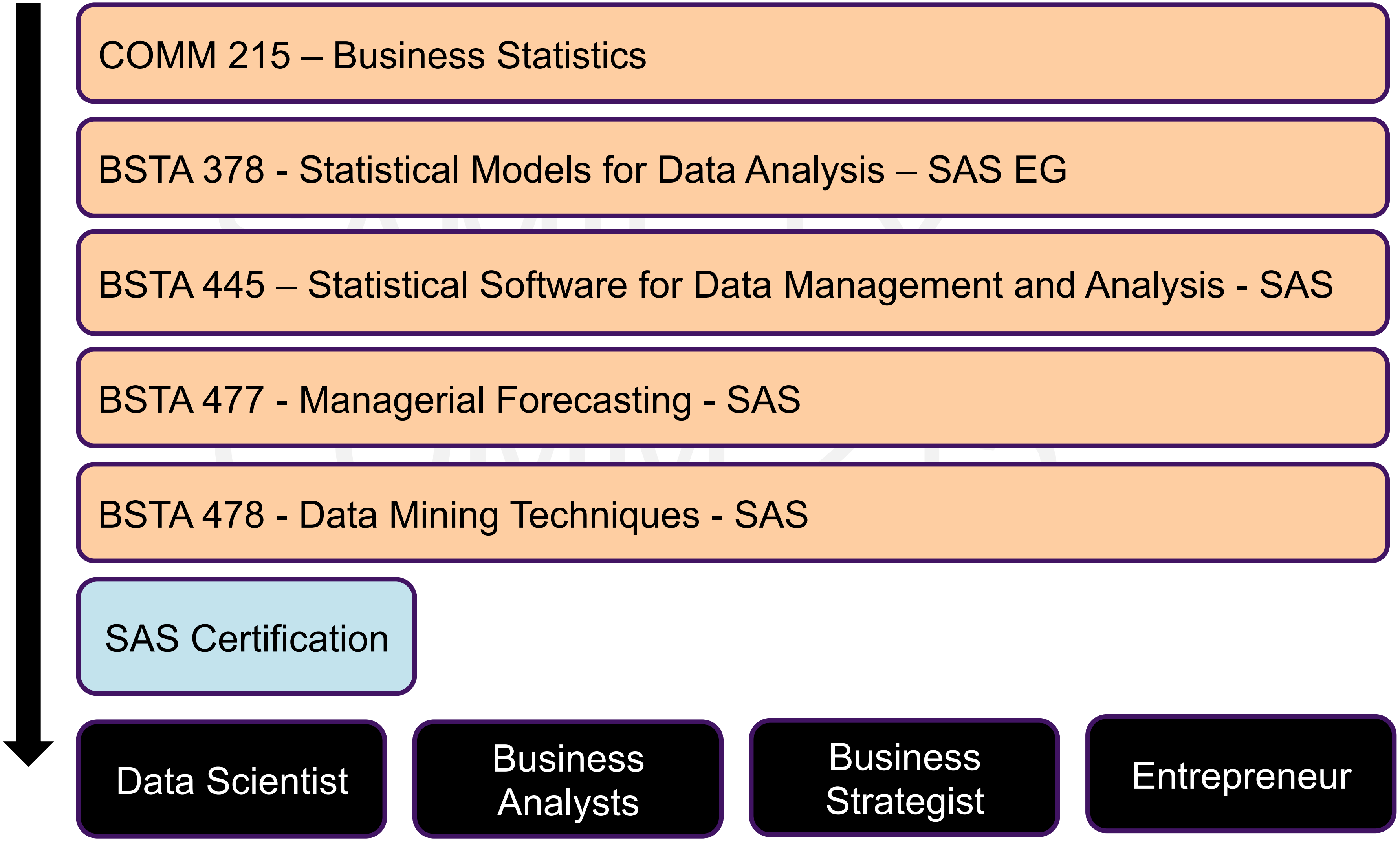
COMM215

First the Foundation, then Innovation

LESSON ONE –INTRODUCTION

SAMIE L.S. LY

AFTER THIS CLASS



COMM 215 – Business Statistics

BSTA 378 - Statistical Models for Data Analysis – SAS EG

BSTA 445 – Statistical Software for Data Management and Analysis - SAS

BSTA 477 - Managerial Forecasting - SAS

BSTA 478 - Data Mining Techniques - SAS

SAS Certification

Data Scientist

Business Analysts

Business Strategist

Entrepreneur

COMM215
First the Foundation, then Innovation

DATA SOURCES LINKS

http://www.tweetbeam.com/show?query=business%20analytics

https://www.pubnub.com/developers/realtime-data-streams/twitter-stream/

http://www.darkhorseanalytics.com/

AGENDA FOR TODAY

1. Classroom Logistics
2. Getting Started
3. Introduction to Statistics (Chapter 1)

- ● Classroom Logistics
- Getting Started
- Introduction to Statistics (Chapter 1)

Course Outline

Course Book

Course Components

Plagirism

House Rules

Moodle

Connect

COURSE OUTLINE

COMM 215 SECTION I, J, DD (3 Credits)

Instructor: Samie Li Shang Ly

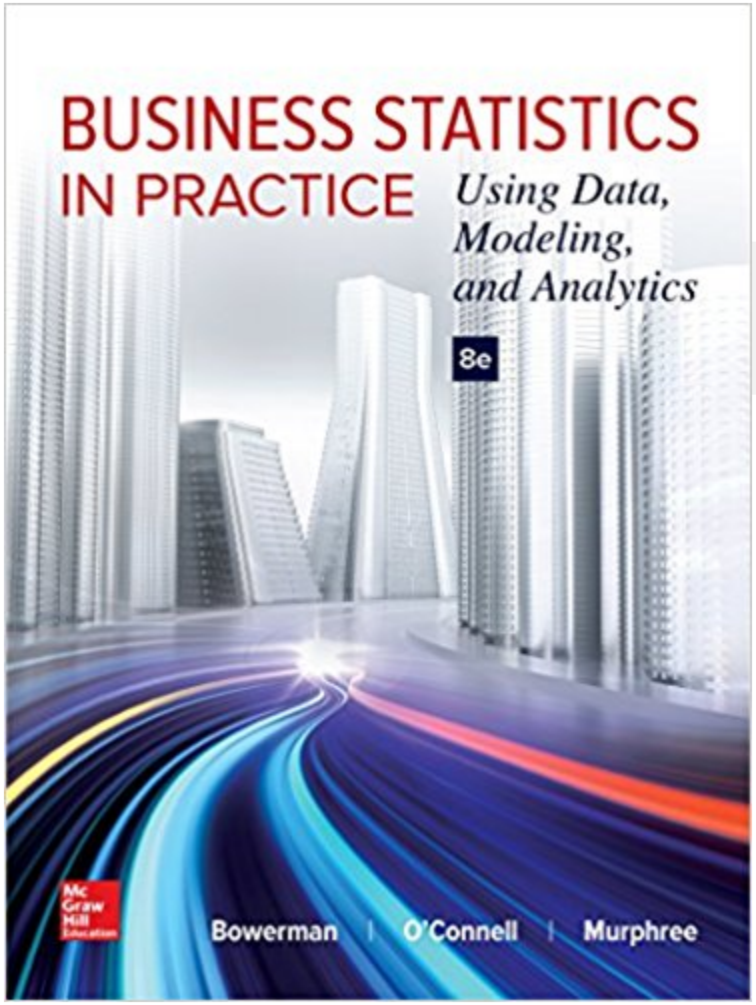
Office: MB 12.107

Office Hours: Wednesday 15:00-16:00 and

by appointment.

Email: samie.ly@concordia.ca

COURSE BOOK


BUSINESS STATISTICS
IN PRACTICE
Using Data, Modeling, and Analytics
8e
Mc Graw Hill Education
Bowerman | O'Connell | Murphree

Bookstore – Loose Leaf Format

- Hard Copy of the Book
- Soft Copy of the Book
- Megastat
- Connect Activities

Bowerman, B. L., O'Connell, R. T., Murphree, E., Huchendorf, S. C., & Porter, D. C. (2003).
Business statistics in practice
(pp. 728-730). New York: McGraw-Hill/Irwin.

COURSE COMPONENTS

Evaluation	Weight	Notes
Quizzes (Best 4 of 6)	10%	Best 4 out of 6 on Moodle
Midterm Exam	25%	February 18th, 14:00-17:00
Final Exam	50%	Minimum of 45% to pass the course
Study Activities	5%	On Connect
Case Analysis	10%	Groups of 1 to 5 person

COURSE EVALUATION – PLAGIRISM

Plagiarism

- Offense under the Academic Code of Conduct “ the presentation of the work of another person as one’s own or without proper acknowledgement.”

DO NOT COPY, PARAPHRASE OR TRANSLATE ANYTHING FROM ANYWHERE WITHOUT SAYING FROM WHERE YOU OBTAINED IT!

(Source: The Academic Integrity Website: http://provost.concordia.ca/academicintegrity/plagiarism/)

HOUSE RULES

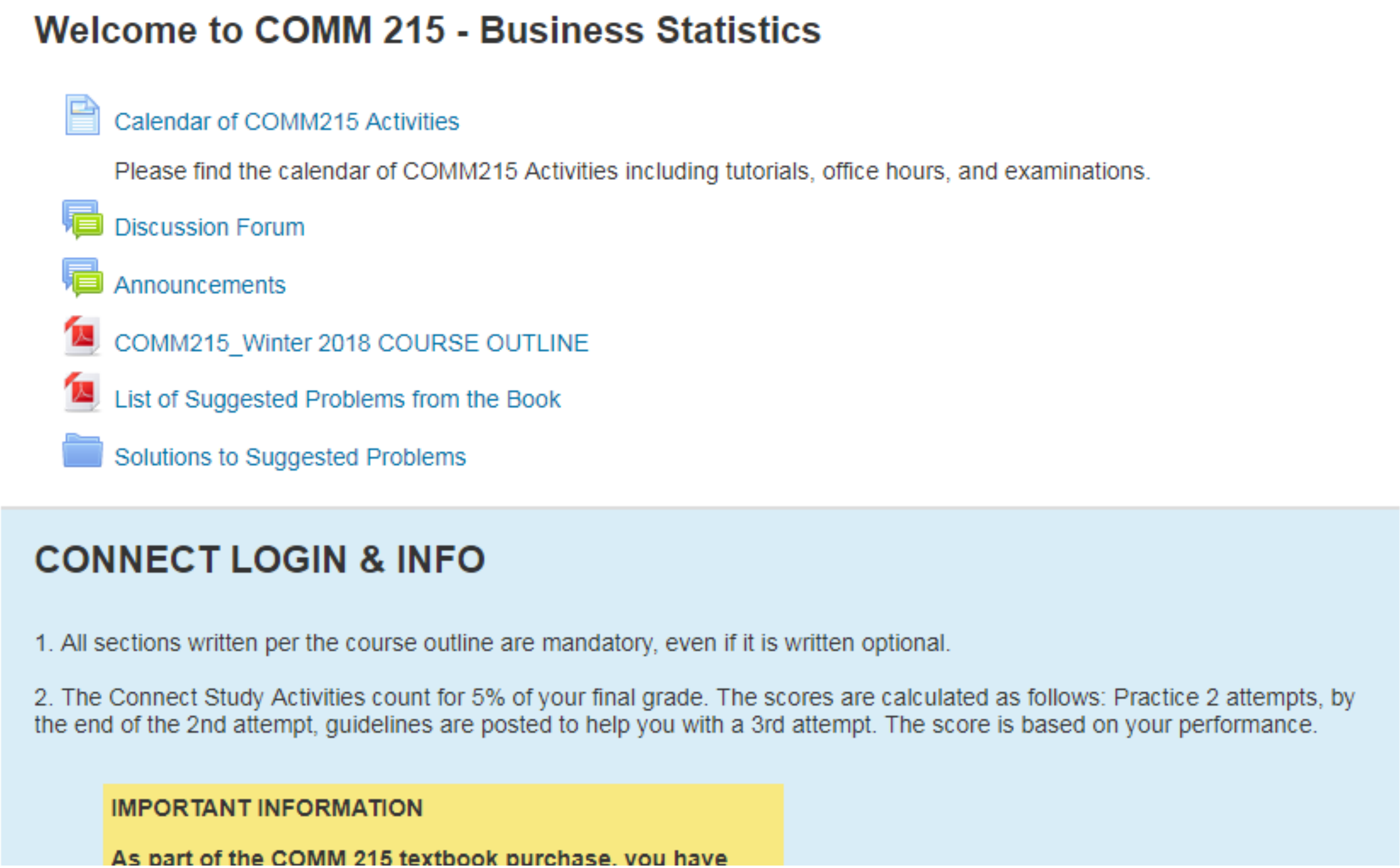
Please arrive on time for class

No Laptops*

No Cellphones*

Be careful with food consumption

MOODLE



Welcome to COMM 215 - Business Statistics

Calendar of COMM215 Activities

Please find the calendar of COMM215 Activities including tutorials, office hours, and examinations.

