Lecture 1 & 2 Notes, pt. 2

Young Researchers Fellowship Program

Daniel Sánchez Pazmiño

September 2024

Table of contents

Session 4: Distributions and Correlation

Measures of Distributional Shape

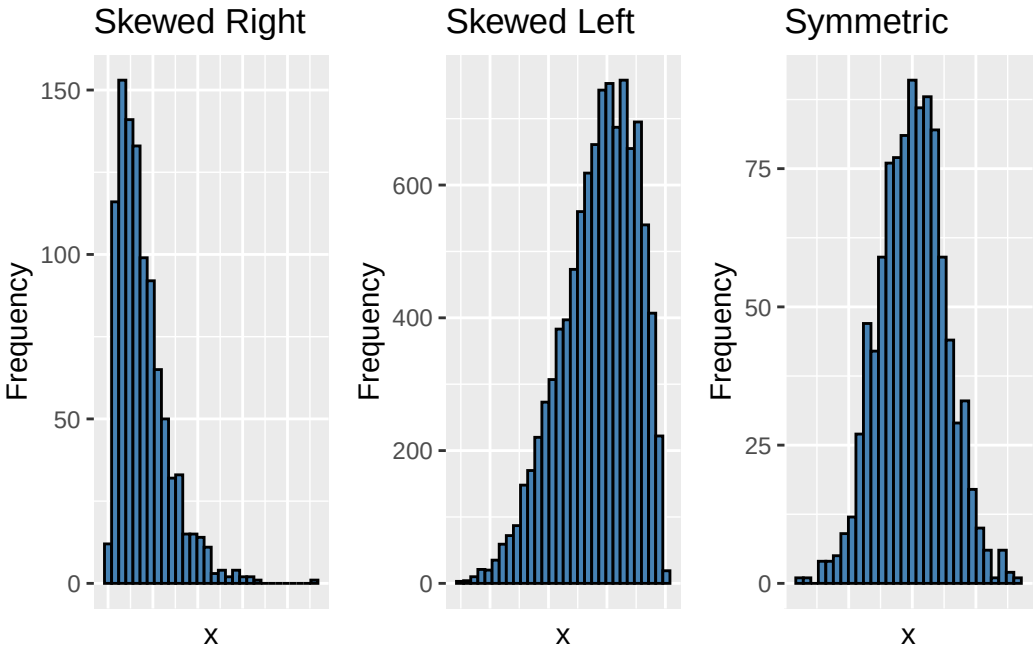
Before, we talked about how it is important to know the central tendency and the variability of the set of data that we're trying to understand. However, there is one extra thing that we should try to understand, which is the shape of the data, or the distribution of the data. The distribution means how the data is spread out. For instance, if we have a set of data that is normally distributed, it means that the data is spread out symmetrically around the mean. If we have a set of data that is skewed to the right, it means that the data is spread out asymmetrically to the right of the mean. If we have a set of data that is skewed to the left, it means that the data is spread out asymmetrically to the left of the mean.

A histogram tells us if the data is symmetric or not. However, we may want to characterise this by a numerical measure. There are two measures of distributional shape that we will talk about here: skewness and kurtosis. Skewness measures the asymmetry of the data, while kurtosis measures the thickness of the tails of the data.

Skewness

Skewness measures the symmetry of the data. When a dataset is symmetric, you can draw a line in the center and find the same amount of observations on both sides of the line. When a dataset is not symmetric, you can draw a line in the center and find a different amount of observations on both sides of the line. The skewness measures how different the amount of observations on both sides of the line is.

For a dataset which is symmetric, we see a histogram with higher bars nearing the center. For a dataset which is not symmetric, we see a histogram with higher bars on one side of the center than on the other side of the center.



There is a formula to calculate the skew, which is the following:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

where n is the number of observations, $\bar{x}$ is the sample mean, and s is the sample standard deviation. The numerator is the average of the cubed distance from the mean, and the de-

nominator is the cube of the standard deviation. However, it is unlikely that you ever need to calculate the skewness by hand, as it is a built-in function in most statistical software.

