Musterlösungen zu Serie 10

Lösung 10.1

a) Die Daten werden wie folgt als Data Frame in **Python** eingelesen : (zu **R**)

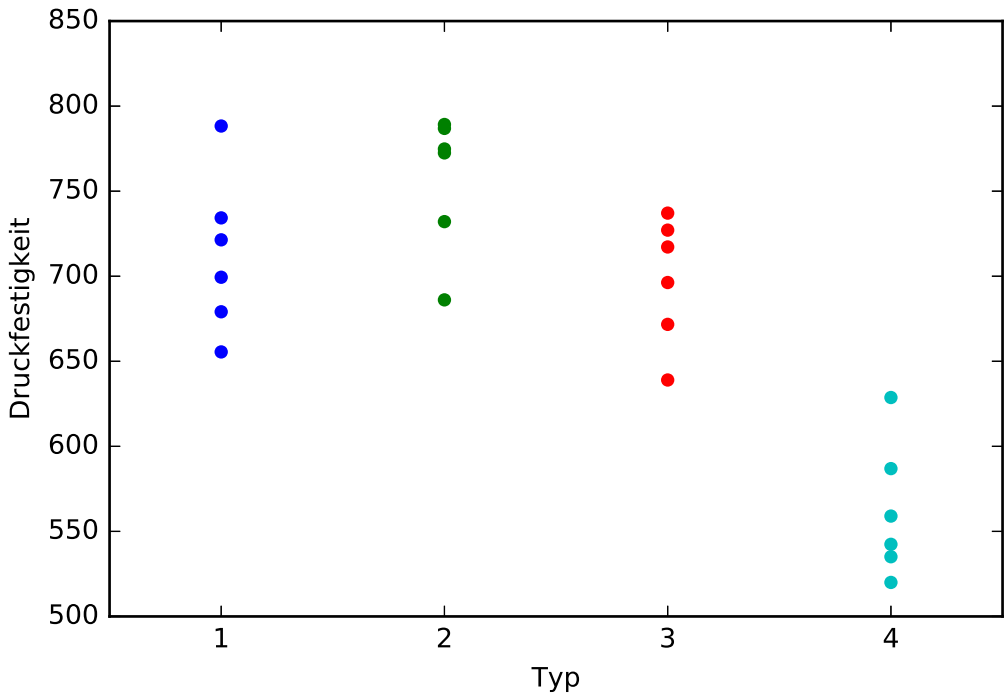
```
df=DataFrame({
  "Typ": np.repeat(["T1", "T2", "T3", "T4"], [6, 6, 6, 6]),
  "Druckfestigkeit" : [655.5, 788.3, 734.3, 721.4, 679.1, 699.4,
                       789.2, 772.5, 786.9, 686.1, 732.1, 774.8,
                       737.1, 639.0, 696.3, 671.7, 717.2, 727.1,
                       535.1, 628.7, 542.4, 559.0, 586.9, 520.0]
})
print(df)
```

```
##    Typ  Druckfestigkeit
## 0   T1            655.5
## 1   T1            788.3
## 2   T1            734.3
## 3   T1            721.4
## 4   T1            679.1
## 5   T1            699.4
## 6   T2            789.2
## 7   T2            772.5
## 8   T2            786.9
## 9   T2            686.1
## 10  T2            732.1
## 11  T2            774.8
## 12  T3            737.1
## 13  T3            639.0
## 14  T3            696.3
## 15  T3            671.7
## 16  T3            717.2
## 17  T3            727.1
## 18  T4            535.1
## 19  T4            628.7
## 20  T4            542.4
## 21  T4            559.0
## 22  T4            586.9
## 23  T4            520.0
```

Wir erzeugen eine Stripchart Graphik wie folgt:

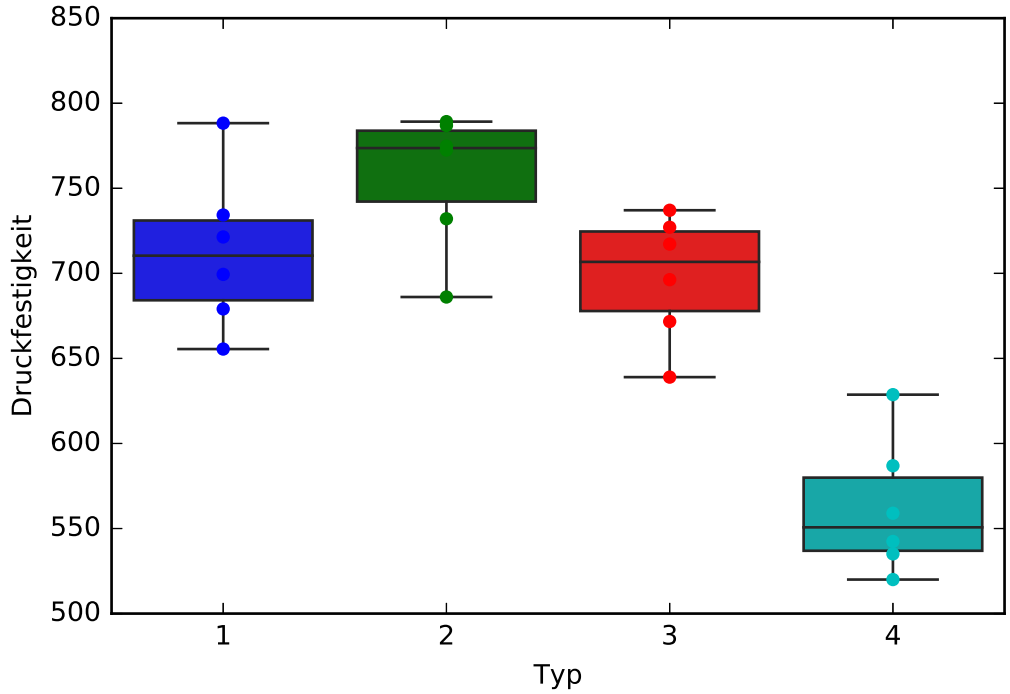
```
sns.stripplot(x="Typ", y="Druckfestigkeit", data=df)
plt.xlabel("Typ")
plt.ylabel("Druckfestigkeit")
```

```
plt.show()
```



Die entsprechenden Boxplots sind:

```
sns.boxplot(x="Typ", y="Druckfestigkeit", data=df)
plt.xlabel("Typ")
plt.ylabel("Druckfestigkeit")
plt.show()
```



Man sieht deutliche Unterschiede in der Lage der vier Stichproben. Vor allem die Stichprobe für den Typ 4 hat deutlich tiefere Werte als die drei anderen. Bezüglich der Streuung sind sich alle in etwa gleich (d.h. die Boxhöhe ist bei allen etwa gleich). Stichprobe zu Typ 2 ist linksschief, während Stichprobe zu Typ 4 rechtsschief ist.

b) Ein Gruppenmittelmodell lautet :

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

wobei μ einen globalen Parameter bezeichnet, den alle Gruppen miteinander teilen, τ_i bezeichnet die behandlungsspezifische Abweichung vom globalen Parameter μ und ε_{ij} ist der Fehlerterm. Wir wählen die Parametrisierung $\mu = \mu_1$, d.h., die behandlungsspezifische Abweichung τ_1 ist 0. Mit **Python** ergibt sich dann folgende Parameterschätzung für diese Modell: (zu **R**)

```
df=DataFrame({
  "Typ": np.repeat(["T1", "T2", "T3", "T4"], [6, 6, 6, 6]),
  "Druckfestigkeit" : [655.5, 788.3, 734.3, 721.4, 679.1, 699.4,
                       789.2, 772.5, 786.9, 686.1, 732.1, 774.8,
                       737.1, 639.0, 696.3, 671.7, 717.2, 727.1,
```

```
                        535.1, 628.7, 542.4, 559.0, 586.9, 520.0]
})
fit = ols("Druckfestigkeit~Typ", data=df).fit()

fit.params


## Intercept    713.000000
## Typ[T.T2]     43.933333
## Typ[T.T3]    -14.933333
## Typ[T.T4]   -150.983333
## dtype: float64
```

