

Math 107: An Introduction to Statistics

Dr. Richard Mikula

January 2008

What is Statistics?:

Statistics deals with collecting informative data, interpreting these data, and drawing conclusions about the subject being studied, from the collected data.

The principles and methodology of the subject of statistics are useful when considering the following questions:

- What kind of data do we need to collect? (experiment design)
- How much data do we need to collect? (experiment design)
- Once collected, how do we organize and interpret these data? (descriptive statistics)

- How do we analyze these data, and in particular, how do we draw general conclusions from the collected data? (inferential statistics)
- How do we assess the strength of the conclusions we draw and how do we measure uncertainty? (inferential statistics)

Statisticians have three major roles:

- Make predictions about a population under study. Where here a **population** can actually be a population of people, things, etc.
- State how large the error margin for the prediction made.
- Design the experiment which collects the data to make these predictions.

The subject of statistics is the science of experiment design, combined with the art of drawing the most reasonable conclusions possible about a population on a whole, from a sample of measurements taken from the entire population. Generally, it is infeasible, or even impossible to study the entire population.

Statisticians thus collect data from a population, and then must organize the data and interpret the data. In doing this, they must also determine a measure of the reliability of the predictions that they make.

Some terms we shall use:

- A **unit** is a single entity, usually a person or an object, whose characteristics are of interest to us.
- The **population of units** is the complete collection of units about which information is sought.
- The **statistical population** is the set of all the potential measurements from the population of units if we were able to take measurements from everyone in the group.
- A **sample** is the set of measurements from a small collection of units in the population.

The statistician's most important job is the design of the experiment, which will gather the data to make predictions about the population of interest. Unfortunately, often the population (of units) is too large to collect data for every unit, and the statistician collects data from a sample of units. A danger in collecting sample data is that personal or regional biases can creep into the testing

We distinguish between **Anecdotal Evidence** and **Empirical Evidence**. Anecdotal evidence is data or conclusions that are not drawn scientifically and empirical evidence is evidence which is collected in a careful scientific manner.

Chapter 2: Understanding Data from Samples

Assuming that an experiment was designed and then implemented to collect data from a sample taken from a population, the next step a statistician must address is organizing the data in order to then use the data to make predictions about the population under study.

At this point, we shall discuss

- Different types of data (e.g. Qualitative and Numerical data)
- Ways to summarize the data set (e.g. graphs, tables)
- Measurements of the center and the dispersion of the data set (e.g. mean, standard deviation, etc.)

Data sets are divided into two primary types

- Qualitative or Categorical Data: Data that is not measured numerically
- Numerical Data

Some examples of **qualitative data** are: Gender, eye color, political party affiliation.

Some examples of **numerical data** are: Annual income, weight, G.P.A.

When we are looking at numerical data, we call the particular characteristic we are measuring a **random variable**. Thus, a random variable is a value that is associated to each member of the population of units.

For example, if we were studying students at Lock Haven University (here the population of units is the set of LHU students), then we may be interested in the students' G.P.A.'s. Thus, the random variable is the G.P.A. of a student at LHU.

For the same population of units, the students at LHU, another possible random variable is number of semesters the student at LHU has completed thus far.

Again, for the same population of units (students at LHU) we may define another random variable, the student's height.

Essentially, we classify random variables as either **discrete random variables** or as **continuous random variables**. In our first 2 examples above, the random variables (namely the students G.P.A. and the number of semesters completed at LHU) are examples of discrete random variables.*

In the third example (the student's height) is a continuous random variable.

*In the text, Dr. Morgan comments that a G.P.A. can be modeled by a continuous random variable. Although the set of possible G.P.A. values is discrete (there are 401 possible G.P.A. values, namely any value in the set $\{0.00, 0.01, 0.02, \dots, 3.99, 4.00\}$), when you plot all possible G.P.A. values on the number line, notice that the gaps between any two possible consecutive values are so small, and thus a continuous approximation is warranted.

The way we distinguish between a discrete random variable and a continuous random variable is by the possible values the random variable may take on. Essentially, if the value of the random variable can be any value possible value in an interval (or union of intervals), we say it is continuous, otherwise we shall consider it a discrete random variable.

Describing Data:

We shall consider first an example involving qualitative data:

The data in the following table represents the educational attainment of residents of the United States 25 years or older in 2003, based on data obtained from the U.S. Census Bureau. The data are in thousands.

