

1.3 Das Testen von Hypothesen am Beispiel des Einstichproben t-Tests

Statistische Tests dienen dem Testen von Vermutungen, so genannten Hypothesen, über Eigenschaften der Gesamtheit aller Daten („Grundgesamtheit“ oder „Population“), aus denen man eine Stichprobe entnommen hat. Diesen Bereich der Statistik zählt man zur schließenden Statistik (Inferenz-Statistik, induktive Statistik), da man von einer Stichprobe auf die Grundgesamtheit, das heißt auf die unbekannt Parameter oder die unbekannt theoretische Verteilung schließt. Man unterscheidet:

- Hypothesen über die unbekannt Parameter eines bekannten Verteilungstyps (parametrische Tests).
- Hypothesen über das Symmetriezentrum der Verteilung bei unbekanntem Verteilungstyp (nichtparametrische Tests),
- Hypothesen über die Art einer Verteilung (Anpassungstests)
- Hypothesen über die Abhängigkeit von Zufallsvariablen (Unabhängigkeitstests).

Bei einem statistischen Test geht man von einer so genannten Nullhypothese „ H_0 “ aus. Die Alternativhypothese nennt man „ H_A “ oder „ H_1 “. Ziel ist es anhand statistischer Schlussweisen die Nullhypothese zu widerlegen und damit die Alternative statistisch nachzuweisen. Man berechnet dazu mit Hilfe einer Stichprobe eine Prüfgröße oder Teststatistik z (diese wird später auch mit t , t^+ , ... bezeichnet werden). Diese ist Realisierung einer Zufallsvariablen Z , deren theoretische Verteilung (z.B. Normalverteilung, t-Verteilung, usw.) man kennt, unter der Voraussetzung, dass die Nullhypothese richtig ist (kurz: „unter H_0 “). Wenn in diesem und den nächsten Kapiteln die Verteilung der Zufallsvariablen, deren Realisierung die Prüfgröße ist, spezifiziert wird, dann ist immer die Verteilung unter H_0 gemeint! Mit dem über die Stichprobe berechneten konkreten Wert z wird dann eine Entscheidung zugunsten von H_0 oder von H_1 getroffen. Wenn die Prüfgröße z extreme, d.h. eigentlich der Nullhypothese

widersprechende Werte annimmt, dann wird die Nullhypothese verworfen. Die Wahrscheinlichkeit dafür, dass solche extreme der Nullhypothese widersprechenden Werte auftreten, kann man berechnen, da man die Verteilung unter der Nullhypothese kennt. Dies ist dann der maximale Fehler, den man beim Verwerfen einer richtigen Nullhypothese macht.

Statistische Tests gibt es als einseitige oder zweiseitige Tests. Bei einem einseitigen Test zum Niveau α , wobei $0 < \alpha < 1$, zerfällt der Wertebereich von Z in zwei Teilbereiche. In einen dieser Teilbereiche fällt z bei Gültigkeit der Hypothese H_0 mit einer Wahrscheinlichkeit von $1-\alpha$, in den anderen Bereich, der auch kritischer Bereich oder Ablehnungsbereich genannt wird, fällt z mit einer Wahrscheinlichkeit α . Die von uns vor Beginn des Tests zu treffende Wahl von α ist abhängig von den Konsequenzen einer möglichen Fehlentscheidung. Meist wählt man $\alpha = 0.05 = 5\%$ oder $\alpha = 0.01 = 1\%$.

