Probability and Statistics Project

Part A – Perform a Hypothesis Test Using Data Collected from the Real World

Part B – Create a Regression Model using Bivariate Data Collected from the Real World

Important Directions to Follow for your Project

- Create a document file in Microsoft Word, Open Office, or some other word processing software which will serve as your final report for this project.
- Include your name and a title for each part of your project.
- The project will be graded on content of your answers and comprehension of the material. Walk the reader through what you did as if they know nothing about your investigation. You will be graded on clarity, neatness, completeness and visual presentation in your report.
- Insert all your graphs into the report file or print and attach them.
- Your project should be put together in the same order that you see in the original directions. Any different order will result in a loss of points in the overall neatness and clarity of your project.
- This project is due no later than **the day of the final exam**. No exceptions will be considered.
- Submit your project electronically through email or give it to me in person at the final exam.

Taking Screen Shots for your Project

Use the following steps to obtain screenshots of any computer generated graphs you create. You can print your images or insert them into the document file for your project.

Taking Screen Shots on Mac

To capture a portion of the screen, press **Command-Shift-4**. A cross-hair cursor will appear and you can click and drag to select the area you wish to capture. When you release, the screen shot will be automatically saved as a PNG file on your desktop.

To copy a portion of the screen to the clipboard, press **Command-Control-Shift-4**. A cross-hair cursor will appear and you can click and drag to select the area you wish to capture. When you release the mouse button, you can paste the screen shot to another application.

To capture a specific application window, press and hold **Command-Shift-4 then tap on the Spacebar**. The cursor will change to a camera, and you can move it around the screen. As you move the cursor over an application window, the window will be highlighted. The entire window does not need to be visible for you to capture it. When you have the cursor over a window you want to capture, just click the mouse button and the screen shot will be saved as a PNG file on your desktop.

To copy a specific application window, press and hold **Command-Control-Shift-4 then tap on the Spacebar**. The cursor will change to a camera, which you can move around the screen. As you move the cursor over an application window, the window will be highlighted. The entire window does not need to be visible for you to capture it. When you have the cursor over a window you want to capture, just click the mouse button and you can paste the screen shot into another application.

Taking Screen Shots on Windows

To take a screenshot of only one window click on the title bar of the window that you want to capture. **Press "Alt + PrtScn**". A screenshot of your currently active window will be copied to the clipboard. Paste it into an image editor or document editor. Note: On some laptops and other devices, you may need to press the "**Alt + Fn + PrtScn**" keys instead.

To take a screenshot of part of your screen press "**Windows + Shift + S**". Your screen will appear grayed out and your mouse cursor will change. Click and drag on your screen to select the part of your screen you want to capture. A screenshot of the screen region you selected will be copied to your clipboard. You can paste it in any application by selecting **Edit > Paste** or pressing **Ctrl + V**, just as you'd paste a full-screen shortcut taken with the Print Screen key. This only works in Windows 10's Creators Update. On older versions of Windows, this shortcut is part of Microsoft's OneNote application. With the Creators Update, Microsoft integrated this shortcut into Windows 10 itself.

Another option is typing "**Snipping Tool**" into the start menu in the lower left. Click "**New**" to allow yourself to click and drag a selection to take a snapshot of.

Part A – Compute Statistics for a One Variable Sample Taken from the Real World

Section A1 – Choose two related populations and make a claim (10 pts)
First, decide what the topic of your study should be. Who or what are you interested in gathering statistics on? There are many, many possibilities for topics. In this part of the project we want to eventually use statistical methods to compare two similar populations. You may wish to generate evidence for an experiment of some kind.

