ADM 2303 STATISTICS FOR MANAGEMENT I

Week 1 – Introduction, Data and Probability

INTRODUCTION

What is Statistics?

- Statistics is the science of making effective use of numerical data relating to groups of individuals or experiments. It deals with all aspects, including not only the collection, analysis and interpretation of data, but also the planning of data collection, in terms of the design of surveys and experiments.
- Statistics helps to quantify uncertainty; critical for making decisions in the face of uncertainty
  - Uses concepts of **probability** to help model uncertainty

Uses of Statistics in Business

- Forecast profits under different scenarios
- Predict demand and supply
- Compare or evaluate performance
- Obtain information on customer satisfaction, opinions, etc.
- Identify factors that affect viability or success
- Avoid being fooled by evidence
- Assess the likelihood of key events

5

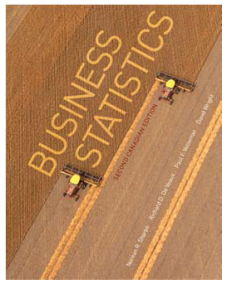
Why study Statistics?

- Statistics is more than just..
  - The ability to calculate the mean or
  - To understand the normal distribution
- Statistics is…
  - A tool for making sense of the world
  - Provides evidence for competing claims (Statistical hypothesis)
  - Helps in the art of reasoning; used to understand the current state of affairs, and predict (forecast) future events
  - Decision theory

6

Administration Info

- Course Outline
- Office Hour: Wednesdays 16:30 – 17:30 DMS 5140
- Textbook: *Business Statistics*, 2nd edition by Sharpe, De Veaux, Velleman, Wright
  - Many resources within MyStatLab
  - Textbook emphasizes Plan, Do, Report



7

Plan, Do, Report

- PLAN: Chart out where you are headed and why. Clearly defining and understanding your objective will save a lot of work.
- DO: The mechanics.
- REPORT: Report what you have learned. Until you have explained your results in a context that can be understood, the work is incomplete.

8

Administrative Information

- Familiarize yourself with computer and internet resources
- Blackboard Learn
  - Downloading course material.
  - Submitting Part II of assignments
- MyStatLab
  - For Part I of assignments, homework and learning aids

9

DGD Labs

- DGD Labs (Teaching Assistant Labs)
  - Section E: Fridays 17:30 – 19:00
  - Location:
    - **September 18th** - labs will be held in DMS 2130/2140
      - Labs will introduce computer resources and software (VERY IMPORTANT)
      - MINITAB, MyStatLab
    - **September 25th onwards** – labs will be held in LEE B163
      - Regular labs will consider practice problems, review assignment and quiz solutions and respond to student questions

10

Evaluation

- Assignments 28%
- Midterm Quiz (October 17th) 20%
- Final Examination 52%

- There will be four assignments, worth 7% each.
- The Midterm Quiz will be held on Saturday October 17th in the afternoon – more details to follow.
- The Final examination will be held during the regular final exam period.

11

Assignments

- Part I
  - Automated submission in MyStatLab. There will be a maximum of four tries – be sure you are ready to respond to the questions before you use up a "try." The best mark is retained.
- Part II
  - Exam-style questions. More detail to follow.

12

MINITAB

- Exams and Assignments will use the statistical software package MINITAB
- MINITAB is a platform for analyzing and graphing data
  - Students are expected to learn MINITAB or some similar statistics package (R for example)
  - Students are expected to be familiar with MINITAB output for assignments and exams

Course Design

	In class examples, the first time a topic is taught.	Homework exercises	Assignment Exercises	In class previous years final exam questions.
Level	Easy	Progressing from easy to more complex	Progressing to final exam level	Final exam
If you have difficulty …	Ask the Professor during/after class.	Attend a lab/DGD and ask the TA.	Before due date, send an email. After due date: Prof will review in class, if necessary. Ask a TA during a lab/DGD.	Ask the Professor during/after class.

Course Topics

- Data analysis
  - The gathering, display and summarizing of data
- Probability
  - The laws of chance
  - Provides the "grammar" (rules) for exploring the implications that follow from a set of premises
- Statistical Inference
  - The science of drawing statistical conclusions from data using knowledge of probabilty

DATA

Textbook – Chapter 2

What are Data?

- The building blocks of Statistics are Data
- Data can be numbers, names, etc.
- Data may be either numerical or categorical
  - Numerical - e.g. Salary (in dollars)
  - Categorical – e.g. Smoking status (smoker, non-smoker)
- It is important to clarify data in terms of their context.

