Lecture 1: Statistical methods for linguistic research: Advanced Tools

Shravan Vasishth

Department of Linguistics
University of Potsdam, Germany

August 9, 2015

Motivating this course

- In psychology and linguistics, we usually use frequentist methods to analyze our data.
- The R library I use most is `lme4`.
- In recent years, very powerful programming languages have become available that make Bayesian modeling relatively easy.
- Bayesian tools have several important advantages over frequentist methods, but they require some very specific background knowledge.
- My goal in this course is to try to provide that background knowledge.

Motivating this course

In this introductory lecture, my goals are to

1. motivate you look at Bayesian Linear Mixed Models as an alternative to using frequentist methods.
2. make sure we are all on the same page regarding the moving parts of a linear mixed model.

Prerequisites

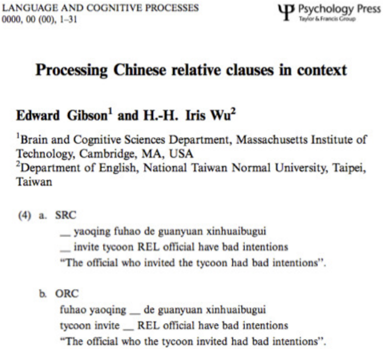
1. Familiarity with fitting standard LMMs such as:

   ```
   lmer(rt~cond+(1+cond|subj)+(1+cond|item),dat)
   ```

2. Basic knowledge of R.
3. A little bit of fearlessness in the face of basic algebra and some very simple calculus derivations (kept to a minimum).
4. For detailed derivations and more background theory, please see:
   1. my linear modeling notes: https://github.com/vasishth/LM
   2. my Bayesian modeling notes: to-do

Linear mixed models

Example: Gibson and Wu data, Language and Cognitive Processes, 2013



LANGUAGE AND COGNITIVE PROCESSES
0000, 00 (00), 1–31

Psychology Press
Taylor & Francis Group

Processing Chinese relative clauses in context

Edward Gibson[1] and H.-H. Iris Wu[2]

[1]Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of English, National Taiwan Normal University, Taipei, Taiwan

(4) a. SRC
__ yaoqing fuhao de guanyuan xinhuaibugui
__ invite tycoon REL official have bad intentions
"The official who invited the tycoon had bad intentions".

b. ORC
fuhao yaoqing __ de guanyuan xinhuaibugui
tycoon invite __ REL official have bad intentions
"The official who the tycoon invited had bad intentions".

Linear mixed models

Example: Gibson and Wu data, Language and Cognitive Processes, 2013

```
headnoun<-subset(data,region=="headnoun")
## no. of subjects:
length(unique(headnoun$subj))

## [1] 37

## no. of items:
length(unique(headnoun$item))

## [1] 15

## no. of rows in data frame:
dim(headnoun)[1]

## [1] 547
```

Linear mixed models

Example: Gibson and Wu data, Language and Cognitive Processes, 2013

```
head.means<-aggregate(rt~subj+type,
                      mean,data=headnoun)
```

Linear mixed models

Example: Gibson and Wu data, Language and Cognitive Processes, 2013

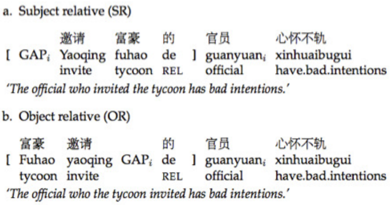
```
t.test(log(subset(head.means,type=="subj-ext")$rt),
       log(subset(head.means,type=="obj-ext")$rt),
       paired=T)

##
##  Paired t-test
##
## data:  log(subset(head.means, type == "subj-ext")$rt) a
## t = 1.9826, df = 36, p-value = 0.05509
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -0.003513238  0.309397834
## sample estimates:
## mean of the differences
##               0.1529423
```

Linear mixed models

Example: Gibson and Wu data, Language and Cognitive Processes, 2013

- Subject vs object relative clauses in Chinese, self-paced reading.
- The critical region is the head noun.
- The goal is to find out whether ORs are harder to process than SRs at the head noun.



