Grundlagen des statistischen Testens

Dr. Mariana Nold

Institut für Soziologie,
Fakultät für Sozial- und Verhaltenswissenschaften,
Lehrstuhl für empirische Sozialforschung und Sozialstrukturanalyse

13. November 2017

Übersicht

1. Ziel der heutigen Veranstaltung

2. Signifikanter Unterschied zwischen den Gruppen
   - Kann das Zufall sein?
   - Das Modell grafisch
   - Die Grundidee des Signifikanztest

3. Die Teststatistik und die Testentscheidung
   - Die Teststatistik
   - Der Ablehnbereich und der p-Wert

4. Ausblick Übung: Häufige Testprobleme

Ziel der heutigen Veranstaltung ...

ist es die folgenden Fragen beantworten zu können:

Zielfragen für heute

1. Was bedeuten die Begriffe Nullhypothese und Forschungshypothese?
2. Welche der beiden Hypothesen kann man nachweisen?
3. Wie sind der α-Fehler und der β-Fehler definiert?
4. Was ist das Signifikanzniveau bzw. die Irrtumswahrscheinlichkeit?
5. Was versteht man unter einer Teststatistik?
6. Welche Rolle spielt der kritische Wert?
7. Was versteht man unter einem p-Wert?

Die Grundidee des Signifikanztests

- Wir werden uns heute an einem Beispiel ansehen, welcher Logik der klassische Signifikanztest folgt.
- Unser Fokus liegt hierbei weniger auf die mathematischen Zusammenhänge, als vielmehr auf die wissenschaftstheoretische Legitimation dieser Vorgehensweise.
- Die Darstellung orientiert sich an den lesenswerten Kapiteln
  - *4.3.1: Die Grundidee von Signifikanztests (+Vorwort, ab S. 136)*
  - *4.3.2: Die Praxis von Signifikanztests am Beispiel des Testens von Mittelwertunterschieden (nur bis S. 151)*
  - *4.3.2 Problem statistischen Testens*

  aus dem Buch "Statistik-Eine Einführung für Sozialwissenschaftler" von Ludwig-Mayerhofer, Liebeskind, Geißler, (im Folgenden mit LMLG abgekürzt.)

Gibt es einen Unterschied hinsichtlich eines Merkmals in zwei Gruppen?

Ich werde heute anhand eines Beispiels darstellen, was die Behauptung: " Der Unterschied hinsichtlich eines Merkmals in zwei Gruppen ist signifikant" bedeutet. Zu dem Beispiel habe ich keine Daten.

Reales Beispiel: Kann das Zufall sein?

In einer Einrichtung zur stationären Drogenfreien Therapie, hat ein Klient den Eindruck zu wenig Gulasch auf seinem Teller zu haben. Er fragt jemand vom Personal, ob es Zufall sein kann, dass er und die anderen Klientinnen und Klienten, wenn es Gulasch gibt, immer nur zwei oder drei Stücke Fleisch auf dem Teller haben und die Mitarbeiterinnen und Mitarbeiter deutlich mehr.

Die Vermutung des Klienten

- Er hat die Vermutung, dass der Koch, der das Gulasch ausgibt tiefer schöpft, wenn ein Mitarbeiter oder eine Mitarbeiterin vor ihm steht.
- Wenn seine Vermutung zutrifft, dann sind es zwei Unterschiedliche Zufallsmechanismen.
- In diesem Fall, kann es vorkommen, dass eine Klientin bzw. ein Klient mehr Fleisch auf dem Teller hat als ein Mitarbeiter oder eine Mitarbeiterin, **aber nach Wahrscheinlichkeit, haben die Angestellten mehr Fleisch auf dem Teller, als die die dort eine Therapie machen?**
- Die Vermutung kann daher überprüft werden, indem man die Mengen an Fleisch vergleicht.
- Dabei sind verschieden Vorgehensweisen denkbar.

Kann der Sachverhalt mit statistischen Methoden geklärt werden?

- Im Prinzip nicht, denn es bleibt immer eine Unsicherheit. Man kann allerdings die Unsicherheit quantifizieren.
- Desto mehr empirische Information uns vorliegt, desto sichere wird unsere Aussage. Mit anderen Worten: Desto länger wird diesen Ausgabeprozess beobachten, desto werden wir in der Beurteilung der Vermutung
- Dabei gilt die Annahme, dass der Prozess sich nicht verändert. Das bedeute, wir gehen davon aus, dass das Modell mit dem wir arbeiten, die ganze Zeit über gleicher Maßen gut passt.

Ein geeignetes Modell I

Wir brauchen ein Modell, um auf der Modellebene arbeiten zu können.

- Das Modell ist folgendes: Ich bezeichne mit der Zufallsvariable X, die Menge Fleisch auf den Tellern der Klientinnen und Klienten und mit Y entsprechen die Menge Fleisch auf den Tellern der Mitarbeiterinnen und Mitarbeiter.
- In einem Gedankenexperiment, wird der Prozess ein Jahr lang beobachtet.
- Immer, wenn es Gulasch wird aus jeder Gruppe eine einfache Zufallsstichprobe gezogen. Die am Ende erreichten Stichprobenumfänge werden mit n_1 (für X) und n_2 (für Y) bezeichnet.
- In der List der Klienten und Klientinnen sehen die Daten formal so aus: $x_1, x_2, \ldots, x_{n_1}$, dabei entspricht jedes x_i einer beobachteten Zahl.

Ein geeignetes Modell II

- In formaler Schreibweise haben wir n_1 Realisation der Zufallsvariable X_1 und n_2 Realisation der Zufallsvariable Y.
- Ich mache jetzt die folgenden Modellnahmen X_i iid wie

