Skript
VO Statistik
Sommer 2023

Dominik Duell

Universität Innsbruck

Contents

Sitzung 1: Einführung

Literatur
Fields, Kapitel 1
de Mesquita/Fowler, Kapitel 1

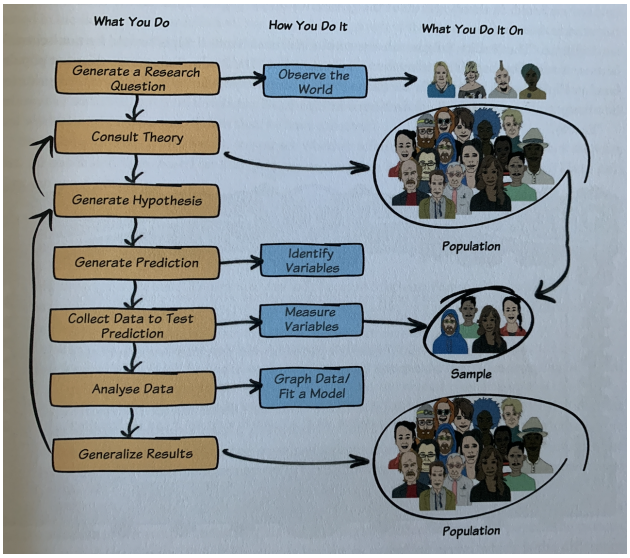
Zunächst beginnen wir mit der sehr grundlegenden Frage, was machen wir eigentlich hier und was ist das, Statistik.

Was ist Statistik und wozu brauche ich sie? Was sollte ich aus dieser Vorlesung für meine weitere Ausbildung mitnehmen?

Die Welt sammelt Daten wie wahnsinnig. Viele von euch sind an die Unversität gekommen, um etwas zu studieren, dass euch helfen soll, die soziale und politische Welt zu verstehen. Beide Beobachtungen zusammen, sollten klar machen, dass wir die Statistik brauchen. Statistik ist das Werkzeug, um Daten zu organisieren und zu analysieren.

Ok, also Statistik hilft uns, die Welt besser zu verstehen. Großartig. Ist das alles? Natürlich nicht. **Der** Weg, um die Welt zu verstehen, ist die **wissenschaftliche Methode.** Um Wissenschaft zu betreiben, sind wir nicht nur an empirischen Daten interessiert, aber oft. Wenn wir einer Forschungsfrage wissenschaftlich nachgehen und dazu empirische Daten erheben, dann ist es sehr oft die Statistik, die uns hilft sich der wissenschaftlichen Frage mit den erhobenen Daten anzunähern. Wie und wo genau.

Fields zeigt uns, wie der Forschungsprozess aussieht (die Darstellung ist abstrahiert, versteht sich).



Sobald wir empirische Vorhersagen aus unserer Theorie abgeleitet haben, müssen wir uns Gedanken machen, welche Variablen die theoretischen Konzepte am Besten abbilden. Wie das funktioniert, behandelt ihr in anderen Kursen. Sobald wir uns Variablen auserkoren haben und Daten zu diesen Variablen dasitzen, da wird erstmal die **Deskriptive Statistik** (Sitzung 2-4) nüztlich. Wir können mit dem Handwerkszeug der Deskriptiven Statistik die Verteilung der Werte unserer Variablen in unserem Datensatz erkunden. Wozu? Das sehen wir uns in Sitzung 3 an.

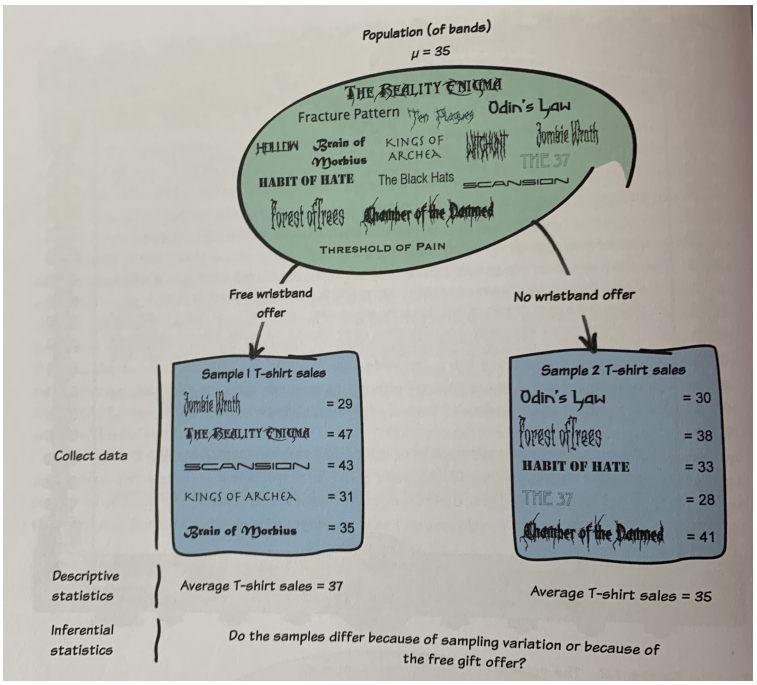
Oftmals werden unsere empirischen Vorhersagen schon mit Deskriptiver Statistik evaluierte werden können. Etwa eine Vorhersage wie, höhere Löhne in einer Industrie gehen mit größeren Gewerkschaften einher. Hier handelt sich um eine Aussage über eine Korrelation, und um solch eine Vorhersage zu testen, würde es genügen, deskriptiv, zum Beispiel die Löhne in Industriezweigen mit großen Gewerkschaften, den Industriezweigen mit kleinen Gewerkschaften gegenüber zustellen.

Stichprobe (sample) und Grundgesamtheit (population)

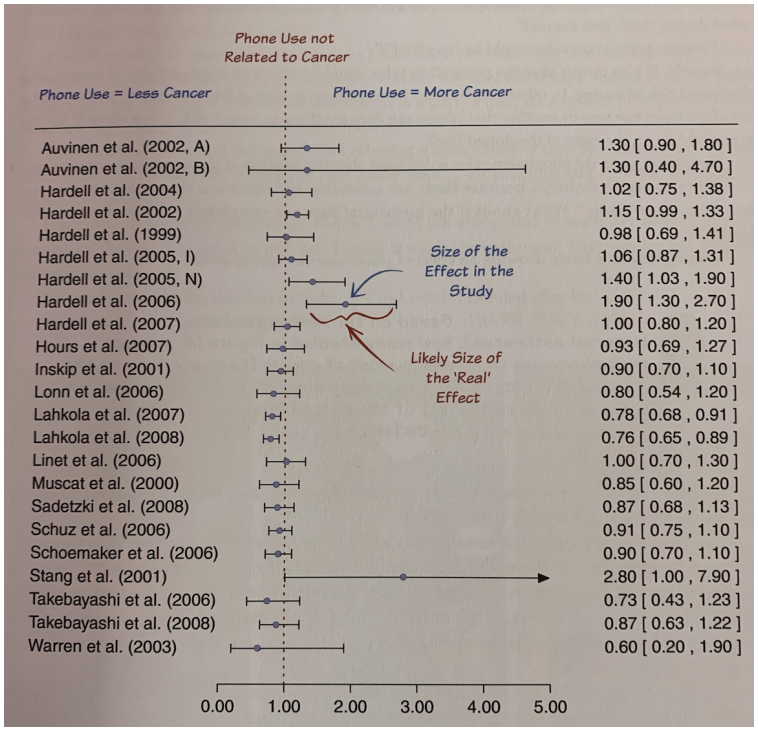
Die wissenschaftliche Methode hilft uns Wissen aufzubauen, da in einem iterativen Prozess, immer wieder die Aussagen einer Theorie mit Daten überprüft und aus dieser Überprüfung Theorien fallen gelassen oder weiter entwickelt werden. Zur Wissenschaft gehört also ein fortlaufender Prozess der empirischen

Überprüfung. In der Regel geschieht das, in dem wir immer wieder Stichproben aus der Grundgesamtheit unserer Untersuchungsobjekte ziehen, und mit Hilfe von (statistichen) Testverfahren bestehende theoretische Vorhersagen abklopfen – ich gehe hier schon von einem bestimmten philosohpischen Ansatz, wie Wissenschaft Erkenntnis schafft, aber das schaut ihr euch in der VO Wissenschaftliches Arbeiten und den PS Angewandte Methoden genauer an.

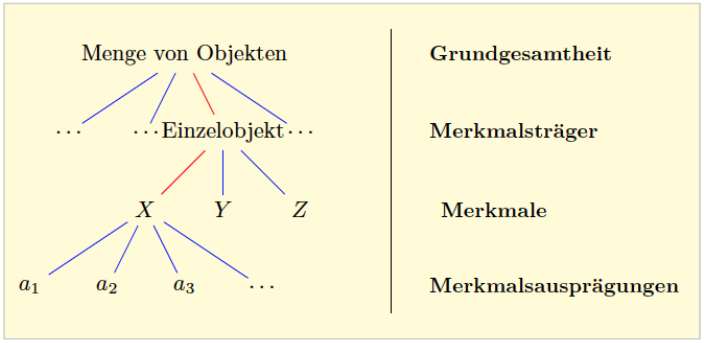
Das wiederkehrende ziehen einer Stichprobe ist ein Grundpfeiler unserer Arbeitsweise. Fields zeigt hier am Beispiel der Frage, verkauft eine Lärm-Band mehr T-Shirts wenn die Band auch kostenlose Armbänder verteilt. Dazu wird die Grundgesamtheit der Bands als diejenige definiert, die in der Wolke oben abgebildet sind. Aus dieser Grundgesamtheit wird dann eine Stichprobe gezogen und die Bands aus der Stichprobe mit Armbad, mit denen ohne Armband verglichen. Wir würden sicherlich gerne immer alle Einheiten in der Grundgesamtheit zur empirischen Analyse heranziehen, oft geht das aber nicht. Wir können etwa aus Resourcen- und ethischen Gründen nicht alle österreichischen Wähler nach ihren politischen Einstellungen befragen, daher ziehen wir eine representative Stichprobe. Es sind auch oft einfach die Daten für einige Forschungseinheiten nicht verfügbar. Im Beispiel der Bands unten, mag es durchaus sein, dass es ein paar Bands gab, die Armbänder verteilt haben, aber sich kurz danach im Drogen-induzierten Streit getrennt haben, bevor wir Daten erheben konnten.



Um zu erfahren, ob Wissen etabliert ist, werden oft so viele Studien zur selben Fragestellung herangezogen. Das nennt man eine **Meta-Studie**. Im Beispiel unten sind mehr als ein Dutzend Studien zur Frage, ob Telefonnutzung Krebs verursacht (ich nehmen mal an, die Frage bezieht sich auf die Benutzung von Mobiltelefonen). Eine jede dieser Studien, zieht eine ander Stichprobe von der Grundgesamtheit "Menschen". Wir sehen, dass nur in einer Studie Belege für die Wirkung von Telefonnutzung auf Krebshäufigkeit gefunden wurden. Wir sollten daher eine Zusammenhang nicht vermuten.



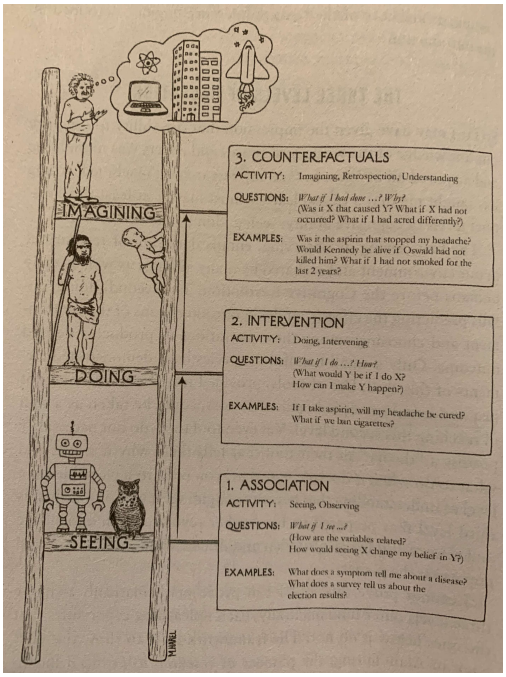
Mittag/Schüller behandeln auch die Konzepte Stichprobe und Grundgesamtheit. Deren Darstellung beinhaltet auch, wie sich das einzelne Forschungsobjekt, die Forschungseinheit, in den Prozess einbindet. Aus der Menge an möglichen Forschungsobjekten, der Grundgesamtheit, ziehen wir eine Stichprobe an Einzelobjekte und studieren deren Ausprägungen in den Merkmalen (Variablen), die uns interessieren.



Korrelation und Kausalzusammenhänge

Ob wir **valide** und **robuste** Erkenntnisse aus unserer empirischen Analyse erhalten können, wird an der Anwendung geigneter statistischer Testverfahren liegen. Das ist es was wir in dieser Vorlesung behandeln. Es wird auch in der korrekten Definition der Grundgesamtheit und einem angemessen Verfahren des Ziehens einer Stichprobe liegen. Das werden wir hier nicht behandlen. Was wir allerdings inhaltlich lernen können, ist eine weitereichende Frage. Statistik alleine sagt uns nichts darüber, wir müssen verstehen, wie wir statistische Ergebnisse interpretieren. Eine der wichtigsten Fragen, die wir uns stellen werden ist, ob ein Zusammenhang (zwischen zwei oder mehreren Variablen) eine Korrelation oder Kausalität beinhaltet. Eine Korrelation ist schnell berechnet und erkannt. Belege für einen Kausalzusammenhang können allerdings nur aus einem angemessen Forschungsdesign kommen. Solch ein Forschungsdesign wird in irgendeiner Weise beinhalten, dass

wir uns angemessene Vergleichgruppen vorstellen und bilden. Pearl illustriert diesen Punkt in interessanter Weise in der Graphik unten.



Die Vergleichsgruppen, die wir benötigen, beinflussen dann, welche Daten wir erheben wollen. Oft ist es schwer, angemessen Vergleichsgruppen zu bilden. Sagen wir, wir sind daran interessiert zu lernen, ob die Flüchtlingswelle von 2015 Wahlerfolge von populistischen Parteien beeinflusst haben. Wir würden also gerne eine Welt ohne Flüchtlingswelle mit der existierenden Welt mit Flüchtlingswelle vergleichen. Das ist natürlich nicht möglich, da erstere nicht existiert. Also versuchen wir meist eine Vergleichsgruppe zu bilden, die sich einer Welt ohne Flüchtlingswelle annähert. In diesem Beispiel, könnten wir etwa in die Zeit vor 2015 sehen oder in Länder/Regionen, in die weniger Flüchtlinge gekommen sind. So ein Vergleich ist sicherlich nicht perfekt. Die Zeit vor 2015 ist in viellerlei Hinsicht anders als die Zeit danach, nicht nur mit Bezug auf Flüchtlinge. Und Länder/Regionen sind verschieden und gerade die Länder, in die viele Flüchtlinge gekommen sind, sind schon mal anders, als die Länder in die keine gekommen sind. In einen PSs und VUs Angewandte Methoden werdet ihr mehr über Experimente und Kausalinferenz lernen, wie man solche Fragen dennoch sehr valide und robust studieren kann.

Daten

Wie man Wissen schaft, was die Statistik dazu beiträgt, sind alles gute Fragen. Habt ihr euch aber auch schon einmal gefragt, was wir eigentlich meinen, wenn wir sagen “Daten”. Hier sind erstmal ein paar gute Gedanken von Keller: “Data are what we hear, see, smell, taste, touch, and more. Data can even be what we sense. Data can represent anything and everything that we can discriminate well enough to distinguish from something else. In short, if it can be perceived, it can be coded and used as data (Keller, 2016, p. 7).” Dann fassen wir das ganze mal kürzer, siehe die Box unten für eine Definition.

Definitionen:
Statistik: Organisation und Analyse von Daten
Daten : Organisierte, erhebbare Information
Variablen: Merkmale (variable) mit Merkmalsausprägung (variable value) der beobachteten Forschungseinheit (research unit)
Korrelation: Zusammenhang zwischen zwei Variablen
Kausalzusammenhang: Eine Veränderung in einer Variable zieht eine Veränderung in einer anderen Variable nach sich, während alle anderen Umstände gleich bleiben (Wooldridge, p.798).
Experiment: Ein Forscher wirkt manipulativ darauf ein, wie Daten generiert werden.

Sitzung 2: Nimm Maß: Deskriptive Statistik 1

Literatur
Fields, Kapitel 2 und 3
Mittag/Schüller, Kapitel 2
Sibbertsen/Lehne, Kapitel 1.2

Statistik ist dazu da Daten zu organisieren und zu analysieren. Aber was für Daten? Woher wissen wir welche Daten wir erheben sollten. Das sagt uns unsere Forschungsfrage, das Ziel unseres Erkenntnisfeldzuges. Der Inhalt definiert hier die Methode.

Sagen wir, wir haben uns klar gemacht, welches Theme uns interessiert, was unsere Fragestellung ist. Diese Fragestellung bezieht sich immer auf ein Phänomene, das wir beschreiben oder erklären wollen. Mehrheitlich werden das Phänomenee sein, die sich mit der Gesellschaft oder Politik befassen. Diese Phänomenee haben Charakteristiken, die wir mit Variablen beschreiben können. Aber welche Variablen? Dafür müssen wir wissen, welche Charakteristiken uns interessieren. Wir brauchen erstmal die **Konzepte**, welche das Phänomene beschreiben. Dann können wir uns fragen, was ist ein geeignetes **Messverfahren**, um diese Konzepte zu repräsentieren. Darum geht es heute.

Konzepte und deren Operationalisierung

Beginnen wir mit einem Beisipel. Als Sozialwissenschaftler beschäftigen wir uns oft mit Regierungssystemen und dabei oft mit Fragen wie, was ist eine Demokratie, wie funktioniert die Demokratie oder ist die Demokratie in Gefahr? Wir brauchen also eine Definition von Demokratie? Wir müssen das Konzept **Demokratie** detailiert erfassen.

Ich habe euch in der Sitzung gefragt, was euch einfällt, wenn ihr den Begriff "Demokratie" hört. Hier sind eure Antworten:

Definiere Demokratie:
"Absetzbarer Anführer, Vom Volk fürs Volk, Staatsform, in der die Macht vom Volk ausgeht, Meinungsfreiheit, die Existenz einer Opposition und Rechtsstaatlichkeit, Volksherrschaft, Freie Wahlen, Mehrere Parteien, Parlament vom Volk gewählt, Pressefreiheit, Alle Macht geht vom Volk aus, Mitbestimmung, Free and fair elections, Civil liberties, Gewaltentrennung, Repräsentation vom ganzen Volk , Gleichwertigkeit der Stimmen, Wahlrecht, Volk entscheidet Gleichberechtigung, Meinungsfreiheit"

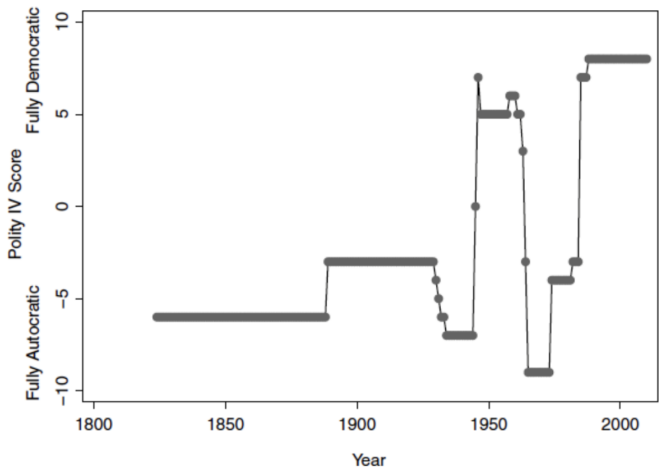
Ihr seht, Demokratie ist ein Konzept mit vielen Element, vielen Dimensionen. Sagen wir nun, wir möchten wissen, ob Demokratie und wirtschaftlicher Wohlstand zusammenhängen. Wir brauchen also ein Maß, dass uns für jeden Staat angibt, wie "demokratisch" der Staat ist. Hier sind eure Antworten, wie man den Status der Demokratie messen könnte:

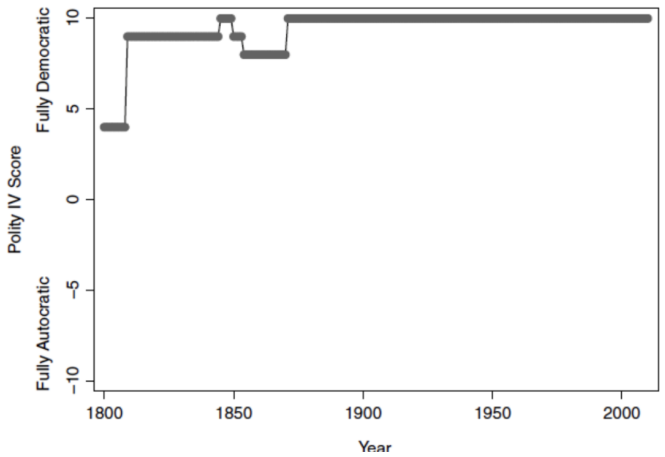
Messe Demokratie:
"Gibt es faire Wahlen? Wahlbeteiligung, Berichterstattung der Medien, Stärke des Rechtssystems, Stärke der Regierung (Exekutive und Premierminister/Präsident) in Relation zur Parlamentsmacht, Existenz von Volksabstimmungen, Existenz einer kritische Presse, Gewaltenteilung bzw. Inwiefern die Gewalten sich gegenseitig kontrollieren, Level an Diskriminierung, Untergraben der Rechtsstaatlichkeit durch Exekutive, Zufriedenheit mit dem System."

Das sind auch alles gute und umsetzbare Messverfahren, um verschiedenste Elemente des Konzepts "Demokratie" empirisch abzubilden. Uns muss nur bewusst sein, dass wir als Analysten, im Moment wenn wir ein Konzept, und dann ein Messverfahren entwickeln, viele Entscheidungen treffen müssen. Wie wir das Konzept, dass uns interessiert, definieren und wie wir es in einem Messverfahren operationalisieren, beeinflusst welche Statistik wir anwenden können und welche Rückschlüsse wir ziehen können.

Definition
Ein **Konzept** beschreibt eine Sammlung oder Klasse an Dingen, die gleichartig sind, da sie gemeinsame Charakteristiken oder Verhalten aufweisen.

Viele Konzepte, die uns interessieren, wurden schon behandelt und es gibt gute Ansatzpunkte in der Forschung für uns, um an detailierte Konzepte und erprobte Messverfahren zu kommen. Hier sind Abbildungen, die den Demokratie-Score des Polity-Projektes wiedergeben. Diese Konzeptualisierung und Operationalisierung ist weit verbreitet, vielfach validiert und diskutiert.





Sicherlich, die Polity Scores sind nicht unumstritten und teilweise unbrauchbar. Das etwas die USA in den 1950 Jahren so demokratisch wie heute sein soll, ist beim damaligen Ausschluss großer Teile der schwarzen Bevölkerung doch eher zu bezweifeln. Übrigens, → hier ist der Link zum Codebook des Polity-Projekts. Es enhält die sehr ausführliche Abhandlung darüber, wie Demokratie gemessen wird. Nichtsdestotrotz, alles brauchbar.

Zusammenfassend sei gesagt, ein **gutes Konzept** ist durch folgende Merkmale charakterisiert: **umfassend, präzise, representativ.**

Messverfahren

Was sind nun gute Messverfahren für die Konzepte, die uns interessieren. Erstmal wieder, hier eine längere Definition: "If you can perceive it, you can measure it. A measurement is an assigned value for a single characteristic. The way a characteristic is captured and, therefore, the way its data should be interpreted determine the measure being used to address the question at hand (Keller, 2016, p. 11)." Klingt richtig, obwohl ich die Reihenfolge ändern würde. Unsere Fragestellung sollte beeinflussen wie wir die Merkmalsausprägung (assigned value for a single characteristic) erheben sollten. Sicher ist wahr, unser Messverfahren bestimmt mit, wie wir die Daten überhaupt interpretieren können.

Ok, hier ist meine Definition:

Definition
Ein **Messverfahren** erhebt Ausprägungen von Variablen, welche gemeinsamen Charakteristiken oder Verhalten beschreiben.

Messarten: "Quantitative" oder "Qualitative" Daten?

Wir können erst einmal zwischen quantitativen und qualitativen Daten unterscheiden. Quantitative Messungen sind dadurch gekennzeichnet, dass die Ausprägung eines Merkmals in form einer Zahl (oder auch einer Kategorie) wiedergegeben werden kann. Qualitative Daten sind in der Regel verbale, visuelle, oder ton-basierte Beschreibungen. Ich setze beide Termine in Anführungszeichen, da sie für mich keine gegensätzlichen oder leicht abgrenzbare Datenarten sind. Jede qualitative Beschreibung kann quantitativ gecoded werden (ob es Sinn für den Erkenntnisgewinn macht, sei mal dahin gestellt) und hinter jedem quantitativen Maß steckt ein detailiertes Konzept oder Beschreibung, was wir als qualitative Information ansehen sollten.

