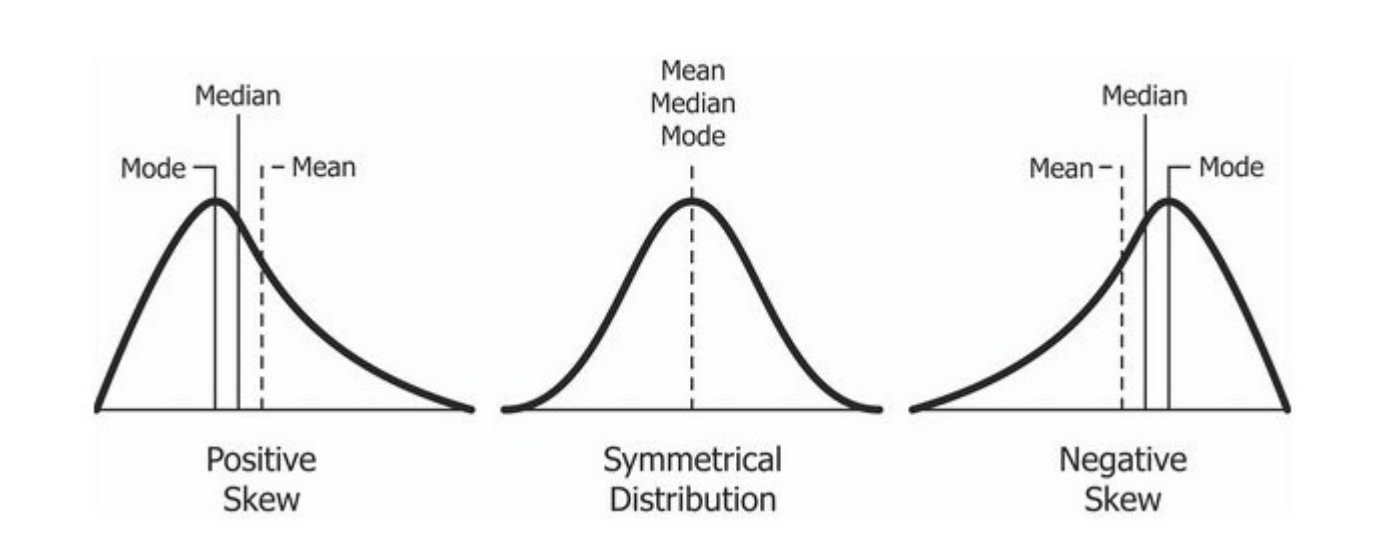
CSE303

Lecture 4: Different Data Distributions

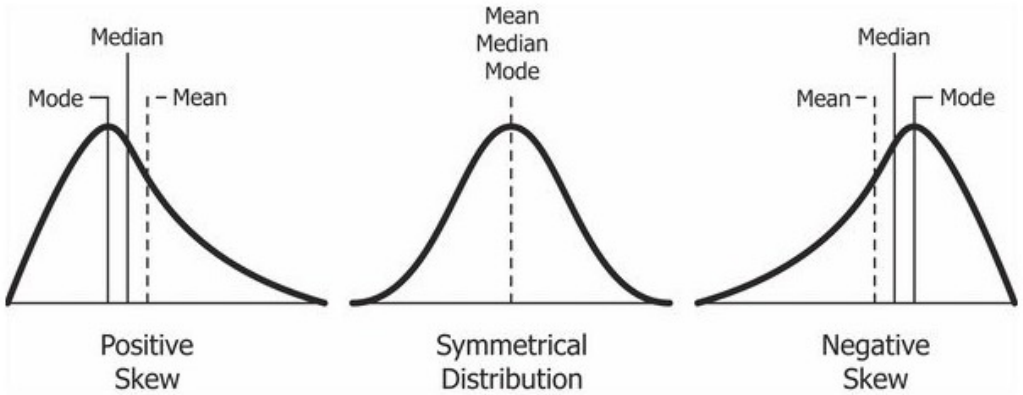
Skewness

- It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution.
- It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.



