Anova – two way

Peter Büchel

HSLU I

Stat: Block 10

Anova – two way

- Einfache Varianzanalyse (Anova): Zielgrösse Y hängt von *einem* Faktor ab
- Zweifache Anova: Y hängt von *zwei* Faktoren ab
- Überlegungen bei zweifacher Anova sind sehr ähnlich wie im einfachen Fall
- Details sind allerdings erheblich
- Hier rein graphisch
- p-Wert mit `Python` berechnen
- Vorgehen mit zwei künstlichen Beispielen

Beispiel

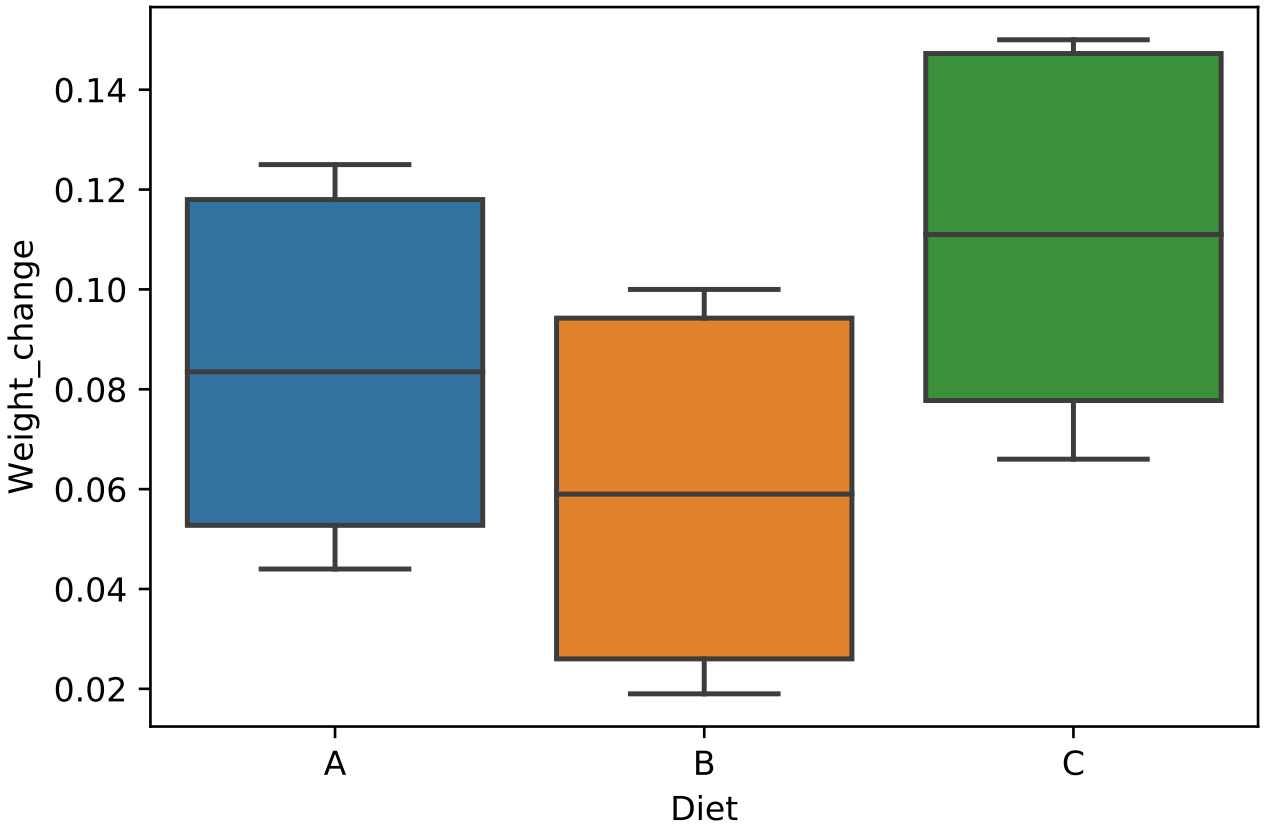
- Experiment: Personen machen eine von drei Diäten, um Gewicht zu verlieren
- Gewichtsverlust: Abhängige Variable
- Unabhängigen Variablen: Drei verschiedenen Diäten und Land, indem die entsprechenden Personen leben
- Daten:

```
import pandas as pd
df = pd.read_csv("../../Data/weight.csv")

df.head()
```

```
##    Unnamed: 0 Diet Country  Weight_change
## 0           0    A     USA          0.120
## 1           1    A     USA          0.125
## 2           2    A     USA          0.112
## 3           3    A      UK          0.052
## 4           4    A      UK          0.055
```