What is important to know is that if the data is symmetric, the skewness is zero. If the data is skewed to the right, the skewness is positive. If the data is skewed to the left, the skewness is negative. Being skewed to the right is also called being *positively skewed*, while being skewed to the left is also called being *negatively skewed*. The way that a histogram looks when the data is skewed to the right is that the right side of the graph holds a bit less data than the left side of the graph. The way that a histogram looks when the data is skewed to the left is that the left side of the graph holds a bit less data than the right side of the graph.

Kurtosis

Kurtosis measures the thickness of the tails of the data. When a dataset has thin tails, it means that the data is concentrated around the mean. When a dataset has thick tails, it means that the data is spread out away from the mean. The kurtosis measures how thick the tails of the data are.

For a dataset which has thin tails, we see a histogram with lower bars on the sides and higher bars in the center. For a dataset which has thick tails, we see a histogram with higher bars on the sides and lower bars in the center.

The kurtosis is calculated as follows:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

where n is the number of observations, $\bar{x}$ is the sample mean, and s is the sample standard deviation. The numerator is the average of the fourth power of the distance from the mean, and the denominator is the fourth power of the standard deviation. However, it is unlikely that you ever need to calculate the kurtosis by hand, as it is a built-in function in most statistical software.

Kurtosis is often used as a means to test whether or not the distribution is normally distributed. If the kurtosis is close to zero, it means that the distribution is close to normal. If the kurtosis is greater than zero, it means that the distribution has thicker tails than the normal distribution. If the kurtosis is less than zero, it means that the distribution has thinner tails than the normal distribution. We will talk later about what the normal distribution is, but for now, it is enough to know that it is a symmetric distribution with thin tails, like the one in the symmetric data histogram above.

Relative location

There are also measures which can be calculated for any particular data point in order for us to tell how far they are from the center of the data. These measures are called relative location measures. The most common relative location measures are the z-score and the percentile. We already know what percentiles are, but it is useful to review how to calculate them. The z-score is a measure of how many standard deviations away from the mean a particular data point is.

Percentiles

Percentiles are a measure of relative location. They tell us what percentage of the data is below a particular data point. As you know, if a data point is at the 75th percentile, it means that 75% of the data is below that data point. If a data point is at the 50th percentile, it means that 50% of the data is below that data point. If a data point is at the 25th percentile, it means that 25% of the data is below that data point. If a data point is at the 100th percentile, it means that 100% of the data is below that data point, which means that it is the maximum value in the dataset. If a data point is at the 0th percentile, it means that 0% of the data is below that data point, which means that it is the minimum value in the dataset.

If you are asked to calculate the ninth percentile of a dataset with 350 observations on the variable of interest, you need to find the observation for which 9% of the data is below. We need to know it in terms of the rankings of the dataset, so we can use the following formula to calculate the *location* of any percentile p:

$$\frac{p}{100} \times (n+1)$$

where p is the percentile, and n is the number of observations in the dataset. For example, if we want to calculate the location of the 9th percentile in a dataset with 350 observations, we would use the following formula:

$$\frac{9}{100} \times (350+1) = 31.59$$

This means that the 9th percentile is the 32nd observation in the dataset. After ranking the dataset, you have to find the 32nd observation from smallest to largest yourself.

You may also need to calculate the percentile of any given point on the dataset. For instance, you are given a dataset of 15 observations like the one below:

Table 1: Calculating percentiles by hand

x	rank
5.57	1
8.33	2
8.36	3
8.75	4
8.76	5
9.39	6
10.37	7
10.66	8
10.78	9
10.97	10
11.15	11
11.48	12
12.25	13
13.02	14
13.19	15

If you want to calculate the percentile of the observation $x = 9.39$, you will need to order the dataset from smallest to largest, and then use the following formula:

$$\frac{rank - 1}{n - 1} \times 100$$

where $rank$ is the rank of the observation, and n is the number of observations in the dataset. For example, if we want to calculate the percentile of the observation $x = 9.39$, we would use the following formula:

$$\frac{6 - 1}{15 - 1} \times 100 = 42.86$$

This means that 42.86% of the data is below the observation $x = 9.39$.