Discussion Forum

Announcements

COMM215_Winter 2018 COURSE OUTLINE

List of Suggested Problems from the Book

Solutions to Suggested Problems

CONNECT LOGIN & INFO

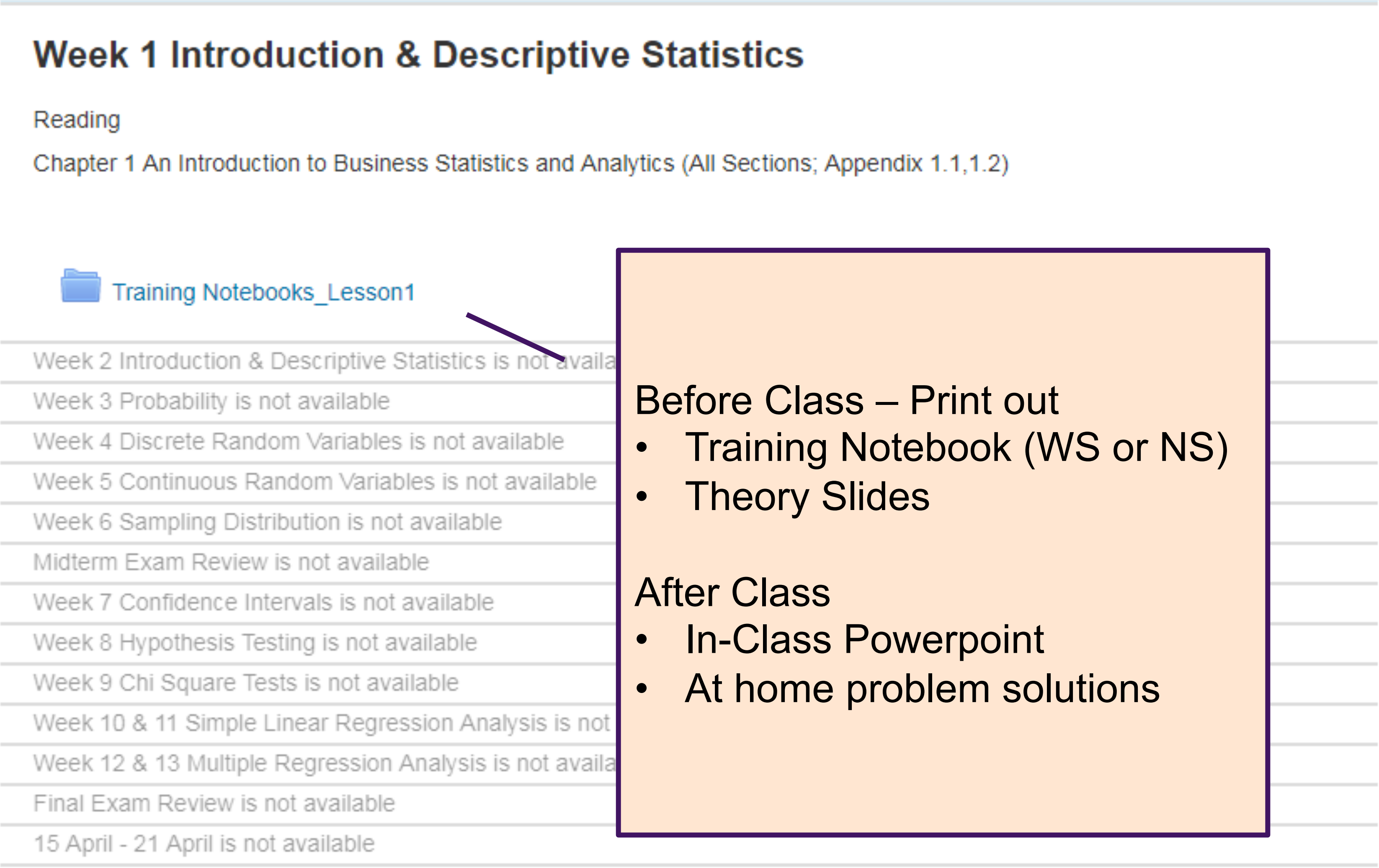
1. All sections written per the course outline are mandatory, even if it is written optional.

2. The Connect Study Activities count for 5% of your final grade. The scores are calculated as follows: Practice 2 attempts, by the end of the 2nd attempt, guidelines are posted to help you with a 3rd attempt. The score is based on your performance.

IMPORTANT INFORMATION

As part of the COMM 215 textbook purchase, you have

MOODLE - LESSONS



Week 1 Introduction & Descriptive Statistics

Reading

Chapter 1 An Introduction to Business Statistics and Analytics (All Sections; Appendix 1.1,1.2)

Training Notebooks_Lesson1

Week 2 Introduction & Descriptive Statistics is not availa

Week 3 Probability is not available

Week 4 Discrete Random Variables is not available

Week 5 Continuous Random Variables is not available

Week 6 Sampling Distribution is not available

Midterm Exam Review is not available

Week 7 Confidence Intervals is not available

Week 8 Hypothesis Testing is not available

Week 9 Chi Square Tests is not available

Week 10 & 11 Simple Linear Regression Analysis is not

Week 12 & 13 Multiple Regression Analysis is not availa

Final Exam Review is not available

15 April - 21 April is not available

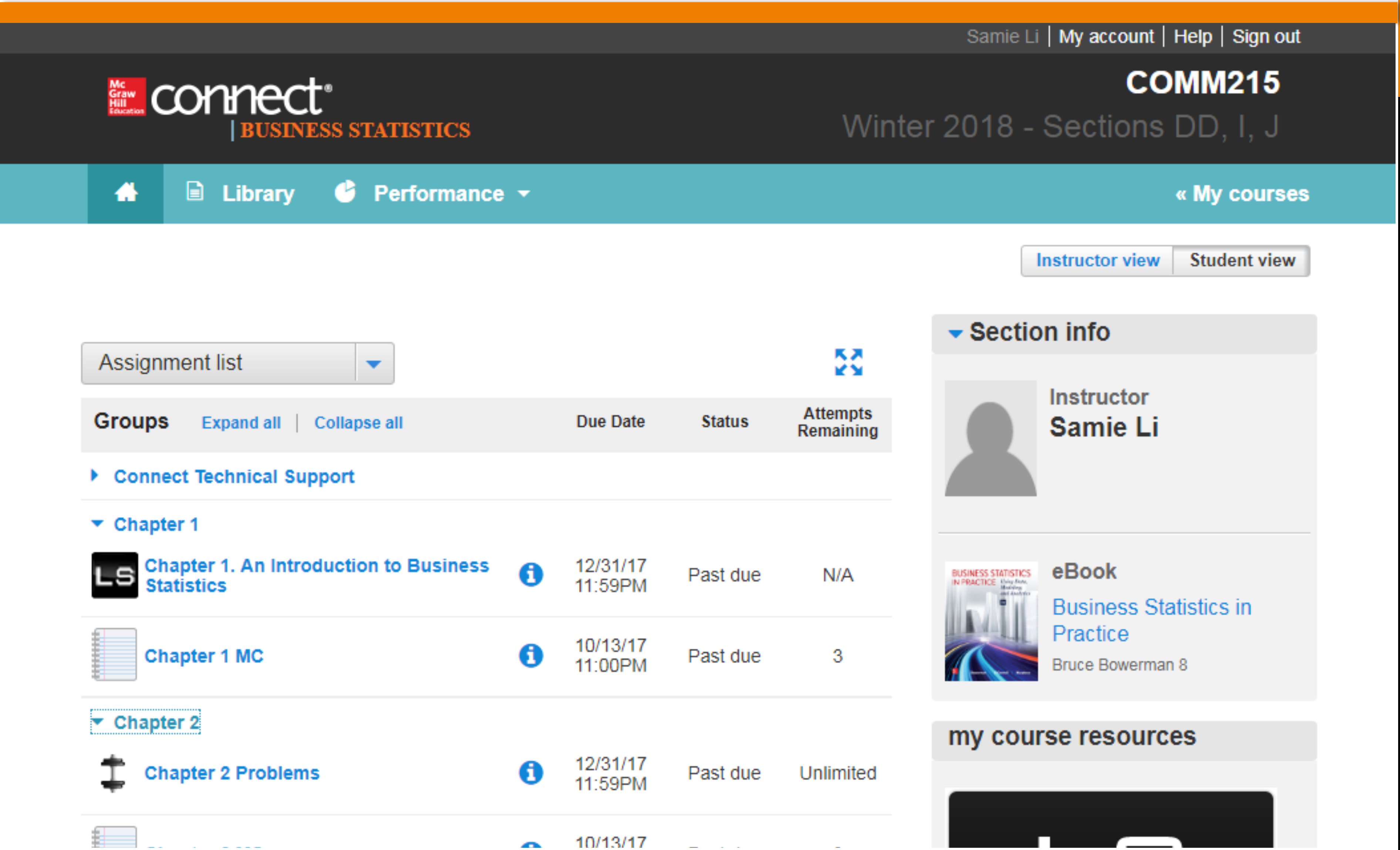
Before Class – Print out

- Training Notebook (WS or NS)
- Theory Slides

After Class

- In-Class Powerpoint
- At home problem solutions

CONNECT



Samie Li | My account | Help | Sign out

Mc Graw Hill Education connect® | BUSINESS STATISTICS

COMM215

Winter 2018 - Sections DD, I, J

Groups Expand all | Collapse all		Due Date	Status	Attempts Remaining
Connect Technical Support				
Chapter 1				
Chapter 1. An Introduction to Business Statistics	ℹ	12/31/17 11:59PM	Past due	N/A
Chapter 1 MC	ℹ	10/13/17 11:00PM	Past due	3
Chapter 2				
Chapter 2 Problems	ℹ	12/31/17 11:59PM	Past due	Unlimited
		10/13/17		

