MATH 134 - ELEMENTARY STATISTICS

In this class we will focus on the ideas and the concepts behind good statistical thinking more than in the actual methods.

We will consider some descriptive techniques that are used to summarize the information contained in data sets.

We will then consider the different types of studies that give rise to statistical information and how their design affects the analysis and conclusions that can be drawn from them.

We will consider Probability Theory and its use in statistical reasoning. We will then develop some statistical inference to draw conclusions on population characteristics using samples.

Statistics is the study of variability or uncertainty: Question is, how do we measure uncertainty and what do we do about it?

Suppose we want to estimate the percentage, \( \Theta \), of the deer who live in the Yosemite Park as of May 30, 2012, who have chronic wasting disease (a transmissible neurological disease of deer that produces small lesions in the brain.)

What do we know about this percentage? Well, we don’t know the value of \( \Theta \) exactly, however, let’s assume that most of the deer appear healthy, so it’s probably rather small. We do have substantial uncertainty about its precise value.

We can reduce my uncertainty by gathering data on the disease status of deer. How shall we collect the data?

The set:

\[
P = \{ \text{the deer who live in Yosemite Park as of May 30, 2012} \}
\]

is a population, which is a collection of elements of interest to us.

The aspect of each of these population elements we’re interested in is, for any deer encountered – does the deer have chronic wasting disease or not?

Things that can be measured on population elements are called variables. In this case, the variable of interest takes on only two values, \{yes, no\}. Such variables are called dichotomous or binary.
In principle we could perform a complete census of the entire population. Instead, it’s natural to choose a subset, $S$ of $P$ and evaluate the variables of interest only on the population elements in the subset.

Such a subset is called a sample from the population, $P$. If the sample is chosen well, it seems like a good idea to use the data in the sample to make an estimate of (i.e. and educated guess at) the population parameter, $\Theta$, of interest.

An estimate, $\hat{\Theta}$ of a population parameter, $\Theta$, is also sometimes called a statistic.

---

**Data Types**

Ex 1: Genetic phenotype. Suppose eye color is a variable you’re studying and may take on only two values (brown, blue). Eye color has no unique place on the number line, and is classified as dichotomous or binary. Similarly, hair color might take on four values (brown, black, red or white). Variables like this are said to occur on a **nominal scale** of measurement.

Ex 2: Success in running a maze might be recorded as

1 = very slow  
2 = slow  
3 = moderate  
4 = fast  
5 = very fast

There are still no unique places on the number line for such values, HOWEVER, (unlike example 1), there’s a natural ordering. Variables like this are said to occur on an **ordinal scale**. Some other names for nominal and ordinal variables are qualitative and categorical.

---

Ex 3: Size of a plant: Two measures of the size of a plant would include its height and the number of leaves it has.

Unlike the situations in Ex 1 & 2, the values taken on by these variables do have unique places on the number line.

**Important characteristics:**

- **Constant size interval** so that the concept of 1 unit on measurement scale stays consistent.
- **True Zero** with a direct physical meaning exists.

Variables like these allow us to make meaningful statements about ratios (for example, plant $C$ is 4.1 times taller than plant $B$) and are said to occur on a **RATIO** scale.

---

Ex 4: Temperature: (measured either in °C or °F) does have a constant size interval but lacks a true zero. (When it’s 80°F outside you can’t correctly say that it’s twice as hot as when it’s 40°F.)

Variables like these are said to occur on an **interval** scale. Other examples of interval variables include time of day, time of year, and compass point, all of which are said to occur on circular scales.

Some other names for **RATIO** and **interval** variables are quantitative and numerical.
Data Types
Variables can be classified as:

- **Quantitative data.** Correspond to observations measured on a numerical scale. **Important distinction:**
  - **Discrete** when the values can differ by fixed amounts like in **number of leaves**.
  - **Continuous** differences in values can be arbitrarily small like in **age**.

