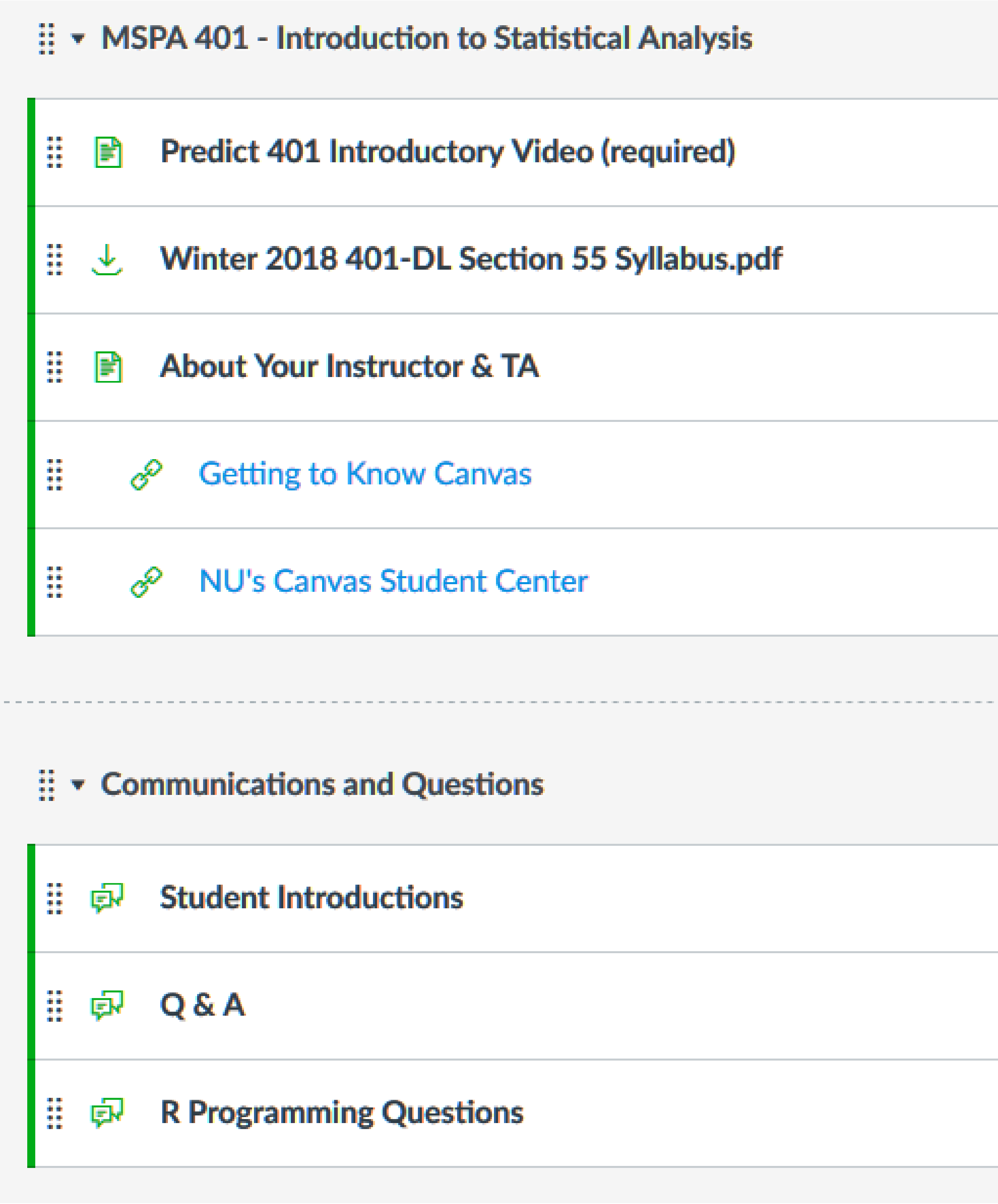
Session Agenda Sync #1

- Academic Integrity
- Introductory Video
- Statistical thinking
- Misleading graphics
- Measures of Shape and Location
- EDA using R
  - Displays
  - Outliers
- Things to remember
- Test #1
- Using Rstudio Demo

Academic Integrity at Northwestern

Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., using material from readings without citing or copying another student's paper). **Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program.**

Preliminaries



MSPA 401 - Introduction to Statistical Analysis

Predict 401 Introductory Video (required)

Winter 2018 401-DL Section 55 Syllabus.pdf

About Your Instructor & TA

Getting to Know Canvas

NU's Canvas Student Center

Communications and Questions

Student Introductions

Q & A

R Programming Questions

Predict 401 Course Procedures

- **Each week there are two discussion topics. One involves subject matter, the second personal reflections. Class participation is expected. (5 points for the subject matter topic, 3 points for the reflections topic plus 2 points for class participation making 10 points overall.)**

- **Discussion topics come available each Monday at 12:01 am and close the following Sunday at 8:00 pm. It is necessary to post your initial comments before the system reveals the postings made by the class.**

- **There are four tests on the reading assignments. These come available Monday at 12:01 am CST and close the following Sunday two weeks later at 11:55 pm CST.**

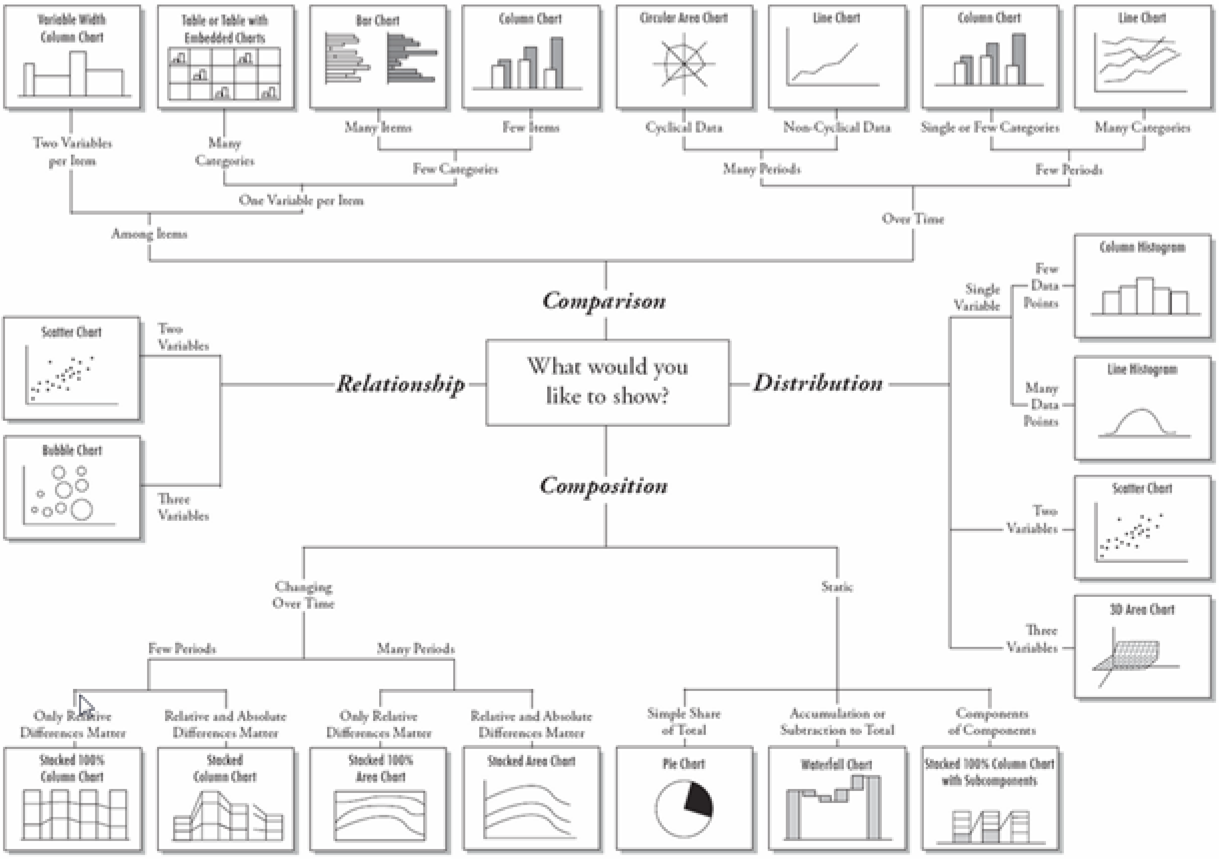
- **The proctored final exam comes available prior to the end of the course. Check the syllabus for specific dates and time. Information regarding arranging with the proctoring organization for scheduling your final exam is posted on the course site in the module for Week Ten.**

