

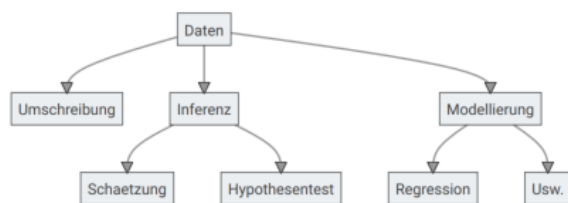
1 Einführung

Wissenschaft ist Methode: der systematische Versuch, Muster zu erkennen. Etwas zu entdecken, was nicht so einfach sichtbar ist.

Ziele in der Politikwissenschaft:

- Politische Welt **beschreiben**
- Statistische **Inferenz**: Aus wenigen Datenpunkten auf Allgemeines schliessen (statist. Relevanz, W'keitstheorie, Hypothesentests, Validität)
- Theoretische **Modelle** für Politik entwickeln (und empirische Überprüfung derselbigen durch Umwandlung in statistische Modelle)
- Suche nach **Ursachen** (Erforschung Kausalität)
- **Vorhersagen** (schwierig), z.B. mit Data Science

Strukturierung Inhalt:



2 Konzepte, Daten, Messungen

Daten sind eine Sammlung von Beobachtungen, die Informationen über Phänomene enthalten.

Datensätze enthalten

- Zeilen: **Forschungseinheiten**
- Spalten: **Variablen** (Merkmale der FE)

Arten von Variablen:

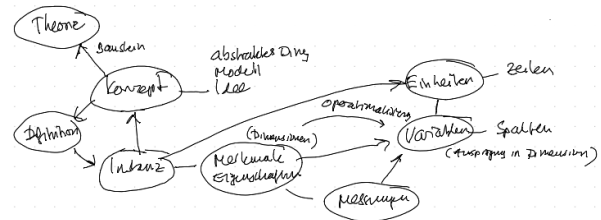
- Numerisch
- Faktoren (Enums, d.h. diskrete Kategorien codiert als numerischer Wert)
- Strings

Datensätze in einem *Codebook* beschrieben. Variablen sind oft Charakteristika aus einer bestimmten Theorie (Strukturalismus, Modernismus etc).

Konzepte¹ sind abstrakte Begriffe, mentale Konstrukte. Umfasst verschiedene empirische, messbare Phänomene und kann abgegrenzt werden (welcher Fall gehört dazu, welcher nicht). Konzepte sind die Bausteine von Theorien und die Grundlage der Datenerfassung (Messung).

¹Der deutsche Begriff *Konzept* steht eigentlich für etwas anderes: für einen Plan, eine strukturierte Vorgehensweise. Hier ist der englische Begriff *Concept* gemeint.

Die Politikwissenschaft kennt sehr viele Konzepte, auch geteilt mit anderen Disziplinen (Ökonomie, Soziologie, Psychologie etc.) Debatten: Ist ein neues Konzept nötig, um ein Phänomen zu beschreiben, oder genügen alte Konzepte? Konzepte können „ausser Mode“ geraten und nicht mehr passen.



Woher kommen Konzepte?

- Deduktion: abgeleitet (allgemein → spezifisch). Ähnlichkeiten beobachten und als Konzept zusammenfassen
- Induktion: hergeleitet (spezifisch → allgemein). Idee formulieren, nach Elementen mit spezifischen Merkmalen suchen.

Konzepte können auch zu umfassend sein, dann sind sie kein wirksamer Begriff und müssen mit weiteren Diskriminatoren genauer spezifiziert werden.

Ontologien Beschreiben, welche Objekte, Konzepte, Strukturen und Beziehungen in der Welt es gibt, über die wir Wissen erlangen können.

Arten von Konzeptualisierungen:

Konzepte können durch **notwendige** und **hinreichende** Merkmale oder Dimensionen beschrieben werden. Beispiel: „freie Wahlen“ sind ein notwendiges Merkmal für eine Demokratie, aber sie sind noch nicht hinreichend.

Operationalisierung Zuordnung von Konzepten zu empirischen Phänomenen, welche diese Konzepte erfassen. In der Regel werden dafür **Indikatoren** gesucht, die messbare Werte haben, und **Datenquellen**, welche diese Indikatoren liefern können.

Die Operationalisierung bildet also Brücke zwischen Theorie und empirischer Forschung.

Beispiel: Messung von populistischen Einstellungen in der Bevölkerung.

- Datenquelle = Umfrage
- Indikatoren = Fragen zu populist. Dimensionen, z.B. als **Likert-Items** mit Antwortskala: 1=agree completely, ..., 5=disagree completely

Manchmal kann eine Dimension mit einem **einfachen Indikator** beschrieben werden (Alter), bei komplexen Konzepten sollten **mehrere Indikatoren** pro Dimension definiert werden (mehrere Fragen zur gleichen Dimension). Damit lässt sich die Bedeutung eines Konzepts besser abbilden und es ist möglich, Skalen zu erstellen, die zuverlässiger sind.

Messungen und Variablen können verschiedene **Skalentypen** haben:

nominal Klassen und Kategorien werden durch untersch. Zahlwerte repräsentiert

ordinal Kategorien mit Rangordnung, können sortiert werden

interval Gleiche Abstände zwischen Werten, Intervalle: $A-B = C-D$. Aussagen wie „gleiche Dauer wie“ sind möglich.

ratio Gleiche Abstände zwischen Werten, Gleiche Verhältnisse $A:B = C:D$, absolute Null, keine negativen Werte. Aussagen wie „n-mal länger als“ sind möglich.

Messungen sind **zuverlässig** (reliable) wenn sie bei Wiederholungen konsistent sind. **Messfehler** können die Zuverlässigkeit beeinträchtigen. Wenn wir mehrere Indikatoren benutzen, können wir **additive Skalen** erstellen, welche zuverlässiger sind als ihre Einzelteile. Ein **systematischer Fehler** (z.B. falsch eingestellte Uhr) beeinträchtigt die Zuverlässigkeit einer Messung nicht, sondern die Gültigkeit.

Cronbach Alpha todo

Validität (Gültigkeit) bezeichnet den Umstand, ob wir messen was wir zu messen beabsichtigen. Ein guter Indikatorensatz (dh. eine gute Operationalisierung) misst den gesamten Inhalt eines Konzepts und nichts anderes.