Section info

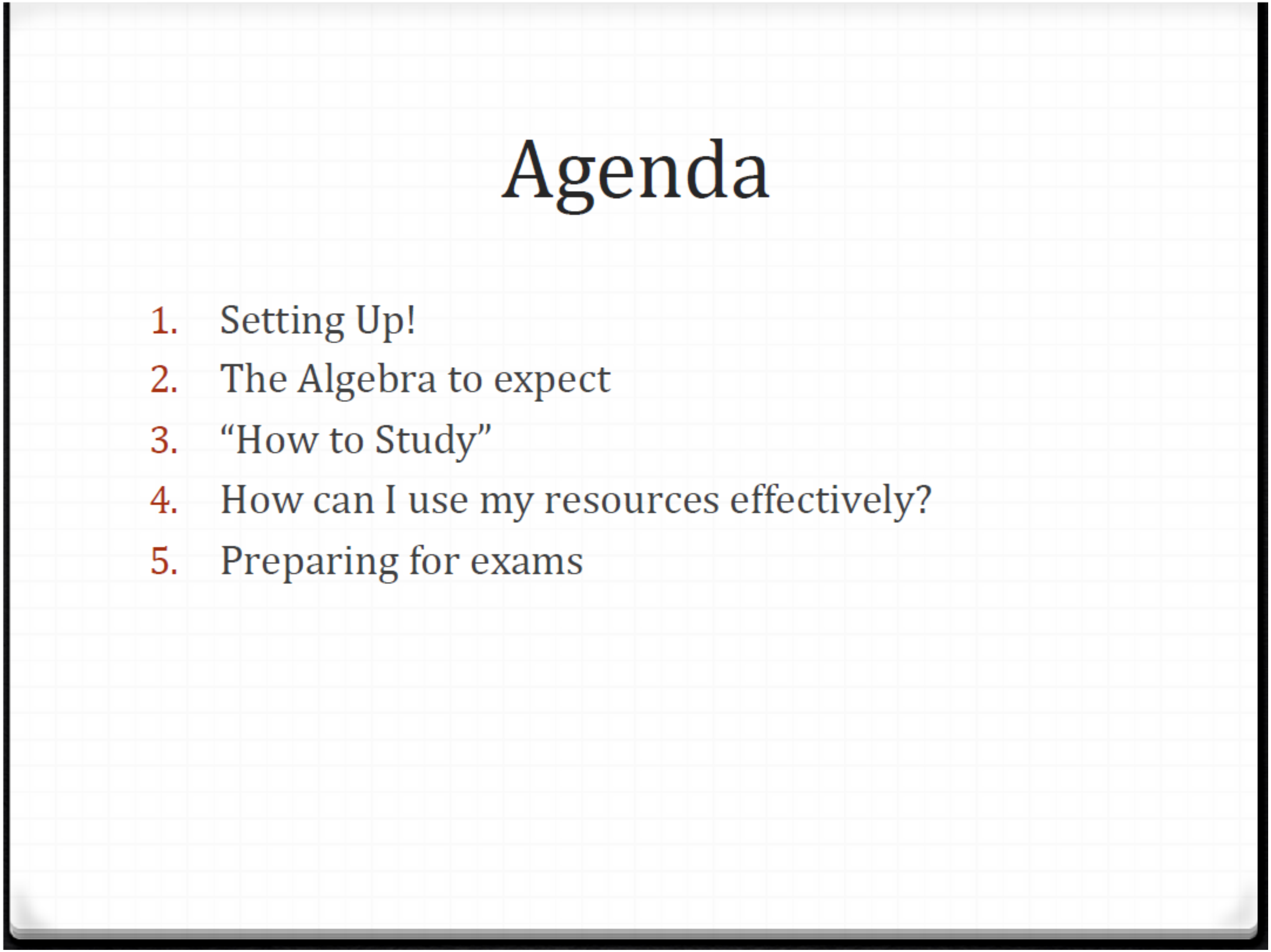
Instructor
Samie Li

eBook
Business Statistics in Practice
Bruce Bowerman 8

my course resources

- Getting Started Workshop

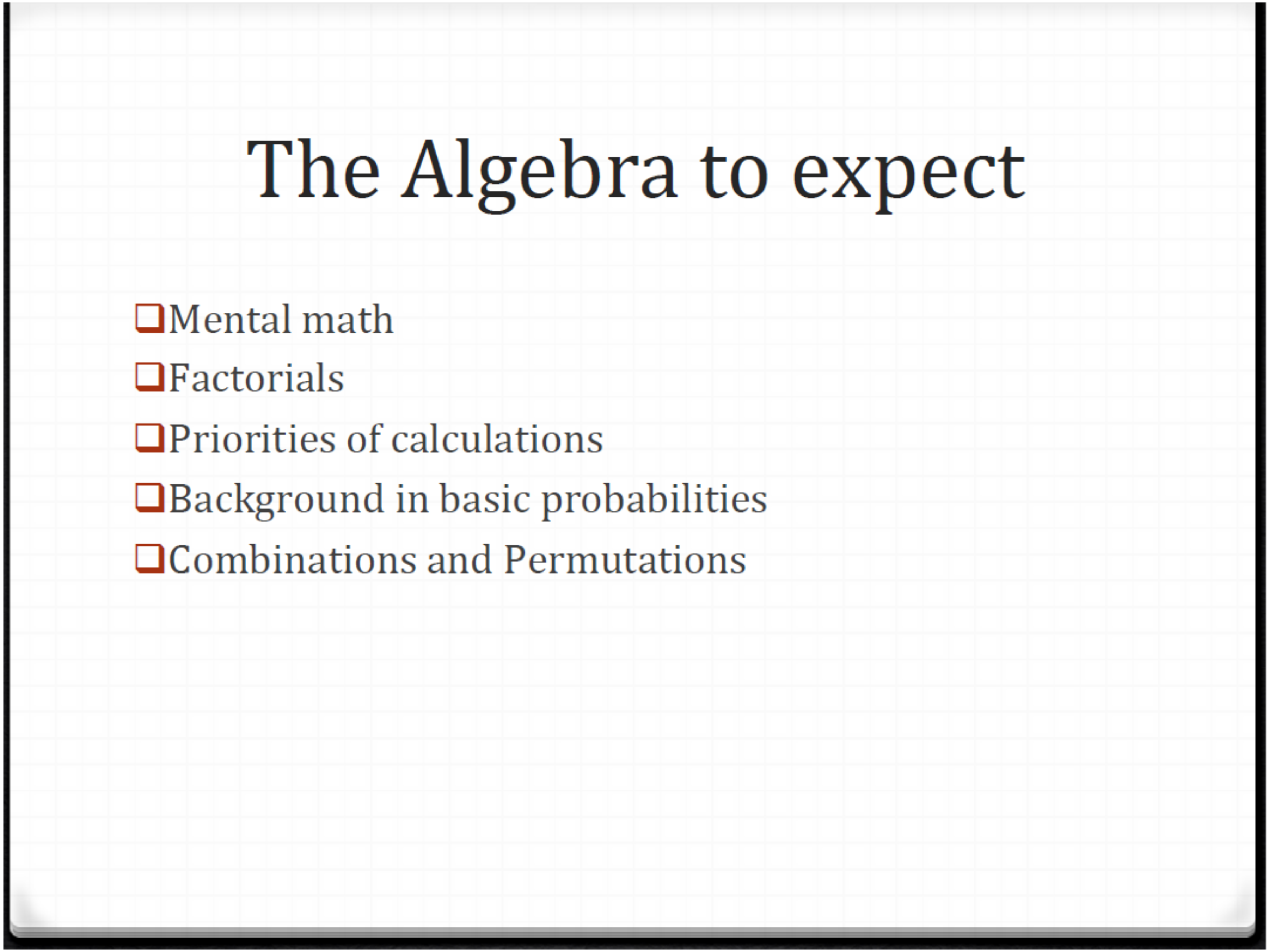
Introduction to Statistics (Chapter 1)



Agenda

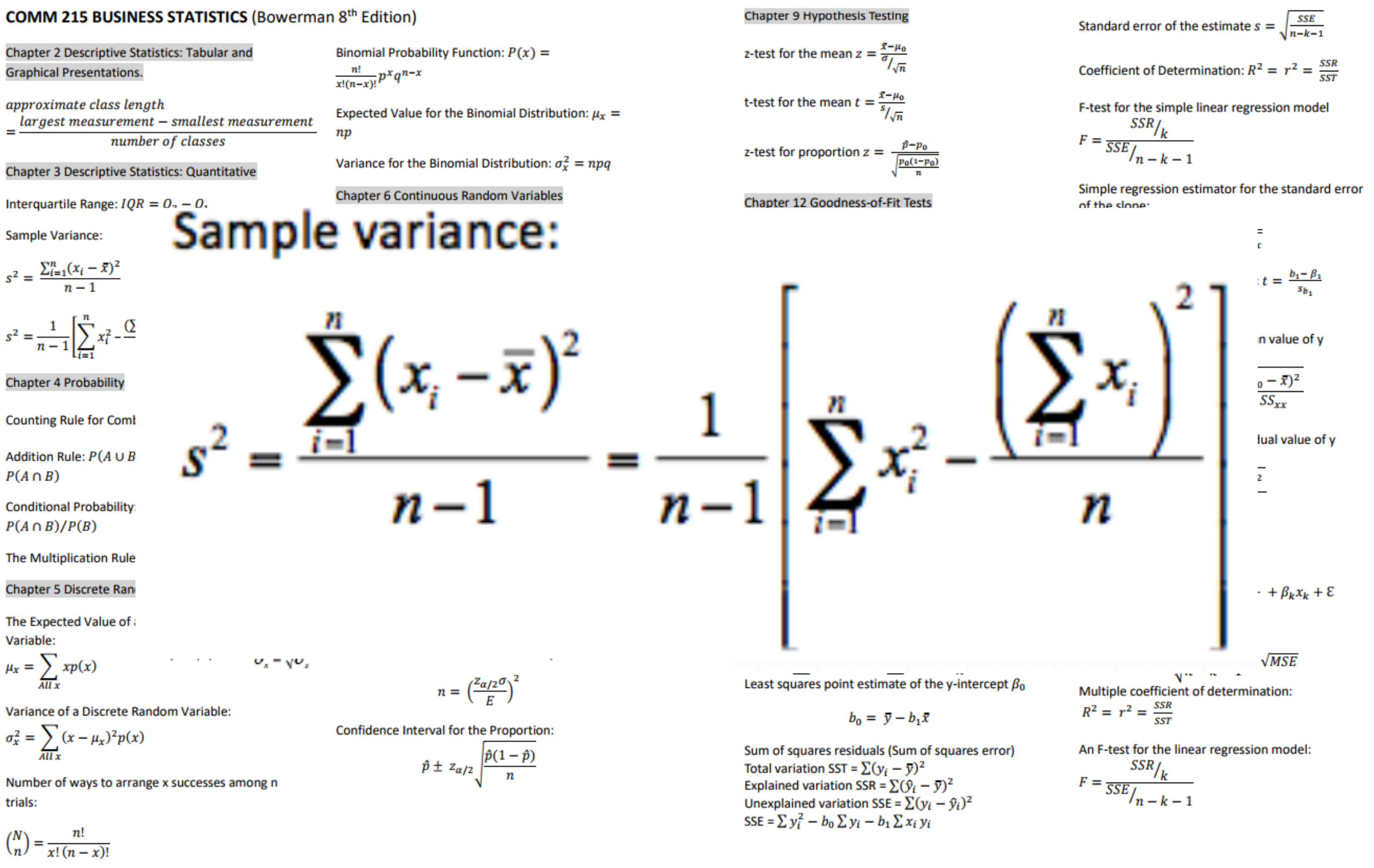
1. Setting Up!
2. The Algebra to expect
3. “How to Study”
4. How can I use my resources effectively?
5. Preparing for exams

Setting Up!



The Algebra to expect

- Mental math
- Factorials
- Priorities of calculations
- Background in basic probabilities
- Combinations and Permutations



COMM 215 BUSINESS STATISTICS (Bowerman 8th Edition)

Chapter 2 Descriptive Statistics: Tabular and Graphical Presentations.

$$approximate\ class\ length = \frac{largest\ measurement - smallest\ measurement}{number\ of\ classes}$$

Chapter 3 Descriptive Statistics: Quantitative

Interquartile Range: $IQR = Q_3 - Q_1$

Sample Variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\Sigma$$

Chapter 4 Probability

Counting Rule for Comb

Addition Rule: $P(A \cup B$

$P(A \cap B)$

Conditional Probability:

$P(A \cap B)/P(B)$

The Multiplication Rule

Chapter 5 Discrete Ran

The Expected Value of a

Variable:

$$\mu_x = \sum_{All\ x} xp(x)$$

$\sigma_x = \sqrt{\sigma_x}$

Variance of a Discrete Random Variable:

$$\sigma_x^2 = \sum_{All\ x}(x - \mu_x)^2 p(x)$$

Number of ways to arrange x successes among n trials:

$$\binom{N}{n} = \frac{n!}{x!\,(n-x)!}$$

Binomial Probability Function: $P(x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}$

Expected Value for the Binomial Distribution: $\mu_x = np$

Variance for the Binomial Distribution: $\sigma_x^2 = npq$

Chapter 6 Continuous Random Variables

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

Confidence Interval for the Proportion:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Chapter 9 Hypothesis Testing

z-test for the mean $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

t-test for the mean $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

z-test for proportion $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Chapter 12 Goodness-of-Fit Tests

Least squares point estimate of the y-intercept β_0

$$b_0 = \bar{y} - b_1\bar{x}$$

Sum of squares residuals (Sum of squares error)
Total variation SST = $\sum(y_i - \bar{y})^2$
Explained variation SSR = $\sum(\hat{y}_i - \bar{y})^2$
Unexplained variation SSE = $\sum(y_i - \hat{y}_i)^2$
SSE = $\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i$

Standard error of the estimate $s = \sqrt{\frac{SSE}{n-k-1}}$

Coefficient of Determination: $R^2 = r^2 = \frac{SSR}{SST}$

F-test for the simple linear regression model

$$F = \frac{SSR/_{k}}{SSE/_{n-k-1}}$$

Simple regression estimator for the standard error of the slope:

$t = \frac{b_1 - \beta_1}{s_{b_1}}$

n value of y

$\frac{{}_0 - \bar{x})^2}{SS_{xx}}$

lual value of y

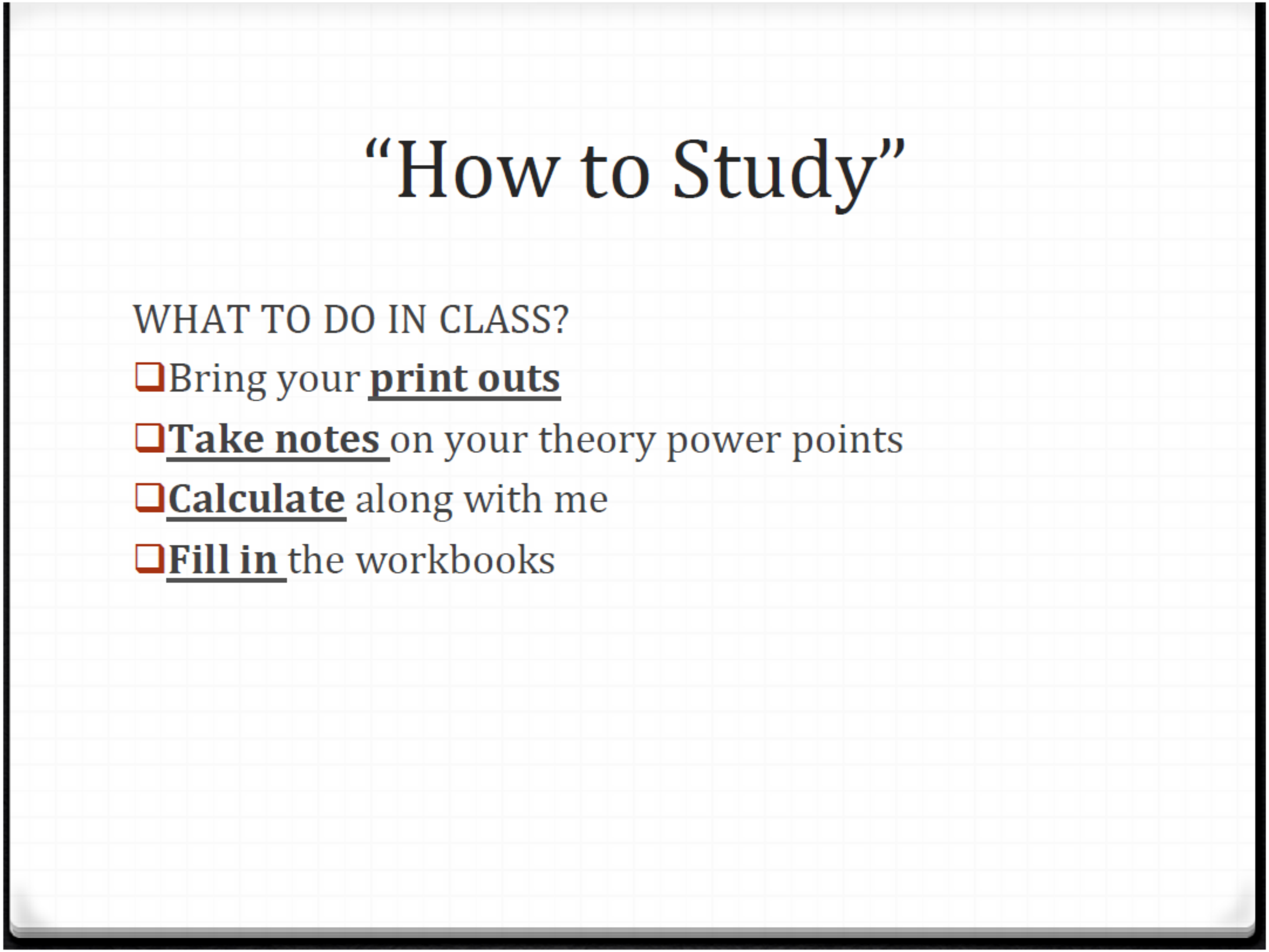
$+ \beta_k x_k + \varepsilon$

$\sqrt{MSE}$

Multiple coefficient of determination:
$R^2 = r^2 = \frac{SSR}{SST}$

An F-test for the linear regression model:

$$F = \frac{SSR/_{k}}{SSE/_{n-k-1}}$$



“How to Study”

WHAT TO DO IN CLASS?

