Using R in Undergraduate and Graduate Probability and Mathematical Statistics Courses*

Amy G. Froelich

Michael D. Larsen

Iowa State University

*The work presented in this talk was partially supported by a grant from the ISU College of Liberal Arts and Sciences Computer Advisory Committee

Stat 341 at ISU

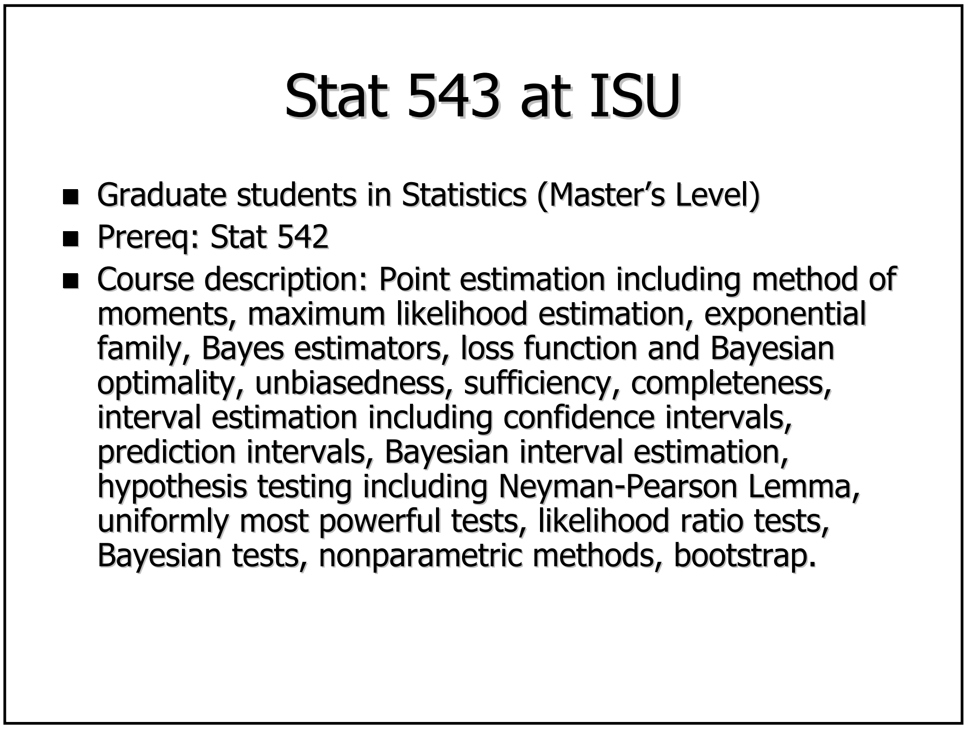
- Statistics, Math, and Math Ed Majors
- Prereq: Calculus I, II and III
- Course Description: Probability; distribution functions and their properties; classical discrete and continuous distribution functions; moment generating functions, multivariate probability distributions and their properties; transformations of random variables.

Stat 342 at ISU

- Statistics majors (some math majors)
- Prereq: Stat 341, Linear Algebra
- Course Description: Sampling distributions; confidence intervals and hypothesis testing; theory of estimation and hypothesis tests; linear model theory, enumerative data.

