

Using R in Undergraduate Probability and Mathematical Statistics Courses*

Amy G. Froelich
Iowa State University
September 24, 2007

*Part of this seminar was presented at the 2007 UseR! Conference, Ames, IA and was supported by a grant received by Michael D. Larsen and Amy G. Froelich from the ISU College of Liberal Arts and Sciences Computer Advisory Committee.

Statistics Education

- Materials Development
- Use of Technology
- Research into Effectiveness of Materials in Student Learning
- Assessments
- Training TAs to Teach Statistics
- Training Math Teachers to Teach Statistics
- Statistics for Mathematics and Statistics Majors

Materials Development

- Development of course materials for introductory statistics.
 - ISU Miller Faculty Fellowship.
 - NSF Grant (2003-2005).
 - Froelich & Stephenson – *How much does an M&M weigh?*
 - Invited Breakout Session at USCOTS, 2007.

Use of Technology

- How can technology help students learn and understand statistical concepts?
 - Development of JMP scripts to teach concepts in introductory statistics.
 - Invited to work with team from JMP/SAS.
 - Designing and testing JMP scripts over next year.
 - Student Edition of JMP (18-24 months).

Research into Effectiveness

- Do developed materials and technology make a difference in student conceptual learning and understanding?
 - Froelich, Stephenson & Duckworth – *Assessment of Materials for Engaging Students in Statistical Discovery*.
 - Testing JMP Scripts for sampling distributions and inference next semester.

Assessments

- How do you measure student conceptual learning and understanding of course materials and students' attitudes in introductory statistics?
 - Survey of Attitudes Toward Statistics (SATS) – Schau
 - Comprehensive Assessment of Outcomes in a first Statistics Course (CAOS) – Garfield, delMas & Chance

Training TAs

- How can we train statistics graduate students to teach introductory statistics?
 - Two departmental teaching documents
 - Froelich, Duckworth & Stephenson – *Training Statistics Teachers at Iowa State University*
 - Invited session at JSM 2006 & JSM 2008
 - Future: New Orientation Seminar Course for Graduate Students

Training Math Teachers

- What statistical content and pedagogy training in statistics should HS and CC math teachers receive?
 - Faculty member – Master of School Mathematics Program
 - Requirement of 6 graduate credit hours in Statistics to teach CC courses. (Iowa DE)
 - Stat 410X – Statistical Methods for Mathematics Teachers
 - Pending Regents proposal for Stat 410X

Statistics for Math and Stat Majors

- What should the introductory and follow-up courses for statistics and mathematics majors look like?
 - Introductory course – Stat 101L
 - Applied Courses – Stat 401/402
 - Theory Courses – Stat 341/342

Undergraduate Probability and Mathematical Statistics at ISU

- STAT 341
 - Probability
 - Discrete Random Variables
 - Continuous Random Variables
 - Multivariate Distributions
- STAT 342
 - Transformations of Random Variables
 - Sampling Distributions and Inference
 - Theory of Estimation and Hypothesis Tests
 - Linear Models
 - Categorical Data Analysis
- Textbook: Mathematical Statistics with Applications by Wackerly, Mendenhall & Scheaffer

Why Use R in STAT 341 & 342?

- Curriculum: Exposure to R
- Content: Model nature of statistics
- Pedagogy: Ground theoretical concepts in understanding obtained from earlier applied statistics courses.
 - I. Connect
 - II. Explore
 - III. Visualize
 - IV. Expand

*The purpose of computing is
insight, not numbers.*

(Tukey and/or Haming)