- **Qualitative data.** Correspond to observations classified in groups or categories like in **sex** and **marital status**. Sometimes the groups have some ordering, as in the case of **grades**. Of particular importance are binary variables that can take only two values.

### Classify the following variables as either: qualitative - (nominal scale or ordinal), quantitative-(interval or ratio), discrete or continuous

**VISUALIZE THE DATA**

- Records of whether an electrical device is working or not.
- The depth of the snow pack at a monitoring station in the Sierras.
- The number of female students in MATH 134.
- The final grade of a student in MATH 134.
- The State where a given car is registered.
- Temperature in °K.
- The number of calls to 911 in a given month.

Graphical Descriptive Methods
When a large sample is collected there is a need to produce summaries of the information contained in it.

There are graphical and numerical tools that are commonly used by statisticians. Simple but powerful graphical tools to produce summaries of data sets include **stem plots** and **histograms**.

**Stem Plots**
The 1987 *Farmer’s Almanac* gives the growing season for selected U.S. cities as reported by the National Climatic Center. The growing season is defined as the average number of days between the last frost in the spring and the first frost in the fall.

See handout
Histograms

Case Study 2: Chicago Civil Engineer Test Scores

Information is available from 131 hospitals. We show a histogram of the average length of stay measured in days for each hospital. The area of each block is proportional to the number of hospitals in the corresponding class interval.

In this example all the intervals have the same length, so the heights of the blocks give all the information about the number of hospitals in each class.

There are 7 class intervals corresponding to

- 6 to 8 days
- 8 to 10 days
- 10 to 12 days
- 12 to 14 days
- 14 to 16 days
- 16 to 18 days
- 18 to 20 days

Note that the class that corresponds to 14 to 16 days is empty and that the class with the highest count of hospitals is the one of 8 to 10 days.

<table>
<thead>
<tr>
<th>Income level in $</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 1,000</td>
<td>1</td>
</tr>
<tr>
<td>1,000 – 2,000</td>
<td>2</td>
</tr>
<tr>
<td>2,000 – 3,000</td>
<td>3</td>
</tr>
<tr>
<td>3,000 – 4,000</td>
<td>4</td>
</tr>
<tr>
<td>4,000 – 5,000</td>
<td>5</td>
</tr>
<tr>
<td>5,000 – 6,000</td>
<td>5</td>
</tr>
<tr>
<td>6,000 – 7,000</td>
<td>5</td>
</tr>
<tr>
<td>7,000 – 10,000</td>
<td>15</td>
</tr>
<tr>
<td>10,000 – 15,000</td>
<td>26</td>
</tr>
<tr>
<td>15,000 – 25,000</td>
<td>26</td>
</tr>
<tr>
<td>25,000 – 50,000</td>
<td>8</td>
</tr>
<tr>
<td>50,000 and over</td>
<td>1</td>
</tr>
</tbody>
</table>
If we graph the percentages, the resulting histogram is called a **relative frequency histogram**.

Notice that the class interval of incomes above $50,000 has been ignored.

Yet, this is misleading, because it looks like there’s a many more families with incomes over $25,000 than under $7,000.

It is a mistake to set the heights of the blocks equal to the percentages in the table. This is because some class intervals are longer than others.

To figure out the height of a block divide the percentage by the width of the interval. The resulting histogram is called a **density histogram**.

In a density histogram, the areas of the blocks represent percentages

The table needed to calculate the heights of the blocks looks like:

<table>
<thead>
<tr>
<th>Income level in $</th>
<th>percent</th>
<th>( \div ) width ( \times $1,000 )</th>
<th>= height</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 1,000</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1,000 – 2,000</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2,000 – 3,000</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3,000 – 4,000</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4,000 – 5,000</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5,000 – 6,000</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6,000 – 7,000</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>7,000 – 10,000</td>
<td>15</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>10,000 – 15,000</td>
<td>26</td>
<td>5</td>
<td>5.2</td>
</tr>
<tr>
<td>15,000 – 25,000</td>
<td>26</td>
<td>10</td>
<td>2.6</td>
</tr>
<tr>
<td>25,000 – 50,000</td>
<td>8</td>
<td>25</td>
<td>.32</td>
</tr>
<tr>
<td>50,000 and over</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This is the resulting **Density Histogram**

Distribution of family income in the US in 1973
**Vertical scale**

What is the meaning of the vertical scale in a density histogram?