Arten von Validität: todo

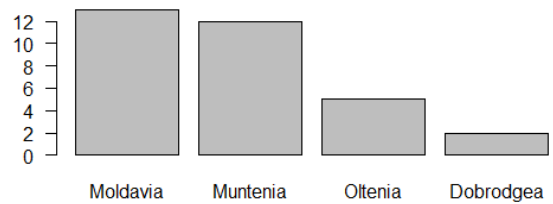
- Face validity
- Inhaltliche Gültigkeit
- Konstruktivität

3 Datenvisualisierung

Mächtiges Werkzeug mit Potential für Missbrauch. Erlaubt Einsicht in Muster, die in Daten enthalten sind, zu erhalten. Wichtigste Visualisierungstypen:

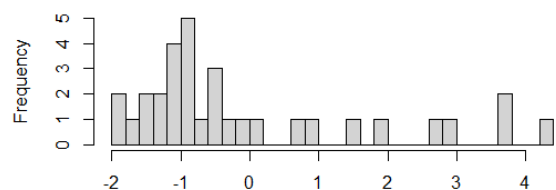
Balkendiagramm (Barchart) Zählt Vorkommen von Kategorien, 1 Balken pro Kategorie. Nicht numerisch, Faktoren (R). Skalen: nom, ord

```
freq <- table(peasant$area)
barplot(freq)
```



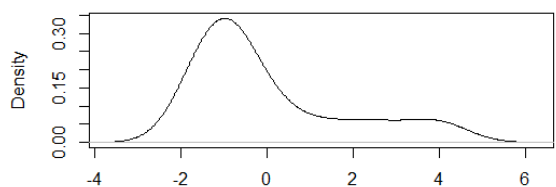
Histogramm Zählt numerische Variablen, in Intervalle gruppiert (Bins/Buckets). Bins segmentieren die Daten. Skalen: intv, ratio

```
hist(peasant$I, breaks=30)
```



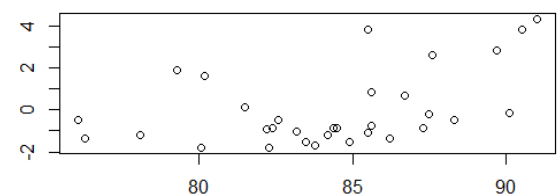
Dichtediagramm (Density) Glatte Annäherung an Histogramm. Numerische Variablen, kontinuierlich visualisiert. Bandwidth definiert die Fitness der Kurve. Skalen: intv, ratio

```
density <- density(peasant$I)
plot(density, bandwidth=0.7)
```



Streudiagramm (Scatterplot) Beziehung zwischen zwei Variablen (**bivariat**), zeigt Wertepaare in 2D Raum. Gleichgrosse Vektoren! ⚠️

```
t <- peasant$T; i <- peasant$I;
plot(t, i, xlab=., ylab=.)
```



Tricks/Fehlerhafte bei Visualisierung (**the bad**):

- Rosinenpickerei bei Daten / Bewusstes Weglassen von Daten („Find the data that fits the narrative“)
- Zahlen, die nicht zusammenpassen (% , Units)

- Falsche oder fehlleitende Skalierung
- Mehrere Achsen in unterschiedl. Massstäben
- Hochrechnungen ohne Angabe von Unsicherheit

Grundsätze der Visualisierung (**the good**):

- Unnötiges weglassen
- Nüchternheit
- Zu den Daten passende Darstellung, Elemente und Konfiguration (Symbole, Farben) wählen
- Datenstruktur folgen, Muster hervorheben
- Angemessene Skalen, keine Über-/Untertreibung
- Vergleich erleichtern
- Klarheit: Hinweise, Beschriftungen, Legenden
- Angabe von Kontext, wo nötig und sinnvoll (z.B. Ausgangswerte, Quelle, Jahr)

4 Deskriptive Statistik

Die deskriptive Statistik dient dazu, **Eigenschaften eines Datensatzes** zu verstehen (z. B. Verteilung, Zentrum, Streuung). Mittels **Kennzahlen** (Mittelwert, Median, Varianz, Korrelation) oder **Visualisierungen** (Histogramme, Streudiagramme).

Was ist eine **Statistik**? Ein numerischer Wert, der aus einer Stichprobe von Daten berechnet wird.

4.1 Univariate Statistik

Die univariate Statistik bezieht sich auf die **Analyse einer einzigen Variablen**.

Ziel Eigenschaften und Verteilungen der Variablen untersuchen, z.B. mittels **Lagemassen** wie Mittelwert, Median, Modus oder mit **Streuungsmassen** wie Varianz, Spannweite, Standardabweichung.

Beispiel Untersuchung der *Körpergrösse* in einer Population.

4.1.1 Masse der zentralen Tendenz

Beschreiben die zentrale oder typische Beobachtung einer Verteilung.

Modus Wert, der am häufigsten erscheint.

Median Wert in der Mitte sortierter Daten. 50% Daten grösser, 50% kleiner. Auch: **2. Quartil**.

Mittelwert Arithmetisches Mittel (mean) ist Summe der Werte geteilt durch Stichprobengrösse:

$$\bar{y} = \frac{1}{n} \sum_{i=0}^n y_i$$

Quantile Das p-te Quantil ist der Wert, bei dem p% der Daten einen kleineren Wert haben, z.B. 10%-Quantil. Viertelquantile sind nummeriert:

1. Quartil: 25% kleiner, 75% grösser
2. Quartil: Median
3. Quartil: 75% kleiner, 25% grösser

4.1.2 Streuungsmasse

Beschreiben die Heterogenität von Daten.

Range Variationsbreite der Werte einer Variable

$$R = \max(y) - \min(y)$$

Varianz Durchschnitt der *quadrierten* Abweichungen vom Mittelwert. Quadrieren, um Ergebnis 0 durch gegenseitige Aufhebung zu vermeiden.

$$s^2 = \frac{\sum_{i=0}^n (y_i - \bar{y})^2}{n-1}$$

Warum $n - 1$ und nicht n ? Wir kennen den Mittelwert (ist Teil der Formel), deshalb haben wir einen **Freiheitsgrad** verloren.

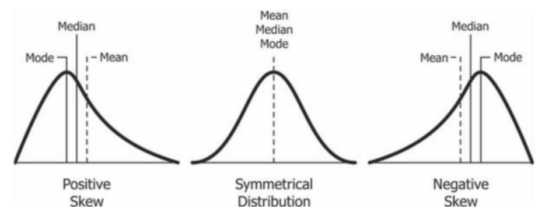
Standardabweichung Varianz ist *quadrierte* Einheit. Rückkehr zur ursprünglichen Messskala durch Ziehen der Quadratwurzel

$$s = \sqrt{s^2}$$

Skewness (Schiefe) Schiefe einer Verteilung.

