Wahrscheinlichkeitsrechnung und
Statistik für Biologen

Faktorielle Varianzanalyse

Dirk Metzler & Noémie Becker

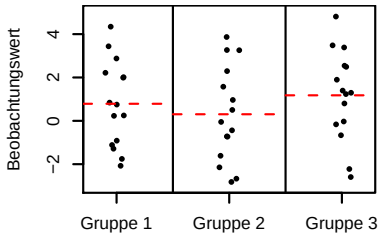
23. Mai 2016

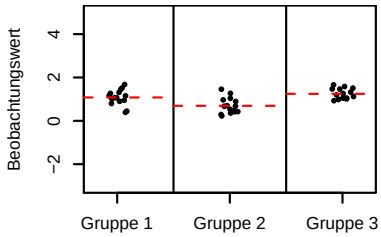
Inhaltsverzeichnis

1 Die einfaktorielle Varianzanalyse und der F-Test

Grundidee der Varianzanalyse

[-0.8cm]
Wir beobachten unterschiedliche Gruppenmittelwerte:





[-0.8cm] Variabilität innerhalb der Gruppen groß

[-0.8cm] Variabilität innerhalb der Gruppen klein

Sind die beobachteten Unterschiede der Gruppenmittelwerte ernst zu nehmen — oder könnte das alles Zufall sein?

Das hängt vom Verhältnis der Variabilität der Gruppenmittelwerte und der Variabilität der Beobachtungen innerhalb der Gruppen ab: die Varianzanalyse gibt eine (quantitative) Antwort.

Beispiel: Blutgerinnungszeiten

Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

Gruppe	Beobachtung							
1	62	60	63	59				
2	63	67	71	64	65	66		
3	68	66	71	67	68	68		
4	56	62	60	61	63	64	63	59

Globalmittelwert $\overline{x}_{..} = 64$,
Gruppenmittelwerte $\overline{x}_{1.} = 61$, $\overline{x}_{2.} = 66$, $\overline{x}_{3.} = 68$, $\overline{x}_{4.} = 61$.

Vorsicht: Der Globalmittelwert ist in diesem Beispiel auch der Mittelwert der Gruppenmittelwerte. Das muss aber nicht immer so sein!

Beispiel

Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

Gr.	$\overline{x}_{i.}$	Beobachtung							
1	61	62	60	63	59				
		$(62-61)^2$	$(60-61)^2$	$(63-61)^2$	$(59-61)^2$				
2	66	63	67	71	64	65	66		
		$(63-66)^2$	$(67-66)^2$	$(71-66)^2$	$(64-66)^2$	$(65-66)^2$	$(66-66)^2$		
3	68	68	66	71	67	68	68		
		$(68-68)^2$	$(66-68)^2$	$(71-68)^2$	$(67-68)^2$	$(68-68)^2$	$(68-68)^2$		
4	61	56	62	60	61	63	64	63	59
		$(56-61)^2$	$(62-61)^2$	$(60-61)^2$	$(61-61)^2$	$(63-61)^2$	$(64-61)^2$	$(63-61)^2$	$(59-61)^2$

Globalmittelwert $\overline{x}_{..} = 64$,
Gruppenmittelwerte $\overline{x}_{1.} = 61$, $\overline{x}_{2.} = 66$, $\overline{x}_{3.} = 68$, $\overline{x}_{4.} = 61$.

Die *roten* Werte (ohne die Quadrate) heißen *Residuen*: die „Restvariabilität“ der Beobachtungen, die das Modell nicht erklärt.