Bei einem zweiseitigen Test gibt es 3 Teilbereiche, da hier der kritische Bereich nochmals in zwei Teilbereiche zerlegt wird. Der kritische Bereich beim einseitigen oder die kritischen Bereiche beim zweiseitigen Test ergeben sich durch die Formulierung der Alternativhypothese. So wird, wie wir gleich beim t-Test sehen werden, die Nullhypothese beim einseitigen Test entweder bei zu großen oder zu kleinen Werten der Prüfgröße z verworfen, je nachdem wie die Alternativhypothese formuliert wird. Beim zweiseitigen Test wird die Nullhypothese stets bei zu großen oder zu kleinen Werten der Prüfgröße z verworfen. Liegt nun z in dem Teilbereich der zu einer Wahrscheinlichkeit kleiner oder gleich α gehört, so wird die Hypothese H_0 verworfen. Man sagt dann: "Die Alternative H_1 ist zum Niveau α signifikant". Der Fehler bei dieser Entscheidung, d.h. H_0 fälschlicherweise zu verwerfen, hat gerade eine Wahrscheinlichkeit kleiner oder gleich dem Niveau α des Tests. Man spricht auch vom α -Fehler oder vom Fehler 1. Art und nennt α auch Irrtumswahrscheinlichkeit oder Signifikanzniveau. Fällt z in den anderen Bereich, so bleibt man bei der Hypothese H_0 .

Kann man H_0 nicht verwerfen, ist diese noch nicht bewiesen, da man im praktischen Fall (wo nur die Verteilung unter H_0 bekannt ist) keine Aussage über den so genannten Fehler 2. Art β machen kann, das heißt den Fehler H_0 anzunehmen obwohl H_0 falsch ist. Im Beispiel 1 des Kapitels 1.4 berechnen wir bei einer „speziellen“ Alternative den Fehler 2. Art. Allgemein gilt für Tests, dass mit steigendem Stichprobenumfang der Fehler 2. Art abnimmt, wobei sich dann die Teststärke $1 - \beta$ (Power) vergrößert. Die Teststärke ist somit die Wahrscheinlichkeit, eine falsche Nullhypothese zu erkennen.

Die Aussage, die man eigentlich nachweisen möchte, formuliert man immer in der Alternativhypothese (soweit dies möglich ist, denn bei Anpassungstests ist dies im Allgemeinen nicht möglich). Aus diesem Grund möchte man also zum Verwerfen der Hypothese H_0 gelangen. Arbeitet man z.B. auf 5%-igem Signifikanzniveau, so kann man, falls es gelingt H_0 zu verwerfen, behaupten, dass diese Entscheidung in höchstens 5% der Fälle falsch ist (Fehler 1. Art). Man nimmt somit beim Verwerfen der Nullhypothese maximal einen Fehler von 5% in Kauf.

Bemerkung: Es wird an dieser Stelle darauf hingewiesen, dass es üblich ist, bei Anpassungstests ein Signifikanzniveau von meist 20% bzw. 25% zugrunde zu legen. Dadurch wird in diesem Fall der kritische Bereich vergrößert und man kommt eher zum Verwerfen der Nullhypothese. Dies ist erforderlich, da man bei Anpassungstests die Nullhypothese gerne nachweisen würde. Daher könnte man sagen: Wenn man trotz dieses hohen Fehlerniveaus nicht zum Verwerfen der Nullhypothese kommt, spricht nichts gegen diese. Man kennt aber in diesem Fall trotzdem nicht den Fehler 2. Art.

Stellvertretend für die zahlreichen Tests, die man in der Statistik kennt, wollen wir den t-Test (für eine Stichprobe) besprechen und vorführen. Den t-Test gibt es als folgende Varianten: den t-Test für eine einzelne Stichprobe (Einstichproben t-Test, one-sample t-test), für zwei

verbundene abhängige Stichproben (paired t-test) und für zwei unabhängige Stichproben (Zweistichproben t-Test, two-sample t-test) für gleiche und ungleiche Varianzen.