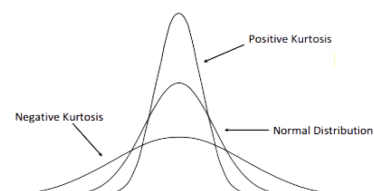
$b_1=0$: *symmetrisch*, $b_1>0$: *positiv* (rechtsschief, linkssteil), $b_1<0$: *negativ* (linksschief, rechtssteil).

$$b_1 = \frac{1}{n} \frac{\sum_{i=1}^n (y_i - \bar{y})^3}{s^3}$$



Kurtosis (Wölbung) Wird relativ zur Normalverteilung gemessen ($NV = 3$). $b_2<3$: *Platykurtosis* (flach), $b_2>3$: *Leptokurtosis* (spitz).

$$b_2 = \frac{1}{n} \frac{\sum_{i=1}^n (y_i - \bar{y})^4}{s^4}$$



Interquartilabstand (IQR) Abstand zwischen 1. und 3. Quartil, wird als Mass für Streuung ordinalskalierter Variablen verwendet.

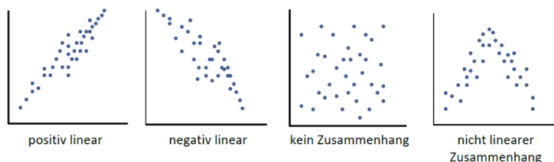
4.2 Bivariate Statistik

Die bivariate Statistik untersucht die **Beziehung zwischen zwei Variablen**.

Ziel Abhängigkeiten, Zusammenhänge oder Unterschiede zwischen den beiden Variablen analysieren, z.B. mittels **Korrelation**, d.h. Untersuchung der Stärke und Richtung eines linearen Zusammenhangs (z. B. Pearson-Korrelation) oder mittels **Regressionsanalyse**, der Untersuchung, wie eine Variable durch die andere vorhergesagt werden kann.

Beispiel Zusammenhang zwischen *Körpergrösse* und *Gewicht* untersuchen.

Kovarianz (Linearer Zusammenhang) Zusammenhang der Werte zweier Variablen. *positiv*: tiefe Werte x mit tiefen Werten y, *negativ*: tiefe Werte x mit hohen Werten y.



Mass ergibt sich aus durchschnittlichem Kreuzprodukt der Abstände zu den Mittelwerten.

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n - 1}$$

Interpretation: 0 = kein Zusammenhang
(positiv, negativ) = positive/negative Assoz.

Korrelation Normalisierte Kovarianz im Bereich (-1,1) mit **Pearson-Korrelationskoeffizient**:

$$r_{xy} = \frac{\text{Kovarianz}(X,Y)}{\text{SD}(X) \cdot \text{SD}(Y)} = \frac{s_{xy}}{s_x \cdot s_y}$$

Interpretation: (1,-1) = perfekt, 0 = kein

⚠ Korrelation kann 0 sein bei perfektem nicht-linearem Zusammenhang!

5 Statistische Inferenz

Wenn wir aus Stichprobendaten auf eine Gesamtpopulation schliessen, dann nennt man dies **Inferenz**.

Um eine **Zufallsstichprobe** zu ziehen, muss

1. ein **Auswahlrahmen** gewählt werden, der möglichst die gesamte Grundpopulation abdeckt (Telefonbuch, Steuerdaten, Fahrzeugregister, etc.)

2. eine **Auswahl** (Sample) mit **Grösse n** gewählt werden. Jedes Auswahlitem muss gleiche W'keit haben, gewählt zu werden, kein Item mit $p=0$!

Probleme:

- Auswahlrahmen deckt sich nicht mit Zielpop
- Over-coverage: Auswahlrahmen \notin Zielpop
- Under-coverage: Zielpop \notin Auswahlrahmen
- Non-response: Einheiten liefern keine Daten

D.h. eine (grössere) Teilmenge hat $p=0$, um in die Probe aufgenommen zu werden.

Beispiel Telefonverzeichnis Kaum mehr Festnetzanschlüsse, enthält ggf. Personen, die nicht Abstimmungsberechtigt sind, oder angerufene wollen keine Auskunft geben.

Warum dürfen wir Stichproben verwenden (bei grosser Zielpopulation und genügend grossem n)?

Zentraler Grenzwertsatz Mittelwert der Mittelwerte aller Stichproben ist normalverteilt (gestreut) um den Mittelwert der Zielpopulation.

💡 Stichproben beinhalten immer **Unsicherheit**. Bei kleiner Zielpopulation sollte wenn möglich eine Vollerhebung gemacht werden!

5.1 Unsicherheit

Quellen für Unsicherheit sind:

- Unvollständige Theorie, limitiertes Modell
- Verwendung von Stichproben
- Messfehler

Zufall, resp. Unsicherheit kann durch **Wahrscheinlichkeiten** quantifiziert werden. Zwei Arten:

Frequentismus W'keit als relative Häufigkeit bei (unendlich) vielen Wiederholungen $\lim_{n \rightarrow \infty}$, muss also durch Experimente bestimmt werden.

Bayesianismus W'keit als subjektiver Grad der Überzeugung. Kann auch für Ereignisse mit wenig Wiederholungen (n) definiert werden, braucht keine Wiederholungen.

5.2 Wahrscheinlichkeitsrechnung

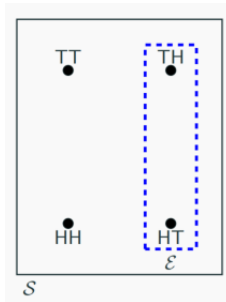
Terminologie:

Experiment Vorgang unter kontr. Bedingungen

Stichprobenpunkt (auch: **Ergebnis**) Ein Wert im Stichprobenraum, z.B. K bei Münzwurf, oder KZ bei zweimaligem Münzwurf.

Stichprobenraum \mathcal{S} (auch: **Ergebnisraum**) Die Menge aller möglichen Ergebnisse eines Experiments, z.B. $\{K, Z\}$ für einen Münzwurf, oder $\{KK, KZ, ZK, ZZ\}$ bei zweimaligem Münzwurf.