$$X \sim N(\mu_X, \sigma_X) \tag{1}$$

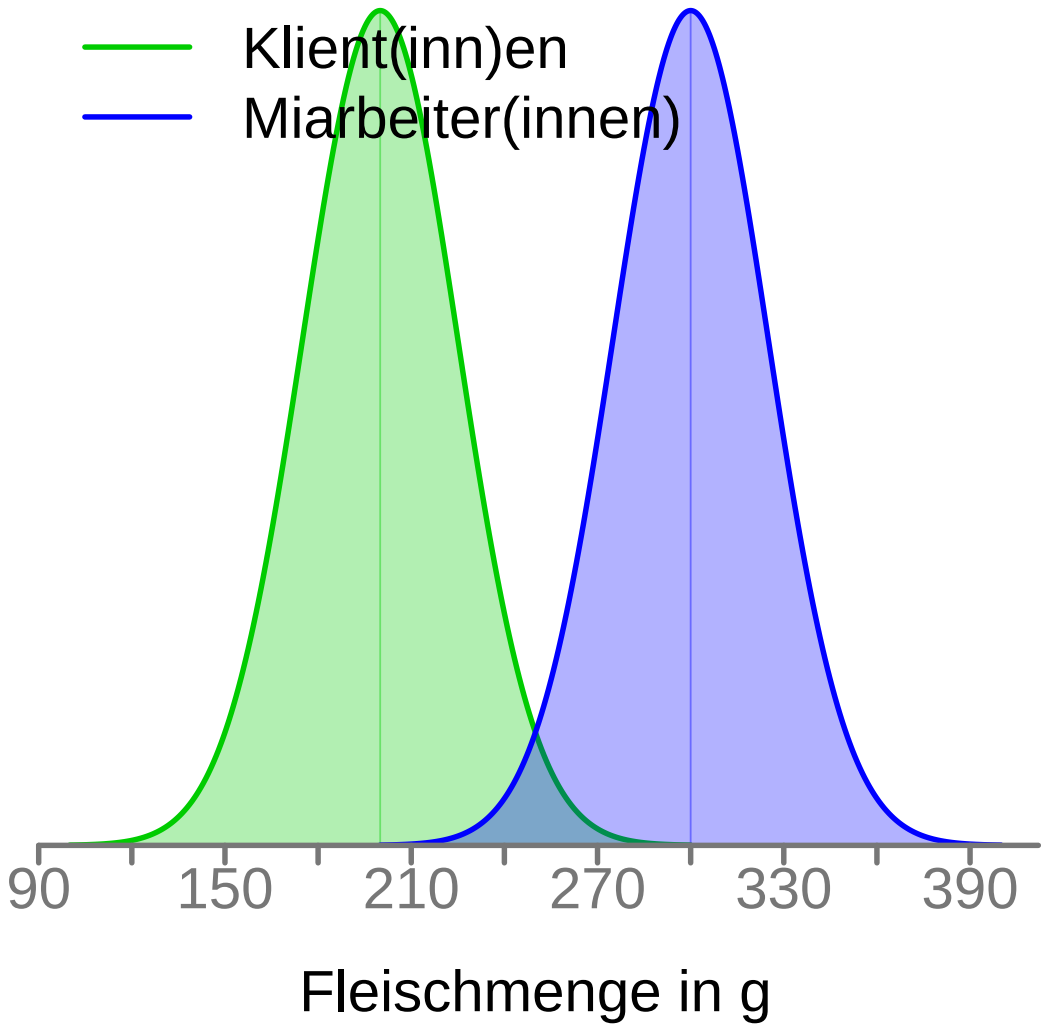
und Y_i iid wie

$$Y \sim N(\mu_Y, \sigma_Y). \tag{2}$$

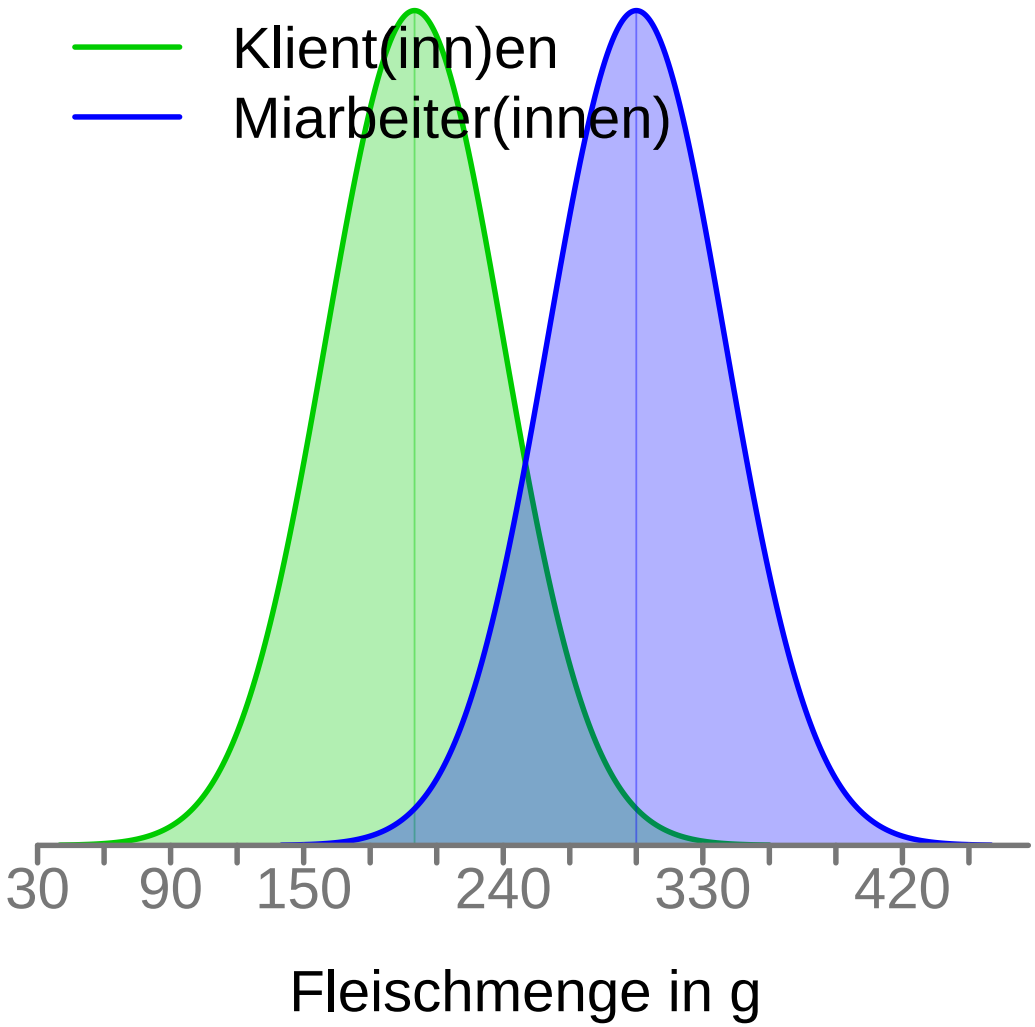
Bitte fassen Sie die Folgenden Bilder in Worte

- Die folgenden Bilder zeigen in gelb die Modell-Dichte für X und in grün die Modell-Dichte für Y.
- In den grafisch dargestellten Modellen gilt jeweils: $\sigma_X = \sigma_Y$
- Der Überlappungsbereich ist je in Weinrot dargestellt.
- Wenn die Modell so der Realität entsprechen würden, was würde das bedeuten?

