Outline

Maximum Likelihood (ML) estimation

Likelihood: probability of the data given a specific model and parameter values.
Model, for the corn example:

$$y_i = \mu + \alpha_{j[i]} + e_i \quad \text{with } \alpha_j \sim \mathcal{N}(0, \sigma^2_\alpha), e_i \sim \mathcal{N}(0, \sigma^2_e),$$

where $j[i] =$ site #, $i =$ plot #.

Marmoset example or more generally: fixed effects β and one or more level(s) of random effects:

$$y_i = X\beta + \gamma_{k[j[i]]} + \alpha_{j[i]} + e_i$$

$$\text{with } \gamma_k \sim \mathcal{N}(0, \sigma^2_\gamma), \alpha_j \sim \mathcal{N}(0, \sigma^2_\alpha), e_i \sim \mathcal{N}(0, \sigma^2_e),$$

Maximum Likelihood estimation

We estimate fixed effects (μ or β) and variances of random effects (σ_γ, σ_α, σ_e, etc.) by minimizing

$$-2 \log \mathrm{Lik}_{\mathrm{ml}} = n \log(2\pi) + \underbrace{\log(|V|)}_{n \log(\sigma_e^2)} + \underbrace{R^t V^{-1} R}_{\mathrm{RSS}/\sigma_e^2 \text{ in standard regression}}$$

where

- $V = V(\sigma_e, \sigma_\alpha, \sigma_\gamma)$ is a matrix that depends on the variance components, and
- $R = Y - X\hat{\beta}$ is the vector of residuals when the predicted values are based on fixed effects only.

Maximum Likelihood estimation

- If we know the variance components with certainty, then $\hat{\beta}$ has a normal distribution if the model is true. We use $\text{var}(\hat{\beta}) = (X^t \hat{V}^{-1} X)^{-1}$ to test hypotheses and get CI.

  Corn example: $\hat{\mu} = 4.29$ with estimated SE of 0.56.
  Marmoset: $\gamma_{\text{male}} = -7.1$ s with estimated SE of 22.7 s.

- The estimates $\hat{\sigma}^2_e$, $\hat{\sigma}^2_\alpha$, etc. are not normally distributed.

- Biased estimates of σ^2_e, σ^2_α, etc.

  Most simple case: one random sample $y_i = \mu + e_i$. With ML, $\hat{\sigma}^2_e = \sum(y_i - \bar{y})^2/n$, which is too small on average. We usually use:

Restricted Maximum Likelihood (REML) estimation

- Provides unbiased estimates of variance components if the model is correct.
- Roughly: restrict the data to $n - p$ modified observations, which are independent of β. Then maximize the likelihood of these restricted modified observations.

Most simple case: one random sample, $y_i = \mu + e_i$, $e_i \sim \mathcal{N}(0, \sigma_e^2)$.
We can restrict the data to the $n - 1$ observations

$$y_2 - y_1, y_3 - y_2, \ldots, y_{n-1} - y_n$$

Their distribution is independent of μ, depends on σ_e^2 only. The ML estimate of σ_e^2 based on these restricted data is

$$\hat{\sigma}_e^2 = \sum (y_i - \bar{y})^2 / (n - 1)$$

Restricted Maximum Likelihood (REML) estimation

Standard linear regression: we get $\hat{\sigma}^2 = \sum(y_i - \bar{y})^2/(n-p)$ where p is the number of coefficients, $n-p =$ residual df.

We've been using REML for a long time.

Mixed models: we minimize the REML criterion

$$-2\log \text{Lik}_{\text{reml}} = (n-p)\log(2\pi) + \log(|V|) + R^t V^{-1} R + \log(|X^t V^{-1} X|)$$

where

- the matrix $V = V(\sigma^2_\alpha, \sigma^2_e)$ is as before, residuals $R = Y - X\hat{\beta}$ as before.
- X are the predictors with fixed effects β.

ML versus REML

REML-based variances:

- good for **unbiased** variance estimates.
- same residual variance estimate as we have used before on fixed-effect only models.
- same variance estimates as ANOVA methods when the design is balanced (more later).

