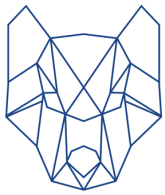
ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES



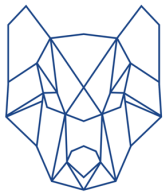
Informatik
& Security

Hypothesentests

Angewandte Statistik

Dipl.-Ing. Hubert Schölnast, BSc

Stand: 3. Juli 2023



Inhaltsverzeichnis

1 Hypothesentests

1.1 Hypothese und Theorie

In der Wissenschaft ist eine Hypothese eine wohlbegründete, aber noch unbewiesene Annahme über die genaue Art des Zusammenhangs zwischen einer Ursache und der sich daraus ergebenden Wirkung. Sehr oft gibt es mehrere einander widersprechende Hypothesen, die versuchen, denselben Zusammenhang zu erklären. Aus diesen Hypothesen kann man bestenfalls einige als falsch entlarven, es ist aber prinzipiell unmöglich, zu beweisen, dass eine Hypothese den Zusammenhang korrekt erklärt. Das Widerlegen einer falschen Hypothese kann gelingen, wenn eine Hypothese Vorhersagen macht, die über den ursprünglichen Zusammenhang hinausgehen. Wenn sich diese Vorhersagen als falsch erweisen, ist die Hypothese zu verwerfen. Wenn nicht nachgewiesen werden kann, dass die Vorhersagen falsch sind, bleibt die Hypothese als möglicher Kandidat für die korrekte Beschreibung der Realität im Rennen.

Erst wenn viele Hypothesen formuliert, getestet und verworfen worden sind, kann man davon ausgehen, dass jene Hypothesen, die allen Bemühungen, sie zu widerlegen standgehalten haben, möglicherweise die Wahrheit beschreiben. Erst wenn die Gemeinschaft der Wissenschaftler*innen gemeinsam zu dem Schluss kommt, dass eine bestimmte Hypothese diesen Grad an Verlässlichkeit aufweist, erst dann nennt man die betreffende Hypothese eine Theorie.

1.1.1 Beispiel Gravitation:

Dass die meisten Dinge nach unter fallen, wenn man sie loslässt, ist eine Erkenntnis, die seit jeher den Menschen so vertraut ist, dass man es in jenen Zeiten, in denen man versuchte, die Welt durch das Wirken von Göttern zu beschrieben, nicht für notwendig hielt, sie durch ein solches göttliches Wirken zu begründen. Weder bei den Ägyptern noch bei den Mesopotamiern, nicht bei den German, nicht bei den Azteken und auch nicht bei den Ureinwohnern Australiens gibt es einen Gott der Gravitation. Auch in der Bibel wird die Kraft, die Dinge nach unten fallen lässt, nicht erklärt.

Aber die Menschen dachten trotzdem über die Ursache dieser allgegenwärtigen nach unten gerichteten Kraft nach, und kamen zu der Auffassung, dass »unten« der natürliche Ort aller Dinge ist, und dass alle Dinge daher nach unten streben, wenn man ihnen die Möglichkeit dazu gibt. Bei Wolken und Vögeln war das anders, weil deren natürlicher Ort »oben« war.

Sogar der Astronom Nikolaus Kopernikus schrieb im Jahr 1543 über die Schwerkraft: *»Ich bin wenigstens der Ansicht, dass die Schwere nichts Anderes ist, als ein von der göttlichen*

Vorsehung des Weltenmeisters den Theilen eingepflanztes, natürliches Streben, vermöge dessen sie dadurch, dass sie sich zur Form einer Kugel zusammenschließen, ihre Einheit und Ganzheit bilden. Und es ist anzunehmen, dass diese Neigung auch der Sonne, dem Monde und den übrigen Planeten innewohnt.«

Johannes Keppler (um 1600) und Galileo Galilei (um 1640) teilten im Wesentlichen diese Meinung, steuerten aber auch Ideen bei, die dazu betrugen, das Ausmaß dieser Kraft abzuschätzen.

Erst im Jahr 1687 begann Isaac Newton von *Massen* zu sprechen die aufeinander eine anziehende Kraft ausüben und er war auch der erste, der die Gravitation mit einer exakten Formel beschrieb. Diese *Hypothese* Newtons wurde zur Newtonschen Gravitations*theorie* und löste die älteren Gravitationshypothesen ab, und zwar in dem Sinn, dass durch Newtons Theorie die alten Hypothesen als Näherungen der neuen Theorie anzusehen waren.

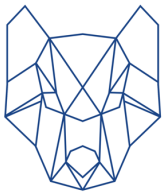
Abgelöst wurde diese Theorie erst 1916 durch Einsteins Allgemeine Relativitätstheorie, welche die Gravitation als eine Wirkung der Verformung des Raumes beschreibt, wobei diese Verformung wiederum eine Folge der Anwesenheit von Masse ist. Die Newtonsche Gravitation erscheint nun selbst als Näherung der Einstein'schen Gravitation.

Die Allgemeine Relativitätstheorie und die Quantenmechanik sind heute die Basistheorien der gesamten Physik, und alle Maschinen, die wir heute bauen, beruhen darauf, dass wir diesen beiden Theorien vertrauen können. Aber es ist schon seit fast hundert Jahren bekannt, dass sie einander in bestimmten Bereichen eklatant widersprechen (nämlich bei der Beschreibung schwarzer Löcher und bei der Beschreibung der Zustände während und kurz nach dem Urknall). Mindestens eine der beiden Theorien (sehr wahrscheinlich aber beide) sind daher selbst auch nur Näherungen einer noch nicht verfügbaren besseren Theorie.

1.2 Nullhypothese und Alternativhypothese

Eine Möglichkeit, eine Hypothese zu widerlegen, stellt die Statistik in Form von Hypothesentests bereit. Dabei wird eine Hypothese in Form einer Gleichung oder Ungleichung formuliert. Eine andere Hypothese beschreibt genau das Gegenteil. Jene der beiden Hypothesen, die entweder die ursprüngliche ist, oder der man im Zweifel lieber vertraut, bekommt die Rolle der Nullhypothese (H_0) zugewiesen. Ihre Verneinung ist die Alternativhypothese (H_1). Welche der beiden Hypothesen man als Nullhypothese und welche man als Alternativhypothese betrachten will, hängt von der Aufgabenstellung ab. In der Statistik hat man sich dabei auf folgende Aufteilung geeinigt:

- Nullhypothese: Es ist alles so, wie es sein soll, oder es bleibt alles so, wie es bisher war.
- Alternativhypothese: Es ist nicht so wie es sein soll oder wie es bisher war.