Voraussetzung für die Anwendung des t-Tests ist, dass die Stichprobe aus Realisierungen von unabhängig und identisch normalverteilten Zufallsvariablen (mit dem Erwartungswert μ und der Varianz σ^2) besteht, denn nur dann ist die Prüfgröße des Tests Realisierung einer (unter H_0) t-verteilten Zufallsvariablen. Beim Zweistichproben t-Test gilt diese Voraussetzung jeweils für die erste und zweite Stichprobe, wobei die Parameter der Normalverteilung bei der ersten Stichprobe natürlich nicht notwendigerweise die gleichen sein müssen wie bei der zweiten Stichprobe. Ist diese Voraussetzung nicht erfüllt, so ist der Test nicht anwendbar. Es wäre also zunächst ein Test auf Normalverteilung durchzuführen.

Beim t-Test für eine Stichprobe geht es darum, Hypothesen über den Erwartungswert μ anhand einer Stichprobe zu überprüfen. Dabei ist die Varianz ebenso wie der Erwartungswert der zugrunde liegenden Normalverteilung unbekannt. Der entsprechende Test bei bekannter Varianz heißt Gaußtest, wobei die Prüfgröße hier $N(0, 1)$ -verteilt bzw. standardnormalverteilt (d.h. normalverteilt mit Erwartungswert 0 und Varianz 1) wäre.

Der Erwartungswert μ wird bei den Hypothesen des t-Tests mit einem konkret festgelegten Wert μ_0 verglichen. Ein mögliches einseitiges Testproblem wäre, dass die Nullhypothese

$$H_0: \mu \leq \mu_0$$

gegen die Alternative

$$H_1: \mu > \mu_0$$

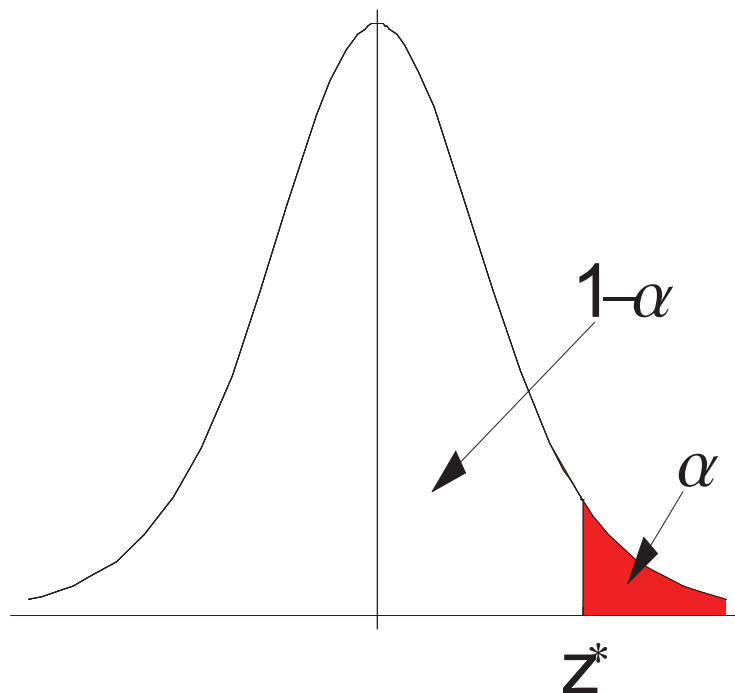
getestet wird.

Die Prüfgröße z des t-Tests, die (unter H_0) Realisierung einer t-verteilten Zufallsvariablen Z mit $n - 1$ Freiheitsgraden ist, wird wie folgt berechnet:

$$z = \sqrt{n} \frac{\bar{x} - \mu_0}{s}$$

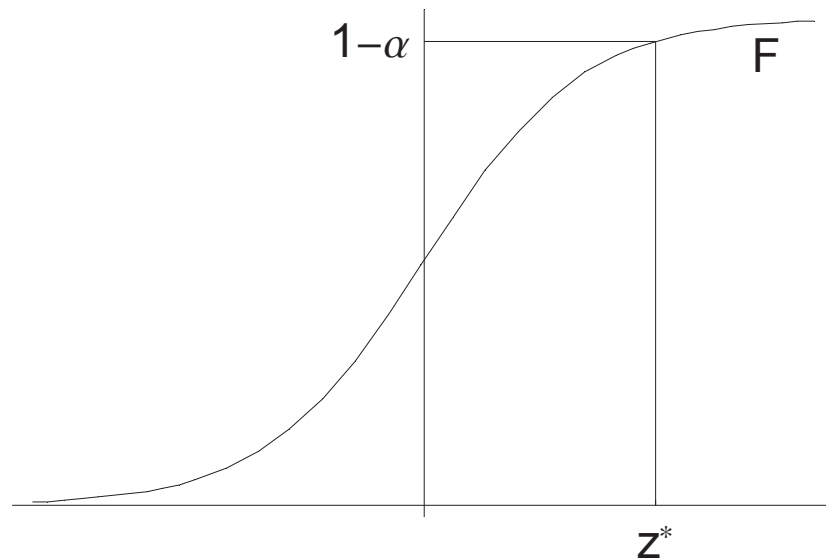
Es ist s die empirische Standardabweichung, \bar{x} das arithmetische Mittel und n der Stichprobenumfang (siehe Erläuterung des Outputs).

Mit den obigen Hypothesen wird die Nullhypothese verworfen, falls die Prüfgröße z einen zu großen Wert aufweist, d.h. falls diese größer als das (oder gleich dem) $(1-\alpha)$ -Quantil z^* der t-Verteilung mit $n - 1$ Freiheitsgraden ist. Es ist $z^* = F_{t_{n-1}}^{-1}(1-\alpha)$. Dabei ist $F_{t_{n-1}}$ die Verteilungsfunktion der t-Verteilung mit $n - 1$ Freiheitsgraden.



Die obige Grafik zeigt die Dichtefunktion einer t-Verteilung zusammen mit dem kritischen Wert z^* und den Flächen mit dem Wert α beziehungsweise $1 - \alpha$, die die Wahrscheinlichkeiten repräsentieren, dass eine Realisierung der Zufallsvariable Z in das entsprechende Intervall auf der x-Achse fällt. Die nächste Grafik zeigt diesen Sachverhalt anhand der Verteilungsfunktion F einer t-Verteilung. Hier wird der Zusammenhang zwischen den Quantilen und den kritischen Werten deutlich.

Es gilt $F(z^*) = 1 - \alpha$ bzw. $F^{-1}(1 - \alpha) = z^*$.



Demnach wird die Nullhypothese verworfen, wenn:

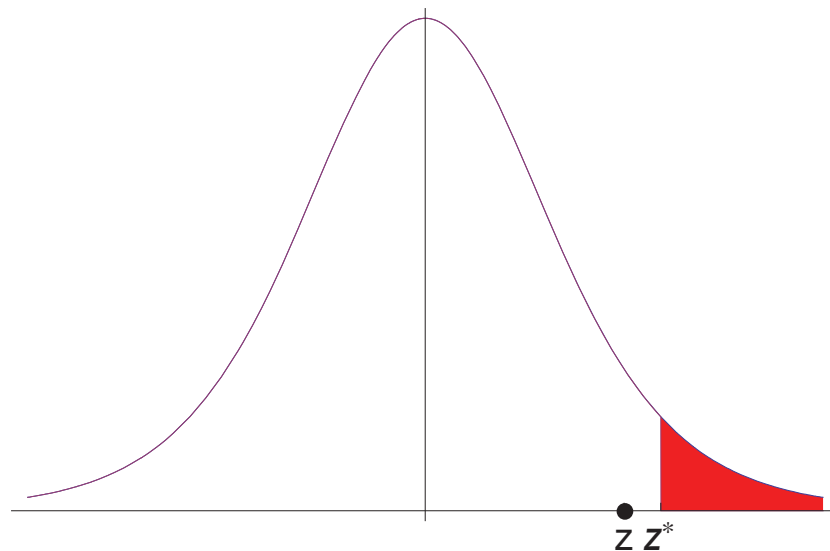
$$z \geq z^* = F_{t_{n-1}}^{-1}(1 - \alpha) \Leftrightarrow F_{t_{n-1}}(z) \geq 1 - \alpha \Leftrightarrow \alpha \geq 1 - F_{t_{n-1}}(z) = p\text{-Wert}$$

