiert.

---

[24] zitiert aus Wikipedia: `http://en.wikipedia.org/wiki/Type_I_error`; Quelle: Mosteller, F., *A k-Sample Slippage Test for an Extreme Population*, The Annals of Mathematical Statistics, Vol.19, No.1, (March 1948), pp.58-65.

[25] Ein Vorschlag für einen Typ IV Fehler stammt von dem Harvard Ökonom Howard Raiffa, *"solving the right problem too late"*.

[26] zitiert aus `http://science.orf.at/science/news/149922`, [05.11.2007].

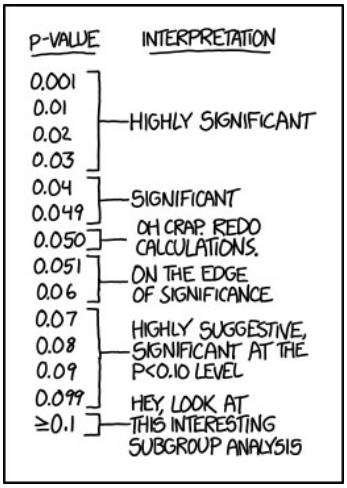


Abbildung 5.22: p-values Quelle: XKCD, `https://xkcd.com/1478/`.

Nach Leeb and Pötscher (2008) führt die Auswahl von Modellen anhand von vorausgehenden Tests (z.B. Modellselektionskriterien wie das R^2 oder Aikaike Informationskriterium) dazu, dass die Schätzung von Parametern aus solchen Modellen nicht einmal konsistente Ergebnisse liefert, wenn die Schätzung auf der gleichen Datengrundlage erfolgt wie die Modellselektion.[27]

Publizierten Resultaten von Hypothesentests kann man nicht ansehen, wie sie zustande gekommen sind, und es ist offensichtlich, dass der *'publish or perish'* Druck viele Forscher verleitet, gezielt nach signifikanten Ergebnissen zu suchen.

Dies ist natürlich ein altbekanntes Problem, so warnt die *American Statistical Society* in ihren *'Ethical Guidelines for Statistical Practice'*

> "Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one 'significant' result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading."
> (`http://www.amstat.org/about/ethicalguidelines.cfm`)

Ein verwandtes Problem ist das so genannte *'outcome switching'*. Im Juli 2012 verhängten amerikanische Behörden eine Rekordstrafe von drei Milliarden (!) US-Dollar über den Pharmakonzern GlaxoSmithKline (GKS). Was war geschehen? In einer Versuchsreihe – der mittlerweile berühmten Studie 329 – wurde ein Psychopharmaka (Paxil) auf die Wirksamkeit in Bezug auf auf acht a priori festgelegte

[27] "We consider the problem of estimating the unconditional distribution of a post-model-selection estimator. The notion of a post-model-selection estimator here refers to the combined procedure resulting from first selecting a model (e.g., by a model-selection criterion such as the Akaike information criterion [AIC] or by a hypothesis testing procedure) and then estimating the parameters in the selected model (e.g., by least squares or maximum likelihood), all based on the same data set. We show that it is impossible to estimate the unconditional distribution with reasonable accuracy even asymptotically. In particular, we show that no estimator for this distribution can be uniformly consistent (not even locally)" (Leeb and Pötscher, 2008).