1.2.1 Teststatistik und p-Wert

Die Teststatistik ist eine reelle Zahl, die man aus den Daten der Stichprobe berechnet, wobei die genaue Rechenvorschrift vom gewählten Testverfahren abhängt. Diese Zahl ist umso größer, je weniger die analysierte Stichprobe der wahrscheinlichsten Stichprobe entspricht, die zu erwarten wäre, wenn die Nullhypothese gilt.

Im Fall eines t-Tests heißt diese Teststatistik »t-Wert«. Wenn Sie einen Chi-Quadrat-Test machen, heißt diese Zahl »Chi-Quadrat-Statistik«. Andere Testverfahren haben andere Namen dafür, und in jedem Verfahren gibt es eine andere Rechenvorschrift dafür. Details finden Sie bei der Beschreibung der jeweiligen Testverfahren.

Die Teststatistik ist aber nur ein Zwischenergebnis. In einem weiteren Schritt wird daraus der p-Wert berechnet, der weiter unten noch genauer erklärt wird.

Sehr oft finden Sie vor allem in älteren Lehrbüchern die Formulierung, dass man den p-Wert aus einer vorgefertigten Tabelle auslesen soll. Wenn Sie ein Statistik-Programm oder eine Statistik-Bibliothek einer Programmiersprache verwenden, macht das aber der Computer für Sie. Mathematisch ist der Schritt von der Teststatistik zum p-Wert nämlich bei fast allen Testverfahren jener Schritt, der die komplexesten Berechnungen erfordert. Allerdings wird dabei sehr oft mit genau den gleichen oder mit sehr ähnlichen Werten gerechnet, daher machte es in der Vergangenheit Sinn, diese aufwändigen Rechnungen von sogenannten »Computern« ausführen zu lassen und die Ergebnisse dann in Form von Tabellenbüchern zu publizieren. »Computer« war bis in die Mitte des 20. Jahrhunderts im englischsprachigen Raum die Berufsbezeichnung für Menschen (fast ausschließlich Frauen), die dafür bezahlt wurden, langweilige, aber wichtige Berechnungen durchzuführen.

Vereinfacht gesagt ist der p-Wert eine Zahl, die angibt, wie überraschend bzw. wie unwahrscheinlich es ist, die gerade vorliegende Stichprobe zu erhalten, falls die Nullhypothese zutrifft.

Der p-Wert und die Teststatistik sind beides Maße für die **Extremität** der Stichprobe. Die Stichprobe, die man mit der größten Wahrscheinlichkeit erwartet, falls die Nullhypothese zutrifft, ist die am wenigsten extreme Stichprobe. Ihr p-Wert ist 1 und ihre Teststatistik hat den kleinstmöglichen Wert (meist 0). Je mehr eine Stichprobe von dieser wahrscheinlichsten Stichprobe abweicht, desto extremer ist sie, und desto kleiner ist der p-Wert.

Wenn der p-Wert unterhalb einer vorher festgelegten Grenze liegt (wenn also die Stichprobe zu extrem ist), wird man sich entschließen zu glauben, dass der Grund dafür, dass man diese Stichprobe erhalten hat, darin besteht, dass die Nullhypothese nicht zutrifft. Diese Grenze nennt man das **Signifikanzniveau**.

1.2.2 Beibehalten, verwerfen und annehmen

Man sollte sich aber vor der Durchführung eines statistischen Hypothesentests darüber im Klaren sein, dass das Ergebnis des Hypothesentests eines dieser beiden Ereignisse ist:

1. Der p-Wert ist groß genug: Die Hinweise, die gegen die Nullhypothese sprechen, reichen nicht aus, um sie zu verwerfen. Sie könnte gültig sein.
   H_0 wird *beibehalten*
   H_1 wird *nicht angenommen*
2. Der p-Wert ist zu klein: Die Nullhypothese erweist sich mit hoher Wahrscheinlichkeit als falsch, daher sieht man sich gezwungen, davon auszugehen, dass die Alternativhypothese zutrifft.
   H_0 wird *abgelehnt* bzw. verworfen
   H_1 wird *angenommen*

Mit den beiden Hypothesen passen also unterschiedliche Dinge:

- H_0: beibehalten oder ablehnen
- H_1: annehmen oder nicht annehmen

Man geht also vor dem Test immer davon aus, dass die Nullhypothese stimmt, und weicht von dieser Annahme nur dann ab, wenn schwerwiegende Gründe dafür vorliegen.

Beachte, dass es kein Szenario gibt, in dem die Nullhypothese bewiesen wird. Entweder man verwirft sie, weil die Daten nicht zu ihr passen, oder man behält sie bei, weil man keine ausreichenden Gegenargumente hat.

Beachte auch, dass die Alternativhypothese im Fall 2 nur »notgedrungen« angenommen wird (man würde ja eigentlich lieber bei der Nullhypothese bleiben), dass es aber im Fall 1 keine konkrete Aussage über die Alternativhypothese gibt. Sie gilt im Fall 1 keineswegs als ausgeschlossen. Es gibt nur nicht genügend Argumente, die für sie sprechen.