Wie dem auch sei, hier einige Beispiele aus der Vorlesung, was eurer Ansicht nach quantitative oder qualitative Daten wären:

"Quantitativ"	**"Qualitativ"**
Abstimmungsverhalten im Parlament Antworten von einer Umfrage Offizielle Statistiken Abstimmungsverhalten im Parlament, *Reden* *Offizielle Statistiken*	Parteiprogramme Reden Berichte der Regierung Tagebücher Feldnotizen aus einer teilnehmenden Beobachtung Parteiprogramme Politische positionen der parteien Interview Mitschriften Verhalten in einem Experiment Reden

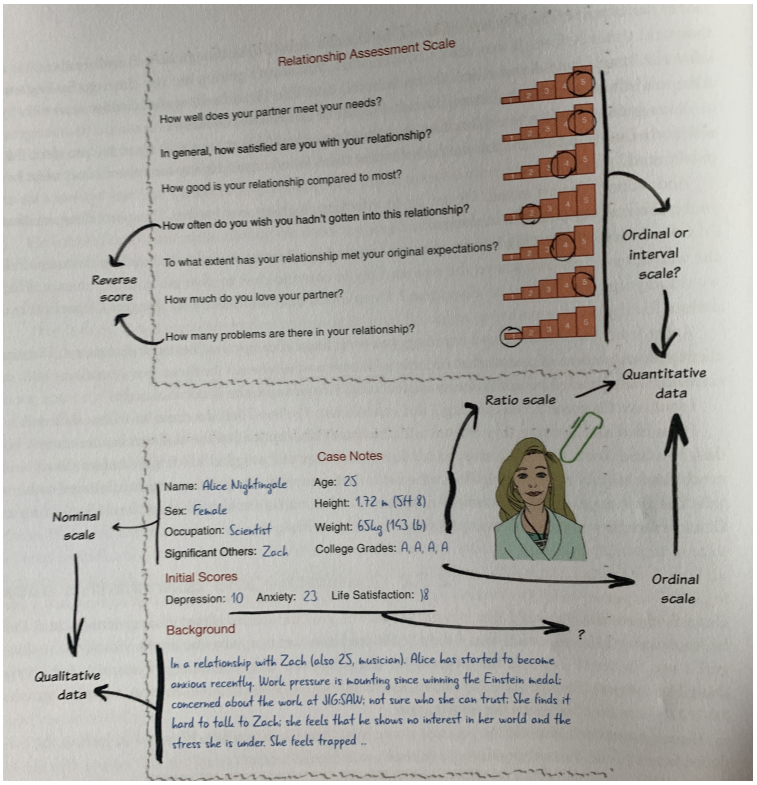
Messskalen

Messverfahren unterscheiden sich auch nach der Skala, die sie benutzen. Hier sind die Definintion, der Skalen, die du kennen solltest:

Definitionen:
Nominalskala: Merkmalsausprägung ist Namen oder Kategorie
Ordinalskala oder Rangskala: Merkmalsausprägung ist Rangordnung ohne sinnvolle Differenz- oder Quotientenbildung
Metrische Skala oder Kardinalskala: Abstände zwischen Merkmalsausprägung interpretierbar – Untertypen: Intervallskala, Verhältnis-/Ratioskala

Nominalskala und Rangskale sind immer **diskrete** Skalen. Die Ausprägungen der gemessenen Merkmale können immer abgezählt werden. Metrische Skalen sind manchmal diskret und manchmal **stetig**. In einer stetigen Skala, können die Ausprägungen als jegliche reelle Zahlen beschrieben werden. Euer Gewicht etwa, wird auf einer stetigen Skala gemessen (wobei wir meist bei einem oder zwei Ziffern hinter dem Komme aufhören). Euer Alter hingegen wird in Umfragen meist als diskrete Zahl erhoben (20 Jahre, 21 Jahre usw. aber nicht als 20,45 Jahre, 20,46 Jahre usw.).

Unten ist eine interessante Beschreibung von Fields, wie verschiedene Messarten und Messskalen die erhobenen Daten abbilden können:



Weiterhin, zur Hilfe, um die Messskalen zu unterscheiden gibt uns Mittag/Schüller diese Tabelle:

Skala		sinnvolle Operationen			
		auszählen	ordnen	Differenz bilden	Quotienten bilden
Nominalskala		ja	nein	nein	nein
Ordinalskala		ja	ja	nein	nein
Metrische Skala	Intervall-skala	ja	ja	ja	nein
	Verhältnis-skala	ja	ja	ja	ja
	Absolut-skala	ja	ja	ja	ja

Zusammenfassend, wie können wir festlegen, welche Messskala wir hier vor uns haben? Stellt euch folgende Fragen:

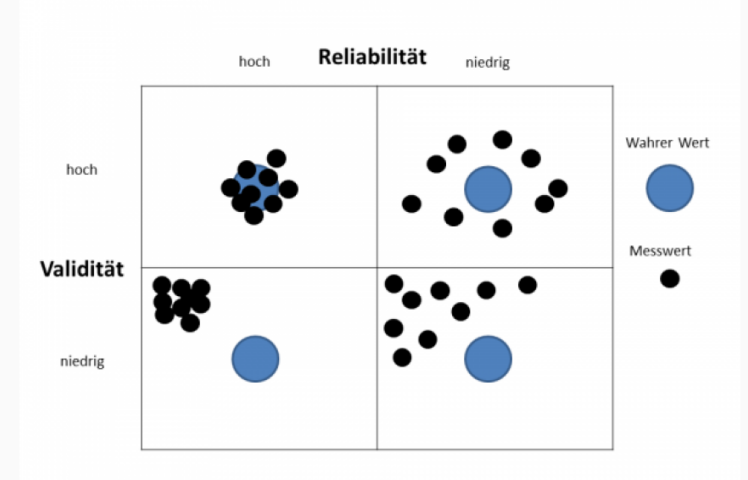
1. Haben die Kategorien eine Rangfolge? Nein, dann Nominalskala
2. Ja, dann: Bilden die Merkmalsausprägungen sinnvolle Differenzen? Nein, dann Ordinalskala
3. Ja, dann: Sind die Merkmalsausprägungen immer positive? Nein, dann Intervallskala - sonst Verhältnisskala

Gute Messverfahren

Was sind nun gute Messverfahren? Ein brauchbares Messverfahren weist folgende Merkmale auf:

0. **Keine Messfehler**: Keine Differenz zwischen tatsächlichen und gemessenen Wert
1. **Objektivität**: Intersubjektive Nachvollziebarkeit
2. **Reliabilität**: Messgenauigkeit (robust)
3. **Validität**: Gültigkeit

Reliabilität und Validität sind hierbei miteinander zu betrachten. Sicherlich wollen wir beides, hohe Reliabilität und Validität, aber wir können uns viele Situationen vorstellen, wo wir das eine oder andere nicht erreichen.



Was solltet ihr von heute mitnehmen?

1. Den Unterschied zwischen Korrelation und Kausalzusammenhang, sowie zwischen experimentellen Daten und Beobachtungsdaten verstehen.
2. Wissen, was ein Konzept ist und wie es in ein Messverfahren übertragen werden kann.
3. Erkennen, was ein gutes Messverfahren ist.

Sitzung 3: Gut verteilt: Deskriptive Statistik 2

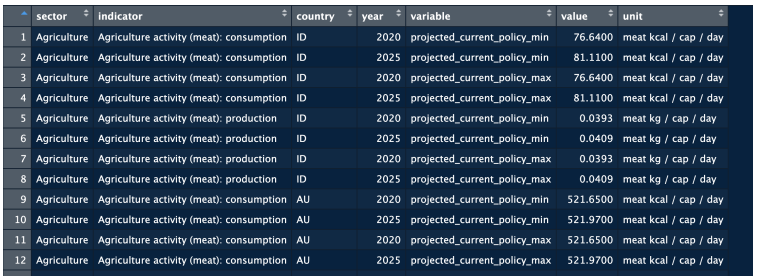
Literatur
Fields, Kapitel 2 und 3
Mittag/Schüller, Kapitel 4 und 5
Sibbertsen/Lehne, Kapitel 2 und 3

Jetzt sind wir an dem Punkt angekommen, wo wir unsere Stichprobe gezogen haben und die Daten vor uns liegen. Wir nehmen mal weiter an, dass wir valide und reliable Messungen durchgeführt haben. Was nun. Jetzt wollen wir etwas über die Daten erfahren. Wie schon erwähnt, sehen wir uns in dieser Vorlesung zwei Wege, die mit der Stichprobe generierten Daten zu erkunden: **Beschreibung** und **Inferenz**.

Wir beginnen mit Ersterem, der Beschreibung unserer Daten. Wie sollen wir das machen? Wir wollen eine Beschreibung, welche die Menge an Information, die in Rohdaten stecken, reduziert. Dazu nutzen wir einige gängige Methoden: **Häufigkeitsverteilungen**, **Lagemaße** und **Streuugsmaße**.

Häufigkeitsverteilung (Frequency distributions)

Beginnen wir mit den Daten, die unsere Stichprobe erzeugt hat. Als Beispiel nehmen wir Emissionsdaten und unsere Stichprobe umfasst das Jahr 2020 und alle Länder. Unten seht ihr erstmal die Tabelle dieser Daten, zumindest die ersten paar Zeilen:



	sector	indicator	country	year	variable	value	unit
1	Agriculture	Agriculture activity (meat): consumption	ID	2020	projected_current_policy_min	76.6400	meat kcal / cap / day
2	Agriculture	Agriculture activity (meat): consumption	ID	2025	projected_current_policy_min	81.1100	meat kcal / cap / day
3	Agriculture	Agriculture activity (meat): consumption	ID	2020	projected_current_policy_max	76.6400	meat kcal / cap / day
4	Agriculture	Agriculture activity (meat): consumption	ID	2025	projected_current_policy_max	81.1100	meat kcal / cap / day
5	Agriculture	Agriculture activity (meat): production	ID	2020	projected_current_policy_min	0.0393	meat kg / cap / day
6	Agriculture	Agriculture activity (meat): production	ID	2025	projected_current_policy_min	0.0409	meat kg / cap / day
7	Agriculture	Agriculture activity (meat): production	ID	2020	projected_current_policy_max	0.0393	meat kg / cap / day
8	Agriculture	Agriculture activity (meat): production	ID	2025	projected_current_policy_max	0.0409	meat kg / cap / day
9	Agriculture	Agriculture activity (meat): consumption	AU	2020	projected_current_policy_min	521.6500	meat kcal / cap / day
10	Agriculture	Agriculture activity (meat): consumption	AU	2025	projected_current_policy_min	521.9700	meat kcal / cap / day
11	Agriculture	Agriculture activity (meat): consumption	AU	2020	projected_current_policy_max	521.6500	meat kcal / cap / day
12	Agriculture	Agriculture activity (meat): consumption	AU	2025	projected_current_policy_max	521.9700	meat kcal / cap / day

In diesem Datensatz finden sich mehrere Variablen: “sector”, “indicator”, “country”, “year”, “variable”, “value”, “unit”. Wir haben für alle (annerkannten) Länder dieser Welt sowie einiger Regionen (Zusammenfassung von Ländern), Emmissionsdaten aus dem Jahr 2020 (und einer Projektion für das Jahr 2025) für mehrere Wirtschaftssektioren und Indikatoren. Wie viele Beobachtungen haben wir pro Land? Das kann uns die **absolute Häufigkeit** in einer **Häufigkeitstabelle** sagen:

```
emissionsData %>%
  freq(country)
```

```
## Frequencies
## emissionsData$country
## Type: Factor
##
##                          Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------- ------- --------- -------------- --------- --------------
##                    AE     871     2.259          2.259     2.259          2.259
##                Africa       0     0.000          2.259     0.000          2.259
##                    AR    1030     2.672          4.931     2.672          4.931
##       Asia-Pacific-40       0     0.000          4.931     0.000          4.931
##                    AU    1188     3.081          8.012     3.081          8.012
##                    BR    1855     4.811         12.824     4.811         12.824
##                    BT     189     0.490         13.314     0.490         13.314
##                    CA    1245     3.229         16.543     3.229         16.543
##                    CH    1106     2.869         19.412     2.869         19.412
##                    CL    1126     2.921         22.332     2.921         22.332
##                    CN    2041     5.294         27.626     5.294         27.626
##                    CR     919     2.384         30.010     2.384         30.010
##                    DE    1229     3.188         33.198     3.188         33.198
##                    ET     798     2.070         35.267     2.070         35.267
##                    EU       0     0.000         35.267     0.000         35.267
##                    FR    1249     3.240         38.507     3.240         38.507
##                    GB    1271     3.297         41.804     3.297         41.804
##                    GM     229     0.594         42.398     0.594         42.398
##                    ID    1565     4.059         46.457     4.059         46.457
##                    IN    1740     4.513         50.970     4.513         50.970
##                    IT    1150     2.983         53.953     2.983         53.953
##                    JP    1303     3.380         57.333     3.380         57.333
##                    KR    1028     2.666         59.999     2.666         59.999
##                    KZ     999     2.591         62.590     2.591         62.590
```

```
##       Latin America-31       0     0.000         62.590     0.000         62.590
##                     MA      20     0.052         62.642     0.052         62.642
##           Middle East       0     0.000         62.642     0.000         62.642
##                     MX    1554     4.031         66.673     4.031         66.673
##                     NO    1208     3.133         69.806     3.133         69.806
##         Non-EU Europe       0     0.000         69.806     0.000         69.806
##                     NP     803     2.083         71.889     2.083         71.889
##                     NZ    1136     2.947         74.835     2.947         74.835
##                     PE     924     2.397         77.232     2.397         77.232
##                     PH     977     2.534         79.766     2.534         79.766
##                     RU    1162     3.014         82.780     3.014         82.780
##                     SA     873     2.264         85.044     2.264         85.044
##                     SG      62     0.161         85.205     0.161         85.205
##                     TR    1153     2.991         88.196     2.991         88.196
##                     UA    1006     2.609         90.805     2.609         90.805
##                     US    2288     5.935         96.740     5.935         96.740
##                  World       0     0.000         96.740     0.000         96.740
##                     ZA    1257     3.260        100.000     3.260        100.000
##                   <NA>       0                                0.000        100.000
##                  Total   38554   100.000        100.000   100.000        100.000
```

Die Häufigkeitstabelle oben zeigt uns für jede Ausprägung (ZA=Zambia, US=United States, etc) des Merkmals "country", wie viele Beobachtungen wir in unsere Stichprobe haben. Wir sehen diese **absolute Häufigkeit** in der zweiten Spalte von links ("Freq").

Exkurs: Notation

Statistiken sind in mathematischer Notation definiert. Ihr müsst also gewisse Rechenwege beherrschen, um diese Definitionen verstehen zu können. Folgendes brauchen wir:

- Rechnen mit Klammern: immer von innen nach außen: $2 \times (3 + 1) = 2 \times 4 = 8$
- Rechnen mit Exponenten: erst die Exponenten: $2 \times 4^2 = 2 \times 16 = 32$
- Grundrechenarten: Erst Division/Multiplikation, dann Addition/Subtraktion $10 + \frac{32}{4} = 10 + 8 = 18$
- Summen und Produktzeichen: $\sum_{i=1}^{N}$ und $\prod_{i=1}^{N}$

Absolute Häufigkeit

Ok, dann können wir mit dieser Notation jetzt die absolute Häufigkeit definieren.

Definition: Bei einer Stichprobe vom Umfang N wird ausgezählt, wie häufig jeder der k Ausprägungen auftritt. Diese Anzahl bezeichnet man als **absolute Häufigkeit** einer Ausprägung n_j (wobei j für eine bestimmte der k Ausprägungen steht).

$$\sum_{j=1}^{k} n_j = N$$

Die Summe der absoluten Häufigkeit (frequency) von allen k Ausprägungen ergibt die Stichprobengröße N. Quelle: Sibbertsen/Lehne, S.13/14

Relative Häufigkeit

Wenn wir uns die Tabelle oben noch einmal ansehen, ergeben sich noch weitere Wege, die Variable "country" zu beschreiben. Wie oft sehen wir eine bestimmte Merkmalsausprägung relative zu wie oft wir alle anderen

Merkmalsausprägungen sehen. Das ist die **relative Häufigkeit**. Wir finden die relative Häufigkeit in der zweiten Spalte von rechts ("% Total").

Definition: Die absolute Häufigkeit bezogen auf den Stichprobenumfang

$$\frac{n_j}{N}$$

ist die **relative Häufigkeit** (relative frequency) einer Ausprägung.
Die Summe der relativen Häufigkeiten von allen k Ausprägungen ergibt 1.

$$\sum_{j=1}^{k} \frac{n_j}{N} = 1$$

Quelle: Sibbertsen/Lehne, S.13/14

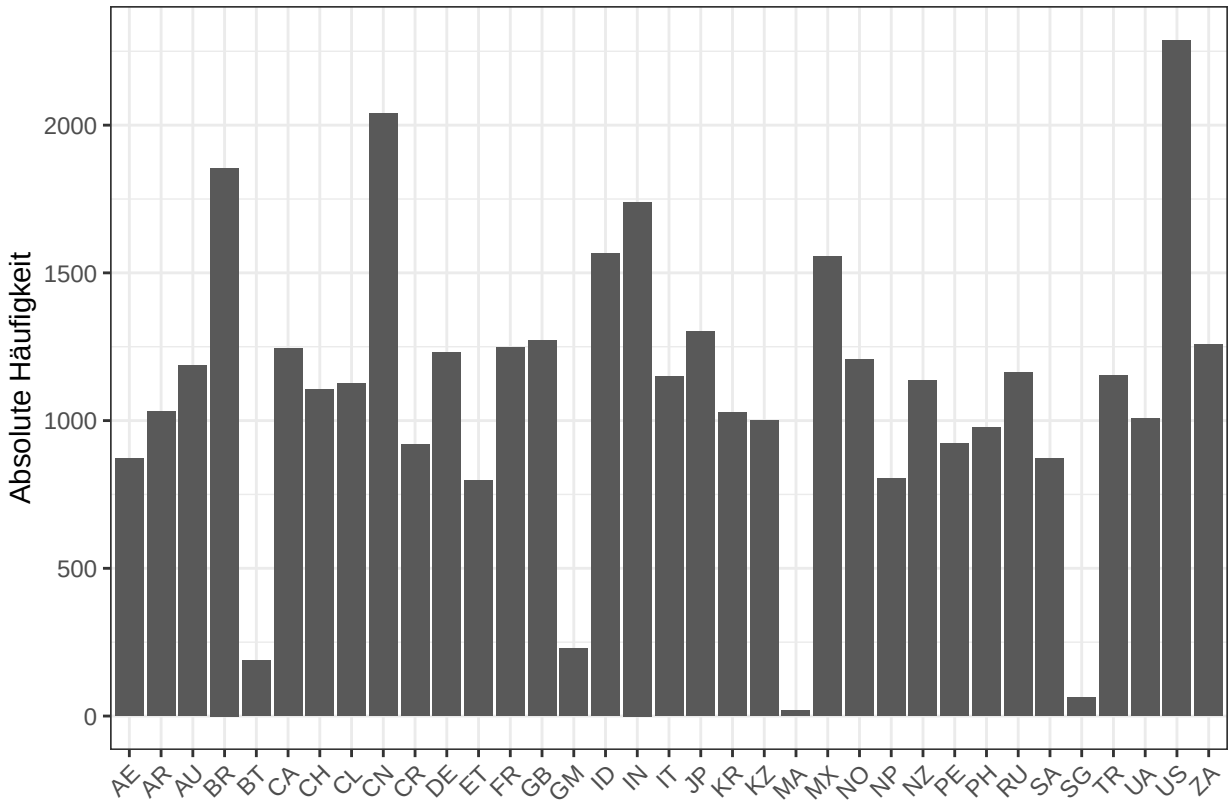
Eine weitere Häufigkeit ist oft interessant für uns, etwa um noch mehr über ungleiche Verteilungen sagen zu können. Das ist die Kummulierte Relative Häufigkeit (Spalte ganz rechts, "% Total Cum."). Wir haben Daten für 35 Länder. Wenn wir wissen wollen, ob wir die selbe Anzahl an Beobachtungen für jedes Land haben, können wir leicht die absolute oder relative Häufigkeit zwischen den Ländern vergleichen (Spalte "Freq" oder "% Total"). Wenn wir etwa wissen wollen, ob wir genau 20% der Beobachtungen sich auf die ersten 20% der Länder (die ersten 7 Länder von oben) verteilen, dann könnten wir checken ob die kummulierte relative Häufigkeit nach Land 7 von oben gezählt bei 20% ist. Die kummulierte relative Häufigkeit ist aber nur bei 16.5%. Wir haben also relative mehr Beobachtungen für die anderen Länder. Das ganze ist hier jetzt nicht so interessant, aber wenn wir uns zum Beispiel die Einkommensverteilung ansehen, dann ist es wohl interessant ob die 50% der Bevölkerung in der unteren Hälfte der Einkommensverteilung auch 50% des Einkommens besitzt. Mehrheitlich ist das in westlichen Gesellschaften nicht der Fall, also hätten wir mehr Beobachtungen in den unteren Einkommen, als in den höheren Einkommen.

Graphische Darstellung von Häufigkeit

Oft werden absolute und relative Häufigkeit graphisch dargestellt. Diese Art an Graph ist ein **Histogram**. Für diskrete Variablen, ist ein Histogram eine sehr aussagekräftige Darstellungsform. Hier ist die absolute und relative Häufigkeit der Beobachtungen in den verschiedenen Regionen der Welt.

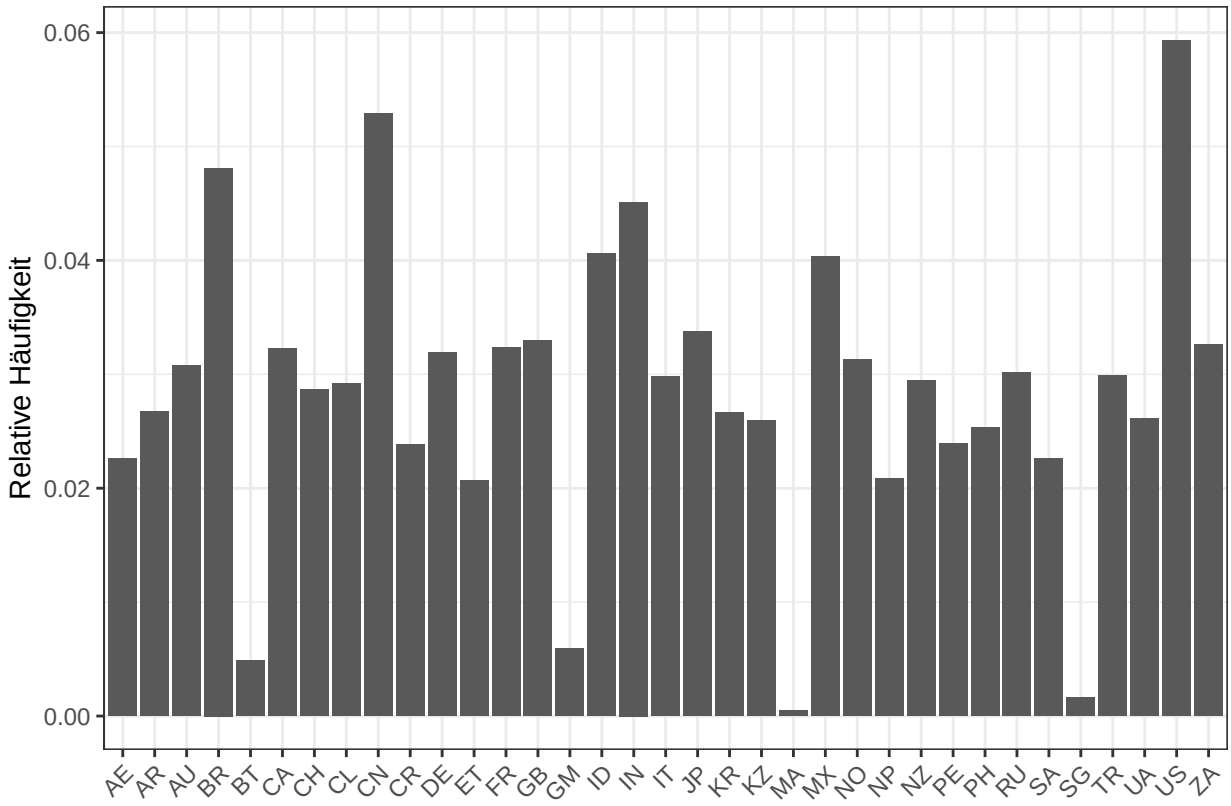
Erstmal die absolute Häufigkeit:

```
emissionsData %>%
  ggplot(aes(x=country)) +
  geom_histogram(stat="count") +
  labs(x='',y='Absolute Häufigkeit') +
  theme_bw() +
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1))
```



Dann die relative Häufigkeit:

```
emissionsData %>%
  ggplot(aes(x=country,y=(..count..)/sum(..count..))) +
  geom_bar() +
  labs(x='',y='Relative Häufigkeit') +
  theme_bw() +
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1))
```

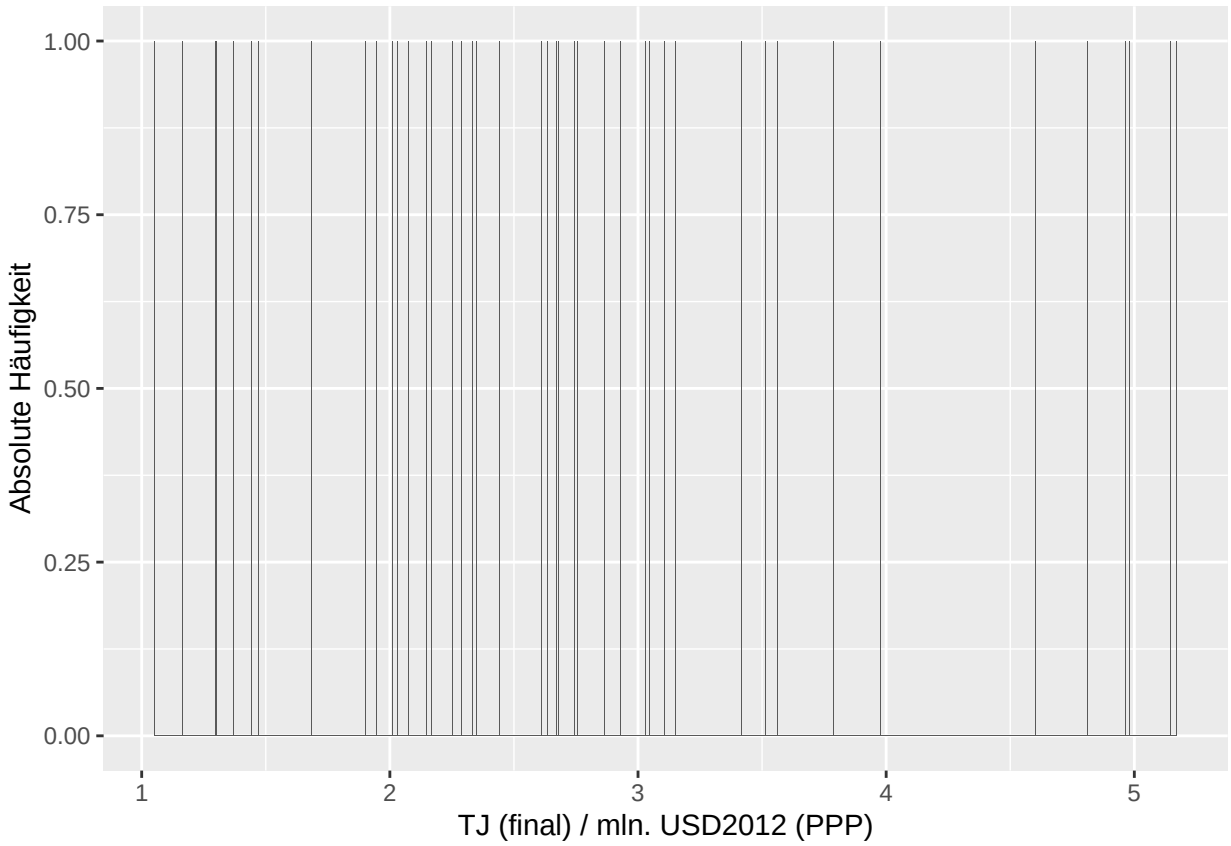


Die Verteilung sieht gleich aus, wie im Graphen der absoluten Häufigkeit, aber die y-Achse ist nun von 0 bis 1 skaliert.