a. Subject relative (SR)

		邀请	富豪	的		官员	心怀不轨
[	GAP$_i$	Yaoqing	fuhao	de	]	guanyuan$_i$	xinhuaibugui
		invite	tycoon	REL		official	have.bad.intentions

'The official who invited the tycoon has bad intentions.'

b. Object relative (OR)

	富豪	邀请		的		官员	心怀不轨
[	Fuhao	yaoqing	GAP$_i$	de	]	guanyuan$_i$	xinhuaibugui
	tycoon	invite		REL		official	have.bad.intentions

'The official who the tycoon invited has bad intentions.'

Linear mixed models

Example: Gibson and Wu 2012 data

```
head(data[,c(1,2,3,4,7,10,11)])

##    subj item     type pos   rt        rrt  x
## 7     1   13  obj-ext   6 1140 -0.8771930  1
## 20    1    6 subj-ext   6 1197 -0.8354219 -1
## 32    1    5  obj-ext   6  756 -1.3227513  1
## 44    1    9  obj-ext   6  643 -1.5552100  1
## 60    1   14 subj-ext   6  860 -1.1627907 -1
## 73    1    4 subj-ext   6  868 -1.1520737 -1
```

lme4 model of Gibson and Wu data

Crossed varying intercepts and slopes model, with correlation

```
## Loading required package:  Matrix
```

This is the type of "maximal" model that most people fit nowadays, following Barr et al 2013's Keep it Maximal paper:

```
m1 <- lmer(rrt~x+(1+x|subj)+(1+x|item),
          subset(data,region=="headnoun"))

m2 <- lmer(rrt~x+(1+x|subj)+(1|item),
          subset(data,region=="headnoun"))
```

lme4 model of Gibson and Wu data

Crossed varying intercepts and slopes model, with correlation

I will now show two major (related) problems that occur with the small datasets we usually have in psycholinguistics:

- The correlation estimates either lead to degenerate variance covariance matrices, and/or
- The correlation estimates are wild estimates that have no bearing with reality.

This is not a failing of `lmer`, but rather of the researcher: the model is overparameterized. The researcher is demanding too much of `lmer`.

lme4 model of Gibson and Wu data

Typical data analysis: Crossed varying intercepts and slopes model, with correlation

```
round(summary(m1)$coefficients,digits=3)

##             Estimate Std. Error t value
## (Intercept)   -2.672      0.138 -19.369
## x             -0.039      0.046  -0.835
```

The “best” model

One way to decide on the “best” model is to find the simplest model using the Generalized Likelihood Ratio Test (Pinheiro and Bates 2000). Here, this is the **varying intercepts** model, not the maximal model.

```
m1<- lmer(rrt~x+(1+x|subj)+(1+x|item),
          headnoun)
m1a<- lmer(rrt~x+(1|subj)+(1|item),
          headnoun)
```

The “best” model

```
anova(m1,m1a)

## refitting model(s) with ML (instead of REML)

## Data: headnoun
## Models:
## m1a: rrt ~ x + (1 | subj) + (1 | item)
## m1: rrt ~ x + (1 + x | subj) + (1 + x | item)
##     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq
## m1a  5 1603.5 1625.0 -796.76   1593.5
## m1   9 1608.5 1647.3 -795.27   1590.5 2.9742      4     0.562
```

Another approach is shown in Bates, Kliegl, Vasishth, Baayen: ArXiv preprint http://arxiv.org/abs/1506.04967.

How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data

Here, we simulate data with the same structure, sample size, and parameter values as the Gibson and Wu data, except that we assume that the correlations are 0.6. Then we analyze the data using `lmer` (maximal model). **Can lmer recover the correlations**?

How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data

We define a function called `new.df` that generates data similar to the Gibson and Wu data-set. For code, see accompanying lecture1.R file.

How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data

Next, we write a function that generates data for us repeatedly with the following specifications: sample size for subjects and items, and some correlation between subject intercept and slope, and item intercept and slope.