Z-scores

The z-score is a measure of relative location. It tells us how many standard deviations away from the mean a particular data point is. The formula for calculating the z-score is as follows:

$$z = \frac{x - \bar{x}}{s}$$

where x is the data point, $\bar{x}$ is the sample mean, and s is the sample standard deviation. For example, if we want to calculate the z-score of the observation $x = 9.39$ in the dataset above, we would use the following formula:

$$z = \frac{9.39 - 10.20}{2.04} = -0.4$$

This means that the observation $x = 9.39$ is 0.4 standard deviations *below* the mean. If the z score is positive, it means that the observation is above the mean. If the z score is negative, it means that the observation is below the mean. If the z score is zero, it means that the observation is equal to the mean.

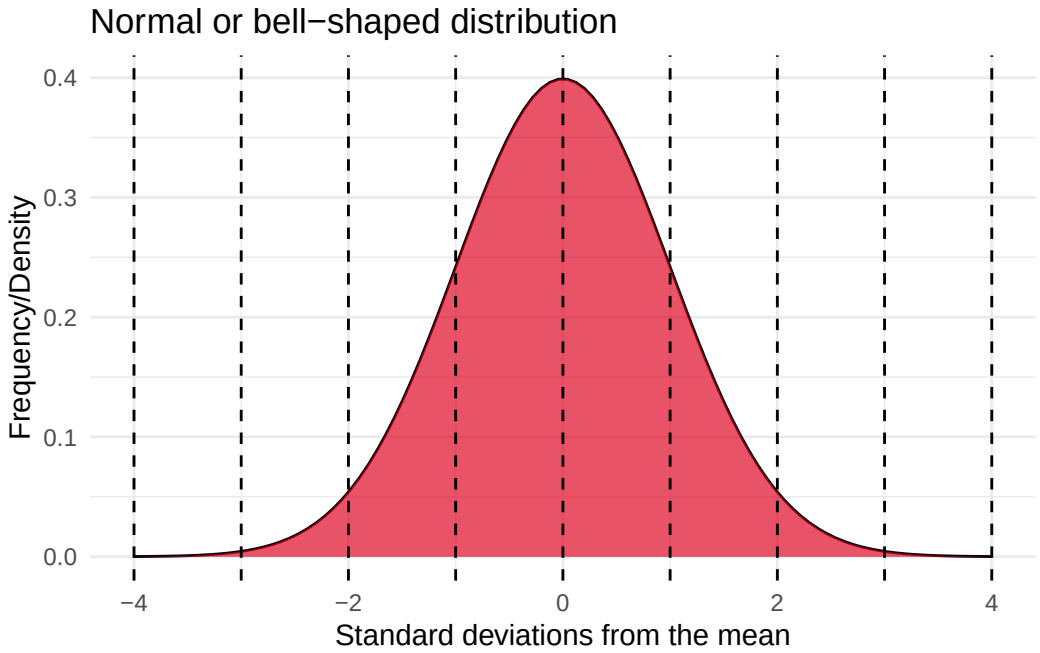
Z-scores are also called standardized values. They are useful because they allow us to compare observations from different datasets. For example, if we have two datasets, one with a mean of 10 and a standard deviation of 2, and another with a mean of 20 and a standard deviation of 5, we can compare the observations from both datasets by calculating their z-scores. It kind of means that we are "evening out" the datasets so that we can compare them.

Outliers, which are observations that are very far away from the rest of the data, can be identified by looking at their z-scores. If an observation has a z-score that is greater than 3 or less than -3, it is considered an outlier.

