HCAI5DS02 – Data Analytics and Visualization.
Lecture – 08
Statistical Inference and Hypothesis Testing – Part -II.
Chi – Square and ANOVA

Siman Giri

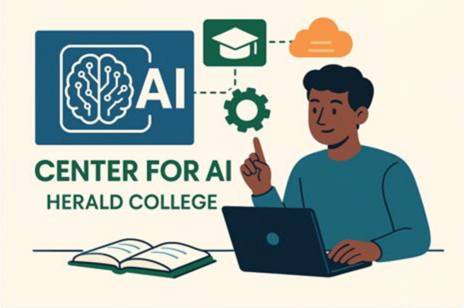


image generated via copilot.

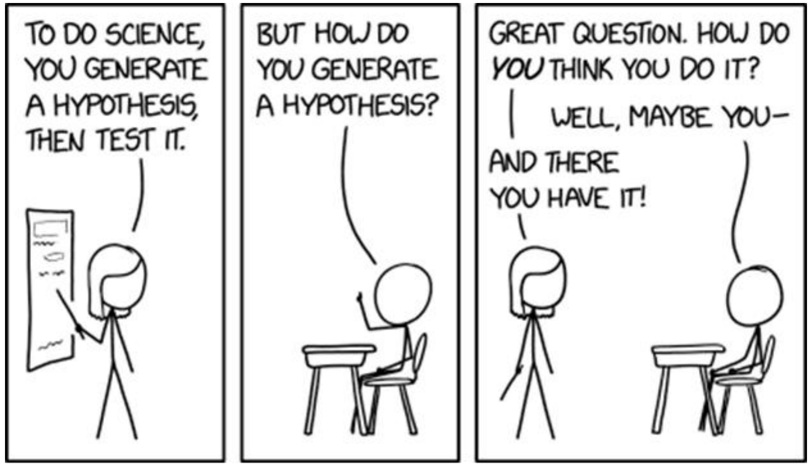
HERALD COLLEGE | CENTER FOR AI.

From Last Week …

{A. Basics of Hypothesis Testing …}

A.1 Statistical Inference and Hypothesis Testing.

- **Statistical inference** allows us to **evaluate claims** about a **population parameter** using **observed data**.
  - This is done **through hypothesis testing**, where we:
    1. Formulate a null hypothesis (e.g. no effect, no difference)
    2. Collect sample data
    3. Calculate a test statistic (e.g. t – value, z – score)
    4. Evaluate the evidence against the null hypothesis using probability (p – values)
    5. **Make a decision: reject or fail to reject the null.**



©Explain xkcd.

A.2 Setting up Hypothesis …

1. **Setup the null Hypothesis:**
   - The **null Hypothesis** always states that there is **no difference**, **no relationship** between **variables in a study**.
2. **Setup Alternate Hypothesis:**
   - The **alternative hypothesis** (denoted as $\mathbf{H_1}$ **or** $\mathbf{H_a}$) is the statement that researchers
     - **aim to provide evidence for in a statistical hypothesis test.**
   - It is the opposite of the null hypothesis and represents:
     - the **presence of an effect**, **difference, or relationship that the researcher expects or hopes to find**.

Example 1:	The new drug has no effect on the disease compared to a placebo.		
Scenario	**Null Hypothesis**	**Alternate Hypothesis**	**In Practice**
Clinical Trial: New Drug vs. Placebo	• No difference in effect. • $H_0: \mu_{drug} = \mu_{placebo}$	• $H_a: \mu_{drug} \neq \mu_{placebo}$ • **Two-sided test.**	• $H_1: \mu_{drug} > \mu_{placebo}$ • We are only interested in improvement. • **One – sided alternative.**

A.3 Hypothesis Testing → Making Decisions.

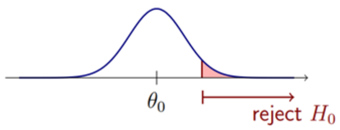
- The **hypothesis** we want to test is **whether H_a is likely true**.
  - So, there are two possible outcomes:
    - **Reject H_0 and accept H_a** because of **sufficient evidence** in the sample in favor of **H_a**.
    - Do not **reject H_0** because of **insufficient evidence** to **support H_a**.

Very important!!!
Note that failure to **reject H_0** does not mean the null hypothesis is true. There is no formal outcome that says "**accept H_0**".It only means that we do not have sufficient evidence to **support H_a**.

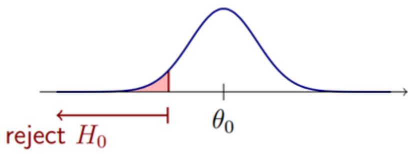
A.4 One vs. Two Tailed Test

One – Tailed Test:

- A **one tailed test** is used when you are interested in
  - **detecting a difference in a specific direction.**
- You **hypothesize** that:
  - **the true parameter is**
    - **either greater than or less than a certain value**,
    - **but not both**.



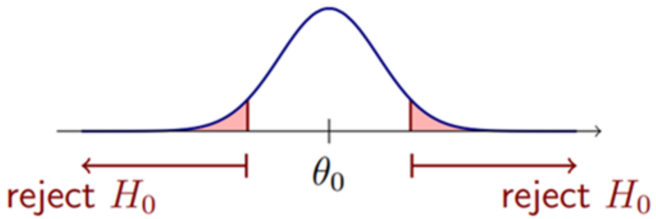
- **Right Tailed:**
  - $\mathbf{H_0: \mu \leq \mu_0}$
    $\mathbf{H_a: \mu > \mu_0}$



- **Left Tailed:**
  - $\mathbf{H_0: \mu \geq \mu_0}$
    $\mathbf{H_a: \mu < \mu_0}$

Two – Tailed Test:

- A **two tailed test** is used when you are interested in
  - **detecting any significant difference.**
  - From **null hypothesis**, **regardless of the direction**.



- **Two Tailed:**
  - $\mathbf{H_0: \mu = \mu_0}$
    $\mathbf{H_a: \mu \neq \mu_0}$

A.5 Understanding What a Hypothesis Can Test?

1. **Single – Sample Hypothesis (One Group vs. a value):**
   - You are testing whether the population mean of one group is equal to some fixed number or benchmark.
   - *Example: Is the average delivery time less than 30 minutes?*
   - **Hypothesis:**
     - **null $\rightarrow$ $H_0: \mu = 30$; alternate $\rightarrow$ $H_a: \mu < 30$**
2. **Two – Sample Hypothesis (Between two Groups):**
   - You are testing whether two different groups have the same mean, proportion etc.
   - *Example: Do male and female employees earn the same average salary?*
   - **Hypothesis:**
     - **null $\rightarrow$ $H_0: \mu_1 = \mu_2$; alternate $\rightarrow$ $H_a: \mu_1 \neq \mu_2$**
3. **Hypothesis about relationships (Between Variables):**
   - You are testing whether two variables are related (not necessarily groups).
   - *Example: Is there a correlation between study time and exam score?*
   - **Hypothesis:**
     - **null $\rightarrow$ $H_0: \rho = 0$ (no correlation) ; alternate $\rightarrow$ $H_a: \rho \neq 0$**
   - This tests for **association**, not difference between groups.

