

STAT6171001

Basic Statistics

Analysis of Variance & Chi Square
Session 11

Session Learning Outcomes

Upon completion of this session, students are expected to be able to

- LO 2. Analyze a problem by using the basic concept of descriptive and inferential statistics
- LO 3. Design a descriptive and inferential statistics solution to meet a given set of computing requirements in the context of computer science
- LO4. Produce descriptive and inferential statistics solutions

Topics

- Analysis of Variance
- Chi Square

Analysis of Variance

Introduction

- Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means.
- It may seem odd that the technique is called “Analysis of Variance” rather than “Analysis of Means.”
- As you will see, the name is appropriate because inferences about means are made by analyzing variance.

When Can ANOVA Be Used?

- ANOVA can be used when the dependent variable in an experiment or a quasi-experiment is measured on an interval scale or a ratio scale.
- Analysis of variance would not be the right technique when the dependent variable is measured on an ordinal scale or a categorical scale.
- The interval scale of measurement is a type of measurement scale that is characterized by equal intervals between scale units. Example: the Fahrenheit scale to measure temperature.

When Can ANOVA Be Used?

- The ratio scale of measurement is a type of measurement scale. It's characterized by equal intervals between scale units and a minimum scale value of zero.
- The ordinal scale is a type of measurement scale. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale. Example: the results of a horse race, reported as "win", "place", and "show". We can't tell whether it was a close race or whether the winning horse won by a mile.
- Categorical. Categorical variables take on values that are names or labels. The color of a ball (red, green, blue) or the breed of a dog (collie, shepherd, terrier) would be examples of categorical variables.

How Does ANOVA Work?

- Analysis of variance refers to a set of techniques for interpreting differences between groups (in true experiments or in quasi-experiments).
- These techniques have similarities and differences.
- Note: Quasi-experiments are studies that aim to evaluate interventions but that do not use randomization

ANOVA Techniques

- Specify a **mathematical model** to describe how the independent variable(s) and the dependent variable are related.
- Write statistical **hypotheses** to be tested by experimental data.
- Specify a **significance level** for a hypothesis test.
- Compute **sums of squares** for each effect in the model.
- Find **degrees of freedom** for each effect in the model.
- Based on sums of squares and degrees of freedom, compute **mean squares** for each effect in the model.
- Find the **expected value** of the mean squares for each effect in the model.
- Compute **test statistics**, based on observed mean squares and their expected values.
- Find a **P value** for each observed test statistic.
- **Accept or reject the null hypothesis**, based on the P value and the significance level.
- Assess the **magnitude of the effect** of the independent variable(s), based on sums of squares.

One-Way Analysis of Variance

One-Way Analysis of Variance: Example

- A pharmaceutical company conducts an experiment to test the effect of a new cholesterol medication. They selects 15 subjects randomly from a larger population. Each subject is randomly assigned to one of three treatment groups. Within each treatment group, subjects receive a different dose of the new medication. In Group 1, subjects receive 0 mg/day; in Group 2, 50 mg/day; and in Group 3, 100 mg/day.
- The treatment levels represent all the levels of interest to the experimenter, so this experiment used a fixed-effects model to select treatment levels for study.
- After 30 days, doctors measure the cholesterol level of each subject.

Example

- The results for all 15 subjects appear in the table:
- In conducting this experiment, the experimenter had two research questions:
 - Does dosage level have a significant effect on cholesterol level?
 - How strong is the effect of dosage level on cholesterol level?

Dosage			
Group 1, 0 mg	Group 2, 50 mg	Group 3, 100 mg	
210	210	180	
240	240	210	
270	240	210	
270	270	210	
300	270	240	

Example

- To answer these questions, the experimenter intends to use one-way analysis of variance.
- But before you crunch the first number in one-way ANOVA, you must be sure that one-way ANOVA is the correct technique.
- That means you need to ask two questions:
 - Is the experimental design compatible with one-way ANOVA?
 - Does the data set satisfy the critical assumptions required for one-way ANOVA?
- Let's address both of those questions.

Example

- Experimental Design
 - One-way ANOVA is only appropriate with one experimental design - a completely randomized design.
 - That is exactly the design used in our cholesterol study, so we can check the experimental design box.

Example

- Critical Assumptions, one-way ANOVA makes 3 critical assumptions:
 - **Independence.** The dependent variable score for each experimental unit is independent of the score for any other unit.
 - **Normality.** In the population, dependent variable scores are normally distributed within treatment groups.
 - **Equality of variance (homogeneity of variance or homoscedasticity).** In the population, the variance of dependent variable scores in each treatment group is equal.
- Therefore, for the cholesterol study, we need to make sure our data set is consistent with the critical assumptions.

Example

Independence of Scores

- The assumption of independence is the most important assumption.
- When that assumption is violated, the resulting statistical tests can be misleading.
- The independence assumption is satisfied by the design of the study, which features random selection of subjects and random assignment to treatment groups.
- Randomization tends to distribute effects of extraneous variables evenly across groups.

