Data Science and Machine Learning

Prof. Eric A. Suess
July 2019

机器学习
Machine Learning is exciting!!!

- If you are not already super excited about Machine Learning and Artificial Intelligence I hope you will be at the end of the weekend.
- I want you to become fluent in ML terminology and become more aware of newer ML methods.
- And I hope to give you the opportunity to find a new idea where you can see the use of ML algorithms to transform your organization and a business process.

第一天　Traditional Data

- **Statistics:** Understanding data and the value of data for improved decision making. Simulation. Bayesian thinking.
- **Data Mining:** Methods for data analysis and discovery in databases/spreadsheets.
- **Business Analytics:** Methods for solving important business questions.

第二天　Data Now

- **Time Series Analysis:** Trends, Patterns, Predictions and Forecasts
- **Change Point Analysis:** Change in level
- **Anomaly Detection:** Outliers
- **Regression** and **Logistic Regression** Modeling: Model the relationship between inputs and output(s)

第二天　Data Now

- **Data Visualization:** Maps and dynamic visualizations
- **Big Data:** Database Storage, Data reduction
- **Unsupervised and Supervised Learning:**
- **Clustering:** Grouping or Segmenting
- **Market Basket Analysis:** Co-purchased items
- **Machine Learning (ML):** Prediction and Classification, Pattern recognition, Generative Models
- **Artificial Intelligence (AI):** Bigger collection of ideas that brings together many types of data and many algorithms and machines.
- **Streaming data:** Sensor data that is being collected in real time for many locations or collecting many different types of data.

教育 Education

- University of California, Berkeley
  - B.A. Statistics & Economics, with a minor in Demography
- California State University, East Bay
  - M.S. Statistics
- University of California, Davis
  - Ph.D. Statistics

终身教授 Professor

- **Full Professor** with tenure at California State University, East Bay, Department of Statistics and Biostatistics, College of Science, jointly appointed in the Department of Engineering, College of Engineering
- **Taught** courses for the California State University, East Bay, Economics Department, Marketing Department, Analytics Department, College of Business
- **Visiting Professor** at Stanford University, Time Series Analysis

系主任 Department Chair

- Three terms as Department Chair, 2006-2015.
- Hired 7 new professors and many lecturers.
- I am now the most senior professor in our Department.
- My focus for the last 5 years has been the **modernization** of our curriculum.
- Course development in **Machine Learning** and **Data Science**.

创业公司 Start-ups

- Currently I am the **Chief Statistician** at machineVantage
- Machine Learning and Artificial Intelligence Applied to Marketing and Product Innovation
- Text data and social media data
- Product sales data

创业公司 Start-ups

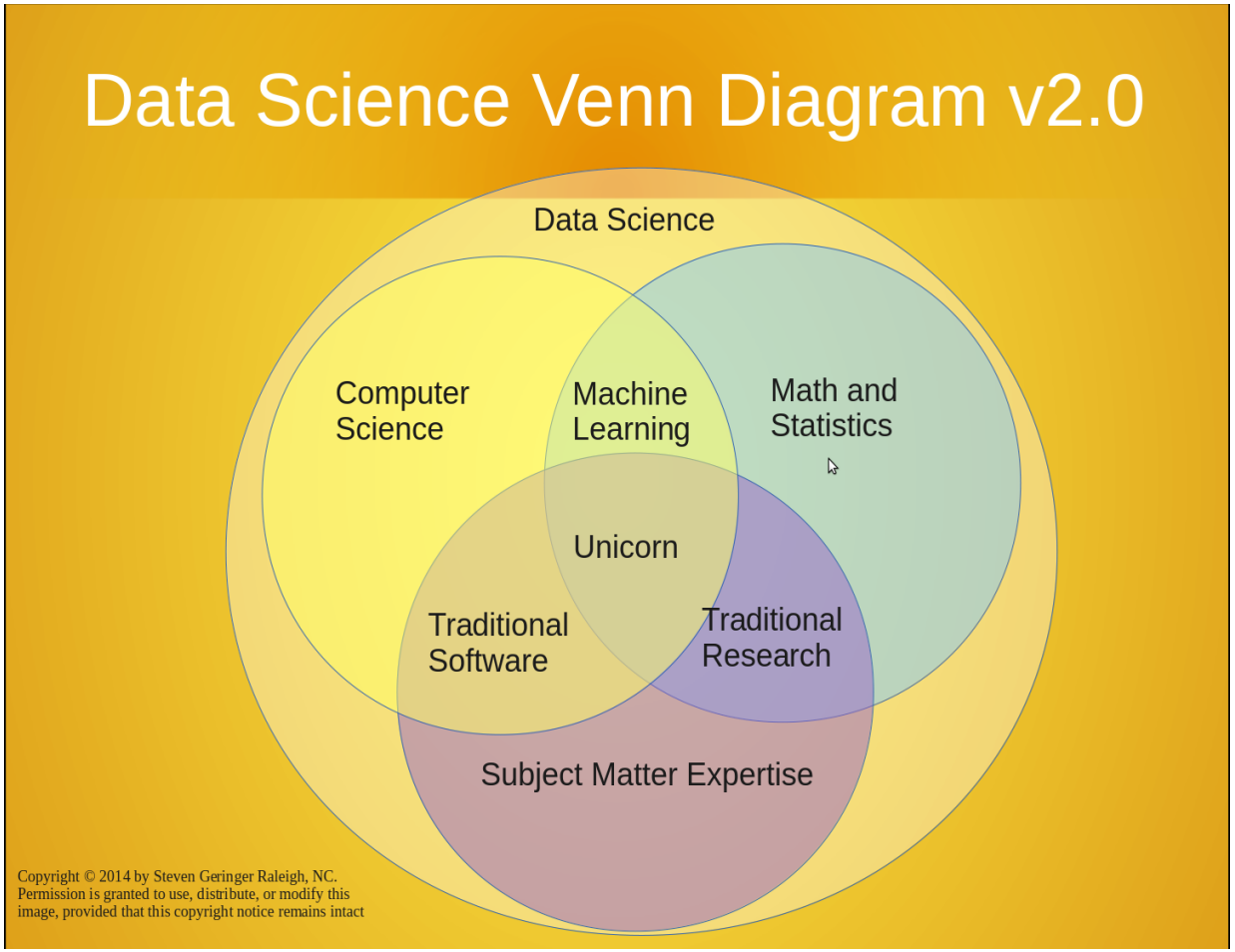
- Consulted for GameChanger a marketing firm focusing on Retail Labs. This was the testing of new product designs in stores with real customer.
- Helped my father with his Real Estate Appraisal business, BayHill Appraisal Associates, for 20+ years. Most of my work was with the use of computers, networking, and software to run his business.
- Helped my brother with his Air Quality and Energy Regulation business, DSG Solutions related to start-up and shut-down emission levels for power plans.
- I currently have my own company Hampton Consulting where I do Data Science work.

研究项目 Research Areas

- Probability Theory, Probability Simulation
- Statistical Modeling
- Bayesian Statistics
- Biostatistics
- Time Series Analysis
- Data Science
- Machine Learning
- Natural Language Processing (NLP)
- Deep Learning

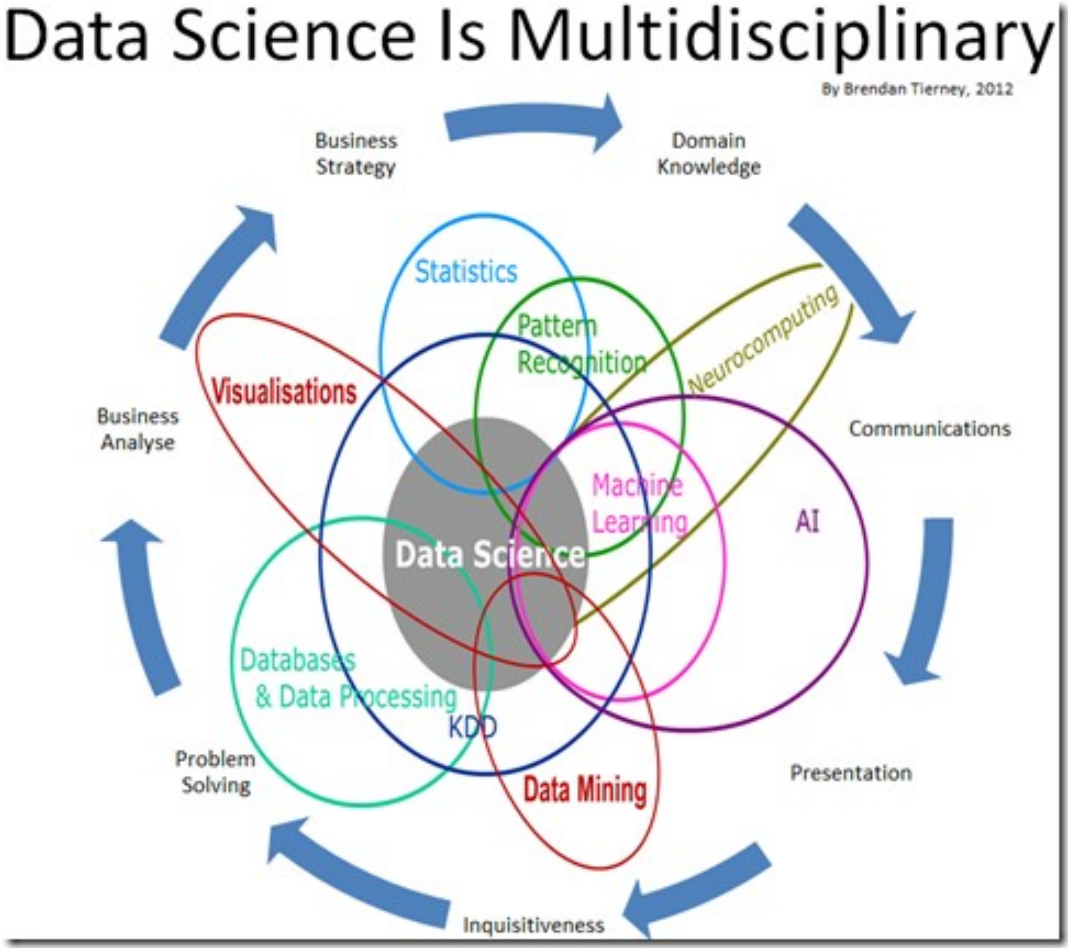
数据科学 Data Science

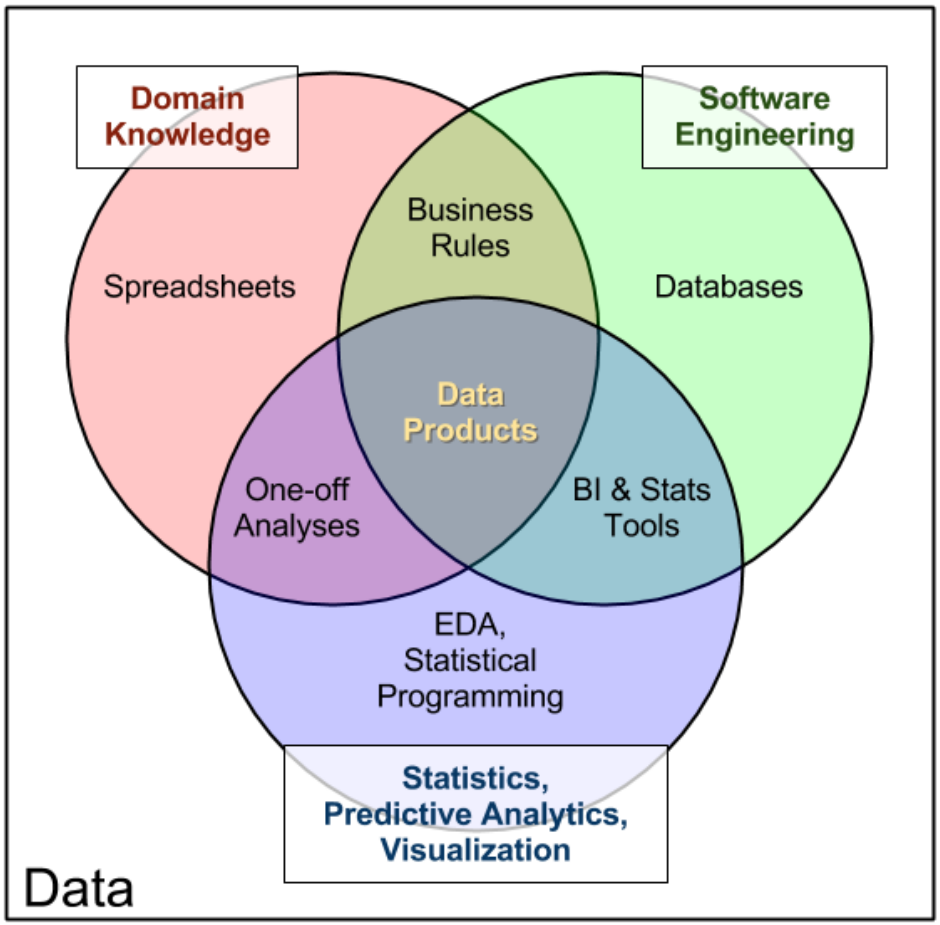
- What is Data Science?
- The Data Science Venn Diagram(s).



KDNuggets: Battle of the Data Science Venn Diagrams

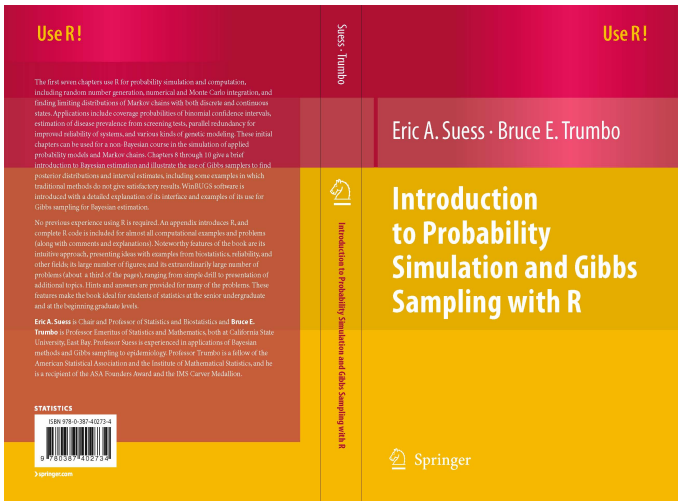
Data Science Is Multidisciplinary




Domain Knowledge
Software Engineering
Business Rules
Spreadsheets
Databases
Data Products
One-off Analyses
BI & Stats Tools
EDA, Statistical Programming
Statistics, Predictive Analytics, Visualization
Data

书 Book

- Wrote a book with my co-author on Probability Simulation, Stochastic Processes, Markov Chain Monte Carlo, and Bayesian Statistics.
- The focus of the book was on Gibbs Sampling for Bayesian Statistics.



数据 Data

- **Seismic array** data analysis – Ripple fired mining explosions
- **Animal Epidemiology** – Estimates of disease in a country, farm, animal.
- **Sports Analytics** – Football playing ability
- **Air Quality** – Is a power plant's start up emissions too high?
- **Weather prediction** – Probability of rain tomorrow.
- **Social Media** Data – On-line reviews from Yelp.

Open Source Computing

- I use Linux extensively. Arch and Debian
- I like and use BSD also. OpenBSD
- I like embedded system. Replacing firmware on routers with OpenWRT and DDWRT is a hobby.
- I like working OrangePI, RasberryPI, Arduino systems for data collection.
- I work with Cloud computing Linux systems.

你是谁？ Who are you?

- You are the *leaders* and *visionaries* of your respective organizations and companies.
- You have the *big picture* of what **data** is available in your company.
- I believe you are interested in *improving* the functioning of your company and in the end the profitability of your company.
- I hope to give you some *new ideas* about the use of **Machine Learning**, **Data Science**, and **Data Technologies** that can be used with your data to answer important questions you are interested in answering.
- What we discuss today and tomorrow could be ideas you include in *new job descriptions* or *changes made to current roles* some of your employees working for your company.
- May change the way you think about the data you have and how it is used. Or give some ideas about *new data to collect*.

问题 Questions?

- **I like questions and discussion.**
- Please ask.
- One of my goals is to try to answer your questions and follow up when I don’t know the answer.

Questions to you.

3 Questions:

1) In your work do you have a business question that could be answered with data already being collected?

2) What is your biggest question about ML?

3) Do you have ML algorithms working today in your company?

第一天 Lets get started: Outline

- Statistics
- Data Mining
- Business Analytics
- Biostatistics, Quality Control, Actuarial Science, etc.