How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data

```
gendata<-function(subjects=37,items=15){
  dat<-new.df(nsubj=subjects,nitems=items,
              rho.u=0.6,rho.w=0.6)
  dat <- dat[[1]]
  dat<-dat[,c(1,2,3,9)]
  dat$x<-ifelse(dat$cond==1,-0.5,0.5)

return(dat)
}
```

How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data

Set number of simulations:

```r
nsim<-100
```

Next, we generate simulated data 100 times, and then store the estimated subject and item level correlations in the random effects, and plot their distributions.
We do this for two settings: Gibson and Wu sample sizes (37 subjects, 15 items), and 50 subjects and 30 items.

How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data

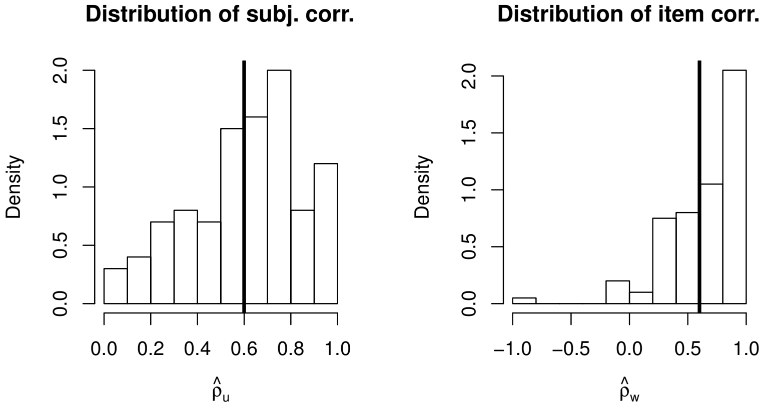
37 subjects and 15 items

```
library(lme4)
subjcorr<-rep(NA,nsim)
itemcorr<-rep(NA,nsim)

for(i in 1:nsim){
dat<-gendata()
m3<-lmer(rt~x+(1+x|subj)+(1+x|item),dat)
subjcorr[i]<-attr(VarCorr(m3)$subj,"correlation")[1,2]
itemcorr[i]<-attr(VarCorr(m3)$item,"correlation")[1,2]
}
```

How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data



How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data

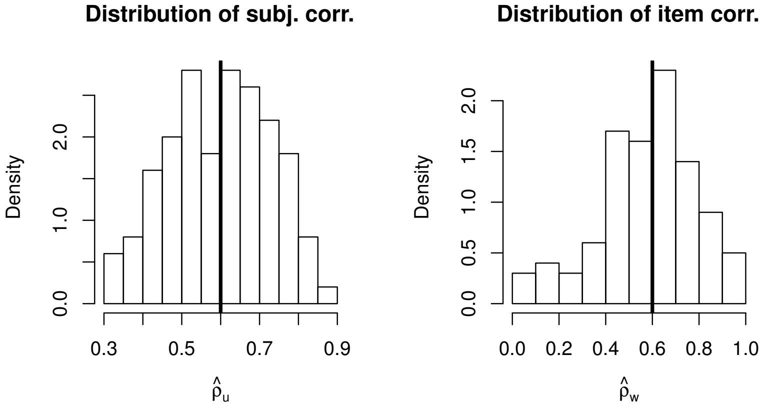
50 subjects and 30 items

```r
subjcorr<-rep(NA,nsim)
itemcorr<-rep(NA,nsim)

for(i in 1:nsim){
dat<-gendata(subjects=50,items=30)
m3<-lmer(rt~x+(1+x|subj)+(1+x|item),dat)
subjcorr[i]<-attr(VarCorr(m3)$subj,"correlation")[1,2]
itemcorr[i]<-attr(VarCorr(m3)$item,"correlation")[1,2]
}
```

How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data



How meaningful were the lmer estimates of correlations in the maximal model m1?

