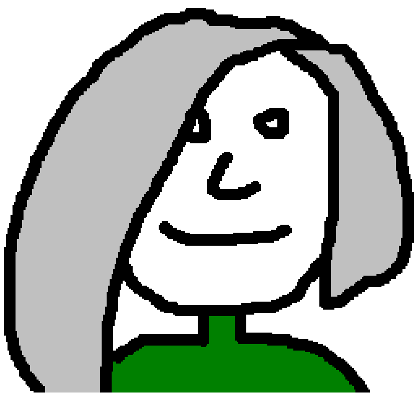
Testing Linguistic Theories Using Logistic Regression

Peter Graff

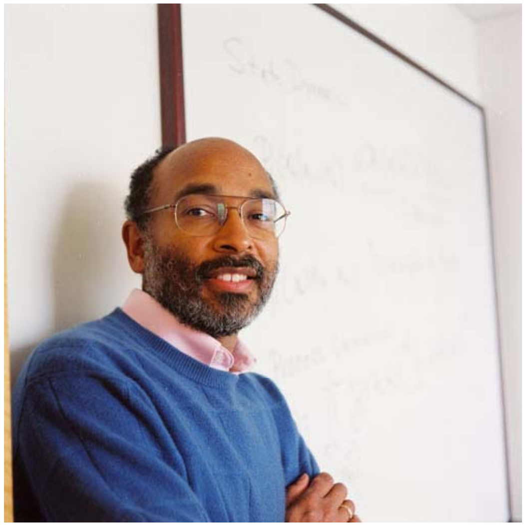
MIT, Linguistics and Philosophy

Acknowledgements



Ellen Gurman Bard

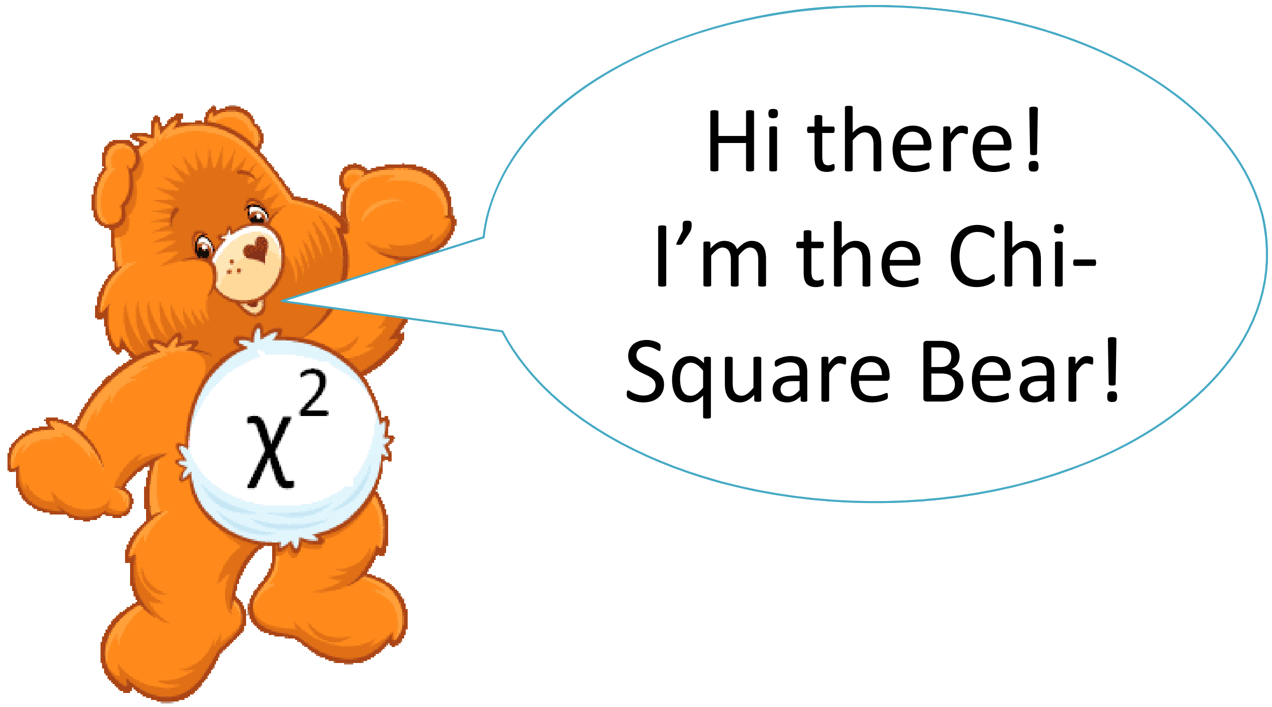
T. Florian Jaeger



Emory Brown

All errors are my own!

Introduction



- My furry friend and helper throughout this lecture.

Plan for today

- Part 1
  - What is Logistic Regression
  - How to fit a Logistic Regression
  - How to compare Logistic Regression models
- Part 2
  - Plural Comparison
  - How to use Logistic Regression to decide between theories of Plural Comparison

What is Logistic Regression?
Limitations of Linear Models

- Assumptions of Linear Models
  - ***Linearity in Coefficients***
  - ***Normally distributed outcome (or error)***
- But many/most of the outcomes of interest to linguists are categorical!
  - ***Non-continuous outcomes are usually not normally distributed***

What is Logistic Regression?
Categorical outcomes

- Grammaticality
  - #kn (attested/unattested)
- Syntactic Variation:
  - Dative alternation (NP NP/NP PP)
- Phonological Variation
  - t-Deletion (t/∅)
- Experimental Data:
  - Forced Choice, Eye-tracking, ...

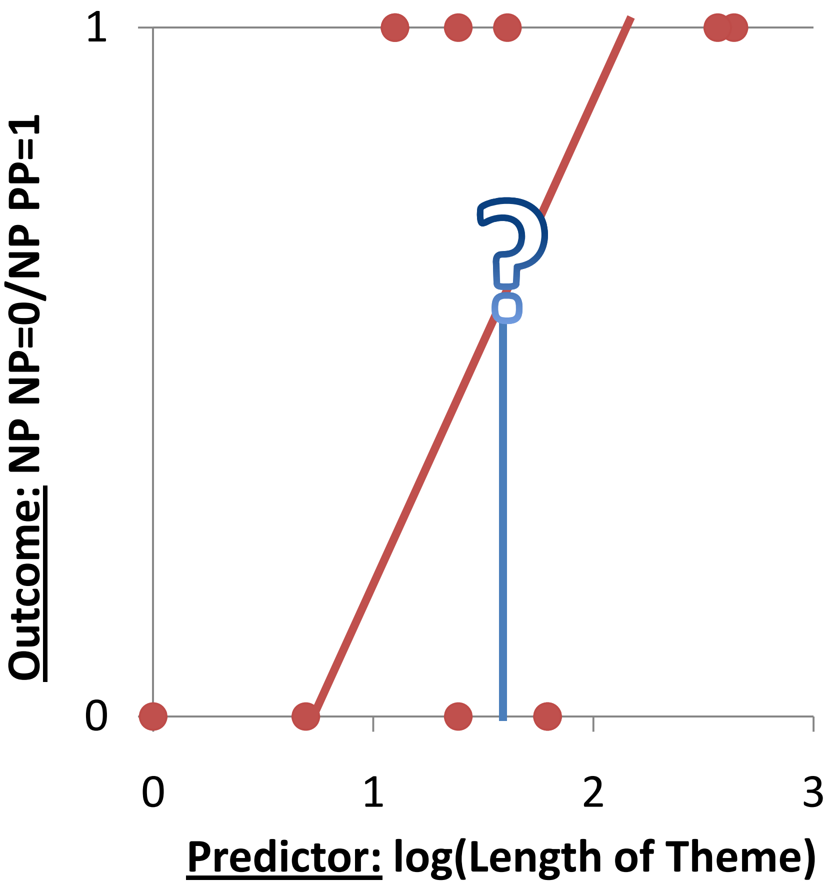
What is Logistic Regression?
Categorical outcomes

- Grammaticality
  - #kn (attested/unattested)
- Syntactic Variation:
  - Dative alternation (NP NP/NP PP)
- Phonological Variation
  - t-Deletion (t/∅)
- Experimental Data:
  - Forced Choice, Eye-tracking, ...

Can you think of some more?