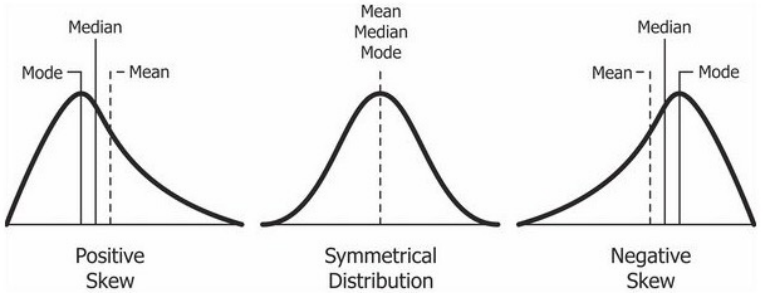
Skewness

- **Positive Skewness** means when the tail on the right side of the distribution is longer. The mean and median will be greater than the mode.
- **Negative Skewness** is when the tail of the left side of the distribution is longer than the tail on the right side. The mean and median will be less than the mode.



Skewness: Example

- Let us take a very common example of house prices. Suppose we have house values ranging from $100k to $1,000k with the average being $500k.
- If the peak of the distribution was left of the average value, portraying a *positive skewness* in the distribution. It would mean that many houses were being sold for less than the average value, i.e. $500k. This could be for many reasons, but we are not going to interpret those reasons here.
- If the peak of the distributed data was right of the average value, that would mean a *negative skew.* This would mean that the houses were being sold for more than the average value.

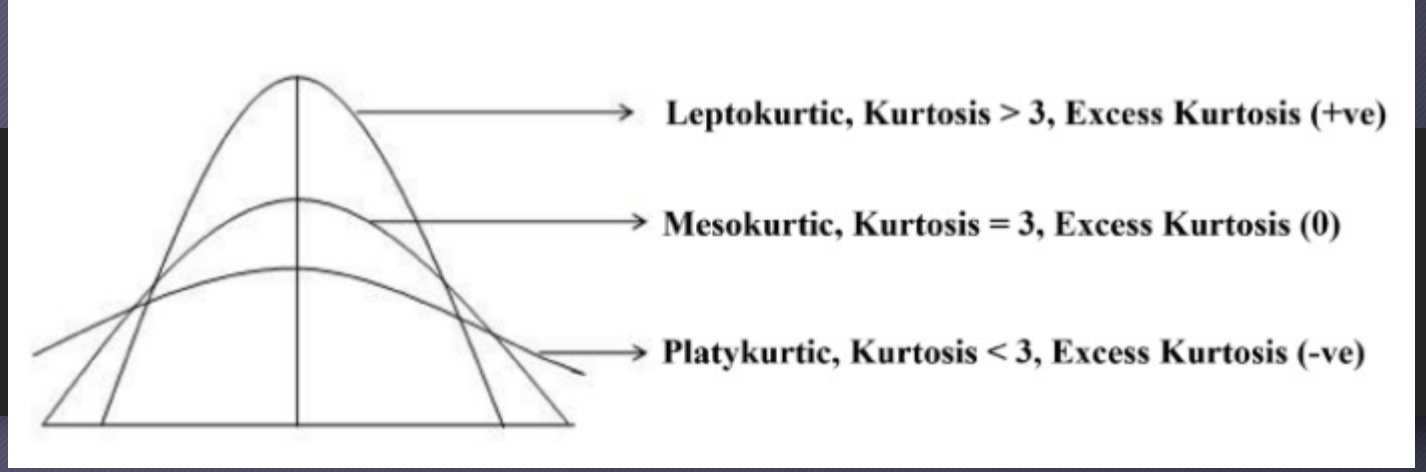


Skewness

So, when is the skewness too much?

- The rule of thumb seems to be:
- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
- If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.

kurtosis



- Like skewness, kurtosis is a statistical measure that is used to describe distribution.
- Whereas skewness differentiates extreme values in one versus the other tail, kurtosis measures extreme values in either tail.
- Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). Distributions with low kurtosis exhibit tail data that are generally less extreme than the tails of the normal distribution.