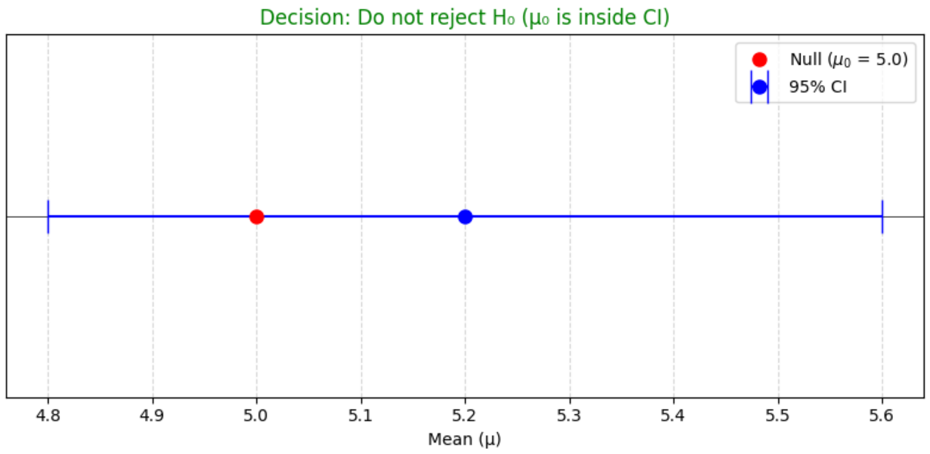
B. Various Approaches to Hypothesis Testing.

B.1 Confidence Interval Approach.

- We can use a confidence interval to help us **weigh the evidence against the null hypothesis**.
  - A confidence interval gives us a **range of** ***plausible*** **values for μ.**
    - If the **null value** is **in the interval**, then μ_0 is *plausible value* for **μ**.
    - If the **null value** is **not in the interval**, then μ_0 is *not a plausible value* for **μ**.

1. **State Hypotheses:**
   - $\mathbf{H_0}: \mu = \mu_0$ (NullHypothesis)
   - $\mathbf{H_1}: \mu \neq \mu_0$ (AlternativeHypothesis)
2. **Choose Significance Level:**
   - $\alpha = 0.05$(95%CI) or $\alpha = 0.01$(99%CI)
3. **Check Assumptions:**
   - Sample size, normality, data conditions for CI method
4. **Find Critical Value:**
   - Using z^* (normal) or t^* (t-distribution) for given α
5. **Compute Confidence Interval:**
   - $\mathrm{CI} = \bar{x} \pm (\text{critical value} \times \mathrm{SE})$
6. **Decision Rule:**

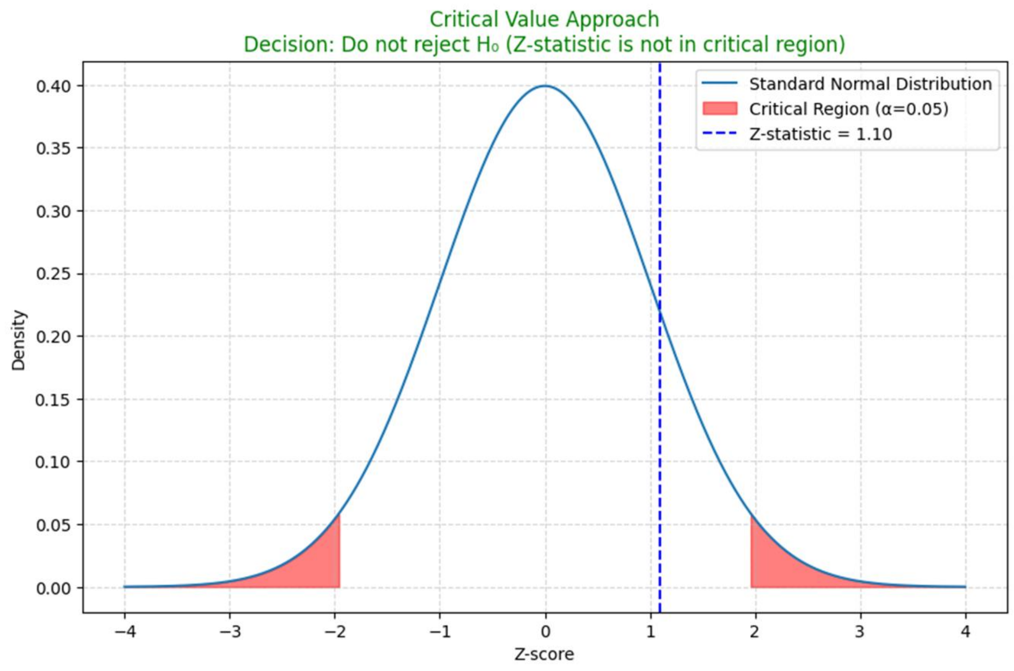
   if $\mu_0 \notin CI$ then
   Reject H_0
   ⇒ Evidence for H_1 ($\mu \neq \mu_0$)
   else
   Fail to reject H_0
   ⇒ No significant difference found
7. **Interpretation:**
   - State conclusion in context of the research question



B.2 Critical Value Approach.

1. **State Hypotheses:**
   - $\mathbf{H_0} : \mu = \mu_0$
   - $\mathbf{H_1} : \mu \neq \mu_0$
2. **Choose Significance Level:**
   - $\alpha = \mathbf{0.05}$ or $\mathbf{0.01}$
3. **Check Assumptions:**
   - **Normality, sample size requirements**
4. **Calculate Test Statistic:**
   - $\mathbf{z} = \dfrac{\bar{\mathbf{x}} - \mu_0}{\mathbf{SE}}$ or **t**-statistic
5. **Find Critical Value:**
   - $\mathbf{z}^*$ or $\mathbf{t}^*$ for given α
6. **Decision Rule:**