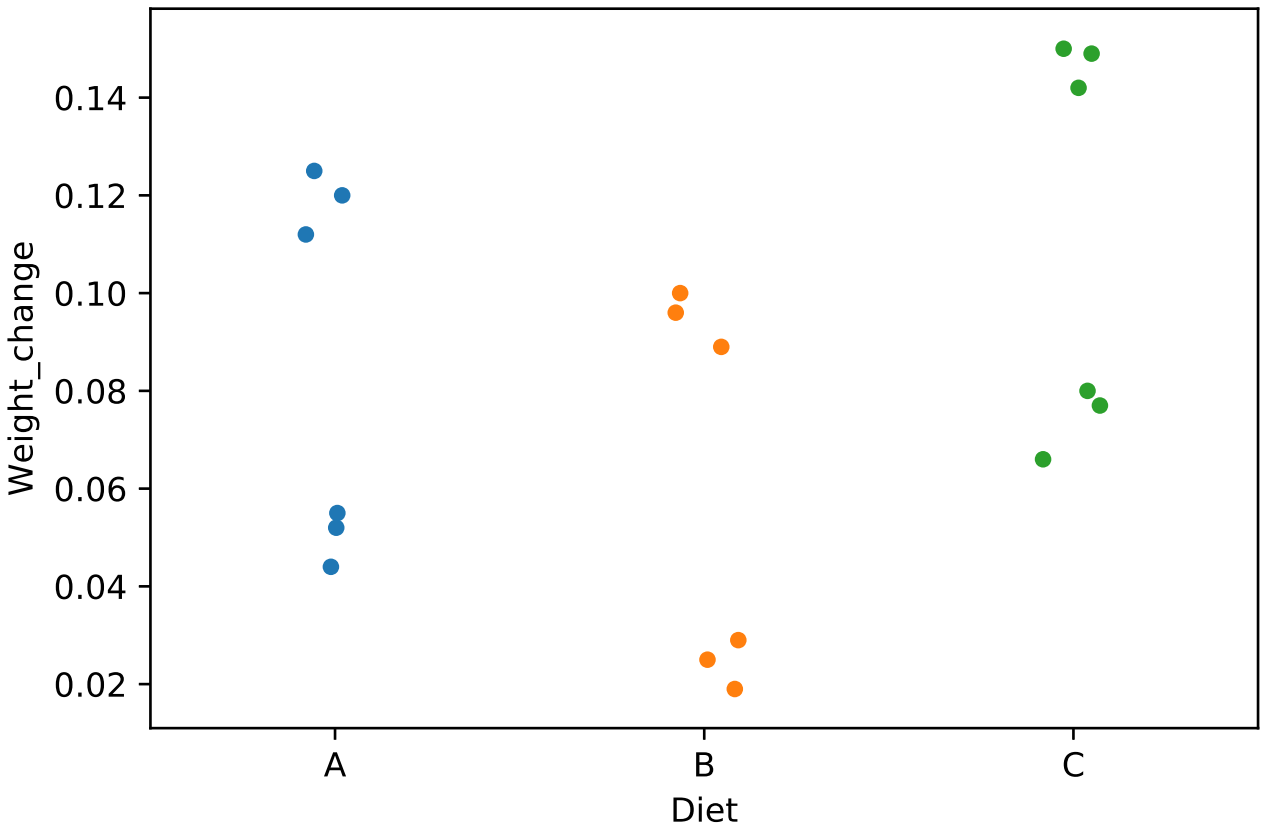
- Gewichtsverlust kann von Diät abhängen oder vom Land in dem die Personen leben
- Boxplot:

```
import seaborn as sns

sns.boxplot(x="Diet", y="Weight_change", data=df)
```



- Stripchart:

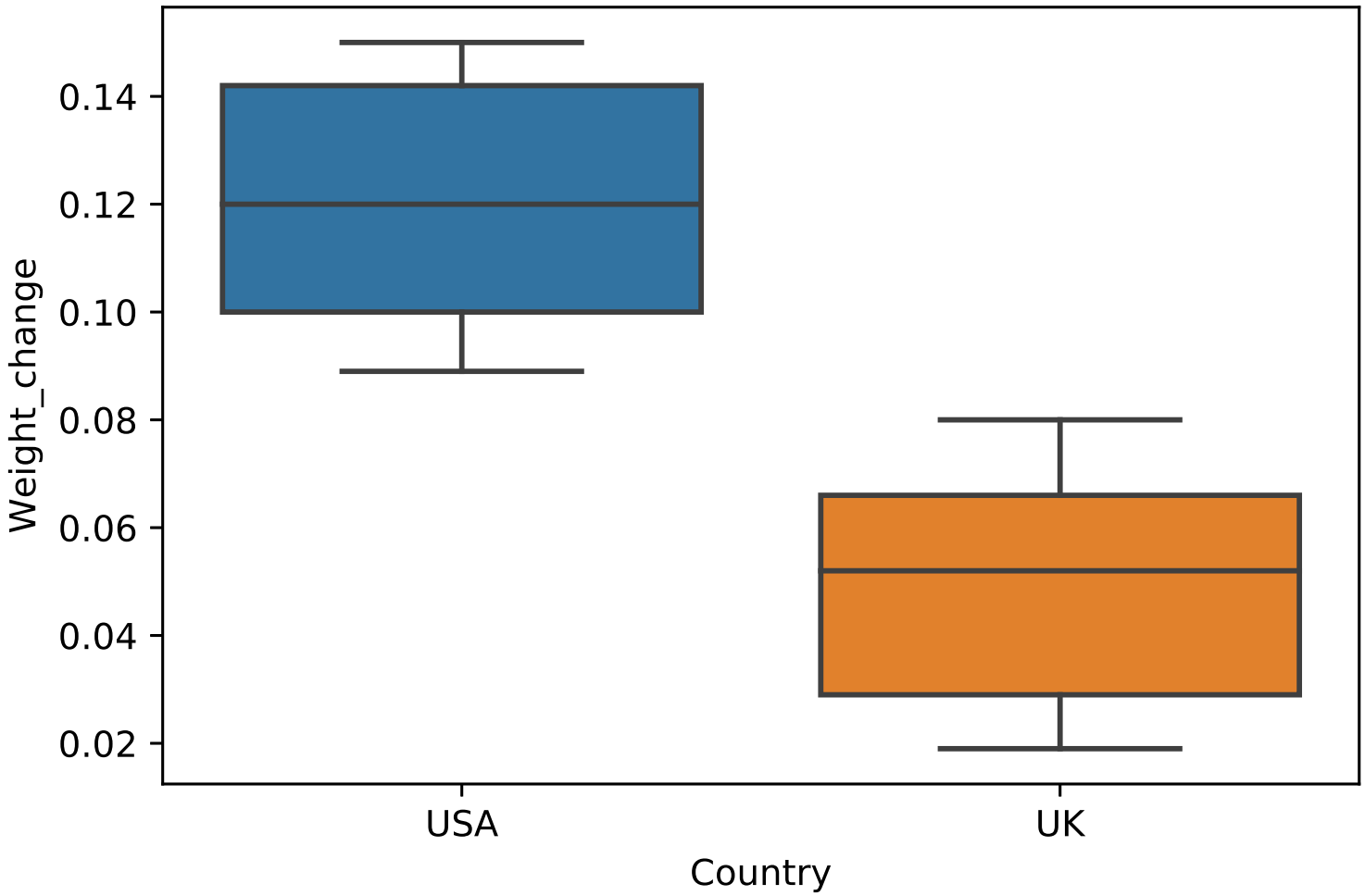
```
sns.stripplot(x="Diet", y="Weight_change", data=df)
```



- Frage: Liegen Gruppenmittelwerte wesentlich auseinander?
- Vorgehen wie bei einfacher Anova
- Aber noch ein Faktor vorhanden (Land)
- p-Werte gemeinsam berechnen

