Math 243: Introduction to Methods of Probability and Statistics

Statistics, while originally developed for the physical sciences, now pervade the social sciences, politics, sports, and business.

Statistics, while originally developed for the physical sciences, now pervade the social sciences, politics, sports, and business. Most articles in the main section of the newspaper (as well as the sports section) contain statistics (especially now that polling data fills the news).

Statistics, while originally developed for the physical sciences, now pervade the social sciences, politics, sports, and business. Most articles in the main section of the newspaper (as well as the sports section) contain statistics (especially now that polling data fills the news).

But statistics can be misleading

Statistics, while originally developed for the physical sciences, now pervade the social sciences, politics, sports, and business. Most articles in the main section of the newspaper (as well as the sports section) contain statistics (especially now that polling data fills the news).

But statistics can be misleading in a couple of ways,

Statistics, while originally developed for the physical sciences, now pervade the social sciences, politics, sports, and business. Most articles in the main section of the newspaper (as well as the sports section) contain statistics (especially now that polling data fills the news).

But statistics can be misleading in a couple of ways, and one of the main goals of this course is for you to be equipped to judge the mathematical significance of statistics you encounter.

Statistics, while originally developed for the physical sciences, now pervade the social sciences, politics, sports, and business. Most articles in the main section of the newspaper (as well as the sports section) contain statistics (especially now that polling data fills the news).

But statistics can be misleading in a couple of ways, and one of the main goals of this course is for you to be equipped to judge the mathematical significance of statistics you encounter.

Before going into the mathematics, let us clarify a common non-mathematical way in which statistics can be misleading, namely by telling only part of the story.

Before going into the mathematics, let us clarify a common non-mathematical way in which statistics can be misleading, namely by telling only part of the story.

For example, in the Register-Guard one can on occasion read arguments which bring up the fact that Oregon has the third-highest income tax in the country to imply that our taxes are extraordinarily high.

Before going into the mathematics, let us clarify a common non-mathematical way in which statistics can be misleading, namely by telling only part of the story.

For example, in the Register-Guard one can on occasion read arguments which bring up the fact that Oregon has the third-highest income tax in the country to imply that our taxes are extraordinarily high. But in total tax bill

Before going into the mathematics, let us clarify a common non-mathematical way in which statistics can be misleading, namely by telling only part of the story.

For example, in the Register-Guard one can on occasion read arguments which bring up the fact that Oregon has the third-highest income tax in the country to imply that our taxes are extraordinarily high. But in total tax bill (including for example sales tax),

Before going into the mathematics, let us clarify a common non-mathematical way in which statistics can be misleading, namely by telling only part of the story.

For example, in the Register-Guard one can on occasion read arguments which bring up the fact that Oregon has the third-highest income tax in the country to imply that our taxes are extraordinarily high. But in total tax bill (including for example sales tax), Oregon ranks 34th according to the Tax Foundation.

We will *not* be talking about ways in which to spot incomplete, misleading statistics,

We will *not* be talking about ways in which to spot incomplete, misleading statistics, which is more a matter of common sense and experience.

We will *not* be talking about ways in which to spot incomplete, misleading statistics, which is more a matter of common sense and experience.

But even statistics which should tell the whole story can be misunderstood unless we understand the "shape" of the data.

For example, suppose hypothetically that in total tax bill the data was:

For example, suppose hypothetically that in total tax bill the data was:
NY 15%; CA 14%; OR 10.05%; ... lots of other states between 9.5 and 10 %; ... MS 7% etc.

For example, suppose hypothetically that in total tax bill the data was:
NY 15%; CA 14%; OR 10.05%; ... lots of other states between 9.5 and 10 %; ... MS 7% etc. Then one could rightly say that Oregon had the third-highest amount of taxes, but would that imply that its taxes are out of line?