Example

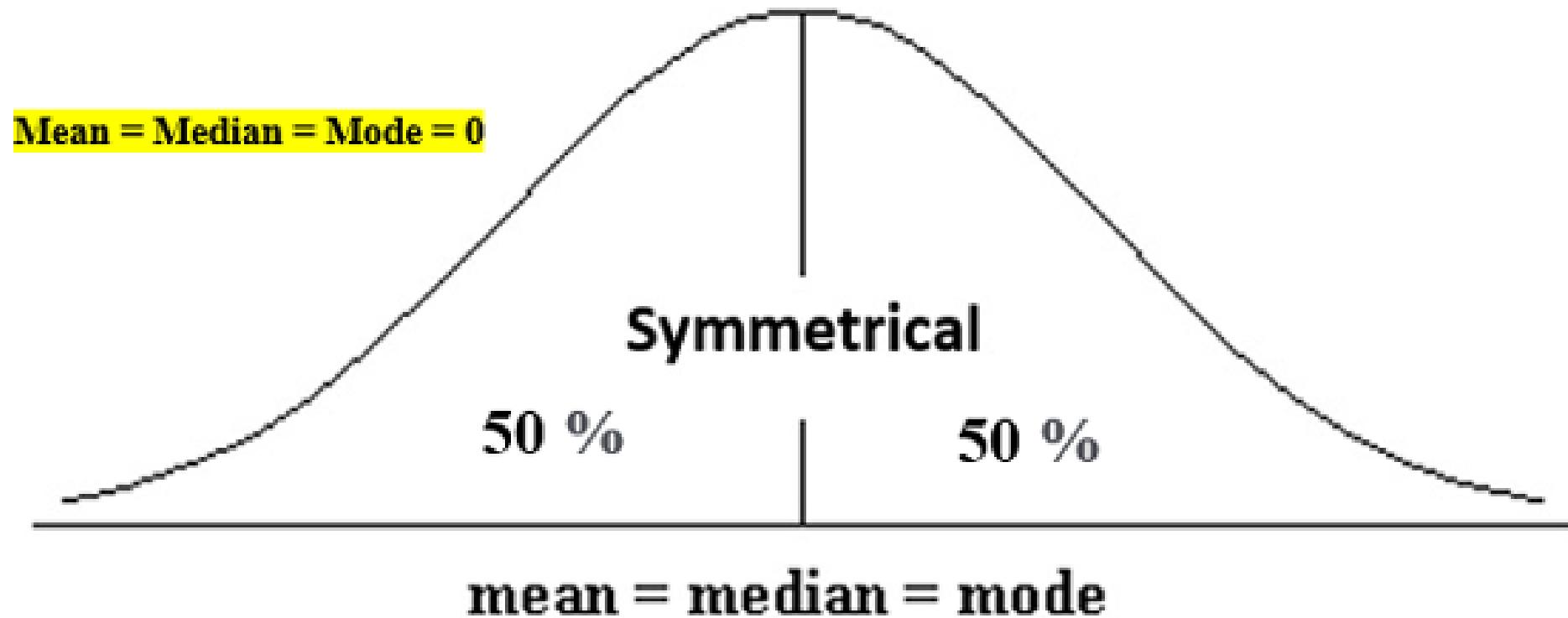
Normal Distributions in Groups

- Violations of normality can be a problem when sample size is small, as it is in this cholesterol study.
- Therefore, it is important to be on the lookout for any indication of non-normality.
- There are many ways to check for normality.
- Given the small sample size, our best option for testing normality is to look at the following descriptive statistics

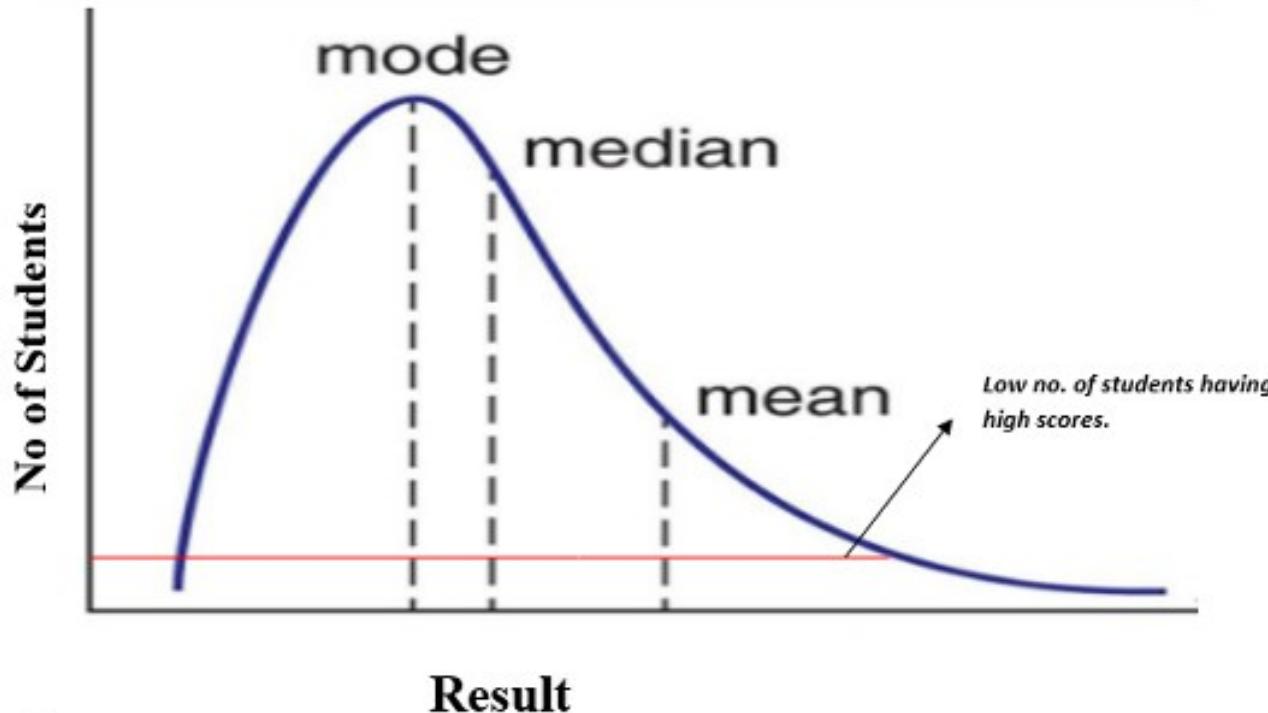
Example

- **Central tendency.** The mean and the median are summary measures used to describe central tendency. With a normal distribution, the mean = the median
- **Skewness.** Skewness is a measure of the asymmetry of a probability distribution. If observations are equally distributed around the mean, the skewness value is zero; otherwise, the skewness value is positive or negative. Skewness between -1 and +1 is consistent with a normal distribution.
- **Kurtosis.** It's a measure of whether observations cluster around the mean of the distribution or in the tails of the distribution. The normal distribution has a kurtosis value of zero. Kurtosis between -2 and +2 is consistent with a normal distribution.