The Chebyshev Theorem

The Chebyshev Theorem is a theorem which tells us how many observations are within a certain number of standard deviations from the mean. It is useful because it allows us to make statements about the data without knowing anything about the distribution of the data. The Chebyshev Theorem states that at least $1-\frac{1}{k^2}$ of the data is within k standard deviations from the mean. For example, if we want to know how many observations are within 2 standard deviations from the mean, we can use the Chebyshev Theorem to calculate that at least $1 - \frac{1}{2^2} = 0.75$ of the data is within 2 standard deviations from the mean. This means that at least 75% of the data is within 2 standard deviations from the mean. This is very useful, because we don't need to know the data distribution to make this statement.

When the dataset is normally distributed, meaning that its histogram follows a bell-shaped distribution like the one below, the Chebyshev Theorem implications take a very specific form, which is called the empirical rule. You might have heard about it before.



The empirical rules tells us that *if the data is normally distributed*, then: - 68% of the data is within 1 standard deviation from the mean - 95% of the data is within 2 standard deviations from the mean - 99.7% of the data is within 3 standard deviations from the mean

The empirical rule is widely used, however, it is important to remember that not all datasets are bell-shaped. A well known example that we covered before is wages. In any given city, country or continent, wages are not normally distributed. They are skewed to the right, meaning that they have a long tail to the right (more people make little money). This means that the empirical rule does not apply to wages, and we must use other methods to make statements about the data, such as the general Chebyshev Theorem or applying transformations to the data to make it approximate the normal distribution. We will talk about the latter in other lessons.

Correlation

So far we have only talked about ways to analyse data for one variable at a time. However, in the social sciences we are often interested in the relationship between variables. For example, we might want to know if there is a relationship between the number of hours that a student studies and their grades. Or we might want to know if there is a relationship between the number of hours that a student studies and their income after graduation. Or we might want to know if there is a relationship between the number of hours that a student studies and their happiness.

In order to answer these questions, we need to learn how to analyse data for two variables at a time. Covariance and correlation are two measures that we can use to analyse data for two variables at a time, but the concept of correlation in general is more broad than the numerical measure of correlation. In general, correlation means that there is a relationship between two variables. For example, if we say that there is a correlation between the number of hours that a student studies and their grades, we mean that there is a relationship between the number of hours that a student studies and their grades. This relationship may be studied using the numerical measure of correlation, covariation, data visualisation, or other methods, such as regression.

Covariance

Covariance is a measure of how much two variables vary together. Covariance is calculated as the average of the product of the deviations of each variable from its mean. The formula for covariance is:

$$\sigma^2_{x,y} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)$$

where N is the number of observations, x_i is the value of the ith observation of the variable x, μ_x is the mean of the variable x, y_i is the value of the ith observation of the variable y, and μ_y is the mean of the variable y.

It is useful to hand-calculate covariance for a small dataset to understand what it means. As with the variance, we proceed by dividing the covariance formula.

Table 2: Calculating covariance by hand

study_hours	grades	deviations_study_hours	deviations_grades	product_deviations
4.719762	86.67980	-0.3175506	13.326594	-4.2318684
4.884911	74.70667	-0.1524016	1.353469	-0.2062707
5.779354	80.68770	0.7420413	7.334490	5.4424950
5.035254	71.87177	-0.0020586	-1.481440	0.0030497
5.064644	62.05025	0.0273310	-11.302960	-0.3089217
5.857532	101.94889	0.8202197	28.595686	23.4547439
5.230458	78.85051	0.1931453	5.497299	1.0617774
4.367469	36.70556	-0.6698434	-36.647647	24.5481862
4.656574	78.45978	-0.3807392	5.106574	-1.9442730
4.777169	61.57114	-0.2601438	-11.782063	3.0650308

Summing the product deviations gives us 50.88 which we then divide by the number of observations to get the covariance of 10 observations, which is 5.0883949.

