



Lecture 14

Statistics II

STAT-221 • UG

Dr. Rizwan Jahangir

NUST Business School

National University of Sciences & Technology

- **1. One-Sample Test of a Population Proportion:**
 - ▶ Nominal Data Transition & The Rule of Thumb ($n\pi \geq 5$, $n(1 - \pi) \geq 5$).
 - ▶ Step-by-Step Solved Business Example.
- **2. Two-Sample Test of Population Proportions:**
 - ▶ Comparing Multi-Group Proportions & Labeled pooled proportion (p_c).
 - ▶ Step-by-Step Solved Business Example.
- **3. Nonparametric Methods: Goodness-of-Fit:**
 - ▶ Hypothesis Test of Equal Expected Frequencies.
 - ▶ Hypothesis Test of Unequal Expected Frequencies.
- **4. Contingency Table Analysis (Test for Independence):**
 - ▶ Expected Frequencies in standard grid $f_e = (RT \times CT)/GT$.
 - ▶ Degrees of freedom $df = (r - 1)(c - 1)$.



One-Sample Test of a Proportion

Transition to Nominal Data Hypotheses

In previous lectures, we tested claims about population means using interval- or ratio-scaled data. Now, we shift to nominal-level data.

- **Nominal Data:** Observations are classified into mutually exclusive categories with no inherent ordering (e.g., Yes/No, Accept/Reject, Diesel/Gasoline).
- Instead of looking at means (μ), we look at the **proportion** of successes in the population (π).
- The sample proportion is calculated as:

$$p = \frac{x}{n}$$

Where x is the number of successes in a sample of size n .



Required Assumptions for the Proportion Test

Before applying a standard normal z-test, we must ensure our sample size is large enough to justify the normal approximation to the binomial distribution.

The Rule of Thumb Check:

The sample size n must be large enough such that both of the following conditions are met:

$$n\pi \geq 5 \quad \text{and} \quad n(1 - \pi) \geq 5$$

Where π is the hypothesized population proportion.



Required Assumptions for the Proportion Test

Before applying a standard normal z-test, we must ensure our sample size is large enough to justify the normal approximation to the binomial distribution.

The Rule of Thumb Check:

The sample size n must be large enough such that both of the following conditions are met:

$$n\pi \geq 5 \quad \text{and} \quad n(1 - \pi) \geq 5$$

Where π is the hypothesized population proportion.

If these conditions are satisfied, the sampling distribution of p is approximately normal, allowing us to use standard normal z-tables.



The One-Sample Proportion Test Statistic

To test a claim about a population proportion, we calculate the standard normal z-test statistic:

One-Sample Proportion z-Test Statistic:

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Where:

- p is the calculated sample proportion (x/n).
- π is the hypothesized population proportion.
- n is the sample size.



Step-by-Step Example 1: Employee Willingness to Relocate

Textbook Example: Willingness to Relocate

The Struthers Wells Corporation employs more than 10,000 white-collar workers in its sales and manufacturing offices. Historically, 35% of workers were willing to relocate outside the United States ($\pi = 0.35$).

A recent random sample of $n = 300$ workers revealed that $x = 120$ would accept a transfer to a location outside the U.S.

Task: At the **0.05 significance level**, can we conclude that there has been an **increase** in the proportion of workers willing to relocate?



Relocation Example: Steps 1 & 2

Step 1: State the Null and Alternate Hypotheses.

- $H_0 : \pi \leq 0.35$ (No increase in the willingness to relocate)
- $H_1 : \pi > 0.35$ (There is an increase in willingness)
- *Note: The keyword "increase" makes this a right-tailed test.*



Relocation Example: Steps 1 & 2

Step 1: State the Null and Alternate Hypotheses.

- $H_0 : \pi \leq 0.35$ (No increase in the willingness to relocate)
- $H_1 : \pi > 0.35$ (There is an increase in willingness)
- *Note: The keyword "increase" makes this a right-tailed test.*

Step 2: Select the Level of Significance.