EXAMPLES OF DATA THAT COULD BE FOUND ON THE INTERNET:

1. Square footage of a house and price (restrict to a particular location – zip code)
   www.homesteadnet.com , www.nothnagle.com , www.realtor.com
2. Weather and climate data
3. Sports statistics
4. Miles on a particular car (same model, year, options) and its value
   www.Edmunds.com , www.kelleybluebook.com
5. Health and Disease Statistics
6. Crime Statistics
7. Science Data
8. Astronomy Data
9. Economic Statistics
10. Political Data

Examples of populations that you might wish to study:

Heights of living things
Population sizes in particular regions
Number of births in a given time period
Peoples' use of safety interventions (seatbelts, airbags, masks, personal protective gear, etc)
Cigarette or vaping use
Cost of a service or product
Time spent working out
Number of times per week spent exercising
Amount of money people spend on a service or product
GDP (Gross Domestic Product) of a country
Debts or deficits of governments
Number of deaths due to a particular condition
Number of crimes of a specific type committed
Presence of a pollutant in the environment
Distance an object travels, such as a car
Mileage of a vehicle
Distance people drive to work or school
Time that people spend commuting to work or school
Number of police in a locality
Rent prices
Amount students spend on course materials
Credit hours students are taking
Salaries at a particular type of job
Number of hours spent sleeping

Number of hours spent working at a job
Number of hours spent studying
Amount people spent on gasoline in a time period
Amount spent on groceries per week
Number of people living in a household
Amount spent on car repairs (and/or maintenance) per year
Age of person
Car insurance costs

* The above are just a few ideas for populations, but the possibilities are, of course, virtually endless.

In one example, I may be interested in studying how often people are leaving their homes and going out into the community in December 2020, during the COVID-19 pandemic.

In another example, I may be interested in studying how often people were wearing masks on the street during the pandemic in December 2020, during the COVID-19 pandemic.

Keep in mind, that you'll be selecting your topic and samples at a time when you likely just know the basics of data collection, but before you know all the statistical techniques that will be required for analysis. In order to collect good data for future analysis, and to be able to answer all future questions, pick <u>one topic</u> and then collect <u>two distinct samples</u> regarding that same topic. The purpose of having two distinct samples is so you may make a claim about the difference between the two different population parameters. Your study will either involve population means $(\mu_1 \text{and } \mu_2)$ or population proportions $(p_1 \text{and } p_2)$ whichever are appropriate. Then you'll say how you think they compare. Be sure to state your claim in both words and symbolic form.

In one example I might say, "**I claim that in the month of December 2020, those who live in the city of Buffalo are leaving their residence and going out into the community more often than those who live in Williamsville.**"
So, I could take one sample of responses from the city of Buffalo, and a separate sample from Williamsville. When I do the analysis, I will compare the results that I had got from the different groups. The frequency of how often people are leaving their homes in a month during the pandemic is an example of **numerical quantitative data**. Each person I ask the question will be giving me a numerical value, a number of occurrences, as a response. In this case, μ should be used. I'll let μ_1 stand for the population mean for Buffalo, and μ_2 will stand for the population mean for Williamsville. In symbolic terms my claim is $\mu_1 > \mu_2$

For a different project I might say, "**I claim that in the month of December 2020, those who live in the city of Buffalo are wearing masks on the street more often than those who live in Williamsville.**"
Whether people are wearing a mask on the street or not, could be considered a binomial "yes or no" type of question, depending on how you collect the data. If you aren't measuring a quantitative property, such as time spent wearing the mask, but instead are simply assigning a "yes or no" to the people you observe on the street, then this is an example of binomial qualitative data. In this case p should be used. I'll let p_1 stand for the city of Buffalo population proportion, and I'll use p_2 to stand for the Williamsville population proportion. In symbolic terms my claim is $p_1 > p_2$

Section A2 – Gather the data (10 pts)

Come up with a plan for gathering your data. You have the option of performing a survey. Discuss whether your method of sampling is along the lines of simple random sampling, stratified sampling, systematic sampling, cluster sampling, or perhaps convenience sampling. Describe very clearly how you collected your data. If the data is gathered from the internet or some other source, provide the appropriate information showing where you got the data from. If someone else gathered the data explain what methods they used to get the data. **Be sure to display your raw data in your project**.