Das Merkmal "country" ist eine diskrete Variable, im Gegensatz zu einer stetigen Variable, mit Ländernamen als Ausprägung. Die Verteilung der Beobachtungen über diese Länder hinweg können wir mit einem Histogramm gut für eine stetige Variable abbilden. Stetige Variablen könnten wir auch mit Hilfe eines Histogramms abbilden. Von einem Histogramm erhoffen wir uns eine **visuelle** Beschreibung. Sobald wir eine stetige Variable durch ein Histogramm beschreiben wollen, laufen wir in das Problem, dass es möglicherweise sehr viele verschiedene Werte gibt.

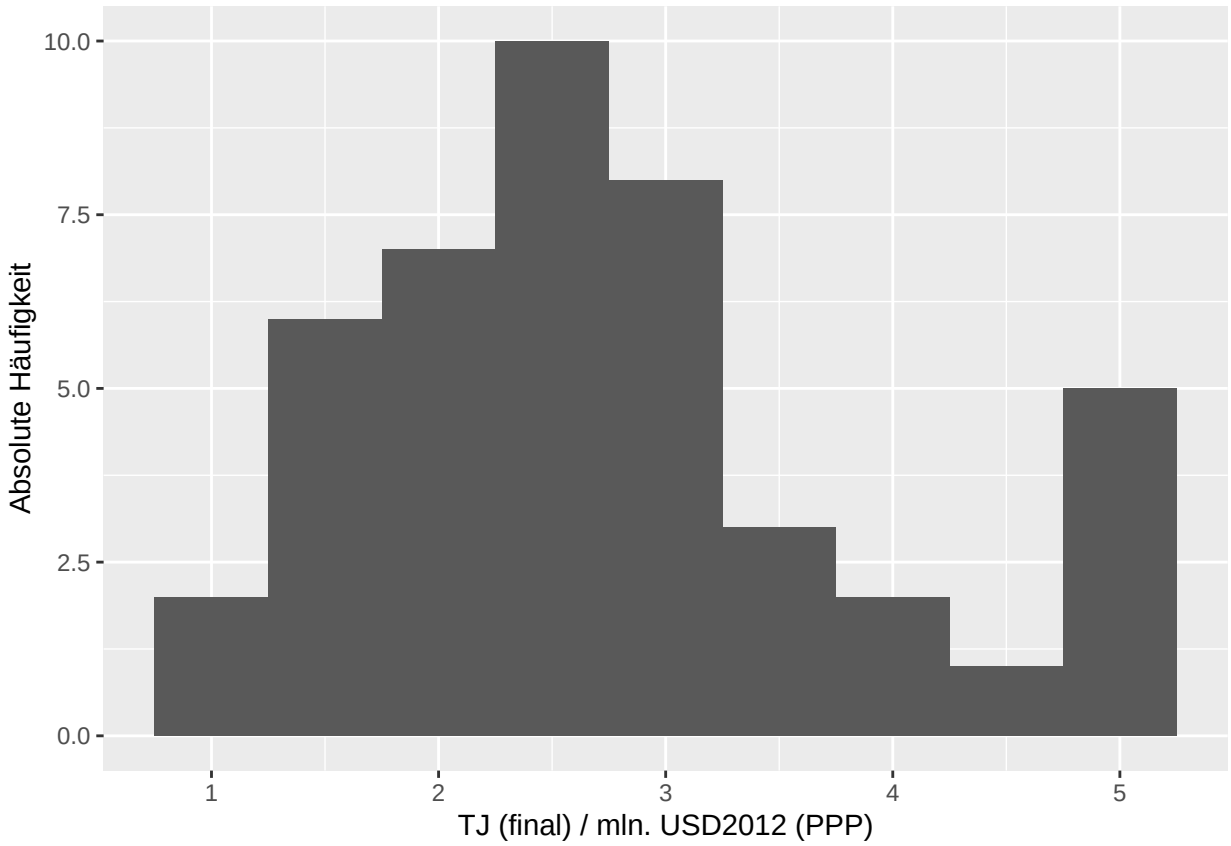
Die Phänomene für die wir uns interessieren sind natürlich oftmals durch viele Merkmale charakterisiert. Wenn wir etwa die Energieintensität der Volkswirtschaft, eine stetige Variable, als Histogramm abbilden, so dass wir einen Balken für jede Ausprägung der Variable bekommen (wie wir das auch für jedes Land, die Ausprägung der Variable "country" getan haben), dann lernen wir etwas weniger als zuvor, wo in den Histogramm die Verteilung eine deutliche Form angenommen hat.

```
emissionsData %>%
  filter(indicator=='Final energy intensity of GDP'&
           variable=='projected_current_policy_min') %>%
  ggplot(aes(x=value)) +
  geom_histogram(binwidth=.001) +
  labs(x='TJ (final) / mln. USD2012 (PPP)',y='Absolute Häufigkeit')
```

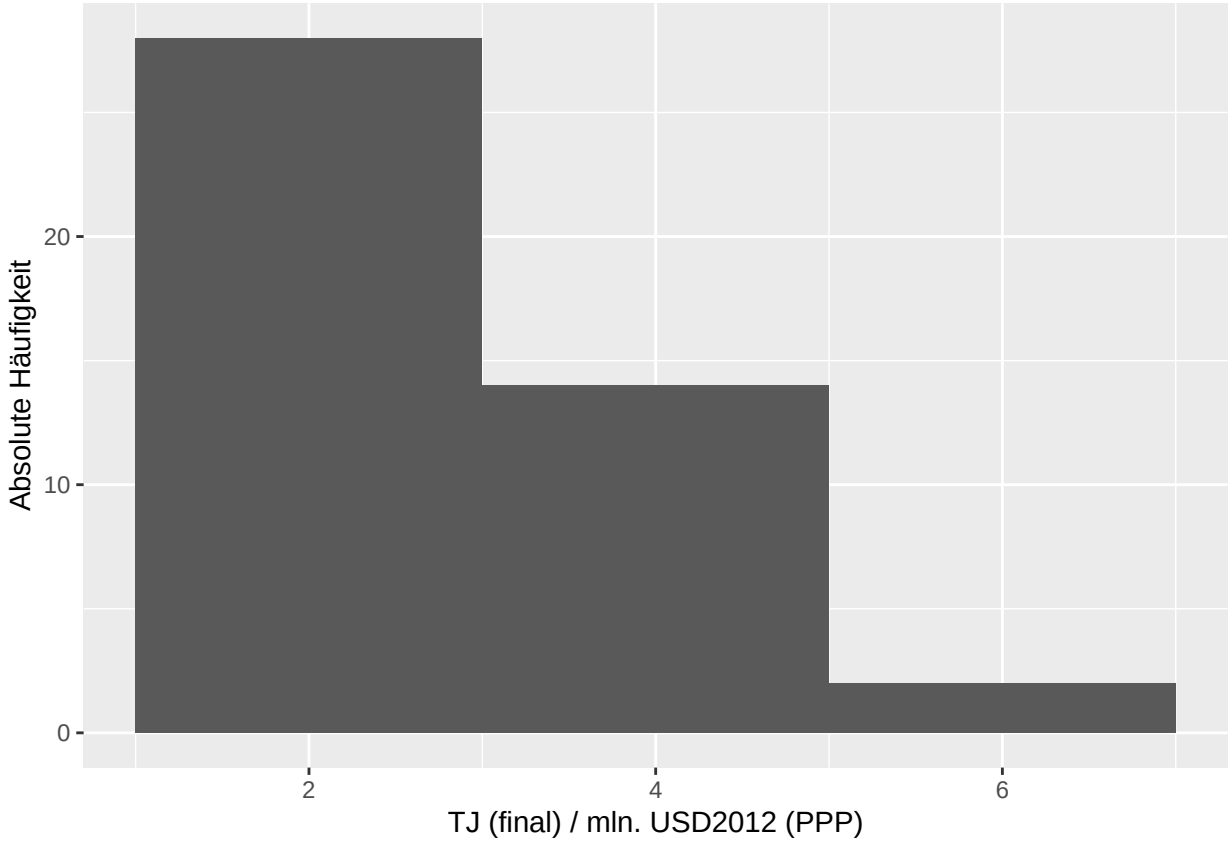


Für eine stetige Variable macht es oft mehr Sinn, nicht alle Ausprägungen bis zur letzten Nachkommastelle in einer Verteilung abzubilden. Es ist oft besser, Ausprägungen in einem Balken zusammen zunehmen (mit anderen Worten, die Werte zu runden). Nun die Frage, wie viele Werte sollten wir zusammen nehmen? Unten seht ihr zwei Graphen, beide basieren auf den selben Daten, aber der erste Graph nimmt weniger Werte in einem Balken zusammen, der andere mehr. Das ist wichtig: eure Entscheidung, wie ihr die Daten darstellt, mag uns einen sehr unterschiedlichen Eindruck der Verteilung vermitteln.

```
emissionsData %>%
  filter(indicator=='Final energy intensity of GDP'&
           variable=='projected_current_policy_min') %>%
  ggplot(aes(x=value)) +
  geom_histogram(binwidth=.5) +
  labs(x='TJ (final) / mln. USD2012 (PPP)',y='Absolute Häufigkeit')
```

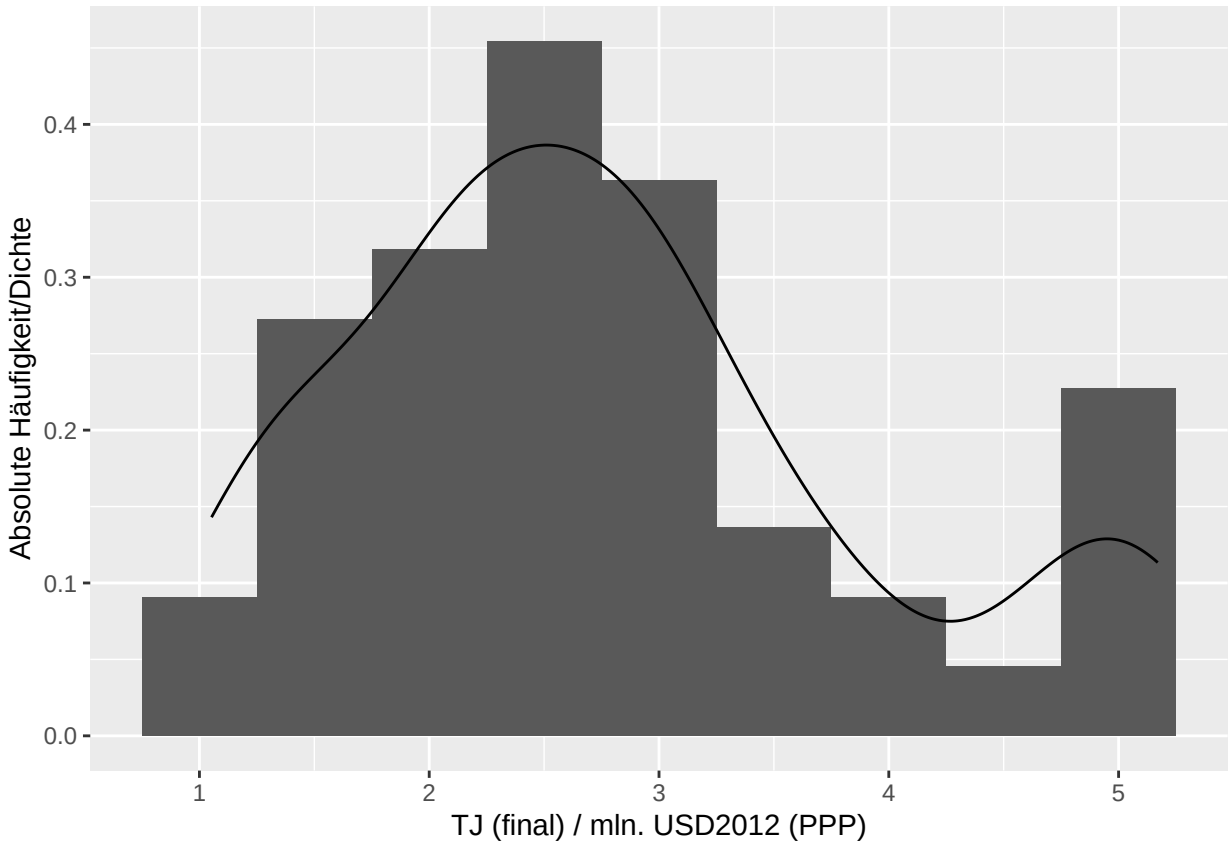



```
emissionsData %>%
  filter(indicator=='Final energy intensity of GDP'&
           variable=='projected_current_policy_min') %>%
  ggplot(aes(x=value)) +
  geom_histogram(binwidth=2) +
  labs(x='TJ (final) / mln. USD2012 (PPP)',y='Absolute Häufigkeit')
```

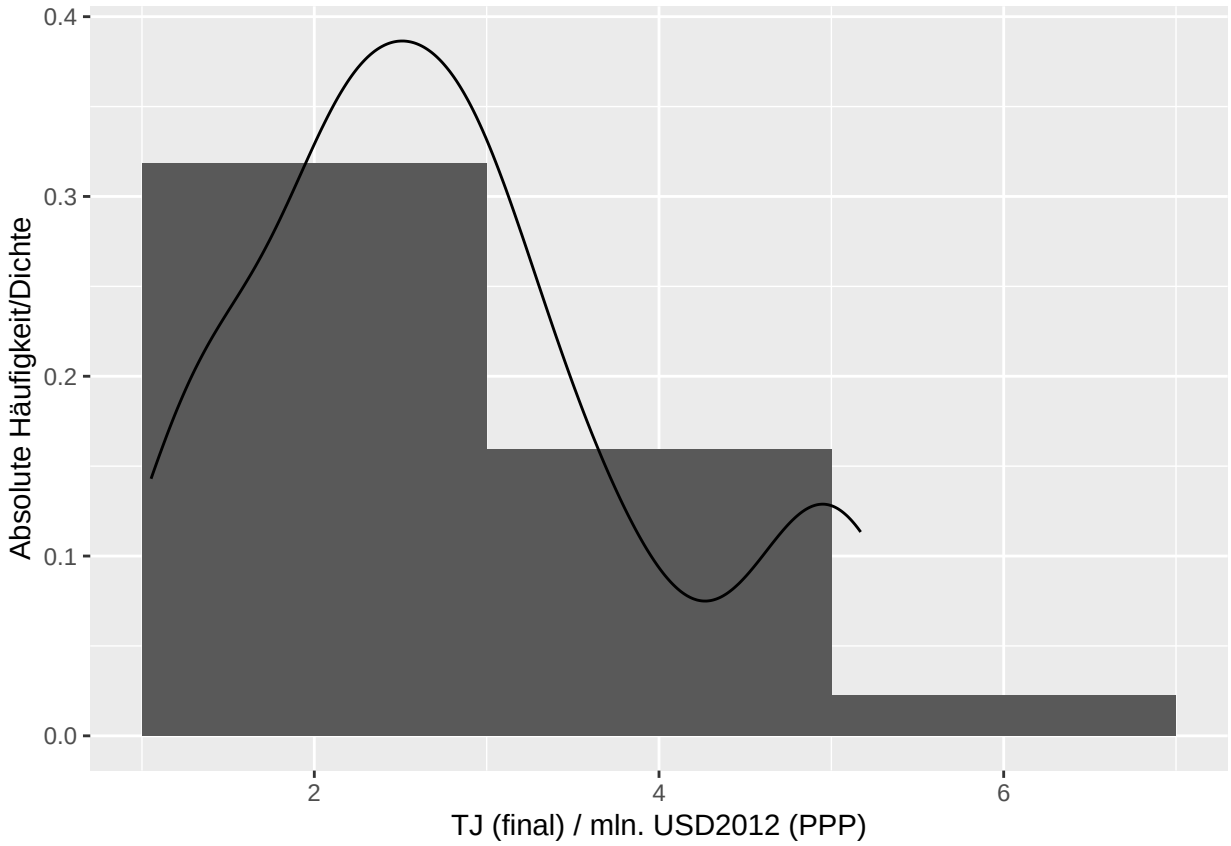


Eine Alternative zum Histogramm ist die Darstellung der **Dichte** der Werte einer Variable. Wir gehen in der nächsten Sitzung noch einmal auf die diskrete und stetige Wahrscheinlichkeitsverteilungen ein. Ein Histogramm zeigt die empirische Wahrscheinlichkeitsfunktion, ein Dichteplot die Annäherung an die Dichte einer empirischen Verteilung (es ist immer nur eine Annäherung, weil wir nie unendlich viele Ausprägungen einer Variable habe auf jedem möglichen Werte). Hier sind noch einmal die Histogramme von eben, nun mit einem Dichteplot darüber gelegt.

```
emissionsData %>%
  filter(indicator=='Final energy intensity of GDP'&
           variable=='projected_current_policy_min') %>%
  ggplot(aes(x=value,y=..density..)) +
  geom_histogram(binwidth=.5) +
  geom_density() +
labs(x='TJ (final) / mln. USD2012 (PPP)',y='Absolute Häufigkeit/Dichte')
```



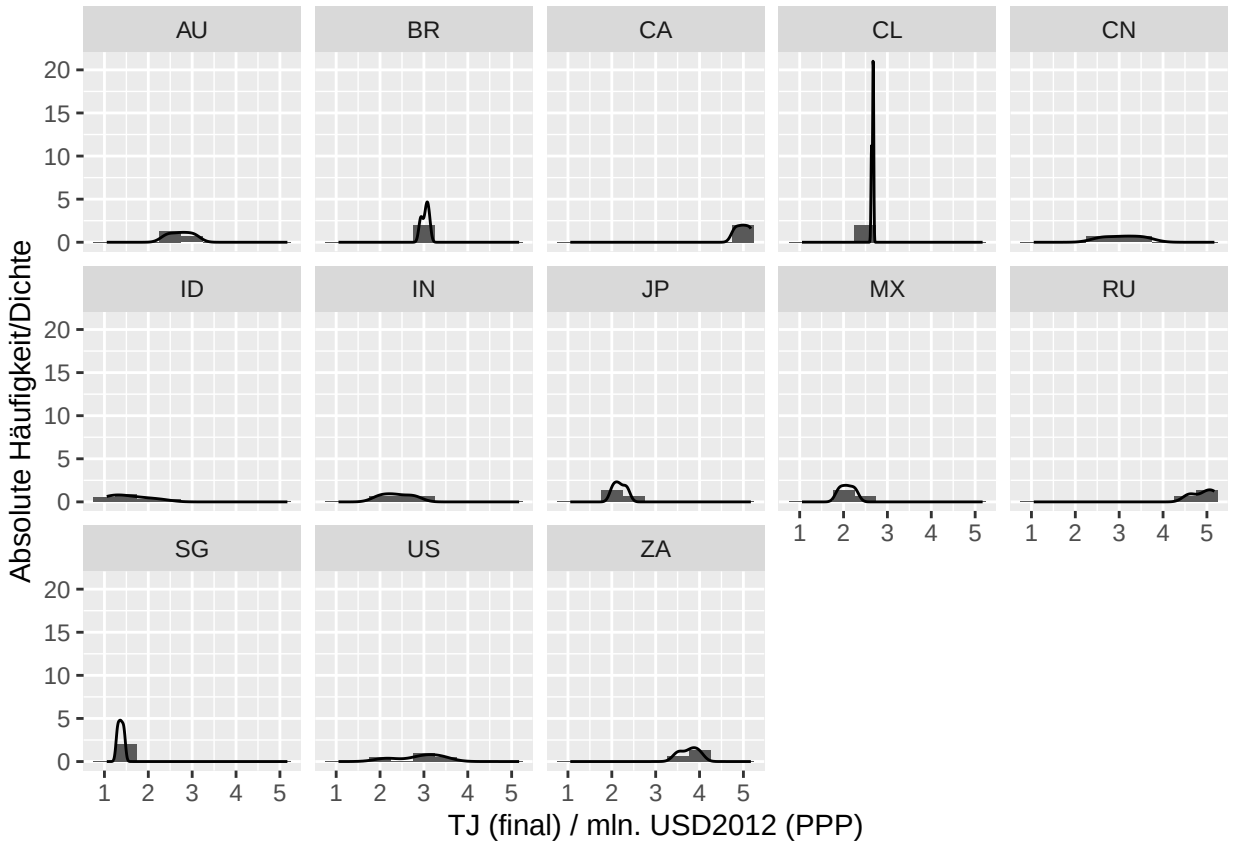

```
emissionsData %>%
  filter(indicator=='Final energy intensity of GDP'&
           variable=='projected_current_policy_min') %>%
  ggplot(aes(x=value,y=..density..)) +
  geom_histogram(binwidth=2) +
  geom_density() +
  labs(x='TJ (final) / mln. USD2012 (PPP)',y='Absolute Häufigkeit/Dichte')
```



Was ihr hier sehen solltet, ist dass der Densityplot die gleiche Verteilung angibt, also die Werte der Variable bis in die letzte Nachkommastellung abbildet. Das ist besser für eine stetige Variable.

Wenn wir jetzt die Verteilung der Emissionsvariable für jedes Land abbilden wollen bekommen wir eine **bedingte Häufigkeitsverteilung**. Solch eine Darstellung dient immer noch der Beschreibung, aber enthält sicherlich mehr Information aus dem Vergleich der Länder.

```
emissionsData %>%
  filter(indicator=='Final energy intensity of GDP'&
           variable=='projected_current_policy_min') %>%
  ggplot(aes(x=value,y=..density..)) +
  facet_wrap(~country,ncol=5) +
  geom_histogram(binwidth=.5) +
  geom_density() +
labs(x='TJ (final) / mln. USD2012 (PPP)',y='Absolute Häufigkeit/Dichte')
```



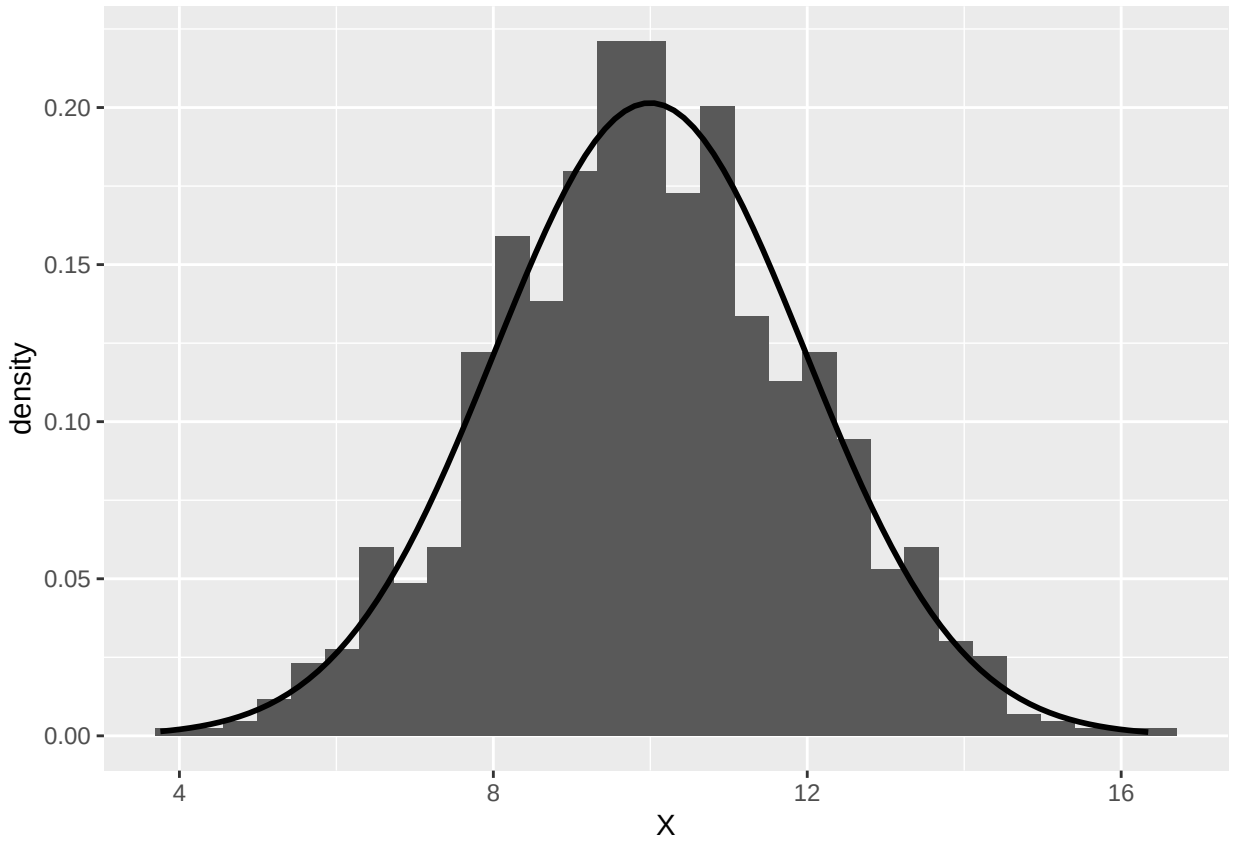
Wichtige Verteilungen

Wir haben uns bisher **empirische Häufigkeitsverteilungen** angesehen. Das ist die Verteilung der Ausprägung eines Merkmals. Viele Merkmale, welche die Phänomenee für die wir uns interessieren beschreiben, folgen ähnlichen Verteilungen. Hätten wir unendliche viele Beobachtungen für solche Merkmale, würden sich die Verteilungen fast perfekt annähern. Was sind das für Verteilungen und welche Variablen folgen solchen Verteilungen? Die wichtigste, die wir uns jetzt gleich ansehen, ist die **Normalverteilung** und die allgemeinere Übermutter der Normalverteilung, die **Standartnormalverteilung**.