For example, suppose hypothetically that in total tax bill the data was:
NY 15%; CA 14%; OR 10.05%; ... lots of other states between 9.5 and 10 %; ... MS 7% etc. Then one could rightly say that Oregon had the third-highest amount of taxes, but would that imply that its taxes are out of line?

One can loosely say that statistics is a set of tools which measure, in various ways, the shape of data.

For example, suppose hypothetically that in total tax bill the data was:
NY 15%; CA 14%; OR 10.05%; ... lots of other states between 9.5 and 10 %; ... MS 7% etc. Then one could rightly say that Oregon had the third-highest amount of taxes, but would that imply that its taxes are out of line?

One can loosely say that statistics is a set of tools which measure, in various ways, the shape of data. Let's get started putting together those tools.

Review of chapter 1

Review of chapter 1

We typically describe the distribution of a variable

Review of chapter 1

We typically describe the distribution of a variable (which always represents some quantity we want to measure)

Review of chapter 1

We typically describe the distribution of a variable (which always represents some quantity we want to measure)by a **bar graph** or **pie chart** or **table** if it is a categorical variable.

Review of chapter 1

We typically describe the distribution of a variable (which always represents some quantity we want to measure)by a **bar graph** or **pie chart** or **table** if it is a categorical variable.

We describe the distribution of a quantitative variable by a list of values or, more often, a **histogram.**

Review of chapter 1

We typically describe the distribution of a variable (which always represents some quantity we want to measure)by a **bar graph** or **pie chart** or **table** if it is a categorical variable.

We describe the distribution of a quantitative variable by a list of values or, more often, a **histogram.** A description of the data usually includes

Review of chapter 1

We typically describe the distribution of a variable (which always represents some quantity we want to measure)by a **bar graph** or **pie chart** or **table** if it is a categorical variable.

We describe the distribution of a quantitative variable by a list of values or, more often, a **histogram.** A description of the data usually includes

- *Center*: This is the **median** value. It is either the middle value (if there are an odd number of values), or

half-way between the two middle values if there are an even number of values.

half-way between the two middle values if there are an even number of values.

- *Spread*: We'll talk more about this in the next section. A crude measure of spread is given by two numbers: the distance of the lowest value from the center, and the distance of the largest value from the center.

half-way between the two middle values if there are an even number of values.

- *Spread*: We'll talk more about this in the next section. A crude measure of spread is given by two numbers: the distance of the lowest value from the center, and the distance of the largest value from the center.

- *Shape*.

half-way between the two middle values if there are an even number of values.

- *Spread*: We'll talk more about this in the next section. A crude measure of spread is given by two numbers: the distance of the lowest value from the center, and the distance of the largest value from the center.

- *Shape*.
  The data is **symmetric** if the histogram to the left of the center looks like a mirror image of the histogram to the right of the center. Histograms will almost never

be perfectly symmetric - we are only worried about approximate symmetry here.

be perfectly symmetric - we are only worried about approximate symmetry here.

The data is **left skewed** if the data extends further to the left of center than to the right of center. (And of course **right skewed** is *vice versa*.)

be perfectly symmetric - we are only worried about approximate symmetry here.

The data is **left skewed** if the data extends further to the left of center than to the right of center. (And of course **right skewed** is *vice versa*.)

The data is in **clusters** if there is more than one region in which a lot of the data is concentrated.

be perfectly symmetric - we are only worried about approximate symmetry here.

The data is **left skewed** if the data extends further to the left of center than to the right of center. (And of course **right skewed** is *vice versa*.)

The data is in **clusters** if there is more than one region in which a lot of the data is concentrated.

It is usually interesting to try to understand *why* the data has the shape that it has.

be perfectly symmetric - we are only worried about approximate symmetry here.

The data is **left skewed** if the data extends further to the left of center than to the right of center. (And of course **right skewed** is *vice versa*.)

The data is in **clusters** if there is more than one region in which a lot of the data is concentrated.