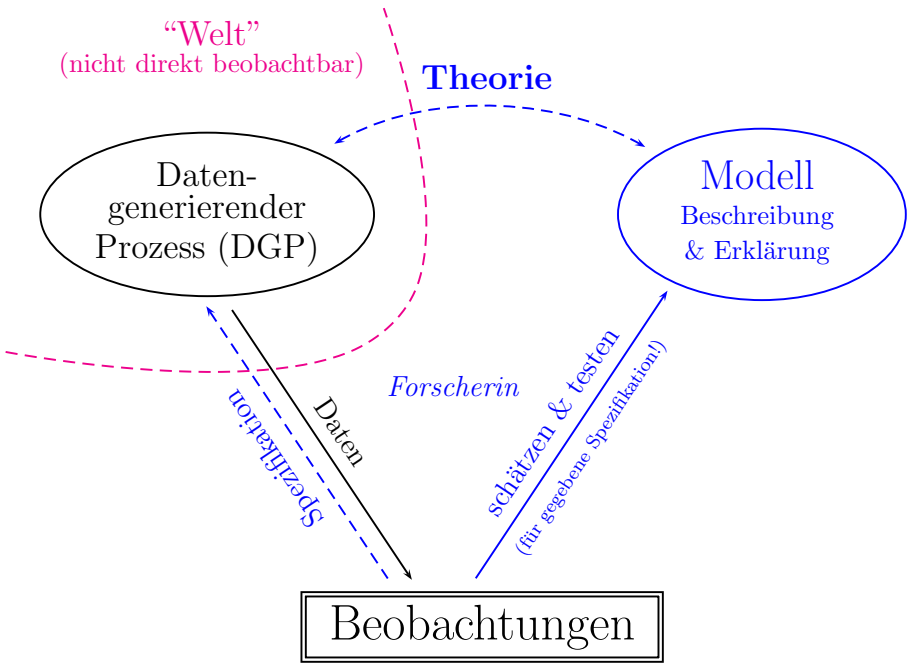


Abbildung 5.23: Die übliche Schätz- und Testtheorie setzt eine korrekte Spezifikation voraus. Wenn die Spezifikationssuche datengetrieben ist und auf den gleichen Daten wie die Schätzung beruht können die üblichen Verfahren fehlerhafte Ergebnisse liefern; siehe Diskussion zu *'pretest estimators'*, z.B. Danilov and Magnus (2004)

.

Ergebnis-Variablen gemessen. Es zeigte sich, dass das Medikament in keiner dieser acht Variablen signifikant bessere Ergebnisse erzielte als Placebos.

Darauf hin entschieden sich die Forscher einen Zusammenhang mit 19 weiteren Variablen zu testen, und fanden in vier Fällen tatsächlich signifikante Ergebnisse. In der Publikation wurden die früheren Versuchsreihen verschwiegen und vorgegeben, dass diese vier signifikanten Ergebnisse von vornherein festgelegt worden wären.

Später zeigte sich, dass dieses Medikament nicht nur wirkungslos in Bezug auf diese vier selektierten Ergebnisse war, sondern obendrein schwere Nebenwirkungen hatte. (The Economist, Mar 26th 2016, Clinical trials: For my next trick...).[28]

Eine spezielle Gefahr des *'data mining'* besteht darin, dass ex post fast jedes Resultat 'plausibel' erklärt werden kann. Nach F. Nietzsche sind *Überzeugungen die größeren Feinde der Wahrheit als Lügen!*

Der Psychologe und Wirtschaftsnobelpreisträger Kahneman (2013, 85) nennt unser schnelles, intuitives Denken *"a machine for jumping to conclusions"*. Dies ist besonders gefährlich, wenn Ergebnisse kausal interpretiert werden. Hypothesentests alleine können nie Kausalitäten bestätigen, sondern bestenfalls Assoziationen.

Auf diese Gefahr wird manchmal mit dem Schlagwort *'avoid HARKing'* verwiesen, wobei der Begriff HARKing ein englisches Akronym für *"Hypothesizing After the Results are Known"* ist.

[28] http://www.economist.com/news/science-and-technology/21695381-too-many-medical-trials-move-their-goalposts-halfway-through-new-initiative