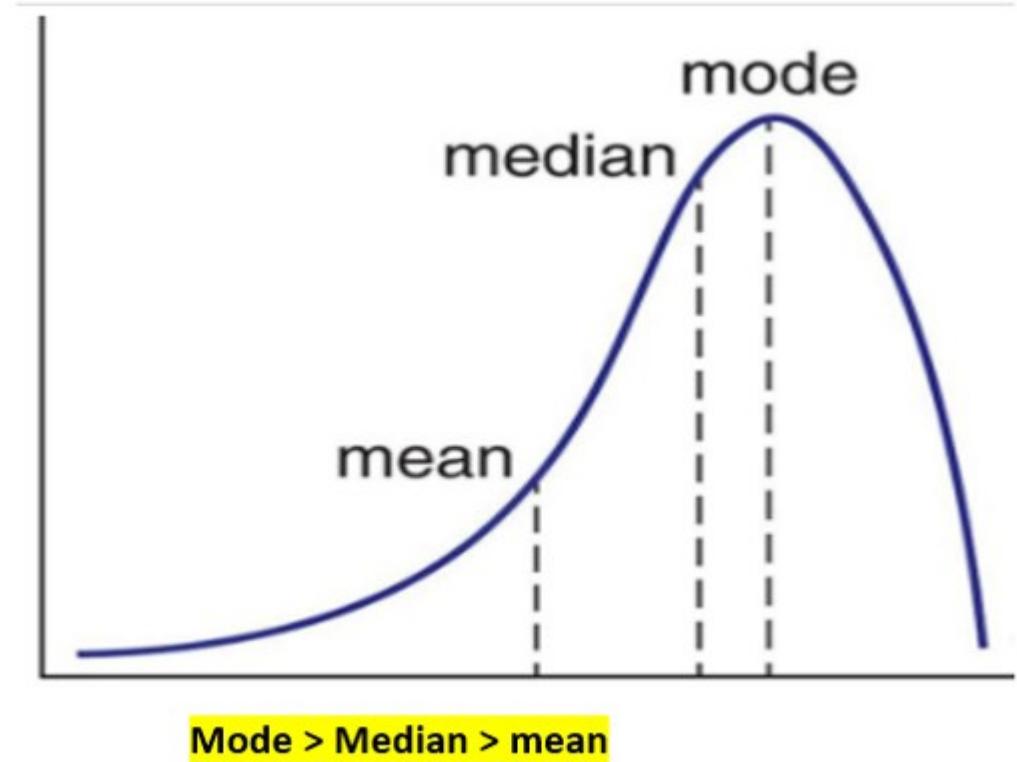
Central Tendency



Positive Skewed and Negative Skewed



Mean > Median > Mode



Calculate the Skewness Coefficient

- Skewness can be calculated using various methods, whereas the most commonly used method is Pearson's coefficient.

Pearson's First Coefficient of Skewness

- To calculate skewness values, subtract the mode from the mean, and then divide the difference by standard deviation.

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

- Pearson's first coefficient of skewness is helping if the data present high mode.
- But, if the data have low mode or various modes, Pearson's first coefficient is not preferred, and Pearson's second coefficient may be superior, as it does not rely on the mode.

Pearson's Second Coefficient of Skewness

- To calculate skewness values, subtract the median from the mean, multiply the difference by 3, and divide the product by the standard deviation.

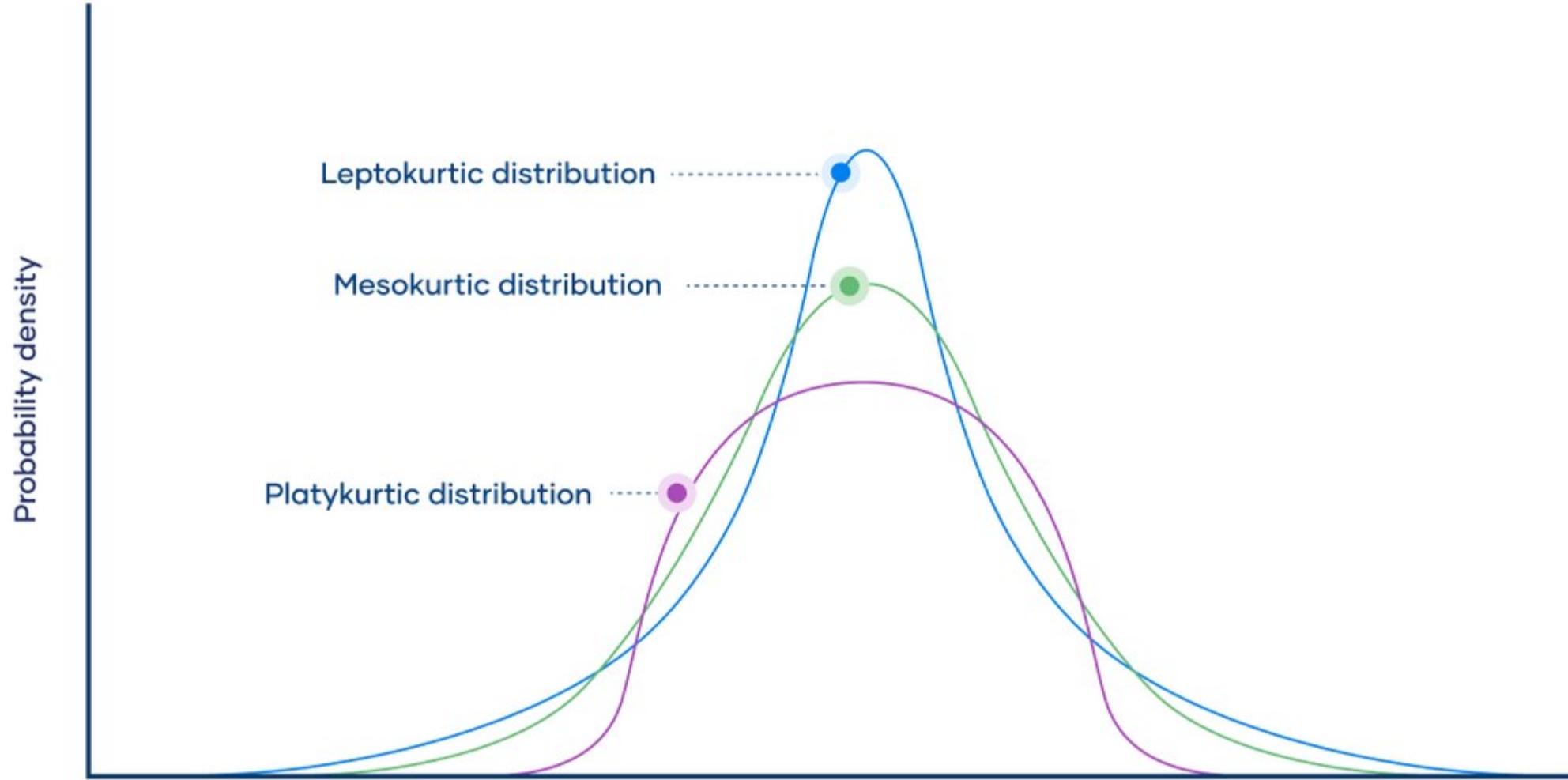
$$\text{Pearson's second coefficient} = \frac{3 (\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\text{Mean} - \text{Mode} \approx 3 (\text{Mean} - \text{Median})$$

What Is Kurtosis?

- Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution.
- Kurtosis helps in analyzing the characteristics and outliers of a dataset.

Types of Excess Kurtosis



Types of Excess Kurtosis

1. Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution, Kurtosis > 3)
2. Mesokurtic (kurtosis same as the normal distribution, Kurtosis = 3)
3. Platykurtic or short-tailed distribution (kurtosis less than normal distribution, Kurtosis < 3)

The Formula for Kurtosis

- The **sample excess kurtosis** is calculated as follows:

$$K = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$K_{\text{sample}} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- n denotes the number of observations in the sample, \bar{X} is the sample mean, whereas s is the sample standard deviation.

The Formula for Kurtosis

- The **population raw kurtosis** is calculated as follows:

$$K = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2\right)^2}$$

- n denotes the number of observations in the sample, μ is the population mean.

The Formula for Kurtosis

- If you subtract 3 from the kurtosis, you get excess kurtosis, which compares the dataset to a normal distribution.
- The **population excess kurtosis** is calculated as follows:

$$K_{\text{excess}} = K - 3$$

- For a normal distribution $K_{\text{excess}} = 0$.

Raw Kurtosis

Use raw kurtosis when:

- **Absolute Tailedness Matters:** You want to measure the overall degree of tailedness in a distribution without comparing it to a normal distribution.
- **Mathematical/Technical Applications:** Certain statistical models or formulas might require the raw fourth moment rather than a deviation from normality.
- **Comparing Across Datasets:** If you're comparing the kurtosis of multiple datasets but not necessarily against the normal distribution baseline.

Excess Kurtosis

Use excess kurtosis when:

- Comparison to Normality is Key: You want to assess how "different" the data's shape is from a normal distribution.
- Practical Interpretation: It simplifies the interpretation:
 - $K_{\text{excess}} = 0$: Same tailedness as normal.
 - $K_{\text{excess}} > 0$: Heavier tails (leptokurtic).
 - $K_{\text{excess}} < 0$: Lighter tails (platykurtic).
- Risk Analysis: In fields like finance, where fat tails indicate the potential for extreme events, excess kurtosis is often preferred.

When to Use Minus 3

- Use raw kurtosis if you're calculating the actual tailedness of a dataset.
- Use excess kurtosis if you need to compare a dataset to the normal distribution.

Example Kurtosis

- The new data points are 27, 13, 17, 57, 113, and 25.
- First, calculate the mean; add up the numbers and divide by 6 to get 42.

$$s^2 = \sum (y_i - \bar{y})^2$$

$$s^4 = \sum (y_i - \bar{y})^4$$

where:

- $y_i = i^{\text{th}}$ variable of the sample
- $\bar{y} = \text{Mean of the sample}$

source: <https://www.investopedia.com/>

Example Kurtosis

- To get s^2 , use each variable, subtract the mean, and then square the result.
- Add all of the results together:

$$(27 - 42)^2 = (-15)^2 = 225$$

$$(13 - 42)^2 = (-29)^2 = 841$$

$$(17 - 42)^2 = (-25)^2 = 625$$

$$(57 - 42)^2 = (15)^2 = 225$$

$$(113 - 42)^2 = (71)^2 = 5,041$$

$$(25 - 42)^2 = (-17)^2 = 289$$

$$225 + 841 + 625 + 225 + 5,041 + 289 = 7,246$$

Example Kurtosis

- To get s^4 , use each variable, subtract the mean, and raise the result to the fourth power.
- Add all of the results together:

$$(27 - 42)^4 = (-15)^4 = 50,625$$

$$(13 - 42)^4 = (-29)^4 = 707,281$$

$$(17 - 42)^4 = (-25)^4 = 390,625$$

$$(57 - 42)^4 = (15)^4 = 50,625$$

$$(113 - 42)^4 = (71)^4 = 25,411,681$$

$$(25 - 42)^4 = (-17)^4 = 83,521$$

$$\begin{aligned} 50,625 + 707,281 + 390,625 + 50,625 + 25,411,681 \\ + 83,521 = 26,694,358 \end{aligned}$$

Example Kurtosis

- So, our sums are:

$$s^2 = 7,246$$

$$s^4 = 26,694,358$$

- Now, calculate m_2 and m_4 , the second and fourth moments of the kurtosis formula:

$$\begin{aligned}m_2 &= \frac{s^2}{n} \\&= \frac{7,246}{6} \\&= 1,207.67\end{aligned}$$

$$\begin{aligned}m_4 &= \frac{s^4}{n} \\&= \frac{26,694,358}{6} \\&= 4,449,059.67\end{aligned}$$

Example Kurtosis

$$k = \frac{m^4}{m^2^2} - 3$$

- So, the kurtosis is:

$$\frac{4,449,059.67}{1,458,466.83} - 3 = .05$$

Example

- The table shows the mean, median, skewness, and kurtosis:
- In all three groups, the difference between the mean and median looks small.
- Skewness and kurtosis measures are consistent with a normal distribution.
- These are crude tests, but they provide some confidence for the assumption of normality in each group.

	Group 1, 0 mg	Group 2, 50 mg	Group 3, 100 mg
Mean	258	246	210
Median	270	240	210
Range	90	60	60
Skewness	-0.40	-0.51	0.00
Kurtosis	-0.18	-0.61	2.00

Example

Homogeneity of Variance

- When the normality of variance assumption is satisfied, you can use Hartley's *Fmax test* to test for homogeneity of variance.
- **Step 1.** Compute the sample variance (s^2_j) for each group

$$s^2_j = \frac{\sum_{j=1}^k (X_{i,j} - \bar{X}_j)^2}{(n_j - 1)}$$

- Where $X_{i,j}$ is the score for observation i in Group j, \bar{X}_j is the mean of Group j, and n_j is the number of observations in Group j.

Example

- Here is the variance (s_j^2) for each group in the cholesterol study

Group 1, 0 mg	Group 2, 50 mg	Group 3, 100 mg
1170	630	450

Example

Homogeneity of Variance

- **Step 2.** Compute an F ratio from the following formula:

$$F_{\text{RATIO}} = s^2_{\text{MAX}} / s^2_{\text{MIN}}$$

$$F_{\text{RATIO}} = 1170 / 450$$

$$F_{\text{RATIO}} = 2.6$$

- where s^2_{MAX} is the largest group variance, and s^2_{MIN} is the smallest group variance.

