Basic Methods

# Module 3 – Review of Basic Data Analytic Methods Using Python




1

Introduction | Analytics Lifecycle | Basic Methods | Adv. Methods | Tools | Lab

## Module 3: Review of Basic Data Analytic Methods Using Python

Upon completion of this module, you should be able to:

- Use basic analytics methods such as distributions, statistical tests and summary operations to investigate a data set.
- Use R as a tool to perform basic data analytics, reporting and basic data visualization.




2

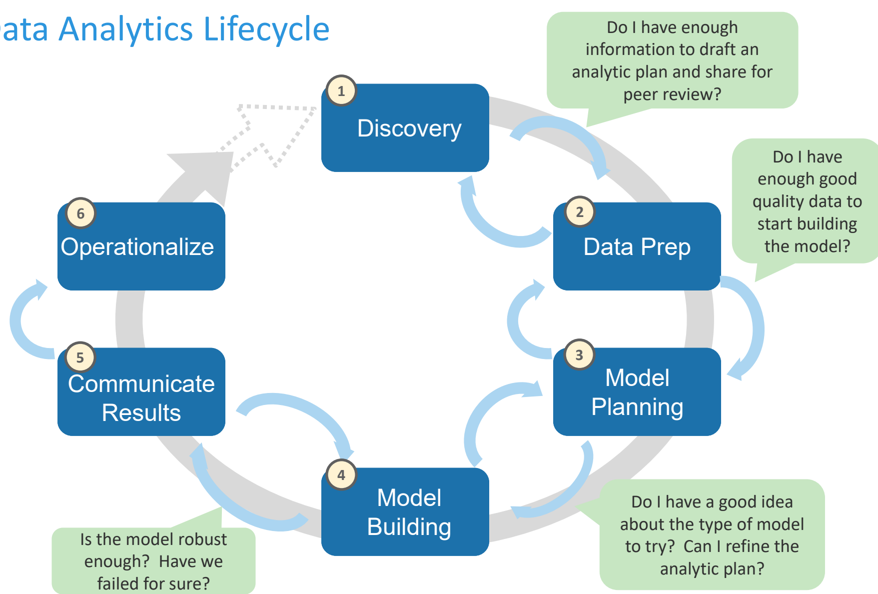
## Putting the Data Analytics Lifecycle into Practice

- From Module 2 you learned a strategy to approach any data analytics problem:
    - **Phase 1: Discovery**
    - **Phase 2: Data Preparation**
    - **Phase 3: Model Planning** *(covered in Module 4)*
    - Phase 4: Model Building
    - Phase 5: Communicate Results
    - Phase 6: Operationalize
- To begin to analyze the data you need:
    - 1. A tool that allows you to look at the data – that is “Python or R”.
    - 2. Skill in basic statistics – we’re providing a refresher.




3

## Data Analytics Lifecycle







4

Introduction | Analytics Lifecycle | **Basic Methods** | Adv. Methods | Tools | Lab

## Module 3: Review of Basic Data Analytic Methods Using R

### Lesson 1: Using R to Look at Data – Introduction to R

During this lesson the following topics are covered:

- Using the R Graphical User Interface
- Overview: Getting Data into (and out of) R
- Data Types Used in R
- Basic R Operations
- Basic Statistics
- Generic Functions

GETTING A HANDLE ON THE DATA



 Module 3: Basic Data Analytic Methods Using R 5

5

Introduction | Analytics Lifecycle | **Basic Methods** | Adv. Methods | Tools | Lab

## Module 3: Review of Basic Data Analytic Methods Using R

### Lesson 1: Using R to Look at Data – Introduction to R

R, also called the Language for Statistical Computing:

- Developed **Ross Ihaka** and **Robert Gentleman** at the University of Auckland in the nineties.
- Open Source Implementation of the S Language, so GNU is S.
- S Language was developed by John Chambers in the Bell Laboratories in the eighties.
- R provides a wide variety of statistical techniques and visualization capabilities.



 Module 3: Basic Data Analytic Methods Using R 6

6

Introduction | Analytics Lifecycle | Basic Methods | Adv. Methods | Tools | Lab

## Module 3: Review of Basic Data Analytic Methods Using R

### Lesson 1: Using R to Look at Data – Introduction to R

- Another very important feature about R is that it is highly extensible. Because R is open source.
- It actually was the vehicle to bring the power of S language to a larger community.
- Like every programming languages, there are pros and cons.



 Module 3: Basic Data Analytic Methods Using R 7

7

Introduction | Analytics Lifecycle | Basic Methods | Adv. Methods | Tools | Lab

Module 3: Review of Basic Data Analytic Methods Using R

Lesson 1: Using R to Look at Data – Introduction to R

Start with Pros first ,

- It is open source, so it is free.
- R's graphical capabilities are top notch and it is very easy to build publications quality plots.
- In comparison to many other statistical software packages, R uses a command R uses a command line interface, which means that you have to actually code things in your console and in scripts.
- While this might be frustrating at first, it makes your work reproducible. You can now wrap your work in R scripts, which you can then easily share with your colleagues. Because of these advantages, R appeals to a large audience both in academia and in business.



 Module 3: Basic Data Analytic Methods Using R 8

8

Introduction | Analytics Lifecycle | Basic Methods | Adv. Methods | Tools | Lab

## Module 3: Review of Basic Data Analytic Methods Using R

### Lesson 1: Using R to Look at Data – Introduction to R

Pros (Cont.) ,

- Moreover, it's fairly easy to create R packages, which are extensions of R, aimed at solving particular problems.
- R's very active community has created thousands of well-documented R packages for a very broad range of applications in the financial sector, health care and for cutting edge research.



 Module 3: Basic Data Analytic Methods Using R 9

9

Introduction | Analytics Lifecycle | Basic Methods | Adv. Methods | Tools | Lab

## Module 3: Review of Basic Data Analytic Methods Using R

### Lesson 1: Using R to Look at Data – Introduction to R

However, as with anything, there are also some disadvantages,

- R seems to be relatively easy to learn at first, but it is hard to really master it.
- Also the fact that R is command-based is a frightening detail for statisticians that are used to the typical point and click programs out there.
- This steep learning curve sometimes results in poorly written R code that can be hard, both to read and to maintain.
- Furthermore, poorly written R code can become slow if you're working with large data sets.



 Module 3: Basic Data Analytic Methods Using R 10

10

Module 3: Review of Basic Data Analytic Methods Using R

## Lesson 1: Using R to Look at Data – Introduction to R

However, as with anything, there are also some disadvantages,

- But fear not! This course is here to help you master R in no time!
- Use this course to get a grip on R's fundamental concepts, and you can always consult R Documentation for documentation on all publicly available R packages.
- Next up: your first steps in R!



