Die Varianzanalyse

Analysis of Variance (ANOVA)

Jonathan Harrington

```
pfad = "Verzeichnis wo Sie anova1 gespeichert haben"
attach(paste(pfad, "anova1", sep="/"))
```

Variablen, Faktoren, Stufen

Faktoren oder Unabhängige Variablen (kategorial)

Gibt es signifikante Unterschiede zwischen den Stufen eines Faktors...

Faktor	Stufen
Dialekt	Bayern, Hessen, Sachsen
Geschlecht	M, W
Silbenposition	initial, medial, final

Eine abhängige Variable (kontinuierlich)

...in F2, oder in der Dauer oder in der Kieferposition usw?

Der statistische Test

t-test oder ANOVA	Ein Faktor hat 2 Stufen
ANOVA	Ein Faktor mit mehreren Stufen
	Mehrere Faktoren

Was ist die Varianzanalyse?

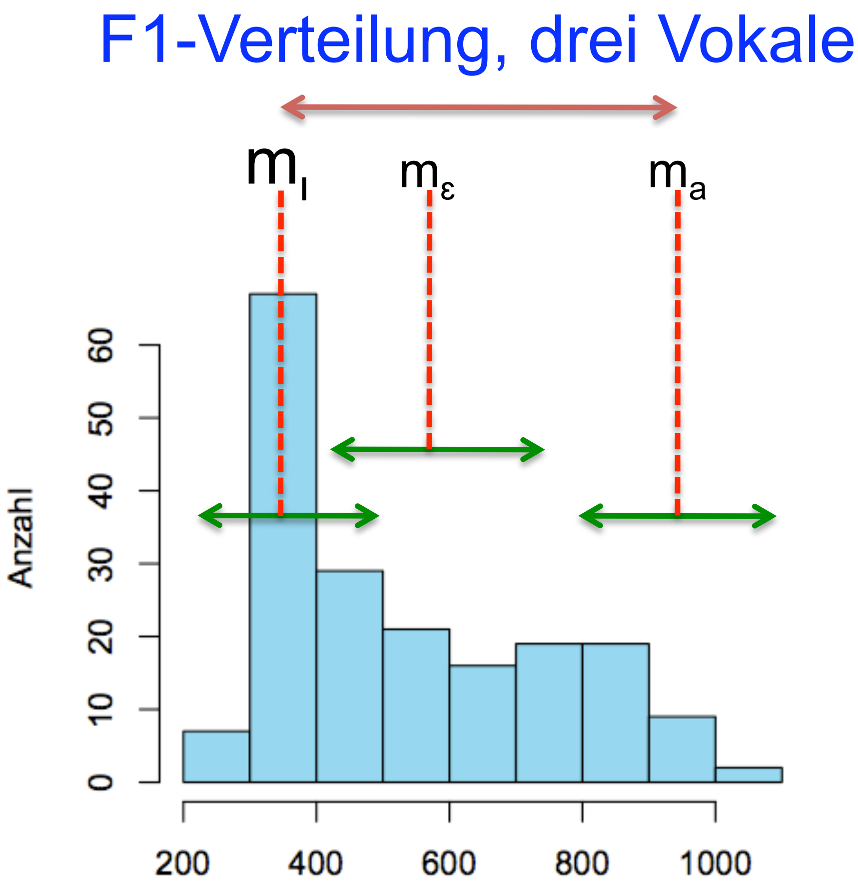
Mit der Varianzanalyse wird (durch ein F-Test) ein Verhältnis zwischen zwei Varianzen berechnet: **innerhalb von Stufen** und **zwischen Stufen**.

z.B. F1 von drei Vokalkategorien, /ɪ,ɛ,a/.

innerhalb: Es gibt eine **randomisierte Variation von F1** innerhalb jeder Stufe (F1 von /I/ variiert, F1 von /E/ variiert, F1 von /a/ variiert).

zwischen: F1 variiert, weil es eine **systematische Variation** zwischen den Verteilungen der Vokalkategorien gibt: die Werte von /I/, /ɛ/, und /a/ liegen in ganz unterschiedlichen F1-Bereichen, und je unterschiedlicher sie sind, **umso größer wird diese Varianz im Verhältnis zu der willkürlichen, randomisierten Varianz innerhalb der Stufen sein**.

