



**SUMMARY OF PART 1**

* Collect data and construct a well detailed histogram, making sure to describe the shape and any unusual features (if any exist)
* Calculate mean, median, mode, range, 5-number summary, standard deviation.
* Check your data for any outliers.
* Sketch and label a normal model for your data regardless of what your distribution looks like.
* Check to see what percentage of your data is within 1, 2, and 3 standard deviations of the mean and compare those percentages to the empirical rule.
* Make sure to describe your distribution (histogram) Shape, Center, Spread and explain why you chose those measures.
* Do the statistics calculated (mean/median etc.) support the shape of your histogram or would you have expected the shape to be different.

**Project Part 2: Linear Regression**

**The Project:** You will perform a linear regression analysis with residual diagnostics using a data set that you collect. With this data you will create and analyze a scatter plot and answer several questions about the data

**Gathering Data:**

* Select a sample size of at least 25 (x,y) values
* Make sure to specify your explanatory and response variables.

**Analyzing Data**

* Explain if you believe your two quantitative variables should have an underlying (cause/effect) relationship
* Use this data to create and analyze a scatter plot and a residual plots using DESMOS (or some similar spreadsheet program) and the steps we discussed in class.

**Summary**

* Explain your data:
* Who/What it represents?
* Where/How did you collect it?
* What units is the data in?
* Analysis of the scatter plot (to include):
  + An image of the scatter plot with appropriate scales and labels, including a regression line and equation and r2 value.
  + A description of the scatter plot (Shape, Strength, Type of Correlation)
  + A calculation and interpretation of r
  + A calculation and interpretation of the slope and r2
  + A residual plot and an analysis of whether a linear model is appropriate for the data
  + Identification of outliers and influential data points
  + Explanation as to how you feel about using your regression equation to make predictions.

|  |  |
| --- | --- |
| **Rubric for Statistics Project Part 2** | **Points Possible** |
| The process of collecting data from the web or other was completed correctly | 6 |
| The amount of data gathered was appropriate (at least 25 points) | 4 |
| Quantitative and not categorical variables were used | 4 |
| The data hypothesized to have or not have a relationship is reasonable | 12 |
|  | **26** |
| Scatter plot is drawn correctly | 6 |
| LSRL lines are calculated correctly and included on the scatter plot | 8 |
| R-squared and “r” are calculated correctly and included | 6 |
| Residual plots are drawn correctly | 6 |
|  | **26** |
| Slope is interpreted correctly for the graph | 8 |
| R squared is interpreted correctly for the graph | 8 |
| The appropriateness of a linear model is discussed correctly | 8 |
| Individual residuals are calculated and interpreted correctly | 8 |
| Comments on possible outliers or influential points | 8 |
| Explanation about using equation for predictions | 8 |
|  | **48** |
| **Total:** | 100 |

**Part III of Project**

**M&M Colors**

The M&M/Mars Corporation makes a variety of M&M candies. In 1995, they decided to replace the tan colored M&Ms with a new color. After conducting an extensive national preference survey, they decided to replace the tan M&Ms with blue M&Ms. The company’s Consumer Affairs Department announced:

*On average, the new mix of colors of M&M’s Plain Chocolate Candies will contain 13% browns, 20% oranges, 14% yellows, 13% reds, 16% greens and 24% blues.*

*While we mix the colors as thoroughly as possible, the above ratios may vary somewhat, especially in the smaller bags. This is because we combine the various colors in large quantities for the last production stage (printing). The bags are then filled on high-speed packaging machines by weight, not by count.*

The purpose of this activity is to compare the color distribution in a sample of M&Ms with the advertised distribution.

1. What is the model for the M & M distribution? (4 pts)

2. What is the null and alternative hypotheses? (In context) (6 pts)

**Data collection & summary**

3. Open your bag and count your M&Ms & record your color totals in the chart below, then calculate the third row. (6 pts)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Color** | Brown | Yellow | Red | Orange | Green | Blue |
| **My Counts** |  |  |  |  |  |  |
| Expected  (E) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

4. In a group of 3 or 4 collect the individual counts from each student, then take the sum for each color and put that total in the observed row of the chart below. Finally fill in the rest of the chart below. (6 pts)

Group Counts of M&Ms

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Color | Brown | Yellow | Red | Orange | Green | Blue |
| Observed (O) |  |  |  |  |  |  |
| Expected  (E) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

5. What was the group’s total sample size? *n* = \_\_\_\_\_\_\_\_\_\_ (2 pt)

**Analysis & conclusions**

6. Test your hypotheses using your individual data at a 5% significance level. Be sure you include ALL aspects of an appropriate hypothesis test, even if stated prior. (60 pts)

7. Now test your hypotheses for the group counts at the 5% significance level. You only need to include the Chi-Square value and p-value of your test, if you have the same conclusion as the individual counts just write “same conclusion” if different, then write the new conclusion. (8 pts)

8. Discuss your results:

* Which p-value do you think is more reliable to base your conclusion on? (The group or the individual one and why?) (6 pts)
* Why can we assume that the random sample check is met even though it doesn’t state it in the opening paragraphs? (2 pts)