Ereignis \mathcal{E} Eine Teilmenge des Ereignisraums, d.h. kombinierte Ergebnisse, z.B. $\{KK, ZZ\}$ (nur Köpfe oder nur Zahlen) oder $\{KZ\}$ (zuerst Kopf, dann Zahl)

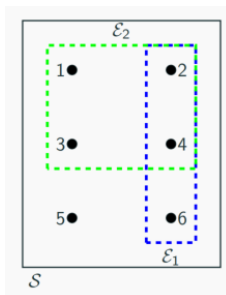


Die Wahrscheinlichkeit über den gesamten Stichprobenraum ist $P(\mathcal{S}) = 1$.

Randwahrscheinlichkeit Die W'keit eines Ereignisses unabhängig von anderen Faktoren ist $P(\mathcal{E}_1)$, z.B. $\frac{1}{2}$ für $\{KZ, ZK\}$ oben.

Bedingte Wahrscheinlichkeit W'keit für Ereignis \mathcal{E}_2 , gegeben $\mathcal{E}_1 = P(\mathcal{E}_2|\mathcal{E}_1)$, z.B. W'keit für Zahl < 5 wenn Zahl gerade ist $\frac{2}{3}$ (unten).

Überlegung: Erstes Ereignis schränkt Stichprobenraum auf \mathcal{S}' ein, zweites Ereignis ist dann noch Teilmenge in \mathcal{S}' .



Multiplikationsregel (und) Für gleichzeitig vorkommende Ereignisse gilt: $P(\mathcal{E}_1 \cap \mathcal{E}_2) = P(\mathcal{E}_2|\mathcal{E}_1) \times P(\mathcal{E}_1) = P(\mathcal{E}_1|\mathcal{E}_2) \times P(\mathcal{E}_2)$, z.B. W'keit für Zahl gerade *und* auch <5 also $\frac{1}{3}$ (oben).

Additionsregel (inklusive-oder) Es gilt $P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1 \cap \mathcal{E}_2)$, doppelt gezählte Ergebnisse müssen wieder abgezogen werden. Bei sich ausschliessenden Ereignissen somit $P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2)$.

Zwei Ereignisse sind **statistisch unabhängig** wenn \mathcal{E}_1 keine Information zu \mathcal{E}_2 liefert. Dann gilt $P(\mathcal{E}_2|\mathcal{E}_1) = P(\mathcal{E}_2)$ und somit auch $P(\mathcal{E}_1 \cap \mathcal{E}_2) = P(\mathcal{E}_2) \times P(\mathcal{E}_1)$.

5.3 Zufallsvariablen

Mit Zufallsvariablen bezeichnen Werte, die unsicher sind, z.B. Variablen in Untersuchungen oder Experimenten. Wir kennen zwar den Wertebereich, können aber nicht mit Sicherheit *a priori* wissen, welchen Wert die Variable schlussendlich annimmt.

Die möglichen Werte einer **Zufallsvariable** werden mit Hilfe einer **Wahrscheinlichkeitsverteilung** beschrieben. Der tatsächliche Wert hängt vom Zufall ab. Name= $\{X, Y, \dots\}$, Wert= $\{x, y, \dots\}$.

Die Menge aller möglichen Werte einer Zufallsvariable Y wird als **Unterstützung** bezeichnet.

Beispiel Eine Variable Wahlbeteiligung Y. Mögliche Werte: Y=1 gewählt, Y=0 nicht gewählt, mit sich zu 1 ergänzenden Wahrscheinlichkeiten: p, 1-p. Berechnung mit *Bernoulli-Verteilung*.

Der Oberbegriff **Wahrscheinlichkeitsfunktion $p(y)$** bezeichnet die Funktion für eine Zuordnung der Wahrscheinlichkeit an Werte von Y. Abhängig von der Art der Zufallsvariable (diskret oder stetig) ist entweder als Masse- oder Dichtefunktion.

W'keitsmassefunktion Bei **diskreten** Zufallsvariablen (dh. mit abzählbaren Werten) gibt die Wahrscheinlichkeitsfunktion direkt die W'keit an, dass Wert y beobachtet wird: $P(Y=y) = p(y)$

Die Summe aller W'keiten ergibt 1.

W'keitsdichtefunktion Bei einer **stetigen** Zufallsvariable (dh. mit kontinuierlichem Wertebereich) gibt es keine direkte W'keit für einen Wert, es muss über ein Intervall zwischen zwei möglichen Werten integriert werden:

$$P(Y=y) = P(a \leq y \leq b) = \int_a^b p(y) dy$$

Die Fläche unter der Kurve ist immer 1.

Dabei steht $P(Y=y)$ oder $\Pr(Y=y)$ für die Wahrscheinlichkeit, dass Y den Wert y annimmt und $p(y)$ steht für die Wahrscheinlichkeitsfunktion, welche einem Wert y eine Wahrscheinlichkeit zuordnet.

Unter **kumulative Wahrscheinlichkeit** versteht man die Wahrscheinlichkeit $P(Y \leq y)$.

5.3.1 Verteilungen

Es gibt verschiedene Wahrscheinlichkeitsverteilungen. Wie auswählen? Muss zu Art der Daten passen!

- Empirisch: Für einige Phänomene bekannt
- Theoretisch: z.B. Annahme Messfehler norm' vert.
- Pragmatisch: Beginne mit einfacher Verteilung

Vergleich mit empirisch erhobenen Daten zeigt, ob die gewählte Verteilung als Prädiktor taugt.

Bernoulli-Verteilung Ein einziges Experiment mit genau zwei möglichen Ausgängen 1 und 0 (Erfolg und Misserfolg). Der Parameter p ist die W'keit für Erfolg, d.h. $P(Y = 1) = p$, somit W'keit für Misserfolg $P(Y = 0) = 1 - p$.

Bernoulli-Verteilung ist ein **Spezialfall der Binomial-Verteilung** mit $n=1$. Vereinfacht:

$$P(Y = y) = \begin{cases} p & \text{if } y=1 \\ 1-p & \text{if } y=0 \end{cases}$$

Beispiel: Münzwurf mit fairer Münze $p=0.5$.

Binomial-Verteilung Y diskret, Verteilungsfunktion als Histogramm. Modelliert n unabhängige Experimente (Bernoulli-Experimente), bei denen jedes Experiment die gleiche Erfolgswahrscheinlichkeit p hat.

$$P(Y=y) = \binom{n}{y} \cdot p^y \cdot (1-p)^{n-y}$$

Beispiel: Serie von Münzwürfen, also z.B. W'keit für 3x Kopf bei 4 Würfeln, $p=0.5$.

