

Logistische Regression und die Analyse von Proportionen

Jonathan Harrington

```
library(lme4)  
library(lattice)  
library(multcomp)  
source(file.path(pfadu, "phoc.txt"))
```

Logistische Regression und Proportionen

Mit der logistischen Regression wird geprüft, inwiefern die proportionale Verteilung in binären Kategorien (also 2 Stufen) von einem (oder von mehreren) Faktoren beeinflusst wird.

Der abhängige Faktor ist **immer binär**, zB:

glottalisiert vs. nicht-glottalisiert

lenisiert vs. nicht-lenisiert

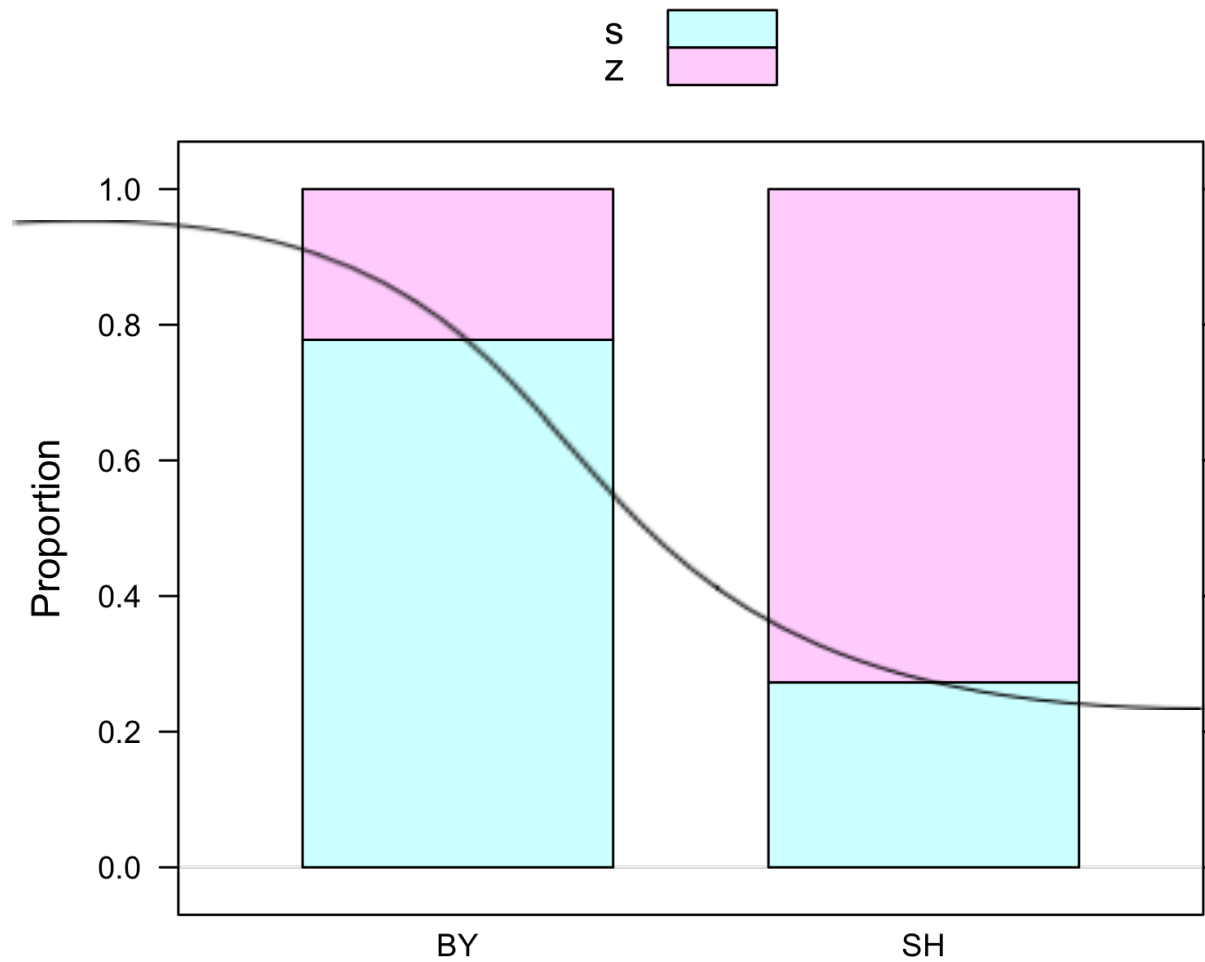
geschlossen vs. offen

ja vs. nein

True vs. False usw.