Quadratsumme innerhalb der Gruppen: $ss_{\text{innerh}} = 112$, 20 Freiheitsgrade

Quadratsumme zwischen den Gruppen: $ss_{\text{zw}} = 4 \cdot (61-64)^2 + 6 \cdot (66-64)^2 + 6 \cdot (68-64)^2 + 8 \cdot (61-64)^2 = 228$, 3 Freiheitsgrade

$$F = \frac{ss_{\text{zw}}/3}{ss_{\text{innerh}}/20} = \frac{76}{5{,}6} = 13{,}57$$

Beispiel: Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

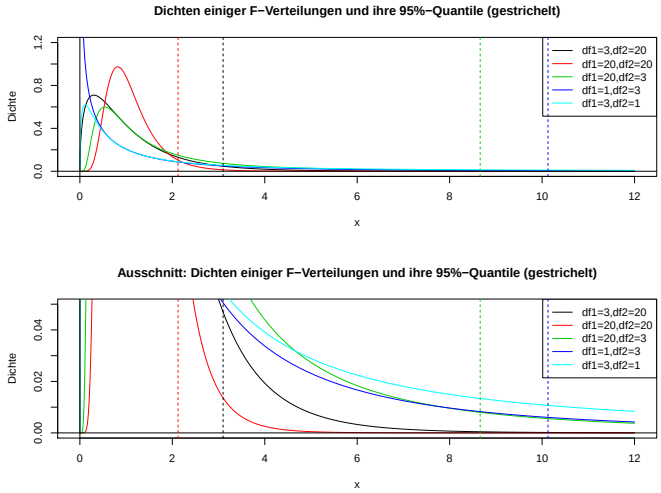
ANOVA-Tafel („ANalysis Of VAriance“)

	Freiheitsgrade (DF)	Quadratsumme (SS)	mittlere Quadratsumme (SS/DF)	F-Wert
Gruppe	3	228	76	13,57
Residuen	20	112	5,6	

Unter der Hypothese H_0 „die Gruppenmittelwerte sind gleich“ (und einer Normalverteilungsannahme an die Beobachtungen)
ist F Fisher-verteilt mit 3 und 20 Freiheitsgraden, $p = \text{Fisher}_{3,20}([13{,}57, \infty)) \leq 5 \cdot 10^{-5}$.

Wir lehnen demnach H_0 ab.

Dichte der F-Verteilung



Das 95%-Quantil der F-Verteilung mit $df_1 = 3$ und $df_2 = 1$ passte leider nicht in diese Abbildung. Es beträgt 215.7

F-Test

$n = n_1 + n_2 + \cdots + n_I$ Beobachtungen in I Gruppen,
$X_{ij} = j$-te Beobachtung in der i-ten Gruppe, $j = 1, \ldots, n_i$.
Modellannahme: $X_{ij} = \mu_i + \varepsilon_{ij}$,
mit unabhängigen, normalverteilten ε_{ij}, $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$
(μ_i ist der „wahre“ Mittelwert innerhalb der i-ten Gruppe.)

$\overline{X}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} X_{ij}$ (empirisches) „Globalmittel“
$\overline{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ (empirischer) Mittelwert der i-ten Gruppe

$$SS_{\text{innerh}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2 \quad \text{Quadratsumme innerhalb d. Gruppen, } n - I \text{ Freiheitsgrade}$$

$$SS_{\text{zw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 \quad \text{Quadratsumme zwischen d. Gruppen, } I - 1 \text{ Freiheitsgrade}$$

$$F = \frac{SS_{\text{zw}}/(I-1)}{SS_{\text{innerh}}/(n-I)}$$

F-Test

$X_{ij} = j$-te Beobachtung in der i-ten Gruppe, $j = 1, \ldots, n_i$,
Modellannahme: $X_{ij} = \mu_i + \varepsilon_{ij}$. $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

$$SS_{\text{innerh}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2 \quad \text{Quadratsumme innerhalb d. Gruppen, } n - I \text{ Freiheitsgrade}$$

$$SS_{\text{zw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 \quad \text{Quadratsumme zwischen d. Gruppen, } I - 1 \text{ Freiheitsgrade}$$

$$F = \frac{SS_{\text{zw}}/(I-1)}{SS_{\text{innerh}}/(n-I)}$$

Unter der Hypothese $H_0 : \mu_1 = \cdots = \mu_I$ („alle μ_i sind gleich“) ist F Fisher-verteilt mit $I - 1$ und $n - I$ Freiheitsgraden

(unabhängig vom tatsächlichen gemeinsamen Wert der μ_i).

