# MATH 3070 Introduction to Probability and Statistics
## Lecture notes
## Introduction

### The Dichotomy in Statistics

Statistics can be broadly divided into two separate categories : descriptive and inferential. **Descriptive** statistics are used to tell us about the data and convey information in a manner which has as little pain as possible. Descriptive statistics often take the form of graphs or charts, but can include such things as stem-and-leaf displays and pictograms. Descriptive statistics, however, summarize the data and reduce it to a form which is not convertible back into the original data. We can only see what the presentor wants us to see.

**Inferential** statistics make generalizations about data or make decisions when comparing datasets. The goal is to determine if the effects observed are random or can be attributed to the phenomenon under study. It is inferential statistics with which we will concern ourselves in this class.

### Basic Concepts for Statistics

There are several terms which any student of statistics needs to be familiar with. These terms form the vocabulary of statistics, and are necessary to the understanding of the statistical techniques and the results.

An individual number is a data point. If you are studying the temperature and you record today's temperature at noon, then that is a data point. Multiple data points form a data set. A collection of temperature readings for the month of January is a data set. Each line of data in the data set is referred to as an observation. The data points on an observation can be of different types and can measure different things. As an example, if you are studying the physical fitness of a group of college sophomores, then the observation may have data points for the gender, height, weight, percent body fat, and body mass index for each person. Each person would represent an observation in the data set.

Data points can be one of two types. **Continuous** data points are usually decimal, or real, numbers and represent phenomenon that can be measured at any level of precision the researcher wants and the equipment will allow. Height is a continuous type of data as is air pollution measured in parts-per-million. Only the precision of the instrument doing the measurement restricts the precision of the measurement. **Discrete** data points are whole, or integer, numbers and represent phenomenon which are measured in whole units. The number of children or cars a family has are examples of discrete data points. Despite the assertion of the Census Bureau that the average American household has 2.3 children, no one has found 0.3 of a child.

Data points are measurements of phenomena, and phenomena can be classified into four categories. The categories are : **nominal**, **ordinal**, **interval**, and **ratio**. Data from a **nominal** phenomenon have no order and are purely categorical. The color of the car you drive is a nominal data point. Nominal data's only purpose is to group observations in some manner without regard to hierarchy. **Ordinal** data, however, does have a hierarchy. Each value in an ordinal data set has a higher or lower value relative to another data point. An example of ordinal data is the alphabet. The letter

'a' is the lowest while the letter 'z' is the highest, ranking from 'a' to 'z'. Ordinal data, therefore, groups observations within some hierarchy.

The next category, **interval**, has categories that are ordered the same as ordinal, but in this case the distance between categories has meaning. A Likert scale is a prime example of interval data. A Likert scale is used primarily in survey research. We've all seen one, we just may not have known what it was we were seeing. The scale usually has four or five responses, each with a number assigned to it. The responses range from lowest to highest, or worst to best. The distance between responses in a quantifiable difference and is of importance. One person's response of '5' and another's response of '2' means a difference of opinion, and a measurable difference. By asking several questions about a particular subject and summing the responses the researcher can build a scale to measure opinion about a subject. Temperature is another example of interval data. Each degree is an indication of more or less heat. The difference between the temperature today and the temperature yesterday is a measurable, meaningful quantity.

The last category, **ratio**, is often lumped together with interval, but there are differences between then two. Ratio data is scalar, just like interval, and the difference between data points is a quantifiable difference, but the ratio data starts at some point recognized as zero. The zero point means the absence of the phenomenon rather than, as in a Likert scale 'no opinion'. An example of ratio data is the weight of an object. If an object has a weight of zero, we say it has no weight. Similarly, the number of children in a family is a ratio measure. If the number is zero, then there are no children in the family.

The difference in the types of data is important, especially in later statistical studies. You use the different types of data in different ways to design studies and to do analysis. A nominal data value is used to classify the observations into categories (treatment and non-treatment, for example) while the analysis is done on a ratio data value (increase in growth).