Hier ist erstmal ein Beispiel einer Normalverteilung einer Variable, nennen wir sie mal X, mit Mittelwert 10 und Standardabweichung 2 (ich generiere hier 1000 Beobachtungen einer normal-verteilten Variable zur Illustration):

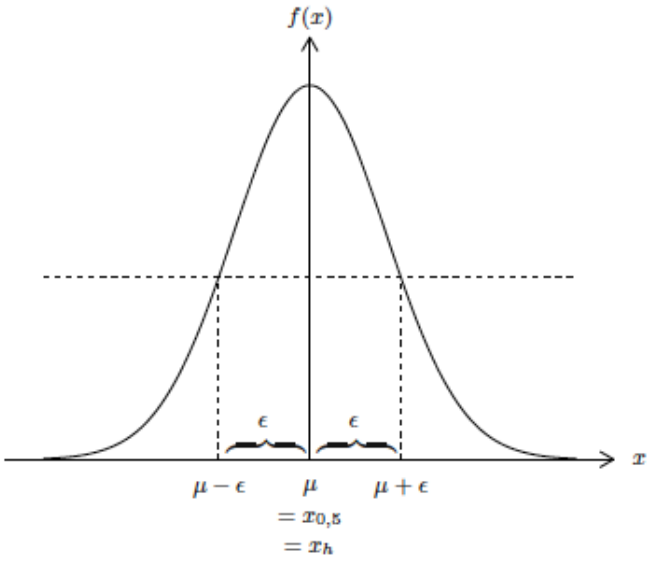
```
data <- data.frame(X=rnorm(1000,10,2))
ggplot(data,aes(X)) +
  geom_histogram(aes(y=..density..)) +
  stat_function(fun=dnorm,args=list(mean=mean(data$X),sd=sd(data$X)),size=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

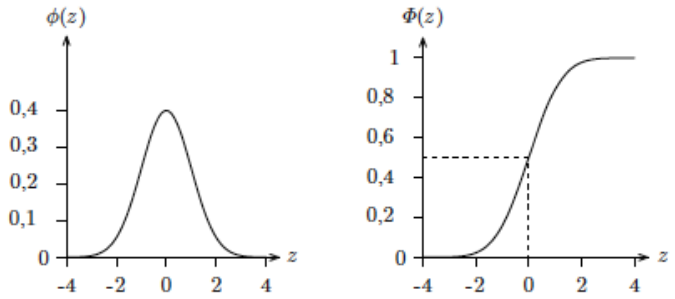


Im Graphen oben, seht ihr eindeutig wie sich die Verteilung der endlichen Werte des Merkmals an die Normalverteilung annähnern. Die Werte sind endlich, da wir hier nur 1000 Beobachtungen haben und eben nicht unendliche viele Beobachtungen. Die schwarze Linie gibt die Dichte der Normalverteilung einer solchen Variable für jeden möglichen Wert. Da für eine stetige Variable die möglichen Werte unendliche viele sind, z.B. ist ein Wert 8 möglich, aber eben auch 8.01, 8.001, 8.00001, etc., ist die schwarze Linie durchgezogen.

Was ist jetzt so toll an einer Normalverteilung? Sie ist symmetrisch, und der **Mittelwert**, der **Median** und der **Modalwert** fallen auf den selben Wert der Variable. Die Abbildung in Sibbertsen/Lehne illustriert diese Charakteristiken. Zur Notation: Die Dichte ist eine Funktion der Werte von x, daher ist auf der y-Achse $f(x)$ zu sehen. Der Mittelwert ist hier mit μ (liest man "mü"), der Median mit $x_{0.5}$ (aka 0.5-Quartile) und der Modalwert mit x_h.



Brilliant. Moving on. Wir können jetzt jede normalverteilte Variable nehmen und deren Standardnormalverteilung abbilden. Man **standartisiert** eine Variable, in dem man von jedem ihrer Werte den Mittelwert der Variable abzieht und dann durch die Standardabweichung der Variable (und die Anzahl der Beobachtungen) teilt. Wir machen das, weil damit können wir die Verteilung einer jeden normalverteilten Variable vergleichen. Die Standartisierung bringt alle Variablen auf die gleiche Skala (mit Hilfe der Teilung durch die Standarabweichung der Variable) und zentriert sie alle bei 0 (durch Abziehen des Mittelwerts). Sibbertsen/Lehne zeigen uns wie solch eine Standardnormalverteilung aussieht. Links ist die Dichte gezeigt (≈ relative Häufigkeit)., rechts die kumulative Dichte (≈ kummulierte relative Häufigkeit).

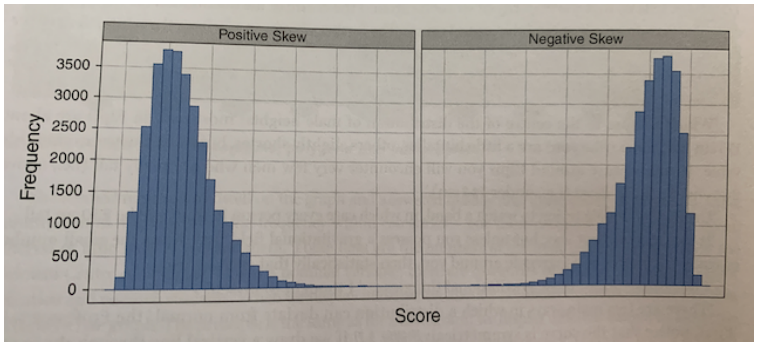


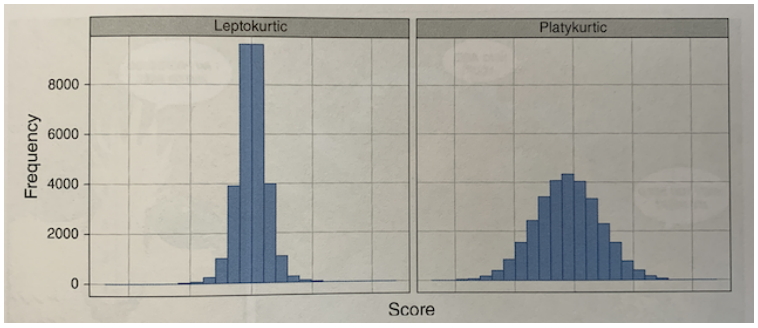
::: {.center data-latex=“ “}

Wir werden noch mehr über diese und andere gänge Verteilungen erfahren, wenn wir uns statistischen Tests widmen.

Schiefmaße: links- und rechtssteile Verteilung (skew)

So, nicht alle Verteilungen sind so hübsch symmetrisch. Hier sind weitere Maße, welche wir benutzen, um Verteilungen zu charakterisieren: **Skeweness** und **Kurtosis**, oder auf Deutsch: Linkssteile und rechtssteile Verteilungen sowie unterschiedliche Grade an **Wölbung**. Warum die so heißen, wir beim Blick auf die Abbildungen in Fields klar:





Lagemaße

In der Diskussion der Normalverteilung, habe ich **Mittelwert**, **Median** und **Modalwert** genannt. Kennt wahrscheinlich schon jeder, aber hier nochmal, um vollständig zu sein, ein paar Definitionen:

Modalwert (mode)

Definition: Die Ausprägung, die am häufigsten im Datensatz enthalten ist, bezeichnet man als **Modus**.
Quelle: Sibbertsen/Lehne, S.54

Median

Definition: Der Median teilt die der Größe nach geordneten Ausprägungen eines Merkmals in zwei gleich große Teile.

Arithmetische Mittel (mean)

Definition:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Quelle: Sibbertsen/Lehne, S.54

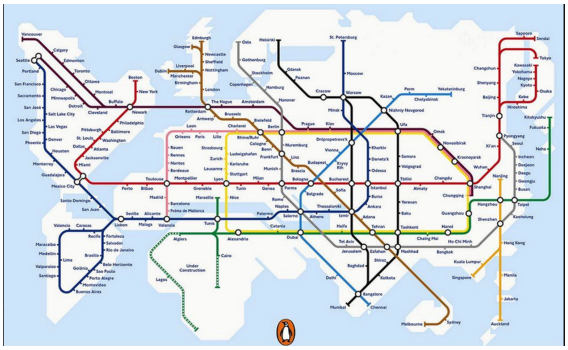
Statistische Modelle

Wenn wir unsere Daten nehmen, um die Welt zu beschreiben, machen wir das niemals so gut, dass wir die Welt perfekt beschreiben. Der Hauptgrund, warum unsere Daten dafür nicht ausreichen ist, dass wir fast nie alle Beobachtungen, die es in der Grundgesamtheit gibt, in unserer Stichprobe haben. Und, selbst wenn wir alle haben, es gibt immer sehr viele Faktoren, welche das Phänomene, das wir studieren, beschreiben. Es ist sehr unwahrscheinlich in der Sozialwissenschaft, dass wir Daten zu all diesen Faktoren haben. Was machen wir nun?

Wir bilden ein **statistisches Modell**, das versucht so nah wie möglich an die "Realität" kommt, damit wir etwas lernen können. Bevor wir auf diese statistischen Modell eingehen und uns ansehen, wie wir sie bilden, braucht es eine kurze Exkursion mit der Frage, was ist ein Modell.

Modelle

Was zum Teufel sind Models? Modelle sind Werkzeuge, sie können nicht getestet werden, sie sind nicht wahr oder falsch, sie sind nur nützlich (oder nicht nützlich), um uns zu helfen, ein Phänomen zu verstehen, zu erklären und dann empirische Daten zu analysieren, die wir über das Phänomen gesammelt haben. Wir sollten über Modelle wie **Karten** nachdenken. Seht euch diese beiden Karten unten an:



Keine dieser Karten ist ein genaues Abbild der Realität. Sicher nicht. Beides sind Abstraktionen, aber nur

eine davon ist nützlich, wenn ihr mit der Erstellung der Karte beabsichtigt haben, den Menschen zu helfen, sich in den öffentlichen Verkehrsmitteln in London zurechtzufinden. Das ist eindeutig die erste Karte, sie ist eine **Darstellung** des echten Londoner U-Bahn-Systems. Das zweite ist ein Kunstwerk. Wir könnten sagen, das zweite ist auch eine Darstellung, möglicherweise von Reiserouten zwischen Städten der Welt. Wir wissen jedoch über die Welt, dass dies keine gute Darstellung solcher Reiserouten ist.

Ein Modell, wie eine Karte, kann nicht von der Realität getrennt werden kann. Es muss nahe genug an der Realität sein, um nützlich zu sein. Ein Modell muss der Realität **ähnlich** genug sein, um nützlich zu sein. Die erste Karte wäre ebenso wie die zweite nicht hilfreich, um die geografische Entfernung zwischen Stationen genau einzuschätzen, aber nur die erste hilft euch dabe herauszufinden, wie man von Mile End im Osten der Karte zur Waterloo Station kommt. Wenn ihr euch jedoch die zweite Karte anseht, kommt ihr wahrscheinlich in der Realität nicht mit dem Zug von London nach Izmir, um dann nach Caracas weiterzufahren, wie auf der Karte angegeben. Die zweite Karte ist der Realität nicht ähnlich genug, um nützlich zu sein. Bei der ersten, nützlicheren Karte sollten wir beachten, dass diese nicht mit Informationen überfüllt ist. Es gibt keine Straßen in der Karte, keine Erhebungen, keine Gebäude. Zur Orientierung gibt es die Themse, aber das war es auch schon mit Orientierungspunkten.

Merkt euch einfach folgendes:

1. Modelle sind weder wahr noch falsch.
2. Modelle haben eine begrenzte Genauigkeit.
3. Modelle beziehen sich auf Teilaspekte.
4. Modelle sind zweckbezogen.

Hier ist ein Beispiel für ein weit verbreitetes Modell in der Gesetzgebungsforschung.

Example
Baron and Ferejohn (1989) Das Divide-the-Dollar-Modell modelliert Verhandlungen, bei denen ein Vorschlagender, der zufällig aus allen Gesetzgebern ausgewählt wird, einen Vorschlag zur Aufteilung eines festen Betrags öffentlicher Ausgaben unterbreitet. In seiner einfachsten Form wird, wenn das Angebot abgelehnt wird, ein neuer Anbieter ausgewählt, und dieser Prozess wird fortgesetzt, bis eine Einigung erzielt wird. Das Modell bietet Einblicke in die Rolle des Agenda-Setters und über die Regeln, wie eine Gesetzesvorhaben im parlamentarischen Prozess verändert werden kann. Das Modell konzentriert sich auf eine ganz bestimmte Art von politischem Kampf und sagt wenig über andere wichtige Phänomene in der Gesetzgebungspolitik aus.
Clarke/Primo S.61/62

Beachtet, was dieses Modell hier tut. Es legt eine Definition fest, wie wir über ein Phänomen denken sollten, trifft bestimmte Annahmen darüber, wie sich beispielsweise Akteure verhalten, fügt einschränkende Bedingungen hinzu und macht dann Vorhersagen, die aus dem Modell als theoretische Hypothese hervorgehen, die empirisch überprüft werden könnte. Eine Theorie besteht aus potentiell mehreren Modellen.

Grundstruktur eines statistischen Modells

Fahren wir fort mit statistischen Modellen und wie man diese bildet, um etwas über die Welt zu lernen. Lasst uns das statistische Modell erst einmal abstrakt formulieren. Unser Modell lautet: **Ausprägung in der Grundgesamtheit = Modell + Fehler.** Oder,

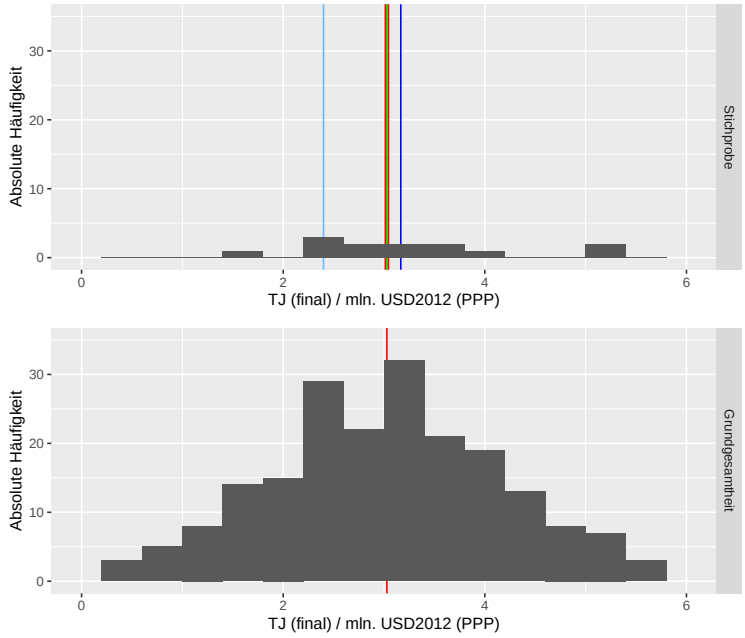
$$\text{outcome}_i = model + error_i$$

Die Frage ist nun, was ist das richtige Modell, welches unsere Daten am besten beschreibt und sich der Ausprägung in der Grundgesamtheit am besten annähert? Als Beispiel, fangen wir mal mit einem wirklich guten Modell der mittleren Lage an! Kennt ihr alle. Genau, der Mittelwert. Sagen wir, wir wollen schätzen,

welchen Wert die Ausprägungen einer Variable, für die wir uns interessieren, so mehrheitlich annehmen. Wir haben eine Stichprobe erhoben und die beste Schätzung, finden wir gerade, welchen Wert die Ausprägung für die Beobachtung i annimmt, ist der Mittelwert aller Ausprägungen in unserer Stichprobe.

$$\text{outcome}_i = \text{Sample mean} + \text{error}_i$$

Klingt vernünftig, allerdings werden wir bei fast jeder Beobachtung mit dieser Schätzung einen kleineren oder größeren Fehler machen. Warum? Die **Grundgesamtheit** $\neq$ **Stichprobe**. Die Abbildung unten illustriert das. Die obere Abbildung zeigt die Verteilung einer Variable in der Grundgesamtheit, die untere Abbildung die Verteilung in der Stichprobe.



Wir sehen, dass die Verteilung in der Grundgesamtheit etwas weiter ist. Das liegt daran, dass wenn wir eine Stichprobe ziehen, wenig extreme Werte werden einfach öfters gezogen als mehr extreme Werte. Das hat eine geringere Standardabweichung der Stichprobe in dieser Variable zur Folge. Wir sehen aber auch, dass die Lagemaße, welche die Grundgesamtheit für diese Variable beschreiben, nicht mit denen der Stichprobe übereinstimmen. Das könnte ein Problem sein, wenn wir von der Stichprobe auf die Grundgesamtheit schließen wollen.

Noch einmal, dieses Problem ist in der nun schon bekannten Gleichung ausgedrückt:

$$outcome_i = model + error_i$$

Wir machen für jede Beobachtung i einen Fehler in unserem Modell, wenn wir den Wert der Variable, die wir erklären wollen, aus einer Stichprobe schätzen.

Wenden wir das mal auf das Beispiel der Emissionen an. Wir wollen die Emissionen eines bestimmten Landes i schätzen und sagen, der Mittelwert der Emissionen aller Länder, könnte eine ganz gute Schätzung sein.

$$emissions_i = \overline{emissions} + error_i$$

Oder eine abstraktere Darstellung ist oft eine wo die Schätzung als allgemeiner Parameter, hier b_0 angegenben wird.

$$emissions_i = b_0 + error_i$$

Mit solch einer Schreibweise, können wir auch leicht komplexere Modelle ausdrücken, zum Beispiel ein Modell der Emissionen, welche von mehreren Faktoren abhängen könnten. Unsere Schätzung bezieht sich dann auf zwei Parameter.

$$emissions_i = b_0 + b_1 GDP_i + error_i$$

Übrigens, bei den Fehlern, $error_i$ sprechen wir auch von **Abweichungen**. Wir können die Fehler als Abweichung von dem Schätzziel, den Werten der Variable *emissions* ausdrücken:

$$error_i = emissions_i - \overline{emissions}$$

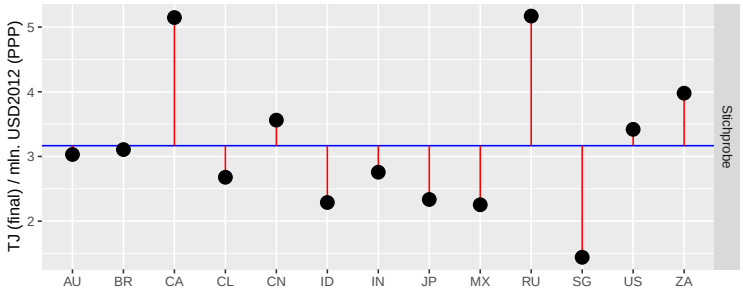
oder abstrakter,

$$error_i = emissions_i - b_0$$

oder für das komplexere Modell,

$$error_i = emissions_i - b_0 - b_1 GDP_i$$

Wenn wir nun die unsere Stichprobe nehmen und sagen, ist denn der Mittelwert ein guter Schätzer, dann weichen unsere eigentlichen Beobachtungen immer ein bisschen ab. Das ist in der Abbildung unten illustriert:



Ist jetzt der Mittelwert ein guter Schätzer, um unsere Daten in dieser Stichprobe zu beschreiben? Man beachte, wir fragen uns gerade noch nicht, ob wir mit den Daten unserer Stichprobe und diesem Schätzer etwas über die Grundgesamtheit sagen können. Also, erstmal die Frage, beschreibt unser Schätzer, der Mittelwert in diesem Beispiel, die Daten in der Stichprobe gut? Was wir wollen ist das die Abweichungen vom Schätzer über alle Datenpunkte hinweg so gering wie möglich ist. Wir berechnen also zunächst die Summe dieser Abweichungen.

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} error_i$$

oder für unser Beispiel in den Emissionsdaten

$$\text{Summe der Abweichungen} = \sum_{i=1}^{N} (emissions_i - \overline{emissions})$$

Es ist uns auch egal ob wir nun in die eine, positive Richtung, einen Fehler machen oder in die andere, negative Richtung. Daher schauen wir uns die quadrierte Summe der Abweichungen an, da fallen die Vorzeichen der Abweichung weg.

$$\text{Summe der quadrierten Abweichungen} = \sum_{i=1}^{N}(emissions_i - \overline{emissions})^2$$

Diese **Summe der quadrierten Abweichung** ist die Kenngröße schlechthin, um Schätzer zu beurteilen. Merkts euch. Ok, aber was ist jetzt ein gutes Modell, gegeben dieser Kenngröße? Wir müssen Modelle, unsere Schätzer, vergleichen können. Dafür können wir uns die Mittlere quadrierte Abweichung (Mean squared error, MSE) ansehen. Diese Kenngröße erlaubt uns Schätzer, die auf verschieden Stichprobengrößen beruhen zu vergleichen. Umso kleiner der MSE, umso besser der Schätzer.

$$\text{Mittlere quadrierte Abweichung} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{N}$$

Bei einer größeren Stichprobe, wird der MSE natürlich kleiner. Umso mehr Information wir über die Welt haben, umso besser unsere Schätzung.

Übrigens, wenn unser Modell, unser Schätzer der Mittelwert ist, dann ist dies die **Varianz** der Variable *emissions*.

Statistiken für Grundgesamtheit und Stichprobe

Im Beispiel oben, in der Diskussion darüber was ein statistisches Model ist, habe ich ein Lagemaß, den Mittelwert, als Schätzer der Werte eines Merkmals eingeführt. Wir können natürlich uns auch Streuungsmaße in der Grundgesamtheit ansehen und uns fragen, wie wir diese doch am besten Schätzen könnten.

Varianz und Standardabweichung

Die Streuungsmaße, denen ihr ständig begegnen werde sind **Varianz** und **Standardabweichung** mit folgenden Definitionen:

Definition
Varianz der Variable X

$$\sigma^2 = var(X) = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}$$

Standardabweichung (standard deviation) der Variable X

$$\sigma = \sqrt{\sigma^2} = \text{sd(X)}$$

Mit der Standardabweichung kommen wir zur ursprünglichen Einheit der Variable zurück.

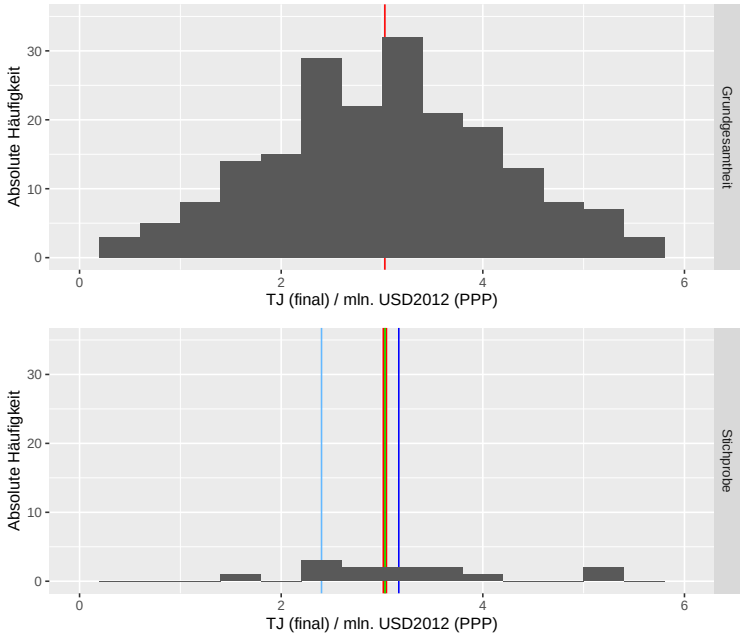
Die beiden Streuungsmaße, Varianz σ^2 und Standardabweichung σ, sind Statistiken, welche die Grundgesamtheit beschreiben. Wie sich zeigen wird, sind die folgenden Statisiken, s^2 und s gute Schätzer, um mit den Daten aus einer Stichprobe, nahe an die Varianz und Standardabweichung in der Grundgesamtheit heranzukommen. Hier sind die Schätzer s^2 und s für σ^2 und σ.