F-Test: Wir lehnen H_0 zum Signifikanzniveau α ab, wenn $F \geq q_\alpha$, wobei q_α das $(1-\alpha)$-Quantil der Fisher-Verteilung mit $I-1$ und $n-I$ Freiheitsgraden ist.

Berechnung der Signifikanz mit R

Wie muss man q wählen, damit $\Pr(F \leq q) = 0.95$ für Fisher(6,63)-verteiltes F?

```
> qf(0.95,df1=6,df2=63)
[1] 2.246408
```

p-Wert-Berechnung: Wie wahrscheinlich ist es, dass eine Fisher(3,20)-verteilte Zufallsgröße einen Wert ≥ 13.57 annimmt?

```
> pf(13.57, df1=3, df2=20, lower.tail=FALSE)
[1] 4.66169e-05
```

Tabelle der 95%-Quantile der F-Verteilung

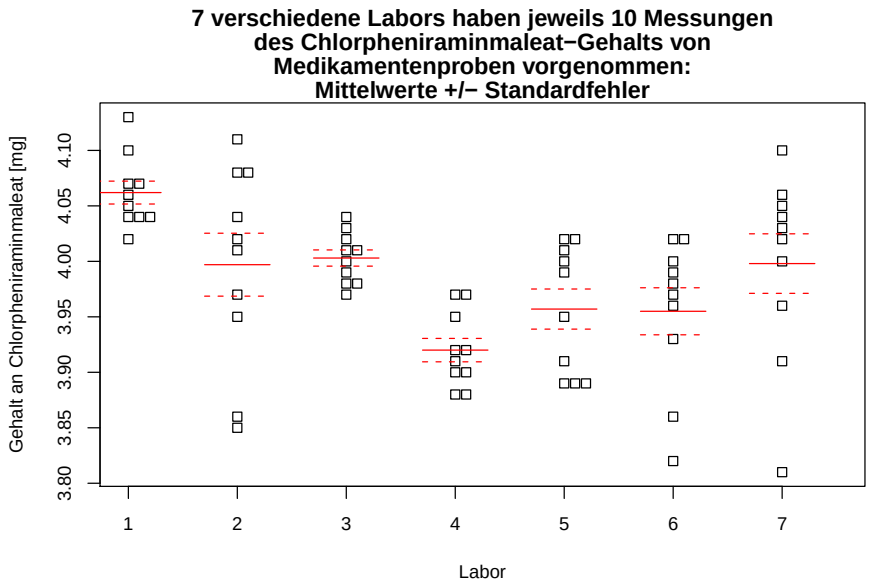
Die folgende Tabelle zeigt (auf 2 Nachkommastellen gerundet) das 95%-Quantil der Fisher-Verteilung mit k_1 und k_2 Freiheitsgraden (k_1 Zähler- und k_2 Nennerfreiheitsgrade)

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.4
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.7
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.6
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.1
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.72
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.57
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.51
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.41
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.34
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.31

Varianzanalyse komplett in R

Die Text-Datei gerinnung.txt enthält eine Spalte "bgz" mit den Blutgerinnungszeiten und eine Spalte "beh" mit der Behandlung (A,B,C,D).

```
> rat<-read.table("gerinnung.txt",header=TRUE)
> rat.aov <- aov(bgz~beh,data=rat)
> summary(rat.aov)
            Df Sum Sq Mean Sq F value    Pr(>F)
beh          3    228    76.0  13.571 4.658e-05 ***
Residuals   20    112     5.6
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```



Beachte: Die Labore sind sind mit Zahlen nummeriert. Damit R das nicht als numerische Werte sondern als Nummern der Labore auffasst, müssen wir die Variable "Labor" in einen sog. Factor umwandeln:

```
> chlor <- read.table("chlorpheniraminmaleat.txt")
> str(chlor)
'data.frame': 70 obs. of  2 variables:
 $ Gehalt: num  4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4.1 4.04 ...
 $ Labor : int  1 1 1 1 1 1 1 1 1 1 ...
> chlor$Labor <- as.factor(chlor$Labor)
> str(chlor)
'data.frame': 70 obs. of  2 variables:
 $ Gehalt: num  4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4.1 4.04 ...
 $ Labor : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Nun können wir die Varianzanalyse durchführen:

```
> chlor.aov <- aov(Gehalt~Labor,data=chlor)
> summary(chlor.aov)
            Df  Sum Sq  Mean Sq F value    Pr(>F)
Labor        6 0.12474 0.020789  5.6601 9.453e-05 ***
Residuals   63 0.23140 0.003673
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

2 Paarweise Vergleiche und Exkurs über multiples Testen

Die Varianzanalyse zeigte, dass es signifikante Unterschiede zwischen den Laboren gibt.

Aber welche Labore unterscheiden sich signifikant?

p-Werte aus paarweisen Vergleichen mittels t-Tests:

```
        Lab2    Lab3    Lab4    Lab5    Lab6    Lab7
Lab1 0.05357 0.00025 0.00000 0.00017 0.00055 0.04657
Lab2         0.84173 0.02654 0.25251 0.25224 0.97985
```

```
Lab3                0.00001 0.03633 0.05532 0.86076
Lab4                        0.09808 0.16280 0.01944
Lab5                                0.94358 0.22336
Lab6                                        0.22543
```

Wir haben 21 paarweise Vergleiche; auf dem 5%-Niveau zeigen einige davon Signifikanz an.

Problem des Multiplen Testens: Wenn die Nullhypothese ("alles nur Zufallsschwankungen") stimmt, verwirft man im Schnitt bei 5% der Tests die Nullhypothese zu Unrecht. Testet man mehr als 20 mal und gelten jeweils die Nullhypothesen, wird man im Schnitt mehr als eine Nullhypothese zu Unrecht verwerfen. Daher sollte man bei multiplen Tests mit korrigierten p-Werten arbeiten.

Eine Möglichkeit bei Varianzanalysen: Tukey's Honest Significant Differences (HSD).

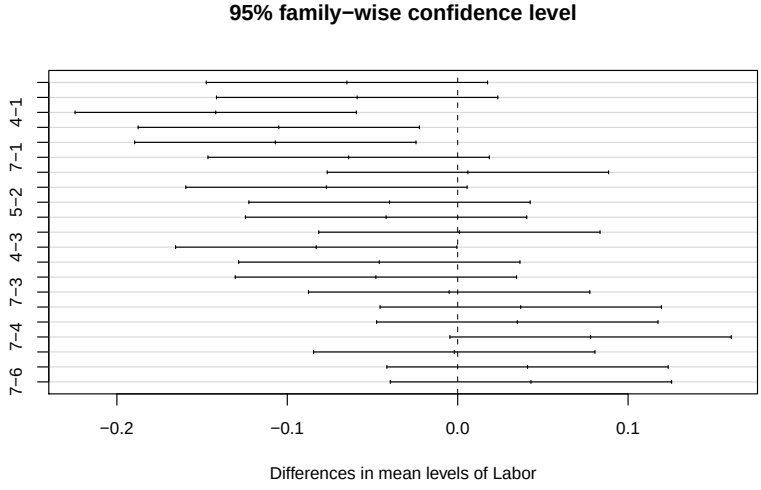
```
> TukeyHSD(chlor.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Gehalt ~ Labor, data = chlor)