Comparing ML and REML:

- The extra term in REML (last one) depends on X. The REML criterion can only be compared across models that have the same X, to ensure we compare models based on the same restricted data.
- ML: the only method to perform **likelihood ratio tests**, when testing fixed effects β.

ML versus REML in R

REML is the default:

```
> lmer(harvwt ~ (1 | site), data = corn)

Linear mixed model fit by REML
...
   AIC   BIC logLik deviance REMLdev
 194.9 201.4 -94.47    189.6   188.9
Random effects:
 Groups   Name        Variance Std.Dev.
 site     (Intercept) 2.41652  1.55452
 Residual             0.76477  0.87451
Number of obs: 64, groups: site, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   4.2917     0.5603   7.659
```

ML versus REML in R

Variances and SE are slightly smaller with ML:

```
> lmer(harvwt ~ (1 | site), data = corn, REML=FALSE)

Linear mixed model fit by maximum likelihood
...
   AIC   BIC logLik deviance REMLdev
 195.5   202 -94.77    189.5     189
Random effects:
 Groups   Name        Variance Std.Dev.
 site     (Intercept) 2.10251  1.45000
 Residual             0.76477  0.87451
Number of obs: 64, groups: site, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   4.2917     0.5242   8.188
```

Testing random effects

Corn example. We might want to test the presence of site effects:

$$H_0\text{: } \sigma_\alpha = 0 \text{ versus the alternative } H_A\text{: } \sigma_\alpha > 0.$$

This question is about the whole population of sites, not just the 8 sampled sites.

Warning: variance estimates are usually not normally distributed.

- For one thing, they are always positive.
- That's why `lmer` does not output any SE for variance components: $\hat{\sigma}^2_\alpha \pm 2\text{SE}$ would a bad confidence intervals.

Testing random effects with a Likelihood ratio test

Compare a simple, null model with an alternative model, which has k more variance parameters. Test statistic:

$$X^2 = -2 \log \text{Lik}(\text{null}) + 2 \log \text{Lik}(\text{alternative})$$

What is its null distribution?
Chi-square-based p-value: compare X^2 to $\chi^2_{\text{df}=k}$. **But:**

- the chi-square distribution is inappropriate when we test a 'borderline' parameter. $\sigma^2_\alpha = 0$ is borderline. The resulting p-value is too large, the conclusion is **conservative**.
- In simple cases, the appropriate distribution is $X^2 = 0$ with 50% chance and $X^2 \sim \chi^2_1$ with 50% chance. If so, the appropriate p-value is half that obtained by comparing x^2 to χ^2_1.

Parametric-bootstrap-based p-value: simulate data under the null model ($\sigma^2_\alpha = 0$) to get the true distribution of X^2 under H_0. More computer intensive, but **more accurate**.

Flowering time: Testing the inventory effect

- We can use the REML likelihood **or** the ML likelihood here.
- `anova()` uses ML on `lmer` models.
- `anova()` provides the chi-square-based p-value.

```
> fit = lmer(logdtf ~ (1|subspecies)+(1|inventoryID), data=brassica2)
> fit.noinventory = update(fit, .~.- (1|inventoryID))
> anova(fit, fit.noinventory)

Data: brassica2
Models:
fit.noinventory: logdtf ~ (1 | subspecies)
fit: logdtf ~ (1 | subspecies) + (1 | inventoryID)

                Df     AIC     BIC logLik Chisq Chi Df Pr(>Chisq)
fit.noinventory  3  -23.33  -14.66  14.66
fit              4 -117.97 -106.40  62.98 96.64      1  < 2.2e-16 ***
```

Since $p < 2.2 * 10^{-16}$ is conservative (too high), we can confidently say that there is strong evidence for an inventory effect: very strong evidence that $\sigma^2_{\text{inventory}} > 0$.

Corn: testing the site effect

- Need to use `lm` instead of `lmer` for the null model, because it has no random effect.
- make sure we use the same criterion in both
- LRT with p-value based on the chi-square distribution:

```
> corn.lmer = lmer(harvwt ~ (1 | site), data = corn)
> corn.lm   = lm(  harvwt ~  1        , data = corn)
> x2 = -2*logLik(corn.lm, REML=T) +2*logLik(corn.lmer, REML=T)
> x2
[1] 61.368
> pchisq(x2, df=1, lower.tail=F)
[1] 4.734755e-15
```

The appropriate p-value would be even lower. Strong evidence that $\sigma^2_{\text{site}} > 0$.

