Probability & Statistics Workbook

Regression

krista king
MATH

SCATTERPLOTS AND REGRESSION

■ 1. The table gives weight in pounds and length in inches for 3-month-old baby girls. Graph the points from the table in a scatterplot and describe the trend.

Weight (lbs)	Length (in)
9.7	21.6
10.2	22.1
12.4	23.6
13.6	25.1
9.8	22.4
11.2	23.9
14.1	25.8

■ 2. The following values have been computed for a data set of 14 points. Calculate the line of best fit.

$$\sum x = 86$$

$$\sum y = 89.7$$

$$\sum xy = 680.46$$

$$\sum x^2 = 654.56$$

■ 3. For the data set given in the table, calculate each of the following values:

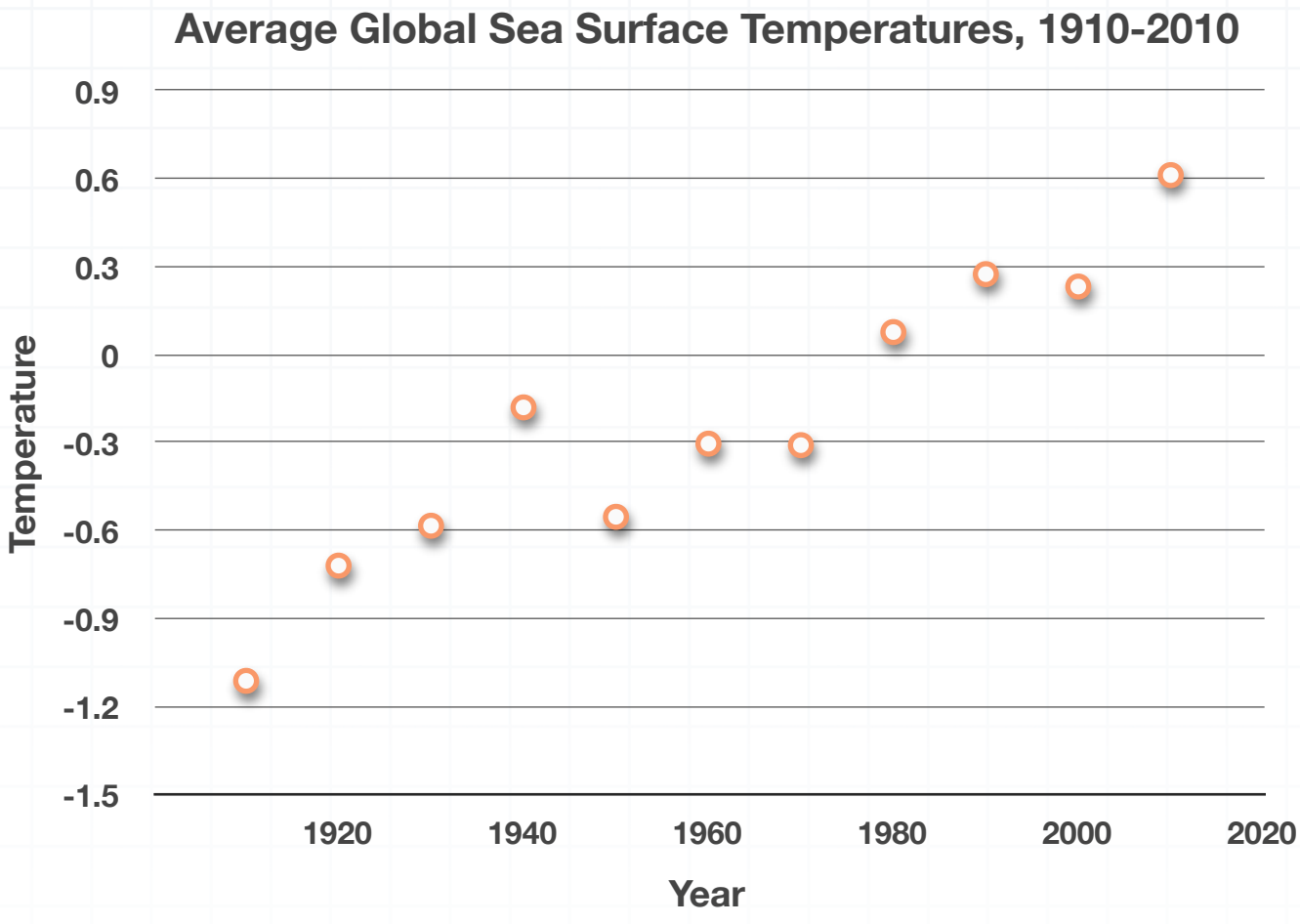
$$n,\ \sum x,\ \sum y,\ \sum xy,\ \sum x^2,\ \left(\sum x\right)^2$$

