

Varianzanalyse mit Messwiederholungen (Repeated-measures (ANOVA))

Jonathan Harrington

Messwiederholungen: der gepaarte t-test

8 französische Vpn. erzeugten /pa/ und /ba/. Die VOT-Werte (ms) für diese 8 Vpn. sind wie folgt. Wir wollen prüfen, ob sich diesbezüglich /pa/ und /ba/ unterscheiden.

		ba	pa	
8 verschiedene	[1,]	10	20	
Vpn, zwei	[2,]	-20	-10	
Messung pro	[3,]	5	15	
Vpn, einmal	[4,]	-10	0	←
fuer /pa/, einmal	[5,]	-25	-20	
fuer /ba/	[6,]	10	16	
	[7,]	-5	7	
	[8,]	0	5	

VOT für Vpn 4 ist
-10 ms für /ba/, 0
ms für /pa/.

Ist der VOT-Unterschied zwischen /ba, pa/ signifikant?

Messwiederholungen: der gepaarte t-test

	ba	pa
[1,]	10	20
[2,]	-20	-10
[3,]	5	15
[4,]	-10	0
[5,]	-25	-20
[6,]	10	16
[7,]	-5	7
[8,]	0	5

Vielleicht ein t-test?

```
voice = read.table(file.path(pfad, "voice.txt"))  
t.test(vot ~ Stimm, var.equal=T, data = voice)
```

```
data: vot by Stimm
```

```
t = -1.2619, df = 14, p-value = 0.2276
```

Nicht signifikant

Messwiederholungen: der gepaarte t-test

```
      ba  pa
[1,]  10  20
[2,] -20 -10
[3,]   5  15
[4,] -10   0
[5,] -25 -20
[6,]  10  16
[7,]  -5   7
[8,]   0   5
data:  vot by Stimm
t = -1.2619, df = 14, p-value = 0.2276
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
  -22.94678    5.94678
sample estimates:
mean in group ba mean in group pa
      -4.375          4.125
```

Mit einem konventionellen t-Test wird jedoch nicht berücksichtigt, dass die Werte **gepaart sind**, d.h. Paare von /pa, ba/ sind **von derselben Vpn.** Genauer: der Test vergleicht einfach **den Mittelwert von /pa/ (über alle 8 Vpn) mit dem Mittelwert von /ba/**, ohne zu berücksichtigen, dass z.B. VOT von Vpn. 2 insgesamt viel kleiner ist als VOT von Vpn. 6.

Messwiederholungen: der gepaarte t-test

Ein **gepaarter t-test** klammert die Sprechervariation aus und vergleicht **innerhalb von jedem Sprecher** ob sich /pa/ und /ba/ unterscheiden

```
t.test(vot ~ Stimm, var.equal=T, paired=T, data = voice)
```

```
Paired t-test
```

```
data: vot by Stimm
```

```
t = -8.8209, df = 7, p-value = 4.861e-05
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-10.778609 -6.221391
```

```
sample estimates:
```

```
mean of the differences
```

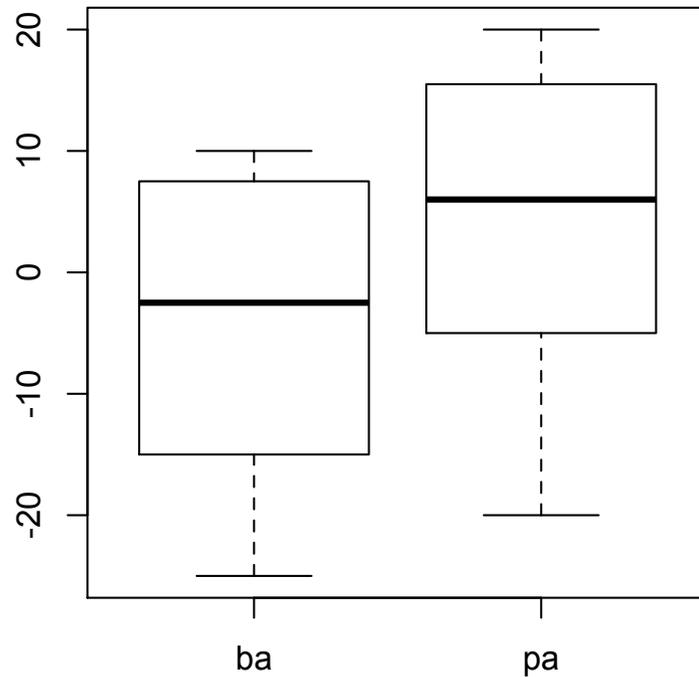
```
-8.5
```

Signifikant, $t = -8.82$, $df = 7$, $p < 0.001$

t-test (Anova)

prüft ob sich die Mittelwerte
der Verteilungen
unterschieden (hier falsch)

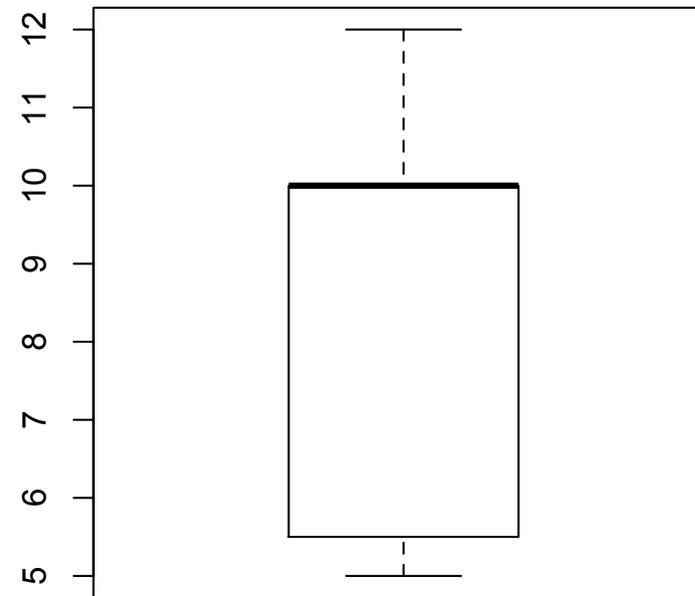
falsch



gepaarter t-test (RM-Anova)

prüft ob die Unterschiede
zwischen Paaren im selben
Sprecher von 0 (Null) abweichen