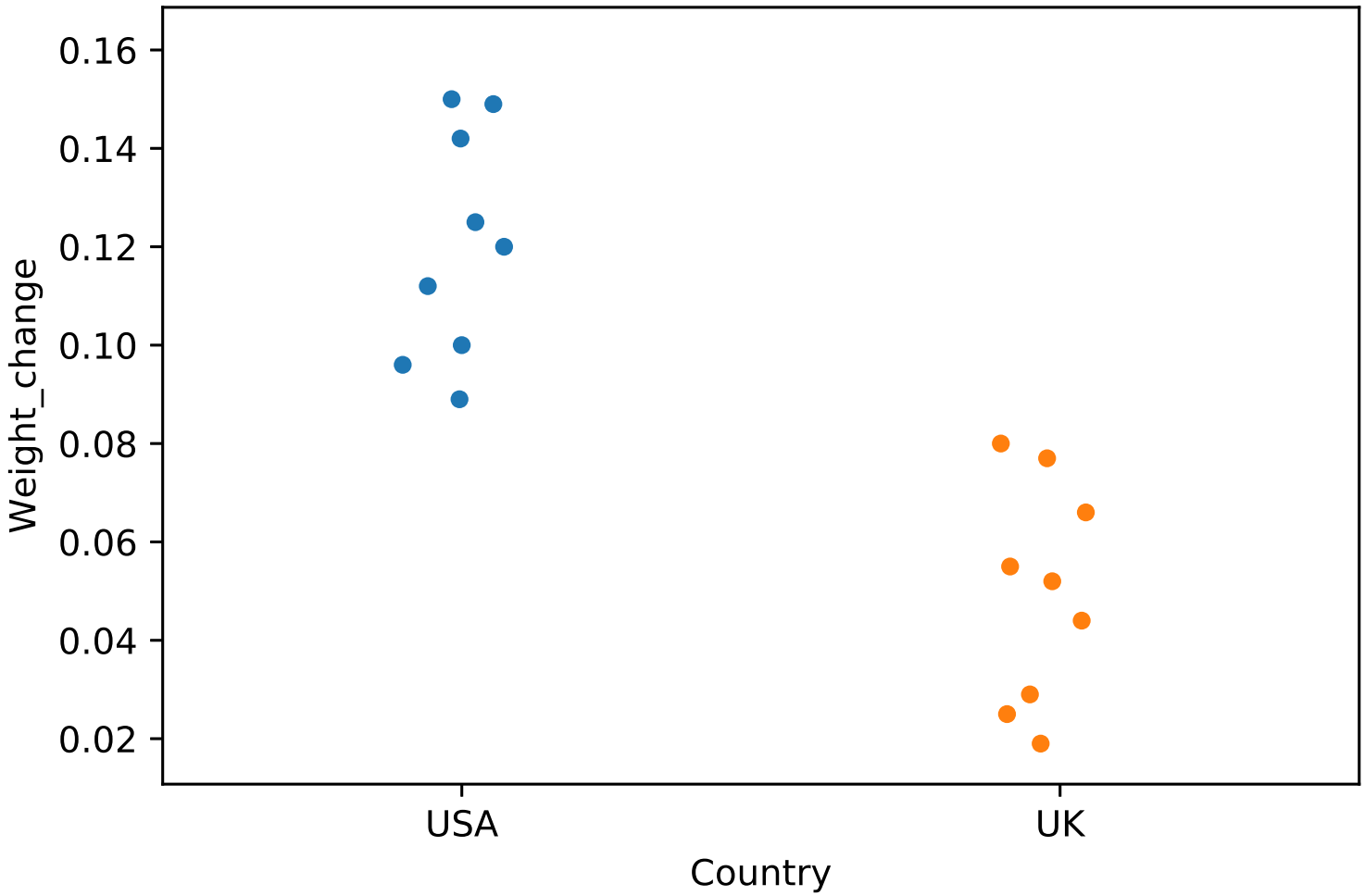
- Boxplot von Land:

```
sns.boxplot(x="Country", y="Weight_change", data=df)
```



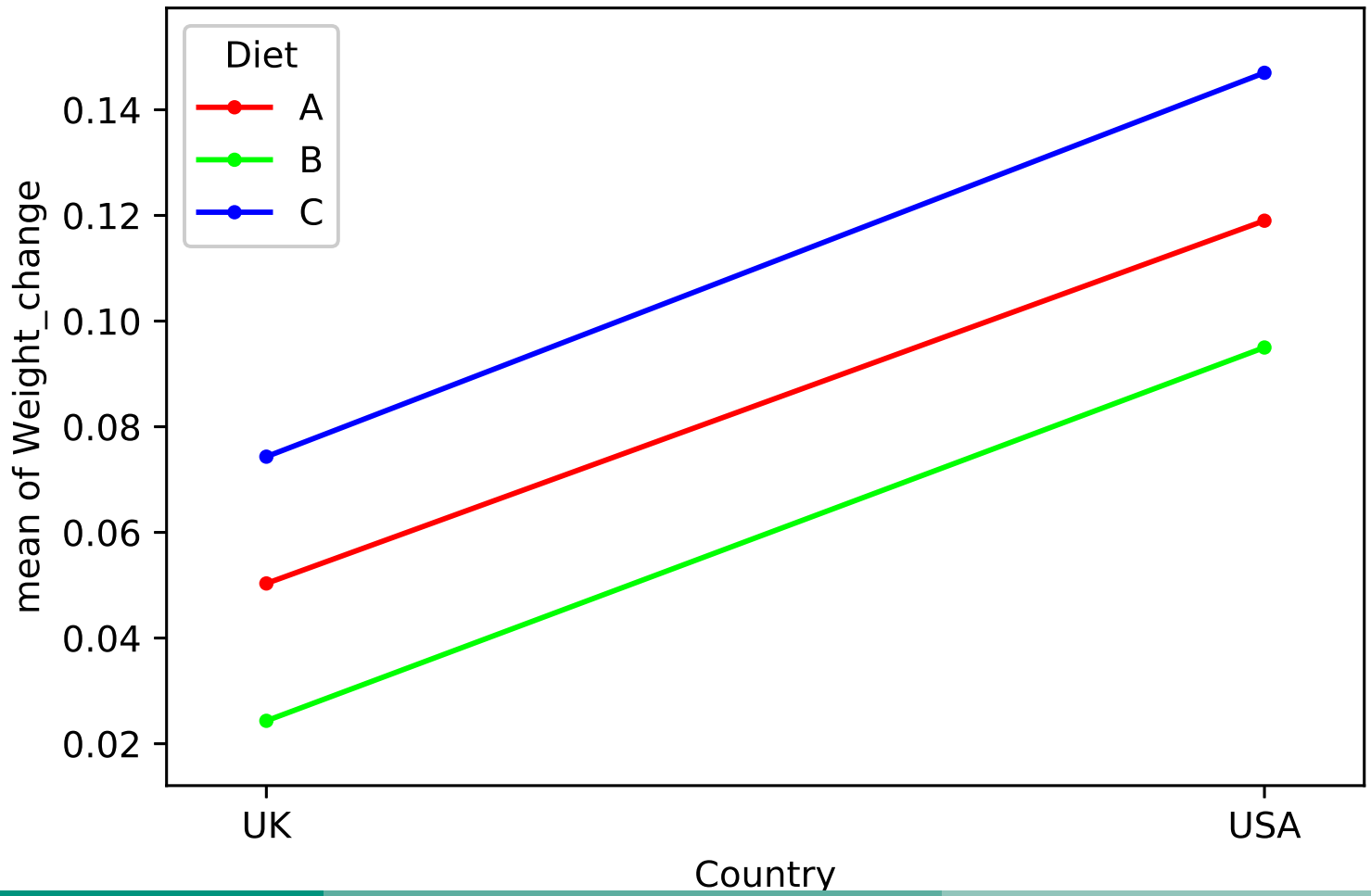
- Stripchart:

```
sns.stripplot(x="Country", y="Weight_change", data=df)
```