- **Tests on the readings and the final are automatically graded by Canvas. Do not submit until you are completely finished. There is no retest.**

- **The two programming tests are graded by the instructor. There are two data analysis assignments that require use of R. Both are graded by the instructor. Study the "Quick Start Guide for R" in advance to prepare.**

- **Without prior arrangement, any late assignment may receive a 1% point deduction for each hour late totaling to a maximum 25% deduction. (For example, ten hours late means a 10% deduction.)**

Chapters 1 & 2 Black
Introduction, Charts and Graphs



http://www.labnol.org/software/find-right-chart-type-for-your-data/6523/

Predict 401 will be concerned primarily with distributions, comparisons and relationships. Frequently used displays will include histograms, scatter plots, box-and-whisker plots, line charts, frequency charts, bar charts, stem-and-leaf plots and Q-Q charts.

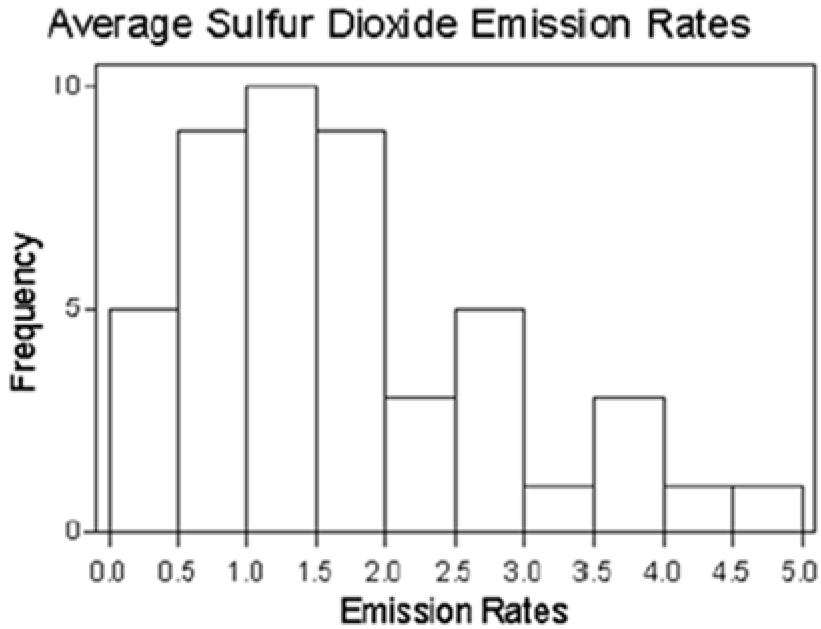
However simple, statistical charts and summaries can be misleading.

Test Question

Dealing with Statistical Thinking

The following histogram shows average SO_2 (sulfur dioxide) boiler emission rates from utility companies. The data were collected from a voluntary response sample of utility companies.

Does the distribution depicted in the histogram reflect the true distribution of the population? Why or why not?



"How to Spot Spin?"

http://researchbriefings.parliament.uk/ResearchBriefing/Summary/SN04446

Statistical Thinking

Prepare

1. Context and Goals
2. Source of the Data
3. Methods of Collection
4. Relevant Variables

Analyze

1. Assess Data Quality
2. Explore the Data
3. Apply Statistical Methods

Conclude

1. Determine Results
   a. Statistical Significance
   b. Practical Significance
2. Communicate Results

Data Quality

ISO 9000: 2015 definition:

Data quality is the degree to which a set of characteristics of data fulfills requirements.

Examples of characteristics are:

- **Validity**
- **Reliability**
- **Accuracy**
- **Completeness**
- **Timeliness**

Requirements are determined explicitly or implicitly by the intended use of the data for planning, decision making and operations. Data fit for their intended usage are worthy of analysis.

Osborne, Jason W., "Best Practices in Data Cleaning", Sage (2013) ISBN 978-1-4129-8801-8.

Exploratory Data Analysis (EDA)

Tukey: **“The value of a picture is when it forces us to notice what we never expected to see.”**

Ratner: “Let your data be your guide.”

EDA is investigative detective work requiring an attitude of skepticism, openness, sharp-sightedness and flexibility to uncover and understand what’s there.

EDA should be performed BEFORE any classical statistical analysis or predictive model building.

Characteristics of Data

- Center (estimated by measures of location)
- Variation (estimated by s^2, MAD, etc.)
- Distribution (illustrated by histograms, Q-Q plots, etc.)
- Outliers (revealed with modified boxplots, etc.)
- Time (changes in properties noted over time)

Judgment and critical thinking are needed to make practical sense of data. Real data are usually not perfect. The presence of outliers is a case in point.

Which Measure of Central Location?

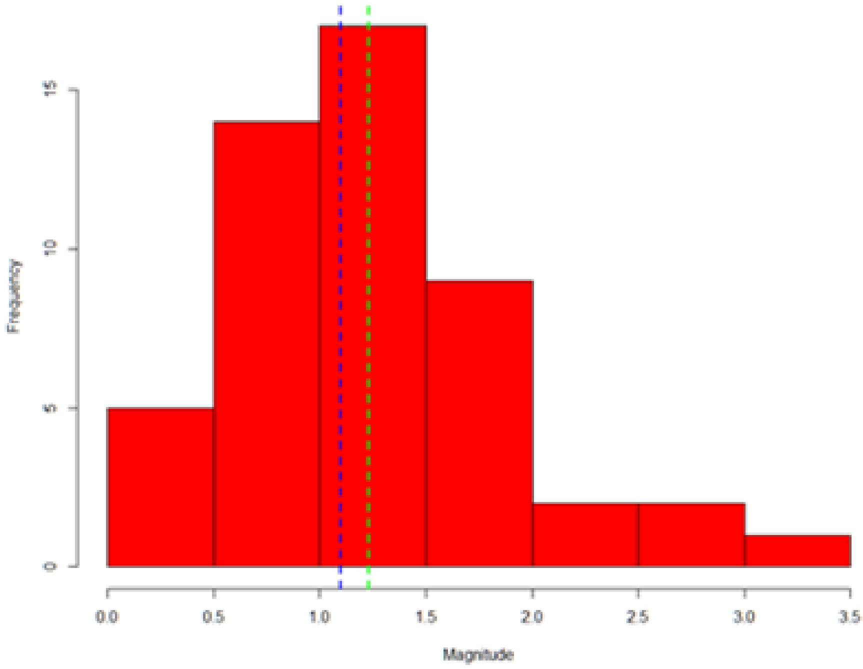
The scale of measurement and questions being asked must be considered.