❑Bring your **print outs**

❑**Take notes** on your theory power points

❑**Calculate** along with me

❑**Fill in** the workbooks

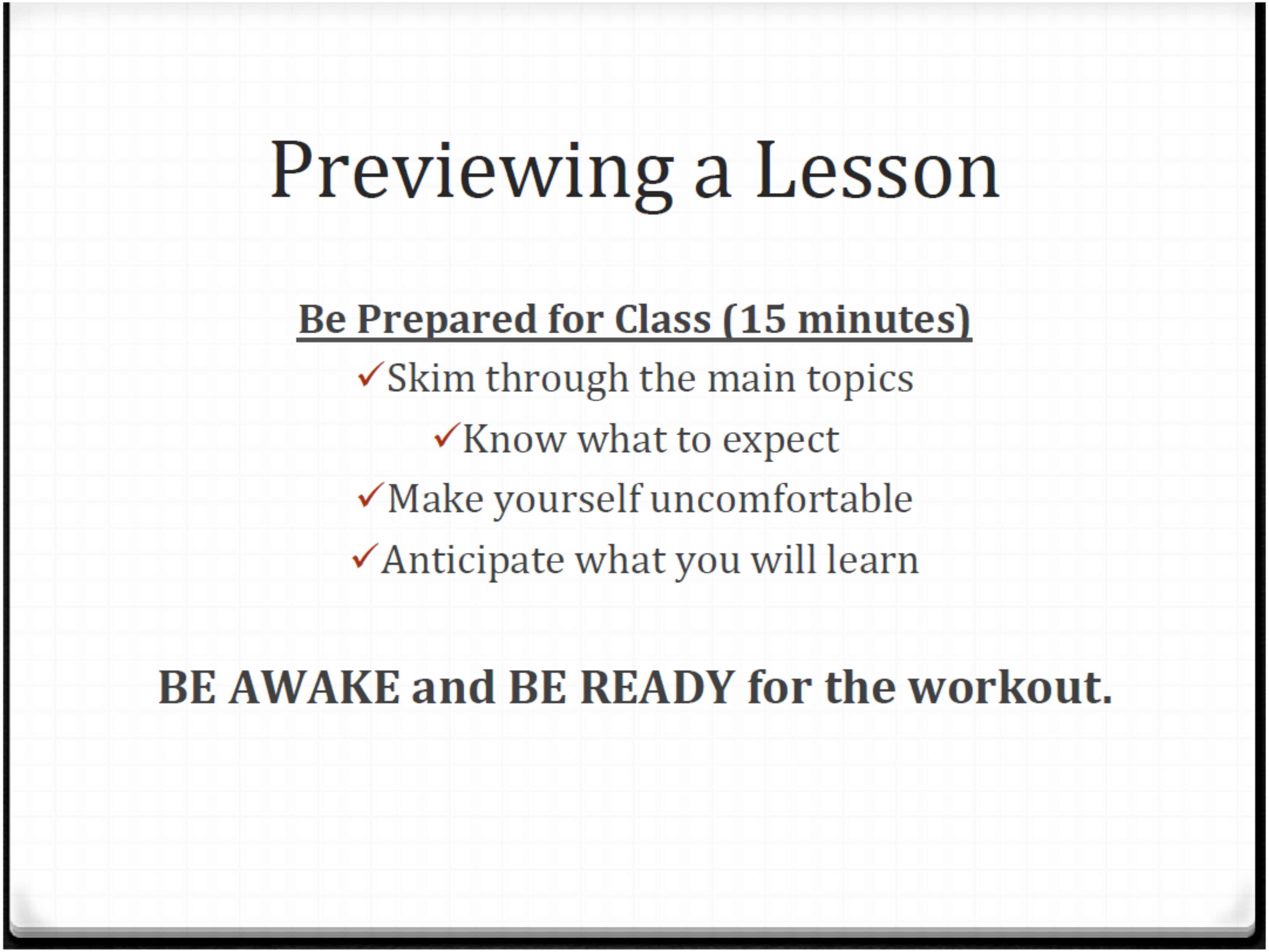
I understand most of it, but I am not sure about the difference between a Ratio and Interval Scale.

Figure it out!.. How?

After Lesson 1...

I understand everything in Lesson 1!

Test yourself!



Previewing a Lesson

Be Prepared for Class (15 minutes)

✓Skim through the main topics

✓Know what to expect

✓Make yourself uncomfortable

✓Anticipate what you will learn

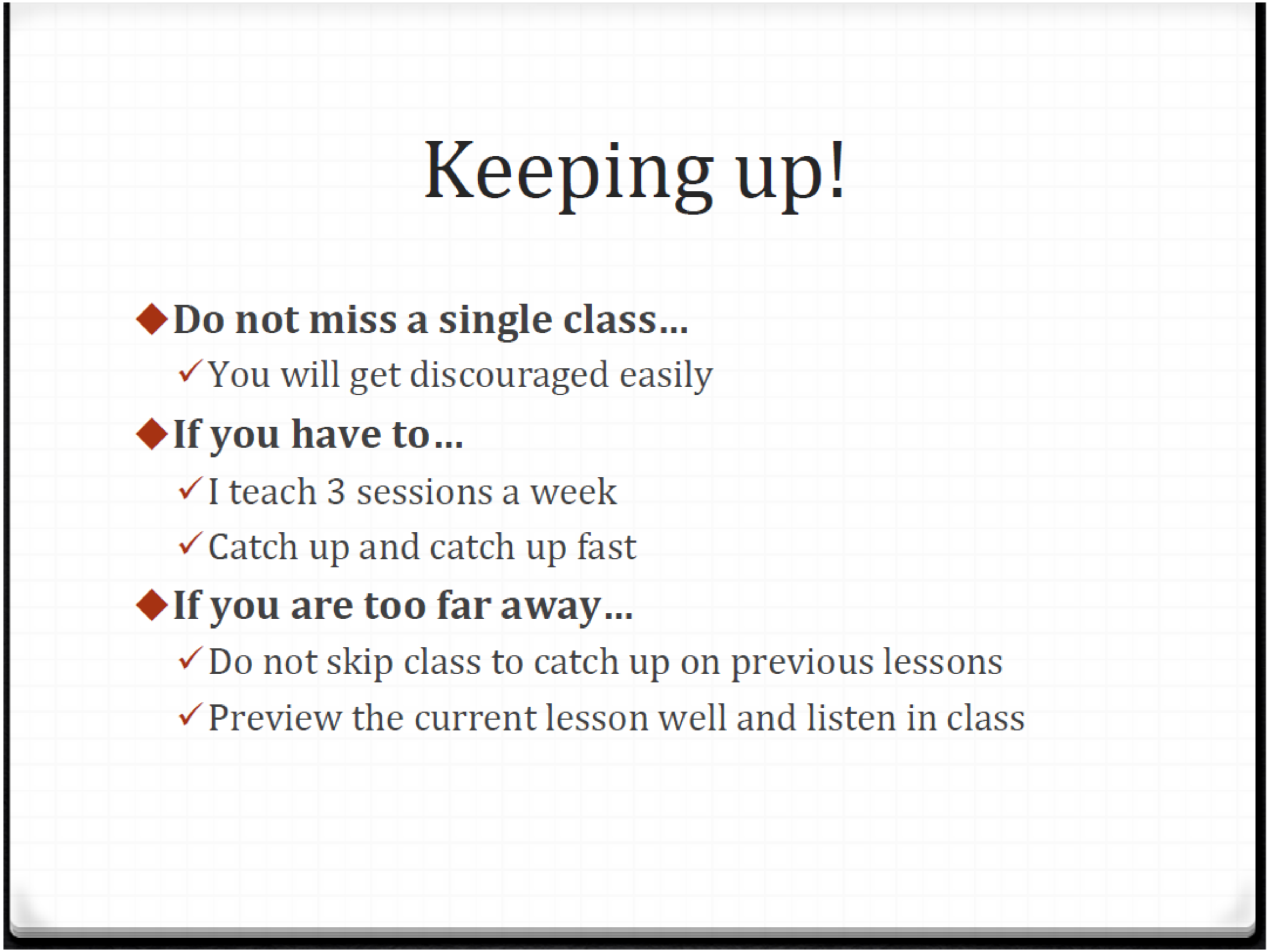
BE AWAKE and BE READY for the workout.

How to use my resources effectively?

- ✓Teacher Assistants (TA) Office Hours
- ✓Asking questions on the General Forum
- ✓Tutorials held a few weeks before exams
- ✓Scheduling with the Counseling and Development Center

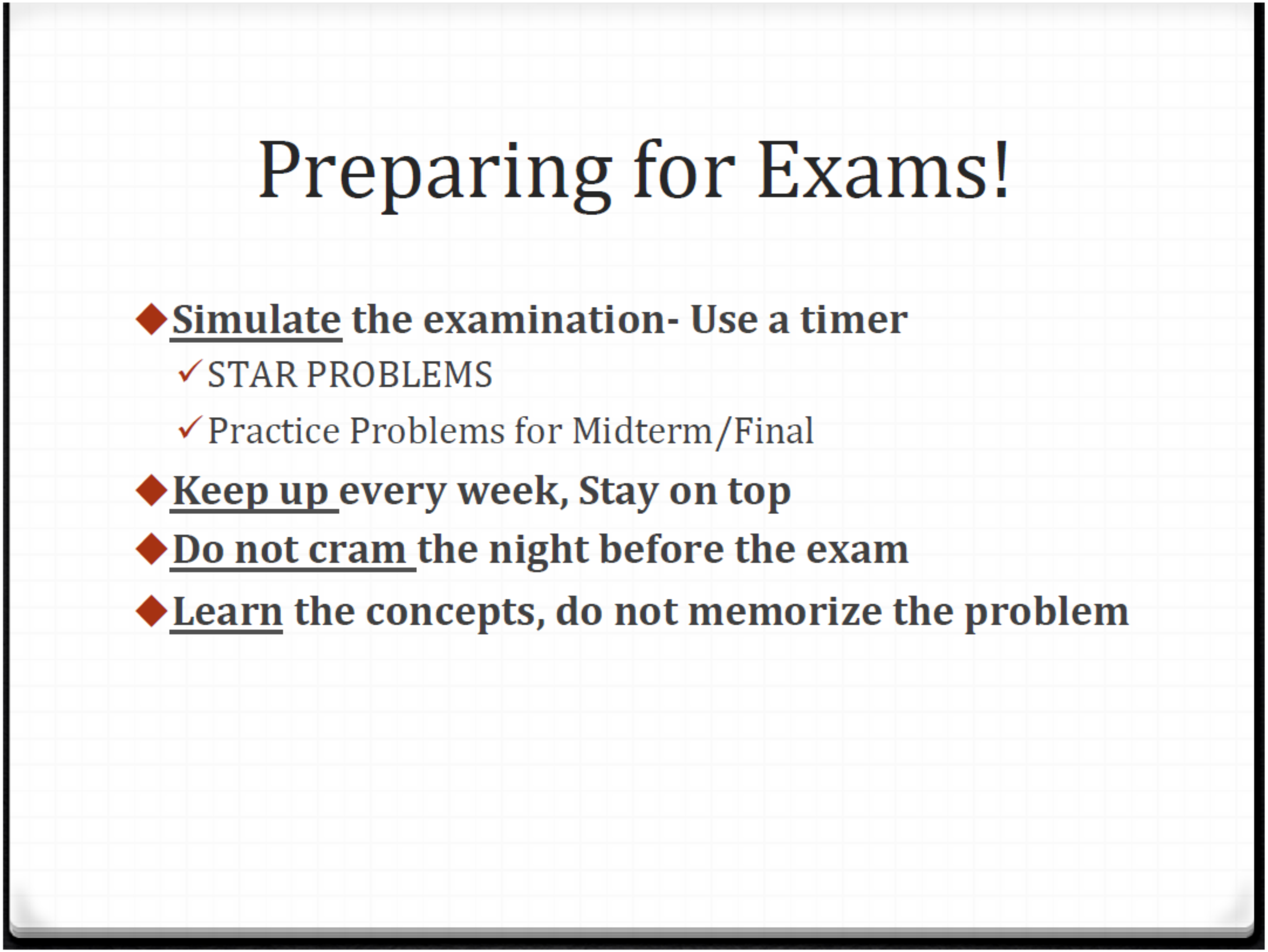
What else?

- ✓Create your own study groups!



Keeping up!

- **Do not miss a single class…**
  - ✓ You will get discouraged easily
- **If you have to…**
  - ✓ I teach 3 sessions a week
  - ✓ Catch up and catch up fast
- **If you are too far away…**
  - ✓ Do not skip class to catch up on previous lessons
  - ✓ Preview the current lesson well and listen in class



Preparing for Exams!

- **Simulate the examination- Use a timer**
  - ✓ STAR PROBLEMS
  - ✓ Practice Problems for Midterm/Final
- **Keep up every week, Stay on top**
- **Do not cram the night before the exam**
- **Learn the concepts, do not memorize the problem**

Take Away

- ✓Work smart (working hard blindly is just as ineffective)
- ✓Be Efficient

Introduction to Statistics (Chapter 1)

Content Structure

Application

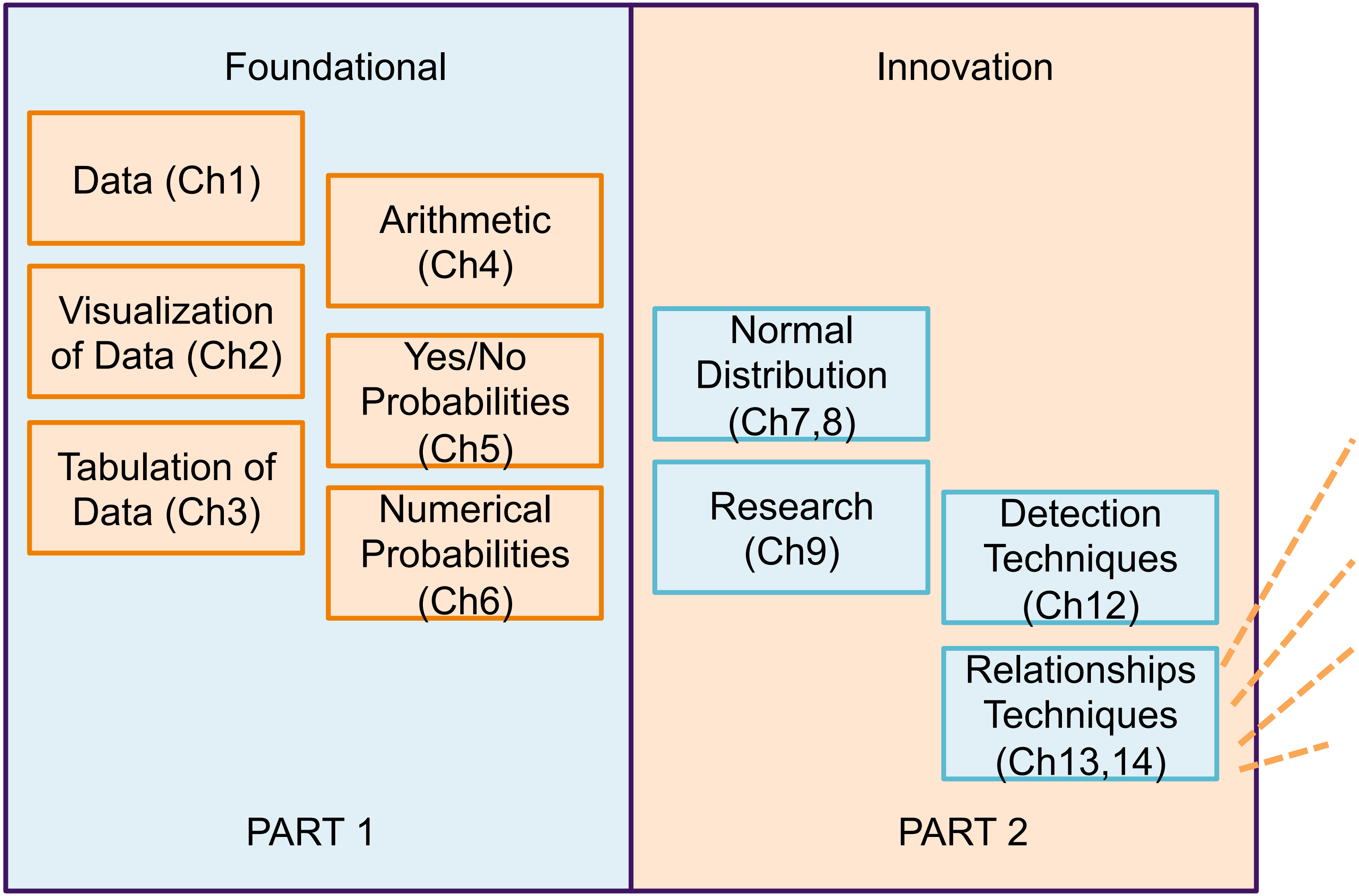
Data Collection

Data Sources

Data Analytics-Mining

Ethics

CONTENT STRUCTURE



Foundational

Data (Ch1)