1.3 Ein- und zweiseitige Tests

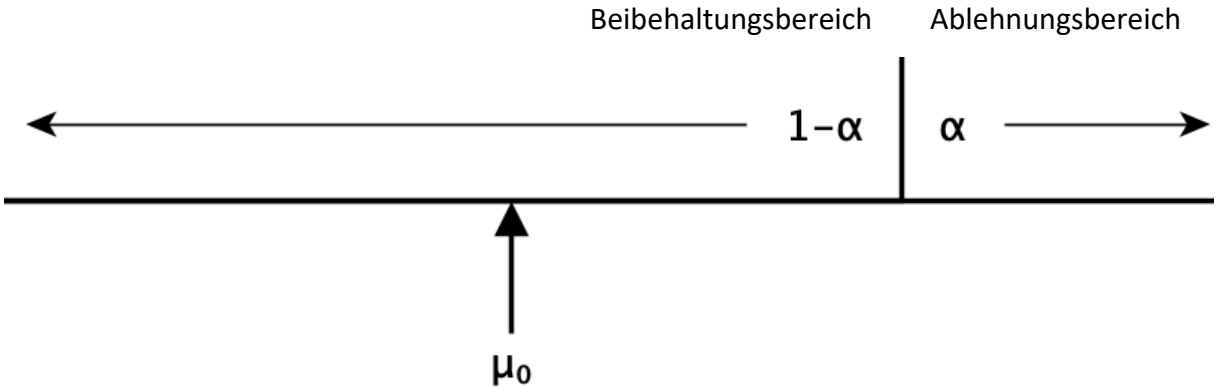
1.3.1 Einseitige Tests (Bereichstests)

Beim einseitigen Test behauptet die Nullhypothese, dass eine bestimmte statistische Kennzahl (meist ein Lagemaß, z.B. das arithmetische Mittel, manchmal aber auch die Varianz oder ein anderes Maß), größer oder kleiner als ein bestimmter Grenzwert ist.

Beispiele:

- Das mittlere Gewicht der Kartoffelsäcke, die in einem Supermarkt zum Verkauf angeboten werden, beträgt mindestens 2,0 kg.
- Der mittlere Anteil fehlerhafter CPUs auf den Silizium-Wafers eines bestimmten Herstellers ist kleiner als 10%.

Die Alternativhypothese behauptet dann, dass die Säcke im Schnitt leichter als 2 kg sind, bzw. dass der Anteil im Schnitt größer als 10% ist. Beide Hypothesen definieren also Bereiche (Intervalle).



Nullhypothese $H_0: \mu \leq \mu_0$ Alternativhypothese $H_1: \mu > \mu_0$

1.3.2 Zweiseitige Tests (Punkttests)

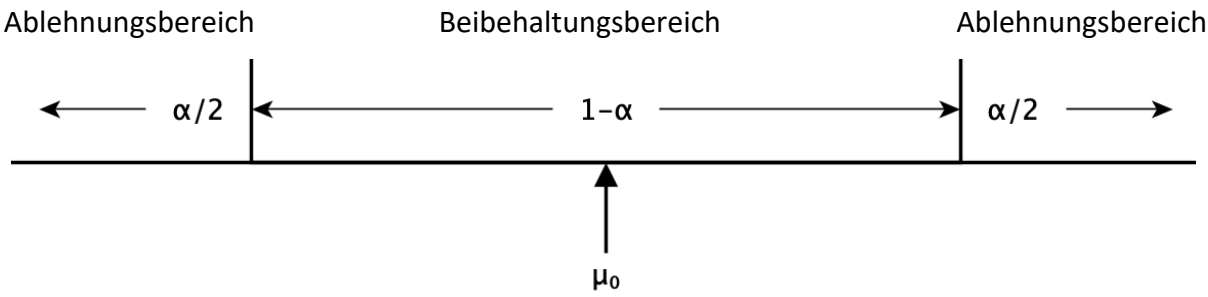
Beim zweiseitigen Test behauptet die Nullhypothese, dass die Kennzahl einen genauen Wert hat. Die Alternativhypothese behauptet dann, dass die Kennzahl einen anderen Wert hat, der sowohl größer als auch kleiner sein kann.

Beispiele:

- Männliche Österreicher im Alter zwischen 18 und 25 Jahren sind im Schnitt 178 cm groß[1]
- Das Pro-Kopf-Bruttonationaleinkommen Österreichs beträgt 44.610 Euro pro Jahr[2]

[1] https://www.laenderdaten.info/durchschnittliche-koerpergroessen.php

[2] https://www.laenderdaten.info/durchschnittseinkommen.php



Nullhypothese H_0: $\mu = \mu_0$
Alternativhypothese H_1: $\mu \neq \mu_0$

Die Nullhypothese definiert also einen Punkt, keinen Bereich. Trotzdem gibt es einen Bereich (den Beibehaltungsbereich), innerhalb dessen die Nullhypothese beibehalten wird.

1.4 Signifikanzniveau

Bei einem Hypothesentest definiert man **vor dem Test** ein bestimmtes Vertrauensintervall. Kommt die Schätzgröße der Stichprobe in diesem Intervall zu liegen, vertraut man darauf, dass die Differenz zwischen der Schätzgröße μ und dem wahren Wert der Grundgesamtheit μ_0 auf erwartbare zufällige Schwankungen bei der Stichprobenziehung zurückzuführen ist. Kommt μ außerhalb des Vertrauensintervalls zu liegen, erscheint die Annahme, dass die Abweichung zwischen μ und μ_0 rein zufällig ist, nicht mehr glaubwürdig.

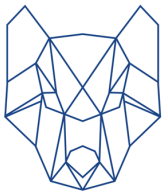
Die Breite dieses Vertrauensniveaus hängt von der Varianz der Mittelwerte vieler Stichproben ab und entspricht genau dem Konfidenzintervall von Intervallschätzungen und wird auch genau wie dort beschrieben berechnet (siehe Skriptum Schätzverfahren, Abschnitt 3).

Was bei der Intervallschätzung noch »Irrtumswahrscheinlichkeit« geheißen hat, heißt jetzt »Signifikanzniveau«. Sie wird in beiden Fällen mit dem Buchstaben α bezeichnet.

Bei Hypothesentests gibt das Signifikanzniveau an, wie groß die Wahrscheinlichkeit dafür ist, die Nullhypothese irrtümlicherweise abzulehnen, obwohl sie in Wahrheit zutrifft.

1.4.1 α-Fehler (Fehler 1. Art; falsch positiv)

Der eben beschriebene Fehler (H_0 trifft in Wahrheit zu, wird aber irrtümlich abgelehnt) heißt »α-Fehler« oder »Fehler 1. Art«. Manchmal, vor allem in einem medizinischen Kontext, bezeichnet man Ereignisse, die diesem Fehler entsprechen, auch als »falsch positiv«. Als »positiv« gilt dabei ein Testergebnis, das eine Abweichung von der Norm (also meist eine Erkrankung) anzeigt.