Compare with the ANOVA F-test on fixed effects:

```
> anova(lm(harvwt ~ site , data=corn))
          Df  Sum Sq Mean Sq F value    Pr(>F)
site       7 140.679 20.0969  26.278 1.551e-15 ***
Residuals 56  42.827  0.7648
```

Pygmy marmosets: Testing group effects

5 populations, 2-3 groups/population, 2 individuals/group (M,F).

Response: we will look at the number of notes/call in J calls.

```
> fit.spg = lmer(notes ~ sex + (1|population) + (1|group), data=jcalls)
> summary(fit.spg)
   AIC   BIC logLik deviance REMLdev
 131.2 137.9 -60.61    122.8   121.2

Random effects:
 Groups     Name        Variance   Std.Dev.
 group      (Intercept) 3.4310e+00 1.85228248
 population (Intercept) 6.7060e-09 0.00008189
 Residual               2.6839e+00 1.63825737

Fixed effects:
            Estimate Std. Error t value
(Intercept)  18.4214     0.6609  27.873
sexM          0.3000     0.6192   0.484

> fit.sp = update(fit.spg, .~. - (1|group))
> summary(fit.sp)
   AIC   BIC logLik deviance REMLdev
 134.1 139.5 -63.07    128.1   126.1

Random effects:
 Groups     Name        Variance   Std.Dev.
 population (Intercept) 5.2577e-10 2.2930e-05
 Residual               6.1149e+00 2.4728e+00

Fixed effects:
            Estimate Std. Error t value
(Intercept)  18.4214     0.6609  27.874
sexM          0.3000     0.9346   0.321
```

Pygmy marmosets: Testing group effects

Remarks on the output from the `summary()`

- deviance = -2*logLik. The “deviance” value corresponds to the ML criterion, the “REMLdev” value is the deviance corresponding to the REML criterion.
- the “logLik” value corresponds to the criterion that was optimized: the REML log-likelihood here. We can check that -2*logLik = $-2 * (-60.61) = 121.2$ is the REML deviance.
- Just from this output, we can compare the two models
  based on ML: $x^2 = 128.1 - 122.8 = 5.3$
  based on REML: $x^2 = 126.1 - 121.2 = 4.9$
  based on AIC: 131.2 (full) versus 134.1 (no group effect)

Testing group effects with LRT, chi-square based

Chi-square based p-value for the LRT:

```
> anova(fit.sp, fit.spg)
Models:
fit.sp: notes ~ sex + (1 | population)
fit.spg: notes ~ sex + (1 | population) + (1 | group)
        Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fit.sp   4 136.09 141.41 -64.043
fit.spg  5 132.79 139.45 -61.396 5.2932      1    0.02141 *
```

We know the correct p-value should be smaller than 0.02141, perhaps half: 0.0107 ?

Testing group effects with LRT, simulation based

- `simulate()` simulates one (or more) sets of observed responses under the model we provide.

Here, we want to simulate under H_0: no group effect. Will use our fitted model *without* group effect, `fit.sp`.

```
> y = simulate(fit.sp)
> y
      [,1]
 [1,] 19.7
 [2,] 20.0
 [3,] 19.8
 [4,] 16.8
...
[26,] 19.1
[27,] 20.5
[28,] 19.4
```

Testing group effects with LRT, simulation based

We simulate data and X^2 under H_0: do this once first.

```
> y  = simulate(fit.sp)
> f.null = lmer(y ~ sex + (1 | population)             , data=jcalls)
> f.alt  = lmer(y ~ sex + (1 | population) + (1 | group), data=jcalls)
> anova(f.null, f.alt)
       Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
f.null  4 131.47 136.80 -61.737
f.alt   5 133.28 139.94 -61.639 0.1958      1     0.6581
> str(anova(f.null, f.alt))
 $ Df         : int  4 5
 $ AIC        : num  131 133
 $ BIC        : num  137 140
 $ logLik     : num  -61.7 -61.6
 $ Chisq      : num  NA 0.196
 $ Chi Df     : int  NA 1
 $ Pr(>Chisq): num  NA 0.658
 - attr(*, "heading")= chr  "Data: jcalls" ...
> anova(f.null, f.alt)$Chisq
[1]        NA 0.1958155
> anova(f.null, f.alt)$Chisq[2]
[1] 0.1958155
```

