Stochastik für die Naturwissenschaften

Dr. C.J. Luchsinger

9. TESTEN VON HYPOTHESEN

Ziele dieses Kapitels:

Das gesamte Setting (die Philosophie, 9.1) rund um das Testen muss verstanden werden. Ausgehend von den in der Vorlesung behandelten Situationen sollten Sie auch fähig sein, Aussagen über statistische Tests in komplizierteren Modellen, welche Sie (noch) nicht kennen, nachzuvollziehen. Zuerst ein paar Vorbemerkungen:

9.1 Philosophie hinter den Tests / Problemstellung

9.1.1 Philosophie hinter den Tests

Frage an's Publikum: (abgesehen von ethischen Fragen), was ist die Grenze der akademischen Forschungsfreiheit? Genauer: Welche Theorien dürfen Sie in den *Naturwissenschaften* vertreten und welche nicht?

Wie forschen wir in den Naturwissenschaften?

Von Montag bis Samstag betreiben Sie bitte seriöse, konkrete Arbeit - aber am Sonntag darf es auch ein bisschen Philosophie, hier Erkenntnis- und Wissenschaftstheorie (Theory of Knowledge), sein.

Es folgen zusammengefasst die Hauptideen von Popper und ein paar anderen Leuten. Wem das zu theoretisch ist: die Kriterien von Bradford-Hill sind im Gegensatz zu unterem praktisch direkt anwendbar für die Forschung: https://en.wikipedia.org/wiki/Bradford_Hill_criteria . Wer radikal anders denken will, fängt am Besten mit Paul Feyerabend an: https://de.wikipedia.org/wiki/Paul_Feyerabend . In kürzester Zusammenfassung sagte er als Methodenanarchist: "anything goes". Feyerabend bezeichnete das Nachfolgende auch als "Law-and-Order-Rationalismus".

Grundlagen

* saubere Fundierung durch Karl Raimund Popper (1902-1994); ausserhalb der Naturwissenschaften umstritten, innerhalb der Naturwissenschaften unbestritten.

* unbekannte Naturgesetze; gehen davon aus, dass sie existieren, konstant in Raum und Zeit und unabhängig von Beobachter.

* "Mein Teddybär kann reden - aber nur, wenn Du nicht im Zimmer bist" - Kinder sind Kinder! Wir untersuchen stattdessen messbare Grössen (in Kilogramm, Sekunden, Metern, etc) und diese sind damit allgemein nachvollziehbar kommunizierbar (historische Fachbegriffe: "intersubjektive Kontrollierbarkeit und Mitteilbarkeit"). Das muss in den Naturwissenschaften erfüllt sein.

* aus obigem folgt die Forderung, dass die Resultate/Daten von Experimenten von Anderen reproduziert werden können. Berechtigte Einwände: statistische Messungenauigkeiten, Mangel an Zeit und Geld, Unvollkommenheit der technischen Hilfsmittel, schlechte Protokollierung. Zudem können verwendete Begriffe im Wandlungsprozess der Wissenschaften eine Abänderung erfahren. Diesen Problemkreis, der eng mit der verwendeten Sprache verwoben ist, nennt man "Basisproblem": wir wollen eine gemeinsame, dauerhafte Basis haben, auf der wir die Resultate formulieren können.

* Gravitation:

  "Jeder Stein fällt nach unten" > nicht *verifizierbar*

  "Jeder Stein fällt nach oben" > *falsifizierbar*

Dieses Prinzip der Falsifikation kannten Forschende bereits spätestens in der 2. Hälfte des 19. Jahrhunderts (überliefert) wenn nicht gar vorher.

* wir kennen nur Theorien und die Resultate von objektiven Experimenten (kontrolliert) und Beobachtungen (ohne Kontrolle).

* Beobachtungen oder Idee > Theorie > ergebnisoffenes Experiment zur Überprüfung > entweder falsifiziert oder *bestätigt* (aber nicht *bewiesen* - das ist ein Problem in der Wissenschaftskommunikation, zum Beispiel in der Klimaforschung).

Jargon

* wir beobachten, dass jeder Stein bis jetzt nach unten fällt ⇒ Theorie (=Hypothese): "Jeder Stein fällt nach unten". Dieser Schritt ist eine sogenannte *Induktion* (vom Einzelnen zum Allgemeinen, Entwicklung einer Theorie auf Basis empirischer Befunde). Falsch, sobald auch nur ein Stein nicht nach unten fällt (falsifizierbar).

"Jeder Stein fällt nach unten" ist ab dann eine Hypothese, mit der wir arbeiten, solang sie nicht widerlegt wurde. Wir schliessen damit vom allgemeinen Grundsatz "Jeder Stein fällt nach unten" für jeden Einzelfall erneut, dass ein konkreter Stein auch nach unten fallen wird. Das ist eine sogenannte *Deduktion* (vom Allgemeinen zum Besonderen, Ableitung einer überprüfbaren Hypothese aus der Theorie).

* vorläufige Resultate, solange nicht widerlegt. Wissenschaftliche Debatten sind *in den Naturwissenschaften* deshalb *nie* "abgeschlossen".

* wir suchen nach Gesetzen, *Kausalität*, also Ursache und Wirkung.

* Falsifizierung bedeutet Fortschritt in der Forschung (Wissenschaft als Fehlerentsorgungsunternehmen; "Die Theorien von heute sind die Irrtümer von morgen"). Das hat psychologisch auch den Aspekt eines Vater-Sohn-Konflikts; das Alte vertreten und verteidigen die Alten und die Jungen wollen Neues beweisen - ein Treiber des Fortschritts!

* Mathematik ist die einzige Wissenschaft, in der man etwas wirklich sicher beweisen kann. Alles andere ist unsicher; wir arbeiten stattdessen mit Hypothesen. Mathematik/Logik ist eine sogenannte *a priori*-Wissenschaft (a priori = von vornherein, unsere Aussagen können sich nicht durch Erfahrung auf einmal als falsch erweisen). Wir haben in der Mathematik gut gewählte Axiome, welche wir nicht mehr bei jedem Schritt weiter hinterfragen. Sie gelten von Start weg (a priori).

* Die Naturwissenschaften bauen auf der Erfahrung auf, man spricht von *Empirie, empirische Wissenschaft* (vom Griechischen etwa für "Erfahrungswissen"); es ist eine *a posteriori*-Wissenschaft. A posteriori = im Nachhinein, weil diese Aussagen sich nur durch Erfahrung bewähren können (Bestätigung bzw. Falsifikation).

Bemerkungen

* Akzeptieren Sie Ihr Unwissen und diese Unsicherheit mit Demut; im Gegensatz zu Ingenieur/innen und Mechaniker/innen, welche eine Maschine bauen und genau wissen, was sie warum dort einbauen, wissen wir nicht, was alles von der Biochemie im menschlichen Körper oder von der Physik in der Atmosphäre am wirken ist (und das *warum* macht eventuell kaum Sinn - *Ziele* haben wir in den Naturwissenschaften erst recht nicht!).

* eine kleine Warnung: Wissenschaftler wie alle Menschen neigen dazu, nicht nur allgemeine Hypothesen, sondern auch ihre Intuitionen und vorgefassten Meinungen durch wissenschaftliche Methoden zu bestätigen - das ist nicht per se falsch.

* Hypothesen müssen falsifizierbar, überprüfbar sein (sonst ist es Meta-Physik) und die Beweislast liegt beim Hypothesen-Ersteller und nicht umgekehrt (Bringschuld)! Es kann aber sein, dass die zur Überprüfung notwendigen Anlagen noch nicht gebaut oder zu teuer sind oder aus anderen Gründen (noch) nicht zur Verfügung stehen. Dann muss mindestens angegeben werden können, wie die Theorie überprüft werden *könnte*.

* Religion und Naturwissenschaft / Popper: Die Existenz einer übersinnlichen Gottheit oder die Frage eines freien Willens sind mit obigem Setting wohl nicht entscheidbar. Für

die Naturwissenschaften ist der sogenannte "Deismus" als Glaubensrichtung am unproblematischsten: Gott hat "die Welt" erschaffen, greift jetzt aber nicht mehr weiter ein - und verletzt damit nicht (die von ihm geschaffenen) Naturgesetze. In Religionsfragen bleibt nur noch *glauben zu müssen statt wissen zu können.* Deshalb gibt es auch zwei Verben dazu: *glauben* versus *wissen*!

Bei der Frage des freien Willens *kann* man zudem aus naturwissenschaftlicher Sicht kritisieren, dass die Frage falsch/unpräzis gestellt wurde. "Freier Wille" ist ja weder messbar noch brauchbar kommunizierbar. In diesem Sinne gibt es dann zwar falsche Fragestellungen ("Gibt es einen freien Willen?"), aber ansonsten keine unlösbaren Rätsel in den Naturwissenschaften: wenn sich eine Frage überhaupt stellen lässt, so kann sie grundsätzlich auch beantwortet werden.

* Popper lieferte Wissenschaftstheorie, "wie sollen wir forschen?", zumindest in den Naturwissenschaften (Buchtitel: Logik der Forschung). Der Film "Johannes Kepler, der Himmelsstürmer" 2020, zeigt auch obige Probleme im Wissenschaftsbetrieb.