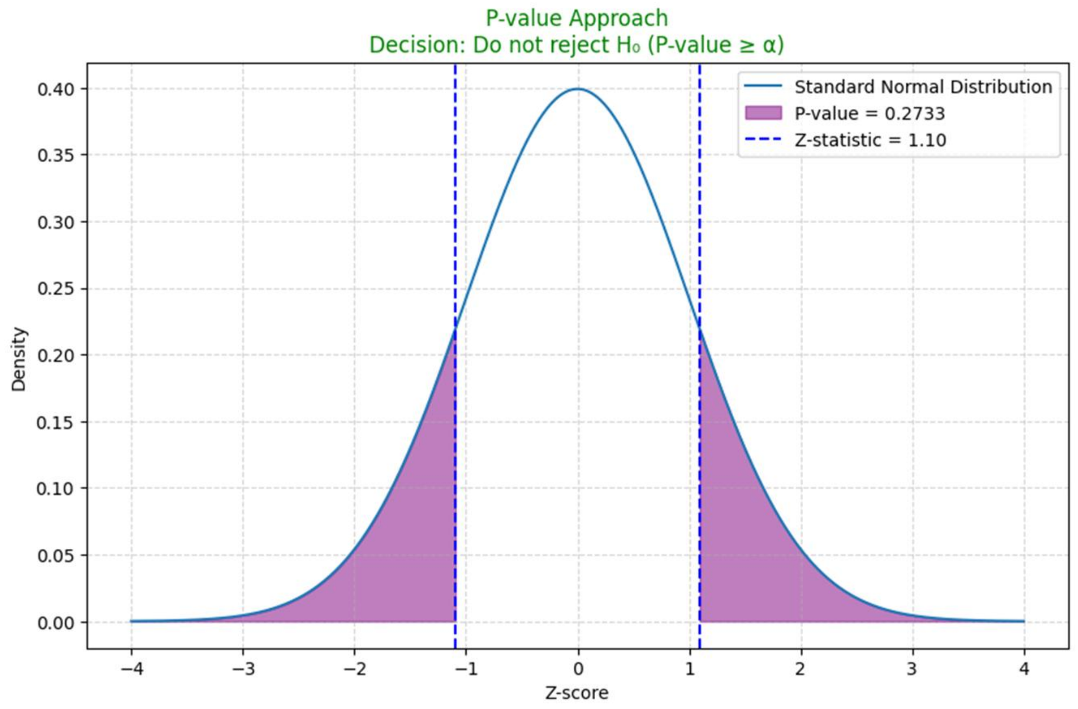
   if $|z| > z^*$ *or* $|t| > t^*$ **then**
   Reject H_0
   ⇒ Statistically significant
   else
   Fail to reject H_0
   ⇒ Not significant
7. **Interpretation:**
   - **Compare test statistic to critical value**



B.3 p – value Approach.

1. **State Hypotheses:**
   - $\mathbf{H_0}: \mu = \mu_0$
   - $\mathbf{H_1}: \mu \neq \mu_0$
2. **Choose Significance Level:**
   - $\alpha = \mathbf{0.05}$ or $\mathbf{0.01}$
3. **Check Assumptions:**
   - **Normality, sample size requirements**
4. **Calculate Test Statistic:**
   - $\mathbf{z} = \dfrac{\bar{\mathbf{x}} - \mu_0}{\mathbf{SE}}$ or **t**-statistic
5. **Find p-value:**
   - $\mathbf{P}(|\mathbf{Z}| > |\mathbf{z}|)$ or $\mathbf{P}(|\mathbf{T}| > |\mathbf{t}|)$
6. **Decision Rule:**

   if *p-value* $< \alpha$ **then**
   Reject H_0
   $\Rightarrow$ **Statistically significant**
   else
   Fail to reject H_0
   $\Rightarrow$ **Not significant**
7. **Interpretation:**
   - **p-value represents strength of evidence against** H_0



1. Variance Based Testing.

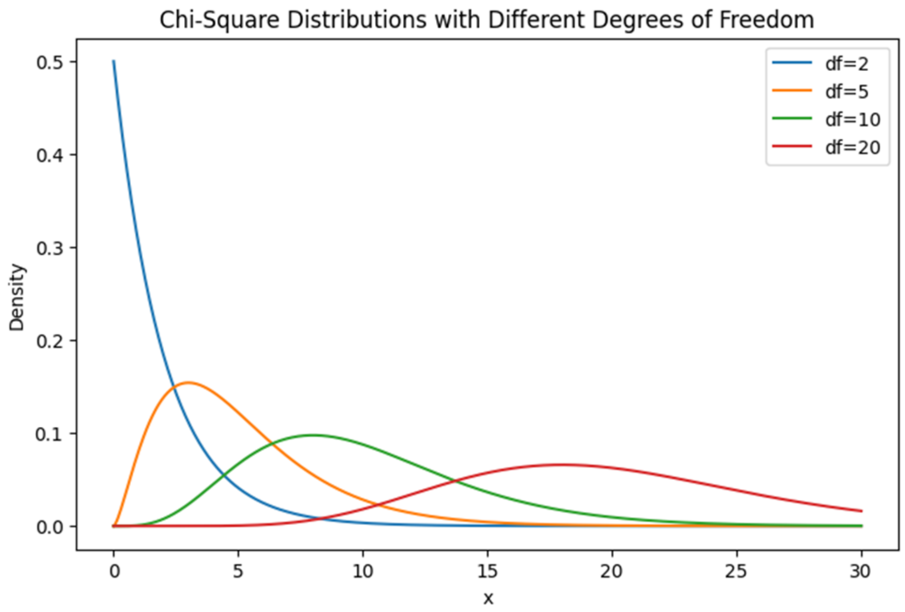
{ Chi – Square Distribution }

1.1 Motivation: Why Test Variance?

- In business, we rarely care only about the ***average***:
  - we **often also want to know** whether **differences** between groups or categories are
    - **real** or **just due to chance**.
  - **Example 1 – Customer Satisfaction:**
    - You run a survey for 3 service channels (Phone, Email, Live Chat)
    - Average ratings: **Phone** $= \mathbf{4.3}$; **Email** $= \mathbf{4.0}$, **Live Chat** $= \mathbf{4.1}$;
    - Question: Are these differences meaningful or random?
  - **Example 2 – Market Segmentation:**
    - You test **if product preference depends on region**.
    - Data is categorical: **North**, **South**, **East**, **West**.
    - Question: Is there an association between region and product choice?
- Sometimes differences in sample data are due to *natural random variation*.
  - We need **statistical tools** to decide whether **these differences are significant**.