Die behandlungsspezifischen Abweichung lauten somit

$$\tau_{\text{TypT1}} = 0 \quad \tau_{\text{TypT2}} = 43.93333 \quad \tau_{\text{TypT3}} = -14.93333 \quad \tau_{\text{TypT4}} = -150.98333$$

c) Die Null-Hypothese lautet, dass sich die Typen nicht unterscheiden, also dass die Gruppenmittelwerte

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

sind, oder die Behandlungseffekte (eng. *treatment effects*)

$$\tau_2 = \tau_3 = \tau_4 = 0$$

Die Alternative besagt, dass sich mindestens ein Gruppenpaar i und j im Gruppenmittelwert unterscheidet, d.h., $\mu_i \neq \mu_j$. (zu **R**)

```
df=DataFrame({
  "Typ": np.repeat(["T1", "T2", "T3", "T4"], [6, 6, 6, 6]),
  "Druckfestigkeit" : [655.5, 788.3, 734.3, 721.4, 679.1, 699.4,
                       789.2, 772.5, 786.9, 686.1, 732.1, 774.8,
                       737.1, 639.0, 696.3, 671.7, 717.2, 727.1,
                       535.1, 628.7, 542.4, 559.0, 586.9, 520.0]
})
fit = ols("Druckfestigkeit~Typ", data=df).fit()

anova_lm(fit)

##              df         sum_sq      ...                F        PR(>F)
## Typ         3.0  127374.754583      ...       25.094289  5.525450e-07
## Residual   20.0   33838.975000      ...              NaN           NaN
##
## [2 rows x 5 columns]
```

Da die transformierte Teststatistik einen P-Wert von $5.5e-7$ hat und somit kleiner als das Niveau 5 % ist, wird die Nullhypothese verworfen und es gilt die Alternative. Dies ist ja schon ersichtlich aus dem Boxplot: Typ 4 unterscheidet sich wesentlich von den anderen drei Typen; evt. auch Typ 2 von den Typen 1 und 3.

Lösung 10.2

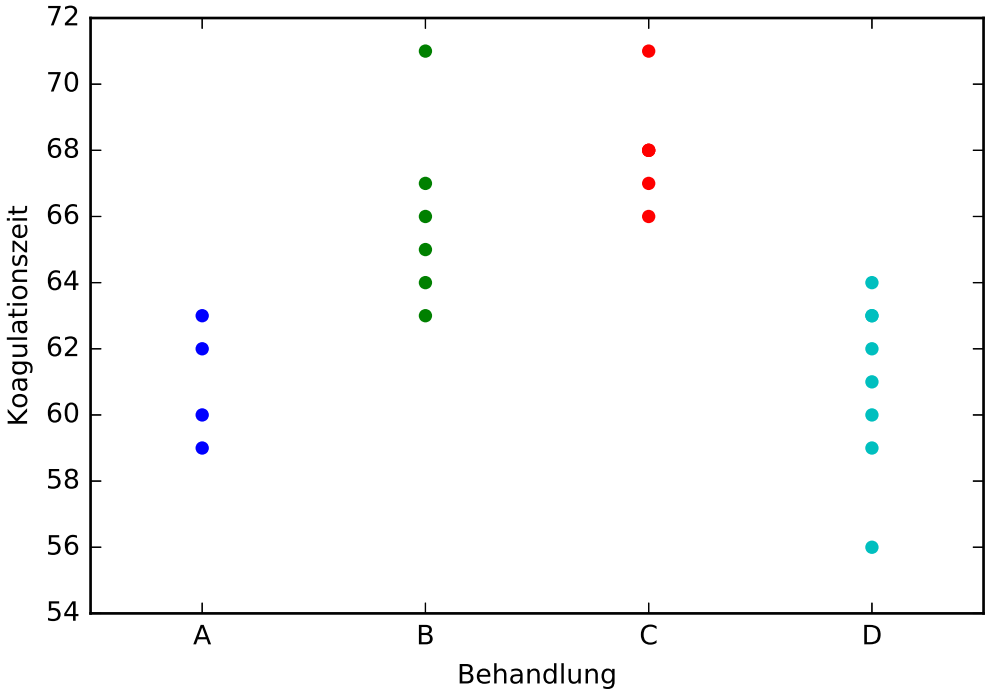
a) Die Daten werden wie folgt als Data Frame in **Python** eingelesen : (zu **R**)

```
df=DataFrame({
  "Behandlung": np.repeat(["A", "B", "C", "D"], [4, 6, 6, 8]),
  "Koagulationszeit" : [62, 60, 63, 59, 63, 67,
                        71, 64, 65, 66, 68, 66,
                        71, 67, 68, 68, 56, 62,
                        60, 61, 63, 64, 63, 59]
})
print(df)
```

```
##    Behandlung  Koagulationszeit
## 0           A                62
## 1           A                60
## 2           A                63
## 3           A                59
## 4           B                63
## 5           B                67
## 6           B                71
## 7           B                64
## 8           B                65
## 9           B                66
## 10          C                68
## 11          C                66
## 12          C                71
## 13          C                67
## 14          C                68
## 15          C                68
## 16          D                56
## 17          D                62
## 18          D                60
## 19          D                61
## 20          D                63
## 21          D                64
## 22          D                63
## 23          D                59
```