统计学 Statistics

- Classical Statistics uses **data** to estimate means and standard deviations. Or uses **data** to estimate proportions and percentages.

- *Hypothesis Testing* can be used to make *comparisons*.

- *Confidence Intervals* are used to be error bars on *estimates*.

统计学 Statistics

- Measures or Metric
- Total or Sum
- Average/Mean
- Counts
- Proportion
- Standard Deviation

$$X$$

$$T = \sum X_i$$

$$\bar{X} = \frac{\sum X_i}{N}$$

$$\hat{p} = \frac{number\ of\ successes}{N}$$

$$S = \sqrt{\frac{\sum \left(X_i - \bar{X} \right)^2}{N - 1}}$$

Simulation: Example

- I like to use **probability simulation** to think about data problems.
- Here we will start with a simulation of Yes, No outcomes. Yes = 1, No = 0.
- The first question is can we estimate the “true” proportion from the simulated data?

Computer Simulation:

I use computer simulation to help think about what may happen with randomly sampled data and to verify that modeling methods will detect what is expected in the data.

Parameter estimation is checked.

Computer Simulation used the *Law of Large Numbers*, convergence of means.

Computer Simulation uses the *Central Limit Theorem*, accuracy of means.

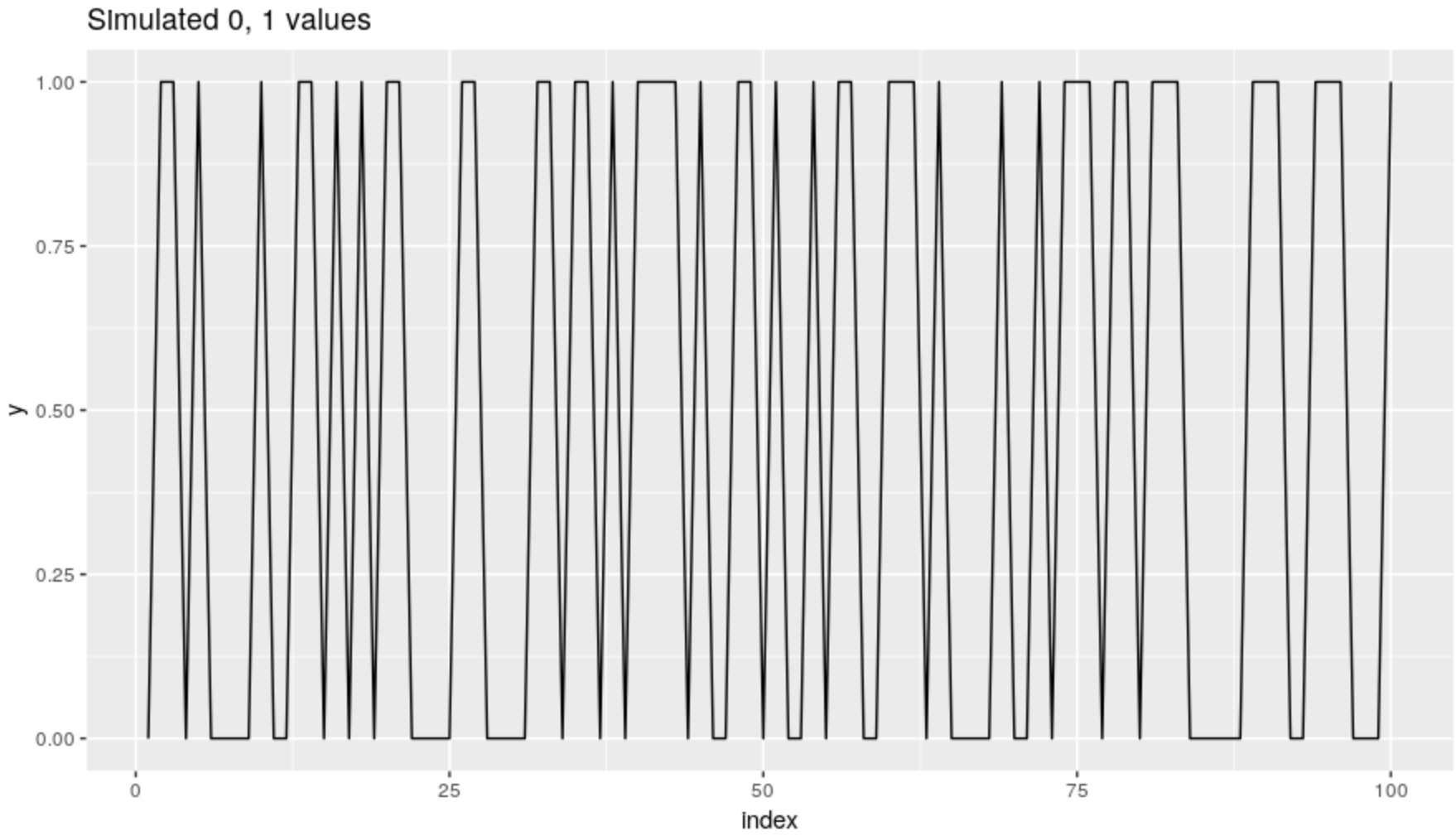
Simulation: Example 1

Experiment: Flip a fair coin 100 of times.

```
> Reps <- 100  # replications of the simulation
> n <- 1             # sample size
> p <- 0.5           # true value of the parameter
```

Truth: About half of the flips will be 1.

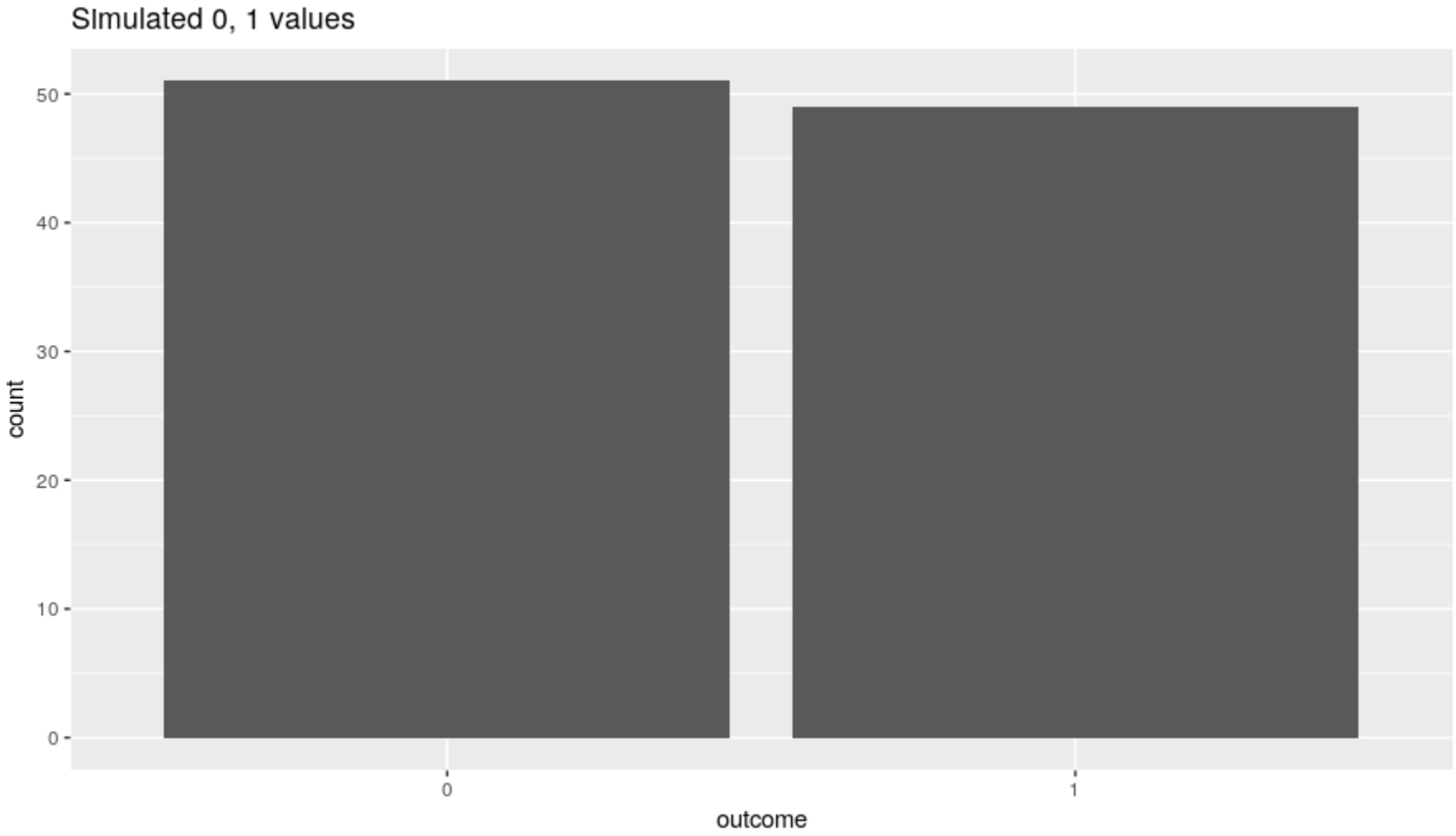
Simulation: Time Plot



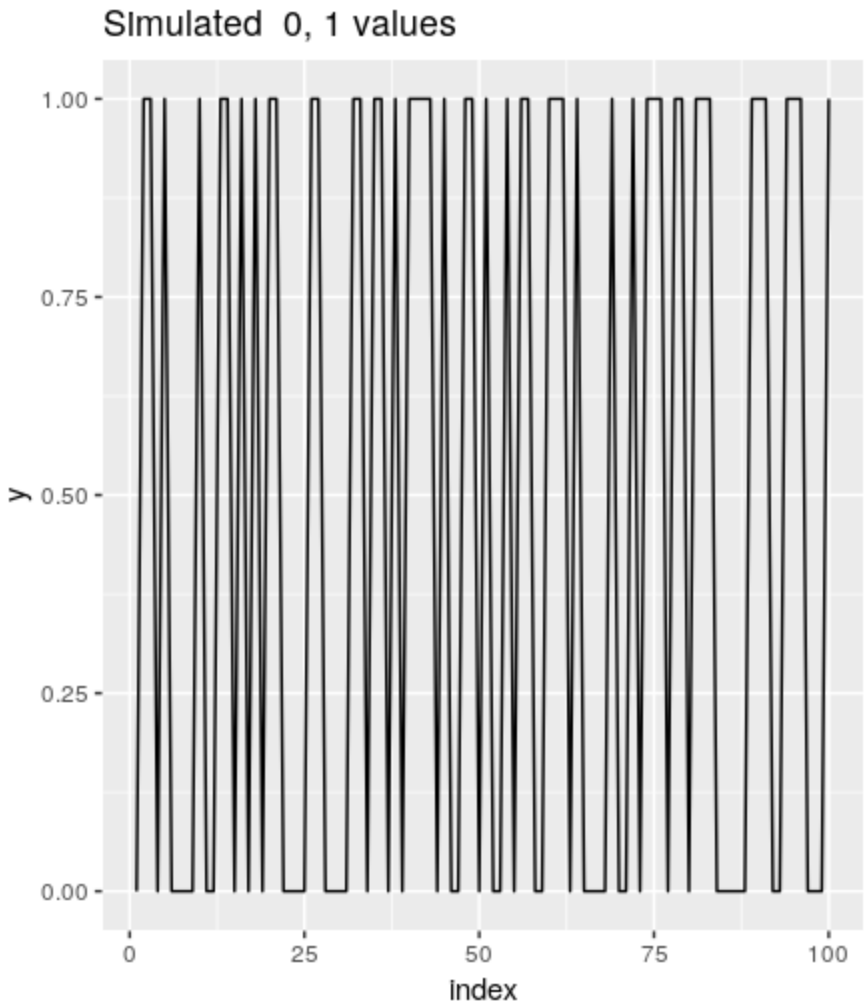
Simulation: Example 1

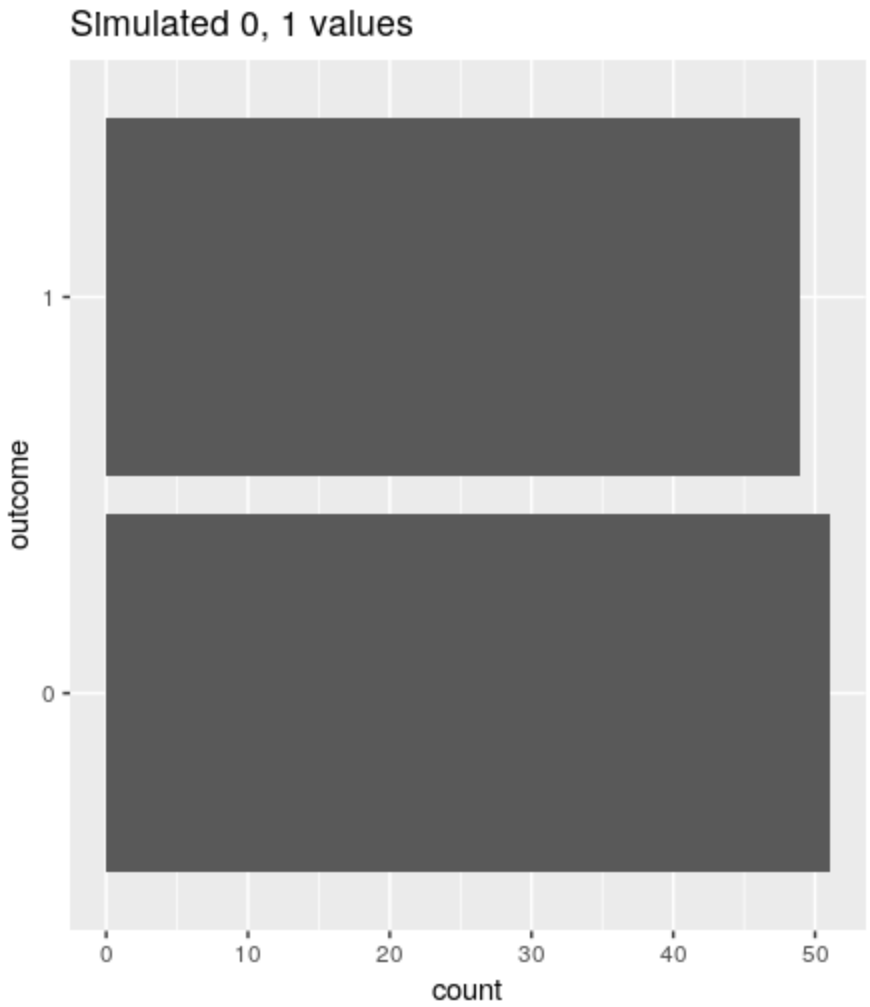
- Can you see any patterns in the simulated data?
- **Answer:** It is a bit hard to see. It looks like there are more 1s than 0s.
- Does the data look balanced between 0 and 1?
- **Answer:** Nearly balanced, but not equal. A graph would be helpful for answering this question.

Simulation: Bargraph



Simulation: Visualization





Simulation: Example 1

What have we learned?

- Simulated 100 numbers.
- Simulated data has random patterns. Here repeated values of 0s or 1s occur.
- We do not see the data flip back and forth.
- The distribution of values is theoretically equal with p = 0.5
- The time plot and horizontal bargraph is useful.