- $\alpha = 0.05$.



Relocation Example: Steps 3 & 4

Step 3: Select the Test Statistic.

- We first verify our assumptions:

$$n\pi = 300 \times 0.35 = 105 \geq 5 \quad \text{and} \quad n(1 - \pi) = 300 \times 0.65 = 195 \geq 5$$

- Since both conditions are met, we use the z-statistic: $z = \frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}}$.



Relocation Example: Steps 3 & 4

Step 3: Select the Test Statistic.

- We first verify our assumptions:

$$n\pi = 300 \times 0.35 = 105 \geq 5 \quad \text{and} \quad n(1 - \pi) = 300 \times 0.65 = 195 \geq 5$$

- Since both conditions are met, we use the z-statistic: $z = \frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}}$.

Step 4: Formulate the Decision Rule.

- From the Standard Normal table, the critical value for a right-tailed test at $\alpha = 0.05$ is **+1.645**.
- **Decision Rule:** Reject H_0 if computed $z > +1.645$.



Relocation Example: Step 5 (Calculations)

Let's calculate our sample proportion and computed test statistic:

① **Compute Sample Proportion (p):**

$$p = \frac{x}{n} = \frac{120}{300} = \mathbf{0.40}$$



Relocation Example: Step 5 (Calculations)

Let's calculate our sample proportion and computed test statistic:

① Compute Sample Proportion (p):

$$p = \frac{x}{n} = \frac{120}{300} = \mathbf{0.40}$$

② Compute Standard Error (σ_p):

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} = \sqrt{\frac{0.35 \times 0.65}{300}} = \sqrt{\frac{0.2275}{300}} = \sqrt{0.0007583} = \mathbf{0.0275}$$



Relocation Example: Step 5 (Calculations)

Let's calculate our sample proportion and computed test statistic:

① Compute Sample Proportion (p):

$$p = \frac{x}{n} = \frac{120}{300} = \mathbf{0.40}$$

② Compute Standard Error (σ_p):

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.35 \times 0.65}{300}} = \sqrt{\frac{0.2275}{300}} = \sqrt{0.0007583} = \mathbf{0.0275}$$

③ Compute Test Statistic (z):

$$z = \frac{p - \pi}{\sigma_p} = \frac{0.40 - 0.35}{0.0275} = \frac{0.05}{0.0275} = \mathbf{1.82}$$



Relocation Example: Step 6 (Decision & Interpretation)

Mathematical Decision:

- Computed $z = 1.82$.
- Critical $z = 1.645$.
- Since $1.82 > 1.645$, it falls into the rejection region.



Relocation Example: Step 6 (Decision & Interpretation)

Mathematical Decision:

- Computed $z = 1.82$.
- Critical $z = 1.645$.
- Since $1.82 > 1.645$, it falls into the rejection region.

We **Reject** H_0 at the 5% significance level.



Relocation Example: Step 6 (Decision & Interpretation)

Mathematical Decision:

- Computed $z = 1.82$.
- Critical $z = 1.645$.
- Since $1.82 > 1.645$, it falls into the rejection region.

We **Reject** H_0 at the 5% significance level.

Business Interpretation:

- There is a statistically significant increase in the proportion of white-collar workers willing to accept a transfer outside the United States.
- The current willingness has risen significantly from the historical baseline of 35%, now reaching 40% in our sample. Ms. Wanda Carter can update human resource plans accordingly!



Two-Sample Test of Proportions

Comparing Proportions Across Two Populations

Often, a business analyst needs to determine whether there is a significant difference between the proportions of two independent populations.

- Are baby-boomers and millennials equally likely to purchase a Malibu?
- Is the proportion of acceptable parts from Supplier 1 equal to Supplier 2?
- We want to test:

$$H_0 : \pi_1 = \pi_2 \quad \text{vs.} \quad H_1 : \pi_1 \neq \pi_2$$



Comparing Proportions Across Two Populations

Often, a business analyst needs to determine whether there is a significant difference between the proportions of two independent populations.