/ba-pa/ Unterschiede im selben Sprecher



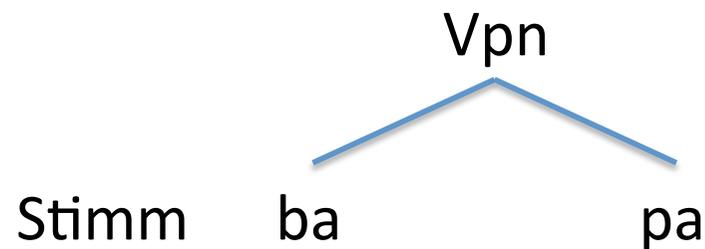
Within- and between-subjects factors

within-subject factor

Für das letzte Beispiel war Stimm (Stufen = ba, pa) ein **within-subjects Faktor**, weil es **pro Versuchsperson für jede Stufe von Stimm einen Wert gab** (einen Wert für ba, einen Wert für pa).

	ba	pa
[1,]	10	20
[2,]	-20	-10
[3,]	5	15
[4,]	-10	0
[5,]	-25	-20
[6,]	10	16
[7,]	-5	7
[8,]	0	5

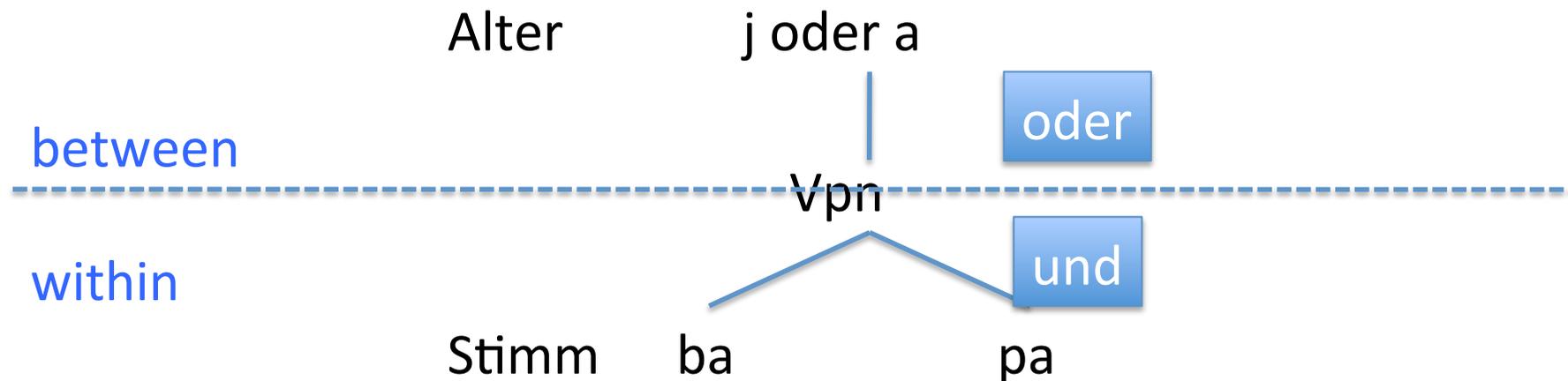
Stimm ist ein Faktor
mit 2 Stufen (ba, pa)



2 Stufen pro Vpn

Within- and between-subjects factors

Ein **Between subjects factor** beschreibt meistens eine kategorische Eigenschaft pro Vpn. Z.B. Sprache (englisch oder deutsch oder französisch), Geschlecht (m oder w), Alter (jung oder alt) usw.



Within- and between-subjects factors

	ba	pa		
[1,]	10	20		
[2,]	-20	-10		
[3,]	5	15	Between	keine
[4,]	-10	0		
[5,]	-25	-20	Within	Stimm
[6,]	10	16		
[7,]	-5	7		
[8,]	0	5		

Die Kieferposition wurde in 3 Vokalen /i, e, a/ und jeweils zu 2 Sprechtempi (langsam, schnell) gemessen. Die Messungen (3 x 2 = 6 pro Vpn) sind von 16 Vpn erhoben worden, 8 mit Muttersprache spanisch, 8 mit Muttersprache englisch.

Inwiefern haben Sprache, Sprechtempo, oder Vokale einen Einfluss auf die Kieferposition?

Between

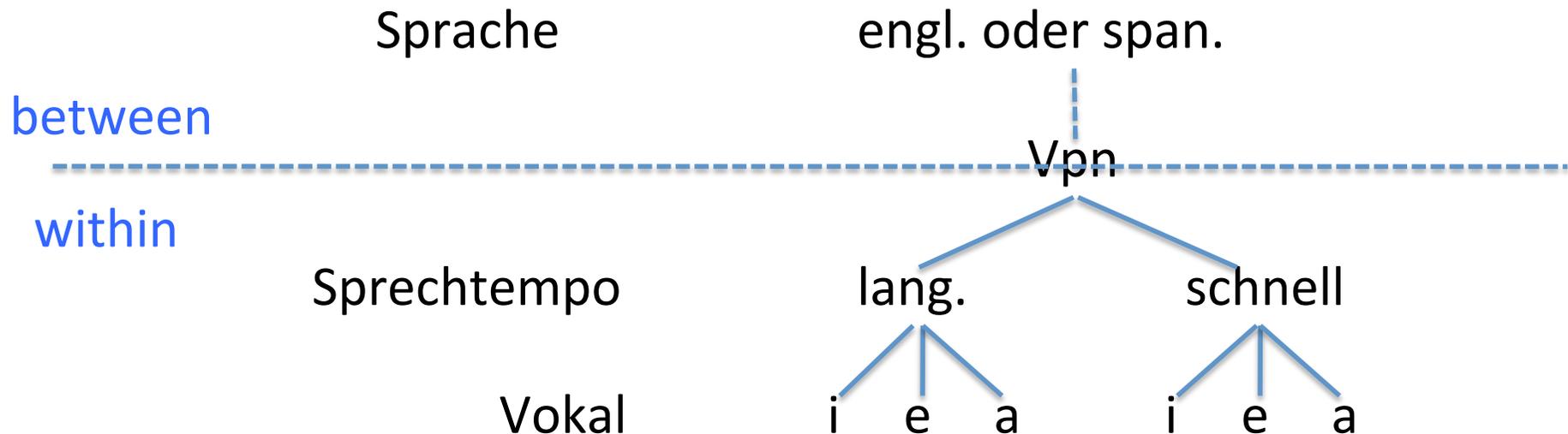
Sprache

Within

Sprechtempo, Vokal

Within- and between-subjects factors

Die Kieferposition wurde in 3 Vokalen /i, e, a/ und jeweils zu 2 Sprechtempi (langsam, schnell) gemessen. Die Messungen sind von 8 mit Muttersprache spanisch, 8 mit Muttersprache englisch aufgenommen worden.