I. Connect

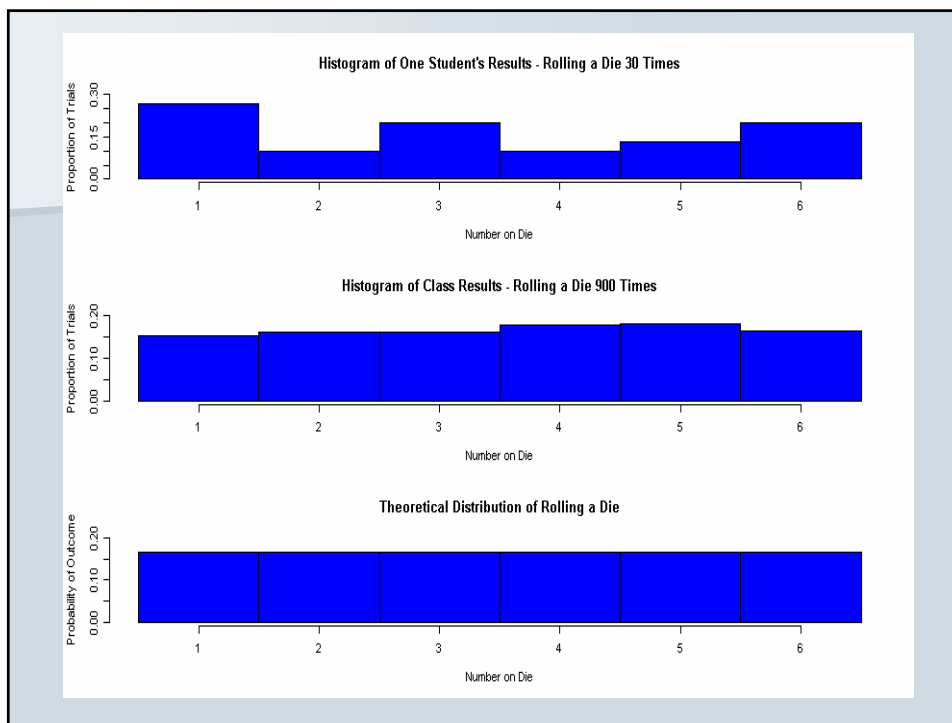
- Observed vs. Theoretical Probabilities*
- Observed vs. Theoretical Distributions
- Observed vs. Theoretical Moments
- Univariate Normal Distribution to Multivariate Normal Distribution
 - Univariate to linear regression
 - Multivariate to linear regression
 - Models versus simulation versus data

I. Observed and Theoretical Probabilities

- Use R to explore basics of data analysis.
- Use R to explore probabilities.
- Develop methods for determining theoretical probabilities of events.
- Return later in the course.
 - Common Discrete and Continuous Distributions
 - Transformations of Random Variables
 - Order Statistics

I. Example 1 - Dice

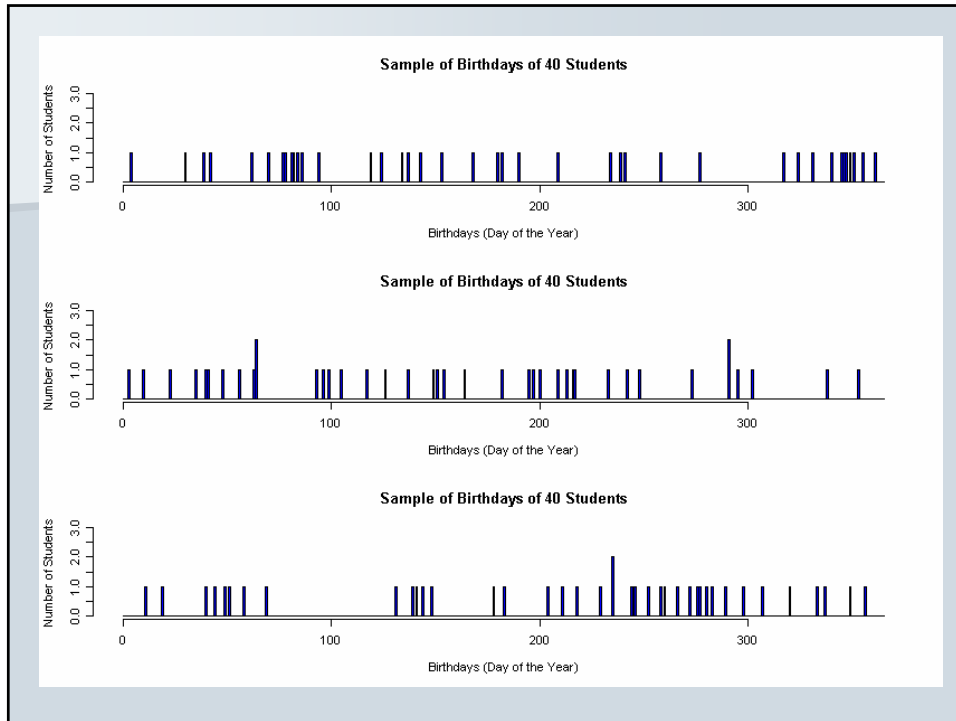
- Opening Activity: Roll a die 30 times. Keep track of the number of 1s, 2s, 3s, etc.
- Use R to plot results from one group.
- Use R to plot results from entire class.
- Compare observed distributions of number on die to theoretical probability distribution.



