UseR! 2007 Conference
Iowa State University, Ames, Iowa

"Experiences using R to teach undergraduate statistics courses"

Philip Turk

Department of Mathematics & Statistics

NORTHERN ARIZONA UNIVERSITY

Flagstaff, Arizona

Background

- I've taught a variety of undergraduate statistics courses at three different medium-sized state universities
  - These have included elementary and intermediate statistics, probability, sampling, ANOVA, regression, time series, and statistical methods
- My goal has been to integrate the use of R as a key structural component of these courses
- In the next 20 minutes, I will briefly show you how I have tried to accomplish this goal

The Introduction to R

- I've written a handout that contains the following sections:
  - Obtaining and Installing R
  - Getting Help in R
  - Ending an R Session
  - Enhancing and Saving Graphics
  - Summary of R Commands
  - Creating a Data File
  - Elementary R Programs
  - Hints and Reminders on R
- I'd be happy to email you a copy of this:
  - Philip.Turk@nau.edu

Recommended Books

- “Introductory Statistics with R”, Peter Dalgaard
- “Using R for Introductory Statistics”, John Verzani
- Honorable Mention:
  - “Linear Models with R”, Julian Faraway
  - “An R and S-PLUS Companion to Applied Regression”, John Fox

Generic Example From My Lecture Notes

8.3.1 R Program - Measures of Central Tendency

The following data are from Zar (p. 22), 4th edition. They consist of a sample of 24 from a population of butterfly wing lengths.

Table 8.9 displays an R program that computes descriptive statistics that are measures of central tendency using the previous data set.

Table 8.9: R Program to Obtain Measures of Central Tendency

```
Length <- c(3.3, 3.5, 3.6, 3.6, 3.7, 3.8, 3.8, 3.8, 3.9, 3.9, 3.9, 4.0,
            4.0, 4.0, 4.0, 4.1, 4.1, 4.1, 4.2, 4.2, 4.3, 4.3, 4.4, 4.5)
which.max(table(Length)) # Finding the mode.
median(Length)
mean(Length)
mean(Length, trim = 1/24) # Trimmed mean.
```

Generic Example From My Lecture Notes

The output produced by the code now follows.

```
---------------------------------------------------------------------

4 # Output tells us the mode is 4, which is the seventh unique value
7 # of the sorted data.


4 # Median
3.958333 # Mean
3.963636 # Trimmed Mean
---------------------------------------------------------------------
```

Generic Example From Homework

3. Load the following data set using the R code below:

```
rainfall.data <- read.table('http://jan.ucc.nau.edu/~stapjt-p/STA570/data/rainfall.txt', header = TRUE)
```

The data are also available as a text file on the course web page.

The data represent values of March precipitation for Minneapolis-St. Paul over a period of 30 years. The first column of the text file contains the year index values and is labeled `Year`; the second column contains the corresponding precipitation amount and is labeled `Precipitation`.

(a) Using R, construct a normal probability plot.

(b) Using your plot, interpret the shape of the distribution of rainfall values. Also, comment on the appropriateness of a normal probability distribution model.

