4. LECTURE 4

Objectives

- ☐ I can determine the precision of a measurement and reflect that precision in my work.
- ☐ I understand that not all data is ideal for determining a relationship. When possible, I can find better data.
- ☐ By visual inspection, I can isolate an outlier and justify its removal.
- ☐ I understand that every time I use data to determine a relationship, I am making a set of assumptions.

4.1. **Precision.** In the last section, we looked at the frequency of chirps per second versus temperature. Both the temperature and the chirp frequency had no more than three digits; however, when we calculated our best fit line, the calculator displayed numbers containing many digits.

Do we need that many digits? No, we don't need them. More importantly, they communicate a level of precision that we do not have based on out data.

Precision reflects the extent to which our data is accurate. In the case of temperature, we were given data that was three digits, like $17.2°$ C. This means that the equipment used could only measure the accuracy of the temperature up until the tenths place. We say that this measurement has 3 significant figures. If our precision allows for three significant figures but the final digit is zero, we keep that digit there to reflect the precision. For example, we could have the measurement $20.0°$. As it is written, it still has 3 significant figures.

When we find a function describing how two measurements are related, we will choose coefficients to reflect that precision. Therefore, we write

$$y = 3.41x + 22.8$$

to keep our three significant figures.

If we have two measurements with a different number of significant figures, we will always choose the *smaller* of the two.

Check your Understanding

The figures $\$\ 21,000.10$ and $45°$F have how many significant figures? If there was a function related these two measurements, how many digits should our coefficients have?

4.2. **Improving Your Data.** Not all data is created equal. In much of your future work, you will likely have access to *lots* of data, but much of it will be useless to you. We therefore need to spend some time discussing what constitutes good data. All the data presented here can be found at `https://migbirdapps.fws.gov`, a website managed by the US Fish & Wildlife Service.

Let's suppose you want to know if the frequency of mourning dove calls is affected by temperature. Why would you want to know this? Much of population estimation depends on indirect measurements like samples of animals seen or heard. If temperature impacts how much an animal is heard, we may overestimate or underestimate their population based on our observations. So understand how mourning dove calls are influenced by temperature can help us better predict their populations from sound.

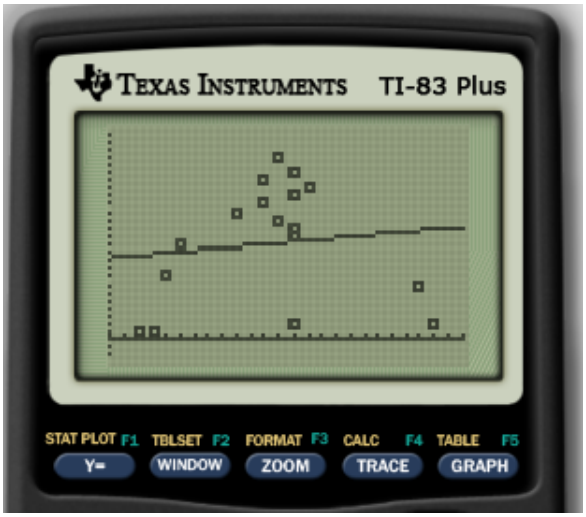
For a small project like this, you'll need to work off the data available. First, you decide to look at the temperature versus bird calls heard throughout Pennsylvania on selected days in May 2010. Below is the data.

Table 4.1:

Temperature (F)	Mourning Dove Call-Count
60	14
60	18
50	1
52	12
58	17
59	23
70	2
60	2
51	8
61	19
60	13
59	15
69	7
58	20
49	1
56	16
60	21

FIGURE 1. This data is taken from the US Fish & Wildlife Service, Division of Migratory Bird Management. The samples listed were taken throughout PA in May 2010.

When you graph it, you see no obvious relationship, especially when compared with the best fit line.



Does that mean no relationship exists in real life? Not necessarily. Although the data suggests there is no relationship, we need to consider whether this data is appropriate for the question we are asking.

We want to know whether temperature impacts mourning dove vocalizations. The data is a collection of observed mourning dove calls in various regions throughout Pennsylvania. These observations were taken in each area on one day in May 2010. What potential problems exist with using data like this?

(1) Different regions will likely have different population sizes of mourning doves.
(2) These differences in population may have a stronger impact on dove vocalizations than temperature, which would hide the relationship in the data.
(3) If we only look at measurements in the same month, we may not observe dramatic changes in temperature. That makes it harder to detect the relationship.