Dies bedeutet, dass die Nullhypothese verworfen wird, falls die Prüfgröße z größer als das (oder gleich dem) $(1-\alpha)$ -Quantil z^* der t-Verteilung mit $n - 1$ Freiheitsgraden z^* ist, was äquivalent dazu ist, dass der p-Wert kleiner als das (oder gleich dem) gewählte

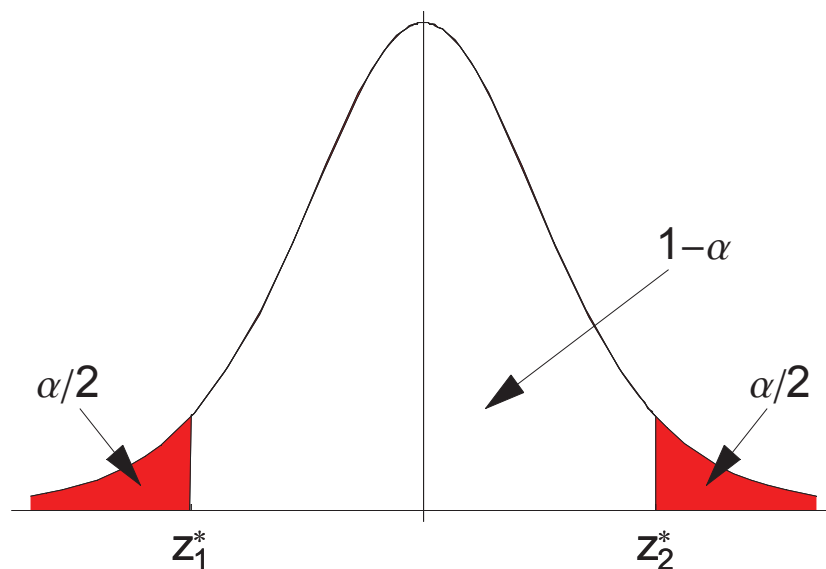
Signifikanzniveau α ist. Der p-Wert wird von den meisten Statistikprogrammpaketen ausgegeben, wobei hier aber meist der zweiseitige t-Test durchgeführt wird.

Bemerkung: Da die t-Verteilung eine stetige Verteilung ist, können in der oberen Gleichung zur Herleitung des p-Wertes auch alle „ \geq “-Zeichen durch „ $>$ “ ersetzt werden, denn hier gilt $P(Z = z) = 0$. Dies gilt nicht bei diskreten Verteilungen!

Im folgenden ist eine Grafik zu sehen, in der die Fläche zwischen dem kritischen Bereich (dies ist das Intervall $[z^*, \infty)$ auf der x-Achse) und dem Graph der Dichtefunktion eingefärbt ist. Der Punkt auf der x-Achse soll die Prüfgröße z darstellen, die man aus einer konkreten Stichprobe vom Umfang n gewonnen hat. Nach der unteren Grafik käme man nicht zum Verwerfen der Nullhypothese, da $z < z^*$ ist, also z in einen „mit der Nullhypothese nicht verträglichen“ Bereich fällt, in dem eine Realisierung von Z mit der Wahrscheinlichkeit $1-\alpha$ auftritt. Je nachdem wie groß man α wählt, wird der kritische Bereich größer (für größere α) oder kleiner.



Beim zweiseitigen Test gibt es, wie bereits beschrieben, zwei kritische Bereiche, für die gilt, dass die Fläche zwischen Kurve und den beiden kritischen Bereichen insgesamt α ist.