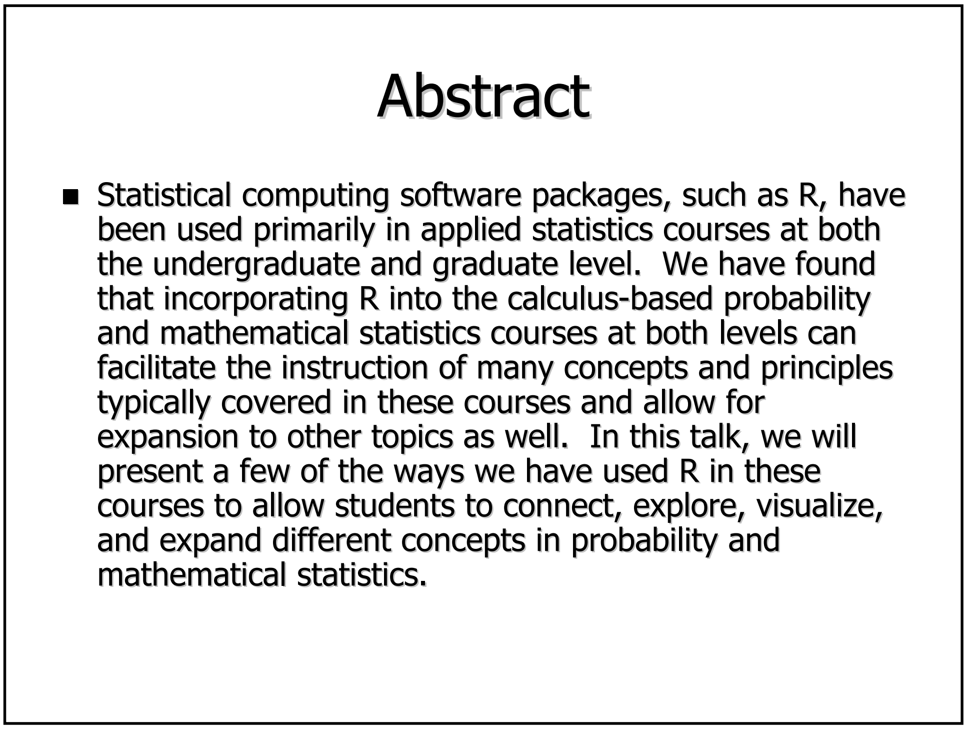
Stat 542 at ISU

- Graduate students in Statistics (Master's Level)
- Prereq: Stat 341, Real Analysis or Advanced Calculus
- Course description: Sample spaces, probability, conditional probability, random variables, univariate distributions, expectation, moment generating functions; common theoretical distributions; joint distributions, conditional distributions and independence, covariance; probability laws and transformations, introduction to the multivariate normal distribution, sampling distributions, order statistics, convergence concepts, the central limit theorem and delta method, basics of stochastic simulation.



Stat 543 at ISU

- Graduate students in Statistics (Master’s Level)
- Prereq: Stat 542
- Course description: Point estimation including method of moments, maximum likelihood estimation, exponential family, Bayes estimators, loss function and Bayesian optimality, unbiasedness, sufficiency, completeness, interval estimation including confidence intervals, prediction intervals, Bayesian interval estimation, hypothesis testing including Neyman-Pearson Lemma, uniformly most powerful tests, likelihood ratio tests, Bayesian tests, nonparametric methods, bootstrap.



Abstract

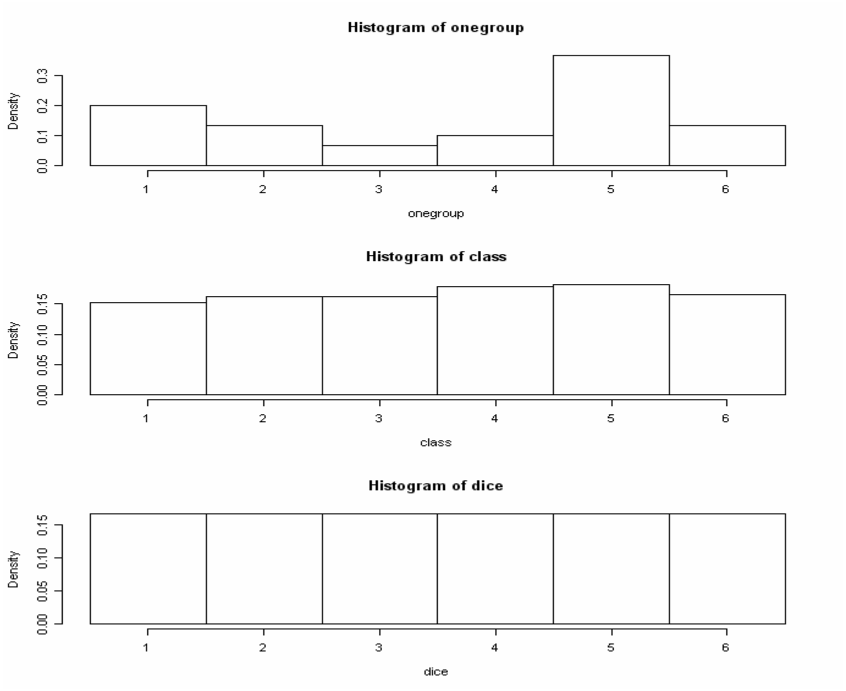
- Statistical computing software packages, such as R, have been used primarily in applied statistics courses at both the undergraduate and graduate level. We have found that incorporating R into the calculus-based probability and mathematical statistics courses at both levels can facilitate the instruction of many concepts and principles typically covered in these courses and allow for expansion to other topics as well. In this talk, we will present a few of the ways we have used R in these courses to allow students to connect, explore, visualize, and expand different concepts in probability and mathematical statistics.