Data Questions: The 5 W's

- To assess the context of data, we ask the following questions of the data:
  - Who
  - What (and in what units)
  - When
  - Where
  - Why (if possible)
  - How

Example

- Data is often arranged in a table, which shows the context of the data under consideration (Exercise 2 page 25):

Transaction ID	Customer ID	Date	ISBN Number of Purchase	Price	Coupon?	Gift?	Quantity
297843209 12	4J438	11/12/2009	345-23-2355	$29.95	N	N	1
26483589001	3K729	9/30/2009	983-83-2739	$16.99	N	N	1
26483589002	3K729	9/30/2009	102-65-2332	$9.95	Y	N	1
36429489305	3U034	12/5/2009	295-39-5884	$35.00	N	Y	1
36429489306	3U034	12/5/2009	183-38-2957	$79.95	N	Y	1

- The table also describes the What (column) and the Who (row) for the data set.

Who

- The Who from a dataset identifies the individuals or cases for which (or whom) we have collected data
  - Individuals who answer survey questions are called **respondents**
  - Individuals on whom we experiment are called **participants** or **subjects**
  - Non-human or inanimate objects are called **experimental units**
- Alternatively, we may refer to data values as observations

What

- The characteristics recorded about each individual or case are called Variables.
  - The Variables indicate what has been measured
- There are two main types of variables:
  - Quantitative (numerical) – income, height, weight
  - Qualitative (categorical) – gender, field of study, smoking status
- For quantitative variables, the units should always be provided
  - Income (dollars); Height (cm); Weight (kg)

What

- Qualitative variables may be classified as ordinal or nominal
  - Ordinal – categories display a natural ordering
    - E.g. **Satisfaction level** (Very satisfied, Somewhat satisfied, Unsatisfied, Very Unsatisfied)
  - Nominal – categories do not display a natural ordering
    - E.g. **Method of payment** (Cash, Credit card, Paypal, Debit card, Other)

Where, When, How and Why

- The Who and What of the data are necessary to analyze the dataset
- The other W's and the How provide further insight into the data
- When and Where provide important information about the data context
  - Knowing information on the target population and time of data collection help to place the conclusions in context.
  - Example: In investigating incidences of the flu, knowing the data are from the Flu Epidemic of 1918 provides useful information about the data.

Where, When, How and Why

- How the data are collected can make the difference between insight and nonsense.
  - Data should be collected according to sound statistical procedures
  - Samples should be **randomly selected** and **representative** of the population under consideration
  - Such principles reduce bias and allow conclusions to be generalized
    - E.g. Results from Internet surveys are often biased
- Why reveals the most about the data, providing insight into the Who, What and How of the data.
  - Example: Knowing that King Cola company conducted a taste test of colas might make the reader skeptical of the results of the study.

What can go wrong?

- Even though the variable values are numbers, do not assume the variable is quantitative
- Always be skeptical – don't take data for granted.

Key Concepts

- Variables may be classified as quantitative or qualitative (categorical).
- Qualitative variables with a natural ordering are called ordinal variables
- In addition to the data itself, the context of the data is very important.
- Consider the "W's":
  - Who, What (and in what units), When, Where, Why (and How) when examining a set of data

Example

Consumer Reports Health routinely compares drugs in terms of effectiveness and safety. Suppose that in the summer of 2010 they reviewed drugs used to treat arthritis. Information was reported on convenience of use (how many pills required each day), possible side effects (e.g., dizziness, stomach upset), cost, and ratings of effectiveness in relieving symptoms (very effective, somewhat effective, not effective).

1. Describe the W's for the information given.
   Who:
   What:
   When:
   Where:
   How:
   Why:

2. List the variables. Indicate whether each variable is categorical or quantitative. If the variable is quantitative, list the units.

DISPLAYING CATEGORICAL DATA

Textbook - Chapter 3

The Three Rules of Data Analysis

- Your text describes the three rules of data analysis:

1. Make a picture – ideas about the data may be revealed that are not obvious in the raw data.
2. Make a picture – important features and patterns in the data will be displayed.
3. Make a picture – the best way to communicate results is with a well-chosen graph.

Graphical Analysis is important in all phases of statistical analysis.

