CS544 Module1

Suresh Kalathur

Course Outline

- Module1
  - Review basics in statistics and probability
  - R - Data types and structures
- Module2
  - Probability, Random variables, R – Programming constructs
- Module3
  - Data – Univariate, Bivariate, Multivariate
- Module4
  - Distributions – Discrete, Continuous
- Module5
  - Central Limit Theorem, Sampling, Errors
- Module6
  - Confidence intervals, hypothesis testing

Grading

- Programming assignments – 30%
- Midterm Exam – 20%
- Project – 20%
- Final Exam – 20%
- Class Participation and Quizzes – 10%

Statistics

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Measures of Variation
  - Range
  - Standard deviation
  - Inter-quartile range

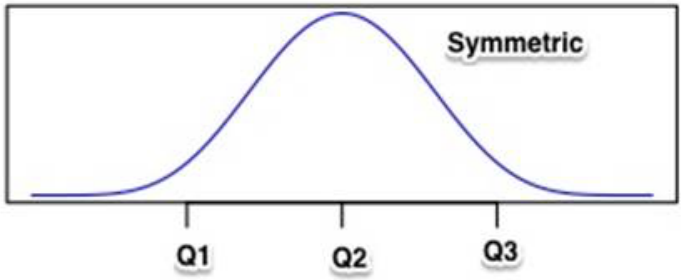
- Percentiles
  - Divide data into 100 equal parts
  - http://money.cnn.com/calculator/pf/income-rank/
- Quartiles
  - Divide data into 4 equal parts
    - Q1 – bottom 25% from the top 75%
    - Q2 – bottom 50% from the top 50% (Median)
    - Q3 – bottom 75% from the top 25%

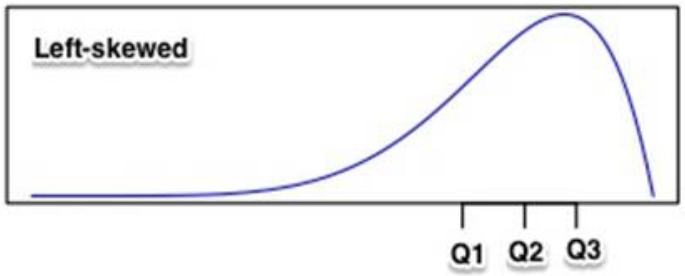
- IQR – Inter Quartile Range
  - Q3 – Q1
- Five Number Summary
  - Min, Q1, Q2, Q3, Max
- Variations in each quarter
- Outliers
  - Outside the range
    - (Q1 – 1.5*IQR, Q3 + 1.5*IQR)
    - (Mean – 3*SD, Mean + 3*SD)

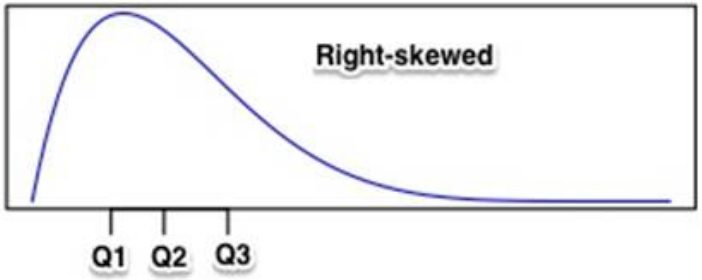
- Population versus Sample
- Standardized Variables
  - Mean 0 and Standard Deviation 1
  - *z*-score for variables
    - Negative score – below the mean
      - How many SD below the mean
    - Positive score – above the mean
      - How many SD above the mean
  - Most values in the range -3 to 3
    - Otherwise, outliers

Shape of Data

- Distribution of the data
  - Symmetric
    - Mean and median are the same
  - Left-skewed (negatively skewed)
    - An easy quiz/exam
    - Mean is less than the median
  - Right-skewed (positively skewed)
    - A hard quiz/exam
    - Mean is greater than the median







Probability

- Events
  - Chance that a particular event will occur
- Based on prior knowledge
  - Priori probabilities
- Based on observed data
  - Empirical probabilities
- Sample Space
  - Collection of all possible outcomes
- Conditional probability

R

- A language and environment for statistical computing and graphics
- GNU General Public License
- Initially written by Robert Gentleman and Ross Ihaka (University of Auckland)
- http://www.r-project.org
- Base version of R
- Rstudio
- Jupyter Notebooks