Visualization of Data (Ch2)

Tabulation of Data (Ch3)

Arithmetic (Ch4)

Yes/No Probabilities (Ch5)

Numerical Probabilities (Ch6)

PART 1

Innovation

Normal Distribution (Ch7,8)

Research (Ch9)

Detection Techniques (Ch12)

Relationships Techniques (Ch13,14)

PART 2

STATISTICAL APPLICATIONS

Business needs a record of its past history

- with respect to sales, costs, sources of materials, market facilities, etc.

Statistics are used to measure progress, financial standing, and economic growth.

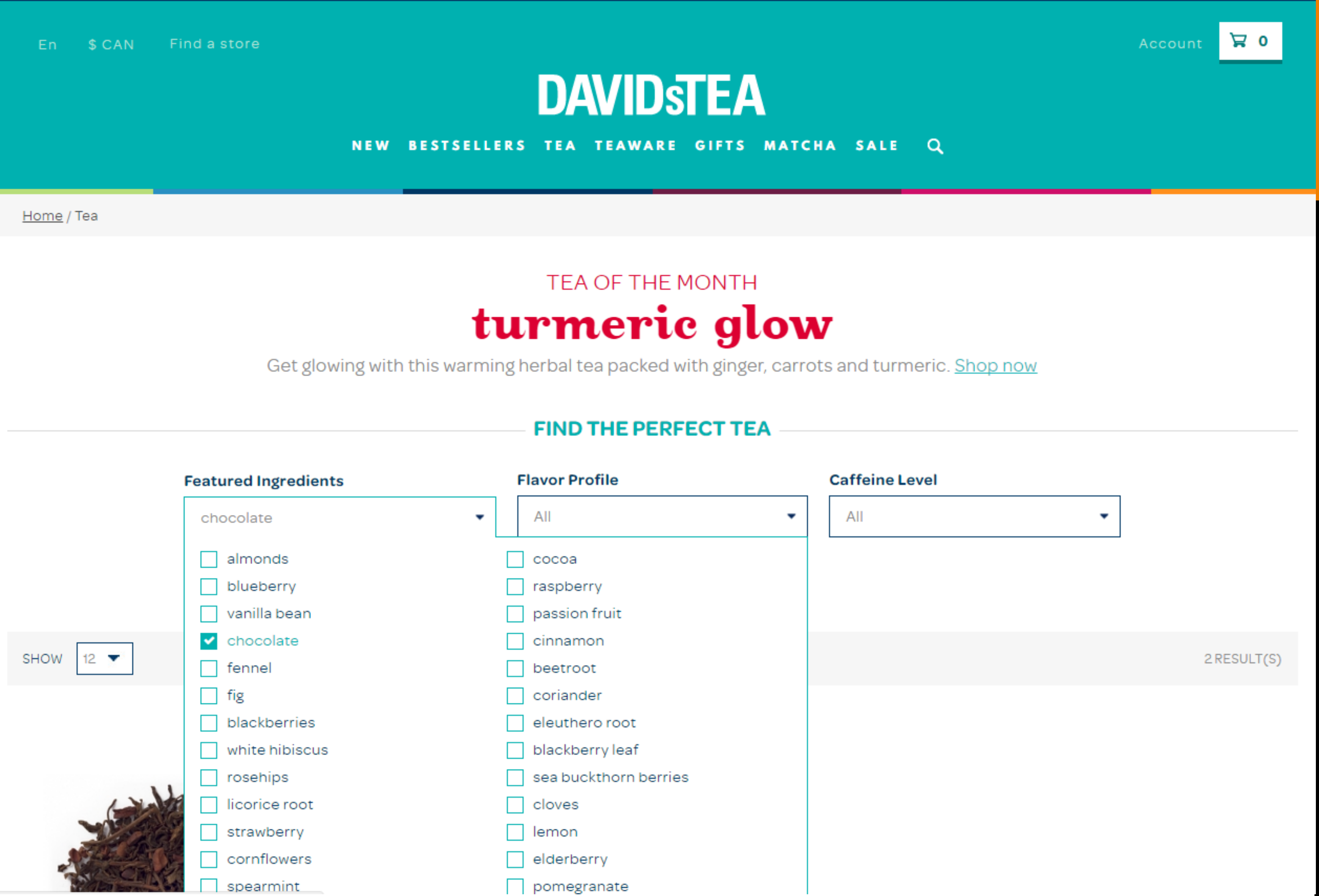
- A record of business changes- of its rise and decline and of the sequence of forces influencing it- it necessary for estimating future developments.

STATISTICAL APPLICATIONS

Our behavior in the marketplace help companies make decisions on products to be retained, dropped, or modified.

Opening Hours

**7 Days from
8:00 to 21:00**



En $ CAN Find a store

Account 0

DAVIDsTEA

NEW BESTSELLERS TEA TEAWARE GIFTS MATCHA SALE

Home / Tea

TEA OF THE MONTH

turmeric glow

Get glowing with this warming herbal tea packed with ginger, carrots and turmeric. Shop now

FIND THE PERFECT TEA

Featured Ingredients

chocolate

- □ almonds
- □ blueberry
- □ vanilla bean
- ☑ chocolate
- □ fennel
- □ fig
- □ blackberries
- □ white hibiscus
- □ rosehips
- □ licorice root
- □ strawberry
- □ cornflowers
- □ spearmint
- □ cocoa
- □ raspberry
- □ passion fruit
- □ cinnamon
- □ beetroot
- □ coriander
- □ eleuthero root
- □ blackberry leaf
- □ sea buckthorn berries
- □ cloves
- □ lemon
- □ elderberry
- □ pomegranate

Flavor Profile

All

Caffeine Level

All

SHOW 12

2 RESULT(S)

DATA COLLECTION

Tell a story

Student Name	Gender	Status	University Year	Age Bracket	Hours of Sleep in a day
Sandra	F	Full Time	1	20-23	9-11
Eric	M	Full Time	2	20-23	6-8
Brad	M	Part Time	1	27+	3-5

Elements

Variables

SCALES OF MEASUREMENT

Scales of Measurement

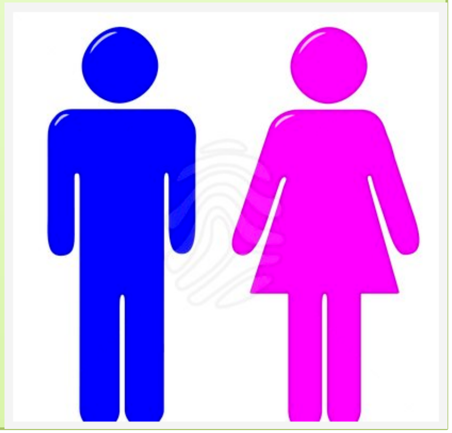
- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

SCALES OF MEASUREMENT

Nominal scale

- A variable consist of labels or names used to identify an attribute of the element

- 1: Male, 2: Female
- 1:Full-Time, 2:Part-Time



SCALES OF MEASUREMENT

Nominal + Rank

Ordinal Scale

- The data exhibits the properties of nominal data
- The order or rank of the data is meaningful.

[1] **Strongly agree**

[2] **Agree**

[3] **No Opinion**

[4] **Disagree**

[5] **Strongly disagree**

SCALES OF MEASUREMENT

No Absolute Zero

Interval Scale

- All the properties of ordinal data
- The interval between values is expressed in terms of a fixed unit of measure.
- Interval data are always numeric.
  - E.g: GPA scores, Temperature

[4.00]	[3.00]	[2.00]	[1.00]	[0.00]
A	**B**	**C**	**D**	**Fail**

SCALES OF MEASUREMENT

Absolute Zero

Ratio scale

- All the properties of interval data
- The ratio of two values is meaningful.
  - E.g: distance, height, weight .

- Bob is 140lbs, Mary is 70lbs.
- Bob is __________ heavier than Mary.

SCALES OF MEASUREMENT

Eric has a GPA is 3.00, Sam has a GPA of 1.50.

Can you say Eric is two-times smarter than Sam?

PROBLEM # 1.1

For each of the following, indicate the scale of measurement that best describes the information.

In 2008, Dell corporation had approximately 78,000 employees.

ANSWER: Ratio scale; there is an absolute zero point associated with the number of employees.

Source: Fortune, May 4, 2009, p.F-48

PROBLEM # 1.1

For each of the following, indicate the scale of measurement that best describes the information.

USA Today reports that the previous day’s highest temperature in the United States was 105 degrees in Death Valley, California.

ANSWER: Interval scale; there is no absolute zero point for temperature.

Source: USA Today, June 19, 2009, p.12A

PROBLEM # 1.1

For each of the following, indicate the scale of measurement that best describes the information.

An individual respondent answers “yes” when asked if TV contributes to violence in Canada.

ANSWER: Nominal scale; we could use "1" to identify yes and "0" for no.

PROBLEM # 1.1

For each of the following, indicate the scale of measurement that best describes the information.

In a comparison test of family sedans, a magazine rates the Toyota Camry higher than the VW Passat.

ANSWER: Ordinal scale; the cars are ranked but there is no measure for the distance between them.

RESEARCH

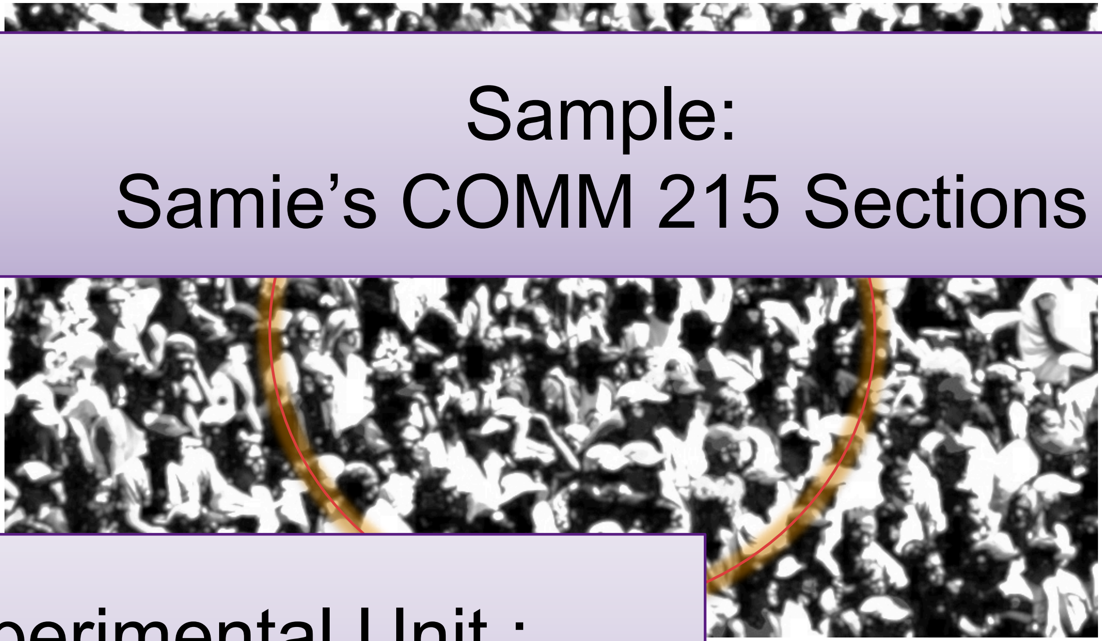
Population:
All students who have taken COMM 215

A census: An examination of all units in a population.