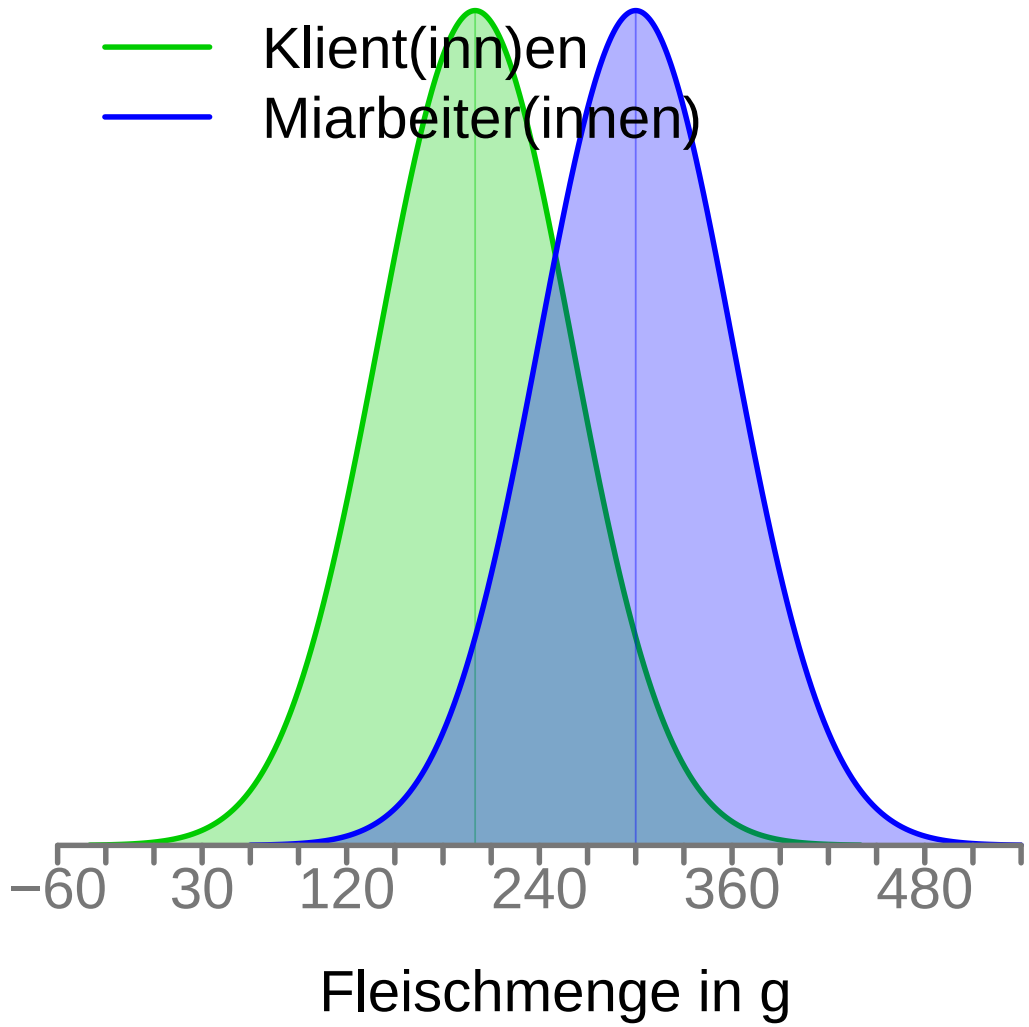
$\mu_X = 200$, $\mu_Y = 300$, $\sigma_X = \sigma_Y = 25$



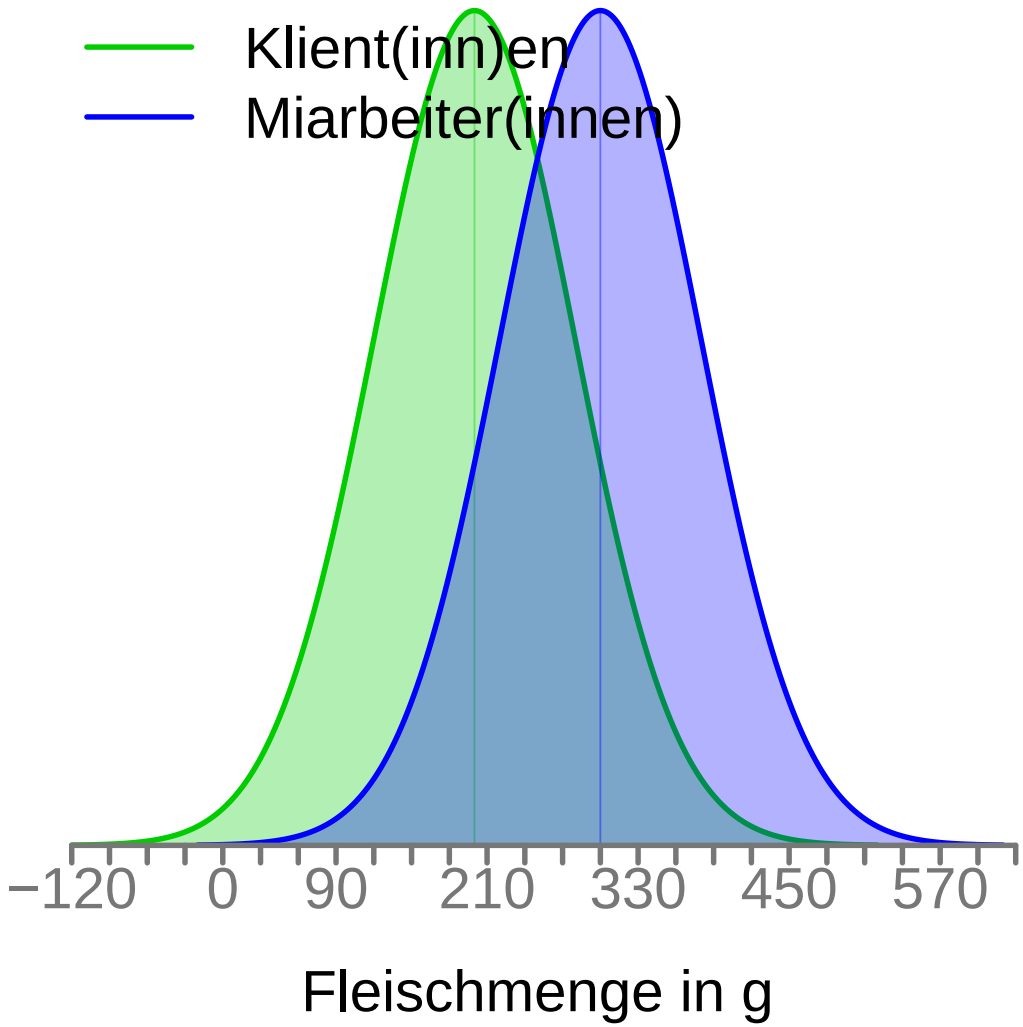
$\mu_X = 200$, $\mu_Y = 300$, $\sigma_X = \sigma_Y = 40$