Example

Homogeneity of Variance

- **Step 3.** Compute degrees of freedom (df).

$$df = n - 1$$

$$df = 5 - 1$$

$$df = 4$$

- where n is the largest sample size in any group.

Example

Homogeneity of Variance

- **Step 4.** Based on the degrees of freedom (4) and the number of groups (3), Find the critical F value from the Table of Critical F Values for Hartley's Fmax Test. From the table, we see that the critical Fmax value is 15.5.
- Note: The critical F values in the table are based on a significance level of 0.05.

Example

DF (n-1)	Number of treatments (k)										
	2	3	4	5	6	7	8	9	10	11	12
2	39.0	87.5	142	202	266	333	403	475	550	626	714
	199	448	729	1036	1362	1705	2063	2432	2813	3204	3605
3	15.4	27.8	39.2	50.7	62.0	72.9	83.5	93.9	104	114	124
	47.5	85.0	120	151	184	21	24	28	31	33	36
4	9.6	15.5	20.6	25.2	29.5	33.6	37.5	41.1	44.6	48.0	51.4
	23.2	37.0	49.0	59	69	79	89	97	106	113	120
5	7.2	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5	28.2	29.9
	14.9	22.0	28.0	33	38	42	46	50	54	57	60
6	5.82	8.28	10.4	12.1	13.7	15.2	16.2	17.5	18.6	19.7	20.7

Example

Homogeneity of Variance

- **Step 5.** Compare the observed F ratio computed in Step 2 to the critical F value recovered from the Fmax table in Step 4.
- If the F ratio is smaller than the Fmax table value, the variances are homogeneous.
- Otherwise, the variances are heterogeneous.
- Here, the F ratio (2.6) is smaller than the Fmax value (15.5), so we conclude that the variances are **homogeneous**.

Example - Analysis of Variance

- Having confirmed that the critical assumptions are tenable, we can proceed with a one-way analysis of variance. That means taking the following steps:
 - Specify a mathematical model to describe the causal factors that affect the dependent variable.
 - Write statistical hypotheses to be tested by experimental data.
 - Specify a significance level for a hypothesis test.
 - Compute the grand mean and the mean scores for each group.
 - Compute sums of squares for each effect in the model.
 - Find the degrees of freedom associated with each effect in the model.
 - Based on sums of squares and degrees of freedom, compute mean squares for each effect in the model.
 - Compute a test statistic, based on observed mean squares and their expected values.
 - Find the P value for the test statistic.
 - Accept or reject the null hypothesis, based on the P value and the significance level.
 - Assess the magnitude of the effect of the independent variable, based on sums of squares.

Example

1. Mathematical Model

- In this experiment, the dependent variable (X) is the cholesterol level of a subject, and the independent variable (β) is the dosage level administered to a subject.
- Here is the fixed-effects model for a completely randomized design:

$$X_{ij} = \mu + \beta_j + \varepsilon_{i(j)}$$

- where X_{ij} is the cholesterol level for subject i in treatment group j , μ is the population mean, β_j is the effect of the dosage level administered to subjects in group j ; and $\varepsilon_{i(j)}$ is the effect of all other extraneous variables on subject i in treatment j .

Example

2. Statistical Hypotheses

- For fixed-effects models, it is common practice to write statistical hypotheses in terms of the treatment effect β_j .
- Null hypothesis: the independent variable (dosage level) has no effect on the dependent variable (cholesterol level) in any treatment group.

$$H_0: \beta_j = 0 \text{ for all } j$$

- Alternative hypothesis: the independent variable has an effect on the dependent variable in at least one treatment group.

$$H_1: \beta_j \neq 0 \text{ for some } j$$

- If the null hypothesis is true, mean scores in the k treatment groups should be equal. If the null hypothesis is false, at least one pair of mean scores should be unequal.

Example

3. Significance Level

- The significance level (also known as alpha or α) is the probability of rejecting the null hypothesis when it is actually true.
- Experimenters often choose significance levels of 0.05 or 0.01.
- For this experiment, let's use a significance level of 0.05.

Example

4. Mean Scores

- Analysis of variance begins by computing a grand mean and group means
- Grand mean. The grand mean (\bar{X}) is the mean of all observations, computed as follows

$$n = \sum_{j=1}^k n_j = 5 + 5 + 5 = 15$$

$$\bar{X} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij})$$

$$\begin{aligned}\bar{X} &= (1 / 15) * (210 + 210 + \dots + 270 + 240) \\ \bar{X} &= 238\end{aligned}$$

Example

- Group means. The mean of group j (\bar{X}_j) is the mean of all observations in group j, computed as follows:

$$\bar{X}_j = \left(\frac{1}{n_j} \right) \sum_{i=1}^{n_j} (X_{ij})$$
$$\bar{X}_1 = 258$$
$$\bar{X}_2 = 246$$
$$\bar{X}_3 = 210$$

- In the equations above, n is the total sample size across all groups; and n_j is the sample size in Group j .

Example

5. Sums of Squares

- A sum of squares is the sum of squared deviations from a mean score. One-way analysis of variance makes use of three sums of squares:
 - a. Between-groups sum of squares
 - b. Within-groups sum of squares
 - c. Total sum of squares

Example

a. Between-groups sum of squares (SSB)

- SSB measures variation of group means around the grand mean.
- It can be computed from the following formula:

$$SSB = \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{X}_j - \bar{X})^2 = \sum_{j=1}^k n_j (\bar{X}_j - \bar{X})^2$$

$$SSB = 5 * [(238-258)^2 + (238-246)^2 + (238-210)^2]$$

$$SSB = 6240$$

Example

b. Within-groups sum of squares (SSW)

- SSW measures variation of all scores around their respective group means.
- It can be computed from the following formula:

$$SSW = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

$$SSW = (210 - 258)^2 + \dots + (240 - 210)^2 = 2304 + \dots + 900 = 9000$$

Example

c. Total sum of squares (SST)

- SST measures variation of all scores around the grand mean.
- It can be computed from the following formula:

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

$$SST = 784 + 4 + 1084 + \dots + 784 + 784 + 4$$

$$SST = 15,240$$

Example

- It turns out that the total sum of squares is equal to the between-groups sum of squares plus the within-groups sum of squares, as shown below:

$$SST = SSB + SSW$$

$$15,240 = 6240 + 9000$$

Example

6. Degrees of Freedom

- The term degrees of freedom (df) refers to the number of independent sample points used to compute a statistic minus the number of parameters estimated from the sample points.
- To illustrate what is going on, let's find the degrees of freedom associated with the various sum of squares computations:
 - a. Between-groups sum of squares
 - b. Within-groups sum of squares
 - c. Total sum of squares

Example

a. Between-groups sum of squares (SSB)

- The between-groups sum of squares formula appears below:

$$SSB = \sum_{j=1}^k n_j (\bar{X}_j - \bar{X})^2$$

- The formula uses k independent sample points, the sample means \bar{X}_j . And it uses one parameter estimate, the grand mean \bar{X} , which was estimated from the sample points.
- So, the between-groups sum of squares has $k-1$ degrees of freedom (df_{BG}).

$$df_{BG} = k - 1 = 3 - 1 = 2$$

Example

b. Within-groups sum of squares (SSW)

- The within-groups sum of squares formula appears below:

$$SSW = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

- The formula uses n independent sample points, the individual subject scores X_{ij} . And it uses k parameter estimates, the group means \bar{X}_j , which were estimated from the sample points. So, the within-groups sum of squares has $n - k$ degrees of freedom (df_{WG}).

$$n = \sum n_i = 5 + 5 + 5 = 15$$

$$df_{WG} = n - k = 15 - 3 = 12$$

Example

c. Total sum of squares (SST)

- The total sum of squares formula appears below:

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

- The formula uses n independent sample points, the individual subject scores X_{ij} . And it uses one parameter estimate, the grand mean \bar{X} , which was estimated from the sample points. So, the total sum of squares has $n - 1$ degrees of freedom (df_{TOT}).

$$df_{TOT} = n - 1 = 15 - 1 = 14$$

Example

- The degrees of freedom for each sum of squares are summarized in the table below:

Sum of squares	Degrees of freedom
Between-groups	$k - 1 = 2$
Within-groups	$n - k = 12$
Total	$n - 1 = 14$

Example

7. Mean Squares

- A mean square is an estimate of population variance.
- It is computed by dividing a sum of squares (SS) by its corresponding degrees of freedom (df), as shown below:

$$MS = SS / df$$

- To conduct a one-way analysis of variance, we are interested in two mean squares:
 - a. Between-groups sum of squares
 - b. Within-groups sum of squares

Example

a. Between-groups sum of squares (SSB)

- The between-groups mean square (MS_{BG}) refers to variation due to differences among experimental units within the same group plus variation due to treatment effects. It can be computed as follows:

$$MS_{BG} = SSB / df_{BG}$$

$$MS_{BG} = 6240 / 2 = 3120$$

Example

b. Within-groups sum of squares (SSW)

- The within-groups mean square (MS_{WG}) refers to variation due to differences among experimental units within the same group. It can be computed as follows:

$$MS_{WG} = SSW / df_{WG}$$

$$MS_{WG} = 9000 / 12 = 750$$

Example

8. Test Statistic

- Suppose we use the mean squares to define a test statistic F as follows:

$$F(v_1, v_2) = MS_{BG} / MS_{WG}$$

$$F(2, 12) = 3120 / 750 = 4.16$$

- where MS_{BG} is the between-groups mean square, MS_{WG} is the within-groups mean square, v_1 is the degrees of freedom for MS_{BG} , and v_2 is the degrees of freedom for MS_{WG} .

Example

- The F ratio is a convenient measure that we can use to test the null hypothesis. Here's how:
 - When the F ratio is close to one, MS_{BG} is approximately equal to MS_{WG} . This indicates that the independent variable did not affect the dependent variable, so we cannot reject the null hypothesis.
 - When the F ratio is significantly greater than one, MS_{BG} is bigger than MS_{WG} . This indicates that the independent variable did affect the dependent variable, so we must reject the null hypothesis.
- What does it mean for the F ratio to be *significantly* greater than one? To answer that question, we need to talk about the **P-value**.

Example

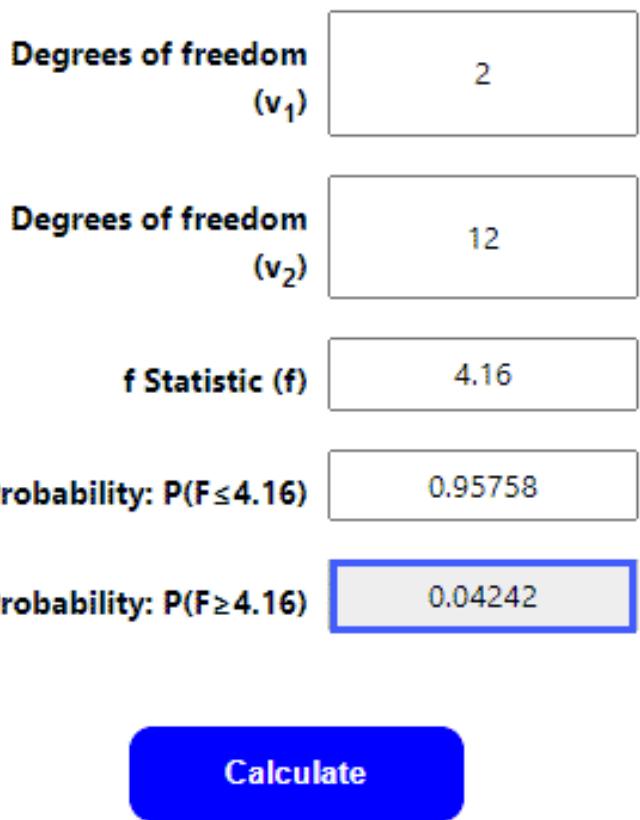
9. P-Value

- In an experiment, a P-value is the probability of obtaining a result more extreme than the observed experimental outcome, assuming the null hypothesis is true.
- With ANOVA, the F ratio is the observed experimental outcome that we are interested in.
- So, the P-value would be the probability that an F statistic would be more extreme (i.e., bigger) than the actual F ratio computed from experimental data.

Example

- We can use Stat Trek's F Distribution Calculator to find the probability that an F statistic will be bigger than the actual F ratio observed in the experiment.
- <https://stattrek.com/online-calculator/f-distribution>
- From the calculator, we see that the $P(F > 4.16)$ equals about 0.04.
- Therefore, the P-Value is 0.04.