It is usually interesting to try to understand *why* the data has the shape that it has. For example, let's think about the shape of a histogram of heights of people.

be perfectly symmetric - we are only worried about approximate symmetry here.

The data is **left skewed** if the data extends further to the left of center than to the right of center. (And of course **right skewed** is *vice versa*.)

The data is in **clusters** if there is more than one region in which a lot of the data is concentrated.

It is usually interesting to try to understand *why* the data has the shape that it has. For example, let's think about the shape of a histogram of heights of people.

Final note: will not stemplots or time plots in class, but will expect you to be able to make them when necessary.

Chapter 2. Summarizing distributions numerically.

Chapter 2. Summarizing distributions numerically.

Mean:

We assume we have a list of n values for some variable

$$x_1, \ldots, x_n.$$

The **mean** of this set of values is

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This is another possible measure of the center (besides the median, which we've already discussed).

This is another possible measure of the center (besides the median, which we've already discussed).

EXAMPLE: State populations. From our data,

$$\begin{aligned} \text{median} = M &= 4.01 \text{ million} \\ \text{mean} = \overline{x} &= 5.51 \text{ million} \end{aligned}$$

This is typical of a distribution skewed to the **right**; in that case the mean will be further to the **right** than the median.

This is another possible measure of the center (besides the median, which we've already discussed).

EXAMPLE: State populations. From our data,

$$\begin{aligned} \text{median} = M &= 4.01 \text{ million} \\ \text{mean} = \overline{x} &= 5.51 \text{ million} \end{aligned}$$

This is typical of a distribution skewed to the **right**; in that case the mean will be further to the **right** than the median.

In general the mean is much more sensitive to outliers than the median is. Because of this the median is called a **resistant measure** in that it resists distortion by outliers. Whereas the mean is *not* a resistant measure.

In general the mean is much more sensitive to outliers than the median is. Because of this the median is called a **resistant measure** in that it resists distortion by outliers. Whereas the mean is *not* a resistant measure.

Example 1. *Suppose we pick 10 people at random and find out what their annual incomes are in thousands of dollars.*

$$45, 18, 25, 29, 37, 12, 82, 130, 21, 19.$$

What are the median and mean?

In general the mean is much more sensitive to outliers than the median is. Because of this the median is called a **resistant measure** in that it resists distortion by outliers. Whereas the mean is *not* a resistant measure.

Example 1. *Suppose we pick 10 people at random and find out what their annual incomes are in thousands of dollars.*

$$45, 18, 25, 29, 37, 12, 82, 130, 21, 19.$$

What are the median and mean?

Now suppose we pick one more person at random, and this person happens to be Bill Gates, whose annual income is something like 1.9 billion dollars. Now our distribution is (still in thousands of dollars)

$$45, 18, 25, 29, 37, 12, 82, 130, 21, 19, 1.9 \times 10^6$$

What are the median and mean now?

Now suppose we pick one more person at random, and this person happens to be Bill Gates, whose annual income is something like 1.9 billion dollars. Now our distribution is (still in thousands of dollars)

$$45, 18, 25, 29, 37, 12, 82, 130, 21, 19, 1.9 \times 10^6$$

What are the median and mean now?

We see that although we shifted the median only slightly, the mean has been shifted dramatically.

Now suppose we pick one more person at random, and this person happens to be Bill Gates, whose annual income is something like 1.9 billion dollars. Now our distribution is (still in thousands of dollars)

$$45, 18, 25, 29, 37, 12, 82, 130, 21, 19, 1.9 \times 10^6$$

What are the median and mean now?

We see that although we shifted the median only slightly, the mean has been shifted dramatically. (Does this imply that the mean is always the right thing to measure?)

Now suppose we pick one more person at random, and this person happens to be Bill Gates, whose annual income is something like 1.9 billion dollars. Now our distribution is (still in thousands of dollars)

$$45, 18, 25, 29, 37, 12, 82, 130, 21, 19, 1.9 \times 10^6$$

What are the median and mean now?

