Module 3: Association: Contingency, Correlation, Regression

Often we are interested in whether or not two variables are related.

- Do smokers have shorter life-spans than non-smokers?
- Do highschools with higher per-student funding tend to have higher mean SAT scores for their students?

Response and Explanatory Variables

The **response variable** is the outcome variable on which comparisons are made. The values of the response variable *depend on* or can be *explained* by the **explanatory** variable.

A categorical explanatory variable will define groups to be compared with respect to the response variable. When an explanatory variable is quantitative, we will look at whether increasing or decreasing the explanatory variable is associated with changes in the response variable.

Be careful: Just because a variable is an explanatory variable this does not mean that this variable *causes* the response variable to change.

Example

For each example, identify the response variable and the explanatory variable. Are these variables categorical or quantitative

- Over one thousand women were followed for 20 years. Were non-smokers more likely than smokers to be alive at the end of the study? Response variable: survival status (alive or deceased); explanatory variable: smoking status (yes or no). Both variables are categorical.
- Do highschools with higher per-student funding tend to have higher mean SAT scores for their students? Response variable: SAT score (quantitative); explanatory variable: funding level (quantitative)
- Do females study more hours per week than males? Response variable: studying (hours); explanatory variable: gender. Studying is quantitative while gender is categorical.

Association between two variables

An **association** exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.

For example, it is more likely that someone with a high school GPA of 4.0 will have a college GPA above 3.5 than a student who has a high school GPA of 3.0. Therefore, there is an *association* between high school GPA and college GPA.

The analysis we use to identify the association between two variables depends on the types of variables we are looking at.

- If one variable is *quantitative* (such as height) and the other variable is *categorical* then we can compare the categories (such as females vs. males) using summaries of center and variability and graphics such as side-by-side box plots.
- If both variables are *categorical*, then we construct a contingency table and compare conditional proportions.
- If both variables are *quantitative*, we construct scatterplots and calculate the correlation.

Contingency table

A **contingency table** is a display for two categorical variables. Its rows list the categories of one variable and its columns list the categories of the other variable. Each entry in the table is the number of observations in the sample at a particular combination of categories of the two categorical variables.

Each row and column combination in a contingency table is called a **cell**. The process of taking a data file and finding the frequencies for each cell in a contingency table is called **cross-tabulation** of the data.

Example

A study by the USDA and the state of California looked at the association between pesticide status and food type for organically grown and conventionally grown foods, which generated the following data.

	Pesticide Status		
Food type	**Present**	**Not Present**	**Total**
Organic	29	98	127
Conventional	19,485	7,086	26,571
Total	**19,514**	**7,184**	**26,698**

- How many organic foods had pesticides present? 29
- How many conventional foods had pesticides present? 19,485
- Is there an association between food type and pesticides status? In order to answer this question we have to look at the *conditional proportions*.

Conditional proportion

A **conditional proportion** is the proportion of observations that occur for each possible value of a variable (typically the explanatory variable).

	Pesticide Status			
Food type	**Present**	**Not Present**	**Total**	**n**
Organic	0.23	0.77	1.00	127
Conventional	0.73	0.27	1.00	26,571

Let's construct the contingency table and plot the conditional proportions in *R*

It should be clear that there is an *association* between food type and pesticide status because the proportion of food items with pesticides differs greatly between the two groups (.23 for organic vs. .73 for conventional).

There would be *no association* if there was no difference between groups. For example, if in each group 60% of foods had pesticides present and 40% of foods did not, then we would conclude that pesticide status is *independent* of food type.

Note that we are *not* comparing the conditional proportions across each category of the response variable. Instead we must compare the conditional proportions between each category of the *explanatory* variable.
Also note that we must compare *conditional* proportions, as illustrated by the following example.

Example

AT&T used to claim to be the network with the fewest dropped calls...so?

Carrier	**# dropped calls (per min)**
AT&T	10
Verizon	15

Is there a relationship between carrier and # of dropped calls?
Can we calculate the conditional proportions?

Let's look at some data in *R*. We want to see if there is an association between Internet Penetration (or Internet Use) and Facebook Penetration. First, let's look at these variables one at a time.

- What is the mean percentage of Internet and Facebook penetration?
- Describe the shape of the distributions for Internet and Facebook penetration?
- Twenty-five percent of the countries surveyed have Internet penetration rates above what percent?

Now let's look at the relationship between Internet and Facebook Penetration.

Scatterplot

A **scatterplot** is a graphical display for two quantitative variables using the horizontal (x) axis for the explanatory variable x and the vertical (y) axis for the response variable y. For each subject, a point is drawn on the scatterplot corresponding to its (x, y) value.

Let's use *R* to construct the scatterplot and answer a few questions:

- Can you find the point corresponding to the United States?
- Is there a trend in the data? Are there any potential outliers?

Positive and negative associations

Two quantitative variables have a **positive association** if high values of x tend to occur with high values of y, and low values of x tend to occur with low values of y. As x increases, y tends to increase.

Two quantitative variables have a **negative association** if high values of one variable tend to pair with low values of the other. As x increases, y tends to decrease.

Questions to consider

- Is there an association? If so, is it positive or negative? Is it linear?
- Are there any potential outliers (unusual observations). Are these informative?