- Enter values for degrees of freedom (v_1 and v_2).
- Enter a value for one, and only one, of the other textboxes.
- Click **Calculate** to compute a value for the last textbox.



Degrees of freedom (v_1)

Degrees of freedom (v_2)

f Statistic (f)

Probability: $P(F \leq 4.16)$

Probability: $P(F \geq 4.16)$

Calculate

Example

10. Hypothesis Test

- We specified a significance level 0.05 for this experiment.
- Here's the decision rule for accepting or rejecting the null hypothesis:
 - If the P-value $> \alpha$, accept the null hypothesis.
 - If the P-value $\leq \alpha$, reject the null hypothesis.
- Since the P-value (0.04) $< \alpha$ (0.05), we reject the null hypothesis that drug dosage had no effect on cholesterol level.
- And we conclude that the mean cholesterol level in at least one treatment group differed significantly from the mean cholesterol level in another group.

Example

11. Magnitude of Effect

- The hypothesis test tells us whether the independent variable in our experiment has a statistically significant effect on the dependent variable, but it does not address the magnitude of the effect.
- Here's the issue:
 - When the sample size is large, you may find that even small differences in treatment means are statistically significant.
 - When the sample size is small, you may find that even big differences in treatment means are not statistically significant.

Example

- It is customary to supplement analysis of variance with an appropriate measure of effect size.
- Eta squared (η^2) is one such measure.
- Eta squared is the proportion of variance in the dependent variable that is explained by a treatment effect.
- The eta squared formula for one-way analysis of variance is:
 - $\eta^2 = SSB / SST$
- where SSB is the between-groups sum of squares and SST is the total sum of squares.

Example

- Given this formula, we can compute eta squared for this drug dosage experiment, as shown below:
 - $\eta^2 = SSB / SST = 6240 / 15240 = 0.41$
- Thus, 41 percent of the variance in our dependent variable (cholesterol level) can be explained by variation in our independent variable (dosage level).
- It appears that the relationship between dosage level and cholesterol level is significant not only in a statistical sense; it is **significant** in a practical sense as well.

Example - ANOVA Summary Table

- This ANOVA table allows any researcher to interpret the results of the experiment, at a glance.

Analysis of Variance Table

Source	SS	df	MS	F	P
BG	6,240	2	3,120	4.16	0.04
WG	9,000	12	750		
Total	15,240	14			

Chi Square

What Are Categorical Variables?

- Categorical variables belong to a subset of variables that can be divided into discrete categories. Names or labels are the most common categories.
- Known as qualitative variables.
- Categorical variables can be divided into two categories:
 - Nominal Variable: A nominal variable's categories have no natural ordering. Example: Gender, Blood groups
 - Ordinal Variable: A variable that allows the categories to be sorted is ordinal variables. Customer satisfaction (Excellent, Very Good, Good, Average, Bad, and so on) is an example.

What Is a Chi-Square Test?

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

Chi-Square Test Definition

- 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.
- A chi-square test is required to test a hypothesis regarding the distribution of a categorical variable.
- Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal.

Formula for Chi-Square Test

- Where:
 - c = Degrees of freedom
 - O = Observed Value
 - E = Expected Value
- The degrees of freedom in a statistical calculation represent the number of variables that can vary in a calculation.
- The Observed values are those you gather yourselves.
- The expected values are the frequencies expected, based on the null hypothesis.

$$\chi_c^2 = \frac{\sum (O_i - E_i)^2}{E_i}$$

Chi-Square Test

- Chi-Squared Tests are most commonly used in hypothesis testing.
- The Chi-Square test estimates the size of inconsistency between the expected results and the actual results when the size of the sample and the number of variables in the relationship is mentioned.
- These tests use degrees of freedom to determine if a particular null hypothesis can be rejected based on the total number of observations made in the experiments.
- Larger the sample size, more reliable is the result.

Chi-Square Test

- There are two main types of Chi-Square tests namely -
 - Independence
 - Goodness-of-Fit

Independence

- It's 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 and is considered as non-parametric test.
- A relatively large sample size and independence of observations are the required criteria for conducting this test.

Independence

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 unrelated.
- If this is true, the movie genres don't impact snack sales.

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.

Goodness-of-Fit

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.

Goodness-of-Fit

Example:

- A die is rolled 60 times. You want to test if the die is fair (i.e., each side has an equal probability of appearing).

Null Hypothesis (H_0): The die is fair, so the observed frequencies match the expected frequencies.

Alternative Hypothesis (H_1): The die is not fair, so the observed frequencies do not match the expected frequencies.

Goodness-of-Fit

Example:

The observed frequencies (from rolling the die 60 times):

Face	1	2	3	4	5	6
Observed Frequency (O_i)	8	10	12	11	9	10

If the die is fair, each face should appear $\frac{1}{6}$ of the time.

Expected frequency for each face (E_i) = Total Rolls \times Probability of each face = $60 \times \frac{1}{6} = 10$.

Face	1	2	3	4	5	6
Expected Frequency (E_i)	10	10	10	10	10	10

Conclusion Statement

- Depends on whether the calculated chi-square value is greater than or less than the critical value.
- If greater than the critical value: reject the null hypothesis
- If less than the critical value: fail to reject the null hypothesis

Example

- We want to know if gender has anything to do with political party preference. Poll 420 voters in a simple random sample to find out which political party they prefer. The results of the survey are shown in the table below:

	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	420

- To see if gender is linked to political party preference, perform a Chi-Square test of independence using the steps below.

Example

- Step 1: Define the Hypothesis
 - H_0 : There is no link between gender and political party preference.
 - H_1 : There is a link between gender and political party preference.

Example

- Step 2: Calculate the Expected Values
 - Now you will calculate the expected frequency, for example, the expected value for Male Republicans is:

$$\text{Expected Value} = \frac{(\text{Row Total}) * (\text{Column Total})}{\text{Total Number Of Observations}} = \frac{(240) * (200)}{420} = 114,29$$

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

Expected Values				
		Republican	Democrat	Independent
Male	Republican	114,29	61,90	23,81
	Democrat	61,90	68,10	26,19
Total	Independent	240	130	50
Total	Total	200	220	420

Example

- Step 3: Calculate $(O - E)^2 / E$ for Each Cell in the Table
 - Where: O = Observed Value and E = Expected Value

		$(O - E)^2 / E$		
		Republican	Democrat	Independent
		Male	Female	Male
Male		1,79	1,06	1,61
Female		1,62	0,96	1,46

Example

- Step 4: Calculate the Test Statistic X^2

X^2 is the sum of all the values in the last table

$$= 1.79 + 1.06 + 1.61 + 1.62 + 0.96 + 1.46$$

$$= \mathbf{8.50}$$

- Before conclude, first determine the critical statistic, which requires determining our degrees of freedom (df).
- 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 $(c-1)(r-1) \rightarrow (3-1)(2-1) = 2$

Example

- Compare obtained statistic with the chi-square table.
- For $\alpha = 0.05$ and $df = 2$, the critical statistic is 5.991, less than our obtained statistic of 8.50.
- You can reject our null hypothesis because the obtained statistic is higher than your critical statistic, means it has sufficient evidence to say that there is an association between gender and political party preference.