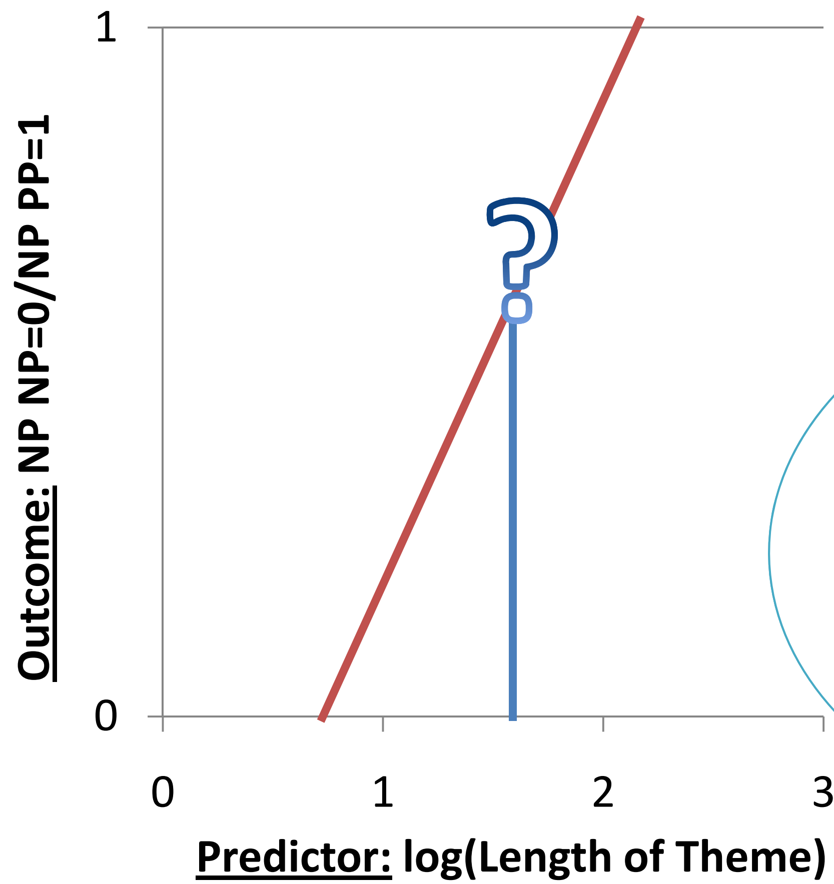
What is Logistic Regression?

Can a linear model do the job?



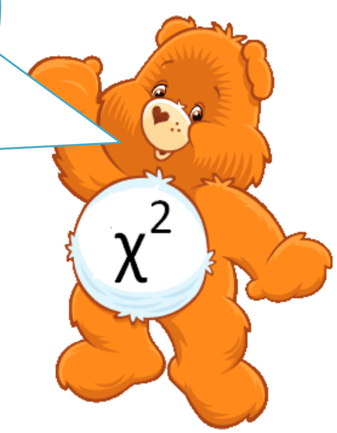
- Predicting the Realization of Dative from length of theme.

What is Logistic Regression?
Can a linear model do the job?



- Predicting the Realization of Dative from length of theme.

Why can't we use a linear model to predict a dichotomous variable?



What is Logistic Regression?
Can a linear model do the job?

- The linear model makes impossible predictions
  - ***Values of Y>1***
  - ***Values of Y<0***
  - ***Values of Y>0 and Y<1***
- The linear model is meaningless if its assumptions are violated

What is Logistic Regression?
Generalized Linear Models

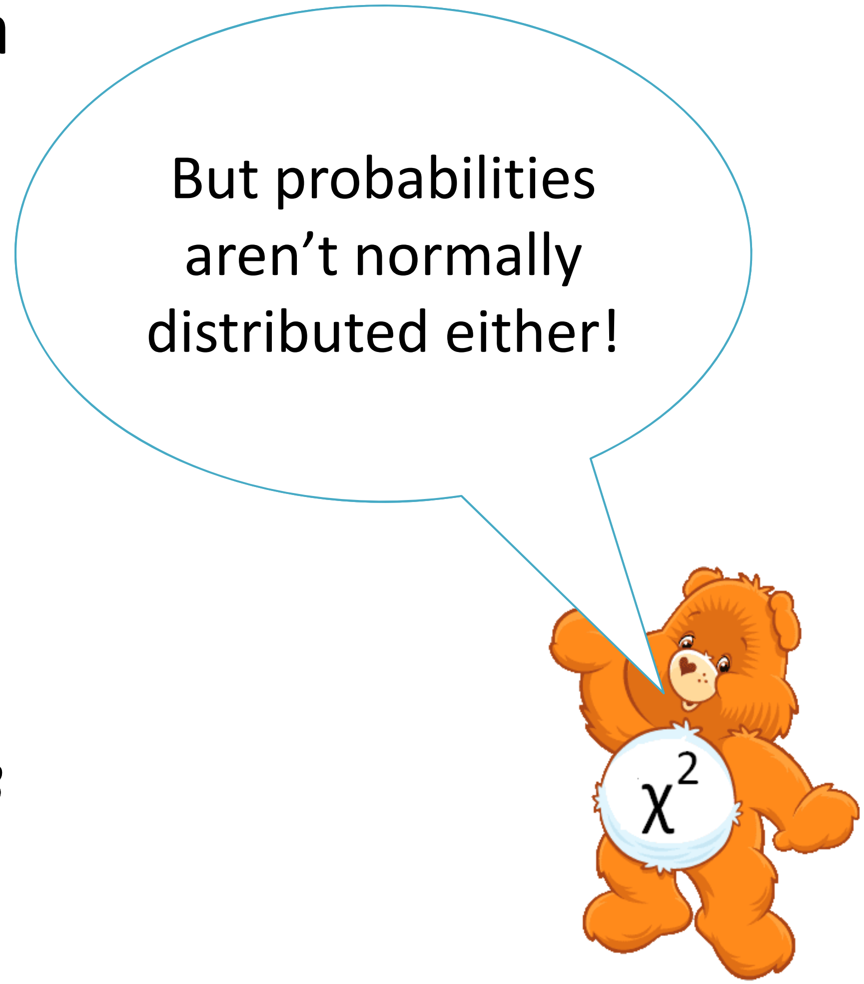
- Transform non-normally distributed variables into a linear space.

- Fit a line in to predict the transformed variable.

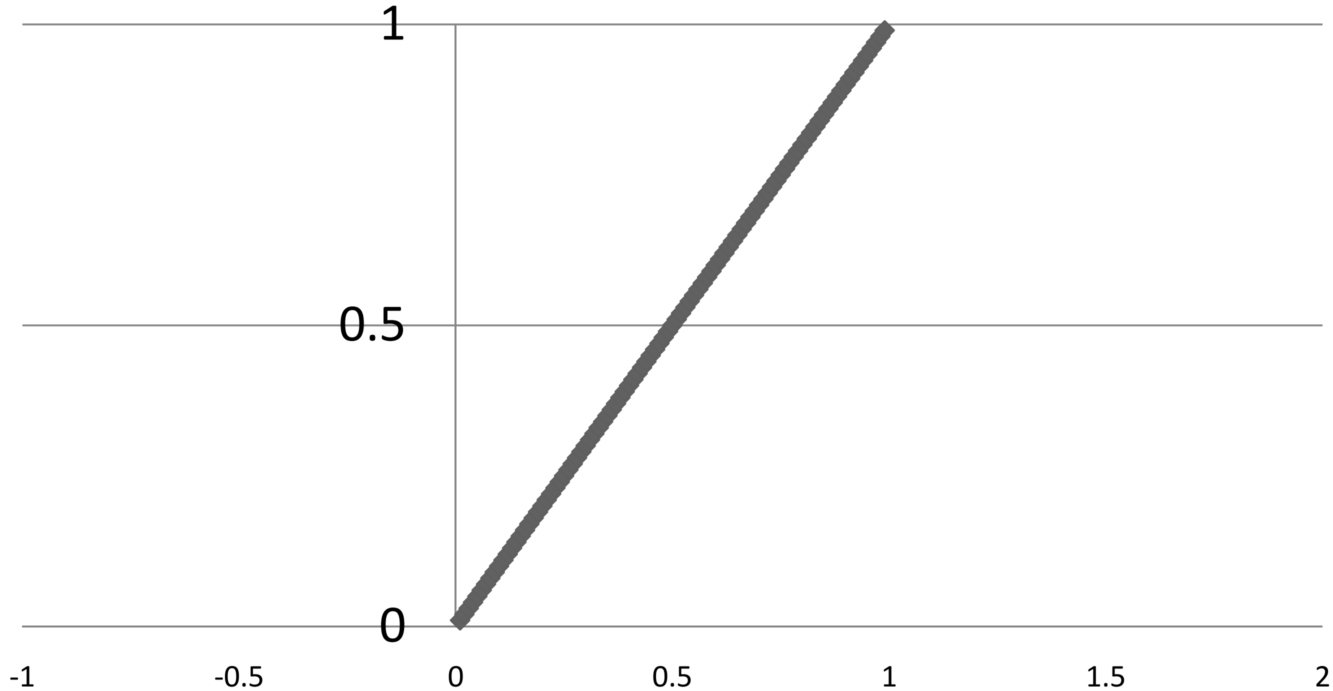
- What do we do for binary outcomes?
  - ***The probability of outcome A over outcome B***