Connect

- *Observed and Theoretical Probabilities and Distributions
  - – Simulation versus theory
  - – Theory versus data
  - – Simulation versus data
- Observed Moments to Theoretical Moments
- Univariate Normal Distribution to Multivariate Normal Distribution
  - – Univariate to linear regression
  - – Multivariate to linear regression
  - – Models versus simulation versus data

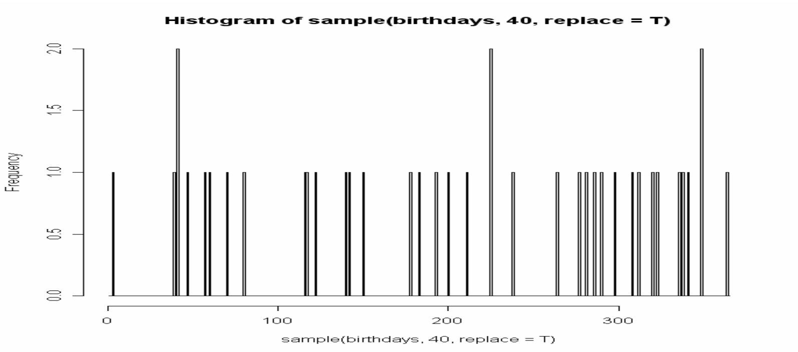
Observed and Theoretical Probabilities and Distributions

- 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 distribution.



Observed and Theoretical Probabilities and Distributions

- Birthday Problem – Guess probability for class of 40 students.
- Simulate: birthdays<- c(1:365)
  sample(birthdays, 40, replace = T)
- Make histogram of birthdays using breaks to count number of students with each birthday.

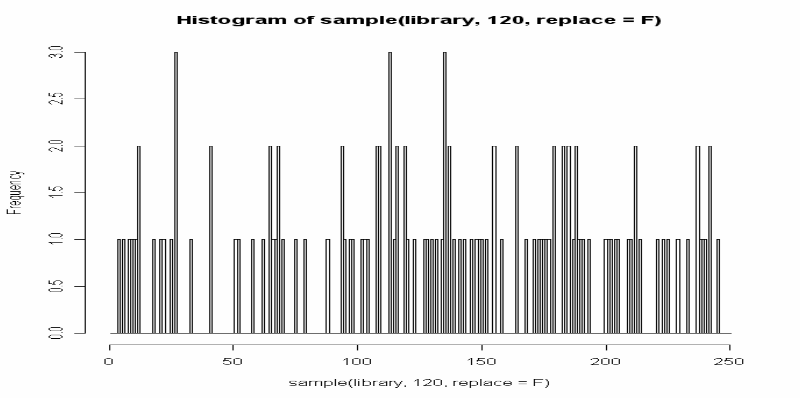


- In this sample, there are 3 birthdays that appear twice in the 40 students.
- Simulate: In 8897 of 10000 samples there is at least one shared birthday.

Observed and Theoretical Probabilities and Distributions

- iPod shuffle feature – Is it 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.

- Is Levy correct? Is probability of 3 or more songs from ANY one album small?
- Probability of 3 or more songs from ONE SPECIFIC album is small – approx 1%.
- Assume
  - 3000 songs
  - 250 albums
  - 12 songs per album.
- library<- rep(1:250, 12)
- Simulate selection of 120 songs from library.



- In this sample, there were 3 songs selected from 3 albums.
- Simulate: In 9453 out of 10000 samples at least 3 songs were selected from at least one album.

