Wie funktionieren statistische Tests?

Henry Haustein

In den letzten Jahren sind rund 35% der Studenten durch die Statistik-Prüfung gefallen. Während Corona sank diese Quote auf rund 15%. Ist diese Änderung zufällig oder war das Online-Lehre des Lehrstuhls so gut, dass plötzlich so viele Studenten die Inhalte verstanden haben[1]?

Genau solche Fragen können statistische Tests beantworten. Wir brauchen dazu

- 2 Hypothesen, die sich gegenseitig ausschließen
- eine Teststatistik
- einen kritischen Wert

Die Hypothesen dienen dazu, den Test zu formalisieren und sich bewusst zu machen, was man eigentlich testen möchte. Durch eine sprachliche Beschreibung ist dies nicht immer präzise möglich. Aufgrund der stochastischen Hintergründe eines Tests kann man die Nullhypothese nur annehmen oder nicht annehmen (aber nicht ablehnen!), weswegen man die zu testende Aussage immer als Alternativhypothese auffasst (*Testen Sie, ob er Mittelwert kleiner als 15 ist!* $\Rightarrow H_1 : \mu < 15$ und damit $H_0 : \mu \geq 15$). Sollen wir auf Gleichheit testen, so ist Gleichheit immer in der Nullhypothese und die Ungleichheit immer in der Alternativhypothese.
Im obigen Beispiel wollen wir ja testen, ob bei einer Durchfallquote von 35% es sein kann, dass nur 15% durchgefallen sind. Ist also die Durchfallquote nur 15%? Die Hypothesen sind also

$$H_0 : \hat{p} = 0.15$$
$$H_1 : \hat{p} \neq 0.15$$

Die Teststatistik hängt vom Test ab, der durchgeführt werden soll und steht immer in der Formelsammlung. Für unser Beispiel oben ist die Teststatistik (ich habe einfach mal $n = 400$ angenommen)

$$\begin{aligned} T &= \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \\ &= \frac{0.15 - 0.35}{\sqrt{\frac{0.35 \cdot 0.65}{400}}} \\ &= -8.3863 \end{aligned}$$

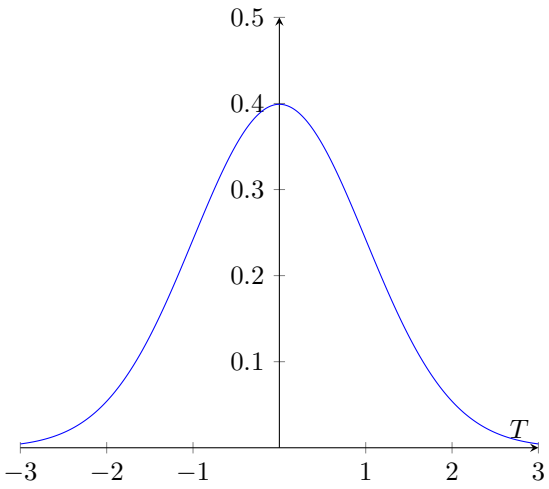
[1] Betrug während der Klausur ist natürlich völlig ausgeschlossen...

Welche Bedeutung hat die Teststatistik? Sie ist ein Maß für den Abstand zwischen dem, was wir erwarten (35% Durchfallquote) und dem, was wir beobachten (15% Durchfallquote). Eine sehr hohe oder sehr niedrige Teststatistik deutet also darauf hin, dass Erwartung und Beobachtung auseinander fallen und ist damit ein Zeichen dafür, dass die Nullhypothese zweifelhaft ist.
Aber warum ist dann die Berechnung der Teststatistik so kompliziert? Die Teststatistik

$$\begin{aligned}\tilde{T} &= \hat{p} - p \\ &= 0.15 - 0.35 \\ &= -0.2\end{aligned}$$

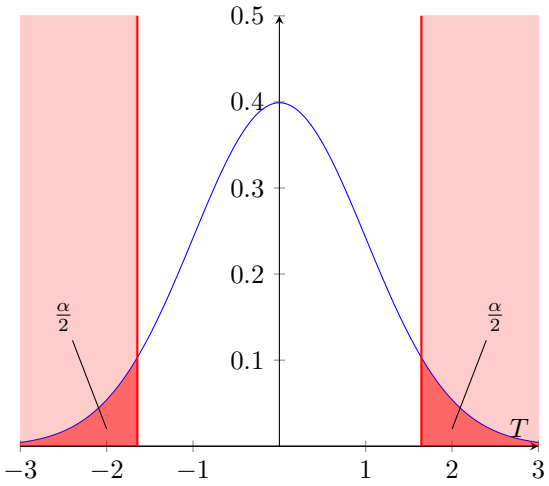
ist ja auch ein Maß für den Abstand zwischen Erwartung und Beobachtung.
Das hat mit den kritischen Werten zu tun. Die Teststatistik T hat die weitere Eigenschaft, dass sie approximativ standardnormalverteilt ist. Diese Eigenschaft brauchen wir, wenn wir die kritischen Werte einfach und schnell ermitteln wollen. Deren Berechnung ist noch viel komplizierter[2], aber sie hängen (fast) nicht mehr von den Parametern des Tests ab.