Section A3 – Calculate statistics for each of your two samples (10 pts)

If you are using μ for your project then calculate these sample statistics for both of your samples:

sample mean, median, sample standard deviation

To do this you can calculate these statistics easily using the website www.desmos.com/calculator

- Go to the site and click plus (+) in the upper lefthand corner. Select "add table."
- Enter your first sample data into the x_1 column.
- Create an x_2 column and enter your second sample.
- Then type the following in some new fields:

If you are using p for your project then calculate these sample statistics for both of your samples:

sample proportion $\hat{p}$

The first sample proportion will have the formula $\hat{p}_1 = \frac{x_1}{n_1}$

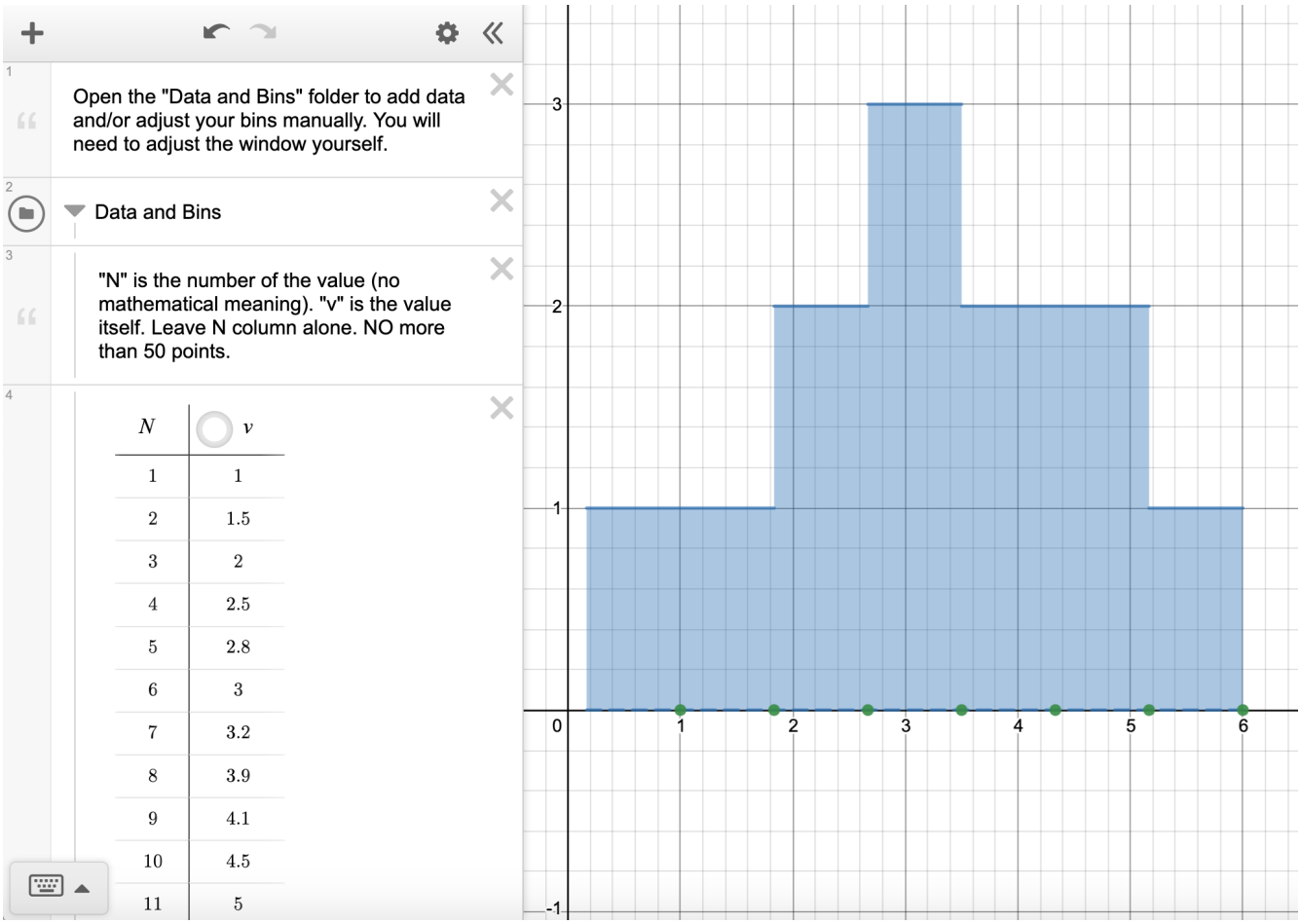
The second sample proportion will have the formula $\hat{p}_2 = \frac{x_2}{n_2}$

Present all the statistics you get in your final report.
Comment on the meaning of each of these values as they relate to your data.