Correlation Analysis (Nominal Data) / feature selection

- Chi-Square test:
  - The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. This test can also be used to determine whether it correlates to the categorical variables in our data. It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them.

- To know how to apply this, first we need to understand the concepts of Hypothesis test

Hypothesis testing

- Hypothesis testing is a technique for interpreting and drawing inferences about a population based on sample data.

- Null Hypothesis (H0) - The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.

- Alternate Hypothesis(H1 or Ha) - The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. H1 is the symbol for it.

Hypothesis testing

- The two claims needs to be mutually exclusive, meaning only one of them can be true.
- The alternative hypothesis is typically what we are trying to prove.
- The null hypothesis, H0 is the commonly accepted fact; it is the opposite of the alternate hypothesis. Researchers work to reject, nullify or disprove the null hypothesis. Researchers come up with an alternate hypothesis, one that they think explains a phenomenon, and then work to reject the null hypothesis.

How to State the Null Hypothesis from a Word Problem

- You'll be asked to convert a word problem into a hypothesis statement in statistics that will include a null hypothesis and an alternate hypothesis. Breaking your problem into a few small steps makes these problems much easier to handle.

Hypothesis testing

- Example Problem: A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. Average recovery times for knee surgery patients is 8.2 weeks.
- Step 1: Figure out the hypothesis from the problem. The hypothesis is usually hidden in a word problem and is sometimes a statement of what you expect to happen in the experiment. The hypothesis in the above question is "I expect the average recovery period to be greater than 8.2 weeks."
- Step 2: Convert the hypothesis to math. Remember that the average is sometimes written as μ.
  - H1: $\mu > 8.2$
  - Broken down into (somewhat) English, that's H1 (The hypothesis): μ (the average) > (is greater than) 8.2
- Step 3: State what will happen if the hypothesis doesn't come true. If the recovery time isn't greater than 8.2 weeks, there are only two possibilities, that the recovery time is equal to 8.2 weeks or less than 8.2 weeks.
  - H0: $\mu \leq 8.2$
  - Broken down again into English, that's H0 (The null hypothesis): μ (the average) ≤ (is less than or equal to) 8.2

Hypothesis testing: But what if the researcher doesn't have any idea what will happen?

- Example Problem: A researcher is studying the effects of radical exercise program on knee surgery patients. There is a good chance the therapy will improve recovery time, but there's also the possibility it will make it worse. Average recovery times for knee surgery patients is 8.2 weeks.
- Step 1: State what will happen if the experiment doesn't make any difference. That's the null hypothesis--that nothing will happen. In this experiment, if nothing happens, then the recovery time will stay at 8.2 weeks.
  - H0: $\mu = 8.2$
  - Broken down into English, that's H0 (The null hypothesis): μ (the average) = (is equal to) 8.2
- Step 2: Figure out the alternate hypothesis. The alternate hypothesis is the opposite of the null hypothesis. In other words, what happens if our experiment makes a difference?
  - H1: $\mu \neq 8.2$
  - In English again, that's H1 (The alternate hypothesis): μ (the average) ≠ (is not equal to) 8.2

Chi-Square test

- Recap: A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.
- For example, a meal delivery firm in India wants to investigate the link between gender, geography, and people's food preferences. They can use chi-square test to find it out.

Chi-Square test

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

- Where
- c = Degrees of freedom
- O = Observed Value
- E = Expected Value

- The Observed values are those you gather yourselves.
- The expected values are the frequencies expected, based on the null hypothesis.
- The degrees of freedom in a statistical calculation represent the number of variables that can vary in a calculation. The degrees of freedom can be calculated to ensure that chi-square tests are statistically valid. These tests are frequently used to compare observed data with data that would be expected to be obtained if a particular hypothesis were true.

Why Do You Use the Chi-Square Test?