- Für zweifache Anova: `interaction.plot`:

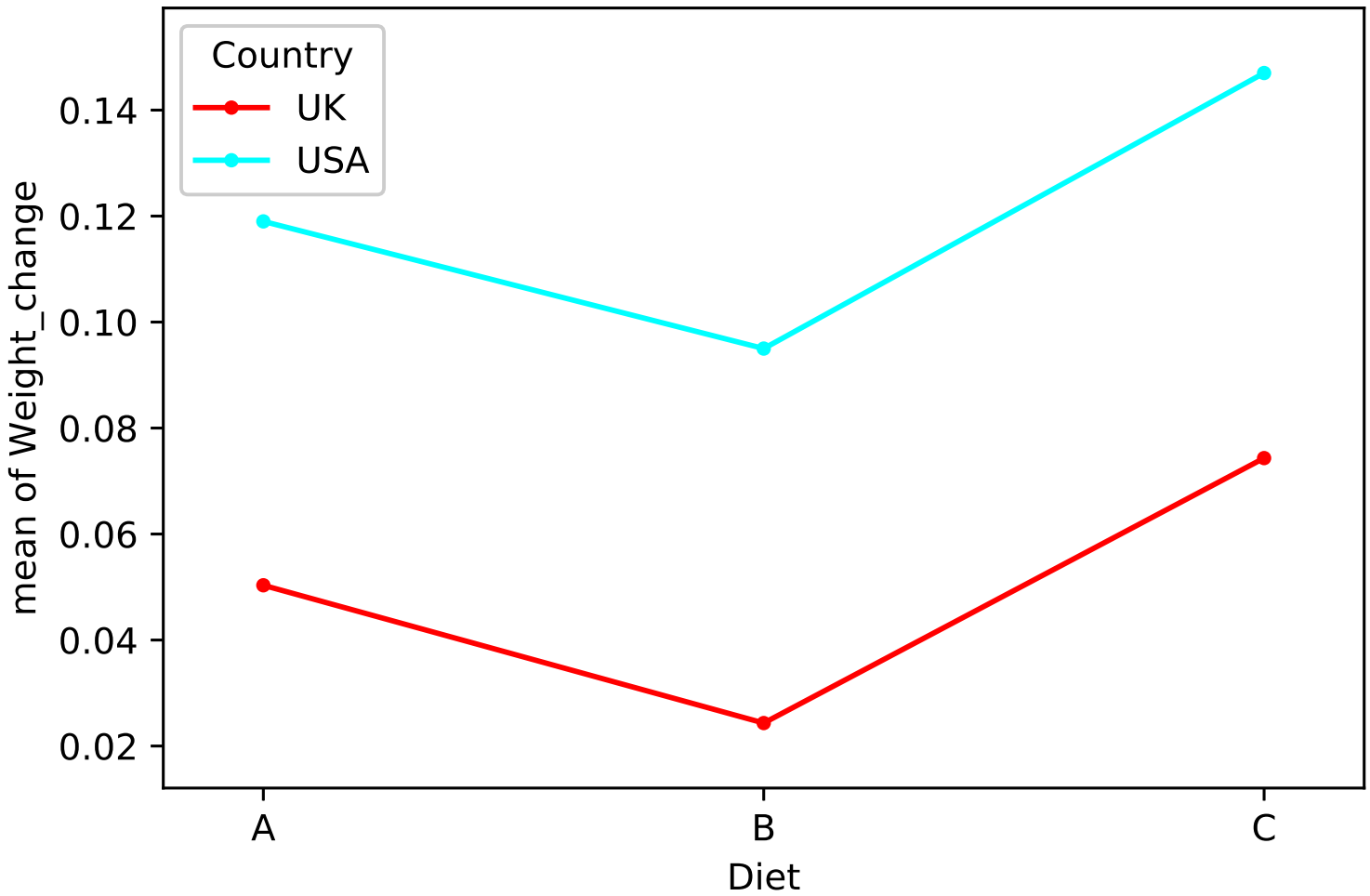
```
from statsmodels.graphics.factorplots import interaction_plot

interaction_plot(x=df["Country"], trace=df["Diet"],
response=df["Weight_change"])
```