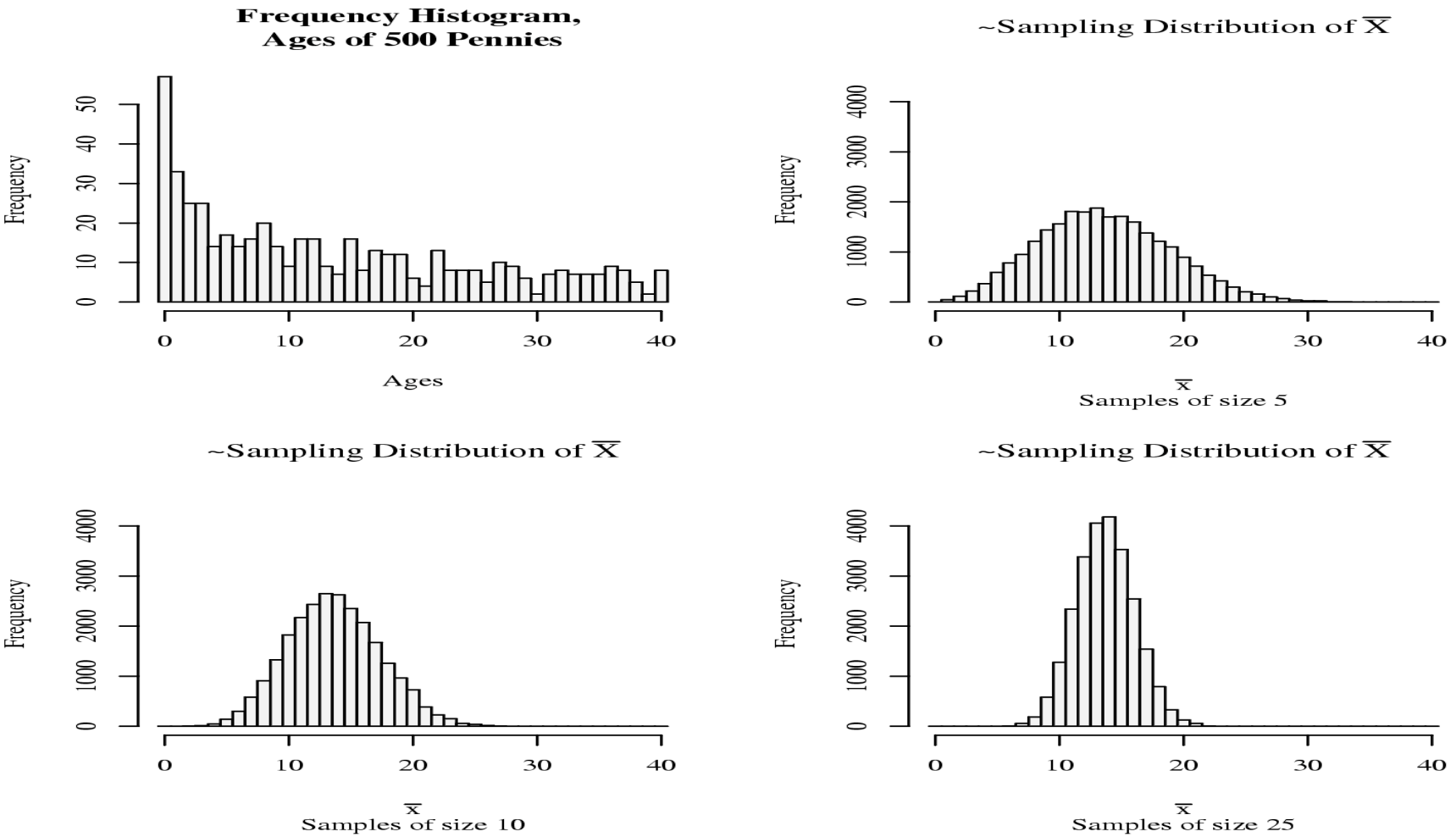
- Give them the R code in the solutions
- Take-home exams were set up in the same fashion

Course Web Page

- There are five important links:
  - “How to Install R” (pdf file)
  - Data Sets (text files)
  - R Code (text files)
  - Link to “Simple R” notes
    - http://www.math.csi.cuny.edu/Statistics/R/simpleR
  - Link to CRAN