- Are baby-boomers and millennials equally likely to purchase a Malibu?
- Is the proportion of acceptable parts from Supplier 1 equal to Supplier 2?
- We want to test:

$$H_0 : \pi_1 = \pi_2 \quad \text{vs.} \quad H_1 : \pi_1 \neq \pi_2$$

Since the null hypothesis assumes that the two proportions are identical ($\pi_1 = \pi_2$), we combine the data from both samples to create a single shared estimate.



The Pooled Proportion (p_c)

To estimate the assumed common population proportion under H_0 , we combine the successes and sample sizes from both groups.

Pooled Proportion Formula (p_c):

$$p_c = \frac{x_1 + x_2}{n_1 + n_2}$$

Where:

- x_1, x_2 are the number of successes in Sample 1 and Sample 2.
- n_1, n_2 are the sample sizes of Sample 1 and Sample 2.



The Two-Sample Proportions Test Statistic

Using the pooled proportion (p_c), we compute our z-statistic:

Two-Sample Proportions z-Test Statistic:

$$z = \frac{p_1 - p_2}{\sqrt{p_c(1 - p_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where:

- $p_1 = x_1/n_1$ and $p_2 = x_2/n_2$ are the two sample proportions.
- p_c is the pooled proportion.



Step-by-Step Example 2: Manufacturing Defects

Textbook Example: Supplier Defect Rates

An operations manager is comparing the proportion of acceptable lightbulbs delivered by two suppliers: **Supplier 1** and **Supplier 2**.

- A sample of $n_1 = 100$ bulbs from Supplier 1 revealed that $x_1 = 19$ were unacceptable.
- A sample of $n_2 = 200$ bulbs from Supplier 2 revealed that $x_2 = 62$ were unacceptable.

Task: At the **0.05 significance level**, can the manager conclude that there is a **difference** in the proportion of unacceptable bulbs between the two suppliers? ($H_0 : \pi_1 = \pi_2$ vs. $H_1 : \pi_1 \neq \pi_2$).



Supplier Example: Steps 1 to 4

Step 1: State the Hypotheses.

- $H_0 : \pi_1 = \pi_2$ (No difference in the unacceptable rates)
- $H_1 : \pi_1 \neq \pi_2$ (There is a difference in unacceptable rates)

Step 2: Level of Significance. $\alpha = 0.05$.



Supplier Example: Steps 1 to 4

Step 1: State the Hypotheses.

- $H_0 : \pi_1 = \pi_2$ (No difference in the unacceptable rates)
- $H_1 : \pi_1 \neq \pi_2$ (There is a difference in unacceptable rates)

Step 2: Level of Significance. $\alpha = 0.05$. **Step 3: Select Test Statistic.** z-statistic.



Supplier Example: Steps 1 to 4

Step 1: State the Hypotheses.

- $H_0 : \pi_1 = \pi_2$ (No difference in the unacceptable rates)
- $H_1 : \pi_1 \neq \pi_2$ (There is a difference in unacceptable rates)

Step 2: Level of Significance. $\alpha = 0.05$. Step 3: Select Test Statistic. z-statistic. Step 4: Formulate the Decision Rule.

- This is a two-tailed test, so we divide α by 2: $\alpha/2 = 0.025$ in each tail.
- Critical values = ± 1.96 .
- **Decision Rule:** Reject H_0 if computed $z > +1.96$ or $z < -1.96$.