Was ist die Varianzanalyse?



$$F = \frac{\text{Varianz zwischen den Stufen}}{\text{Varianz innerhalb der Stufen}}$$

Ist F signifikant größer als 1?

Berechung der Varianzen, innerhalb und zwischen

Diese Berechung erfolgt über die sogenannte **Quadratsumme** oder **sum-of-squares**, die sich von der Varianz ableiten lässt (1) oder die durch die Formel (2) direkt berechnet werden kann

(1) $ss_x = (n-1)\sigma_x^2$

(2) $ss_x = \sum (x - m_x)^2$

(Quadratsumme von *x* gleicht die Varianz von *x* mal *n*-1 (*n* ist die Anzahl der Stichproben).

Bestätigen

```
x = 1:6
n   = length(x)
v = var(x)
v * (n-1)
```

```
m = mean(x)
ssx    = sum((x - m)^2)
```

Berechung der Varianzen, innerhalb und zwischen

d.h. wenn wir die Quadratsummen wissen, gelangen wir zu den Varianzen, und wenn wir die Varianzen wissen, können wir den erwünschten F-Test durchführen.

Warum aber diese Schiene über die Quadratsummen?

Wegen einer Beziehung zwischen 3 Quantitäten, die auf eine sehr ähnliche Weise in der Regression vorkam.

SSY	**=**	**SSR**	**+**	**SSE**
Die Quadratsumme über die gesamte Verteilung berechnet	=	Die Quadratsummen zwischen den Stufen	+	Die Quadratsummen innerhalb der Stufen

Berechung der Varianzen, innerhalb und zwischen

y 20 F2-Werte, 10 /I/, 10 /E/, ein Wert pro Person (also 20 Werte von 20 unterschiedlichen Personen)

vokal = factor(names(y))

table(vokal)

```
 E  I
10 10
```

Berechung der Varianzen, innerhalb und zwischen

Quadratsummen gesamt (SSY) =

SSY = var(y) * (length(y) – 1)

$$ss_y = (n-1)\sigma_y^2$$

Q-Summen innerhalb (SSE) +

$$SSE = ss_I + ss_E$$

Q-summen **innerhalb der Stufen** gleicht die Quadratsumme von /I/ plus die Q-Summe von /E/

Quadratsumme von /I/ Quadratsumme von /E/

temp = vokal=="I"

SSE = var(y[temp]) * 9 + var(y[!temp]) * 9

In einer Zeile

SSE = sum(tapply(y, vokal, var) * 9)

Q-Summen zwischen (SSR)

SSR = SSY - SSE

Berechung der Varianzen, innerhalb und zwischen

$$\text{Fratio} = \frac{\text{Varianz zwischen den Stufen}}{\text{Varianz innerhalb der Stufen}} = \frac{\text{MSR}}{\text{MSE}}$$

Ist Fratio signifikant größer als 1?

(1) $ss_x = (n-1)\sigma_x^2$ $\qquad$ $\sigma_x^2 = \dfrac{ss_x}{(n-1)}$

MSE = SSE / 18

18 weil n-1 = 9 pro Stufe

MSR = SSR/1

weil 2 Stufen (/I, E/),
2 – 1 = 1

Fratio = MSR/MSE

```
[1] 18.77464
```

ANOVA Berechnung in R

Wir bekommen SSY, SSR, SSE mit derselben lm() Funktion, die wir für die Regression eingesetzt haben.

reg = lm(y ~ x)

Regression: x ist eine **(numerische) Variable**

Varianzanalyse: x ist **ein Faktor** (zB Vokal mit 2 Stufen ("I", "I", "E", "I", "E" usw).

reg = lm(y ~ vokal)