11

## Five Things to Remember About R

1. (Almost) everything is a *object*
2. (Almost ) everything is a *vector*
   - Example: `x <- 3` -- `x` is a vector of length 1
     `v <- c(2,4,6,8,10)` -- `v` is a vector of length 5
3. All commands are functions
   - Example: `quit()` or `q()`, not q
4. Some commands produce different output depending…
5. Know your default arguments!



12

## Five Things to Remember About R (Continued)

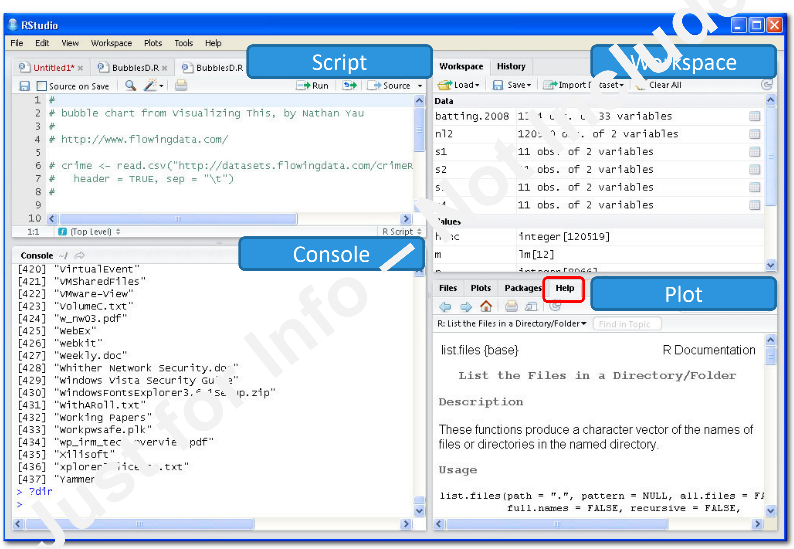
1. (Almost) everything is a *object*
2. (Almost ) everything is a *vector*
   - Example: `a <- 3 ---` *a* is a 1x1 vector, `v <- c(1,2,3,4,5)` is a 1x5 vector
3. All commands are functions
   - Example: `quit()` or `q()`, not q
4. Some commands produce different output depending…
5. Know your default arguments!






13

## Using the RStudio Graphical User Interface









14

## Overview: Getting Data Into (and Out of) R

- Getting Data Into R
  - Type it in (if it's small)!
  - Read from a data file
  - Read from a database
- Getting Data Out of R
  - Save in a workspace
  - Write a text file
  - Save an object to the file system
  - You can save plots as well!

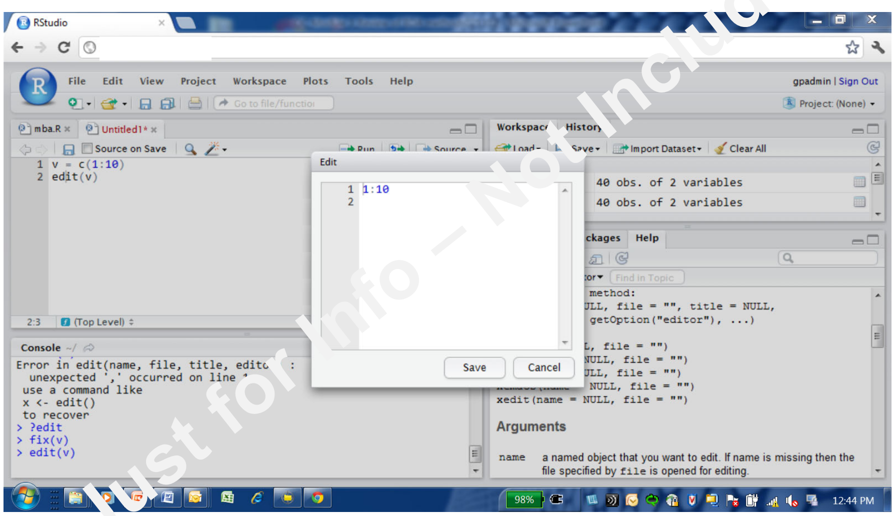








## Typing Data Into R



## Getting Data Into R: External Sources

- R supports multiple file formats
  - `read.table()` is the main function
- File name can be a URL
  - `read.table("http://ahost/file.csv", sep=",")` is the same as `read.csv(…)`
- Can read directly from a database via ODBC interface
  - `mydb <- odbcConnect("MyPostgresDB", …)`
- R packages exist to read data from Hadoop or HDFS (more later)

**Note! R always uses the forward-slash ("/") character in full file names "C:/users/janedoe/My Documents/Script.R"**



 Module 3: Basic Data Analytic Methods Using R 17

17

## Getting Data Out of R

| Options | R Code |
|---|---|
| Save it as part of your workspace (or a different workspace) | `save.image(file="dfm.Rdata")`<br>`save.image() # a .Rdata file`<br>`load.image("dfm.Rdata")` |
| Save it as a data file | `write.csv(dfm, file="dfm.csv")` |
| Save it as an R object | `save(Mydata,`<br>`     file="Mydata.Rdata")`<br>`load(file="Mydata.Rdata")` |
| Plots can be saved as images | `saveplot(filename="`*`filename.ext`*`",`<br>`    type="`*`type`*`")` |



 Module 3: Basic Data Analytic Methods Using R 18

18

## Data Classification: A Quick Review

| Data "Noir" | Examples |
|---|---|
| **N**ominal | condo, house, rental |
| **O**rdinal | hates < dislikes <neutral < likes < loves |
| **I**nterval | 10F colder tomorrow than today |
| **R**atio | 5342 > 4321 |

Some statistical tests require data at the interval level or higher. Other tests assume ordinal or nominal. Make sure you check.