Ultimately, the data we used was not ideal for the problem. We'd like to use data more appropriate to our question. Below is data also taken from the US Fish and Wildlife Service. All the observations were taken in one park region in Chester County, PA. The measurements were taken on one day in May from 2001 to 2010.

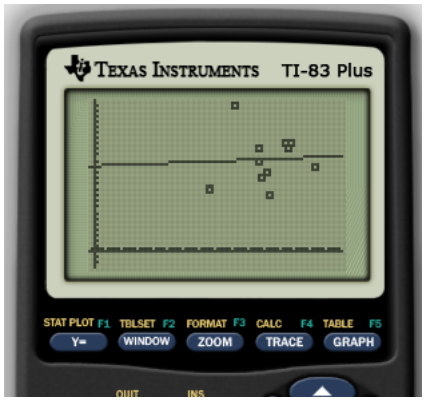
This data lacks some of the problems of the previous one. We are restricting our observation to the same park land and the measurements are taken across many years so we can get a variety of temperatures. The data still is not ideal. We do not know if the population of mourning doves remained constant during those years in that park. We also do not have many samples. The smaller your sample size, the more likely you are to see a pattern that is not there.

Still, we can do some analysis on the data we have to determine if there is a relationship. Below is a print out of the data along with a best fit line.

Table 4.2:

Temperature (F)	Mourning Dove Call-Count
54	10
59	19
52	13
68	15
51	18
36	11
51	16
43	26
53	14
61	19
60	18

FIGURE 2. This data is taken from the US Fish & Wildlife Service, Division of Migratory Bird Management. The samples listed were taken from Chester County, PA in May from 2001 to 2010.



Upon visual inspection, the data appears to have a nice linear relationship. Unfortunately, there appears to be one point, $(43, 26)$, which is very far from the rest of the data. Furthermore, the best fit line does not match nicely with the data we observe. In the image below, the point $(43, 26)$ is circled and a red line is drawn to highlight the trend of the rest of the data.

Points that are considered too far from the data set are called **outliers**. There are a variety of possible reasons for an outlier to occur:

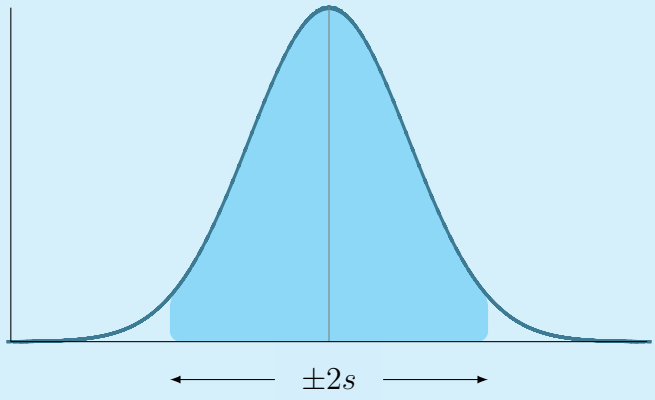
(1) **Human Error**: Those doing the measuring—in this case, counting the bird calls heard—can make mistakes. It is possible that the count of 26 bird calls was inaccurate.
(2) **Equipment Error**: Sometimes equipment is faulty. In data where each sample could be taken using different equipment (like measurements that consist of one sample per year), it is possible that a mismeasurement resulted from a technical issue.
(3) **A Fluke**: The measurement could be accurate but something unusual could be happening in the physical situation that could skew the data, like having someone nearby feeding the doves.
(4) **Nature**: Finally, it could be an accurate measurement with no weird circumstances. It is possible that the outlier really reflects what can happen in certain instances.

Because of that last reason, we need to be very careful when we consider removing a data point. If we can't definitively point to circumstances that call the measurement into question, then we run the risk of coming to inaccurate conclusions. That is, if we don't know the data point is bad, we should be careful about removing it and possibly seeing a pattern that is not there.

In this case, we did not collect the data nor do we know the circumstances under which it was collect. So if we remove the outlier, we run some risk. To reduce our risk, we will define a mathematical procedure to determine if an outlier should be removed.

