MATH 1070 Introductory Statistics
Lecture notes
Descriptive Statistics and Graphical Representation

Objectives:

1. Learn the meaning of descriptive versus inferential statistics
2. Identify bar graphs, pie graphs, and histograms
3. Interpret a histogram
4. Compute individual items for a five number summary
5. Compute and use standard variance and deviation
6. Understand and use the Empirical Rule

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 phenomena 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).

Graphical Display of Data

It is easiest to interpret and absorb data when it is presented in a graphical format. There are several different ways to display data depending upon the purpose.

A bar graph/bar chart is a simple graphical technique for illustrating a categorical data set. (Hayter) The displayed value is either the count or percent of items/individuals who fall in a particular category. Each category has a bar whose length is proportional to the frequency associated with that category. The bars may be either vertical or horizontal.

A pie chart divides a complete circle into slices, one corresponding to each category with the central angle of the slice proportional to the category relative frequency. A pie chart emphasizes the proportion of the total data set that is taken up by each of the categories. If a data set of n observations has r observations in a specific category, then that category receives a slice of the pie with an angle of $\frac{r}{n} \times 360$.

A histogram is similar to a bar chart, but the values represented are numerical rather than categorical. Whereas in a bar chart the values on the horizontal, or x-axis, list the various

categories in a histogram the x-axis is numerical. A histogram consists of a number of bands whose length is proportional to the number of data observations that take a value within that band. It is important to carefully construct the width of the bands in a histogram. (Hayter)

The width of the bands is related to the number of bands. A suggested method for choosing how many bands is

Less than 25 observations	5 or 6 bands
25 to 50 observations	7 to 14
More than 50	15 to 20

There are about 5 steps to constructing a histogram (Mendenhall).

What are we looking for in a histogram? The shape of the data, basically, but also a few other characteristics.

- Overall pattern
- Striking, noticeable deviations from that pattern
- Its shape, center, and spread of the data
- Any observation that is significantly apart, an outlier

An important question about the shape is, "Is it symmetric?" If the left and right halves of the distribution are approximately mirror images, then the data are said to be symmetric. If the data aren't symmetric, then we talk about the skewness of the data.

- Right, positively, skewed right-hand tail is longer and flatter
- Left, negatively, skewed left-hand tail is longer and flatter

It might appear that there isn't a single, central point to the data. There might be two (or more). If there are two separate peaks, or humps, in the data then we say it is bimodal. What does this mean? It may be that the data are actually from two different groups or populations. For example, a data set measuring some attribute of people may more usefully be separated into one data set for men and one for women.

An outlier is a data point that appears to be separate from the rest of the data set. What do we do with an outlier? It may be a misrecorded value, or a typo. It could be a valid data point that is just that extreme. (Bill Gates is an outlier in a sample of CEO salaries, but he's not an error.)

For small datasets, a stem and leaf plot is another way to display data. In a stem and leaf plot, the "stem" and "leaf" are parts of the data points. The leaf is usually the rightmost digit, in the one's place. The stem is the remaining digits. This isn't a hard and fast rule, you can make the stem the leftmost digit, if you prefer. The rightmost is more common.

List the stems in a column, smallest to largest. Then list each leaf value, to the right of the stem, in increasing order. If there are two values with the same value, list them together.

If you rotate the stem and leaf plot on its side, you will get the same type of bar graph as with a histogram.

Measures of Central Tendency

Graphical descriptions of data sets are only so useful. If the data set is large, constructing the graph (or any type) becomes tedious. Though we can see the shape, center, and spread of a distribution from a histogram or stem and leaf plot, we can't quantify it. We need to figure out a numerical method for describing data.

Important terms to learn at this point **statistic** and **parameter**. A statistic is a numerical descriptive measure computed from sample data. A parameter is a numerical descriptive measure of a population. Since we can't always get a complete population (N), we have to rely on samples (n) and the statistics computed from them. We use different symbols to represent statistics than for parameters.

Our three most common measures of central tendency are arithmetic mean, median, and mode.

The arithmetic mean, or average, is the sum of the data values in a data set divided by the number of items. (show formula) If the mean is for the population, the symbol used is μ ("mu"). If the mean is for a sample, the symbol is $\bar{x}$ ("x-bar") (or "y-bar").

In statistics we use several Greek letters to represent various concepts or operations. Here we use the "big sigma" (Σ) to represent the summation of a series of items.

The median is the middle number of a series of data points when the data points are arranged in ascending, or descending, order. If there are an odd number of items, the median is the middle item $[\frac{(n+1)}{2}\ ]$. If there are an even number of items, the median is the average of the two items in the middle $[\ \frac{n}{2} + (\frac{n}{2} + 1)\ ]$. The median is the point where there is one-half (50%) of the data set on either side, similar to the median in a road.

Quartiles

We can measure the spread of the data by identifying the quartiles. As the word implies, a quartile is one-quarter (25%) of the data. We begin by arranging the data points in order. We then find the middle, or median, point. The median is, conveniently, the second quartile since it is at the 50% (one-half) point in the data. The median point between the true median and the lower end of the list of data items is the first quartile. This is where 25% of the data is below this point. The median point between the true median and the upper end of the list of data items is the third quartile. This is where 75% of the data is below this point.