What is Logistic Regression?
Generalized Linear Models

- Transform non-normally distributed variables into a linear space.
- Fit a line in to predict the transformed variable.
- What do we do for binary outcomes?
  - ***The probability of outcome A over outcome B***



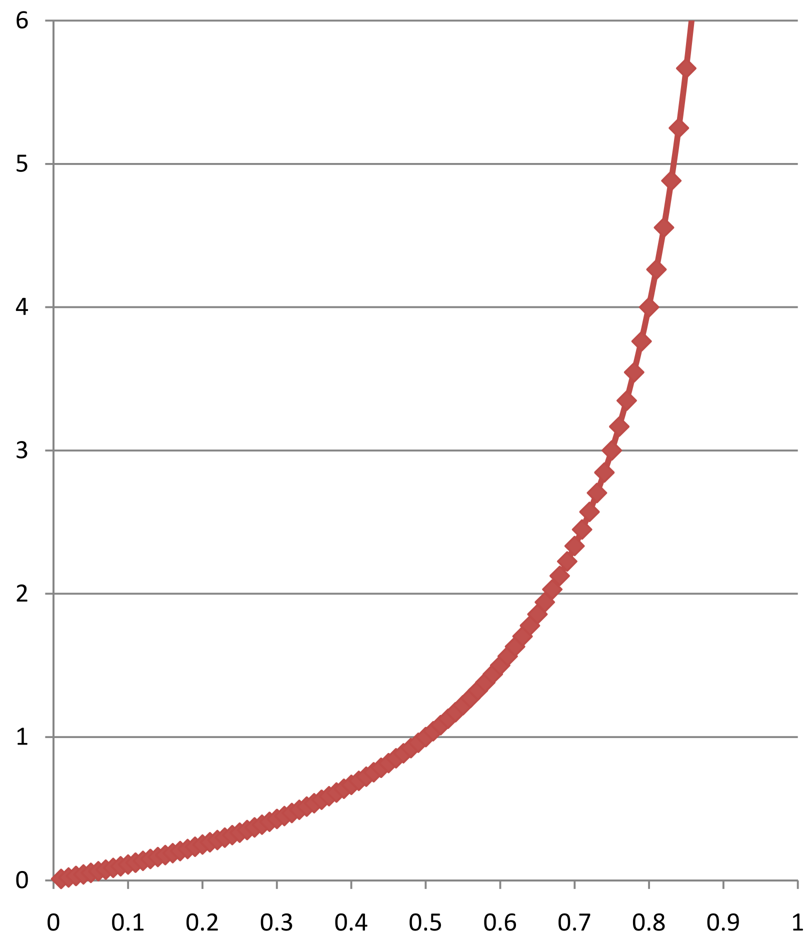
What is Logistic Regression?
Transforming Probabilities



- Probabilities have an upper and a lower bound
- Changes in probability around .5 mean something different from changes around 0 and 1.

What is Logistic Regression?
Transforming Probabilities

- Probabilities range between 1 and 0
- Odds range from 0 to ∞

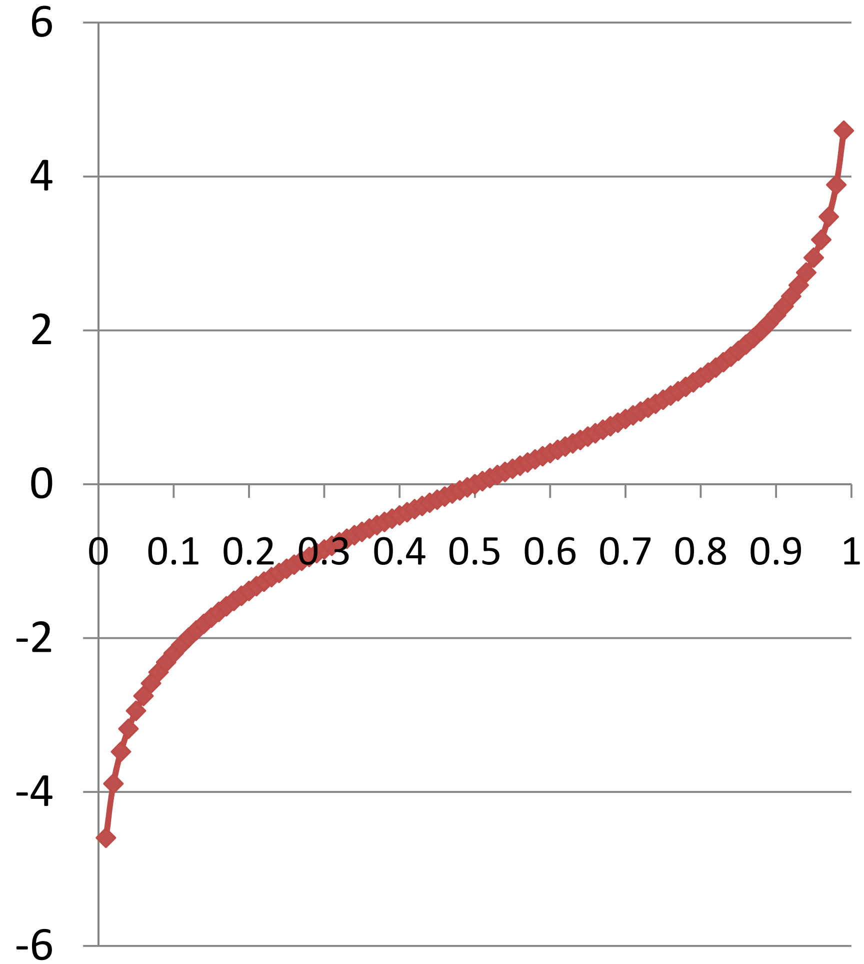


$$\boldsymbol{o = (p/1-p)}$$

- $p<.5$, $0 < o < 1$
- $p=.5$, $o=1$
- $p>.5$, $o > 1$

What is Logistic Regression?
Transforming Probabilities

- Logged Odds range from -∞ to ∞
- Natural logarithm of the odds ratio (a.k.a. **logit**)
- 0 at p=.5
- Probabilities with the same distance from .5 have the same logits but different signs.



How to fit a Logistic Regression
Input Data

- lrm(formula)

```
Trial/Case        0/1    IV1    IV2    ...
Trial/Case        0/1    IV1    IV2    ...
Trial/Case        0/1    IV1    IV2    ...
```

- glm(formula, familiy = “binomial”)

```
Cell      #of0   #of1   IV1    IV2    ...
Cell      #of0   #of1   IV1    IV2    ...
Cell      #of0   #of1   IV1    IV2    ...
```

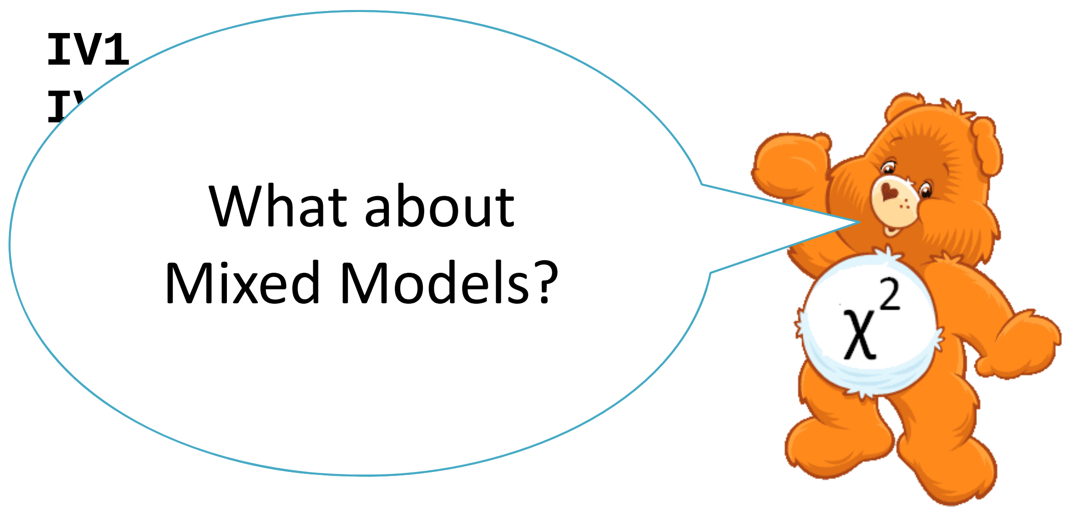
How to fit a Logistic Regression
Input Data