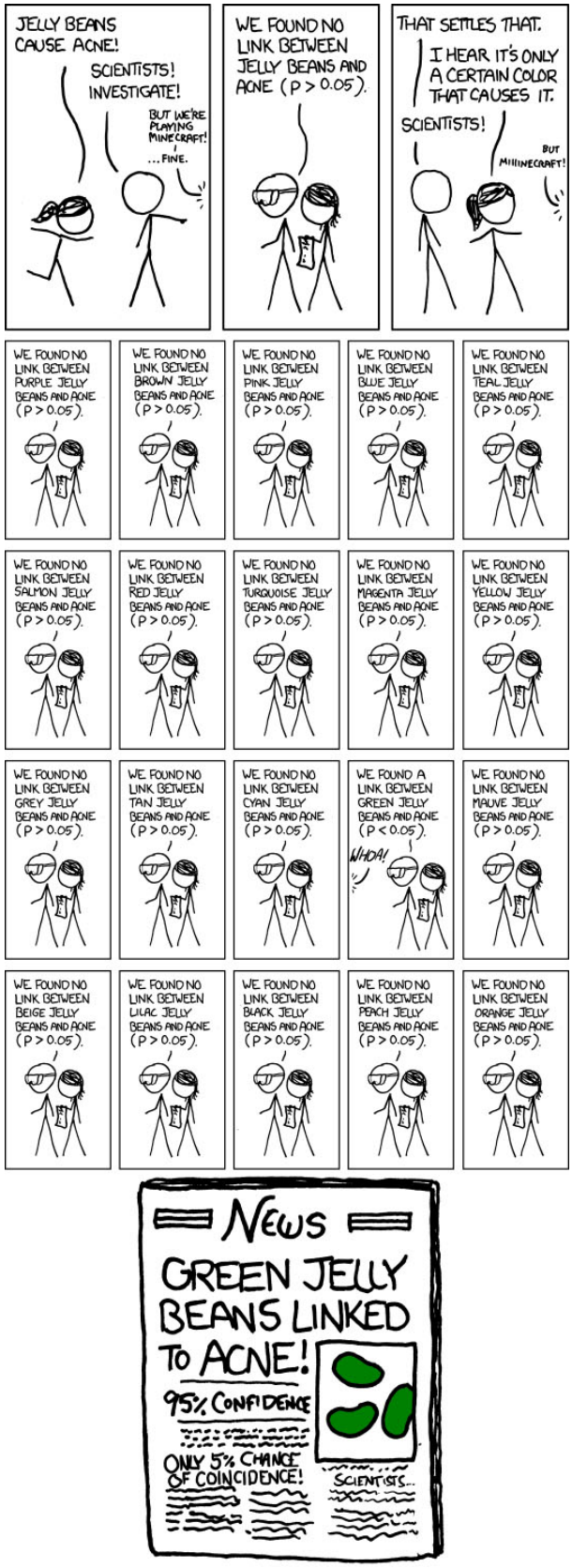


Abbildung 5.24: Data-mining, *p*-hacking
Quelle: XKCD, `https://xkcd.com/882/`

.

Erschwert wird dieses Problem noch dadurch, dass Forscher wie alle anderen Menschen 'Fehlwahrnehmungen' unterliegen, wie z.B. dem berühmten *'confirmation bias'*. Nach George Box neigen Forscher manchmal dazu, sich wie Künstler in ihre Modelle zu verlieben. Überflüssig zu erwähnen, dass dies den klaren Blick trüben kann.

Bereits dem Pionier des empirischen Denkens, Sir Francis Bacon (1561 – 1626), war dies offensichtlich nicht fremd, er schreibt

> "The human understanding when it has once adopted an opinion [...] draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects, in order that by this great and pernicious predetermination the authority of its former conclusion may remain inviolate." (Francis Bacon, Novum Organon, XLVI, 1620)

5.5.3 Niedrige Power und Publikationsbias

Ein spezielles Problem besteht darin, dass viele Menschen ein Problem haben bedingte Wahrscheinlichkeiten richtig zu interpretieren.

In einem korrekt durchgeführten Hypothesentest gibt uns der Typ I Fehler die Wahrscheinlichkeit dafür an, dass die empirische Teststatistik in den Verwerfungsbereich fällt, *wenn die Nullhypothese richtig ist.*

In der Regel interessieren wir uns aber für eine ganz andere Frage, nämlich wie groß ist die Wahrscheinlichkeit, dass die Nullhypothese falsch ist, *wenn die empirische Teststatistik in den Verwerfungsbereich fällt*! Das ist offensichtlich eine ganz andere Frage, auf die uns ein einfacher Hypothesentest keine Antwort gibt.

Der Economist (Oct 19th 2013) erklärt das Problem anhand eines einfachen Zahlenbeispiels, siehe Abbildung 5.25.

Nehmen wir an es gibt 1000 zu testende Hypothesen, und 100 dieser (Alternativ-) Hypothesen seien tatsächlich wahr, die restlichen 900 falsch.