If we divide the data set into 100 equal slices, we call these percentiles. "Per cent" means "of 100" so each percentile represents 1/100th of the data set. The quartiles are, then, the 25th, 50th, and 75th percentiles, respectively.

Another definition of percentile is

> The 100pth percentile of a data set is a value of y located so that 100p% of the data lies to the left of the 100pth percentile and 100(1-p)% of the data lies to the right.

Example: If your grade in an industrial engineering class was located at the 84th percentile, then 84% of the grades were lower than your grade and 16% were higher. (Mendenhall)

Five Number Summary

A quick way to summarize a data set is by creating a five number summary. This is the minimum data point, the first, second, and third quartile, and the maximum data point.

Sample Variance and Standard Deviation

While a five number summary is convenient, there is a simpler way to describe a data set. We can compute the standard deviation to measure the distance from the mean, or central point, of the data set. This gives us a measure of the spread of the data set. First we have to compute the variance of the data set, then we take the square root to find the standard deviation.

We must pay attention to how we compute the variance, however, because the data will make this value "disappear."

The variance is computed by summing the square of the difference between each item in the data set and the mean of the data set ($x\bar{x}$) and dividing by ($n-1$). We have to sum the square of the difference because the number of observations greater than the mean is the same as those less than the mean, and the total difference of the greater observations is exactly equal to the total difference of the lesser observations.

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

```
Example from Moore:

Metabolic rates for participants in dieting study (in calories)

1792   1666   1362   1614   1460   1867   1439

mean = 1600 standard deviation = 189.24
```

The symbol used to represent the sample variance is little s squared (s^2). For the population variance we use little sigma squared (σ^2).

The standard deviation of the data is the square root of the variance. The symbol used is either "little s" (s) or "little sigma" (σ), depending upon whether it is the sample deviation or the population deviation.

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

So, to find the standard deviation (and variance) of a data set we follow these steps (Cryer and Miller):

1. Compute the deviations for the data set ($x_i - \bar{x}$).
2. Square all the deviations and sum them.
3. Divide the sum by $n-1$.
4. Finally, take the positive square root to get the standard deviation s.

It is important to note a few things about the standard deviation:

- The units for the standard deviation are the same as the units for the observations.
- The standard deviation is never negative. It is zero if, and only if, all the values are equal.
- The standard deviation is strongly influenced (not resistant) by extreme observations and outliers.

Degrees of Freedom

Note that we divide by $n-1$ to compute the variance. Why would we do that? It is because not all the data items are "free" to choose their value. Remember that all the deviations sum to zero? If that is the case, the last data point to be summed **has** to make up any difference (positive or negative) between the sum so far and zero. The other data points $(1, 2, 3, \ldots, (n-1))$ effectively made the decision for the last, or n^{th}, data point.

Another way to think about this is to consider what happens when you go out with a group of friends. Eventually the bill comes and it's time to pay up. The bill starts around the table and each person puts in what they think is their share of the bill, or what they have (usually the case). When the bill gets to the last person (you, maybe) the difference between what is owed and what has been contributed (don't forget the tip!) has to be made up. If you're lucky there is more than what you owe and your share is less. Otherwise, you get to pay more.

The Empirical Rule (68-95-99.7)

If the distribution of the data is approximately symmetrical with a single peak we usually refer to this as a **normal distribution**. The distribution is centered around the mean and has noticeable changes in slope at various points related to the standard deviation. We can say this about the data:

1. 68% of the observations fall within 1 standard deviation of the mean.
2. 95% of the observations fall within 2 standard deviations of the mean.
3. 99.7% of the observations fall within 3 standard deviations of the mean.

How can we know this? It depends upon the curve and the characteristics of a symmetrical curve, or density curve. The area underneath the curve is 1 and always positive. The curve bends at each multiple of the standard deviation. These bending points are called inflection points.

Standardization

The phrase, "Comparing apples to oranges," is used to imply that two things being compared really can't be. An example would be SAT to ACT scores. The tests have different measuring scales and it's hard to know if a score of 620 on the SAT verbal section is equal to a score of 18 on the ACT verbal. So how can we do this?

If the data being compared come from normal distributions, we can transform the data so that the two sets are equivalent. We call this transformation **standardization**. By doing this we change the values from their original units into standard deviation units. We also change the starting measuring point to be the mean of the distribution and convert that value to zero.

The result of this operation is called a **z-score**. The z-score tells us how far above or below the mean (in units of standard deviations) an observation is. It also allows us to compare the values since both now are measured the same (in terms of standard deviations) and measured from the same starting point (the mean, zero). Taking advantage of the fact that the area under the normal curve is equal to one (1), we can measure the percentage area for each value.

The formula for this transformation is

$$z = \frac{x - \bar{x}}{s}$$

To compute the area, or percentage, at a given z-score, we must always remember that the mean (0) is at 50%. This means that any value to the left (below) the mean will have less than 50% area below it and any value to the right (above) the mean will have more than 50% area below it. The area above the z-score will be the inverse.

We find the area under the curve using a table. This table can be either one sided or two sided, depending upon the author. Ours is two sided.

First we compute the z-score to two decimal places. Then we look at the table. The column lists the values for the gross measurement, the one's digit and the tenth's. The columns list the fine measurement, the hundredth's. Where the two intersect is the area under the curve for that z-score.