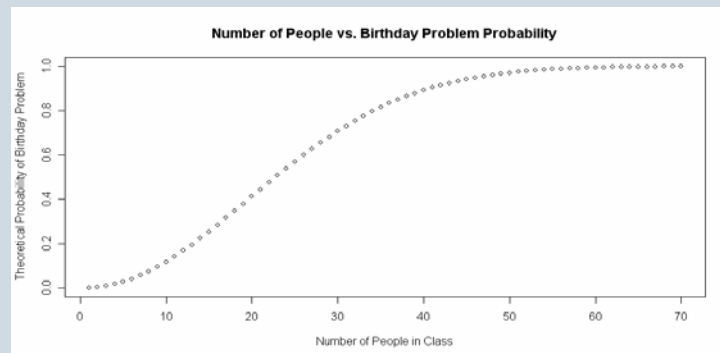
- Difference between results
 - from one person and from class?
 - from one person and from dice?
 - from class and from dice?
- Return later in the course
 - Observed vs. Simulated vs. Theoretical Distributions
 - Goodness of fit

I. Example 2 – Birthday Problem

- Guess probability for class of 40 students.
 - $1/365$, $40/365$, 10%, 50%, 85%
- Simulate birthdays for a class of 40 students.
 - `daysofyear<- c(1:365)`
 - `birthdays<- sample(daysofyear, 40, replace = T)`
- Look for repeated birthdays.
 - `daysbreaks<- c(0:365) + 0.5`
 - `hist(birthdays, breaks = daysbreaks)`



- Out of 10000 samples of 40 students each, 8897 had at least one repeated birthday.
- Theoretical Calculation = 0.8912318