Once again, there is a correction that needs to be applied to the covariance formula to make it unbiased if we are working with sample data. The unbiased covariance divides by `n-1` instead of `N`. The formula for the unbiased covariance is:

$$s_{x,y} = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

where n is the number of observations in a sample of a population of size N. x_i is the value of the ith observation of the variable x, $\bar{x}$ is the sample mean of the variable x, y_i is the value of the ith observation of the variable y, and $\bar{y}$ is the sample of the variable y.

If the covariance is positive, then the two variables are positively correlated, meaning that they move in the same direction. If the covariance is negative, then the two variables are negatively correlated, meaning that they move in opposite directions. If the covariance is zero, then the two variables are not correlated, meaning that there is no relationship between the two variables. The number itself does not really mean anything, it is the sign that matters. This is because the number is affected by the units of the variables. For example, if we are looking at the relationship between the number of hours that a student studies and their grades, the covariance will be different if we measure the number of hours in hours, days, or weeks. However, the sign will be the same, meaning that the two variables are either positively or negatively correlated.

It is important to take into consideration that covariance measures the relationship between two variables in a linear fashion only. Specifically, the covariance will be positive if when one variable is above its average, the other is too. The covariance will be negative if when one variable is above its average, the other is below its average. If the relationship is nonlinear, which is something that sometimes appears in real world data, covariance is not useful.

Correlation coefficient

The numerical measure of correlation, also called correlation coefficient, is a measure whose magnitude has an actual use. There are many types of correlation measures, but the most common one is the Pearson correlation coefficient, or the perason product moment correlation coefficient.

The correlation coefficient is a measure of the strength of the linear relationship between two variables. The correlation coefficient is calculated as the covariance divided by the product of the standard deviations of the two variables. The formula for the correlation coefficient is:

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y}$$

where $s_{x,y}$ is the covariance between X and Y, s_x is the standard deviation of X, and s_y is the standard deviation of Y. Notice that if we are calculating the population correlation coefficient, we use $\sigma_{x,y}$, σ_x, and σ_y instead of $s_{x,y}$, s_x, and s_y, as follows:

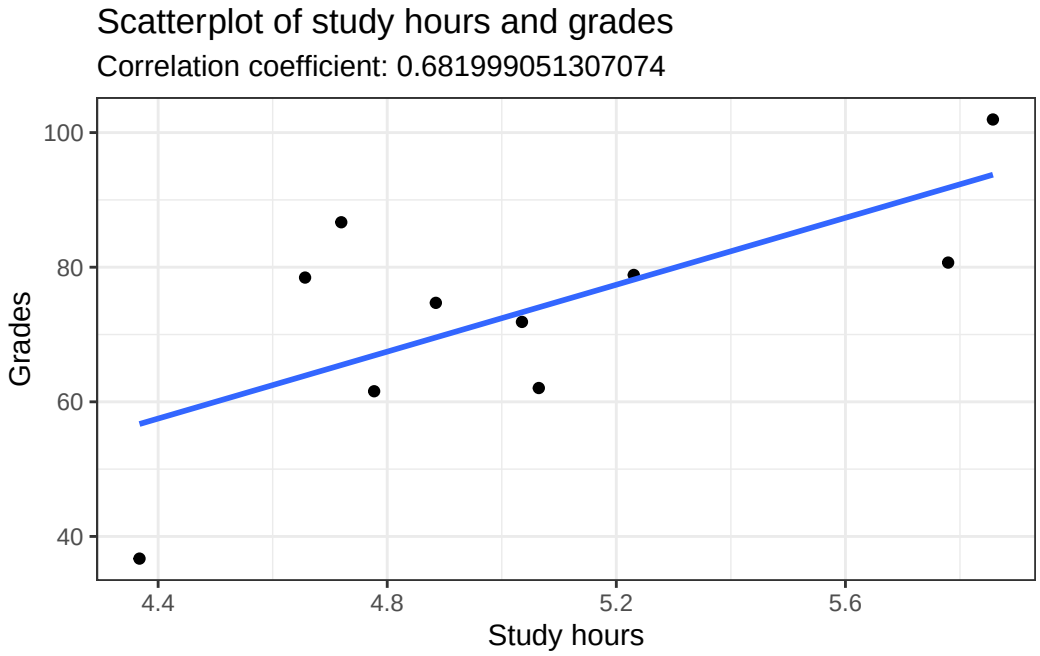
$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \cdot \sigma_y}$$

The correlation coefficient for sample data is denoted as $r_{x,y}$, while the correlation coefficient for population data is denoted as $\rho_{x,y}$, the greek letter *rho*.