$$s^2 = var(emissions)^{\text{Stichprobe}} = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions})^2}{\mathbf{N-1}}$$

$$s = \sqrt{s} = \text{Standardabweichung}^{\text{Stichprobe}}$$

s, s^2 sind unverfälschte (unbiased) Schätzer für σ und σ^2 wenn wir eine Zufallsstichprobe gezogen haben

Zur Illustration, wie würden verschieden große Standardabweichungen aussehen? Well, die Standardabweichung in der Grundgesamtheit ist größer als die in der Stichprobe.



Weiter Streuungsmaße: Spannweite

Varianz und Standardabweichung sind natürlich nicht die einzigen Streuungsmaße, welche wir benutzen, um die Verteilung in der Grundgesamtheit und Stichprobe zu beschreiben und, als nächsten Schritt, die Statistik in der Grundgesamtheit mit Hilfe eines Modells und Daten aus der Stichprobe zu schätzen.

So, zurück zu den weiteren Streuungsmaßen. First up, **Spannweite** oder

$$max(emissions_i) - min(emissions_i)$$

Sagen wir, wird ordnen die Werte der Variable *emissions* der Größe nach und nennen den kleinsten Wert $emissions_1$ und den größten Wert $emissions_N$ für N Werte, dann kriegen wir Maximum und Minimum-Wert der variable *emissions*. Wir könnten das dann auch so schreiben:

$$emissions_N - emissions_1$$

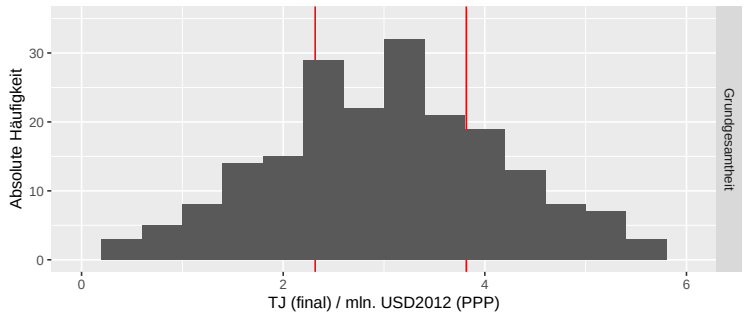
Next, **Interquartilsabstand**. Was ist das? Wir gehen die folgenden Schritte, zur Berechnung des Interquartilsabstand:

- Nach Ordnung der Werte der Variable von kleinstem zu größten Wert bestimmen wir **Quartile**:
  - **0.25 Quartile**: der Wert der Variable so dass 25% der Werte kleiner und 75% der Werte größer sind: $emissions_i^{0.25}$
  - **0.75 Quartile**: der Wert der Variable so dass 75% der Werte kleiner und 25% der Werte größer sind: $emissions_i^{0.75}$

Somit ist der **Interquartilsabstand**:

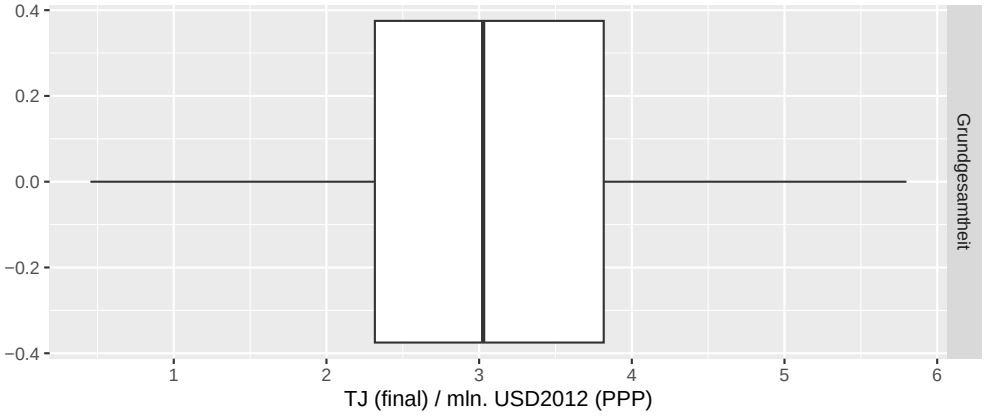
$$emissions_i^{0,75} - emissions_i^{0,25}$$

Und so sieht Interquartilsabstand graphisch aus:



Wir brauchen den Interquartilsabstand, für eine sehr hilfreiche graphische Darstellung der Verteilung einer Variable, dem **Boxplot**:

Der, oder vielleicht auch das, Boxplot, zeigt uns die Verteilung einer Variable, ausgedrückt in Median und Interquartilsabstand. Wir finden den Boxplot gut, weil diese beiden Statistiken nicht leicht durch extreme Werte zu beeinflussen.



Wie visualisiert man welche Statistiken?

Zum Abschluss dieser Sitzung, möchte ich euch in Erinnerung rufen, dass wir Daten, wie etwa die Verteilung einer Variable, auf viele Arten und Weisen visualiserien können. Unterschiedliche Visualiserugen sind hilfreich für verschiedene Variablen oder Lernzielen, aber wir haben große Auswahl. Unter diesem Link findet ihr eine ganze Reihe an graphischen Darstellungen verschiedener Daten, implenentiert in R → Viele, viele Graphen

Was solltet ihr aus dieser Sitzung mitnehmen?

1. Verstehen, wie die in einer Stichprobe erhobenen Ausprägungen einer Variable eine Häufigkeitsverteilung ergeben.
2. Wissen, welche wichtigen Lagemaße existieren und für welche Arten von Merkmale diese aussagekräftig sind.
3. Erklären können, wie das ziehen eine Stichprobe zur empirischen Häufigkeitsverteilung führt und wann wir aus einer solchen Verteilung lernen können.
4. Wissen warum wir von einem statistischen Modell sprechen, wenn wir Stichprobenstatistiken berechnen (schätzen).
5. Ideen haben, wie man Daten am besten visuell beschreibt.

Sitzung 4: Aus dem Zusammenhang: Deskriptive Statistik 3

Literatur
Fields, Kapitel 4 und 5 + Kapitel 13.3.1, 13.3.2 und 14.1
de Mesquita/Fowler, Kapitel 2
Mittag/Schüller, Kapitel 10
Sibbertsen/Lehne, Kapitel 5 und 6

Modelle statistischer Zusammenhänge

In der vorherigen Sitzung, haben wir uns statistische Modelle als Konzept generell angesehen und Modelle der Lage und der Streuung im Speziellen. Wir können die Lage und Streuung einer Variable in unseren Daten mit Statistiken wie Mittelwert oder Varianz beschreiben. Und wir können einen Schätzung mit Hilfe dieser Statistiken abgeben, um etwa von den Daten der Stichprobe etwas über die Grundgesamtheit zu lernen oder etwa von den Daten die wir haben, unseren Beobachtungen, etwas über Daten zu sagen, die wir nicht haben (etwa in der Grundgesamtheit oder über zukünftige Beobachtungen).

In any case, bisher haben wir uns mehrheitlich eine Variable angesehen, und wie wir diese am besten beschreiben können, oftmals sind wir aber in Zusammenhänge zwischen Variablen interessiert. Was sind also Maße, Statistiken, die einen Zusammenhang beschreiben? Wir beginnen mit dem einfachsten Fall, wir studieren den Zusammenhang zwischen zwei Variablen.

Kovarianz

Das erste Maß, das wir uns ansehen, ist die **Kovarianz.** Ihr habt schon die Varianz einer Variable kennengelernt; die Varianz sagt uns etwas darüber, wie weit sind den Abweichungen der einzelnen Werte einer Variable von Mittelwert der Variable. Die Kovarianz nimmt dieses Konzept und fragt, wie weit gehen den die Abweichungen der Werte der einen Variable vom Mittelwert der Variable einher mit den Abweichungen der Werte der anderen Variable vom Mittelwert der anderen Variable. Also ein ähnliches Konzept.

Die Kovarianz von *emissions* und GDP ergibt sich damit folgender Maßen:

$$cov(emissions, GDP) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions_i})(GDP_i - \overline{GDP_i})}{N}$$

Nur nebenbei, die Kovarianz einer Variable mit sich selbst, ist die Varianz, nämlich so:

$$cov(emissions, emissions) = \frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions_i})(emissions_i - \overline{emissions_i})}{N} =$$

$$\frac{\sum_{i=1}^{N}(emissions_i - \overline{emissions_i})^2}{N}$$

Wenn wir eine Schätzung basierend auf den Daten der Stichprobe abgeben wöllten, dann bitte N im Nenner mit $N-1$ ersetzen.

Was lernen wir aus der Kovarianz? Wenn die Kovarianz groß ist, dann variiert die eine Variable, hier *emissions*, stark um den Mittelwert der Variable genau dann und genau in die Richtung, wie die andere Variable um deren Mittelwert variiert.

Pearson's Korrelationskoeffizent

Die Kovarianz ist super, aber beeinflusst von der Skala der einbezogenen Variablen. Damit ist die Kovarianz nicht vergleichbar über Variablen hinweg. Da hilft uns der Koerrelationskoeffizent. Der bekannteste Korrelationskoeffizent ist **Pearson's Korrelationskoeffizent** gegeben als:

$$\rho_{emissions,GDP} = \frac{cov(emissions, GDP)}{s_{emissions} s_{GDP}}$$

Was lernen wir aus dieser Statistik. Wenn ρ nahe an -1 oder 1 ist, dann ist die standardisierte Kovarianz zwischen zwei Variablen groß und Bewegungen in der einen Variable, gehen mit Bewegungen in der anderen Variable einher. Merkt euch auch, ρ ist immer eine Statistik zwischen -1 und 1 für die Korrelation zwischen jeglichen Variablen.

Einfaches lineares Regressionsmodell

Korelationskoeffizienten sind gut, aber was ist wenn wir denken die Welt ist komplex und der Zusammenhang zwischen zwei Variablen hängt eventuell von anderen Variablen ab? Ein Korelationskoeffizienten ist immer nur ein Maß des bedingungslosen Zusammenhangs zwischen zwei Variablen. Da kommen wir zum Arbeitstier der statischen Analyse in der Sozialwissenschaft, dem **linearen Regressionsmodell**. Warum **linear**? Weil wir ein statistisches Modell der Welt bilden, wo die eine Variable, eine lineare Funktion der Parameter der anderen Variablen ist. Wir schreiben das für unsere Emissions-Beispiel so:

$$emissions_i = b_0 + b_1 GDP_i + u_i$$

Der Zusammenhang zwischen *emissions* und GDP ist modelliert als die Summe der Parameter b_0 und b_1. Linear impliziert, dass eine Änderung um eine Einheit in GDP (die Einheit in der GDP notiert ist, könnte 1 Euro, 1000 Euro, etc. sein) immer eine Änderung in *emissions* nach sich zieht. Was bedeuten die einzelnen Buchstaben hier?

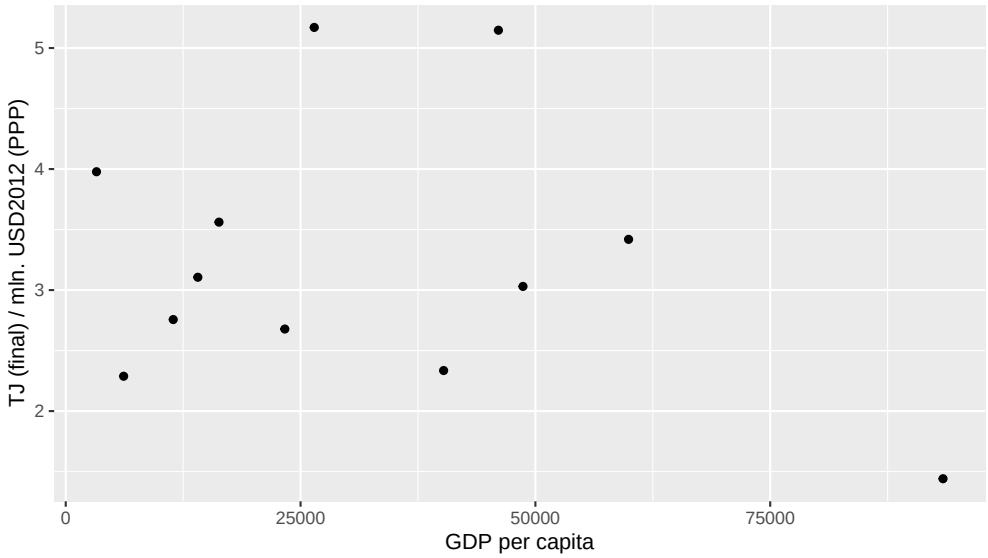
- b_0 sagt uns wo unsere Modell sagt die Werte von *emissions* liegen sollten, im Verhältnis zu den tatsächlichen Werten von *emissions*.
- b_1 gibt uns die Form, wie *emissions* und GDP zu einander stehen.
- b_0 und b_1 bezeichnen wir als Regressionskoeffizient.

Definition: Das lineare Modell

$$y_i = b_0 + b_1 x_i + u_i$$

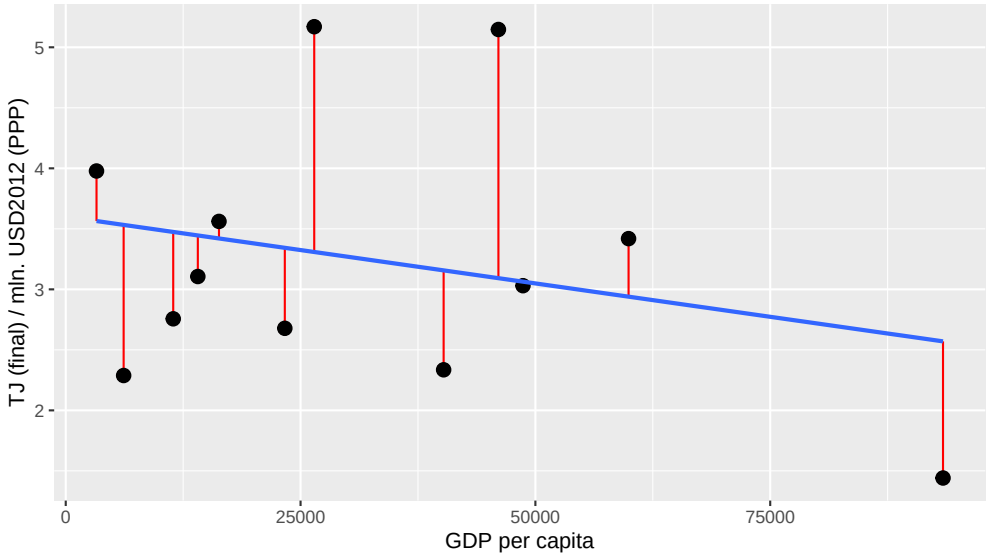
heißt einfaches lineares Regressionsmodell. Dabei bezeichnet die Variable X die unabhängige Variable, auch Regressor oder erklärende Variable genannt. Die abhängige Variable Y heißt Regressand, erklärte oder zu erklärende Variable. Die Fehler u_i beschreiben die möglichen Abweichungen der Gerade von den Beobachtungen, da bis auf wenige Ausnahmen die Beobachtungen nicht auf der Geraden liegen werden. Quelle: Sibbertsen/Lehne, S.137

Schauen wir uns mal ein Beispiel an. Unten seht ihr eine Graphik, welche die Werte von *emissions* für jeden Wert von GDP in den Ländern im Datensatz anführt.



Nebenbei, die Graphik ist ein **Streudiagramm**. Ok, was ist der Zusammenhang hier? Erstmal, die Kovarianz ist -7701.71. Also ein negativer Zusammenhang. Das sagt auch der Korrelationkoeffizient, der liegt bei -0.26.

Was wäre nun unsere Schätzung der Parameter im linear Regressionsmodell, der Regressionskoeffizienten, die am besten diese Daten und den Zusammenhang beschreiben. Die beste Schätzung ist durch eine Linie gegeben, wie unten abgebildet.



Warum ist es gerade diese Linie, welche den Zusammenhang am besten beschreibt? Es ist die Linie durch die Datenwolke, welche die Summe der quadrierten Fehler minimiert. Das ist mal wieder die **Methode der kleinsten Quadrate**.

Für unser lineares Modell

$$emissions_i = b_0 + b_1 GDP_i + \epsilon_i$$

ist die Steigung dieser Linie (unsere beste Schätzung von b_1) immer gegeben als:

$$\hat{b}_1 = \frac{Cov(emissions, GDP)}{Var(emissions)}$$

Und der beste Schätzer for den Achsenabschnitt b_0 ist

$$\hat{b}_0 = \hat{y} - \hat{b}_1 \overline{emissions}$$

Wo kommt diese beste Schätzung her?

Die Antwort? Wir wenden die **Methode der kleinsten Quadrate** auf unser Regressionsmodel an. Hier ist das einfache Regressionsmodel von oben mit einer erklärenden Variable.

$$emissions_i = b_0 + b_1 GDP_i + \epsilon_i$$

Was wir jetzt brauchen, ist dass die Summe der quadrierten Fehler in unserer Schätzung so klein wie möglich ist. Dafür müssen wir folgenen Term, die Summer der quadrierten Fehler, so klein wie möglich machen.

$$\sum_{i=1}^{N} \epsilon_i^2$$

Aber was ist ϵ_i. Erinnert euch, wir können den Fehler den wir mit unserem Model in jeder Beobachtung machen (aka die Abweichung unserer Schätzung vom tatsächlichen Wert der Variable für diese Beobachtung), auch anders schreiben:

$$\sum_{i=1}^{N} (emissions_i - b_0 - b_1 GDP_i)^2$$

Aha, da sind also diese Regressionskoeffizienten und unsere Variablen drin. Wir wissen die Werte für diese Variablen, die sind fixiert in unserem Datensatz. Die Daten für *emissions* und GPP sind gegeben durch unsere Stichprobe. Was wir dann machen müssen, ist die Regressionskoeffizienten so zu wählen, dass die Gleichung unten so klein wie möglich wird. Das ist ein klassisches Optimierungsproblem und Ableitungen einer Funktion kommen zur Anwendung.

Dann optimieren wir doch mal diese Funktion unten, d.h., lass uns die Werte von b_0 und b_1 finden, welche die Funktion minimieren.

$$\sum_{i=1}^{N} (emissions_i - b_0 - b_1 GDP_i)^2$$

Machen wir mal b_1:

Erst die Funktion ableiten nach b_1 und gleich 0 setzen. Die erste Ableitung ist 0, wenn wir einen Extremwert erreicht haben. Um die Schreibweise etwas zu vereinfachen, nennen wir jetzt mal die erklärende Variable GDP x und die zu erklärende Variable *emissions* y.

$$\frac{\partial \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2}{\partial b_1} = 0$$

Die erste Ableitung dieses quadrierten Terms oben nach b_1 wird dann irgendwie so aussehen (errinert euch an Ableitungsregeln, zum Beispiel ist die erste Ableitung der Funktion x^2, $2x$, etc.). Damit wollen wir uns aber nicht im Detail beschäftigen.

$$\sum_{i=1}^{n} y_i^2 - 2b_0 y_i - \ldots = 0$$

Wichtig ist nur, dass wir am Ende mit einer ersten Ableitung enden und die gleich 0 gesetzt haben, um nach b_1 lösen zu können.

$$\sum_{i=1}^{n} -2(y_i - b_0 - b_1 x_i)x_i = 0$$

Wenn wir nach b_1 lösen, kommt dies hier heraus:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Voila, wir haben hier die Kovarianz der erklärenden Variable x und der zu erklärenden Variable y geteilt durch die Varianz von x. Das ist der dieser beste Schätzer, der die Linie durch die Datenwolke beschreibt, welche die Summe der quadrierten Fehler minimiert.

Wir beschäftigen uns etwas später mit Regressionsmodellen mit mehr als nur einer erklärenden Variable und unter welchen Annahmen, wir wirklich diese beste Schätzung kriegen. Für den Moment, merkt euch, der Regressionskoeffizient einer unabhängigen Variable in so einem Model, gibt uns die Steigung des Zusammenhangs mit der abhängigen Variable. Mit anderen Worten, der Regressionskoeffizient sagt uns, um wieviel ändert sich die abhängige Variable, wenn sich die unabhängige Variable um eine Einheit ändert.

Was solltet ihr von dieser Sitzung mitnehmen?

1. Verstehen, dass wir ein Modell brauchen, um mit den Daten aus unserer Stichprobe Statistiken der Grundgesamtheit zu schätzen
2. Wissen was grundlegene Modelle sind, welche ganz gute Annäherungen der Grundgesamtheit sind (solange wir eine Zufallsstichprobe gezogen und genug Beobachtungen zu haben)
3. Kenntniss grundlegender Streuungsmaße
4. Wissen, wie man einen (linearen) Zusammenhang gut schätzen kann

---

Sitzung 5: Da bin ich mir unsicher: Wahrscheinlichkeit

Literatur
Englisch: Fields, Kapitel 7
Mittag/Schüller, Kapitel 11
Sibbertsen/Lehne, Kapitel 8

Unsicherheit spielt in der Gesellschaft und Politik eine große Rolle. Menschen ziehen ständig Schlussfolgerungen aus den verfügbaren und oft nicht vollständigen Information über die Welt. Statistik ist nichts anderes als das Studium solcher Schlussfolgerungen. Wir brauchen eine präzise Sprache, um über diese Schlussfolgerungen, diese **Inferenzen** unter Unsicherheit zu sprechen, und alles beginnt mit der Wahrscheinlichkeitstheorie. Wie kommt die Wahrscheinlichkeit in die statistische Analyse? Wir sind solchen Einschätzungen über Unsicherheit ausgedrückt als Wahrscheinlichkeit schon begegnet. In den Regressionstabellen, die ich zu Beginn des Semsters gezeigt habe. Hier ist ein weiteres Beispiel. Seht euch die Tabelle unten an, die aus einem Datensatz mit verschiedene Wirtschaftsindikatoren über 20 Jahre für EU-Mitglieder und europäische Nicht-EU-Länder stammt. Hier sehen wir das Ergebnis einer linearen Regression des Pro-Kopf-BIP ($GDPPerCapita$ auf einen

Indikator, ob ein Land der EU bis 2004 beigetreten ist oder nicht (Die Referenzkategorie hier ist "nicht beigetreten").

```
read.csv('../data/indicators.csv') %>%
  lm(GDPPerCapita~euJoin2004,data=.) %>%
  summary()
```

```
##
## Call:
## lm(formula = GDPPerCapita ~ euJoin2004, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8960.1 -2741.9  -474.1  2538.8 16372.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               11128.9      335.3   33.19   <2e-16 ***
## euJoin2004No EU Member  -7607.7      443.6  -17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4694 on 455 degrees of freedom
##   (56 observations deleted due to missingness)
## Multiple R-squared:  0.3926, Adjusted R-squared:  0.3912
## F-statistic: 294.1 on 1 and 455 DF,  p-value: < 2.2e-16
```

An mehreren Stellen kommt Wahrscheinlichkeit vor: in den Hypothesentests über Statistiken wie dem Regressionskoeffizienten oder der F-Statistik, die uns etwas über allgemeinen Modelfit sagt. Wahrscheinlichkeit kommt auch in den Standardfehlern vor, da diese auf der Grundlage der Wahrscheinlichkeitsverteilung der Zufallsvariablen berechnet werden. Wenn euch das alles gerade fremd vorkommt, keine Sorgen, ihr werdet am Ende des Semesters fähig sein genauestens alle Statistiken in der Tabele zu definieren und zu verstehen, was diese bedeuten. Ihr sollt nur gerade mitnehmen, dass ohne die dazugehörigen Wahrscheinlichkeitsaussagen, könnte man in dieser Tabelle nicht viel über die Daten erfahren. Dann bauen wir mal eine Sprache auf, um über Wahrscheinlichkeit zu sprechen.

Wahrscheinlichkeitsmaß

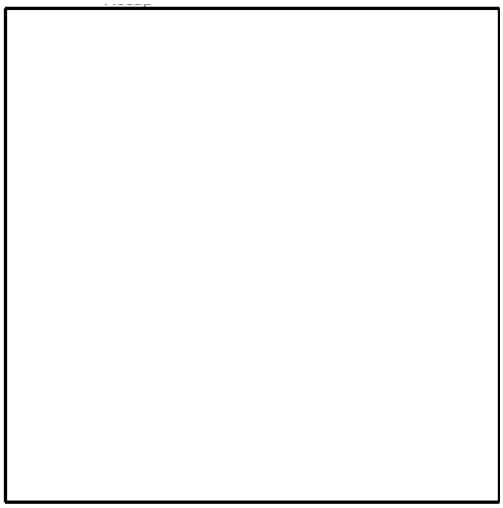
Definitionen

Zunächst einige Definitionen aus der Mengenlehre, um die Umgebung zu beschreiben, in der wir Daten generieren:

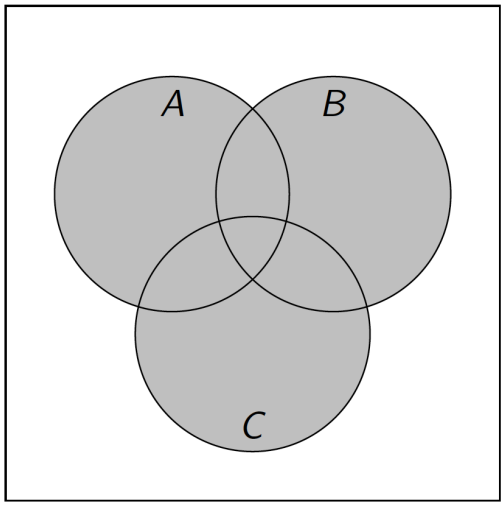
- **Mengen**: Begrenzte Sammlungen, die durch ihren Inhalt definiert sind: Menge an Bundesländern {Tirol, Voralberg, Salzburg, ...}
- **Elemente**: die in einer menge enthalten sind
- **Experiment**: spezifische Momentaufnahme der Welt, die viele Male wiederholt werden kann
- **Ergebnis**: Alles, was in einem bestimmten Experiment passieren kann
- **Ergebnisraum eines bestimmten Experiments**: Menge aller möglichen Ergebnisse eines Experiments
- **Ereignis**: jede Sammlung möglicher Ergebnisse eines Experiments
- **Ereignisraum**: jede sich gegenseitig ausschließende, kollektiv erschöpfende Sammlung von Ereignissen eines Experiments
- **Zusammengesetzte Ereignisse**: bestehen aus zwei oder mehr einfachen Ereignissen – entweder **unabhängig** oder **bedingt** voneinander.

Diese Konzepte können wir auch graphisch, in Form eines Venn-Diagramms darstellen. Zunächst veran-

schaulichen wir den **Stichprobenraum** oder **universelle Menge** eines Experiments, also die Menge aller möglichen Ergebnisse eines Experiments, durch eine abgegrenzte weiße Fläche.



In diesem Raum finden wir drei Ereignisse A, B und C. Ereignisse sind eine beliebige Sammlung möglicher Ergebnisse eines Experiments. Wir zeichnen sie als Kreise, die eine Reihe individueller Ergebnisse beinhalten.



Mal ein Beispiel, und ja es müssen Würfel und Münzen sein: Sagen wir, wir würfeln einmal mit einem 6-seitigen Würfel, das ist unser Experiment, das ist die Aktion, die uns Beobachtungen liefert. Was ist die Menge aller möglichen Ergebnisse eines solchen Experiments? Wir werfen den 6-seitigen Würfel nur einmal, der Würfel hat 6 Seiten, also haben wir 6 mögliche Ergebnisse

$$1, 2, 3, 4, 5, 6$$

Was wären nun mögliche Ereignisse bei einem solchen Experiment? Ein Ereignis könnte “eine 1 würfeln” oder “eine 3 würfeln” sein. Es könnte auch “eine Zahl größer als 4 würfeln” sein. Das letztgenannte Beispiel zeigt, dass Ergebnisse und Ereignisse unterschiedliche Konzepte sind. Dieses letztere Ereignis würde die Ergebnisse 5 und 6 kombinieren. Hier ist ein weiteres, noch allgemeineres Beispiel, nur um den Punkt zu verdeutlichen. Angenommen, unser Experiment besteht darin, eine Münze dreimal zu werfen. Was ist die Ergebnismenge? Wir haben eine Münze, die zwei Seiten hat, “Kopf” und “Zahl”, und wir werfen sie dreimal, damit sie bei

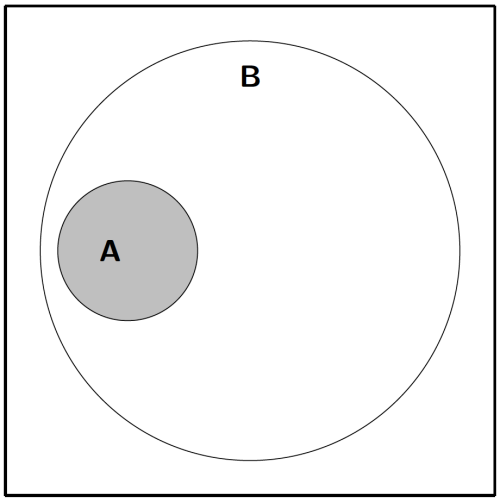
jedem dieser Würfe als Kopf oder Zahl herauskommt, was uns $3^2 = 8$ mögliche Ergebnisse gibt

$$TTT, TTH, THT, THH, HTT, HTH, HHT, HHH.$$

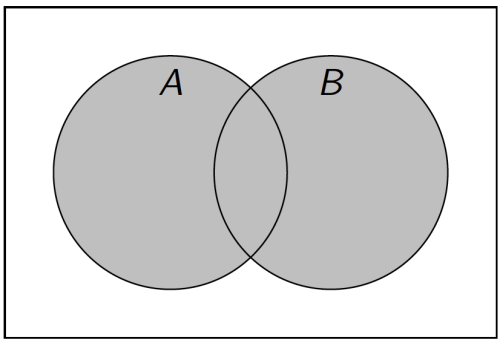
Ein Ereignis eines solchen Experiments könnte sein "Wirf drei mal Zahl" oder "Wirf beim ersten Wurf ein Zahl und beim dritten Wurf einen Kopf."

Und hier sind Operationen, die man mit Mengen durchführen kann: - **Gegenereignis:**: $A' = \{X : X \notin A\}$ - **Teilmenge**: A ist eine Teilmenge von B wenn jedes Element von A auch in B ist: $A \subset B \Leftrightarrow \forall X\, X \in A,\, X \in B$ - **Gleichheit**: $A = B \Leftrightarrow A \subset B,\, B \subset A$ - **Vereinigung von A, B**: $A \cup B = \{X : X \in A \,\text{or}\, X \in B\}$ $A_1 \cup A_2 \cup \ldots A_n = \bigcup_{i \leq n} A_i$ - **Durchschnitt (oder Schnittmenge) von A,B**: $A \cap B = \{X : X \in A \,\text{and}\, X \in B\}$ $A_1 \cap A_2 \cap \ldots A_n = \bigcap_{i \leq n} A_i$ - **Leere Menge**: enthält keine Elemente, $\emptyset$

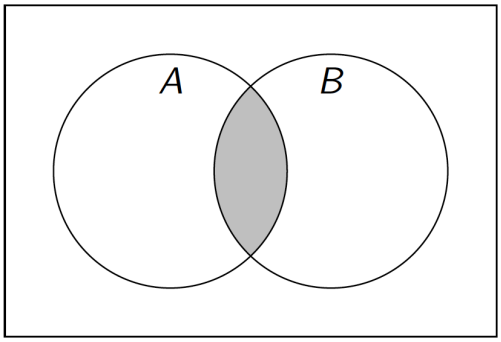
Mit Hilfe von Venn-Diagrammen, können wir diese Operationen auch graphisch darstellen. Hier gibts nichts zu sehen, außer A als Teilmenge von B.



Dann die Vereinigung von A und B, die wir auch als $A \cup B$ oder $A + B$ schreiben können,



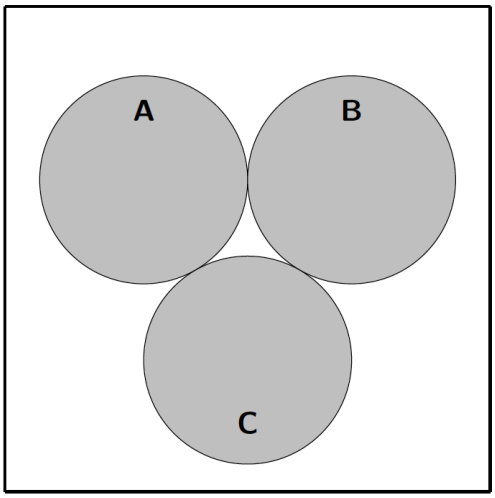
und der Durchschnitt von A und B. Das kann man so schreiben: $A \cap B$ oder AB



Who cares? Nun, Vereinigung und Schnittmenge etc erlauben uns weitere Konzepte zu definieren, die bereits in der obigen Definition des Ereignisraums aufgetaucht sind: **gegenseitig ausschließend** und **kollektiv erschöpfend**:

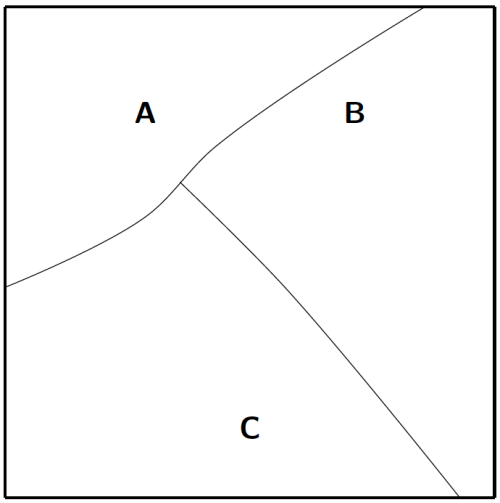
- k Mengen schließen sich gegenseitig aus, wenn ihre Schnittmenge leer ist.
- k-Mengen sind kollektiv erschöpfend, wenn ihre Vereinigung gleich dem Stichprobenraum ist.

In einem Venn-Diagramm sehen Mengen, die sich gegenseitig ausschließend folgendermaßen aus.



In dieser Darstellung sind A, B un C gegenseitig ausschließend. k Mengen $A_1, A_2, \ldots, A_k$ sind gegenseitig ausschließend wenn und nur wenn, $A_1 \cap A_j = \emptyset \; \forall i \neq j$.

Mengen, die kollektiv erschöpfend sind, sehen wie folgt aus:



k Mengen $A_1, A_2, \ldots, A_k$ sind kollektiv erschöpfend wenn und nur wenn, $\bigcup_{k=1}^{K} A_k = S$. Übrigens, A, B und C in der obigen Darstellung sind auch gegenseitig ausschließend. Wie würde das Ganze aussehen, wenn A, B, und C kollektiv erschöpfend, aber nicht gegenseitig ausschließend sind? Siehe unten:

! (../images/collectivelyExhaustiveNotMutuallyExclusive.png){width=40%}

Zusammengefasst, ein Ergebnis ist alles, was in einem Experiment passieren kann. Ein Ereignis ist jede Sammlung möglicher Ergebnisse eines Experiments, der Ereignisraum ist jede sich gegenseitig ausschließende, kollektiv erschöpfende Sammlung von Ereignissen eines Experiments, und der Stichprobenraum ist die feinskörnigste, sich gegenseitig ausschließende, kollektiv erschöpfende Menge aller möglichen Ergebnisse eines Experiments. Was heißt feinkörnigst? Die Menge kann nicht weiter unterteilt werden.

Dann lass uns mal diese Definitionen verwenden.

Example
Stellt euch folgendes Experiment vor: Wir fragen 2000 Personen nach ihren Rauch- und Trinkgewohnheiten und ihrem Alter. Hier ist eine Reihe von Ereignissen, die in diesem Experiment auftauchen: 612 sind Raucher, 960 Trinker, 670 sind älter als 25, 86 trinken und rauchen, 290 trinken und sind älter als 25, 158 rauchen und sind älter als 25, 44 trinken, rauchen, und sind älter als 25, 248 trinken nicht, rauchen nicht und sind jünger als 25. Stellen wir diese Ereignisse in einem Venn-Diagramm dar und definieren Ereignis A: “Rauchen”, Ereignis B: “Trinken”, Ereignis C: “Älter als 25”. Jetzt können wir alle Schnittmengen von Ereignissen und den Teil jedes Ereignisses, der sich nicht mit anderen Ereignissen überschneidet, mit der Anzahl der Personen füllen, für die das Ereignis gilt (d.h. wie viele Personen rauchen, trinken aber nicht, wie viele Menschen trinken, rauchen und sind älter als 25 usw.):

Wahrscheinlichkeitsmaß

Wo kommt nun die Wahrscheinlichkeit ins Spiel? Wie fügen sich die obigen Konzepte und Wahrscheinlichkeit zusammen. Um Wahrscheinlichkeitsaussagen machen zu können, müssen wir jedem Punkt im Ereignisraum eine Wahrscheinlichkeit zuordnen (Zur Vollständigkeit, wir betrachten hier eine diskrete Welt mit einer diskreten Anzahl von Ereignissen. Da können wir mit Summen arbeiten und brauchen keine Integrale).

Erstmal sehr formell: Betrachten wir ein Experiment mit dem Stichprobenraum S, eine Funktion mit reellen Werten $\mathbb{R}$ für S nennt man **Wahrscheinlichkeitsmaß** wenn es die folgenden Merkmale hat:

1. nicht negative #: für jedes Ereignis A, $p(A) \geq 0$

2. Wahrscheinlichkeit von S ist 1: $p(S) = 1$
3. Für zwei Ergebnisse, die nicht zur gleichen Zeit stattfinden können, gilt: $AB = \emptyset$, $P(A+B) = p(A)+p(B)$

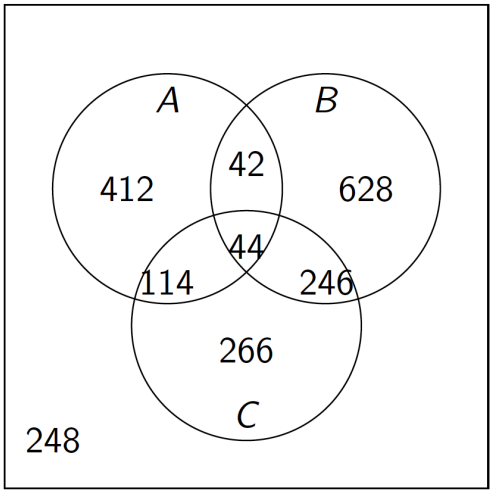
Mal etwas einfacher und grob gesagt:

$$\text{Prob(Ereignis)} = \frac{\text{Anzahl an Möglichkeiten, wie das Ereignis eintreten kann}}{\text{Gesamtanzahl der möglichen Ergebnisse}}$$

Dies ist ein numerisches Maß für die Wahrscheinlichkeit eines Ereignisses und eine Abbildung von definierten Ereignissen auf eine durch 0 und 1 begrenzte Metrik. Man beachte, dies ein theoretisches Konstrukt, wir können uns die Wahrscheinlichkeit auch als empirisches Konstrukt herleiten:

$$\text{Prob(Event)} = \frac{\text{Anzahl der Fälle, eintritt}}{\text{Anzahl der Fälle, in denen ein Ergebnis eintritt}}$$

Kehren wir zum Beispiel mit Münzwurf und Würfel zurück, um zu veranschaulichen, wie das Wahrscheinlichkeitsmaß berechnet wird.



Beispiel
Wir werfen eine Münze einmal, wie groß ist die Wahrscheinlichkeit, dass sie "Kopf" zeigt? Die Gesamtzahl der möglichen Ergebnisse hier ist zwei, entweder "Kopf" oder "Zahl", und die Anzahl der Möglichkeiten, wie das Ereignis "Kopf" auftreten könnte, ist eins. Die Wahrscheinlichkeit für das Ereignis "Kopf" ist also $p(E) = \frac{1}{2}$.
OK, das war einfach, hier noch ein Beispiel: Werfen Sie eine Münze dreimal, wie groß ist die Wahrscheinlichkeit, dass sie "Kopf beim 2. Wurf" zeigt? Die Anzahl der möglichen Ergebnisse beträgt $2^3 = 8$ mögliche Ergebnisse: TTT, TTH, THT, THH, HTT, HTH, HHT, HHH. Das Ereignis "Kopf beim 2. Wurf" kann auf vier verschiedene Arten auftreten (THT, THH, HHT, HHH). Die Wahrscheinlichkeit für das Ereignis "Kopf beim 2. Wurf" ist also $p(E) = 4/8 = 1/2$.

Wenn wir von einer **Verbundwahrscheinlichkeit** sprechen, meinen wir die Wahrscheinlichkeit von **verbundenen Ereignissen**. Verbundene Ereignisse sind entweder **unabhängig**, das Eintreten eines Ereignisses hat keinen Einfluss auf die Wahrscheinlichkeit, dass das andere Ereignis eintritt, oder **bedingt** zueinander.

Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit gibt die Wahrscheinlichkeit eines Ereignisses an, wenn ein anderes Ereignis eingetreten ist.

Betrachten wir nochmal Würfe mit zwei sechsseitigen Würfeln und definieren das Ereignis A: "$R_1 + R_2 < 7$" und Ereignis B: "$R_1 = 1$". Was sind die Wahrscheinlichkeiten das die Ereignisse unabhängig von einander eintreten? Ein Bild hilft:

! (../images/marginalProbability.png){width=40%}

In der obigen Abbildung seht ihr den Ergebnisraum, also alle möglichen Ergebnisse des Experiments "Wirf zwei Würfel". Wir haben einen schwarzen Punkt für jedes mögliche Ergebnis: Würfel 1 (R1) würfelt 1 und Würfel 2 (R2) würfelt 1, R1=1 und R2=2, R1=1 und R2=3, etc. Um an die Wahrscheinlichkeit von Ereignis A, "$R_1 + R_2 < 7$", zu kommen, müssen wir einfach alle Ergebnisse zusammenzählen, bei denen die Augen auf den beiden Würfeln eine Zahl kleiner als 7 ergeben (rotes Vieleck): Es sind 15 verschiedene Ergebnisse. Wie viele mögliche Ergebnisse gibt es insgesamt? $6x6 = 36$. Das bedeutet, dass die Wahrscheinlichkeit für Ereignis A $P(A) = \frac{15}{36} = .42$ ist. Das selbe Verfahren gibt uns die Wahrscheinlichkeit für Ereignis B, "$R_1 = 1$"; es ist $P(B) = \frac{6}{36} = \frac{1}{6}$.

Dann, etwas interessanter, was ist die bedingte Wahrscheinlichkeit von B gegeben A? Was ist $P(B|A)$? Schaut euch sich die obige Abbildung noch einmal an und zählt. Aber was sollt ihr zählen? Wir wollen wissen, wie wahrscheinlich ist B angesichts der Tatsache, dass A auch geschehen ist. Um diese bedingte Wahrscheinlichkeit zu bestimmen, müssen wir zuerst wissen, auf wie viele Arten A passieren könnte. Es gibt 15 Möglichkeiten, wie A passieren könnte. Es gibt 15 Ergebnisse, für die A wahr ist. Dies ist jetzt unser neue Berechnungsgrundlage, unser neuer Nenner, um das Wahrscheinlichkeitsmaß zu berechnen. Um zu bestimmen, wie viele Möglichkeiten es gibt, dass B passieren könnte, wenn A passiert ist, müssen wir die Ergebnisse zählen, bei denen A und B gleichzeitig passieren, das sind die Ergebnisse, die in den roten **und** blauen Vielecken eingeschlossen sind. Diese Anzahl von Ergebnissen ist 5. Daher ist $P(B|A) = 5/15 = 1/3$. Ereignis A wird zur neuen Gesamtanzahl der möglichen Ergebnisse, die wir als Nenner verwenden.

Unabhängigkeit

Zwei Ereignisse sind **unabhängig**, wenn Information über ein Ereignis nichts über das andere Ergebnis aussagt. Manchmal ist dies trivial zu sehen: Werfen wir zwei Münzen und definieren Ereignis A: "erster Kopf" und Ereignis B: "zweite Zahl". Wenn die Münze nicht beim ersten Wurf beschädigt wird, ist der zweite Wurf sicherlich unabhängig vom ersten Wurf, sodass A und B unabhängige Ereignisse sind. Hier ist die formale Definition von **Unabhängigkeit**:

- A und B sind **unabhängig** genau dann, wenn $P(A|B) = P(A)$
- Verallgemeinerung auf N Ereignisse: N Ereignisse $A_1, \ldots, A_N$ sind voneinander unabhängig genau dann, wenn $P(A_i|A_j, A_k, \ldots, A_p) = P(A_i) \, \forall \, i \neq j, k, \ldots, p$ mit $1 \geq i, j, k, \ldots, p \geq N$

Und, daraus folgend: Die Ereignisse A und B sind **bedingt unabhängig** genau dann, wenn $P(AB|C) = P(A|C)P(B|C)$.

Die bedingte Unabhängigkeit ist wichtig in der Statistik. Hier ein Beispiele: Wir gehen davon aus, dass der Fehlerterm (ϵ) im Regressionsmodell bedingt unabhängig von den Regressoren X auf der rechten Seite (auch bekannt als unabhängige Variablen) ist.

Die meisten interessanten Ereignisse in der Politikwissenschaft sind nicht einfach – wir stehen oft Verbundsereignissen gegenüber, die sich oft nicht gegenseitig ausschließen oder von anderen abhängig sind. Ein Beispiel hierfür wäre die Wahlentscheidung. Vergesst mal alles was ihr gehöhrt habt, warum man so wählen geht. Es ist nämlich das Wetter. Drücken wir die Beziehung zwischen Wetter und Entscheidung Wählen zu gehen (aka "turnout") nur zum Spaß als bedingte Wahrscheinlichkeit in Form einer Regressionsfunktion aus: $prob(turnout > .5|weather) = prob(y > .5|\beta, x) = prob((\beta x + \ epsilon)|\beta, x)$. In anderen Worten, wir

brauchen einfach Unabhängigkeit und/oder bedingte Wahrscheinlichkeit, um die Wahrscheinlichkeit von Verbundsereignissen zu bestimmen.

Hier noch ein paar nützliche Formeln:

$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$ $P(B|A) = \frac{P(A \cap B)}{P(A)}$ mit $P(A)$ ungleich Null $P(A|B) = \frac{P(A \cap B)}{P(B)}$ mit $P(B)$ ungleich Null $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bayes-Regel

Forschen heißt immer wieder Beobachtungen sammeln. Wir gehen auch oft von einer Theorie oder Empirie über das Phänomen aus, an dem wir interessiert sind, wir beginnen meistens nicht mit einer leeren Seite. Nun, wie sollte man als Reaktion auf neue Beweise seine Meinung über die Welt ändern? In Bezug auf die obige abstrakte Einführung in die Wahrscheinlichkeit, wie sollte man die Wahrscheinlichkeit ändern, dass ein Ereignis A eintritt, wenn neue Beweise vorliegen, sagen wir Ereignis B ist passiert?

Wir beginnen mit dem **Satz der totalen Wahrscheinlichkeit** aus, der besagt, dass für jedes Ereignis A und B gilt:

$$p(A) = p(A \cap B) + p(A \cap B')$$

Wir könnten diese Formulierung leicht erweitern, um eine Welt zu erfassen, die nicht nur in einen Zustand unterteilt ist, in dem B auftritt, und in den Zustand, in dem B nicht auftritt (geschrieben B'), sondern eine Welt, die in viele Teile von B $B_1, B_2, \ldots B_n$ unterteilt ist. Wie auch immer, wir bleiben vorerst bei B und B' dann leitet sich folgendes aus dem Satz der totalen Wahrscheinlichkeit ab: A kann als Summe bedingter Wahrscheinlichkeiten ausgedrückt werden:

$$p(A) = p(A|B)p(B) + p(A|B')p(B')$$

Was bedeutet der Satz der totalen Wahrscheinlichkeit im Klartext? Die Wahrscheinlichkeit, dass ein Ereignis A eintritt, ist gleich der Wahrscheinlichkeit (1), dass das Ereignis A und das Ereignis B gleichzeitig eintreten ($P(A \cap B)$) und (2) dass das Ereignis A eintritt während das Ereignis B tritt nicht ein ($p(A \cap B')$). Wir werden dies gleich brauchen, wenn wir den Satz von Bayes herleiten. Das ganze hier noch einmal bezogen auf das Beispiel Wetter und Wahlen.

Beispiel Ereignis A: Es regnet am Wahltag Ereignis B: Beliebiger Wähler i geht wählen. Wir wollen wissen, ob die Wahlbeteiligung vom Regen am Wahltag oder von $P(B|A)$ abhängig ist. Wie berechnet man $P(B|A)$? Wir wissen, $P(B|A)P(A) = P(A|B)P(B)$, weil beide gleich $P(A \cap B)$ sind, somit

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Warum ist das hilfreich? Erstmal, in einer empirischen Studie, können wir das Wetter beobachten und daher die Wahrscheinlichkeit berechnen können, dass es $P(A)$ regnet. Wir können auch die Wahlbeteiligung beobachten und $P(B)$ erreichen. Und, das ist der Clou, wir können beobachten, ob es regnet, wenn Wählerin i ihre Stimme abgibt, damit können wir $P(A|B)$ berechnen. Ausgestattet mit diesen empirisch abgeleiteten Größen können wir $P(B|A)$ berechnen, die Wahrscheinlichkeit, dass i in Abhängigkeit davon, ob es geregnet hat, zur Wahl gegangen ist. Was wir hier neu gelernt haben, ist es, dass es geregnet hat (Ereignis A ist passiert). Vielleicht sind wir aber zu diesem Zeitpunkt bereits zu einem Schluss gekommen, vielleicht basierend auf früheren Studien, wie wahrscheinlich es sein würde, dass i bei Regen wählen geht. Jetzt machen wir weitere Beobachtungen zu Regen und Wahlbeteiligung. Wie sollten wir unsere Schlussfolgerung ändern, ob i bei Regen wählen geht? Hier kommt Bayes ins Spiel.

Unsere Schlussfolgerung, unsere Überzeugung bevor neue Beobachtungen hereinkamen, wird **vorherige** oder frühere Überzeugung genannt (prior belief), und sobald wir neue Beweise berücksichtigt haben, kommen wir zu einer **nachträglichen** oder späteren Überzeugung (posterior belief). Was ist unser posterior belief, was ist das endgültige $P(B|A)$, mit dem wir aus unserer Forschung herauskommen? Bayes Ansatz abstrakt: *aktualisiere deine früheren Überzeugungen über das Eintreten von B, indem du $P(A|B)$ mit der gemeinsamen Wahrscheinlichkeit vergleichst, das A eintritt, vorausgesetzt, dass man in Bezug auf B richtig liegt* **und** *vorausgesetzt, man liegt in Bezug auf B falsch.* Say what?!? Betrachten wir das Ganze mal in einer Formel:

Definitionen
Theorem von Bayes Für zwei (binäre) Ereignisse A und B

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B')P(B')}$$

Was ist das jetzt? Das Verhältnis der Verbundswahrscheinlichkeit, dass A und B eintreten, das heißt die Wahrscheinlichkeit, dass A und B gleichzeitig eintreten (dass es regnet und i abstimmt) ist $P(A \cap B) = P(A|B)P(B)$, und die Verbundswahrscheinlichkeit von A und B, $P(A \cap B)$ **plus** die Verbundswahrscheinlichkeit von A und **B'** . Letzteres, $P(A \cap B')$, ist die Verbundswahrscheinlichkeit, dass es regnet, aber i nicht zur Abstimmung geht. Wir teilen die Wahrscheinlichkeit, dass wir bereits richtig liegen (wie wahrscheinlich es ist, dass i bei Regen wählen geht), durch die Wahrscheinlichkeiten, die mit allen anderen möglichen Zuständen der Welt verbunden sind, d.h. die Wahrscheinlichkeit dass wir richtig liegen **und** die Wahrscheinlichkeit, dass wir nicht richtig liegen. Puhwww. Das braucht ein Beispiel.

Beispiel Angenommen, wir wollen wissen, ob die Beobachtung einer Rezession, R, uns etwas darüber aussagt, ob die Regierung über hohe Kompetenz HQ, durchschnittliche Kompetenz AQ oder inkompetente I verfügt. Sagen wir, wir wissen, dass das Auftreten einer Rezession negativ mit dem Kompetenzniveau der Regierung zusammenhängt (aus Theorie oder Empirie): - $P(R|HQ) = 0,1$ - $P(R|AQ) = 0,2$ - $P(R|I) = 0,5$
Diese Zahlen erscheinen vernünftig. Wenn eine Regierung sehr kompetent ist, ist die Wahrscheinlichkeit einer Rezession viel geringer als wenn sie inkompetent ist. Nehmen wir weiter an, dass wir einen *flat prior* darüber haben, welches Kompetenzniveau die Regierung beschreibt, wenn keine anderen Informationen vorhanden sind: $P(HQ) = P(AQ) = P(I) = 1/3$, das heißt, wir denken, dass jeder Niveau der Regierungskompetenz ist ebenso wahrscheinlich, bevor wir neue Empirie gesehen haben. Nun, was ist $P(HQ|R)$? Wenden wir Bayes an:

$$P(HQ|R) = \frac{P(R|HQ)P(HQ)}{P(R|HQ)P(HQ) + P(R|AQ)P(AQ) + P(R|I)P(I)}$$

Super ist, dass wir nicht einmal $P(R)$ kennen müssen, wir müssen nicht einmal wissen, wie wahrscheinlich eine Rezession ist! Setzen wir mal die Zahlen ein:

$$P(HQ|R) = \frac{1/10 \times 1/3}{1/10 \times 1/3 + 1/5 \times 1/3 + 1/2 \times 1/3} = 1/8$$

Die Wahrscheinlichkeit, dass der Politiker ein Idiot ist, gegeben, dass wir eine Rezession beobachtet haben ist 1/8.

Zufallsvariablen

Wir haben bisher über Beobachtungen gesprochen, wir nannten das Experimente, und die Ergebnisse, die in einer bestimmten Beobachtung realisiert werden können. Und wir haben darüber gesprochen, wie man ein Wahrscheinlichkeitsmaß aus Ergebnissen und Ereignissen und verschiedenen Arten von Wahrscheinlichkeiten erstellt und wie wir die Wahrscheinlichkeit bilden und aktualisieren, die wir einem bestimmten Ereignis zuordnen. Jetzt müssen wir all dies mit Variablen verknüpfen, genauer gesagt mit **Zufallsvariablen**, da alle Variablen, mit denen wir in unserer Analyse arbeiten, und alle Statistiken, die wir berechnen, im Grunde Zufallsvariablen sind. Nur um das deutlich zu machen, sie werden nicht Zufallsvariablen genannt, weil sie einfach eine zufällige Ansammlung von Variablen oder Werten sind, nein, sie werden Zufallsvariablen genannt, weil es möglich ist, dass jede Beobachtung, die wir für dieser Variablen machen, jeden möglichen Wert der Variable annehmen könnte. Sobald wir unsere Daten gesammelt haben, sobald wir die Stichprobe erhoben haben, sind die Werte für jede Beobachtung, für jede Variable natürlich festgelegt.

Ok, kommen wir von den allgemeinen Grundsätzen zu den tatsächlichen Zahlen. Eine Zufallsvariable ist eine **Funktion**, die jedem Punkt im Stichprobenraum einen Wert zuweist. Eieieiei, was heißt das? Schaut euch dieses Beispiel an, ja, wir betrachten wieder Münzwürfe:

Beispiel Das Experiment, eine Münze zu werfen, sagen wir 10 Mal, erzeugt Ereignisse, zum Beispiel die "Anzahl der Köpfe". Die "Anzahl der Köpfe" ist nichts anderes als eine Zufallsvariable, nennen wir sie X. X weist jedem Punkt im Ergebnisraum einen Wert zu. Was ist der Ergebnisraum dieses speziellen Experiments? Der Ergebnisraum, die Menge aller möglichen Ergebnisse, ist, dass wir entweder 0 Kopf, 1 Kopf, 2 Kopf, ..., 10 Kopf in 10 Würfen sehen. Wir hätten die Menge aller möglichen Ergebnisse auch mit Bezug auf "Zahl" hätten definieren können. Die Zufallsvariable X, die "Anzahl der Köpfe", könnte nun diesen 11 verschiedenen möglichen Ergebnissen die Werte 0 bis 10 zuweisen oder

$$X = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

Wenn wir das Experiment mit 10 Münzwürfen immer und immer wieder wiederholen würden, würden wir möglicherweise zu einer anderen Realisierung, nennen wir sie x_0, der Zufallsvariablen X führen. Jede Widerholung des Experiments erzeugt nur einen der möglichen Werte von X, wir erhalten entweder 0, 1, ..., 10 Kopf von 10 Münzwürfen in diesem Experiment. Wann immer wir über eine Zufallsvariable sprechen, müssen wir daher klarstellen, ob wir über die Zufallsvariable sprechen, die Funktion selbst, das heißt X, oder eine bestimmte Ziehung von Werten aus dieser Funktion, das heißt x_0.