Remember that the area of the blocks is proportional to the percents. A high height implies that large chunks of area accumulate in small portions of the horizontal scale.

This implies that the density of the data is high in the intervals where the height is large. In other words, the data are more crowded in those intervals.

**Pie Charts are a bad idea!**

From the R manual page for the `pie` function:

*Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.*

Cleveland (1985), page 264: "Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements." This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.

**Other Graphical Methods:** The pie charts correspond to the proportion of ice cream flavors sold annually by a given brand.

Consider the following: Data from the 2003 Census produce the following for housing units in the New York City area that are either occupied by the owner or rented out. Total number of units are shown below.
### Questions:

1. The owner-occupied percents add up to 100.2% and the renter-occupied percents add up to 100.0%, why?

   **Ans:** The answer to the first question is that there is rounding involved in the calculation of the percentages.

2. The percentage of one-room and three-room units is much larger for renter-occupied housing. Is that because there is less renter-occupied housing in total?

   **Ans:** As for the second question, the fact that we are taking percentages accounts for the difference in totals, so a smaller total of renter-occupied units does not explain the difference.

3. Which are larger on the whole: the owner-occupied units or the renter-occupied units?

   **Answer:** see next slide

---

### Table

<table>
<thead>
<tr>
<th>Number of Rooms</th>
<th>Owner Occupied (%)</th>
<th>Renter Occupied (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>2.8</td>
</tr>
<tr>
<td>3</td>
<td>1.4</td>
<td>22.7</td>
</tr>
<tr>
<td>4</td>
<td>9.7</td>
<td>34.5</td>
</tr>
<tr>
<td>5</td>
<td>23.3</td>
<td>22.6</td>
</tr>
<tr>
<td>6</td>
<td>26.4</td>
<td>10.4</td>
</tr>
<tr>
<td>7</td>
<td>17.5</td>
<td>3.6</td>
</tr>
<tr>
<td>8</td>
<td>10.4</td>
<td>1.2</td>
</tr>
<tr>
<td>9</td>
<td>5.0</td>
<td>0.5</td>
</tr>
<tr>
<td>≥ 10</td>
<td>6.4</td>
<td>0.7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100.2</strong></td>
<td><strong>100.0</strong></td>
</tr>
<tr>
<td><strong>Number</strong></td>
<td><strong>72.2 million</strong></td>
<td><strong>33.6 million</strong></td>
</tr>
</tbody>
</table>
Average and spread in a histogram

A histogram provides a graphical description of the distribution of a sample of data. If we want to summarize the properties of such a distribution we can measure the center and the spread of the histogram.

These two histograms correspond to samples with the same center. The spread of the sample on top is smaller than that of the sample in the bottom.

To obtain an estimate of the center of the distribution we can calculate an average.

The average of a list of numbers equals their sum, divided by how many they are

Thus, if 18; 18; 21; 20; 19; 20; 20; 20; 19; 20 are the ages of 10 students in this class, the average is given by

\[
\frac{18 + 18 + 21 + 20 + 19 + 20 + 20 + 20 + 19 + 20}{10} = 19.5
\]

In the hospital data that we considered in a previous slide, the data corresponded to the average length of stay of patients in each hospital in the survey. This means that the length of stay of all patients in a given hospital were added and the sum divided by the number of patients in that hospital.

Mean and median - Measures of Centrality

This histogram corresponds to the rainfall over periods of 10 days in an area of the central plains of Venezuela. The average or mean rainfall is 37.65 mm. We observe that only about 30% of the observations are above the average. Notice that this histogram is not symmetric with respect to the mean.