Die Forscherin weiß dies natürlich nicht, sie testet alle 1000 Hypothesen. Von den tatsächlich falschen 900 Hypothesen wird sie irrtümlich 45 als richtig klassifizieren (5% von 900, Typ I Fehler).

Wenn der gewählte Test eine (sehr gute) Power von 0.8 hat (also 80% der tatsächlich wahren Hypothesen auch als wahr erkennt) wird sie 20% der 100 tatsächlich richtigen Hypothesen irrtümlich verwerfen. Sie glaubt also $80 + 45 = 125$ richtige Hypothesen vorliegen zu haben, davon sind aber 45, das sind 36%, tatsächlich falsch!

Ein besseres Ergebnis würde sie erzielen, wenn sie die nicht signifikanten Ergebnisse ansehen würde. In 875 ($= 1000 - 125$) Fällen wird die Hypothese verworfen, bei einer Power von 0.8 in nur 20 Fällen davon zu unrecht, was einer Trefferquote von fast 98% entspricht. Aber negative Resultate fallen häufig dem *'Publication Bias'* zum Opfer.

Hinweis:* Dieses Resultat kann man auch allgemeiner mit Hilfe des **Satzes von Bayes** zeigen.

Sei

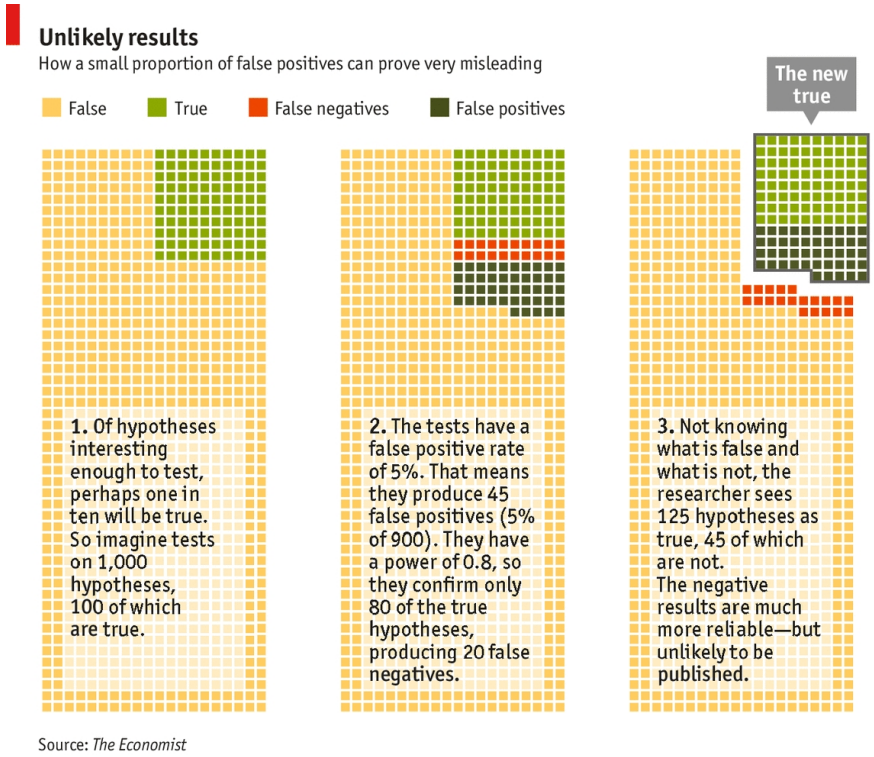


Abbildung 5.25: Unreliable research: Trouble at the lab; Quelle: The Economist, Oct 19th 2013
`https://youtu.be/TosyACdsh-g`

- A das Ereignis die 'Nullhypothese ist falsch' (also ist die interessierende Alternativhypothese richtig),
- B das Ereignis, die 'empirische Teststatistik fällt in den Verwerfungsbereich' (d.h. die Forscherin wird die Nullhypothese verwerfen).

Damit können wir die Wahrscheinlichkeit für einen Typ I Fehler und für die Power definieren.