Supplier Example: Step 5 (Calculations Part 1)

First, calculate individual and pooled proportions:

① Sample Proportions:

$$p_1 = \frac{19}{100} = \mathbf{0.19} \quad \text{and} \quad p_2 = \frac{62}{200} = \mathbf{0.31}$$



Supplier Example: Step 5 (Calculations Part 1)

First, calculate individual and pooled proportions:

① Sample Proportions:

$$p_1 = \frac{19}{100} = \mathbf{0.19} \quad \text{and} \quad p_2 = \frac{62}{200} = \mathbf{0.31}$$

② Pooled Proportion (p_c):

$$p_c = \frac{x_1 + x_2}{n_1 + n_2} = \frac{19 + 62}{100 + 200} = \frac{81}{300} = \mathbf{0.27}$$



Supplier Example: Step 5 (Calculations Part 1)

First, calculate individual and pooled proportions:

① Sample Proportions:

$$p_1 = \frac{19}{100} = \mathbf{0.19} \quad \text{and} \quad p_2 = \frac{62}{200} = \mathbf{0.31}$$

② Pooled Proportion (p_c):

$$p_c = \frac{x_1 + x_2}{n_1 + n_2} = \frac{19 + 62}{100 + 200} = \frac{81}{300} = \mathbf{0.27}$$

Now, let's calculate the standard error of difference on the next slide!



Supplier Example: Step 5 (Calculations Part 2)

Calculate the test statistic z :

$$z = \frac{p_1 - p_2}{\sqrt{p_c(1 - p_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$



Supplier Example: Step 5 (Calculations Part 2)

Calculate the test statistic z :

$$z = \frac{p_1 - p_2}{\sqrt{p_c(1 - p_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$z = \frac{0.19 - 0.31}{\sqrt{0.27 \times 0.73 \left(\frac{1}{100} + \frac{1}{200} \right)}}$$



Supplier Example: Step 5 (Calculations Part 2)

Calculate the test statistic z :

$$z = \frac{p_1 - p_2}{\sqrt{p_c(1 - p_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$z = \frac{0.19 - 0.31}{\sqrt{0.27 \times 0.73 \left(\frac{1}{100} + \frac{1}{200} \right)}}$$

$$z = \frac{-0.12}{\sqrt{0.1971 \times (0.01 + 0.005)}} = \frac{-0.12}{\sqrt{0.1971 \times 0.015}}$$



Supplier Example: Step 5 (Calculations Part 2)

Calculate the test statistic z :

$$z = \frac{p_1 - p_2}{\sqrt{p_c(1 - p_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$z = \frac{0.19 - 0.31}{\sqrt{0.27 \times 0.73 \left(\frac{1}{100} + \frac{1}{200} \right)}}$$

$$z = \frac{-0.12}{\sqrt{0.1971 \times (0.01 + 0.005)}} = \frac{-0.12}{\sqrt{0.1971 \times 0.015}}$$

$$z = \frac{-0.12}{\sqrt{0.0029565}} = \frac{-0.12}{0.05437} = -\mathbf{2.21}$$



Supplier Example: Step 6 (Conclusion)

Mathematical Decision:

- Computed $z = -2.21$.
- Critical $z = \pm 1.96$.
- Since $-2.21 < -1.96$, it falls into the left-tail rejection region.



Supplier Example: Step 6 (Conclusion)

Mathematical Decision:

- Computed $z = -2.21$.
- Critical $z = \pm 1.96$.
- Since $-2.21 < -1.96$, it falls into the left-tail rejection region.

We **Reject** H_0 at the 5% significance level.



Supplier Example: Step 6 (Conclusion)

Mathematical Decision:

- Computed $z = -2.21$.
- Critical $z = \pm 1.96$.
- Since $-2.21 < -1.96$, it falls into the left-tail rejection region.

We **Reject** H_0 at the 5% significance level.

Business Interpretation:

- There is a statistically significant difference in the proportion of unacceptable bulbs between the two suppliers.
- Supplier 1 (19% unacceptable) has a significantly lower defect rate than Supplier 2 (31%).
- **Action:** The purchasing manager should prioritize Supplier 1 to improve factory quality!



15 Minute Break

Nonparametric Statistics

What is Nonparametric Hypothesis Testing?

So far, our test procedures have assumed that the underlying populations follow specific continuous distributions (e.g., standard normal or t-distributions).

- **Parametric Tests:** Require assumptions about the population parameters and distributions (such as assuming normal populations and interval/ratio scales).