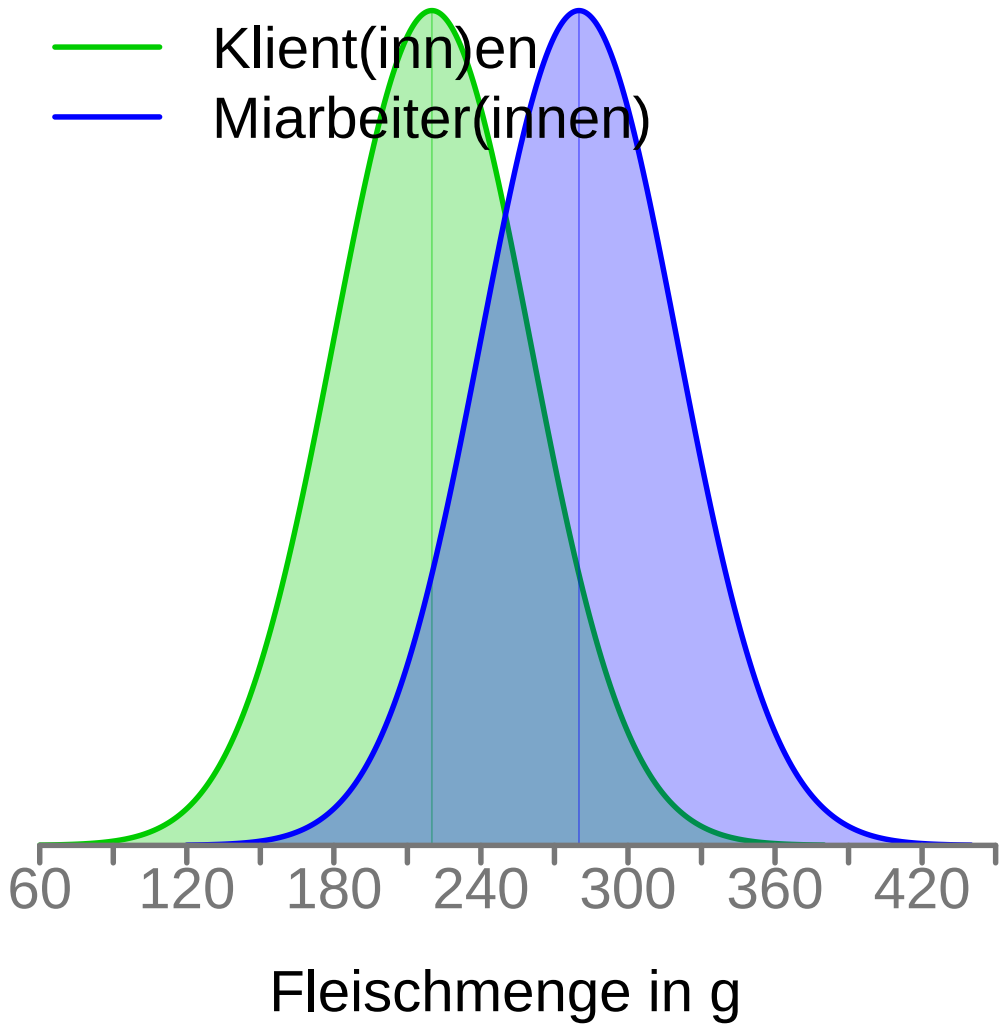
$\mu_X = 200$, $\mu_Y = 300$, $\sigma_X = \sigma_Y = 60$



$\mu_X = 200$, $\mu_Y = 300$, $\sigma_X = \sigma_Y = 80$



$\mu_X = 220$, $\mu_Y = 280$, $\sigma_X = \sigma_Y = 40$



Die Hypothese

Wenn man die Vermutung als statistische Hypothese formuliert, lautet sie

$$\mu_X < \mu_Y.$$

Die Gegenvermutung lautet:

$$\mu_X \geq \mu_Y.$$

Ich nenne die Gegenvermutung im Folgenden Nullhypothese und die Vermutung Forschungshypothese (bzw. Alternativhypothese).

Die Grundidee des Signifikanztest I

- Beim klassischen statistischen Testen, gehen wir davon aus, dass die Nullhypothese gilt.
- Die Nullhypothese ist das Gegenteil von der Vermutung die wir nachweisen möchten. Wir gehen also davon aus, dass die Hypothese die wir nachweisen möchten nicht gilt.
- Die Vorgehensweise erinnert an die Unschuldsvermutung im deutschen Strafrecht aus Perspektive der Staatsanalschaft.
- Die Staatsanwaltschaft möchte die Schuld nachweisen, die Schuld entspricht dann der Forschungshypothese. Sie geht von der Unschuld auch, die Unschuld ist die Nullhypothese.

Die Grundidee des Signifikanztest II (vgl. LMLG S. 139)

- Unsere Nullhypothese (Unschuldsvermutung) ist $\mu_X \geq \mu_Y$. Wir gehen davon aus, dass sie die Realität beschreibt.
- Wenn wir nun in der Stichprobe tatsächlich ein Ergebnis beobachten, das im Lichte der Nullhypothese unwahrscheinlich ist, so spricht das *gegen* diese hypothetische Annahme.
- Wir schließen dann, dass die Nullhypothese nicht haltbar ist, mit Bezug auf die Daten.
- Wenn die Nullhypothese nicht haltbar ist, dann wird sie abgelehnt.
- Wenn wir die Hypothese $\mu_X \geq \mu_Y$ ablehnen, dann folgt logische das wir $\mu_X < \mu_Y$ nachgewiesen haben.