ANOVA Berechnung in R

(Siehe entsprechende Folie, Regression)

anova(reg)   SSR   MSR = SSR/1

fstat = MSR/MSE   1 - pf(fstat, 1, 18)

```
Response: y

          Df Sum Sq Mean Sq F value    Pr(>F)
vokal      1 174658  174658  18.775 0.0004007 ***
Residuals 18 167452    9303
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSE   MSE = SSE/18

F2 wird signifikant vom Vokal beeinflusst
($F(1, 18) = 7.95$, $p < 0.05$).

(MSR, MSE sind die Varianzen zwischen und innerhalb der Stufen – siehe Folie 9)

Beziehung: t-test und ANOVA

Da wir in diesem Fall mit einem Faktor und 2 Stufen zu tun haben, hätten wir das gleiche Ergebnis mit einem t-test bekommen können

t.test(y ~ vokal, var.equal=T)

```
t = -4.333, df = 18, p-value = 0.0004007
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -277.52196  -96.27804
sample estimates:
mean in group E mean in group I
         1695.8          1882.7
```

Die t-Statistik ist die Wurzel vom F-Ratio aus der ANOVA

ANOVA: einige Voraussetzungen

- ähnlich stark besetzte Stufen und Faktoren

  zB 20 initiale, 20 mediale, 20 finale /t/s, um zu messen, ob die Silbenposition (= Faktor) einen Einfluss auf die Dauer hat. Um zusätzlich zu messen, ob Dialekt (Bayern, Hessen) einen Einfluss ausübt: 30 aus Bayern, 30 aus Hessen, jeweils 10 pro Silbenposition.

Silbenposition **Dialekt**	initial	medial	final
Bayern	10	10	10
Hessen	10	10	10

ANOVA: einige Voraussetzungen

Silbenposition Dialekt	initial	medial	final
Bayern	10	10	10
Hessen	10	10	10

- Die Werte sind voneinander unabhängig.

daher müssen entweder **alle Werte von derselben Vpn**. sein, oder **alle Werte sind von unterschiedlichen Personen** (60 Vpn., ein Wert pro Vpn für das obige Beispiel).

Sollte dieselbe Person in verschiedenen Stufen gemessen werden: **ANOVA mit Messwiederholungen** (Repeated-Measures ANOVA).

ANOVA: weitere Voraussetzungen

- Die Varianzen der Stufen eines Faktors sind voneinander nicht signifikant unterschiedlich

Ein Faktor hat 2 Stufen

var.test()

Mehrere Stufen

bartlett.test()

```
var.test(y ~ vokal)
```

```
bartlett.test(y ~ vokal)
```

Wenn die Varianzen innerhalb der Stufen unterschiedlich sind

```
oneway.test(y ~ vokal)
```

Analog zu

```
t.test(y ~ vokal)
```

ANOVA: weitere Voraussetzungen

- Die Verteilung der Werte innerhalb der Stufen weicht nicht signifikant von einer Normalverteilung ab.

  shapiro.test(), pro Stufe

  ```
  tapply(y, vokal, shapiro.test)
  ```

Keine Normalverteilung: Kruskall-Wallis rank sum test

```
kruskal.test(y ~ vokal)
```

Analog zu

```
wilcox.test(y ~ vokal)
```

Zwei Faktoren

```
attach(vok)
names(vok)
"F2"     "Vokal"  "Gen"
```

F2 Daten, 60 Sprecher, 30 m, 30 w, drei Vokale

```
table(Vokal, Gen)