Da es sich um Unabhängige Experimente handelt, entspricht die W'keit einfach der Multiplikation der W'keiten für das Einzelexperiment.

$\binom{4}{3}$ ist der **Binomialkoeffizient** und gibt an, auf wieviele Arten 3x Kopf in 4 Würfeln angeordnet werden können (berechnet als $\frac{n!}{k!(n-k)!}$).

Es gibt 4 Möglichkeiten: KKKZ, KKZK, KZKK, ZKKK. Die Wahrscheinlichkeit für 3x Kopf bei 4 Würfeln (dh. für 3x Erfolg in 4 Versuchen) ist also: $P(X = 3) = 4 \cdot 0.5^3 \cdot (1 - 0.5)^{4-3} = 4 \cdot 0.125 \cdot 0.5^1 = 4 \cdot 0.0625 = 0.25$.

```
# Massefunktion (d*) für 3x Kopf
dbinom(3, size=4, prob=0.5) => 0.25

# Kumulative W'keit für ≤ 1 Köpfe
pbinom(1, size=4, prob=0.5) => 0.3125
```

Normalverteilung Y stetig, Dichtefunktion mit zwei Parametern $\mathcal{N}(\mu, \sigma)^2$, Parameter Mittelwert μ und Standardabweichung σ .

```
# Dichtefunktion: W'keit bei x=1
dnorm(1, mean=0, sd=0.5) => 0.10798

# Kumulative W'keit, dass x ≤ 1
pnorm(1, mean=0, sd=0.5) => 0.97725
```

Die Standardnormalverteilung ist $\mathcal{N}(0, 1)$.

5.4 Schätzer

Ein **Schätzer** ist eine Formel oder Regel, mit der man aus einer Stichprobe einen Wert berechnet, der möglichst gut einen unbekannten Wert der gesamten Population beschreibt, z.B.

- Populationsmittelwert (Durchschnitt μ)
- Populationsvarianz (Streuung σ)

Beispiel Durchschnittliche Körpergrösse? Stichprobe mit $n=100$. Der Durchschnitt der gemessenen Grössen \bar{X} ist ein Schätzer für μ , den wahren Mittelwert der gesamten Population.

Punktschätzung Eine Punktschätzung ist das konkrete Ergebnis, das durch Anwendung eines Schätzers auf eine spezifische Stichprobe erhalten wird. Jede Stichprobe ergibt i.d.R. einen anderen Wert.

Für Schätzer gibt es die folgenden Kenngrössen:

Stichprobenverteilung³ über einem Schätzer ist eine Dichtefunktion, welche die Werte des Schätzers beschreibt, die wir beobachtet hätten, wenn wir wiederholt Stichproben der Grösse n gezogen hätten. Zum Beispiel die Verteilung der durchschnittlichen Körpergrösse in Stichproben mit $n=100$.

Verzerrung eines Schätzers ist die Differenz zwischen Erwartungswert des Schätzers und dem wahren Wert in der Population.

Standardfehler Für die Normalverteilung entspricht der Standardfehler $SE[\bar{y}] = \frac{\sigma}{\sqrt{n}}$

Konfidenzintervall Eine Intervallschätzung, drückt Erwartung (z.B. 95%) aus, dass der wahre Populationsparameter im genannten Intervall liegt. Mit **Konfidenzniveau** 0.05).

²Formula muy complicada, weggelassen weil unwichtig

³Laut Steenbergen wichtigstes Konzept der Vorlesung

Gesucht ist Quantil q für $\mathcal{N}(0, 1)$, so dass $P(-q \leq z \leq q) = 0.95$. Da das Intervall symmetrisch um μ liegt, ist $q = 1 - \frac{0.05}{2}$, also 0.975.

Dieses Quantil liegt bei $\pm 1.96^4$, die Intervallgrenzen sind also $\mu \pm 1.96 \times \sigma$.

Häufige Konfidenzintervalle für $\mathcal{N}(\mu, \sigma)$

- 90%: $\pm 1.645 \cdot \sigma$ um μ herum⁵
- 95%: $\pm 1.960 \cdot \sigma$ um μ herum
- 99%: $\pm 2.576 \cdot \sigma$ um μ herum

Fehlermarge (MoE) Gibt Grösse des Stichprobenfehlers in den Ergebnissen einer Studie an.

$$\text{MoE} = z \times \text{SE}[\bar{y}]$$

Eine Vergrösserung der Stichprobe reduziert die Fehlermarge. Für Halbierung der Fehlermarge muss Stichprobe vervierfacht werden.

Berechnungsbeispiel MoE

Eine ja/nein Umfrage U mit $n=200$ ergibt 51% Zustimmung.

Die Zufallsvariable U ist Bernoulli-verteilt (hat nur 2 mögliche Werte) und der Mittelwert \bar{u} der Stichprobe liegt bei $p = 0.51$.

Bei einem Konfidenzniveau von 0.05 und Normalverteilung ist $z = 1.96$ (aus Tabelle oder berechnet mit `qnorm(...)`, siehe *Konfidenzintervall* oben).

Die Varianz des Stichprobenmittelwertes für eine Bernoulli-verteilte Zufallsvariable ist $\frac{p \times (1-p)}{n}$, d.h. $\text{Var}(\bar{u}) = \frac{0.51 \times 0.49}{200} = 0.0012495$.

Die Varianz ist σ^2 , dh. $\sigma = \sqrt{\sigma^2}$.

$$\text{SE}[\bar{u}] = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{\text{Var}(\bar{u})}{n}}$$

$$\text{MoE} = z \times \text{SE}[\bar{u}] = 1.96 \times \sqrt{\frac{0.00125}{200}} = 0.0049 = 0.49\%$$

5.5 Hypothesentests

Hypothesen sind *vorläufige* Annahmen, die empirisch geprüft werden sollen. Eine Hypothese ist so lange gültig⁶, bis sie widerlegt (falsifiziert) wird.

In der Regel werden nicht *exakte*, sondern *statistische* Hypothesen formuliert (dh. wahrheitsbasierte Hypothesen).

Eigenschaften:

- Population: Worauf bezieht sich die Hypothese?
- Variablen: Welche Beziehung wird postuliert?
- Welcher Art ist der behauptete Zusammenhang?