Simulated data

Conclusion:

1. It seems that lmer can estimate the correlation parameters just in case sample size for items and subjects is "large enough" (this can be established using simulation, as done above).
2. Barr et al's recommendation to fit a maximal model makes sense only if it's already clear that we have enough data to estimate all the variance components and parameters.
3. That is rarely the case in psychology and linguistics, especially when we go to more complex designs than a simple two-condition study.

Keep it maximal?

Gelman and Hill (2007, p. 549) make a more nuanced statement than “Keep it Maximal”:

> *Don’t get hung up on whether a coefficient “should” vary by group. Just allow it to vary in the model, and then, if the estimated scale of variation is small . . . , maybe you can ignore it if that would be more convenient.* **Practical concerns sometimes limit the feasible complexity of a model**–*for example, we might fit a varying-intercept model first, then allow slopes to vary, then add group-level predictors, and so forth. Generally, however,* **it is only the difficulties of fitting and, especially, understanding the models that keeps us from adding even more complexity, more varying coefficients, and more interactions***.*

Advantages of fitting a Bayesian LMM

- For such data, there can be situations where you *really* need to or want to fit full variance-covariance matrices for random effects. Bayesian LMMs will let you fit them even in cases where `lmer` would fail to converge or return nonsensical estimates (due to too little data).
  The way we will set them up, Bayesian LMMs will return a zero correlation with a wide credible interval, unless there is enough data pointing to a non-zero correlation.

Advantages of fitting a Bayesian LMM

- A direct answer to the research question can be obtained by examining the posterior distribution given data.
- We will abandon the traditional hard binary decision associated with frequentist methods: $p < 0.05$ *implies reject null, and* $p > 0.05$ *implies "accept" null*. **We are more interested in quantifying our uncertainty about the parameter estimate in question**.
- Prior knowledge can be included in the model.

Disadvantages of doing a Bayesian analysis

- You have to invest effort into specifying a model; unlike `lmer`, which involves a single line of code, JAGS and Stan model specifications can extend to 20-30 lines.
  A lot of decisions have to be made.
- There is a steep learning curve; you have to know a bit about probability distributions, MCMC methods, and of course Bayes' Theorem.
- It takes much more time to fit and assess a complicated model in a Bayesian setting than with `lmer`.

But I will try to demonstrate to you in this course that it's worth the effort, especially when you don't have a lot of data (usually the case in psycholinguistics).

Linear models

$$\underset{\substack{\uparrow\\ \textit{response}}}{y_i} = \overset{\substack{\textit{parameter}\\ \downarrow}}{\beta_0} \quad + \quad \overset{\substack{\textit{parameter}\\ \downarrow}}{\beta_1}\underset{\substack{\uparrow\\ \textit{predictor}}}{x_i} \quad + \quad \underset{\substack{\uparrow\\ \textit{error}}}{\varepsilon_i} \tag{1}$$

where

- ε_i is the residual error, assumed to be normally distributed: $\varepsilon_i \sim N(0, \sigma^2)$.
- Each response y_i (i ranging from 1 to I) is independently and identically distributed as $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.
- **Point values for parameters**: β_0 and β_1 are the parameters to be estimated. In the frequentist setting, **these are point values, they have no distribution**.
- **Null Hypothesis Significance Test (NHST)**: Usually, β_1 is the parameter of interest; in the frequentist setting, we test the null hypothesis that $\beta_1 = 0$.

Repeated measures data

- Linear mixed models are useful for correlated data (e.g., repeated measures) where the responses y are not independently distributed.
- A key difference from linear models is that the intercept and/or slope vary by subject $j = 1, \dots, J$ (and possibly also by item):

$$Y_{ijk} = \beta_j + b_{ij} + \varepsilon_{ijk} \tag{2}$$

$i = 1, \dots, 10$ is subject id, $j = 1, 2$ is the factor level, k is the number of replicates. $b_i \sim N(0, \sigma_b^2), \varepsilon_{ijk} \sim N(0, \sigma^2)$. $b_{ij} \sim N(0, \sigma_b^2)$. The variance σ_b^2 must be a 2×2 matrix:

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{3}$$

Unpacking the lme4 model

Crossed varying intercepts and slopes model, with correlation

$$\underset{\substack{\uparrow\\ \textit{response}}}{y_i} = [\beta_0 + \overset{\substack{\textit{varying intercepts}\\ \downarrow}}{u_{0j}} + w_{0k}] \; + \; [\beta_1 + \overset{\substack{\textit{varying slopes}\\ \downarrow}}{u_{1j}} + w_{1k}]\underset{\substack{\uparrow\\ \textit{predictor}}}{x_i} \; + \; \underset{\substack{\uparrow\\ \textit{error}}}{\varepsilon_i} \tag{4}$$

This is the “maximal” model we saw earlier:

```r
m1 <- lmer(rrt~x+(1+x|subj)+(1+x|item),
           headnoun)
```

Unpacking the lme4 model

Crossed varying intercepts and slopes model, with correlation

```
round(summary(m1)$coefficient,digits=3)

##             Estimate Std. Error t value
## (Intercept)   -2.672      0.138 -19.369
## x             -0.039      0.046  -0.835
```

Unpacking the lme4 model

Crossed varying intercepts and slopes model, with correlation

$$\texttt{rrt}_{ijk} = \underbrace{\beta_0 + u_{0j} + w_{0k}}_{\text{varying intercepts}} + \underbrace{\beta_1 + u_{1ij} + w_{1ik}}_{\text{varying slopes}} + \varepsilon_{ijk} \tag{5}$$

$$\Sigma_u = \begin{pmatrix} \sigma_{u0}^2 & \rho_u \sigma_{u0} \sigma_{u1} \\ \rho_u \sigma_{u0} \sigma_{u1} & \sigma_{u1}^2 \end{pmatrix} \quad \Sigma_w = \begin{pmatrix} \sigma_{w0}^2 & \rho_w \sigma_{w0} \sigma_{w1} \\ \rho_w \sigma_{w0} \sigma_{w1} & \sigma_{w1}^2 \end{pmatrix} \tag{6}$$

$$\begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_u \right), \quad \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_w \right) \tag{7}$$

The bivariate distribution

Independent (normal) random variables

If we have two random variables U0, U1, and we examine their joint distribution, we can plot a 3-d plot which shows, u0, u1, and f(u0,u1). E.g., $f(u0, u1) \sim (N(0, 1))$, with two independent random variables:

```
## Error in library(mvtnorm):  there is no package
called 'mvtnorm'
## Error in FUN(X, Y, ...):  could not find function
"dmvnorm"
## Error in persp.default(u0, u1, z, theta = -30, phi
= 30, ticktype = "detailed"):  object 'z' not found
```

The bivariate distribution

Correlated random variables ($\rho = 0.6$)

The random variables U, W could be correlated positively...

```
## Error in FUN(X, Y, ...):  could not find function
"dmvnorm"
## Error in persp.default(u0, u1, z, theta = -30, phi
= 30, ticktype = "detailed"):  object 'z' not found
```

The bivariate distribution

Correlated random variables ($\rho = -0.6$)

. . . or negatively:

```
## Error in FUN(X, Y, ...):  could not find function
"dmvnorm"
## Error in persp.default(u0, u1, z, theta = -30, phi
= 30, ticktype = "detailed"):  object 'z' not found
```

Bivariate distributions

This is why, when talking about two normal random variables U0 and U1, we have to talk about

1. U0's mean and variance
2. U1's mean and variance
3. the correlation ρ between them

A mathematically convenient way to talk about it is in terms of the variance-covariance matrix we saw for the Gibson and Wu data:

$$\Sigma_u = \begin{bmatrix} \sigma_{\mathrm{u}0}^2 & \rho_u\,\sigma_{u0}\sigma_{u1} \\ \rho_u\,\sigma_{u0}\sigma_{u1} & \sigma_{u1}^2 \end{bmatrix} = \begin{bmatrix} 0.61^2 & -.51 \times 0.61 \times 0.23 \\ -.51 \times 0.61 \times 0.23 & 0.23^2 \end{bmatrix} \tag{8}$$