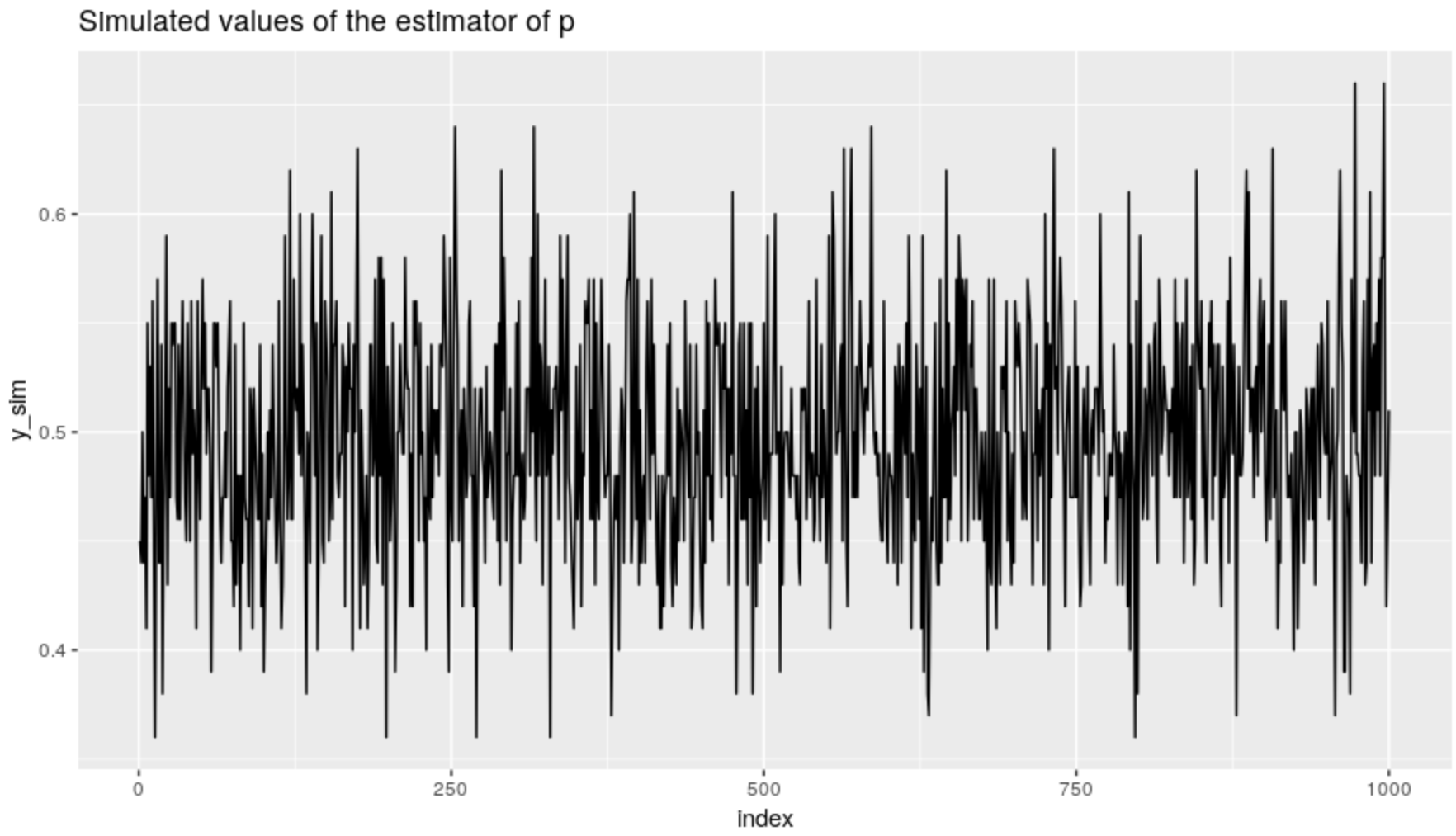
Simulation: Example 2

Experiment: Flip a fair coin 100 of times. Do the simulation 1000 times. Compute 1000 estimates of the proportion of 1s. What is the distribution of the estimator?

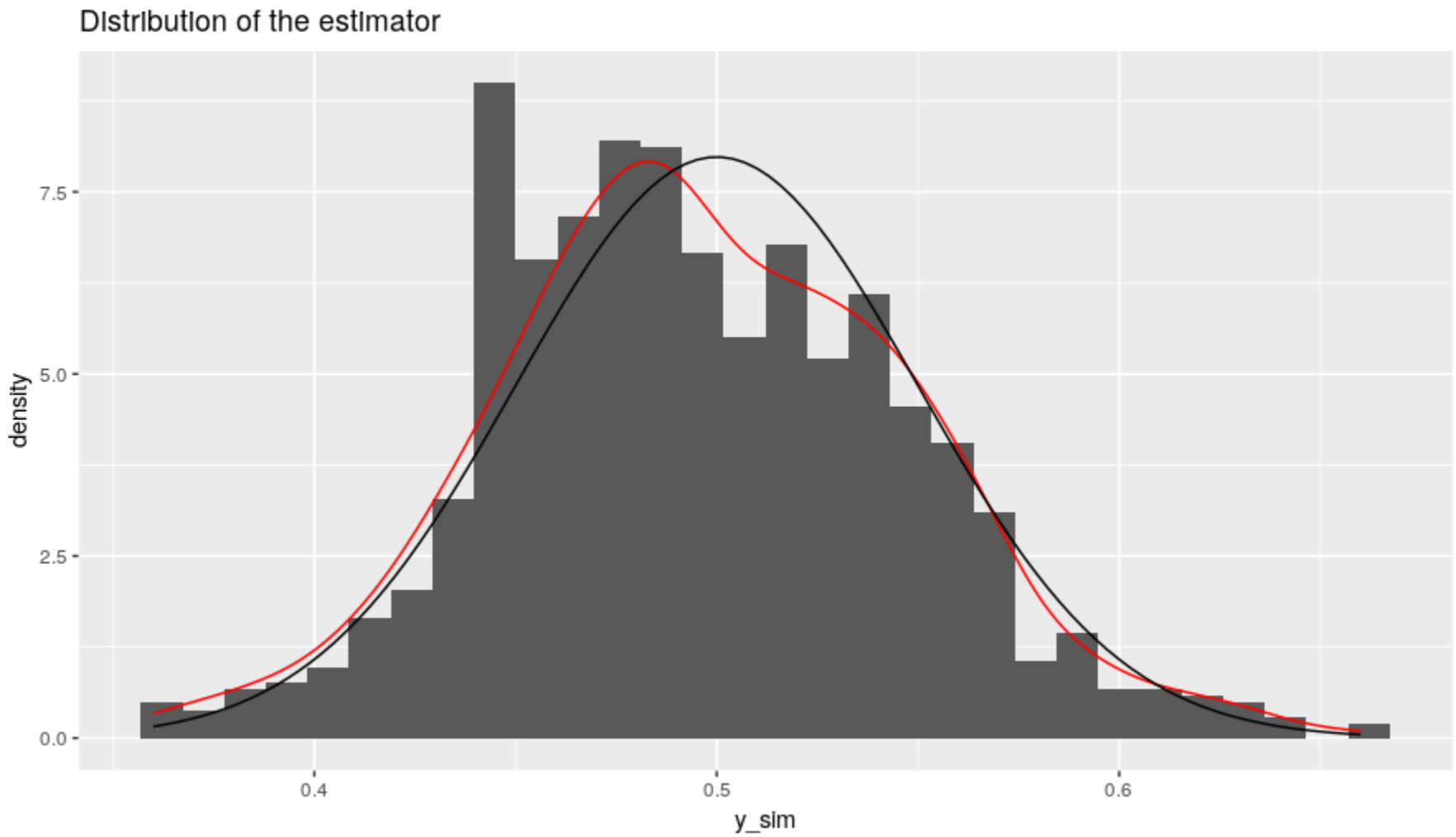
```
> B  <- 1000        # replications of the simulation
> Reps <- 100       # sample size
> n <- 1            # sample size
> p <- 0.5          # true value of the parameter
```

Truth: The proportion of 1s should be about half of the sample size.

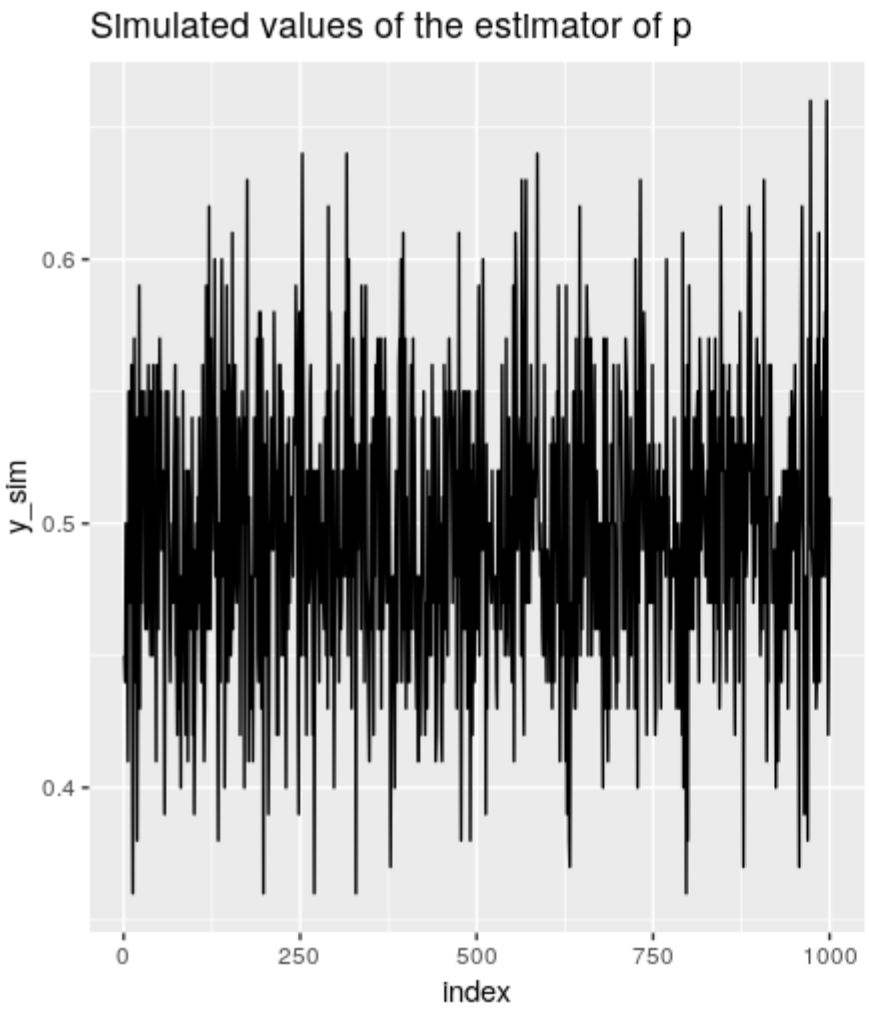
Simulation: Time Plot

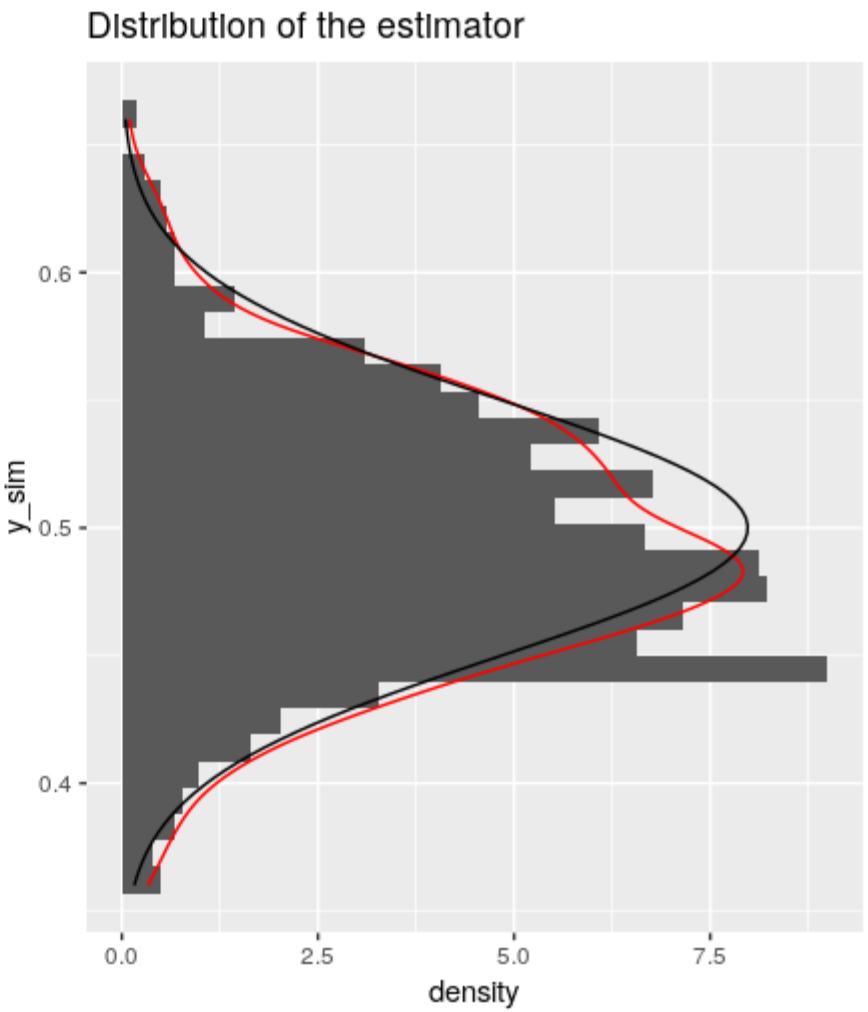


Simulation: Histogram

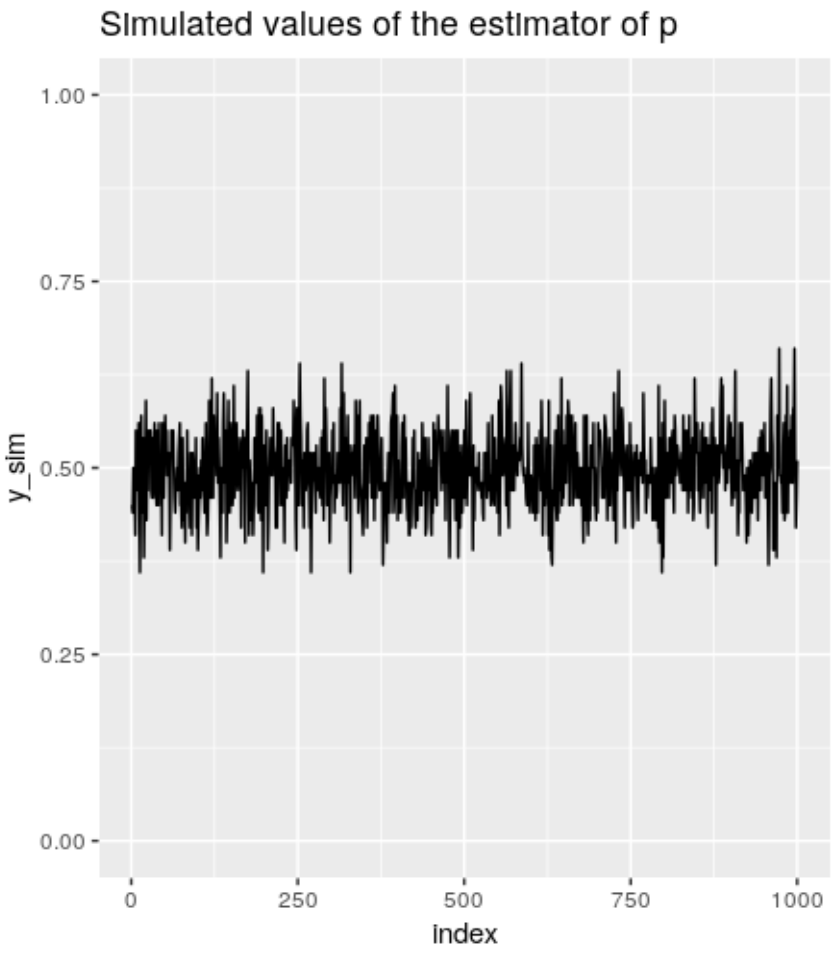


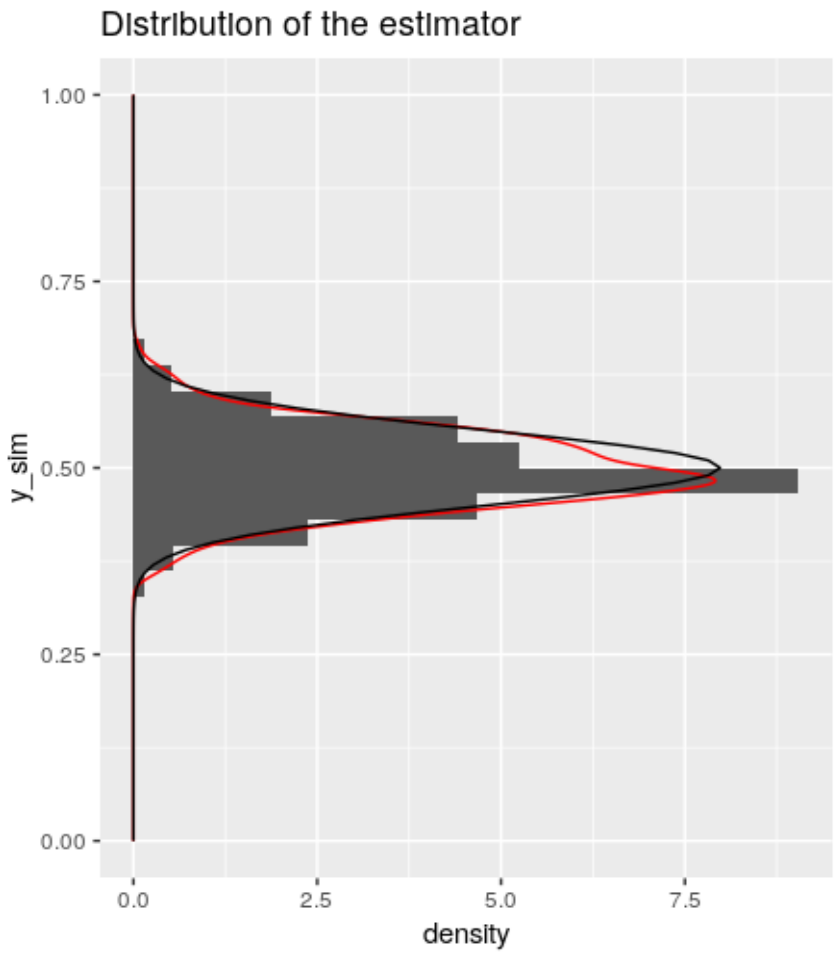
Simulation: Visualization





Simulation: Example 2





Simulation: Example 2

What have we learned?