## Data Types Used in R

| Data Types | R Code |
|---|---|
| Numbers, Strings | n <- 3<br>s <- "columbus, ohio" |
| Vectors | levels <- c("Wow", "Good","Bad")<br>ratings <- c("Bad", "Bad", "Wow") |
| Factors and Lists | f <- factor(ratings, levels)<br>l <- list(ratings=ratings,<br>critics=c("Siskel","Ebert")) |
| Functions | stdev <- function(x) {sd(x)} |

## R Structured Types

| Data Types | R Code |
|---|---|
| Matrix - (n*m numeric data frame) | m <- matrix( c(1:3, 11:13), nrow = 2, ncol = 3, byrow = TRUE) |
| Table – contingency table | t <- table(dfm$factor_variable) |
| Data frames – data sets | dfm <- read.csv("CrimeRatesByStates2005.csv") |
| Extracting data | xdfm <- dfm[1:3,]<br>ydfm <- dfm[, 3:5]<br>s <- dfm$state |









## Basic R Operations on Vectors

| Function | R Code |
|---|---|
| Operations on Vectors | v <- c(1:10); w <- c(15:24) ; nv <- v * pi ; nw <- w * v |
| Vector transformations | radius <- sqrt( d$area)/ pi)<br>t <- as.table(dfm$factor_variable)<br>pct <- t/sum(t)* 100 |
| Logical Vectors | v[ v < 1000 ]<br>ndf <- subset(dfm, d$population < 10000)<br>nv <- v[c(1,2,3,5,8,13)] |
| Examining data structures | dim(dfm); attributes(dfm) ; class(dfm); typeof(dfm) |

## Descriptive Statistics

| Function | R Code |
|---|---|
| View the data | head(x); tail(x) |
| View a summary of the data | summary(x) |
| Compute basic statistics | sd(x); var(x); range(x); IQR(x) |
| Correlation | cor(x); cor(d$var1, d$var2) |






23

## Generic Functions

- Also known as method overriding in OO-land
- Specific actions that differ based on the class of the object :

| Code | Function |
|---|---|
| Plot the variable x | plot (x) |
| Histogram of x | hist (x) |
| Internal structure of x | str (x) |

- Good for initial data exploration (more later)






24

## Check Your Knowledge

- Which data structures in R are the most used? Why?
- Consider the cbind() function and the rbind() function that bind a vector to a data frame as a new column or a new row. When might these functions be useful?



 Module 3: Basic Data Analytic Methods Using R 25

25

Introduction Analytics Lifecycle Basic Methods Adv. Methods Tools Lab

## Module 3: Review of Basic Data Analytic Methods Using R

### Lesson 1: Summary

During this lesson the following topics were covered:

- How to use the R Graphical User Interface
- How to get data into (and out of) R
- Data Types used in R, and the basic R operations
- Basic descriptive statistics
- Using generic functions



 Module 3: Basic Data Analytic Methods Using R 26

26

## Lab Exercise 2: Introduction to R

This lab is designed to investigate and practice working with R and using it to examine data.

- After completing the tasks in this lab you should able to:
  - Read data sets into R, save them, and examine the contents







27

## Lab Exercise 2: Introduction to R

1. Invoke the R environment
2. Examine the Workspace
3. Getting Familiar with R
4. Read in the Lab Script
5. Working with R : reading external data
6. Verify the contents of the tables
7. Manipulating data frames in R
8. Investigate your data
9. Save the data sets
10. Continue investigating the data
11. Exit R







28


## Module 3: Review of Basic Data Analytic Methods Using R

### Lesson 2: Analyzing and Exploring the Data

During this lesson the following topics are covered:

- Why visualize?
- Examining a single variable
- Examining pairs of variables
- Indications of dirty data.
- Data exploration vs. presentation



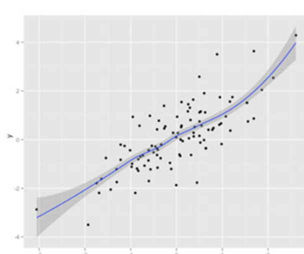

29

## Why Visualize?

Summary statistics give us some sense of the data:

- Mean vs. Median.
- Standard deviation.
- Quartiles, Min/Max.
- Correlations between variables.

```
summary(data)
       x                   y
 Min.   :-3.05439   Min.   :-3.50179
 1st Qu.:-0.61055   1st Qu.:-0.75968
 Median : 0.04666   Median : 0.07340
 Mean   :-0.01105   Mean   : 0.09383
 3rd Qu.: 0.56067   3rd Qu.: 0.88114
 Max.   : 2.60614   Max.   : 4.28693
```



Visualization gives us a more holistic sense




30

## Anscombe's Quartet

4 data sets, characterized by the following. Are they the same, or are they different?

| Property | Values |
|---|---|
| Mean of x in each case | 9 |
| Exact variance of x in each case | 11 |
| Exact mean of y in each case | 7.5 (to 2 d.p) |
| Variance of Y in each case | 4.13 (to 2 d.p) |
| Correlations between x and y in each case | 0.816 |
| Linear regression line in each case | Y = 3.00 + 0.500x (to 2 d.p and 3 d.p resp.) |

i

| x | y |
|---|---|
| 10.00 | 8.04 |
| 8.00 | 6.95 |
| 13.00 | 7.58 |
| 9.00 | 8.81 |
| 11.00 | 8.33 |
| 14.00 | 9.96 |
| 6.00 | 7.24 |
| 4.00 | 4.26 |
| 12.00 | 10.84 |
| 7.00 | 4.82 |
| 5.00 | 5.68 |

ii

| x | y |
|---|---|
| 10.00 | 9.14 |
| 8.00 | 8.14 |
| 13.00 | 8.74 |
| 9.00 | 8.77 |
| 11.00 | 9.26 |
| 14.00 | 8.10 |
| 6.00 | 6.13 |
| 4.00 | 3.10 |
| 12.00 | 9.13 |
| 7.00 | 7.26 |
| 5.00 | 4.74 |

iii

| x | y |
|---|---|
| 10.00 | 7.46 |
| 8.00 | 6.77 |
| 13.00 | 12.74 |
| 9.00 | 7.11 |
| 11.00 | 7.81 |
| 14.00 | 8.84 |
| 6.00 | 6.08 |
| 4.00 | 5.39 |
| 12.00 | 8.15 |
| 7.00 | 6.42 |
| 5.00 | 5.73 |

iv

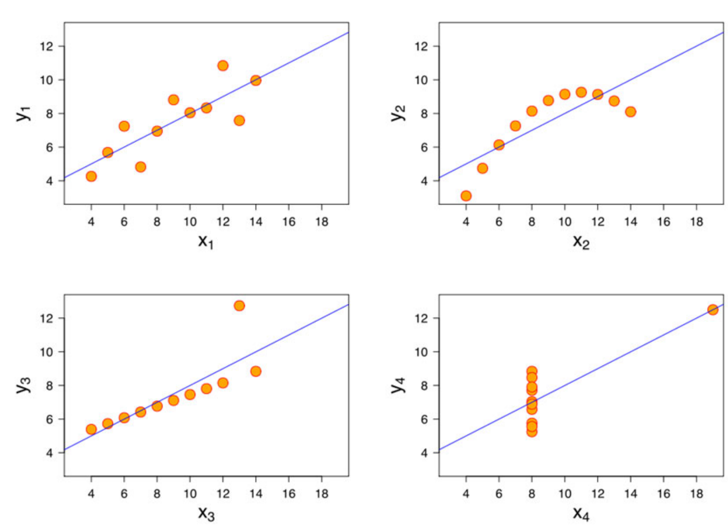
| x | y |
|---|---|
| 8.00 | 6.58 |
| 8.00 | 5.76 |
| 8.00 | 7.71 |
| 8.00 | 8.84 |
| 8.00 | 8.47 |
| 8.00 | 7.04 |
| 8.00 | 5.25 |
| 19.00 | 12.50 |
| 8.00 | 5.56 |
| 8.00 | 7.91 |
| 8.00 | 6.89 |