- lrm(formula)

```
Trial/Case      0/1    IV1    IV2    ...
Trial/Case      0/1    IV1    IV2    ...
Trial/Case      0/1    IV1    IV2    ...
```

- glm(formula, familiy = “binomial”)

```
Cell      #of0   #of1   IV1
Cell      #of0   #of1   I
Cell      #of0   #of1
```



How to fit a Logistic Regression
Input Data

- lmer(formula, family = “binomial”)

```
Trial/Case        0/1    IV1    IV2    ...
Trial/Case        0/1    IV1    IV2    ...
Trial/Case        0/1    IV1    IV2    ...
```

How to fit a Logistic Regression
The Formula

- Formula in R:

DV ~ IV+...+IV

- '+' crosses IV's
- ':' denoted the interaction of 2 IV's
- '*' cross and interaction
- '|' grouping operator
- '(IV+...+IV)^n' all interactions up to level n
- ***for glm() DV must be entered as cbind(#of0,#of1)***

How to fit a Logistic Regression
The Output

```
> lrm(RealizationOfRec~AnimacyOfRec+AnimacyOfTheme+LengthOfTheme,data=verbs)
Logistic Regression Model
lrm(formula = RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme + LengthOfTheme, data = verbs)
Frequencies of Responses
NP  PP
555 348
Obs Max Deriv Model L.R. d.f.  P    C        Dxy      Gamma      Tau-a    R2      Brier
903 2e-07     144.52     3     0    0.726    0.452    0.486      0.214    0.201   0.203

                          Coef     S.E.   Wald Z P
Intercept                  0.01976 1.1435   0.02 0.9862
AnimacyOfRec=inanimate     0.49402 0.2544   1.94 0.0522
AnimacyOfTheme=inanimate   0.94931 1.1358   0.84 0.4032
LengthOfTheme             -1.04129 0.1005 -10.36 0.0000
```

Here is the lrm() output, summary(glm()) contains the same information.

How to fit a Logistic Regression
The Output

```
> lrm(RealizationOfRec~AnimacyOfRec+AnimacyOfTheme+LengthOfTheme,data=verbs)
Logistic Regression Model
lrm(formula = RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme + LengthOfTheme, data = verbs)
Frequencies of Responses
NP  PP
555 348
Obs Max Deriv Model L.R. d.f.  P    C        Dxy      Gamma      Tau-a    R2      Brier
903 2e-07     144.52     3     0    0.726    0.452    0.486      0.214    0.201   0.203

                          Coef     S.E.    Wald Z P
Intercept                 0.01976 1.1435   0.02 0.9862
AnimacyOfRec=inanimate    0.49402 0.2544   1.94 0.0522
AnimacyOfTheme=inanimate  0.94931 1.1358   0.84 0.4032
LengthOfTheme            -1.04129 0.1005 -10.36 0.0000
```

Base probability of outcome=1 in logged odds

How to fit a Logistic Regression
The Output

```
> lrm(RealizationOfRec~AnimacyOfRec+AnimacyOfTheme+LengthOfTheme,data=verbs)
Logistic Regression Model
lrm(formula = RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme + LengthOfTheme, data = verbs)
Frequencies of Responses
NP  PP
555 348
Obs Max Deriv Model L.R. d.f.  P    C        Dxy      Gamma       Tau-a    R2       Brier
903 2e-07     144.52     3     0    0.726    0.452    0.486       0.214    0.201    0.203

                            Coef    S.E.   Wald Z P
Intercept                   0.01976 1.1435   0.02 0.9862
AnimacyOfRec=inanimate      0.49402 0.2544   1.94 0.0522
AnimacyOfTheme=inanimate    0.94931 1.1358   0.84 0.4032
LengthOfTheme              -1.04129 0.1005 -10.36 0.0000
```

How P(outcome=1) changes depending on the setting of the independent variables in logged odds

How to fit a Logistic Regression
The Output

```
> lrm(RealizationOfRec~AnimacyOfRec+AnimacyOfTheme+LengthOfTheme,data=verbs)
Logistic Regression Model
lrm(formula = RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme + LengthOfTheme, data = verbs)
Frequencies of Responses
NP  PP
555 348
Obs Max Deriv Model L.R. d.f.  P    C       Dxy     Gamma      Tau-a   R2      Brier
903 2e-07     144.52     3     0    0.726   0.452   0.486      0.214   0.201   0.203

                               Coef    S.E.   Wald Z P
Intercept                     0.01976 1.1435   0.02 0.9862
AnimacyOfRec=inanimate        0.49402 0.2544   1.94 0.0522
AnimacyOfTheme=inanimate      0.94931 1.1358   0.84 0.4032
LengthOfTheme                -1.04129 0.1005 -10.36 0.0000
```

Standard Error of the Coefficient

How to fit a Logistic Regression
The Output

```
> lrm(RealizationOfRec~AnimacyOfRec+AnimacyOfTheme+LengthOfTheme,data=verbs)
Logistic Regression Model
lrm(formula = RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme + LengthOfTheme, data = verbs)
Frequencies of Responses
NP  PP
555 348
Obs Max Deriv Model L.R. d.f.  P    C        Dxy      Gamma      Tau-a    R2      Brier
903 2e-07     144.52     3     0    0.726    0.452    0.486      0.214    0.201   0.203
```

```
                              Coef    S.E.   Wald Z P
Intercept                   0.01976 1.1435   0.02 0.9862
AnimacyOfRec=inanimate      0.49402 0.2544   1.94 0.0522
AnimacyOfTheme=inanimate    0.94931 1.1358   0.84 0.4032
LengthOfTheme             -1.04129 0.1005 -10.36 0.0000
```

WaldZ = Coef/SE. This is distributed as z and gives us a P value for P(Coef=0) i.e. IV has no effect

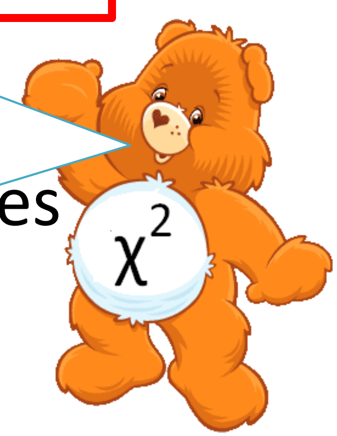
How to fit a Logistic Regression
The Output

```
> lrm(RealizationOfRec~AnimacyOfRec+AnimacyOfTheme+LengthOfTheme,data=verbs)
Logistic Regression Model
lrm(formula = RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme + LengthOfTheme, data = verbs)
Frequencies of Responses
NP  PP
555 348
Obs Max Deriv Model L.R. d.f.  P    C       Dxy     Gamma      Tau-a   R2      Brier
903 2e-07     144.52     3     0    0.726   0.452   0.486      0.214   0.201   0.203
```

	Coef	S.E.	Wald Z	P
Intercept	0.01976	1.1435	0.02	0.9862
AnimacyOfRec=inanimate	0.49402	0.2544	1.94	0.0522
AnimacyOfTheme=inanimate	0.		84	0.4032
LengthOfTheme				0.0000

But what if I want to compare two competing theories?

WaldZ = Coef/SE. This ... ves
us a P value for P(Coef- ... t



Model Comparison
Introduction

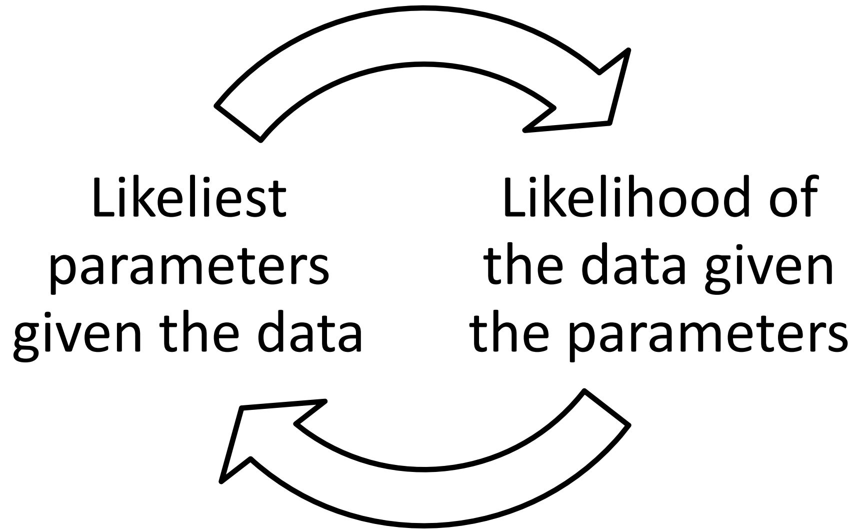
- Often we want to compare two theories in terms of how well they predict our data

- We need to take into account the relative complexity of the theories as more complex theories (theories with more free parameters) will necessarily always do better.