Testing the inventory effect: LRT with simulations

Now write a function to do all this in one easy step:

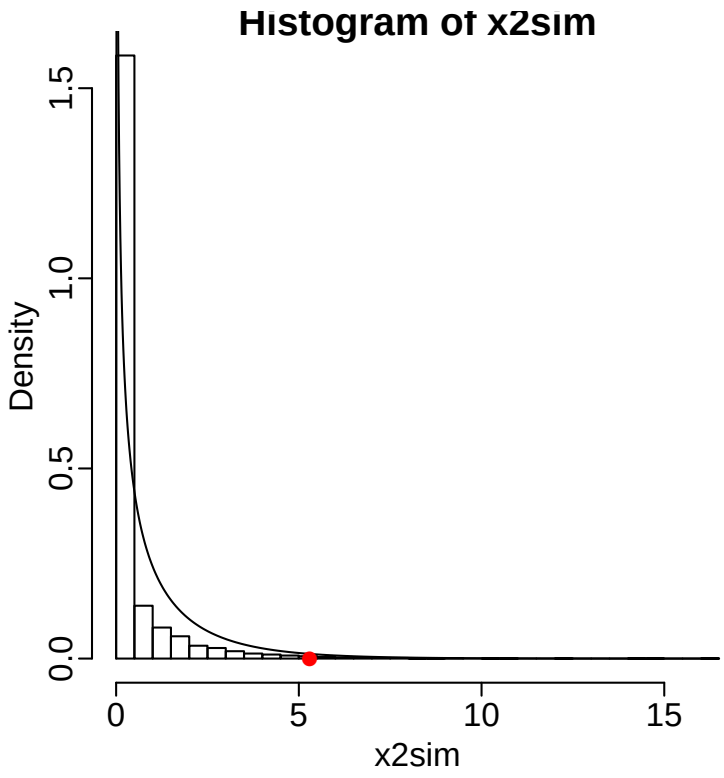
```
oneX2 = function(){
 y = simulate(fit.sp)
 f.null = lmer(y ~ sex + (1 | population)                , data=jcalls)
 f.alt  = lmer(y ~ sex + (1 | population) + (1 | group), data=jcalls)
 return( anova(f.null, f.alt)$Chisq[2] )
}
```

```
> oneX2()
[1] 0
> oneX2()
[1] 0.1842820
> oneX2()
[1] 0.05587135
> oneX2()
[1] 2.315413
> oneX2()
[1] 0
> oneX2()
[1] 4.063094
```

Then repeat this simulation under the null model 10,000 times...

Testing the inventory effect: LRT with simulations

```
> x2sim = replicate(10000, oneX2())
> hist(x2sim, freq=F, breaks=30)  # histogram whose area is 1
> curve(dchisq(x,df=1), add=T, n=1001) # option n to get more precisio
> points(x=5.2932, y=0, col="red", pch=16)
> sum(x2sim>5.2932) / 10000
[1] 0.0089
```



χ^2_1 is a conservative approximation of the null distribution. The real p-value is smaller than that provided by χ^2_1.

Conclusion: strong evidence for a group effect: $\sigma^2_{\text{group}} > 0$ (p=.009). Individuals from the same group use a more similar number of notes than individuals from different groups.

Outline

Testing fixed effects with LRT, chi-square distribution

$$X^2 = -2 \log \text{Lik}(H_0) + 2 \log \text{Lik}(H_A)$$

- Approximately, with large sample sizes, X^2 has a **chi-square distribution** on $df = k$ if H_0 is true.
- $k = \#\ \text{coef}(H_A) - \#\ \text{coef}(H_0)$
- Only the ML likelihood is appropriate. The REML is not: the data are not restricted the same way under H_0 and under H_A, so the REML likelihoods are likelihoods of different data sets: they are not comparable.
- Warning: again, the χ^2_k distribution is not always a very good approximation of the null distribution. It tends to be anti-conservative: the χ^2_k p-value tends to be too small.