Der Typ I Fehler gibt die Wahrscheinlichkeit dafür an, dass bei einer neuerlichen Ziehung die empirische Teststatistik in den Verwerfungsbereich fällt, *wenn die Nullhypothese richtig ist*. Da eine richtige Nullhypothese das Komplementärereignis zur falschen Nullhypothese ist können wir schreiben

$$P(\text{Typ I Fehler}) = P(B|\overline{A})$$

Ein Typ II Fehler gibt die Wahrscheinlichkeit dafür an, dass die empirische Teststatistik *nicht* in den Verwerfungsbereich fällt, wenn die Nullhypothese falsch ist, also $P(\overline{B}|A)$. Die Power eines Tests ist die Gegenwahrscheinlichkeit für einen Typ II Fehler, also

$$\text{Power} = 1 - P(\overline{B}|A) = P(B|A)$$

Die Power gibt also die Wahrscheinlichkeit dafür an, dass bei einer tatsächlich falschen Nullhypothese die empirische Teststatistik in den Verwerfungsbereich fällt, ist also ein Maß für die Treffsicherheit eines Tests.

Wir interessieren uns aber meist nicht für die Wahrscheinlichkeit eines Typ I Fehlers, das heißt die Wahrscheinlichkeit, dass die empirische Teststatistik in den Verwerfungsbereich fällt, *wenn die Nullhypothese richtig ist* $P(B|\overline{A})$, sondern für die Wahrscheinlichkeit dafür, dass die Nullhypothese tatsächlich falsch ist, *wenn die empirische Teststatistik in den Verwerfungsbereich fällt*, also für $P(A|B)$.

In anderen Worten, wir möchten wissen wie groß die Wahrscheinlichkeit dafür ist, dass die Nullhypothese wirklich falsch ist, *gegeben wir beobachten eine signifikante Teststatistik*. Dies ist offensichtlich eine komplett andere Frage!

Wenn wir zusätzliche Information haben, nämlich die Power des Tests und die a priori Wahrscheinlichkeit $P(A)$ kennen, können wir diese Frage mit Hilfe des Satzes von Bayes beantworten[29]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}$$

Dazu benötigen wir zusätzlich die Information $P(A)$, d.h. die a priori Wahrscheinlichkeit dafür, dass eine zufällig gezogene Nullhypothese falsch ist (bzw. den Anteil aller falschen Nullhypothesen an der Grundgesamtheit aller 'testwürdigen' Nullhypothesen). Im folgenden Beispiel werden wir (wie vorhin im Beispiel des Economist) $P(A) = 0.1$ annehmen.

Wenn wir außerdem übliche Werte für die Wahrscheinlichkeit eines Typ I Fehlers und die Power annehmen,

$$\begin{aligned} P(\text{Typ I Fehler}) = P(B|\overline{A}) &= 0.05 \\ \text{Power} = P(B|A) &= 0.8 \\ P(A) &= 0.1 \end{aligned}$$

und einsetzen erhalten wir

$$P(A|B) = \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.05 \times 0.9} = 0.64$$

Das bedeutet, dass gegeben die empirische Teststatistik fällt in den Verwerfungsbereich die Wahrscheinlichkeit dafür, dass die Nullhypothese tatsächlich falsch ist, lediglich 64 Prozent beträgt.

Im Umkehrschluss bedeutet dies, dass unter obigen Annahmen die Wahrscheinlichkeit dafür, dass bei einer gegebenen signifikanten Teststatistik die Nullhypothese trotzdem richtig ist, 36% $(1 - 0.64 = 0.36)$ beträgt!

Dabei ist eine Power von 0.8 und eine a priori Wahrscheinlichkeit $P(A) = 0.1$ noch ziemlich optimistisch, wenn die Power nur 0.6 ist und $P(A) = 0.05$ zeigen bereits 61%

[29] Die einfache Version des Satzes von Bayes folgt unmittelbar aus der Definition der bedingten Wahrscheinlichkeit

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{P(A \cap B)}{P(A)} P(A)}{P(B)} = \frac{P(B|A)\, P(A)}{P(B)}$$

Da A und $\overline{A}$ eine Partition bilden folgt aus dem Satz der Totalen Wahrscheinlichkeit $P(B) = P(B|A)\, P(A) + P(B|\overline{A})\, P(\overline{A})$.

aller Hypothesentest ein irreführendes Ergebnis (d.h. die Nullhypothese ist richtig obwohl die empirische Teststatistik in den Verwerfungsbereich fällt). ■

Dieses Problem wird durch den *'Publication Bias'* noch vergrößert. Der 'Publication Bias' besteht darin, dass Referees und Zeitschriften häufig nur signifikante Ergebnisse als publikationswürdig erachten. Dies kann dazu führen, dass ein völlig verzerrtes Bild der signifikanten Ergebnisse entsteht.