A symmetric histogram will look like this. In this case 50% of the data are above the average.

The median of a histogram is the value with half the area to the left and half to the right. In a symmetric histogram the median and the average coincide.

The relationship between the mean and the median determines the shape of the tails of a histogram.
The mean is very sensitive to **extreme observations**, so when dealing with variables like income or rainfall, that exhibit very long tails, it is preferable to use the median as a measure of centrality.

**Measures of Spread**

As we saw before, two samples can have the same center and be scattered along their ranges in different ways. To measure the way a sample is spread around its average we can use the **standard deviation**, or SD.

The SD of a list of numbers measures how far away they are from their average.

Thus a large SD implies that many observations are far from the overall average.

Most observations will be one SD from the average. Very few will be more than two SDs away.

**Calculating the SD**

To calculate the **standard deviation** of a sample follow the steps:

- Calculate the average
- Calculate the list of deviations from the average by taking the difference between each datum and the average.
- Calculate the r.m.s. size of the resulting list.

$$SD = \text{r.m.s. deviation from average}.$$ 

Consider the list 1, 2, 9. Then

$$\text{average} = \frac{1 + 2 + 9}{3} = 4$$

The list of deviations is -3, -2, 5. Then

$$SD = \sqrt{\frac{(-3)^2 + (-2)^2 + 5^2}{3}} = \sqrt{\frac{38}{3}} \approx 3.6$$

Let’s think about how to calculate the SD which measures how far away the data elements are from their mean.

Consider the tiny fake data set with $n = 3$ values, $(y_1, \ldots, y_3) = \{1, 2, 9\}$, whose mean, $\bar{y}$, is 4 and whose histogram looks like this:

How should we quantify the spread of a list of numbers? Let’s discuss the possibilities.
Using a calculator

Most scientific calculators will have a function to calculate the average and the SD of a sample. The steps needed to obtain those values vary from model to model.

The important fact is that most calculators do not produce the SD as we have defined it here. They consider the sum of the squares of the deviations over the total number of data minus one. So, if you obtain the SD from your calculator (or spreadsheet), say SD*, then

\[ SD = \sqrt{\frac{\text{number of entries} - \text{one}}{\text{number of entries}}} \times SD^* \]

Some calculators have both, SD and SD*. Please read the manual of your calculator regarding this fact.

Notice that the units of SD are the same as the original data. So if the data were measured in years, SD is also in years.

Problems

Problem 1: Both the following lists have the same average of 50. Which one has the smaller SD and why? (Do no computations)

1. 50,40,60,30,70,25,75
2. 50,40,60,30,70,25,75,50,50,50

Repeat for the following two lists

1. 50,40,60,30,70,25,75
2. 50,40,60,30,70,25,75,99,1

Solutions

Problem 1: Both the following lists have the same average of 50. Which one has the smaller SD and why? (Do no computations)

The second list has more entries at the average, so the SD is smaller.

Repeat for the following two lists

1. 50,40,60,30,70,25,75
2. 50,40,60,30,70,25,75,99,1

The second list has two wild observations, 99 and 1, which are away from the average, so the SD is larger.

Empirical Rule

SDs are a pain to compute by hand or with a calculator, so it would be good to have a simple way to roughly approximate the SD of a list of numbers.

- Roughly 68% of the observations are within one SD of the average.
- Roughly 95% of the observations are within two SDs of the average.
- Roughly 99% of the observations are within three SDs of the average.

