Lecture 2:
Lesson and Activity Packet

MATH 232: Introduction to Statistics

September 9, 2016

Last time, we discussed:

- Measures of central tendency
  - Mean
  - Median
  - Mode
  - Midrange
- Mathematical concepts
  - Existence
  - Uniqueness
  - Counterexample

Today, we will go over:

- Measures of dispersion
  - Range
  - Variance
  - Standard deviation
- What is the difference between a population and a sample?

Individual Exercise 1

Suppose a basketball player scores 22, 26, and 24 points in her first three games. A teammate of hers has scored 41, 13, and 18 in these games. Compute the mean score of each player.

Group Exercise 2

What is the difference between the two players? Which one is more consistent? Would you prefer to have the first, or the second, as a teammate? Does the measure of central tendency *(in this case, the mean) tell the whole story about this data?*

Sometimes, the measure of central tendency is, indeed, insufficient to describe a data set. In this case, we need a way to measure the *spread* or *dispersion* of the data—that is, how far the data is, in general, from the central tendency.

Definition 1 (*Measures of Dispersion*)

A measure of dispersion *describes the spread, or dispersion, of a data set.*

There are several *measures of dispersion* at our disposal, and we will add three of them to our "toolbox" today:

- Range
- Variance
- Standard deviation

Definition 2 (*Range*)

For a data set S, the range of S is equal to $\max(S) - \min(S)$. It is exactly the spread in data.

Individual Exercise 3

What are the ranges of the scores of the basketball players in the previous exercise?

The *range* gives us some information about the dispersion, but it doesn't tell us everything.

Group Exercise 4

Compute the ranges of the following three data sets:

$$S_1 := \{5, 20, 20, 20, 20, 20, 20, 20\}$$
$$S_2 := \{5, 5, 5, 5, 20, 20, 20, 20\}$$
$$S_3 := \{5, 7, 9, 12, 15, 17, 19, 20\}.$$

However, the range is easy to compute, and industries can use it as a "quick and dirty" check on quality control, for example.

Let's take a break from measures of deviation, and ask the difference between a *sample* and a *population*.

Definition 3 (*Population*)

A population *is an entire set of all conceivably possible observations of a given phenomenon, and may be finite or infinite.*

Example 1 (*Population*)

The weights of all grains of sand on the beach is a population of infinite size (given we believe the assertion that there are infinitely many grains of sand on the beach).

If the total order that MCLA places for lightbulbs from a supplier is 10,000 units, then this set is a population of finite size.

Definition 4 (*Sample*)

On the other hand, a sample *is a set of data that consists of observations of a given phenomenon taken from only part of a group. A sample is always finite.*

Example 2 (*Sample*)

Let's say that MCLA was interested in determining the lifespan of the lightbulbs it orders. It would be supremely impractical to order 10,000 lightbulbs and then burn them all out just for the sake of the test. So we would instead test a sample of them—say, we test 20—and we infer *a result about the longevity of the population, knowing the longevity of the sample we studied.*

Group Exercise 5

If 5 lightbulbs out of the 20 we test burn out within the first 5,000 hours, then how many lightbulbs out of next year's order of 10,000 would we expect to also burn out within the first 5,000 hours?

The distinction between a population and a sample depends somewhat on the context of the experiment.

Example 3

The complete figures for a recent year, giving the rate of retention of a class of students at MCLA, can be seen as a population if we are interested only in that particular year at MCLA. But if we want to use this data to infer something about the retention rates at colleges and universities in the entire state of Massachusetts, then the MCLA data would be considered just a sample.

There is a link in the Canvas module that explains this distinction a little more deeply, with some supporting examples.

One of the biggest practical hurdles that applied statisticians need to climb is to find a way of choosing a sample that accurately represents the entire population they make inferences about.

Group Exercise 6

The worst failure of public election polling in history happened just this March, in the Democratic primary election held in the state of Michigan. On the morning of the election, all of the nation's leading pollsters and analysts, including Nate Silver[a]*, predicted that Hillary Clinton would beat Bernie Sanders by huge margins, ranging from 21 to 37 percentage points; in the actual contest that followed, Sanders would actually defeat Clinton by 1.5 points.*

It is evident that the pollsters' data samples were not representative of the population of Democratic voters in Michigan. The beginning of one possible explanation is that federal law permits political polling to be conducted only via landline telephones. Why would this skew the data in this sample in the favor of Hillary Clinton? (That is, please explain the logic a little.) Can you see any other factors that might have skewed the data?

[a]Silver was a hobby statistician made famous in 2008 by his web site `fivethirtyeight.com`, which used statistical weighting methods to accurately predict the results of 49 out of 50 states in the general presidential election, and whose forecasts for **all** senate races that year were also accurate. He became interested in statistics because of a childhood love for baseball; it turned out to be very profitable for him when the New York Times subsequently hired him as a columnist, and later when ESPN purchased rights to his blog. He now has a team of employees working permanently at `fivethirtyeight.com`, which promises to be one of the enduring major predictors of political events in our country.