Und jetzt noch ein paar Bemerkungen zu unseren speziellen Freunden

* Wir richten unsere Theorien an der Welt aus (a posteriori); Verschwörungstheoretiker bauen sich die Welt nach ihrer Theorie (a priori), also umgekehrt (grosse Phantasie - wenig Selbstkritik). Verschwörungstheoretiker kann man meist einfach reden lassen, sie schiessen sich normalerweise selber ab. Es gibt aber noch den viel clevereren *Denialismus*: im Einklang mit Popper und dem hier vorgestellten Ansatz streichen sie einseitig und mit einer hidden Agenda Unsicherheit in der Forschung hervor (die gibt es ausserhalb der Mathe immer!), fordern immer mehr Forschung (schön); es ist ein *interessengeleitetes* Zersetzen des Forschungskonsenses (Klima- und Covidforschung). Wenn Sie mit solchen Leuten zu tun haben, machen Sie mindestens im Stillen den sogenannten "ÄrztInnentest": würde die Person auch so argumentieren, wenn sie selber eine potentiell lebensbedrohliche Krankheit hätte (Krebs, Herzprobleme) und es um deren Behandlung ginge?

* Merke: "Metaphysische Termini trennen - wissenschaftliche Termini verbinden." (Otto Neurath; Mitglied des Wiener Kreises (1922-1936))

https://schweizermonat.ch/beweisen-wie-die-alten-griechen/
https://schweizermonat.ch/bewiesenermassen-unmoeglich/
https://schweizermonat.ch/der-beweis-nach-158-jahren/
https://schweizermonat.ch/donald-das-very-stable-mathematik-genie/
https://www.math.uzh.ch/vorkurs , VK 2 - da lernen Sie, was ein Beweis ist...

9.1.2 Problemstellung

Stellen wir uns vor, wir sind in einer Quizshow: Gegeben sind zwei *mögliche* Hypothesen, wir nennen sie Nullhypothese $\mathcal{H}_0$ (zB $p = 0.5$) und Alternativhypothese (oder Gegenhypothese) $\mathcal{H}_1$ ($p \neq 0.5$) . Entweder ist die Hypothese $\mathcal{H}_0$ richtig oder die Hypothese $\mathcal{H}_1$. Wir wissen leider nicht, welche Hypothese richtig ist. Wir erhalten dann eine Realisation x_1 einer Zufallsgrösse. Je nachdem, welche Hypothese richtig ist, wird die Zufallsgrösse aber anders verteilt sein (z.B. anderer Mittelwert). Wir müssen uns dann entscheiden, ob wir sagen wollen: "wir nehmen **vorläufig / provisorisch** $\mathcal{H}_0$ an" oder "wir nehmen **vorläufig / provisorisch** $\mathcal{H}_1$ an". 4 Situationen sind möglich; auswendig:

1. $\mathcal{H}_0$ ist richtig und wir nehmen **vorläufig / provisorisch** $\mathcal{H}_0$ an.
2. $\mathcal{H}_0$ ist richtig und wir nehmen **vorläufig / provisorisch** $\mathcal{H}_1$ an (Fehler 1. Art).
3. $\mathcal{H}_1$ ist richtig und wir nehmen **vorläufig / provisorisch** $\mathcal{H}_1$ an.
4. $\mathcal{H}_1$ ist richtig und wir nehmen **vorläufig / provisorisch** $\mathcal{H}_0$ an (Fehler 2. Art).

Entscheid \ Realität	$\mathcal{H}_0$	$\mathcal{H}_1$
ich sage "$\mathcal{H}_0$"		
ich sage "$\mathcal{H}_1$"		

analog Rechtsprechung:

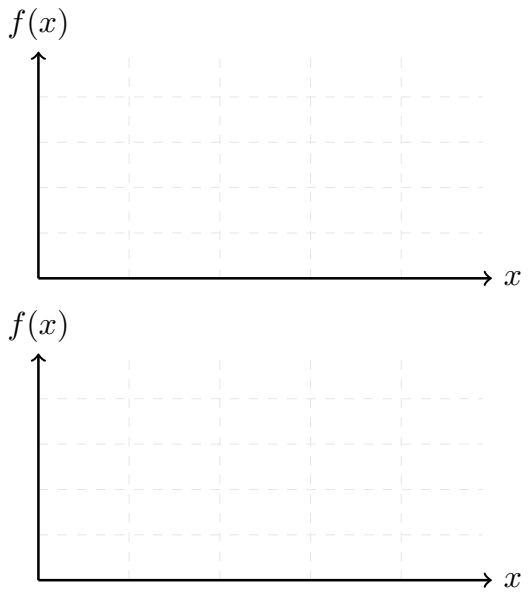
* real: ist Mörder vs ist nicht Mörder
* Gericht: schuldig vs Nicht-Schuldig
* "Nicht-Schuldig" präferiert wegen "in dubio pro reo"

"Nehmen $\mathcal{H}_0$ an" ist eigentlich, später in der Forschung, heikel als Formulierung. Grund: $\mathcal{H}_0$ wird nie *bewiesen*. Hier aber ok, da Spiel und pädagogisch sinnvoll.

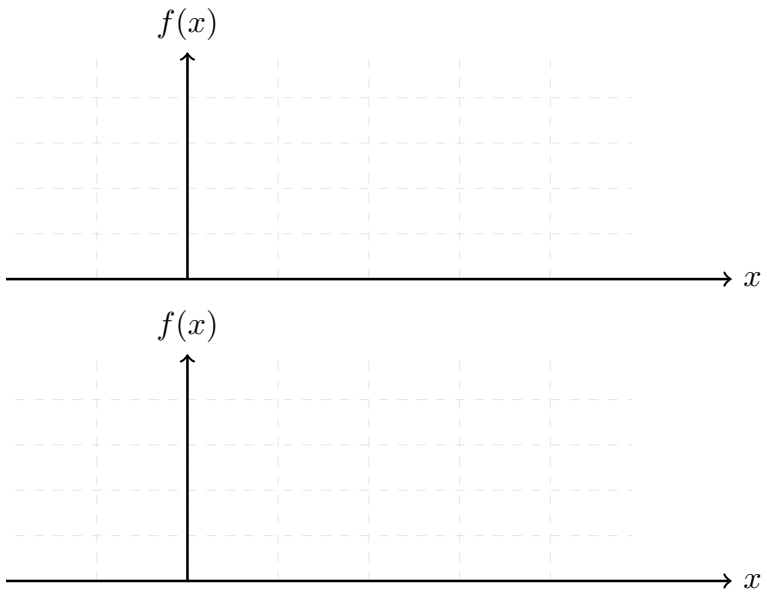
Beispiele für $\mathcal{H}_i$ sind:

* $\mathcal{H}_0$: Würfel fair (1/6) gegen $\mathcal{H}_1$: Würfel verfälscht (viele Möglichkeiten)
* $\mathcal{H}_0$: Münze fair (1/2) gegen $\mathcal{H}_1$: Münze verfälscht (viele Möglichkeiten)
* $\mathcal{H}_0$: Medikament lässt Blutdruck unverändert gegen $\mathcal{H}_1$: Medikament ändert Blutdruck
* (Lohn-)Umfrage unter 10'000 Personen, schliesse auf Gesamtpopulation: $\mathcal{H}_0$: Durchschnittslohn gleich geblieben gegen $\mathcal{H}_1$: Durchschnittslohn anders

Erste Aufgabe Gegeben sei eine Realisation x_1 einer Zufallsgrösse aus einer $U[a, b]$-Verteilung. Im Fall $\mathcal{H}_0$ ist dies eine $U[0, 1]$-Verteilung, im Fall $\mathcal{H}_1$ ist es eine $U[2, 3]$-Verteilung. Wir müssen uns jetzt mit Beobachtung x_1 entscheiden, ob wir im Fall $\mathcal{H}_0$ oder $\mathcal{H}_1$ sind:



Zweite Aufgabe (Ziel: **Denken in Hypothesen**) Gegeben sei eine Realisation x_1 einer Zufallsgrösse aus einer $\mathcal{N}(\mu, \sigma^2)$-Verteilung. $\sigma^2 = 1$ gelte auf jeden Fall; aber: im Fall $\mathcal{H}_0$ ist dies eine $\mathcal{N}(0, 1)$-Verteilung, im Fall $\mathcal{H}_1$ ist es eine $\mathcal{N}(1, 1)$-Verteilung. Wir müssen uns jetzt in dieser Quizshow mit dieser einzigen Beobachtung x_1 entscheiden, ob wir im Fall $\mathcal{H}_0$ oder $\mathcal{H}_1$ sind:



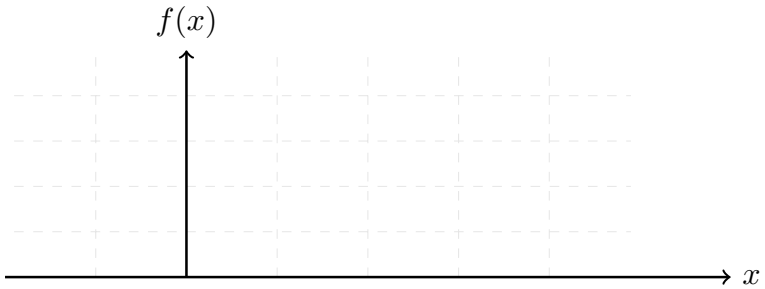
Dritte Aufgabe Wir behandeln nochmals die zweite Aufgabe mit folgenden Auflagen:

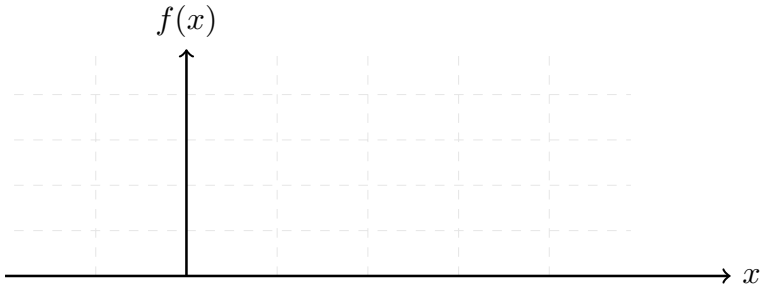
a) Wenn wir uns für $\mathcal{H}_1$ entscheiden, obschon $\mathcal{H}_0$ richtig ist, verlieren wir all unser Vermögen (in den 3 anderen Möglichkeiten geschieht nichts). Wie sollten wir uns jetzt entscheiden?

b) Wenn wir uns für $\mathcal{H}_0$ entscheiden, obschon $\mathcal{H}_1$ richtig ist, verlieren wir all unser Vermögen (in den 3 anderen Möglichkeiten geschieht nichts). Wie sollten wir uns jetzt entscheiden?

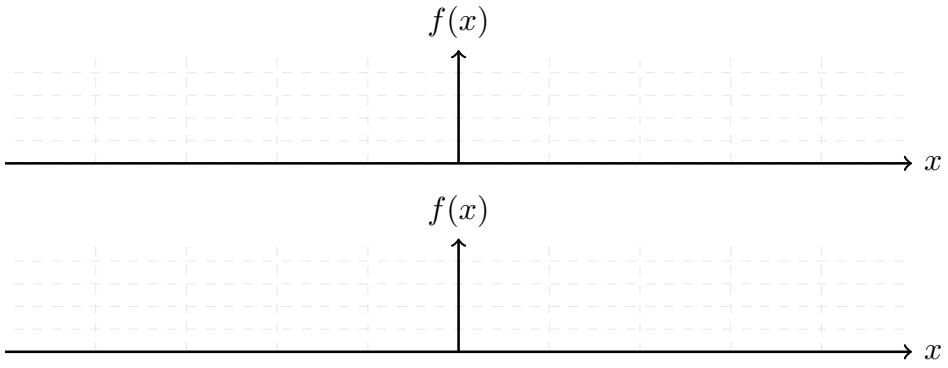
Vierte Aufgabe Philosophisch geht es in dieser Aufgabe um folgendes: konservativ, die bisherigen Resultate wertschätzend, glauben wir aktuell noch, dass $\mathcal{H}_0$ richtig ist, sind aber offen für neues, für Fortschritt, den wir aber genau prüfen wollen. Fortschritt ($\mathcal{H}_1$) geht nicht ohne Risiko (Fehler 1. Art); konservativer Stillstand ist aber auch Risiko (Fehler 2. Art). Wir lassen also einen Spalt offen (10 %), wo Fortschritt passieren kann - aber möglichst sinnvoll, damit sich das eingegangene Risiko auch lohnt.

Wir behandeln nochmals die zweite Aufgabe mit folgender Auflage: Wenn $\mathcal{H}_0$ richtig ist, dürfen wir (ohne Strafe) mit $\alpha = 10\%$ Wahrscheinlichkeit eine falsche Entscheidung treffen (sagen: "$\mathcal{H}_1$ ist richtig", Fehler 1. Art). Wie sollten wir uns jetzt entscheiden?

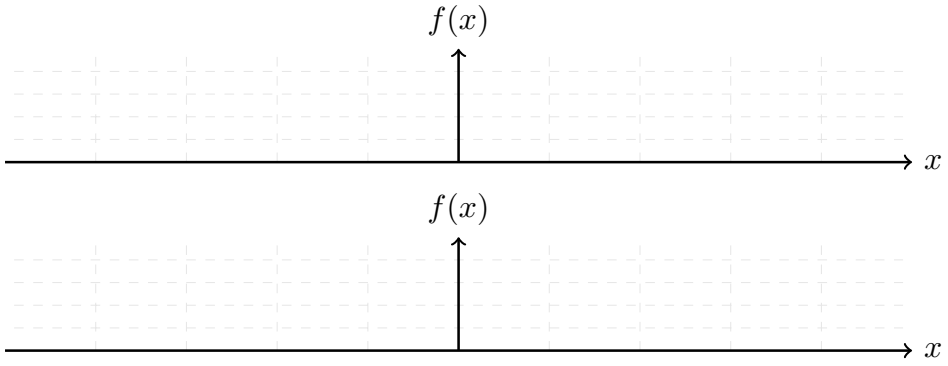




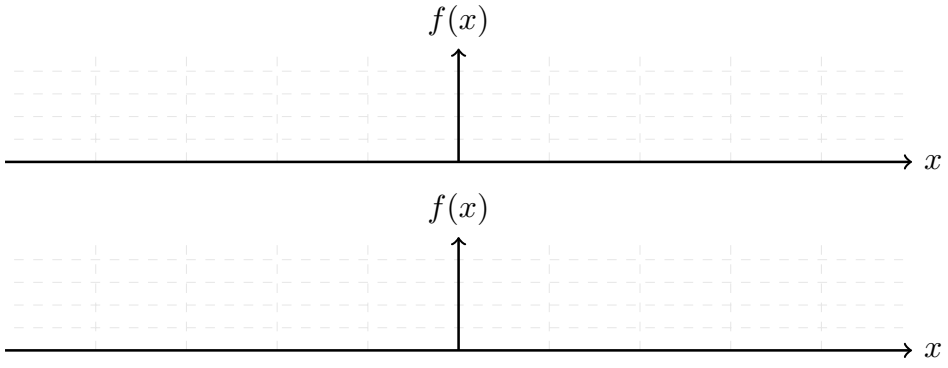
Alternative Vorschläge:



***



***



Fünfte Aufgabe Wir wollen die vierte Aufgabe zu einer Optimierungsaufgabe umformulieren. Wir haben gesehen, dass wir ∞-viele Möglichkeiten haben, diese 10 % der Fälle "auf der x-Achse zu platzieren" (= sog. Ablehnungsbereich (engl. Critical Region), wir lehnen dort die $\mathcal{H}_0$-Hypothese ab). Wir sollten deshalb noch den Fehler 2. Art mit einbeziehen. Ziel ist es, "**pro Fehler 1. Art, den man macht, ein Maximum an Fehler 2. Art vernichten!**" Wie sollten wir uns jetzt entscheiden?

Mathematische Statistik: sogenanntes Neyman-Person-Lemma sagt uns: rechts aussen ist optimal (1. Vorschlag). Alle 4 Vorschläge haben gleiche W'keit für Fehler 1. Art, sogenanntes $\alpha = 10\%$, damit fairer Vergleich, aber 1. Vorschlag hat kleinste W'keit für Fehler 2. Art β (siehe p+1 für α, β-Definition).

***

ganz MAT183, Prüfung: Luchs sagt, was optimal und angesagt!

***

Wichtige Warnung: Folgende Aussage ist (obschon häufig gehört) schwachsinnig: "Mit 90 % Wahrscheinlichkeit ist die Hypothese $\mathcal{H}_0$ die richtige." Richtig ist: Entweder ist $\mathcal{H}_0$ die richtige Hypothese oder $\mathcal{H}_1$. Wir werden aber nie (Ausnahmen sind triviale Fälle wie die erste Aufgabe) wirklich wissen, in welcher Situation wir sind. Wir werden uns in 10 % der Fälle, wo eigentlich $\mathcal{H}_0$ richtig wäre, für $\mathcal{H}_1$ entscheiden, weil uns dies erlaubt wurde. Wir machen dies aber möglichst sinnvoll (vgl. fünfte Aufgabe).

Die bisherigen Aufgaben waren sehr künstlich, weil wir nur $n = 1$ gehabt haben. In der Praxis haben wir aber hoffentlich ein grosses n. Wenn wir eine Münze 1'000 mal werfen und die Anzahl Kopf zählen, so *erwarten* wir eine Zahl wie 500. Aber es gibt Variabilität in den Daten! *Selbst bei einer fairen Münze* (d.h. $\mathcal{H}_0$) wird es um die 500 herum schwanken. **Wie weit darf es aber schwanken, bis wir stutzig werden (sollten)?** Diese letzten paar Sätze stammen aus dem Intro zu dieser Vorlesung. Schauen Sie doch mal, wieviel Sie mittlerweile gelernt haben:

$\sum_{i=1}^{n} x_i$ ist Anzahl Kopf, wenn $x_i = 0$ bei Zahl und $x_i = 1$ bei Kopf.