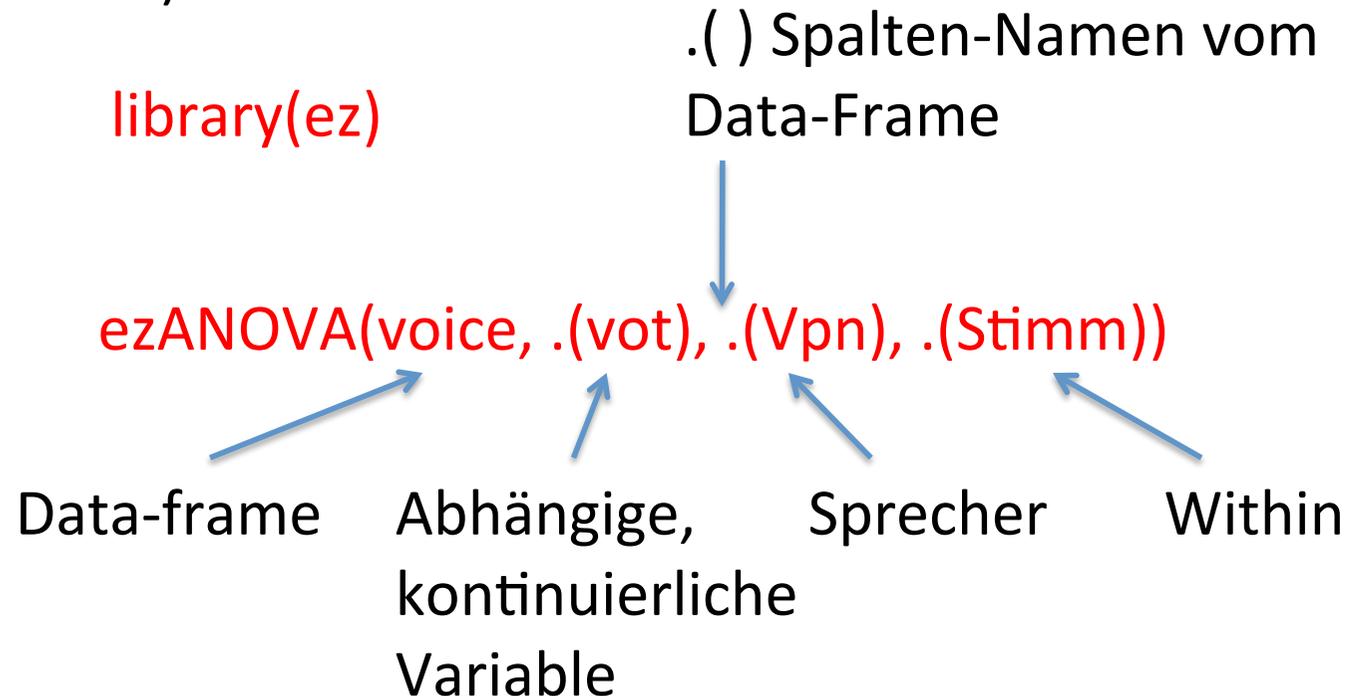


(6 Werte pro Vpn)

ANOVA mit Messwiederholungen und der gepaarte t-test

Die Generalisierung eines gepaarten t-tests ist die **Varianzanalyse mit Messwiederholungen** (RM-ANOVA, repeated measures ANOVA).

```
      ba  pa
[1, ] 10  20
[2, ] -20 -10
[3, ]  5  15
[4, ] -10  0
[5, ] -25 -20
[6, ] 10  16
[7, ] -5   7
[8, ]  0   5
```



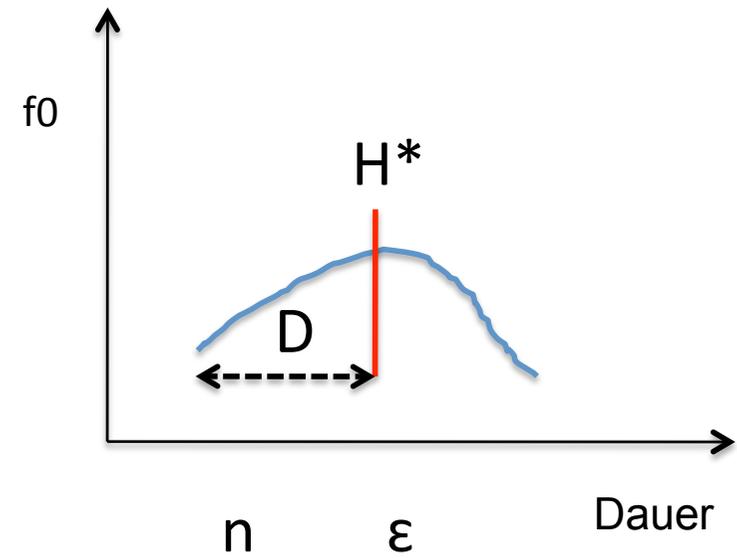
```
$ANOVA
```

	Effect	DFn	DFd	SSn	SSd	F	p	p<.05	pes
1	(Intercept)	1	7	0.25	2514.75	6.958942e-04	9.796907e-01		9.940358e-05
2	Stimm	1	7	289.00	26.00	7.780769e+01	4.860703e-05	*	9.174603e-01

Vot wird von Stimmhaftigkeit beeinflusst ($F[1,7] = 77.8, p < 0.001$)

RM-Anova: between and within

Die Dauer, D , (ms) wurde gemessen zwischen dem Silbenonset und dem H^* Tonakzent in äußerungsinitialen Silben (zB nächstes) und -finalen Silben (demnächst) jeweils von 10 Vpn., 5 aus Bayern (B) und 5 aus Schleswig-Holstein (SH).