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

d	Probability of exceeding the critical value				d				
	0.05	0.01	0.001			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	

What the Difference?

- Chi-Square test is used when we perform hypothesis testing on two categorical variables from a single population or we can say that to compare categorical variables from a single population.
- T-test is an inferential statistic that is used to determine the difference or to compare the means of two groups of samples which may be related to certain features. It is performed on continuous variables.
- Analysis of variance is used to compare multiple (three or more) samples with a single test. It is used when the categorical feature has more than two categories.

Z-TEST

- In a z-test, we assume the sample is normally distributed.
- A z-score is calculated with population parameters such as population mean and population standard deviation.
- We use this test to validate a hypothesis that states the sample belongs to the same population.

T-TEST

- We use a t-test to compare the mean of two given samples.
- Like a z-test, a t-test also assumes a normal distribution of the sample.
- When we don't know the population parameters (mean and standard deviation), we use t-test.

CHI-SQUARE TEST

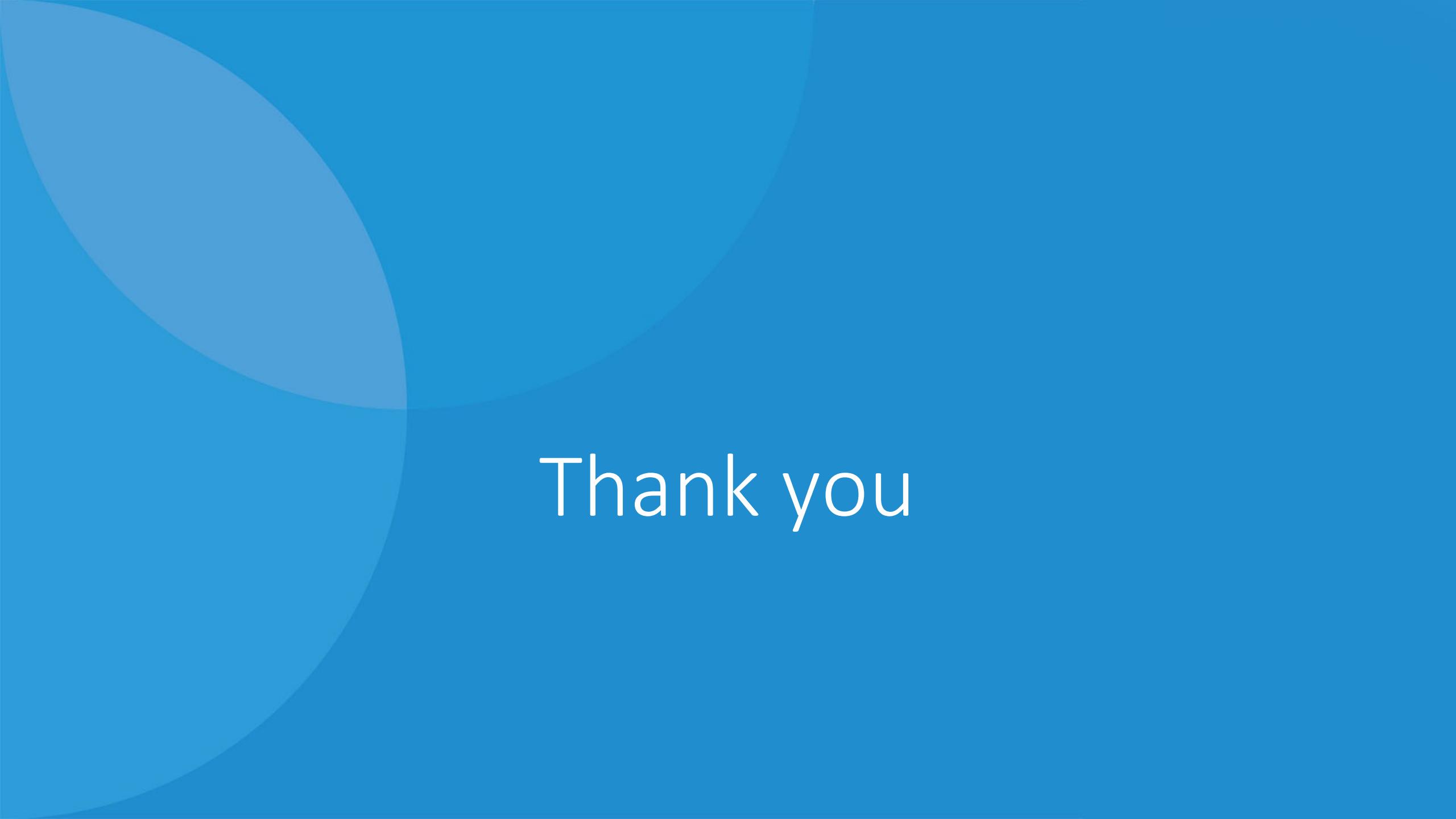
- We use the chi-square test to compare categorical variables.
- We use a t-test to compare the mean of two given samples, but we use the chi-square test to compare categorical variables.

ANOVA

- We use analysis of variance (ANOVA) to compare three or more samples with a single test.

References

- Witte, R.S.&Witte, J.S. (2017). Statistics (11th ed.). Wiley. ISBN: 978-1119386056.
- Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D.& Zimmer, H. (2003). Introduction to Statistics. Online edition at <https://open.umn.edu/opentextbooks/textbooks/459>
- Levine, D.M., Stephan, D.F. & Szabat, K.A. (2017). Statistics for Managers Using Microsoft Excel (8th ed.). Pearson. ISBN: 978-0134566672

The background features a minimalist design with three overlapping circles in varying shades of blue. A large, semi-transparent light blue circle is positioned in the upper left, while two smaller, darker blue circles overlap in the lower right.

Thank you