Da die Dichtefunktion einer t-Verteilung symmetrisch ist, genügt es hier, einen kritischen Wert z^* anstelle von zwei kritischen Werten z_1^* und z_2^* zu berechnen. Denn hier gilt $-z_1^* = z_2^* = z^*$.

Beim zweiseitigen Test wird die Hypothese

$$H_0: \mu = \mu_0$$

gegen

$$H_1: \mu \neq \mu_0$$

getestet.

Wir verwerfen die Nullhypothese zugunsten der Alternativhypothese, falls die Prüfgröße z „zu große“ oder „zu kleine“ Werte annimmt, d.h. wir kommen zum Verwerfen, falls z größer als das (oder gleich dem) $(1-\alpha/2)$ -Quantil z_2^* oder kleiner als das (oder gleich dem) $\alpha/2$ -Quantil

z_1^* der t-Verteilung mit $n - 1$ Freiheitsgraden ist. Hier würde die Nullhypothese also verworfen werden, falls gilt:

$$z \leq z_1^* = F_{t_{n-1}}^{-1}(\alpha/2) \quad \text{oder} \quad z \geq z_2^* = F_{t_{n-1}}^{-1}(1 - \alpha/2)$$

Dieses Kriterium ist wegen der beschriebenen Symmetrie der t-Verteilung äquivalent zu

$$|z| \geq z^* = F_{t_{n-1}}^{-1}(1 - \alpha/2).$$

Um auf den p-Wert zu kommen, der von vielen Statistiksystemen ausgegeben wird, kann man die obere Gleichung durch Äquivalenzumformungen auf die folgende Form bringen:

$$F_{t_{n-1}}(|z|) \geq 1 - \alpha/2 \Leftrightarrow \alpha \geq 2(1 - F_{t_{n-1}}(|z|)) = p - \text{Wert}$$

Also wird die Nullhypothese verworfen, wenn gilt: p-Wert $\leq \alpha$.

Sie können mit diesem p-Wert und der Prüfgröße aus dem zweiseitigen t-Test auch einen einseitigen t-Test durchführen. Der p-Wert ist zu halbieren, da beim einseitigen t-Test nicht das $(1-\alpha/2)$ -Quantil, sondern das $(1-\alpha)$ - bzw. α -Quantil der entsprechenden t-Verteilung verwendet wird, je nachdem wie die Alternativhypothese formuliert wurde.

Ist dann die Hälfte des p-Wertes aus dem zweiseitigen t-Test kleiner als das gewählte Signifikanzniveau (oder gleich diesem) und gilt für die Prüfgröße $z < 0$ (hier muss natürlich auch das Vorzeichen der Prüfgröße z beachtet werden, da in der Formel zur Berechnung des zweiseitigen p-Wertes, wie oben zu sehen ist, nur der Betrag von z verwendet wird), so kann die einseitige Nullhypothese

$H_0: \mu \geq \mu_0$ zugunsten der Alternativhypothese $H_1: \mu < \mu_0$

verworfen werden.

Ist die Hälfte des p-Wertes aus dem zweiseitigen t-Test kleiner als das gewählte Signifikanzniveau (oder gleich diesem) und gilt für die Prüfgröße $z > 0$, so kann die einseitige Nullhypothese

$H_0: \mu \leq \mu_0$ zugunsten der Alternativhypothese $H_1: \mu > \mu_0$

verworfen werden.

Achtung: Ein „sauberes“ Vorgehen verlangt, dass man vor der Interpretation des zweiseitigen p-Wertes sich für einen einseitigen oder zweiseitigen t-Test entscheidet. Hat man zuerst einen zweiseitigen t-Test durchgeführt und sich nach der Interpretation des p-Wertes bereits für eine Hypothese entschieden, so sollte man sich erst einen neuen Datensatz besorgen, mit dem man dann zusätzlich den einseitigen t-Test durchführt.