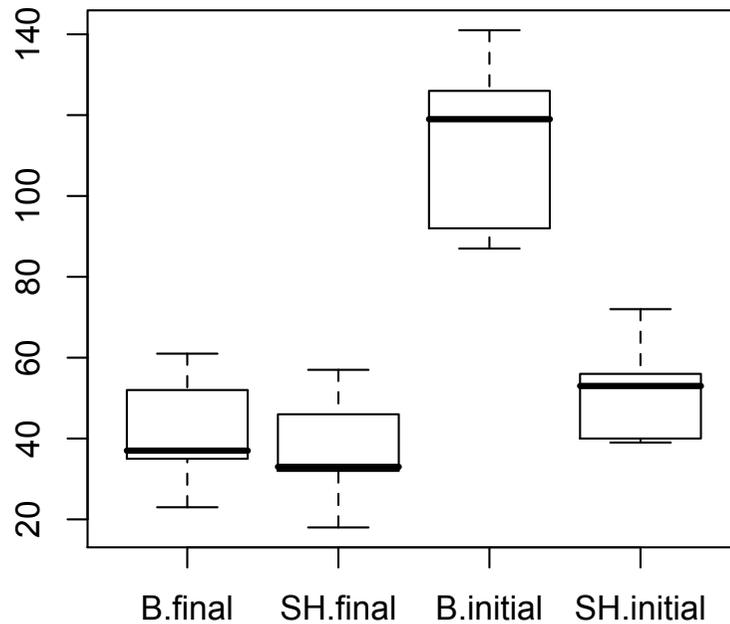
Inwiefern wird die Dauer von der Position und/oder Dialekt beeinflusst?

```
dr = read.table(file.path(pfad, "dr.txt"))
```

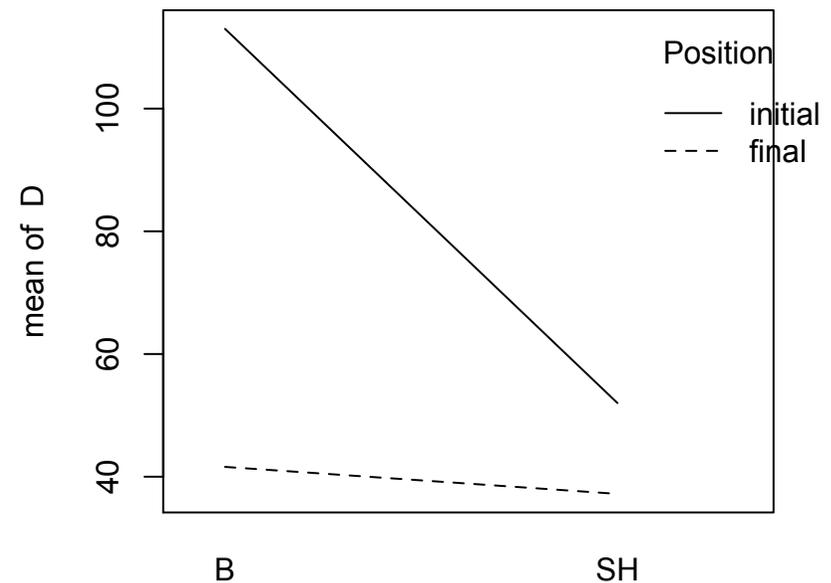
RM-Anova: between and within

Abbildungen

```
boxplot(D ~ Dialekt * Position,  
data=dr)
```



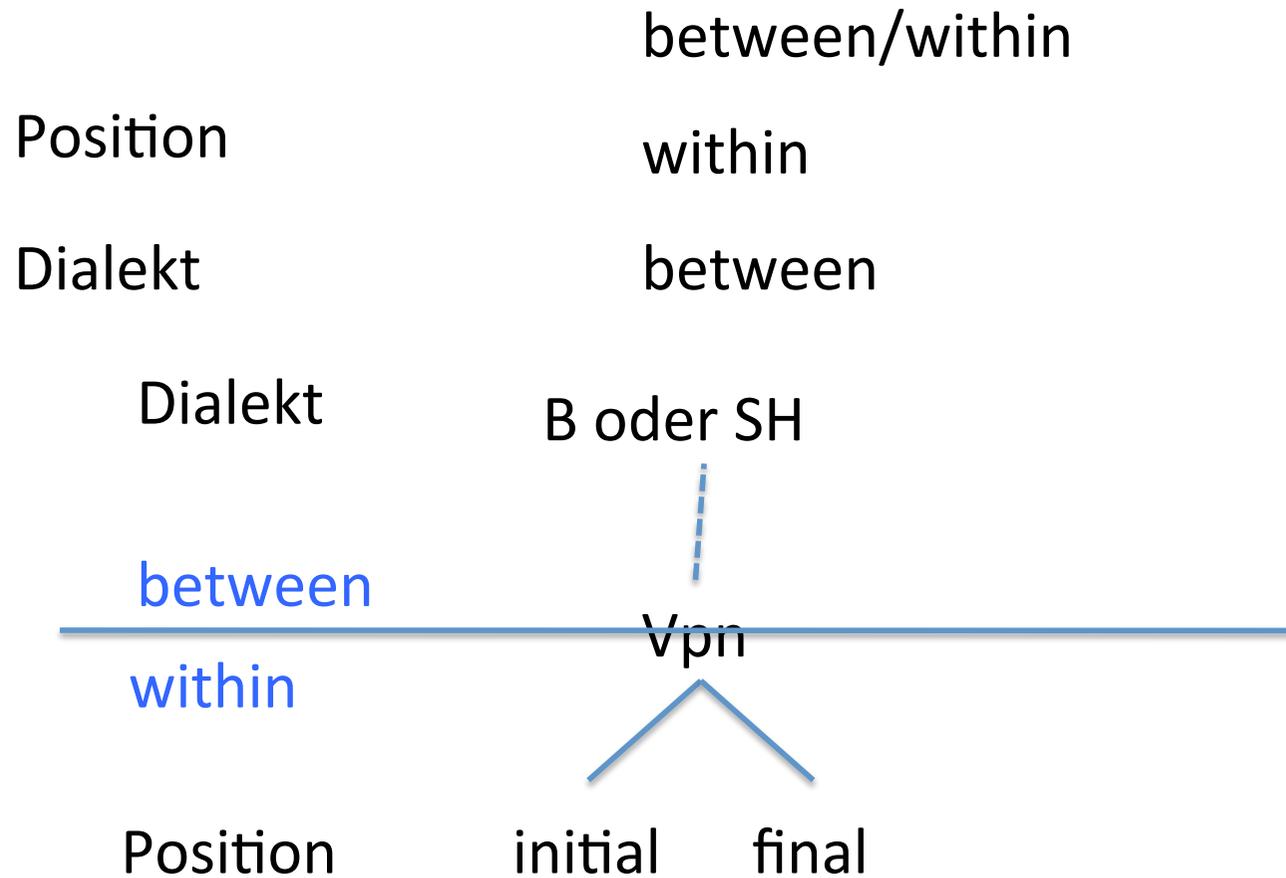
```
with(dr, interaction.plot(Dialekt,  
Position, D))
```



Position signifikant? Dialekt signifikant?