Um es noch einmal zu wiederholen, eine Zufallsvariable, die normalerweise mit Großbuchstaben bezeichnet wird, z. B. X, ist eine Funktion, die den Ergebnisraum auf eine Teilmenge der reellen Zahlen abbildet. Zufallsvariablen gehen darüber aber auch noch hinaus. Wir haben auch eine **Wahrscheinlichkeitsverteilung** über den Bereich der zulässigen Werte der Zufallsvariablen definiert. Was ist das nun? Sehte euch noch einmal das Münzwurfbeispiel an.

Beispiel Welche Wahrscheinlichkeiten sind mit jeder Realisierung der Zufallsvariablen X verbunden, der Anzahl von "Kopf" bei 10 Münzwürfen? Erinnern wir uns an die Definition des Wahrscheinlichkeitsmaßes? Es ist die Anzahl der Möglichkeiten, wie ein bestimmtes Ereignis über die Gesamtzahl der möglichen Ergebnisse auftritt. Beginnen wir mit $X = 0$, das heißt, nach 10 Münzwürfen kommt kein Kopf. Auf wie viele Arten kann es passieren, dass keine Köpfe auftauchen? Nur auf eine Weise! Bei jedem der 10 Münzwürfe darf kein Kopf zu sehen sein. Ok, aber was ist die Gesamtzahl der möglichen Ergebnisse dieser 10 Münzwürfe? Für jeden Wurf gibt es 2 mögliche Ergebnisse, {Kopf, Zahl}, dann erzeugen zwei mögliche Ergebnisse in 10 Würfen $2^{10} = 1024$ mögliche Ergebnisse. Wir könnten Kopf beim ersten Münzwurf und dann 9 mal "Zahl" haben, wir könnten Kopf beim zweiten Münzwurf haben und ansonsten Zahl usw. Es gibt 1024 verschiedene Möglichkeiten, wie dieses Experiment ablaufen könnte. Zurück zur Wahrscheinlichkeit von keinem Kopf in 10 Münzwürfen. Es gibt nur einen Weg, bei dem kein Kopf in 10 Münzwürfen auftauchen kann, nämlich "Zahl" in allen Münzwürfen. Also

$$p(X = 0) = \frac{1}{1024} = 0.0009765625$$

Kommen wir nun zu $X = 1$. Wie viele Möglichkeiten gibt es, dass wir nach 10 Münzwürfen nur 1 mal Kopf haben? Es gibt 10 Wege. Wir könnten Kopf auf Münzwurf 1 oder Münzwurf 2 oder Münzwurf 3 usw. sehen. Damit wissen wir das

$$p(X = 1) = \frac{10}{1024} = 0.009765625$$

Wir könnten mit $X = 2$, $X = 3$ usw. fortfahren, aber es wird immer mühsamer, die Anzahl der Möglichkeiten zu berechnen, auf die 2, 3, 4 usw. Kopf in 10 Münzwürfen kommen könnte. Wir haben dafür eine Formel, der wir uns gleich zuwenden. Zunächst, hier ist die **Wahrscheinlichkeitsverteilung** der Zufallsvariablen X, die Anzahl der Köpfe bei 10 Würfen einer Münze:

X	0	1	2	3	4	5	6	7	8	9	10
P(X)	.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010

Wahrscheinlichkeitsverteilung von Zufallsvariablen

Bevor wir uns einen einfacheren Weg zur Berechnung der mit den Werten einer Zufallsvariablen verbundenen Wahrscheinlichkeit ansehen, ohne jedesmal einzeln berechnen zu müssen, wie viele Wege es gibt, wie ein bestimmter Wert der Zufallsvariablen auftreten kann, führe ich einige allgemeine Gedanken und Definitionen im Bezug auf Wahrscheinlichkeitsverteilungen an.

Erstens, was ist eine Wahrscheinlichkeitsverteilung? Eine Wahrscheinlichkeitsverteilung verwendet die Gesetze der Wahrscheinlichkeit, um Erwartungen darüber zu entwickeln, wie die Werte einer Zufallsvariablen – ein Konzept, eine Funktion, das durch eine Variable operationalisiert wird – verteilt sind. Wir werden hauptsächlich mit **empirischen Häufigkeitsverteilungen** und **theoretischen Wahrscheinlichkeitsverteilungen** arbeiten. Als es um Häufigkeitsverteilungen ging, haben wir uns schon empirische Häufigkeitsverteilungen angesehen. Wenn wir mit einer diskreten Variablen arbeiten, also einer Variablen, deren Werte endlich und zählbar sind, sprechen wir von einer **Wahrscheinlichkeitsfunktion** (Probabilitz mass function oder PMF). Ein PMF ordnet jedem Wert, der zufällig aus einer Population gezogen wird, Wahrscheinlichkeiten zu, oder anders ausgedrückt. Ein PMF ist die Funktion, die die erwartete relative Häufigkeitsverteilung jedes Werts einer Zufallsvariablen beschreibt.

Definitionen
Wahrscheinlichkeitsfunktion (PMF) Eine PMF ordnet jedem Ereignis im Ergebnisraum eine Wahrscheinlichkeit zu. Die PMF ordnet jedem Wert einer Zufallsvariablen eine Wahrscheinlichkeit zu. Wir bezeichnen die PMF der Zufallsvariable X als

$$P(X = x)$$

Es gelten die Wahrscheinlichkeitsaxiome.

Manchmal möchten wir die Verbundswahrscheinlichkeit von Werten zweier (oder mehrerer) Zufallsvariablen ausdrücken. Zum Beispiel die Wahrscheinlichkeit, dass 10 Würfe mit einer Münze 4 mal Kopf ergeben, wenn gleichzeitig 10 Würfe einer anderen Münze 0 mal Kopf ergeben. Formal würde das $P_{X,Y}(x, y)$ lauten. Das ist die Wahrscheinlichkeit, dass X, Anzahl der Köpfe auf einer Münze, und Y, Anzahl der Köpfe auf einer anderen Münze, in einem bestimmten Experiment die Werte x bzw. y annehmen. Um die Verbundswahrscheinlichkeit $P_{X,Y}(x, y)$ von den Wahrscheinlichkeiten zu unterscheiden, die jeder Zufallsvariablen separat zugeordnet sind – $P_X(x)$ und $P_Y(y)$ (oder $P(X)$ und $P(Y)$ als Abkürzung) – letzteres nennen wir **Randwahrscheinlichkeiten**. Manchmal sagen wir auch **einfache** Wahrscheinlichkeiten, aber ich mag diesen Begriff nicht.

Für die gemeinsame Wahrscheinlichkeit gelten die üblichen Wahrscheinlichkeitsaxiome:

1. $\sum_x \sum_y P_{X,Y}(x, y) = 1$
2. $\sum_x P_{X,Y}(x, y) = p_Y(y)$
3. $\sum_y P_{X,Y}(x, y) = p_X(x)$
4. $0 \leq P_{X,Y}(x, y) \leq 1$

Oder wir wollen von bedingten Realisierungen der Werte zweier (oder mehrerer) Zufallsvariablen sprechen. Dazu benötigen wir die **bedingte Wahrscheinlichkeitsverteilung**, die äquivalent zu den bedingten Wahrscheinlichkeiten von Ereignissen definiert ist.

$P_{X,Y}(x|y)$ ist die bedingte Wahrscheinlichkeit, dass X den Wert x annimmt **gegeben** oder **bedingt**, dass Y den Wert y annimmt. Formell,

$$P_{X,Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}$$

Beachtet auch, dass $P_{X,Y}(x, y) = P_X(x)P_{X,Y}(y|x) = P_Y(y)P_{X,Y}(x|y)$.

Wir arbeiten nicht nur mit diskreten Zufallsvariablen, sondern auch mit kontinuierlichen Zufallsvariablen. Können wir auch eine Wahrscheinlichkeitsverteilung über die Werte einer solchen kontinuierlichen Variablen definieren? Sicher. Das ist eine **Wahrscheinlichkeitsdichtefunktion**.

Definitionen
Wahrscheinlichkeitsdichtefunktion (Probability density function, PDF) Ein PDF ordnet jedem Ereignis im Stichprobenraum eine Wahrscheinlichkeit zu. Ein PDF ordnet jedem Wert einer Zufallsvariablen eine Wahrscheinlichkeit zu. Wir beziehen uns auf das PDF der Zufallsvariablen X als

$$f_X(x)$$

Es gelten die Wahrscheinlichkeitsaxiome.

Unabhängigkeit und bedingte Unabhängigkeit von Zufallsvariablen

Die realisierten Werte zweier Zufallsvariablen könnten nicht nur bedingt zusammenhängen, sondern auch unabhängig sein, oder sie könnten bedingt unabhängig sein.

Wir sagen, die Zufallsvariablen X und Y sind unabhängig genau dann, wenn

$$P_{X,Y}(y|x) = P_Y(y) \, \forall x, y$$

Die Zufallsvariablen X und Y sind bedingt unabhängig vom Ereignis A genau dann, wenn

$$P_{X,Y}(x, y|A) = P_{X|A}(x|A)P_{Y|A}(y|A) \, \forall x, y$$

X und Y sind bedingt unabhängig, wenn das Festhalten von Ereignis A (oder das Festhalten einer anderen Zufallsvariablen auf einem bestimmten Wert) impliziert, dass die gemeinsame Wahrscheinlichkeit von X und Y das Produkt der Randwahrscheinlichkeiten von X ist und Y.

Bedingte Unabhängigkeit spielt bei dem, was wir tun werden, eine große Rolle. Zum Beispiel, wir gehen davon aus, dass der Fehlerterm im Regressionsmodell bedingt ist unabhängig von X (hier wird bedingte Unabhängigkeit zu einer Annahme zum Identifizieren einer unverfälschten Schätzung des Regressionskoeffizienten).

Erwartung der Zufallsvariablen

Wir interessieren uns oft nicht nur für die Wahrscheinlichkeitsverteilung einer Zufallsvariablen, sondern für Statistiken, die eine solche Variable zusammenfassen. Statistiken, die etwas über die **zentrale Tendenz** oder Merkmale der **Verteilung** einer Variablen aussagen. Denken Sie an Mittelwert und Median oder Varianz und Standardabweichung.

Erst einmal ist es wichtig zu wissen, dass Funktionen von Zufallsvariablen selbst Zufallsvariablen sind. Das heißt, sie haben eine Verteilung, und obwohl wir normalerweise bestimmte Realisierungen von Werten dieser Zufallsvariablen beobachten, wären andere Realisierungen möglich gewesen (z. B. wenn wir eine neue Stichprobe von Beobachtungen der Werte einer Zufallsvariablen gezogen hätten). Bestimmt sind euch Funktionen von Zufallsvariablen wie das Quadrat der Werte einer Variablen $f(h) = h^2$ oder eine Exponentialfunktion $f(h) = v^r$ begegnet.

Die wichtigste Funktion einer Zufallsvariablen ist jedoch die **Erwartung einer Zufallsvariablen** oder **Erwartungswert einer Zufallsvariablen**:

$$E[X] = \sum_{i=1}^{N} x_i p_X(x_i) = \overline{X}$$

Was ist das? Wir mitteln hier über N Realisierungen einer Zufallsvariablen X, aber es ist ein bestimmter Durchschnitt. Es ist ein **wahrscheinlichkeitsgewichteter Durchschnitt**. Wir möchten einen solchen Durchschnitt nehmen, da einige Werte von X wahrscheinlicher auftreten, sodass wir bei der Mittelwertbildung auf jeden Fall ihre höhere Häufigkeit berücksichtigen müssen. Eine Zufallsvariable X könnte jede Funktion $f(x)$ sein, in der obigen Definition betrachten wir die einfache Funktion $f(x) = x$.

Der Erwartungswert ist der **Mittelwert** einer Verteilung von Werten der Zufallsvariablen, auch bekannt als **erstes Moment**.

Wir haben mehr Momente, um die Verteilung von Zufallsvariablen zu charakterisieren. Das **n-te zentrale Moment** um den Mittelwert ist

$$E[(x - \overline{x})^n].$$

Wenn wir $n = 0$ einsetzen, erhalten wir den **ersten zentralen Moment**

$$E[(x - \overline{x})^1] = 0.$$

Im Klartext: Die wahrscheinlichkeitsgewichtete Summe der Abweichungen jedes Werts der Zufallsvariablen vom Mittelwert ist 0.

Was ist dann der **zweite zentrale Moment** einer Verteilung einer Zufallsvariablen?

$$E[(x - \overline{x})^2] = E[x^2] - E[x]^2$$

Ja! Das ist die **Varianz.** Dies ist **das** Maß für die Streuung einer Funktion und wird normalerweise mit σ^2 bezeichnet.

Wir haben mehr zentrale Momente. Die **Schiefe** einer Verteilung ist das dritte zentrale Moment und die **Wölbung**, die Dicke der Ausläufer der Verteilung, ist das vierte zentrale Moment. Das sind nur die zentralen Momente, die bekannte Namen abbekommen haben, wir könnten Momente mit höheren n definieren.

Abschließend noch eine Zusammenfassung einiger hilfreicher Gleichungen in Bezug auf Erwartungen: $E(X + Y) = E(X) + E(Y)$ $E(aX + b) = aE(X) + b$
Wenn X und Y unabhängig sind, dann gilt: $Var(X + Y) = Var(X) + Var(Y)$ $Var(X) = E(X^2) - E(X)^2$

Wir würden, mit mehr Zeit in der Einführungsvorlesung, diese Gleichungen bei der Ableitung von Regressionsschätzern verwenden.

Binomialwahrscheinlichkeitsfunktion

Geiles zusammengesetztes Wort erstmal. Dann, wir beenden den Überblick über die grundlegende Wahrscheinlichkeitstheorie mit einem Blick auf eine weit verbreitete bestimmte Wahrscheinlichkeitsmassenfunktion, die **binomiale PMF**. Warum ist diese so weit verbreitet? Weil viele Phänomene, die wir betrachten, leicht durch eine Zufallsvariable beschrieben werden können, die einen binomialen PMF hat. Jede Zufallsvariable, die auf einem **binären Ergebnis** basiert, folgt einem binomialen PMF. Erinnert euch ans Münzwerfen. Hat ein binäres Ergebnis. Oder allgemeiner und relevanter, alle Phänomene, die zu einem binären Ergebnis führen: Erfolg/Misserfolg im Konflikt, Protest/kein Protest, Staatsstreich/kein Staatsstreich, Wahlsieg/-verlust. Wir nennen solche Phänomene mit binären Ergebnissen **Bernoulli-Experimente** oder **Bernoulli-Versuche**.

Hier ist die Grundidee: Ergebnis 1 tritt mit Wahrscheinlichkeit p und Ergebnis 0 mit Wahrscheinlichkeit $1 - p$ auf. Diese beiden Ergebnisse bilden eine Unterteilung des Ergebnisraums. Es ist kein anderes Ergebnis möglich. Das heißt, wir können die Wahrscheinlichkeit jedes Ergebnisses aufschreiben als

$$P(x) = p \quad \text{und} \quad P(x') = 1 - p.$$

Das nennt man **Bernoulli-PMF**. Wie kommen wir zur binomialen PMF? Wir führen mehrere Bernoulli-Experimente durch. Wir werfen eine Münze mehrmals, wir schauen uns mehrere Wahlen zwischen zwei Kandidaten an usw. Die Zufallsvariable ist die Summe der beobachteten binären Bernoulli-Ereignisse

$$Y = \sum_{i=1}^{n} X_i.$$

Angenommen, wir betrachten 3 Münzwürfe. Wie viele Möglichkeiten, um y-Ereignisse zu erhalten? Solche Ereignisse könnten $Y = 2$ Kopf oder $y = 3$ Kopf sein: Es gibt $\binom{3}{y}$ Wege. Was ist das Doppeldecker-Ding in Klammern? Das ist der **Binominalkoeffizient**. Er wird im Beispiel aus 3 Bernoulli-Versuchen berechnet

$$\binom{3}{y} = \frac{3!}{y!(3-y)!}$$

wobei die **Fakultät** 3! $3! = 3 \times 2 \times 1$ bedeutet. Angenommen, wir suchen danach, wie viele Möglichkeiten es gibt, bei 3 Versuchen mit $Y = 2$ zu enden. Es gibt

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3 \times 2 \times 1}{2 \times 1 \times 1} = 3$$

Möglichkeiten, in 3 Versuchen mit $Y = 2$ zu enden. Erinnert euch an das Beispiel des Münzwurfs. Wie viele Möglichkeiten gibt es, bei 3 Münzwürfen von 8 möglichen Ergebnissen 2 Kopf zu bekommen? Das wären 3, HHT, HTH und THH.

Was ist jetzt mit der Wahrscheinlichkeitsverteilung über Zufallsvariablen. In diesem Beispiel interessieren wir uns für $P(Y = 2)$. Wenn wir im Beispiel des Münzwurfs davon ausgehen, dass die Münze fair ist und die Würfe voneinander unabhängig sind, wissen wir, dass die Wahrscheinlichkeit, dass Kopf kommt, $p = 0,5$ beträgt und die Wahrscheinlichkeit, dass Zahl kommt, $1 - p = .5$ beträgt. Für die drei verschiedenen Arten, bei denen 2 mal Kopf, $Y = 2$, auftaucht, könnten wir die Wahrscheinlichkeit berechnen, mit der jedes mögliche Ergebnis auftritt:

P(Ergebnis)	Ergebnis
$p \times p \times (1 - p)$	HHT
$p \times (1 - p) \times p$	HTH
$(1 - p) \times p \times p$	THH

Wenn wir $P(Ergebnis)$ für einige der Ergebnisse leicht umstellen, erhalten wir die gleiche Wahrscheinlichkeit für alle drei Ergebnisse: $p \times p \times (1 - p)$. Wir könnten sogar noch weiter vereinfachen, indem wir die Exponentialschreibweise $p^2(1 - p)^1$ verwenden. Um $P(Y = 2)$ zu berechnen, müssen wir diese Wahrscheinlichkeit dreimal nehmen, weil es drei Möglichkeiten gibt, zu $Y = 2$ zu kommen. Wir wissen, dass es drei Möglichkeiten gibt, weil wir das mit dem Binomialkoeffizienten berechnet haben. Das gibt uns

$$P(Y = 2)\binom{3}{2}p^2(1 - p)^{3-2} = 3 \times 0,25 \times 0,5 = 0,375 = 3/8.$$

Haste gesehen, wie ich $(1 - p)^1$ durch $(1 - p)^{3-2}$ ersetzt habe? Weil wir so leicht vom Spezialfall $Y = 2$ auf $Y = y$ und von $N = 3$ auf $N = n$ schließen können.

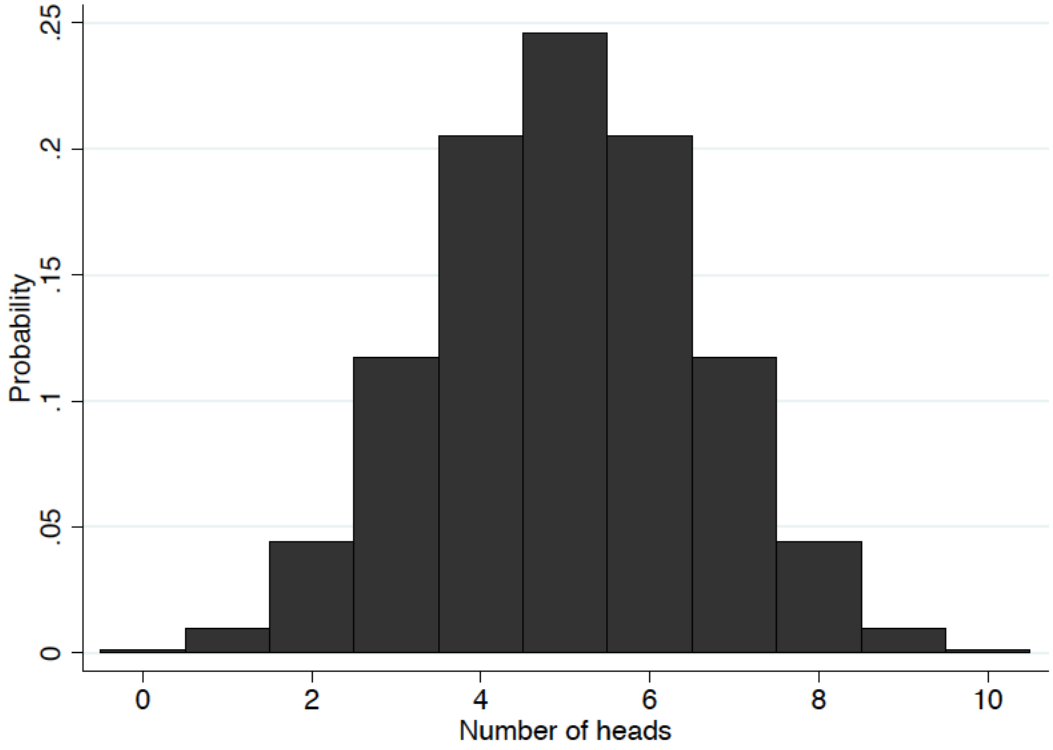
Hier ist die allgemeine **binomiale PMF**

$$p(Y = y|n, p) = \binom{n}{y}p^y(1 - p)^{n-y}.$$

Erinnert euch an die PMF von X: "die Anzahl der Köpfe in 10 Würfen einer fairen Münze"? Ich habe aufgehört zu illustrieren, wie man bei $P(X = 1)$ zum PMF kommt, weil es langweilig wird. Nun, das Einfügen in die Formel der binomialen PMF macht das alles so einfach (sicherlich würde jede Statistiksoftware die Berechnung auch für euch übernehmen, aber es von Hand zu tun, ist für wenige n auch nicht schwer). Wie auch immer, hier ist noch einmal der binomiale PMF dieses 10-Münzwurf-Beispiels:

X	0	1	2	3	4	5	6	7	8	9	10
P(X)	.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010

Grafisch würde das übrigens so aussehen



Was solltet ihr von dieser Sitzung mit nach Hause nehmen?

Ihr werdet noch einige anderem gängigen Wahrscheinlichkeitsmodelle, wie hier die binomiale PMF, kennen lernen. Ok, das ist das Ende der Überlegungen zur Wahrscheinlichkeitstheorie. Was mitnehmen? Wahrscheinlichkeit ist der Weg, um Ungewissheit in unserer Datenanalyse zu formalisieren. Wir benötigen ein Wahrscheinlichkeitsmodell, um die Ergebnisse von Zufallsereignissen (d. h. die realisierten Werte von Zufallsvariablen) zu beschreiben und müssen eine Wahrscheinlichkeitsverteilung über die Realisierungen der Zufallsvariablen definieren. All diese Dinge brauchen wir, um statistische Inferenz zu betreiben.

Sitzung 6: Gut getestet: Inferenzstatistik I

Literatur
Englisch: Fields, Kapitel 8 und 10
Mittag/Schüller, Kapitel 15 und 16
Sibbertsen/Lehne, Kapitel 11 und 14

Erstmal zur Erinnerung, die deskriptive Statistik ist enorm wichtig. Wir müssen unsere Daten kennen, bevor wir mit aussagekräftige statistische Inferenz beginnen können. Sich in sinnvoller Weise die Daten zusammenzufassen und gute grafische Darstellungen zu erstellen braucht etwas Übung. Deskriptive Statistiken und Grafiken könnnen uns da auch in die Irre führen. Gut, jetzt zurück zur Wahrscheinlichkeit, Zufallsvariablen, etc.

Grundlagen der statistischen Inferenz

Nehmen wir an, wir gute Daten gesammelt (d.h. wir haben valide und robuste Maße aus theoretisch fundierten Operationalisierungen unserer Variablen gebildet) und wir haben uns unsere Daten genau angesehen. Wir wissen wie die Werte unserer Variablen verteilt sind und was die Lagemaße und Streuungsmaße sind. Dann versuchen wir jetzt mal von den Daten, die wir haben, etwas über die Daten zu lernen, die wir nicht haben: **Inferenz**.

Wir brauchen wieder ein paar Definitionen (manche kennt ihr schon):

Definitionen

- Grundgesamtheit: Beobachtungen für jeden mögliche, relevante Einheit mit Bezug zu unserer Forschungsfrage.
- Stichprobe: eine Teilmenge von Einheiten, die aus der zugrunde liegenden Grundgesamtheit gezogen wurden.
- Statistische Inferenz: Der Prozess, bei dem wir Rückschlüsse aus der Stichprobe über die Grundgesamtheit ziehen.
- Punktschätzung: Schätzung einer einzelnen Statistik der Grundgesamtheit aus unserer Stichprobe.
- Intervallschätzung: Schätzung einer Reihe von Statistik der Grundgesamtheit aus unserer Stichprobe.
- Hypothesentest: Wahrscheinlichkeitsgestütze Aussagen über die Grundgesamtheit treffen.
- Stichprobenverteilung: hypothetische Verteilung der Stichprobenstatistik
- Standardfehler der Statistik: Standardabweichung der Stichprobenverteilung von der Stichprobenstatistiken
- Deskriptive Inferenz: Verwendung von vorhandenen Daten, um etwas über nicht vorhandene Daten zu lernen
- Kausale Inferenz: Lernen über die Ursachen des Beobachteten.

Nun, schaut euch die Abbildung unten mal an. Sie veranschaulicht, wie der Prozess der statistischen Inferenz funktioniert. Wir beschäftigen uns hier mit sogenannten **Frequenzstatistiken,** d.h., wir bauen unsere Schlussfolgerung auf der Häufigkeit von Statistiken in unserer Stichprobe auf, um Aussagen über den wahren Wert der Statistik in der Grundgesamtheit zu treffen. Etwas später im Semester werden wir diese Art der statistischen Inferenz der **Bayes'schen Statistik** gegenüberstellen (aber gaaaaanz kurz nur).

Ok, zurück zur Abbildung. Uns interessiert, wie die Welt tatsächlich aussieht, d.h., wir interessieren uns für die tatsächlichen Statistiken oder eine tatsächliche Verteilung von Werten unserer Variablen in der Grundgesamtheit. Sagen wir mal, es interessiert uns das arithmetische Mittel einer Variable in der Grundgesamheit. So ein Durchschnitt wir meist mit Hier interessiert uns der Durchschnitt, μ, einer Variablen in der Grundgesamtheit. Eine Grundgesamtheit für die wir uns beispielsweise interessieren könnten, wäre etwa die österreichische Wählerschaft, die Länder Afrikas südlich der Sahara oder die Monate Apirl 1990 bis April 1991. Mehrheitlich werden wir nicht alle Einheiten dieser Grundgesamtheit beobachten können, daher die Stichprobe ... sagte ich bereits häufig. In der Abbildung ist die Grundgesamtheit mit F bezeichnet.

Wenn wir statistische Schlussfolgerungen ziehen, stützen wir uns normalerweise auf eine Stichprobe, die wir aus dieser Grundgesamtheit gezogen haben. Mit hilfe eine Reihe von Annahmen, behaupten wir dann, wenn wir immer wieder Stichproben aus derselben Grundgesamtheit genommen hätten und wenn wir dieselbe Statistik für jede dieser Stichproben berechnet hätten (diese Stichproben sind blau dargestellt), generieren wir eine Verteilung von Statistiken. Das ist die Stichprobenverteilung. Hier ist die Statistik der Durchschnitt von $\overline{X}$, den wir in jeder Stichprobe berechnent. Die Verteilung von $\overline{X}$s, die Stichprobenverteilung, ist rechts dargestellt.

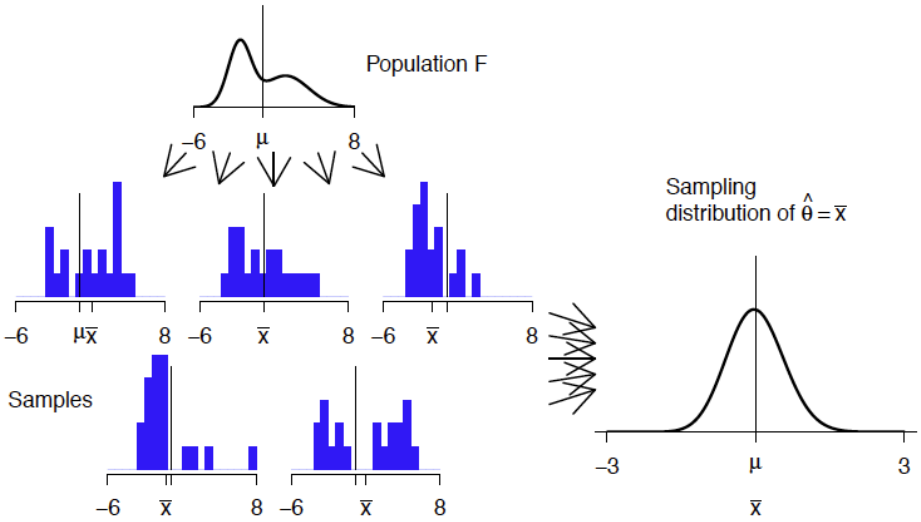


Figure 4: *Ideal world.* Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution.

Die Stichprobenverteilung hat eine zentrale Tendenz, die wir durch ein Lagemaß beschreiben können, wie wir das auch mit den zugrunde liegenden Werten einer Variable in der deskriptiven Statistik gemacht haben. Und die Stichprobenverteilung hat natürlich auch eine Streuung. Für statistische Inferenz würden wir nun den Mittelwert dieser Verteilung als unsere Punktschätzung des wahren Durchschnitts μ in der Grundgesamtheit nehmen.

Mit einigen Annahmen über die Verteilung von Daten in der Grundgesamtheit und die Merkmale der Statistik, an der wir interessiert sind, leiten wir diese Stichprobenverteilung ab. Nur nochmal zum mitschreiben: wir haben nie mehrere Stichproben, wir haben nur eine, wir sagen einfach: So sieht die Verteilung der Statistik, würden wir mehrere Stichproben aus derselben Grundgesamtheit). Wozu brauchen wir diese Verteilung, warum ist die eine Statistik, die wir für unsere Stichprobe berechnet haben, nicht genug? Wir werden diese abgeleitete Stichprobenverteilung verwenden, um einen Hypothesentest über die Punktschätzung durchzuführen.

Seht euch nun mal dieses Gadget unten an. Es handelt sich um eine Simulation, mit der ihr Stichproben aus verschiedenen Grundgesamtheiten ziehen könnt. Ich hatte euch das eher im Semester schon einmal gezeigt.

Um statistische Inferenz zu betreiben, die auf immer wieder neuen Stichproben aus derselben Grundgesamtheit aufbauen, muss unsere Stichprobe bestimmte Eigenschaften aufweisen. Es muss eine **Zufallsstichprobe** sein, d.h. sie muss **unabhängig und identisch verteilt** sein. What's that?

Definitionen
Eine Stichprobe von n Beobachtungen einer oder mehreren Variablen, $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ ist eine **Zufallsstichprobe**, wenn die n Beobachtungen **unabhängig** voneinander aus derselben Grundgesamtheit mit der Wahrscheinlichkeitsverteilung $F(\mathbf{Y}, \theta)$ gezogen werden und die Wahrscheinlichkeit das ein bestimmter Wert gezogen wird die gleiche ist (**identisch**).

Was ist θ? Ein unbekannter Parameter, eine Statistik, welche die Grundgesamtheit beschreibt. Ich sage hier nur θ als Platzhalter für jegliche Parameter oder Statistiken, die uns interessieren könnten. Was wäre so ein θ, das uns interessieren könnte? Für eine Bernoulli-verteilte Variable, so was haben wir uns letzte Woche angesehen, dass sind all Variablen die auf binären Ereignissen basieren, wäre das etwa die Erfolgswahrscheinlichkeit p. Für eine normalverteilte Variable, wäre eine interessante Statistik der Mittelwert oder die Standardabweichung sein.

So, nochmal ein paar Defintionen der statistischen Inferenz, jetzt bezogen auf diese Statistik, θ:

Definitionen
- Statistik: jede Funktion, die aus den Daten in einer Stichprobe berechnet wird – da diese Statistik eine Funktion einer oder mehrerer Zufallsvariablen ist, ist sie auch eine Zufallsvariable mit einer Wahrscheinlichkeitsverteilung.
- Stichprobenverteilung: Wahrscheinlichkeitsverteilung einer Statistik.
- Punktschätzung: Statistik, die einen einzelnen Wert für θ liefert.
- Intervallschätzung: Bereich von Werten, die θ enthalten, mit vorab zugewiesener Wahrscheinlichkeit.
- Standardfehler der Schätzung: Standardabweichung der Stichprobenverteilung.
- Schätzer: Regel zur Verwendung der Daten zur Schätzung von θ.

Was sind Beispiele für solche Schätzer? Schätzer für Mittelwert der Grundgesamtheit μ ist der Durchschnitt der Zufallsstichprobe $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. Diesen Schätzer hatten wir schon einmal im Thema: statistische Modelle kennengelernt.

Das waren jetzt wieder viele theoretische Konzepte, das müssen wir erstmal in einem Beispiel verdauen.

Hier sind die ersten paar Beobachtungen einer Stichprobe, es tut gerade nicht zur Sache, für welche Fragestellung diese Daten gesammelt wurden. Nur soviel, da ist eine Variable "var" drin und wir haben Beobachtungen für zwei Gruppen (variable "cat"). Wir interessieren hier dafür, ob wir Unterschiede in "var" zwischen den Gruppen sehen. Die Beobachtungen in diesen beiden Gruppen wurden unabhängig von einander erhoben (z.B. innerhalb eines Experiments, wo eine Gruppe die Behandlungsgruppe ist, die andere die Kontrollgruppe).

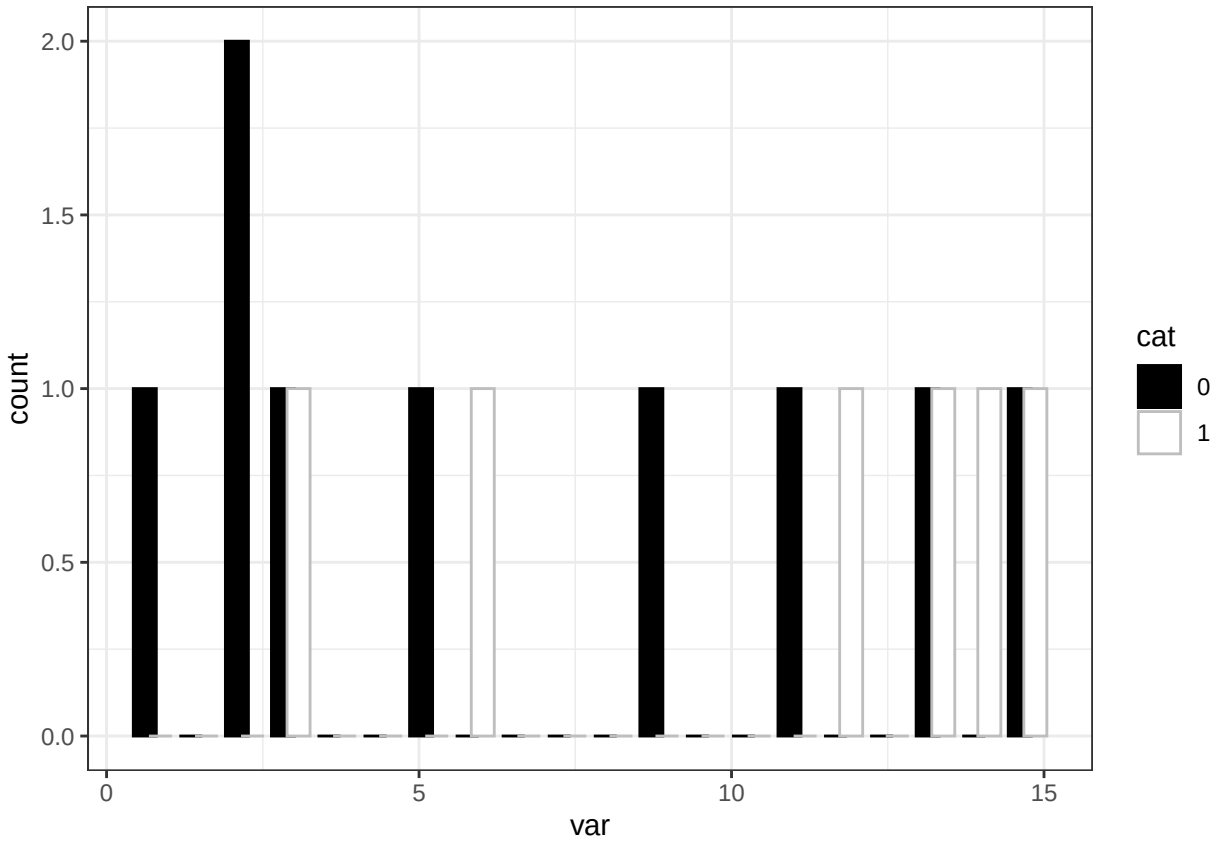
```
data <- read.csv('../data/gv300_data_fakeData.csv') %>% mutate(cat=factor(cat))
data %>% head()
```