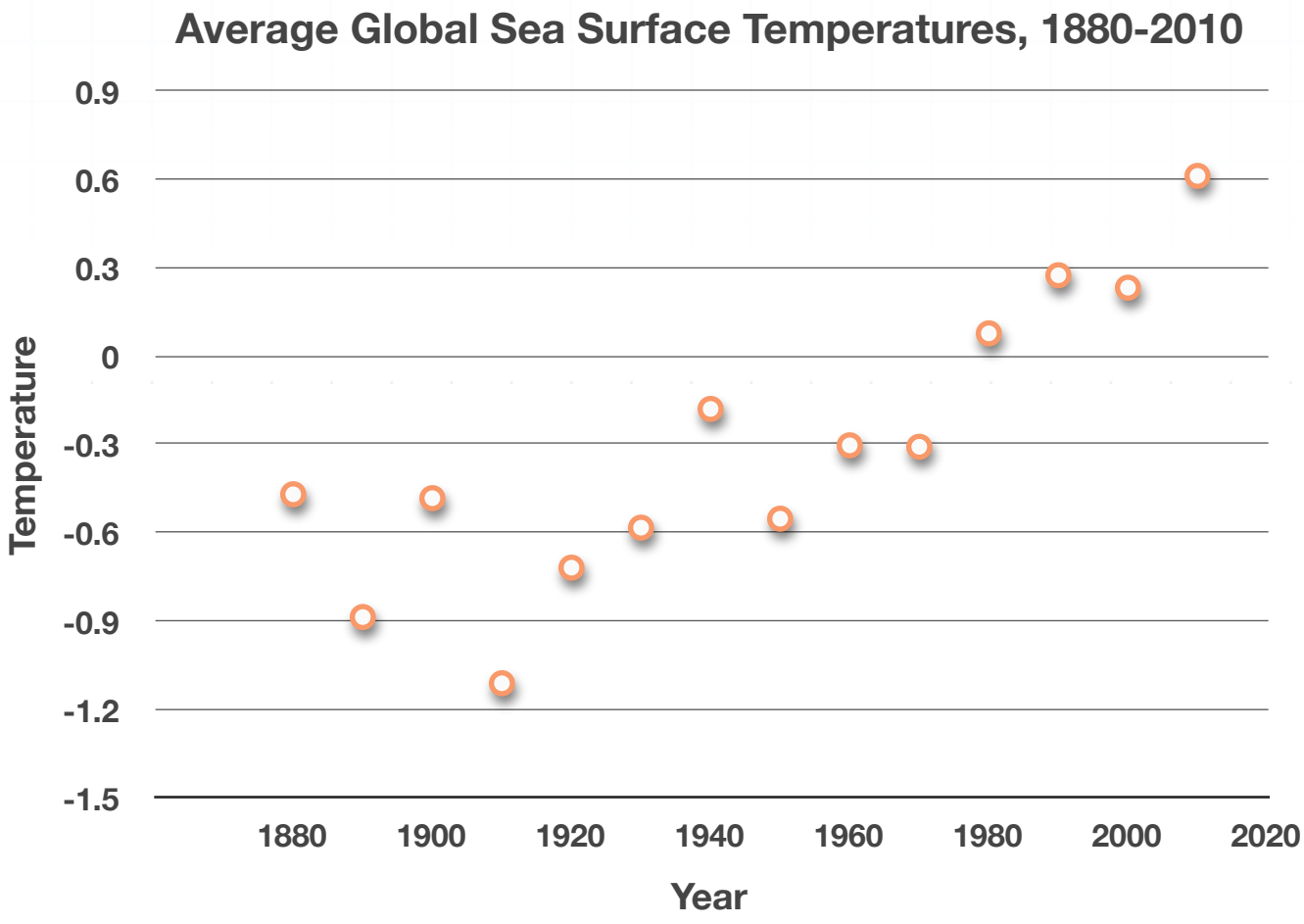
Month	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	73	73	75	75	77	79	79	81	81	81	77	75

■ 4. Use the Average Global Sea Surface Temperatures data shown in the table to create a line of best fit for the data. Consider 1910 as year 10. Use the equation to predict the average global sea surface temperature in the year 2050.

Year	Temperature, F
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

5. Compare the scatterplots. The second graph includes extra data starting in 1880. How does this compare to the plot that only shows 1910 to 2010? Explain trends in the data, and how the regression line changes by adding in these extra points. Which trend line would be best for predicting the temperature in 2050?



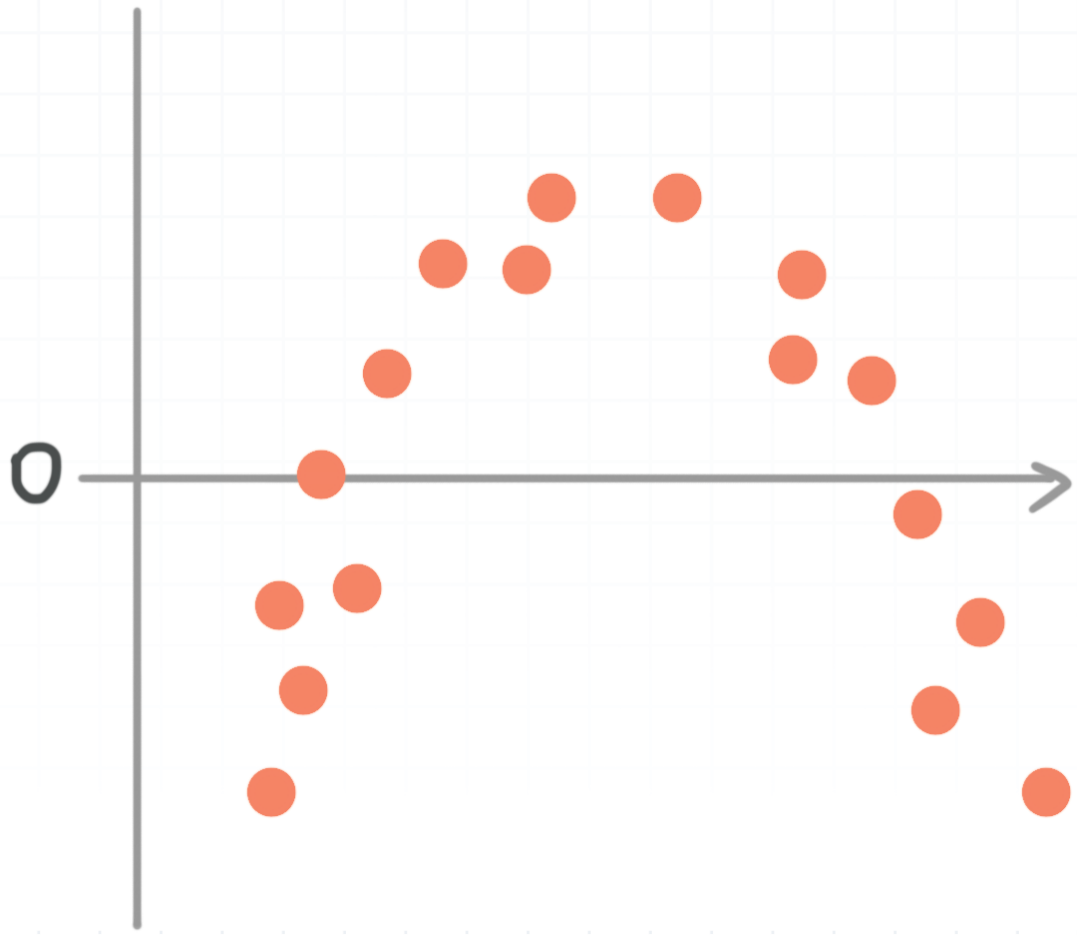


6. A small coffee shop wants to know how hot chocolate sales are affected by daily temperature. Find the rate of change of hot chocolate sales, with respect to temperature.

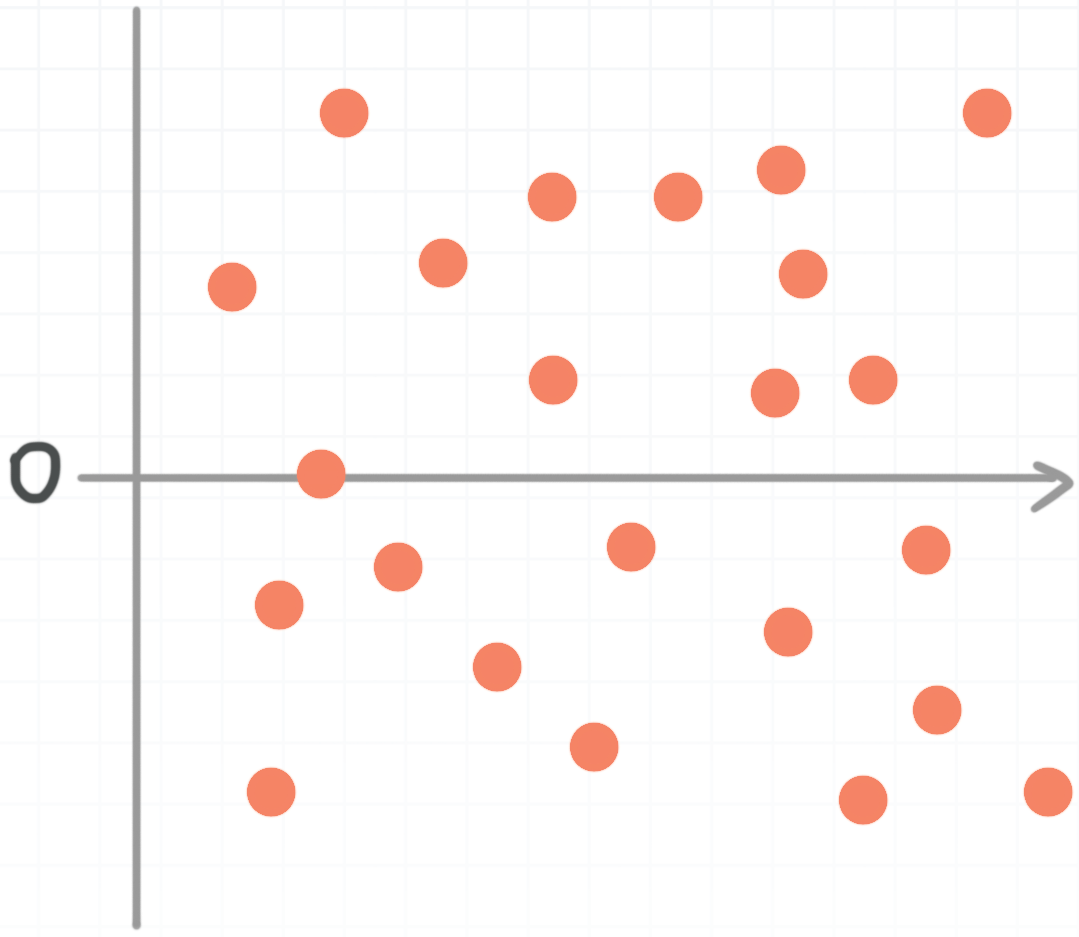
Daily Temperature, F	Hot Chocolate Sales
28	110
29	115
31	108
33	103
45	95
48	93
55	82
57	76

CORRELATION COEFFICIENT AND THE RESIDUAL

1. What does the shape of this residual plot tell us about the line of best fit that was created for the data?



2. What does the shape of this residual plot tell us about the line of best fit that was created for the data?



3. Calculate and interpret the correlation coefficient for the data set.

x	y
54	0.162
57	0.127
62	0.864
77	0.895
81	0.943
93	1.206

■ 4. Calculate the residuals, draw the residual plot, and interpret the results. Compare the results to the r-value in the previous problem. The equation of the line of best fit for the data is

$$\hat{y} = 0.0257x - 1.1142$$

x	y
54	0.162
57	0.127
62	0.864
77	0.895
81	0.943
93	1.206

■ 5. The table shows average global sea surface temperature by year. Calculate and interpret the correlation coefficient for the data set. Leave the years as they are.

Year	Temperature, F
1880	-0.47001
1890	-0.88758
1900	-0.48331
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

■ 6. Calculate the residuals and create the residual plot for the data in the table. Compare this with the r-value we calculated in the last question and interpret the results. Use the equation for the regression line $\hat{y} = 0.0097x - 19.1539$.

Year	Temperature, F
1880	-0.47001
1890	-0.88758
1900	-0.48331
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

COEFFICIENT OF DETERMINATION AND RMSE

■ 1. Linda read an article about the predictions of high school students and their GPA. The article studied three factors, the number of volunteer organizations each student participated in, the number of hours spent on homework, and the student's individual scores on standardized tests.

The article concluded that the number of hours spent on homework are the best predictor of GPA, because they found $24\,\%$ of the variance in GPA to be from hours spent on homework, $15\,\%$ from the number of volunteer organizations, and $11.5\,\%$ from individual scores on standardized tests.

What is the coefficient of determination for the line-of-best-fit that has y-values of high school GPA and x-values of hours spent on homework? Is the line of best fit a good predictor of the data? Why or why not?

■ 2. For the data in the table, calculate the sum of the squared residuals based on the mean of the y-values.

x	y
1	3.1
2	3.4
3	3.7
4	3.9
5	4.1

■ 3. Use the same data as the previous question to calculate the sum of the squared residuals based on the least squares regression line, $\hat{y} = 0.25x + 2.89$.

■ 4. Based on the previous two questions, in which we found the sum of the squared residuals based on the mean of the y-values and then the line of best fit, what percentage of error did we eliminate by using the least squares line? What is the term for this error?

■ 5. What is the RMSE of the data set and what does it mean?

x	y
1	3.1
2	3.4
3	3.7
4	3.9
5	4.1

■ 6. Calculate the RMSE for the data set, given that the least squares line is $\hat{y} = 0.0028x + 1.2208$.

x	y
5	1.25
10	1.29
12	1.17
15	1.24
17	1.32

CHI-SQUARE TESTS

1. We want to know whether a person's geographic region of the United State affects their preference of cell phone brand. We randomly sample people across the country and ask them about their brand preference. What can we conclude using a chi-square test at $95\,\%$ confidence?

	iPhone	Android	Other	Totals
Northeast	72	33	8	113
Southeast	48	26	7	81
Midwest	107	50	10	167
Northwest	59	33	10	102
Southwest	61	27	9	97
Totals	347	169	44	560

2. A beverage company wants to know if gender affects which of their products people prefer. They take a random sample of fewer than $10\,\%$ of their customers, and ask them in a blind taste test which beverage they prefer. What can the company conclude using a chi-square test at $\alpha = 0.1$?

	Beverage			
	A	B	C	Totals
Men	35	34	31	100
Women	31	33	36	100
Totals	66	67	67	200

3. A coffee company wants to know whether or not drink and pastry choice are related among their customers. The company randomly sampled fewer than $10\,\%$ of their customers, and recorded their drink and pastry orders. What can the restaurant conclude using a chi-square test at $99\,\%$ confidence?

	Bagel	Muffin	Totals
Coffee	38	34	72
Tea	25	29	54
Totals	63	63	126

4. A school district wants to know whether or not GPA is affected by elective preference. They randomly sampled fewer than $10\,\%$ of their students, and recorded their elective preference an GPA. What can the school district conclude using a chi-square test at $\alpha = 0.1$?

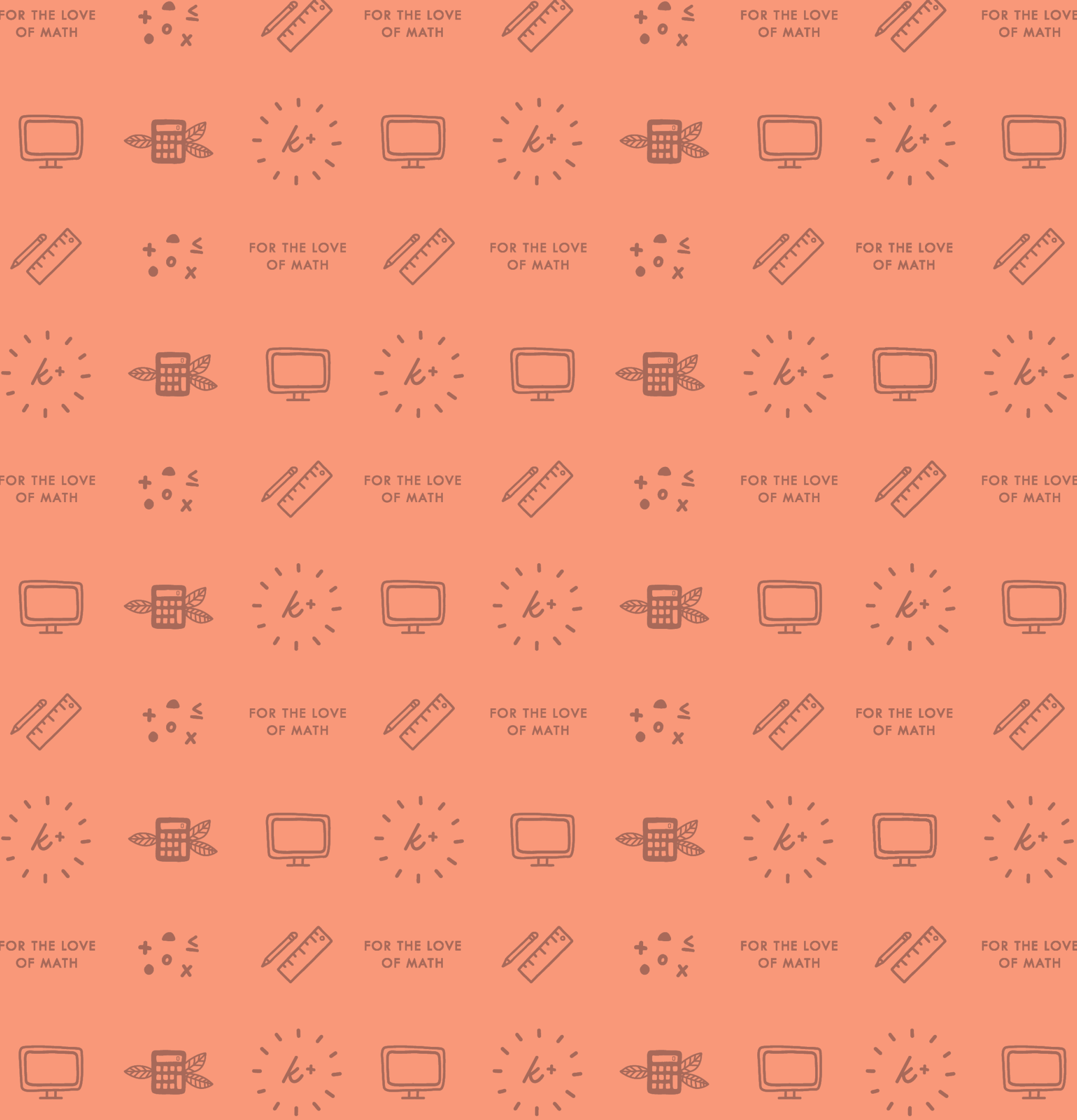
	GPA range				
	<2	2	3	4+	Totals
Music	12	26	31	34	103
Theater	21	22	23	21	87
Art	36	29	29	32	126
Totals	69	77	83	87	316

■ 5. An airline wants to know if people travel constantly throughout the year, or if travel is more concentrated at specific times. They recorded flights taken each quarter, and recorded them in a table (in hundreds of thousands). What can the airline conclude using a chi-square test at 95 % confidence?

Quarter	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Total
Flights	3.97	4.58	4.73	5.14	18.42

■ 6. A sandwich company wants to know how their sales are affected by time of day. They recorded sandwiches sold during each part of the day. What can the sandwich company conclude using a chi-square test at $\alpha = 0.1$?

Time of day	Midday	Afternoon	Evening	Total
Sales	213	208	221	642


FOR THE LOVE
OF MATH