31

## Moral: Visualize Before Analyzing!








32

## Visualizing Your Data

- Examining the distribution of a single variable
- Analyzing the relationship between two variables
- Establishing multiple pair wise relationships between variables
- Analyzing a single variable over time
- Data exploration versus data presentation

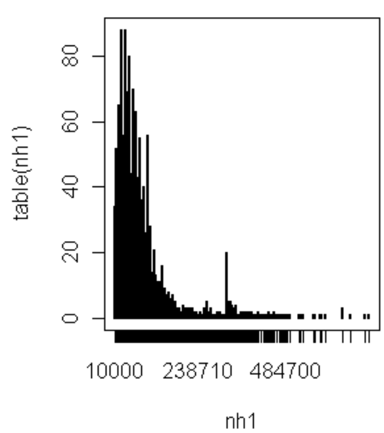



33

## Examining the Distribution of a Single Variable

Graphing a single variable

- plot(sort(.)) – for low volume data
- hist(.) – a histogram
- plot(density(.)) – densityplot
  - A "continuous histogram“
- Example
  - Frequency table of household income





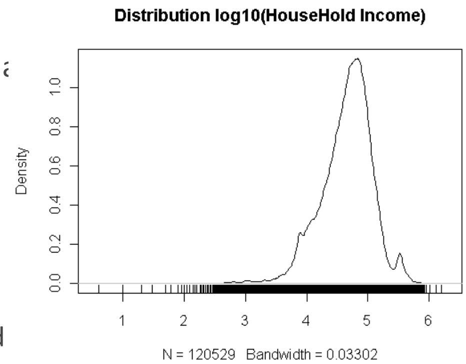


34

## Examining the Distribution of a Single Variable

Graphing a single variable

- plot(sort(.)) – for low volume da
- hist(.) – a histogram
- plot(density(.)) – densityplot
  - A "continuous histogram"
- Example
  - Frequency table of household income
    - rug() plot emphasizes distribution







35

## What are we looking for?

**A sense of the data range**

- If it's very wide, or very skewed, try computing the log

**Outliers, anomalies**

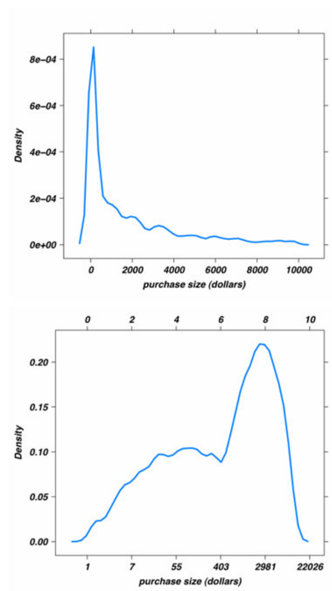
- Possibly evidence of dirty data

**Shape of the Distribution**

- Unimodal? Bimodal?
- Skewed to left or right?
- Approximately normal? Approximately lognormal?

**Example - Distribution of purchase size ($)**

- Range from 0 to > $10K, right skewed
- Typical of monetary data
- Plotting log of data gives better sense of distribution
- Two purchasing distributions
  - ~ $55
  - ~ $2900

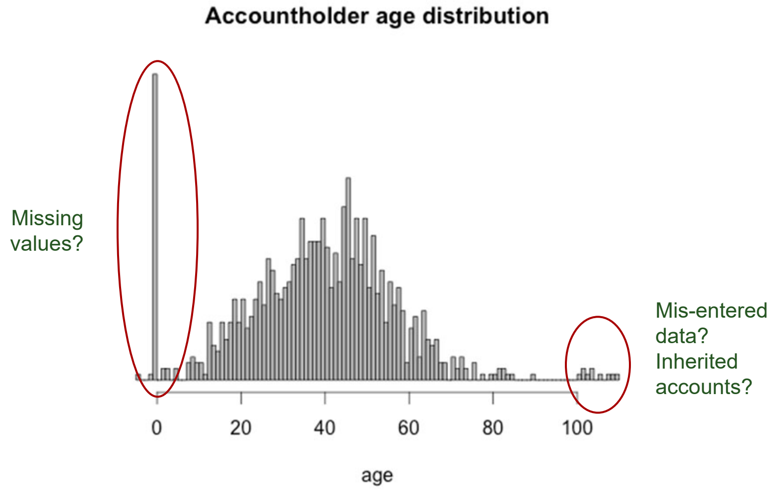






36

## Evidence of Dirty Data





Module 3: Basic Data Analytic Methods Using R 37

37

## "Saturated" Data


Do we really have no mortgages older than 10 years?

Or does the year 2004 in the origination field mean "2004 or prior"?

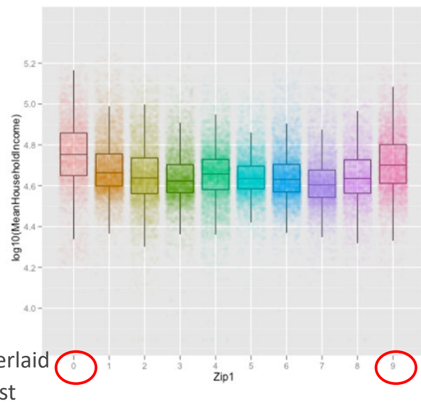

Module 3: Basic Data Analytic Methods Using R 38

38

## Analyzing the Relationship Between Two Variables

### How?

- Two Continuous Variables (or two discrete variables)
  - Scatterplots
  - LOESS (fit smoothed line to the data)
  - Linear models: graph the correlation
  - Binplots, hexbin plots
    - More legible color-based plots for high volume data
- Continuous *vs.* Discrete Variable
  - Jitter, Box and whisker plots, Dotplot or barchart




### Example:

- Household income by region (ZIP1)
- Scatterplot with jitter, with box-and-whisker overlaid
- New England (0) and West Coast (9) have highest mean household income