The variance components associated with subjects

```
Random effects:
 Groups    Name         Variance Std.Dev. Corr
 subj      (Intercept)     0.37   0.61
           so              0.05   0.23     -0.51
```

$$\Sigma_u = \begin{bmatrix} \sigma_{u0}^2 & \rho_u \sigma_{u0}\sigma_{u1} \\ \rho_u \sigma_{u0}\sigma_{u1} & \sigma_{u1}^2 \end{bmatrix} = \begin{bmatrix} 0.61^2 & -.51 \times 0.61 \times 0.23 \\ -.51 \times 0.61 \times 0.23 & 0.23^2 \end{bmatrix} \tag{9}$$

The variance components associated with items

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 item     (Intercept)    0.11  0.33
          so             0.01  0.10     1.00
```

Note the by items intercept-slope correlation of $+1.00$.

$$\Sigma_w = \begin{bmatrix} \sigma_{\mathrm{w0}}^2 & \rho_w\, \sigma_{w0}\sigma_{w1} \\ \rho_w\, \sigma_{w0}\sigma_{w1} & \sigma_{w1}^2 \end{bmatrix} = \begin{bmatrix} 0.33^2 & 1 \times 0.33 \times 0.10 \\ 1 \times 0.33 \times 0.10 & 0.10^2 \end{bmatrix} \tag{10}$$

The total number of parameters

The parameters are $\beta_0, \beta_1, \Sigma_u, \Sigma_w, \sigma$. Each of the matrices Σ has three parameters. So we have 9 parameters.
Note that this model is overparameterized; there is simply not enough data to fit the correlation parameters.

Summary so far

- Linear mixed models allow us to take all relevant variance components into account; LMMs allow us to describe how the data were generated.
- However, maximal models should not be fit blindly, especially when there is not enough data to estimate parameters.
- For small datasets we often see degenerate variance covariance estimates (with correlation ± 1). Many researchers ignore this degeneracy but they should not.
- E.g., if the correlation is theoretically interesting, one should not ignore the degeneracy of the variance matrices.

Summary so far

- You will find a split in attitude between psychologists and linguists, who are taught to only look to see if the p-value is less than 0.05 or not, and statisticians, who want to express, using the most parsimonious model possible, how the data were generated.
- As I showed in week 1, the p-value is not only a completely useless measure (it answers a question we don't even care about), it is actually harmful (leads to incorrect inferences—think low power, Type S and M errors).
- One point I want to get across in this course is that our goal is to build the best model of the data that we can. "Hypothesis testing" will be based on our best estimates of the parameter of interest, given data and whatever prior knowledge we can bring to the table.

The frequentist approach

1. In the frequentist setting, we start with a dependent measure y, for which we assume a probability model.
2. In the above example, we have reading time data, rt, which we assume is generated from a normal distribution with some mean μ and variance σ^2; we write this $\mathrm{rt} \sim N(\mu, \sigma^2)$.
3. Given a particular set of parameter values μ and σ^2, we could state the probability distribution of rt given the parameters. We can write this as $p(\mathrm{rt} \mid \mu, \sigma^2)$.

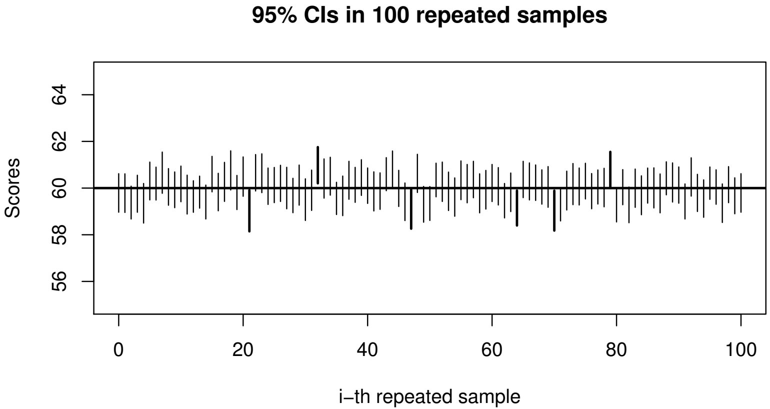
The frequentist approach

1. In reality, we know neither μ nor σ^2. The goal of fitting a model to such data is to estimate the two parameters, and then to draw inferences about what the true value of μ is.
2. The frequentist method relies on the fact that, under repeated sampling and with a large enough sample size, the sampling distribution the sample mean $\bar{X}$ is distributed as $N(\mu, \sigma^2/n)$.
3. The standard method is to use the sample mean $\bar{x}$ as an estimate of μ and given a large enough sample size n, we can compute an approximate 95% confidence interval $\bar{x} \pm 2 \times (\hat{\sigma}^2/n)$.

