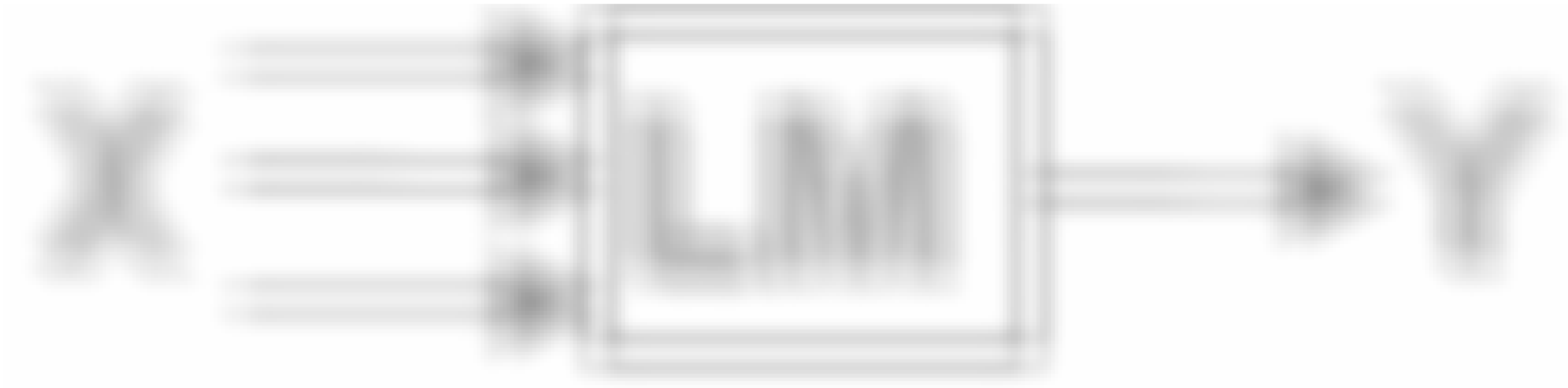


Intro to categorical data analysis in R

Anova over proportion vs. ordinary and mixed logit models

T. Florian Jaeger (tiflo@bcs.rochester.edu)
Brain & Cognitive Sciences, University of Rochester

Today

- Recap of ANOVA's assumption
- Example for ANOVA over proportions
- Example for Logistic Regression
- Example for Mixed Logit Model

Assumes:

- Normality of dependent variable within levels of factors
- Linearity
- (Homogeneity of variances)
- Independence of observations → leads to F1, F2

Designed for balanced data

- Balanced data comes from balanced designs, which has other desirable properties

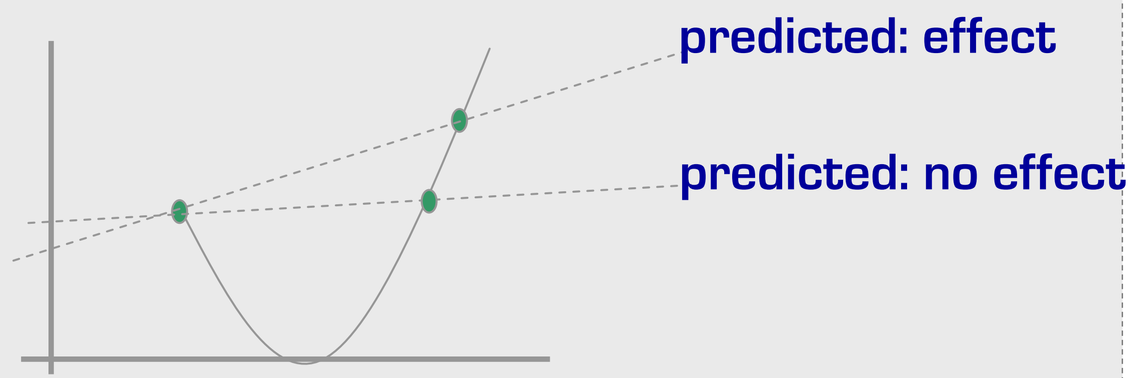
- ANOVA can be seen as a special case of linear regression

- Linear regression makes more or less the same assumptions, but does not require balanced data sets
  - Deviation from balance brings the danger of collinearity (different factors explaining the same part of the variation in the dep.var.) → inflated standard errors → spurious results
  - But, as long as collinearity is tested for and avoided, linear regression can deal with unbalanced data

- Unbalanced data sets are common in corpus work and less constrained experimental designs

- Generally, more naturalistic tasks result in unbalanced data sets (or high data loss)

- ANOVA designs are usually restricted to categorical independent variables → binning of continuous variables (e.g. high vs. low frequency) →
  - Loss of power (Baayen, 2004)
  - Loss of understanding of the effect (is it linear, is it log-linear, is it quadratic?):

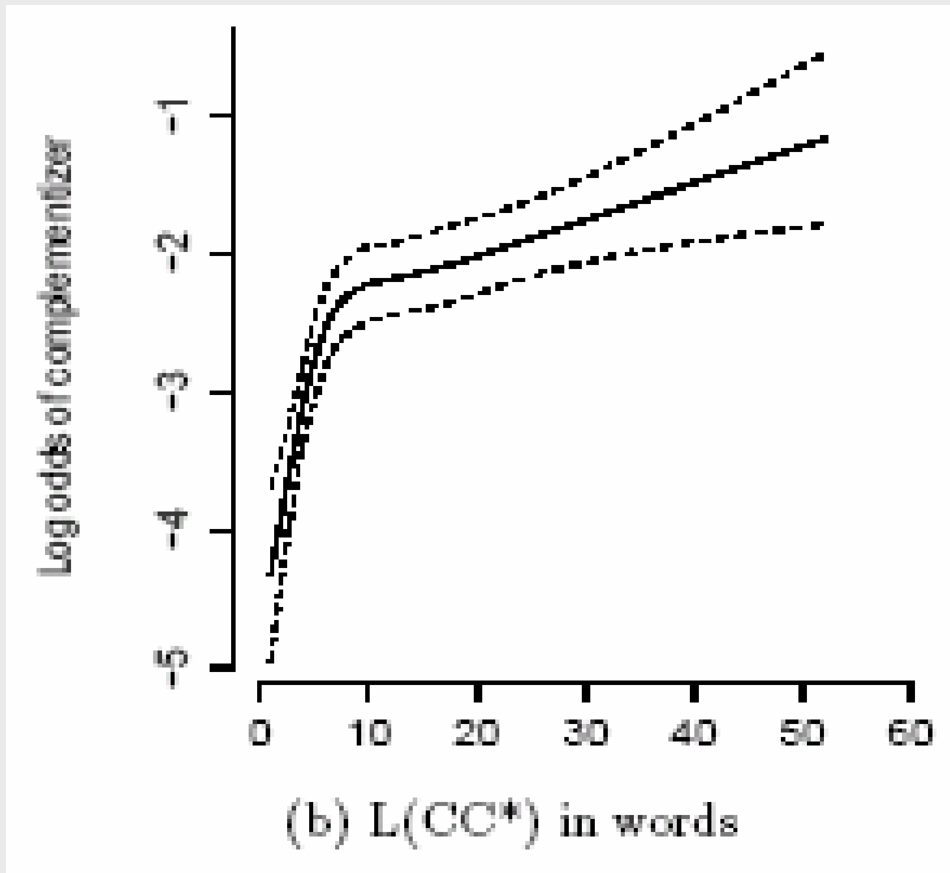


  - E.g. speech rate has a quadratic effect on phonetic reduction; dual-route mechanisms lead to non-linearity

- Regressions (Linear Models, Generalized Linear Models) are well-suited for the inclusion of *continuous predictors*
- R comes with tools to test linearity (e.g. `rcs()`, `pol()` in Design library)

- Example: effect of CC-length on *that*-mentioning:

He really believes (that) he's not drunk.