How do we interpret the correlation coefficient? The correlation coefficient is a number between -1 and 1. If the correlation coefficient is positive, then the two variables are positively correlated, meaning that they move in the same direction. If the correlation coefficient is negative, then the two variables are negatively correlated, meaning that they move in opposite directions. If the correlation coefficient is zero, then the two variables are not correlated, meaning that there is no relationship between the two variables. The magnitude of the correlation coefficient indicates the strength of the relationship. The closer the correlation coefficient is to 1 or -1, the stronger the relationship. The closer the correlation coefficient is to 0, the weaker the relationship.

For the example above, where we calculated the covariance, we can also calculate the correlation coefficient. The sample covariance for that dataset is 6 and the standard deviations are 0.48 and 17.38. Therefore, the correlation coefficient is 0.6819991. Once again, know that the correlation coefficient is a measure of the strength of the linear relationship between two variables, nothing more.

Often, the strength of the relationship is shown with a scatterplot. The scatterplot below shows the relationship between the number of hours that a student studies and their grades. The correlation coefficient for this dataset is 0.6819991.



In the graph above, each point represents an observation. A blue line is drawn through the points, which is usually called the best-fit line or regression line. We will talk more about it in the regression sessions, but for now, know that the best-fit line is the line that best fits the data. If we square the correlation coefficient, we get a measure called the *coefficient of determination*, or R^2 (read "r squared") which is the proportion of the variance in one variable that is explained by the other variable. This coefficient is calculated as follows:

$$R^2_{x,y} = \frac{s^2_{x,y}}{s^2_x \cdot s^2_y}$$

Its interpretation is as follows: if the coefficient of determination is 0, then the two variables are not correlated. If the coefficient of determination is 1, then the two variables are perfectly correlated. If the coefficient of determination is between 0 and 1, then the two variables are correlated, but not perfectly. The closer the coefficient of determination is to 1, the stronger the relationship. The closer the coefficient of determination is to 0, the weaker the relationship. We do not infer positive or negative relationships from this measure, only the strength of the relationship.

In the example above, the coefficient of determination is 0.465223. This means that 46.52 of the variance in the grades is explained by the number of hours that a student studies.

Causality and validity

Causality or causation means that one variable causes another variable. A causality statement often comes from our own understanding of reality or from what the scientific theory in our field has found. For instance, we can be pretty sure that in any circumstance, studying leads to have better grades, and those students who study more will have better grades. This is a causality statement.

Scientific theory in the natural sciences makes a lot of causality statements based on experimental studies that they can perform. In those, they can control the experimental environment. For instance, a study in biology can control the environment of a plant to see how it grows. They can control the amount of water, the amount of sunlight, the amount of nutrients, and so on. They can define *treatments* and vary the amount of water that they give to the plant, so that they can find out what is the effect of giving water to the plant, *holding everything else constant*. When they perform their data analysis and find that plants that received more water grew more, they can say that water **causes** plants to grow, because they controlled everything else.

However, in the social sciences it is often difficult to perform controlled experiments like the ones in the natural sciences, so we rely a lot on observational studies. In observational studies, however, we cannot control the environmen.

A classical study in the social sciences is the effect of using phones during class. Teachers hypothesise that high school students who use their phones during class will have lower grades. If you were to quietly observe students who used phones during class, and compared grades between those who used phones and those who did not, you would find that those who used phones had lower grades on average.