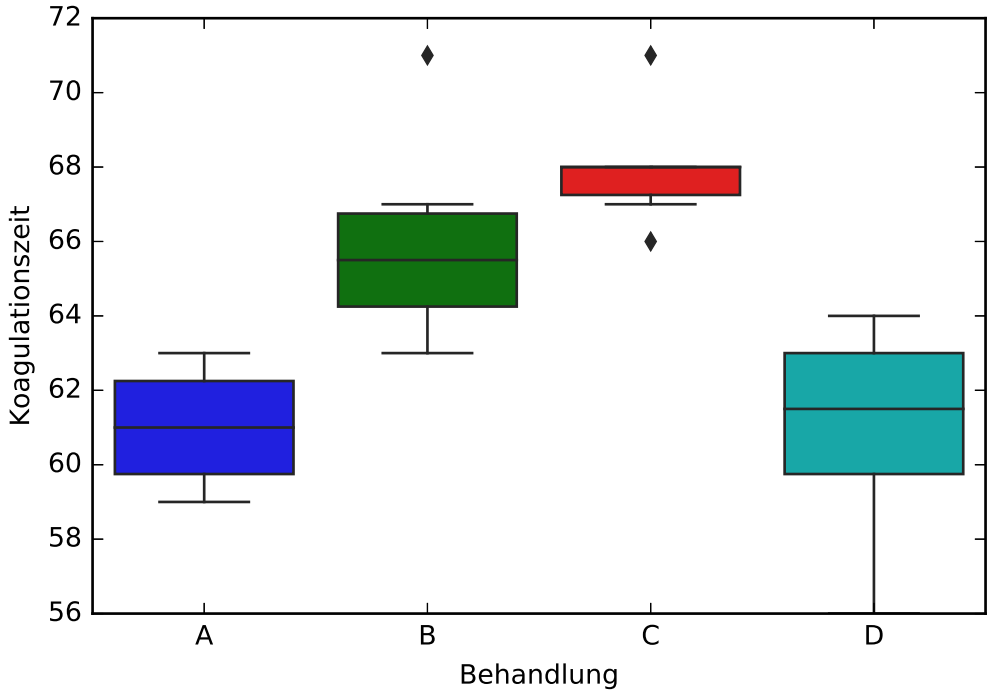
Wir erzeugen eine Stripchart Graphik wie folgt:

```
sns.stripplot(x="Behandlung", y="Koagulationszeit", data=df)
plt.xlabel("Behandlung")
plt.ylabel("Koagulationszeit")
plt.show()
```