A **population** is a set of units (usually people, objects, transactions, or events) that we are interested in studying. E.g. All students who have taken COMM 215.

http://www.cartoonstock.com/newscartoons/cartoonists/hkh/lowres/hkhn17l.jpg

RESEARCH

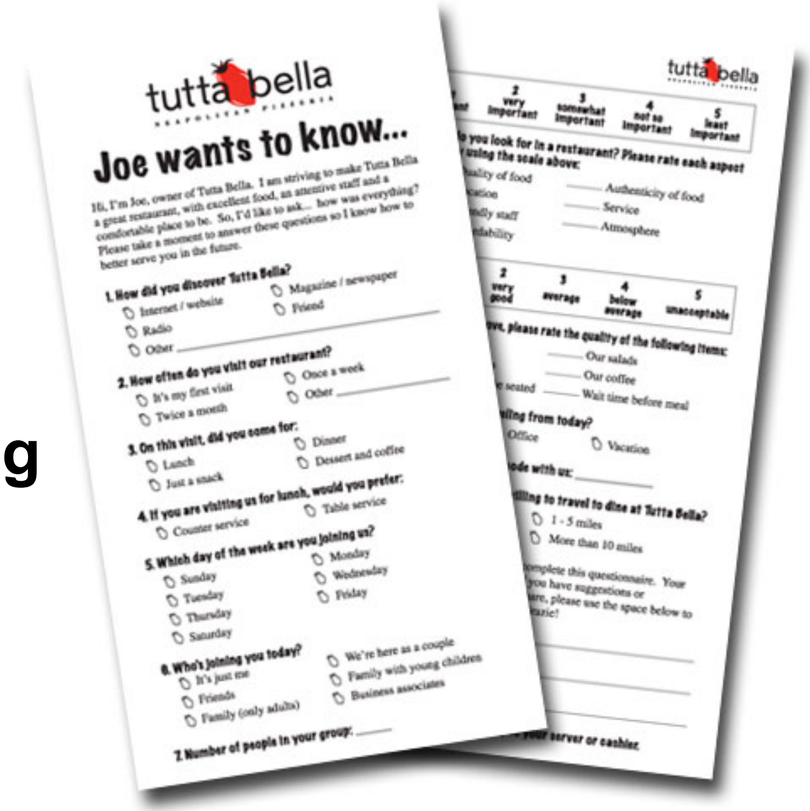


Sample:
Samie's COMM 215 Sections

Experimental Unit :
Samie is going to collect data from her sections.

SAMPLING

Voluntary response sampling



tutta bella
NEAPOLITAN PIZZERIA

Joe wants to know...

Hi, I'm Joe, owner of Tutta Bella. I am striving to make Tutta Bella a great restaurant, with excellent food, an attentive staff and a comfortable place to be. So, I'd like to ask... how was everything? Please take a moment to answer these questions so I know how to better serve you in the future.

1. How did you discover Tutta Bella?
○ Internet / website
○ Radio
○ Other ______
○ Magazine / newspaper
○ Friend

2. How often do you visit our restaurant?
○ It's my first visit
○ Twice a month
○ Once a week
○ Other ______

3. On this visit, did you come for:
○ Lunch
○ Just a snack
○ Dinner
○ Dessert and coffee

4. If you are visiting us for lunch, would you prefer:
○ Counter service
○ Table service

5. Which day of the week are you joining us?
○ Sunday
○ Tuesday
○ Thursday
○ Saturday
○ Monday
○ Wednesday
○ Friday

6. Who's joining you today?
○ It's just me
○ Friends
○ Family (only adults)
○ We're here as a couple
○ Family with young children
○ Business associates

7. Number of people in your group: ______

tutta bella

	2 very important	3 somewhat important	4 not so important	5 least important

...o you look for in a restaurant? Please rate each aspect ...using the scale above:

...uality of food
...cation
...ndly staff
...ndability
______ Authenticity of food
______ Service
______ Atmosphere

2 very good	3 average	4 below average	5 unacceptable

...ve, please rate the quality of the following items:

______ Our salads
______ Our coffee
...seated ______ Wait time before meal

...ling from today?
...Office
○ Vacation

...ode with us: ______

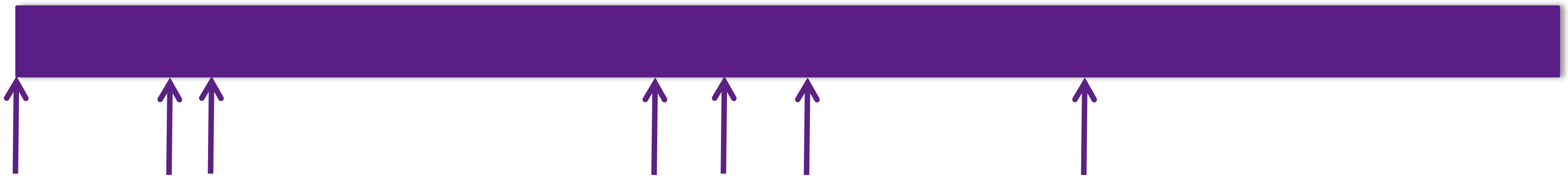
...lling to travel to dine at Tutta Bella?
○ 1 - 5 miles
○ More than 10 miles

...omplete this questionnaire. Your ...you have suggestions or ...are, please use the space below to ...azie!

______
______
______

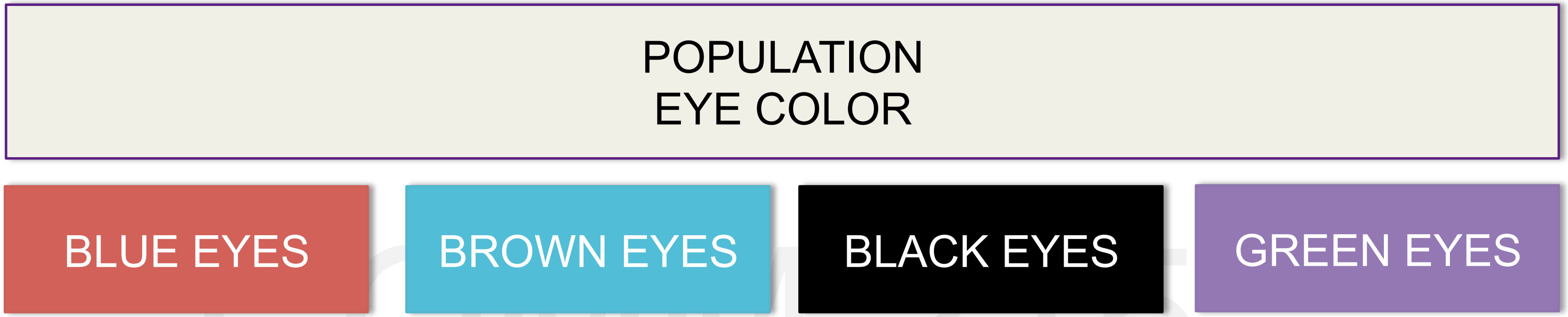
... your server or cashier.

Simple Random sampling



SAMPLING

Stratified random sampling



POPULATION
EYE COLOR

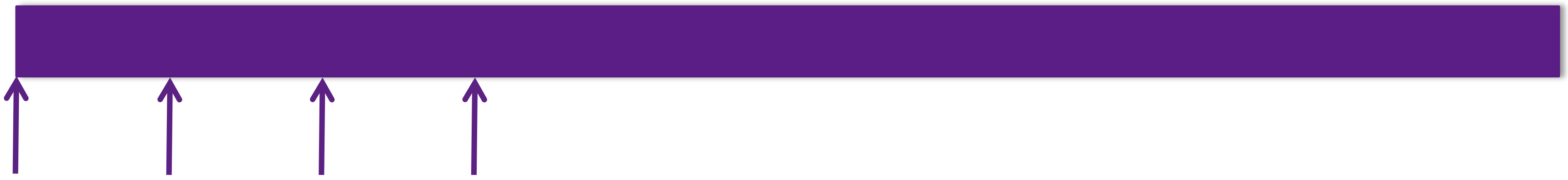
BLUE EYES

BROWN EYES

BLACK EYES

GREEN EYES

Systematic sampling



TYPES OF DATA

Categorical Data vs Quantitative Data

Cross-Sectional vs Time Series

TYPES OF DATA-CATEGORICAL

Categorical Data

- Grouped by specific categories.
- Categorical data are obtained using either the ___________ or ____________ scale of measurement.

Nominal

Ordinal

TYPES OF DATA-QUANTITATIVE

Quantitative Data

- Numeric values
- Quantitative data are obtained using either the __________ or ___________ scale of measurement.

Ratio

Interval

TYPES OF DATA

Cross-Sectional Data- same point in time

- Today, I will be collecting data from all COMM 215 students at the same time.

Time series data- different times

- Through out the semester, I will be collecting data every week from a few specific students.

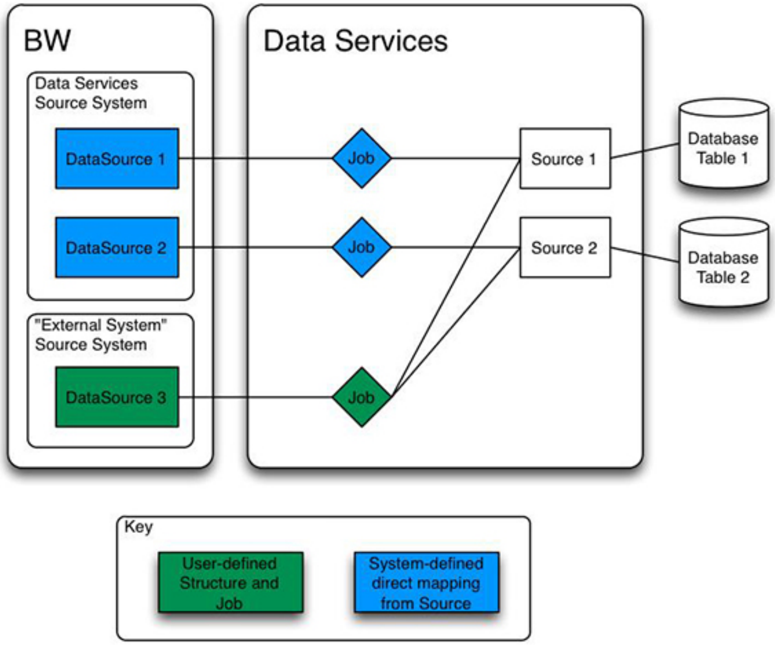
DATA SOURCES

Existing Sources

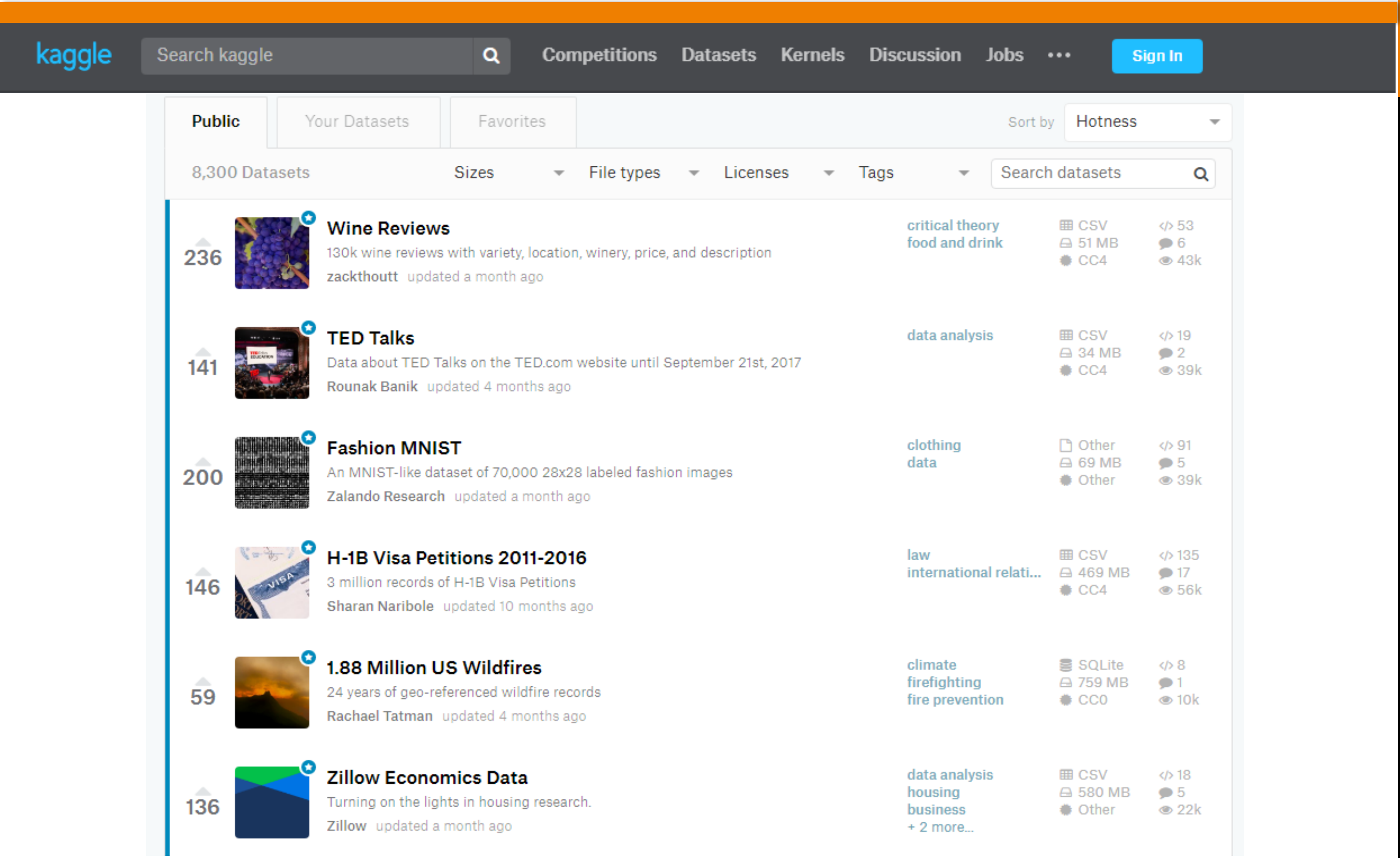
- Primary sources – SAP, ORACLE
- Secondary sources - Statistics Canada (www.statcan.gc.ca)

Data Acquisition Errors

- Making errors during data collection
- Writing 24-year-old as 42 year-old
- Asking ambiguous questions
- Inconsistency
- Spotting outliers
- Selection bias


BW
Data Services Source System
DataSource 1
DataSource 2
"External System" Source System
DataSource 3
Data Services
Job
Job
Job
Source 1
Source 2
Database Table 1
Database Table 2
Key
User-defined Structure and Job
System-defined direct mapping from Source