Ein Irrtum ist möglich, ...

genauer zwei Irrtümer sind möglich:

1) Wir lehnen die Nullhypothese ab, obwohl sie wahr ist. Man nennt diesen Fehler α-Fehler oder Fehler 1. Art.
2) Wir behalten die Nullhypothese bei, obwohl sie falsch ist. Man nennt diesen Fehler β-Fehler oder Fehler 2. Art.

Vorsicht!

Es ist nicht möglich die Nullhypothese nachzuweisen. Man kann sie nur beibehalten, weil nicht genug empirische Evidenz dagegen spricht.

Was bedeutet signifikant?

- Signifikant ist gleichbedeutend mit über-zufällig.
- Wenn die Daten so stark gegen die Nullhypothese sprechen, dass man sich nicht mehr vorstellen kann, dass die Nullhypothese war ist und nur durch Zufall ein so stark widersprüchliches Ergebnis zu Stande kommt.
- Als Maß dafür zu entscheiden, ob das Ergebnis über-zufällig ist, dient der α-Fehler.
- Damit der Test gültig ist, muss man vor der Durchführung des Tests festlegen, wie groß der α-Fehler höchstens sein darf. Üblich Werte sind 1%, 5% oder 10%.

Irrtumswahrscheinlichkeit als Schutz von H_0

- Wie gehen davon aus, dass der Koch den Klientinnen und Klienten mindestens so viel Fleisch auf den Teller schöpft, wie den Mitarbeiterinnen und Mitarbeitern. In Formalsprache:

$$H_0 : \mu_X \geq \mu_Y$$

- In Analogie: Das ist entspricht der Unschuldsvermutung
- Wir schützen diese Nullhypothese bzw. Unschuldsvermutung durch das Signifikanzniveau.
- Wir können nicht ausschließen, dass wir einen Irrtum begehen. Wir können aber festlegen, dass die Wahrscheinlichkeit des α-Fehlers höchstens 5% beträgt.

Die Bedeutung des Signifikanzniveaus

- Der Begriff Signifikanzniveau ist gleichbedeutend mit dem Begriff Irrtumswahrscheinlichkeit.
- Desto niedriger das Signifikanzniveau (bzw. die Irrtumswahrscheinlichkeit), desto stärker müssen die Daten gegen die Nullhypothese sprechen damit sie verworfen wird.
- Sehr häufig wählt man das Signifikanzniveau 5%.
- Wenn wir noch mehr Sicherheit möchten, um möglichst sicher zu sein, dass wir nicht H_0 verwerfen, obwohl es in Wirklichkeit gilt, dann können wir auch ein Signifikanzniveau von 1% wählen.

Nullhypothese verwerfen

Wenn die Analyse der Daten, also signifikant gegen die Nullhypothese spricht mit einem Signifikanzniveau von 5%, dann können wir dem Koch sagen:

Inhaltliche Bedeutung

Lieber Koch, es kann sein, dass $H_0 : \mu_X \geq \mu_Y$ gilt, aber es ist sehr unwahrscheinlich, dass wir unter Gültigkeit von H_0 die Daten so wie sind beobachten können. Wir gehen davon aus, dass es kein Zufall ist und verwerfen die Nullhypothese. Die Wahrscheinlichkeit für einen Irrtum beträgt 5%.

Wie kann man entscheiden, ab wann man H_0 verwirft?

- Für einen (klassischen) parametrischen statistischen Test, braucht man ein Teststatistik.
- Ich möchte Ihnen heute eine Teststatistik vorstellen, um zu zeigen, wie der Test abläuft.
- Es gibt viele statistische Tests, sie brauchen die entsprechenden Teststatistiken nicht zu kennen.
- Für die Herleitung unserer Teststatistik brauchen wir die Modellannahme der Normalverteilung, also $X \sim N(\mu_X, \sigma_X)$ und $Y \sim N(\mu_Y, \sigma_Y)$. Wir nehmen an X_i iid wie X und Y_i iid wie Y.
- Im Folgenden mache ich die zusätzliche Annahme $\sigma_X = \sigma_Y$.

Die zusätzliche Annahme $\sigma_X = \sigma_Y$.

- Je nachdem ob man diese Annahme macht oder nicht, gibt es zwei unterschiedliche Tests, die auch auf unterschiedlichen Teststatistiken beruhen.
- Ich mache diese Annahme jetzt, um anhand der entsprechenden Teststatistik, zu zeigen, wie man darüber entscheidet, ob man die Nullhypothese beibehält oder nicht.
- Es gibt auch einen stat. Test mit der Nullhypothese $H_0 : \sigma_X = \sigma_Y$.
- Manchmal rechnet man zuerst diesen Test um dann zu entscheiden, ob man von der Gleichheit der Standardabweichungen ausgehen kann oder nicht.
- Natürlich ist auch hier ein Irrtum möglich.

Unsere Teststatistik

Leseempfehlung: (vglm LMLG S. 141- 152)

- Die Nullhypothese H_0 bzw. die Forschungshypothese H_1 beinhalten eine Beziehung zwischen den Erwartungswerten der Normalverteilung μ_X und μ_Y.
- Wir hatten schon im letzten Semester gesehen, dass das arithmetische Mittel $\bar{X}$ bzw. $\bar{Y}$ ein guter Schätzer für μ_X bzw μ_Y ist.
- Mit den Gesetzen der Wahrscheinlichkeitstheorie lässt sich auch nachweisen, dass das arithmetische Mittel als Schätzer für den Erwartungswert bei normalverteilten Daten eine hohe Güte hat.

Die Teststatistik entwickeln I

Die Nullhypothese lautet:

$$H_0 : \mu_X \geq \mu_Y$$

Gleichbedeutend damit ist:

$$H_0 : \mu_X - \mu_Y \geq 0$$

Aus den gerade genannten Gründen ist die Differenz der Mittelwerte

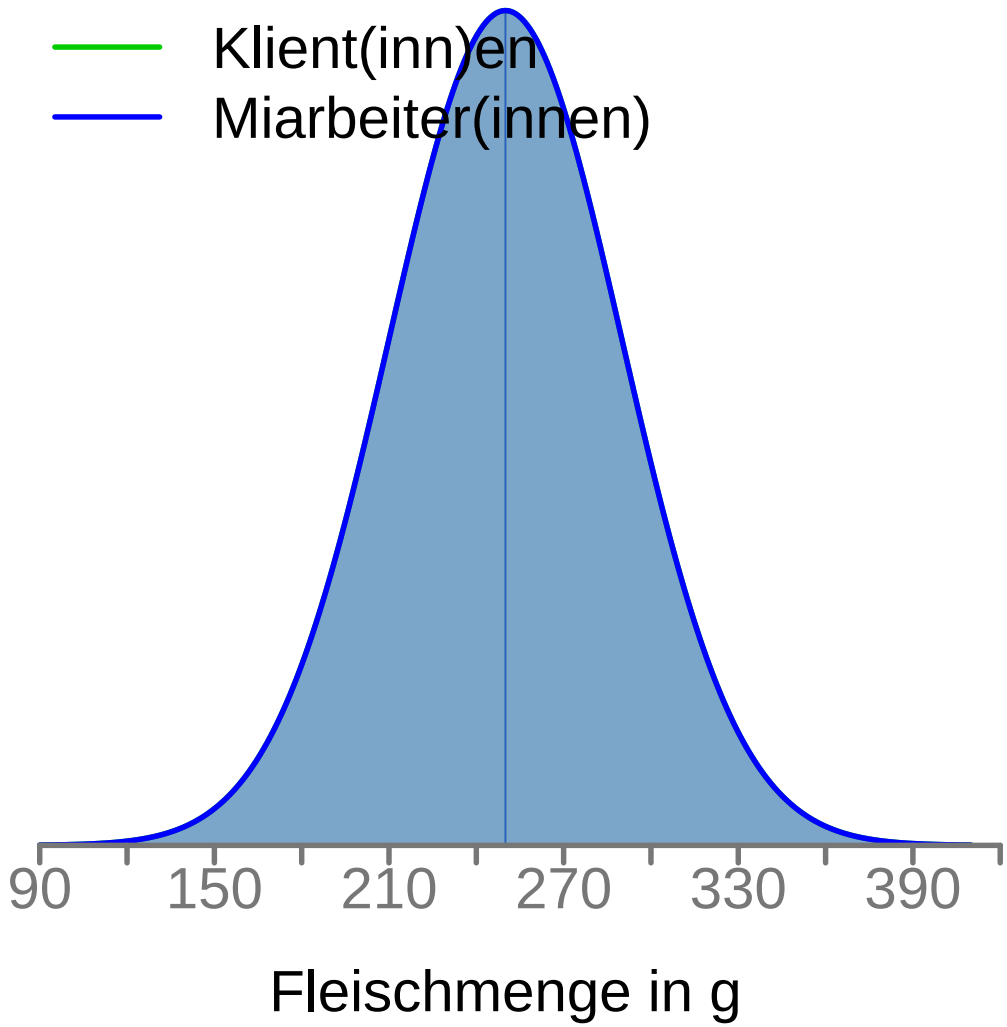
$$\bar{X} - \bar{Y}$$

gut geeignet um die Differenz der Modellparameter zu schätzen.

Die Teststatistik entwickeln II

- Wir werden also aus den Daten $Z := \bar{X} - \bar{Y}$ berechnen und auf das Ergebnis dieser Rechnung unsere Entscheidung über die Ablehnung bzw. Beibehaltung der Nullhypothese stützen.
- Wir müssen dafür wissen, wie sich die Zufallsvariable Z verhält, denn wir müssen wissen, welche Werte von Z gegen die Nullhypothese sprechen und auch wie stark sie dagegen sprechen.
- Ein Problem dabei ist, dass unsere Nullhypothese ein ganzes Intervall abdeckst. Wir ersetzen die Nullhypothese für den Moment durch $\tilde{H}_0 : \mu_X = \mu_Y$ bzw. analog $\tilde{H}_0 : \mu_X - \mu_Y = 0$
- Im Klartext bedeutet die modifizierte Nullhypothese, dass wir von völlig identischen Verteilungen der Fleischmengen pro Portion Gulasch ausgehen.

Die modifizierte Nullhypothese grafisch



Wir gehen davon aus, dass die Nullhypothese zutrifft...

...also davon, dass der Schöpfmechanismus beim Gulasch Schöpfen sich in den beiden Personengruppen gar nicht unterscheidet.

$$Z := \bar{X} - \bar{Y}$$

- Ich simuliere die Datenerhebung . In der Simulation, wird die Fleischmenge von je 100 Personen der beiden Gruppen erfragt. In der Simulation gehe ich davon aus, dass kein Unterschied in der Verteilung der Fleischmenge besteht.
- Ich gehe davon aus, dass die erwarte Fleischmenge 250 g beträgt, mit einer Standartabweichung von 40 g. (Das entspricht der Abbildung auf der vorigen Folie)

Die Verteilung der Teststatistik $Z := \bar{X} - \bar{Y}$

- Wir haben jetzt Daten, von 10 Personen jeder Gruppe, wir berechnen jeweils den Mittelwert.
- Die mittlere Fleischmenge $\bar{x}$ ergibt sich zu 252.1 g.
- Die mittlere Fleischmenge $\bar{y}$ ergibt sich zu 247.8 g.
- Wir beobachten daher $z = \bar{x} - \bar{y} = 4.3$ g.
- Finden Sie dass ein Unterschied von 4.3 g als Hinweis darauf gesehen werden kann, dass der Koch unfair schöpft?
- Welche Werte würden dafür sprechen, dass der Koch unfair schöpft? Werte die größer oder kleiner sind als Null?
- Um wie viel muss der Wert unter Null liegen, dass sie sagen würden: "Das kann kein Zufall sein."?

Wir wiederholen das Experiment

- Wir stellen uns vor, dass wir einen weiteren Monat die Gulasch-Ausgabe beobachten. Wieder gilt die Nullhypothese. Der Koch schöpft also gerecht.
- Wenn wir die Daten simulieren, dann wissen wir, dass der Koch fair schöpft. In der Realität beobachten wir einfach die Daten und wir wissen nicht, ob die Nullhypothese gilt.
- Die mittlere Fleischmenge $\bar{x}$ ergibt sich zu 237.6 g.
- Die mittlere Fleischmenge $\bar{y}$ ergibt sich zu 252.0 g.
- Wir beobachten daher $z = \bar{x} - \bar{y} = -14.4$ g.
- Wenn Sie nicht wüssten, dass die Daten unter Gültigkeit der Nullhypothese entstanden sind, wie würden sie diesen Wert beurteilen?

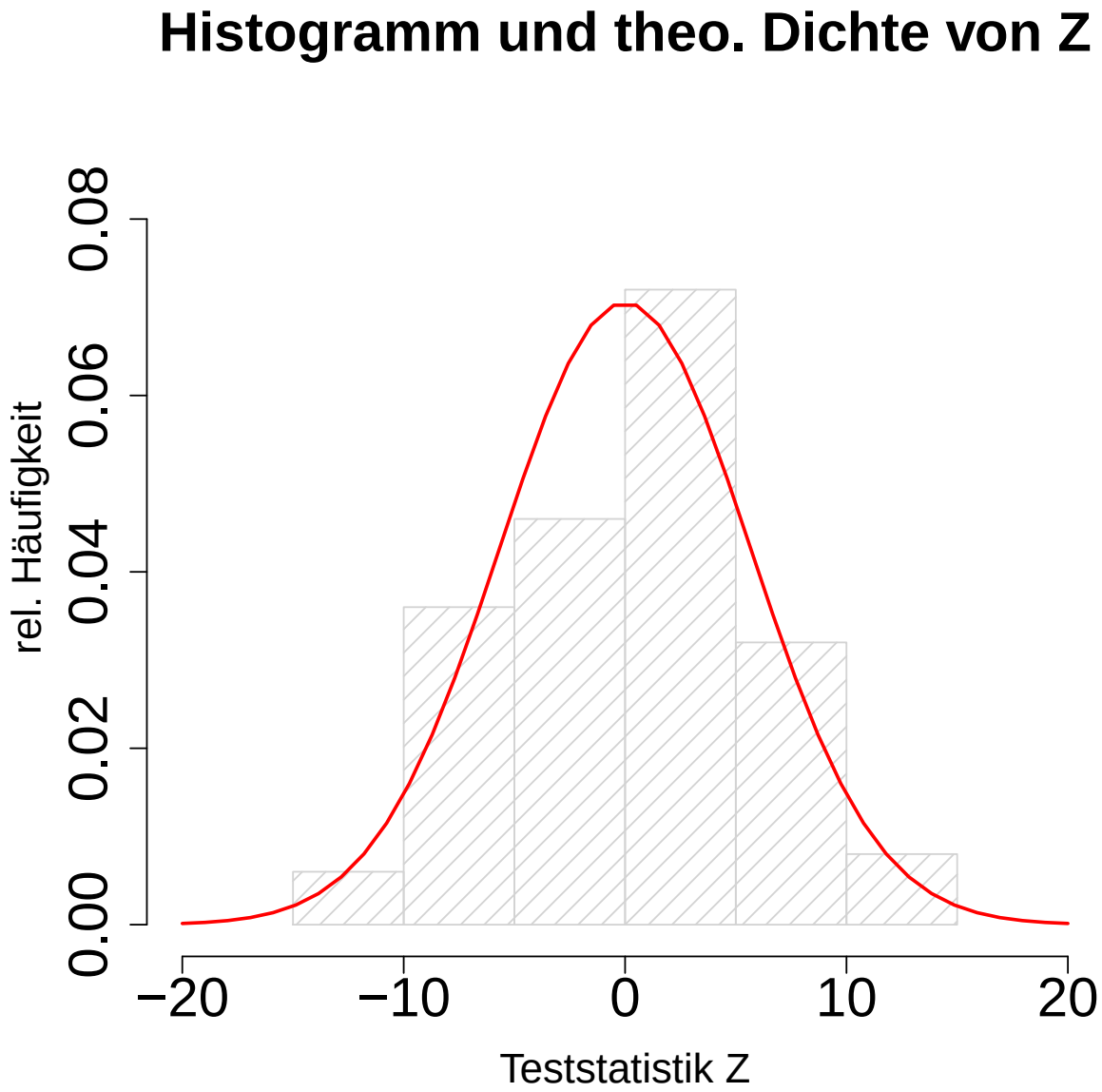
Ein kritischer Wert wird gebraucht

- Wir haben gesehen, auch, wenn die Nullhypothese gilt, kann es passieren, dass Z Werte annimmt, die deutlich unter Null liegen und uns nahe legen, zu glauben, dass der Koch unfair schöpft.
- Wenn wir dann die Nullhypothese verwerfen, begehen wir den Fehler 1. Art (α-Fehler)
- Für diesen Fehler hatten wir vereinbart, dass er einen Irrtumswahrscheinlichkeit von 5% nicht übersteigen soll.
- Wir müssen den kritischen Wert also so wählen, dass die Wahrscheinlichkeit die Nullhypothese abzulehnen, wenn der Koch in Wirklichkeit fair ist, höchsten 5% beträgt.

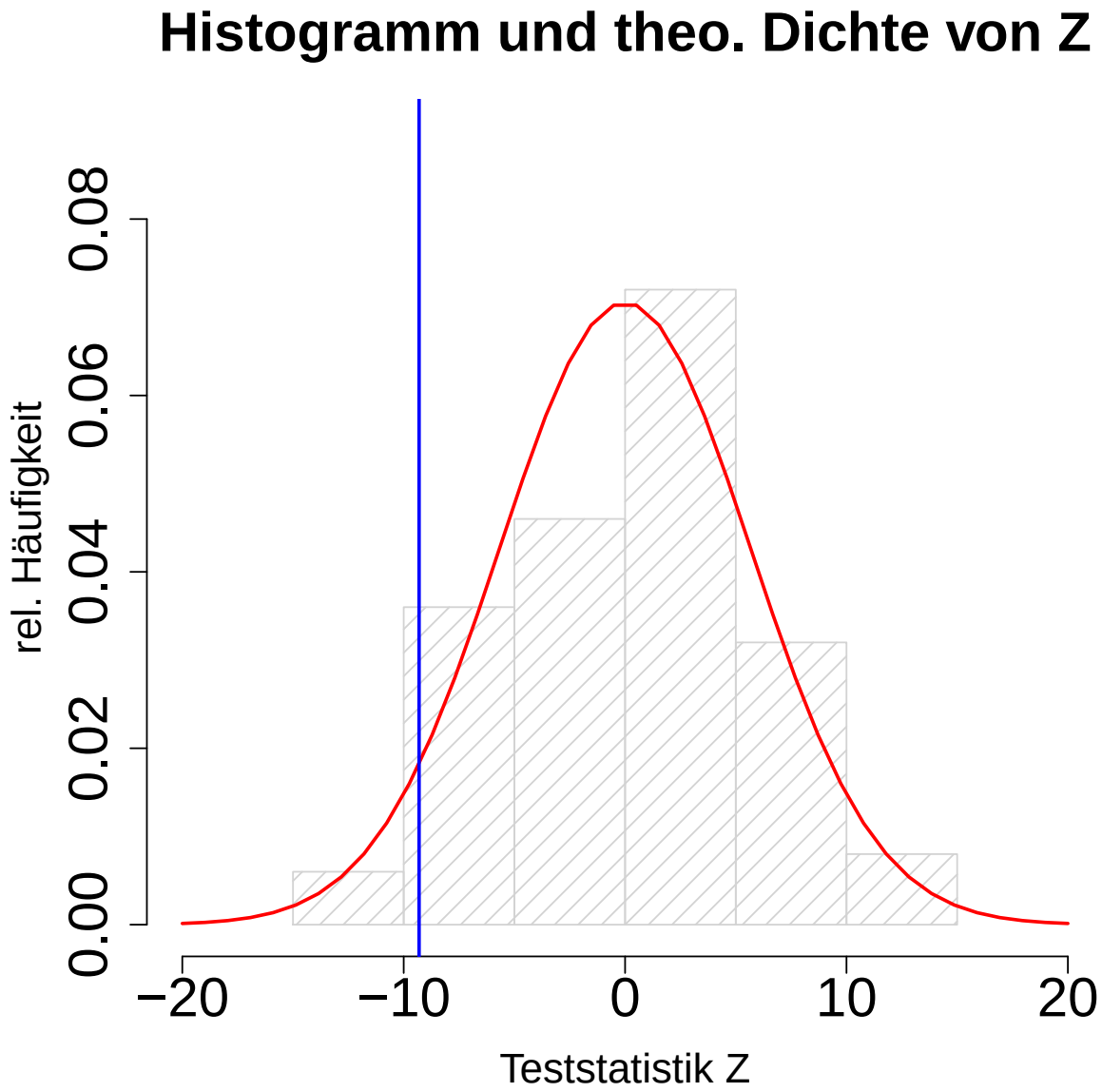
Der kritische Wert ist $k := -9.3$

- Der Wert, der die Eigenschaft hat, in unserem Beispiel, dass die Irrtumswahrscheinlichkeit gleich 5% beträgt ist -9.3
- In Worten: Unter der Annahmen dass der Koch gerecht schöpft die die Wahrscheinlichkeit, dass bei je 10 Portionen Gulasch pro Gruppe, die Differenz der Mittelwerte kleiner als -9.3 ist, höchstens 5%
- Was sie verstehen sollten ist:
  - Durch die Irrtumswahrscheinlichkeit ist der kritische Wert festgelegt.
  - Sie brauchen ihn nicht bestimmen zu können.
  - Sie sollten verstehen, welche Bedeutung er hat.

Verteilung der Teststatistik Z



Verteilung der Teststatistik Z und kritischer Wert k



Die Rolle der Stichprobenumfänge n_1 und n_2.

- Desto mehr Informationen man hat, desto kleiner sind die Unterschiede die der Test aufdecken kann.
- Nehmen wir an, die Differenz $\delta := \mu_x - \mu_y$ beträgt in Wirklichkeit nur $10g$, dann braucht man einen große Stichprobenumfänge n_1 und n_2 um diese kleine Differenz aufdecken zu können.
- Wenn δ einen Wert von 100 g hat wird bereits ein kleiner Stichprobenumfang von z. B. 10 Personen pro Gruppe zu einer Ablehnung der Nullhypothese führen.
- Merksatz: Desto stärker die Nullhypothese verletzt ist, desto weniger empirische Information braucht man, um die Forschungshypothese nachzuweisen.

Von der modifizieren Nullhypothese zu Nullhypothese

Gulasch-Beispiel

Ein paar Anmerkungen:

- Die Nullhypothese und die Forschungshypothese sind in ihrer Bedeutung unterschiedlich.
- Es gibt zwei Möglichkeiten sich zu irren: Den Fehler 1. Art und den Fehler 2. Art. Auch diese werden unterschiedlich behandelt.
- Der Fehler 1. Art wird a-priori festgelegt und durch ihn wird der kritische Wert bestimmt.
- Beruhend auf dem kritischen Wert können wir entscheiden, ob wir glauben, dass der Koch frair ist oder nicht. Wir wissen wir hoch die Wahrscheinlichkeit ist, ihn zu unrecht als ungerecht einzustufen.
- Die Verteilung der Teststatistik Z hängt von den Stichprobenumfängen n_1 und n_2 ab.

Zusammenfassung: statistisches Testen allgemein

1) Wenn wir die Modellannahmen formuliert haben, und die Hypothesen aufgestellt, dann suchen wir nachdem Test, der für dieses Problem geeignet ist. Das Signifikanzniveau wird festgelegt.
2) Dieser Test hat eine Teststatistik. Die Teststatistik, legt fest, welcher Wert aus den Daten berechnet wird, um die Testentscheidung zu treffen.
3) Durch die Festlegung der Irrtumswahrscheinlichkeit (=Signifikanzniveau) und der des Auswahl der Teststatistik, ergibt sich ein kritischer Wert.
4) Die Statistik-Software berechnet diesen Wert und informiert uns darüber ob wir die Nullhypothese beibehalten, oder ob sich die Forschungshypothese nachweisen lässt.

Übung

Wir werden für einige Testproblem die häufig auftreten die geeigneten Test auswählen, mit STATA rechnen und die Ergebnisse interpretieren. Wir behandeln:

- Approximative Binominaltest
- χ^2 Unabhängigkeitstest
- Test zu Lagealternativen im Ein- und Zwei-Stichprobenfall:
  - Ein-Stichprobenfall: einfacher Gaußtest und t-Test
  - Zwei-Stichprobenfall (unabhängige Stichproben): doppelter Gauß- und t-Test, Welch-Test