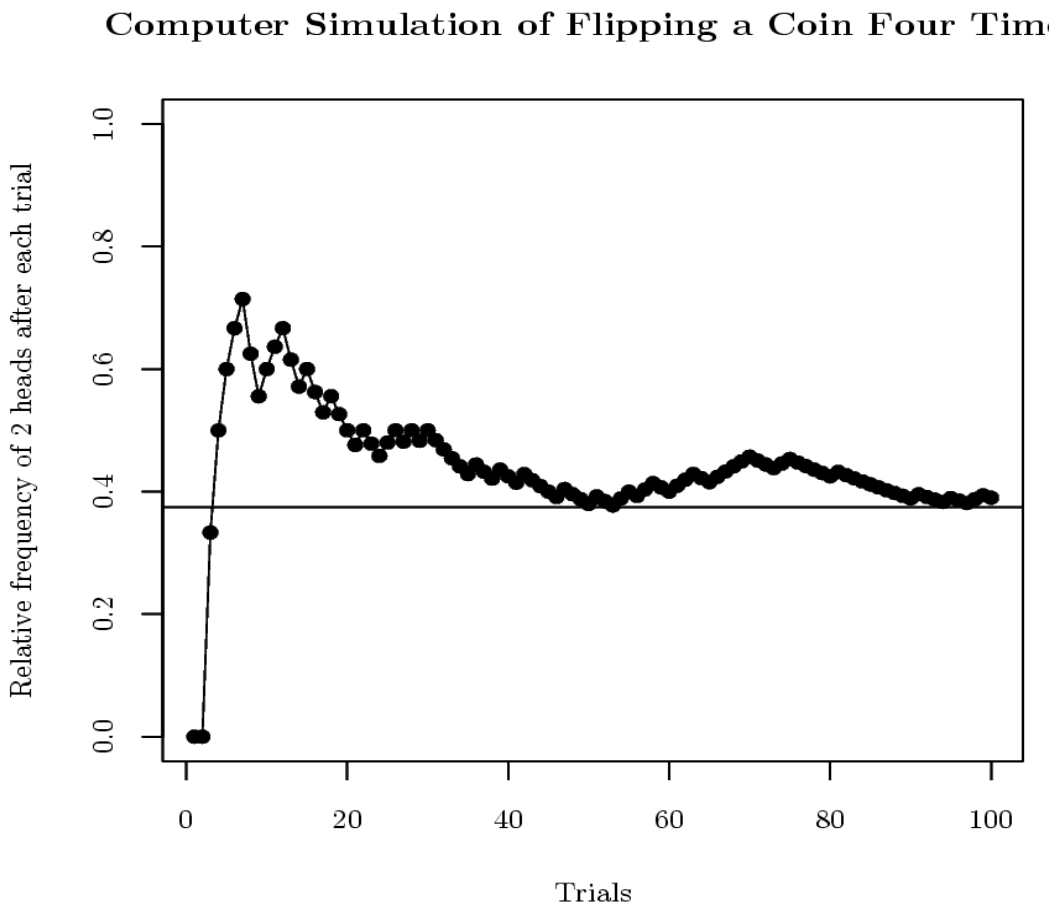
Example – Introductory Statistics

Figure 9.8: Sampling Distributions of $\overline{X}$: Example 2

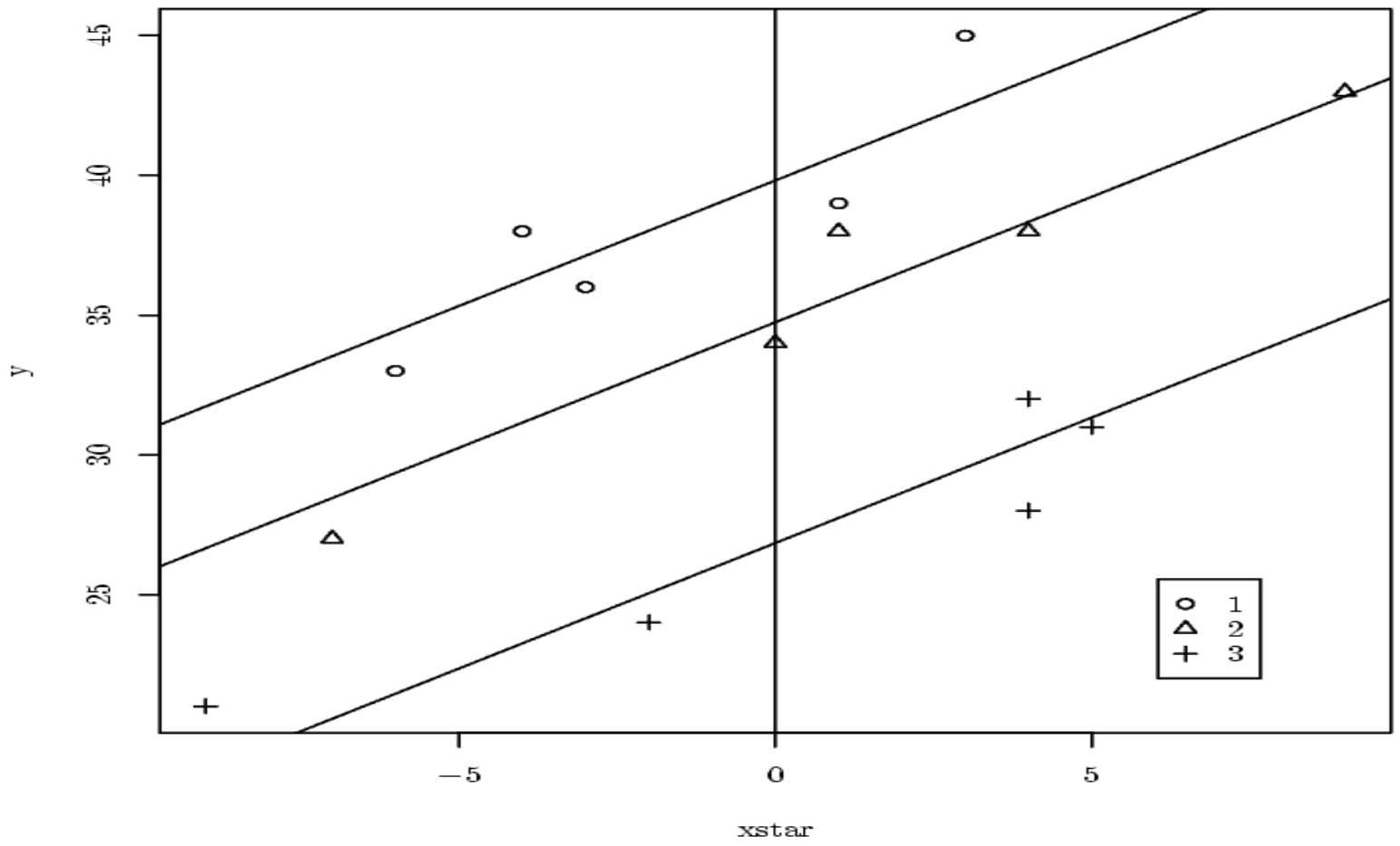