What is Nonparametric Hypothesis Testing?

So far, our test procedures have assumed that the underlying populations follow specific continuous distributions (e.g., standard normal or t-distributions).

- **Parametric Tests:** Require assumptions about the population parameters and distributions (such as assuming normal populations and interval/ratio scales).
- **Nonparametric Tests:** Do **not** require any assumptions regarding the shape of the population. They are distribution-free tests!
- These tests are highly suitable when working with nominal or ordinal scales, or when sample sizes are too small to verify normality.



Characteristics of the Chi-Square (χ^2) Distribution

To perform nonparametric goodness-of-fit tests, we utilize a new probability distribution: the **Chi-Square (χ^2) Distribution**.

Five Key Characteristics of the χ^2 Distribution:

- ❶ **Non-Negative:** Values are always zero or positive because calculations involve squared differences.
- ❷ **Positively Skewed:** The distribution curve has a long tail pointing to the right.
- ❸ **Family of Curves:** Shaped by a single parameter: **Degrees of Freedom (df)**.
- ❹ **Asymptotic:** As degrees of freedom increase, the distribution approaches a symmetric bell-shape.
- ❺ **One-Tailed:** Goodness-of-fit and independence rejection regions are located strictly in the **right tail**.

Goodness-of-Fit Tests

Introducing Goodness-of-Fit Tests

Definition: Chi-Square Goodness-of-Fit Test

A nonparametric test used to compare an observed set of frequencies (sample data) with an expected frequency distribution.



Introducing Goodness-of-Fit Tests

Definition: Chi-Square Goodness-of-Fit Test

A nonparametric test used to compare an observed set of frequencies (sample data) with an expected frequency distribution.

Core Concepts:

- **Observed Frequencies (f_o):** The actual number of observations falling into each category from our sample data.
- **Expected Frequencies (f_e):** The number of observations we would expect to see in each category under our null hypothesis.
- **Equal expected frequencies:** The null hypothesis claims that observations are distributed equally across all categories.



The Chi-Square Test Statistic

To evaluate how well our observed data fits the expected distribution, we compute the χ^2 test statistic:

Chi-Square (χ^2) Test Statistic Formula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Where:

- f_o represents the observed frequency in each category.
- f_e represents the expected frequency in each category.
- The sum is taken over all k categories.
- **Degrees of Freedom (df):** $k - 1$.



Expected Frequencies in the Equal Distribution Model

When our null hypothesis assumes that all categories are equally utilized, the expected frequency for each of the k classes is simply:

Expected Frequency for Equal Classes:

$$f_e = \frac{n}{k}$$

Where:

- n is the total sample size across all categories ($n = \sum f_o$).
- k is the number of categories.



Step-by-Step Example 3: Building Entrance Traffic

Textbook Example: Equal Expected Frequencies

A building maintenance supervisor at a government center wants to know if the building's four entrances are utilized equally ($k = 4$ categories).

A sample of $n = 400$ people are observed entering the building:

- Main Street Entrance: $f_o = 140$
- Broad Street Entrance: $f_o = 120$
- Cherry Street Entrance: $f_o = 90$
- Walnut Street Entrance: $f_o = 50$

Task: At the **0.01 significance level**, is there a significant difference in the utilization of the four entrances?



Entrance Traffic: Steps 1 to 4

Step 1: State the Hypotheses.

- H_0 : The four entrances are equally utilized ($\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$).
- H_1 : The four entrances are NOT equally utilized (Frequencies differ).

Step 2: Level of Significance. $\alpha = 0.01$.



Entrance Traffic: Steps 1 to 4

Step 1: State the Hypotheses.

- H_0 : The four entrances are equally utilized ($\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$).
- H_1 : The four entrances are NOT equally utilized (Frequencies differ).

Step 2: Level of Significance. $\alpha = 0.01$. **Step 3: Select Test Statistic.** Chi-Square ($\chi^2 = \sum(f_o - f_e)^2/f_e$).



