1 Correlation and Causality

1.1 Seeking Correlation

Definition 1.1: A **correlation** exists between two variables when higher values of one variable consistently go with higher values of another variable, or when higher values of one variable consistently go with lower values of another variable.
Ex:

- There is a correlation between the variables "height" and "weight" for people. That is, taller people tend to weigh more than shorter.
- There is a correlation between the variables "demand for apples" and "price of apples". That is demand tends to decrease as prices increase.
- There is a correlation between "practice time" and ""skill among piano players. That is, those who practice more tend to be more skilled.

Definition 1.2: A **scatterplot** is a graph in which each point represents the value of two variables.
Ex:

1. Create a scatter plot for the following data.

Movie	Production cost in millions	U.S. Gross Receipts in millions
John Carter	300	66
Pirates of the Caribbean: At Worlds End	300	309
The Hobbit: An Unexpected Journey	300	305
Tangled	270	201
Spiderman 3	260	337
Harry Potter and the Half Blood Prince	258	302
Pirates of the Caribbean: On Stranger Tides	250	241
The Dark Knight Rises	250	448
Avatar	237	761
Superman Returns	232	200
Quantum of Solace	230	169
The Chronicles of Narnia: Price Caspian	225	142
Pirates of the Caribbean: Dead Man's Chest	225	423
Marvel's The Avengers	220	623
Men in Black 3	215	180

Definition 1.3: **Positive correlation:** Both variables tend to increase (or decrease) together.

Definition 1.4: **Negative Correlation:** The two variables tend to change in opposite directions, with one increasing while the other decreases.

Definition 1.5: **No correlation:** there is no apparent relationship between the two variable.

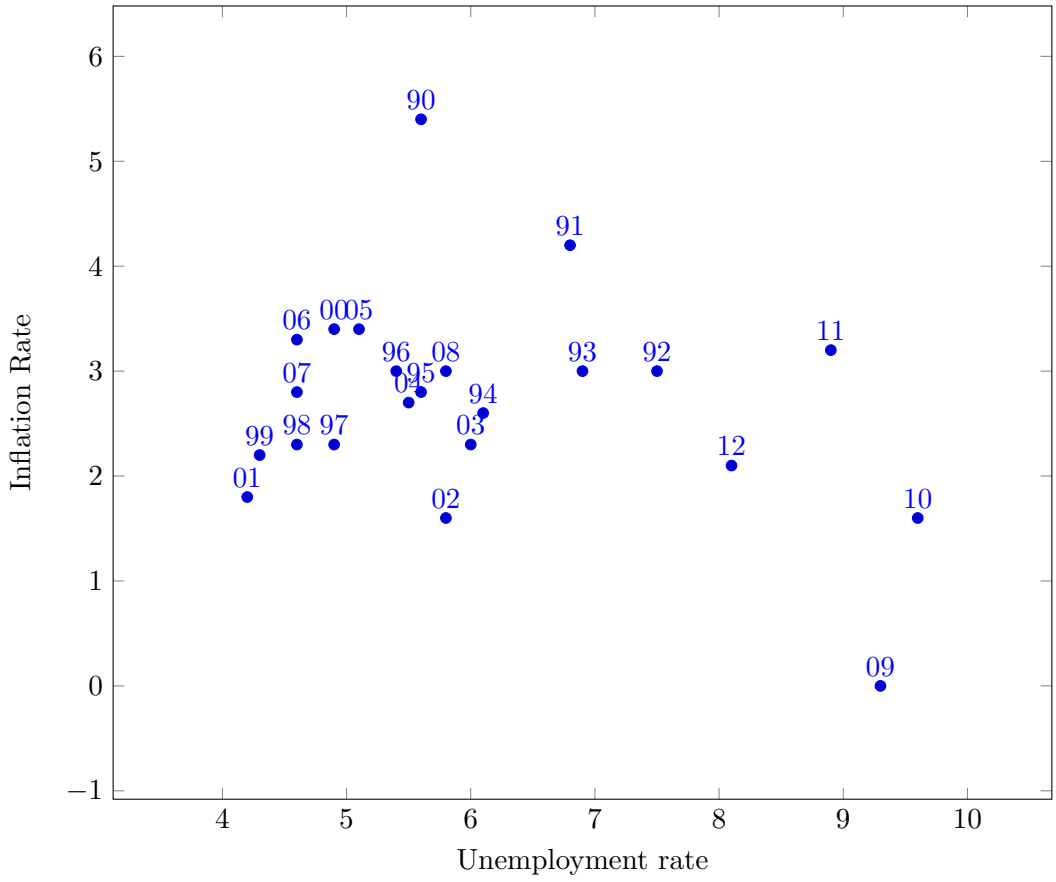
Definition 1.6: **Strength of a correlation:** The more closely two variables follow the general trend, the stronger the correlation (which may be either positive or negative). In a perfect correlation, all data points lie on a straight line.

Year	Unemployment rate	Inflation rate	Year	Unemployment rate	Inflation rate
1990	5.6	5.4	2002	5.8	1.6
1991	6.8	4.2	2003	6.0	2.3
1992	7.5	3.0	2004	5.5	2.7
1993	6.9	3.0	2005	5.1	3.4
1994	6.1	2.6	2006	4.6	3.3
1995	5.6	2.8	2007	4.6	2.8
1996	5.4	3.0	2008	5.8	3.0
1997	4.9	2.3	2009	9.3	0*
1998	4.6	2.3	2010	9.6	1.6
1999	4.3	2.2	2011	8.9	3.2
2000	4.0	3.4	2012	8.1	2.1
2001	4.2	1.8			

Table 1: *The 2009 inflation rate was actually negative but we have set it to zero here.