Connecting Back to Past Content

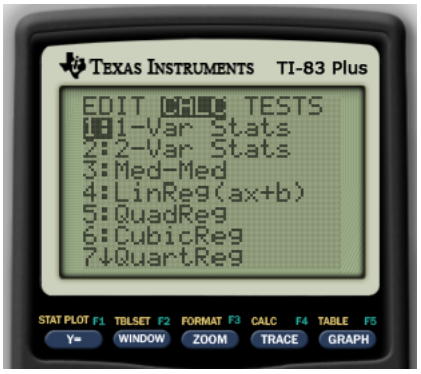
If you have taken statistics, then you understand that all measurements are just samples of a larger distribution. We will use this underlying fact to try to reduce our risk of throwing away good data to 5%.
The following procedure will require we find the mean, μ, and the sample standard deviation, s, of the dependent variable. Then we consider any values of the dependent variable smaller than $\mu - 2s$ or larger than $\mu + 2s$ to be outliers. If most of the data is reliable, the points removed are less than 5% likely to happen, meaning these points are more likely to come from an error.



Note that if most of the data is bad, this technique will not work.

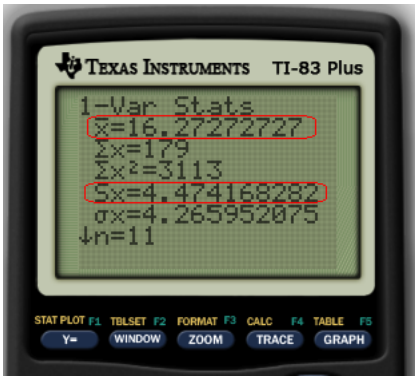
4.2.1. *Determining Outliers.* To determine outliers, we will calculate the average (or **mean**) of the dependent variable's measurements and the **sample standard deviation** (something that measures the width of the data's spread).

To calculate the mean (μ) and the sample standard deviation (s) using a TI-83 or TI-84, click STAT and select the CALC menu at the top. Then select the first entry 1:1-Var Stats by pressing ENTER.



You will now see the phrase 1-Var Stats on the home screen of the calculator. Before hitting enter, type the list of the *dependent* variable after. Here, our dependent variable is list L_2. Then press ENTER.

You'll see a display with a set of statistics. The ones we need are the mean, denoted as $\bar{x}$, and the sample standard deviation, denoted Sx. Although the calculator uses different symbols, we will refer to the mean as μ and the sample standard deviation as s.

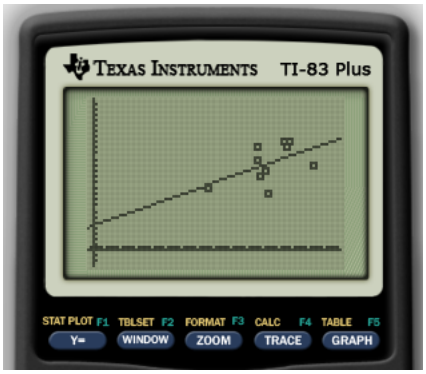


The print out above is for the data regarding dove calls in Chester County, PA. Based on the data, we will consider a point an outlier if the dependent variable is

- *smaller* than $\mu - 2s = 16 - (2 \times 4.5) = 7$, or
- *larger* than $\mu + 2s = 16 + (2 \times 4.5) = 23$.

The point we suspected to be an outlier is (43, 26). Notice that the dependent variable's value for this point is 26, which is bigger than 23! Therefore, we can remove this point with small risk.

Once we remove this point, we get a line that appears to better explain the relationship between the two measurements.



4.3. **Assumptions.** When making conclusions from raw data, it is important that we can identify our underlying assumptions. This is a list of what we need to be true in order for our findings to be meaningful. The precise statements will change from situation to situation, so here we will list the assumption for the problem above (dove calls versus temperature).

Assumptions

(1) The data is reliable; it was measured by attentive researchers with working equipment.
(2) The data was collected in roughly the same location every year.
(3) The population of mourning doves did not change significantly from year to year.
(4) The data set is large enough to define a relationship well.
(5) The data follows a "normal distribution" (a statistical assumption that means our method of selecting outliers is a good one).

Assumptions 1, 4 and 5 are needed in every problem involving raw data.

Summary of Ideas: Lecture 4

- Calculations should represent the precision of the measurements taken by keeping the same number of significant figures.
- *How* data is collected matters.
- Data can contain extreme values called **outliers**. Outliers can come about from an error or from nature.
- Any value that is **larger** than the $\mu + 2 \times s$ or **smaller** than $\mu - 2s$ is considered to be an **outlier**. They are often removed to perform calculations better.
- Every time we draw information from raw data, we are employing assumptions including that the data can be relied upon and that the data set is sufficiently large.