Scale of Measurement	Descriptive Measures of Central Tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed, outliers rare)	Mean
Interval/Ratio (skewed, outliers not rare)	Median or Trimmed Mean

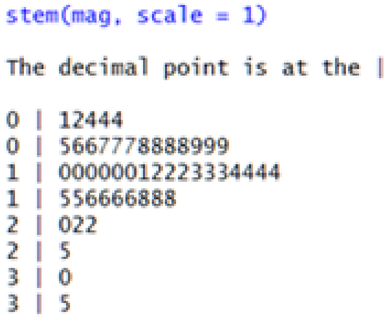
- For describing the "typical" value of a data set when the mean and median disagree (i.e. outliers), the median is better. This assumes there is an unambiguous center to describe.
- For a symmetric distribution (outliers rare), the mean will be more efficient. For a heavy-tailed symmetric distribution (outliers) the median or trimmed mean will be more efficient.
- There are situations with asymmetric distributions for which describing a typical value is not desired. Here the mean may be preferred. For example:
    - Total revenue can be computed from a sample of customers using the mean of a sample multiplied by the total number of customers. The median of a sample is not useful for this purpose.

Location and Shape Examples

Histogram of Earthquake Magnitudes



Stem and Leaf



```
stem(mag, scale = 1)

The decimal point is at the |

0 | 12444
0 | 5667778888999
1 | 00000012223334444
1 | 556666888
2 | 022
2 | 5
3 | 0
3 | 5
```

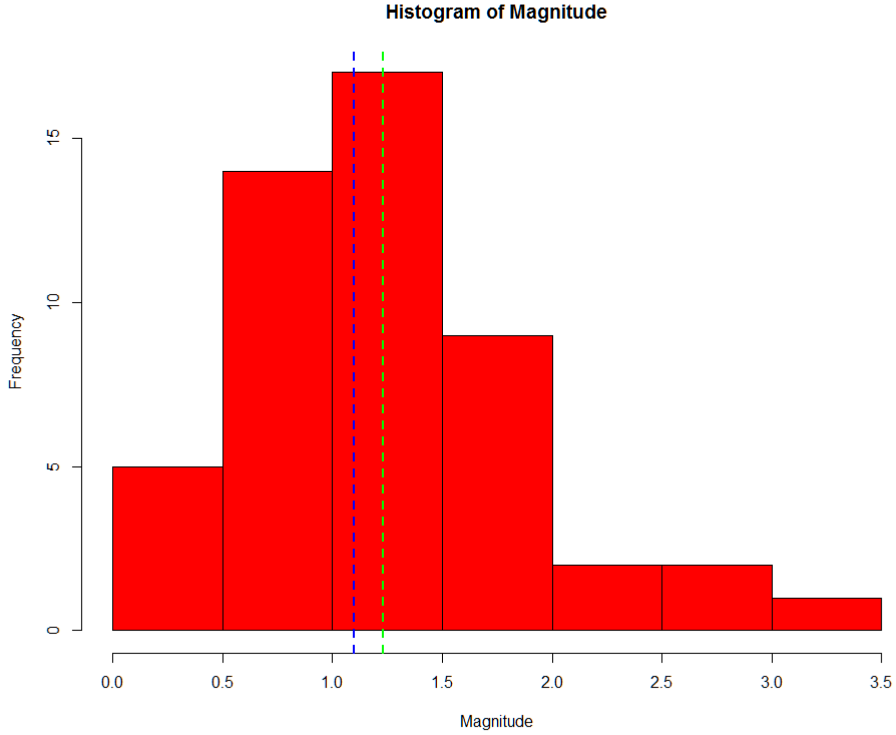
0.10 0.20 0.39 0.40 0.40 0.54 0.64 0.64 0.70 0.70 0.74 0.79 0.79 0.83 0.84 0.90 0.92 0.93

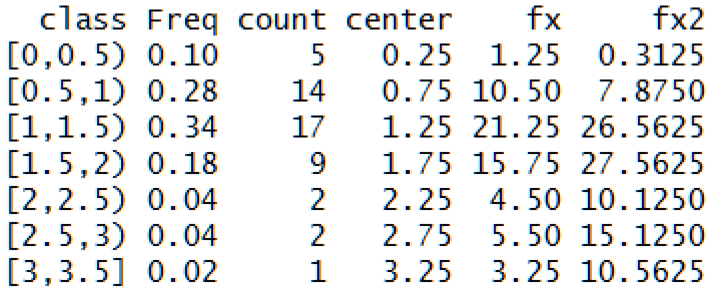
0.99 1.00 1.00 1.00 1.00 1.02 1.05 1.15 1.22 1.24 1.25 1.25 1.28 1.39 1.42 1.42 1.44

1.49 1.54 1.56 1.56 1.62 1.64 1.76 1.79 1.83 1.98 2.20 2.24 2.50 2.95 3.45

Measures of Location		**Measures of Spread (Dispersion)**	
Mean	**1.234**	**Range**	**3.35**
Median	**1.1**	**Standard Deviation**	**0.663**
20% Trimmed Mean	**1.154**	**20% Winsorized Standard Deviation**	**0.349**

Histograms and Frequency Tables

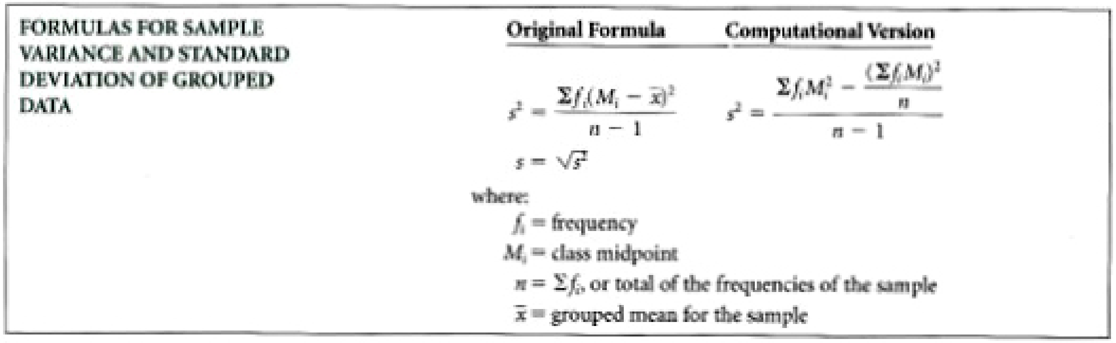




```
    class Freq count center    fx     fx2
[0,0.5) 0.10     5   0.25  1.25  0.3125
[0.5,1) 0.28    14   0.75 10.50  7.8750
[1,1.5) 0.34    17   1.25 21.25 26.5625
[1.5,2) 0.18     9   1.75 15.75 27.5625
[2,2.5) 0.04     2   2.25  4.50 10.1250
[2.5,3) 0.04     2   2.75  5.50 15.1250
[3,3.5] 0.02     1   3.25  3.25 10.5625
```