- Simulated 100*1000 = 100,000 numbers.
- Simulated data has random patterns.
- Proportions are normally distributed.
- Larger sample sizes lead to more accuracy.
- An estimator of a proportion is more accurate with a larger sample size.
- Axes in visualizations need to be the same for direct comparison.

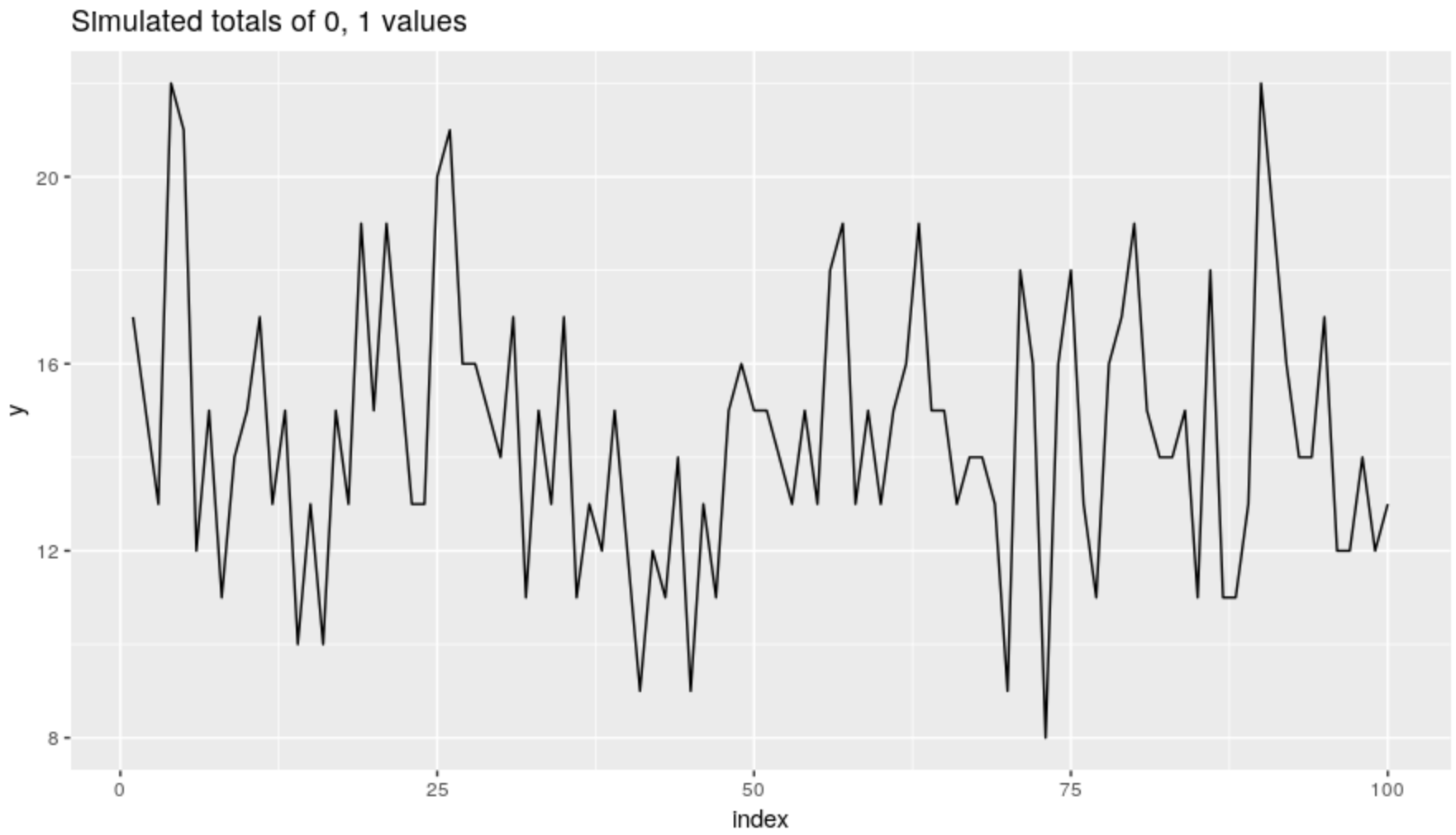
Simulation: Example 3

Experiment: Flip 30 fair coins 100 of times. What is the distribution of the total?

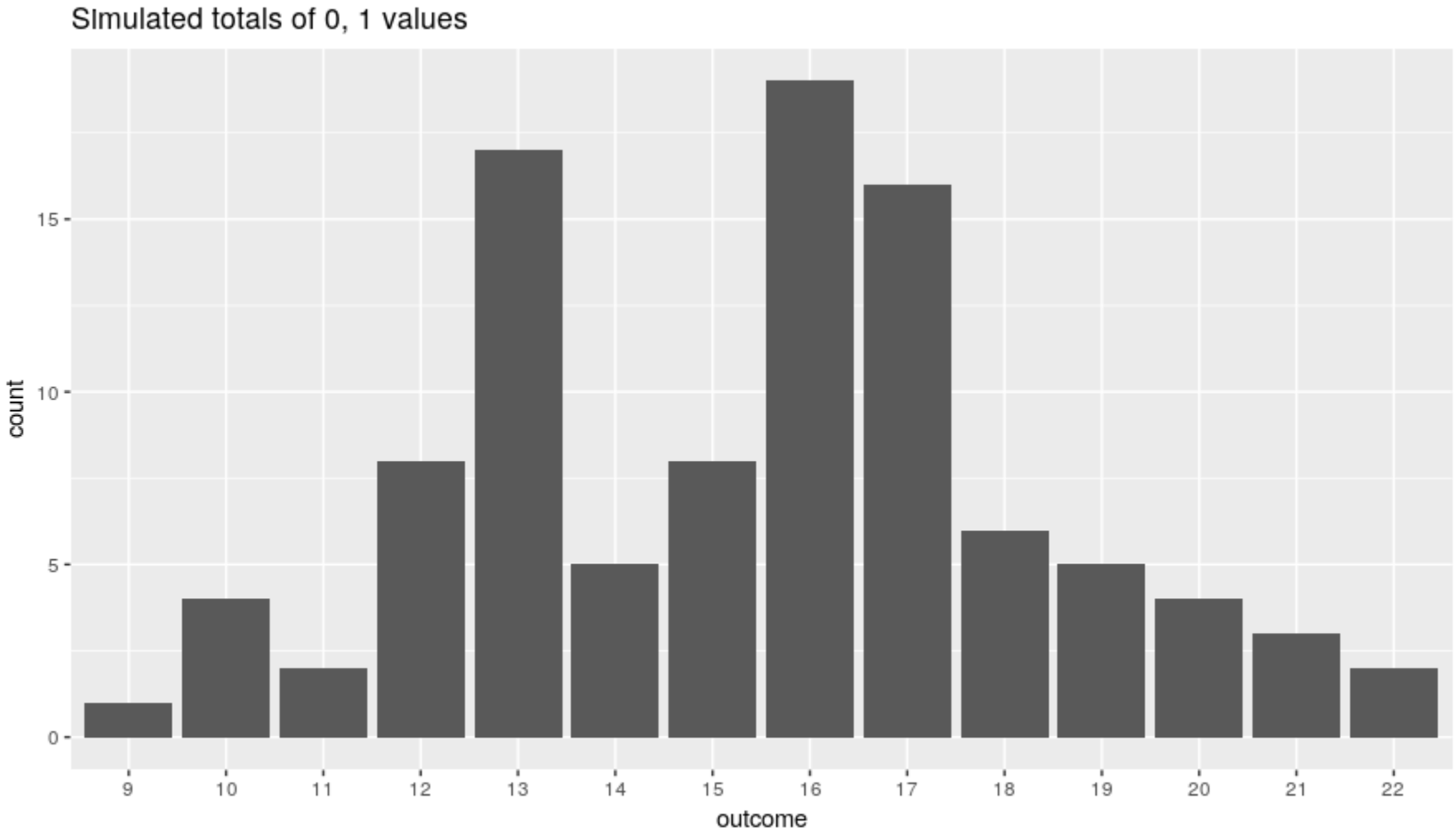
```
> Reps <- 100      # sample size
> n <- 30          # sample size
> p <- 0.5         # true value of the parameter
```

Truth: The proportion of 1s should be about half of the sample size, so the *expected value* is 15.

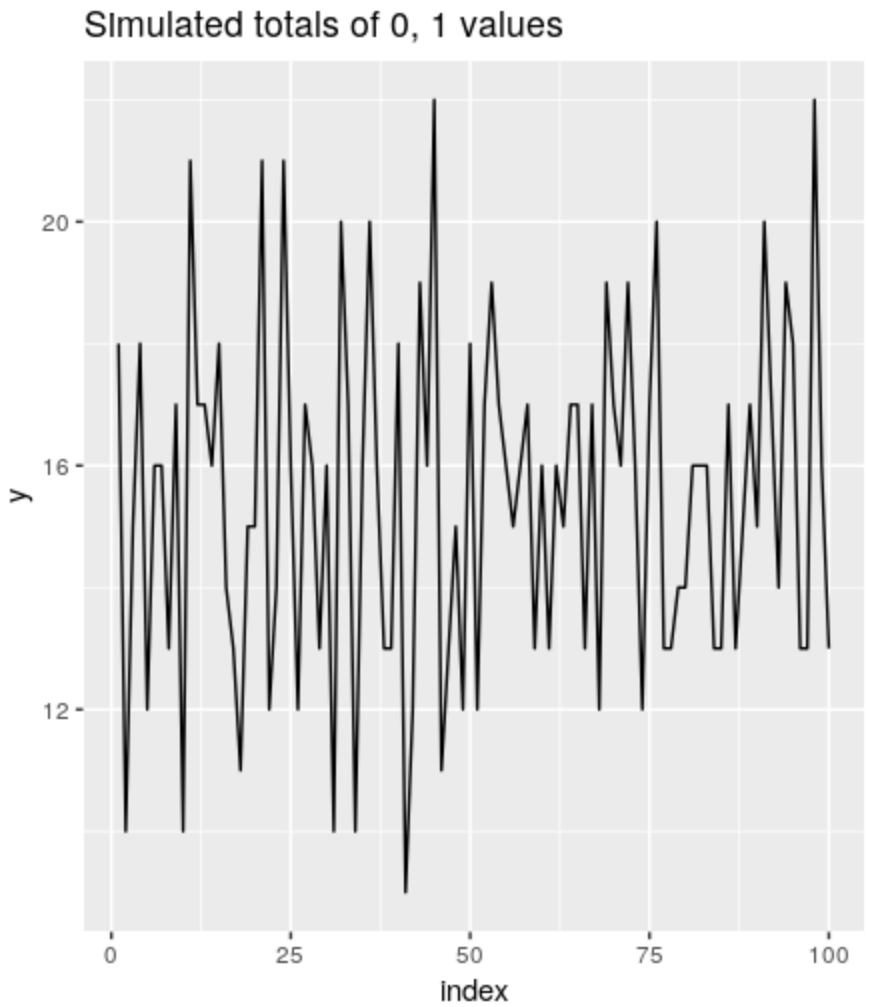
Simulation: Time Plot

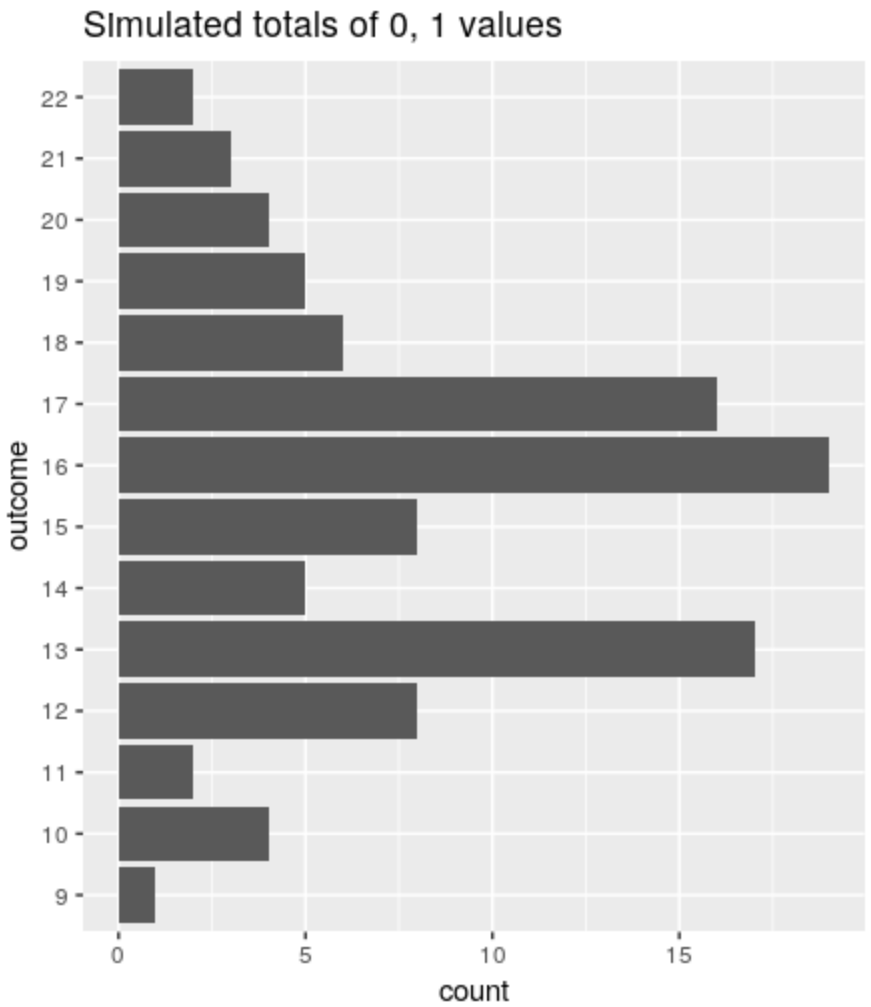


Simulation: Bargraph



Simulation: Visualization





Simulation: Example 3

What have we learned?

- Simulated 30*100 = 300 numbers.
- Simulated data has random patterns.
- Totals are normally distributed.

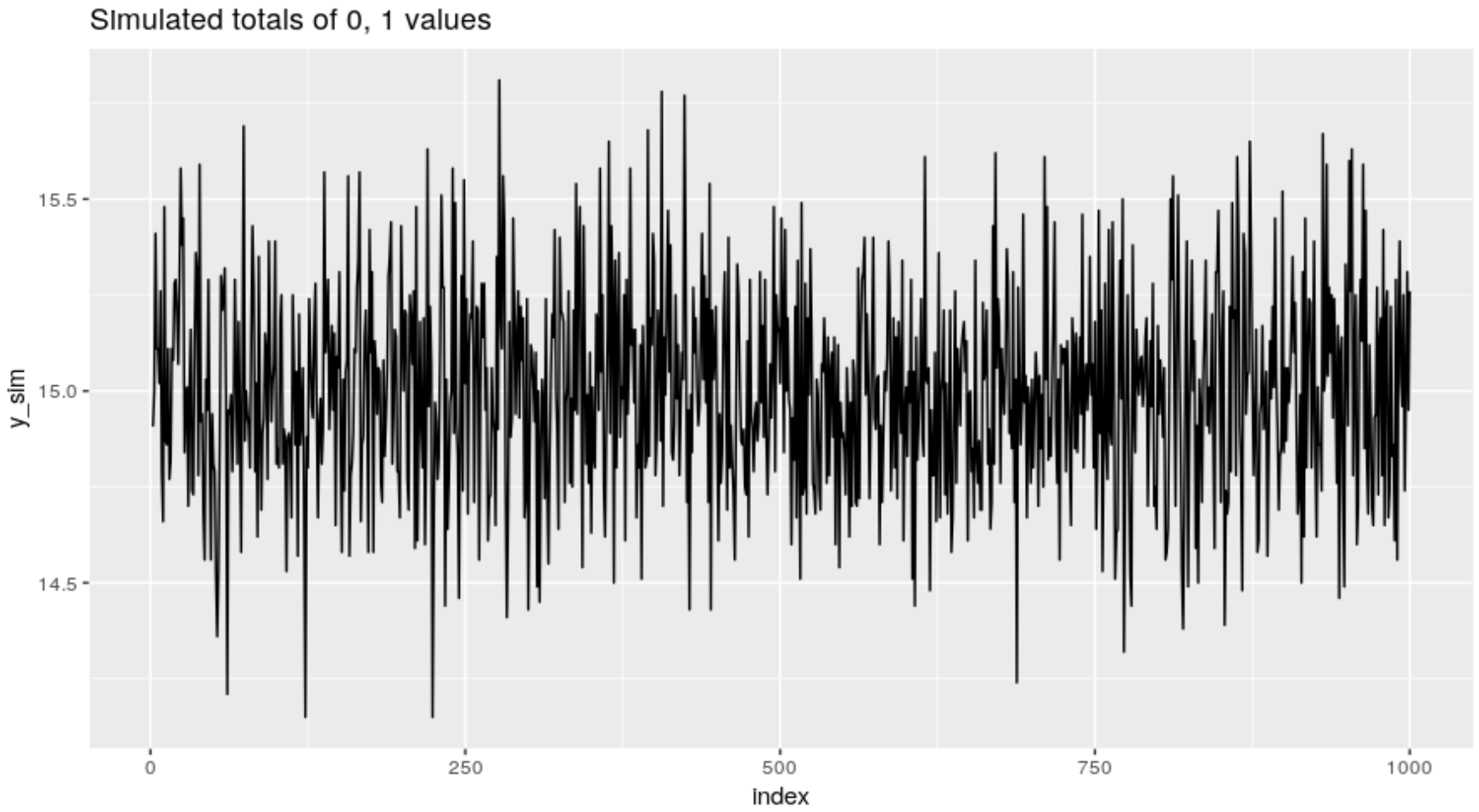
Simulation: Example 4

Experiment: Flip 30 fair coins 100 of times. Do the simulation 1000 times. Compute 1000 estimates of the total of 1s. What is the distribution of the estimator?

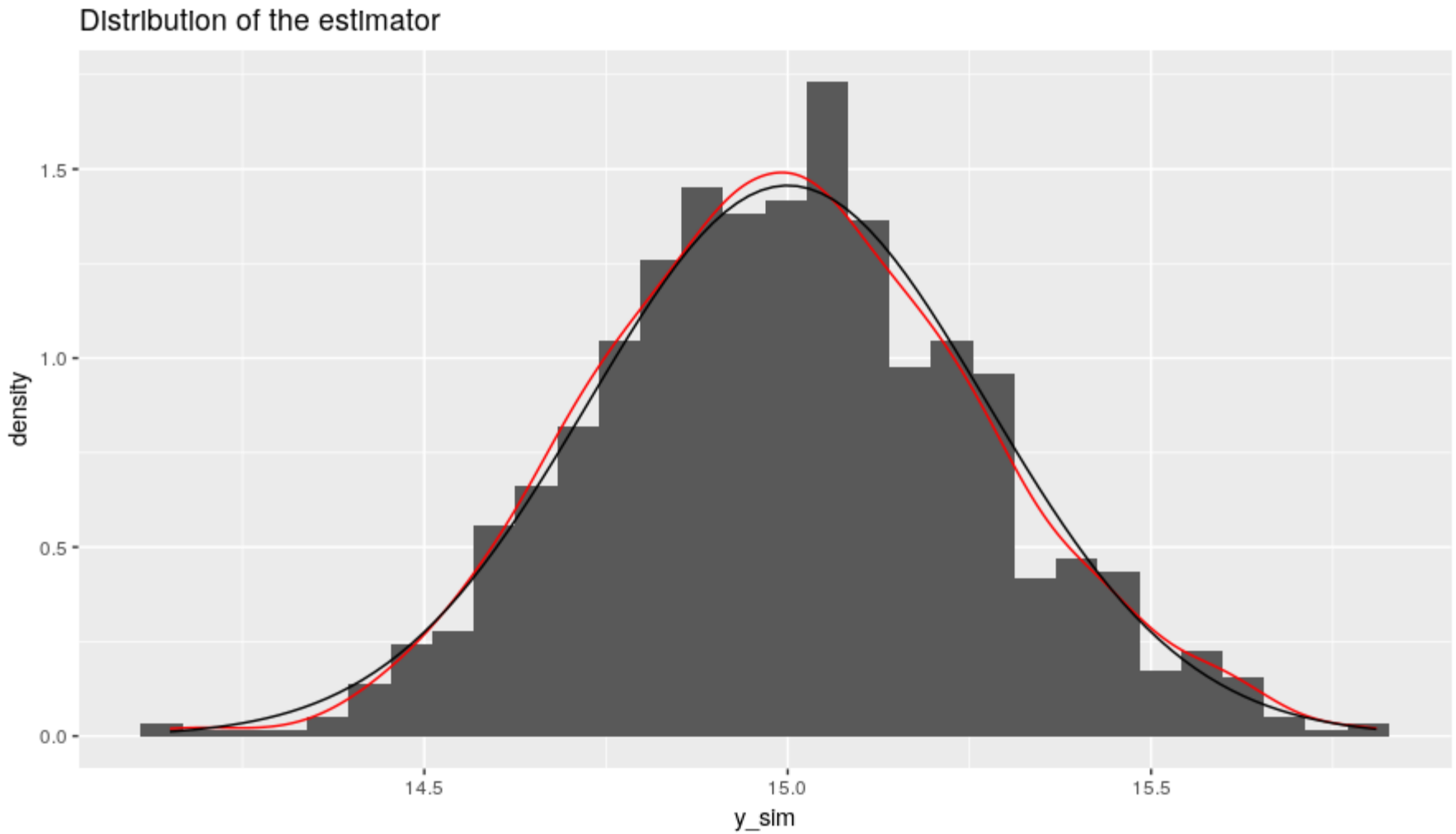
```
> B  <- 1000        # replications of the simulation
> Reps <- 100       # sample size
> n <- 30           # sample size
> p <- 0.5          # true value of the parameter
```

Truth: The proportion of 1s should be about half of the sample size, so the *expected value* is 15.

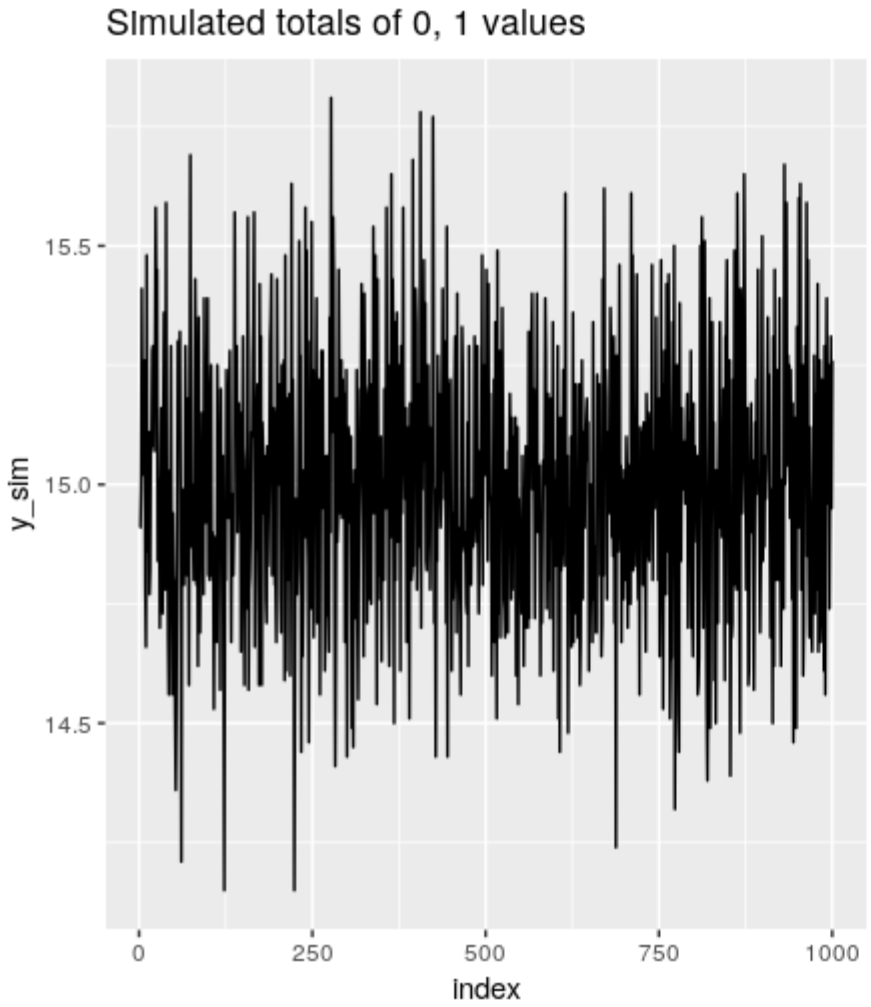
Simulation: Time Plot

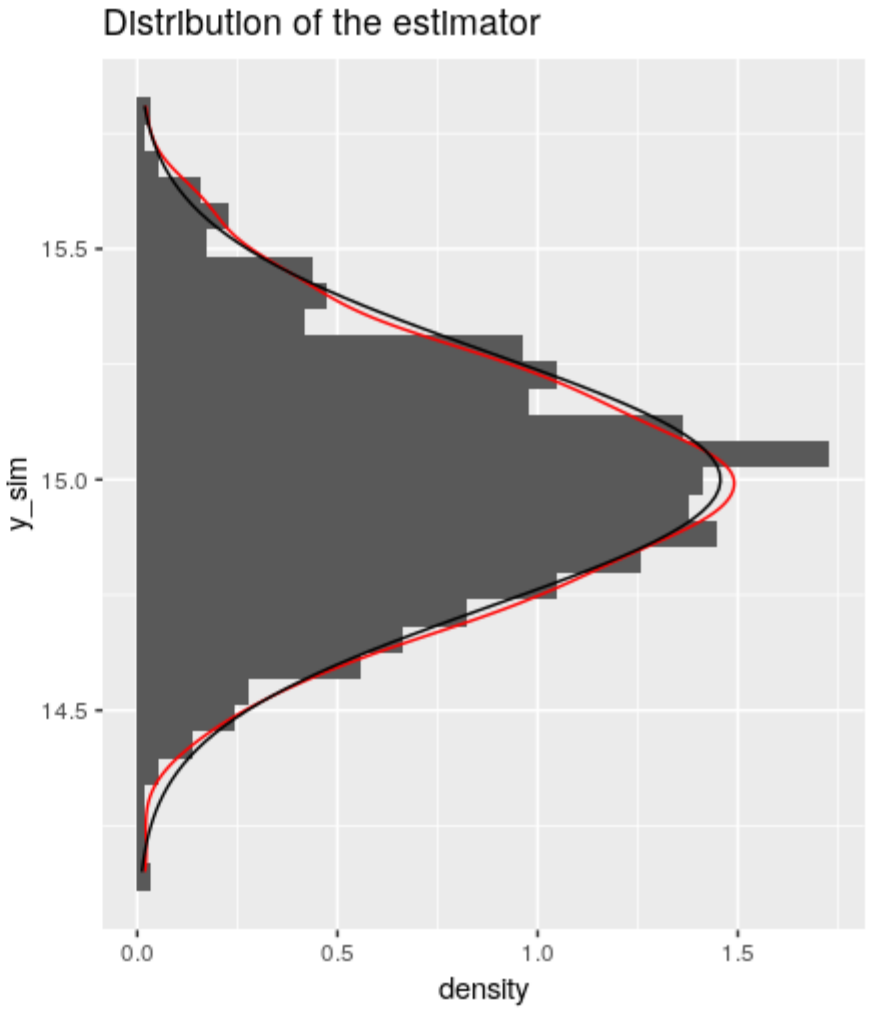


Simulation: Example 4

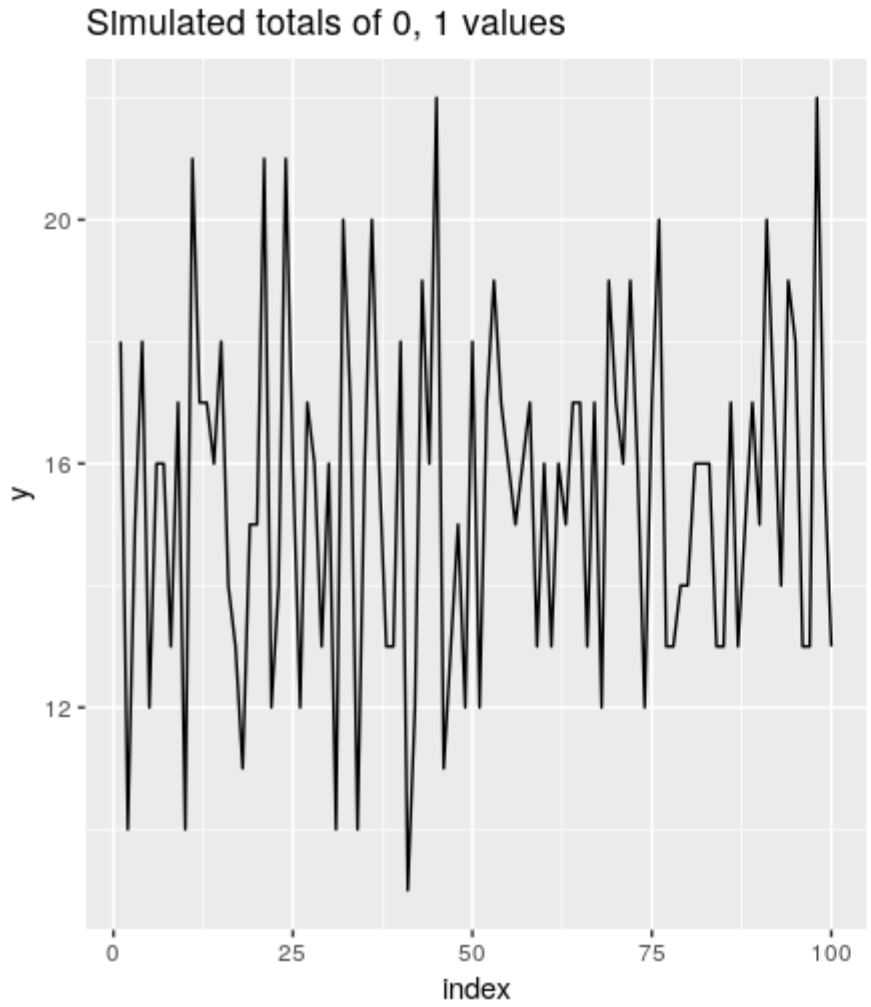


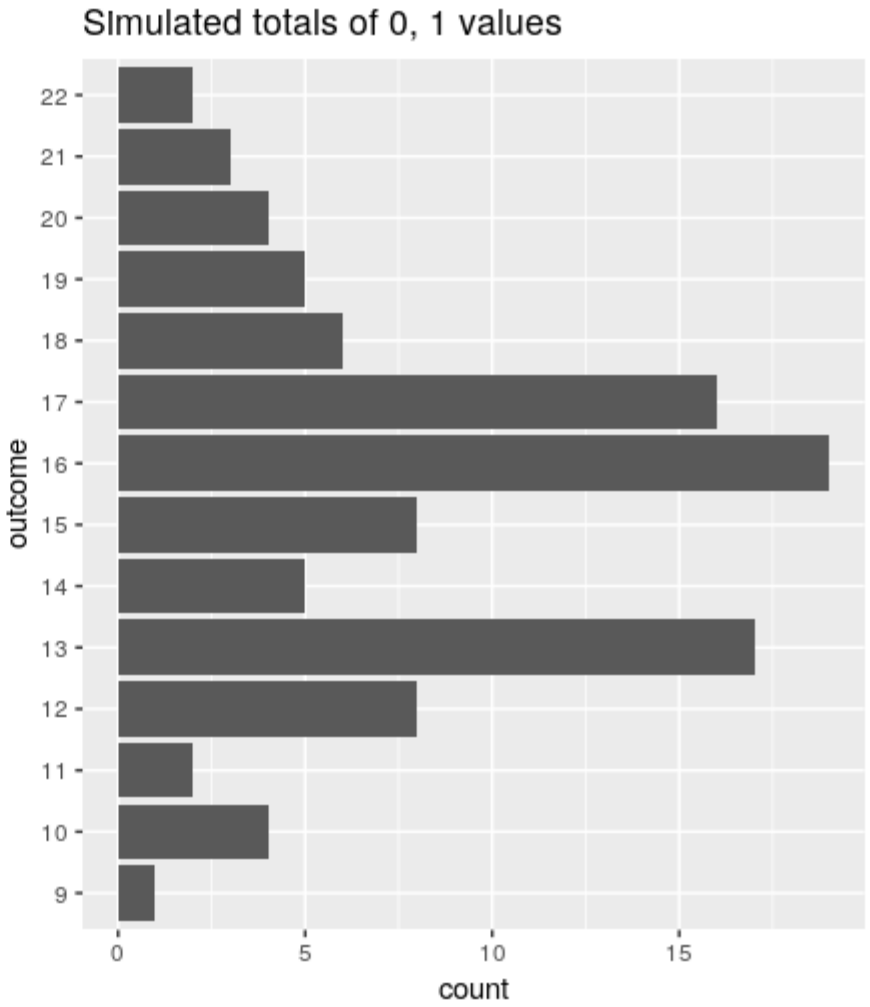
Simulation: Example 4



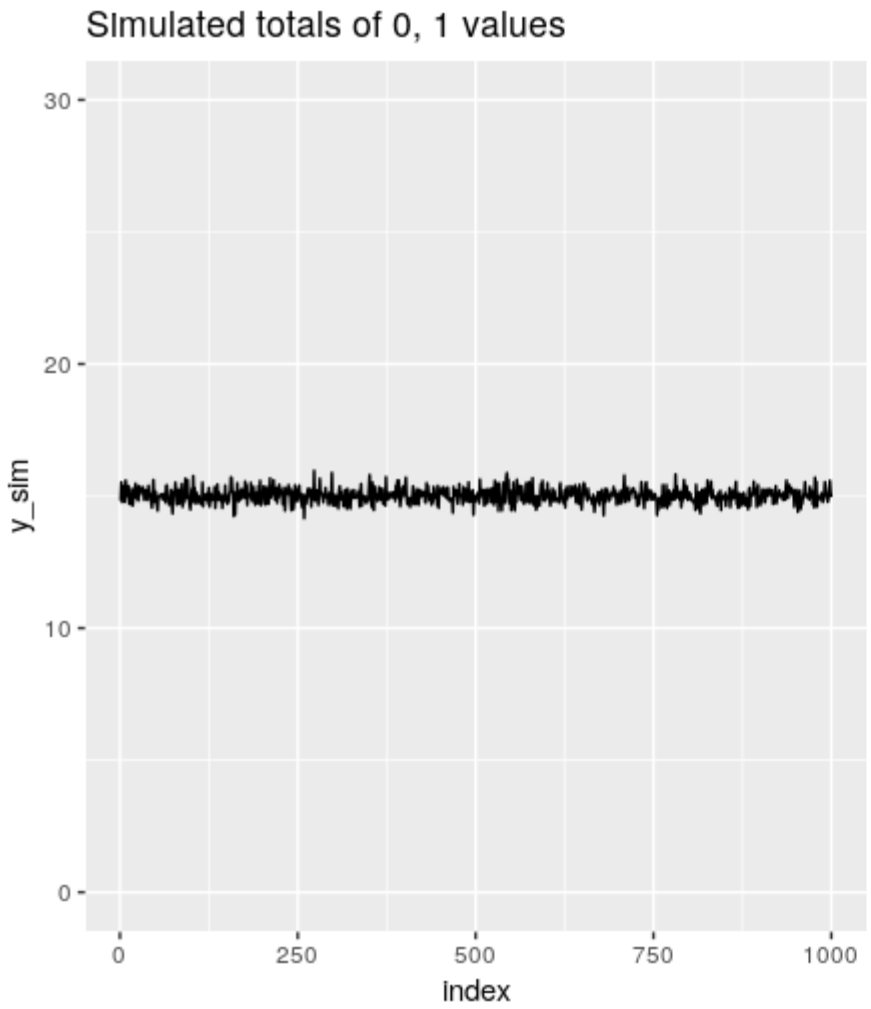


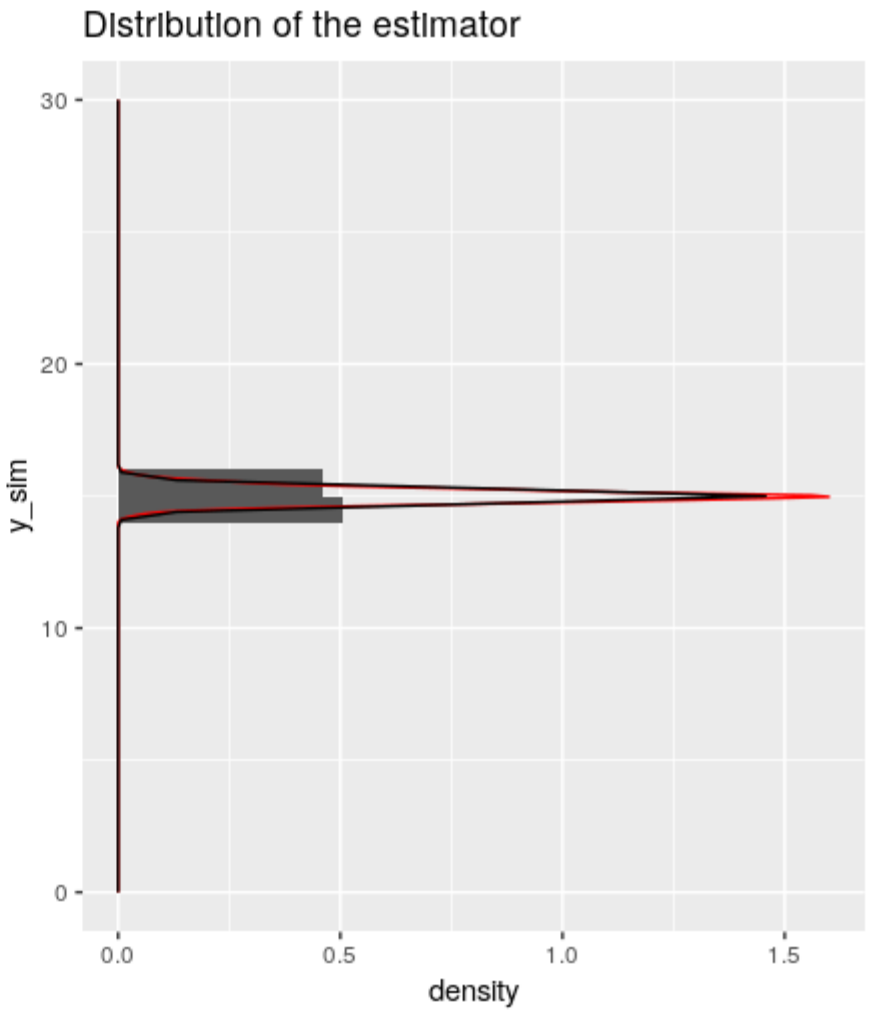
Simulation: Example 4





Simulation: Example 4





Simulation: Example 4

What have we learned?

- Simulated 30*100*1000 = 300,000 numbers.
- Simulated data has random patterns.
- Totals are normally distributed.
- Larger sample sizes lead to more accuracy.
- An estimator of a total is more accurate with a larger sample size.
- Axes in visualizations need to be the same for direct comparison.

Law of Large Numbers

- Replicating an experiment a larger number of time leads to more accurate estimates.
- Using less data may result in less accuracy of estimates.

$$\bar{x} = \frac{\sum x_i}{Reps} \rightarrow \mu$$

Central Limit Theorem

- The Distribution of an estimator will be normally distributed.
- For larger sample sizes.
- In the simulation the Reps needs to be large.

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{Reps}\right)$$

Statistics: The main point!

- Good estimators are unbiased or close to unbiased.
- Good estimators have improved accuracy with larger sample sizes.

Bayesian Statistics

- Bayesian Statistics includes the use of prior information, summarized in a prior distribution, along with the data to produce posterior distributions.

- *Prior information* can be based on the current knowledge about a topic of interest. $P(\theta)$

- *Posterior distribution(s)* represent the updated distribution which includes the prior and the data. $P(Y|\theta)$

Bayesian Statistics

Bayes Theorem

$$P(\theta|Y) = \frac{P(\theta)P(Y|\theta)}{P(\theta)} \propto P(\theta)P(Y|\theta)$$

Bayes rule can be used to update prior knowledge of a parameter of interest given observed data.

Bayes Theorem

- From Wikipedia Bayes Theorem
- Example: Medical testing
- Posterior probability of having a disease given a positive test result always seems lower that it should.

$$
\begin{aligned}
P(\text{User} \mid +) &= \frac{P(+ \mid \text{User})P(\text{User})}{P(+)} \\
&= \frac{P(+ \mid \text{User})P(\text{User})}{P(+ \mid \text{User})P(\text{User}) + P(+ \mid \text{Non-user})P(\text{Non-user})} \\
&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\
&\approx 33.2\%
\end{aligned}
$$

Bayesian Statistics

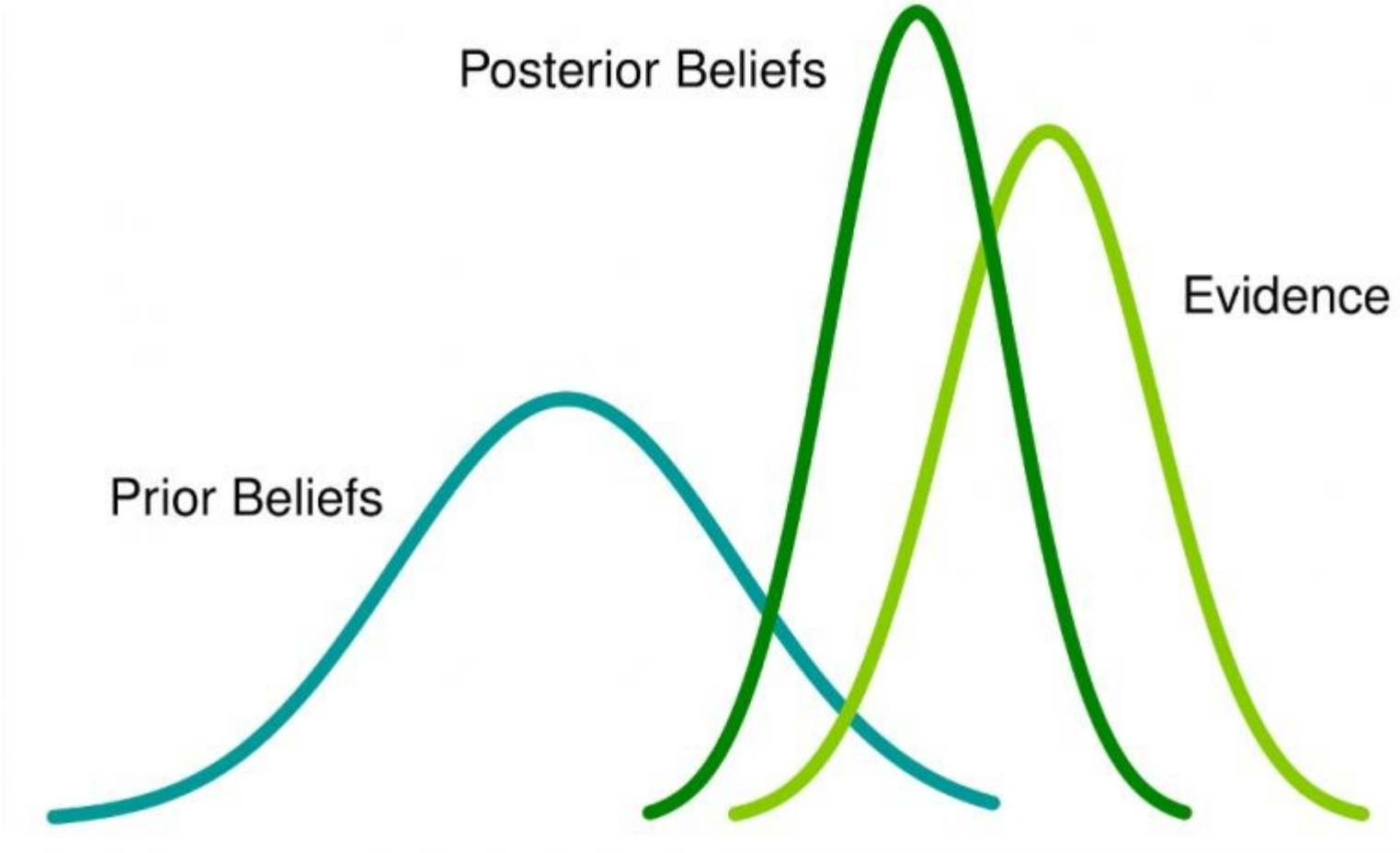
- Formula

$$P(\theta|y) \propto P(\theta) \times P(y|\theta)$$

- *Posterior* is proportional to the *Prior* x *Likelihood*
- To estimate the posterior Calculus can be used for simple problems.
- For higher dimensional problems computational optimization algorithms need to be used.
- One answer if MCMC using Gibbs Sampling.

Bayesian Statistics

- I appreciate Bayesian Statistics for it use of Bayes Theorem.
- Bayes Theorem updates prior opinions, represented by prior distributions, with current data to produce posterior distributions.
- The posterior distributions can be used as the current best opinions about the state of parameters of interest.



Analytics Vidhya:
Bayesian Statistics explained to Beginners in Simple English

数据挖掘 Data Mining

- The process of finding patterns in large databases. The goal is to find useful information for actionable decisions to be made. Transform data into usable structures that connects silos of data.

- KDD Knowledge Discovery in Databases
- Connected data tables within databases.
- Unsupervised Learning, ML

Technology Tools

- SQL
- Cloud Computing
- In-house Data Server

商业分析 Business Analytics

- The use of data and statistical methods to understand business processes and to develop new insights.

- Visual Analytics
- Descriptive Analytics
- Predictive Analytics, Supervised Learning, ML
- Prescriptive Analytics
- Customer Analytics
- Talent Analytics
- Preventative Maintenance

Technology Tools

- Knime
- RapidMiner

- Cloud Computing

- Amazon
- Google
- Microsoft
- IBM

- Alibaba
- Baidu
- Tencent
- UCloud

Businesses using Data

Business sectors that are using Statistics, Data Mining, Business Analysis.
Examples:

1) Health Care, Biostatistics
2) Manufacturing, Quality Control
3) Insurance, Actuarial Science
4) Finance, Risk Management

Case Studies

- Biostatistics
- Quality Control
- Actuarial Science
- Risk Management

Case Study: Biostatistics

- In the pharmaceutical industry clinical trials are used to determine the effectiveness of new medial treatments and drug therapies.
- *Designed Experiments* are used to collect data.
- Patients are *randomly assigned* to Treatment and Control groups to balance variability within the groups.
- *Hypothesis Testing* is used to test the effectiveness of the treatment compared to chance.

Case Study: Quality Control

- In the manufacturing industry *quality control methods* are used to monitor manufacturing processes to make sure they are producing products that meet the required specification.
- Control Charts can be used to watch for periods of production that are out of the required specification.
- Control Charts use *Confidence Intervals*.

Case Study: Actuarial Science

- As the Finance industry has grown, the insurance industry has grown.
- Valuing different types of insurance based on measured risk is a growing opportunity.
- Rates (or estimated probabilities) of disease, accidents, fire, flood, etc.

Case Study: Risk Management

- Related to insurance is the overall area of Risk Management.
- All financial institutions deal with Risk Management of investments.
- Banks are interested in risk management for making loans.
- Do crypto-currencies have risk?

Examples

- All of these examples collect 0 and 1 data.
- Estimates of proportions are calculated.
- Estimates of a total are calculated.
- Are the estimates accurate?
- **Simulations** can inform decision making.