Counts and Frequency Tables

- We can "pile" the data by counting the number of data values in each category of interest
- We can organize these counts into a frequency table, which records the totals and the category names
- A relative frequency table is similar, but gives the percentages (instead of the counts) for each category

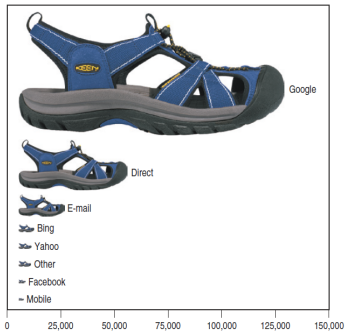
Frequency Tables

- Frequency tables and relative frequency tables describe the distribution of a categorical variable by naming the possible categories and telling how frequently each category occurs.
- Example: A sandal manufacturer has different forms of advertisements, and has measured the number of visits by customers for each type of advertisement in a given month. The frequency counts (Page Visits) and relative frequency (Visits by %) tables are presented here:

Source	Visits	Visits by %
Google	130,158	57.36
Direct	52,969	23.34
E-mail	16,084	7.09
Bing	9,581	4.22
Yahoo	7,439	3.28
Facebook	2,253	0.99
Mobile	1,701	0.75
Other	6,740	2.97
Total	**226,925**	**100.00**

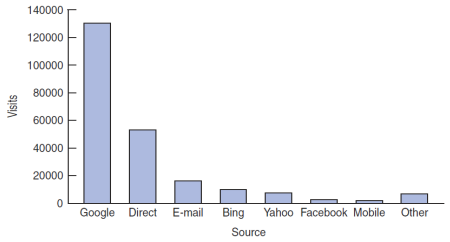
Problems with Graphs

- One possible way to display the sandal data is below:



The Area Principle

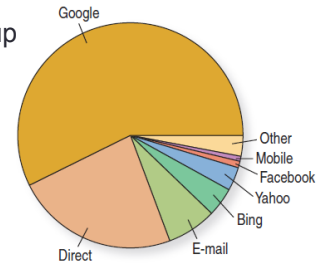
- The sandal display violates the area principle: the area occupied by a part of the graph should correspond to the magnitude of the value it represents.
- Thus a better display is:



33

Pie Chart

- When you are interested in parts of the whole, a pie chart might be the preferred graphical display.
- Pie charts display the whole group of cases as a circle.
- They slice the circle into pieces whose size is proportional to the fraction of the whole in each category.



34

Contingency Tables

- A contingency table allows us to look at two categorical variables together
  - Example: We can examine the class of ticket and whether a person survived the Titanic:

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	202	118	178	212	710
	Dead	123	167	528	673	1491
	Total	325	285	706	885	2201

- The totals in the margins of the table represent the marginal distribution of the respective variables.

35

Conditional Distributions

- A distribution of one variable for only those individuals or cases satisfying some condition on another variable is called a conditional distribution.
- In a contingency table, variables are independent when the distribution of one variable is the same for all categories of another.

36

Conditional Distributions

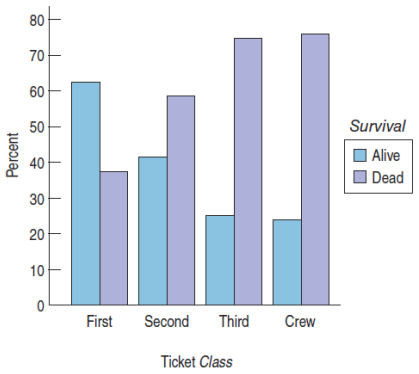
- Consider the following two pie charts:

- These pie charts show the ticket class of the passengers conditional on survival status. We note differences in the distributions – ticket class and survival are not independent.

37

Side-by-side bar chart

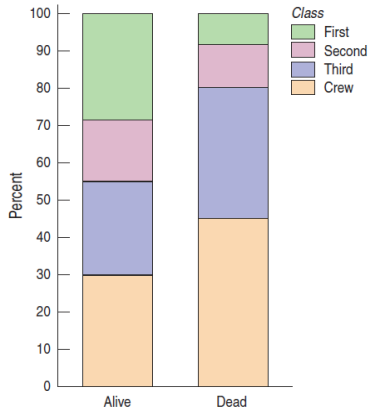
- We could also construct a side-by-side bar chart to compare the survival for each class
- A side-by-side bar chart shows the conditional distribution of *Survival* for each category of ticket *Class*



38

Segmented Bar charts

- A segmented bar chart displays the same information as a pie chart, but in the form of bars instead of circles.
- Here is the segmented bar chart for ticket class by survival status:



39

Key concepts

- Categorical (qualitative) variables can be summarized in frequency or relative frequency tables.
- Categorical variables can be displayed with bar charts and/or pie charts – just make sure to follow the area principle.
- A contingency table summarized two variables at a time.
  - From a contingency table we can find the marginal distribution for each variable or the conditional distribution for one variable conditional on the other variable.