- Another shortcoming of ANOVA is that it is limited to continuous outcomes

- Often ignored as a minor problem → ANOVAs performed over percentages (derived by averaging over subjects/items)

Proportion ← Categorical variable (e.g. either 0 or 1)

```
i.F1<- aggregate(i[,c('CorrectResponses')],
  by= list(subj= …, condition= …),
  FUN= mean)

F1<- aov(CorrectResponses ~ condition +
      Error(subj/(condition)), i.F1)
```

This is unsatisfying:

- Doesn't scale to categorical dependent variables with multiple outcomes (e.g. multiple choice answers; priming: no prime vs. prime structure A vs. prime structure B)

- Violates assumption of homogeneity of variances
  - Leads to **spurious results**, because percentages are not the right space

- **Logistic regression**, a type of Generalized Linear Model (a generalization over linear regressions), addresses these problems

$E[Y] = X\beta \leftrightarrow Y = X\beta + e$

$Y \sim N(X\beta, \sigma^2) \leftrightarrow Y = X\beta + e,\ e \sim N(0, \sigma^2)$

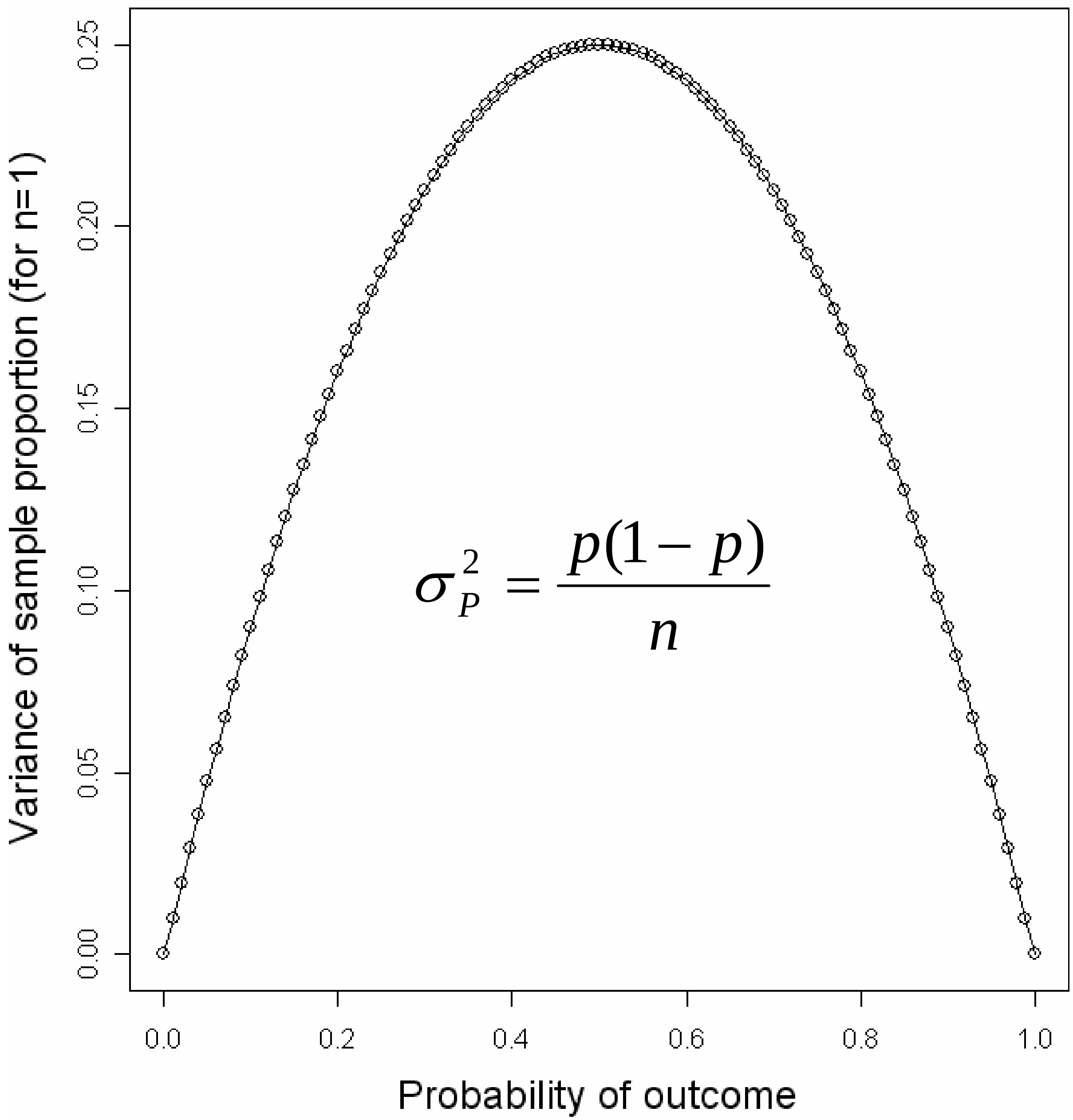
- Intuitively, why aren't percentages the right space?
  - Can lead to **un-interpretable results**: below or above 0 ... 100% (b/c CIs lie outside [0,1])
  - Simple question: how could a 10% effect occur if the baseline is already 95%?

- Change in percentage around 50% is less of a change than change close to 0 or 100%
  - E.g., going from 50 to 60% correct answers is only **20% error reduction**, but going from 85 to 95% is a **67% error reduction**

→effects close to 0 or 100% are underestimated, those close to 50% are overestimated

More formally,

- ANOVA over proportions of violate the assumption of homogeneity of variances



$$\sigma_P^2 = \frac{p(1-p)}{n}$$

- In what space can we avoid these problems?

  → **odds** = p / (1 – p) from [0; ∞];