Die entsprechenden Boxplots sind:

```
sns.boxplot(x="Behandlung", y="Koagulationszeit", data=df)
plt.xlabel("Behandlung")
plt.ylabel("Koagulationszeit")
plt.show()
```



Man sieht deutliche Unterschiede in der Lage der vier Stichproben. Vor allem die Stichprobe der Behandlung C hat deutlich höhere Werte als die drei anderen. Bezüglich der Streuung gibt es auch Unterschiede : Behandlung C weist eine kleine innere Streuung auf. Ansonsten ist aber die Streuung innerhalb der Gruppen klein im Vergleich zur Streuung zwischen den Gruppen.

b) Der grand mean ist gegeben durch

$$\begin{aligned}\overline{y}_{..} &= \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij} \\ &= \frac{1}{24}\sum_{i=1}^{4}\sum_{j=1}^{n_i} y_{ij} \\ &= 64\end{aligned}$$

Die Gruppenmittelwerte

$$\mu_i = \overline{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

sind dann

$$\mu_A = 61 \quad \mu_B = 66 \quad \mu_C = 68 \quad \mu_D = 61$$

c) Die empirischen Gruppenvarianzen berechnen sich durch:

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} r_{ij}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - y_{i\cdot})^2$$

und lauten

$$s_A^2 = 3.333 \quad s_B^2 = 8 \quad s_C^2 = 2.8 \quad s_D^2 = 6.85$$

d)

$$\mathrm{SS}_E = \sum_{i}^{g} \sum_{j}^{n_i} (y_{ij} - \hat{y}_{i\cdot})^2 = 112$$

$$\mathrm{MS}_E = \frac{1}{n - g} \mathrm{SS}_E = \frac{1}{24 - 4} 112 = 5.6$$

e)

$$\mathrm{SS}_G = \sum_{i}^{g} \sum_{j}^{n_i} (Y_{ij} - \hat{Y}_{\cdot\cdot})^2 = 228$$

$$\mathrm{MS}_G = \frac{1}{g - 1} \mathrm{SS}_G = 76$$

Wir stellen fest, dass die geschätzte Varianz zwischen den Gruppen MS_E viel grösser ist als die geschätzte Varianz innerhalb der Gruppen MS_E. Dies könnte auf einen Effekt der Diät auf die Koagulationszeit hinweisen.

f) (zu **R**)

```
df=DataFrame({
  "Behandlung": np.repeat(["A", "B", "C", "D"], [4, 6, 6, 8]),
  "Koagulationszeit" : [62, 60, 63, 59, 63, 67,
                        71, 64, 65, 66, 68, 66,
                        71, 67, 68, 68, 56, 62,
                        60, 61, 63, 64, 63, 59]
})
fit = ols("Koagulationszeit~Behandlung", data=df).fit()

fit.params

anova_lm(fit)

## Intercept          6.100000e+01
```

```
## Behandlung[T.B]    5.000000e+00
## Behandlung[T.C]    7.000000e+00
## Behandlung[T.D]    3.552714e-15
## dtype: float64
##                df  sum_sq  mean_sq          F    PR(>F)
## Behandlung   3.0   228.0     76.0  13.571429  0.000047
## Residual    20.0   112.0      5.6        NaN       NaN
```

g) Die Null-Hypothese lautet, dass sich die Behandlungsgruppen nicht unterscheiden, also dass die Gruppenmittelwerte

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

sind, oder die Behandlungseffekte (eng. *treatment effects*)

$$\tau_2 = \tau_3 = \tau_4 = 0$$

Da die transformierte Teststatistik mit $F = 13.57$ einen P-Wert von $5e - 5$ hat und somit kleiner als das Niveau 5 % ist, wird die Nullhypothese verworfen und es gilt die Alternative. Dies war ja schon ersichtlich aus dem Boxplot.

R-Code