Grouped Data **mean = 1.24** **std. dev. = 0.658**

Ungrouped Data **mean = 1.23** **std. dev. = 0.663**



FORMULAS FOR SAMPLE VARIANCE AND STANDARD DEVIATION OF GROUPED DATA

Original Formula	Computational Version
$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1}$	$s^2 = \frac{\Sigma f_i M_i^2 - \frac{(\Sigma f_i M_i)^2}{n}}{n - 1}$
$s = \sqrt{s^2}$	

where:

f_i = frequency

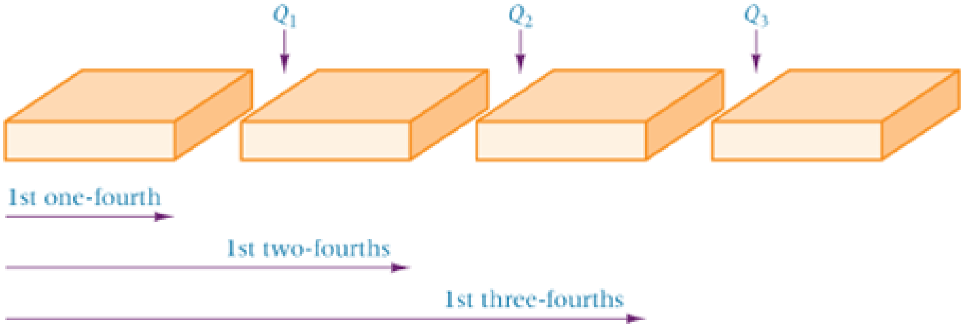
M_i = class midpoint

$n = \Sigma f_i$, or total of the frequencies of the sample

$\bar{x}$ = grouped mean for the sample

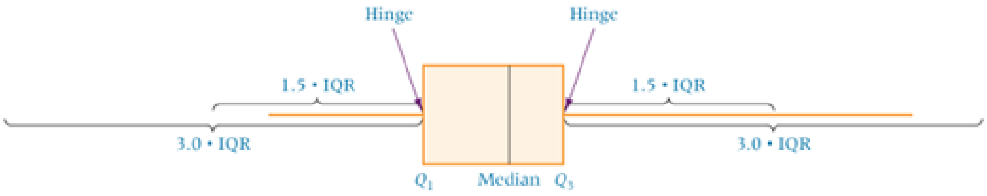
Location and Shape

FIGURE 3.2 Quartiles



The IQR equals Q_3 minus Q_1. Data points which fall outside the box beyond 1.5*IQR are defined as outliers. Beyond 3.0*IQR, they are considered to be extreme outliers.

FIGURE 3.13 Box-and-Whisker Plot

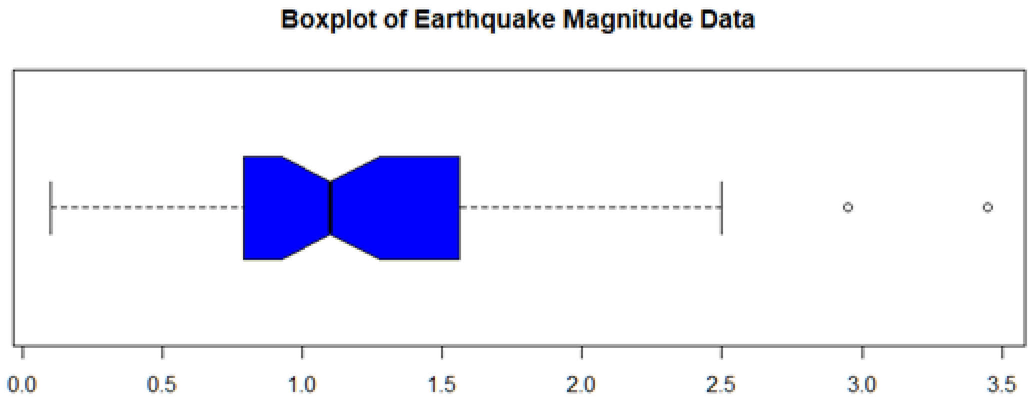


(There is no universal agreement on the determination of quantiles (and hence quartiles). Different formulas have been developed to address specific distributional considerations. R provides nine options with the function quantile(). Black uses type 2 and the default in R is type 7. SAS uses type 3. As sample sizes grow larger, the numerical difference between the methods becomes unimportant.)

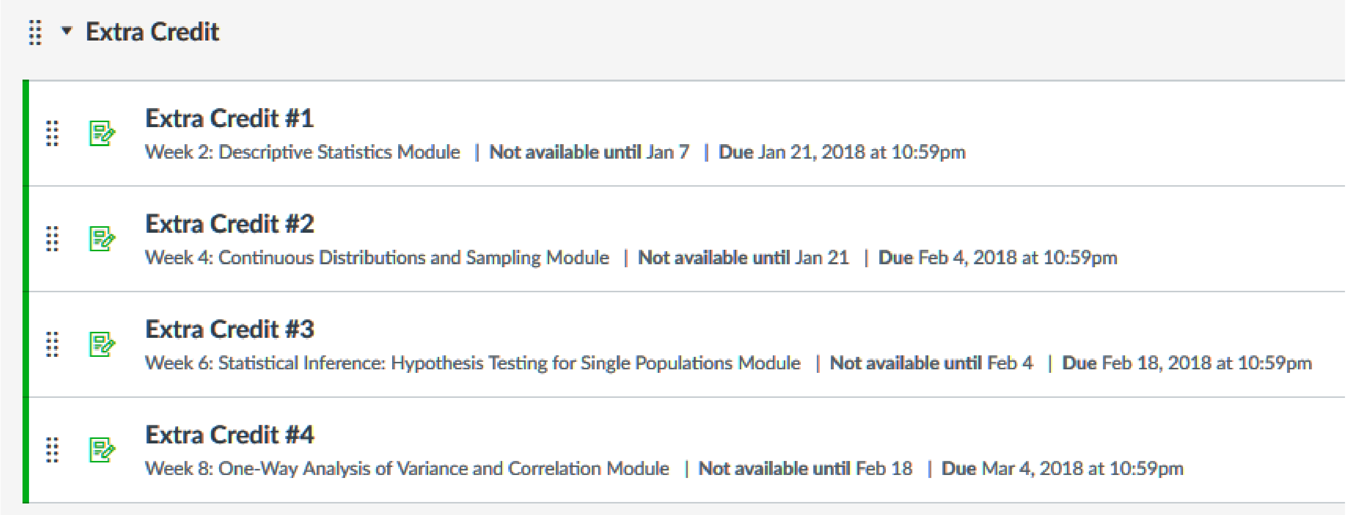
Illustration of Quartile Calculations

Use the R default (type 7) with quantile() except for duplicating calculations in Business Statistics. For that purpose use the type 2 option.

```
> # Earthquake magnitude data
> x <- c(0.70, 0.74, 0.64, 0.39, 0.70, 2.20, 1.98, 0.64, 1.22, 0.20, 1.64, 1.02,
+       2.95, 0.90, 1.76, 1.00, 1.05, 0.10, 3.45, 1.56, 1.62, 1.83, 0.99, 1.56,  0.40,
+       1.28, 0.83, 1.24, 0.54, 1.44, 0.92, 1.00, 0.79, 0.79, 1.54, 1.00, 2.24, 2.50,
+       1.79, 1.25, 1.49, 0.84, 1.42, 1.00, 1.25, 1.42, 1.15, 0.93, 0.40, 1.39)
> p <- c(0.25, 0.5, 0.75)
>
> quantile(x, probs = p, na.rm = FALSE, names = TRUE, type = 2)
 25%  50%  75%
0.79  1.10  1.56
>
> quantile(x, probs = p, na.rm = FALSE, names = TRUE, type = 7)
  25%    50%    75%
0.800  1.100  1.555
>
> boxplot(x, range = 1.5, notch = TRUE, col = "blue", horizontal = TRUE,
+       main = "Boxplot of Earthquake Magnitude Data")
> boxplot.stats(x, coef = 1.5 )
$stats
[1] 0.10 0.79 1.10 1.56 2.50
$n
[1] 50
$conf
[1] 0.9279468 1.2720532
$out
[1] 2.95 3.45
```