Kommen wir nun zu unserem Beispiel. Hier möchten wir die folgenden Hypothesen testen:

$H_0: \mu = 175$

gegen

$H_1: \mu \neq 175$

In unserem Beispiel verwenden wir die folgenden Daten:

v1
167
163
155
167
161
177
173
179

Wenn Sie diese eingeben und dann →**Univariate Statistik** wählen, können Sie neben dem Button t-Test den Wert für μ_0 eintragen, also hier 175. Danach können Sie auf →**t-Test** klicken und erhalten den folgenden Output:

Einstichproben t-Test

H0: $\mu = 175$
 gegen
 H1: $\mu < 175$

Stichprobenumfang n	8
arithmetisches Mittel	167.75
geschätzte Varianz	67.357142857143
geschätzte Standardabweichung	8.2071397488493
Prüfgröße t (Freiheitsgrade der t-Verteilung: 7)	-2.4985679885961
p-Wert	0.0411

Im Folgenden erklären wir den durchgeführten t-Test. Es gilt:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 167,75$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \approx 8,20714$$

$$\mu_0 = 175$$

$$t = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \approx -2,49857$$

$$p\text{-Wert} = 2(1 - F_{t_{n-1}}(|t|)) \approx 0,041078$$

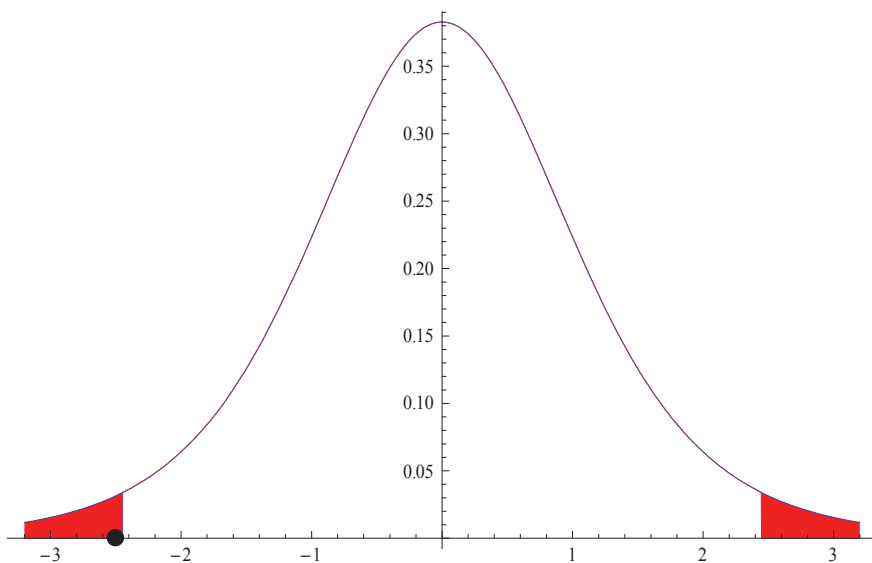
Aufgrund des p-Wertes von 0,041.. ($\leq 0,05 = \alpha$) kann die Nullhypothese ($H_0: \mu = 175$) zugunsten der Alternativen ($H_1: \mu \neq 175$) verworfen werden, wenn man ein Signifikanzniveau von 5% verwendet. Man kann nun sagen, dass der Erwartungswert μ sich signifikant vom Wert 175 unterscheidet. Dabei nehmen wir maximal einen Fehler von 5% in Kauf. Der p-Wert ist somit auch eine Untergrenze für das Signifikanzniveau, ab dem man noch die Nullhypothese verwerfen kann. Wie wir bereits beschrieben haben, verlangt aber ein „sauberes Vorgehen“ zuerst die Wahl des Signifikanzniveaus, bevor der p-Wert betrachtet wird.

Wir wollen die Lage der Prüfgröße bezüglich des kritischen Bereichs zusammen mit der Dichtefunktion der t-Verteilung (mit $n - 1 = 7$ Freiheitsgraden) in einer Grafik darstellen. Dazu berechnen wir das 0,025-Quantil (da $\alpha = 0,05$).