R

- Statistical techniques
  - Linear and nonlinear modeling
  - Classical statistical tests
  - Time-series analysis
  - Classification
  - Clustering, …
- Graphical techniques

RStudio

Go to file/function

Project: (None)

DemoTalk.R

Source on Save

Run

Source

1

1:1 (Top Level) R Script

Environment History

Import Dataset Clear List

Global Environment

Environment is empty

Console ~/

```
s in publications.

Type 'demo()' for some demos, 'help()' for
on-line help, or
'help.start()' for an HTML browser interfa
ce to help.
Type 'q()' to quit R.

>
```

Files Plots Packages Help Viewer

The R Language Find in Topic

Statistical Data Analysis

Manuals

An Introduction to R
Writing R Extensions
R Data Import/Export
The R Language Definition
R Installation and Administration
R Internals

Data in R

- *Data types frequently used in R*
  - numeric
  - integer
  - logical
  - character
  - complex

...Data in R

- *Data structures*
  - *vector* – a collection of values of the same type
  - *factor* – a collection of values from a fixed set of possible values
  - *matrix* – a two-dimensional collection of values of the same type
  - *list* – a collection of any of the data structures
  - *data frame* – a collection of vectors all of the same length

...R

Reading a CSV file

```
athlete.info <- read.csv(
  "http://kalathur.com/athletedata.csv",
  header = TRUE)
```

```
> athlete.info
        Name Salary Endorsements      Sport
1 Mayweather  105.0             0     Boxing
2    Ronaldo   52.0            28     Soccer
3      James   19.3            53 Basketball
4      Messi   41.7            23     Soccer
5     Bryant   30.5            31 Basketball
```

...R

Read as a data frame

Access column information

```
> athlete.info$Salary
[1] 105.0  52.0  19.3  41.7  30.5
> athlete.info$Sport
[1] Boxing     Soccer     Basketball Soccer     Basketball
Levels: Basketball Boxing Soccer
```

...R

Summary of data frame columns

```
> summary(athlete.info$Salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   19.3    30.5    41.7    49.7    52.0   105.0
> summary(athlete.info$Sport)
Basketball     Boxing     Soccer
         2          1          2
```

...R

Slicing columns

```
> athlete.info[c("Name", "Sport")]
        Name      Sport
1 Mayweather     Boxing
2    Ronaldo     Soccer
3      James Basketball
4      Messi     Soccer
5     Bryant Basketball
```

...R

Slicing rows

```
> athlete.info[c(2,4), ]
     Name Salary Endorsements  Sport
2 Ronaldo   52.0           28 Soccer
4   Messi   41.7           23 Soccer
```

```
> athlete.info[athlete.info$Sport == "Soccer", ]
     Name Salary Endorsements  Sport
2 Ronaldo   52.0           28 Soccer
4   Messi   41.7           23 Soccer
```

...R

Slicing rows and columns

```
> athlete.info[c(2,4), c(1,2)]
     Name Salary
2 Ronaldo   52.0
4   Messi   41.7
```

```
> athlete.info[athlete.info$Sport == "Soccer",
+              c("Name","Salary")]
     Name Salary
2 Ronaldo   52.0
4   Messi   41.7
```

...R

Subset of a data frame

```
> subset(athlete.info, Sport == "Soccer")
     Name Salary Endorsements  Sport
2 Ronaldo   52.0           28 Soccer
4   Messi   41.7           23 Soccer
```

```
> subset(athlete.info,
+        Sport == "Soccer" & Salary > 50)
     Name Salary Endorsements  Sport
2 Ronaldo     52           28 Soccer
```

```
> subset(athlete.info, Sport == "Soccer",
+        select = c(Name, Salary))
     Name Salary
2 Ronaldo   52.0
4   Messi   41.7
```

...R

Modifying the data frame

```
> athlete.info$Pay <-
+     athlete.info$Salary + athlete.info$Endorsements
>
> athlete.info
       Name Salary Endorsements      Sport   Pay
1 Mayweather  105.0             0     Boxing 105.0
2    Ronaldo   52.0            28     Soccer  80.0
3      James   19.3            53 Basketball  72.3
4      Messi   41.7            23     Soccer  64.7
5     Bryant   30.5            31 Basketball  61.5
```