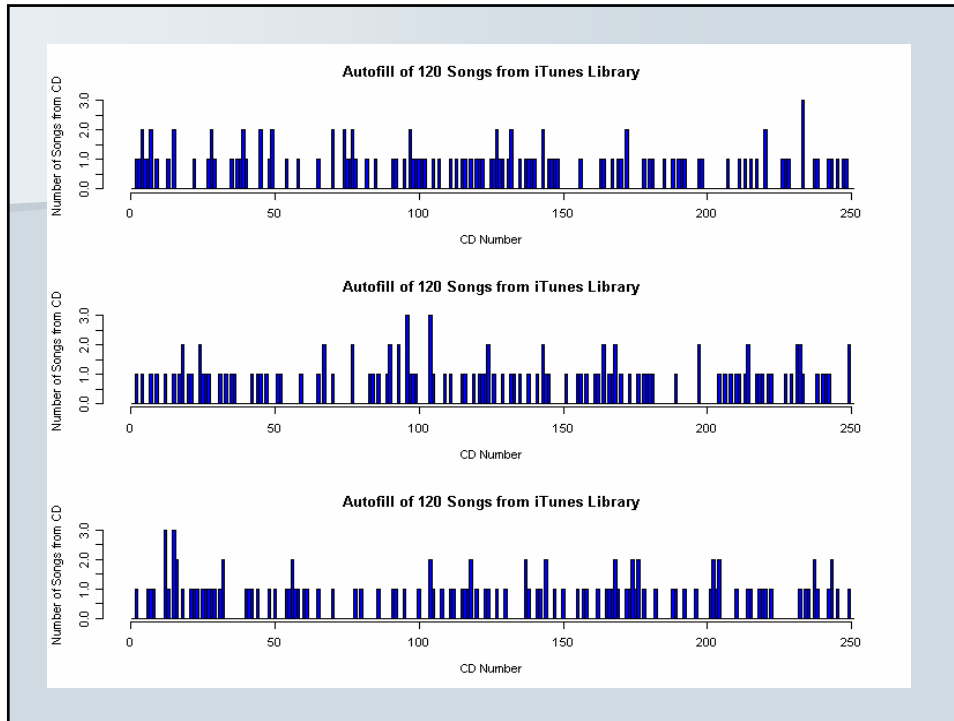


- Role of Assumptions

I. Example 3 – iPod Shuffle

- Is the iPod shuffle feature random?
- Autofill feature in iTunes fills smallest iPod with approx. 120 songs.
- "The first few times. . ., I found some disturbing clusters in the songs chosen. More than once the 'random' playlist included three tracks from the same album! Since there are more than 3000 tunes in my library, this seemed to defy the odds." Steven Levy, Newsweek Magazine, January 31, 2005.

- Probability of 3 or more songs from ANY one album in autofill.
- 250 albums, 12 songs per album, 3000 songs.
 - `library<- rep(1:250, 12)`
- Simulate an autofill = selection of 120 songs from library.
 - `autofill<- sample(library, 120, replace = F)`
- Look for number of songs per album in autofill.
 - `albumbreaks<- c(0:250) + 0.5`
 - `hist(autofill, breaks = albumbreaks)`



- Out of 10000 autofills, 9453 had 3 or more songs from any one album.
- Theoretical Calculation

$$1 - \frac{\sum_{x=0}^{60} \binom{250}{x} \binom{250-x}{120-2x} \binom{12}{2}^x \binom{12}{1}^{120-2x}}{\binom{3000}{120}} = 0.9445$$

- Role of Assumptions
- Maximum Observation from Multivariate Hypergeometric Distribution

II. Explore

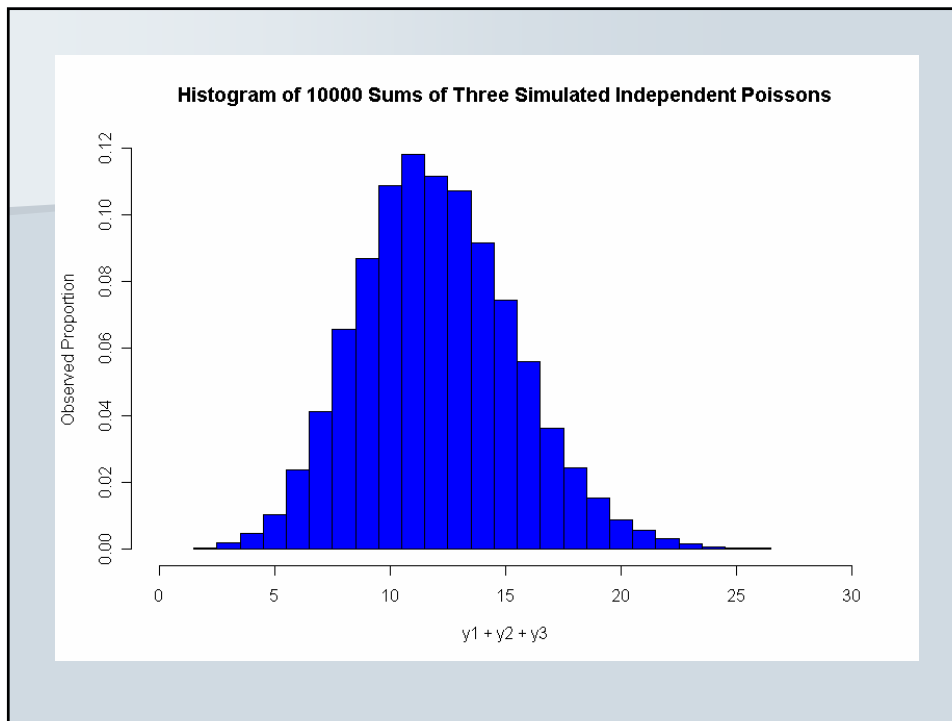
- Law of Large Numbers
- Transformations of Random Variables*
- Central Limit Theorem
- Sampling Distributions
- Confidence Intervals
- Hypothesis testing - Type I and Type II error rates and test procedures