Es gilt

$$z_1 = F_{t_{n-1}}^{-1}(\alpha/2) \approx -2,44691$$

und somit ist $z_2 \approx 2,44691$.



Wie zu sehen ist, liegt die Prüfgröße (als Punkt auf der x-Achse dargestellt) mit einem Wert von $\approx -2,49857$ im kritischen Bereich, womit die Nullhypothese (wie bereits beschrieben) verworfen werden kann. Der p-Wert ist hier mit 0,0410 kleiner als unser übliches Signifikanzniveau $\alpha = 0,05$ (= 5%). Somit ist der Erwartungswert signifikant vom Wert 175 verschieden.

Eine Kurze Zusammenfassung zum Thema „Tests“:

Wie in diesem Kapitel gezeigt wurde, genügt es bei den Tests, die man mit Statistiksystemen durchführen kann

- 1) Die Voraussetzungen des Testes zu kennen, damit die bei der Berechnung des p-Wertes zu Grunde gelegte Verteilung richtig ist.
- 2) Die Nullhypothese und Alternativhypothese des Testes zu kennen.

Danach wählt man ein Signifikanzniveau α und vergleicht dieses mit dem p-Wert. Ist der p-Wert kleiner oder gleich α , so kann die Nullhypothese (H_0) zugunsten der Alternativhypothese (H_1) verworfen werden. Der p-Wert ist somit das kleinste Signifikanzniveau, mit dem man H_0 gerade noch verwerfen könnte. Es ist dabei zu beachten, dass das System den p-Wert auf 4 Nachkommastellen rundet. Damit kann auch, falls eine Null als p-Wert ausgegeben wird, nicht gleichzeitig auf jedem Signifikanzniveau α die Nullhypothese verworfen werden, allerdings auf jedem gängigen Signifikanzniveau (z.B. 10%, 5% oder 1%).

Umsetzung des t-Tests in SAS:

```
data dat1;
input x;
cards;
167
163
155
167
161
177
173
179
run;

proc univariate data = dat1 mu0=175;
var x;
run;
```

SAS-Output zur Prozedur UNIVARIATE

Die Prozedur UNIVARIATE
Variable: x

Momente			
N	8	Summe Gewichte	8
Mittelwert	167.75	Summe Beobacht.	1342
Std.abweichung	8.20713975	Varianz	67.3571429
Schiefe	-0.0441899	Kurtosis	-0.8864415
Unkorr. Qu.summe	225592	Korr. Quad.summe	471.5
Variationskoeff.	4.89248271	Stdfh. Mittelw.	2.90166209

Grundlegende Statistikmaße			
Lage		Streuung	
Mittelwert	167.7500	Std.abweichung	8.20714
Median	167.0000	Varianz	67.35714
Modalwert	167.0000	Spannweite	24.00000
		Interquartilsabstand	13.00000

Tests auf Lageparameter: $\mu_0=175$			
Test	Statistik	p-Wert	
Studentsches t	t -2.49857	Pr > t 	0.0411
Vorzeichen	M -2	Pr >= M 	0.2891
Vorzeichen-Rang	S -13.5	Pr >= S 	0.0703

Alternativ ist auch die Umsetzung mit der Prozedur TTEST empfehlenswert

```
proc ttest data = dat1 H0=175;  
  var x;  
run;
```

SAS-Output zur Prozedur TTEST

Die Prozedur TTEST
Variable: x

N	Mittelwert	Std.abw.	Std.fehler	Minimum	Maximum
8	167.8	8.2071	2.9017	155.0	179.0

Mittelwert	95% CL Mittelwert	Std.abw.	95% CL Std Dev
167.8	160.9	174.6	8.2071 5.4263 16.7038

DF	t-Wert	Pr > t
7	-2.50	0.0411