Educational Attainment	Frequency
Less than 9th grade	12,276
9th-12th grade, but no H.S. diploma	16,323
High School diploma	59,292
Some College, but no degree	31,762
Associate's degree	15,147
Bachelor's degree	33,213
Graduate or Professional degree	17,169
Total	185,182

We may refer to the above table as a **frequency table**. The following table is a **relative frequency table**. We obtain this from the previous table by taking the numbers in the right column (the frequencies of occurrence) and divide them by the total number of observations, or units in our sample.

Educational Attainment	Rel. Freq.
Less than 9th grade	0.0663
9th-12th grade, but no H.S. diploma	0.0881
High School diploma	0.3202
Some College, but no degree	0.1715
Associate's degree	0.0818
Bachelor's degree	0.1794
Graduate or Professional degree	0.0927

Observe that in the previous table the numbers in the right column add up to one. In general this should be true*.

*Or at least add up to a number very close to one, which may differ from one due to rounding.

The next example which we shall consider involves numerical data. In particular, it represents a discrete data set as well.

The following data represents the number of pot holes on 24 random 1 mile stretches of Interstate 80 in Pennsylvania:

1	0	1	2	1	3
1	1	0	1	0	0
2	1	3	8	2	4
1	3	2	3	1	1

Some questions we may consider about this data set are:

- What is the most frequent number in this table?
- How many pot holes does there seem to be along a typical 1 mile stretch of highway?

To answer these questions, or any other question, we should first organize our data in some way. Below is a frequency and relative frequency table for our data set:

Number of Pot Holes	Frequency	Rel. Freq
0	4	0.167
1	10	0.417
2	4	0.167
3	4	0.167
4	1	0.042
8	1	0.042

Observe here that the most frequent number of pot holes observed is 1. To answer the second question, we need to discuss various notions of *center of a data set*. One such notion is what is called the **mean** or **average** of the numerical data set. We obtain this by adding all the values in our data set, and then dividing that sum by the total number of items in our data set. Thus, in this example the mean is

$$\begin{aligned} & \frac{4(0) + 10(1) + 4(2) + 4(3) + 1(4) + 1(8)}{24} \\ &= \frac{42}{24} \\ &= 1.75. \end{aligned}$$

We generalize this as follows:

Given a sample of numerical values

$$a_1, a_2, a_3, \dots, a_{n-1}, a_n,$$

the sample **mean** is the number \bar{x} defined by

$$\bar{x} = \frac{a_1 + a_2 + a_3 + \dots + a_{n-1} + a_n}{n}.$$

To simplify this formula, we introduce what is commonly referred to as **sigma notation** for sums. In this notation, the mean is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n a_i.$$

Another measure of the center of a data set is the so-called **median** of a data set. If we arrange the data set in increasing order, this is the value in the middle of our list if there are an odd number of items in our list, and it is the average of the middle two numbers if there are an even number of items in our list.

In the above example, the median is

$$\begin{aligned} &= \frac{1 + 1}{2} \\ &= 1. \end{aligned}$$

In the lecture, we will examine what is referred to as a **density histogram**, which is a graph that organizes the discrete data. We say that this graph is **skewed to the right** if the mean is larger than the median. We say that this graph is **skewed to the left** if the mean is less than the median. In our pot hole example, the density histogram is skewed to the right.

In the above (pot hole) example, the **range** of the data set is 8. The **range** is the largest value in the data set minus the smallest value in the data set. In this case, the largest value of the data set is 8, and the smallest value is 0.

Notice that the values in the data set are integers, in this case either 0,1,2,3,4, or 8. We consider 8 an **outlier** for the data set. This is because it only appears once, and it is "far away" from all the other values. If we were to simply throw away this value, then the relative frequency table becomes:

Number of Pot Holes	Frequency	Rel. Freq
0	4	0.174
1	10	0.435
2	4	0.174
3	4	0.175
4	1	0.043

The mean for this new data set is

$$\begin{aligned}\bar{x} &= \frac{4(0) + 10(1) + 4(2) + 4(3) + 1(4)}{23} \\ &= \frac{34}{23} \\ &\approx 1.478.\end{aligned}$$

The median for this new data set is 1. Notice that discarding outliers results in a new data set whose median and mean are closer together than they were in the original data set.