II. Transformations of Random Variables

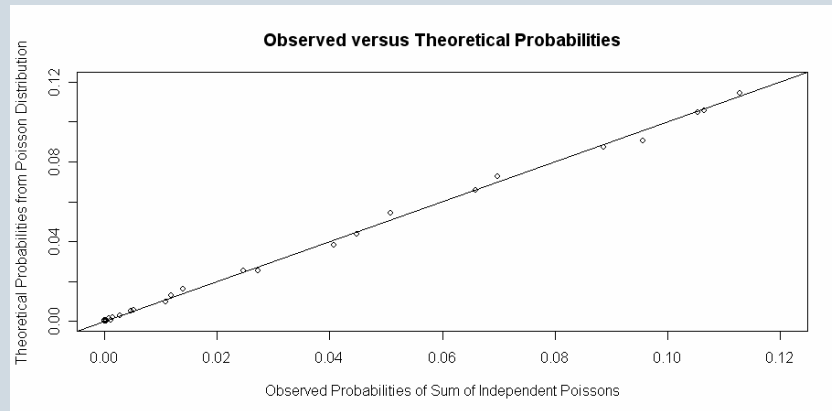
- Use R to explore distributions of independent sums
- Develop methods for proving ideas developed using R.
- Return to ideas later in the course.
 - Central Limit Theorem
 - Sampling Distributions
 - Inference
 - Goodness of Fit

II. Example 1 - Poisson

- Y_1, Y_2 and Y_3 are independent Poisson r.v.s with means 3, 4, and 5 respectively.
- What is the distribution of $Y_1 + Y_2 + Y_3$?
- Simulate 10000 obs. of the sum in R
 - Obs. Mean = 12.0842
 - Obs. Var = 12.13852

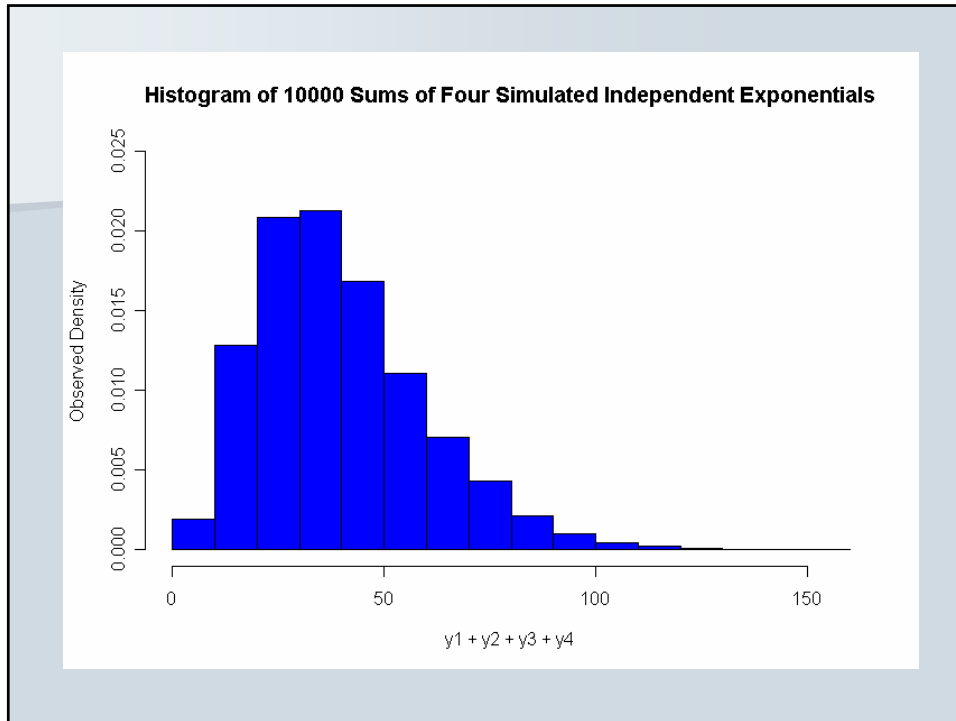


- Is the distribution of sum Poisson?
 - Mean and Variance roughly equal?
 - Do observed probabilities match Poisson probabilities?