1.2 Chi – square Distribution.

- **Numerically**, **the variance is different from the mean because**
  - it can not be negative, so it won't make sense to use **a normal or t distribution.**
- In order to do **hypothesis testing for a variance**,
  - we need to learn a little bit about a new distribution, the **chi-square distribution**.
- Key properties of the **chi – square Distribution**:
  - If $\mathbf{X_1, X_2, \dots, X_n}$ are **independent** **standard normal variables** **then**:
    - $\mathbf{Chi\ -Square\ (\chi^2) = X_1^2 + X_2^2 + \cdots + X_k^2}$
    - follows a **chi – square distribution with df = k.**
  - Notations → χ^2_{df}
  - **Shape Right skewed**, **starts at 0, extends to +∞**
  - Parameters:
    - **Degrees of freedom (df) – determines the shape.**
    - **Mean: μ = df**
    - **Variance: $\sigma^2 = 2 \times$ df**



1.3 Two Main Tools.

Test	When to Use	Data Type
Chi – square Test	**Relationship between categorical variables**	Categorical
ANOVA	**Compare means across 3 + groups**	• Continuous (numerical) outcome, • categorical group variable.

2. Variance Based Testing.

{ Chi – Square Test }

2.1 Chi – Square Test – Core Idea.

- **Motivation:**
  - The **chi – square** (χ^2) **test** measures **how different observed counts** are **from expected counts**.
  - If **differences** are **too large** to be **explained by chance**,
    - **we conclude** that **the distribution or relationship is statistically significant**.
- **Types of Chi – Square Test:**
  1. **Goodness of Fit Test:**
     - **Purpose: Tests if sample data matches a hypothesized distribution**.
     - **When to use:**
       - Comparing **one categorical variable** to a **theoretical distribution**.
       - Testing **proportions** against **benchmarks.**
     - **Examples:**
       - **Marketing:** *"Do our website visitors follow the expected 40%/30%/30% split by age group?*
       - **Operations:** *"Is our defect rate (3%) consistent across all products?*
       - **HR:** *" Does our hiring ratio match the applicant pool's demographics?"*
     - **In General: "Does our observed data fit what we expected?"**

2.1 Chi – Square Test – Core Idea.

- **Motivation:**
  - The **chi – square (χ^2) test** measures **how different observed counts** are **from expected counts**.
  - If **differences** are **too large** to be **explained by chance**,
    - **we conclude** that **the distribution or relationship is statistically significant**.
- **Types of Chi – Square Test:**
  2. Test of Independence:
     - **Purpose:** Tests if two categorical variables are related.
     - **When to use:**
       - Analyzing **two categorical variables** simultaneously.
       - Checking **for dependencies between two factors**.
     - **Examples:**
       - **Retail:** *"Is purchase frequency (low/med/high) related to customer region?"*
       - **Heatlcare:** *"Does treatment outcome (success/failure) depend on clinic location?"*
     - *In General: "Are these two factors associated.?"*

2.2 Case: Goodness of Fit

- **Derivation of Chi – square Statistic:**
  - **In Goodness – of – fit :**
    - We want to compare **observed frequencies O_i** and **expected frequencies E_i** for **k categories**.
      - **Step 1 – Measure difference for each category:**
          - $\mathbf{Difference = O_i - E_i}$
        - But difference can be positive or negative.
        - To avoid them cancelling out → square them.
          - $\mathbf{Difference = (O_i - E_i)^2}$
      - **Step 2 – Weight by expected count:**
        - Large differences in categories with small expected counts are more unusual than in large counts.
        - So, we normalize by E_i, i.e.
          - $\frac{(O_i - E_i)^2}{E_i}$

2.2.1 Case: Goodness of Fit

- **Step 3 – Sum across Categories:**
  - The chi – square statistic is:
    - $\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}}$
- **Step 4 – Sampling distribution:**
  - Under the null hypothesis, χ^2 follows a chi – square distribution with:
    - $\mathbf{df = k - 1}$

2.2.2 Example – Goodness – of – fit Test.

- **Scenario:**
  - An **e – commerce business claims that**
    - **50% of purchase come from mobile**, **30% from desktop**, and **20% from tablet.**
  - You collect a **random sample of 400 purchases**:

Device	Observed (O_i)
Mobile	**220**
Desktop	**120**
Tablet	**60**

  - **Null Hypothesis (H_0):**
    - The distribution of purchases by device matches the company's claim:
      - $p_{Mobile} = 0.50;\ p_{Desktop} = 0.30;\ p_{Tablet} = 0.20$
  - **Alternative Hypothesis (H_a):**
    - The distribution of purchases by device does not match the company's claim:
      - **At least one of $p_{Mobile}, p_{Desktop}, p_{Tablet}$ differs from the claim.**

2.2.2 Solutions – Goodness – of – Fit Test:

1. **Data & Claimed proportions:**
   - Claimed proportions:
     - $\mathbf{p_{Mobile} = 0.50;\ p_{Desktop} = 0.30;\ p_{Tablet} = 0.20}$
   - Sample Size: $\mathbf{n = 400}$
   - Observed Counts: $\mathbf{O_i}$: $\mathbf{Mobile = 220; Desktop = 120; Tablet = 60.}$
2. **Expected counts under H_0:**
   - $\mathbf{E_i = n \times p_i}$ **(Why ?)**

Device	Observed (O_i)	Expected $(E_i = n \times p_i)$
Mobile	220	$E_{Mobile} = 400 \times 0.50 = 200.$
Desktop	120	$E_{Desktop} = 400 \times 0.30 = 120.$
Tablet	60	$E_{Tablet} = 400 \times 0.20 = 80.$

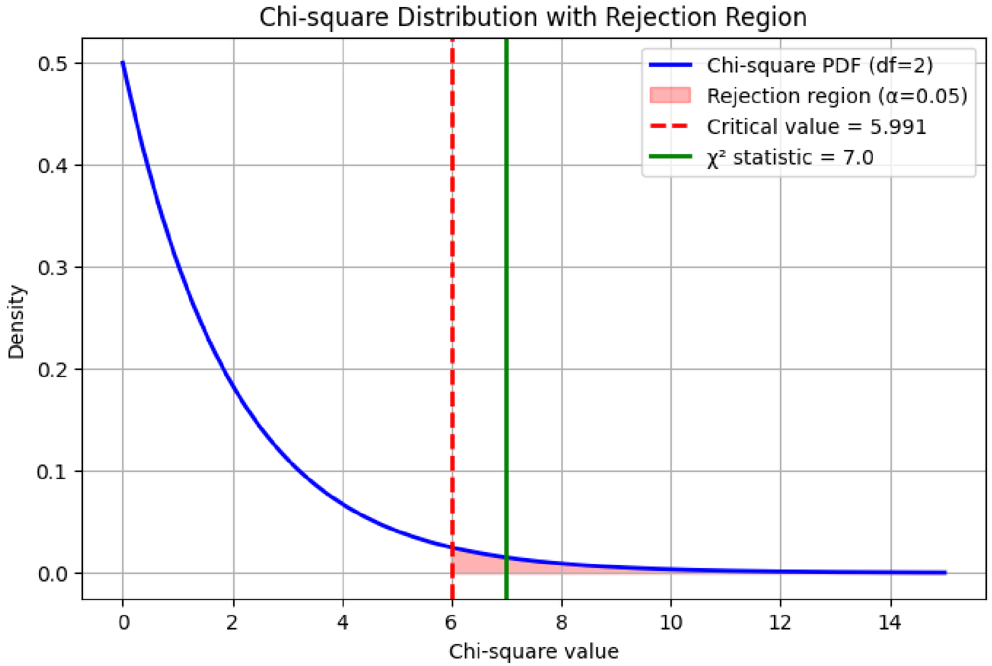
2.2.2 Solutions – Goodness – of – Fit Test:

3. Compute the **chi – square statistic**:
   - $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
- **Compute Each term:**

Device	Observed (O_i)	Expected $(E_i = n \times p_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Mobile	220	$E_{Mobile} = 400 \times 0.50 = 200.$	$(220 - 200)^2 = 400$	$\frac{400}{200} = 2$
Desktop	120	$E_{Desktop} = 400 \times 0.30 = 120.$	$(120 - 120)^2 = 0$	0
Tablet	60	$E_{Tablet} = 400 \times 0.20 = 80.$	$(60 - 80)^2 = 400$	$\frac{400}{80} = 5$
			$\chi^2 \Rightarrow$	$2 + 0 + 5 = 7.0$

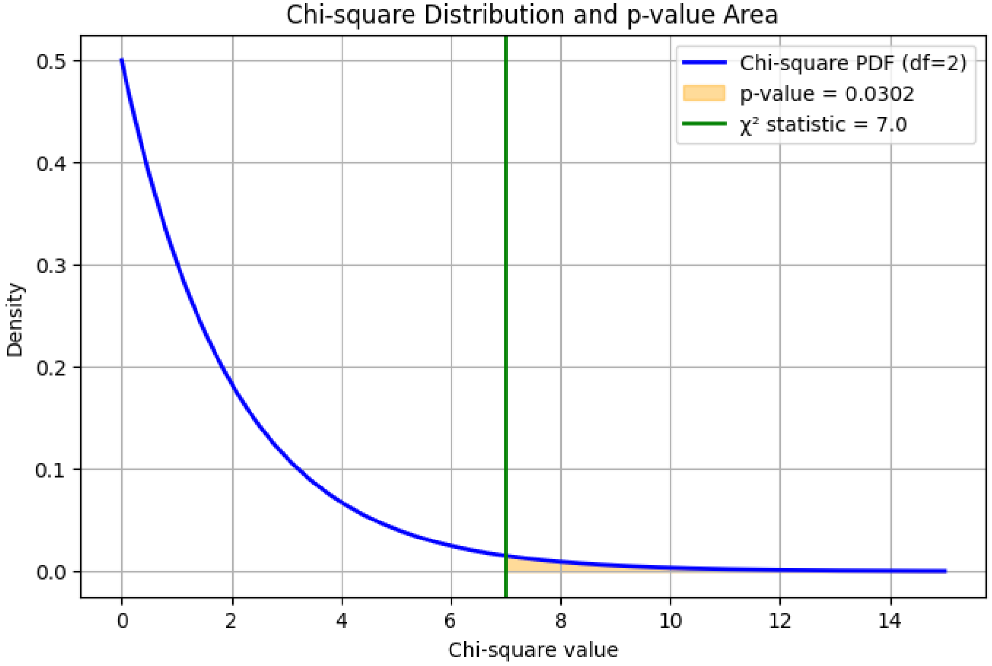
2.2.2 Solutions – Goodness – of – Fit Test:

4. **Degrees of Freedom:**
   - For goodness – of – fit with k categories:
     - $\mathbf{df = k - 1 = 3 - 1 = 2}$
5. **Critical value ($\alpha = 0.05$) and decision rule:**
   - **Critical value** from $\chi^2_{df=2}$ **distribution** at **0.05 significance**:
     - $\boxed{\chi^2_{0.95,df=2} \approx 5.991}$
   - **Decision rule** @ $\alpha = 0.05$:
     - **If** $\chi^2 > 5.991 \rightarrow$ **reject** H_0
       - **Here** $\boxed{\chi^2 = 7.0 > 5.991 \rightarrow \text{reject } H_0}$



2.2.2 Solutions – Goodness – of – Fit Test:

6. **p – value :**
   - p-value is the probability of observing a chi-square at least as **large a 7 under H_0**
     - $\mathbf{p - value = P\left(\chi^2_{df=2} \geq 7.0\right) \approx 0.0302}$
   - Because $\boxed{0.0302 < 0.05}$ this agrees with the **decision to reject H_0**.



2.2.2 Solutions – Goodness – of – Fit Test:

7. **Conclusion & Business Interpretation:**
   i. **Statistical conclusion:**
      - **Reject the null hypothesis**.
        - The **observed device distribution (220,120,60)** is
          - **significantly different** from the **claimed (50%, 30%, 20%) at the 5% level**.

   ii. **Business insight:**
      - Mobile purchases are higher than claimed (220 vs 200 expected);
      - tablet purchases are lower (60 vs 80 expected).

   iii. **The company should investigate:**
      - *Why mobile appears stronger (better UX, targeting, or promotion)?*
      - *Why tablet underperforms (poor UI, tracking issues, or lower traffic)?*

   iv. **Actionable next steps:**
      - drill down by segments (region, time, campaign),
      - check tracking,
      - consider reallocating optimization resources.

2.3 Case: Test of Independence.

- **Derivation of Chi – square Statistic:**
  - While the **Goodness-of-Fit** test checks whether a **single categorical variable** matches an expected distribution,
    the **Test of Independence** examines whether **two categorical variables are related**.
  - **Step 1 – Setup:**
    - We **collect data** in an **r × c contingency table:**

	Category 1	Category 2	⋯	Total
Group 1	O_{11}	O_{12}	⋯	$\sum O_{11} + O_{12} + \cdots$
Group 2	O_{21}	O_{22}	⋯	$\sum O_{21} + O_{22} + \cdots$
⋯	⋯	⋯	⋯	⋯
Total	$\sum O_{11} + O_{21} + \cdots$	$\sum O_{12} + O_{22} + \cdots$	⋯	

Table: r × c contigency Table.

2.3.1 Case: Test of Independence.

- **Step 2 – Expected Counts:**
  - If the **variables are independent:**
    - $\mathbf{E_{ij}} = \frac{(\text{Row Total}_i)\times(\text{Column Total}_j)}{\text{Grand Total}}$
  - If the **variables are not independent:**
    - **then the above formula no longer represents the expected counts**,
      - because this formula is derived under the null hypothesis $\mathbf{H_0}$
  - **Why?**
    - **Under independence**, the probability of being in **cell**$(\mathbf{i},\mathbf{j})$ is simply:
      - $\mathbf{P(Row_i) \times P(Column_j)}$
    - Which leads to the **above multiplication formula**.
  - If the **variables are dependent**,
    - the **probability of being in** $(\mathbf{i},\mathbf{j})$ is not **the product of marginal probabilities**
      - **it depends on the joint distribution**.