We see that although we shifted the median only slightly, the mean has been shifted dramatically. (Does this imply that the mean is always the right thing to measure?)

Quartiles:

The mean and median give measures of the center of the distribution. The quartiles give partial information about the spread of the distribution.

Quartiles:

The mean and median give measures of the center of the distribution. The quartiles give partial information about the spread of the distribution.

The **first quartile**, $\mathbf{Q}_1$ is the median of the data below the overall median. In other words, $\frac{1}{4}$ or 25% of the list should be below $\mathbf{Q}_1$, and $\frac{3}{4}$ of the list should be above $\mathbf{Q}_1$.

Quartiles:

The mean and median give measures of the center of the distribution. The quartiles give partial information about the spread of the distribution.

The **first quartile**, $\mathbf{Q}_1$ is the median of the data below the overall median. In other words, $\frac{1}{4}$ or 25% of the list should be below $\mathbf{Q}_1$, and $\frac{3}{4}$ of the list should be above $\mathbf{Q}_1$.

The **second quartile** is the median.

Quartiles:

The mean and median give measures of the center of the distribution. The quartiles give partial information about the spread of the distribution.

The **first quartile**, $\mathbf{Q}_1$ is the median of the data below the overall median. In other words, $\frac{1}{4}$ or 25% of the list should be below $\mathbf{Q}_1$, and $\frac{3}{4}$ of the list should be above $\mathbf{Q}_1$.

The **second quartile** is the median.

The **third quartile**, $\mathbf{Q}_3$ is the median of the data above

the overall median. In other words, $\frac{3}{4}$ of the list should be below $\mathbf{Q}_3$ and $\frac{1}{4}$ of the list should be above $\mathbf{Q}_3$.

Five-number summary and boxplots

The **five-number summary** for a list of observations of a variable is:

Five-number summary and boxplots

The **five-number summary** for a list of observations of a variable is:

Minimum,

Five-number summary and boxplots

The **five-number summary** for a list of observations of a variable is:

$$\text{Minimum}, Q_1,$$

Five-number summary and boxplots

The **five-number summary** for a list of observations of a variable is:

$$\text{Minimum}, Q_1, \text{Median},$$

Five-number summary and boxplots

The **five-number summary** for a list of observations of a variable is:

$$\text{Minimum}, Q_1, \text{Median}, Q_3,$$

Five-number summary and boxplots

The **five-number summary** for a list of observations of a variable is:

$$\text{Minimum}, Q_1, \text{Median}, Q_3, \text{Maximum}.$$

Five-number summary and boxplots

The **five-number summary** for a list of observations of a variable is:

$$\text{Minimum}, Q_1, \text{Median}, Q_3, \text{Maximum}.$$

A **boxplot** is a graphical representation of the five number summary.

Five-number summary and boxplots

The **five-number summary** for a list of observations of a variable is:

$$\text{Minimum}, Q_1, \text{Median}, Q_3, \text{Maximum}.$$

A **boxplot** is a graphical representation of the five number summary.

Example 2. *Calculate the five-number summary for the price of a quart of milk in the 30 largest economies (in US dollars), arranged in order:*

1.20, 1.24, 1.32, 1.33, 1.35, 1.38, 1.43, 1.47, 1.48, 1.62, 1.84, 1.84, 1.86, 1.86, 1.95, 1.96, 2.01, 2.18, 2.18, 2.19, 2.21, 2.21, 2.29, 2.32, 2.34, 2.71, 2.71, 2,97, 3.14, 3.60, 4.10, 4.59

1.20, 1.24, 1.32, 1.33, 1.35, 1.38, 1.43, 1.47, 1.48, 1.62, 1.84, 1.84, 1.86, 1.86, 1.95, 1.96, 2.01, 2.18, 2.18, 2.19, 2.21, 2.21, 2.29, 2.32, 2.34, 2.71, 2.71, 2,97, 3.14, 3.60, 4.10, 4.59