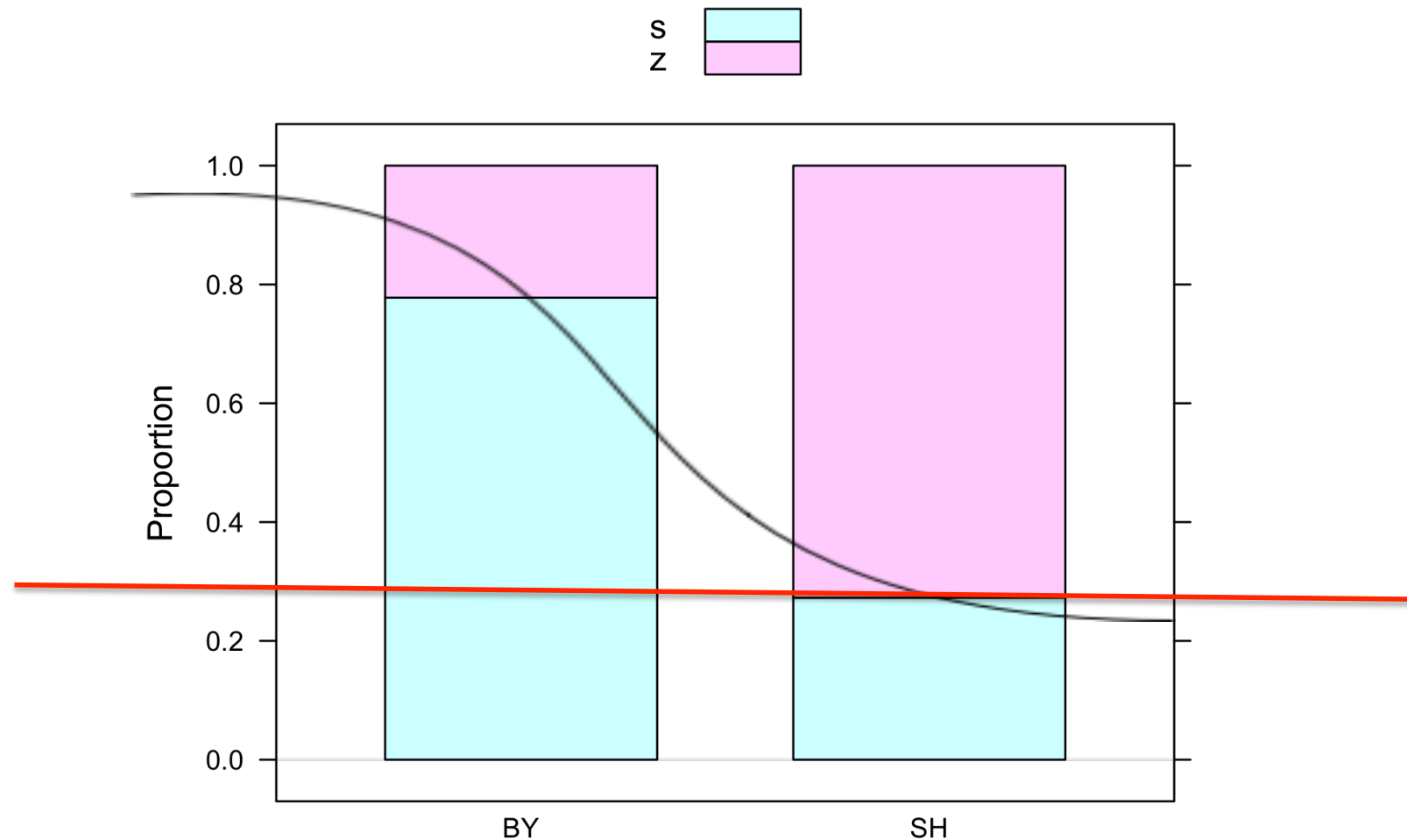
Logistische Regression und Wahrscheinlichkeiten