II. Example 2 – Exponential

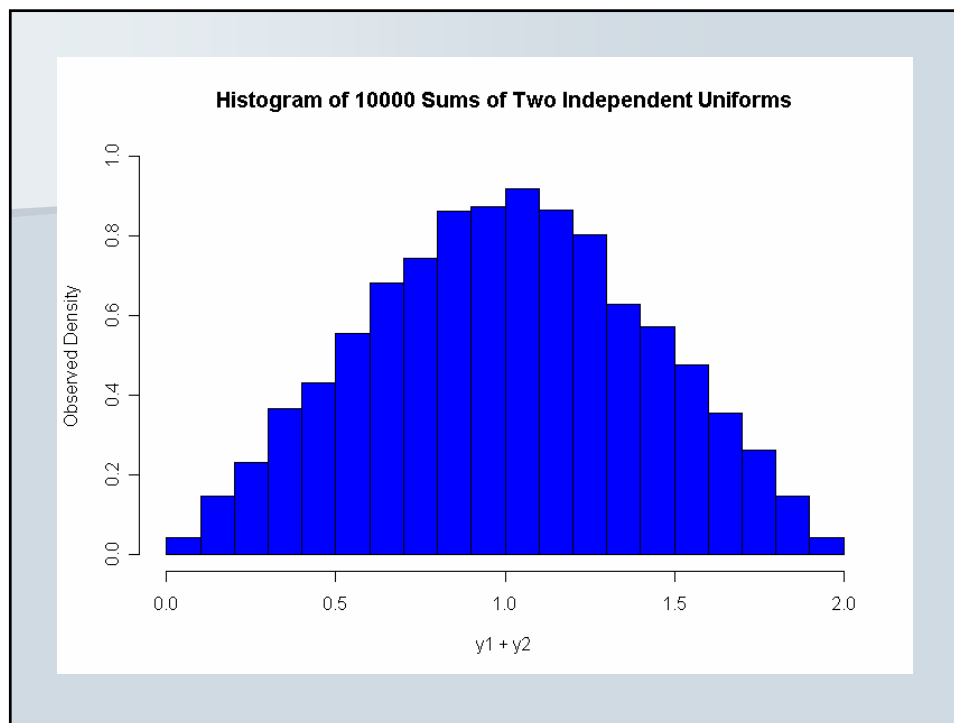
- Y_1, Y_2, Y_3 and Y_4 are independent exponential r.v.s with mean 10.
- What is the distribution of $Y_1 + Y_2 + Y_3 + Y_4$?
- Simulate 10000 obs. of sum in R
 - Obs. Mean = 40.08191
 - Obs. Var = 393.3244



- Is the distribution of sum Exponential?
 - Right-skewed distribution
 - Mean approx. 40
 - Variance approx. 400
- Is the distribution something related?
 - Right-skewed distribution
 - Mean approx. $40 = 4 \cdot 10$
 - Variance approx. $400 = 4 \cdot 10^2$

II. Example 3 – Uniform

- Y_1 and Y_2 are independent Uniform (0,1) r.v.s.
- What is the distribution of $Y_1 + Y_2$?
- Simulate 10000 obs. of sum in R
 - Obs. Mean = 1.004535
 - Obs. Var = 0.1671598



- Is the distribution of sum Uniform?
 - Flat Histogram?
 - Min = 0, Max = 2
 - Mean approx. 1
 - Variance approx. 0.17
- Is the distribution something we have seen?
 - Triangle-shaped histogram