$Labor
      diff          lwr          upr     p adj
2-1 -0.065 -0.147546752  0.017546752 0.2165897
3-1 -0.059 -0.141546752  0.023546752 0.3226101
4-1 -0.142 -0.224546752 -0.059453248 0.0000396
5-1 -0.105 -0.187546752 -0.022453248 0.0045796
6-1 -0.107 -0.189546752 -0.024453248 0.0036211
7-1 -0.064 -0.146546752  0.018546752 0.2323813
3-2  0.006 -0.076546752  0.088546752 0.9999894
4-2 -0.077 -0.159546752  0.005546752 0.0830664
5-2 -0.040 -0.122546752  0.042546752 0.7578129
6-2 -0.042 -0.124546752  0.040546752 0.7140108
7-2  0.001 -0.081546752  0.083546752 1.0000000
4-3 -0.083 -0.165546752 -0.000453248 0.0478900
5-3 -0.046 -0.128546752  0.036546752 0.6204148
6-3 -0.048 -0.130546752  0.034546752 0.5720976
7-3 -0.005 -0.087546752  0.077546752 0.9999964
5-4  0.037 -0.045546752  0.119546752 0.8178759
6-4  0.035 -0.047546752  0.117546752 0.8533629
7-4  0.078 -0.004546752  0.160546752 0.0760155
6-5 -0.002 -0.084546752  0.080546752 1.0000000
7-5  0.041 -0.041546752  0.123546752 0.7362355
7-6  0.043 -0.039546752  0.125546752 0.6912252
```

Wir erhalten Konfidenzintervalle [lwr,upr] für die Unterschiede zwischen den Labormittelwerte und p-Werte für die Nullhypothese, dass diese Unterschiede 0 sind; alles bereits korrigiert für multiples Testen.

Mit `plot(TukeyHSD(chlor.aov))` bekommt man die Konfidenzintervalle für die Unterschiede auch visualisiert:



Einschränkung: Tukeys HSD-Methode ist streng genommen nur für *balancierten Versuchsplänen* (engl. *balanced design*) anwendbar, d.h. wenn in jeder Gruppe die selbe Anzahl von Messungen vorliegt. (Außerdem geht HSD wie die ANOVA selbst von gleichen Varianzen in allen Gruppen aus.)

Das ist bei dem Laborvergleich der Fall, da jedes Labor 10 Messungen durchgeführt hat.Die Blutgerin-

nungsdaten sind jedoch nicht balanciert, da Behandlung 1 an vier Ratten und Behandlung 2 an 8 Ratten erprobt wurde.

Was können wir verwenden, wenn die Bedingungen für Tukeys HSD nicht erfüllt sind?

Eine ganz allgemeine Korrektur für multiples Testen ist die **Bonferroni-Methode**: Multipliziere jeden p-Wert mit der Anzahl n der durchgeführten Tests.

Beispiel: Paarweise Vergleiche (mittels t-Test) für die Blutgerinnungszeiten bei vier verschiedenen Behandlungen, zunächst ohne Korrektur für multiples Testen:

```
        B       C       D
A 0.00941 0.00078 1.00000
B         0.17383 0.00663
C                 0.00006
```

Nun mit Bonferroni-Korrektur (alle Werte mit 6 multiplizieren):

```
        B       C       D
A 0.05646 0.00468 6.00000
B         1.04298 0.03978
C                 0.00036
```

Nach Bonferroni-Korrektur führen folgende Paare von Behandlungen zu jeweils signifikant unterschiedlichen Ergebnissen: A/C, B/D sowie C/D. (Der Bonferroni-korrigierte p-Wert von 6.0 für den Vergleich der Behandlungen A und D ist natürlich nicht als echter p-Wert zu interpretieren.)

Die Bonferroni-Methode ist sehr *konservativ*, d.h. um auf der sicheren Seite zu sein, lässt man sich lieber die eine oder andere Signifikanz entgehen.

Eine Verbesserung der Bonferroni-Methode ist die **Bonferroni-Holm-Methode**: Ist k die Anzahl der Tests, so multipliziere den kleisten p-Wert mit k, den zweitkleinsten mit $k-1$, den drittkleinsten mit $k-2$ usw.

In R gibt des den Befehl p.adjust, der p-Werte für multiples Testen korrigiert und dabei defaultmäßig Bonferroni-Holm verwendet:

```
> pv <- c(0.00941, 0.00078, 1.00000, 0.17383,
+         0.00663, 0.00006)
> p.adjust(pv)
[1] 0.02823 0.00390 1.00000 0.34766 0.02652 0.00036
> p.adjust(pv, method="bonferroni")
[1] 0.05646 0.00468 1.00000 1.00000 0.03978 0.00036
```

Für paarweise t-Tests gibt es ebenfalls eine R-Funktion, die per default die Bonferroni-Holm-Korrektur verwendet:

```
> pairwise.t.test(rat$bgz,rat$beh,pool.sd=FALSE)