Welche Rolle spielen die kritischen Werte? Sie geben die Grenze an, bis wann eine Abweichung vom Erwarteten noch ok ist. Die unsere Teststatistik ist approximativ standardnormalverteilt, also



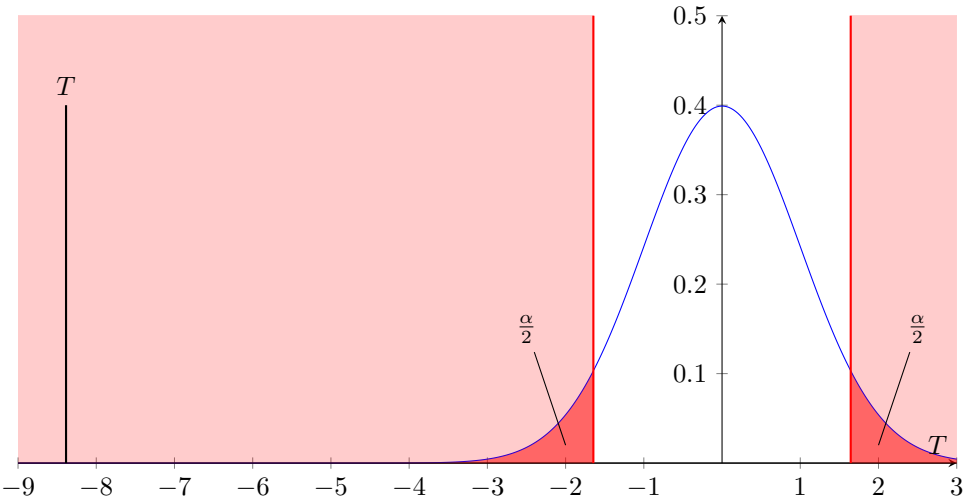
Bei einem Test brauchen wir auch immer eine Irrtumswahrscheinlichkeit α. Sie gibt an, wie stark wir Abweichungen vom Erwarteten tolerieren. α heißt deswegen Irrtumswahrscheinlichkeit, weil es angibt, wie wahrscheinlich es ist, dass die Abweichung so groß ist, dass wir sie nicht mehr tolerieren, aber das Erwartete immer noch stimmt.

[2] Hier muss man Integrale berechnen, die keine Stammfunktion haben und häufig braucht man noch die Gamma-Funktion. Beides sind Rechenoperationen, die heutzutage nur effizient ein Computer lösen machen kann, aber als diese Tests entwickelt wurden gab es noch keine Computer und die Berechnung war sehr aufwendig. Die kritischen Werte wurden dann nur einmal berechnet und die Ergebnisse wurden weitergegeben. Klar, dass man lieber eine etwas aufwendigere Teststatistik berechnet als die kritischen Werte jedes mal neu zu bestimmen.



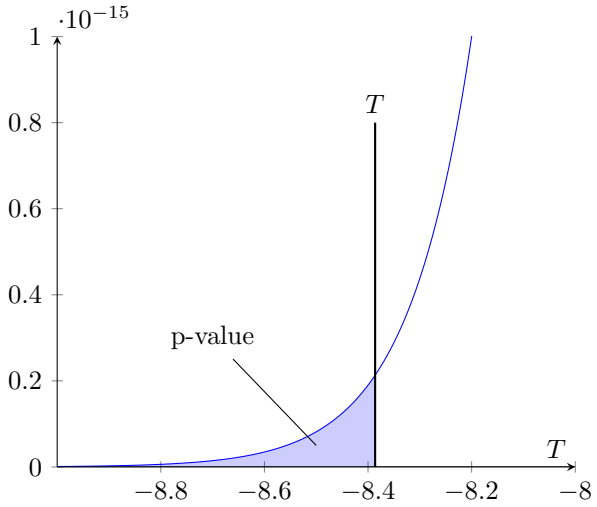
Die Grafik ist für einen zweiseitigen Test. Es gibt auch einseitige Tests, aber da ist der Ablehnungsbereich nur entweder links oder rechts, die Grenzen sind damit andere. Sonst ändert sich nichts, weswegen ich im Folgenden immer von zweiseitigen Tests ausgehen werde.

Der rote Bereich ist der Ablehnungsbereich, das heißt sollte die Teststatistik einen Wert annehmen, der in diesem Bereich liegt, so nimmt man H_0 nicht an, sondern entscheidet sich für die Alternativhypothese. Der Ablehnungsbereich ist so konstruiert, dass die Fläche unter der Verteilung der Teststatistik im Ablehnungsbereich genau α ist. Die Grenzen des Ablehnungsbereiches sind die kritischen Werte.
Für unser Beispiel bedeutet dies, dass das Bild so aussieht (mit $\alpha = 10\%$ ergeben sich die kritischen Werte zu ± 1.6449):



Unsere Teststatistik T ist im Ablehnungsbereich, wir nehmen also H_0 nicht an und entscheiden uns für H_1.

Besonders im angelsächsichen Raum wird häufig nicht mit kritischen Werten, sondern mit p-values gearbeitet. Das ist die Fläche unter der Dichtefunktion bis zur Teststatistik T:



In unserem Beispiel beträgt dieser $\Phi(-8.3863) = 2.50841 \cdot 10^{-17}$. Sind die p-values sehr klein, so liegt T weit weg vom Erwarteten und damit säht das Zweifel an H_0. Hat eine Teststatistik einen p-value von $> \frac{\alpha}{2}$, so liegt T im Nicht-Ablehnungsbereich und damit wird H_0 angenommen.

Zusammenhang mit Konfidenzintervallen

Statistische Tests und Konfidenzintervalle hängen stark zusammen und die Aussage, die sie machen, ist die selbe. Man könnte den Nicht-Ablehnungsbereich als Konfidenzintervall für die Teststatistik bezeichnen. Falls die Teststatistik in dieses Konfidenzintervall fällt, so wird H_0 angenommen.