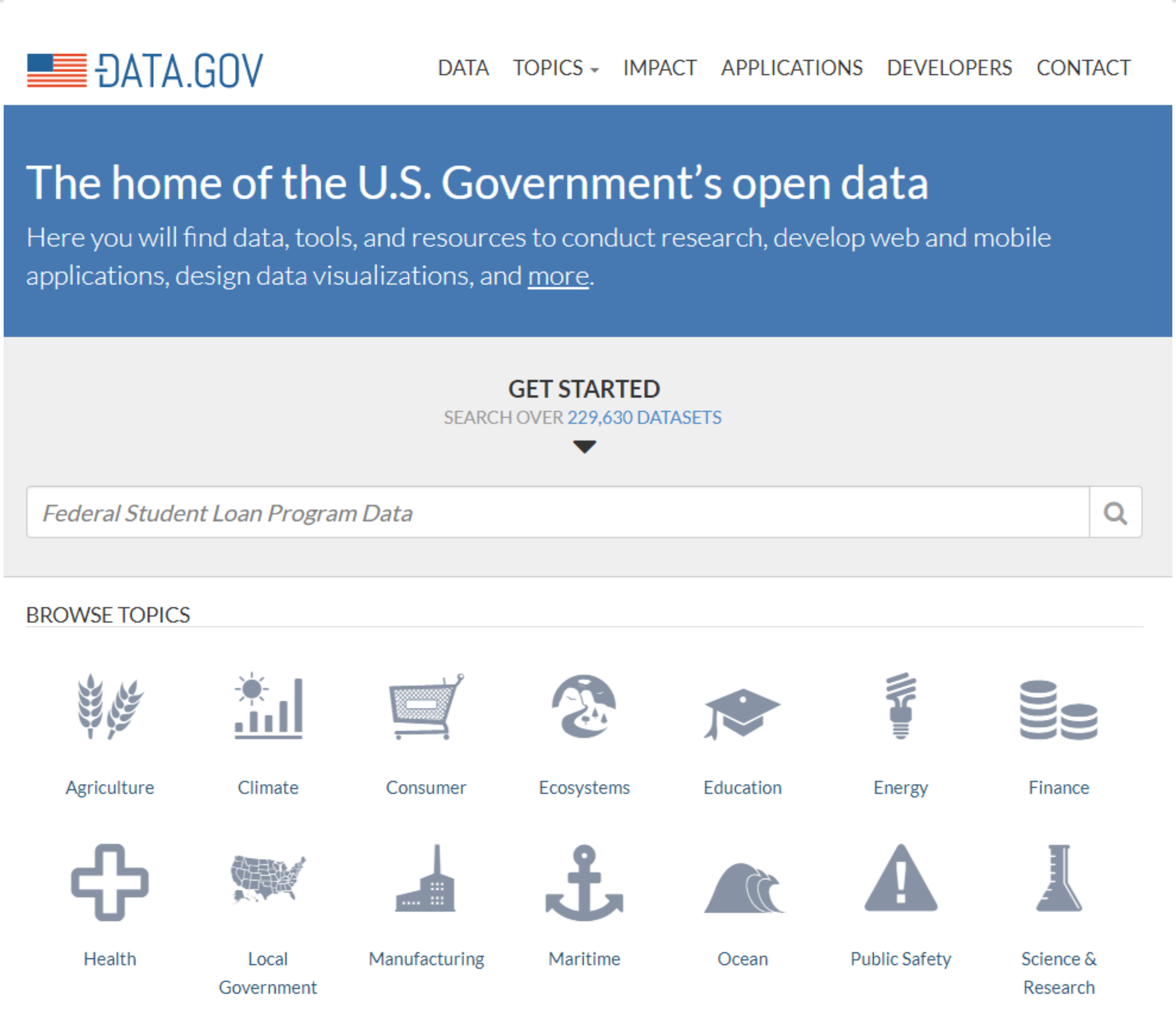
KAGGLE DATA COMPETITION



236	**Wine Reviews** 130k wine reviews with variety, location, winery, price, and description zackthoutt updated a month ago	critical theory food and drink	CSV 51 MB CC4	</> 53 💬 6 👁 43k
141	**TED Talks** Data about TED Talks on the TED.com website until September 21st, 2017 Rounak Banik updated 4 months ago	data analysis	CSV 34 MB CC4	</> 19 💬 2 👁 39k
200	**Fashion MNIST** An MNIST-like dataset of 70,000 28x28 labeled fashion images Zalando Research updated a month ago	clothing data	Other 69 MB Other	</> 91 💬 5 👁 39k
146	**H-1B Visa Petitions 2011-2016** 3 million records of H-1B Visa Petitions Sharan Naribole updated 10 months ago	law international relati...	CSV 469 MB CC4	</> 135 💬 17 👁 56k
59	**1.88 Million US Wildfires** 24 years of geo-referenced wildfire records Rachael Tatman updated 4 months ago	climate firefighting fire prevention	SQLite 759 MB CC0	</> 8 💬 1 👁 10k
136	**Zillow Economics Data** Turning on the lights in housing research. Zillow updated a month ago	data analysis housing business + 2 more...	CSV 580 MB Other	</> 18 💬 5 👁 22k

Source: https://www.kaggle.com/datasets

DATA GOVERNMENT



DATA.GOV

DATA TOPICS ▾ IMPACT APPLICATIONS DEVELOPERS CONTACT

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER 229,630 DATASETS

Federal Student Loan Program Data

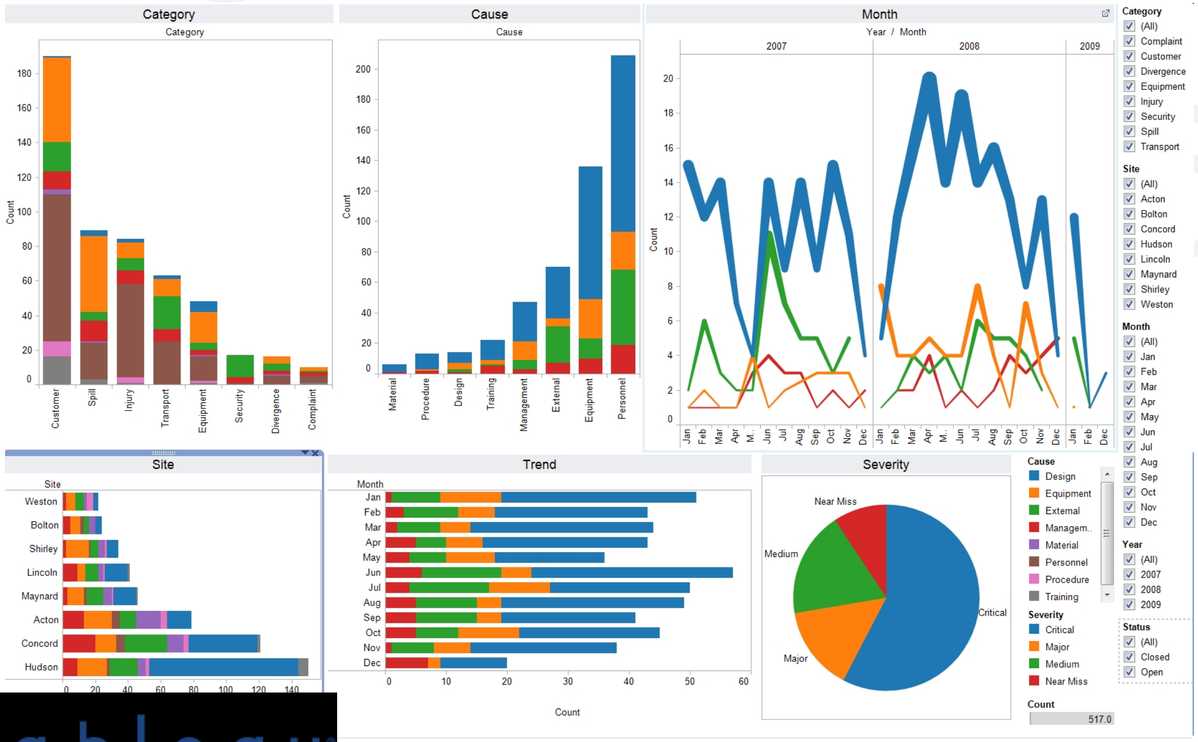
BROWSE TOPICS

Source: https://www.data.gov/

DESCRIPTIVE STATISTICS

Descriptive Statistics
utilizes numerical graphical methods

- to look for patterns in a data set
- to summarize the information revealed in a data set
- to present the information in a convenient form.



Category
☑ (All)
☑ Complaint
☑ Customer
☑ Divergence
☑ Equipment
☑ Injury
☑ Security
☑ Spill
☑ Transport

Site
☑ (All)
☑ Acton
☑ Bolton
☑ Concord
☑ Hudson
☑ Lincoln
☑ Maynard
☑ Shirley
☑ Weston

Month
☑ (All)
☑ Jan
☑ Feb
☑ Mar
☑ Apr
☑ May
☑ Jun
☑ Jul
☑ Aug
☑ Sep
☑ Oct
☑ Nov
☑ Dec

Year
☑ (All)
☑ 2007
☑ 2008
☑ 2009

Status
☑ (All)
☑ Closed
☑ Open

Cause
🟦 Design
🟧 Equipment
🟩 External
🟥 Managem..
🟪 Material
🟫 Personnel
▨ Procedure
▨ Training

Severity
🟦 Critical
🟧 Major
🟩 Medium
🟥 Near Miss

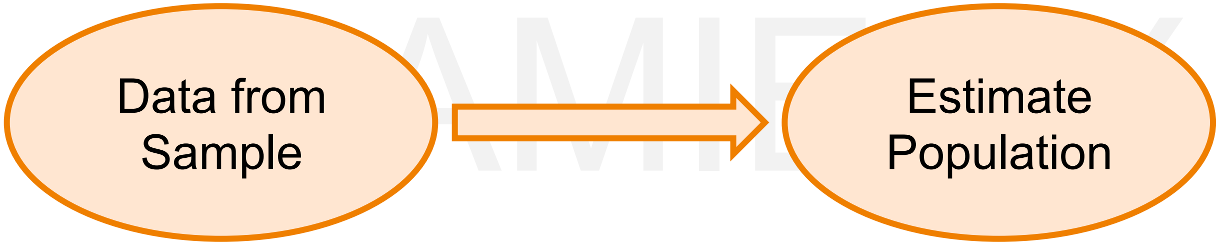
Count
517.0

COMM215
First the Foundation, then Innovation

STATISTICAL INFERENCE

Inferential Statistics utilizes sample data

- to make estimates, decisions, predictions, or other generalizations about a larger set of data.



Data from Sample

Estimate Population

Sample of 500 Students in COMM215	
Year	COMM215 Failure Rate
2011	11.5 %
2015	12.3 %
2016	12.0 %
2017	11.0%
2018	**?**

PROBLEM # 1.2
KEY ELEMENTS OF STATISTICAL PROBLEM

Cola wars is the popular term for the intense competition between Coca-Cola and Pepsi displayed in their marketing campaigns. Their campaigns have featured movie and television stars, rock videos, athletic endorsements, and claims of consumer preferences based on taste tests. Suppose, as part of a Pepsi marketing campaign, 1,000 cola consumers are given a blind taste test. Each consumer is asked to state a preference for brand A or brand B.

Describe the population.

Describe the variable of interest.

Describe the sample.

Describe the inference.

PROBLEM # 1.2
KEY ELEMENTS OF STATISTICAL PROBLEM

Population of interest: all cola consumers

Variable of interest: cola preference

Sample: 1,000 cola consumers selected

Inference: generalization of the cola preference of 1,000 sampled consumers to the population of all cola consumers.