Pygmy marmosets: Testing a gender effect with LRT

Test statistic: $X^2 = -2 \log \text{Lik}(H_0) + 2 \log \text{Lik}(H_A)$ where
H_0: model with population and group as random effects
H_A: population, group as random effects, sex as fixed effect.

Two ways to obtain the p-value

- Compare X^2 to a chi-square distribution. Quick and easy: `anova( fit.simple.Ho, fit.complex.Ha)`
- Compare X^2 to the distribution of X^2 simulated under H_0. Computationally more intensive, but could be more accurate.

Testing the gender effect with anova

```
> fit.spg = lmer(notes ~ sex + (1|population)+(1|group), data=jcalls)
> summary(fit.spg)
Linear mixed model fit by REML ...
   AIC   BIC logLik deviance REMLdev
 131.2 137.9 -60.61    122.8   121.2
Fixed effects:
            Estimate Std. Error t value
(Intercept)  18.4214     0.6609  27.873
sexM          0.3000     0.6192   0.484

> fit.pg = update(fit.spg, .~. - sex)
> fit.pg
Linear mixed model fit by REML ...
   AIC   BIC logLik deviance REMLdev
 130.3 135.6 -61.16    123.1   122.3

> anova(fit.pg, fit.spg)
...     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fit.pg   4 131.06 136.39 -61.531
fit.spg  5 132.79 139.45 -61.396 0.2698      1     0.6035
```

What likelihood was used by anova?

Testing the gender effect with anova

anova makes sure to use the ML criterion, even though the models were fit with REML:

```
> anova(fit.pg, fit.spg)
        Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fit.pg   4 131.06 136.39 -61.531
fit.spg  5 132.79 139.45 -61.396 0.2698      1     0.6035
> logLik(fit.pg)
'log Lik.' -61.15946 (df=4)
> logLik(fit.spg)
'log Lik.' -60.61334 (df=5)
```

Recall AIC$= -2 \log \text{Lik} + 2p$ where $p =$# parameters.

- What is the ML and REML AICs for these models?
- Which AIC values can we compare here?

- Based on AIC, which model would we prefer?

Testing fixed effects with LRT and simulations

Alternatively: simulate new data and their X^2 values under H_0, compare the original X^2 to the simulated X^2 to get a p-value.

```
> y = simulate(fit.pg)  # should be the null model: assumes no fixed e
> f.alt  = lmer(y ~ sex + (1|population) + (1|group), data=jcalls)
> f.null = update( f.alt, .~. - sex)
> anova(f.null,f.alt)
f.null: y ~       (1 | population) + (1 | group)
f.alt: y ~ sex + (1 | population) + (1 | group)
       Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
f.null  4 129.74 135.07 -60.872
f.alt   5 130.47 137.13 -60.236 1.2706      1     0.2597
> anova(f.null,f.alt)$Chisq[2]
[1] 1.270615
```

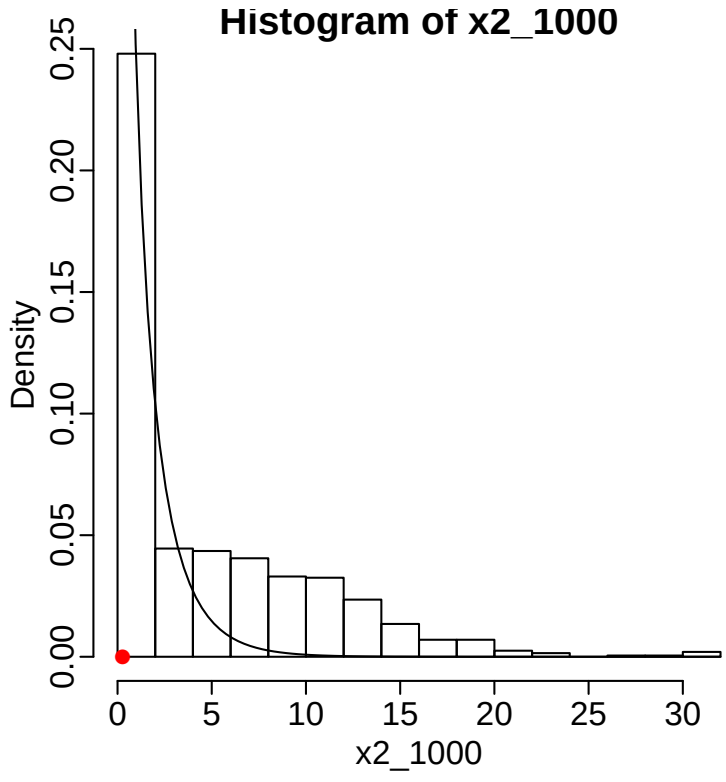
Our simulation worked once, Now make it a function:

```
oneLR = function(){
 y = simulate(fit.pg)  # y simulated under the null model
 f.alt  = lmer(y ~ sex + (1|population) + (1|group), data=jcalls)
 f.null = update( f.alt, .~. - sex)
 return( anova(f.null,f.alt)$Chisq[2] )
}
> oneLR()
[1] 3.549157
```