III. Visualize

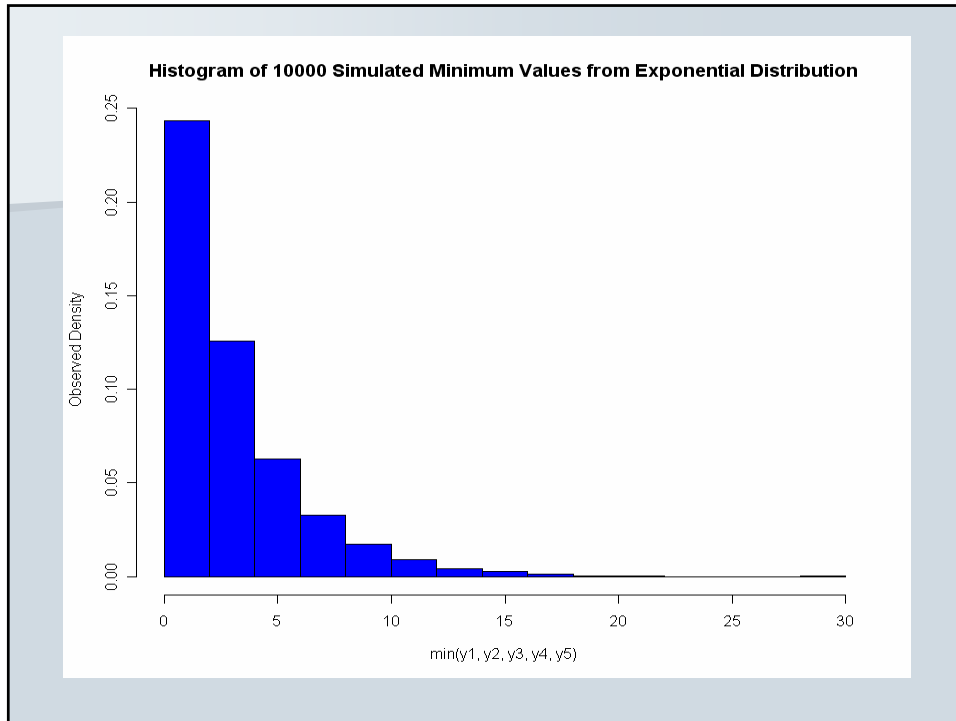
- Distribution Functions of Discrete and Continuous Random Variables
- Order Statistics*
- Likelihood Functions
- Asymptotic Normality of Maximum Likelihood Estimators

III. Order Statistics

- Use R to explore the meaning of an order statistic.
- Use R to explore the distribution of common order statistics.
 - Minimum, Maximum, Median
- Develop theoretical distributions of order statistics.
- Return to ideas later in the course.
 - Properties of Estimators.
 - MVUEs and MLEs.

III. Example 1 – Minimum of Exponentials

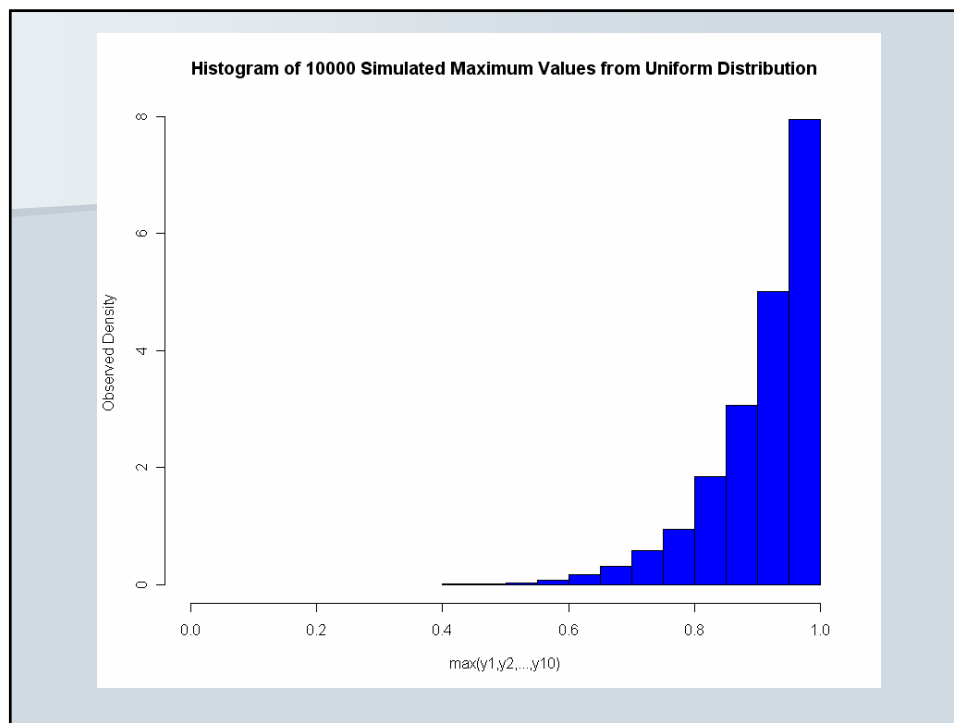
- $X = \text{Min}(Y_1, Y_2, \dots, Y_5)$ where Y_i are independent exponential r.v.s with mean 15.
- Distribution of X ?
- Simulate 10000 obs of minimum in R.
for (i in 1:10000) x[i]<- min(rexp(5, 1/15))
 - Obs. Mean = 3.003846
 - Obs. Var = 8.923922