Toward the end of October, we will discuss several methods of collecting sample data from populations so that the samples are fairly representative of the entire group. But, as you can see, there will always be room for improvement of these methods. This is why statisticians and data analysts are in such high demand in industry. (Seriously: search on `indeed.com` for 'applied mathematician' or 'data analyst' and see how many jobs come up.)

When we are talking about *samples*, we tend to use Roman letters to represent the mean, variance, and standard deviation; with *populations*, though, we use Greek letters.

Example 4

We learned last time that the mean of a sample set $\{x_1, \ldots, x_n\}$ is denoted $\bar{x}$. If the set is instead interpreted as a finite population, then the mean is computed in exactly the same way, except it is denoted μ.

The median, mode, and midpoint values for populations do not have standardized notation in Greek letters.

While the mean is computed the same way for populations and samples, **the standard deviation is computed differently for populations and for samples**, as we will shortly see.

Definition 5 (*Deviation of a Data Point from the Mean*)

If x_i is a data point in the set $S := \{x_1, \ldots, x_n\}$, then the deviation of x_i from the mean of S *is the quantity $x_i - \bar{x}$ in case S is a sample, and is $x_i - \mu$ in case S is a population.*

Group Exercise 7

What does this definition mean in "real words"? Would you expect the deviations of data points to be positive, negative, or both?

Group Exercise 8

Show mathematically that for any data set, the sum of all deviations from the mean must be zero. [Hint: Start by writing out the sum of all deviations, and then substitute the sum that defines $\bar{x}$. If the proof is too difficult at first, then try seeing how it works with a small data set (with three points, for example). Then go back and try to revisit the general theory. Since you need to become comfortable with sigma notation for finite sums, please take this as an opportunity to use it, and check for agreement with your groupmates.]

Definition 6 (*Sample Variance*)

For data ***in a sample****, the* sample variance *is*

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

Definition 7 (*Population Variance*)

For data ***in a population****, the* population variance *is*

$$\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}.$$

The reason these are computed differently is beyond the scope of this lesson, but in general is a correction for the bias of the sample variance as an estimator of the population variance. Don't worry about this if it is confusing (we don't know what a "bias" is, yet!).

Definition 8 (*Sample Standard Deviation*)

For data ***in a sample****, the* sample standard deviation *is*

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

Definition 9 (*Population Standard Deviation*)

For data ***in a population****, the* population standard deviation *is*

$$\sigma := \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}.$$

The variance is typically not used in data analysis, except as an intermediate calculation on the way to the standard deviation. The fact that the units of variance are different than the units of the original measurement limits their practical meaning (for example, if our data set consisted of points measured in units of dollars, or in units of years, the variances would be in units of square dollars or square years (?!)).

Individual Exercise 9

Suppose that we have the sample set $\{1, 3, 14\}$. *Find the variance and the standard deviation.*

Group Exercise 10

Explain the standard deviation in words.

Group Exercise 11

Can the standard deviation be negative? Under which circumstances (if any) is it exactly zero?

A general rule of thumb: For most data sets, the vast majority of data points lie within 2 standard deviations of the mean.

Example 5

The Wechsler Adult Intelligence Scale involves an IQ test designed so that the mean score is 100 and the standard deviation is 15[a]*. Use the general rule of thumb to find the minimum and maximum "usual" IQ scores. Would an IQ of 135 be considered "unusual"?*

According to the rule of thumb, the minimum "usual" data point will be two standard deviations below the mean, and the maximum "usual" point will be two standard deviations above the mean. That is,

$$\begin{aligned} \textit{minimum "usual"} &= \mu - 2\sigma \\ &= 100 - 2 \cdot 15 \\ &= 70, \end{aligned}$$

and

$$\begin{aligned} \textit{maximum "usual"} &= \mu + 2\sigma \\ &= 100 + 2 \cdot 15 \\ &= 130. \end{aligned}$$

An IQ score of 135 would therefore indeed be considered unusual.

[a] This test is shown to have both racial and socioeconomic bias. Don't put too much faith in its results in general.

Recap

- Measures of Dispersion
  - Range
    - Not terribly enlightening, but easy to compute
  - Variance
    - Useful as an intermediate step to compute standard deviation
    - Is zero only when all data points are identical
  - Standard Deviation
    - Tells how far the data points generally are from their mean
    - For many data sets, the vast majority of points fall within two standard deviations of the mean—a quick way of identifying "unusual" data
- Difference between population and sample

Homework

- Please find the syllabus module on Canvas, and complete the quiz there. This quiz will not contribute to your grade, but it is required before you complete Homework 1.
- Homework 1 has been posted to Canvas, and will be due September 12.