Example: Suppose that in a class of 25 students we have the following exam scores for the final exam.

15, 25, 55, 56, 56, 60, 61

62, 62, 63, 65, 70, 71

72, 73, 74, 74, 76, 78

80, 81, 86, 92, 95, 100

Here we first examine a histogram for these exam scores. When we do this, we will not plot all the scores individually, but instead we first group the scores in the following table

Range	Freq.	Rel. Freq.
0-9	0	0
10-19	1	0.04
20-29	1	0.04
30-39	0	0
40-49	0	0
50-59	3	0.12
60-69	6	0.24
70-79	8	0.32
80-89	3	0.12
90-100	3	0.12

Next, we compute the average exam score

$$\begin{aligned}\bar{x} &= \{15 + 25 + 55 + 56 + 56 + 60 \\ &\quad + 61 + 62 + 62 + 63 + 65 \\ &\quad + 70 + 71 + 72 + 73 + 74 + 74 + 76 \\ &\quad + 78 + 80 + 81 + 86 + 92 + 95 + 100\}/25 \\ &= \frac{1702}{25} \\ &= 68.08.\end{aligned}$$

Also, the median exam score is 71.

Thus, we see that since the mean is less than the median, and thus the histogram is skewed to the left. If we consider the scores 15 and 25 outliers, then the median for the new data set (the 23 remaining exam scores) is 72 and the mean becomes

$$\bar{x} = \frac{1662}{23}$$
$$\approx 72.261.$$

Notice that the median as a measure of the center of a data set is less sensitive to outliers than the mean is. Thus, when we throw away outliers, the mean of the data set without the outliers seems to be closer to the median of the data set without the outliers.

At this point, we will discuss **measures of dispersion** in a data set or **measures of variability** in the data set. One measure of dispersion of a data set is the so-called **range** of a data set. This is simply the largest value in our data set minus the smallest value in the data set. In the last example, the range is

$$\begin{aligned} \text{range} &= 100 - 15 \\ &= 85. \end{aligned}$$

When we throw out the outliers

15, 25

the range becomes

$$\text{range} = 100 - 55 = 45.$$

Another very important measure of dispersion or variability of a data set is the so called **standard deviation** of the data set.

For a data set

$$a_1, a_2, a_3, \dots, a_n$$

we define the **standard deviation** s of the data set by the formula

$$s = \sqrt{\frac{(a_1 - \bar{x})^2 + (a_2 - \bar{x})^2 + \dots + (a_n - \bar{x})^2}{n - 1}}$$
$$= \sqrt{\frac{1}{n - 1} \sum_{k=1}^n (a_k - \bar{x})^2}.$$

The standard deviation s of a data set

$$a_1, a_2, \dots, a_n$$

which has a mean \bar{x} is a measure of the average distance of a point in the data set to the mean. The standard deviation is zero if and only if the data set's values are all \bar{x} . In general, the standard deviation $s \geq 0$ for any data set.

To compute the standard deviation for data set in our example above (the exam scores example), we need to do some computations, which we shall do below:

a_i	$(a_i - \bar{x})^2$
15	$(15 - 68.08)^2 = 2817.4864$
25	$(25 - 68.08)^2 = 1855.8864$
55	$(55 - 68.08)^2 = 171.0864$
56	$(56 - 68.08)^2 = 145.9264$
56	$(56 - 68.08)^2 = 145.9264$
60	$(60 - 68.08)^2 = 65.2864$
61	$(61 - 68.08)^2 = 50.1264$
62	$(62 - 68.08)^2 = 36.9664$
62	$(62 - 68.08)^2 = 36.9664$
63	$(63 - 68.08)^2 = 25.8064$
65	$(65 - 68.08)^2 = 9.4864$
70	$(70 - 68.08)^2 = 3.6864$
71	$(71 - 68.08)^2 = 8.5264$
72	$(72 - 68.08)^2 = 15.3664$
73	$(73 - 68.08)^2 = 24.2064$
74	$(74 - 68.08)^2 = 35.0464$
74	$(74 - 68.08)^2 = 35.0464$
76	$(76 - 68.08)^2 = 62.7264$
78	$(78 - 68.08)^2 = 98.4064$
80	$(80 - 68.08)^2 = 142.0864$
81	$(81 - 68.08)^2 = 166.9264$
86	$(86 - 68.08)^2 = 321.1264$
92	$(92 - 68.08)^2 = 572.1664$
95	$(95 - 68.08)^2 = 724.6864$
100	$(100 - 68.08)^2 = 1018.8864$