Example – Statistical Methods

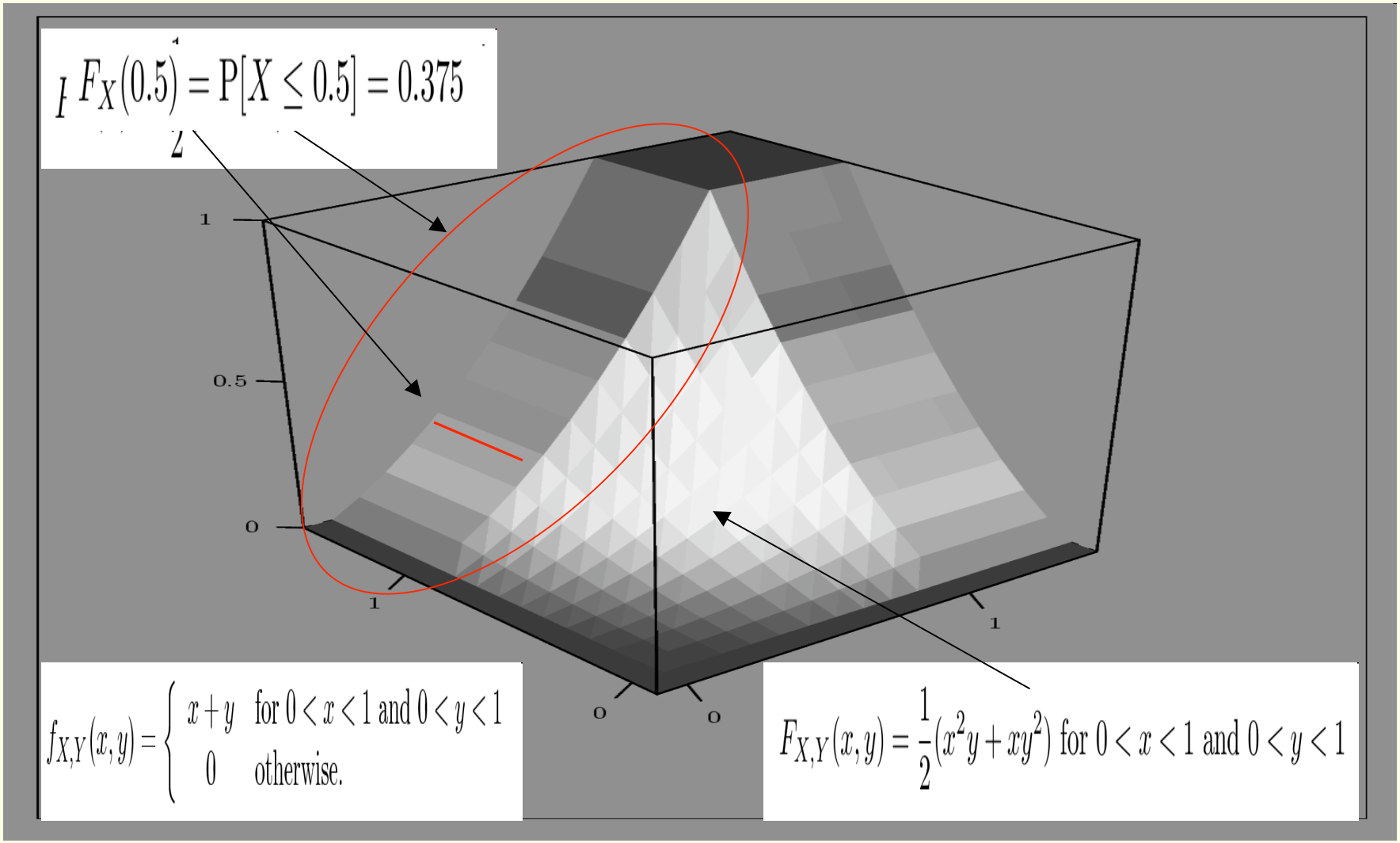
Figure 9.2: Computer Simulation of Flipping a Coin Four Times (100 Trials)