Correlation

The **correlation** summarizes the direction of the association between two quantitative variables and the strength of its linear (straight line) relationship. The correlation, r, can take on values between -1 and $+1$.

- A positive value for r indicates a positive association and a negative value for r indicates a negative association.
- The closer r is to ± 1 the closer the data points fall to a straight line, and the stronger the linear relationship is. The closer r is to 0, the weaker the linear association. If $r = 0$, there is no *linear* association.

The correlation is calculated using the formula

$$r = \frac{1}{n-1} \sum z_x z_y$$

where z_x is the z-score for each x observation and z_y is the z-score for each y observation. Recall that

$$z = \frac{\text{observation - mean}}{\text{standard deviation}}.$$

Each observation contributes to the correlation value in the following way.

z_x	z_y	contribution $(z_x z_y)$
$+$	$+$	$+$
$+$	$-$	$-$
$-$	$+$	$-$
$-$	$-$	$+$

Properties of correlation

- The correlation r is always between -1 and $+1$. The closer r is to 1 (positive or negative), the stronger the linear relationship and the closer the points are to falling on a straight line.
- A positive correlation indicates a positive association; a negative correlation indicates a negative association
- The value of the correlation does not depend on the units used (e.g., weight in pounds vs. weight in kilograms)

Properties of correlation (continued)

- The correlation between two variables is the same regardless of the order of the variables (e.g., which is the response variable and which is the explanatory variable).
- The correlation is meaningless if the relationship between variables is not linear. A correlation of $r = 0$ does not mean there is no relationship between the variables, just that there is no *linear* relationship. Always plot the data first to assess whether the correlation is appropriate.

In addition to constructing a scatterplot and calculating the correlation, we are often interested in finding an equation that predicts the variable of the response variable from the value of an explanatory variable

Regression Line: An Equation for Predicting the Response Outcome

The **regression line** predicts the value for the response variable y as a linear (straight-line) function of the explanatory variable x. We use $\hat{y}$ to denote the **predicted** value of y. The equation for the regression line has the form

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

where a denotes the **y-intercept** and b denotes the **slope**

Interpretation of the $y-$intercept and slope

- $y-$intercept: this is the predicted value of y when $x = 0$. Note that this may or may not make sense in the context of the problem.
- slope: The slope represents the change in y for any 1-unit increase in x

Example

Anthropologists can predict how tall an individual was based on femur (thighbone) lengths. The regression line has the equation $\hat{y} = 61.4 + 2.4x$, where $\hat{y}$ is the predicted height and x is the length of the femur, both in centimeters.

- What is the predicted height of an individual with a femur length of 50 cm? $\hat{y} = 61.4 + 2.4(50) = 181.4$.
- Identify and interpret the $y-$intercept. The $y-$intercept is a in the equation $\hat{y} = a + bx$. In this example, $a = 61.4$, which means for femurs that are 0 cm, the predicted height would be 61.4. However, this does not make sense in the context of this problem because a femur with length 0 cm would mean that the femur does not exist
- Identify and interpret the slope. The slope is b in the equation $\hat{y} = a + bx$. In this example, $b = 2.4$. This means that each 1 cm increase in the length of a femur is associated with a 2.4 cm increase in height of the individual

Finding the regression line

A **residual** is the difference between an observation and its predicted value: $y - \hat{y}$.

A regression line is found by minimizing the sum of squared residuals, i.e., by minimizing $\sum (y - \hat{y})^2$. For this reason the method is sometimes called the **least squares method**.

Properties of the least squares regression line

- The sum of the residuals is zero.
- The line always passes through the point $(\bar{x}, \bar{y})$.
- The slope is equal to $r \frac{s_y}{s_x}$.
  Why doesn't the slope tell us about the *strength* of the association between two variables?

How well does the regression line fit the data?

The **coefficient of determination**, often denoted R^2, is equal to r^2, where r is the correlation.

R^2 is the *proportion of the variation in y that is accounted for by the linear relationship of y with x.*

Extrapolation is dangerous

As mentioned previously, **extrapolation** involves using a regression line to predict y values for x values outside the observed range of data.

Example: Forecasting Climate Change

We will use *R* to look at annual mean temperature from 1869-2010 for Central Park, New York City.

- Find the regression line.
- Identify and interpret the slope.
- Predict the annual mean temperature in the year 1999.
- Predict the annual mean temperature in the year 3000.

Be cautious of influential outliers

A **regression outlier** is an observation that does not follow the trend of the rest of the data. A regression outlier may or may not (and is usually not) an outlier in its x or y value alone.

An **influential** observation has a large effect on the results of a regression (or correlation) analysis. An observation will be *influential* if

- Its x value is relatively low or high compared to the rest of the data
- The observation is a regression outlier

Correlation Does Not Imply Causation

Correlation (or association) between two variables never implies causation.

There are three possibilities if an association is found between variables x and y.

- There is a cause and effect relationship betwen x and y. NEVER ASSUME THIS unless the data comes from an experimental study
- A **lurking variable** may be influencing the association between x and y. **Confounding** occurs when two explanatory variables that are associated with a response variable are also associated with one another.
- It may appear from the study that the two variables are associated, even though they are not (i.e., the association is due to *chance*). More on this later in the semester...