第二天 Lets get started: Outline

1) Time Series Analysis
2) Change Point Analysis
3) Anomaly Detection
4) Linear Regression Analysis
5) Logistic Regression Analysis

Time Series Analysis and Forecasting

- Additive Models

$$Y_t = T_t + S_t + I_t$$

- Multiplicative Models

$$Y_t = T_t * S_t * I_t$$

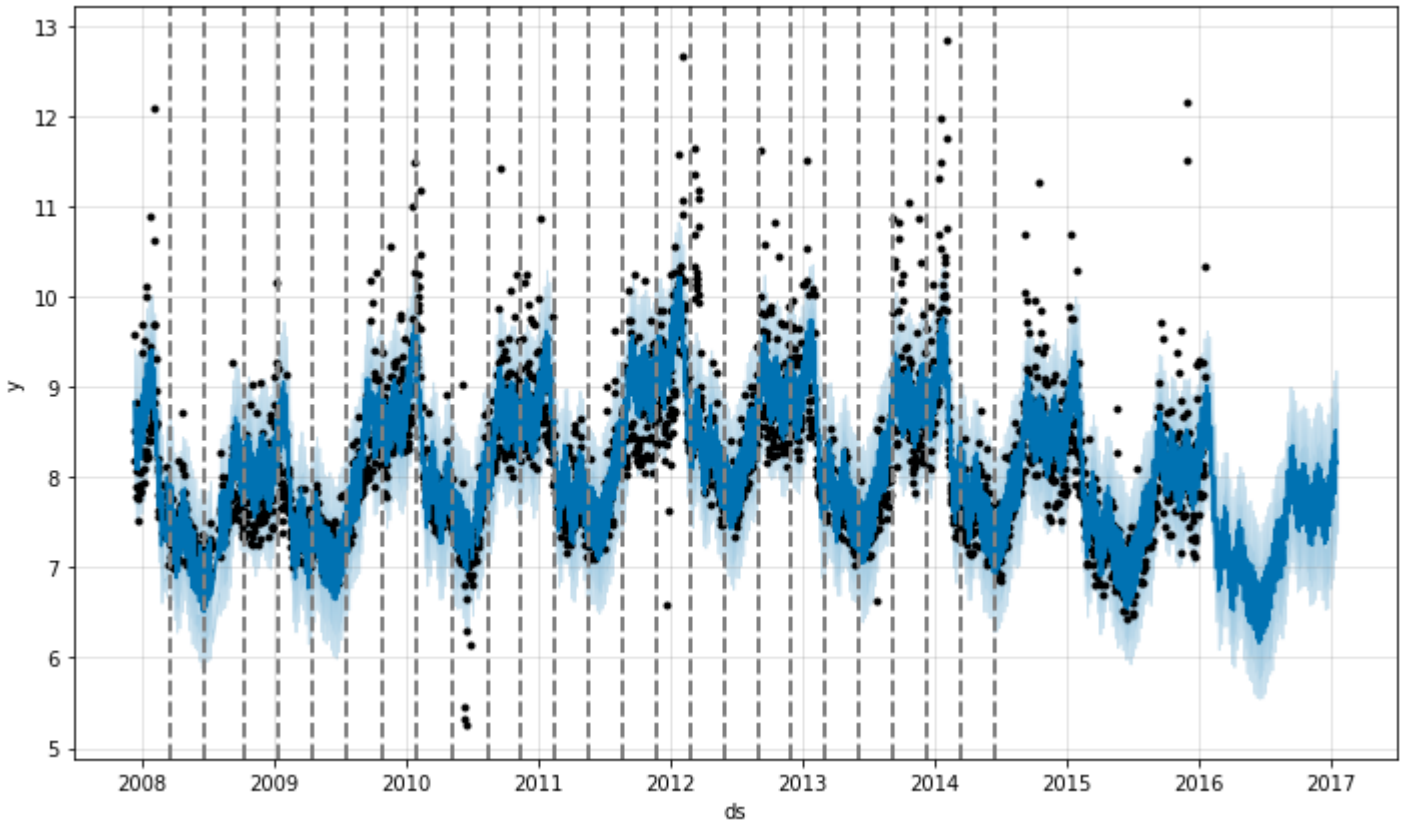
- Trends, Seasonality, Error
- Visualization
- Yahoo's open source Prophet package for Python and R
- Prediction and Forecasting

Example: Yahoo Prophet

- **Example:** The example from the Prophet website is for a time series of the log of the daily views for the Wikipedia page for Peyton Manning. He is a famous Football player in the United States.
- He has been to the Super Bowl and the playoffs many times over his career.

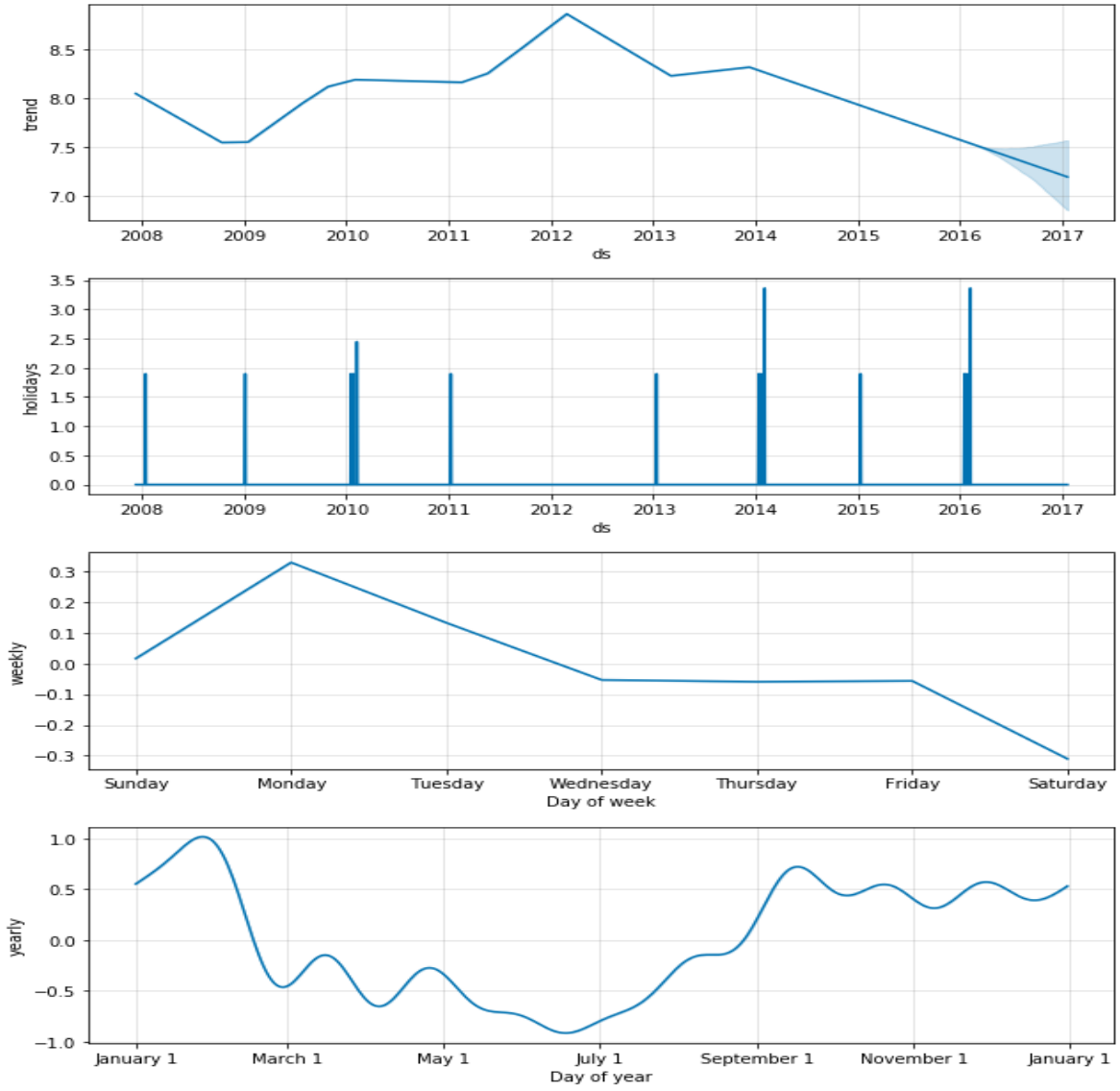
Change Point Analysis

Automatic changepoint detection in Prophet



Anomaly Detection

Modeling Holidays and Special Events



Regression and Logistic Regression

- Linear Regression, numeric dependent variable

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Logistic Regression, binary dependent variable

$$logit\left(p_i\right) = \beta_0 + \beta_1 x_i + \epsilon_i$$

Case Study: Regression

- **Example:** Suppose we are developing a model to predict the strength of concrete based on the amounts of various inputs.
- Assuming the input measurement are also numeric then a linear model and be used to predict the strength of concrete.
- A *Linear Regression* model could be used.
- Alternatively, *CART* or *Random Forests*.

Case Study:Logistic Regression

- **Example:** Suppose the failure of a type of engine water pump is to be modeled in terms of other input variables.
- Assume there are input variables, then the response variable, failure, could be modeled using *Logistic Regression*.

机器学习 Machine Learning

What is Machine Learning (ML)?

- Holdout Method
- Supervised Learning
  - Regression, Classification
- Unsupervised Learning
  - Clustering
- Deep Learning

机器学习 Machine Learning

- According to Wikipedia, “Machine Learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.”

Holdout Method

- Training Data
- Validation Data
- Testing Data
- Training and Testing accuracy
- Over-fitting
- Balance between Variance and Bias

预测 Prediction

- Machine Learning algorithms can be used to make numeric prediction and forecasts.
- Example: **Predict** concrete strength after a period of time.
- **Linear Regression** is the most basic algorithm.
- *Decision Trees* and *Random Forests* are commonly used.
- *Neural Networks* and *Deep Learning* can be used.

类别 Classification

- Machine Learning algorithms can be used to make predictions of classes.
- Example: **Classification** of patients as having a disease given various clinical measures.
- **Logistic Regression** is the most basic algorithm.
- *Decision Trees* and *Random Forests* are commonly used.
- *Neural Networks* and *Deep Learning* can be used.

Types of ML

Unsupervised Learning

- Clustering
- Big Data and Principle Component Analysis
- Market Basket Analysis

Supervised Learning

- Machine Learning (ML)
- Artificial Intelligence (AI)
- Streaming data

Unsupervised Learning

- Clustering
- PCA
- Market Basket Analysis

Clustering

- *Unsupervised Machine Learning algorithms* for discovering segments or groups in data.
- Or used to subset dataset into target groups.
- Distance

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

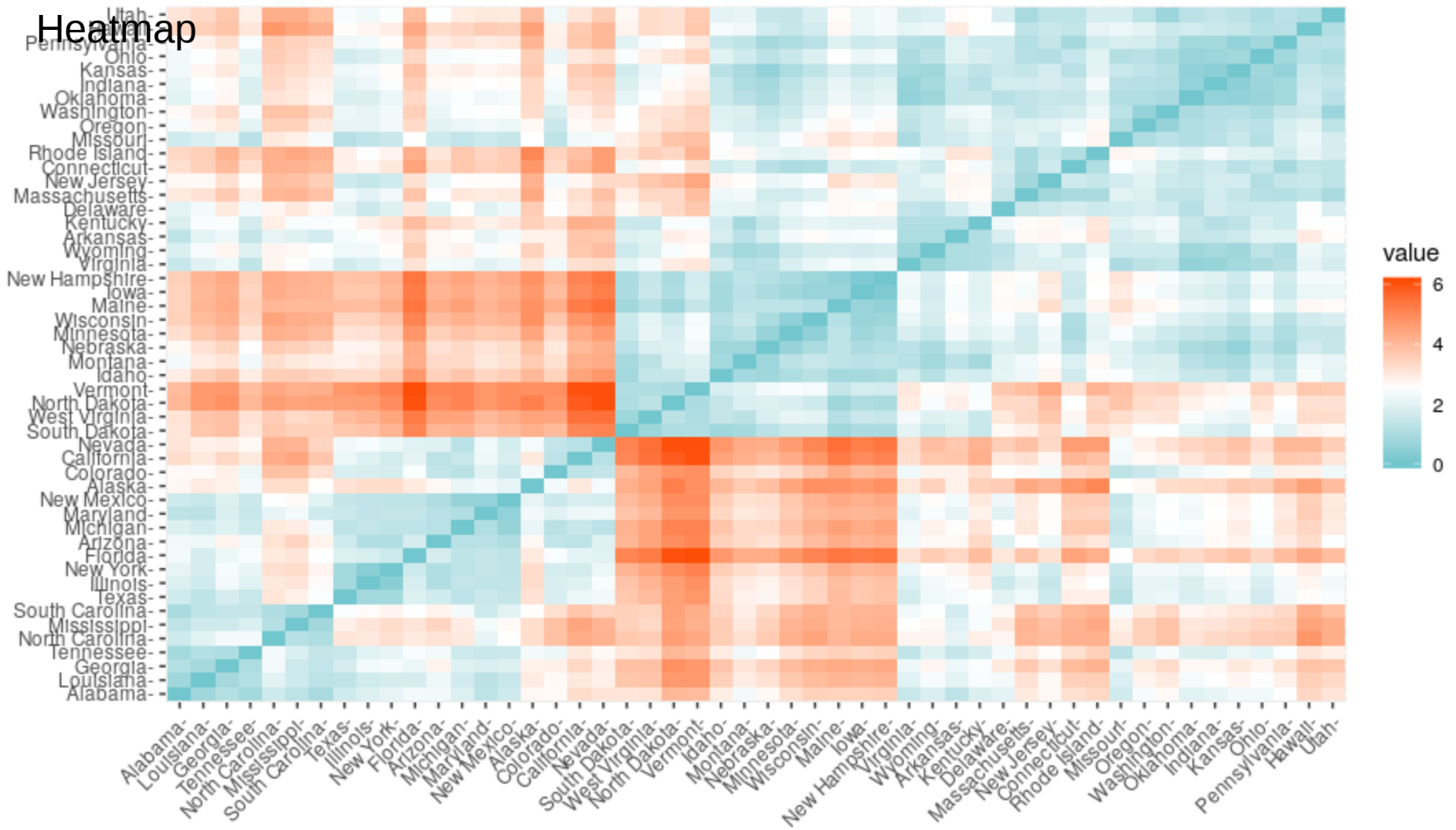
- Works with numeric features.
- Nice reference: UC Business Analytics

Clustering

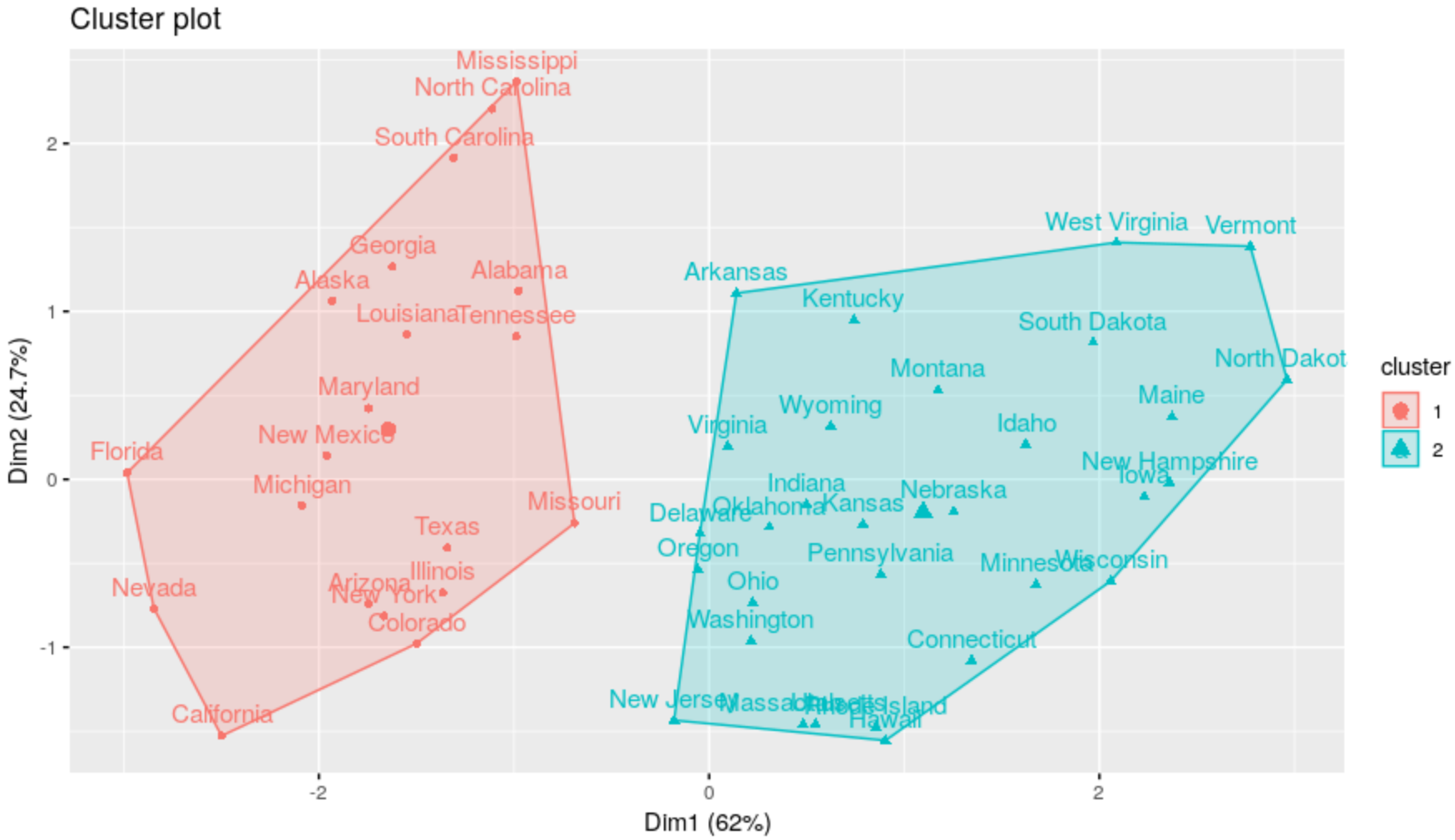
- Example: Differences in crime by region in the US.
- To use k-Means the data needs to be
  - Cleaned, remove NAs
  - Scaled, center at zero and scale to standard deviation 1.
  - Select the number of clusters.