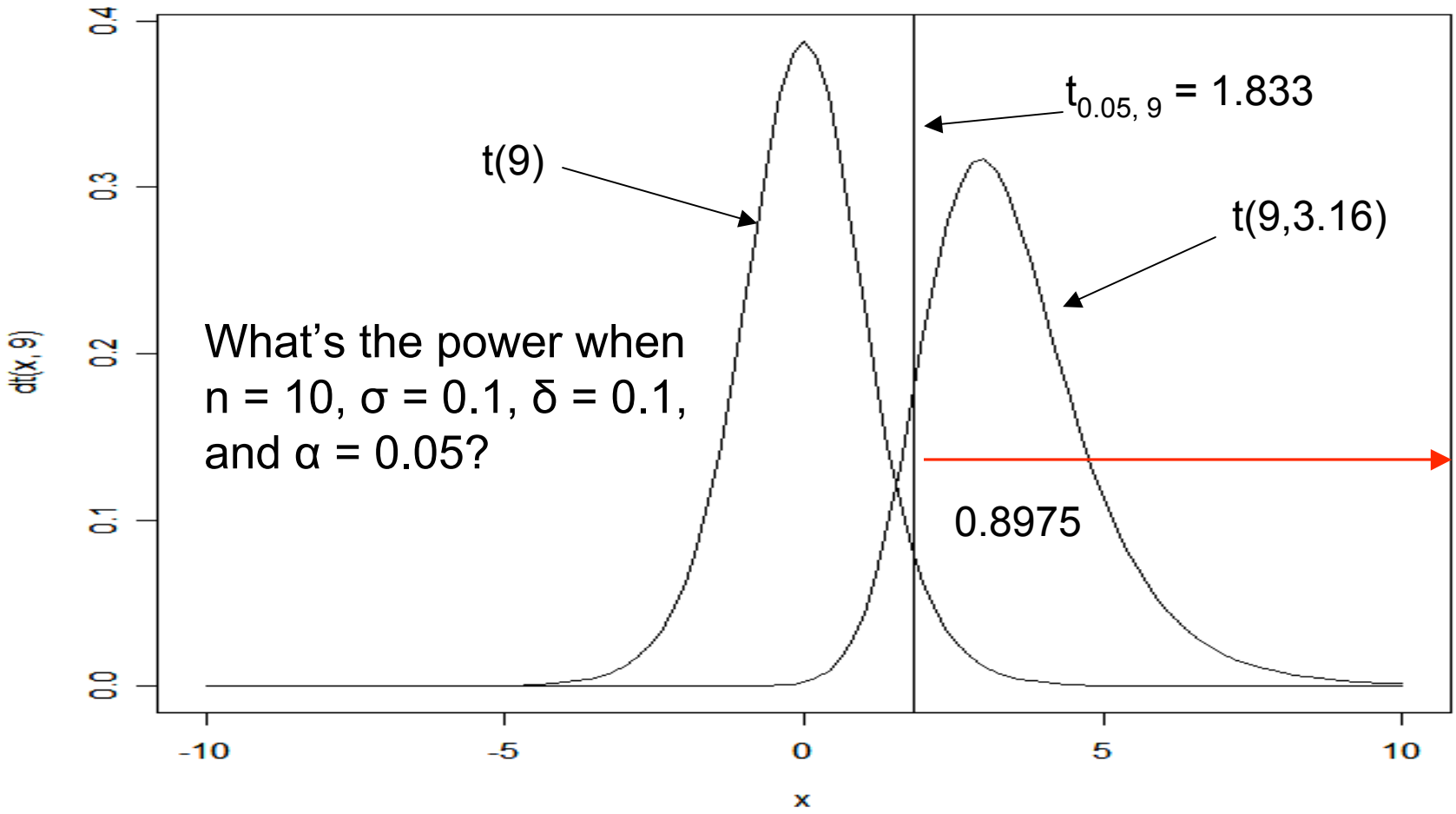
Example – ANOVA



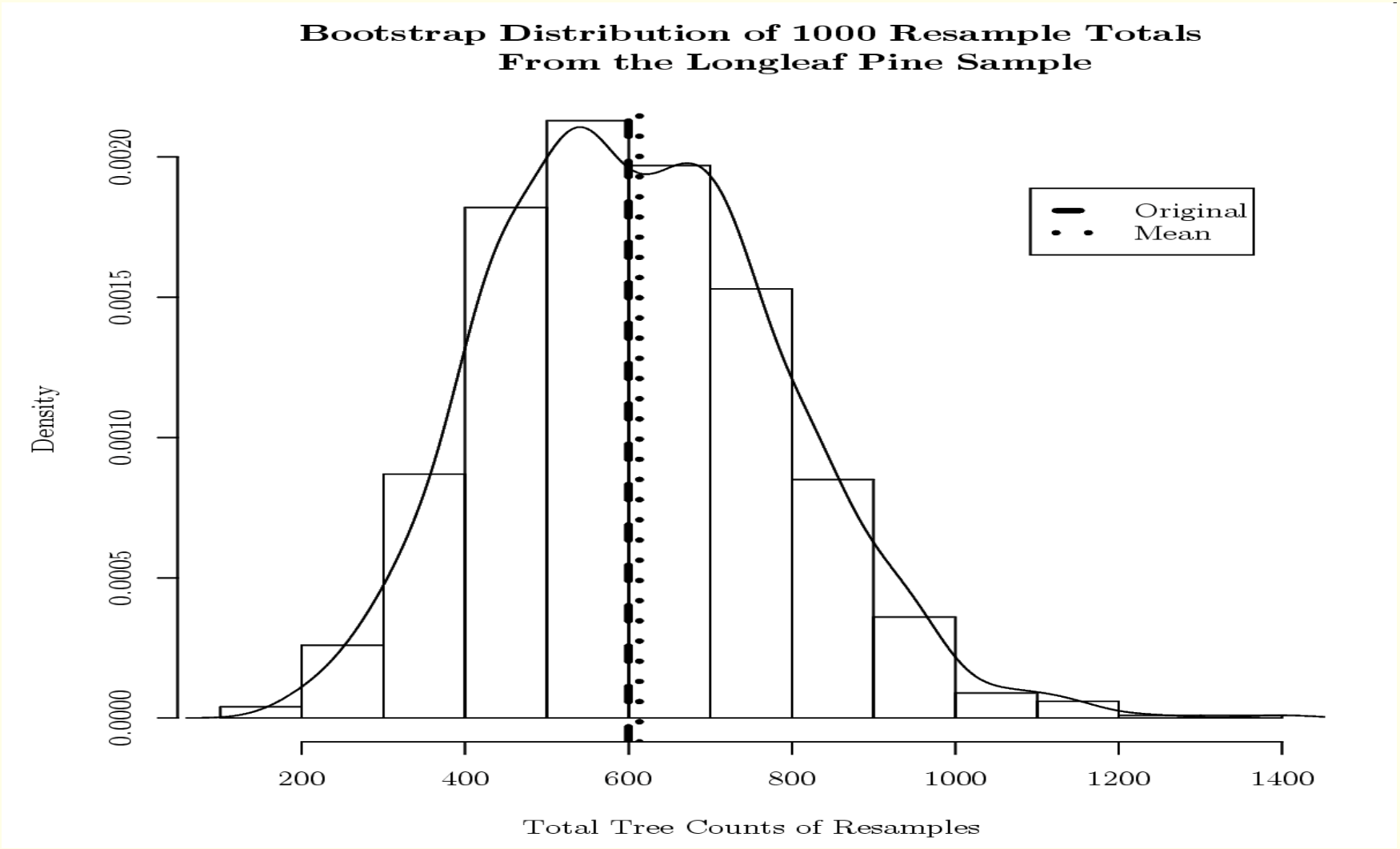
Example – Probability



Example – Intermediate Statistics



Example - Sampling



Example – Regression Analysis

The $\mathbf{X}$ matrix and the $\mathbf{Y}$ vector