- Punkte: Gruppenmittelwerte
- Punkt links unten: Gruppenmittelwert der Diät B, der in UK lebenden Personen
- Alle drei Diäten: Gruppenmittelwerte für den Gewichtsverlust in USA grösser als in UK
- Sind sie aber statistisch signifikant grösser?
- Keine Unterschied.Geraden parallel zur horizontalen Achse
- Die drei Diäten liegen auseinander
- Liegen sie aber statistisch signifikant auseinander?
- Alle Diäten gleich wirksam.Die drei Geraden liegen aufeinander

- **interaction.plot**:

```
interaction_plot(x=df["Diet"], trace=df["Country"],
response=df["Weight_change"])
```



- Punkte: Gruppenmittelwerte
- Punkt links unten: Gruppenmittelwert der Diät A, der in UK lebenden Personen
- Bei den beiden Länder liegen „weit" auseinander
- Aber liegen sie statistisch signifikant auseinander?
- Wenn die Diäten in beiden Ländern gleich wirksam wären, dann würde die beiden Linien aufeinanderliegen
- In beiden Ländern sind die Gruppenmittelwerte unterschiedlich
- Sind sie aber statistisch signifikant unterschiedlich?
- Gäbe es keinen Unterschied in den Gruppenmittelwerten, so müssten diese auf einer zur horizontalen Achse parallelen Gerade liegen

- Die letzten beiden Fragen mit Hypothesentest beantworten
- Nullhypothese jeweils
  - Die drei Diäten haben alle denselben Einfluss auf den Gewichtsverlust
  - In beiden Ländern gibt es denselben Gewichtsverlust
- Machen Hypothesentest mit R auf Signifikanzniveau von 5 %:

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from patsy.contrasts import Sum

fit = ols("Weight_change~Diet+Country", data=df).fit()
anova_lm(fit)
```

```
##                df     sum_sq   mean_sq           F        PR(>F)
## Diet          2.0   0.007804  0.003902  129.450237  9.351176e-10
## Country       1.0   0.022472  0.022472  745.516588  1.526843e-13
## Residual     14.0   0.000422  0.000030         NaN           NaN
```

- Beide Hypothesen: p-Wert weit unter dem Signifikanzniveau von 0.05
- Die Nullhypothesen werden also verworfen:
  - Die Diäten sind statistisch signifikant unterschiedlich wirksam
  - In beiden Ländern ist der Gewichtsverlust statistisch signifikant unterschiedlich

Beispiel

- Ist die Ausbeute von guten Cookies abhängig von der Backtemperatur und von der Zeit im Ofen?

- Tabelle: Resultate von 8 Bachblechen aufgeführt (`yield` in %):

