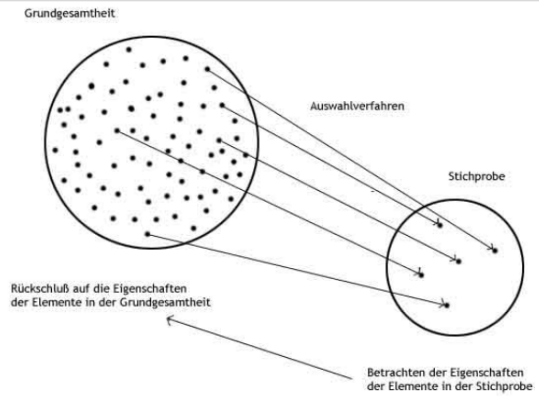
Die Grundgesamtheit

- Dazu gehören alle Personen, über die man (als Gesamtheit) Aussagen machen möchte: `räumlich`, `zeitlich`, `sachlich` eingegrenzt.
- Diese soll genau festgelegt werden. Idealerweise (wenn auch selten) ist eine vollständige Liste der Personen vorhanden.
- Beispiele:
  - Alle für die nächste Nationalratswahlen Wahlberechtigten in Österreich
  - Alle Absolventen der HTL Anichstraße ab einem bestimmten Jahrgang
- Ist es möglich und sinnvoll alle Personen der Grundgesamtheit zu befragen, spricht man von einer **Vollerhebung**
  - Meist nur bei kleinen Populationen möglich (Ausnahme: Befragungen im Rahmen von Volkszählungen)
  - Vorteil: Daten können beliebig weit aufgesplittet werden und über alle Teile können Aussagen gemacht werden. Alle statistischen Aussagen, die darauf basieren, können ohne Schwankungsbreite interpretiert werden
  - Nachteil: Meist sehr teuer bzw. unmöglich
- Wenn es nicht möglich oder unwirtschaftlich ist, alle Personen zu befragen: Dann verwendet man eine **Stichprobe**.

Die Stichprobe



- Ist eine Auswahl von Personen der Grundgesamtheit
- die Auswahl muss **repräsentativ** sein:
  - Jede Person muss die gleiche Chance haben, in die Stichprobe mitaufgenommen zu werden
  - Keine Person darf mehrmals teilnehmen/darin vorkommen
- Eine hohe Anzahl an Teilnehmern allein garantiert nicht die Repräsentativität:
  - Beispiel: Eine Onlinebefragung, wo jeder mehrmals teilnehmen kann ist **nicht** repräsentativ und läßt somit keine statistischen Aussagen über die Grundgesamtheit zu, auch wenn 5000 Teilnahmen verzeichnet sind.
- Diese Auswahl kann getroffen werden durch:
  - **Zufallsauswahl**: z.B. durch Auflistung aller Personen der Grundgesamtheit und ziehen von Zufallszahlen zur Ermittlung der Personen. Man erhält daraus eine Zufallsstichprobe.
  - **Quota-Verfahren**: Auswahl entsprechend der Verteilung in der Gesamtbevölkerung (z.B. Alter, Geschlecht, Wohnbezirk)
- Jeder Stichprobe ist mit einem Fehler behaftet:
  - **Standardfehler** : Das Verwenden der Stichprobe allein führt schon zu fehlerhaften Schätzungen der wahren Werte
  - **Systematische** Fehler: Wenn die Stichprobe falsch gezogen wurde, wird das Ergebnis verzerrt. Bsp.: Schüler der 4. und 5. Klassen sind bei einer Schülerbefragung in der Stichprobe stark überrepräsentiert.

Stichprobenberechnung

Bei einer sehr großen Grundgesamtheit (größer 100.000) kann man die vereinfachte Formel verwenden:

$$n = \frac{(t^2 * p * q)}{e^2}$$

wobei....

- **n** = Stichprobengröße
- **t** = Konfidenzstufe (t = 1 = 68,3% Sicherheit, t = 2 = 95,5% Sicherheit und t = 3 = 99,7% Sicherheit bzw. t= 1.96 für 95%)
  Sie gibt an mit welcher Wahrscheinlichkeit die Ergebnisse der Stichprobe die wahren Werte der Grundgesamtheit widerspiegeln.
- **p** = Erwartete Variabilität: Sie bezieht sich auf die Wahrscheinlichkeit eines bestimmten Merkmals in der Grundgesamtheit. Bei maximaler Unsicherheit (z. B. ein Ja/Nein-Merkmal mit einer Wahrscheinlichkeit von 50 % für beide Antworten) ist die benötigte Stichprobengröße am größten
- **q** = 1 - p
- **e** = Gewünschte Fehlermarge (zB +/- 5 %)

Bei kleinerer Grundgesamtheit wird diese (**N**) in der Formel berücksichtigt:

$$n = \frac{(t^2 * p * q * N)}{(t^2 * p * q + e^2 * (N-1))}$$

Was ist deskriptive Statistik?