In der logistischen Regression wird eine sogenannte Sigmoid-Funktion an Proportionen angepasst:



Logistische Regression und Wahrscheinlichkeiten

Die Wahrscheinlichkeit wird geprüft, dass die Sigmoid-Neigung **0** (Null) sein könnte



Denn wenn die Neigung 0 ist (= eine gerade Linie) unterscheiden sich auch nicht die Proportionen

Logistische Regression in R

af, FF, RF: Abhängiger/Fixed/Random Faktor

(N.B: af ist immer binär also mit 2 Stufen)

1. Abbildung

```
tab = table(FF, af)
```

```
prop = prop.table(tab, 1:n)
```

```
barchart(prop, auto.key=T)
```

2. Modell

ohne RF

mit RF

```
glm(af ~ FF, family=binomial)      lmer(af ~ FF + (1 | RF), family=binomial)
```

3. Signifikanz FF

```
anova()
```

(4. post-hoc Test)

FF hat mehr als 2 Stufen;
oder mehr als 1 FF

```
glht()
```

1. Abbildung

```
sz = read.table(file.path(pfadu, "sz.txt"))
```

Inwiefern beeinflusst Dialekt die Wahl zwischen/s, z/?

Abbildung: Häufigkeiten

...oder Proportionen

FF



af (immer an letzter Stelle)

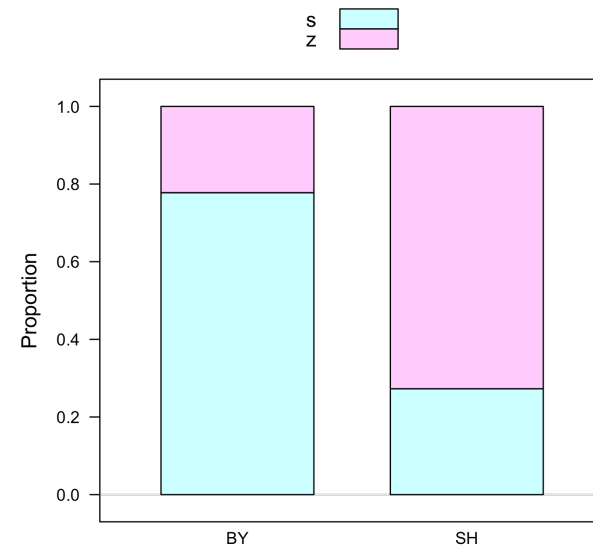
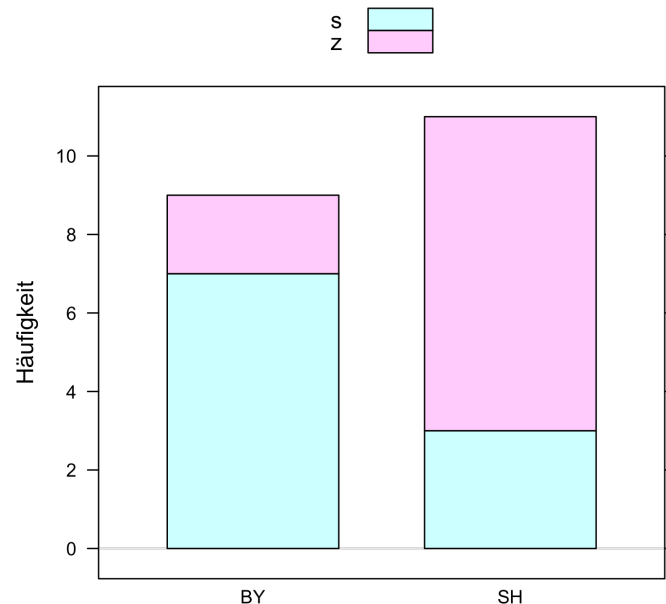


```
tab = with(sz, table(Dialekt, Frikativ))
```

```
prop = prop.table(tab, 1)
```

```
barchart(tab, auto.key=T,  
horizontal=F, ylab="Häufigkeit")
```

```
barchart(prop, auto.key=T,  
horizontal=F, ylab="Proportion")
```



2. Test: hat FF (Dialekt) einen Einfluss auf die Proportionen?

```
o = glm(Frikativ ~ Dialekt, family=binomial, data = sz)
anova(o, test="Chisq")
```