5.5.4 Statistische Signifikanz und Kausalität

Anfänger machen manchmal den Fehler, statistisch signifikante Zusammenhänge mit Kausalität zu verwechseln. Dies ist natürlich völliger Unsinn, ein Hypothesentest alleine kann uns nichts über mögliche Kausalitätsbeziehungen verraten.

Hypothesentests wurden ausschließlich entwickelt um Stichprobenfehler (*'sampling errors'*) zu berücksichtigen, und dies vor allem für Experimente. Bereits R.A. Fisher war klar, dass die Rechtfertigung einer möglichen Kausalitätsbeziehung aus dem *'Design of Experiments'* (1935) herrühren muss, reine Hypothesentests sind dazu ungeeignet.

Ein Beispiel soll dies verdeutlichen: Angenommen, eine Regression zwischen Werbeausgaben x und Umsätzen y liefert einen positiven und statistisch signifikanten Koeffizienten für die Werbeausgaben.

Können wir daraus schließen, dass höhere Werbeausgaben x höhere Umsätze y *verursachen*?

Dazu überlegen wir, welche mögliche Ursachen für eine beobachtete Korrelation zwischen Werbeausgaben x und Umsätzen y denkbar sind:

- x ist die Ursache für y: höhere Werbeausgaben führen zu höheren Umsätzen.
- y ist die Ursache für x: (*reverse causality*) höhere Umsätze ermöglichen die Finanzierung zusätzlicher Werbeausgaben.

  Auch bei Simultanität, z.B. wenn das Angebot und Nachfrage nach Werbung durch zwei interdependente Gleichungen beschrieben wird, führt die OLS Schätzung einer Einzelgleichung zu falschen Ergebnissen.
- z ist eine gemeinsame Ursache für x und y: (*confounding variable*, Scheinkorrelation), z.B. eine gute Konjunktur führt zu steigenden Umsätzen *und* zu steigenden Werbeausgaben (*'omitted variable'*).
- Die Korrelation könnte durch eine verzerrte Stichprobe, Selbstselektion oder ähnliches zustande gekommen sein.
- Die Korrelation zwischen x und y tritt in einer perfekten Zufallsstichprobe rein zufällig auf: dies – und nur dieser Fall – sollte durch einen statistischen Test erkennbar sein.

Obwohl statistische Signifikanz offensichtlich in keiner Weise eine hinreichende Bedingung für eine mögliche Kausalitätsbeziehung ist, wird man trotzdem statistische Signifikanz fordern, um den letzten der obigen Fälle kontrolliert klein zu halten, nämlich die Möglichkeit eines zufälligen Stichprobenfehlers.