Section A4 – Display your sample data using histograms, only if using μ in your project. (10 pts)

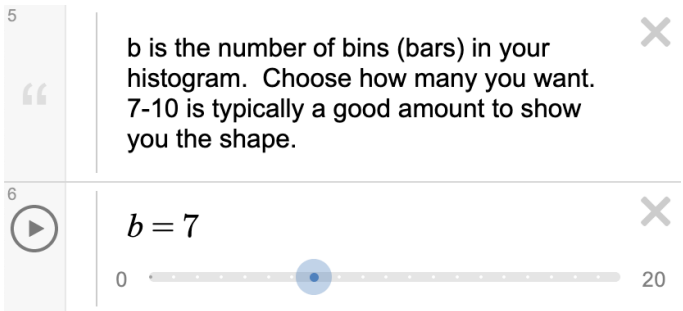
If you are using μ in your project then create a separate histogram for each of your samples.
If you are using p, instead, then the histograms can't be created, so you won't have to do this section.
When creating a histogram use a reasonable number of classes.
This can be done in Desmos at this link: https://www.desmos.com/calculator/lcnnbqh0nc

If the size of your sample is too big for Desmos then you may draw your histogram free hand on some graph or blank paper. (Use a ruler and make it neat)

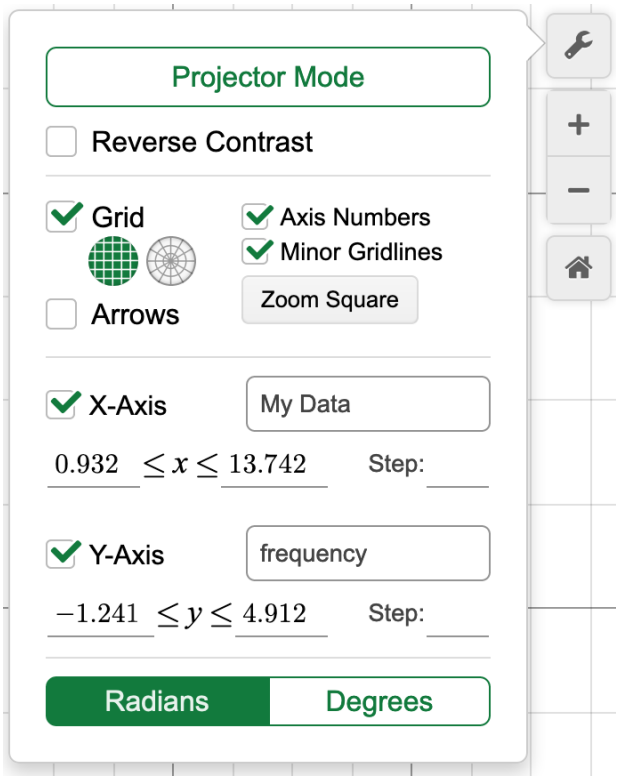


In Desmos as it says, do not change the N column. Enter all your data in the v column,

You can change the number of bars by scrolling down and changing the b value.



You can change the label on the *x* axis by clicking on the wrench icon in the upper right.