2.3.1 Case: Test of Independence.

- **Step 3 – Chi – square Statistic:**
  - **Same formula, but applied to all cells:**
    - $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
  - **Degrees of freedom:**
    - $df = (r-1) \cdot (c-1)$
  - When variables are **not independent**,
    - the actual **observed counts O_{ij}** will **systematically deviate** from these **expected counts.**
    - **The chi – square statistic:**
      - $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
      - ***will be large, leading us to reject H_0.***

2.3.2 Example – Test of Independence.

- **Scenario:**
  - A **retailer wants to know** if **the device type is related to purchase completion**.

Device	Purchased (Yes)	Purchased (No)	Total
Mobile	180	120	300
Desktop	140	160	300
Tablet	60	40	100
Total	380	320	700

  - **Null and Alternative Hypothesis:**
    - **H_0: Device type and purchase completion are independent (no relationship).**
    - **H_a: Device type and purchase completion are not independent (there is a relationship).**

2.3.3 Solution – Test – of – Independence.

1. **Observed Frequencies Table (From Data) ⇒**

Device	Purchased (Yes)	Purchased (No)	Total
Mobile	180	120	300
Desktop	140	160	300
Tablet	60	40	100
Total	380	320	700

- $E_{ij} = \frac{(\text{Row Total})\times(\text{Column Total})}{\text{Grand Total}}$
- Computed for each cell.

2. **Expected Frequencies $\left(E_{ij}\right)$ ⇒**

Device	Purchased (Yes)	Purchased (No)
Mobile	$E_{11} = \frac{300 \times 380}{700} \approx 162.857$	$E_{12} = \frac{300 \times 320}{700} \approx 137.143$
Desktop	$E_{21} = \frac{300 \times 380}{700} \approx 162.857$	$E_{22} = \frac{300 \times 320}{700} \approx 137.143$
Tablet	$E_{31} = \frac{100 \times 380}{700} \approx 54.286$	$E_{32} = \frac{100 \times 320}{700} \approx 45.714$

Expected Table.

2.3.3 Solution – Test – of – Independence.

Device	Purchased (Yes)	Purchased (No)	Total
Mobile	180	120	300
Desktop	140	160	300
Tablet	60	40	100
Total	380	320	700

Device	Purchased (Yes)	Purchased (No)
Mobile	$E_{11} = \frac{300 \times 380}{700} \approx 162.857$	$E_{12} = \frac{300 \times 320}{700} \approx 137.143$
Desktop	$E_{21} = \frac{300 \times 380}{700} \approx 162.857$	$E_{22} = \frac{300 \times 320}{700} \approx 137.143$
Tablet	$E_{31} = \frac{100 \times 380}{700} \approx 54.286$	$E_{32} = \frac{100 \times 320}{700} \approx 45.714$
Expected Table.		

3. Compute Chi – square Statistic ⇒

$$\chi^2 \text{statistic} = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

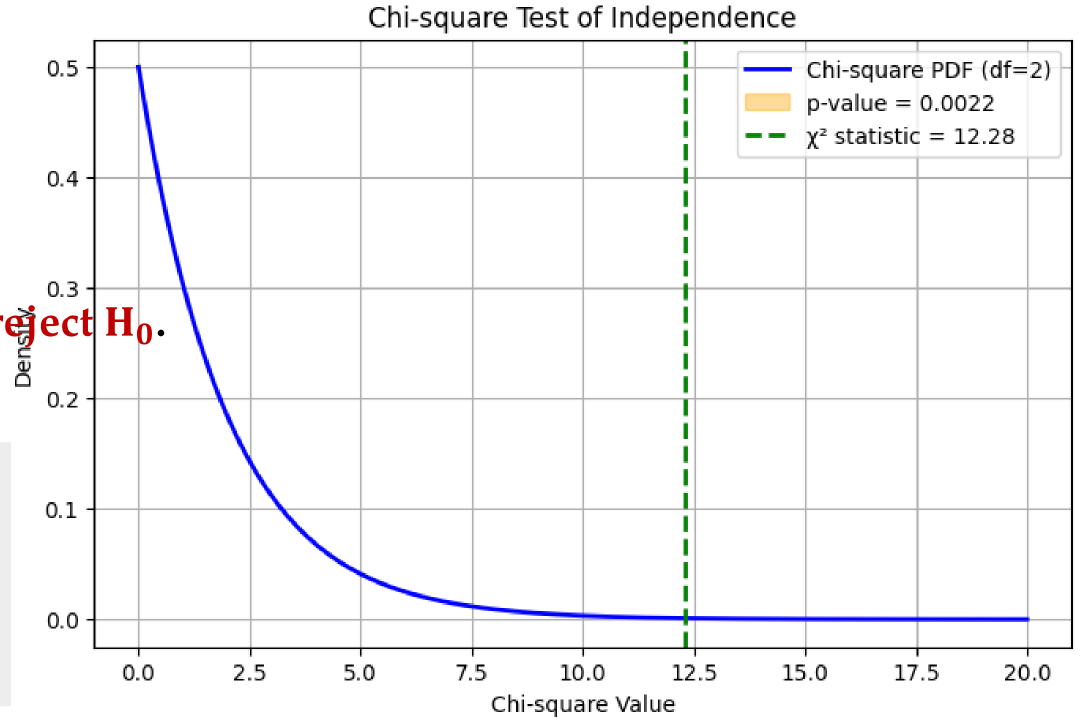
{Repeat for each cell with correct value.}

Device	Purchased (Yes)	Purchased (No)
Mobile	$\frac{(180 - 162.857)^2}{162.857} \approx 1.800$	$\frac{(120 - 137.143)^2}{137.143} \approx 2.145$
Desktop	$\frac{(140 - 162.857)^2}{162.857} \approx 3.207$	$\frac{(160 - 137.143)^2}{137.143} \approx 3.807$
Tablet	$\frac{(60 - 54.286)^2}{54.286} \approx 0.603$	$\frac{(40 - 45.714)^2}{45.714} \approx 0.714$
χ^2 → Each cell value		
Grand Total ⇒ $\chi^2 = 1.800 + 2.145 + 3.207 + 3.807 + 0.603 + 0.714 = 12.276.$		

4. Degrees of Freedom:
   - $df = (r - 1) \times (C - 1) = (3 - 1) \times (2 - 1) = 2.$

2.3.3 Solution – Test – of – Independence.

5. **Critical Value & Decision:**
   - At $\alpha = 0.05$, $\chi^2_{0.95,df=2} \approx 5.991$
   - Since: $\boxed{12.276 > 5.991 \rightarrow \text{we reject } H_0}$
6. **p – value:**
   - Formula:
     - $p = P(\chi^2_{df=2} \geq 12.276) \approx 0.0021$
   - Because $\boxed{0.0021 < 0.05}$ this agrees with the **decision to reject H_0**.