Ich sagte damals, im Intro: $[469, 531]$ ist ok als Schwankungsbreite, denn wenn $p = 0.5$, liegt mit 95 % W'keit die Anzahl Kopf in $[469, 531]$ und mit 5 % W'keit die Anzahl Kopf ausserhalb $[469, 531]$. Dies folgt aus:

CLT:

* 95 % W'keit in $\mu \pm 2\sigma$ wenn $\mathcal{N}(\mu, \sigma^2)$

* gilt etwa auch bei $\text{Bin}(n, p)$ wegen CLT, $n = 1000, p = 0.5$

* aus $\mu \pm 2\sigma$ wird

* bei Bin $np \pm 2\sqrt{np(1-p)} = 500 \pm 2\sqrt{1000 \cdot 0.5 \cdot 0.5} = 500 \pm 2\sqrt{250} \doteq [469, 531]$

Eine solche Testsituation nennt man *zweiseitig*.

Jargon und Bemerkungen:

1. Obiges $\sum_{i=1}^{1000} x_i$, $\overline{x}$, nennt man **Teststatistik**; das ganze einen **statistischen Test**.

2. α ist Wahrscheinlichkeit für Fehler 1. Art (auch Risiko 1. Art oder Grösse des Tests oder Signifikanzniveau, engl. Size). Diese Grösse ist gegeben und eine schöne Zahl. Bei Aufgabe 4 und 5 hatten wir $\alpha = 10\%$; beim Münzwurf $\alpha = 5\%$.

3. β ist Wahrscheinlichkeit für Fehler 2. Art (auch Risiko 2. Art). Macht (engl Power) ist $1 - \beta$. β wollen wir - gegeben α - gerne klein (anders formuliert: wir wollen eine grosse Macht), siehe Aufgabe 4.

4. Signalwörter für $\mathcal{H}_0$ sind: bisherige Stand des Wissens; gerecht/fair $(1/2, 1/6)$, symmetrisch, unspektakulär/langweilig ("konform, uniform, chloroform"). In MAT183 gilt: wenn die eine Hypthese "einfach" ist und die andere "zusammengesetzt" (siehe Punkt 7 unten), dann ist $\mathcal{H}_0$ die einfache Hypothese (zum Beispiel $\mathcal{H}_0 : p = 0.5$ vs $\mathcal{H}_1 : p \neq 0.5$).

5. Bemerkungen zu einseitig vs zweiseitig am Beispiel des Münzwurfs:

Zweiseitig: Wegen 4. werden wir üblicherweise $\mathcal{H}_0$ so wählen, dass dort $p = 0.5$ (symmetrisch, langweilig, gerecht/fair). $\mathcal{H}_1$ ist dann $p \neq 0.5$. Das ist dann eine zweiseitige Testsituation. Signalwörter sind "hat Mondstand einen *Einfluss* auf die Wahrscheinlichkeit" oder "gibt es wegen des Mondstandes eine *Änderung* der Wahrscheinlichkeit"? Ein Einfluss bzw. eine Änderung kann ja auf beide Richtungen sein.

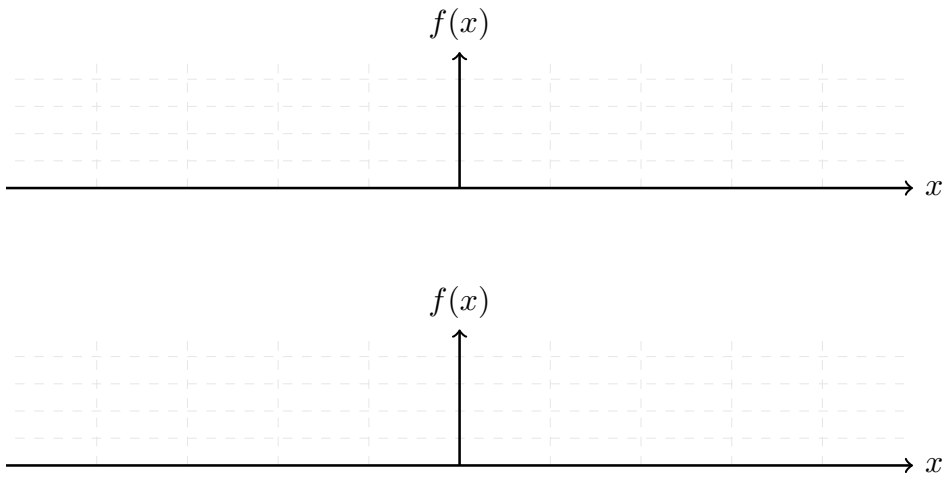
Einseitig: Wenn ich aber in einem Glücksspiel bin (und wenig "Kopf" will), so ist mir eine Verfälschung zu meinen Gunsten egal. Dann kann es Sinn machen, zu wählen: $\mathcal{H}_0$ heisst $p \leq 0.5$ und $\mathcal{H}_1$ ist dann $p > 0.5$. Das ist dann eine einseitige Testsituation. Signalwörter sind Steigerung/Erhöhung bzw Senkung. Hier gibt es dann 2 Möglichkeiten (gegen links oder gegen rechts bzw nach oben oder nach unten).

6. Wenn die Resultate der Teststatistik derart sind, dass wir $\mathcal{H}_0$ verwerfen, sagt man, dass die Resultate *signifikant* sind (zB signifikant von 0 verschieden).

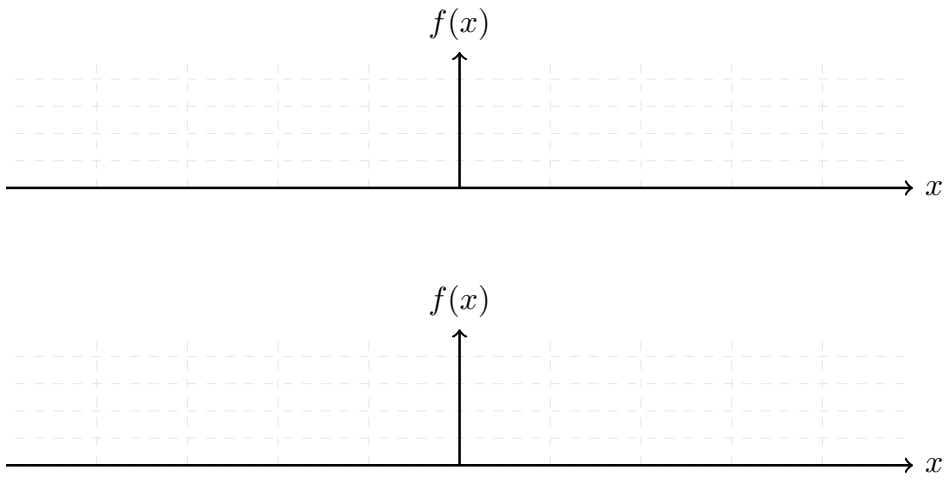
7. Hypothesen können "einfach" sein, aus einem Punkt bestehend: $p = 0.5$ oder "zusammengesetzt" (engl compound): $p > 0.5$, aus ganzen Intervallen.

8. Bemerkungen zu α: Wir haben bis jetzt $\alpha = 0.1$ oder $\alpha = 0.05$ gewählt. Warum nicht (aus Fairness-Gründen) $\alpha = 0.5$ oder besser: Risiko 1. Art = Risiko 2. Art ("Zweite Aufgabe")? In der Praxis hat man meist eine Situation derart, dass $\mathcal{H}_0$ den bisherigen Stand des Wissens repräsentiert. Wir sind sehr konservativ in dem Sinne, dass wir nicht vorschnell zufällige Resultate in gefestigte Lehrmeinungen überführen wollen. Das heisst, wir wählen eher ein kleines α. Genauer gilt folgende Dynamik:

* α gross: nehmen schneller $\mathcal{H}_1$ an (gut, wenn $\mathcal{H}_1$ richtig / schlecht, wenn $\mathcal{H}_0$ richtig):



* α klein: nehmen weniger oft $\mathcal{H}_1$ an (gut, wenn $\mathcal{H}_0$ richtig / schlecht, wenn $\mathcal{H}_1$ richtig):



Da Sie als Forschende spektakuläre Resultate präsentieren wollen (und nicht langweilige / bisherige Stand des Wissens), haben Sie die Neigung, ein eher grosses α zu favorisieren. Das Journal oder die finanzierende Behörde wollen aber keine Schlappe/Skandal (mit sich nachträglich als falsch herausstellenden Resultaten; das gefährdet den guten Ruf) und wollen eher ein kleines α. Schlechter Stil ist dann, nachträglich das α so zu wählen, dass die Resultate signifikant werden (zB α von 1 auf 5 Prozent erhöhen). Wir üben in Vlsg und Ue mit $10, 5, 2, 1\%$.

9. P-Wert:

* kann *vor* dem Versuch im Büro bei Luchsinger nicht berechnet werden; braucht Daten! Damit wird er von Versuch zu Versuch (von Student zu Student) anders ausfallen (wichtig für Leute, welche die R-Aufgaben einfach kopieren).

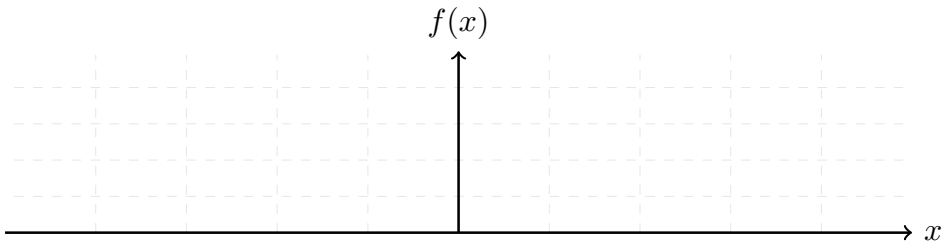
* Anschaulich ist der P-Wert die Wahrscheinlichkeit unter $\mathcal{H}_0$, dass ein so extremer oder noch extremerer Wert der Teststatistik vorkommt. Der P-Wert ist damit ein Mass dafür, wie signifikant das Resultat ist.