Interaktion?

RM-Anova: between and within



`dr.ez = ezANOVA(dr, .(D), .(Vpn), .(Position), .(Dialekt))`

↑
within

↑
between

\$ANOVA

	Effect	DFn	DFd	SSn	SSd	F	p	p<.05
1	Dialekt	1	8	5346.45	3860	11.08073	1.040338e-02	*
2	Position	1	8	9288.05	754	98.54695	8.964643e-06	*
3	Dialekt:Position	1	8	4004.45	754	42.48753	1.845250e-04	*

Dialekt ($F[1, 8]=11.1$, $p < 0.05$) und Position ($F[1, 8] = 98.6$, $p < 0.001$) hatten einen signifikanten Einfluss auf die Dauer und es gab eine signifikante Interaktion ($F[1, 8]=42.5$, $p < 0.001$) zwischen diesen Faktoren.

post-hoc Tests

```
source(file.path(pfad, "phoc.txt"))
```

Für einen RM-Anova kann **ein post-hoc t-test mit Bonferroni Korrektur** angewandt werden.

Je mehr Tests wir post-hoc anwenden, um so wahrscheinlicher ist es, dass wir Signifikanzen per Zufall bekommen werden. Der Tukey (Anova ohne Messwiederholungen) und Bonferroni-adjusted t-Tests (mit Messwiederholungen) sind Maßnahmen dagegen.

Bonferroni-Korrektur: Der Wahrscheinlichkeitswert der individuellen Tests wird mit der **Anzahl der theoretisch möglichen Testkombinationen** multipliziert.

Post-hoc t-test mit Bonferroni Korrektur

1. t-tests aller Stufen-Kombinationen durchführen: als **gepaart** mit denselben Between-Stufen, sonst ungepaart.

SH-initial mit SH-final

g

SH-final mit B-initial

SH-initial mit B-initial

SH-final mit B-final

SH-initial mit B-final

B-initial mit B-final

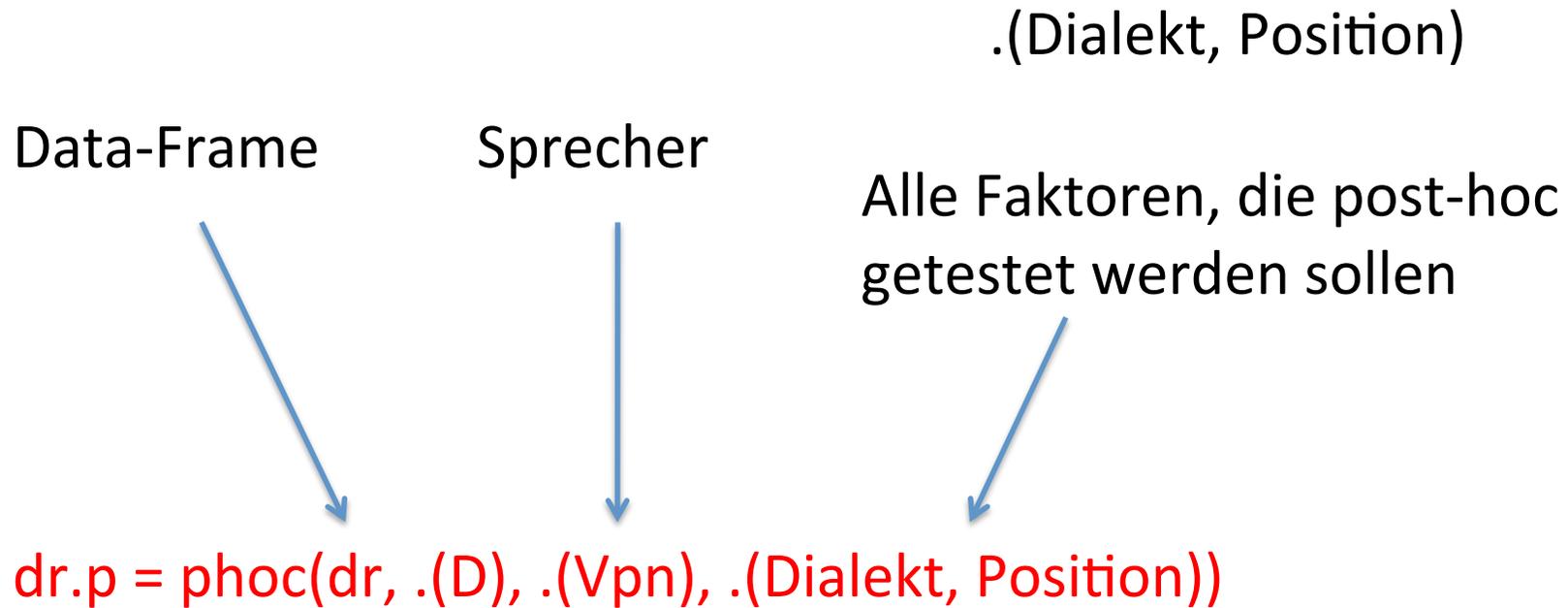
g

2. Bonferroni Korrektur: den Wahrscheinlichkeitswert eines t-tests mit der Anzahl der Tests multiplizieren

zB wenn SH-initial vs SH-final $p = 0.035$, Bonferroni-Korrektur:
 $0.035 * 6 = 0.21$ (weil es 6 mögliche Testpaare gibt).

3. Auswahl: nur die Test-Kombinationen, die sich in **einer Stufe unterscheiden**. Funktion `phsel()`

Post-hoc t-test mit Bonferroni Korrektur



Ergebnisse: auch in dr.p[[1]]

\$res

	t	df	prob-adj
SH:initial-SH:final	2.5709017	4.000000	0.371518380
SH:initial-B:initial	-5.1226150	6.475584	0.010372660
SH:initial-B:final	1.1537054	7.918185	1.000000000
SH:final-B:initial	-6.2006294	6.852279	0.002905609
SH:final-B:final	-0.4666613	7.999611	1.000000000
B:initial-B:final	10.9833157	4.000000	0.002342832

wurde ein gepaarter t-test durchgeführt?

\$paired

[1] TRUE FALSE FALSE FALSE FALSE TRUE

\$bonf

[1] 6

Bonferroni-Multiplikator (alle
Wahrscheinlichkeiten des t-Tests
wurden mit diesem Wert multipliziert)

Ergebnisse auswählen, die sich in einer Stufe unterscheiden

immer [[1]], da die Ergebnisse in dr.p[[1]] sind

phsel(dr.p[[1]]) Position konstant

t	df	prob-adj
SH:initial-B:initial	-5.1226150	6.475584 0.01037266
SH:final-B:final	-0.4666613	7.999611 1.00000000

phsel(dr.p[[1]], 2) Dialekt konstant

t	df	prob-adj
SH:initial-SH:final	2.570902	4 0.371518380
B:initial-B:final	10.983316	4 0.002342832

Inwiefern wird die Dauer von der Position und/
oder Dialekt beeinflusst?

	stat	df	Bonferroni	p
SH.final-B.final	-0.4666613	7.999611	1.000000000	
SH.final-SH.initial	-2.5709017	4.000000	0.371518380	
B.final-B.initial	-10.9833157	4.000000	0.002342832	
SH.initial-B.initial	-5.1226150	6.475584	0.010372660	

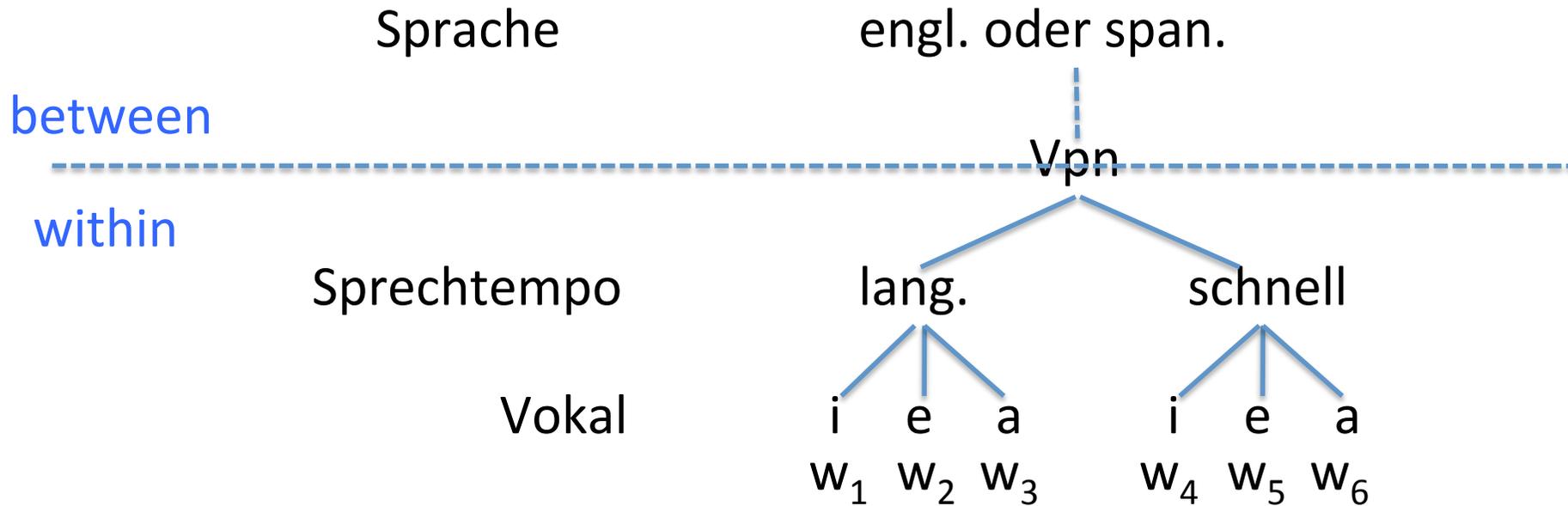
Post-hoc t-Tests mit Bonferroni-Korrektur zeigten signifikante Unterschiede zwischen Bayern und Schleswig-Holstein in initialer ($p < 0.05$) jedoch nicht in finaler Position. Die Unterschiede zwischen initialer und finaler Position waren nur für Bayern ($p < 0.01$) jedoch nicht für Schleswig-Holstein signifikant.