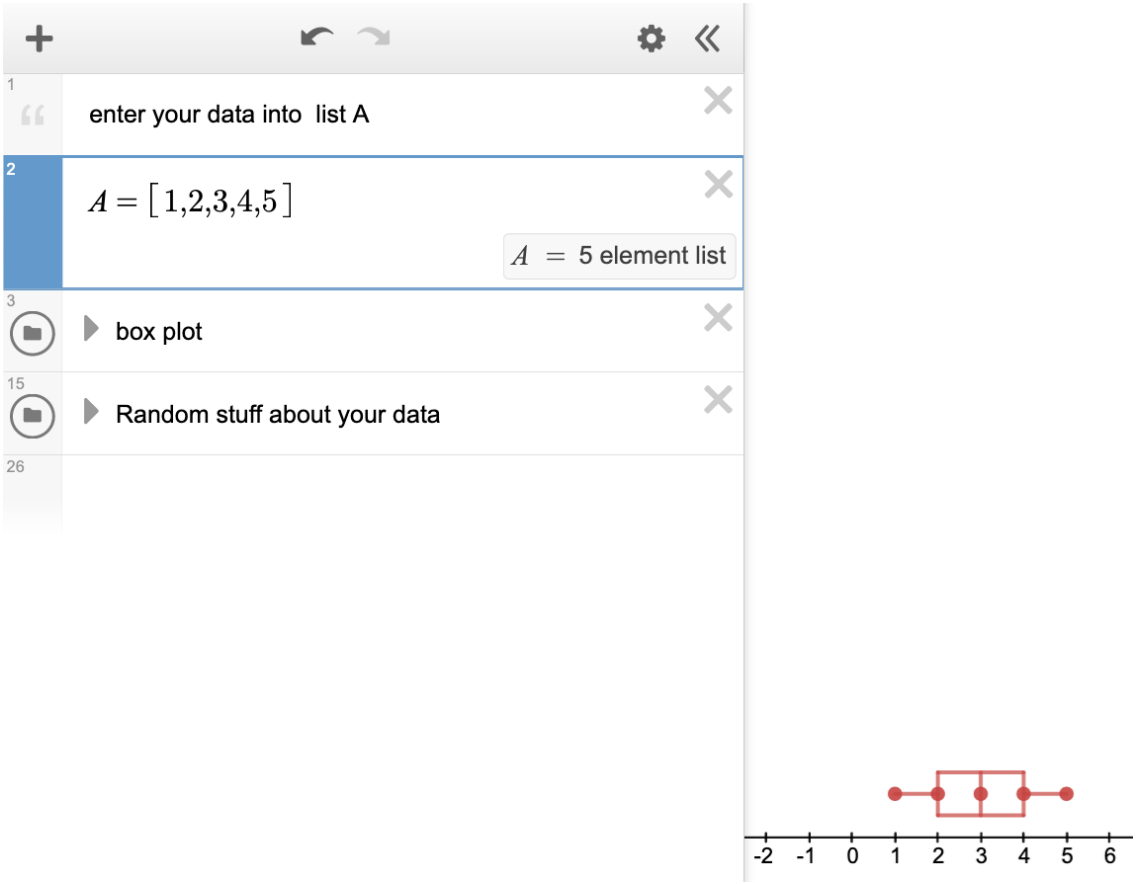
Include your histograms in your final report.
In your report describe what type of distributions, if any, your histograms show.
You can also comment on differences seen between your two histograms.
Why do you think the histograms are similar or why do you think they are different?

Section A5 – Box and Whisker Plot, only if using μ in your project. (10 pts)

If you are using μ in your project then create a separate boxplot for each of your samples.
If you are using p, instead, then the boxplots can't be created, so you won't have to do this section.

A box and whisker plot for your data can be created in Desmos at this link:

https://www.desmos.com/calculator/h9icuu58wn



Include both of your box plots in your final report.
Try to line up the axes visually so they can be easily compared to each other.
State what the box and whisker plots tell you about your data.
For each boxplot, say what is the minimum entry, Q1, Q2, Q3, and the maximum entry.

Section A6 – Create Confidence Intervals for Both Samples (10 pts)

a) Choose a level of confidence to apply to both your intervals.

b) Find the appropriate critical value, which could be a z_c or t_c value.
Explain why you used the value that you used.
You may use the course formula sheets, calculator, or sites such as:
https://stattrek.com/online-calculator/normal.aspx
https://stattrek.com/online-calculator/t-distribution.aspx

c) Find the margin of error for each confidence interval using the appropriate formula:

$$E = z_c \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad E = t_c \frac{s}{\sqrt{n}} \quad \text{or} \quad E = z_c \sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

d) Find the left and right endpoints of each confidence interval.