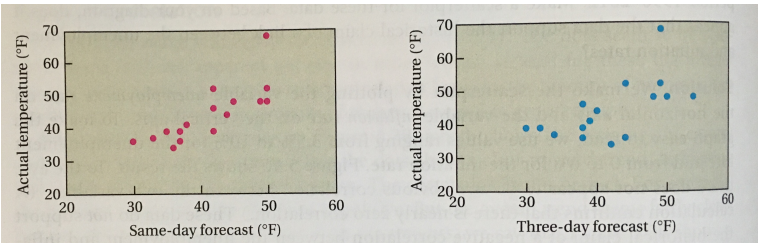
2. Prior to the 1990s, most economists assumed that the unemployment rate and the inflation rate were negatively correlated. That is, when unemployment goes down, inflation goes up, and vice versa. The table above shows unemployed and inflation data for the period 1990-2012. Make a scatterplot for these data. Based on your diagram, does it appear that the data support the historical claim of a link between the unemployment and inflation rates?

Solution. The following is the scatterplot



There does not appear to be any obvious correlation between unemployment and inflation rates. □

3. The scatterplots below show two weeks of data comparing the actual high temperature for the day with the same forecast and the three day forecast. Discuss the types of correlation on each diagram.



Solution. Both scatter plots show that as the predicted temperatures goes up, so does the actual temperature. Both of these scatter plots show positive correlation. Notice that the left hand graph's dots form a straighter line than the right hand graph. This implies that the left hand graph has a stronger positive correlation. □

Possible Explanations for a Correlation

- The correlation may be a coincidence.
- Both variables might be directly influenced by some common underlying cause.
- One of the correlated variables may actually be a cause of the other. Not the that even in the case, it may be only one of several causes.

4. Every financial advisor has a strategy for predicting the direction of the stock market. Most focus on fundamental economic data, such as interest rates and corporate profits. But an alternative strategy relies on a remarkable correlation between the Super Bowl winner in January and the direction of the stock market for the rest of the year: The stock market tends to rise when a team from the old, pre-1970 NFL wins the super Bowl, and tends to fall otherwise. This correlation successfully matched 28 of the first 32 SuperBowls to the stock market, which made the "Super Bowl indicator" a far more reliable predict of the stock market than any professional stock broker during the same period. Suppose that the Super Bowl just ended and the winner was the Detroit Lions, and old NFL team. Should you invest all your spare cash in the stock market?

Solution. Even though the times that the correlation was successful 28 out of 32 times does not mean that there is a causality. This is a case of coincidence. There is no reason to say that Super Bowl outcomes have any affect on the stock market. □

1.2 Establishing Causality

Guidelines for Establishing Causality: If you suspect that a particular variable is causing some effect,

- Look for situations in which the effect is correlated with the suspected cause even while other factors vary.
- Among groups that differ only in the presence or absence of the suspected cause, check the effect is similarly present or absent.
- Look of evidence that larger amounts of the suspected cause, produce larger amounts of the effect.
- If the effect might be produced by other potential causes, make sure that the effect still remains after accounting for these other potential causes.

- If possible, test the suspected cause with an experiment. If the experiment cannot be performed with humans for ethical reasons, consider doing the experiment with animals, cell cultures, or computer models.
- Try to determine the physical mechanism by which the suspected cause produces the effect.

Broad Levels of Confidence in Causality

- **Possible cause:** We have discovered a correlation, but cannot determine whether the correlation implies causality. In the legal system, possible cause (such as thinking that particular suspect possible caused a particular crime) is often the reason for starting an investigation.
- **Probable cause:** We have good reason to suspect that the correlation involves cause, perhaps because some of the guidelines for establishing causality are satisfied. In the legal system, probable cause is the general standard for getting a judge to grant a warrant for a search or wiretap.
- **Cause beyond reasonable doubt:** We have found a physical model that is so successful in explaining how one thing causes another that is seems unreasonable to doubt the causality. In the legal system, cause beyond reasonable doubt is the usual standard for conviction. It generally demand that the prosecution show how and why the suspect committed the crime. Note: Beyond reasonable doubt does not bean beyond all doubt.

Ex:

1. Based on what you know about global warming, do you think that human activity is a possible cause, probable cause, or cause beyond reasonable doubt? Defend your opinion.

   Definition 1.7: **glob·al warm·ing**
   noun
   A gradual increase in the overall temperature of the earth's atmosphere generally attributed to the greenhouse effect caused by increased levels of carbon dioxide, chlorofluorocarbons, and other pollutants.

   Solution. (a) We know that over the years the amount of carbon dioxide, chlorofluorocarbons, and other pollutants in the air has gone up. We also have that the temperature of earth's atmosphere has gone up.

   (b) We know that before the presence of humans (or less humans) the amount of green house gases in the atmosphere was less. The temperature was also lower.

   (c) We can see through out history, the more people there are the more green house gases which cause global warming.

   □