corresponding to a made up example are given below:

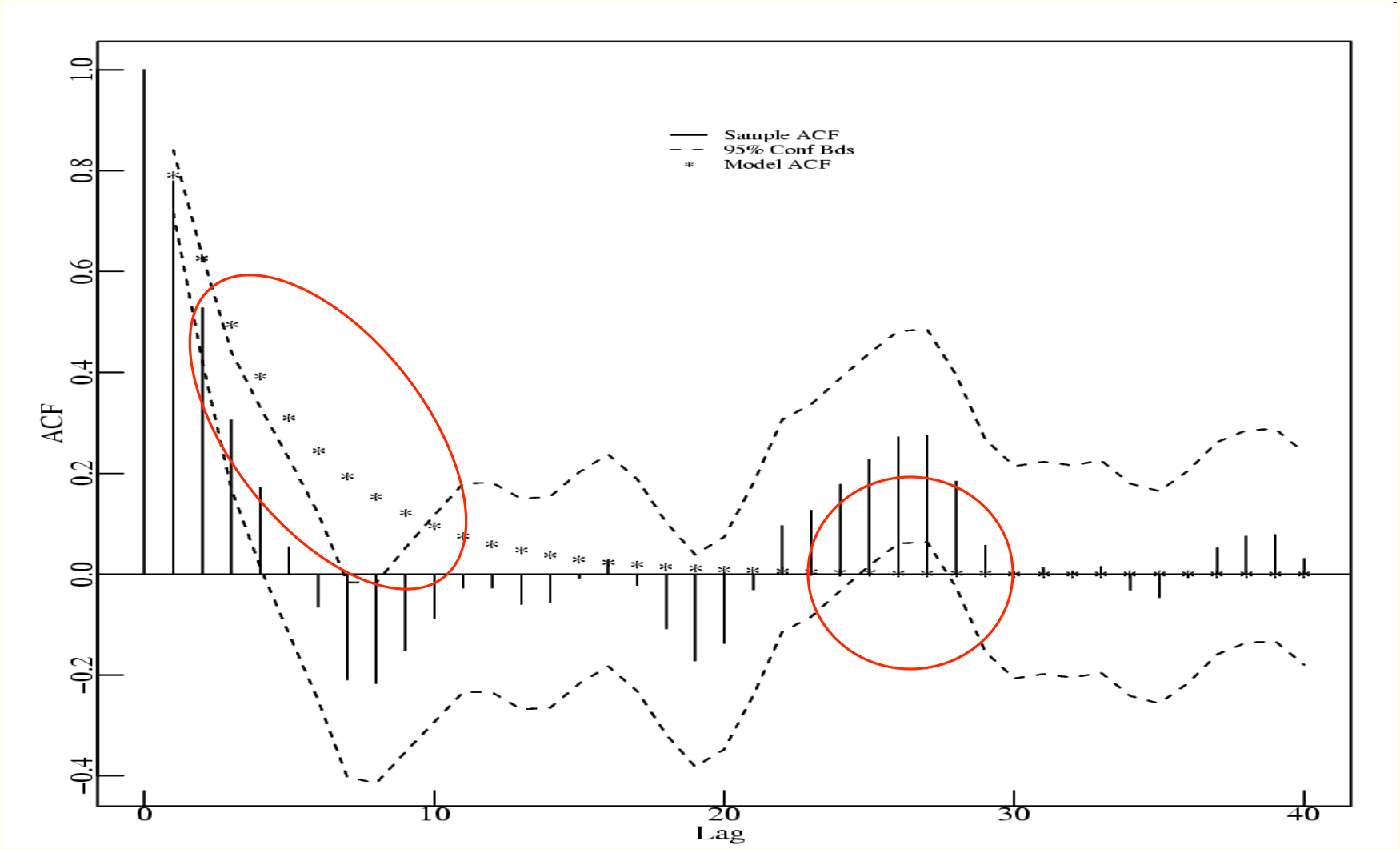
$$\mathbf{Y} = \begin{pmatrix} 3.1 \\ 2.3 \\ 3.0 \\ 1.9 \\ 2.5 \\ 3.7 \\ 3.4 \\ 2.6 \\ 2.8 \\ 1.6 \\ 2.0 \\ 2.9 \\ 2.3 \\ 3.2 \\ 1.8 \\ 1.4 \\ 2.0 \\ 3.8 \\ 2.2 \\ 1.5 \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & 5.5 \\ 1 & 4.8 \\ 1 & 4.7 \\ 1 & 3.9 \\ 1 & 4.5 \\ 1 & 6.2 \\ 1 & 6.0 \\ 1 & 5.2 \\ 1 & 4.7 \\ 1 & 4.3 \\ 1 & 4.9 \\ 1 & 5.4 \\ 1 & 5.0 \\ 1 & 6.3 \\ 1 & 4.6 \\ 1 & 4.3 \\ 1 & 5.0 \\ 1 & 5.9 \\ 1 & 4.1 \\ 1 & 4.7 \end{pmatrix}.$$