The frequentist approach

The 95% confidence interval has a slightly complicated interpretation:
If we were to repeatedly carry out the experiment and compute a confidence interval each time using the above procedure, 95% of those confidence intervals would contain the true parameter value μ (assuming, of course, that all our model assumptions are satisfied).
The **particular** confidence interval we calculated for our **particular** sample does not give us a range such that we are 95% certain that the true μ lies within it, although this is how most users of statistics seem to (mis)interpret the confidence interval.
See Richard Morey's work on confidence intervals for more.

The 95% CI


95% CIs in 100 repeated samples
Scores
56
58
60
62
64
0
20
40
60
80
100
i–th repeated sample

The Bayesian approach

1. The Bayesian approach starts with a probability model that expresses our prior knowledge about the possible values that the parameters μ and σ^2 might have.
2. This probability model expresses what we know so far about these two parameters (we may not know much, but in practical situations, it is not the case that we don't know *anything* about their possible values).
3. Given this prior distribution, the probability model $p(y \mid \mu, \sigma^2)$ and the data y allow us to compute the probability distribution of the parameters given the data, $p(\mu, \sigma^2 \mid y)$.
4. This probability distribution, called the **posterior distribution**, is what we use for inference.

The Bayesian approach

1. Unlike the 95% confidence interval, we can define a 95% **credible interval** that represents the range within which we are 95% certain that **the true value of the parameter lies**, given the prior and the data at hand.
2. Note that in the frequentist setting, the parameters are point values: μ is assumed to have a particular value in nature.
3. In the Bayesian setting, μ has a probability distribution; it has a mean, but there is also some uncertainty associated with its true value.

The Bayesian approach

Bayes' theorem makes it possible to derive the posterior distribution given the prior and the data. The conditional probability rule in probability theory (see Kerns) is that the joint distribution of two random variables $p(\theta, y)$ is equal to $p(\theta \mid y)p(y)$. It follows that:

$$\begin{aligned} p(\theta, y) =& p(\theta \mid y)p(y) \\ =& p(y, \theta) \quad \text{(because } p(\theta, y) = p(y, \theta)\text{)} \\ =& p(y \mid \theta)p(\theta). \end{aligned} \tag{11}$$

The first and third lines in the equalities above imply that

$$p(\theta \mid y)p(y) = p(y \mid \theta)p(\theta) \tag{12}$$

The Bayesian approach

Dividing both sides by $p(y)$ we get:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \tag{13}$$

The term $p(y \mid \theta)$ is the probability of the data given θ. If we treat this as a function of θ, we have the **likelihood function**.
Since $p(\theta \mid y)$ is the posterior distribution of θ given y, and $p(y \mid \theta)$ the likelihood, and $p(\theta)$ the prior, the following relationship is established:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{14}$$

The Bayesian approach

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{15}$$

We ignore the denominator $p(y)$ here because it only serves as a normalizing constant that renders the left-hand side (the posterior) a probability distribution (makes the area under the curve sum to 1).

The above is Bayes' theorem, and is the basis for determining the posterior distribution given a prior and the likelihood.

The rest of this course simply unpacks this idea.

Next week, we will look at some simple examples of the application of Bayes' theorem.