Den Fehler 1. Art kann man in einem Hypothesentest kontrollieren. Man definiert nämlich den Wert von α **vor** dem Test fest und legt somit eine Obergrenze für das Ausmaß dieses Fehlers fest.

1.4.2 β-Fehler (Fehler 2. Art; falsch negativ)

Natürlich kann es auch passieren, dass die Nullhypothese in Wahrheit gar nicht zutrifft, und man sie trotzdem beibehält, weil die Daten der Stichprobe für eine Ablehnung nicht ausreichen. Dann liegt ein »Fehler 2. Art« vor. Das Ausmaß dieses Fehlers gibt man mit der Variablen β an, daher heißt dieser Fehler auch »β-Fehler«. Die Bezeichnung »falsch negativ« kommt daher, dass ein Test keine Normabweichung anzeigt, obwohl in Wahrheit eine vorliegt. (Ein Test zeigt bei einem tatsächlich kranken Menschen keine Erkrankung an.)

Ohne Annahmen über den wahren Wert der Grundgesamtheit ist es nicht möglich, den genauen Wert von β zu berechnen. Ironischerweise ist der Wert von β nämlich umso größer, je näher der wahre Wert und die Schätzung beieinander liegen.

Auch zu beachten ist, dass der β-Fehler größer wird, wenn der α-Fehler kleiner gemacht wird (und umgekehrt).

		In Wahrheit	
		H_0 trifft zu	H_0 trifft nicht zu
Ergebnis des Hypothesentests	H_0 wird beibehalten	richtig negativ P = 1-α	falsch negativ β-Fehler P = β
	H_0 wird abgelehnt	falsch positiv α-Fehler P = α	richtig positiv P = 1-β

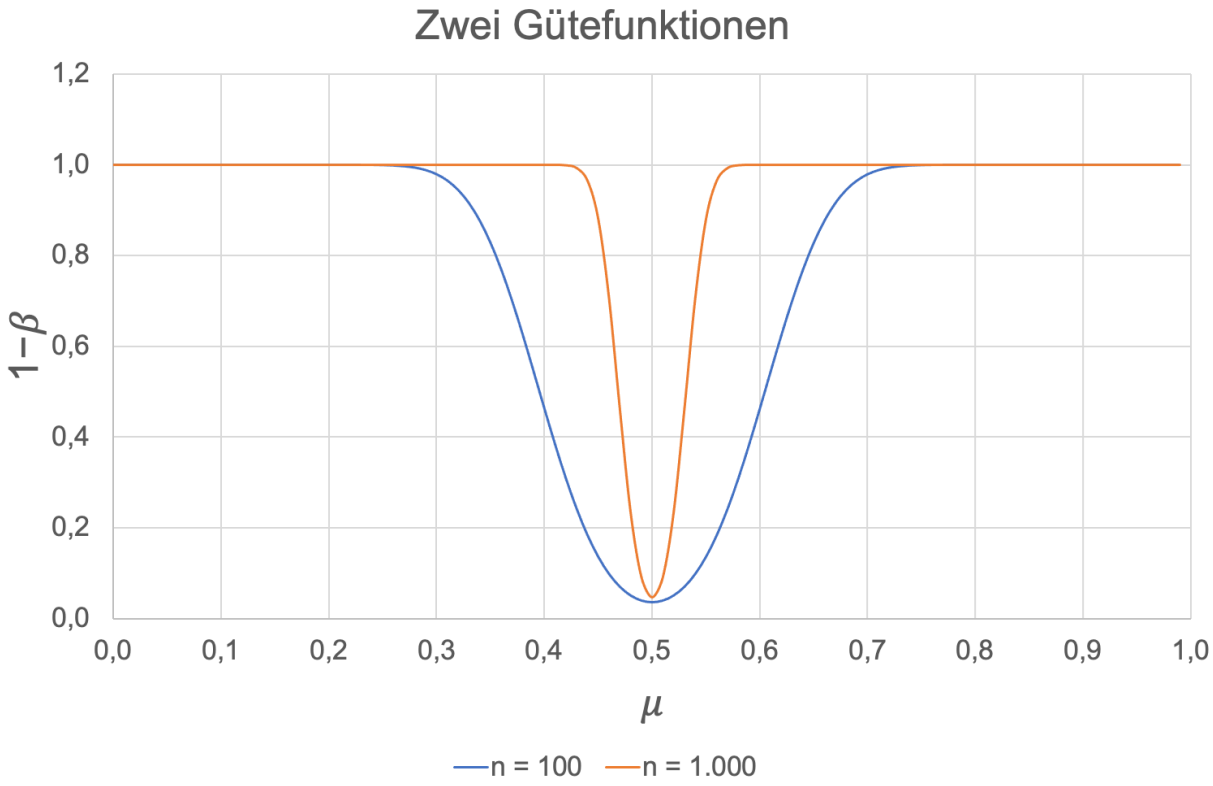
Fehler 1. und 2. Art

1.5 Gütefunktion

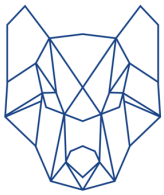
Die Güte eines Tests, oft auch als Teststärke bezeichnet, ist die Wahrscheinlichkeit dafür, dass die Nullhypothese vom Test abgelehnt wird, wenn die Nullhypothese tatsächlich nicht zutrifft (also die Wahrscheinlichkeit eines »richtig positiven« Ergebnisses unter der Bedingung, dass H_0 falsch ist). Die Güte ist also der Wert $1 - \beta$ in der obigen Darstellung.

Das Problem ist aber, dass die Alternativhypothese üblicherweise in folgender Form formuliert wird: »Der wahre Wert der Grundgesamtheit ist **irgendein** anderen Wert als μ_0.«

Um die Güte zu berechnen, muss man eine Alternativhypothese der folgenden Form verwenden: »Der wahre Wert der Grundgesamtheit hat **genau** den Wert μ_1.« (Wobei natürlich gilt: $\mu_0 \neq \mu_1$). Dies erlaubt uns, die Fähigkeit des Tests zu beurteilen, speziell diesen alternativen Zustand zu erkennen. Das kann man aber für beliebige μ-Werte durchrechnen, und man erhält dann eine Funktion, die jedem μ eine bestimmte Güte zuordnet. Das ist dann die Gütefunktion.