Testing fixed effects with LRT and simulations

Repeat many times, summarize these 'null' X^2 values.

```
> x2_1000 = replicate(1000, oneLR())
> hist(x2_1000, freq=F, breaks=20)  # freq=F for a histogram with area
> curve( dchisq(x,df=1), add=T)
> points(x=0.2698, y=0, col="red", pch=16)
> sum(x2_1000>0.2698) / 1000
[1] 0.603
```



We obtain a p-value 0.603 very close to what we had from using a chi-square distribution (0.6035).

But also see that the true null distribution has a heavier tail than the χ^2_1 distribution.

Language and learning flexibility: Case study

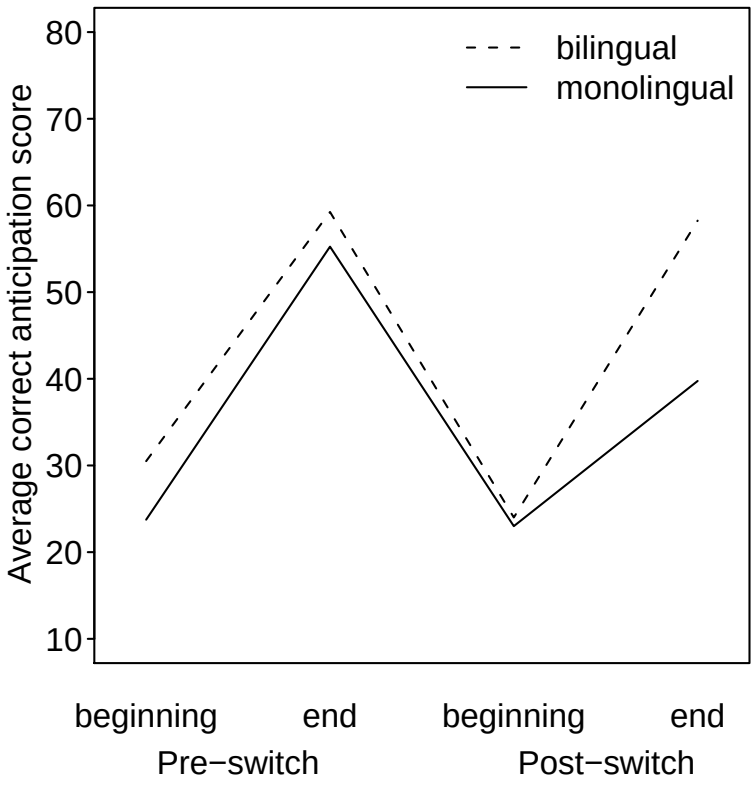
Case study to illustrate:

- Repeated measures
- Testing more complex fixed effects

Kovacs & Mehler (2009) PNAS 106:6556: study to determine whether bilingual infants have an increased capacity to adapt to a new environment. Eye-tracking experiment, 7-month-old infants. Visual reward (cartoon of a dancing puppet) on one side of a computer screen.

1. pre-switch phase: reward given on the same side of the screen several times. Children learned to anticipate and look to the side of the screen where the picture would be coming.
2. post-switch phase: reward repeatedly given on the other side of the screen. How fast do the children suppress their previous learning and “relearn”?