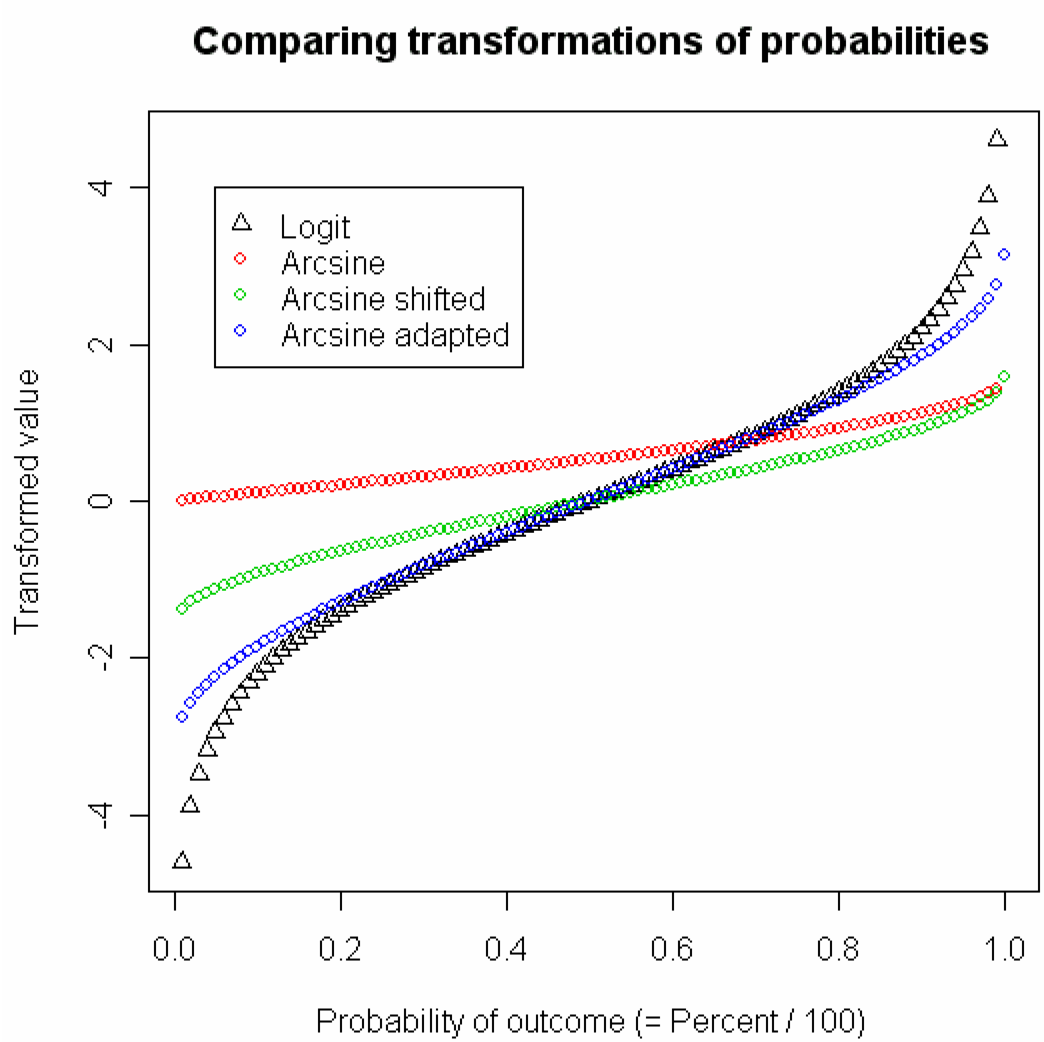
  Multiplicative scale but regressions are based on sums

  → Logit: **log-odds** = log(p / (1 – p)) from [-∞; +∞] centered around 0 (= 50%)

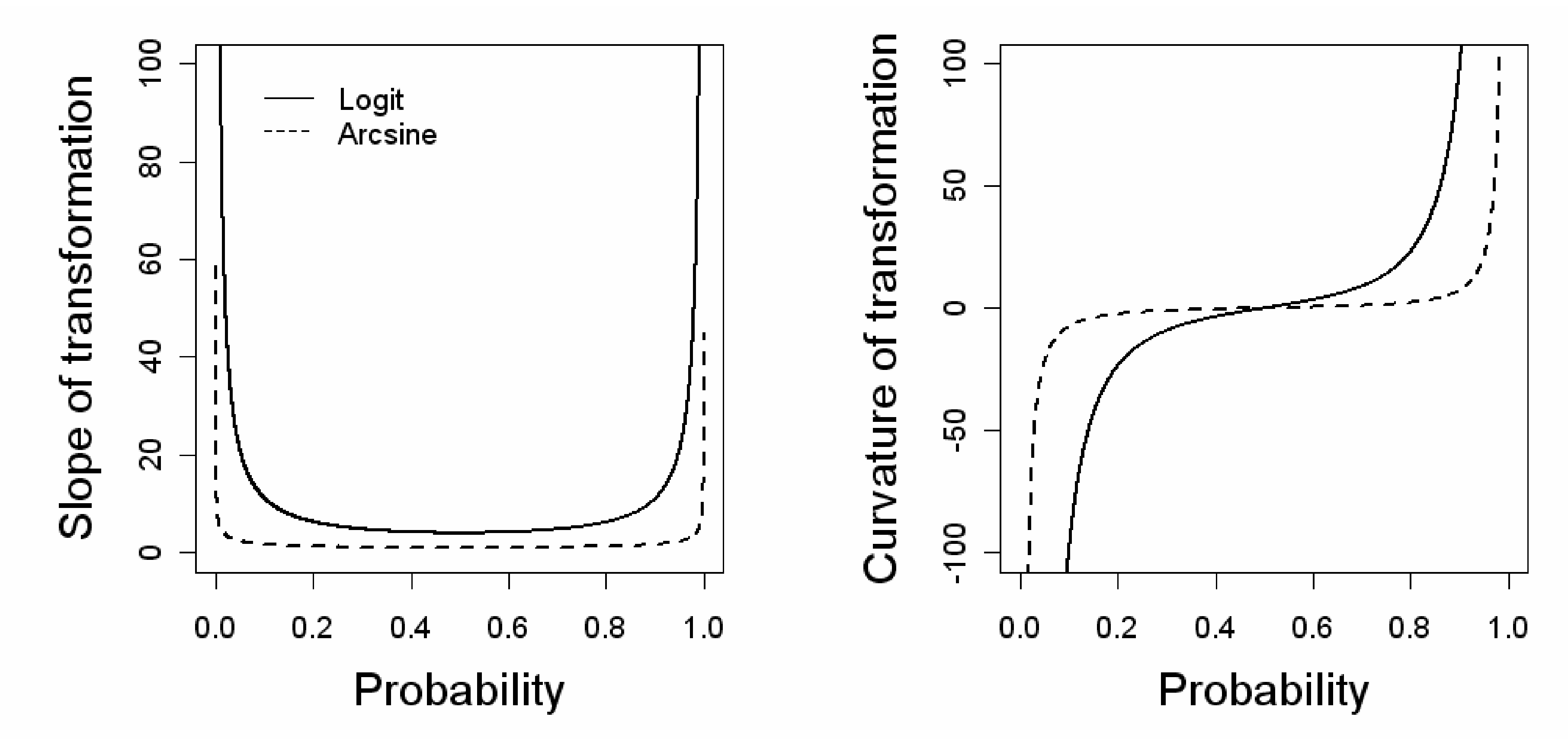
  Logistic regression: linear regression in log-odds space

- Common alternative, ANOVA-based solution: **arcsine transformation, BUT ...**

- Why arcsine at all?
- Centered around 50% with increasing slope towards 0 and 100%
- Defined for 0 and 100% (unlike logit)



- For all probabilities (proportions) the logit has a higher slope and a higher absolute curvature.



An example:

Child relative clause comprehension in Hebrew

(Thanks to Inbal Arnon)

$E[Y] = X\beta \leftrightarrow Y = X\beta + e$

$Y \sim N(X\beta, \sigma^2 I) \leftrightarrow Y = X\beta + e,\ e \sim N(0, \sigma^2 I)$

- Taken from **Inbal Arnon**'s study on child processing of Hebrew relative clauses:

Arnon, I. (2006). *Re-thinking child difficulty: The effect of NP type on child processing of relative clauses in Hebrew*. Poster presented at The 9th Annual CUNY Conference on Human Sentence Processing, CUNY, March 2006

Arnon, I. (2006). *Child difficulty reflects processing cost: the effect of NP type on child processing of relative clauses in Hebrew*. Talk presented at the 12th Annual Conference on Architectures and Mechanisms for Language Processing, Nijmegen, Sept 2006.