(Mehr über die Gütefunktion steht weiter unten, in 2.1.1, wo die Gütefunktion bei einem konkreten Beispiel besprochen wird.)



1.6 p-Wert

Der p-Wert ist das Ergebnis eines Hypothesentests. Dieses Ergebnis wird mit dem α-Wert, der auch Signifikanzniveau heißt, verglichen, um zu entscheiden, ob die Nullhypothese beibehalten werden kann, oder verworfen werden muss.

Eine vereinfachte Definition: Der p-Wert ist die Wahrscheinlichkeit dafür, dass die Stichprobe, die ich gerade untersuche, beim Zutreffen der Nullhypothese rein zufällig zustande gekommen ist. Daher bedeutet ein sehr kleiner p-Wert, dass es sehr unwahrscheinlich ist, dass ich rein zufällig genau diese Stichprobe erhalten habe, obwohl die Nullhypothese zutrifft. Andere Gründe sind viel wahrscheinlicher. Der wahrscheinlichste Grund ist vermutlich, dass die Nullhypothese nicht zutrifft. Ein großer p-Wert hingegen bedeutet, dass es nicht überraschend ist, diese Stichprobe zu erhalten, denn das war mit hoher Wahrscheinlichkeit zu erwarten, wenn die Nullhypothese zutrifft.

Das, was an dieser Formulierung nicht ganz korrekt ist, ist, dass dabei von genau der einen Stichprobe die Rede ist, die untersucht wird. In Wahrheit beschreibt der p-Wert die Wahrscheinlichkeit für das Auftreten dieser einen Stichprobe, plus aller anderen Stichproben, die dieselbe Prüfgröße ergeben, plus alle noch extremeren Stichproben, also alle Stichproben, deren Prüfgröße noch stärker vom wahren Wert der Grundgesamtheit abweicht, als dass bei »meiner« Stichprobe der Fall ist.

Andere, aber gleichwertige Interpretation: Der p-Wert ist das kleinste hypothetische Signifikanzniveau bei dem die Nullhypothese H_0 bereits zu verwerfen wäre. (Der p-Wert markiert genau die Grenze zwischen Beibehalten und Verwerfen.)

Viele Testverfahren sind so konstruiert, dass man für sie gar kein Signifikanzniveau angeben muss. Sie liefern daher auch nicht die Entscheidung zwischen »H_0 beibehalten« und »H_0 ablehnen«, sondern sie geben den p-Wert p aus, den man **nach** dem Test mit dem Signifikanzniveau α vergleichen muss, für das man sich **vor** dem Test entschieden hat.

- $p \leq \alpha$: H_0 ablehnen, H_1 annehmen
- $p > \alpha$: H_0 beibehalten, H_1 nicht annehmen

2 Testverfahren

2.1 Binomialtest

Sie nehmen einem Trickbetrüger eine Münze ab, von der Sie glauben, dass die beiden Werte Kopf und Zahl nicht gleich häufig erscheinen, wenn die Münze geworfen wird. Der Eigentümer der Münze behauptet, die Münze wäre fair, würde also beide Seiten gleich häufig anzeigen.

Das wiederholte Werfen der Münze ist ein Bernoulli-Experiment mit dem Parameter $p = 0{,}5$ (das ist die Wahrscheinlichkeit dafür, dass *Kopf* geworfen wird). Der Erwartungswert nach n Würfen ist $n \cdot p$ (Siehe Skriptum »diskrete Verteilungen«).

Bei einem einseitigen Binomialtest lautet die Nullhypothese, dass nach n Würfen höchstens (oder mindestens) $n \cdot p$-mal *Kopf* erscheint. Beim zweiseitigen Binomialtest lautet die Nullhypothese, dass genau $n \cdot p$-mal *Kopf* erscheint. Die Alternativhypothese behauptet jeweils das Gegenteil.

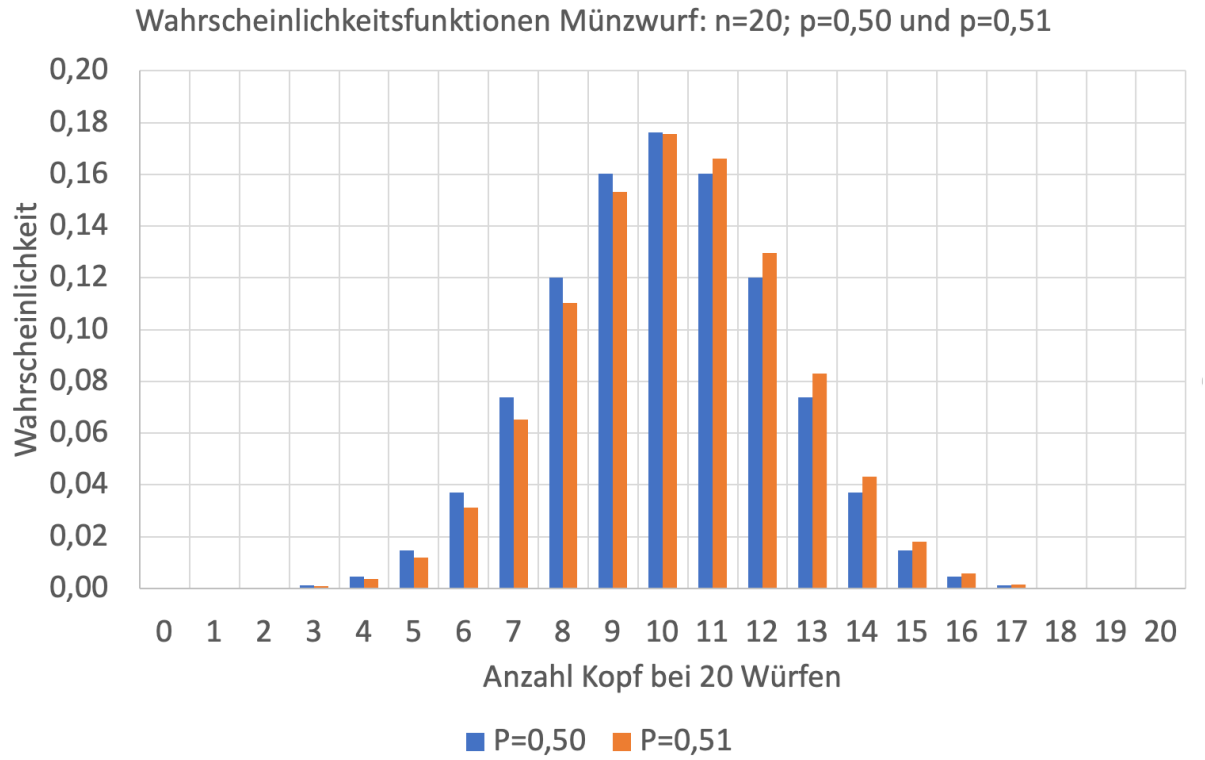
Wir führen im Folgenden einen zweiseitigen Binomialtest durch.

