Correlation and Regression

Pearson Correlation

In summarizing the relationship between two variables, the most common statistic is Pearson's correlation coefficient r. As we've discussed, associations can be qualitatively summarized by their *direction* (positive vs negative), *form* (linear vs nonlinear), and *strength* (strong vs weak). The Pearson correlation varies between -1 and 1 and naturally captures the direction and strength of a linear association.

The Pearson correlation is estimated by

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Breaking down the equation we can separate it into two components - the numerator estimates the covariability, the extent to which X and Y tend to vary together

$$SP = \sum(X - \bar{X})(Y - \bar{Y})$$

The demoniator, uses the sums of squares for X and Y, respectively

$$SS_X = \sum(X - \bar{X})^2 \qquad SS_Y = \sum(Y - \bar{Y})^2$$

Based on this breakdown, we see that the correlation coefficient is a ratio comparing the covarability to the individual variabilities of X and Y...

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

As with the other hypothesis tests, a common step after computing the test statistic is to ask whether or not the association is statistically significant. That is, we want to test the null hypothesis that the population correlation $\rho = 0$. Describing the null distribution exactly is a bit outside the scope of this course, but keep in mind that, as with the t-tests and ANOVA, it's shape depends on a degrees of freedom $df = n - 2$.

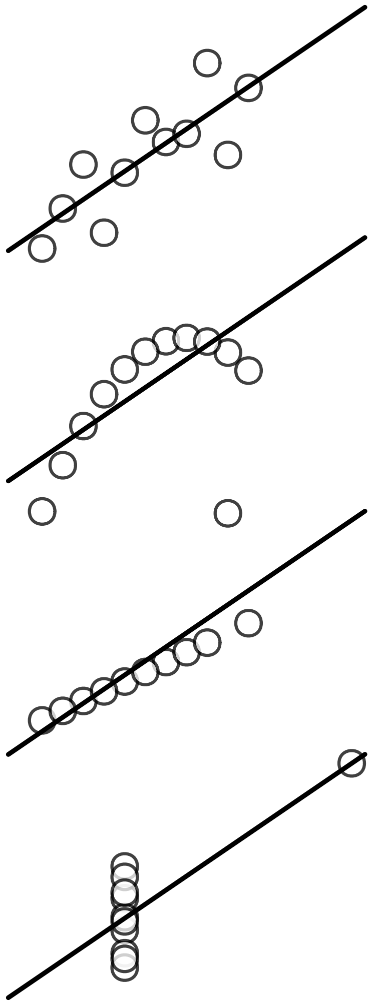

Figure 1: Anscombe's Quartet. Four datasets with the same $\bar{X}$, $\bar{Y}$,s_X,s_Y, and correlation coefficient. We often assume that our data is distributed similarly to the top figure, but non-linear trends, and outliers can dramatically affect the interpretation of data.

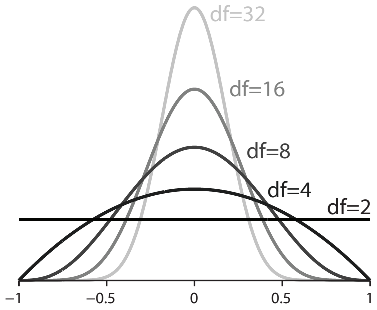

Figure 2: Null distributions for Pearson correlation with different degrees of freedom df.

Spearman Correlation

When the relationship between X and Y is nonlinear or when there are outliers (e.g. the last figure in Anscombe's quartet), Pearson correlation doesn't provide an accurate description of the association. For these trends, a common alternative to the Pearson correlation is the Spearman rank correlation. To compute Spearmann correlation we first convert the X and Y scores to *ranks* (corresponding to the order in a sorted list). The dataset $X : 5, 2, 10, 3$ for instance, would have ranks $x : 3, 1, 4, 2$. Identical scores are assigned an average rank, e.g. $X : 1, 2, 2 \rightarrow x : 1, 2.5, 2.5$. Once we have the ranks, then find differences between ranks in X and Y, $d_i = x_i - y_i$ and we can calculate

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Note that if the X ranks and Y ranks are exactly the same, then $d_i = 0$ for all i and $\rho = 1$. In general there will be some differences between the ranks, but, as with the Pearson corrleation, the Spearman correlation is between -1 and 1 with larger absolute values denoting stronger associations and the same distinction between positive and negative associations.

Linear Regression

Although, correlation does not imply causation, an important feature of correlations is that they allow one variable to be predicted from the other. With only X and Y we can make this prediction concrete by finding a "line of best fit." This will allows us to take a new value of X and predict what the score for Y should be. Recall that a linear equation in one variable can take the form ...

$$\hat{Y} = mX + b$$

where m is the slope of the line and b defines the y-intercept. It turns out that after calculating the Pearson correlation coefficient the line of best fit is given by

$$m = r\frac{s_x}{s_y} \qquad b = \bar{Y} - m\bar{X}$$

We write $\hat{Y}$ (pronounced "y-hat") rather than just Y to denote the fact that this equation is an estimate for the value of Y. When $-1 < r < 1$ there will certainly be some errors. The reason that this these equations are the "best fit" is that these errors are minimized.

Regression is a powerful tool in psychology and many other fields. We will not spend much time on it here, but if you continue in statistics you will learn about models with multiple dimensions (e.g. estimating SAT scores based on high-school GPA, socio-economic status, other predictors) and how to model nonlinear relationships between variables.

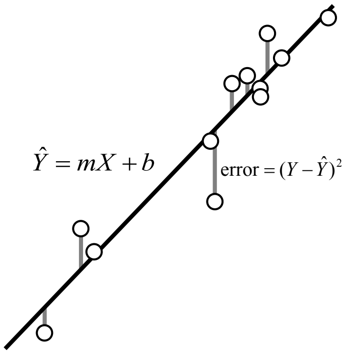

Figure 3: Line of Best Fit. The line of best fit minimizes errors between observed Y and predicted $\hat{Y}$

Example

Solution