BUSINESS STATISTICS DATA MINING

outlier detection

association learning

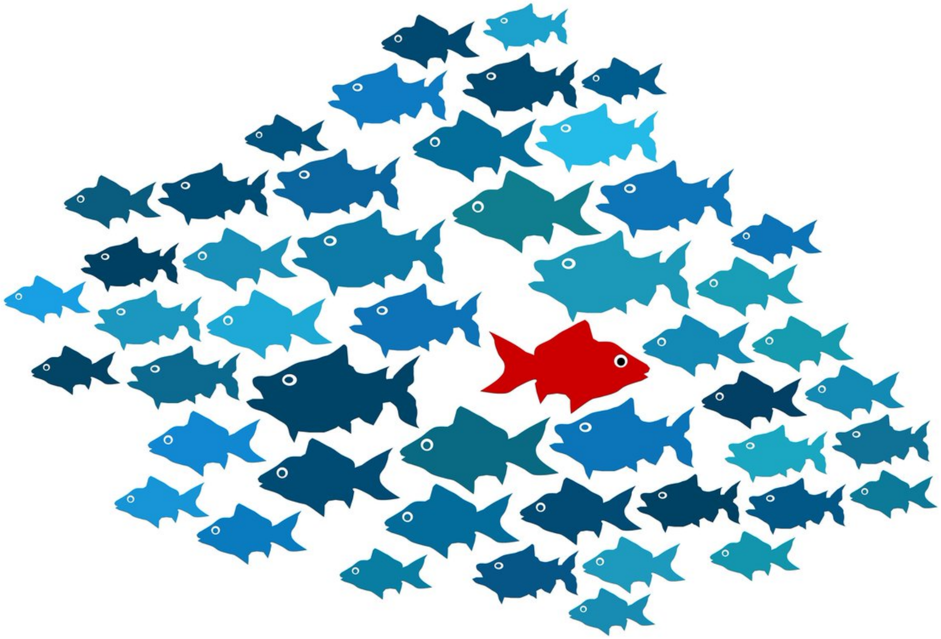
classification

cluster detection

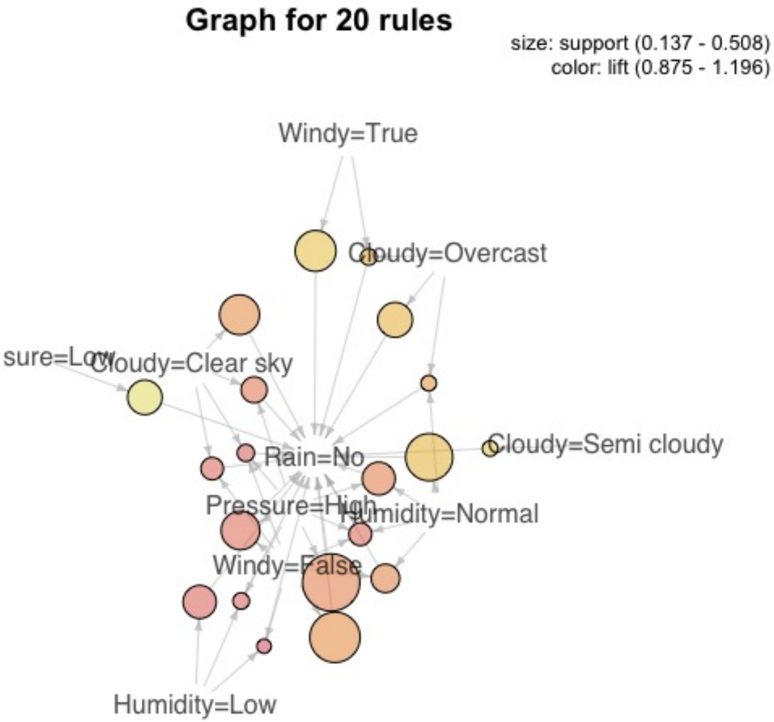
prediction

factor detection

BUSINESS STATISTICS – DATA MINING



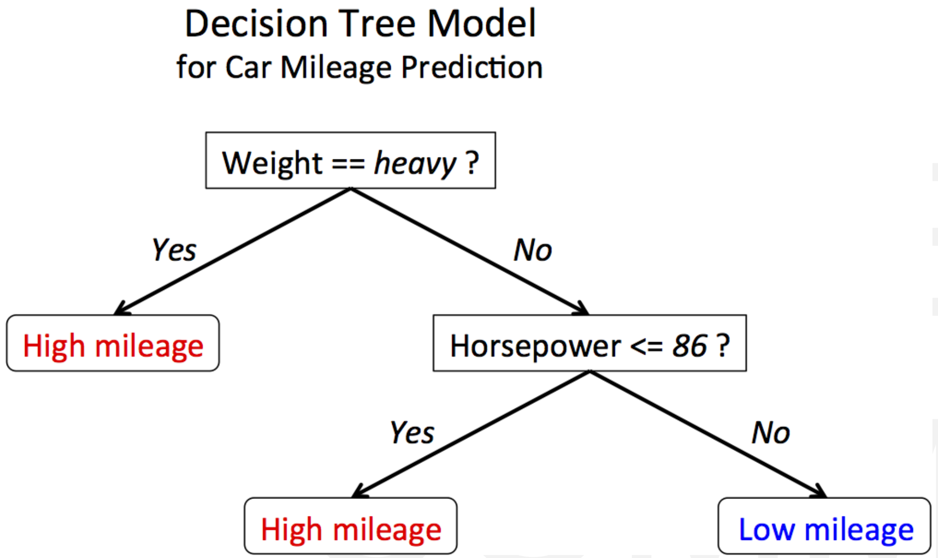
OUTLIER DETECTION



ASSOCIATION LEARNING

COMM215
First the Foundation, then Innovation

BUSINESS STATISTICS – DATA MINING



Decision Tree Model
for Car Mileage Prediction

Weight == *heavy* ?

Yes

High mileage

No

Horsepower <= *86* ?

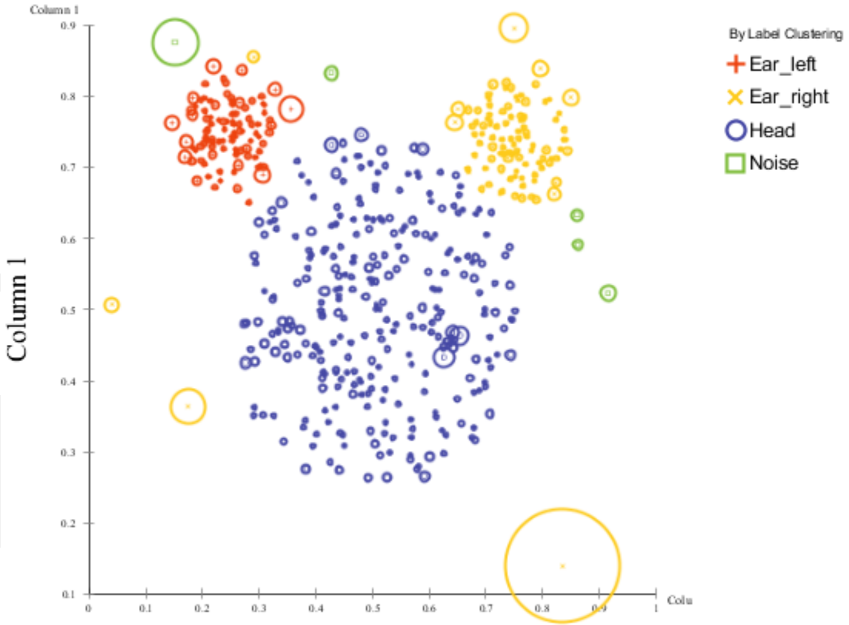
Yes

High mileage

No

Low mileage

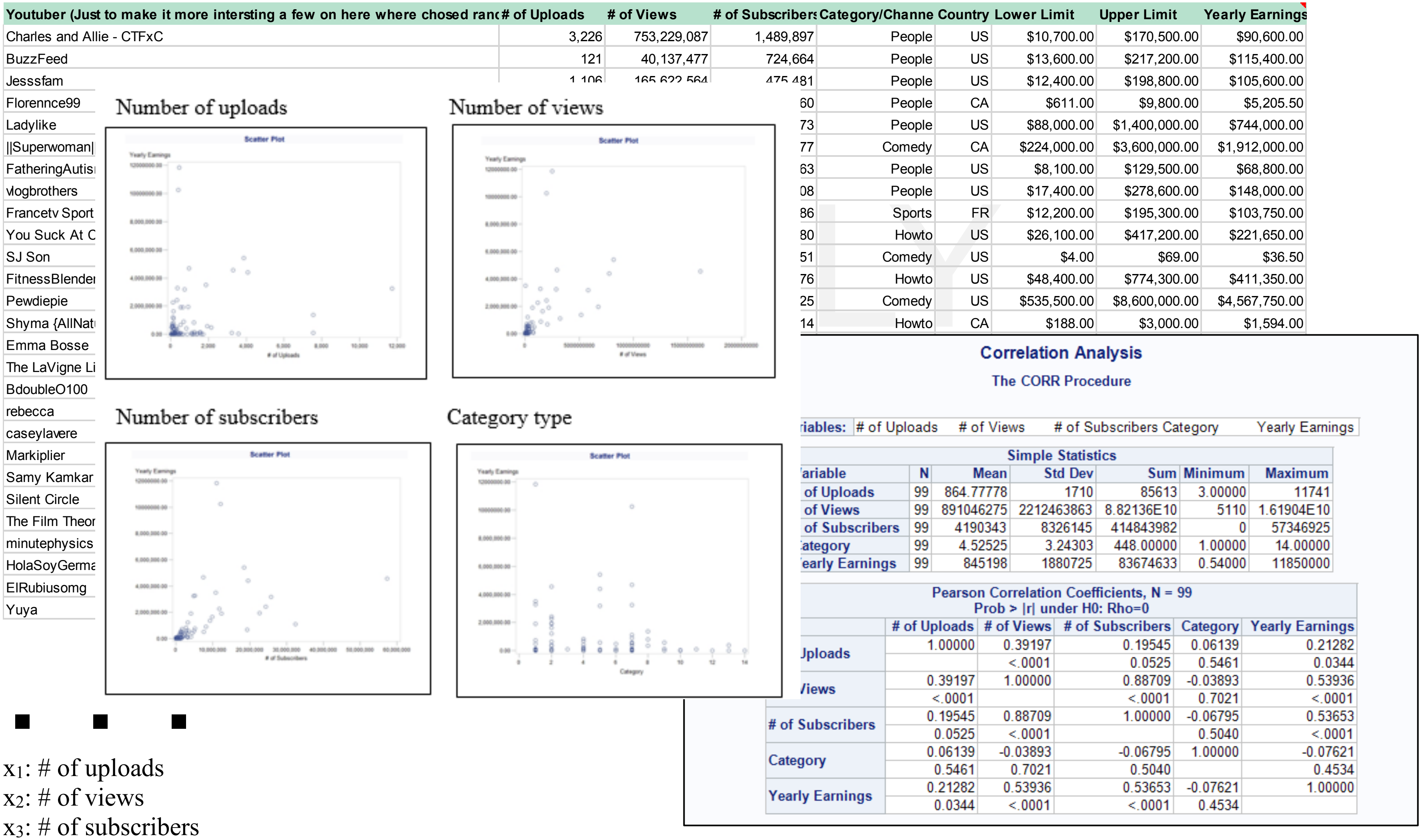
CLASSIFICATION METHODS



outlier detection using LDOF

CLUSTER ANALYSIS

WHAT MAKES A YOUTUBER SUCCEED?



Youtuber (Just to make it more intersting a few on here where chosed ranc	# of Uploads	# of Views	# of Subscribers	Category/Channe	Country	Lower Limit	Upper Limit	Yearly Earnings
Charles and Allie - CTFxC	3,226	753,229,087	1,489,897	People	US	$10,700.00	$170,500.00	$90,600.00
BuzzFeed	121	40,137,477	724,664	People	US	$13,600.00	$217,200.00	$115,400.00
Jesssfam	1,106	165,622,564	475,481	People	US	$12,400.00	$198,800.00	$105,600.00
Florennce99			60	People	CA	$611.00	$9,800.00	$5,205.50
Ladylike			73	People	US	$88,000.00	$1,400,000.00	$744,000.00
||Superwoman|			77	Comedy	CA	$224,000.00	$3,600,000.00	$1,912,000.00
FatheringAutisı			63	People	US	$8,100.00	$129,500.00	$68,800.00
Vlogbrothers			08	People	US	$17,400.00	$278,600.00	$148,000.00
Francetv Sport			86	Sports	FR	$12,200.00	$195,300.00	$103,750.00
You Suck At C			80	Howto	US	$26,100.00	$417,200.00	$221,650.00
SJ Son			51	Comedy	US	$4.00	$69.00	$36.50
FitnessBlendeı			76	Howto	US	$48,400.00	$774,300.00	$411,350.00
Pewdiepie			25	Comedy	US	$535,500.00	$8,600,000.00	$4,567,750.00
Shyma {AllNatı			14	Howto	CA	$188.00	$3,000.00	$1,594.00
Emma Bosse								
The LaVigne Li								
BdoubleO100								
rebecca								
caseylavere								
Markiplier								
Samy Kamkar								
Silent Circle								
The Film Theor								
minutephysics								
HolaSoyGerma								
ElRubiusomg								
Yuya								

Number of uploads

Number of views

Number of subscribers

Category type

Correlation Analysis

The CORR Procedure

riables:	# of Uploads	# of Views	# of Subscribers Category	Yearly Earnings

Simple Statistics						
ariable	N	Mean	Std Dev	Sum	Minimum	Maximum
of Uploads	99	864.77778	1710	85613	3.00000	11741
of Views	99	891046275	2212463863	8.82136E10	5110	1.61904E10
of Subscribers	99	4190343	8326145	414843982	0	57346925
ategory	99	4.52525	3.24303	448.00000	1.00000	14.00000
early Earnings	99	845198	1880725	83674633	0.54000	11850000

Pearson Correlation Coefficients, N = 99
Prob > \|r\| under H0: Rho=0

	# of Uploads	# of Views	# of Subscribers	Category	Yearly Earnings
Jploads	1.00000	0.39197	0.19545	0.06139	0.21282
		<.0001	0.0525	0.5461	0.0344
Views	0.39197	1.00000	0.88709	-0.03893	0.53936
	<.0001		<.0001	0.7021	<.0001
# of Subscribers	0.19545	0.88709	1.00000	-0.06795	0.53653
	0.0525	<.0001		0.5040	<.0001
Category	0.06139	-0.03893	-0.06795	1.00000	-0.07621
	0.5461	0.7021	0.5040		0.4534
Yearly Earnings	0.21282	0.53936	0.53653	-0.07621	1.00000
	0.0344	<.0001	<.0001	0.4534	

■ ■ ■

x_1: # of uploads
x_2: # of views
x_3: # of subscribers
x_4: category

PREDICTION – RELATIONSHIP DETECTION

Project by Kayla V. Michele P. Andrea Cristina S. from BSTA378Fall 2017

ETHICAL GUIDELINES FOR PRACTICE

Statistical thinking

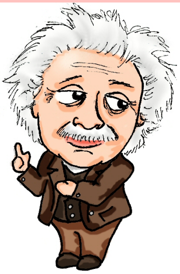
- Know how the data was collected- Is the data from a reliable source?
- Is it a random sample? Or “Self-Selected”?
- Is it possible? Does the data make sense?

If the biased sample was intentional, with the sole purpose to mislead the public the researchers would be guilty of unethical statistical practice.

MISLEADING STATISTICS

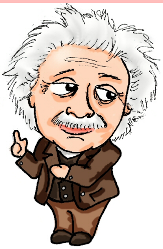
'Assuming that the defendant did not commit this crime, what is the probability that the defendant and the culprit having identical fingerprints?'

'One in several billion.'



'Let me ask you a different question. What is the probability that a fingerprint lifted from a crime scene would be wrongly identified as belonging to someone who wasn't there?'

'Oh, about 1 in 100.'



Forensic DNA Databases and Race: Issues, Abuses and Actions held June 19-20, 2008, at New York University. To link to this paper, visit www.gene-watch.org.

REFERENCES

Statistics for Business and Economics

Keller, G. (2012). *Statistics for management and economics.* Mason: Cengage Learning.

McClave, J. T., Benson, G. P., & Sincich, T. (2008). *Statistics for Business and Economics.* New Jersey: Prentice Hall.

Weiers, R. M. (2011). *Introduction to Business Statistics.* Mason: Cengage Learning.

Bowerman, B. L., O'Connell, R. T., Murphree, E., Huchendorf, S. C., & Porter, D. C. (2003).

Business statistics in practice (pp. 728-730). New York: McGraw-Hill/Irwin.