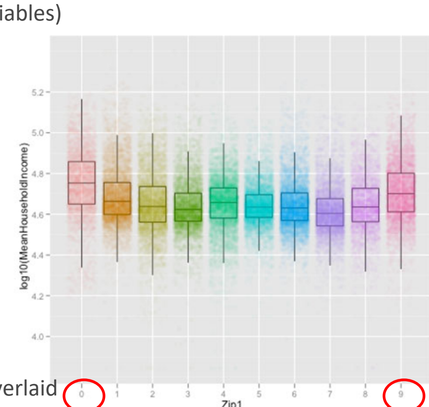
Module 3: Basic Data Analytic Methods Using R 39

39

## Analyzing the Relationship … (Continued)

### How?

- Two Continuous Variables (or two discrete variables)
  - Scatterplots
  - LOESS (fit smoothed line to the data)
  - Linear models: graph the correlation
  - Binplots, hexbin plots
    - More legible color-based plots for high volume data
- Continuous *vs.* Discrete Variable
  - Jitter, Box and whisker plots, Dotplot or barchart




### Example:

- Household income by region (ZIP1)
- Scatterplot with jitter, with box-and-whisker overlaid
- New England (0) and West Coast (9) have highest mean household income

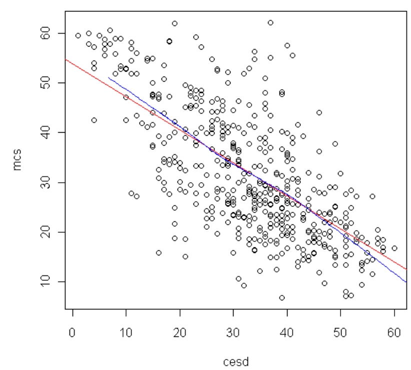


Module 3: Basic Data Analytic Methods Using R 40

40

## Two Variables: What are we looking for?

- Is there a relationship between the two variables?
  - Linear? Quadratic?
  - Exponential?
    - Try semi-log or log-log plots
  - Is it a cloud?
    - Round? Concentrated? Multiple Clusters?
- How?
  - Scatterplots
- Example
  - Red line: linear fit
  - Blue line: LOESS
  - Fairly linear relationship, but with wide variance

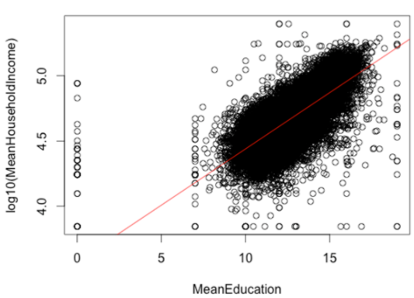




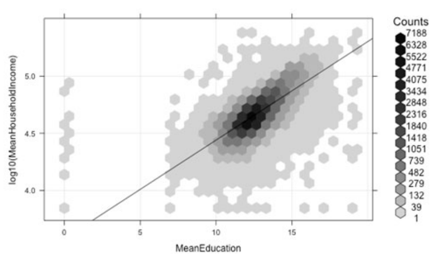


41

## Two Variables: High Volume Data - Plotting




**Scatterplot:**
Overplotting makes it difficult to see structure

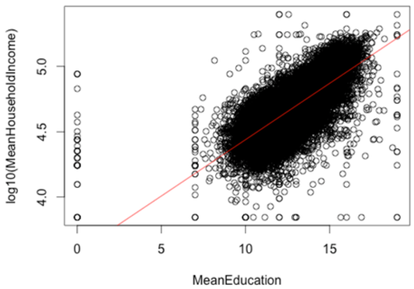



**Hexbinplot:**
Now we see where the data is concentrated.

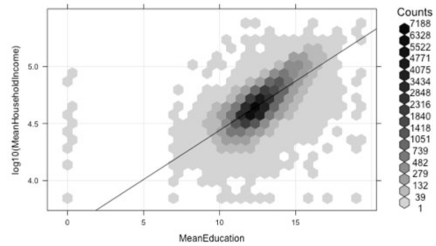



42

## Two Variables: High Volume Data – Plotting (Continued)




**Scatterplot:**
Overplotting makes it difficult to see structure




**Hexbinplot:**
Now we see where the data is concentrated.

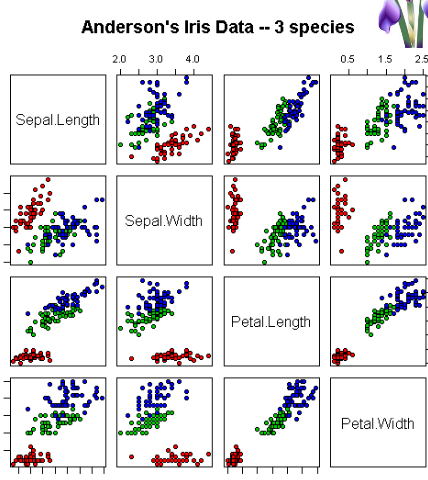


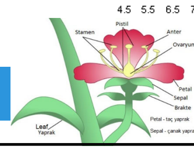

43

## Establishing Multiple Pairwise Relationships Between Variables

freesia

- Why?
  - Examine many two-way relationships quickly
- How?
  - pairs(ds) can generate a plot of each pairs of variables
- Example
  - Iris Characteristics
    - Strong linear relationship between petal length and width
    - Petal dimensions discriminate species more strongly than sepal dimensions






44

## Analyzing a Single Variable over Time

What?

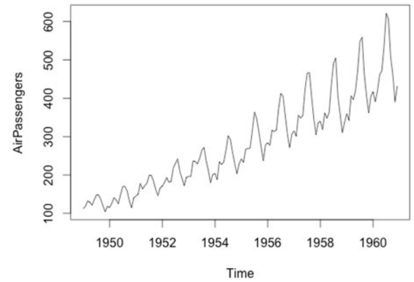
- Looking for …
  - Data range
  - Trends
  - Seasonality

How?

- Use time series plot

Example

- International air travel (1949-1960)
- Upward trend: growth appears superlinear
- Seasonality
  - Peak air travel around Nov. with smaller peaks near Mar. and June

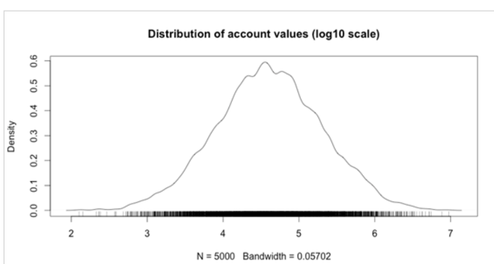






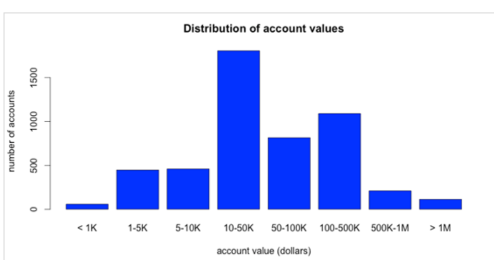
45

## Data Exploration vs. Presentation




Data Exploration:

This tells you what you need to know.