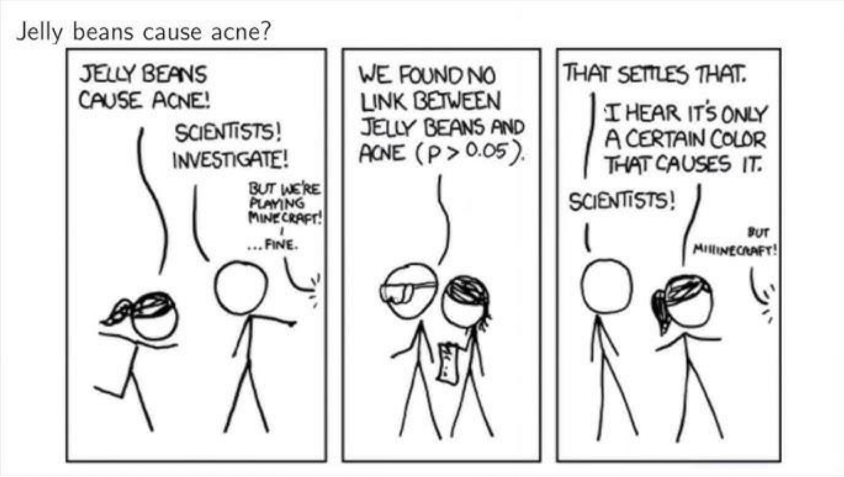
7. **Interpretation (Business Context)**
   - Since $p \approx 0.0021 < 0.05$, there is strong evidence that
     - **device type and purchase completion are related**.
   - From a business standpoint,
     - this means user experience and conversion strategy may need to be tailored
       - **differently** for mobile, desktop, and tablet users.

“Chi-Square works for categorical data, but what if we want to test differences in means across multiple groups?”

{ 3. ANOVA Test}

3.1 Motivation for ANOVA.

- In **business analytics,** you often want to **compare more than two groups**.
  - Does **average sales** differ between regions?
  - Do **conversion rates** differ across multiple website designs?
  - Does **average customer satisfaction** vary between service channels?
- If you only **had two groups**, you could **use a t – test**.
  - But **what if you have 3 or more groups?**
    - Doing **multiple t – tests increase** the **risk of Type – I error (false positives).**
      - **ANOVA** solves this **by testing all groups at once**.



3.2 What is ANOVA?

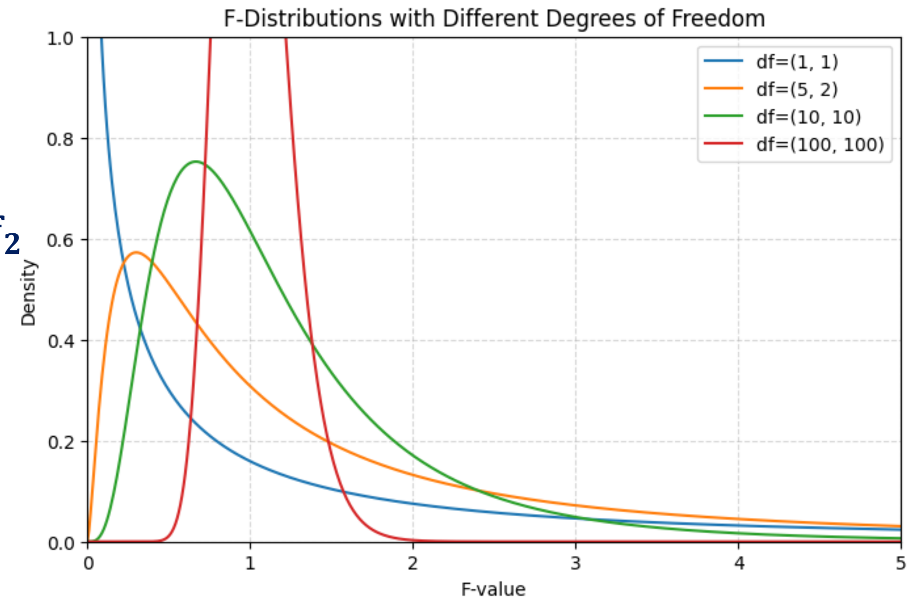
- **ANOVA – Analysis of Variance** checks whether **the means of multiple groups are significantly different**.
  - **Null Hypothesis H_0: All group means are equal.**
  - **Alternative Hypothesis H_a: At least one group is different.**
- **Key Idea:**
  - ANOVA compares:
    - **Between-group variability**: how far group means are from the overall mean.
    - **Within-group variability**: how spread-out values are inside each group.
  - If **between-group variability** is large compared to **within-group variability**,
    - *the means are probably different.*

3.3 Special Distribution for ANOVA.

- **Why not t – distribution?**
  - The **t-distribution** is used to compare **means of two groups** (or a sample mean against a known mean)
    - basically, **one variance estimate** is involved in the denominator (**the estimate of standard error**).
  - ANOVA compares **more than two groups**, so the **simple two-group comparison logic doesn't hold**.
  - Also, the ANOVA test statistic is a **ratio of two variances**,
    - not a **difference of means standardized by a standard error**.
  - The **t-distribution** arises when you have a **single sample variance estimate**
    - and want to **standardize a mean difference**, not when comparing *ratios* of **two variance estimates**.
- **Why not chi square distribution?**
  - The **chi-square distribution** describes the distribution of a **single variance estimate** (sum of squared deviations normalized) based on **normally distributed data**.
  - In ANOVA, we have **two variance estimates**:
    - Variance **between groups** (based on differences of group means from overall mean)
    - Variance **within groups** (based on variability inside groups)
  - The **chi-square distribution** describes **only one** variance estimate at a time, not the ratio of two.