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

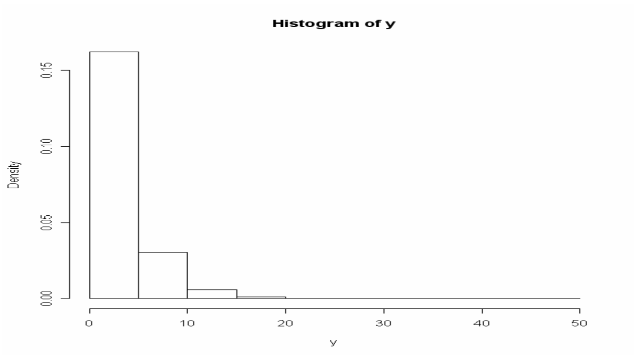
Transformations of Random Variables

- U is uniform(0,1). Y = -3 ln(1-U). What is the distribution of Y?
- Simulate y: u<- runif(10000,0,1);

  y<- -3*log(1-u)
- mean(y)

  [1] 2.98441
- var(y)

  [1] 9.38336



- Right-skewed distribution
- Mean approx. 3
- Variance approx. 9 = 3^2
- Exponential??

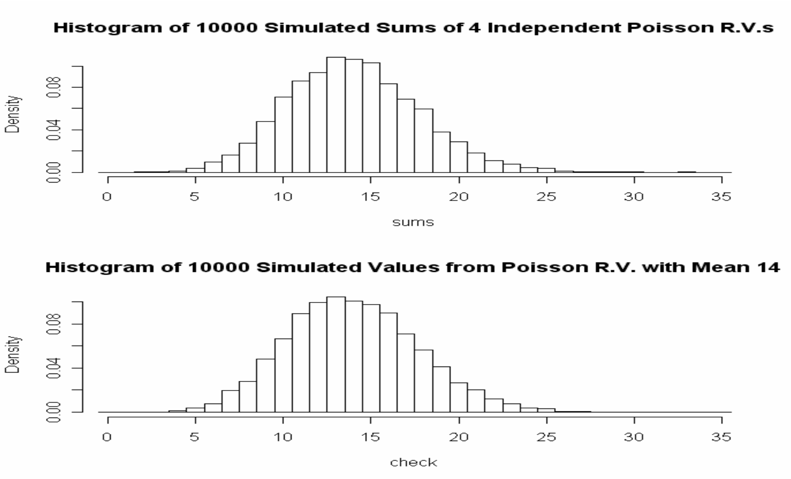
Transformations of Random Variables

- Y1, Y2, Y3 and Y4 are independent Poisson r.v. with means 3, 2, 5 and 4 respectively.
- What is the distribution of Y1 + Y2 + Y3 + Y4?
- Simulate: sums<- rpois(10000, 3) + rpois(10000, 2) + rpois(10000, 5) + rpois(10000, 4)
- What distribution does the sum have? Is it Poisson?
- mean(sums)

  [1] 14.0018
- var(sums)

  [1] 13.97699

- Is the observed distribution of the sum similar to the distribution of a Poisson r.v. with mean 14?

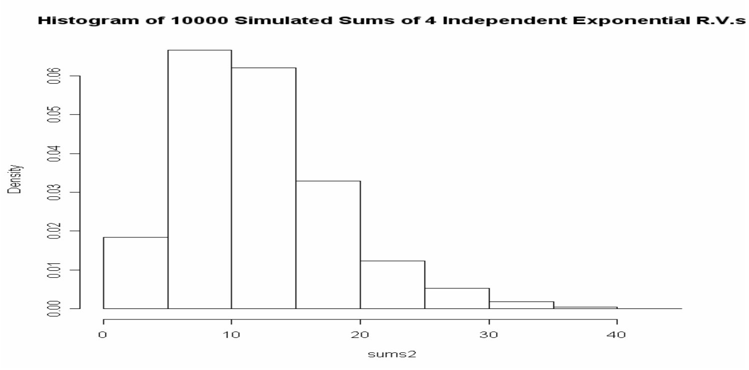


Transformations of Random Variables

- Y1, Y2, Y3 and Y4 are independent exponential r.v.s with mean 3.
- What is the distribution of Y1 + Y2 + Y3 + Y4?
- Simulate: sums2<- rexp(10000, 1/3) + rexp(10000, 1/3) + rexp(10000, 1/3) + rexp(10000, 1/3)
- mean(sums2)

  [1] 12.03085
- var(sums2)

  [1] 36.44002



- Right-skewed distribution
- Mean approx. 12
- Variance approx. 36
- Exponential??
- Other Related Distribution??