 Gen
Vokal  m  w
    E 10 10
    I 10 10
    a 10 10
```

Ist Vokal signifikant?

Ist Gender signifikant?

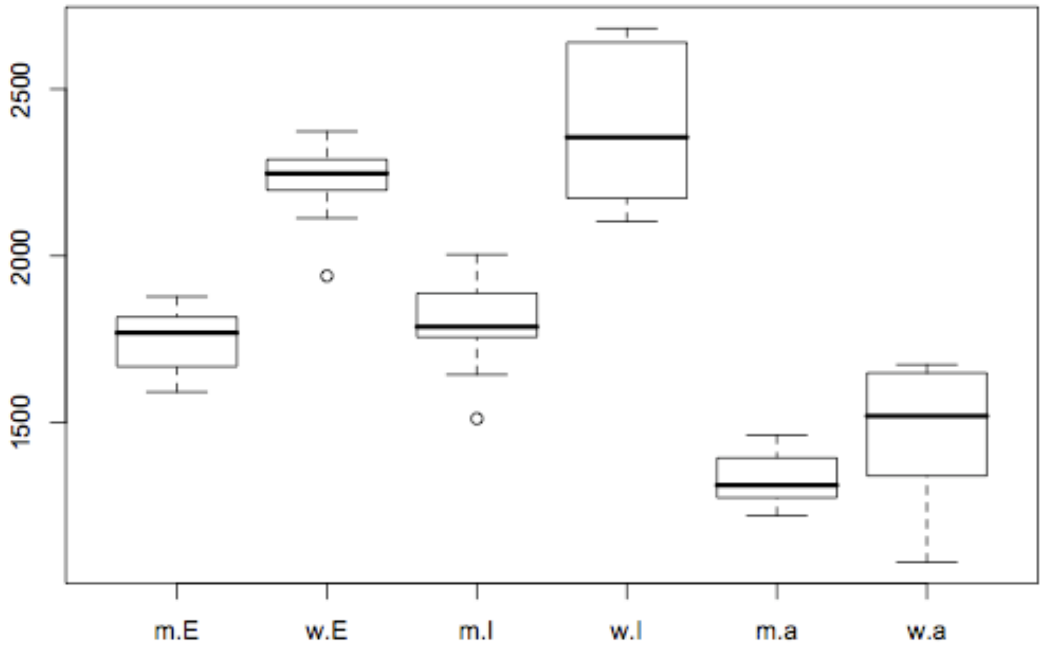
Gibt es eine Interaktion zwischen Vokal und Gender?

= unterscheiden sich Männer und Frauen auf eine ähnliche Weise bezüglich Unterschiede zwischen /I, E, a/?

Zwei Faktoren

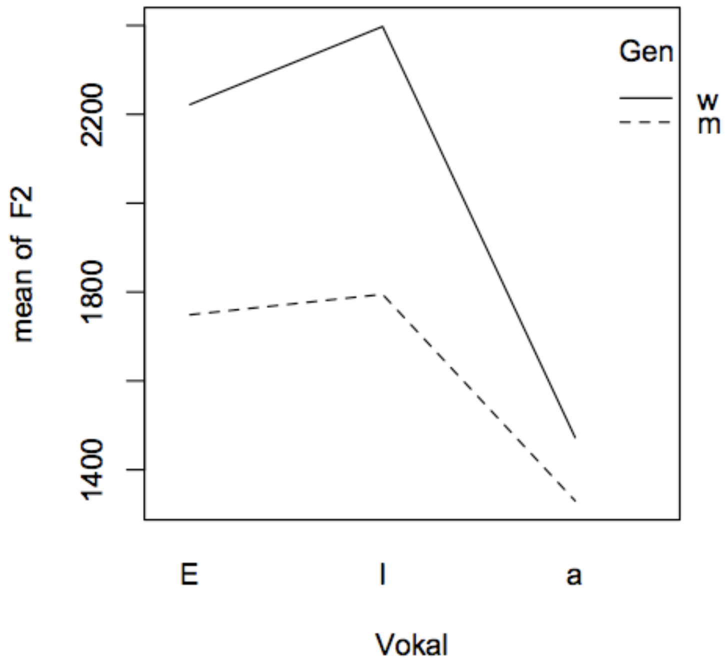
Boxplot Abbildung

boxplot(F2 ~ Gen * Vokal)



Interaktion-Abbildung

interaction.plot(Vokal, Gen, F2)



Ist Vokal signifikant?

Ist Gender signifikant?

Gibt es eine Interaktion zwischen Vokal und Gender?

Zwei Faktoren

genau das Gleiche wie

vok.aov = aov(F2 ~ Vokal * Gen)

anova(vok.aov)

reg = lm(F2 ~ Vokal * Gen)

anova(reg)