* Gute, vorsichtige Einstellung: tiefer P-Wert? Das muss ich genauer anschauen!

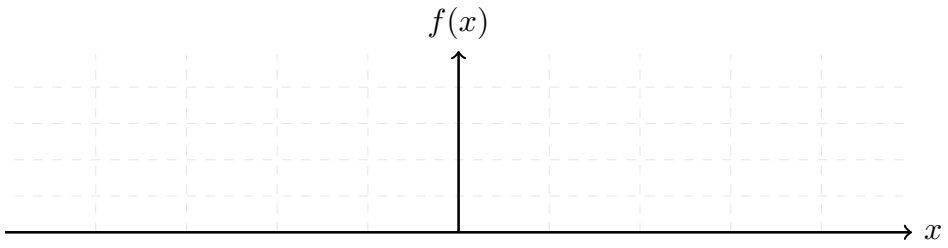
* α ist offenbar eine schöne Zahl (5%); der P-Wert normalerweise nicht ($3.85 \cdot 10^{-7}$). Oft gehörter Blödsinn: "Der P-Wert ist 5%".

* Wegen Punkt 5 oben machen wir *drei* Beispiele mit $n = 1$, $\mathcal{N}(\mu, 1)$ und $\alpha = 5\%$. X_1 sei dazu jeweils $\mathcal{N}(0, 1)$-verteilt (da unter $\mathcal{H}_0$). Didaktik: zuerst zwei einseitige Situationen a) und b), da Quantensprung zu zweiseitig c)!

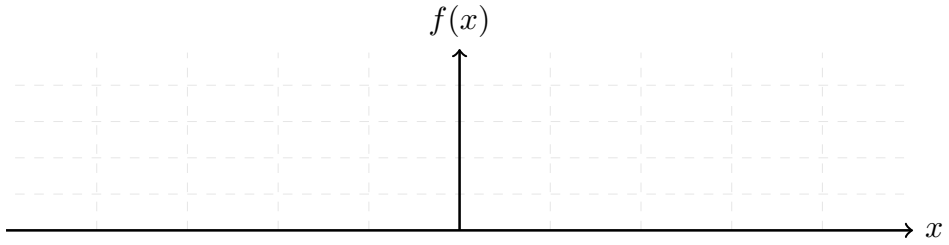
a) $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = 1$ und einziger Wert ergab $x_1 = 2.2$. P-Wert ist dann $P[X_1 \geq 2.2] =$



b) $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = -1$ und einziger Wert ergab $x_1 = -2.9$. P-Wert ist dann $P[X_1 \leq -2.9] =$



c) $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu \neq 0$ und einziger Wert ergab $x_1 = 0.7$. P-Wert ist dann $P[|X_1| \geq 0.7] =$



Richtige Sicht zu P-Wert bei zweiseitiger Situation: *Abweichung* von mindestens 0.7 von 0 weg, das kann aber ± 0.7 und mehr sein!

P-Wert $\leq \alpha : \mathcal{H}_1$ annehmen, Wert Teststatistik eher unwahrscheinlich unter $\mathcal{H}_0$: a), b).
P-Wert $> \alpha : \mathcal{H}_0$ annehmen, Wert Teststatistik gewöhnlich unter $\mathcal{H}_0$: c).

Kochrezept zum Testen von Hypothesen -> zentrale Seite für Prüfung

Um die Mathematik simpel zu halten, wählen wir $\sigma^2 = 1$ bekannt in einem Beispiel der Normalverteilung.

---

1. Hypothesen aufstellen

2. α und (wenn möglich) n wählen

3. gute Teststatistik wählen, Luchs hilft!

4. Verteilung der Teststatistik unter $\mathcal{H}_0$?

5. Kritischer Wert a

Bis hier, ohne Daten, im Büro Luchs, vor Versuch

6. Brauchen Daten

7. $\mathcal{H}_0$ annehmen oder ablehnen

8. Gesunder Menschenverstand: Resultat sinnvoll? links vs rechts vs beidseitig; von Daten her schon offensichtlich (nicht) signifikant; Effekt vs allg Streuung vs Stichprobengrösse?

Beispiel

---

1. $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = 1$
2. $\alpha = 0.05, n$ bekannt
3. $\overline{X}$
4. $\mathcal{N}(0, \frac{1}{n})$
5. $P[\overline{X} > a] = P\left[\frac{\overline{X}}{\frac{1}{\sqrt{n}}} > \frac{a}{\frac{1}{\sqrt{n}}}\right] = P[\sqrt{n}\overline{X} > \sqrt{n}a] = P[\mathcal{N}(0,1) > \sqrt{n}a] = 0.05$
6. $\overline{x}$
7. $\mathcal{H}_0$ annehmen wenn $\overline{x} < a$, sonst ablehnen
8. Wenn $\overline{x} < a$, ist Mittelwert eher klein, dann $\mathcal{H}_0$ anzunehmen ist sinnvoll

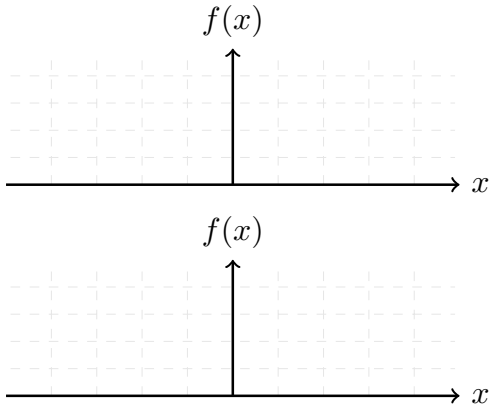
Alternativ ab Schritt 5: P-Wert mit vorhandenen Daten berechnen und mit gängigen α vergleichen.

Beispiel mit konkretem $n = 12$ und $\overline{x} = 0.67$. Was glauben Sie, werden wir bei $\alpha = 0.05$ die Null-Hypothese verwerfen?

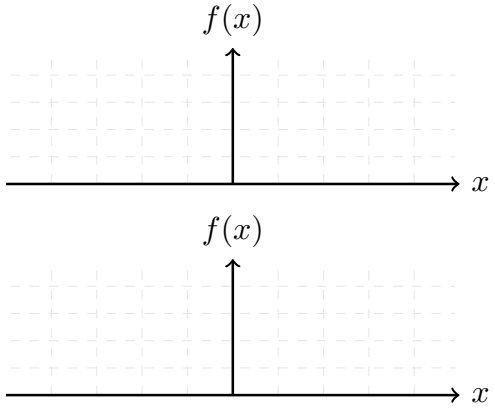
Wir variieren noch n; was ist zu beobachten? Was wenn n sehr gross?

Gute Übung auch wegen Ablesen von Wahrscheinlichkeiten: $\mathcal{H}_0$ sei $\mathcal{N}(0,1)$, $n = 1$, $\alpha = 0.05$. Testen gegen (Test angeben)

a) $\mathcal{H}_1$ ist $\mathcal{N}(1,1)$, bitte auch β berechnen.



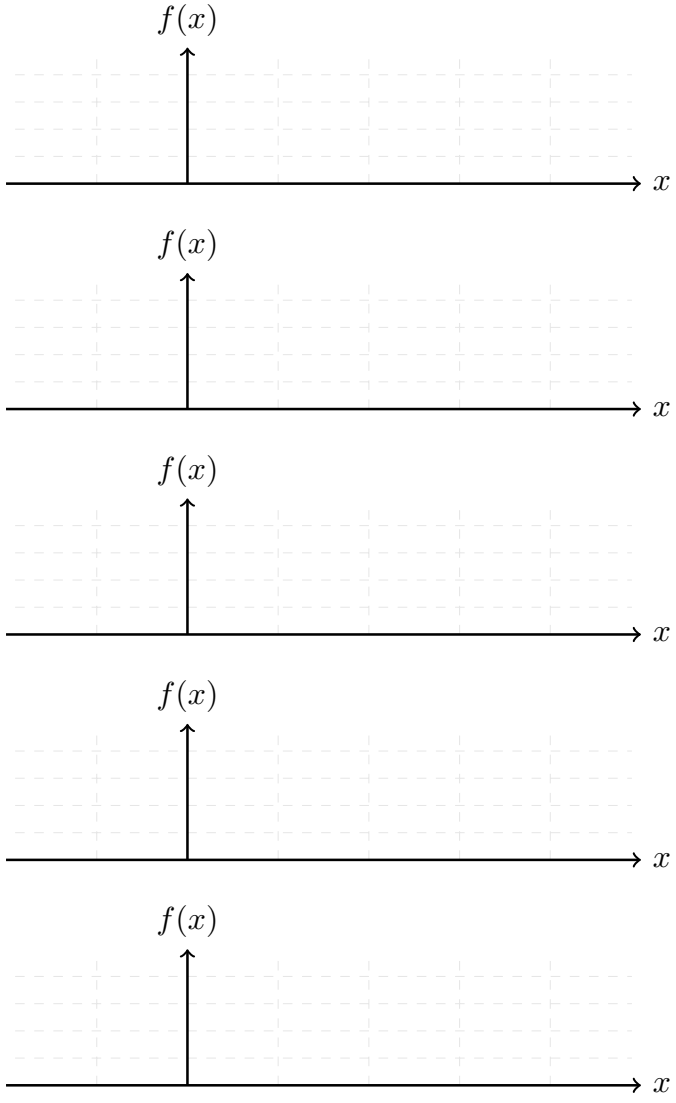
b) $\mathcal{H}_1$ ist $\mathcal{N}(2,1)$, bitte auch β berechnen.



c) $\mathcal{H}_1$ ist $\mathcal{N}(3,1)$, bitte auch β berechnen.

d) $\mathcal{H}_1$ ist $\mathcal{N}(4,1)$, bitte auch β berechnen.

e) Zusammenfassung a)-d). Macht es Sinn?