- Chi-square is a statistical test that examines the differences between categorical variables from a random sample in order to determine whether the expected and observed results are well-fitting.
- Here are some of the uses of the Chi-Squared test:
  - The Chi-squared test can be used to see if your data follows a well-known theoretical probability distribution like the Normal or Poisson distribution.
  - The Chi-squared test allows you to assess your trained regression model's goodness of fit on the training, validation, and test data sets.

Chi-Square test: Types

- There are two main types of Chi-Square tests:
- Independence:
  - The Chi-Square Test of Independence is a derivable (also known as inferential) statistical test which examines whether the two sets of variables are likely to be related with each other or not. This test is used when we have counts of values for two nominal or categorical variables. A relatively large sample size and independence of observations are the required criteria for conducting this test.

  - For Example: In a movie theatre, suppose we made a list of movie genres. Let us consider this as the first variable. The second variable is whether or not the people who came to watch those genres of movies have bought snacks at the theatre. Here the null hypothesis is that the genre of the film and whether people bought snacks or not are unrelatable. If this is true, the movie genres don’t impact snack sales.

Chi-Square test: Types

- Goodness-Of-Fit
  - In statistical hypothesis testing, the Chi-Square Goodness-of-Fit test determines whether a variable is likely to come from a given distribution or not. We must have a set of data values and the idea of the distribution of this data. We can use this test when we have value counts for categorical variables. This test demonstrates a way of deciding if the data values have a “good enough” fit for our idea or if it is a representative sample data of the entire population.

  - For Example: suppose we have bags of balls with five different colors in each bag. The given condition is that the bag should contain an equal number of balls of each color. The idea we would like to test here is that the proportions of the five colors of balls in each bag must be exact.

Chi-Square: Example

	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	440

- Let's say you want to know if gender has anything to do with political party preference. You poll 440 voters in a simple random sample to find out which political party they prefer. The results of the survey are shown in the table.
- To see if gender is linked to political party preference, perform a Chi-Square test of independence using the steps below.

Step 1: Define the Hypothesis

- H0: There is no link between gender and political party preference.
- H1: There is a link between gender and political party preference.

Chi-Square: Example

	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	440

Step 2: Calculate the Expected Values

- Now you will calculate the expected frequency.

$$\text{Expected Value} = \frac{(Row\ Total) * (Column\ Total)}{Total\ Number\ Of\ Observations}$$

- For example, the expected value for Male Republicans is: $= \frac{(240) * (200)}{440} = 109$

- Similarly, you can calculate the expected value for each of the cells.

Expected Values				
	Republican	**Democrat**	**Independent**	**Total**
Male	109	59	22.72	200
Female	120	65	25	220
Total	240	130	50	440

Chi-Square: Example

- Step 3: Calculate $(O-E)^2 / E$ for Each Cell in the Table.
  - Now you will calculate the $(O - E)^2 / E$ for each cell in the table.
    - Where,
      - O = Observed Value
      - E = Expected Value

$(O - E)^2/E$				
	Republican	**Democrat**	**Independent**	**Total**
Male	0.74311927	2.050847	2.332676056	200
Female	3.33333333	0.384615	1	220
Total	240	130	50	440

Chi-Square: Example

Step 4: Calculate the Test Statistic X^2

- X2 is the sum of all the values in the last table
- = 0.743 + 2.05 + 2.33 + 3.33 + 0.384 + 1 = 9.837
- Before you can conclude, you must first determine the critical statistic, which requires determining our degrees of freedom. The degrees of freedom in this case are equal to the table's number of columns minus one multiplied by the table's number of rows minus one, or (r-1) (c-1). We have (3-1)(2-1) = 2.
- Finally, you compare our obtained statistic to the critical statistic found in the chi-square table. As you can see, for an alpha level of 0.05 and two degrees of freedom, the critical statistic is 5.991, which is less than our obtained statistic of 9.83. You can reject our null hypothesis. Obtained statistic > Critical statistic ➔ H_0 can be rejected.
- This means you have sufficient evidence to say that there is an association between gender and political party preference.

The table

- For simplicity, we will use 0.05 as alpha level in all the examples.

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, Table D.1

Thank you