Gute Hypothesen sind:

- spezifisch (d.h. nicht vage)
- klar definiert (Konzepte, Begriffe)
- falsifizierbar (keine Tautologie)
- empirisch überprüfbar (messbare Indikatoren)

Hypothesen *können* nur auf einer Variable basiert sein, das ist allerdings selten der Fall.

5.5.1 Die wissenschaftliche Methode

1. Forschungsfrage formulieren
2. Theorie entwickeln (basierend auf Beobachtung)
3. Hypothese H_a formulieren (Erwartung)
4. Studie entwerfen, Operationalisierung
5. Studie durchführen, Daten sammeln
6. Analyse mit Hypothesentest: H_a gültig?
7. Diskussion der Resultate / Forschungsfrage

Hypothese (inklusive Signifikanzniveau) muss immer **vor Durchführung** der Studie formuliert werden! Sonst gilt Forschung als unseriös.

Teststatistik definieren: Was wird in Experiment (Datenerhebung) gemessen?

5.5.2 p-Wert basierte Hypothesentests

Nach **FISHER** (1990) sollten wir nicht auf Punktwahrscheinlichkeiten fokussieren, sondern auf *kumulative* Wahrscheinlichkeiten.

Wir sollten nicht fragen, wie Wahrscheinlich ein bestimmtes Ereignis ist, sondern wie gross die Wahrscheinlichkeit ist, etwas Extremes als die Daten einer gegebenen Stichprobe zu beobachten, unter Annahme H_0 wahr. Diese bedingte Wahrscheinlichkeit heisst **p-Wert** und bezeichnet, wie gut die beobachteten Daten zu H_0 passen.

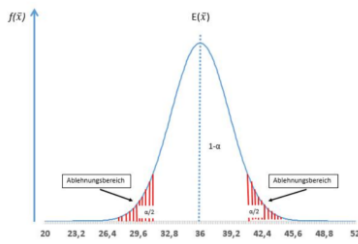
p-Wert $P(\text{Teststatistik mindestens so extrem wie die beobachtete} \mid H_0 \text{ ist wahr})$

Signifikanzniveau α W'keitsschwelle für Ablehnung der Nullhypothese.

⁴Berechnet mit `qnorm(0.975, mean=0, sd=1)`

⁵wobei σ = Standardabweichung und μ = Mittelwert

⁶Nach Popper: Beweis einer Hypothese ist unmöglich, egal wie viele positive Beispiele gefunden werden. Zur Falsifizierung (einer exakten Hypothese) genügt *ein* Gegenbeispiel.



Teststatistik z Funktion der Stichprobendaten, welche die Abweichung der beobachteten Daten von der Nullhypothese quantifiziert.

Vorgehen für Hypothesentest:

1. Formuliere H_0
2. Lege Signifikanzniveau fest
3. Entwerfe Experiment um H_0 zu testen
4. Führe das Experiment durch
5. Berechne den p-Wert aus den erhobenen Daten
6. Entscheide über das Schicksal von H_0

Fehlertypen:

Typ I Fehler H_0 fälschlicherweise abgelehnt (obwohl wahr). Dies entspricht einem **false positive**, denn dann nehmen wir an, dass ein Effekt existiert, obwohl keiner da ist. Die Wahrscheinlichkeit für diesen Fall entspricht dem Signifikanzniveau α (typischerweise 5%).

Typ II Fehler H_0 wird beibehalten (obwohl falsch). Dies entspricht einem **false negative**, weil angenommen wird, dass kein Effekt vorhanden, obwohl einer da wäre. Die Wahrscheinlichkeit für diesen Fall ist β .

Die **Power** eines Tests ist $1 - \beta$ (W'keit *keinen* Typ II Fehler zu begehen, und somit einen Effekt zu erkennen, wenn er da ist. Je höher desto besser.)

Wir wollen primär **Typ I Fehler** vermeiden, d.h. α möglichst klein halten, das geht auf Kosten von β .

Wir **wählen** also als H_0 diejenige Hypothese, deren Ablehnung für die Wissenschaft oder Gesellschaft am problematischsten wäre (Täter ist unschuldig, Medikament wirkt nicht)

Wissenschaftliche Bad Practices:

- **Verzerrung der Publikation:** Nicht signifikante Ergebnisse werden nicht publiziert
- Resultate **geringer statistischer Aussagekraft**⁷
- **p-Wert Hacking:** testen, bis signifikant⁸

⁷Behauptung: Trifft auf 90% der Publikationen zu, darum wäre Angabe der Effektgrösse wichtig.

– **HARK-ing:** Hypothesis after results are known.

Nicht zu vermeiden ist: Bei Signifikanzniveau 5% ist zu erwarten, dass 1 von 20 Studien falsche Schlüsse zieht, im Sinne eines Typ I Fehlers. Kann somit nicht reproduziert werden.

5.6 Angewandte Hypothesentests

5.6.1 \mathcal{T} -Test: Test des Mittelwerts

Für Hypothesen, die sich auf den **Mittelwert einer Stichprobe** (dh. einer Gruppe) beziehen.

Annahme über den Mittelwert einer Stichprobe, z.B. „Menschen sind im Schnitt etwa 1.69 gross“. Erwartung, dass Abweichung normalverteilt. Weil aber Stichprobe nicht so gross ist, verwenden wir \mathcal{T} anstatt \mathcal{N} für den Test.

Alternative Formulierung der Hypothese mit \bar{y} als tatsächliches Mittel der Körpergrösse und Verteilung der Abweichungen um 0: $\frac{\bar{y}-1.69}{SE[\bar{y}]} \sim \mathcal{T}_{n-1}$.

Diese Formel muss nicht explizit spezifiziert werden, in R macht das die `t.test` Funktion automatisch (Standardfehler kann aus Daten geschätzt werden).

```
size <- people$size
t.test(size, alt="two.sided", mu=1.69)
```

Mit `alternative="two.sided"` wird spezifiziert, dass wir die Hypothese „keine Abweichung“ als H_0 testen wollen.

Wenn wir nun einen p-Wert < 0.05 erhalten (z.B. bei einem gemessenen Durchschnitt von 1.73 der Körpergrösse in der Stichprobe), dann können wir H_0 verwerfen.