```
##   var cat    varCorr varWeakCorr   varInd varNonMon  varBiRaw varOutlier
## 1   3   0  6.3203350  17.7186400 4.299648      8.75 0.0316617  -1.000000
## 2   5   0  6.8023790   8.6798990 5.445352     12.75 0.6922094  -1.666667
## 3   3   1  0.0886511   0.7677075 1.029927      8.75 0.1589295  -1.000000
## 4  15   0 12.8066200  18.4330800 9.987334      2.75 0.1257472  50.000000
## 5  12   1 11.5913100   8.5000760 7.071755     11.00 0.6499108  -4.000000
## 6   6   1  1.2054410   1.8113440 3.001570     14.00 0.6948789  -2.000000
```

```
##   varOutlierNoise varBi    varExp
## 1       11.278350     0  1.822119
## 2       15.591280     1  2.718282
## 3       16.707490     0  1.822119
## 4       68.350900     0 20.085540
## 5       -0.429595     1 11.023180
## 6       13.922400     1  3.320117
```

Dann, erstmal die Daten plotten, das wir uns ansehen können, wie die Werte der Variablen verteilt sind. Hier die absolute Häufigkeit in einem Histogramm dargestellt. Keine Erkennbare bekannte Verteilung, aber auch keine Ausreißer. Mit Bezug auf die Frage, ob es da einen Unterschied gibt zwischen den Gruppen (dargestellt in weiß und schwarz), kann man nicht wirklich sagen.

```
data %>% ggplot(aes(x=var,color=cat,fill=cat)) +
  geom_histogram(bins=20,position=position_dodge(width=.5)) +
  scale_color_manual(values=c('black','grey')) +
  scale_fill_manual(values=c('black','white')) +
  theme_bw()
```



Wir wollen aber wissen, ob es einen Mittelwertunterschied in der Variablen `var` zwischen den beiden Gruppen gibt. Und wir wollen wissen ob unsere Stichprobe von einer Grundgesamtheit kommt, wo es auch diesen Unterschied wirklich gibt. Wir brauchen dafür einen Test! Aber welchen? Wir werden uns viele Test ansehen (ihr habt in der Regressionsanalyse auch schon solche Tests gesehen, ich sag euch aber auch noch genau wo). Hier machen wir aber zur Illustration, den absoluten Standardtest von Unterschieden im Mittelwert, wir sehen uns den **t-Test** an. Der t-Test vergleicht die Mittelwerte zweier (unabhängiger) Stichproben. Der t-Test ist hier unser Schätzverfahren für die Differenz in den Mittelwerten. Erstmal den Test laufen lassen:

```
data %>% t.test(var~cat,data=.)
```

```
##
##  Welch Two Sample t-test
##
## data:  var by cat
## t = -1.4019, df = 11.571, p-value = 0.1872
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -9.531011  2.086567
## sample estimates:
## mean in group 0 mean in group 1
##        6.777778       10.500000
```

Da kommt einiges an Information raus. In der letzten Zeile sind die Mittelwerte in Gruppe 0 und 1 aufgeführt. Da ist also ein Unterschied im Mittelwert. Glauben wir aber nun wirklich, dass da ein Unterschied zwischen den Gruppen ist? Wird dieser Unterschied auch in der Grundgesamtheit zu finden sein oder nur zufällig in unserer Stichprobe auftauchen? Nur aus dem Vergleich der Mittelwerte können wir das nicht sagen. Es kann rein zufällig sein, dass die beiden Gruppen in unserer Stichprobe diese Mittelwerte haben. In einer anderen Stichprobe aus der selben Grundgesamtheit mögen die Mittelwerte in den beiden Gruppen ganz andere sein. Was sagt uns dieser t-Test nun darüber ob wir hier Belege für einen Unterschied zwischen den Gruppen haben?

Hier kommt die Stichprobenverteilung ins Spiel. Die Statistik die uns interessiert ist die Differenz in Mittelwerten. Die Statistik, die der t-Test aber verwendet ist eine kleine, aber wichtige, Ableitung dieser Differenz, genannt **t-Statistik**. Hier ist die **t-Statistik** (wir sehen den Wert der t-Statistik für unsere Stichprobe auch in der dritten Zeile oben im Testergebnis aufgelistet):

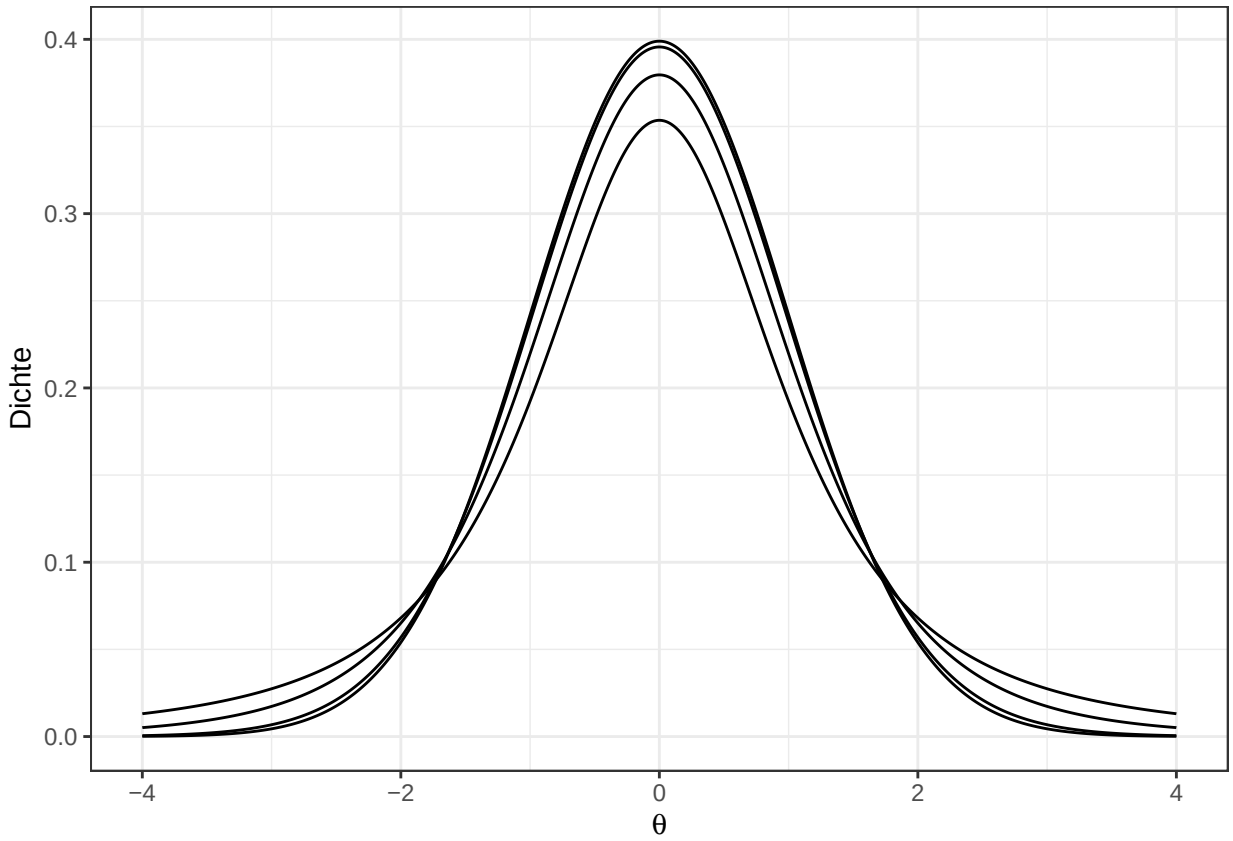
$$t = \frac{|\overline{var}_1 - \overline{var}_0|}{\sqrt{\sigma^2_{var_1,var_0}}\sqrt{\frac{1}{N_1} + \frac{1}{N_0}}}$$

Was ist das denn? Die t-Statistik ist die absolute Differenz zwischen den Mittelwerten ($|\overline{var}_1 - \overline{var}_0|$) geteilt durch die sogenannte **zusammengefasste** Varianz der Werte der Variable "var" in den beiden Gruppen und der Anzahl an Beobachtungen in den beiden Gruppen. Warum nimmt der t-Test nicht einfach die Differenz, sondern diese Ding hier als Stichprobenstatistik? Teilen durch die zusammengefasste Varianz macht die t-Statistik vergleichbar mit dem t-Statistiken aus jedem anderen t-Test mit ganz anderen Variablen (und damit ganz anderen Skalen). Teilen durch die Anzahl an Beobachtungen adjustiert die t-Statistik gegeben der Stichprobengröße: je mehr Beobachtungen, umso größer die t-Statistik bei gleicher Differenz.

Und hier kommt der Kicker, egal welche Variablen wir in diesen Test einfügen, die t-Statistik folgt immer der gleichen Stichprobenverteilung (mit leichten Unterschieden gegeben der Stichprobengröße). Die Abbildung unten zeigt die Verteilung der t-Statistik für verschiedene Stichprobengrößen und auch im Verhältnis zur Normalverteilung:

```
data.frame(theta = seq(-4,4,.001)) %>%
  mutate(
    x1=dnorm(theta,0,1),
    x2=dt(theta,2),
    x3=dt(theta,5),
    x4=dt(theta,30)) %>%
  pivot_longer(cols=x1:x4) %>%
  mutate(name=factor(recode(name,
    'x1'='Normal','x2'='t, df=2','x3'='t, df=5','x4'='t, df=30'),
    level=c('Normal','t, df=2','t, df=5','t, df=30'))) %>%
  ggplot(aes(y=value,x=theta,Farbe=name,Linientyp=name)) +
    geom_line() +
    scale_color_manual(values=c('black','red','red','red')) +
```

```
    scale_linetype_manual(values=c(1,2,3,1)) +
    labs(y='Dichte',x=expression(theta)) +
    theme_bw() +
    theme(legend.position='bottom',legend.title=element_blank())
```



Ok, was sagt uns die t-Statistik darüber, ob wir wirklich sicher sein können, dass die Differenz zwischen den beiden Gruppen in der Variable "var" nicht nur in unsere Stichprobe zufällig auftaucht, sondern auch in der Grundgesamtheit wirklich vorhanden ist? Erstmal, was heißt es das die Differenz vorhanden ist? Der Unterschied zwischen den Mittelwerten in den beiden Gruppen sollte Größer als 0 sein, nicht wahr? Wenn die t-Statistik in unserer Stichprobe nun weit weg von 0 ist, dann sollte das uns Sicherheit geben, dass in anderen Stichproben aus der Grundgesamtheit, die Stichprobe auch nicht 0 sein sollte.

Seht euch nochmal die Abbildung oben an, im Besonderen zeigt die Abbildung die Verteilung der t-Statistik für eine Grundgesamtheit, wo es keine Differenz gibt. Daher würde unsere Stichprobe meistens eine Differenz und daher meistens eine t-Statistik die 0 ist produzieren (aka der Modalwert, Median und Mittelwert ist 0). Stichproben sind zufällig gezogen und wir haben nicht unendlich viele Beobachtungen, damit ist die Differenz in einer Stichprobe manchmal nicht 0, auch wenn die Differenz in der Grundgesamtheit 0 ist.

Jetzt nehmen wir die Differenz in den Mittelwerten und die t-Statistik in unserer Stichprobe, wie oben im Testergebnis ausgegeben, und sehen nach, wo in der Verteilung der t-Statistik unsere Stichprobe landen würde. Der Wert der t-Statistik in unserer Stichprobe ist -1.4019. Unsere Stichprobe ergibt also eine t-Statistik, die schon relativ Weit im linken Ende der Verteilung der t-Statistik einer Grundgesamtheit wo es keine Differenz gibt steckt.

Anders ausgedrückt, in einer Welt (einer Grundgesamtheit) wo es keinen Unterschied zwischen den Gruppen gibt, sollte die Differenz und damit die t-Statstik meist 0 sein und die Verteilung der t-Statistik um 0 herum verteil sein. Das zeigt uns die Abbildung oben. Wir fragen also jetzt: ist die Differenz und damit die t-Statistik in unserer Stichprobe weit genug weg von 0, dass es sehr unwahrscheinlich ist, dass die Stichprobe

aus einer Grundgesamtheit kommt, in der die Differenz 0 ist.

Was heißt unwahrscheinlich? Gängigerweise sagen wir, um zu akzeptieren, dass die die Differenz in der Grundgesamtheit wirklich 0 ist, kann die Wahrscheinlichkeit, dass die t-Statistik aus unserer Stichprobe (das war der Wert -1.4019) in der Verteilung der t-Statistik wie oben abgebildet vorkommt nicht weniger als 0.05 sein (manche nutzen auch 0.1 oder 0.01). Umso weiter der Wert der t-Statistik aus unserer Stichprobe in den Enden der Verteilung oben verschwindet, umso unwahrscheinlich ist dieser Wert in der Verteilung der t-Statistik einer Grundgesamtheit mit einer 0-Differenz.

Unsere statistische Software berechnet diese Wahrscheinlichkeit gleich mal für uns. Schaut euch nochmal das Testergebnis an, ich drucke es hier nochmal ab:

```
data %>% t.test(var~cat,data=.)
```

```
##
##  Welch Two Sample t-test
##
## data:  var by cat
## t = -1.4019, df = 11.571, p-value = 0.1872
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -9.531011  2.086567
## sample estimates:
## mean in group 0 mean in group 1
##        6.777778       10.500000
```

Da wird uns ein $p - value$ von 0.1872 ausgeworfen. Das ist die Wahrscheinlichkeit, dass in der Wahrscheinlichkeitsverteilung der t-Statistik in einer Grundgesamtheit mit einer Differenz zwischen den Gruppen von 0, der Wert der t-Statistik kleiner, gleich -1.4019 ist. Mit einer Wahrscheinlichkeit von 0.1872 kann eine Grundgesamtheit mit 0 Unterschied zwischen den Gruppen eine negative Differenz erzeugen, die einer t-Statistik von -1.4019. Das ist wahrscheinlicher als was wir gängigerweise als unwahrscheinlich bezeichnen (≤ 0.05). Damit können wir nicht ablehnen, dass da keine Unterschied in der Grundgesamtheit ist (und schlussfolgern, dass da wohl ein Unterschied sein müsste).

Voila, das war statistische Inferenz. Nicht einfach, nicht geradlinig, basierend auf vielen Annahmen und Schritten. Aber korrekt. Mit Hilfe diesen Verfahrens können wir Schlüsse über Daten ziehen, die wir nicht haben (etwa alle Werte der Grundverteilung), aus Daten die wir haben (unsere Stichprobe). Das ist super, aber Bedarf halt auch Übung. Daher werden wir uns eine Weile hiermit beschäftigen, wir wollen ja mit vierlei Variablen, vielerlei Statistiken, verschiedenen Tests und unterschiedlichen Stichprobenverteilungen arbeiten. Abhängig von unserern Daten und unserer Fragestellung, unterscheiden sich die Tests.

Noch eines. Wir vergleichen die Stichprobenstatistik immer mit einer Verteilung einer Statistik, die hypothetisch ist. Wie etwa die Verteilung der t-Statistik in der Grundgesamtheit ohne Differenz in den beiden Gruppen. Wir nennen diese Hypothese, dass da keine Differenz ist **Null-Hypothese** und die Verteilung der Statstik gegeben solch eine Hypothese **Null-Verteilung**.

Nächste Woche fahren wir daher fort mit dem was wir allgemeine als **Null-Hypothesentesten** verstehen.

Literatur

Baron, David P., and John A. Ferejohn. 1989. “Bargaining in Legislatures.” *The American Political Science Review* 83 (4): 1181–1206. https://doi.org/10.2307/1961664.

DesignApplause. 2020. “The Ever Evolving London Underground Tube Map,” 2020. https://designapplause.com/design/concept-design/the-ever-evolving-london-underground-tube-map/35186/.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

Transport for London. 2020. “TfL Tube Map,” 2020. http://content.tfl.gov.uk/standard-tube-map.pdf.