40

Key concepts

- Two categorical variables are said to be independent if the conditional distribution of one variable is the same for each category of the other.

PROBABILITY

Chapter 8

Randomization

- A random phenomenon is a situation in which a spectrum of outcomes is known to be possible, but it is uncertain which of these outcomes will occur until a trial has been completed.
- Probability is a language for dealing with many random phenomena.

Probability

- The probability of an event is its long-run relative frequency
- For any random phenomenon, each attempt, or trial, generates an outcome.
  - On each trial, an outcome is observed to occur
- An event consists of a combination of outcomes

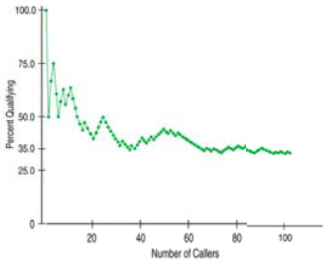
Probability

- When considering the probability of an event involving a combination of outcomes, calculations simplify if the individual trials are independent.
- Trials are independent if the outcome of one trial doesn't influence the outcome of another trial
  - For example, successive flips of a fair coin are independent.

45

The Law of Large Numbers

- The Law of Large Numbers (LLN) says that the long-run relative frequency of repeated independent events gets closer and closer to the true relative frequency as the number of trials increases.



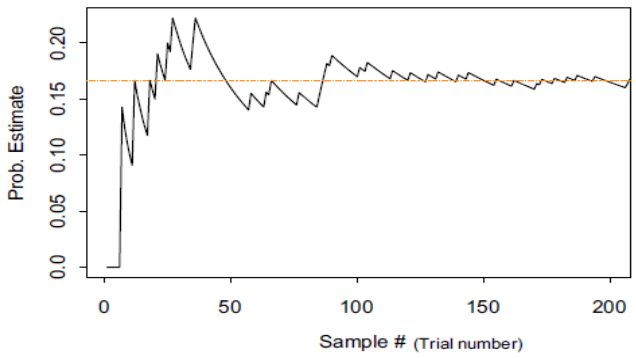
46

Example – Law of Large Numbers

- As a more concrete example, consider the event to be rolling a die and observing a "1" (Snake-eye")

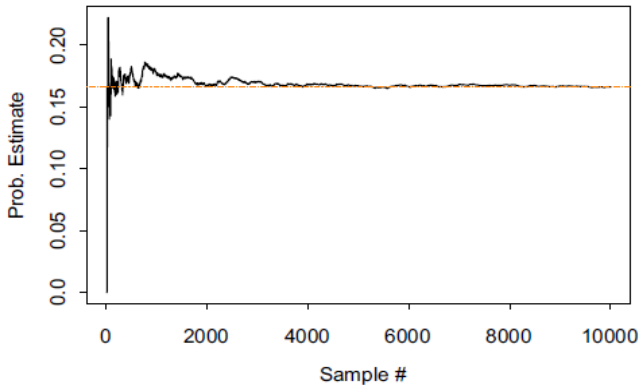
47

Example – Law of Large Numbers



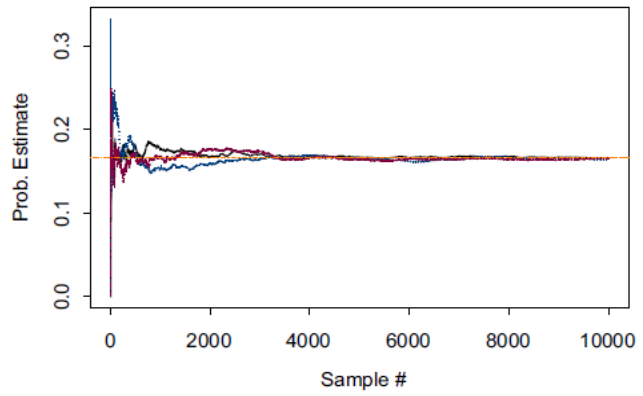
48

Example – Law of Large Numbers



49

Example – Law of Large Numbers



50

The Law of Large Numbers

- A common misconception of the LLN is that random phenomena are supposed to compensate for whatever happened in the past. This is incorrect
  - For example when flipping a fair coin, if a Head is observed on each of the first 10 flips, what is the probability that Tails will be observed on the next flip?
  - The misconception of "I'm due" or "my luck must be running out."