These statements are more accurate when the distribution is symmetric.
Problem 2: Consider the following list of 20 numbers

0.7  1.6  9.8  3.2  5.4  0.8  7.7  6.3  2.2  4.1
8.1  6.5  3.7  0.6  6.9  9.9  8.8  3.1  5.7  9.1

1. Without doing any arithmetic, guess whether the average is around 1, 5 or 10.

2. Without doing any arithmetic, guess whether the SD is around 1, 3 or 6.

Solutions

Consider the list of numbers

0.7  1.6  9.8  3.2  5.4  0.8  7.7  6.3  2.2  4.1
8.1  6.5  3.7  0.6  6.9  9.9  8.8  3.1  5.7  9.1

1. Without doing any arithmetic, guess whether the average is around 1, 5 or 10.
   Only three of the numbers are smaller than 1, none are bigger than 10, so the average is around 5.

2. Without doing any arithmetic, guess whether the SD is around 1, 3 or 6.
   If the SD is 1, then the entries 0.6 and 9.9 are too far away from the average. The entries are too concentrated around 5 for the SD to be 6. So the 3 is the most likely value.

The Normal Curve - Ch. 3 in Moore’s Text

The normal curve plays a fundamental role in probabilities since it can be used as an approximation in a variety of problems.

Actually, there’s not just one normal curve, there are infinitely many of them. For any values \( \bar{y} \) and \( s \) you can imagine for the mean and SD respectively, of a single-variable data set, there’s a corresponding normal distribution with that mean and SD.

Let’s take a look at the histogram for the number of calories from a cereal data set.
The Standard Normal Density

The Gaussian or normal curve corresponds to the following formula

\[ y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \]

where \( e = 2.71828 \ldots \)

and corresponds to the graph

We observe that the curve is symmetric around zero and that most of the area is concentrated between \(-3\) and \(+3\).

The probability of finding a data element within an interval is the corresponding area under the curve.

So, for our cereal data, the probability that any randomly chosen cereal has between 87 and 127 calories that corresponds to \((-1,+1)\) standard units is: about 95% according to the Empirical Rule.

Standard Units - pg. 77

Doing calculations with the normal curve requires the use of a table. Table are available for the standard normal curve and try require that observations be transformed to standard units. Standard units are obtained by calculating the number of SDs that an observation is above or below the average.

Given a list of numbers, we convert to standard units by subtracting the average and dividing by the SD.

Q: Imagine that a data set lists cereals with corresponding calories and one of the cereals has 127 calories. Find the standard units corresponding to 127 and 97 calories. \((\mu = 107\) and the SD, \(s = 20)\)

A:

\[
\frac{127 - 107}{20} = +1 \quad \frac{97 - 107}{20} = -0.5
\]

Table A-690 at the end of the book calculates the areas under the standard Normal curve. For example,

Q: Calculate the probability of a choosing a cereal at random that has less than 87 calories.

A: This corresponds to the entry

\[
\frac{87 - 107}{20} = -1 \quad \text{standard units from the mean}
\]

In the table, the probability from \((-\infty, z)\) is given as \(0.1587 \times 100 = 15.87\%\).

Since the Normal distribution is symmetric, the probability from \((z, +\infty)\) is \(100 - 15.87 = 84.13\%\).

So that the probability of choosing a cereal at random that has more than 87 calories is 84\%.
Q: Calculate the probability of choosing a cereal at random that has less than 117 calories.

A: This corresponds to the entry
\[ \frac{117 - 107}{20} = +0.5 \text{ S.U. from the mean} \]

Since the normal curve is symmetric, the area corresponding to \((-z, 0)\) is equal to the area corresponding to \((0, z)\) for any value of \(z\). We can calculate symmetric intervals around 0 using the table at the end of the book. For example,

Q: Calculate the probability of choosing a cereal at random that has less than 117 calories and more than 97 calories.

A: These values correspond to -0.5 and +0.5 S.U. (because they are 10 calories from the mean and the \(SD = 20\)).

In the table, the probability of randomly choosing a cereal with less than 97 calories is given as \(0.3085 \times 100 = 30.85\%\), whereas, the probability of randomly choosing a cereal with less than 117 calories is given as \(0.6915 = 69.15\%\). Therefore, .... ? .....