Language and learning flexibility: Data



Measured: intensity and frequency with which the children looked to the correct side.

Data: for each child, correct anticipation score on a 0-100 scale at beginning and end of the pre-switch and post-switch phases: 4 anticipation scores for each child.

40 children total: 20 bilingual, 20 monolingual, paired.

Data simulated based on the paper's figure.

Language and learning flexibility: Data

```
> dat = read.table("bilingual.txt", header=T)
> dat$phase = relevel(dat$phase, "pre-switch")
'data.frame':   160 obs. of  6 variables:
 $ pair    : Factor w/ 20 levels "p1","p10","p11",..: 1 1 1 1 1 1 1 1
 $ child   : Factor w/ 40 levels "c1","c10","c11",..: 1 1 1 1 12 12 12
 $ language: Factor w/ 2 levels "bilingual","monolingual": 2 2 2 2 1 1
 $ phase   : Factor w/ 2 levels "pre-switch","post-switch": 1 1 2 2 1
 $ time    : Factor w/ 2 levels "beginning","end": 1 2 1 2 1 2 1 2 1 2
 $ score   : int  25 50 20 25 35 55 15 45 25 35 ...
```

What predictors to include in our model?

- language, phase, time: as fixed effects. Each 2 levels.
- Interactions?

- Repeated measures: each child provides 4 scores. Include child as random effect.
- Pairing: monolingual child paired with bilingual child. Include pair as random effect.

Repeated measures: random child effect

```
> fit = lmer(score ~ phase * time * language + (1|pair) + (1|child),
                                                          data=dat)
> fit
Linear mixed model fit by REML
  AIC  BIC logLik deviance REMLdev
 1195 1228 -586.3     1201    1173

Random effects:
 Groups   Name        Variance Std.Dev.
 child    (Intercept) 26.579   5.1555
 pair     (Intercept)  0.000   0.0000
 Residual             92.590   9.6224
Number of obs: 160, groups: child, 40; pair, 20

Fixed effects:
                                          Estimate Std.Error t value
(Intercept)                                  24.00     2.441   9.832
phasepre-switch                               6.50     3.043   2.136
timeend                                      34.25     3.043  11.256
languagemonolingual                          -1.00     3.452  -0.290
phasepre-switch:timeend                      -5.50     4.303  -1.278
phasepre-switch:languagemonolingual          -5.75     4.303  -1.336
timeend:languagemonolingual                 -17.50     4.303  -4.067
phasepre-switch:timeend:languagemonolingual 20.25     6.086   3.327
```

Testing the 3-way interaction (fixed)