Entrance Traffic: Steps 1 to 4

Step 1: State the Hypotheses.

- H_0 : The four entrances are equally utilized ($\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$).
- H_1 : The four entrances are NOT equally utilized (Frequencies differ).

Step 2: Level of Significance. $\alpha = 0.01$. **Step 3: Select Test Statistic.** Chi-Square ($\chi^2 = \sum(f_o - f_e)^2/f_e$). **Step 4: Formulate the Decision Rule.**

- Degrees of freedom: $df = k - 1 = 4 - 1 = 3$.
- Look up $\alpha = 0.01$, $df = 3$ in Appendix B.7 $\rightarrow$ **Critical** $\chi^2 = 11.345$.
- **Decision Rule:** Reject H_0 if computed $\chi^2 > 11.345$.



Entrance Traffic: Step 5 (Calculations Part 1)

First, let's calculate the expected frequency (f_e) for each entrance:

$$f_e = \frac{n}{k} = \frac{400}{4} = \mathbf{100} \text{ people per entrance}$$



Entrance Traffic: Step 5 (Calculations Part 1)

First, let's calculate the expected frequency (f_e) for each entrance:

$$f_e = \frac{n}{k} = \frac{400}{4} = \mathbf{100} \text{ people per entrance}$$

Now, construct our difference table systematically:

Entrance	Observed (f_o)	Expected (f_e)	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
Main Street	140	100	+40	1,600	16.0
Broad Street	120	100	+20	400	4.0
Cherry Street	90	100	-10	100	1.0
Walnut Street	50	100	-50	2,500	25.0



Entrance Traffic: Step 5 (Calculations Part 2)

Sum the final column to find the computed χ^2 value:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 16.0 + 4.0 + 1.0 + 25.0$$



Entrance Traffic: Step 5 (Calculations Part 2)

Sum the final column to find the computed χ^2 value:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 16.0 + 4.0 + 1.0 + 25.0$$

$$\chi^2 = \mathbf{46.00}$$



Entrance Traffic: Step 5 (Calculations Part 2)

Sum the final column to find the computed χ^2 value:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 16.0 + 4.0 + 1.0 + 25.0$$

$$\chi^2 = \mathbf{46.00}$$

Let's evaluate this against our critical value of **11.345** on the next slide!



Entrance Traffic: Step 6 (Conclusion)

Mathematical Decision:

- Computed $\chi^2 = 46.00$.
- Critical $\chi^2 = 11.345$.
- Since $46.00 > 11.345$, it falls far into the right-tail rejection region.



Entrance Traffic: Step 6 (Conclusion)

Mathematical Decision:

- Computed $\chi^2 = 46.00$.
- Critical $\chi^2 = 11.345$.
- Since $46.00 > 11.345$, it falls far into the right-tail rejection region.

We **Reject** H_0 at the 1% significance level.



Entrance Traffic: Step 6 (Conclusion)

Mathematical Decision:

- Computed $\chi^2 = 46.00$.
- Critical $\chi^2 = 11.345$.
- Since $46.00 > 11.345$, it falls far into the right-tail rejection region.

We **Reject** H_0 at the 1% significance level.

Business Interpretation:

- There is a highly significant difference in the utilization of the four entrances.
- Main and Broad Street entrances are heavily over-utilized, while the Walnut Street entrance is severely under-utilized.
- **Action:** The supervisor should reallocate security guards or maintenance staff to the Main and Broad Street doors!



Goodness-of-Fit with Unequal Expected Frequencies

Sometimes, our null hypothesis does not assume equal distributions. Instead, we compare our observed data to historical, industry-wide, or census proportions.

Unequal Expected Frequencies Formula:

$$f_e = n \times p_i$$

Where p_i is the expected population percentage for category i stated in the null hypothesis.



Goodness-of-Fit with Unequal Expected Frequencies

Sometimes, our null hypothesis does not assume equal distributions. Instead, we compare our observed data to historical, industry-wide, or census proportions.