Notice the skew of this distribution:

1.20, 1.24, 1.32, 1.33, 1.35, 1.38, 1.43, 1.47, 1.48, 1.62, 1.84, 1.84, 1.86, 1.86, 1.95, 1.96, 2.01, 2.18, 2.18, 2.19, 2.21, 2.21, 2.29, 2.32, 2.34, 2.71, 2.71, 2,97, 3.14, 3.60, 4.10, 4.59

Notice the skew of this distribution: The first quarter of values is spread between $1.47 - 1.20$, a difference of 27 cents

1.20, 1.24, 1.32, 1.33, 1.35, 1.38, 1.43, 1.47, 1.48, 1.62, 1.84, 1.84, 1.86, 1.86, 1.95, 1.96, 2.01, 2.18, 2.18, 2.19, 2.21, 2.21, 2.29, 2.32, 2.34, 2.71, 2.71, 2,97, 3.14, 3.60, 4.10, 4.59

Notice the skew of this distribution: The first quarter of values is spread between $1.47 - 1.20$, a difference of 27 cents
The second quarter is between $1.955 - 1.47$, a difference of 48.5 cents

1.20, 1.24, 1.32, 1.33, 1.35, 1.38, 1.43, 1.47, 1.48, 1.62, 1.84, 1.84, 1.86, 1.86, 1.95, 1.96, 2.01, 2.18, 2.18, 2.19, 2.21, 2.21, 2.29, 2.32, 2.34, 2.71, 2.71, 2,97, 3.14, 3.60, 4.10, 4.59

Notice the skew of this distribution: The first quarter of values is spread between $1.47 - 1.20$, a difference of 27 cents
The second quarter is between $1.955 - 1.47$, a difference of 48.5 cents
The third quarter is between $2.34 - 1.955$, a difference of

38.5 cents

38.5 cents

The last quarter is $4.59 - 2.34$, a difference of \$1.25

38.5 cents

The last quarter is $4.59 - 2.34$, a difference of \$1.25

Standard Deviation

The **standard deviation** is another measure of spread.

Standard Deviation

The **standard deviation** is another measure of spread.

Calculating medians and quartiles involves arranging the data in order. Calculating **standard deviation** (like calculating the mean) does not involve doing this, but involves more arithmetic.

Let

$$x_1, .., x_n$$

be a list of values of our variable.

Let

$$x_1, .., x_n$$

be a list of values of our variable. $\overline{x}$ is the mean.

Let

$$x_1, .., x_n$$

be a list of values of our variable. $\overline{x}$ is the mean. The standard deviation is given by

$$\sigma(\text{or } s) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

We will talk more about standard deviation throughout the course.

We will talk more about standard deviation throughout the course. For now just note that if it is small then one expects the quartiles to be close together,

We will talk more about standard deviation throughout the course. For now just note that if it is small then one expects the quartiles to be close together, and if it is large then the quartiles should be spread apart.

We will talk more about standard deviation throughout the course. For now just note that if it is small then one expects the quartiles to be close together, and if it is large then the quartiles should be spread apart. One of our goals will be to understand why the formula above behaves in this way,

We will talk more about standard deviation throughout the course. For now just note that if it is small then one expects the quartiles to be close together, and if it is large then the quartiles should be spread apart. One of our goals will be to understand why the formula above behaves in this way, and in general we will be forging links between intuitive notions such as "spread" and sometimes complicated formulae which we use to measure these notions accurately.

We will talk more about standard deviation throughout the course. For now just note that if it is small then one expects the quartiles to be close together, and if it is large then the quartiles should be spread apart. One of our goals will be to understand why the formula above behaves in this way, and in general we will be forging links between intuitive notions such as "spread" and sometimes complicated formulae which we use to measure these notions accurately.