	Df	Deviance	Resid.	Df	Resid.	Dev	Pr(>Chi)
Dialekt	1	5.3002		18	22.426	0.02132	*

Das gleiche

Ohne FF

```
ohne = glm(Frikativ ~ 1, family=binomial, data = sz)
```

oder

```
ohne = update(o, ~. -Dialekt)
```

Dialekt hat einen signifikanten Einfluss auf Frikativ d.h.
auf die s/z Verteilung ($\chi^2[1] = 5.3, p < 0.05$)

zweites Beispiel: Abbildung

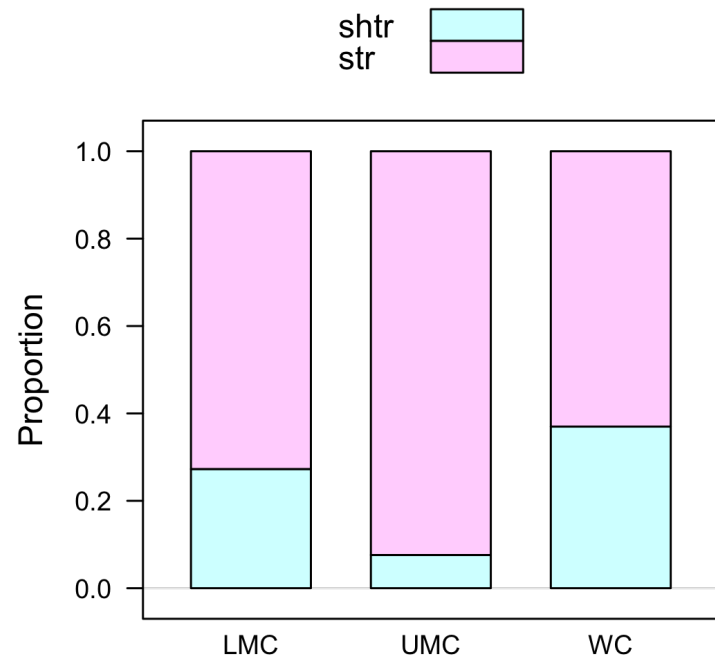
```
coronal = read.table(file.path(pfadu, "coronal.txt"))
```

Inwiefern wird die Verteilung [ftr] vs [str] von der Sozialklasse beeinflusst? (modifiziert aus Johnson, 2008)

```
tab = with(coronal, table(Socialclass, Fr))
```

```
prop = prop.table(tab, 1)
```

```
barchart(prop, auto.key=T, horizontal=F, ylab="Proportion")
```



Test: FF und post-hoc Test

Test

```
o = glm(Fr ~ Socialclass, family = binomial, data = coronal)
```

```
anova(o, test="Chisq")
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
Socialclass	2	21.338	237	241.79	2.326e-05 ***

Post-hoc Test (da FF mehr als 2 Stufen hat)