Visualize

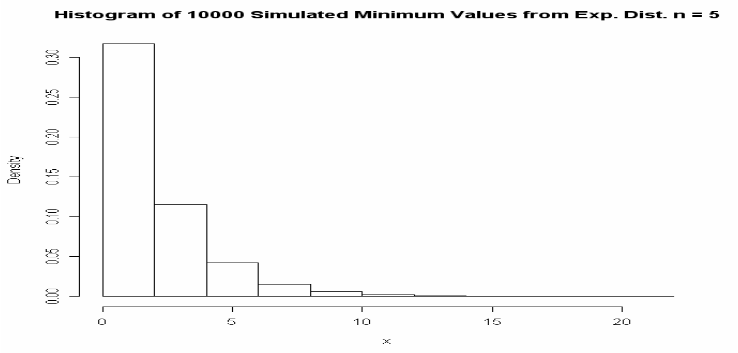
- Distribution Functions of Random Variables
- *Order Statistics
- Likelihood Functions
- Asymptotic Normality of Maximum Likelihood Estimators

Ordered Statistics

- X = Min(Y1, Y2, . . . , Y5) where Yi are independent exponential r.v.s with mean 10.
- Distribution of X?
- Simulate: for (i in 1:10000) x[i]<- min(rexp(5, 1/10))
- mean(x)

  [1] 2.011339
- var(x)

  [1] 4.106261



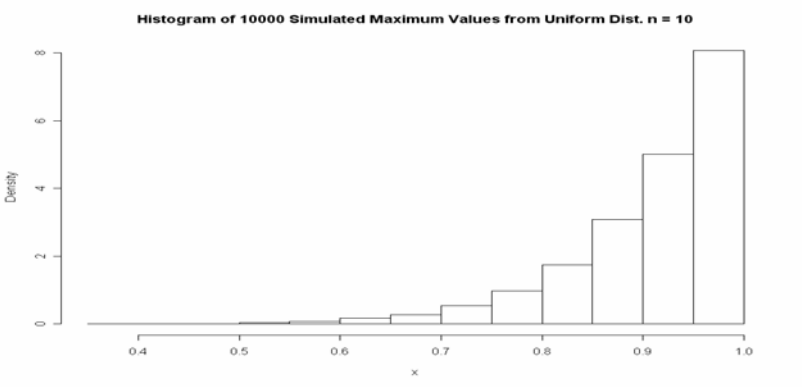
- Right-skewed distribution
- Mean approx. 2
- Variance approx. 4
- Exponential??

Ordered Statistics

- X = max(Y1, Y2, . . . , Y10) where Yi are independent Uniform (0, 1) r.v.s
- Distribution of X?
- Simulate: for (i in 1:10000) x[i]<- max(runif(10, 0, 1))
- mean(x)

  [1] 0.9098818
- var(x)

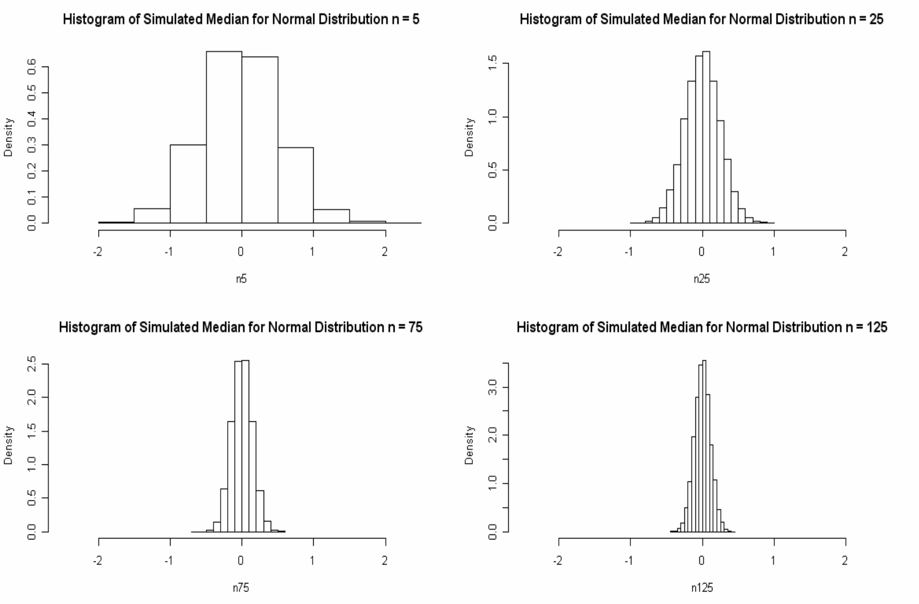
  [1] 0.006769303



- Left-Skewed Distribution
- Mean approx. 0.91
- Variance approx. 0.007
- Distribution???

Ordered Statistics

- X = median(Y1, Y2, . . ., Yn) where Yi are independent Normal r.v.s with mean 0 and variance 1.
- Vary n: n = 5, n = 25, n = 75, n = 125
- Distribution of X? Depends on n?
- Simulate: for (i in 1:10000) xn[i]<- median(rnorm(n, 0, 1))



Observed Means and Variances
of the Simulated Medians

	n = 5	n = 25	n = 75	n = 125
Mean	-0.0075	0.0010	-0.0005	-0.0005
Var	0.2834 (0.20)	0.0612 (0.04)	0.0205 (0.013)	0.0124 (0.008)

Expand

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

Role of Assumptions in Statistical Inference

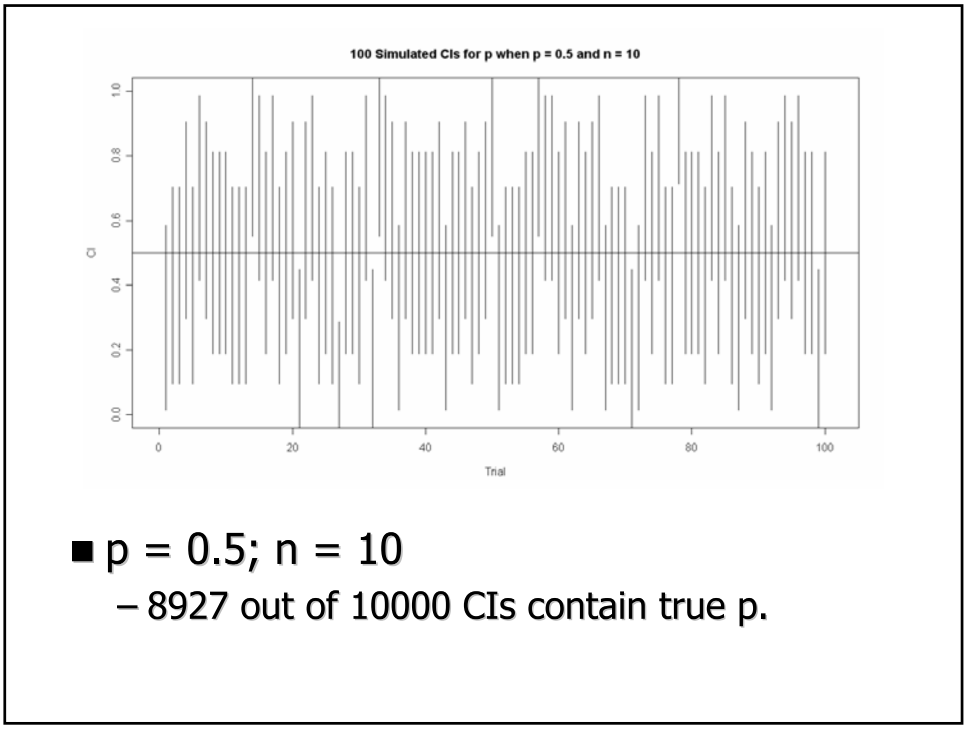
- Y1, Y2, . . ., Yn are independent Bernoulli r.v.s with probability of success p.
- Approx. 95% CI for p

[redacted]

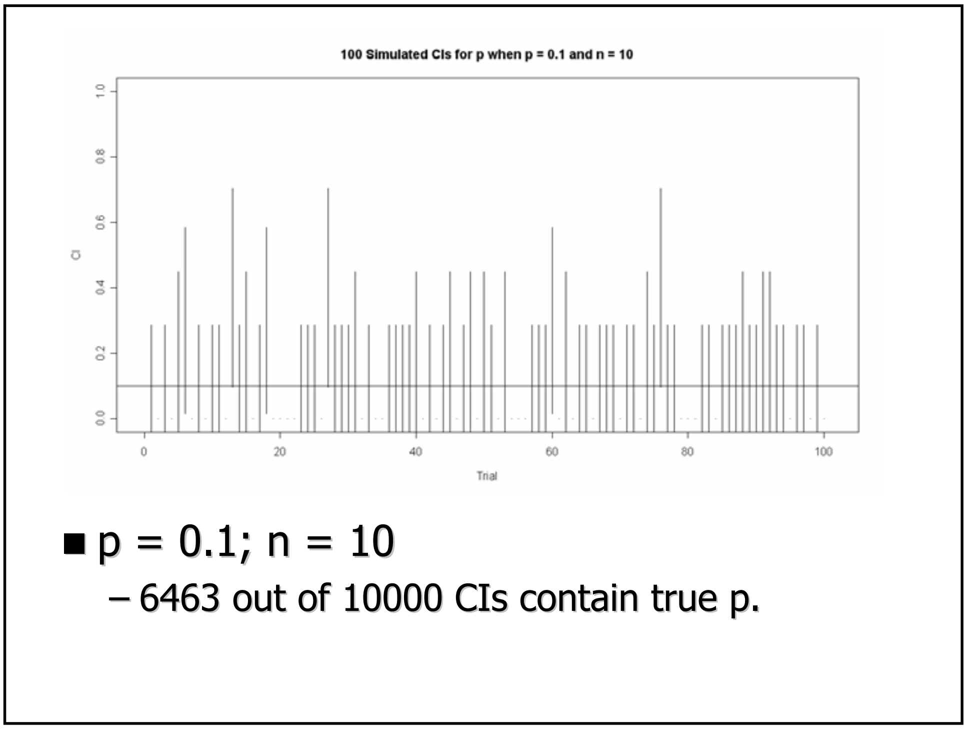
- Why do we say this 95% CI is approx.?

  Based on Normal Distribution.
- When does this approximation work?

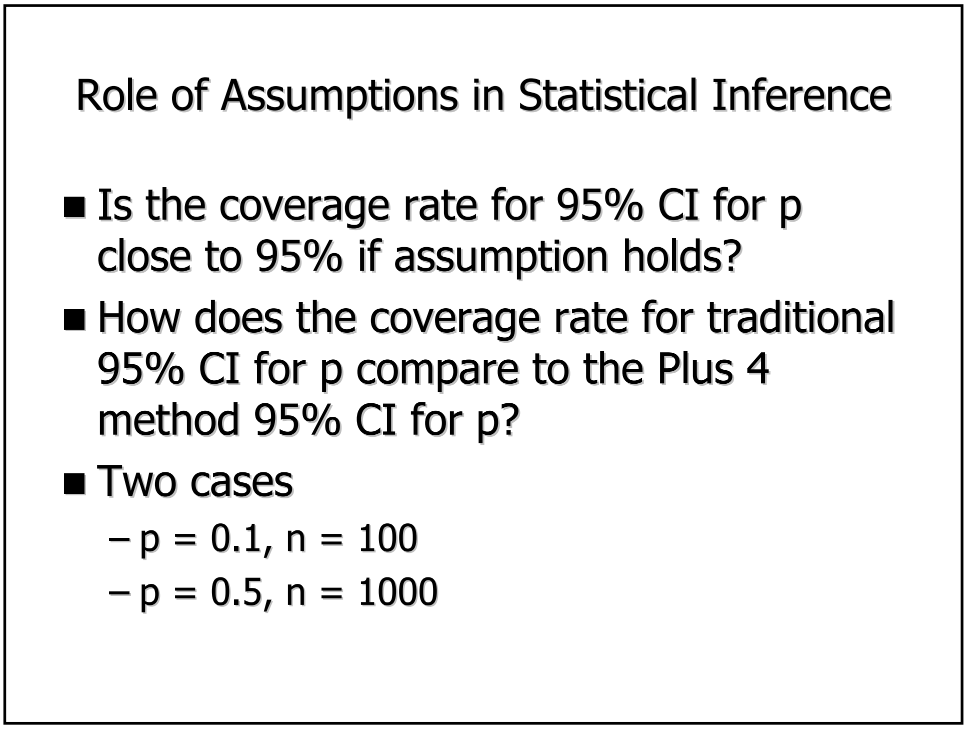
  Assumptions
  - np ≥ 10
  - n(1-p) ≥ 10
- What happens when this assumption is not true?
  - Ex. n = 10; p = 0.5 or p = 0.1



- p = 0.5; n = 10
  - 8927 out of 10000 CIs contain true p.

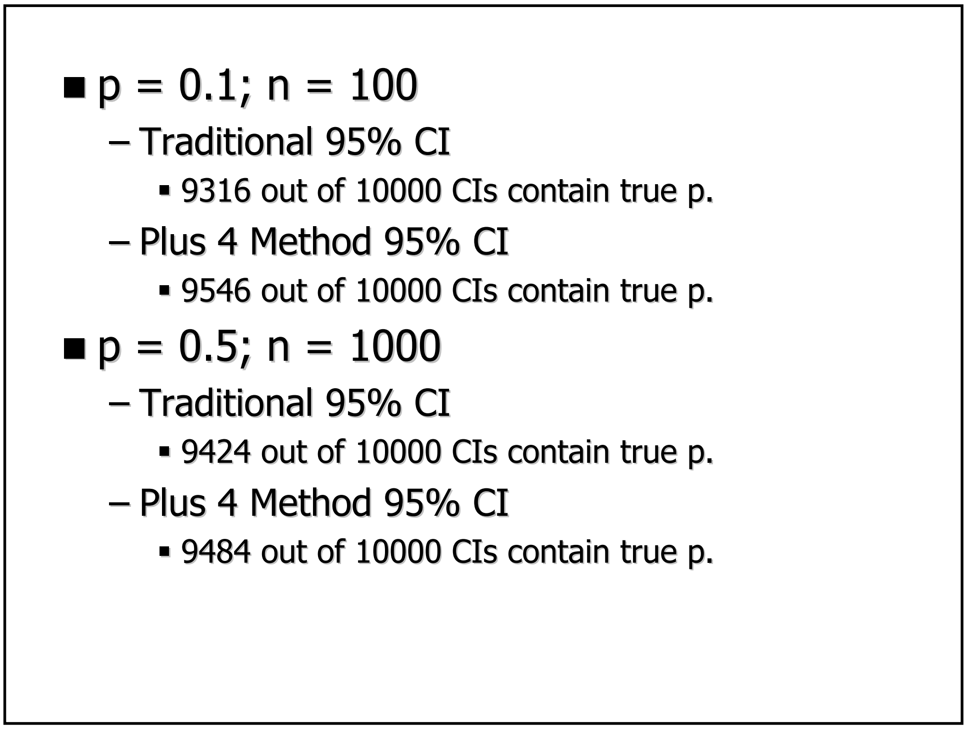


- p = 0.1; n = 10
  - 6463 out of 10000 CIs contain true p.



Role of Assumptions in Statistical Inference

- Is the coverage rate for 95% CI for p close to 95% if assumption holds?
- How does the coverage rate for traditional 95% CI for p compare to the Plus 4 method 95% CI for p?
- Two cases
  - p = 0.1, n = 100
  - p = 0.5, n = 1000



- p = 0.1; n = 100
  - Traditional 95% CI
    - 9316 out of 10000 CIs contain true p.
  - Plus 4 Method 95% CI
    - 9546 out of 10000 CIs contain true p.
- p = 0.5; n = 1000
  - Traditional 95% CI
    - 9424 out of 10000 CIs contain true p.
  - Plus 4 Method 95% CI
    - 9484 out of 10000 CIs contain true p.