1 Exploring Two-Variable Data

1.1 Two Categorical Variables

A side by side bar graph merges two bar graphs into one, in an attempt to compare the distributions of the two categorical variables.

A segmented bar graph is another way to display data, where each group is split by its relative frequency.

A mosaic plot is similar to a segmented bar graph, but it draws attention to the sizes of each group.

Joint relative frequencies are the ratio of the frequency in a cell and the total number of data values.

Marginal relative frequencies is the ratio of the sum in a row or column and the total number of data values.

Conditional relative frequencies are the ratio of a joint relative frequency and related marginal relative frequency.

Basically - joint relative frequency is the cell count divided by the table total, marginal is the row/column total divided by the table total and the conditional relative frequency is the intersection divided by the row/column total.

1.2 Scatterplots and Correlation

We have worked with bar graphs, box plots, and histograms. This is called variate data.

When we compare two variables or bivariate data, we are exploring the relationship between the two.

You should start with a scatter plot. A scatterplot shows the representation between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the plot fixed by the values of both variables for that individual.

A response variable measures the outcome of a study or an observation.

An explanatory variable helps explain or influences change in a response variable.

Commonly, explanatory variables are called independent and respondent are called dependent.

You can describe the overall pattern of a scatterplot by the direction, form, and strength of the relationship. An important kind of deviation is an outlier.

Form is the overall pattern or deviations from the pattern. Direction is whether the graph has a positive or negative slope. Strenght is how close the points lie to a simple form (such as a line).

Here are the steps to creating a scatterplot on a calculator:

- Load the x-values into list 1 and y-values into list 2.
- Using StatPlot - highlight the mini-scatterplot: XList:L1 and YList:L2
- Zoom:9 to fit the scatterplot and then graph

In order to strengthen the analysis when comparing two variables, we can attach a number, called the correlation coefficient (r), to describe the linear relationship between two variables.

The correlation measures the strength and direction of the linear relationship between two quantitative variables.

The formula to find the correlation is:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_1 - \overline{y}}{s_y} \right)$$

Correlation is a number between -1 and 1. The strongest correlations are closer to 1 or -1.

Correlation describes only linear relationship between two variables.

Correlation does not have units and changing units on either axis will not affect correlation.

Switching the x and y axes will not change the correlation.

Correlation is very strongly affected by outliers.

1.3 Linear Regression

Linear regression or least squares regression allows you to fit a line to a scatterplot in order to be able to better interpret the relationship between two variables as well as to make predictions about the response variable.

The fitted line is called the line of best fit and has an equation:

$$\hat{y} = a + bx$$

The way the line is fitted to the data is through a process called the method of least squares. The main idea is that the square of the vertical distance between each data point and the line is minimized.

Using a calculator we can find the slope of the least squares regression line by doing:

$$\text{Slope} = r\left(\frac{S_y}{S_x}\right)$$

Using this, we can find the y-intercept as well:

$$\text{y-intercept} = \overline{y} - \text{slope} \cdot \overline{x}$$

We can get the line by using Stat $\rightarrow$ Calc $\rightarrow$ 8:LinReg(a+bx)

Correlation does not imply causation.

To describe the strength of a prediction, we use the coefficient of determination. Basically we just use r^2, and this gives the proportion of variation in the values of y that is explained by least-squares regression on y on x.

A residual is a vertical distance between an observed value of the response variable and the value predicted by the regression line.

Residual value is the Actual value minus the Predicted value.

A residual plot is plotting residuals against the explanatory variable. Essentially, it turns the regression line horizontal.

In order to draw a residual plot, you must first perform a LinReg. Next, create a StatPlot where XList is L1 and YList is RESID (From 2nd $\rightarrow$ Stat -¿ :(this one depends on the calculator))

1.4 Influential Points and Departure from Linearity

The standard deviation of the residuals gives the approximate size of a "typical" prediction error. Large values means our line is expected to give larger residuals.

An influential point in a data set, is a point that has leverage on the correlation and regression line.

If your scatterplot when graphing data does not show a linear pattern, or the residual plot pattern is not random, consider transforming the graph.

Some of the most common patterns involve transforming the x or y variables by the natural log or a square root.