However, this does not necessarily mean that using phones **causes** lower grades. It could be that those who use phones during class are less interested in the class, and therefore have lower grades because they study less. You might incorrectly conclude that taking away phones in a class will improve the grades of those who previously used phones, but this is not necessarily true. The less interested people will still be less interested, and they will still have lower grades.

So, if you were to actually do this, people would tell you that your conclusion and statistical study is *invalid*. The *validity* of a study is the extent to which the study measures what it is supposed to measure. In the example above, the study is invalid because it does not measure the effect of using phones on grades precisely, as there are things that we cannot control.

Validity, more than a formula or a method, is an issue of critical thinking. So, in exercises where you are asked to evaluate the validity of a study, often you must think about issues regarding causality. The most important principle is that **correlation does not imply causation**. Just because two variables are correlated, it does not mean that one causes the other. Above, we saw that using your phone during class is correlated with lower grades, but it does not

mean that using your phone causes lower grades. We would need more information to make that conclusion.

nother example would be that the number of firefighters that are sent to a fire is correlated with the amount of damage that the fire causes. However, it does not mean that sending more firefighters causes more damage. It could be that the more damage a fire causes, the more firefighters are sent. Sometimes, like the example in this case, we are unsure about the direction of causality.

It is important to keep causality in mind all the time, not just in your statistics course. Politicians, companies and people in general often make statements about causality, when in fact we might just be observing correlation. Correlation is not causation, and when we make decisions based on correlation, we might be making the wrong decisions. Correlation which is not a result of causation is often called *spurious correlation*. You may hear selection bias and omitted variable bias as other reasons why correlation is not causation. We will talk about those in due time.

Summary of Session 4

- The measures of distributional shape are the skewness and kurtosis. Skewness measures the symmetry of the distribution, and kurtosis measures the thickness of the tails of the distribution.
- A positive skew means that the distribution is skewed to the right, and a negative skew means that the distribution is skewed to the left. When the distribution is symmetric, the skew is 0.
- A distribution which is skewed to the right has more values to the left in its histogram. A distribution which is skewed to the left has more values to the right in its histogram.
- Percentiles are values that divide the data into 100 equal parts. To calculate a specific percentile, we need to order the data from smallest to largest, and then find the value that is in the position that corresponds to the percentile. Use the formula given to find which is the ranking that corresponds to the percentile we need to find.
- Z-scores are the number of standard deviations that a value is away from the mean. They are calculated as the difference between the value and the mean, divided by the standard deviation.
- The Chebyshev theorem states that for any distribution, the proportion of values that are within k standard deviations of the mean is at least $1 - 1/k^2$. This means that at least 75% of the values are within 2 standard deviations of the mean, and at least 89% of the values are within 3 standard deviations of the mean.
- The empirical rule states that for a normal distribution, 68% of the values are within 1 standard deviation of the mean, 95% of the values are within 2 standard deviations of the mean, and 99.7% of the values are within 3 standard deviations of the mean. Not all distributions are normal, so this rule does not always apply.

- Covariance is a measure of the strength of the linear relationship between two variables. It is calculated as the sum of the products of the deviations of the two variables from their means, divided by the number of observations.
- The sample covariance is denoted as $s_{x,y}$ and it applies an unbiasedness correction. Instead of dividing by N, it divides by $n-1$, where n is the number of observations in the sample. The population covariance is called $\sigma_{x,y}$ and it divides by N, the total number of observations in the population.
- The correlation coefficient is a measure of the strength of the linear relationship between two variables. It is calculated as the covariance between the two variables divided by the product of the standard deviations of the two variables. The sample correlation, $r_{x,y}$ uses the sample covariance and standard deviations as inputs, and the population correlation $\rho_{x,y}$uses the population covariance and standard deviations as inputs.
- The coefficient of determination is a measure of the proportion of the variance in one variable that is explained by the other variable. It is calculated as the square of the correlation coefficient.
- Correlation does not imply causation. Just because two variables are correlated, it does not mean that one causes the other. The validity of a study is the extent to which the study measures what it is supposed to measure, by taking into account the issue of correlation not implying causation.