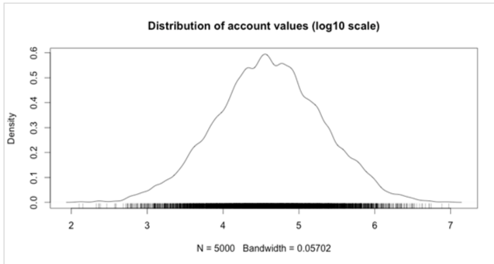

Presentation:

This tells the stakeholders what they need to know.



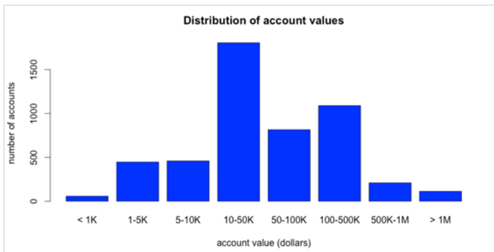

46

## Data Exploration vs. Presentation (Continued)




Data Exploration:

This tells you what you need to know.




Presentation:

This tells the stakeholders what they need to know.





47

## Check Your Knowledge

- Do you think the regression line sufficiently captures the relationship between the two variables? What might you do differently?
- In the Iris slide example, how would you characterize the relationship between sepal width and sepal length?
- Did you notice the use of color in the Iris slide? Was it effective? Why or why not?





48

Introduction | Analytics Lifecycle | Basic Methods | Adv. Methods | Tools | Lab

# Module 3: Review of Basic Data Analytic Methods Using R

## Lesson 2: Summary

During this lesson the following topics were covered:

- Justifying why we visualize data
- Using plots and graphs to determine:
  - Shape of a single variable
  - “dirty” data or “saturated” data
  - Relationship between two or more variables
  - Relationship between multiple variables
  - A single variable over time
- Data exploration *versus* Presentation




49

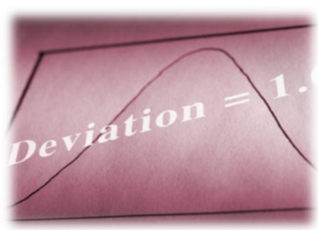
Introduction | Analytics Lifecycle | Basic Methods | Adv. Methods | Tools | Lab

# Module 3: Review of Basic Data Analytic Methods Using R

## Lesson3: Statistics for Model Building and Evaluation

During this lesson the following topics are covered:

- Statistics in the Analytic Lifecycle
- Hypothesis Testing
- Difference of means
- Significance, Power, Effect Size
- ANOVA
- Confidence Intervals







50

## Statistics in the Analytic Lifecycle

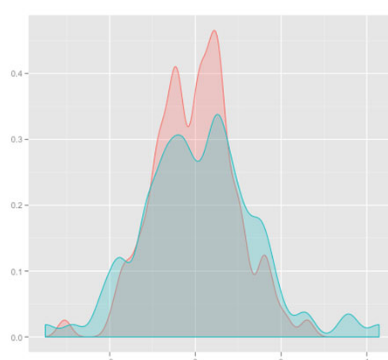
- Model Building and Planning
  - Can I predict the outcome with the inputs that I have?
  - Which inputs?
- Model Evaluation
  - Is the model accurate?
  - Does it perform better than "the obvious guess"
  - Does it perform better than another candidate model?
- Model Deployment
  - Do my predictions make a difference?
    - Are we preventing customer churn?
    - Have we raised profits?

 Module 3: Basic Data Analytic Methods Using R 51

51

## Hypothesis Testing

- Fundamental question: "Is there a difference between the populations based on samples?“
  - Examples : Mean, Variance
- Null hypothesis : There is no difference
- Alternate hypothesis : There is a difference




 Module 3: Basic Data Analytic Methods Using R 52

52

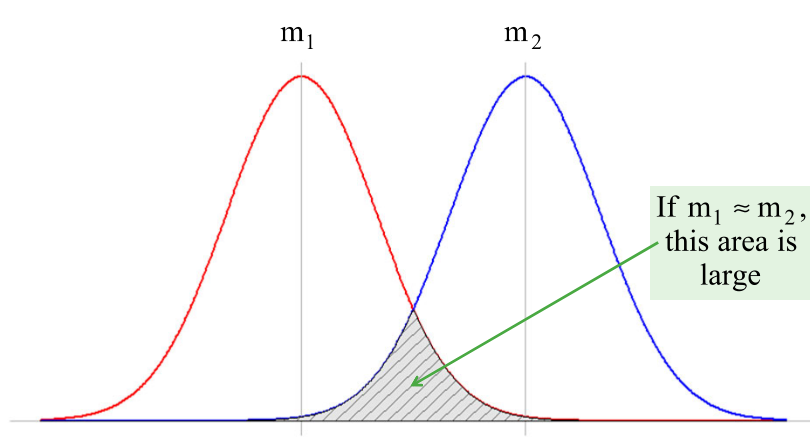
## Null and Alternative Hypotheses: Examples

| Null Hypothesis | Alternative Hypothesis |
|---|---|
| The best estimate of the outcome is the average observed value: • The mean is the "Null Model" | The model predicts better than the null model: • The average prediction error from the model is smaller than that of the null model |
| This variable does not affect the outcome: • The coefficient value is zero | The variable does affect outcome: • Coefficient value is non-zero |
| The model predictions do not improve revenue: • Revenue is the same with or without intervention | Interventions based on model predictions improve revenue: • A/B Testing, ANOVA |




53

## Intuition: Difference of Means





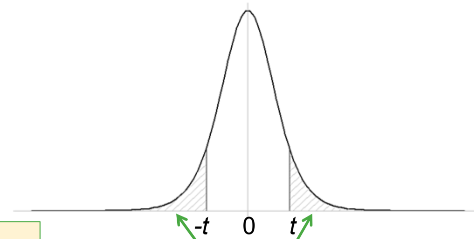


54

## Welch's t-test

t-statistic: $$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

(this is the t-statistic for the Welch t-test)




```
> x = rnorm(10) # distribution centered at 0
> y = rnorm(10,2) # distribution centered at 2
> t.test(x,y)

            Welch Two Sample t-test

data:  x and y
t = -7.2643, df = 15.05, p-value = 2.713e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.364243 -1.291811
sample estimates:
mean of x mean of y
0.5449713 2.3729984
```

p-value: area under the tails of the appropriate student's distribution

if p-value is small (say < 0.05), then reject the null hypothesis and assume that $m_1$ <> $m_2$