51

Probability Revisited

- As a result of the LLN, we know that relative frequencies converge in the long run, so we can officially give the name probability to that value.
- Probabilities must be between 0 and 1 inclusive
  - A probability of 0 indicates an impossible event
  - A probability of 1 indicates certainty.

52

Probability

- In everyday speech, when we express a degree of uncertainty without basing it on the long-run frequencies, we are stating subjective or personal probabilities
- The textbook considers the frequentist definition of probability

53

Formal Probability

1. There are two requirements for the probability:
   - A probability is a number between 0 and 1
   - For any event A, the probability is denoted $P(A)$ and is always $0 \leq P(A) \leq 1$

54

Formal probability (cont'd)

2. "Something has to happen rule"
   - Define S = the set of all possible outcomes (sample space)
   - The probability of the set of all possible outcomes of a trial must be 1.
   - As a result, we have $P(S) = 1$

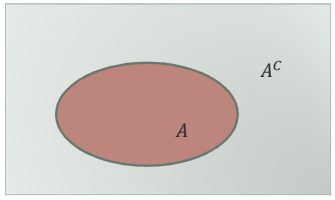
The sample space S

55

Formal probability (cont'd)

3. Complement Rule:
   - Definition: The set of outcomes that are not in the event A is called the complement of A, and is denoted A^C
   - The probability of an event occurring is 1 minus the probability that it does not occur:

$$P(A^C) = 1 - P(A)$$

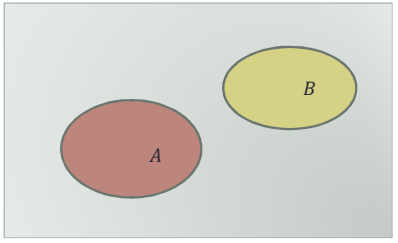


The set A and its complement

56

Formal probability (cont'd)

4. Addition Rule:
   - Definition: Events that have no outcomes in common (i.e. cannot occur together) are called **disjoint** or **mutually exclusive**



Two disjoint sets, A and B

57

Formal Probability (cont'd)

4. Addition Rule

- For two disjoint events A and B, the probability that one occurs OR the other occurs is the sum of the probabilities of the two events

$$P(A\ or\ B) = P(A \cup B) = P(A) + P(B)$$

provided that A and B are disjoint

58

Formal probability (cont'd)

5. Multiplication Rule

- For two independent events A and B, the probability that both A and B occur is the product of the probabilities of the two events

$$P(A\ and\ B) = P(A \cap B) = P(A) \times P(B)$$

provided that A and B are independent.

59

Putting the Rules to Work

- Most efforts to calculate a probability of a defined event will rely on the rules in combination
- The rules can help us move from what we know, to what we would like to know.
- Rules such as the complement rule are often the key to a solution

60

Summary of Probability Rules (in words)

- "Not" involves subtraction from 1;
  - Complement rule
- "or" invokes the Addition Rule
  - Simple addition of probabilities works fine when…
- "and" invokes the Multiplication Rule
  - Simple multiplication of probabilities works fine when…

61

Key Concepts

- Formal probabilities can generally be understood in reference to the Law of Large Numbers
  - The long-run relative frequency of repeated independent events gets closer and closer to the true relative frequency as the number of trials increases
- There are some basic probability principles to keep in mind. If we follow these principles and apply the correct rules, we will be able to find the correct probabilities

62

Notation and Terminology

AND (A and B)	Intersection	∩ (A ∩ B)
OR (A or B)	Union	U (A U B)
NOT (NOT B)	Complement	$\overline{B}$, or B^c

63

Example

Harrison Water Sports has three retail outlets: Vancouver, Victoria, and Nanaimo. The Vancouver store does 50% of the total boat sales in a year, while the Victoria store does 35% of the total boat sales, and Nanaimo the remainder.

a) What fraction of boat sales is produced by the Nanaimo store?

b) For randomly selected boats, find the probability that

i. A boat is not sold in the Vancouver store?

ii. A boat is sold in either Vancouver or Victoria?

iii. A random sample of two boat sales are both from Victoria

iv. None of the three selected boats were sold from Nanaimo

v. At least one of the three boat sales is from Vancouver

vi. The first Vancouver boat sold is the fourth boat selected.