9.2 t-Test

Storrer-Bsp 45.2.A **noch anpassen** zur Qualitätskontrolle: Brote mit Gewichten [g]

$$69, 70, 71, 68, 67, 70, 70, 70, 67, 69$$

Bäcker: "Brote wiegen im Schnitt 70 g." Der Durchschnitt von obigen Broten ist jedoch 69.1 g. Wir einigen uns darauf, das Gewicht der Brote mit einer Normalverteilung zu modellieren. Bild zur Grundproblematik - gehen Sie in solchen Situationen in die Extreme:

9.2.1 σ^2 vorerst bekannt, unrealistisch - noch nicht t-Test, analog Aufbau KI

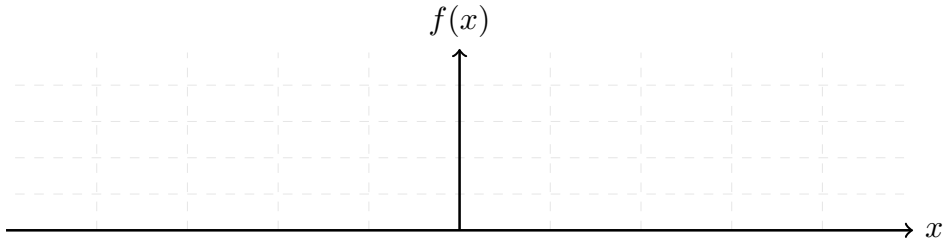
Angenommen, wir haben eine Stichprobe $x_1, \ldots, x_n$, von der wir wissen, dass sie aus einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ stammt. Wir kennen leider nicht den Erwartungswert. Jemand behauptet, der Erwartungswert sei 70 ($\mathcal{H}_0 : \mu = 70, \mathcal{H}_1 : \mu \neq 70$, zweiseitig). Wie könnten wir das testen?

1. Hypothesen aufstellen:

2. α und (wenn möglich) n wählen:

3. gute Teststatistik wählen, 4. Verteilung der Teststatistik unter $\mathcal{H}_0$?

5. Kritischer Wert a



9.2.2 σ^2 unbekannt, realistisch $\Rightarrow$ t-Test

Angenommen, wir haben eine Stichprobe $x_1, \ldots, x_n$, von der wir wissen, dass sie aus einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ stammt. Wir kennen leider weder Erwartungswert noch Varianz. Jemand hat behauptet, der Erwartungswert sei 70 ($\mathcal{H}_0 : \mu = 70, \mathcal{H}_1 : \mu \neq 70$, zweiseitig). Wie könnten wir das testen?

1. Hypothesen aufstellen:

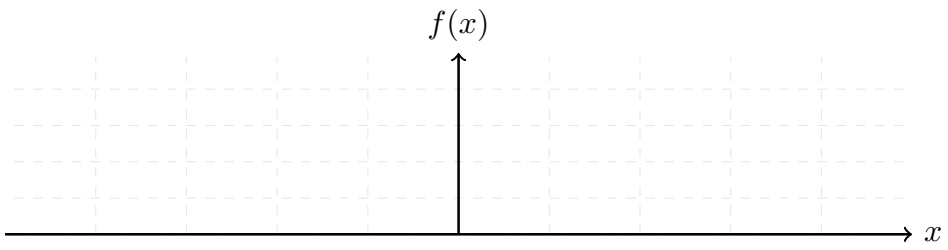
2. α und (wenn möglich) n wählen:

3. gute Teststatistik wählen (vgl 9.2.1), 4. Verteilung der Teststatistik unter $\mathcal{H}_0$?

5. Kritischer Wert a

Bsp p-3 (Brote) fertig besprechen.

Aufgabe 9−1 Der Inhalt von 6 Säcken Puderzucker wurde nachgewogen. Man erhielt folgende Gewichte (in Gramm): 495, 502, 505, 498, 490, 500. Ist die Behauptung, dass die Säcke im Durchschnitt 500 g enthielten, haltbar? Testen Sie mit $\alpha = 5\%$.



Bei Fragen zu unterem, wie immer in R: "?t.test"

```
> b<-c(-2.2, -0.7, 3, 1.5, 0.2)
```
offenbar ist $n = 5$

```
> t.test(b)
One Sample t-test
data:  b
t = 0.4028, df = 4, p-value = 0.7077
```

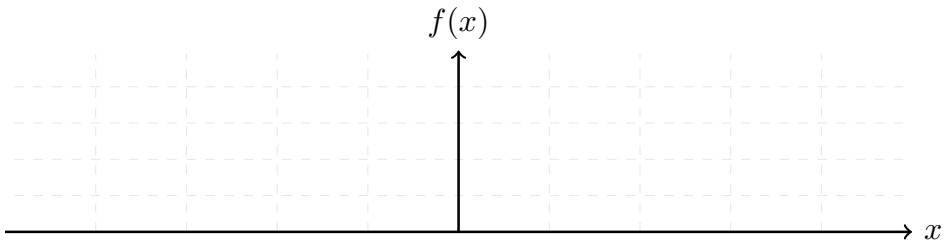
$t =$ *Wert Teststatistik,* $df = n - 1 = 5 - 1 = 4$, *P-Wert ist zweiseitig* $2 \cdot 35.38\%$

```
alternative hypothesis:  true mean is not equal to 0
```

R macht per Default t.test zweiseitig ob $\mu = 0$ *oder nicht*

```
95 percent confidence interval:
```
wird automatisch gleich mitberechnet

```
-2.121154 2.841154
sample estimates:
mean of x
0.36
```
das reicht nicht als Test! Streuung noch beachten liefert dann t oben



Trivia: R kennt weder Euer α noch Eure Hypothesen! Hier aber glasklar für jedes α: $\mathcal{H}_0$ beibehalten da 70.77 % (=P-Wert) $> \alpha$.

2 Tipps für Beruf und Uni, spätere Semester:

* wichtiges erstes Ziel bei R-Ausdrucken: wenn Analyse (von jemand anderem) korrekt gemacht wurde, richtig lesen, interpretieren können. Auch wenn man es selber gar nicht (mehr) kann oder allfällige Fehler nicht erkennen würde. Etwa vergleichbare Kompetenz wie Bilanz und Erfolgsrechnung in Firma oder Verein lesen können.

* nach dem ersten Studienjahr: Kurse an der Uni besuchen zu Software etc. Es ist gratis und Sie haben jetzt noch halbwegs Zeit. Später hat man noch weniger Zeit und der gleiche Kurs kostet schnell mal mehrere Tausend Franken.

9.2.3 2-Stichproben t-Test

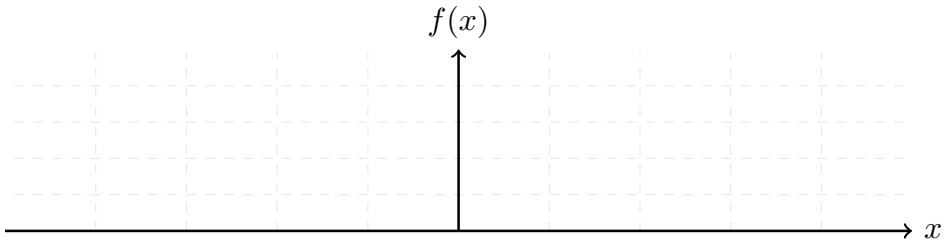
Alte Sorte	410	420	430	440	450	450	480
Neue Sorte	440	450	455	480	490	505	

Angenommen, wir haben eine Stichprobe $x_1, \ldots, x_m$, von der wir wissen, dass sie aus einer Normalverteilung $\mathcal{N}(\mu_1, \sigma^2)$ stammt. Zudem haben wir eine Stichprobe $y_1, \ldots, y_n$, von der wir wissen, dass sie aus einer Normalverteilung $\mathcal{N}(\mu_2, \sigma^2)$ stammt. Die beiden Stichproben seien unabhängig voneinander, die Varianzen seien gleich - aber unbekannt! Wir wollen mit einem Test die beiden Mittelwerte vergleichen.

1. Hypothesen aufstellen:

2. α und (wenn möglich) m, n wählen:

3. gute Teststatistik wählen (vgl 9.2.2), 4. Verteilung der Teststatistik unter $\mathcal{H}_0$?