- Design of comprehension study: 2 x 2
  - Extraction (Object vs. Subject)
  - NP type (lexical NP vs. pronoun)
  - **Dep. variable:** Answer to comprehension question

(1) tasimi madbeka al ha-safta she menasheket et ha-yalda.
Put sticker on the-granny that kisses the-girlACC
'Put a sticker on the granny that kisses the girl'

(2) tasimi madbeka al ha-safta she ha-yalda menasheket.
Put sticker on the-granny that the-girl kisses
'Put a sticker on the granny that the girl kisses'

```
# load data frame
i <-data.frame(read.delim("C:\\Documents and
  Settings\\florian\\Desktop\\R tutorial\\inbal.tab"))


# the data.frame contains data from production and
# comprehension studies. We select comprehension data
  only
# also let's select only cases that have values for all
# variables for interest
i.compr <- subset(i, modality == 1 & Correct != "#NULL!"
  & !is.na(Extraction) & !is.na(NPType))
```

```
# defining some variable values
# we recode (and rename) the two independent variables
  to:
# RCtype :: either "subject RC" or "object RC"
# NPtype :: either "lexical" or "pronoun"
i.compr$RCtype<- as.factor(ifelse(i.compr$Extraction ==
  1, "subject RC", "object RC"))
i.compr$NPtype <- as.factor(ifelse(i.compr$NPType == 1,
  "lexical", "pronoun"))

# in order to average over the categorical dependent
  variable
# we convert it into a number (0 or 1)
i.anova$Correct <-
  as.numeric(as.character(i.anova$Correct))
```

Correct answers	Lexical NP	Pronoun NP
Object RC	68.9% +20.7% +15.4%	84.3% +10.4%
Subject RC	89.6% +6.1%	95.7%

```
# aggregate over subjects
i.F1 <- aggregate(i.anova,
  by= list(subj= i.anova$child, RCtype= i.anova$RCtype,
  NPtype= i.anova$NPtype),
  FUN= mean)
F1 <-  aov(Correct ~ RCtype * NPtype + Error(subj/(RCtype *
  NPtype)), i.F1)
summary(F1)
```

- **RC type :** $F1(1,23) = 30.3$, $p < 0.0001$
- **NP type:** $F1(1,23) = 20.6$, $p < 0.0002$
- **RC type x NP type:** $F1(1, 23) = 8.1$, $p < 0.01$

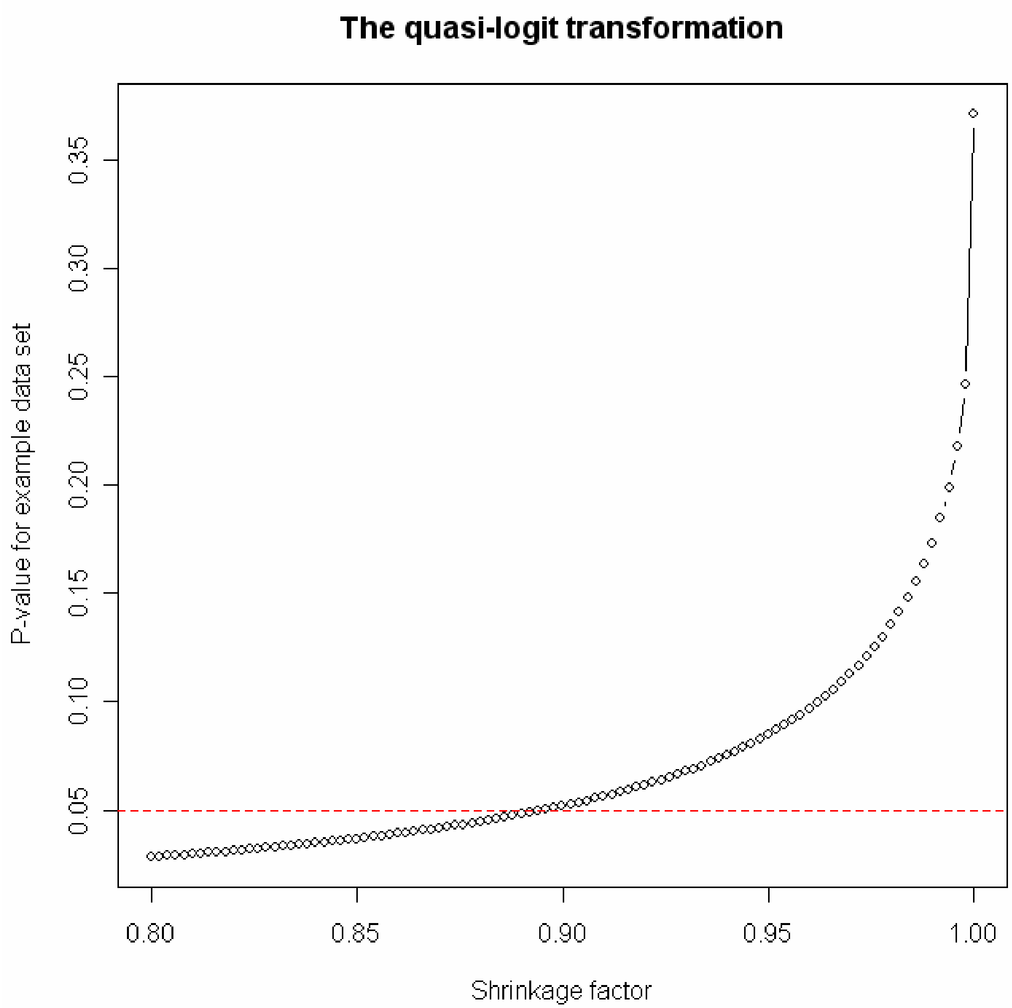
```
# apply arcsine transformation on aggregated data
# note that arcsine is defined from [-1 … 1], not [0 … 1]
i.F1$TCorrect <- asin(sqrt(i.F1$Correct))

F1 <-  aov(TCorrect ~ RCtype * NPtype + Error(subj/(RCtype *
  NPtype)), i.F1)
summary(F1)
```

- **RC type :** $F1(1,23) = 34.3$, $p < 0.0001$
- **NP type:** $F1(1,23) = 19.3$, $p < 0.0003$
- **RC type x NP type:** $F1(1, 23) = 4.1$, $p < 0.054$

```
# apply logit transformation on aggregated data
# use * 0.9999 to avoid problems with 100% cases
i.F1$TCorrect <- qlogis((i.F1$Correct – 0.5) * 0.9999) + .5

F1 <-  aov(TCorrect ~ RCtype * NPtype + Error(subj/(RCtype *
  RCtype)), i.F1)
summary(F1)
```

- **RC type :** $F1(1,23) = 29.0$, $p < 0.0001$
- **NP type:** $F1(1,23) = 13.5$, $p < 0.002$
- **RC type x NP type:** $F1(1, 23) = 0.8$, $p > 0.37$

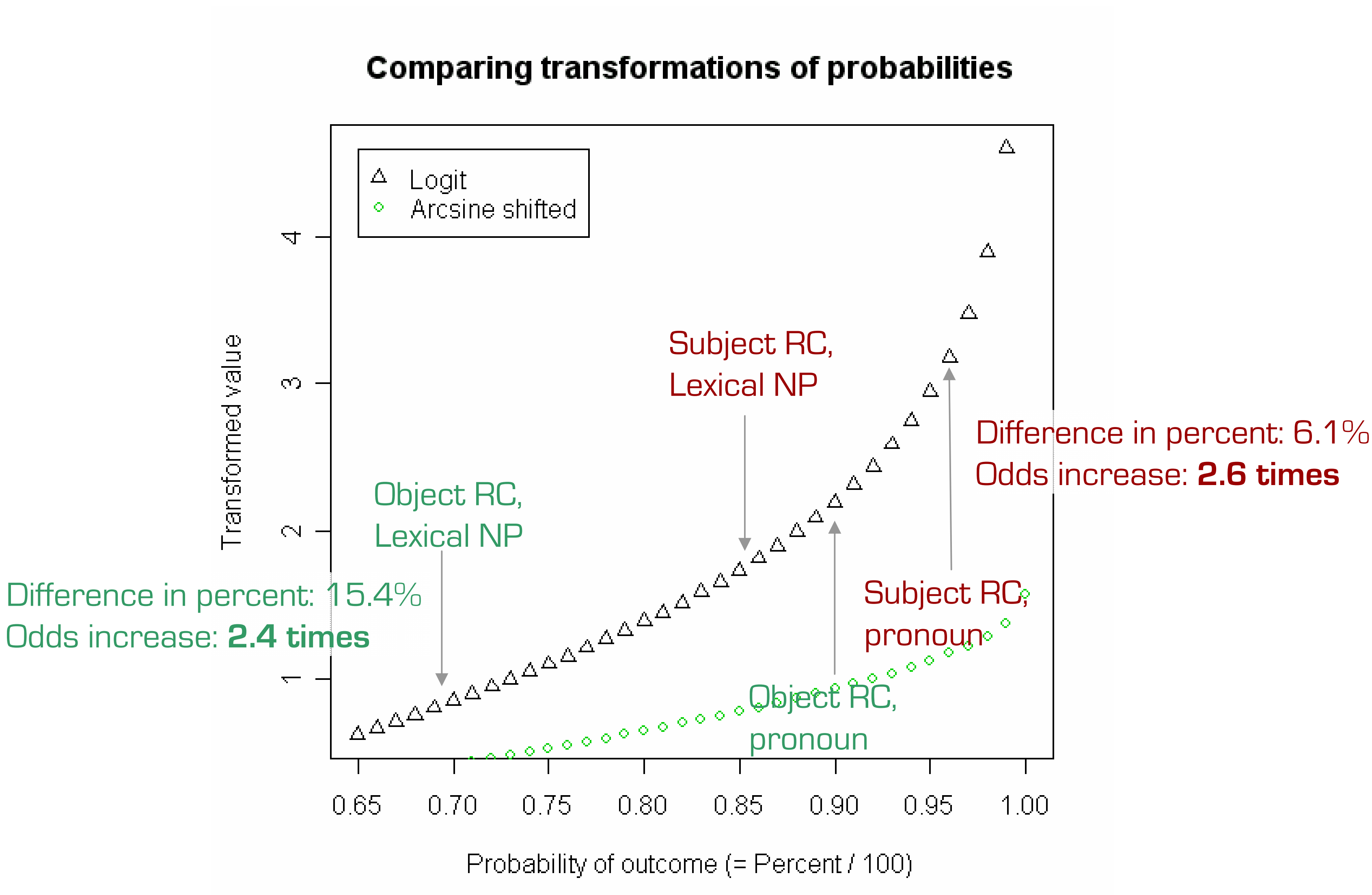
- The significance of the test using the “quasi”-logit transformation depends a lot on how much we “shrink” proportions before applying the logit:



R syntax for previous plot


```
step<- 100
s<- .8
e<- .999999

# rerun anova analysis with different “shrinkage”
for (t in seq(s,e,(e-s) / step)) {
      i.F1$TCorrect <- qlogis(((i.F1$Correct -.5) * t) + .5)
      F1 <-  aov(TCorrect ~ Extraction * NPType +
             Error(subj/(Extraction * NPType)), i.F1)
      # extracting p-value for interaction
      if(t == s) {
         p<- c(as.numeric(
             unlist(
               summary(F1)[4][[1]][[1]]["Pr(>F)"])[1]))
      }
      else {
         p<- append(p, c(as.numeric(
             unlist(
               summary(F1)[4][[1]][[1]]["Pr(>F)"])[1]))) }
}
plot(seq(s,e,(e-s)/step),p,
      xlab="Shrinkage factor",
      ylab="P-value for example data set",
      type="b", main="The quasi-logit transformation")
abline(0.05,0, col=2, lty=2)
```

What's going on???



- For the current sample, ANOVAs over our quasi-logit transformation seems to do the job

- But logistic regressions (or more generally, Generalized Linear Models) offer an alternative
  - more power (Baayen, 2004)
  - easier to add post-hoc controls, covariates
  - easier to extend to unbalanced data

  - nice implementations are available for R, SPSS, …

Logistic regression

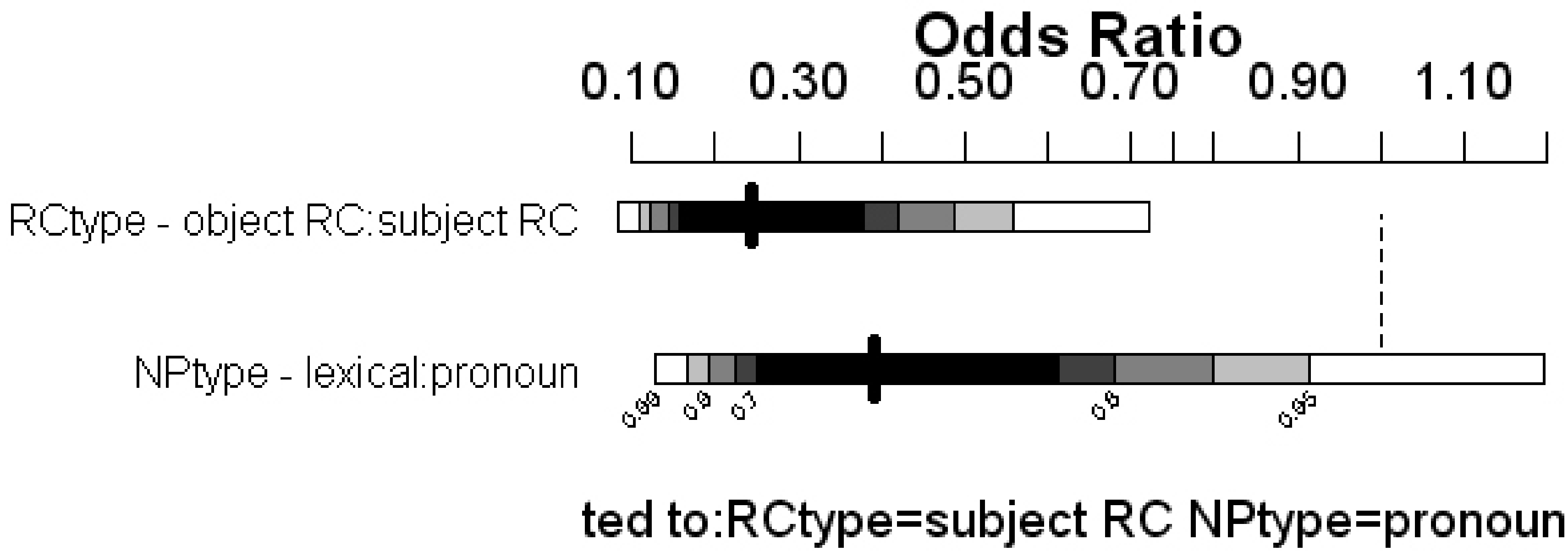
```
# no aggregating
library(Design)
i.d <- datadist(i.compr[,c('Correct','RCtype','NPtype')])
options(datadist='i.d')

i.l <- lrm(Correct ~ RCtype * NPtype, data = i.compr)
```

Children are 3.9 times better at answering questions about **subject RCs**

Children are 2.4 times better at answering questions about **RCs with pronoun** subjects

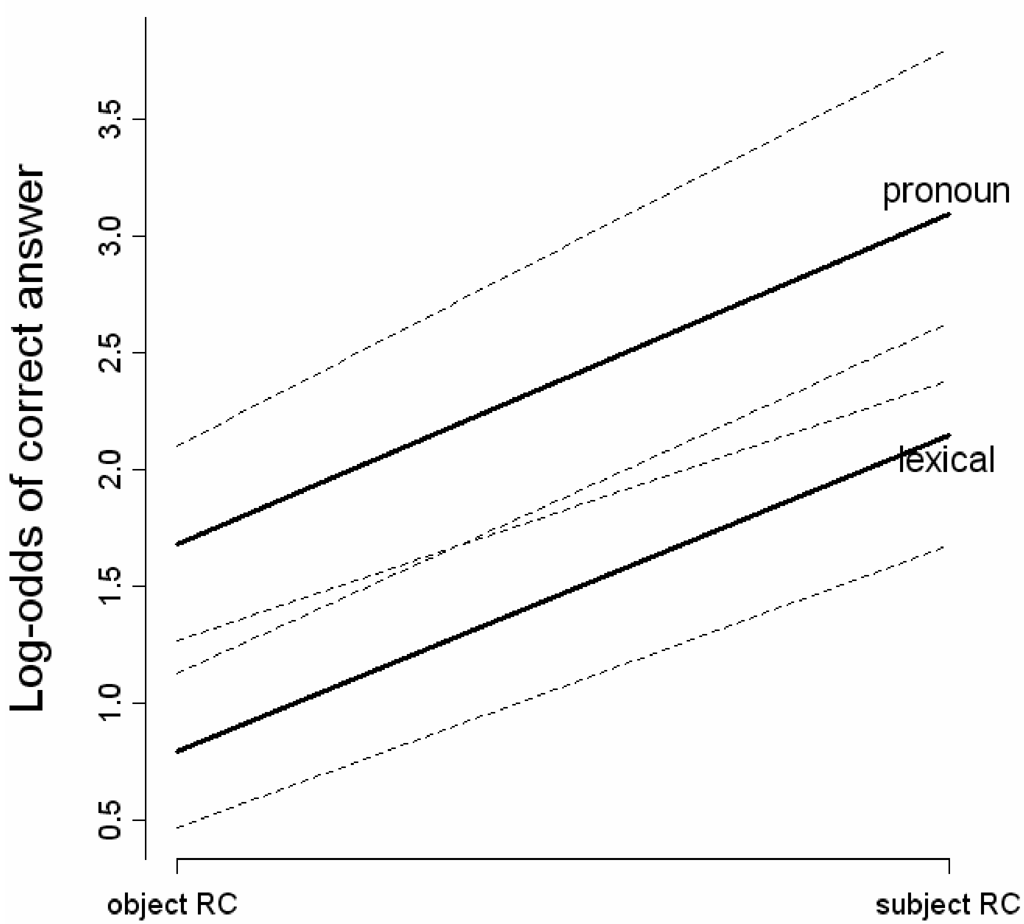
Factor	**Coefficient** (in log-odds)	**SE**	**Wald**	**P**
Intercept	0.80	0.167	4.72	<0.0001
RC type=subject RC	1.35	0.295	4.58	<0.0001
NP type=pronoun	0.89	0.272	3.26	<0.001
RC type * NP type	0.05	0.511	0.10	>0.9

```
par(mar=c(1,1,3,1), cex.lab=1.5, cex=1.2)
plot(summary(i.l), nbar=10)
```



```
par(mar=c(1,1,3,1), cex.lab=1.5, cex=1.2)
plot(summary(i.l), nbar=10)
```

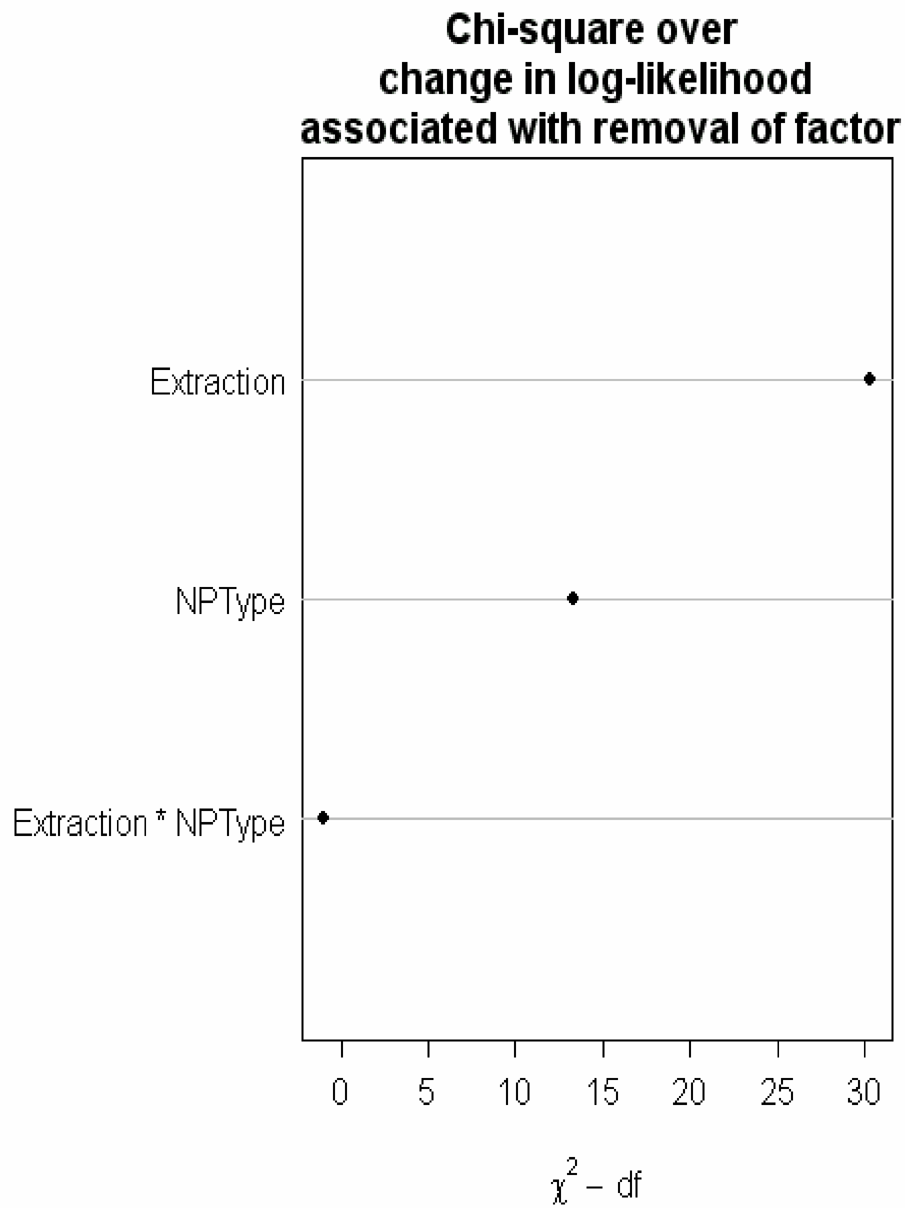
```
plot(i.l, RCtype=NA, NPtype=NA,
  ylab="Log-odds of correct answer")
```



```
# nRCtype and nNPtype are numerically coded
# creating centered interaction variable
i.compr$nInt <- (i.compr$nRCtype - mean(i.compr$nRCtype)) *
               (i.compr$nNPtype - mean(i.compr$nNPtype))

# rerun logistic regression with new terms
i.int.l <- lrm(Correct ~ nRCtype + nNPtype + nInt, i.compr)
i.int.l

# Variance Inflation Factors of new model are lower → nice
vif(i.int.l)
```

Factor	Coefficient (in log-odds)	SE	Wald	P
Intercept	**1.70**	0.200	8.51	<0.0001
RC type=subject RC	1.36	0.257	5.37	<0.0001
NP type=pronoun	0.92	0.263	3.49	<0.001
Centered interaction	0.05	0.513	0.10	>0.9

- Full model: Nagelkerke r^2=0.12
- Likelihood ratio test more robust against collinearity



- Arnon realized post-hoc that a good deal of her stimuli head nouns and RC NPs that were matched in animacy.

- Such animacy-matches can lead to interference

(1) tasimi madbeka al ha-safta she menasheket et ha-yalda.
Put sticker on the-granny that kisses the-girlACC
'Put a sticker on the granny that kisses the girl'

(2) tasimi madbeka al ha-safta she ha-yalda menasheket.
Put sticker on the-granny that the-girl kisses
'Put a sticker on the granny that the girl kisses'

	Match	No Match
S.Lexical	91	91
S.Pronoun	92	92
O.Lexical	95	69
O.pronoun	94	72

- In logistic regression, we can just add the variable
- Matched animacy is almost balanced across conditions, but for more unbalanced data, ANOVA would become inadequate!
- Also, while we're at it – does the children's age matter?

```
i.lc <- lrm(Correct ~ Extraction * NPType + Animacy +
  Age, data = i.compr)

fastbw(i.lc)          # fast backward variable removal
```

Coefficients of Extraction and NP type almost unchanged → good, suggests independence from newly added factor

Lack of animacy-based interference does indeed increase performance, but the other effects persist

Possibly small increase in performance for older children (no interaction found)

Factor	Coefficient (in log-odds)	SE	Wald	P
Intercept	-1.06	0.956	-1.10	>0.25
RC type=subject	1.43	0.300	4.78	<0.0001
NP type=pronoun	0.91	0.275	3.33	<0.001
Animacy=no match	0.64	0.226	2.84	<0.005
Age	0.03	0.018	1.60	<0.11

- Model $r^2 = 0.151$ → quite an improvement

- As we are leaving balanced designs in post-hoc tests like the ones just presented, collinearity becomes an issue
- Collinearity (a and b explain the same part of the variation in the dependent variable) can lead to spurious results

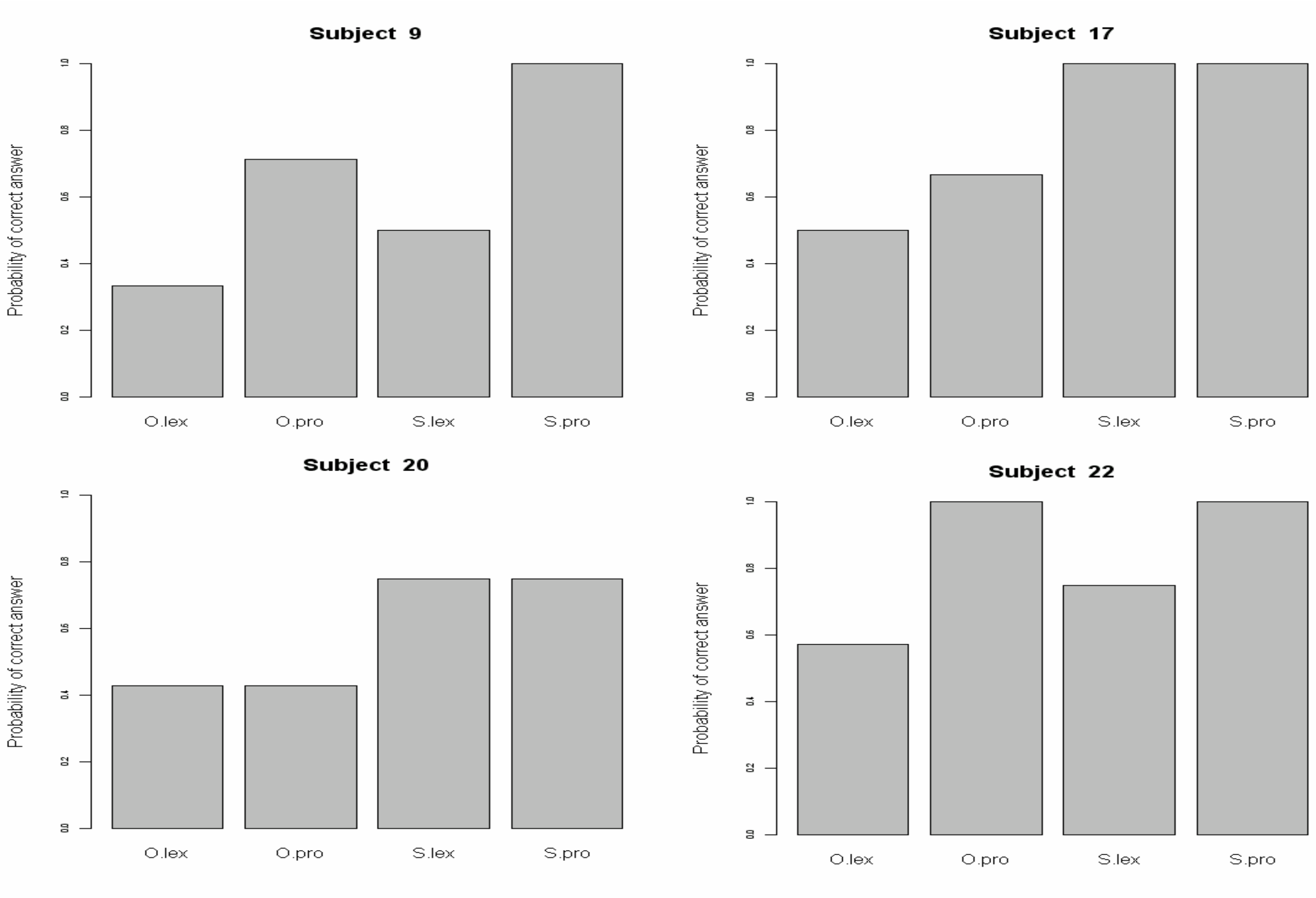
- In this case all VIFs are below 2 (VIFs of 10 means that no absence of total collinearity can be claimed)

```
# Variation Inflation Factor (Design library)

vif(i.lc)
```

- The assumption of independence is violated if clusters in your data are correlated
  - Several trials by the same subject
  - Several trials of the same item

- **Subject/item usually treated as random effects**
  - Levels are not of interest to design
  - Levels represent random sample of population
  - Levels grow with growing sample size
  - Account for variation in the model (can interact with fixed effects!), e.g. subjects may differ in performance

Do subjects differ?



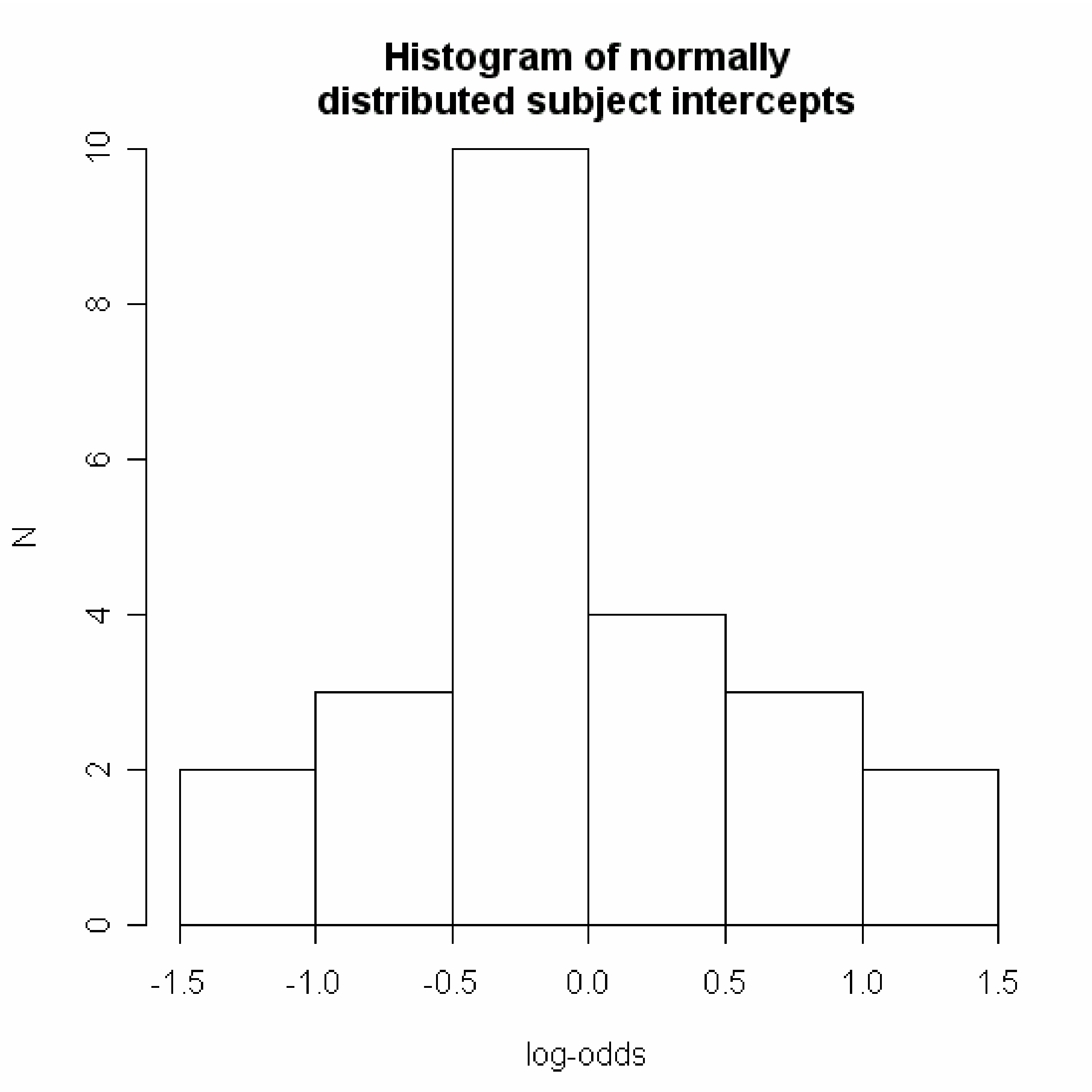
- In ANOVAs, F1 and F2 analyses are used to account for random subject and item effects

- There are several ways that subject and item effects can be accounted for in Generalized Linear Models (GLMs)
  - Run models for each subject/item and examine distributions over coefficients (Lorch & Myers, 1990)
  - Bootstrap with random cluster replacement
  - Incorporate random effects into model → Generalized Linear Mixed Models (GLMMs)

- Random effects are sampled from normal distribution (with mean of zero)
  - Only free parameter of a random effect is the standard deviation of the normal distribution

```
library(lme4)

i.ml <- lmer(Correct ~ RCtype * NPtype + (1 + RCtype *
  NPtype | child), data = i.compr, family="binomial")
summary(i.ml)
```

Factor	**Coefficient** (in log-odds)	**SE**	**Wald**	**P**
Intercept	0.84	0.203	4.12	<0.0001
RC type=subject	1.82	0.368	4.95	<0.0001
NP type=pronoun	1.07	0.289	3.70	<0.0003
RC type * NP type	0.59	0.581	1.02	>0.3


Histogram of normally
distributed subject intercepts
N
0
2
4
6
8
10
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
log-odds

- Using an ANOVA over percentages of categorical outcomes can lead to **spurious significance**
- Using the 'standard' arcsine transformation did **not** prevent this problem

- Our ANOVA over 'adapted' logit-transformed percentages did ameliorate the problem

- Moving to regression analyses has the advantage that imbalance is less of a problem, and extra covariates can easily be added

- Logistic regression provides an alternative way to analyze the data:
  - Gets the right results
  - Coefficients give direction and size of effect
  - Differences in data log-likelihood associated with removal of a factor give a measure of the importance of the factor

- Logit Mixed models provide a way to combine the advantages of logistic regression with necessity of random effects for subject/item
  - subject/item analyses can be done in one model

E.g.

```
l <- lmer(FinalScore ~
  PrimeStrength * log(TargetOdds) +
  Lag +
  PrimingStyle +
  (1 | SuperSubject) +
  (1 | SuperItem),
  data = k,
  family = "binomial")

summary(i.ml)
```

- Intro to R by Matthew Keller http://matthewckeller.com/html/r_course.html [thanks to Bob Slevc for pointing this out to me]
- Intro to Statistic using R by Shravan Vasishth http://www.ling.uni-potsdam.de/~vasishth/Papers/vasishthESSLLI05.pdf; see also the other slides on his website
- Joan Bresnan taught a Laboratory Syntax class in Fall, 2006 on using R for corpus data; ask her for her notes one bootstrapping and mixed models
- Peter Dalgaard. 2002. *Introductory Statistics to R.* Springer, http://staff.pubhealth.ku.dk/~pd/ISwR.html

Mixed model resources

- Harald Baayen. 2004. *Statistics in Psycholinguistics: A critique of some current gold standards.* In Mental Lexicon Working Papers 1, Edmonton, 1-45; http://www.mpi.nl/world/persons/private/baayen/publications/Statistics.pdf
- J.C. Pinheiro & Douglas M. Bates. 2000. *Mixed effect models in S and S-plus*. Springer, http://stat.bell-labs.com/NLME/MEMSS/index.html [S and S+ are commercial variants of R]
- Douglas M. Bates & Saikat DebRoy. 2004. *Linear mixed models and penalized least squares*. Journal of Multivariate Analysis 91, 1–17
- Hugo Quene & Huub van den Bergh. 2004. *On multi-level modeling of data from repeated measures designs: a tutorial.* Speech Communication 43, 103–121