Extra Credit



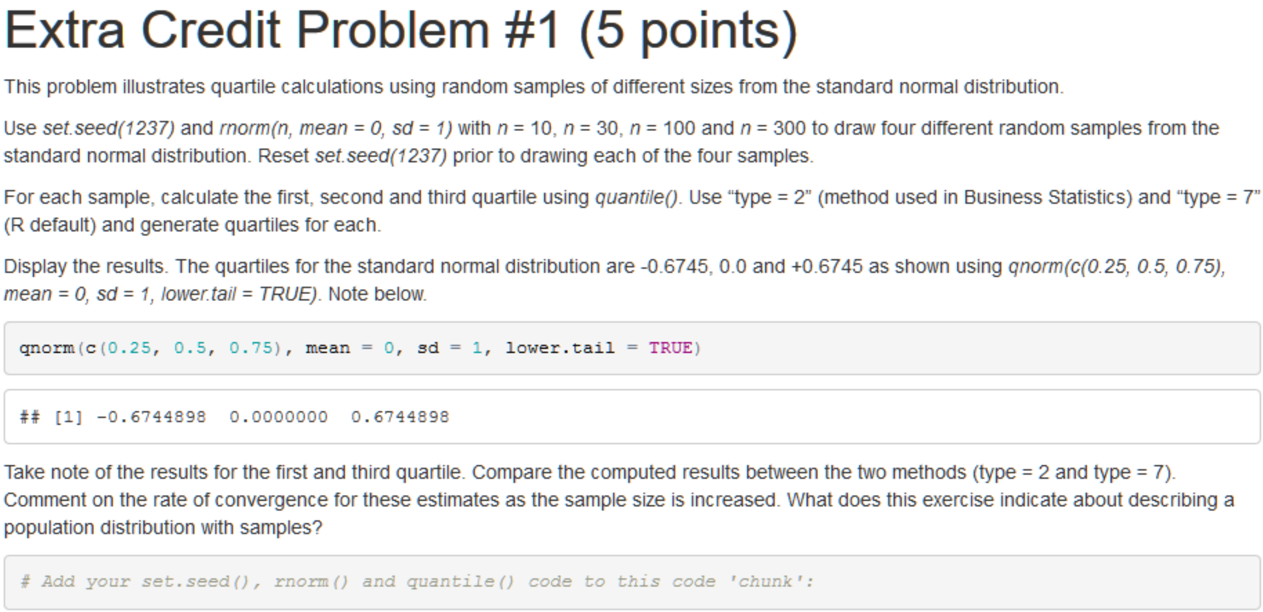
Extra Credit

Extra Credit #1
Week 2: Descriptive Statistics Module | **Not available until** Jan 7 | **Due** Jan 21, 2018 at 10:59pm

Extra Credit #2
Week 4: Continuous Distributions and Sampling Module | **Not available until** Jan 21 | **Due** Feb 4, 2018 at 10:59pm

Extra Credit #3
Week 6: Statistical Inference: Hypothesis Testing for Single Populations Module | **Not available until** Feb 4 | **Due** Feb 18, 2018 at 10:59pm

Extra Credit #4
Week 8: One-Way Analysis of Variance and Correlation Module | **Not available until** Feb 18 | **Due** Mar 4, 2018 at 10:59pm



Extra Credit Problem #1 (5 points)

This problem illustrates quartile calculations using random samples of different sizes from the standard normal distribution.

Use *set.seed(1237)* and *rnorm(n, mean = 0, sd = 1)* with *n* = 10, *n* = 30, *n* = 100 and *n* = 300 to draw four different random samples from the standard normal distribution. Reset *set.seed(1237)* prior to drawing each of the four samples.

For each sample, calculate the first, second and third quartile using *quantile()*. Use "type = 2" (method used in Business Statistics) and "type = 7" (R default) and generate quartiles for each.

Display the results. The quartiles for the standard normal distribution are -0.6745, 0.0 and +0.6745 as shown using *qnorm(c(0.25, 0.5, 0.75), mean = 0, sd = 1, lower.tail = TRUE)*. Note below.

```
qnorm(c(0.25, 0.5, 0.75), mean = 0, sd = 1, lower.tail = TRUE)
```

```
## [1] -0.6744898  0.0000000  0.6744898
```

Take note of the results for the first and third quartile. Compare the computed results between the two methods (type = 2 and type = 7). Comment on the rate of convergence for these estimates as the sample size is increased. What does this exercise indicate about describing a population distribution with samples?

```
# Add your set.seed(), rnorm() and quantile() code to this code 'chunk':
```

Types and Some Causes of Outliers

Anscombe (1960) defined two categories:

1) **Those arising from error in the data, and**
2) **those arising from inherent variability.**

Not all outliers are illegitimate contaminants, and not all illegitimate observations show up as outliers.

Some causes:

- Human error
  - Recording
  - Entry
- Intentional mis-reporting
  - Sabotage the study
- Sampling error
  - Sampled the wrong population
- Standardization failure
  - Intruding factors
- Faulty distributional assumptions
  - Unanticipated outcomes
- Legitimate variation
  - Appearance of extreme values

Nuisance, error or legitimate data?

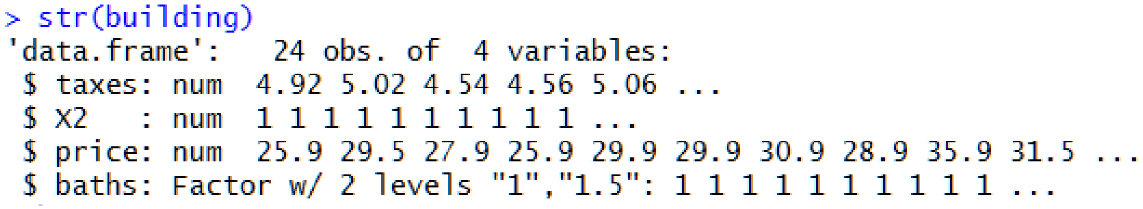
Building Prices Data

Variable	Definition
Y	Sales price of the house in thousands of dollars
X1	Taxes in thousands of dollars
X2	Number of bathrooms (1.0 and 1.5)

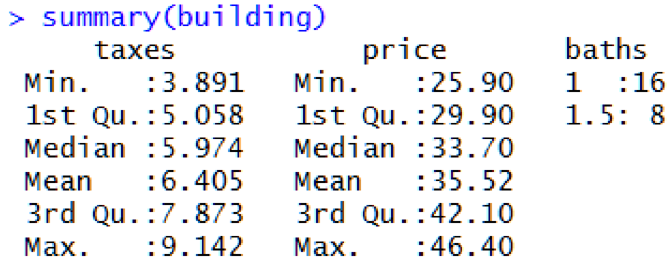
24 observations

Look at sales price relative to taxes and number of bathrooms.