Pairwise comparisons using t tests with non-pooled SD

data:  rat$bgz and rat$beh

  A       B       C
B 0.02823 -       -
C 0.00391 0.34766 -
D 1.00000 0.02654 0.00035

P value adjustment method: holm
```

3 Nichtparameterisch: Der Kruskal-Wallis-Test

Die einfaktorielle Varianzanalyse basiert auf der Annahme, dass die gemessenen Werte unabhängig und normalverteilt sind. Die Mittelwerte $\mu_1, \mu_2, \ldots, \mu_m$ können verschieden sein (das herauszufinden ist Ziel des Tests), aber die Varianzen innerhalb der verschiedenen Gruppen müssen gleich sein.

In Formeln: Ist Y_{ij} die j-te Messung in der i-ten Gruppe, so muss gelten

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

wobei alle ε_{ij} unabhängig $\mathcal{N}(0, \sigma^2)$-verteilt sind, mit demselben σ^2 für alle Gruppen!

Die zu testende Nullhypothese ist $\mu_1 = \mu_2 = \cdots = \mu_m$.
Nicht jede Abweichung von der Normalverteilung stellt ein Problem dar.

Die Anova ist aber nicht robust gegenüber Ausreißern bzw. Verteilungen, die seltene extrem große Werte liefern.

In diesem Fall kann man den **Kruskal-Wallis-Test** verwenden, der wie der Wilcoxon-Test die *Ränge* statt der tatsächlichen Werte verwendet. Es handelt sich also um einen *nicht-parameterischen Test*, d.h. es wird keine bestimmte Wahrscheinlichkeitsverteilung vorausgesetzt.

Nullhypothese des Kruskal-Wallis-Tests: alle Werte Y_{ij} kommen aus derselben Verteilung, unabhängig von der Gruppe.

Grundvoraussetzung ist auch beim Kruskal-Wallis-Test, dass die Werte unanbhängig voneinander sind.

- Sei R_{ij} der Rang von Y_{ij} innerhalb der Gesamtstichprobe.
- Sei
$$\overline{R}_{i\cdot} = \frac{1}{J_i} \sum_{j=1}^{J_i} R_{ij}$$
der durchschnittliche Rang in Gruppe i, wobei J_i die Anzahl der Messungen in Gruppe i ist.
- Der mittlere Rang der Gesamtstichprobe ist
$$\overline{R}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J_i} R_{ij} = \frac{N+1}{2},$$
wobei I die Anzahl der Gruppen ist und N der Umfang der Gesamtstichprobe.
- Unter der Nullhypothese haben die mittleren Ränge der Gruppen denselben Erwartungswert $R_{\cdot\cdot}$.
- Die Abweichung von dieser Erwartung kann man messen mit der Teststatistik
$$S = \sum_{i=1}^{I} J_i \cdot (\overline{R}_{i\cdot} - \overline{R}_{\cdot\cdot})^2.$$
- Um aus S einen p-Wert zu erhalten, muss man die Verteilung von S unter der Nullhypothese kennen. Diese kann man für verschiedene I und J_I in Tabellen finden.
- Für $I \geq 3$ und $J_i \geq 5$ sowie $I > 3$ und $J_i \geq 4$ kann man ausnutzen, dass die folgende Skalierung K von S approximativ χ^2-verteilt ist mit $I - 1$ Freiheitsgraden:
$$K = \frac{12}{N \cdot (N+1)} S = \frac{12}{N \cdot (N+1)} \cdot \left(\sum_{i=1}^{I} J_i \cdot \overline{R}_{i\cdot}^2 \right) - 3 \cdot (N+1)$$

Kruskal-Wallis-Test mit R

```
> kruskal.test(bgz~beh,data=rat)

Kruskal-Wallis rank sum test

data:  bgz by beh
Kruskal-Wallis chi-squared = 17.0154, df = 3,
                                p-value = 0.0007016

> kruskal.test(Gehalt~Labor,data=chlor)

Kruskal-Wallis rank sum test

data:  Gehalt by Labor
Kruskal-Wallis chi-squared = 29.606, df = 6,
                               p-value = 4.67e-05
```