```
summary(glht(o, linfct=mcp(Socialclass="Tukey")))
```

Linear Hypotheses:

	Estimate	Std. Error	z value	Pr(> z)	
UMC - LMC == 0	1.5179	0.4875	3.114	0.00501	**
WC - LMC == 0	-0.4480	0.3407	-1.315	0.38142	
WC - UMC == 0	-1.9659	0.4890	-4.020	< 0.001	***

Sozialklasse hatte einen signifikanten Einfluss auf die [tr] vs [str] Verteilung ($\chi^2[2] = 21.3$, $p < 0.001$). Post-hoc Tukey-Tests zeigten signifikante Unterschiede zwischen UMC und LMC ($p < 0.01$) und zwischen UMC und WC ($p < 0.001$) jedoch nicht zwischen WC und LMC.

Drittes Beispiel: numerischer FF

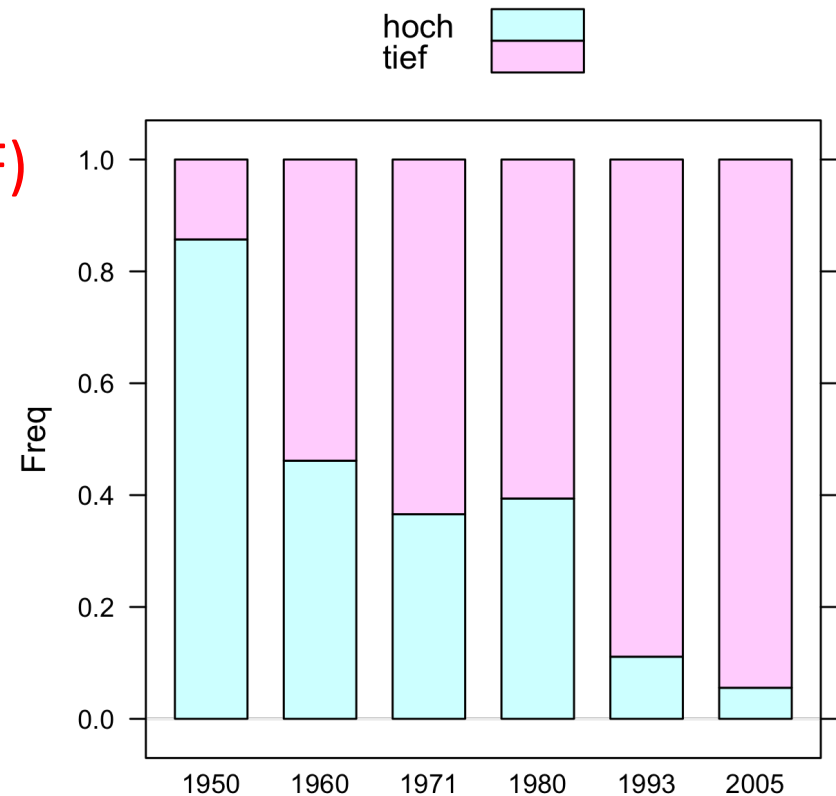
```
ovokal = read.table(file.path(pfadu, "ovokal.txt"))
```

Zwischen 1950 und 2005 wurde der Vokal in *lost* entweder mit hohem /o:/ oder tieferem /ɔ/ gesprochen. Ändert sich diese Proportion mit der Zeit?

```
tab = with(ovokal, table(Jahr, Vokal))
```

```
prop = prop.table(tab, 1)
```

```
barchart(prop, auto.key=T, horizontal=F)
```



Test

```
o = glm(Vokal ~ Jahr, family=binomial, data = ovokal)
anova(o, test="Chisq")
```

Df	Deviance	Resid.	Df	Resid.	Dev	Pr(>Chi)
1	61.121		218	229.45		5.367e-15 ***

Die Wahl (ob /o/ oder /ɔ/) wird signifikant vom Jahr beeinflusst ($\chi^2[1] = 61.1$, $p < 0.001$)

(keine Post-hoc Tests möglich, wenn wie hier der FF numerisch ist)

Viertes Beispiel: mit Random Faktor

daher `lmer()` statt `glm()`

```
pr = read.table(file.path(pfadu, "preasp.txt"))
```

(Daten von Mary Stevens). Es wurde im Italienischen festgestellt, ob vor einem Plosiv präaspiert wurde oder nicht (af = Pre). Inwiefern hat der davor kommende Vokal (FF = vtype) einen Einfluss auf diese Verteilung?

Werte in mehreren Stufen desselben Faktors pro Sprecher