$m_1$ and $m_2$ are "significantly different"






## Wilcoxon Rank Sum Test

- t-test assumes that the populations are normally distributed
  - Sometimes this is close to true, sometimes not
- Wilcoxon Rank Sum test
  - Makes no assumption about the distributions of the populations
  - More robust test for difference of means
  - if p-value is small: reject the null hypothesis (equal means)

```
> mean(x)
[1] 0.5449713
> mean(y)
[1] 2.372998
> wilcox.test(x, y)

            wilcoxon rank sum test

data:  x and y
W = 2, p-value = 4.33e-05
alternative hypothesis: true location shift is not equal to 0
```

## Hypothesis Testing: Summary

- Calculate the test statistic
  - Different hypothesis tests are appropriate, in different situations
- Calculate the p-value on the test statistic
- If p-value is "small" then reject the null hypothesis
  - "small" is often p < 0.05 by convention (95% confidence)
  - Many data scientists prefer a smaller threshold.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$






## Generating a Hypothesis: Type I and Type II Error

| If $H_0$ is X, and we ...: | **Null hypothesis($H_0$) is true** | **Null hypothesis($H_0$) is false** |
|---|---|---|
| Fail to accept the Null Hypothesis → we claim something happened | **Type I error**<br>**False positive**<br>$\alpha$ | **Correct Outcome**<br>**True positive**<br>**We reject the Null hypothesis** |
| Fail to reject the null hypothesis → we claim nothing happened. | **Correct outcome**<br>**True negative**<br>**Accept the NULL hypothesis** | **Type II error**<br>**False negative**<br>$\beta$ |

**Example: Ham or Spam? $H_0$: it's Ham $H_A$: it's Spam**

| It's Really -> we say it's ↓ | Ham | Spam |
|---|---|---|
| Spam | **Type I – false positive** | OK – true positive |
| Ham | OK – true negative | **Type II – false negative** |

- **Goal: Identify Spam**
- **Which error is worse?**

## Generating a Hypothesis: Type I and Type II Error (Continued)

| If $H_0$ is X, and we …: | **Null hypothesis($H_0$) is true** | **Null hypothesis($H_0$) is false** |
|---|---|---|
| Fail to accept the Null Hypothesis → we claim something happened | **Type I error**<br>**False positive**<br>α | **Correct Outcome**<br>**True positive**<br>**We reject the Null hypothesis** |
| Fail to reject the null hypothesis → we claim nothing happened. | **Correct outcome**<br>**True negative**<br>**Accept the NULL hypothesis** | **Type II error**<br>**False negative**<br>β |

**Example: Ham or Spam? $H_0$: it's Ham $H_A$: it's Spam**

| It's Really -><br>we say it's ↓ | Ham | Spam |
|---|---|---|
| Spam | **Type I – false positive** | OK – true positive |
| Ham | OK – true negative | **Type II – false negative** |

- **Goal: Identify Spam**
- **Which error is worse?**

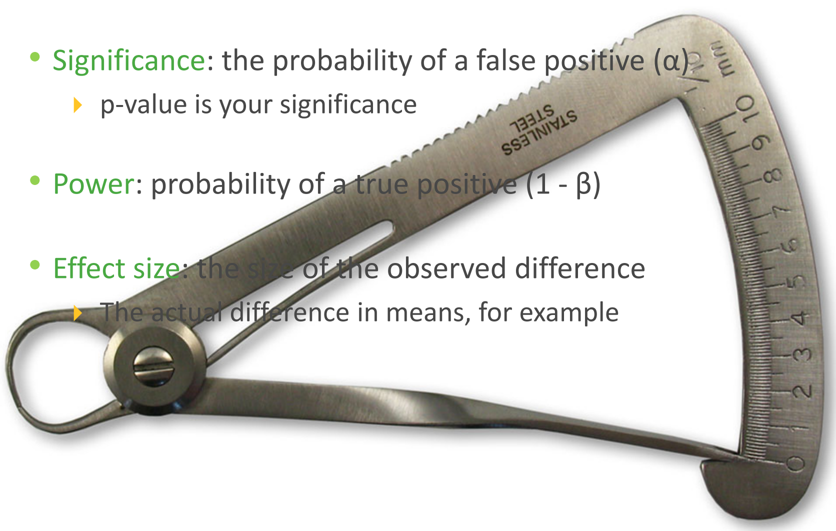




59

## Significance, Power and Effect Size

- Significance: the probability of a false positive (α)
  - p-value is your significance
- Power: probability of a true positive (1 - β)
- Effect size: the size of the observed difference
  - The actual difference in means, for example

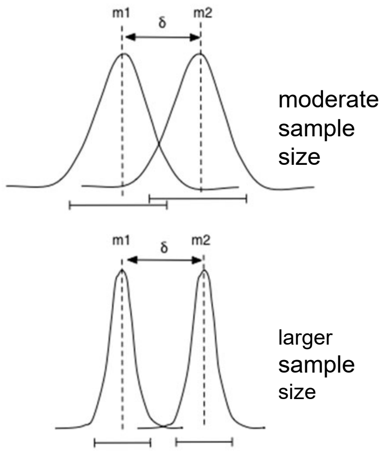







60

## Always Keep Effect Size in Mind!




Both power and significance increase with larger sample sizes.

So you can observe an effect size that is *statistically* significant, but *practically* insignificant!




61

## Hypothesis Testing: ANOVA

ANOVA is a generalization of the difference of means

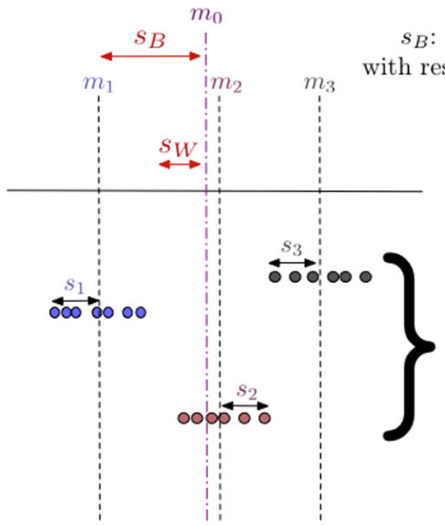
- One-way ANOVA
  - k populations ("treatment groups")
  - $n_i$ samples each – total N subjects
  - Null hypothesis: ALL the population means are equal

| Population | $n_i$: # offers made | $m_i$: avg purchase size |
|---|---|---|
| Offer 1 | 100 | $55 |
| Offer 2 | 102 | $50 |
| No intervention | 99 | $25 |