Alternative Formen:

- Menschen sind im Schnitt > 1.69 (`alt=greater`)
- Liebe zu Spinnen ist < 3 auf Skala 1-10 (`alt=less`)

5.6.2 \mathcal{T} -Test: Differenz der Mittelwerte

Für Hypothesen, die den **Mittelwert zweier Stichproben vergleichen** (Vergleich von zwei Gruppen).

Annahme über den Mittelwert von zwei Gruppen, z.B. dass Holländer im Schnitt grösser sind als Schweizer. H_0 als gegenteilige Annahme, dh. Holländer (\bar{y}_0) \leq Schweizer (\bar{y}_1), mit $\frac{\bar{y}_0-\bar{y}_1}{SE[\bar{y}_0-\bar{y}_1]} \sim \mathcal{T}_v$. Wir nehmen an, dass Varianz der beiden Gruppen ungleich ist.

⁸z.B. indem Zielpopulation aufgeteilt wird in Subgruppen oder indem Kriterien geändert werden

Wie oben schätzt die `t.test` Funktion von R automatisch den Standardfehler und die Varianz (Welch).

```
ch <- pop$size[pop$country == „ch“]
nl <- pop$size[pop$country == „nl“]
# H0 = dutch <= swiss
t.test(nl, ch, alt="gr", mu=0, var.equal=F)
```

Bei einem p-Wert > 0.05 ist der Unterschied nicht signifikant, und wir können H_0 nicht verwerfen.

5.6.3 Cohen's d: Effektgrösse nach T-Test

Wird oft nach einem T -Test zum Vergleich zweier Mittelwerte verwendet, um die Grösse des Unterschieds zu quantifizieren.

Berechnung $d = \frac{\bar{y}_0 - \bar{y}_1}{\sigma}$.

Effektgrössen:

- ab 0.2 \Rightarrow klein
- ab 0.5 \Rightarrow mittel
- ab 0.8 \Rightarrow gross

Berechnung mit R:

```
library(effsize)
cohen.d(nl, ch, pooled=T, conf.level=0.95)
```

Durch Verwendung der direkten Formel kann auch die Effektgrösse nach dem Vergleich gegen eine Konstante (Durchschnittsgrösse = 1.65) geprüft werden.

5.6.4 χ^2 Test: Vergleich kategorialer Variablen

Der χ^2 Test wird angewendet, um festzustellen ob es einen signifikanten Unterschied zwischen n kategorialen Variablen gibt.

Zwei Arten von Test:

Goodness-of-Fit Test Wie gut passt eine erwartete Verteilung?

Unabhängigkeitstest Sind die Werte von 2 (kategorialen) Variablen unabhängig voneinander?

Beispiele:

- Sind Abwesenheiten im Büro unabhängig vom Wochentag (Mo, Di, Mi, Do, Fr)? Erwartung (H_0): gleichverteilt, falls unabhängig.
- Sind Studienrichtung (Infk, Powi, Psych) unabhängig vom Geschlecht (m,w)? Erwartung (H_0): gleichverteilt, falls unabhängig.
- Ist Glaube an die Bibel als Wort Gottes (ja, nein) unabhängig von der politischen Partei? (Dem, Rep, Indep). Erwartung (H_0): gleichverteilt, falls unabhängig.

Dabei werden beobachtete Häufigkeiten paarweise mit erwartete Häufigkeiten verglichen und normalisiert aufsummiert. Der berechnete Wert wird am Ende, abhängig von den Freiheitsgraden df und dem Signifikanzlevel α , gegen einen kritischen Wert verglichen. Alle beobachteten **Häufigkeiten müssen ≥ 5** sein.

Die Freiheitsgrade sind $df = \text{Kategorien} - 1$ bei Fitness-Test und $df = (\text{Rows} - 1) \times (\text{Cols} - 1)$ bei Unabhängigkeitstest.

Für einen χ^2 -Test wird die beobachtete Häufigkeit aller Zellen mit der erwarteten Häufigkeit verglichen, normalisiert und aufsummiert:

$$\sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \sim \chi_v^2$$

Die *erwarteten Häufigkeiten* e_{ij} berechnen durch Multiplikation von Zeilen- und Spaltentotal geteilt durch das Total. Berechnung basiert auf der Annahme *unabhängiger Variablen* und W'keit des gemeinsamen Auftretens: $P(X \cap Y) = P(X) \times P(Y)$

Für eine Tabelle mit R Zeilen und C Spalten ergibt sich die Berechnung des χ^2 -Werts durch Iteration über alle Zellen wie folgt:

$$\sum_{i=1}^R \sum_{j=1}^C \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \sim \chi_v^2$$

Berechnungsbeispiel χ^2

- H_0 : Parteiwahl unabhängig von Geschlecht.
- H_a : Es gibt einen Zusammenhang.

Erhobene Daten:

	Rep	Dem	Other	Total
male	26	13	5	44
female	20	29	7	56
Total	46	42	12	100

Erwartete Frequenzen (zB. für „Republican males“ = $\text{tot}_{\text{rep}} \times \text{tot}_{\text{male}} \div \text{tot}_{\text{all}} = 46 \times 44 \div 100 = 20.24$)

	Rep	Dem	Other	
male	20.24	18.48	5.28	
female	25.76	23.52	6.72	

$$\chi^2 = \frac{(26-20.24)^2}{20.24} + \dots + \frac{(7-6.72)^2}{6.72} = 5.886$$

$$df = (2 - 1) \times (3 - 1) = 2$$

$$\chi_{cv}^2 = 5.991 \text{ (aus Tabelle für } \alpha = 0.05 \text{ und } df=2)$$

Da $5.886 < 5.991$ können wir H_0 nicht verwerfen.

Mit R berechnet:

```
data <- matrix(c(26,13,...,7),nrow=2,byrow=T)
rownames(data) <- c("male", "female")
colnames(data) <- c("Rep", "Dem", "Other")
chisq.test(data)
X-squared = 5.8555, df = 2, p-value = 0.05352
```

Wir erhalten einen p-Wert > 0.05 , H_0 kann also nicht verworfen werden.

5.6.5 Cramér's V: Effektgrösse nach χ^2 -Test

Wird oft nach einem χ^2 -Test durchgeführt, um zu quantifizieren, wie stark der Zusammenhang zwischen den Variablen ist.