- Erste Beschreibungen und einen Überblick über Daten werden gegeben
- Maßzahlen wie Mittelwert, Median, Streuung, ... werden ermittelt
- Grafische Darstellung der Daten wie Histogramme, Boxplot oder Streudiagramme
- Wichtig sind Häufigkeitsverteilungen (`Kontingenztabellen`): Tabellen die zeigen wie oft bestimmte Werte z.B. in einer Spalte bzw. Kategorie vorkommen
- Auch interessant um Ausreißer zu entdecken
- So können erste Plausibilität-Checks (Überprüfung ob die Daten Sinn machen) durchgeführt werden und ev. Fehler in den Daten erkannt werden.
- Als Vorstufe für tiefer gehende statistische Verfahren (inferenzielle Statistik):
  - Erste Trends werden erkannt, die dann statistisch untermauert werden können. diese Trends werden als Hypothesen formuliert.
  - Bestimmt auch ob bestimmte statistische Verfahren zulässig sind.

Beispiele:

- Darstellung der Notenverteilung in einer Klasse mittels Histogramm und Berechnung des Klassendurchschnitts.
- Analyse der monatlichen Umsätze eines Unternehmens, um den Durchschnittsumsatz zu ermitteln und Schwankungen aufzuzeigen.

Was ist inferenzielle Statistik?

- Hier wird versucht Schlussfolgerungen auf die Grundgesamtheit auf Basis der Stichprobe zu machen. Es werden die oben formulierten Hypothesen statistisch untermauert (oder auch nicht bei gegenteiligen Ergebnis).
- Es wird also gezeigt ob ein Zusammenhang oder Unterschied in den Variablen oder Gruppen des Datensatzes statistisch **signifikant** ist.
- Es werden auch bestimmte Punkte oder Intervalle in der Grundgesamtheit geschätzt (s. Regression).

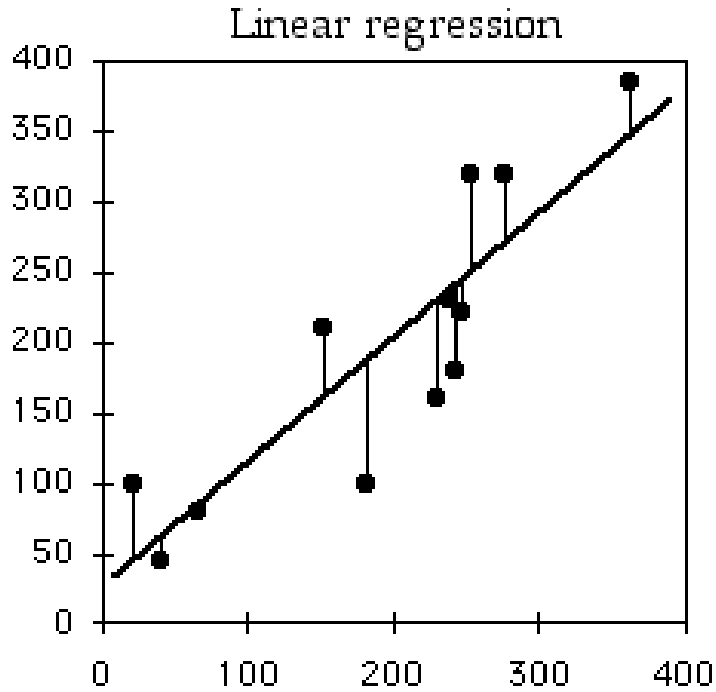
Beispiele:

- Ein neues Medikament wird an 500 Patienten getestet, um Rückschlüsse auf die gesamte Bevölkerung zu ziehen.
  Hypothese: "Das Medikament senkt den Blutdruck signifikant."
- Prüfung, ob Schülergruppen in einer neuen Unterrichtsmethode bessere Noten erzielen als mit der alten Methode

Auf den folgenden Folien werden einzelne inferenzstatistische Verfahren besprochen. Die Messniveaus bestimmen welche Verfahren überhaupt zulässig sind.