- Logistic Regression allows us to do so in a controlled way.

- Three types of model comparison, we will cover today
  - Chi-Square likelihood test
  - Bayesian Information Criterion
  - Akaike Information Criterion

Model Comparison
Chi-Square Likelihood Test



- The performance of a model is evaluated in terms of its data-likelihood.

  ➔ ***The likelihood of the data given the model***

Model Comparison
Data Likelihood and Deviance

- A models data log-likelihood is defined as...

$$\hat{\ell}(\theta \,|\, x_1, \ldots, x_n) = \frac{1}{n} \ln \mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \ln f(x_i|\theta).$$

- A models deviance is defined as...

$$D(y) = -2[\log\{p(y|\hat{\theta}_0)\} - \log\{p(y|\hat{\theta}_s)\}].$$

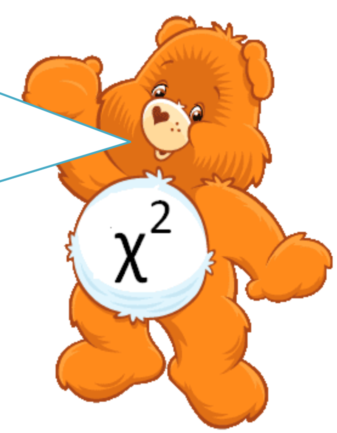
Model Comparison
Chi-Square Likelihood Test

- For nested models, differences in deviance are distributed as chi-square with $d.f. = d.f._{superset} - d.f._{subset}$

Model Comparison
Chi-Square Likelihood Test

- For nested models, differences in deviance are distributed as chi-square with
  d.f. = d.f.$_{superset}$ - d.f. $_{subset}$

That’s me!



Model Comparison
Chi-Square Likelihood Test

- For nested models, differences in deviance are distributed as chi-square with d.f. = $\text{d.f.}_{\text{superset}} - \text{d.f.}_{\text{subset}}$

- If the result of this test is significant we can say that the superset model explains significantly more variance than the subset model considering the additional complexity (degrees of freedom).

Model Comparison

Chi-Square Likelihood Test

```
> lrm(RealizationOfRec~AnimacyOfRec+AnimacyOfTheme+LengthOfTheme,data=verbs)
Logistic Regression Model
lrm(formula = RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme + LengthOfTheme, data = verbs)
Frequencies of Responses
NP  PP
555 348

Obs Max Deriv Model L.R. d.f.  P     C       Dxy     Gamma     Tau-a    R2      Brier
903 2e-07     144.52      3    0    0.726   0.452   0.486     0.214    0.201   0.203

                           Coef     S.E.   Wald Z P
Intercept                  0.01976 1.1435   0.02 0.9862
AnimacyOfRec=inanimate     0.49402 0.2544   1.94 0.0522
AnimacyOfTheme=inanimate   0.94931 1.1358   0.84 0.4032
LengthOfTheme             -1.04129 0.1005 -10.36 0.0000
```

Model L.R. is the likelihood ratio of the model compared to a null-model with no parameters (intercept only).

Because our model has three parameters, degrees of freedom of the model is 3.

Model Comparison
Calculating Model L.R.

- deviance(lrm(...)) returns a vector consisting of the null-models deviance...

-2*ln(likelihood of a model that guesses the majority value for all cases)

- ...and the deviance of your model from the null-model.

- If we put
  deviance(lrm(DV~1))
  the two numbers
  are identical.

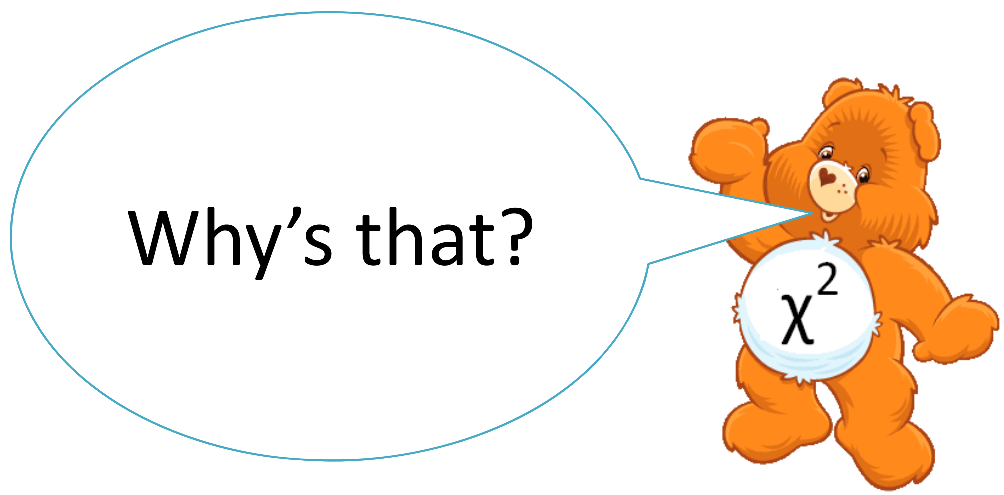
Model Comparison
Calculating Model L.R.

- deviance(lrm(...)) returns a vector consisting of the null-models deviance...

-2*ln(likelihood of a model that guesses the majority value for all cases)

- ...and the deviance of your model from the null-model.

- If we put deviance(lrm(DV~1)) the two numbers are identical.



Model Comparison
Stepwise Regression

- **`anova(lrm(...))`** removes every predictor in the model one by one and lists the difference in deviances of the model with and without that factor.

Factor	Chi-Square	d.f.	P
AnimacyOfRec	3.77	1	0.0522
AnimacyOfTheme	0.70	1	0.4032
LengthOfTheme	107.35	1	<.0001
TOTAL	118.51	3	<.0001

Model Comparison
Nested Model Comparison

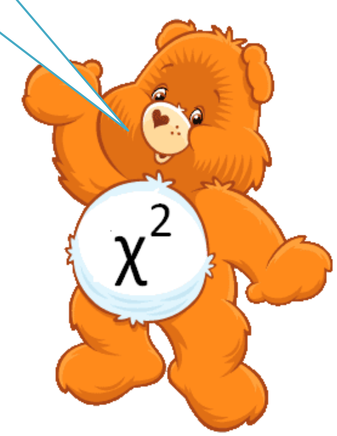
- The following R-code tests whether there is a significant difference in data-likelihood between a subset model A and a superset model B

- **`dchisq(deviance(A)[2]-deviance(B)[2], B$stat[4]-A$stat[4])`**

Model Comparison
Nested Model Comparison

- The following R-code te… …ther there is a significant differenc… …od between a subset … …erset model B

- **`dchisq(deviance(A)[2]-deviance(B)[2], B$stat[4]-A$stat[4])`**



Model Comparison
Nested Model Comparison

```
> anova(lmer,lmer2)
Data: verbs
Models:
lmer2: RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme + (1 | Verb)
lmer: RealizationOfRec ~ AnimacyOfRec + AnimacyOfTheme +
LengthOfTheme + (1 | Verb)
      Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
lmer2  4  852.75  871.97 -422.37
lmer   5  718.06  742.09 -354.03 136.69      1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Comparison
Non-Nested Model Comparison

- Only differences in deviance between nested models are distributed as chi-square.

- When we want to compare non-nested models, we first need to fit a superset model including all parameters and compare it to each subset model in turn.

- If only one of the tests comes out significant we can say that the model that does not significantly differ from the superset model is significantly better than the other model.

Model Comparison

Non-Nested Model Comparison

```
> lrm =
lrm(RealizationOfRec~AnimacyOfRec+AnimacyOfTheme
+LengthOfTheme ,data=verbs)
> lrm.length =
lrm(RealizationOfRec~LengthOfTheme,data=verbs)
>lrm.animac = lrm(RealizationOfRec~AnimacyOfRec+
AnimacyOfTheme,data=verbs)


> dchisq(deviance(lrm.length)[2]-
deviance(lrm)[2],lrm$stat[4]-lrm.length$stat[4])
[1] 0.04972017
> dchisq(deviance(lrm.animac)[2]-
deviance(lrm)[2],lrm$stat[4]-lrm.animac$stat[4])
[1] 3.274109e-30
```

Model Comparison
What if superset models don't converge?

- The Bayesian Information Criterion is defined as...

$$-2 \cdot \ln p(x|k) \approx \mathrm{BIC} = -2 \cdot \ln L + k \ln(n).$$

- The Akaike Information Criterion is defined as...

$$AIC = 2k - 2\ln(L)$$

- Model L.R. is penalized relative to d.f.

Model Comparison
What if superset models don't converge?

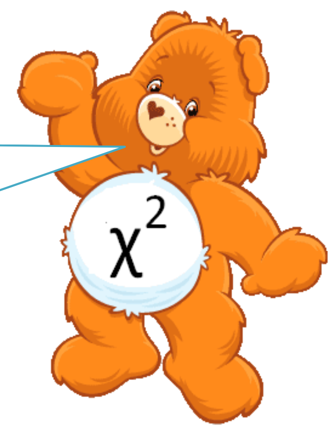
- The Bayesian Information Criterion is defined as...

$$-2 \cdot \ln p(x|k) \approx \mathrm{BIC} = -2 \cdot \ln L + k \ln(n).$$

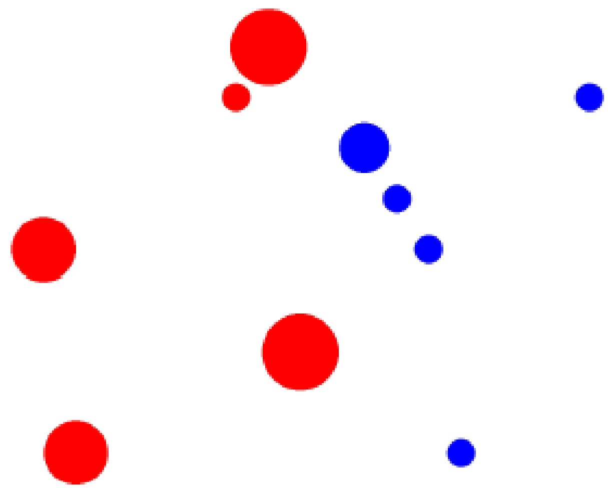
- The Akaike Information Criterion is defined as...

$$AIC = 2k - 2\ln(L)$$

- Model L.R. is penaliz



Plural Comparison
Introduction

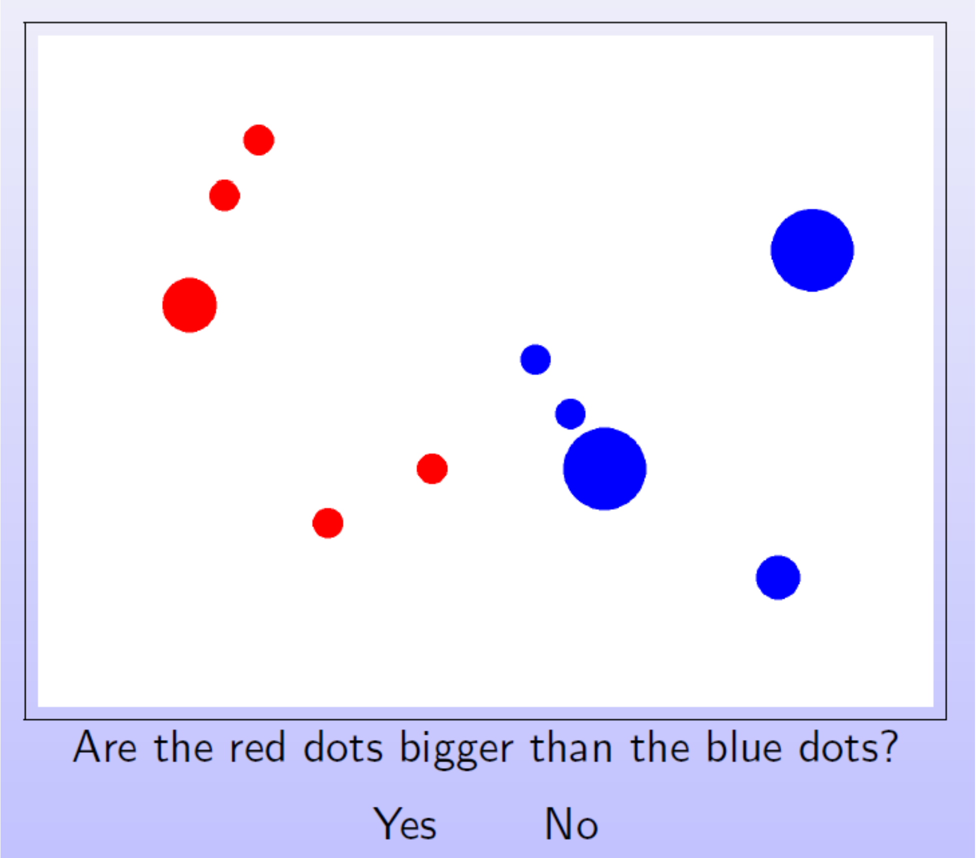


- Are the red circles bigger than the blue circles?

- The intuitions people have about the truth of sentences involving comparison of pluralities does not follow straightforwardly from the semantics of plural and the semantics of comparison.

Plural Comparison
Experiment

- Five red dots and five blue dots differing in size.
- xy-coordinates for the dots chosen at random.
- No blue dot ever appeared to the left of a red dot.
- 32 scenarios where model predictions differed maximally.
- Online questionnaire
- Stimuli presented in 1 of 4 random orders.
- Forced choice task.
- Subjects recruited through Amazon's Mechanical Turk (N=42).

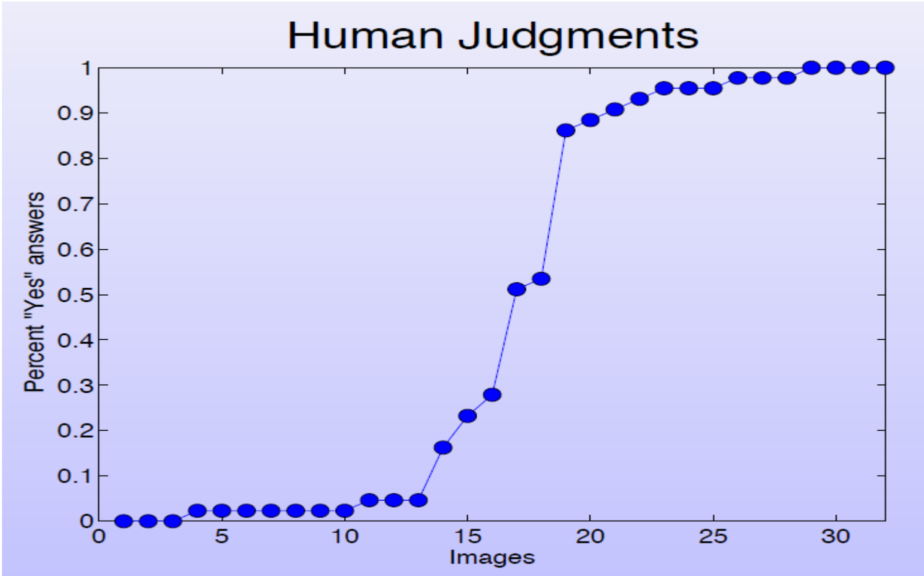


Plural Comparison
Three Models

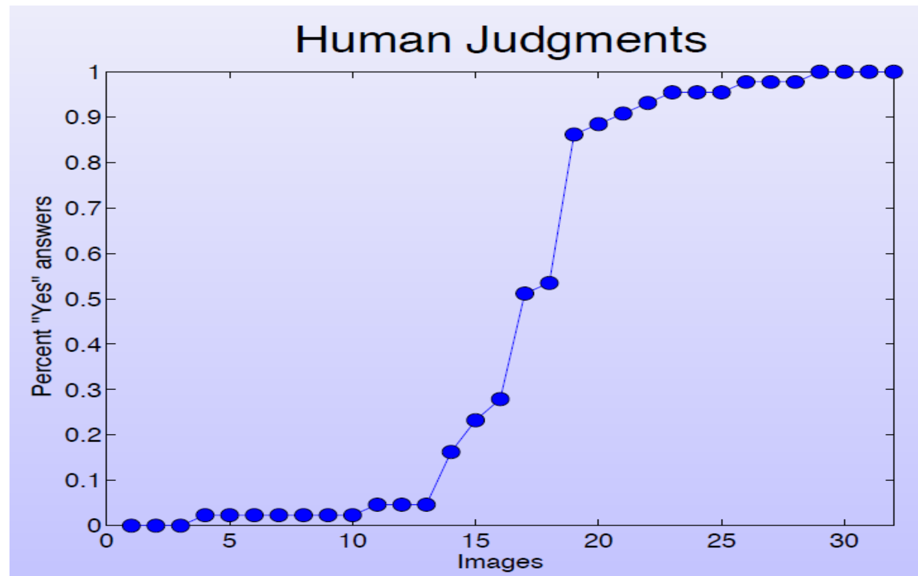
MatuRuys:
X>Y iff each member of X is bigger than some member of Y and each member of Y is smaller than at least one member of X.