Wiederholungen in derselben Zelle

Sphericity Korrektur

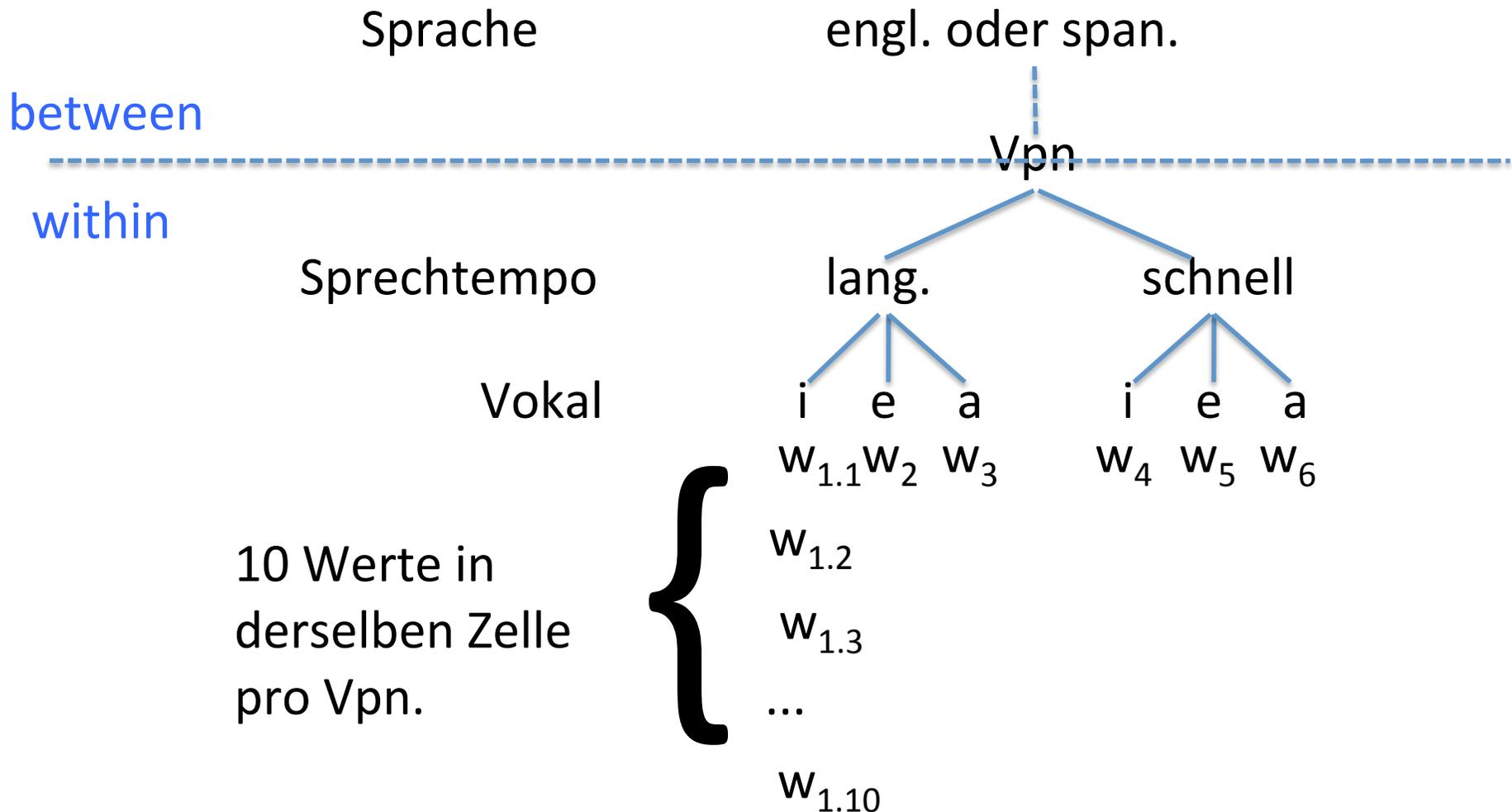
Wiederholungen in derselben Zelle

In allen bislang untersuchten ANOVAs gab es **einen Wert pro Vpn. pro Zelle**. z.B. 2 Faktoren mit 3 und 2 Stufen, dann 6 Werte pro Vpn, also einen Wert pro Stufen-Kombination pro Vpn.

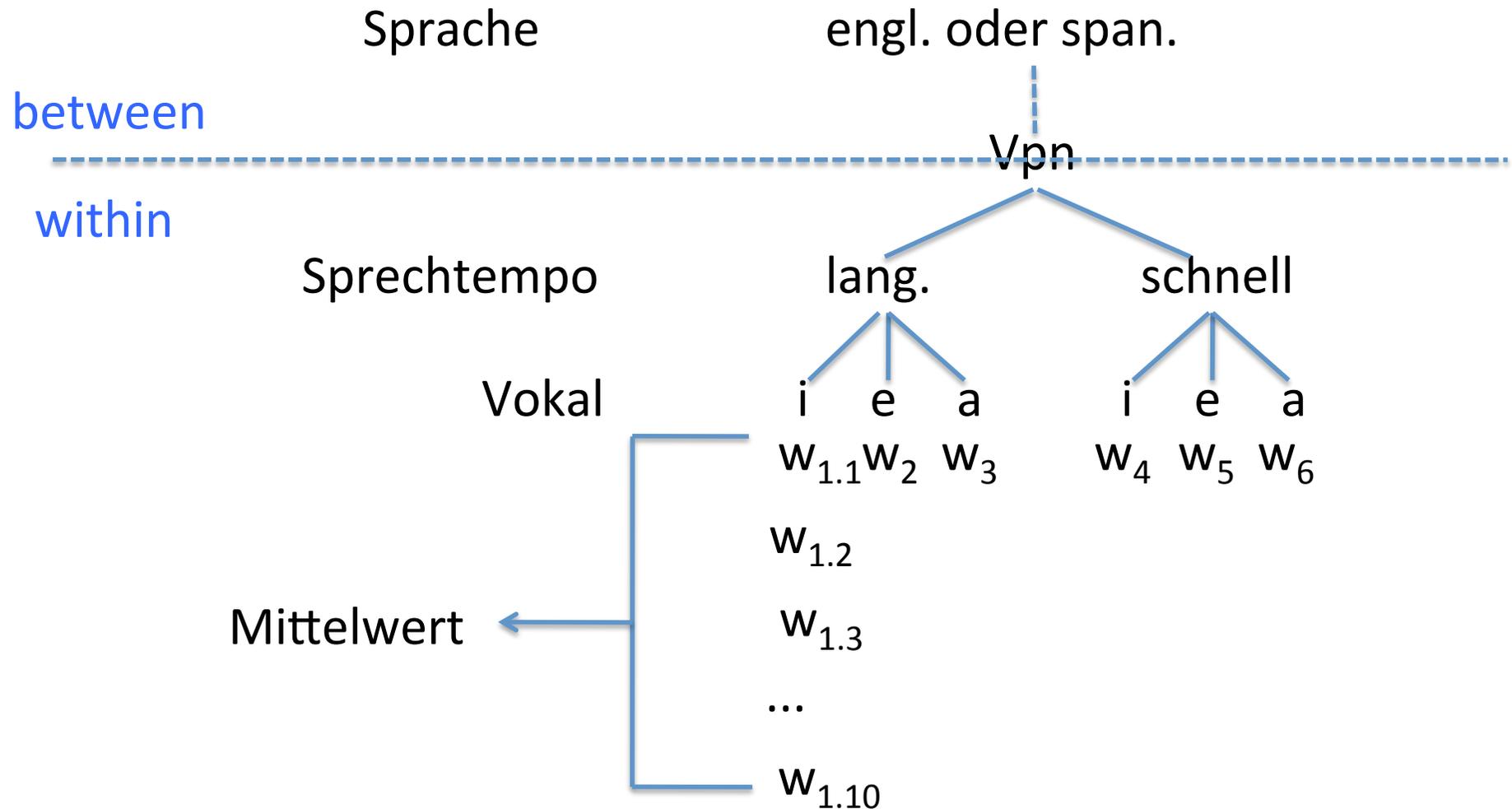


Wiederholungen in derselben Zelle

Jedoch haben die meisten phonetischen Untersuchungen **mehrere Werte pro Zelle**. zB. jede Vpn. erzeugte 'hid', 'head', 'had' zu einer langsamen und schnellen Sprechgeschwindigkeit **jeweils 10 Mal**.