Comment on any differences or similarities in the confidence intervals.
Do you think it is likely that the population means / population proportions are equal, or are they likely different. Why do you think this?

Section A7 – Perform a Hypothesis Test and Make a Decision about your Original Claim (10 pts)

Using the level of confidence that you chose in the previous section, find a significance level for your hypothesis test. Use the formula $\alpha = 1 - c$. Follow all applicable methods covered in chapters 7 and 8 regarding hypothesis tests. Determine what type of hypothesis test needs to be done. Then make sure that you followed all the same applicable steps that were demonstrated in the notes, text and videos.

a) Decide whether your test is about comparing population means or population proportions. Also, decide whether you believe your samples are primarily independent or dependent.

b) Write down the null and the alternative hypotheses. Indicate which of the two is the claim.

c) Write down your significance level.

d) Write down any critical z_c , $-z_c$, t_c or $-t_c$ values.

e) Find your standardized test statistic.

f) Write down whether you reject the null or failed to reject it.

g) Write down a conclusion regarding your original claim.
Be sure to be as specific as possible in this part, and state what you believe your decision means in terms of the population you studied.

Part B – Create a Regression Model using Data Collected from the Real World

Section B1 – Make a claim about the relationship of one numerical variable to another variable (10 pts)

a. Explain what bivariate data you wish to collect and explain what is the population of your study.
Note: Bivariate numerical data will have two variables, x and y, and you may believe the variables have a relationship to each other. Examples of bivariate data are shown below.

b. Describe what type of relationship you expect to find between the two variables.
Choices include but are not limited to: linear, exponential, logarithmic, logistic, power, polynomial

c. Give a reason why you believe the type of mathematical model you selected is most appropriate for this situation. For example, did you look at the scatterplot of your data and see that the trend in the scatterplot visually resembled a particular type of functions we've studied? Is there some other theoretical or intuitive reason why you believe the type of function you selected would be the best fit?

d. What type of correlation do you expect to find? (strong correlation, weak correlation, no correlation) Why?

A FEW EXAMPLES OF BIVARIATE DATA ARE BELOW

The data must be real world data recorded from some naturally occurring source.
The data should not be generated by theoretical means, such as by a formula.

Examples of good real world data pairs to collect are below. You are welcome to choose something different if you wish.

Height of a living thing versus time
Population in a particular region versus time
Number of births versus time
Use of a safety intervention versus time (seatbelts, airbags, masks, personal protective gear)
Cigarette or vaping use versus time
Amount of money spent on a service or product versus time
GDP (Gross Domestic Product) of a country versus time
National debt versus time
Number of deaths due to a particular condition versus time
Number of crimes of a specific type committed versus time
Presence of a pollutant in the environment versus time
Distance an object travels versus time
Volume of liquid in a container evaporated versus time
Distance people drive to work and time that each spends commuting
Number of police in a local area and crime in that same area
Rent prices versus crime in an area
Amount students spend on course materials and number of credit hours registered for
Salaries versus time spent working at a job
Salary versus years of study
Number of hours spent sleeping and number of hours spent working
Height and weight (you may wish to restrict your sample to one gender)
Amount people spent on gasoline in a time period and miles driven in the same period
Amount spent on groceries per week and number of people in the household
Age of vehicle and average amount spent on car repairs (and/or maintenance) per year
Age of person and yearly car insurance cost
Age of person and age of car
Length of a pendulum (object attached to a string) and the period (time) of oscillation
Amount of salt in water versus freezing point of the salt-water solution

Section B2 - Collect 30 bivariate data (10 pts)

Collect at least 30 bivariate quantitative data. **A copy of your raw data needs to be included in your final report.** You have the option of selecting a sample from your population and performing a survey. Describe how you collected your data. If the data is gathered from the internet or some other source, provide the appropriate information showing where you got the data from.