```
with(pr, table(spkr, vtype, Pre))
```

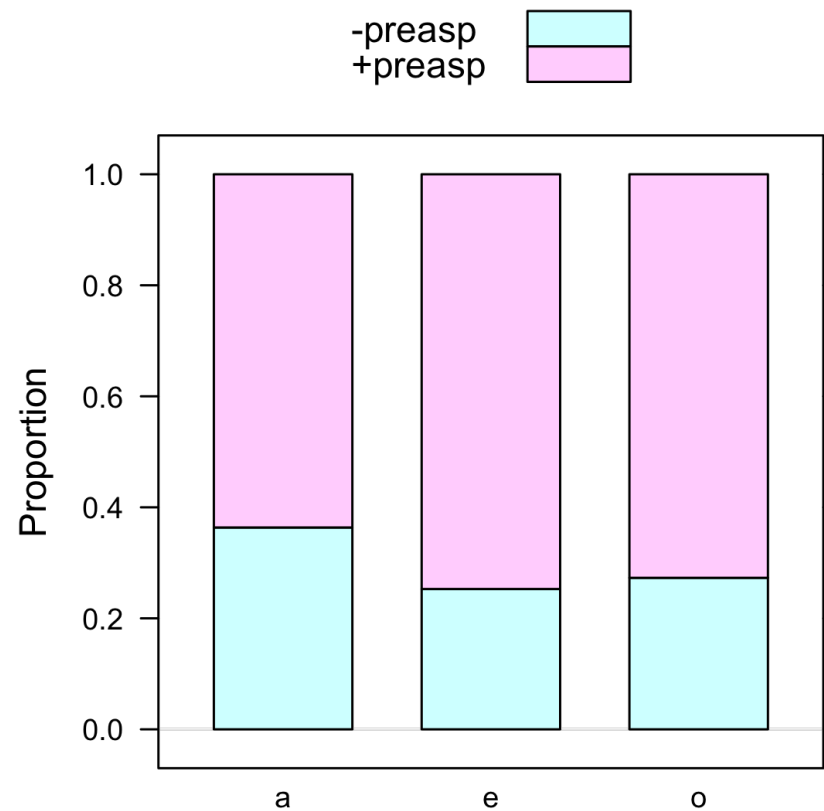
Wir wollen diese Variabilität, die wegen des Sprechers entsteht, herausklammern (daher `lmer(...(1 | spkr))`)

Abbildung

```
tab = with(pr, table(vtype, Pre))
```

```
prop = prop.table(tab, 1)
```

```
barchart(prop, auto.key=T, horizontal=F)
```



Test

```
o = lmer(Pre ~ vtype + (1|spk), family=binomial, data = pr)
ohne = update(o, ~ . - vtype)
anova(o, ohne)
```

	Df	AIC	BIC	logLik	Chisq	Chi	Df	Pr(>Chisq)	
o	4	1060.0	1079.3	-525.98	10.8		2	0.004517	**

post-hoc Test, da $UF > 2$ Stufen hat

```
summary(glht(o, linfct=mcp(vtype="Tukey")))
```

Linear Hypotheses:

	Estimate	Std. Error	z value	Pr(> z)	
e - a == 0	0.6560	0.1979	3.314	0.00269	**
o - a == 0	0.5012	0.1961	2.556	0.02856	*
o - e == 0	-0.1547	0.1848	-0.838	0.67941	

Die Verteilung von $\pm$ Präaspiration wurde vom davor kommenden Vokal signifikant beeinflusst ($\chi^2[2] = 10.8$, $p < 0.01$). Post-hoc Tukey-Tests zeigten signifikante Unterschiede in der $\pm$ Präaspiration-Verteilung zwischen /e, a/ ($p < 0.01$) und zwischen /o, a/ ($p < 0.05$), jedoch nicht zwischen /o, e/.

Zwei Fixed Faktoren

1. Abbildung

table(), prop.table(), barchart()

2. Modell

ohne RF

glm()

mit RF

lmer()

3. Gibt es eine Interaktion?

update()

4: Wenn ja, Faktoren kombinieren

interaction(), glht()

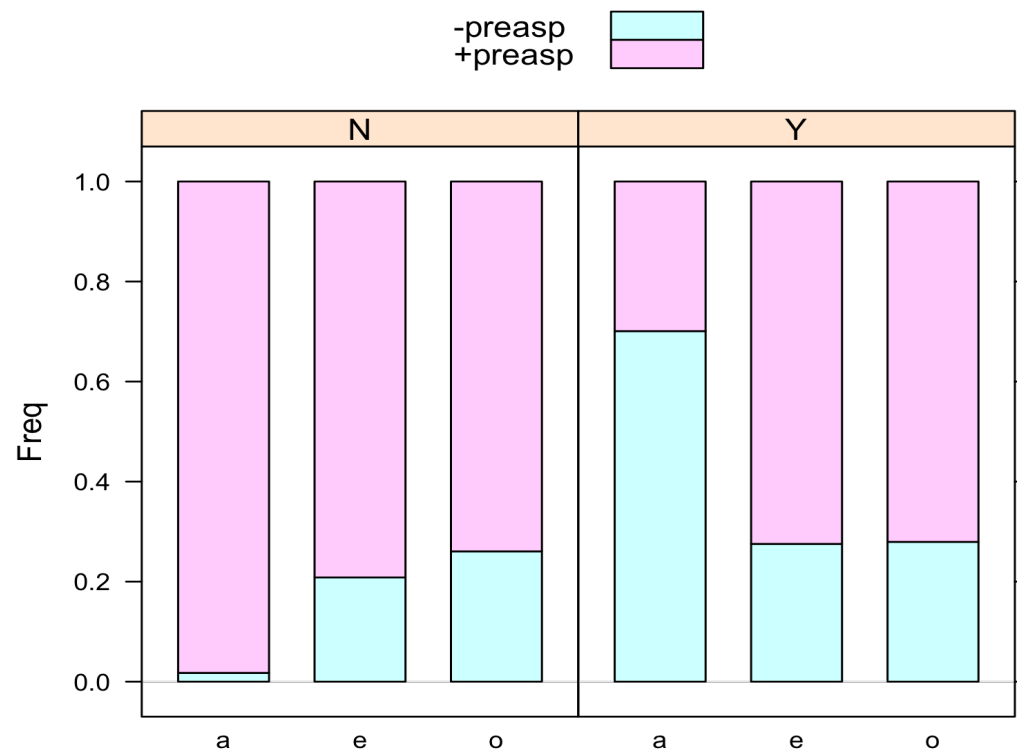
Zwei unabhängige (fixed) Faktoren

Inwiefern wird die Preäspiration vom Vokal und von Pretonic (ob die nächste Silbe betont war oder nicht) beeinflusst?

`tab = with(pr, table(vtype, ptonic, Pre))` (Pre an letzter Stelle)

`prop = prop.table(tab, 1:2)` (1:n bei n Faktoren)

`barchart(tab, auto.key=T, horizontal = F)`



Vokal sig?
Pretonic sig?
Interaktion?

Zwei Fixed Faktoren

1. Interaktion prüfen