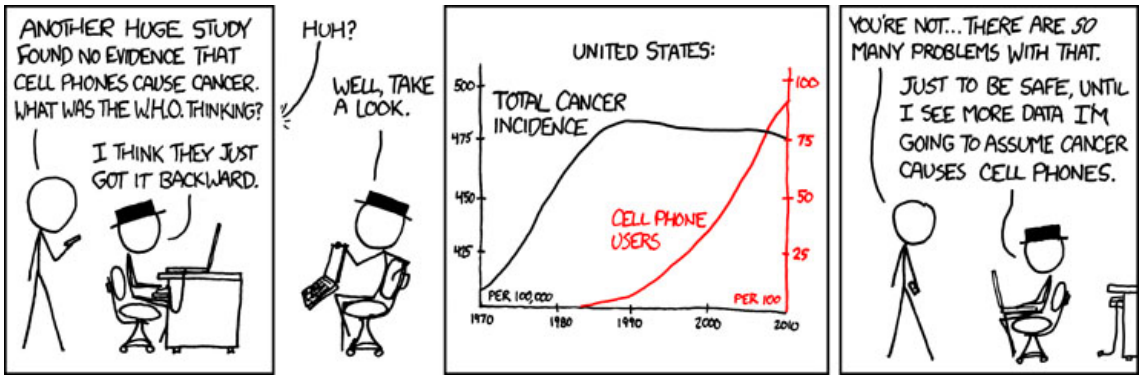


Abbildung 5.26: Kausalität Quelle: XKCD, `https://xkcd.com/925/`

.

5.5.5 Statistische Signifikanz versus 'Relevanz' einer Variablen

Manchmal wird statistische Signifikanz mit der Bedeutung einer Variable verwechselt (*Effektstärke*). Statistische Signifikanz sagt nichts darüber aus, ob die Größe des gemessenen Koeffizienten auch praktisch relevant ist. Da der geschätzte Koeffizient auch von der Dimension abhängt, in der die Variablen gemessen wurden, kann es manchmal nützlich sein, den Koeffizienten einer Variable mit dem Mittelwert dieser Variable zu multiplizieren ($b_h \bar{x_h}$), um einen Eindruck von der quantitativen 'Bedeutung' einer Variable zu bekommen. Ein Zusammenhang kann zwar statistisch hoch signifikant sein, aber für praktische Zwecke trotzdem völlig bedeutungslos sein!

Zu beachten ist auch der Zusammenhang zwischen Stichprobengröße und statistischer Signifikanz. In sehr großen Stichproben sind selbst winzige Unterschiede oft statistisch signifikant, und es kann fast jede Nullhypothese verworfen werden, z.B. für den t-Test

$$\text{für } n \to \infty: \quad t = \frac{\widehat{\beta}_h}{\hat{\sigma}_{\widehat{\beta}_h}} \approx \frac{\beta_h}{0} = \infty$$

Wie bereits erwähnt sind statistische Signifikanz und ökonomische Bedeutung (Relevanz) zwei verschiedene Paar Schuhe.

Die Gültigkeit von Test hängt von einer Reihe von Annahmen ab, die oft schwer zu überprüfen sind, z.B.

- ist die vorliegende Stichprobe tatsächlich eine Zufallsstichprobe, oder ist ein *sample selection bias* zu befürchten?
- ist die der Hypothese zugrunde liegende Kausalitätsvorstellung tatsächlich angebracht, oder ist simultane Kausalität zu befürchten?
- wurden alle relevanten Einflussfaktoren berücksichtigt, oder könnte das Ergebnis durch eine unbeobachtete Variable verursacht worden sein (*omitted variable bias*)?
- usw.

Ein Hypothesentest nach Neyman-Pearson ist in erster Linie eine Entscheidungsregel, aber diese Entscheidungsregel ist nur anwendbar, wenn sie auf zutreffender Information beruht. Ein simpler Hypothesentest alleine kann nicht zwischen Scheinkorrelationen und Kausalität unterscheiden, dazu bedarf es mehr. Erinnern Sie sich, die erste Tugend einer Wissenschaftlerin ist Skepsis. Fragen Sie sich stets *'was könnte das Ergebnis, dass Sie zu sehen glauben, sonst verursacht haben als das, was Sie zu sehen wünschen?'*

Tests können – vernünftig angewandt – ein sehr mächtiges und nützliches Werkzeug sein, aber man kann damit auch ziemlich viel Unfug treiben.

Zusammenfassend: *Blindes Vertrauen ist meistens dumm. Blindes Vertrauen in einen statistischen Test ist davon keine Ausnahme.*

Hinweis: Für einen kritischen Überblick über die Verwendung von Indikatoren für die Unsicherheit von Schätzungen siehe Imbens (2021) (Wirtschaftsnobelpreis 2021) `https://www.aeaweb.org/articles?id=10.1257/jep.35.3.157`.

Quelle: `https://xkcd.com/1781/`

Literaturverzeichnis

Andrews, D. W. K. (1993), 'Tests for parameter instability and structural change with unknown change point', *Econometrica* **61**(4), 821–56.