(lineare) Regression

- Hier wird versucht, eine abhängige Variable mit einer (oder mehreren) unabhängigen Variablen in Beziehung zu bringen
  - Beispiel: Unabhängige Variable: Jahreszahl. Abhängige Variable: Bevölkerung z.B. von Innsbruck.
  - Es wird nun versucht, die Bevölkerungszahl für bestimmte Jahre vorauszusagen
- Es wird hier versucht, eine gerade (lineare Funktion) durch die Punkte zu legen, so dass die quadrierten Abstände der Punkte zur geraden minimal wird.
- Mittels entwickeln einer linearen Funktion ist es dann möglich, Prognosen für die Zukunft zu erstellen (unter der Bedingung das sich die Entwicklung linear verhält (also z.B. daß das Bevölkerungswachstum linear ansteigt.))



(lineare) Regression in Python (1 /2)

- Es gibt mehrere Bibliotheken, die die lineare Regression unterstützen, `statsmodels` ist eigentlich der standard dazu (einfach mit `pip` installieren).
- Bauen es Modells:

```
1  import statsmodels.api as sm
2
3  # Als Eingang dient ein DataFrame. Falls dieser noch nicht in der für die
       Analyse notwendigen Form vorhanden ist, muss man einen erstellen
4  df_reg = pd.DataFrame({"years" : years, "mean_temp" : mean_temp})
5
6  # so wird das Modell gebaut. Links die abhängige, rechts die unabhängige
       Variable
7  model = sm.OLS.from_formula('mean_temp ~ years ', df_reg).fit()
8
9  a = model.params[1] # Die Werte für y = ax+b
10 b = model.params[0]
11
12 # R-quadrat (Wert) von 0-1: Zeigt wie gut die Gerade die Daten repräsentiert
13 rs = model.rsquared
14 fitted = model.fittedvalues # Die Werte auf der Geraden
15 # Abweichungen von Punkten auf der geraden zu den Datenpunkten
16 resid = model.resid
```

(lineare) Regression in Python (2 /2)

- Wie kann man die Vorhersagen machen?

```
# Auch hier bildet ein DataFrame die Basis
df_pred = pd.DataFrame({"years" : np.arange(2020,2040)})
# Wichtig ist dass die Spaltennamen der unabhängigen Variablen denen im Modell
    entsprechen
predictions = model.predict(df_pred)
# Die Predictions sind dann eine einfache Liste, die die Vorhersagen (hief für
    die einzelnen Jahre) beinhalten.
```

- Visualisierung der Ergebnisse.
  - Man kann mehrere Plots auf eine Zeichnung geben, indem man erst am Ende `plt.show()` angibt.
  - In diesem Fall am Besten die Werte und dann die Regressiongserade
  - Für den Blick in die Zukunft muss man den Zeichenbereich erweiteren mit z.B.: `plt.xlim([1980, 2040])`

```
plt.plot(val.years, predictions) # Die Vorhersage
plt.plot(df_reg.years, df_reg.mean_temp) # Die Werte selbst
plt.xlim([1980, 2040])
plt.show() # Erst jetzt show aufrufen, dann wird alles auf eine Grafik
    gezeichnet
```

Kontingenztabelle

- Entsteht durch die Verknüpung von 2 (oder mehr) Merkmalen in einer Tabelle
- Es werden die Häufigkeiten der kombinierten Merkmale angegeben

Beispiel: Kreuzung der Merkmale

- Haben sich während Ihrer Ausbildung Firmen mit Jobangeboten bei Ihnen gemeldet? (Eigentlich sehr unpräzise gefragt)
- Haben Sie studiert/studieren Sie?

	Jobangeboten gemeldet		
Studiert	Ja	Nein	Row Total
FH	12	16	28
	7.389	20.611	
UNI	2	22	24
	6.333	17.667	
Nicht	5	15	20
	5.278	14.722	
Column Total	19	53	72

wobei...