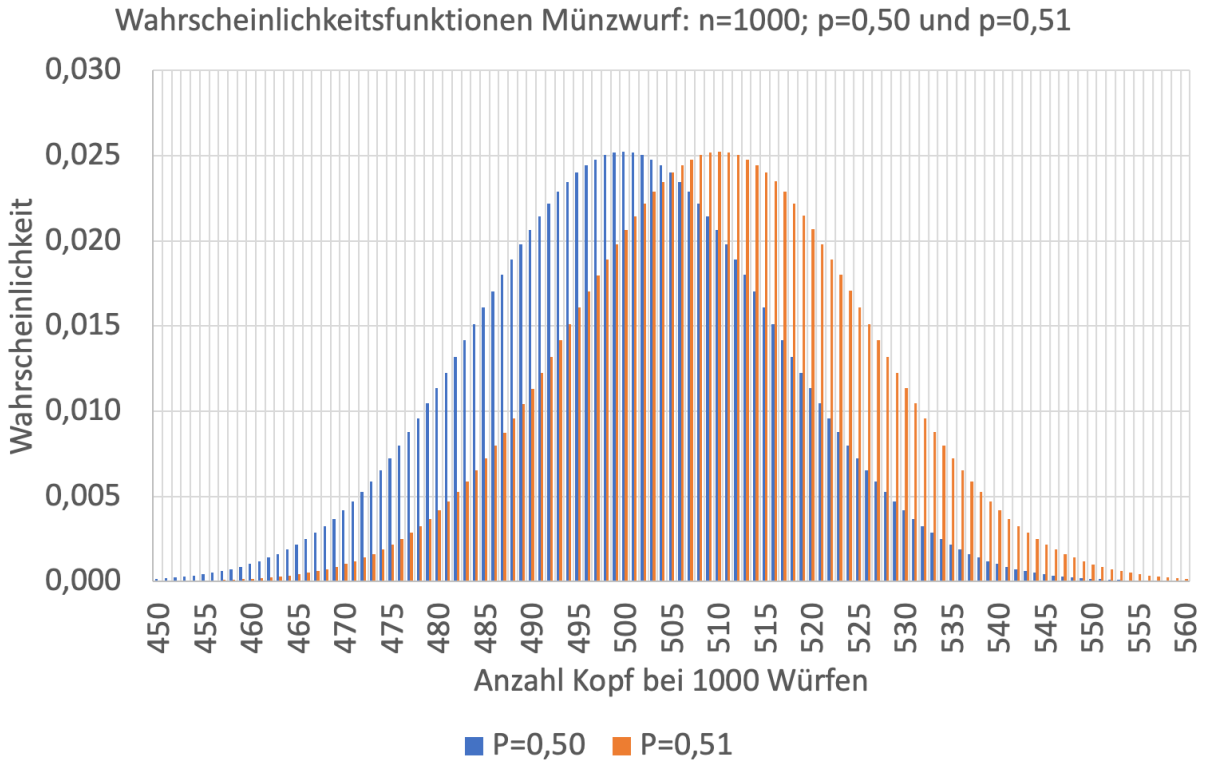
2.1.1 Güte, Gütefunktion

Nehmen wir an, der Falschspieler hätte 2 Münzen. Eine Münze ist so beschaffen, dass sie *immer* so landet, dass *Kopf* erscheint. Die Kopf-Wahrscheinlichkeit w hat also den Wert $1{,}0$. Der Erwartungswert μ_1 dieser Münze ist also n. (Wirft man die Münze n-Mal, erscheint n-Mal Kopf). Die zweite Münze zeigt Kopf nur geringfügig öfter als Zahl. Nehmen wir an, dass bei dieser Münze die Wahrscheinlichkeit für Kopf bei $w = 0{,}51$ liegt. Dann ist $\mu_2 = 0{,}51 \cdot n$ Das heißt, dass bei 100 Würfen der Erwartungswert 51 ist, bei 1000 Würfen 510.

Es ist offensichtlich, dass die völlig einseitige Münze leichter zu enttarnen ist als die mit der kleinen Tendenz zu einer Seite. Bei einer Stichprobe von 20 Würfen wird man die einseitige Münze klar erkennen. Der Test mit dieser Stichprobengröße, kann also klar zwischen »faire Münze« und »völlig einseitige Münze« unterscheiden. Das ist so, weil die Wahrscheinlichkeit, bei einer fairen Münze 20-Mal Kopf zu erhalten, bei ungefähr einem Millionstel liegt. Diese Wahrscheinlichkeit ist der p-Wert.

Die Güte eines Tests ist $1 - p$. Das ist die Wahrscheinlichkeit dafür, bei der gewählten Stichprobengröße die gezinkte Münze zu erkennen. Sie ist bei einer Stichprobe der Größe 20 nahe bei 1. Bei der fast-fairen Münze werden wir keine Chance haben, sie mit nur 20 Würfen von einer tatsächlich fairen Münze zu unterscheiden. Dazu ist die Güte des Tests nicht ausreichend. Der Gütewert für eine Münze mit $w = 0{,}51$ ist nahe bei 0. Sogar 1000 Würfe werden nicht ausreichend sein, um mit Sicherheit sagen zu können, dass die Münze gezinkt ist.