Wiederholungen innerhalb der Zelle in einem ANOVA sind nicht zulässig und müssen gemittelt werden – damit wir pro Vpn. **einen within-subjects Wert pro Kombination der within-subjects Stufen** haben (6 Mittelwerte pro Vpn. in diesem Beispiel).



Wiederholungen in derselben Zelle

```
ssb = read.table(file.path(pfad, "ssb.txt"))
```

In einer Untersuchung zur /u/-Frontierung im Standardenglischen wurde von **12 Sprecherinnen** (6 alt, 6 jung) F2 zum zeitlichen Mittelpunkt in drei verschiedenen /u/-Wörtern erhoben (*used*, *swoop*, *who'd*). Jedes Wort ist von jeder Vpn. 10 Mal erzeugt worden. Ist /u/ in den jungen Vpn. frontierter? (bis zu 60 Werte pro Vpn).

Faktor	within/between	wieviele Stufen?
Word	within	3
Alter	between	2

Wieviele Werte pro Vpn. dürfen in der ANOVA vorkommen? **3**

Wieviele Werte insgesamt in der ANOVA wird es geben? **36**

Wiederholungen in derselben Zelle

1. Anzahl der Wort-Wiederholungen pro Sprecher prüfen

`with(ssb, table(Wort, Vpn))`

Wort	arkn	elwi	frwa	gisa	jach	jeny	kapo	mapr	nata	rohi	rusy	shle
swoop	10	9	10	10	10	10	10	10	10	10	10	10
used	10	10	10	10	10	10	10	10	10	10	10	10
who'd	10	10	10	10	10	10	10	10	10	10	10	10

2. Über die Wort-Wiederholungen mit `aggregate()` mitteln

abhängige Variable alle Faktoren

`ssbm = with(ssb, aggregate(F2, list(Alter, Wort, Vpn), mean))`



Wiederholungen in derselben Zelle

```
ssbm = with(ssb, aggregate(F2, list(Alter, Wort, Vpn), mean))
```

```
dim(ssbm)
```

```
[1] 36 4
```

```
head(ssbm)
```

```
  Group.1 Group.2 Group.3      x  
1      alt  swoop   arkn 10.527359
```

3. Neue Namen vergeben

```
names(ssbm) = c("Alter", "Wort", "Vpn", "F2")
```

4. RM-Anova wie üblich durchführen

```
ezANOVA(ssbm, .(F2), .(Vpn), .(Wort), .(Alter))
```

Sphericity-Korrektur

Sphericity ist die Annahme in einem RM-Anova, dass die Varianzen der Unterschiede zwischen den Stufen eines within-subject-Faktors gleich sind.

Wenn Sphericity nicht gegeben ist, werden die Wahrscheinlichkeiten durch Änderungen in den Freiheitsgraden nach oben gesetzt.

Dieses Problem tritt nur auf wenn ein within-subjects-Faktor mehr als 2 Stufen hat.

Man soll grundsätzlich immer für Sphericity korrigieren, wenn Sphericity-Korrektur in der Ausgabe von `ezANOVA()` erscheint.

Sphericity-Korrektur

\$ANOVA

	Effect	DFn	DFd	SSn	SSd	F	p	p<.05	pes
1	Alter	1	10	61.394752	41.268353	14.876957	3.175409e-03	*	0.5980216
2	Wort	2	20	67.210301	8.561218	78.505534	3.390750e-10	*	0.8870127
3	Alter:Wort	2	20	8.467805	8.561218	9.890888	1.031474e-03	*	0.4972572

\$`Mauchly's Test for Sphericity` (Ignorieren, da es nicht zuverlässig ist)

	Effect	W	p	p<.05
2	Wort	0.5423826	0.06373468	
3	Alter:Wort	0.5423826	0.06373468	

\$`Sphericity Corrections`

	Effect	GGe	p[GG]	p[GG]<.05	HFe	p[HF]	p[HF]<.05
2	Wort	0.6860511	1.340736e-07	*	0.7587667	3.342362e-08	*
3	Alter:Wort	0.6860511	4.370590e-03	*	0.7587667	3.120999e-03	*

1. Die **betreffenden Freiheitsgrade** werden mit dem **Greenhouse-Geisser-Epsilon** multipliziert, wenn er unter 0.75 liegt, sonst mit dem Huynh-Feldt-Epsilon.

Wort: $F[2,20] \rightarrow F[2 * 0.6860511, 20 * 0.6860511] = F[1.37, 13.72]$

Alter x Wort Interaktion: $F[1.37, 13.72]$

Sphericity-Korrektur

\$ANOVA

	Effect	DFn	DFd	SSn	SSd	F	p	p<.05	pes
1	Alter	1	10	61.394752	41.268353	14.876957	3.175409e-03	*	0.5980216
2	Wort	2	20	67.210301	8.561218	78.505534	3.390750e-10	*	0.8870127
3	Alter:Wort	2	20	8.467805	8.561218	9.890888	1.031474e-03	*	0.4972572

\$`Sphericity Corrections`

	Effect	GGe	p[GG]	p[GG]<.05	HFe	p[HF]	p[HF]<.05
2	Wort	0.6860511	1.340736e-07	*	0.7587667	3.342362e-08	*
3	Alter:Wort	0.6860511	4.370590e-03	*	0.7587667	3.120999e-03	*

2. Die neuen damit verbundenen Wahrscheinlichkeiten sind $p[GG]$ (wenn mit GGe multipliziert wurde) sonst $p[HF]$.

Das sind die Wahrscheinlichkeiten mit den korrigierten Freiheitsgraden
z.B. $1 - pf(9.8908882, 2 * 0.6860511, 20 * 0.6860511)$

[1] 0.004370589

Alter ($F[1,10] = 14.9$, $p < 0.001$), Wort ($F[1.37, 13.72] = 78.5$, $p < 0.001$) sowie die Interaktion Wort und Alter ($F[1.37, 13.72] = 9.9$, $p < 0.001$) hatten einen signifikanten Einfluss auf F2.