5. Kritischer Wert a

Falls Hinweis, dass $\sigma_1 \neq \sigma_2$, macht man den sogenannten Welch-Test.

HA: Beispiel aus Buch 9.3.2.A (mit obigen Daten den 2-Stichproben-T-Test machen).

Achtung für die Prüfungseinsicht, falls wir es übersehen haben: man kann statt $\overline{x} - \overline{y}$ auch alles mit $\overline{y} - \overline{x}$ aufziehen - ist gleichwertig und sollte die gleiche Punktzahl geben!

9.3 (vgl 8.3) Die 4 wichtigsten Tests aus der Praxis

Orientierungsschema:

	diskret, Be(p)/Bin(n,p)	stetig, $\mathcal{N}$
Eine Stichprobe		
Zwei Stichproben		

A)

B)

In der Prüfung müssen Sie alle 4 Situationen beherrschen. In den Übungen gibt es aber nicht jedes Semester zu allem ein Beispiel. Schauen Sie dafür die alten Prüfungen im Archiv an.

Achtung für die Prüfungseinsicht, falls wir es übersehen haben: man kann statt $k_1/n_1 - k_2/n_2$ auch alles mit $k_2/n_2 - k_1/n_1$ aufziehen - ist gleichwertig und sollte die gleiche Punktzahl geben!

9.4 χ^2-Tests (Anpassung und Unabhängigkeit)

9.4.1 χ^2-Anpassungstest

Wir motivieren die Grundfrage mit einem historischen Beispiel. Aufgrund der Mendelschen Erbgesetze erwartet man in bestimmten Kreuzungsversuchen, dass Nachkommen in drei verschiedenen Typen mit Wahrscheinlichkeiten 0.25 (aa), 0.5 (aA) und 0.25 (AA) auftreten (https://schweizermonat.ch/in-der-genetik-gibt-es-keine-drittel/).

In einer Untersuchung (Daten aus Stahel) fand man: $s_1 = 29, s_2 = 77$ und $s_3 = 44$ bei 150 Nachkommen. Wäre es exakt "fair" gewesen, so "müsste" es sein: $150/4 = 37.5$, $150/2 = 75$ und wieder $150/4 = 37.5$. Offenbar gibt es im Versuch eine Abweichung gegenüber diesen "fairen" Werten und jedeR WissenschaflerIn, welche mit genau diesen "fairen" Werten aufwarten würde, muss als Statistik-Banause bezeichnet werden (selbst wenn er/sie bei den halben Fällen *je einmal* auf- und abrundet). *Eine gewisse Variation in den Daten ist zu erwarten! Aber wieviel ist noch mit dem Modell von Mendel vereinbar? Ab wann müssen wir von einer signifikanten Abweichung sprechen?*

Wir müssen dazu erstmal eine neue Zufallsgrösse einführen, welche zu Recht an die Binomialverteilung erinnert: die **Multinomialverteilung** (die Binomialverteilung ist ein Spezialfall hiervon). Wenn wir n unabhängige Versuche haben mit m möglichen Ereignissen (in obigem Beispiel: $n = 150, m = 3$ - bei der Binomialverteilung $m = 2$ (Erfolg/Misserfolg)), so ist ein Zufallsvektor $(S_1, \ldots, S_m)$ per Definitionem multinomialverteilt, wenn

$$P[S_1 = s_1, \ldots, S_m = s_m] = \frac{n!}{s_1! \ldots s_m!} p_1^{s_1} \ldots p_m^{s_m}.$$

Dabei muss dann gelten: $s_1 + \ldots + s_m = n$ und $p_1 + \ldots + p_m = 1$.

Die Hypothesen sind derart, dass der Zufallsvektor $(S_1, \ldots, S_m)$ unter $\mathcal{H}_0$ eine Multinomialverteilung mit Wahrscheinlichkeiten $p_1, \ldots, p_m$ hat gegen $\mathcal{H}_1$ einer Multinomialverteilung mit Wahrscheinlichkeiten $q_1, \ldots, q_m$, wo mindestens bei zwei (!) Wahrscheinlichkeitspaaren (p_i, q_i), $1 \leq i \leq m$, gelten muss $p_i \neq q_i$.

Wie bei der Binomialverteilung wird man auch hier die Parameter sinnvollerweise mit

$$\hat{p}_i := s_i/n$$

schätzen. Damit erhalten wir m Abweichungen

$$\hat{p}_i - p_i, \ 1 \leq i \leq m.$$

Positive und negative Abweichungen könnten sich aufheben. Somit ist es sinnvoll, z.B. die absoluten Werte zu betrachten:

$$|\hat{p}_i - p_i| = |s_i - np_i|/n, \ 1 \leq i \leq m.$$

Wir wollen diese m Werte alle zusammen in einer einzigen Testgrösse betrachten und es hat sich herausgestellt, dass

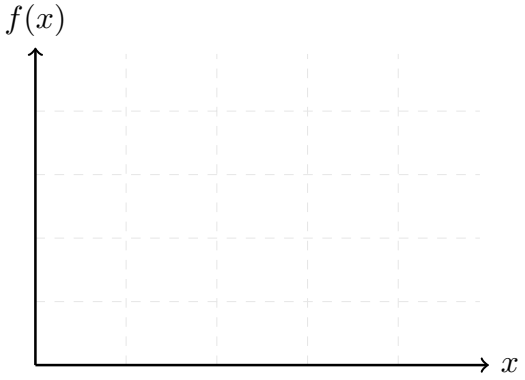
$$u := \sum_{i=1}^{m} \frac{(s_i - np_i)^2}{np_i} \qquad (\chi^2 - \text{Anpassung})$$

eine brauchbare Testgrösse ist. Mit $n \to \infty$ gilt, dass diese Grösse als Zufallsgrösse U unter $\mathcal{H}_0$ eine χ^2_{m-1}-Verteilung besitzt. In Stahel sind Praktikerregeln aufgeführt, ab wann wir n als genügend gross bezeichnen dürfen. Es ist noch beachtenswert, dass in der Referenzverteilung nur m und nicht die einzelnen p_i's vorkommen. Wann werden wir die $\mathcal{H}_0$-Hypothese ablehnen, das heisst, von welcher Form ist der Ablehnungsbereich? Wo war wohl Mendel?

Wir untersuchen jetzt mit diesem Test das einführende Beispiel:

$$\begin{aligned} u &= \frac{(29 - 150 * 0.25)^2}{150 * 0.25} + \frac{(77 - 150 * 0.5)^2}{150 * 0.5} \\ &\quad + \frac{(44 - 150 * 0.25)^2}{150 * 0.25} \\ &= 1.927 + 0.053 + 1.127 \\ &= 3.11. \end{aligned}$$

Es gilt: $P[\chi^2_2 > 5.99] = 0.05$; somit werden wir auf dem $\alpha = 5\%$-Niveau die $\mathcal{H}_0$-Hypothese nicht ablehnen. Die Versuchsresultate sind demnach auf diesem Niveau durchaus mit den Mendel'schen Gesetzen vereinbar.



Im Buch wird noch der Fall behandelt, dass die Werte der Parameter (hier 0.25, 0.5 und 0.25) nicht bekannt sind und aus den Daten geschätzt werden müssen. Zum Beispiel bei einer Po(λ)-Zufallsgrösse muss man noch λ schätzen. Dieser Stoff ist obligatorisch und kam schon in der Prüfung vor; bitte lesen.

9.4.2 χ^2-Test für Unabhängigkeit in Kontingenztafeln

Wir betrachten erfundene Raucher/Krebs-Daten und wollen untersuchen, ob das Auftreten von Krebs unabhängig vom Rauchen ist. Am Besten trägt man dazu die Daten in einer sogenannten Vierfeldertafel ein - und zwar die absoluten Werte!

konkret

Raucher \ Krebs	K	kein K	
R			
NR			

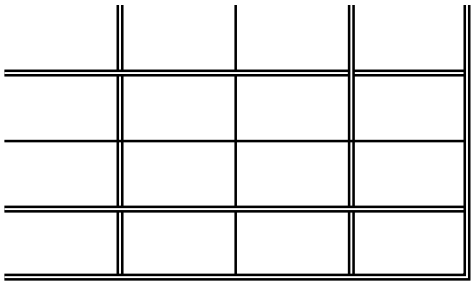
Was könnte hier auf Unabhängigkeit / Abhängigkeit hinweisen? Was müsste bei Unabhängigkeit etwa erfüllt sein?

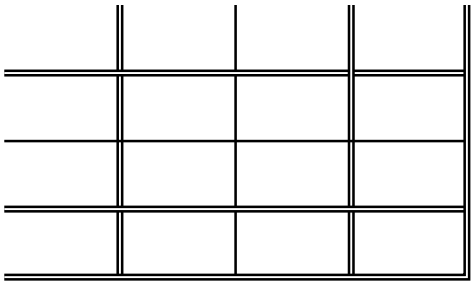
Raucher \ Krebs	K	kein K	
R			
NR			

und abstrakt

Raucher \ Krebs	K	kein K	
R			
NR			

Warum die absoluten Werte und nicht Prozente?