```
import pandas as pd
df = pd.read_csv("../../Data/cookie.csv")
df = df.drop(["Unnamed: 0"], axis=1)
df
```

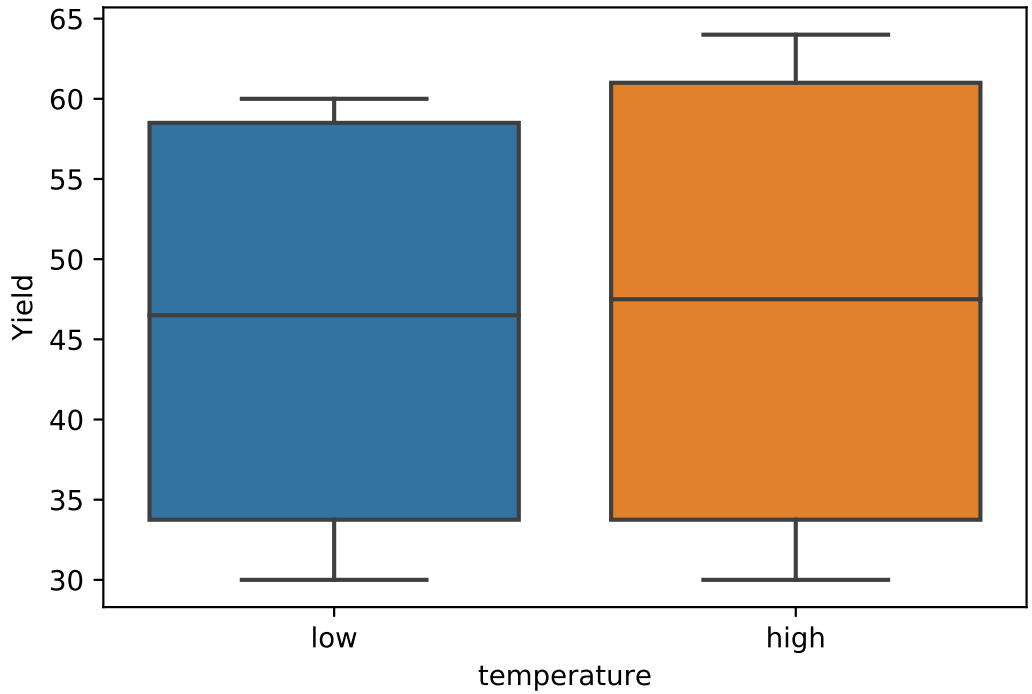
```
##   temperature   time  Yield
## 0         low  short     30
## 1         low  short     35
## 2         low   long     60
## 3         low   long     58
## 4        high  short     60
## 5        high  short     64
## 6        high   long     30
## 7        high   long     35
```

- Bei tiefer Temperatur und kurzer Backzeit ist die Ausbeute schlecht, da die Cookies noch nicht fertig gebacken sind

- Auf der anderen Seite sind bei hoher Temperatur und langer Backzeit die Cookies verbrannt

- Boxplot Temperatur:

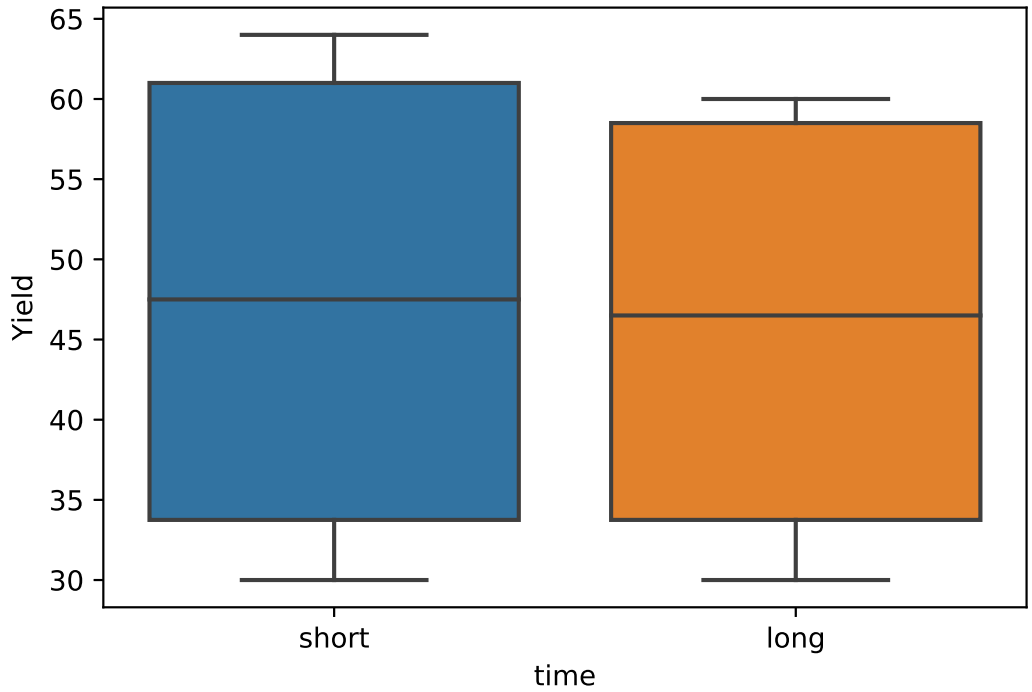
```
sns.boxplot(x="temperature", y="Yield", data=df)
```



- Temperatur hat scheinbar keinen Einfluss auf die Qualität der Cookies
- Komisch

- Boxplot Backzeit:

```
sns.boxplot(x="time", y="Yield", data=df)
```



- Auch hier kaum ein Unterschied
- Seltsam

- Hypothesentest:

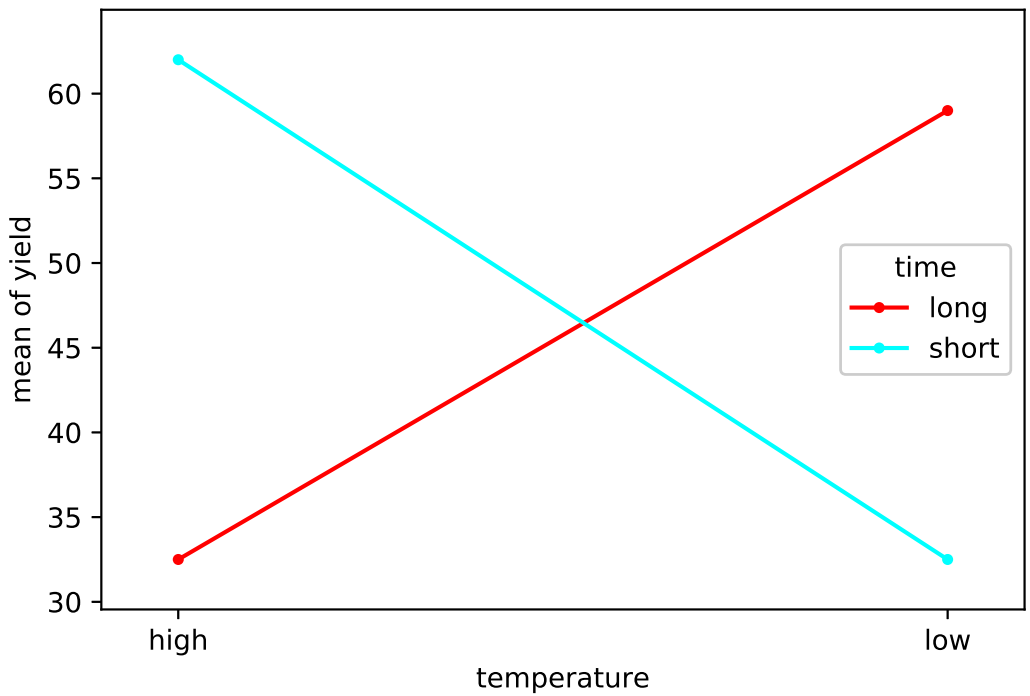
```
fit = ols('Yield~time+temperature', data=df).fit()
anova_lm(fit)
```

```
##                df  sum_sq  mean_sq         F    PR(>F)
## time          1.0     4.5      4.5  0.014036  0.910304
## temperature   1.0     4.5      4.5  0.014036  0.910304
## Residual      5.0  1603.0    320.6       NaN       NaN
```

- Beide p-Werte sind 0.91, also weit über dem Signifikanzniveau von 0.05
- Das heisst, weder Temperatur noch Backzeit haben Einfluss auf Qualität der Cookies
- Das kann ja sicher nicht stimmen

- Interaction-Plot:

```
interaction_plot(x=df["temperature"], trace=df["time"],
response=df["yield"])
```



- Bei hoher Temperatur und kurzer Backzeit oder niedriger Temperatur und langer Backzeit die besten Resultate
- Die beiden Grössen sind als nicht unabhängig voneinander bezüglich der Qualität der Cookies
- Man sagt: Die beiden Faktoren zeigen *Wechselwirkung* (Interaction)
- Wenn die Linien im Interaction-Plot nicht parallel sind, so tritt Wechselwirkung auf

- Hypothesentest für Wechselwirkung
- Nullhypothese: Es kommt keine Wechselwirkung vor
- Mit R p-Wert berechnen (mit einem $*$ statt einem $+$):

```
fit = ols('Yield~time*temperature', data=df).fit()
anova_lm(fit)
```

```
##                      df  sum_sq  mean_sq           F    PR(>F)
## time                1.0     4.5     4.50    0.514286  0.512937
## temperature         1.0     4.5     4.50    0.514286  0.512937
## time:temperature    1.0  1568.0  1568.00  179.200000  0.000180
## Residual            4.0    35.0     8.75         NaN       NaN
```

- p-Wert für die Wechselwirkung `temperature:time` ist 0.00018
- Nullhypothese wird verworfen und es liegt Wechselwirkung vor
- p-Werte für `temperature` und `time` sind hier tiefer (0.51) als weiter oben
- Hier wurden zur Bestimmung des F-Wertes und damit des p-Wertes weitere Grössen verwendet
- p-Werte sagen hier aus, wie stark sie die Zielgrösse Y beeinflussen (mehr dazu bei der Regressionsrechnung)
- Somit hat Temperatur und Backzeit *je alleine* keinen Einfluss auf Qualität der Cookies
- Erst die Wechselwirkung macht die Qualität der Cookies aus