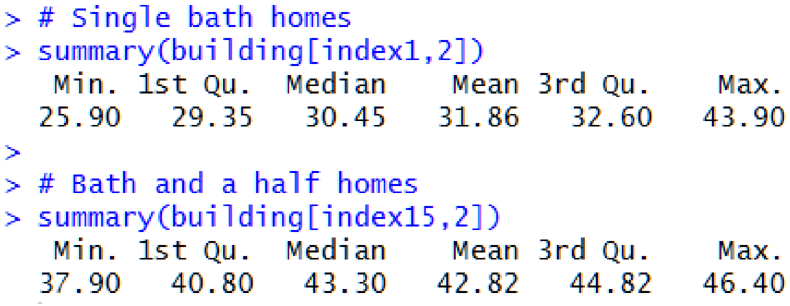
What types of variables are these?



```
> str(building)
'data.frame':   24 obs. of  4 variables:
 $ taxes: num  4.92 5.02 4.54 4.56 5.06 ...
 $ X2   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ price: num  25.9 29.5 27.9 25.9 29.9 29.9 30.9 28.9 35.9 31.5 ...
 $ baths: Factor w/ 2 levels "1","1.5": 1 1 1 1 1 1 1 1 1 1 ...
```



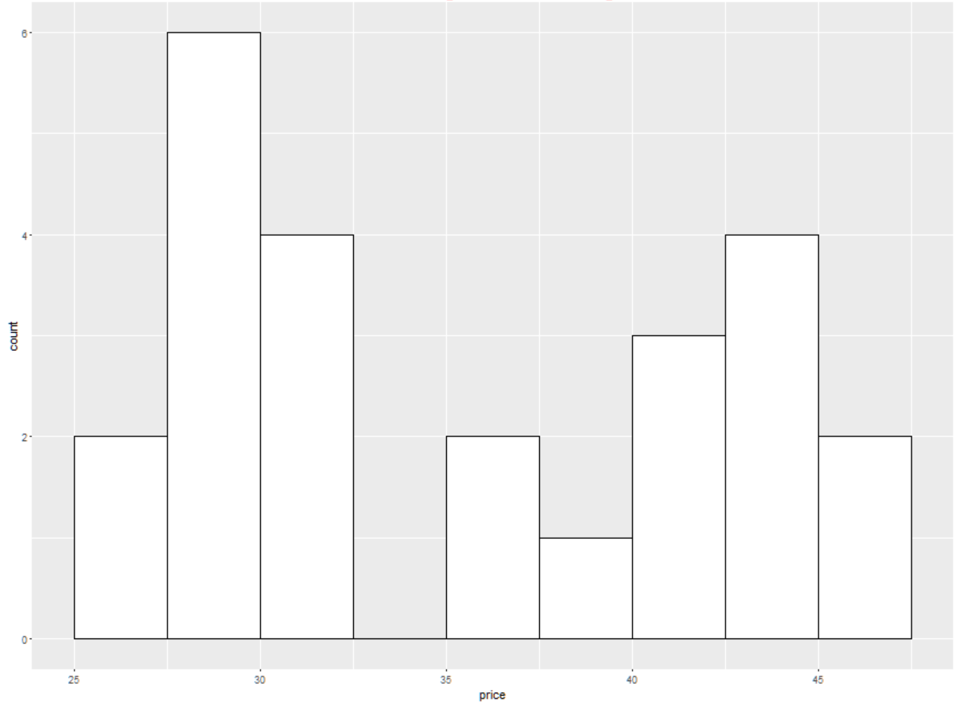
```
> summary(building)
     taxes            price          baths  
 Min.   :3.891   Min.   :25.90   1  :16  
 1st Qu.:5.058   1st Qu.:29.90   1.5: 8  
 Median :5.974   Median :33.70           
 Mean   :6.405   Mean   :35.52           
 3rd Qu.:7.873   3rd Qu.:42.10           
 Max.   :9.142   Max.   :46.40           
```



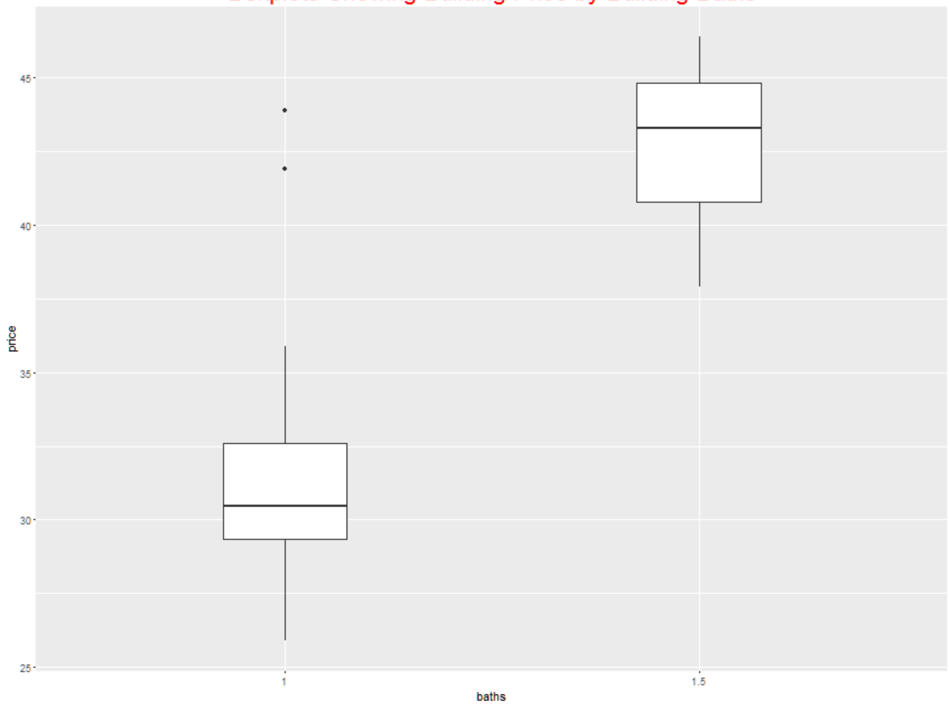
```
> # Single bath homes
> summary(building[index1,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  25.90   29.35   30.45   31.86   32.60   43.90 
> 
> # Bath and a half homes
> summary(building[index15,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  37.90   40.80   43.30   42.82   44.82   46.40 
```