Clustering

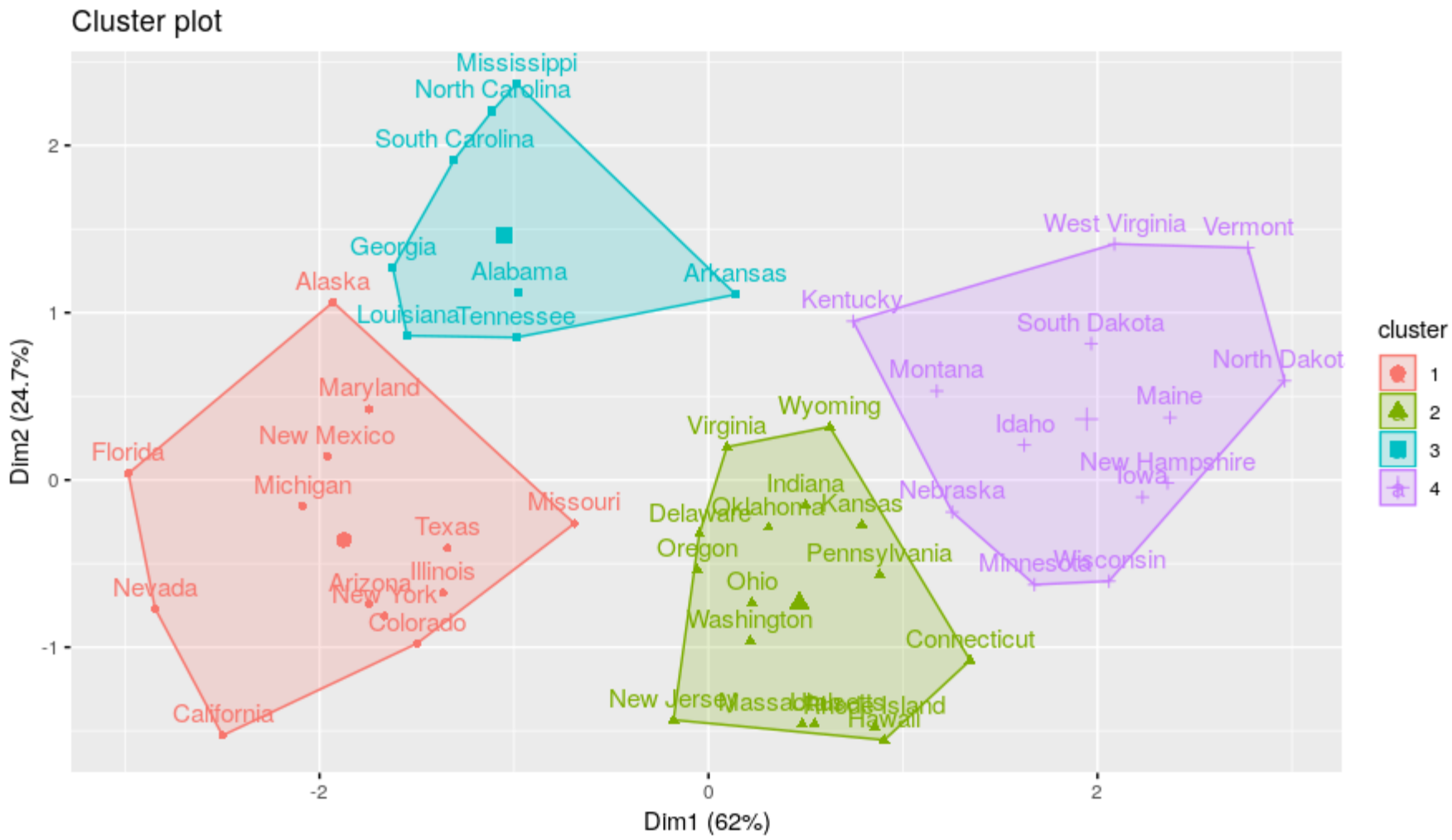
Heatmap



Clustering


Cluster plot
Mississippi
North Carolina
South Carolina
Georgia
Alaska
Alabama
Louisiana
Tennessee
Maryland
New Mexico
Florida
Michigan
Missouri
Texas
Illinois
Nevada
Arizona
New York
Colorado
California
West Virginia
Vermont
Arkansas
Kentucky
South Dakota
North Dakot
Montana
Maine
Wyoming
Virginia
Idaho
New Hampshire
Iowa
Indiana
Nebraska
Delaware
Oklahoma
Kansas
Oregon
Pennsylvania
Minnesota
Wisconsin
Ohio
Washington
Connecticut
New Jersey
Hawaii
cluster
1
2
Dim2 (24.7%)
Dim1 (62%)
2
1
0
-1
-2
0
2

Clustering


Cluster plot
Mississippi
North Carolina
South Carolina
Georgia
Alabama
Arkansas
Alaska
Louisiana
Tennessee
Maryland
New Mexico
Florida
Michigan
Missouri
Texas
Illinois
Nevada
Arizona
New York
Colorado
California
Wyoming
Virginia
Indiana
Delaware
Oklahoma
Kansas
Oregon
Pennsylvania
Ohio
Washington
Connecticut
New Jersey
Hawaii
West Virginia
Vermont
Kentucky
South Dakota
North Dakot
Montana
Maine
Idaho
New Hampshire
Iowa
Nebraska
Minnesota
Wisconsin
cluster
1
2
3
4
Dim2 (24.7%)
Dim1 (62%)
2
1
0
-1
-2
0
2

Big Data, Principle Components

- *Unsupervised Machine Learning algorithms* for reducing the dimension of large collections of features/variables in data.
- Used to make complex datasets more manageable in terms for a smaller set of useful dimensions.
- Eigenvectors, Eigenvalues, Singular Value Decomposition, Factor Analysis
- Also, works with numeric features.

Market Basket Analysis

- *Unsupervised Machine Learning algorithms* for discovering co-purchased items from sparse transaction data.
- Transaction data contains the receipts of customers.
- Many products are for sale and only a few purchased at a time.
- What products are purchased at the same time?
- Coupons, coupons, coupons!

Supervised Learning

- Machine Learning (ML)
- Artificial Intelligence (AI)
- Streaming data

机器学习 Machine Learning

- Natural Language Processing (NLP) and Deep Learning (DL) have become two of the most popular areas in Machine Learning in recent years.
- Two **Unsupervised Learning** algorithms in Natural Language Processing are *Sentiment Analysis* and *Topic Modeling*.
- *Decision Trees* are a **Supervised Learning** algorithm. *Random Forests* are an extension of Decision Trees.
- **Deep Learning** can be used for Supervised Learning, Classification and Prediction. It has been used extensively for Computer Vision problems.
- Deep Learning can be used for Generative applications also.

Technology Tools

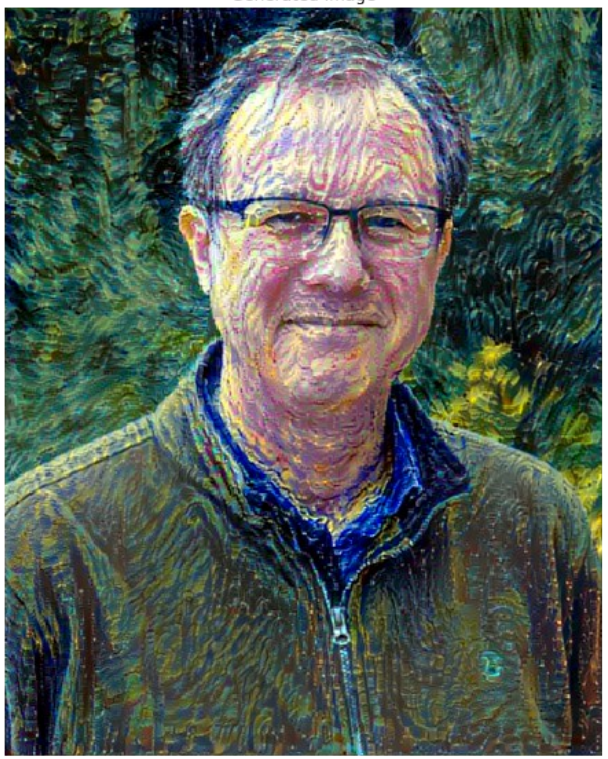
- Various US companies have build their own Deep Learning platforms for fitting deep learning algorithms and made them available in public Cloud computing platforms.
- Google Tensorflow
- Yahoo PyTorch
- Microsoft MxNet
- Amazon and h2O
- Baidu PaddlePaddle

Deep Learning Application Generative Art

- Deep Learning algorithms can used for *generative purposes*, not for prediction.
- One fun application is Style Transfer.
- My photo with Van Gogh’s Starry Night as the style.

Content image

Generated image



Style image

Deep Learning Application Classification

- Simulated data for Logistic Regression.
- Keras sequential model tutorial
- By Pablo Casas

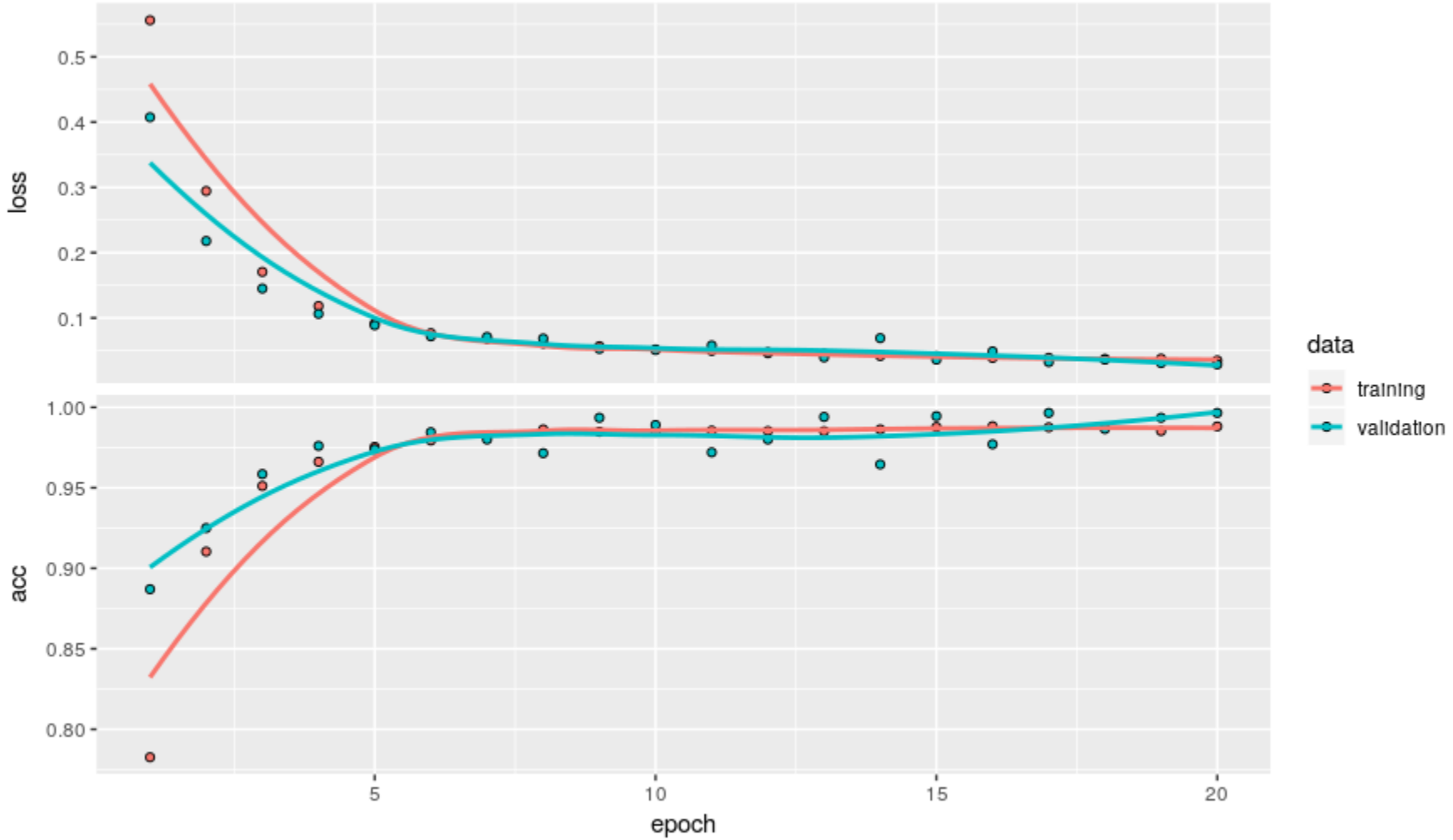
Deep Learning Application Classification

- Simulated data for Logistic Regression.
- By Pablo Casas
- Simulate 3 numeric X variables
- Determine 1 categorical Y variable with two classes 0 or 1
- Build a multilayer Neural Network

Deep Learning Application Classification

```
> model = keras_model_sequential() %>%

  layer_dense(units = 64, activation = "relu", input_shape = ncol(x_data)) %>%

  layer_dense(units = 64, activation = "relu") %>%

  layer_dense(units = ncol(y_data_oneh), activation = "softmax")

> compile(model, loss = "categorical_crossentropy", optimizer = optimizer_rmsprop(), metrics = "accuracy")

> history = fit(model,  x_data, y_data_oneh, epochs = 20, batch_size = 128, validation_split = 0.2)
```

Deep Learning Application Classification



Deep Learning Application Classification

```
> ## Evaluation on training data
> evaluate(model, x_data, y_data_oneh, verbose = 0)
$loss
[1] 0.0295124059
$acc
[1] 0.9959
> ##
> ## Evaluation on Test data (we need the one-hot version)
> evaluate(model, x_data_test, y_data_real_oneh, verbose = 0)
$loss
[1] 0.02922237225
$acc
[1] 0.992
```

Deep Learning Application Classification

- What have we learned?
- Neural Networks can be used and can be fitted easily.
- Neural Networks can be fit to very large networks this is Deep Learning.

仿制 Artificial Intelligence

- Self-driving cars, buses, trucks
- Drones
- They use ML algorithms to automatically make decisions.

Steaming data

- With all of the sensors, video cameras, there is a lot of potential of use of the data.
- **Tensors:** Vectors, Arrays, Picture Collections, Video Collections.
- With Cloud computing the collection of such data is going on very quickly.
- There are many ML algorithms that can be run on the streams of data.

Handout

- Go over the handout.
- Possible Machine Learning algorithms to consider.
- Possible job descriptions to recruit new talent to work on the solutions to these problems.

Future Students

- If any of you have high school age children, then you might suggest Data Science and Statistics as a college major. This is a growing field with good job prospects!
- They can email me directly about majoring in Statistics at CSU East Bay. And I would certainly recommend UC Berkeley, UC Davis, and Stanford University Statistics Department.
- Or for graduate school, consider an MS degree in Data Science and Statistics at CSU East Bay .

Connect

- Email: esuess100@yahoo.com
- Email: esuess@gmail.com
- Website: www.sci.csueastbay.edu/~esuess/

Connect on WeChat