Wahrscheinlichkeitsfunktionen Münzwurf: n=20; p=0,50 und p=0,51
Wahrscheinlichkeit
0,20
0,18
0,16
0,14
0,12
0,10
0,08
0,06
0,04
0,02
0,00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Anzahl Kopf bei 20 Würfen
P=0,50
P=0,51


Wahrscheinlichkeitsfunktionen Münzwurf: n=1000; p=0,50 und p=0,51
Wahrscheinlichkeit
0,030
0,025
0,020
0,015
0,010
0,005
0,000
450 455 460 465 470 475 480 485 490 495 500 505 510 515 520 525 530 535 540 545 550 555 560
Anzahl Kopf bei 1000 Würfen
P=0,50
P=0,51

2.2 t-Test

Es gibt drei Arten des t-Tests:

2.2.1 Ein-Stichproben-t-Test

Voraussetzungen:

Die Stichprobe sollte aus einer normalverteilten Population stammen oder zumindest annähernd normalverteilt sein. Da es um die Verteilung der möglichen Mittelwerte von Stichproben geht, reicht es auch aus, wenn die Stichprobengröße mindestens 30 ist, denn dann sind diese Mittelwerte auch dann fast normalverteilt, wenn die Population eine völlig andere Verteilung hat.

Die Stichprobe soll durch einen Zufallsprozess zustande gekommen sein. (Kein Bias, keine Abhängigkeit der Elemente der Stichprobe untereinander.)

Beispiel:

Der neue Personalchef eines Unternehmens möchte prüfen, ob die Gehälter der Angestellten dem branchenüblichen Durchschnitt entsprechen. Das branchenüblichen Durchschnittsgehalt ist 60.000 Euro pro Jahr. Wir haben eine Stichprobe mit 10 Gehältern:

Stichprobe: {58.000, 62.000, 59.000, 61.000, 57.000, 63.000, 60.000, 62.000, 61.000, 59.000}

Zuerst berechnen wir den Punktschätzer, also den Mittelwert der Stichprobe:

$$\bar{x} = \frac{58000 + 62000 + \cdots + 59000}{10} = 60200$$

Dann die korrigierte Standardabweichung

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

$$s = \sqrt{\frac{1}{9} \cdot (58000 - 60200)^2 + (62000 - 60200)^2 + \cdots + (59000 - 60200)^2}$$

$$s = 1932{,}18\cdots$$

Der hypothesierte Mittelwert der Population (μ) = 60.000

Formuliere die Nullhypothese (H_0) und die Alternativhypothese (H_1).

$$H_0: \mu = 60.000$$

$$H_1: \mu \neq 60.000$$

Wähle ein Signifikanzniveau (α). Häufig wird ein Signifikanzniveau von $\alpha = 0{,}05$ gewählt, und das ist auch der Wert, den wir verwenden wollen.

Berechne den Standardfehler aus der oben berechneten korrigierten Standardabweichung der Stichprobe s und der Stichprobengröße n.

$$SE = \frac{s}{\sqrt{n}} = \frac{1932{,}18\cdots}{\sqrt{10}} = 611{,}01\cdots$$

Berechne die Teststatistik (t) mit der folgenden Formel:

$$t = \frac{\bar{x} - \mu}{SE} = \frac{60200 - 60000}{611{,}01\cdots} = 0{,}3273\cdots$$

Nun muss der kritische t-Wert für $\frac{\alpha}{2} = 0{,}025$ und $(n-1) = (10-1) = 9$ Freiheitsgrade aus einer t-Verteilungstabelle oder mit einer entsprechenden Software ermittelt werden. Der Wert muss für $\frac{\alpha}{2}$ (und nicht für α) gesucht werden, weil wir einen zweiseitigen Test (Punkttest) machen (siehe 1.3.2). Dieser kritische t-Wert ist $2{,}262157\cdots$.

Wenn der zuvor berechnete t-Wert zwischen $-2{,}262$ und $+2{,}262$ liegt, befindet sich der t-Wert im Annahmebereich. Das ist in unserem Beispiel der Fall, daher wird die Nullhypothese angenommen: Die Gehälter haben eine branchenübliche Höhe.

Anmerkung

Wenn die Stichprobe mehr als 30 Werte enthalten hätte, hätte man die t-Verteilung, die für die Varianz gilt, durch eine Standardnormalverteilung annähern können. Dann müsste man zwar ebenfalls in einer Tabelle nachschauen, um den kritischen Wert zu ermitteln, aber das wäre dann die Tabelle für die Standardnormalverteilung, und für diese Tabelle braucht man keine Freiheitsgrade.

Alternativer Rechengang (Berechnen des p-Werts)

Wenn man den kritischen Wert ermittelt, kann man mit Statistik-Programmen aus dem berechneten t-Wert und den Freiheitsgraden den p-Wert berechnen. In Excel gibt es für den zweiseitigen t-Test diese Formel: `=T.VERT.2S(t;f)`. Dabei ist `t` der t-Wert (falls er negativ ist, muss man den Absolutwert nehmen) und `f` ist die Anzahl der Freiheitsgrade. Wenn wir das mit unseren Zahlen machen (`=T.VERT.2S(0,3273;9)`) erhalten wir das Ergebnis 0,7509. Das ist der p-Wert, und er liegt deutlich über 0,05, daher nehmen wir die Nullhypothese an.

2.2.2 Zwei-Stichproben-t-Test

An die Stelle des fest vorgegebenen zu erwartenden Mittelwerts tritt der Mittelwert einer zweiten Verteilung.

Beispiel: Eine Portionierungsmaschine erzeugt vor der Reparatur Portionen mit einem Mittelwert von 352 g. Nach der Reparatur werden wieder Stichproben gezogen, der Mittelwert dieser Stichproben beträgt 346 g. Stellt die Maschine nach wie vor ungefähr gleich große Portionen her? Auch diese Frage kann mit einem t-Test beantwortet werden. Dazu dürfen die beiden Stichproben auch unterschiedlich groß sein.

Auf eine detaillierte Beschreibung des Rechengangs wird hier verzichtet und auf die einschlägige Fachliteratur verwiesen. Zu beachten ist, dass es hier mehrere Berechnungsarten gibt, je nachdem ob die Varianzen der Grundgesamtheiten bekannt oder unbekannt sind, und im Fall von bekannten Varianzen, ob sie gleich oder verschieden sind.