Split Distribution

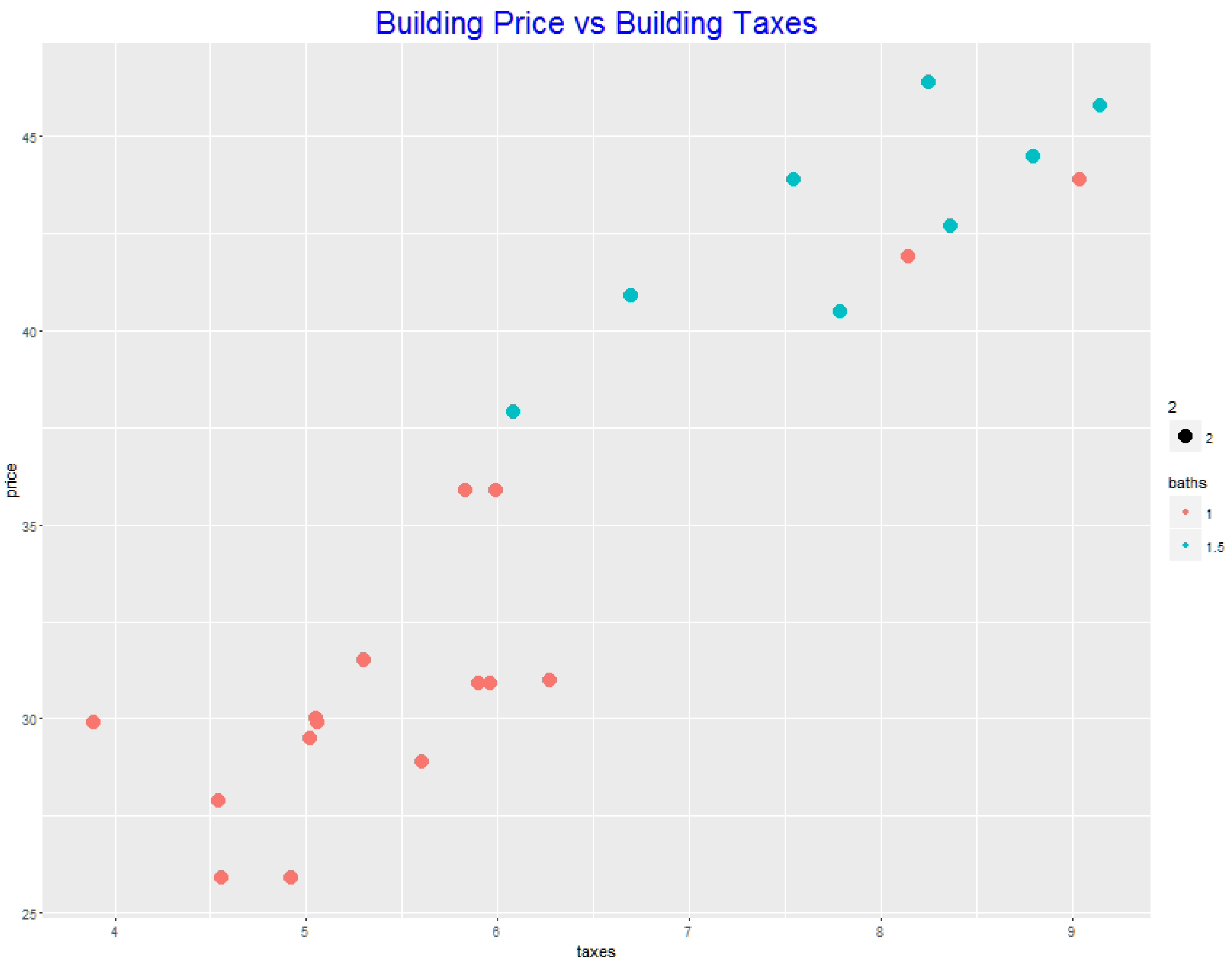
Overall Histogram of Building Price



Boxplots Showing Building Price by Building Baths

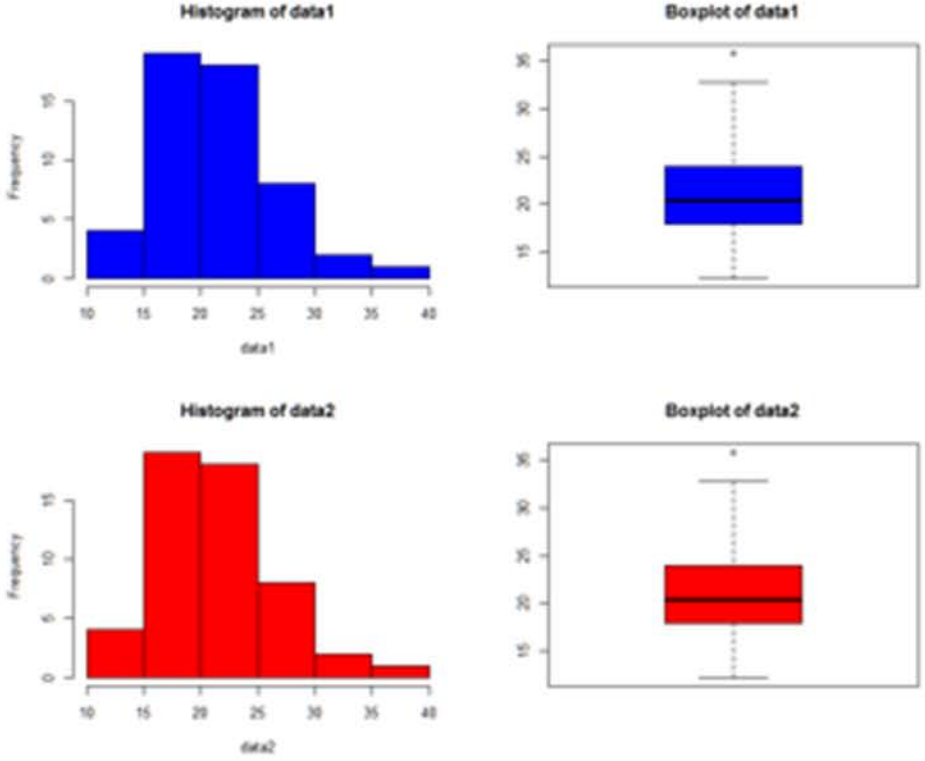


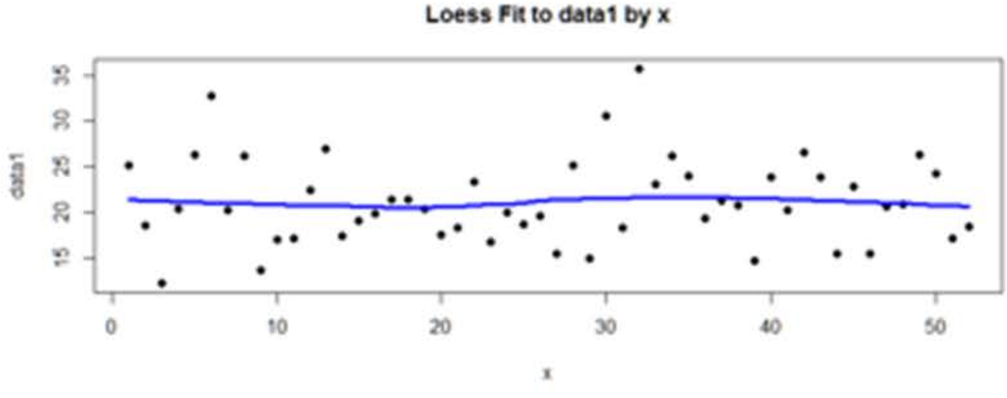
All Variation is Caused!

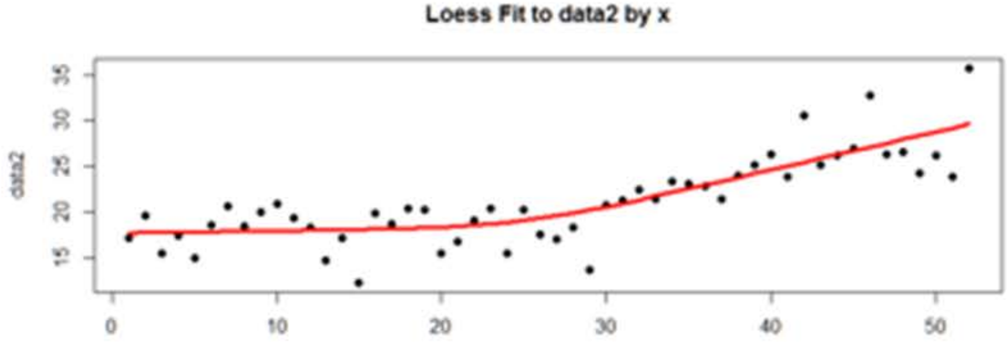


Unusual data points should not be discarded or disregarded without investigation and consideration of possible causes. There could be a simple explanation. Exclusion of data from an analysis should be reported.

Have You Done a Through EDA?







Things to Remember

- If data are not collected properly and lack adequate quality, it may not be possible to salvage them with any statistical analysis.
  - For example, a self-selected voluntary sample may not represent the population under study.