CatMean:
X > Y iff mean(X) > mean (Y)

ProbMean:
½*[1+erf(mean(X)-mean(Y)]



Plural Comparison
Three Models



MatuRuys:

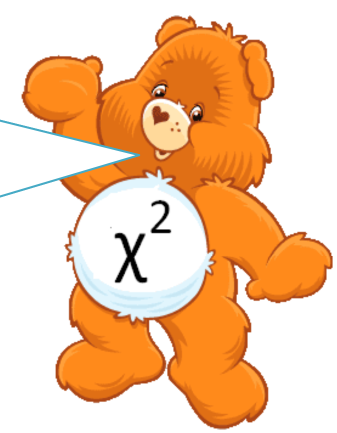
X>Y iff each member of X is bigger than some member of Y and each member of Y is smaller than at least one member of X.

CatMean:

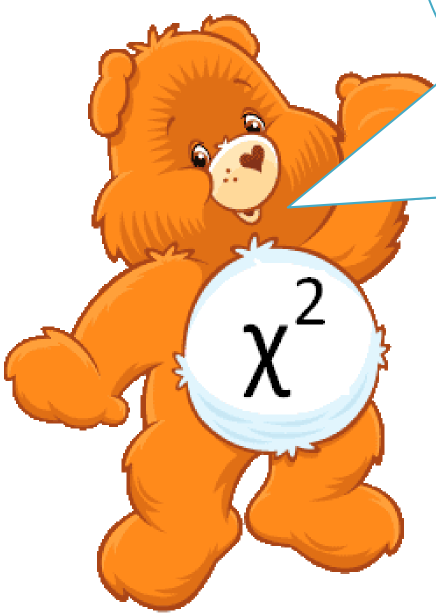
X > Y iff mean(X) > mean (Y)

ProbMean:

½*[1+erf(mean(X)-mean(

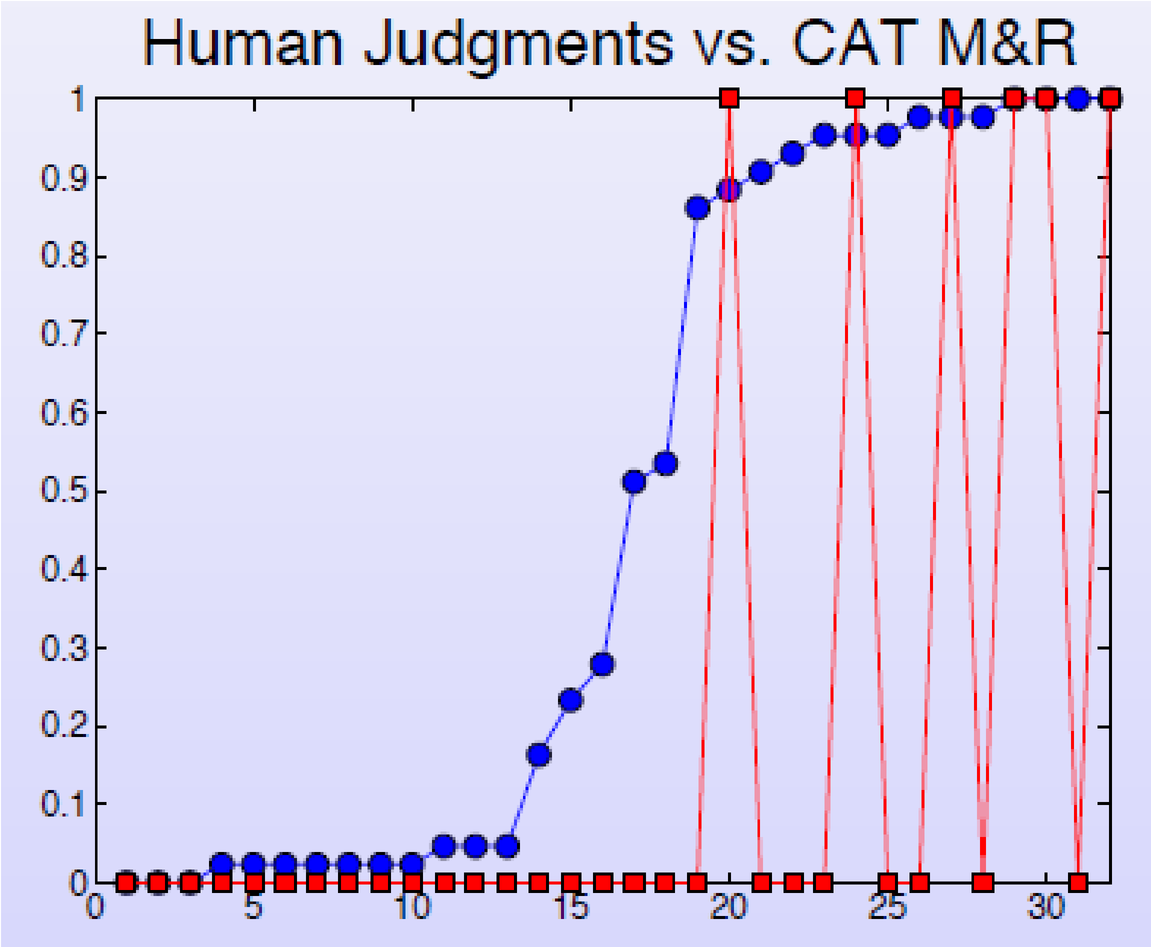


Plural Comparison
Your Turn!

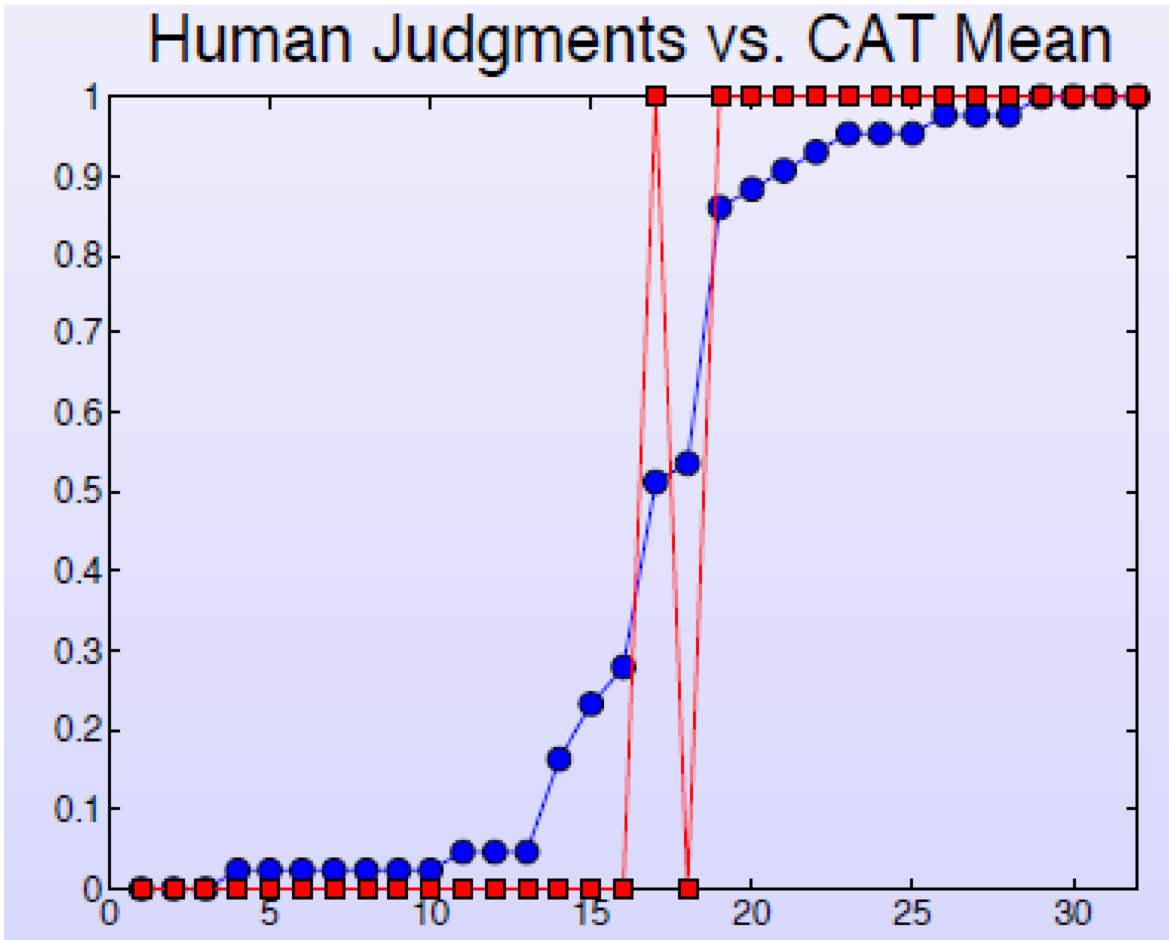


Try to figure out which model best explains the human judgments using Chi-Square Likelihood Test, AIC and BIC!

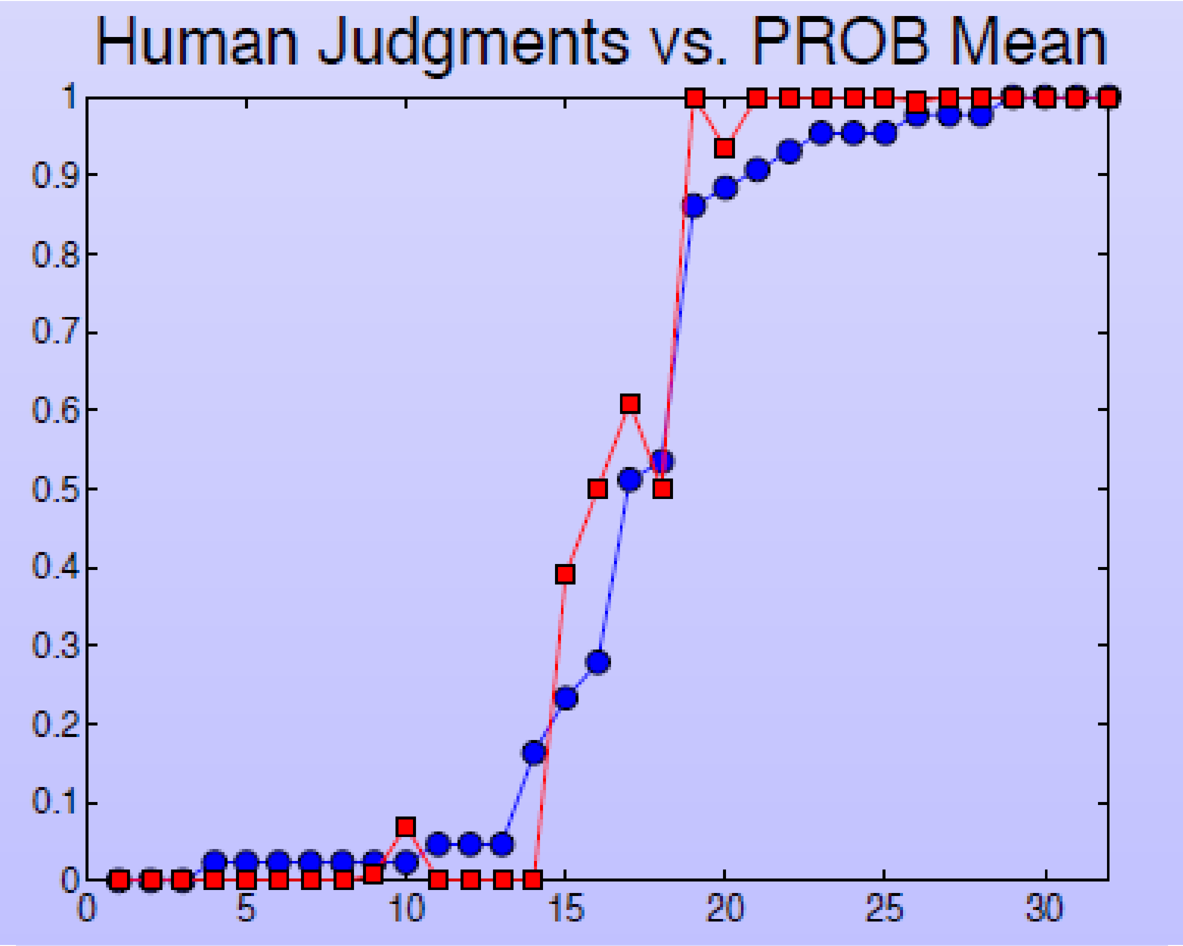
Plural Comparison
Your Turn!



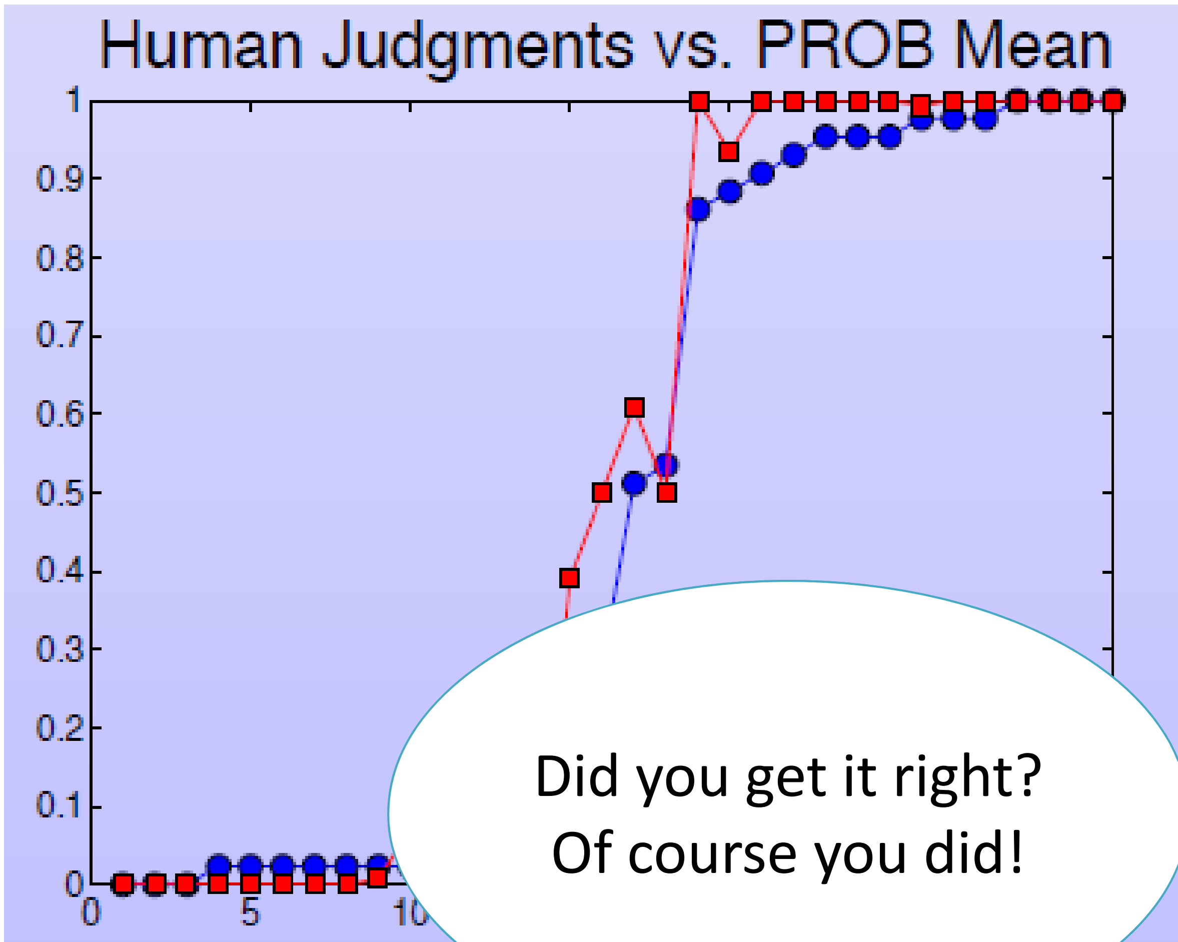
Plural Comparison
Your Turn!



Plural Comparison
Your Turn!



Plural Comparison
Your Turn!



Did you get it right?
Of course you did!