3.3.1 Why the F – distribution?

- The **F – distribution** models the distribution of the ratio of
  - two **independent chi – square variables**,
    - each divided by their degrees of freedom – exactly what the **ANOVA test statistic** is:
  - Formally if:
    - $\mathbf{X \sim \frac{\chi^2_{df_1}}{df_1}}$ **and** $\mathbf{Y \sim \frac{\chi^2_{df_2}}{df_2}}$
    - are independent, then
      - $\mathbf{F = \frac{X}{Y}}$
    - follows an **F – distribution** with **degrees of freedom** $\mathbf{df_1}$ **and** $\mathbf{df_2}$
- This perfectly matches the ANOVA setup where:
  - Numerator = variance estimate between groups (scaled chi – square)
  - Denominator = variance estimate within groups (scaled chi – square)



3.4 Statistic for ANOVA.

- F – statistics Formula:
  - $F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}}$
- Where:
  - Variance Between Groups = MSB = $\frac{SSB}{df_{between}}$
  - Variance Between Groups = MSW = $\frac{SSW}{df_{within}}$

3.5 Types of ANOVA.

- There are **several types** of ANOVA, each suited to **different experimental designs and data structures**.
  - Here's a **quick overview** of the **two main types**:
    1. **One – way ANOVA:**
       - **Purpose:**
         - Compare means across **one categorical independent variable** with **3 or more groups**.
       - **Example:**
         - Comparing **average sales across**
           - **3 marketing channels (Social Media, Email, Organic Search).**
    2. **Two – way ANOVA**:
       - **Purpose:**
         - Compare means across **two categorical independent variables simultaneously**,
           - and check for **interaction effects** between them.
       - **Example:**
         - Comparing ***average sales across marketing channel*** **and** ***region***
           - **(e.g., Social Media vs Email, across North and South regions).**
       - Allows you to see:
         - Main effect of each factor
         - Interaction effect between factors

3.6 Example – One Way ANOVA.

- **Derivation:**
  - **Theory and Formulas for ONE – Way ANOVA:**
    - Suppose you have **k groups** with **sample sizes** $\mathbf{n_1, n_2 \ldots, n_k}$ and **observations** $\mathbf{X_{ij}}$,
      - **where i indexes' groups and j indexes observations within groups**.
  - **Goal:**
    - Test:
      - $\mathbf{H_0: \mu_1 = \mu_2 = \cdots = \mu_k}$
    - Vs.
      - $\mathbf{H_a}$**: At least one group mean differs.**

3.6 Example – One Way ANOVA - Steps.

Steps a) Compute group means and overall mean: fonttitle

1. **Group mean:**

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

2. **Overall mean:**

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$$

where:

$$N = \sum_{i=1}^{k} n_i$$

3.6 Example – One Way ANOVA - Steps.

Steps b) Compute Sum of Squares

1. **Between Groups (SSB):**

$$\mathbf{SSB} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

2. **Within Groups (SSW):**

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

3. **Total (SST):**

$$\mathbf{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \mathbf{SSB} + \mathbf{SSW}$$

3.6 Example – One Way ANOVA - Steps.

Steps c) Degrees of Freedom

1. **Between Groups:**

$$\mathbf{df_{between} = k - 1}$$

2. **Within Groups:**

$$\mathbf{df_{within} = N - k}$$

3. **Total:**

$$\mathbf{df_{total} = N - 1}$$

3.6 Example – One Way ANOVA - Steps.

Steps d) Mean Squares

1. **Mean Square Between (MSB):**

$$\mathrm{MSB} = \frac{\mathrm{SSB}}{\mathrm{df_{between}}}$$

2. **Mean Square Within (MSW):**

$$\mathrm{MSW} = \frac{\mathrm{SSW}}{\mathrm{df_{within}}}$$

3.6 Example – One Way ANOVA - Steps.

Steps e) Compute the F-statistic

$$\mathbf{F} = \frac{\mathbf{MSB}}{\mathbf{MSW}}$$

Steps f) Decision Rule

- Compare **F** to the critical value $F\_\alpha, df\_between, df\_within$ or compute the p-value:

$$\mathbf{p} = \mathbf{P}\left(\mathbf{F}_{\mathbf{df}_{\mathbf{between}},\mathbf{df}_{\mathbf{within}}} \geq \mathbf{F}_{\mathbf{obs}}\right)$$

- Reject $\mathbf{H_0}$ if $\mathbf{p} < \alpha$. Otherwise, do not reject $\mathbf{H_0}$.

3.7 Example – Two Way ANOVA.

Definition and Goal of Two-Way ANOVA

Two-way ANOVA is used to examine the influence of two categorical factors on a continuous response variable, including possible interaction effects between the factors.

Null hypotheses:

$$\begin{cases} H_{0A} : \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{a\cdot} & \text{(No effect of Factor A)} \\ H_{0B} : \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot b} & \text{(No effect of Factor B)} \\ H_{0AB} : \text{No interaction between Factors A and B} \end{cases}$$

Alternate hypotheses: At least one group mean differs for Factor A or Factor B, or there is an interaction effect.

3.7.1 Example – Two Way ANOVA - Steps.

Step a) Compute Marginal Means

- **Row means:**

$$\bar{X}_{i\cdot} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} X_{ijk}$$

- **Column means:**

$$\bar{X}_{\cdot j} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} X_{ijk}$$

- **Overall mean:**

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} X_{ijk}$$

3.7.1 Example – Two Way ANOVA - Steps.

Step b) Compute Sums of Squares

Total SS: $SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(X_{ijk} - \bar{X})^2$

Factor A SS: $SSA = \sum_{i=1}^{a} n_{i.}(\bar{X}_{i.} - \bar{X})^2$

Factor B SS: $SSB = \sum_{j=1}^{b} n_{.j}(\bar{X}_{.j} - \bar{X})^2$

Interaction SS: $SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}(\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2$

Within SS: $SSW = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(X_{ijk} - \bar{X}_{ij})^2$

3.7.1 Example – Two Way ANOVA - Steps.

Step c) Degrees of Freedom

$$
\begin{aligned}
df_A &= a - 1 \\
df_B &= b - 1 \\
df_{AB} &= (a-1)(b-1) \\
df_{within} &= N - ab \\
df_{total} &= N - 1
\end{aligned}
$$

3.7.1 Example – Two Way ANOVA - Steps.

Step d) Compute Mean Squares

$$MS_A = \frac{SSA}{df_A}$$
$$MS_B = \frac{SSB}{df_B}$$
$$MS_{AB} = \frac{SSAB}{df_{AB}}$$
$$MS_{within} = \frac{SSW}{df_{within}}$$

3.7.1 Example – Two Way ANOVA - Steps.

Step e) Compute F-Statistics

$$F_A = \frac{MS_A}{MS_{within}}, \quad F_B = \frac{MS_B}{MS_{within}}, \quad F_{AB} = \frac{MS_{AB}}{MS_{within}}$$

Step f) Decision Rule

- For each factor and the interaction, compare the F statistic to the critical value F_{α,df_1,df_2} or compute the p-value.
- Reject the null hypothesis for that factor or interaction if $p < \alpha$.
- Interpret the main effects and interaction effects accordingly.

Thank You.