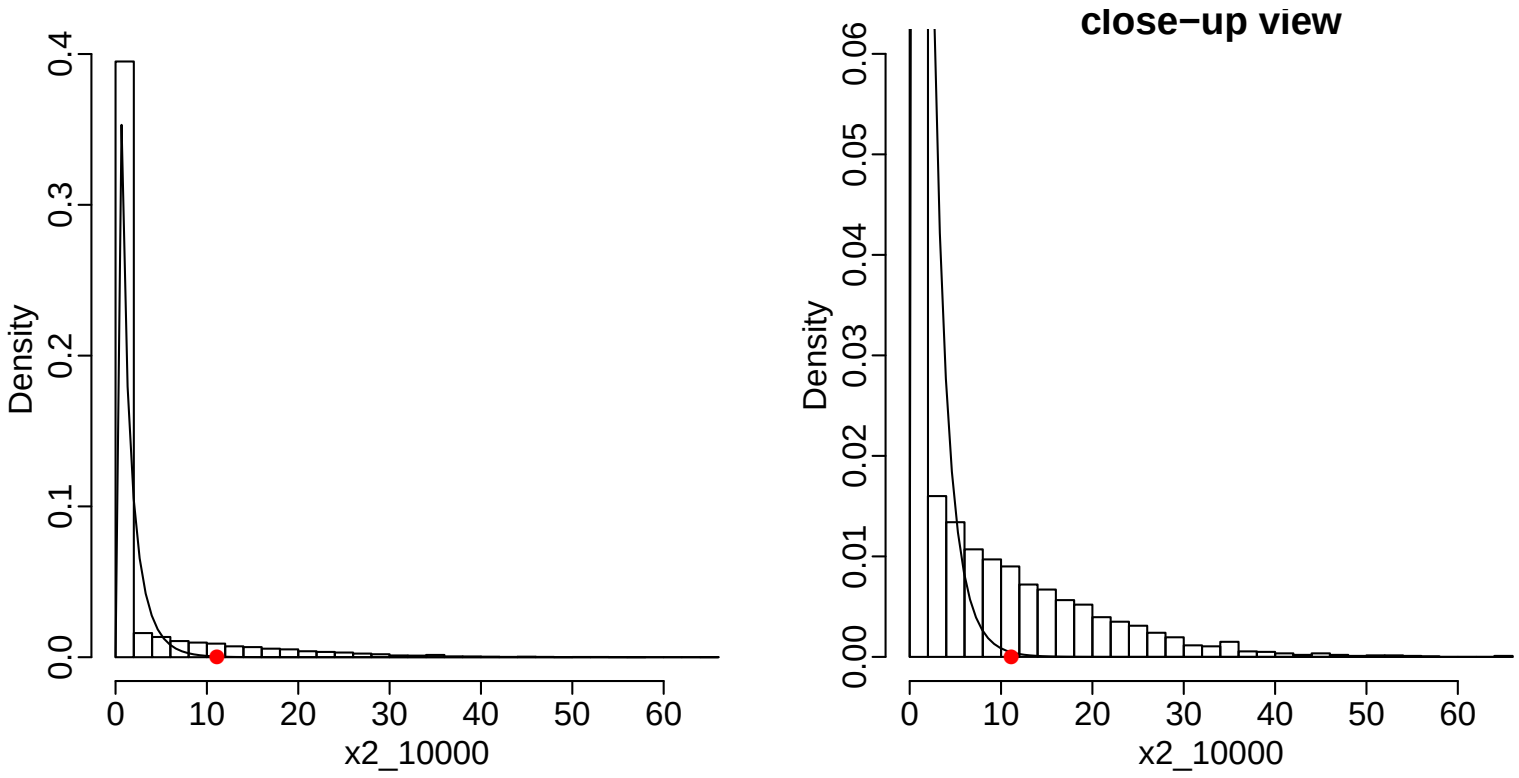
```
> fit.no3way = update(fit, .~. - phase:time:language)
> anova(fit.no3way, fit)
           Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fit.no3way 10 1232.1 1262.8 -606.04
fit        11 1223.0 1256.8 -600.47 11.124      1  0.0008522 ***
```

Assuming the χ^2_1 distribution is a good approximation to the null distribution, very strong evidence for a 3-way interaction.

Testing the 3-way interaction (fixed)

```
oneLR = function(){
 y = simulate(fit.no3way)
 f.alt  = lmer(y ~ phase*time*language +(1|child), data=dat)
 f.null = update( f.alt, .~. - phase:time:language)
 return( anova(f.null,f.alt)$Chisq[2] )
}
> oneLR()
[1] 22.91408
> oneLR()
[1] 0
> oneLR()
[1] 1.907933
> oneLR()
[1] 0
> x2_10000 = replicate(10000, oneLR())
> hist(x2_10000, freq=F, breaks=30)  # freq=F to get area=1
> curve( dchisq(x,df=1), add=T)
> points(x=11.124, y=0, col="red", pch=16)
> sum(x2_10000>11.124) / 10000
[1] 0.0998
```

Testing the 3-way interaction (fixed)



The chi-square χ^2_1 distribution does not approximate the null distribution very well in the tail.

Our test statistic $X^2 = 11.124$ is quite far in the tail...

Testing the 3-way interaction (fixed)

- The χ^2_1-based p-value was VERY anti-conservative. Correct conclusion: The Likelihood-ratio test provides no evidence of a 3-way interaction.
- to test the 3-way interaction, we dropped ‘pair’ from the model, as there was no evidence at all for a random effect of pair ($\hat{\sigma}^2_{\text{pair}} = 0$)
- to specifically use the ML criterion rather than REML, use option `REML=F` in `lmer(formula, data= , REML=F)`
- We will later see an F-test to test for fixed effects, which can be more powerful than the LRT.