```
o = lmer(Pre ~ vtype * ptonic + (1|spk), family=binomial, data=pr)
```

```
ohne = update(o, ~ . -vtype:ptonic)
```

```
anova(o, ohne)
```

Chisq	Chi	Df	Pr(>Chisq)
114.92		2	< 2.2e-16 ***

2. Wenn eine Interaktion vorliegt, dann Faktoren kombinieren

```
plabs = with(pr, interaction(vtype, ptonic))
```

3. Modell

```
beide = lmer(Pre ~ plabs + (1|spk), family=binomial, data=pr)
```

post-hoc Test

```
p = summary(glht(beide, linfct=mcp(plabs = "Tukey")))
```

```
round(phsel(p), 3)          # Faktor 1
```

```
round(phsel(p, 2), 3)      # Faktor 2
```

	Df	AIC	BIC	logLik	Chisq	Chi	Df	Pr(>Chisq)
o	7	886.09	919.98	-436.05	114.92		2	< 2.2e-16 ***

Es gab eine signifikante Interaktion zwischen den Faktoren ($\chi^2[2] = 114.9$, $p < 0.001$)

`round(phsel(p), 3)`

	z value	Adjusted p values
e.N - a.N	-3.691	0.003
o.N - a.N	-4.250	0.000
o.N - e.N	-1.346	0.737
e.Y - a.Y	8.745	0.000
o.Y - a.Y	8.554	0.000
o.Y - e.Y	-0.210	1.000

`round(phsel(p, 2), 3)`

	z value	Adjusted p values
a.Y - a.N	-7.278	0.000
e.Y - e.N	-1.851	0.403
o.Y - o.N	-0.506	0.995

Post-hoc Tukey-Tests zeigten, dass die $[\pm\text{preasp}]$ proportionale Verteilung (d.h. ob präaspiert wurde oder nicht) sich in */e/ vs. /a/* ($p < 0.01$) und in */o/ vs /a/* ($p < 0.001$) aber *nicht in /o/ vs /e/* unterschied. Es gab auch signifikante Unterschiede in der $\pm[\text{preasp}]$ proportionale Verteilung *in /a/* ($p < 0.001$) jedoch nicht in in */e/* noch */o/* Vokalen.