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 passing yards and passing touchdowns in the National Football League last year. For pedagogical purposes, I rearranged the data to make it linear. In the context of physics, you may think of this as setting the "touchdown variable" to be some number and observing the output "passing yards variable." The table of data is in the Appendix to this document. 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

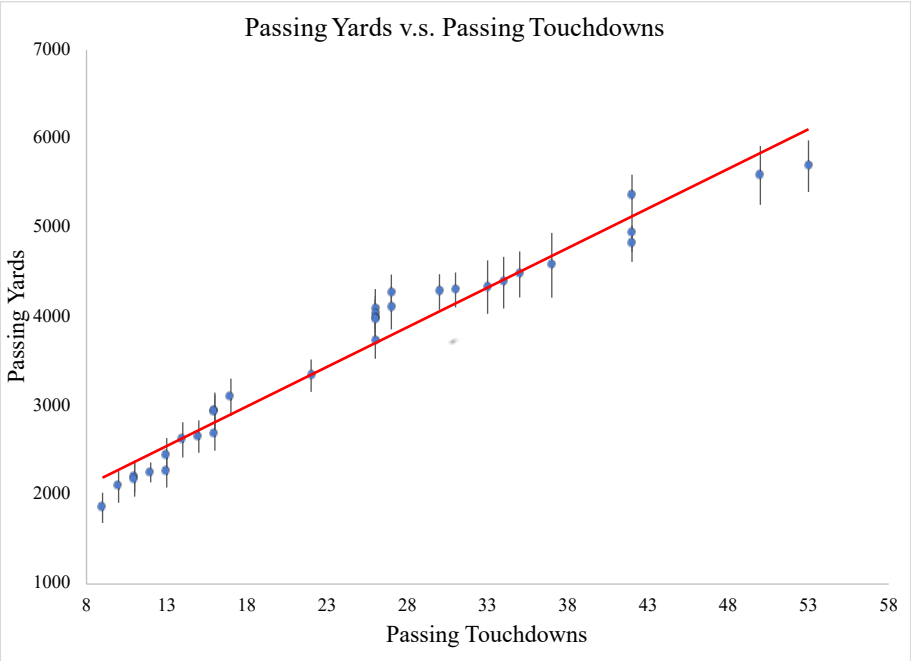

Figure 1: Passing Yards again Touchdowns. This graphs the yardage gained by quarterbacks (on the y-axis) against the number of passing touchdowns achieved by those quarterbacks (x-axis). The data points are in blue with error bars on yardage shown in black. The red line demonstrate a decent linear relationship.

need to have a title, properly labelled axes, and a caption that fully details what is being shown.

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,

$$\begin{aligned} y &= (88.96 \pm 3.47)\, x + (1394 \pm 97)\,, \\ R^2 &= 0.96. \end{aligned} \tag{1}$$

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.

Relating Fit Parameters to Relevant Variables

In order to do this, you need to have an equation you derived to relate the variables in question. In this case, since this isn't physics, I won't derive the equation I will simply state a reasonable guess for how they are related,

$$Y = A \cdot T + 16g, \tag{2}$$

where A is the amount of passing yards gained on touchdown drives and g is the total number of yards gained on non-touchdown drives per game. Here, Y means yards and T means touchdowns. From (1), we see that

$$\begin{aligned} A &= 88.96 \pm 3.47 \text{ yards/touchdown}, \\ g &= 87.12 \pm 6.06 \text{ yards}. \end{aligned} \tag{3}$$

Note that I have to put units! In g, I had to use error propagation. Since g was being multiplied by 16 I simply needed to divide on both sides. Your equations may involve more complicated error propagation! As you can see, we have accomplished the first two steps already, we matched our theoretical equation (2) to our experimental equation (1). Then, after propagating error, we got errors on our variables of physical interest in (3).

Examining Reliability

Now that we have obtained our variables, we should determine whether they are reliable. The first way to do this is to see whether the errors on my parameters are small enough. The rule of thumb is that if the order of magnitude of your errors is smaller than the order of magnitude of the parameter itself, then it is reliable. In both the case of A and g, the errors are indeed smaller by one order of magnitude.

Next, you want to note whether this matches any known results. In this scenario, it turns out that the average number of passing yards per game should be about $z = 1.6 * A + g$. Using your error propagation methods, you should find that

$$z = 229.46 \pm 11.61 \text{ yards.} \tag{4}$$

Upon looking up the actual number, you would find that it is about 227.54 yards. Since this number fits within the range of values of z, this is another indication that our results are reliable.

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 0.96, 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

For each number of passing touchdowns in a season, we measured the associated total number of passing yards. We estimated the error based on the proximity to the yardage lines marked on the field, as one would estimate length error measurements with a ruler. We display the number of passing yards versus the number of passing touchdowns in a season in Figure 2, with data points in blue and error bars in black. The red line demonstrates that the data is well fit by a linear relationship, verified by an R^2 value of 0.96.

The fit parameters are the slope, $m = 88.96 \pm 3.47$ and the y-intercept, $b = 1394 \pm 97$. We can derive the values of the relevant variables stated in the previous section through

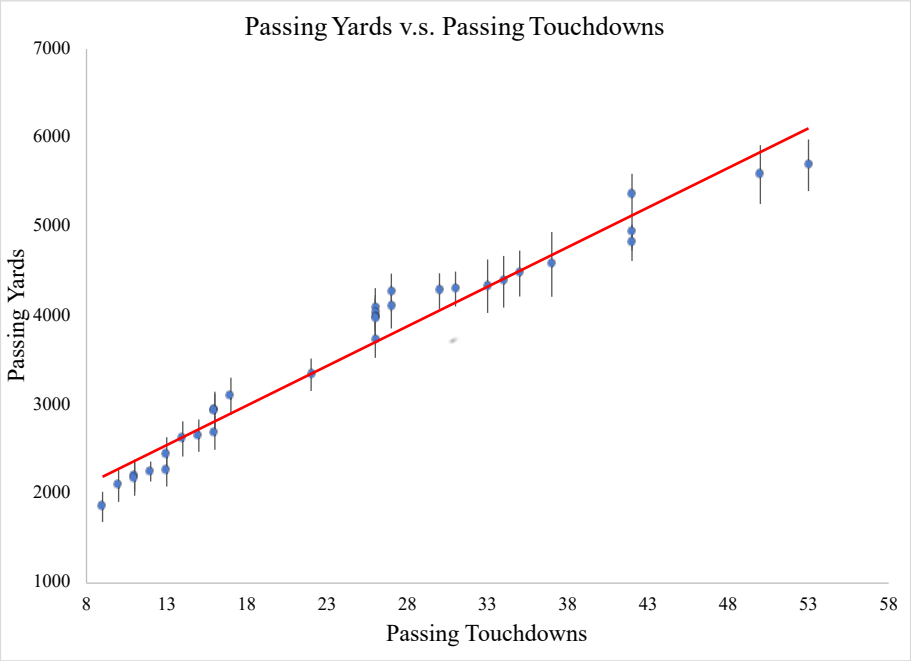

Figure 2: Passing Yards again Touchdowns. This graphs the yardage gained by quarterbacks (on the y-axis) against the number of passing touchdowns achieved by those quarterbacks (x-axis). The data points are in blue with error bars on yardage shown in black. The red line demonstrate a decent linear relationship.

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	

error propagation. First, the number of yards per touchdown, A, is identical to m, so

$$A = 88.96 \pm 3.47 \text{ yards/touchdown}. \tag{5}$$

Next, the y-intercept is related to $16g$. Upon dividing 16 on both sides, we obtain,

$$g = 87.12 \pm 6.06 \text{ yards}. \tag{6}$$

From here, we can find the most interesting variable, yards per game, using $z = 1.6A+g$,

$$z = 229.46 \pm 11.61 \text{ yards}. \tag{7}$$

These parameters and the R^2 value are displayed in Table 1.

Since the errors are sufficiently small compared to the values themselves, these results are indeed good estimates of experimentally measured values. In addition, z matches the literature value of 227.54 yards within error, indicating that our results are consistent with previous expectations.

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

Passing Touch Downs	Passing Yards	Error on Passing Yards
53	5694	290
50	5590	330
42	5361	238
42	4941	208
42	4823	204
37	4581	364
35	4478	258
34	4386	290
33	4336	300
31	4304	196
30	4281	200
27	4265	212
27	4103	242
26	4084	230
26	4030	206
26	3984	210
26	3971	220
26	3733	202
22	3341	182
17	3098	210
16	2943	208
16	2933	194
16	2688	190
15	2657	182
14	2620	198
13	2437	202
13	2259	174
12	2254	112
11	2208	180
11	2170	188
10	2091	178
9	1856	170