Example – Regression Analysis

```
> summary(lm(Y~X[,2]))

Call:
lm(formula = Y ~ X[, 2])

Residuals:
     Min       1Q   Median       3Q      Max
-0.74803 -0.37100  0.01404  0.34792  0.75197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6996     0.7268  -2.338   0.0311 *
X[, 2]        0.8399     0.1440   5.831 1.60e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> solve(t(X) %*% X) %*% t(X) %*% Y

             [,1]
[1,] -1.6995614
[2,]  0.8399123
```

Example – Time Series



Problems I've Encountered

- You have to be diligent about monitoring changes
  - E. g. simint in the multcomp package
  - Fix? Use TukeyHSD
- There is a relatively steep learning curve as opposed to software packages like JMP, Minitab, etc.
  - Fix? A few initial tutorial sessions

Problems I've Encountered

- You may run into instances where more time is spent trying to understand the code than is spent on the actual material!
  - E. g. they never really understood the syntax in nlme
- Be prepared to spend some time debugging code via email and the phone
  - Major dependent

Survey - “Did you like it?”

- “Yes, I did like R, it was hard to figure out at first, but it got easier as time and practice continued.”
- In general, when the dust had settled, most people felt they liked R
- Some liked it so much, they wrote their own functions, alerted me to new books, etc.

Survey - “What were the good points?”

- “I feel like learning to use R is a very valuable skill and in the end I did like it.”
- Some commented on the joy of learning
- Students liked the graphics and felt it facilitated learning course concepts
- Not doing calculations by hand, e.g. ANOVA, etc.

Survey - “What were the bad points?”

- “At times I felt as though we were just thrown into R – an optional R tutorial in a computer lab outside of class would have benefited me.”
- A steep learning curve
  - Giving them code may not be enough
  - Major dependent
- The additional packages

Survey - “Will you use R again?”

- “I do think I will use R again. I would like to get more familiar with the program and I think it’s wonderful to have some basic knowledge of a free statistical package.”
- Specifically, some students felt they would use it in their careers, their theses, etc.
- Any questions?