Section B3 - Display your data with a scatterplot and a regression curve (10 pts)
You may use a variety of standalone or web-based programs. Examples are Minitab, Excel, and StatCrunch. Websites such as www.desmos.com/calculator will display scatter plots along with regression curves. A google search for keywords such as "regression scatterplot" will bring other websites up. Below are the instructions for doing linear regression in Desmos. However, you may decide that a linear model is not the best model for your data. You may use linear, exponential, logarithmic, logistic, power, polynomial, or other type of function.

Visit the website www.desmos.com/calculator

Step 1: Click the + in the upper left and then click "**table**"

Step 2: Enter the x and y values into section 1 which is the table.

Step 3: Click the + in the upper left and then click f(x) expression.

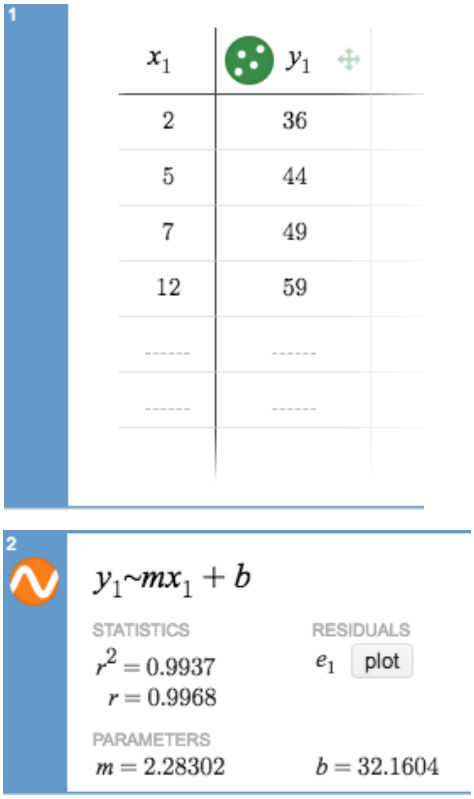
Step 4: Enter the following into section 2:

$$y_1 \sim m x_1 + b$$

r will be the correlation coefficient

m will be the slope of the best fit line.

b will be the intercept of the best fit line.



A link with with a table already set up is **https://www.desmos.com/calculator/jwquvmikhr**
Below are the basic forms of just a few other possible mathematical models:

Exponential: $y_1 \sim a b^{x_1}$

Polynomial: $y_1 \sim a x_1^2 + b x_1 + c$

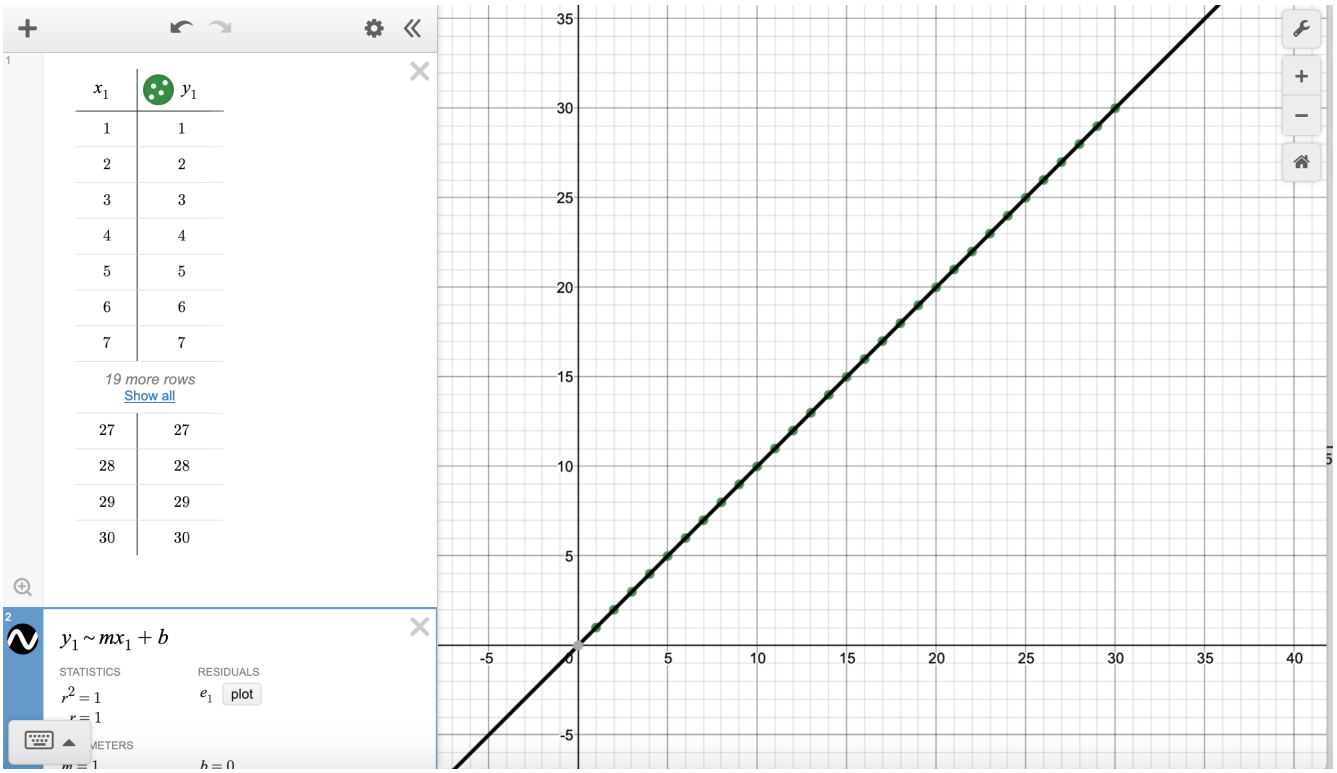
Logarithmic: $y_1 \sim a \ln x_1$

Power: $y_1 \sim a x_1^n$

Power: $y_1 \sim a x_1^n$

Logistic: $y_1 \sim \dfrac{L}{1 + A b^{-x_1}}$

Keep in mind we can introduce shifts of the functions in order to fit the data better if needed.
Scale the scatterplot so that all your data can be clearly seen and it takes up as much of the window as possible. Only one graph is required because the scatterplot and the regression line will appear together and they show the mathematical relationship between the x and y variables.



Copy the values of all your regression constants to your report.
In the example output above we see that $m = 1$ and $b = 0$
The regression line for this example $y = 0 + 1\,x$
The equation gives the slope and intercept of the best fit line.

You will also get an r correlation coefficient.
In this example output $r = 1$ because the scatterplot makes a perfect linear correlation.

Take a screen shot of your scatterplot and regression line and be sure to display your slope, intercept, line equation, and r value in your report.

Final Questions - Answer the following questions in your report. (10 points)

1. Were you surprised by the results of your hypothesis test or were the results what you expected?

2. What sources of error / bias do you believe were present in your hypothesis test investigation?

3. If you had more time / resources what would you do differently in your hypothesis investigation to get better results?

4. Based on your scatter diagram, is there a relationship between the *x* and *y* variables? If yes, describe the relationship. Do you believe there may be association, causation or neither? If you believe a change in one variable causes a change in the other, then say which variable you believe is the dependent and why.

5. Describe any correlation. Is it positive or negative? Strong or weak or moderate? Explain why you think your correlation ended up being what it was.

6. What do each of the constants from your mathematical model mean specifically in relation to the variables you are analyzing? Include units whenever appropriate.

7. Use your model to make a prediction. Pick a value of the horizontal axis (x) variable and use your formula to predict an output (y) value. What is the meaning of your prediction? Is it extrapolation or interpolation and why?

8. Are the results of the regression what you expected? If not, give a possible explanation for why it may be different from what you expected.

9. What would you do to improve the results of your regression if you were to continue this investigation?

10. Give a conclusion of your findings from the report for both parts A and B.