- Simple data displays can be misleading. It is essential to maintain an attitude of healthy skepticism when reviewing data displays and study results. Here are some links.

  https://www.perceptualedge.com/articles/ie/data_presentation.pdf

  http://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/

  http://iase-web.org/islp/documents/Media/How%20To%20Avoid.pdf

  https://cseweb.ucsd.edu/~ricko/CSE3/Lie_with_Statistics.pdf

- "Essentially all models are wrong, but some are useful." George E. P. Box
  - Our goal is development of useful predictive models validating and updating them as new information comes available.

Some Sync Session Learning Points

- EDA is visual, guided by the data, informative, investigative, but is rarely confirmatory.

- EDA includes consideration of changes in data over time.

- The sample mean is a more efficient measure of central tendency than the median for symmetrically distributed data with few outliers.

- The sample median is preferred for describing a typical observation from an asymmetric distribution with many outliers.

- There is no consensus on the best way to estimate a quantile from sample data. *Business Statistics* uses Type 2 in R, and R uses Type 7 as a default.

- A trimmed mean does not estimate the population mean for an asymmetric distribution with or without outliers.

- Box plots are useful for identifying outliers.

- Outliers can arise from inherent variability in the phenomenon being studies. Outliers are not always errors and should be investigated before elimination.

- **All variation is caused.**

Test #1

Ten questions, 25 points based on readings-

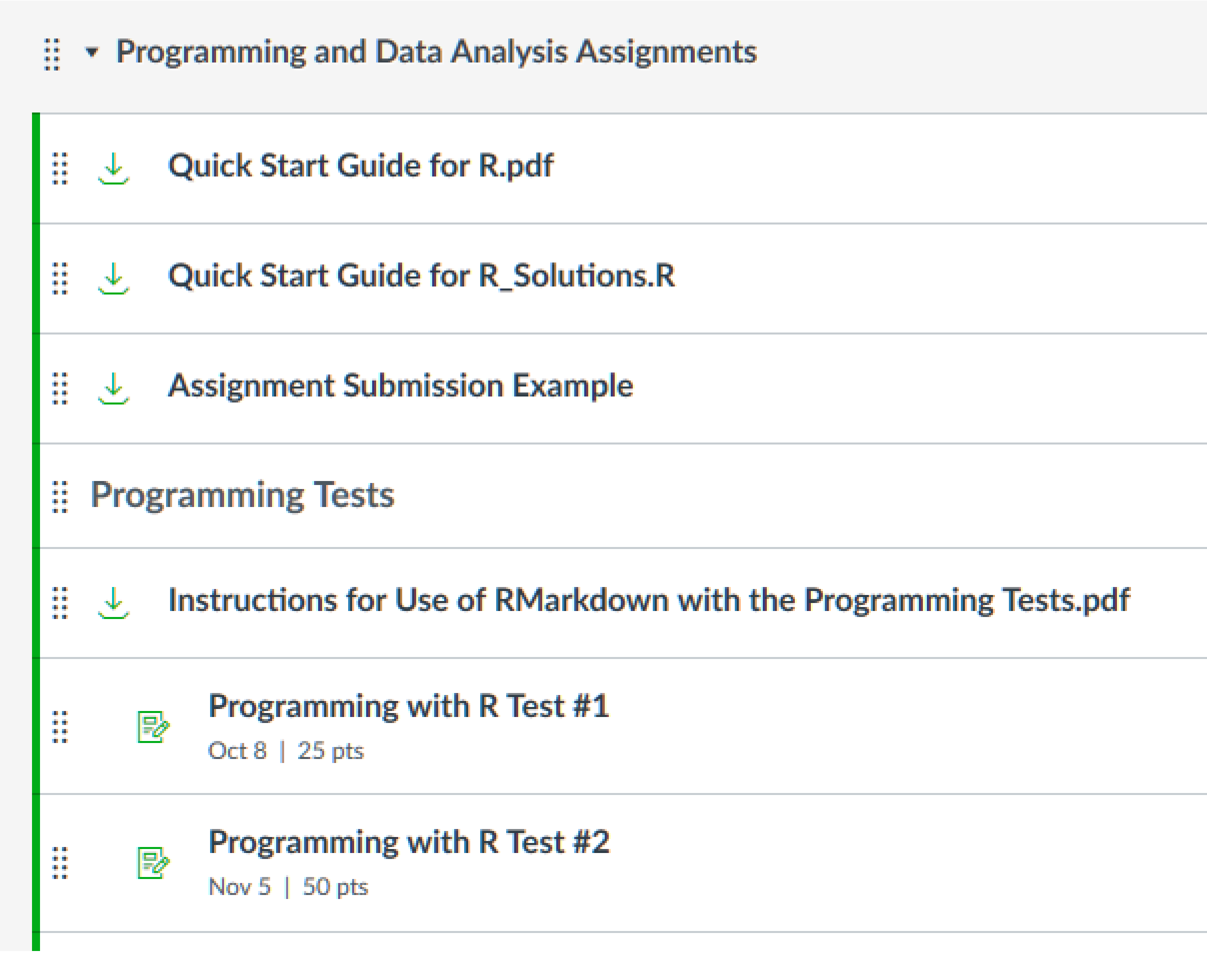
- Measurement
  - Levels of measurement
  - Types of data
- Measures of location
  - Mean
  - Median
  - Mode
  - Trimmed Mean
- Measures of spread
  - Variance
  - Standard deviation
  - Range
  - Mean absolute deviation
  - Coefficient of variation
- Chebyshev's Theorem
- Interpretation of measures

Open book with no time limit. Can leave test (without submitting) and return to continue.

One attempt per question. Correct answer is shown. Automatically graded by Canvas.

First Programming Test

Using RMarkdown



Programming and Data Analysis Assignments

Quick Start Guide for R.pdf

Quick Start Guide for R_Solutions.R

Assignment Submission Example

Programming Tests

Instructions for Use of RMarkdown with the Programming Tests.pdf

Programming with R Test #1

Oct 8 | 25 pts

Programming with R Test #2

Nov 5 | 50 pts

Instructions for use of RMarkdown and a video are included along with each programming test.

Instructions for Use of R Markdown

Programming test answers are to be submitted using R Markdown. R Markdown provides a convenient way to prepare a report saving the author time; and, facilitating review and grading of the work.   R Markdown templates have been prepared, one for each test. All that is required is to place code solutions into the spaces provided on the template, "Knit" the file to prepare an .html document and save both. **The R Markdown program and .html document are to be submitted for grading. Do not submit the .md or other intermediate files. No other formats will be accepted for the programming tests.**

RStudio is strongly recommended for using R Markdown. The R GUI may be used, but this requires making certain arrangements. To use the R GUI, the "R Markdown" package must be downloaded and installed from CRAN, via *install.packages()*, and loaded, via *library()*. The *render()* function may be used to knit the desired output from an existing R Markdown (.Rmd), markdown (.md) or R script (.R) file. The documentation page for *render()* details the required, "input", and optional arguments. Please note that R Markdown (.Rmd) files cannot be opened in the R Editor. If you are not using RStudio, you may use another text editor, e.g. Notepad++, *or* change the .Rmd file to type .R and open with the R Editor. If you choose the latter option, you will need to change the file back to an .Rmd prior to rendering. Neither of these options provides the auto-completion or syntax high-lighting/coloring that RStudio does. Consequently, RStudio is recommended. It may be downloaded from https://www.rstudio.com/.

To familiarize yourself with RStudio, R Markdown and the steps involved, please watch the following six minute Panopto video located at:

https://northwestern.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=099f4232-5267-4064-a414-a86300e9a7e0.

Here are links for those of you who want more information on R Markdown:

rmarkdown.rstudio.com/

https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf