Statistical Simulation: Learning and playing with statistics in R

Prathiba Natesan
Associate Professor
University of North Texas
Statistics

• Extracting scientifically meaningful information from data of all types

• Summarize large amounts of data with a few numbers
  • insight into the process that generated the observed data

• Determining probabilities
  • deductive
  • computing probabilities given a statistic: $pr|s$

• Statistical reasoning
  • inductive
  • guessing the best choice for parameters given the data $s|data$
  • how close our guess is to the real population parameters
Probability Distributions

• All possible events and their respective probabilities

• Univariate:
  • Normal
  • $t$
  • $\chi^2$
  • Skewed normal
  • Uniform

• Multivariate:
  • Multivariate normal
  • Wishart
Statistical Simulation

• Investigate the performance of statistical estimates under varying conditions

• Usually the generating parameters, distributions, and models are known

• Monte Carlo methods used to generate data
  • rely on repeated random sampling
  • Generate draws from a probability distribution
Normal Distribution Example in R

```r
Normal_distribution.R
data <- rnorm(n=5, mean = 0, sd = 1)
#you dont have to specify n, mean, and sd
#instead you can simply type
data <- rnorm(5, 0, 1)
#let us plot the probability density
plot(density(data))
```
Normal Distribution Example in R

• What is the mean of this distribution?
• What is the SD?
• How can I get estimates that more accurately reflect the population?

Normal_distribution_2.R
Rewritten as Normal_distribution_2b.R
Skew_Normal.R
Exercise

• Generate two uniform distributions as follows
• Sample 1 ~ unif(-1,1); n = 5
• Sample 2 ~ unif(-100, 100); n = 5000
• Compare the descriptives
• Plot the densities

Uniform_distribution_2c.R
Autoregression

• Autoregression_example.R
Why Simulation?

• Understand the nuts and bolts of statistical concepts
• Because you already know the true values
• Test the concepts for irregular/idiosyncratic data
• Extend the concepts to newer applications/situations
• Develop new statistical concepts/models
• GREAT teaching tool!
Understanding sampling distribution

• Define sampling distribution
• Distribution of that statistic, when derived from a sample of size n
• Sampling distributions contain statistics and not scores
Sampling distribution of the mean

Algorithm

1. Create a population so you know the “true” parameter values \{y\}
2. Decide on a sample size (or many sample sizes) \{n\}
3. Draw a sub-sample and compute its mean \{\text{sub.sample}\}
4. Store the mean \{\text{averages}\}
5. Repeat steps 3 and 4
6. Averages is the Sampling Distribution of the mean
7. SD of Averages is the standard error: compare with theoretical se
8. Theoretical se = SD(y)/\sqrt{n}
Sampling Distribution

• Sampling_Distribution_a.R
But how close is close enough?
Simulation Diagnostics (1/3)

• RMSE: Root mean squared error

\[
RMSE = \sqrt{\frac{1}{rep} \sum_{i=1}^{rep} (s_i - S)^2}
\]

• Bias
  • Average bias: cancels out
  • Relative bias

• Probabilities (e.g. Natesan et al., under review)

• Bounded vs unbounded
Creating a function in R

```r
averages <- function(){
  S <- sum(vec)  #variables created within the function
  L <- length(vec)  #do not exist outside the function
  A <- S/L
  return(A)  #Asks the function to output A
}
#end of function
Sampling_distribution_b.R
```