Bleymüller, J. (2012), *Statistik für Wirtschaftswissenschaftler*, WiST-Studienkurs, Vahlen.
URL: *https://books.google.at/books?id=xw7xpwAACAAJ*

Chow, G. C. (1960), 'Tests of Equality Between Sets of Coefficients in Two Linear Regressions', *Econometrica* **3**, 591–605.

Danilov, D. and Magnus, J. R. (2004), 'On the harm that ignoring pretesting can cause', *Journal of Econometrics* **122**(1), 27 – 46.
URL: *http://www.sciencedirect.com/science/article/pii/S0304407603002689*

Davidson, R. and MacKinnon, J. G. (2003), *Econometric Theory and Methods*, Oxford University Press, USA.

Edgeworth, F. (1885), *Observations and Statistics: an Essay on the Theory of Errors of Observation and the First Principles of Statistics*, Transactions of the Cambridge Philosophical Society.
URL: *http://books.google.at/books?id=9TesGwAACAAJ*

Fisher, R. (1925), *Statistical Methods For Research Workers*, Cosmo study guides, Cosmo Publications.
URL: *http://books.google.at/books?id=4bTttAJR5kEC*

Fisher, R. (1955), 'Statistical methods and scientific induction', *Journal of the Royal Statistical Society. Series B (Methodological)* **17**(1), pp. 69–78.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. and Kruger, L. (1990), *The Empire of Chance: How Probability Changed Science and Everyday Life (Ideas in Context)*, reprint edn, Cambridge University Press.

Gosset, W. S. (1908), 'The probable error of a mean', *Biometrika* **6**(1), 1–25. Originally published under the pseudonym "Student".
URL: *http://dx.doi.org/10.2307/2331554*

Hansen, B. E. (1997), 'Approximate asymptotic p values for structural-change tests', *Journal of Business & Economic Statistics* **15**(1), 60–67.

Imbens, G. W. (2021), 'Statistical significance, p-values, and the reporting of uncertainty', *Journal of Economic Perspectives* **35**(3), 157–74.

Johnston, J. and Dinardo, J. (1996), *Econometric Methods*, 4 edn, McGraw-Hill/Irwin.

Kahneman, D. (2013), *Thinking, Fast and Slow*, reprint edn, Farrar, Straus and Giroux.

Leeb, H. and Pötscher, B. M. (2008), 'Can one estimate the unconditional distribution of post-model-selection estimators?', *Econometric Theory* **24**(2), 338–376.

Lehmann, E. L. (1993), 'The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?', *Journal of the American Statistical Association* **88**(424), pp. 1242–1249.

Neyman, J. and Pearson, E. S. (1928*a*), 'On the use and interpretation of certain test criteria for purposes of statistical inference: Part i', *Biometrika* **20A**(1/2), 175–240.
URL: *http://www.jstor.org/stable/2331945*

Neyman, J. and Pearson, E. S. (1928*b*), 'On the use and interpretation of certain test criteria for purposes of statistical inference: Part ii', *Biometrika* **20A**(3/4), 263–294.
URL: *http://www.jstor.org/stable/2332112*

Pearson, K. (1900), 'On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling.', *Philosophical Magazine* **50**, 157–175.

Ramsey, J. B. (1969), 'Tests for specification errors in classical linear least squares regression analysis', *Journal of the Royal Statistical Society, Series B* **31**, 350–371.

Salsburg, D. (2002), *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, 1st edition edn, Holt Paperbacks.

Spanos, A. (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press.

Wooldridge, J. (2005), *Introductory Econometrics: A Modern Approach*, 3 edn, South-Western College Pub.

Zeileis, A., Leisch, F., Hornik, K. and Kleiber, C. (2002), 'strucchange: An R Package for Testing for Structural Change in Linear Regression Models', *Journal of Statistical Software* **7**(2), 1–38.
URL: *http://www.jstatsoft.org/v07/i02/*

Ziliak, S. T. (2008), 'Retrospectives: Guinnessometrics: The Economic Foundation of "Student's" t', *Journal of Economic Perspectives* **28**, 199–216.
URL: *http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.22.4.199*