```
Analysis of Variance Table

Response: F2
          Df  Sum Sq Mean Sq F value    Pr(>F)
Vokal      2 5578128 2789064 119.637 < 2.2e-16 ***
Gen        1 2474570 2474570 106.147 2.354e-14 ***
Vokal:Gen  2  563391  281696  12.083 4.603e-05 ***
Residuals 54 1258885   23313
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

post-hoc tests

Wenn eine Interaktion vorliegt, muss geprüft werden, ob die einzelnen Tests zwischen den verschiedenen Stufen-Kombinationen signifikant sind.

Insbesondere wollen wir feststellen, ob die Unterschiede zwischen den Vokalen ähnlich sind für männlich und weiblich.

post-hoc Tukey test: (angenommen, dass die Anzahl der Werte pro Gruppe ziemlich ähnlich ist, wie hier).

post-hoc tests

tk = TukeyHSD(vok.aov)

tk

(innerhalb von Vokal)

```
$Vokal
       diff         lwr       upr     p adj
I-E  110.80   -5.561759  227.1618 0.0650875
a-E -584.25 -700.611759 -467.8882 0.0000000
a-I -695.05 -811.411759 -578.6882 0.0000000
```

Post-hoc Tukey Tests zeigten einen signifikanten Unterschied zwischen /a/ und /E/ ($p < 0.001$) sowie zwischen /a/ und /I/ ($p < 0.001$).

post-hoc tests

```
$Gen
         diff      lwr      upr p adj
w-m 406.1667 327.1282 485.2052     0
```

(innerhalb von Gender) - ist nicht relevant (trägt nichts neues bei), da wir schon aus dem Haupttest wissen, dass es signifikante Unterschiede innerhalb von Gender gibt.

post-hoc tests

alle paarweise Kombinationen von Gender * Vokal

```
$`Vokal:Gen`
           diff         lwr        upr     p adj
I:m-E:m    46.0  -155.74006  247.74006 0.9841188
a:m-E:m  -418.9  -620.64006 -217.15994 0.0000015
E:w-E:m   473.2   271.45994  674.94006 0.0000001
I:w-E:m   648.8   447.05994  850.54006 0.0000000
a:w-E:m  -276.4  -478.14006  -74.65994 0.0021912
a:m-I:m  -464.9  -666.64006 -263.15994 0.0000001
E:w-I:m   427.2   225.45994  628.94006 0.0000010
I:w-I:m   602.8   401.05994  804.54006 0.0000000
a:w-I:m  -322.4  -524.14006 -120.65994 0.0002373
E:w-a:m   892.1   690.35994 1093.84006 0.0000000
I:w-a:m  1067.7   865.95994 1269.44006 0.0000000
a:w-a:m   142.5   -59.24006  344.24006 0.3094441
I:w-E:w   175.6   -26.14006  377.34006 0.1221478
a:w-E:w  -749.6  -951.34006 -547.85994 0.0000000
a:w-I:w  -925.2 -1126.94006 -723.45994 0.0000000
```

F2-Werte von /I/ (z.B. *mitte*) sind von 20 Vpn. (alle männlich) erhoben worden. 10 Vpn. waren aus München, 10 aus Wien; und 10 Vpn. waren jungendliche (< 18 Jahre) und 10 waren ältere Personen (> 50 Jahre). Die F2-Werte dieser 20 Vpn. sind wie folgt (alle Werte in Hz).

München		Wien	
jung	alt	jung	alt
2300	1920	2401	2420
2212	1855	2415	2332
2005	1761	2308	2505
2010	1880	2100	2210
2440	2010	2520	2325

Inwiefern ist die F2-Variation dialekt- und/oder altersbedingt?