Berechnung $V = \sqrt{\frac{\chi^2}{n \times \min(R-1, C-1)}}$

Effektgrössen bei Minimum R-1, C-1:

- 1 \Rightarrow 0.10 (klein), 0.30 (mittel), 0.50 (gross)
- 2 \Rightarrow 0.07 (klein), 0.21 (mittel), 0.35 (gross)
- 3 \Rightarrow 0.06 (klein), 0.17 (mittel), 0.29 (gross)
- ... etc.

```
library(lsr); cramersV(table)
```

6 Modellierung

Alle Modelle sind falsch, aber einige sind nützlich.

— Box, 1976

Holistische Erklärungen versuchen die Welt ganzheitlich und umfassend zu erklären. **Modelle** sind dagegen vereinfachte und idealisierte Abbilder von Systemen, Vorgängen oder Situationen, die einen bestimmten Aspekt zu erklären versuchen.

Je einfacher ein Modell, desto attraktiver ist es aus epistemischer Sicht. **Occam's Razor**: von zwei möglichen Erklärungen ist die einfachere zu wählen.

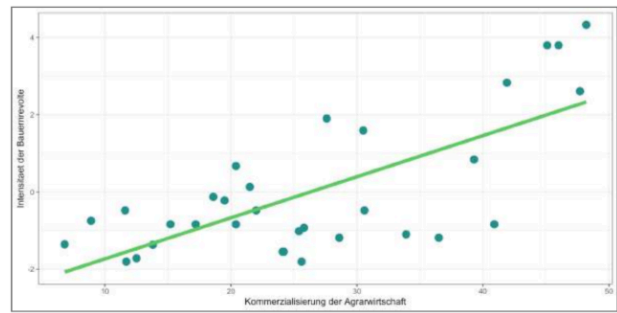
Verbale Modelle haben eine **abhängige Variable Y** und eine Menge von **Prädiktoren** $\{X_1, X_2, \dots\}$ (erklärenden Variablen).

Statistisches Modell Ein mathematisches Modell, das zu erfassen versucht, wie Daten entstanden sind. Bezieht sich auf **empirische Daten** und ist **stochastischer** (probabilistischer) Natur.

Lineare Regression ist ein Beispiel für ein statistisches Modell.

6.1 Regressionsanalyse

Zeigt wie Y von X beeinflusst wird.



Lineares Regressionsmodell: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- i ist eine Einheit auf x/y-Achse
- β_0 **Achsenabschnitt**: Wert von y bei x=0
- β_1 **Steigungskoeffizient**: Inkrement von y pro x
- ε **Fehlerterm**: Enthält alle Faktoren die Y beeinflussen aber nicht mit X ausgedrückt sind, \mathcal{N}

Die **Regressionsgerade** stellt die geschätzte lineare Beziehung zwischen den Variablen dar.

Berechnung einfaches Regressionsmodell mit R:

```
model <- lm(Y ~ X, data=table)
summary(model)
```

Berechnung multiple Regression mit R:

Zentrale Prädiktoren Indikatoren der gewählten Theorie, bzw. des gewählten Ansatzes

Kontrollvariablen Alles was sonst noch bekannt ist (aus der Literatur) auch noch einbeziehen, damit möglichst wenig für den Fehlerterm übrig bleibt.

Ceteri Paribus „Wenn alles andere gleich bleibt“: Annahme, dass sich die Steigung der Kurve ändert, wenn nur der entsprechende Prädiktor verändert wird. Ist in der Realität natürlich in der Regel nicht möglich und auch nicht praktikabel, für statistische Betrachtung aber ok.

```
model <- lm(I ~ C + G + T + M, peasant)
summary(model)
```

Output:

```
Call:
lm(formula = I ~ C + T + G + M, data = peasant)

Residuals:
    Min       1Q   Median       3Q      Max
-2.22890 -0.67181 -0.08113  0.82831  2.37274

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.97799    5.52625  -2.348  0.026428 *
C              0.09112    0.02034   4.481  0.000123 ***
T              0.11920    0.06089   1.958  0.060690 .
G              0.92075    2.86014   0.322  0.749990
M             -0.00415    0.01776  -0.234  0.816967
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Die absolute Grösse des *Estimate* ist irrelevant, da jede Variable auf einer anderen Skala gemessen ist. **Signifikant** ist einzig Prädiktor C.

7 Kausalität

Beweis von Kausalität herzustellen ist schwierig. Korrelation \neq Kausalität.

Scheinkausalität Wir glauben, dass D eine ausreichende Ursache für Y ist.

Aber sobald wir eine bestimmte dritte Variable X kontrolliert verändern, hört die Beziehung zwischen D und Y auf zu existieren. Wir nennen X einen **Confounder**.

7.1 Fallstudien, Systemvergleiche

Grundprinzip Forschungsdesign: Konstanten und Variablen können einander gegenseitig nicht erklären.

Das können wir ausnutzen, in dem wir versuchen soviel wie möglich konstant zu halten und nur einen oder wenige Parameter zu verändern.

MILL (1843), LIJPHART (1971), PRZEWORSKI & TEUNE (1982): Indem wir eine **gezielte Auswahl** von Fällen treffen, die zu einem grossen Teil identische Parameter haben, können wir die Unterschiede zu erklären versuchen.

Wenn:

- ABCD für wxyz verantwortlich
- AEFG für wtuv verantwortlich

Dann:

- muss A die Ursache von w sein (einzige Übereinstimmung)

Most Different Systems Design (MDSD) Beispiele mit möglichst grossen Unterschieden auswählen, ausser in einem Aspekt. Dieser Aspekt wird dann zu erklären versucht.

In der vergleichenden Politikwissenschaft z.B. Länder: Warum haben Australien und Brasilien (grosse Unterschiede) beide eine sehr hohe Wahlbeteiligung? Versuchen herauszufinden, was sonst noch gleich ist \Rightarrow Kandidat für Erklärung/Ursache.

Most Similar Systems Design (MSSD) Beispiele mit möglichst vielen gleichen Eigenschaften (konstante Variablen).

Die nicht-gleichen (d.h. variablen) Eigenschaften sind Kandidaten für Ursache eines Phänomens, das

in beiden Systemen mit unterschiedlichen Werten auftaucht.

Kontrollgruppe Für Kausalforschung braucht es immer auch eine Kontrollgruppe (in der die potentielle Ursache nicht verändert wird)

Keep in Mind:

- Eine Regressionsanalyse kann **nie** eine Kausalität erklären, sie kann nur eine signifikante Korrelation identifizieren.
- Mittel sind nie Ursachen: Thermometer/Barometer korrelieren mit Wetterveränderung, sind aber nur **Symptom**, nicht Ursache.