62

## ANOVA: Understanding the F statistic




$s_B$: how the population means vary with respect to the total mean $m_0$

$$s_B{}^2 = \frac{1}{k-1}\sum_i n_i \cdot (m_i - m_0)^2$$

$$S_W^2 = \frac{1}{N-k}\sum_i^k \sum_j^{n_i} (x_{ij} - m_i)^2$$

$s_W$: the "average" of the $s_i$

Test statistic: $F = s_B{}^2 / s_W{}^2$






## R Example: ANOVA

3 different offers, and their outcomes

Use lm() to do the ANOVA

```
>offers = sample(c("noffer", "offer1", "offer2"),
          size=500, replace=T)
>purchasesize = ifelse(offers=="noffer", rlnorm(500,
meanlog=log(25)), ifelse(offers=="offer1", rlnorm(500,
meanlog=log(50)), rlnorm(500, meanlog=log(55))))
>offertest = data.frame(offer=as.factor(offers),
          purchase_amt=purchasesize)
> model = lm(log10(purchase_amt) ~ as.factor(offers),
             data=offertest)
>summary(model)
Residuals:
    Min      1Q  Median     3Q    Max
-1.1940 -0.2837  0.0135 0.2863 1.3374

Coefficients:
                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)            1.49092    0.03240  46.011  < 2e-16 ***
as.factor(offers)offer1 0.20424   0.04706   4.340 1.73e-05 ***
as.factor(offers)offer2 0.22371   0.04596   4.867 1.52e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4262 on 497 degrees of freedom
Multiple R-squared: 0.05479,    Adjusted R-squared: 0.05098
F-statistic:  14.4 on 2 and 497 DF,  p-value: 8.304e-07
```

offer1-nooffer
offer2-nooffer

**F-statistic: reject the null hypothesis**

Tukey's test: all pair-wise tests for difference of means

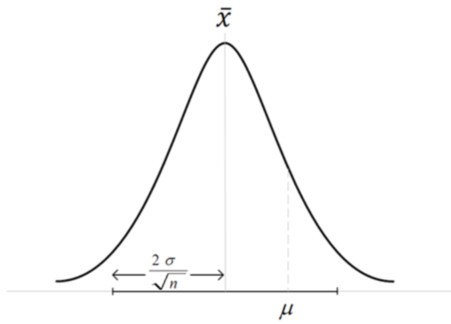
**95% confidence intervals** for difference between means

```
> TukeyHSD(aov(model))
  Tukey multiple comparisons of means
    95% family-wise confidence level
    Fit: aov(formula = model)

$offers
                    diff         lwr       upr     p adj
offer1-noffer 0.20424099  0.09361976 0.3148622 0.0000512
offer2-noffer 0.22370761  0.11566775 0.3317475 0.0000045
offer2-offer1 0.01946663 -0.09146092 0.1303942 0.9104871
```

**.No appreciable difference between offer1 and offer2**

## Confidence Intervals




Example:
- Normal data N(μ, σ)
- x is the estimate of μ
  - based on n samples

μ falls in the interval

$\bar{x} \pm 2\sigma/\sqrt{n}$

with approx. 95% probability ("95% confidence")

If x is your estimate of some unknown value μ, the P% confidence interval is the interval around x that μ will fall in, with probability P.

## Example

The defect rate of a disk drive manufacturing process is within 0.9% - 1.7%, with 98% confidence. We inspect a sample of 1000 drives from one of our plants.

- We observe 13 defects in our sample.
  - Should we inspect the plant for problems?
- What if we observe 25 defects in the sample?

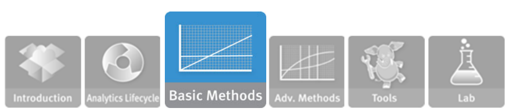
 

## Check Your Knowledge

- Refer back to the ANOVA example on an earlier slide. What do you think? Does the difference between *offer1* and *offer2* make a practical difference? Should we go ahead and implement one of them?
- If yes, and the costs were US $25 for each offer1 and US $10 for *offer2*, would you still make the same decision?
- In our manufacturing plant example, assuming you would check the plant for problems in the manufacturing process, how might you justify this decision financially?




67




## Module 3: Review of Basic Data Analytic Methods Using R

### Lesson 3: Summary

During this lesson the following topics were covered:

- The role of Statistics in the Analytic Lifecycle
- Developing a model and generating the null and the alternative hypothesis
- Difference between means
- Difference between significance, power and effect size, and how they relate to Type I and Type II errors
- Applying ANOVA and determining whether the results are significant
- Defining confidence intervals and applying them




68

## Lab Exercise 3: Basic Statistics, Visualization and Hypothesis Tests

This lab is designed to investigate and practice using R to perform basic statistics and visualization on data and to perform hypothesis testing.

- After completing the tasks in this lab you should able to:
  - Perform basic data analysis
  - Visualize data with R
  - Create and test a hypothesis




69

## Lab Exercise 3: Basic Statistics, Visualization and Hypothesis Tests– Part1 - Workflow

1. Prepare working environment for the Lab and load data files
2. Obtain summary statistics for Household Income and visualize data
3. Obtain summary statistics for number of rooms and visualize data
4. Remove Outliers
5. Stratify Variable – Household Income and plot the results
6. Plot Histogram and Distributions
7. Compute Correlation between income and number of rooms
8. Create a Boxplot – Distribution of income as a factor of number of rooms
9. Exit R




70

## Lab Exercise 3: Basic Statistics, Visualization and Hypothesis Tests - Part 2 - Workflow

1. Define problem – Analysis of Variance (ANOVA)
2. Generate the Data
3. Examine the Data
4. Plot and determine how purchase size varies within the three groups
5. Use lm() to do the ANOVA
6. Use Tukey's test to check all the differences of means
7. Use the lattice package for density plot
8. Plot the Logarithms of the Data
9. Use ggplot() package
10. Generate the example data to perform a Hypothesis Test with manual calculations
11. Create a function to calculate the pooled variance, which is used in the Student's t statistic
12. Examine the Data
13. Calculate the t statistic for Student's t-test
14. Calculate the degrees of freedom
15. Compute the area under the curve
16. Perform Student's t-test directly and compare the results

 Module 3: Basic Data Analytic Methods Using R 71

71

Introduction | Analytics Lifecycle | Basic Methods | Adv. Methods | Tools | Lab

## Module 3: Summary

Key points covered in this module:

- How to use basic analytics methods such as distributions, statistical tests and summary operations to investigate a data set
- How to use R to apply visualization patterns to better understand the data, help develop a model and derive hypotheses, and determine if our actions had a practical affect.

 Module 3: Basic Data Analytic Methods Using R 72

72