- Is the distribution Exponential?
 - Right-skewed distribution
 - Mean approx. 3
 - Variance approx. 9
- Develop theoretical distribution of Minimum

III. Example 2 – Maximum of Uniforms

- $X = \max(Y_1, Y_2, \dots, Y_{10})$ where Y_i are independent Uniform (0, 1) r.v.s
- Distribution of X?
- Simulate 10000 obs. of maximum in R
for (i in 1:10000) x[i]<- max(runif(10, 0, 1))
 - Obs. Mean = 0.9085455
 - Obs. Var = 0.00696515



- Is the distribution Uniform?
 - Left-Skewed Distribution
 - Mean approx. 0.91
 - Variance approx. 0.007
- Develop theoretical distribution of Maximum

IV. Expand

- Additional Probability Distributions
- Non-Central Distributions and Power for Hypothesis Testing
- Randomization Tests
- Role of Assumptions in Statistical Testing*

IV. Role of Assumptions in Statistical Testing

- Use R to explain underlying concepts
 - Coverage rates for Confidence Intervals
 - Type I error rates
 - Power rates
- Use R to study violations of assumptions
 - CI for p
 - Goodness of fit tests

IV. Example 1 – Assumptions for CI for p

- Y is binomial r.v. with parameters n and p.
- Approx. 95% CI for p
- Why Approx.?
- Approx. binomial distribution with normal distribution.
 - $np \geq 10$ and $n(1-p) \geq 10$
- What happens when assumption does not hold?
 - $n = 10$; $p = 0.5$ or $p = 0.1$

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

IV. Example 2 – Traditional vs. Plus 4 method

- Assumption holds (Two cases)
 - $p = 0.1, n = 100$
 - $p = 0.5, n = 1000$
- Coverage rate
 - Traditional 95% CI for p
 - Plus 4 method 95% CI for p

Student Reactions to R

- Course evaluations – Fall 2005.
- Did you like using R? Was it helpful to you?
Did you find it easy or difficult to use?
 - 34 responses – 32 positive; 1 indifferent; 1 negative comments to using R.

Selected Student Comments

- "I liked using R. I found it easy with the examples given in class and the handouts clearly explaining how to use it."
- "I liked using R, it was helpful in doing the difficult distribution problems. I also liked it because we could construct histograms to visual(ize) the distributions."
- "I liked using R, easy to use, somewhat difficult to know what to type in. Continue using it in 341 since it's used in the real world."
- "I thought R was very helpful and easy to use... In most cases, the professor's instructions made the program very understandable."
- "It was nice to use R because it seems really versatile and I like the easy to understand interface."
- "I think R was very helpful, because we didn't have to calculate everything by hand. It was easy to use too, because we got a good explanation about it."

Recent Work in this Area

- 2006 LASCAC Grant
- Froelich, Duckworth & Culhane – *Does your iPod really play favorites?*
- Invited Session at 2007 UseR! Conference
- ISU Statistics Education Materials Repository
<http://stated.stat.iastate.edu/>

Plans for Future Work in this Area

- Statistics Curriculum for Math Majors
 - Stat 101L
 - Stat 341
 - Stat 342 (Prereq: Stat 101L)

Plans for Future Work in this Area

- Further development of materials
 - Froelich & Larsen – *Using R to Teach Undergraduate Probability and Mathematical Statistics.*
 - ISU Statistics Education Materials Repository

Plans for Future Work in this Area

- Research: Do these materials promote student learning and understanding?
 - delMas, Garfield & Chance, 2002
 - Lunsford, Rowell, Goodson-Espy, 2006

Plans for Future Work in this Area

- Invited or Topic-Contributed Session at 2008 JSM

Plans for Future Work in this Area

- Textbook – Modern Probability and Statistics Using R