STAT 2B03

Statistical Methods For Science (Section 03)

Angelo J. Canty

Office : Hamilton Hall 209

Phone : (905) 525-9140 extn 27079

E-mail : cantya@mcmaster.ca

Office Hours : Monday, Wednesday, Thursday 3:30-4:20
or at other times by e-mail appointment.

Lecture Notes : My lecture notes (Section 03) are available at
www.math.mcmaster.ca/canty/teaching/s2b03

Course Web Page :
www.math.mcmaster.ca/childsa/2b03.html

Textbook : *Biostatistics - A Foundation for Analysis in the Health Sciences*, (ninth Edition), W. W. Daniel, Wiley.

Assignment/Lab Manual : Available from the bookstore and is required for all students.

Software : Minitab Version 16. The software is available with the textbook (new only) for use at home.

Labs : Start on January 9. Attendance is optional but strongly encouraged. Bring assignment/lab manual and USB drive.

Chapter 1

What is Statistics?

Statistics is the science of conducting studies to collect, organize, summarize, analyze and draw conclusions from data.

* The first step in statistics is the collection of data.

* Such data must be organized into datasets, usually on a computer.

* The job of the statistician is then to analyze that data with an aim to drawing conclusions from the data.

Statistics is used in many areas such as:

* Government agencies (Statistics Canada, Health Canada etc.)
* Insurance Companies (Actuaries)
* Scientific research (medicine, biology, sociology, chemistry, physics)
* Manufacturing companies (quality control)
* Banks (risk assessment)
* Market research and opinion polling

Data

* The science of statistics requires the collection of numerical facts called data.

* Measurement data results from measuring certain quantities on individuals.

* Count Data usually arises from counting the number of individuals with a certain characteristic or the number of occurrences of a certain event over a time period.

Sources of Data

* Historical records such as medical records in a hospital.
* Surveys such as those carried out by Health Canada and Statistics Canada.
* Results from experiments such as with animal models in medical research.
* Records collected as part of the use of a service such as store loyalty card etc.

Variables

Variable A variable is an attribute or characteristic which can take on different values.

Quantitative Variable This is a variable which results from a measurement. It is always a number.

Qualitative Variable This is a variable which results from putting individuals into classes. It usually is not numeric.

Random Variables

Random Variable A variable whose value is determined by random chance. All random variables are numeric and their exact value cannot be known in advance.

Discrete Random Variable A random variable whose outcome is one of a finite set of possible values or for which there are gaps between each possible value.

Continuous Random Variable A random variable whose value can be any real number in an interval (possibly infinite).

Measurement Scales

Nominal Scale Items are placed in a number of mutually exclusive and exhaustive classes. The names of the classes may be numbers but these are simply labels.

Ordinal Scale This is a nominal scale in which the ordering of the classes is known. Classes are usually given sequential integers as names. Although the ordering is known there is not a defined distance between classes.

Quantitative Scale This is the only scale of measurement in which mathematical operations make sense. Measurements are real values in an interval. Interval and Ratio scale measurements are quantitative. The difference between them is whether or not there is an absolute or arbitrary zero point.

Descriptive and Inferential Statistics

Descriptive Statistics This is the process of using numerical and graphical summaries to describe a given dataset. No attempt is made to extrapolate outside the observations.

Inferential Statistics This type of statistics deals with making conclusions beyond the observed dataset to a wider population without the need to make measurements on every individual in the population.

Populations and Samples

Target Population A collection of individuals about whom we wish to make conclusions.

Sampled Population The population from which we draw a sample. Ideally the sampled population and target population are identical.

Sample A subset of the sampled population. We hope that a sample is representative of the population and so we can make conclusions about the population based on the sample.

Statistical Inference

* Process by which we make conclusions about a population based on the information in a sample taken from that population.

* Much of this course (and the usefulness of statistics in general) is based on the idea of statistical inference.

* Valid statistical inference requires that the sample be representative of the target population.

* Much research has gone into ways of sampling that satisfy this requirement.

Types of Sampling

Simple Random Sampling Every possible set of n individuals in the population has an equal chance of being selected.

Stratified Random Sampling The population is divided into a set of mutually exclusive and exhaustive sub-populations and random samples are taken from each. The final sample is the union of all of these sub-samples.

Systematic Sampling Sampling is not done at random but based on a systematic approach to a list. Important that the list ordering is in no way related to the primary quantity of interest.

Chapter 2

Descriptive Statistics?

* When faced with a dataset the first task is to produce a descriptive summary of the important features of the dataset.

* Descriptive statistics include numerical summaries and graphical displays.

* These summaries are used to describe things like the central location of the datset, how spread out it is and other important characteristics of the dataset.

* In this chapter we shall describe some of the most important descriptive summary measures and graphics.

Data Arrays

* Usually we will be presented with a dataset in the form of an array.

* Each column of the array will correspond to a variable.

* Each row of the array contains the values of the variables for an individual entity.

* For now we shall not worry about relationships between the variables but just look at describing each variable separately.

Example of a Data Set

Name	Sex	Education	Age	Siblings	Income
Abdel	Male	University	38	2	63985
Carolyn	Female	Masters	42	0	51689
Helen	Female	College	57	1	114686
Jacob	Male	High School	28	5	25147
Lamar	Male	Doctorate	47	0	98562
Maria	Female	College	41	1	30250
Pierre	Male	High School	35	0	44485
Robyn	Female	University	26	4	47332
Sarah	Female	College	31	1	21968
Simon	Male	High School	45	4	124637
Tara	Female	Masters	47	0	78521
Thierry	Male	High School	25	7	16852

Frequency Distributions

* Used to summarize raw data into a table.

* Can be used for qualitative or quantitative variables.

* Many questions about the data can be answered more easily from the frequency distribution than from the raw data.

* Frequency distributions usually also include percentage or relative frequency information.

Frequency Distributions for Qualitative Data

These are used for unordered qualitative data.

1. List all of the possible categories that the variable can take on.

2. Count the total number of observations. This is the sample size and is generally denoted n.

3. Count the number of times each category is observed and place the counts in the table.

4. The relative frequencies are the counts from step 3 divided by n. The percentage frequencies are the relative frequencies multiplied by 100. Choose one of these measures and place the values in the table also.

5. Total the frequency column (should equal n) and the relative frequency column (should equal 1) or the percentage column (should equal 100).

Example: The top 20 rated shows in a given week came from the following networks:

CBS	Fox	ABC	Fox	CBS
CBS	ABC	Fox	CBS	ABC
CBS	CBS	NBC	CBS	NBC
NBC	CBS	CBS	NBC	NBC

Construct a frequency and relative frequency distribution for this data.

Frequency Distributions for Ordered Variables

* Essentially the same as frequency distributions for unordered data except that the classes should be listed in increasing order in the table.

* Generally also add cumulative frequency and cumulative percentage columns.

* These are obtained by adding the frequencies (or percentages) for all rows of the table up to and including the current row.

* The final cumulative frequency value should be n and the final cumulative percentage value should be 100.

* For discrete quantitative variables with a small number of possible values we use the same method.

Example

Consider the Education and Siblings variables in the dataset on page 12.

Construct frequency distributions and answer the following questions

1. How many of the subjects do not have a university degree?

2. What percentage have a graduate degree?

3. What percentage were only-children?

4. How many had at least 2 siblings?