Thus, the sum of the right column in this table is

$$\sum_{i=1}^{25} (a_i - \bar{x})^2 = 8589.84.$$

Thus, the standard deviation is

$$\begin{aligned} s &= \sqrt{\frac{1}{24} \sum_{i=1}^{25} (a_i - \bar{x})^2} \\ &\approx \sqrt{\frac{8589.84}{24}} \\ &= \sqrt{357.91} \\ &\approx 18.92 \end{aligned}$$

Recall the pothole example. In this example, the average or mean of the data set is

$$\bar{x} = 1.75.$$

The table for this data set given below is useful when computing the standard deviation s for this data set.

a_i	$(a_i - \bar{x})^2$
0	$(0 - 1.75)^2 = 3.0625$
0	$(0 - 1.75)^2 = 3.0625$
0	$(0 - 1.75)^2 = 3.0625$
0	$(0 - 1.75)^2 = 3.0625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
1	$(1 - 1.75)^2 = 0.5625$
2	$(2 - 1.75)^2 = 0.0625$
2	$(2 - 1.75)^2 = 0.0625$
2	$(2 - 1.75)^2 = 0.0625$
2	$(2 - 1.75)^2 = 0.0625$
3	$(3 - 1.75)^2 = 1.5625$
3	$(3 - 1.75)^2 = 1.5625$
3	$(3 - 1.75)^2 = 1.5625$
3	$(3 - 1.75)^2 = 1.5625$
4	$(4 - 1.75)^2 = 5.0625$
8	$(8 - 1.75)^2 = 39.0625$

Thus summing the rightmost column yields

$$\sum_{i=1}^{24} (a_i - \bar{x})^2 = 68.5.$$

Thus, the standard deviation of this data set is

$$\begin{aligned} s &= \sqrt{\frac{68.5}{23}} \\ &\approx \sqrt{2.9783} \\ &= 1.7258. \end{aligned}$$

Next, we examine the effect of outliers on the standard deviation. Recall that for the pothole example, we considered 8 an outlier. Recall that the mean for the data set without 8 is

$$\bar{x} \approx 1.478.$$

To compute the standard deviation of the data set without 8, we need to modify the previous table, to get:

a_i	$(a_i - \bar{x})^2$
0	$(0 - 1.478)^2 = 2.1845$
0	$(0 - 1.478)^2 = 2.1845$
0	$(0 - 1.478)^2 = 2.1845$
0	$(0 - 1.478)^2 = 2.1845$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
1	$(1 - 1.478)^2 = 0.2285$
2	$(2 - 1.478)^2 = 0.2725$
2	$(2 - 1.478)^2 = 0.2725$
2	$(2 - 1.478)^2 = 0.2725$
2	$(2 - 1.478)^2 = 0.2725$
3	$(3 - 1.478)^2 = 2.3165$
3	$(3 - 1.478)^2 = 2.3165$
3	$(3 - 1.478)^2 = 2.3165$
3	$(3 - 1.478)^2 = 2.3165$
4	$(4 - 1.478)^2 = 6.3605$

Thus, for this data set, the sum of the right-most column is

$$\sum_{i=1}^{23} (a_i - \bar{x})^2 \approx 27.7395^*$$

Thus, the standard deviation is

$$\begin{aligned} s &\approx \sqrt{\frac{27.7395}{22}} \\ &\approx \sqrt{1.2609} \\ &\approx 1.1229. \end{aligned}$$

Thus, we observe that the standard deviation gets smaller when we get rid of outliers.

*We note that the approximate sign \approx was used here because at an earlier stage we approximated \bar{x} by 1.478

More on the Standard Deviation, and Other Measures of Dispersion:

As we saw above, it is somewhat difficult to evaluate s . Moreover, if we throw away outliers, we must start over. It seems as if there is no easy way to use the old standard deviation, or the calculations involved in computing it, to easily compute the new standard deviation for the data set without the outliers. However, the following formula also gives us s , the standard deviation for a data set $a_1, a_2, a_3, \dots, a_n$:

$$s = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n a_i^2 - n\bar{x}^2 \right)}$$

The advantage of this formula over the other formula is in the evaluation of s if we modify our data set.