Wir geben den Merkmalen gleich Bezeichnungen, welche an Zufallsgrössen erinnern sollten: X und Y. Wir stellen jetzt ein Wahrscheinlichkeitsmodell auf:

$$P[X = i, Y = j] =: p_{ij}.$$

Die obigen Anzahlen n_{ij} in den einzelnen Kästchen fassen wir auch als Realisationen von Zufallsgrössen N_{ij} auf $(1 \leq i, j \leq 2)$. Der Zufallsvektor $(N_{11}, \ldots, N_{22})$ ist multinomialverteilt:

$$P[N_{11} = n_{11}, .., N_{22} = n_{22}] = \frac{n!}{n_{11}!..n_{22}!} p_{11}^{n_{11}} ..p_{22}^{n_{22}}.$$

Dabei muss dann gelten: $n_{11} + \ldots + n_{22} = n$ und $p_{11} + \ldots + p_{22} = 1$. Damit sind wir also in der gleichen Situation wie beim χ^2-Anpassungstest.

Was uns hingegen jetzt interessiert, ist, ob die beiden Ausprägungen unabhängig voneinander auftreten (im Beispiel: Es besteht kein Zusammenhang zwischen Rauchen und Krebs). Formalisiert wird das in der Forderung (Nullhypothese)

$$p_{ij} := P[X = i, Y = j] = P[X = i]P[Y = j] =: p_{i.}p_{.j}.$$

Von (χ^2-Anpassung) erhalten wir jetzt folgende Teststatistik

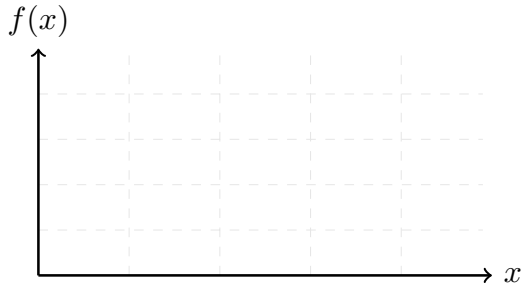
$$u^{(\text{prov})} := \sum_{i,j} \frac{(N_{ij} - np_{i.}p_{.j})^2}{np_{i.}p_{.j}}. \qquad (\chi^2 - \text{Test auf Unabhängigkeit, provisorisch})$$

Die $p_{i.}$ und $p_{.j}$ sind unbekannt und wir müssen sie aus den Daten mit $\hat{p}_{i.} := N_{i.}/n$, $\hat{p}_{.j} := N_{.j}/n$ schätzen. Dadurch wird obiger Ausdruck zur definitiven Variante:

$$u := \sum_{i,j} \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n}. \qquad (\chi^2 - \text{Test auf Unabhängigkeit})$$

Man kann zeigen, dass diese Grösse als Zufallsgrösse U und bei Unabhängigkeit der Merkmale (Nullhypothese) mit $n \to \infty$ eine χ^2-Verteilung mit 1 Freiheitsgrad besitzt.

Im einleitenden Beispiel erhalten wir einen Wert $u = 0.1355$. Der P-Wert von 0.1355 bei einer χ^2-Verteilung mit 1 Freiheitsgrad ist 71.3% (von R); bei $\alpha = 5\%$ erhalten wir einen Ablehnungsbereich der Nullhypothese ab Werten $u \geq 3.841$. Was schliessen wir daraus?



Prüfungstaktik: Im Buch (9.4.5) finden Sie für die Prüfung eine Formel, welche das gleiche Resultat wie (χ^2-Test auf Unabhängigkeit) liefert, nicht anschaulich ist, aber einfacher berechnet werden kann.

Warum *Freiheits*grad?

Raucher \ Krebs	K	kein K	
R			
NR			

9.5 Schlussbemerkungen zu Tests

Vergleich Tests und KI's: Sie haben mittlerweile, vor allem beim t-test, gemerkt, dass bei Tests und KI's die gleichen Zahlen vorkommen. Dazu drei Bemerkungen:

* Prüfungstaktik: Da meist die beiden Berechnungen als Teilaufgaben in einer Aufgabe vorkommen, lohnt es sich, dies zu merken, damit man nicht alles zweimal berechnet - auch in R wird meist automatisch beides angegeben.
* Es gibt operationell einen Unterschied: bei den KI's sind keine Hypothesen involviert; wir arbeiten nur mit den Daten und vergleichen keineswegs mit allenfalls vorhandenen Theorien und ihren Parametern (KI's kamen ja auch schon in Kapitel 8 und die Hypothesen erst in Kapitel 9).
* Das alles ist kein Zufall: Tests sind umgestülpte KI's, mehr dazu im Buch (9.2.6).

9.6 Eine kleine Betrachtung, exemplarisch am t-Test

$$\frac{\overline{x} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

Sicht aus Elektro- und Informationstechnik: Den Zähler kann man als "Signal" und den Nenner als "Rauschen" interpretieren. Für die weitere Untersuchung ist es sinnvoll, das $\sqrt{n}$ in den Zähler zu nehmen:

$$\frac{\overline{x} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}} = \frac{\sqrt{n}(\overline{x} - \mu_0)}{\sqrt{\hat{\sigma}^2}}$$

Jetzt ist im Nenner lediglich die (geschätzte) Standardabweichung (Unsicherheit) einer einzelnen Messung ($\sqrt{\hat{\sigma}^2}$) und nicht mehr die (geschätzte) Standardabweichung (Unsicherheit) des arithmetischen Mittels ($\sqrt{\hat{\sigma}^2/n}$), auch bekannt als Standardfehler. Insbesondere ist die folgende Interpretation einfacher zu plausibilisieren: **Sie wollen als ForscherIn Signifikanz beweisen**. Dazu muss der Absolutbetrag des obigen Ausdrucks "gross" sein (so etwa > 2). Dazu muss der Zähler "gross" sein und der Nenner "klein":

gut: $|\overline{x} - \mu_0|$ gross
gut: n gross (genauer: $\sqrt{n}$ gross)
gut: σ klein.

Beim letzten Punkt kann man noch unterscheiden, dass in der Messung die Unsicherheit von 2 Quellen kommt: natürliche Variabilität des untersuchten Objekts versus Unsicherheit, die von der Mess*methode* kommt. Das erste können Sie nicht beeinflussen (Naturgesetze), das zweite sehr wohl (Technik, besseres Studiendesign).

Kompensation / Zusammenspiel der 3 Effekte:

Wichtig:

1. Lesen Sie jetzt das komplette Kapitel 9 im Buch selber durch.
2. Lösen Sie danach mindestens 5 Aufgaben hinten im Kapitel und vergleichen Sie mit den Lösungen am Schluss des Buches. Bei Bedarf lösen Sie mehr Aufgaben.
3. Gehen Sie in die Übungsstunde. Drucken Sie das Übungsblatt dazu *vorher* aus, lesen Sie *vorher* die Aufgaben durch und machen sich erste Gedanken dazu (zum Beispiel, wie man sie lösen könnte).
4. Dann lösen Sie das Übungsblatt: zuerst immer selber probieren, falls nicht geht: Tipp von Mitstudi benutzen, falls immer noch nicht geht: Lösung von Mitstudi anschauen, 1 Stunde warten, versuchen, aus dem Kopf heraus wieder zu lösen, falls immer noch nicht geht: Lösung von Mitstudi abschreiben (und verstehen - also sollte man insbesondere keine Fehler abschreiben!).
5. Lösen Sie die entsprechenden Prüfungsaufgaben im Archiv.

Lessons Learnt:

$\mathcal{H}_0$ ist richtig und wir nehmen vorläufig / provisorisch $\mathcal{H}_1$ an **(Fehler 1. Art)**.
$\mathcal{H}_1$ ist richtig und wir nehmen vorläufig / provisorisch $\mathcal{H}_0$ an **(Fehler 2. Art)**.

Wichtige Sammlung von Begriffen und Konzepten bei "Jargon und Bemerkungen" und Kochrezept in 9.1.2.

Zentral bei der Konstruktion von Teststatistiken: Differenz (Effekt) in's Verhältnis, also Bruch, zur Streuung setzen.

Z-Test: Teststatistik, σ^2 bekannt: $(\overline{x} - \mu_0)/\sqrt{\sigma^2/n}$

t-Test: Teststatistik, σ^2 unbekannt: $(\overline{x} - \mu_0)/\sqrt{\hat{\sigma}^2/n}$ wo $\hat{\sigma}^2 := \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)$.

2-Stichproben-t-Test: Teststatistik:

$$\frac{\overline{x} - \overline{y}}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\frac{\sum_{i=1}^{m}(x_i - \overline{x})^2 + \sum_{j=1}^{n}(y_j - \overline{y})^2}{m+n-2}}}$$

1-Stichproben-Test einer Wahrscheinlichkeit:

$$\frac{k/n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

2-Stichproben-Test des Vergleichs von 2 Wahrscheinlichkeiten:

$$\frac{k_1/n_1 - k_2/n_2}{\sqrt{\frac{k_1/n_1(1-k_1/n_1)}{n_1} + \frac{k_2/n_2(1-k_2/n_2)}{n_2}}}$$

χ^2-Anpassung, im Normalfall Ablehnung ausschliesslich oben

$$\sum_{i=1}^{m}\frac{(s_i - np_i)^2}{np_i}$$

χ^2-Test auf Unabhängigkeit, im Normalfall Ablehnung ausschliesslich oben

$$\sum_{i,j}\frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n}$$