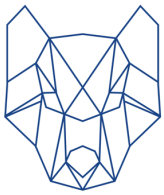
2.2.3 Paarweiser t-Test

Anstatt die Mittwerte zweier Stichproben zu vergleichen, schaut man sich den Mittelwert der Differenzen zweier abhängiger Stichproben an. Was damit gemeint ist, soll ein Beispiel zeigen:

Bei allen Personen, die an einem Medikamententest teilnehmen, wird vor und nach Verabreichung des Medikaments der Blutdruck gemessen. Es gibt also von jeder Person 1 Messungen. Die Messungen vor der Medikamentengabe bilden eine Stichprobe, die Messungen danach bilden die zweite Stichprobe. Beide Stichproben sind exakt gleich groß. Dann bildet man von jeder Person die Differenz der beiden Werte. Die Null-Hypothese bei Medikamententests lautet immer »Das Medikament ist wirkungslos«. Das heißt, dass der Mittelwert der Differenzen genau 0 betragen sollte. Wenn das Medikament den Blutdruck beeinflusst, wird sich der Mittelwert der Differenzen vom Wert 0 signifikant unterscheiden. Auch das kann mit einem t-Test getestet werden.

2.3 Chi-Quadrat-Test und G-Test

Beide Tests können für dieselben Aufgaben verwendet werden. Der schwierigste Rechenschritt in einem Chi-Quadrat-Test ist das Quadrieren von Zahlen. Im G-Test müssen Logarithmen berechnet werden. Daher wurde in den Zeiten, als man Statistik hauptsächlich mit Papier und Bleistift unter Zuhilfenahme von Tabellenbüchern gemacht wurde, praktisch nur der Chi-Quadrat-Test gemacht. Seit es aber Computer und ausgereifte Statistik-Programme gibt, gehört der Chi-Quadrat-Test eigentlich zum alten Eisen, weil der G-Test dieselbe Aufgabe viel genauer löst, aber "intern" eben aufwändige Rechenschritte erfordert. Trotzdem wird der G-Test in Lehrbüchern über Statistik kaum erwähnt. Auch hier wird auf die



genaue mathematische Beschreibung des G-Tests verzichtet. Bitte konsultieren Sie die einschlägige Fachliteratur, wenn Sie mehr darüber wissen wollen.

Beide Tests (Chi-Quadrat- und G-Test) gibt es jeweils in zwei Ausprägungen. In beiden Fällen ist die Grundgesamtheit in Kategorien unterteilt (z.B. Gummibärchen in verschiedenen Geschmacksrichtungen). Der Test ist ungeeignet für stetige Verteilungen.

2.3.1 Chi-Quadrat-Test auf Anpassungsgüte

Es wird behauptet, dass in jeder Packung Gummibärchen alle fünf Geschmacksrichtungen zu je 20% vertreten sind.

Der Chi-Quadrat-Test ist eine Näherung des mathematisch viel komplexeren G-Tests. Der G-Test kann unabhängig von den zu erwartenden Häufigkeiten in den Stichproben gemacht werden, den einfacheren Chi-Quadrat-Test sollte man nur durchführen, wenn in der Stichprobe pro Kategorie jeweils mindestens 5 Exemplare erwartet werden. Wenn eine Stichprobe eine Tüte Gummibärchen ist, sollten also pro Geschmacksrichtung 5 Bärchen erwarten werden, sonst liefert der Chi-Quadrat-Test möglicherweise eine falsche Antwort.

Der Chi-Quadrat-Test wird wie folgt durchgeführt:

Sie ermitteln pro Geschmacksrichtung die Differenz zwischen der erwarteten Anzahl und der Anzahl aus der Stichprobe. Diese Differenzen quadrieren Sie und teilen Sie durch die erwartete Anzahl. Die so erhaltenen Werte addieren sie, um die Prüfgröße zu erhalten. Wenn diese Prüfgröße klein genug ist, kann man davon ausgehen, dass die Verteilung in der Stichprobe der erwarteten Verteilung entspricht.

Beispiel:

Geschmack	Anzahl in der Stichprobe	erwartete Anzahl	Differenz	Quadrat	Quotient
Apfel	18	20	$18-20=-2$	$(-2)^2=4$	$4/20=0{,}2$
Kirsche	25	20	$25-20=5$	$5^2=25$	$25/20=1{,}25$
Orange	12	20	$12-20=-8$	$(-8)^2=64$	$64/20=3{,}2$
Zitrone	22	20	$22-20=2$	$2^2=4$	$4/20=0{,}2$
Traube	23	20	$23-20=3$	$3^2=9$	$9/20=0{,}45$

Die Summe der Werte in der letzten Spalte ergibt 5,3, und dieser Wert muss nun mit einem kritischen Wert verglichen werden, der in einem separaten Verfahren berechnet wird. Darin fließen das zuvor festgelegte Signifikanzniveau (z.B. $\alpha = 5\%$) und eine Anzahl von Freiheitsgraden ein. Diese Anzahl der Freiheitsgrade ist die Anzahl der Kategorien minus 1. Da wir im Beispiel 5 Kategorien hatten, muss man also in die Berechnung als Anzahl der Freiheitsgrade der Wert 4 einsetzen.

2.3.2 Chi-Quadrat-Test auf Unabhängigkeit

Hier analysiert man zwei Verteilungen, die aus zwei Stichproben stammen, und setzt in der oben beschriebenen Berechnung anstelle der theoretisch erwarteten Anteile die Anteile der zweiten Stichprobe ein.

Beispiel:

Sie stellen bei einem Kindergeburtstag Gummibärchen bereit, weisen die Kinder aber an, dass die Mädchen nur Bärchen aus der runden Schüssel auf dem Küchentisch essen dürfen, während die Buben nur Gummibärchen aus der eckigen Schüssel auf dem Couchtisch essen dürfen. Wenn das Fest fertig ist, zählen Sie bei den Mädchen und bei den Buben, wie viele Gummibärchen pro Sorte jeweils gegessen wurden, und führen mit diesen Messergebnissen die ohne beschriebene Rechnung durch. Sie bekommen dann die Antwort auf die Frage, ob Buben und Mädchen dieselben Gummibärchen im selben Ausmaß bevorzugen, oder ob es Unterschieden zwischen den Geschlechtern gibt.