- ...in der ersten Zeile der Zelle der beobachtete Werte
- ...in der zweiten Zeile der Zelle der erwartete Wert steht: $\frac{RowTotal * ColumnTotal}{SumTotal}$, z.B. $\frac{28*19}{72} = 7.389$
- ...Pro Zeile und Spalte die Randsummen stehen
- ...Rechts unten und die Summe der Randsummen stehen (`Sum Total`)

X^2 - Unabhängigkeitstest

- Der Unabhängigkeitstest ist ein Signifikanztest auf Unabhängigkeit in der Kontingenztafel
- Man betrachtet zwei statistische Merkmale X und Y, die beliebig skaliert sein können. Man interessiert sich dafür, ob die Merkmale stochastisch unabhängig sind.
- Es wird die Nullhypothese H_0: Die Merkmale X und Y sind stochastisch unabhängig aufgestellt.
- Berechnung: $X^2 = \sum_{j=1}^{m} \sum_{k=1}^{r} \frac{(n_{jk} - n*_{jk})^2}{n*_{jk}}$ mit (m-1)(r-1) Freiheitsgraden (d.f.)
  wobei...
  - ...n_{jk} die beobachteten und
  - ...$n*_{jk}$ die erwarteten Werte sind.
- Der ermittelte X^2 - Wert wird in einer Tabelle nachgeschlagen, um zu sehen ob das Ergebnis signifikant ist (d.h. es besteht ein Zusammenhang zwischen den Merkmalen). Wenn $\mathbf{p < 0.05}$ dann signifikant, wenn $\mathbf{p < 0.01}$ dann hochsignifikant

Beispiel:

	Jobangeboten gemeldet		
Studiert	Ja	Nein	Row Total
FH	12	16	28
	7.389	20.611	
UNI	2	22	24
	6.333	17.667	
Nicht	5	15	20
	5.278	14.722	
Column Total	19	53	72

```
Statistics for All Table Factors
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  7.956873     d.f. =  2     p =  0.01871487
```

X^2 -Umsetzung in Python/Pandas

```
# Wie bei den meisten Verfahren wird scipy verwendet
from scipy.stats import chi2_contingency

df = pd.read_csv('data/student-mat.csv', sep=";")

# als Basis für den chi^2 werden Kontingenztabellen verwendet:
ct_internet_higher = pd.crosstab(df['internet'], df['higher'])

chi, p, dof, expected = chi2_contingency(ct_internet_higher)
print ("Chi:",chi)
print ("p: %.5f" % p) # < 0.05 : Signifikanter Unterschied in den Gruppen
print ("dof" ,dof)
print ("expected",expected)


#Zum kopieren ist das Beispiel auch im github-Projekt:
# Es sollen die beobachteten als auch die erwarteten Werte (expected)
   dargestellt werden
sns.heatmap(ct_internet_higher, annot=False, cmap="YlGnBu")
sns.heatmap(ct_internet_higher, annot=ct_internet_higher,
   annot_kws={'va':'bottom'}, fmt="", cbar=False , cmap="YlGnBu")
sns.heatmap(ct_internet_higher, annot=expected, annot_kws={'va':'top'},
   fmt=".2f", cbar=False, cmap="YlGnBu")
plt.show()
```

Korrelation

- Eine Korrelation misst die Stärke einer statistischen Beziehung von zwei Variablen **A** und **B** zueinander.
- Der Korrelationskoeffizient bewegt sich zwischen
  - 1: Ja mehr **A**, desto mehr **B** und umgekehrt (direkter Zusammenhang)
  - -1 Ja mehr **A**, desto weniger **B** und umgekehrt (negativer Zusammenhang)
  - 0: Es besteht kein Zusammenhang zwischen den Variablen
- Wann welcher korrelationskoeffizient? (Parameter `method`):
  - `pearson`: Mindestvoraussetzung: Intervallskala
  - `spearman`: Mindestvoraussetzung: Ordinalskala.
    - Ist einfacher zu berechnen
  - `kendall`: Mindestvoraussetzung: Ordinalskala
    - Für kleinere Stichproben robuster
- Befehl dazu:

```
1 # corr ist sogar auf's Dataframe definiert
2 c = df_corr.corr(method ='spearman') #oder pearson
3 print(c) # beinhaltet nur den Korrelationskoeffizienten
```

- Möchte man auch die p-Werte haben muss man `scipy` verwenden:

```
1 from scipy.stats import spearmanr, pearsonr
2 # jetzt kommt auch der p-Wert und ein Korrelationskoeffizient allgemein
3 corr, p = spearmanr(df_corr['G3'], df_corr['Walc'])
4 print("corr: %.6f" % corr)
5 print("p-value: %.6f" % p)
```

- Visualisierung der Werte am Besten wieder mit einem Heatplot:

```
1 sns.heatmap(df_corr, annot=True, cmap="YlGnBu")
```