Consider the data set,

$$-5, 1, 1, 1, 2, 2, 3, 9.$$

This data set has an average

$$\bar{x} = \frac{7}{4} = 1.75.$$

Moreover,

$a_i:$	-5	1	1	1	2	2	3	9
$a_i^2:$	25	1	1	1	4	4	9	81

Summing the second row, we get

$$\sum_{i=1}^8 a_i^2 = 126.$$

Thus, the standard deviation is given by

$$\begin{aligned}
s &= \sqrt{\frac{1}{7}(126 - 8 \cdot 1.75^2)} \\
&= \sqrt{\frac{1}{7}(126 - 8 \cdot 3.0625)} \\
&= \sqrt{\frac{1}{7}(126 - 24.5)} \\
&= \sqrt{\frac{101.5}{7}} \\
&= \sqrt{14.5} \\
&\approx 3.8079.
\end{aligned}$$

Note that if we dispose of the outliers -5, 9 we get:

a_i :	1	1	1	2	2	3
a_i^2 :	1	1	1	4	4	9

The sum of the second row here is 20; thus the standard deviation for the data set without the outliers is given by (where here $\bar{x} = \frac{5}{3} \approx 1.6667$):

$$s \approx \sqrt{\frac{1}{5}(20 - 6 \cdot 1.6667^2)}$$

$$\approx \sqrt{\frac{1}{5}(20 - 6 \cdot 2.7779)}$$

$$= \sqrt{\frac{1}{5}(20 - 16.6674)}$$

$$= \sqrt{\frac{3.3326}{5}}$$

$$= \sqrt{0.6665}$$

$$\approx 0.8164.$$

Another measure of dispersion for a data set is the so-called **variance**. The variance is simply the standard deviation squares s^2 .

Percentiles and Quartiles:

So far we have discussed the **mean** and **median** which are measures of the **center of a data set**. We have also discussed **range**, **standard deviation** and **variance**, which are measures of **dispersion** or **variability in a data set**. The center of a data set tells you the middle or average value. The measures of variability or dispersion measure how spread out the values in the data set are. However, these measures of variability we have discussed so far do not tell us about clumps in the data set, or concentrations in the data set. Thus, it is clear that we are going to need other measures of the dispersion of the data set which will help us see if there are places where the data values are grouped together.

The **100 p -th percentile** of a data set is the number q so that at least $100p$ percent of the data is less than or equal to q and at least $100(1 - p)$ percent of the data is greater than or equal to q .

The **first quartile**, usually denoted Q_1 , is the 25th percentile, and the **third quartile**, usually denoted by Q_3 , is the 75th percentile.

Consider the data set of 25 test scores

15, 25, 55, 56, 56, 60, 61

62, 62, 63, 65, 70, 71

72, 73, 74, 74, 76, 78

80, 81, 86, 92, 95, 100.

Recall that the mean and median for this data set are 68.08 and 71 respectively; and the standard deviation and range are 18.92 and 85 respectively.

We will now compute the first and third quartiles of the data set.

The first quartile is the $25 = 100(0.25)$ th percentile ($p = 0.25$). In the above data set there are $n = 25$ numbers. We compute

$$np = 25(0.25) = 6.25.$$

Round this value $np = 6.25$ up to the number 7, which we call k . Then the 7th value in our ordered list of test scores is 61, and this is Q_1 , the 25th percentile.

The third quartile is the $75 = 100(0.75)$ th percentile ($p = 0.75$). Once again, $n = 25$. We compute $np = 25(0.75) = 18.75$, and round this number up to $k = 19$. Then the 19th value in the data set, namely 78, is the 75th percentile, or Q_3 .

Note that the median 71 is the $100(0.5) = 50$ th percentile.

In general, the procedure for computing the 100 p th percentile is as follows:

NOTE: Suppose a data set has n measurements, and the data has been ordered from smallest to largest.

- Calculate np .
- If np is not an integer, round up to the nearest integer k . The k th value is the 100 p th percentile.
- If np is an integer, which we shall also call k , then compute the average of the k th and the $(k + 1)$ th values in the data set. This average is the 100 p th percentile.