Introduction to Data Analysis

Learning Goals

- Understand how to display data in your lab report
- Study how to analyze your fits and parameters
- Understand how to arrive at conclusions from data

Introduction

In each lab, you spend an abundance of time collecting data. You then take your time to come up with a good way of displaying that data, whether it be through a table of graph. However, this alone is not enough to arrive at physical conclusions. We have to know how to *interpret* that data and convey those ideas to the reader. This process is known as *data analysis*.

In this document, we will discuss how you should analyze your data. We will provide a working example and then walk through the thought-process of analyzing the data before committing the findings to paper.

Setting the Scene

As our guiding example for this document, let's consider the following scenario. I've gathered data on the number of ice cream sales versus the daily high temperature. If we wanted to find a relationship between these two variables, it stands to reason that we would expect these variables to share a linear relationship. If I refer to the amount of sales in USD as S and the temperature in Fahrenheit as T, then it would make sense that they obey a linear relationship,

$$S = m * T + b, \tag{1}$$

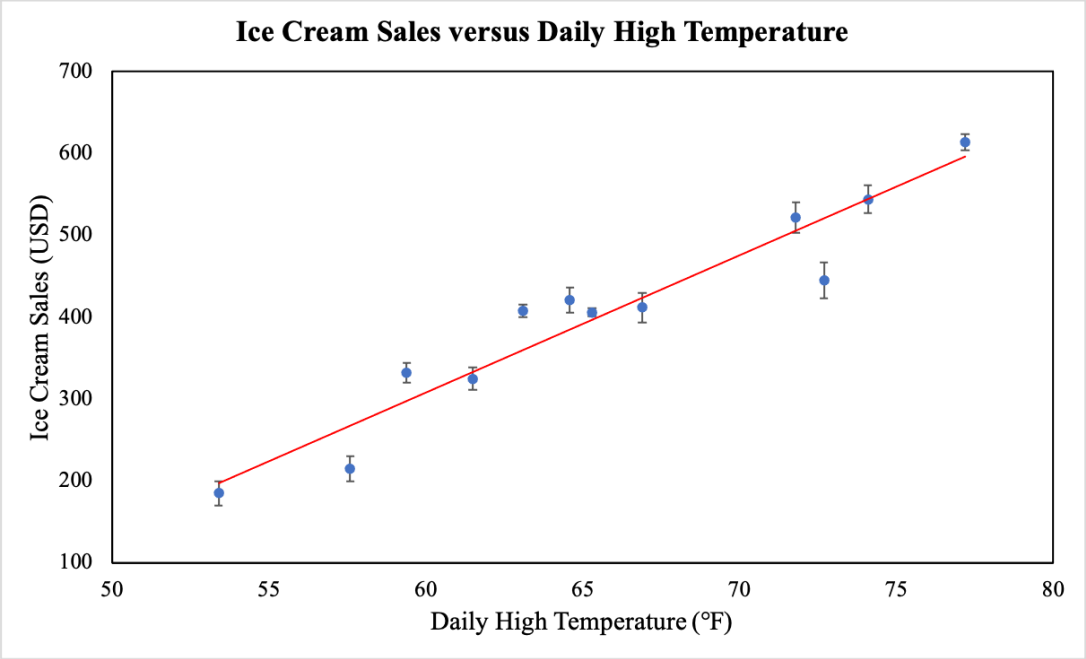

Figure 1: Ice Cream Sales versus Daily High Temperature. This graphs the total number of sales over the span of a week with an average daily high temperature for that week. Both the data and error are shown in blue as points on the figure, while the best fit line is shown in red. The fit demonstrates an almost linear relationship in this temperature regime.

where m is the slope and b is the y-intercept. We will need to understand what they correspond to in the real world, but for now let's leave them be as mathematical variables. The table of data is in the Appendix. The plot I generated using techniques you learned in excel is shown in Figure 1. Note that whenever you put a figure in your report, you need to have a title, properly labeled aces, and a caption that fully details what is being shown. As a rule of thumb, someone should get the gist of what you're trying to say just by reading the captions.

After coming up with your graph, you want to use the LINEST function, as discussed in the previous document on uncertainty and error, to obtain *fit parameters* and *goodness of fit*. For this particular fit, the LINEST function tells me,

$$\begin{aligned} m &= 16.74 \pm 1.58, \\ b &= -696.07 \pm 104.33, \\ R^2 &= 0.91. \end{aligned} \tag{2}$$

It is always a good check to see if things are ok if the error is substantially smaller than the actual value of b and m derived. After you arrive at the equation describing your fit as well as the R^2 value which characterizes the goodness of your fit, you want to conduct the following steps,

1. Understand how these parameters relate to variables of physical interest.
2. Propagate your error in your fit variables to those variables of physical interest.
3. Determine whether the errors on the parameters are small enough for your values to be reliable. Do they match any known results?
4. Determine whether the parameter values are reliable based on the goodness of the fit.
5. Display these results in a table.
6. State your findings in writing.

Let's walk through each of these steps for this example and figure out how to put it into a lab report.

First, you want to study what these parameters mean. Our parameter m is the slope, indicating the amount of sales you would increase by if the temperature rose by a degree. This number seems to be 16.74 and the units are USD/degree. This already seems like a useful quantity, so in the lab report, I may write it down and report as M, the amount of money you'd earn in sales per degree in Fahrenheit,

$$M = (16.74 \pm 1.58)\ USD/^{\circ}\text{F}. \tag{3}$$

The next parameter is b, which is a y-intercept. This tells me that I'd make negative 700 dollars at 0 degrees Fahrenheit. Clearly, something is wrong here since I won't all of a sudden owe people money. This doesn't mean our model is *wrong*, it just means it become a bad model below a certain temperature. In the report, I would report this statistic as indicating that there is some temperature below the temperatures we studied by well above 0 where our model is no longer reliable. I can then explain it by saying there is a point when we get below a certain temperature where no one is buying ice cream anymore. I can also add that at a certain point, no one will buy ice cream when the temperature is too hot (i.e. during heat wave). Therefore, the way we should understand this model is one that is very accurate in this temperature regime. If I went too far above or too far below, this would no longer be a good model.

Lastly, you want to look at the goodness of the fit. As a rule of thumb, if your R^2 value is greater than 0.9, then you have a fit that matches the data well. Since my R^2 value is 0.91, my fit matches the data well. This will be important to note. At this point, we've gathered enough analysis to put it into writing.

Putting Pen to Paper

Now that you have gathered your analysis, it is time to present your argument in the paper. Your job is to

1. Present the data to the reader in a legible fashion.
2. Justify why the results from the data are reliable.
3. Draw connections to physical results and conclusions from the data.
4. Justify to the reader why these physical results can be trusted.

To that end, here is an example data and analysis section,

Data and Analysis

We measured the total sales in ice cream over a week long span and recorded the average versus the average daily high temperature for that week. The error is based on the standard deviation of the measurements we took throughout the week. All data are in Figure 2, with data shown in blue and the best-fit line shown in red. The best fit line is a linear fit and matches the data well as verified by $R^2 - 0.96$.

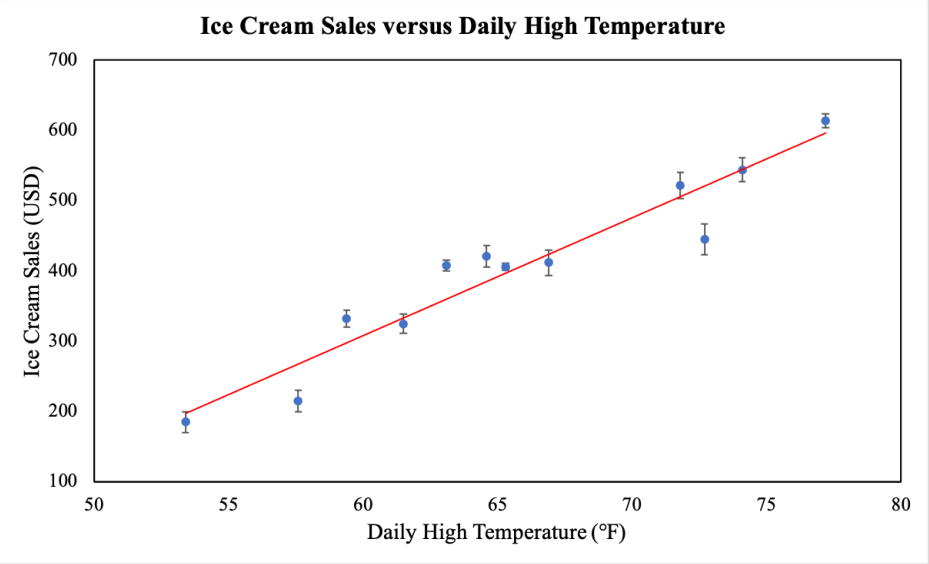

Figure 2: Passing Yards again Touchdowns. This graphs the yardage gained by quarterbacks (on the y-axis) against the number of passing touchdowns achieved by those quarterbacks (x-axis). The data points are in blue with error bars on yardage shown in black. The red line demonstrate a decent linear relationship.

The fit parameters are the slope, m, and the y-intercept, b. The slope itself is related to the amount of sales per unit change in temperature, M. Its value is

$$M = (16.74 \pm 1.58)\ USD/^{\circ}\text{F}. \tag{4}$$

This will be useful in predicting profit margins for the local ice cream business store.

The other fit parameter, b, does not have immediate real-world significance. On the other hand, it implies there is a constraint on the reliability of our model. This fit parameters demonstrates that if we extrapolate a linear fit to *all* temperatures, the ice cream business would be paying *out* money to people not buying ice cream, which is not a situation that would happen in the real-world. Therefore, at lower temperatures we expect the linear fit to no longer work and taper off towards 0 earned without going negative. This is also true in the high temperature regime. At a certain point, it would be too difficult for people to go outside and get ice cream and so it must decrease as temperatures get exceedingly high. Therefore, the relevance of this fit parameter is that it tells us our linear fit model is only relevant in the regime of temperatures of interest. All fit parameters and the fit statistic, R^2 is contained in Table 1.

Table 1: Fit Parameters

Parameter	Value	Error
A	88.96 yards/touchdown	3.47 yards/touchdown
g	87.12 yards	6.06 yards
z	229.46 yards	11.61 yards
R^2	0.96	

Since the errors are sufficiently small compared to the values themselves, these results are indeed good estimates of experimentally measured values. In addition, M matches the literature value of 17 USD per unit temperature, indicating that our results are consistent with previous expectations (Dodd, 2022).

As you can see, we were able to present the data in a graph, the fit parameters in a table, justify why these results are reliable, and then draw conclusions about the physically relevant variables. This is more or less what we expect to see in your lab reports!

Conclusion

In applying this to your own lab reports, remember that you need to understand the important equations that you are applying to your data. Once you collect the data, you should figure out how best to display the data and then match it to the equations you previously derived. Moreover, you will need to justify why we can trust your data. This is done through error analysis. Good luck with your future lab reports!

Appendix

Table 2: Passing Yards and Touchdowns Raw Data

Temperature (°F)	Sales (USD)	Error on Sales (USD)
57.6	215	15
61.5	325	14
53.4	185	15
59.4	332	12
65.3	406	5
71.8	522	19
66.9	412	18
77.2	614	10
74.1	544	17
64.6	421	15
72.7	445	22
63.1	408	8