Unequal Expected Frequencies Formula:

$$f_e = n \times p_i$$

Where p_i is the expected population percentage for category i stated in the null hypothesis.

Degrees of Freedom: Still $df = k - 1$ (if no parameters are estimated from the sample).



Step-by-Step Example 4: Accounts Receivable

Textbook Example: Unequal Expected Frequencies

A national credit agency reports that 60% of credit card accounts are active, 30% are late, and 10% are uncollectible ($\pi_{\text{active}} = 0.60$, $\pi_{\text{late}} = 0.30$, $\pi_{\text{uncol}} = 0.10$).

A regional credit manager reviews a sample of $n = 500$ accounts and finds:

- Active Accounts: $f_o = 320$
- Late Accounts: $f_o = 120$
- Uncollectible Accounts: $f_o = 60$

Task: At the **0.05 significance level**, do the regional accounts reflect the national distribution? (H_0 is as above).



Accounts Receivable: Step-by-Step Calculations

Step 1: Hypotheses. $H_0 : P_A = 0.60, P_L = 0.30, P_U = 0.10$ vs. H_1 : The distribution differs.

Step 2 & 3: Critical Value. $df = 3 - 1 = 2$. Critical $\chi^2_{2,0.05} = \mathbf{5.991}$ (Table B.7).



Accounts Receivable: Step-by-Step Calculations

Step 1: Hypotheses. $H_0 : P_A = 0.60, P_L = 0.30, P_U = 0.10$ vs. H_1 : The distribution differs.

Step 2 & 3: Critical Value. $df = 3 - 1 = 2$. Critical $\chi^2_{2,0.05} = \mathbf{5.991}$ (Table B.7).

Step 4: Compute Expected Frequencies (f_e).

- Active: $f_e = 500 \times 0.60 = \mathbf{300}$
- Late: $f_e = 500 \times 0.30 = \mathbf{150}$
- Uncollectible: $f_e = 500 \times 0.10 = \mathbf{50}$



Accounts Receivable: Step-by-Step Calculations

Step 1: Hypotheses. $H_0 : P_A = 0.60, P_L = 0.30, P_U = 0.10$ vs. H_1 : The distribution differs.

Step 2 & 3: Critical Value. $df = 3 - 1 = 2$. Critical $\chi^2_{2,0.05} = \mathbf{5.991}$ (Table B.7).

Step 4: Compute Expected Frequencies (f_e).

- Active: $f_e = 500 \times 0.60 = \mathbf{300}$
- Late: $f_e = 500 \times 0.30 = \mathbf{150}$
- Uncollectible: $f_e = 500 \times 0.10 = \mathbf{50}$

Step 5: Compute χ^2 Statistic.

$$\chi^2 = \frac{(320 - 300)^2}{300} + \frac{(120 - 150)^2}{150} + \frac{(60 - 50)^2}{50} = \frac{400}{300} + \frac{900}{150} + \frac{100}{50}$$



Accounts Receivable: Step-by-Step Calculations

Step 1: Hypotheses. $H_0 : P_A = 0.60, P_L = 0.30, P_U = 0.10$ vs. H_1 : The distribution differs.

Step 2 & 3: Critical Value. $df = 3 - 1 = 2$. Critical $\chi^2_{2,0.05} = \mathbf{5.991}$ (Table B.7).

Step 4: Compute Expected Frequencies (f_e).

- Active: $f_e = 500 \times 0.60 = \mathbf{300}$
- Late: $f_e = 500 \times 0.30 = \mathbf{150}$
- Uncollectible: $f_e = 500 \times 0.10 = \mathbf{50}$

Step 5: Compute χ^2 Statistic.

$$\chi^2 = \frac{(320 - 300)^2}{300} + \frac{(120 - 150)^2}{150} + \frac{(60 - 50)^2}{50} = \frac{400}{300} + \frac{900}{150} + \frac{100}{50}$$

$$\chi^2 = 1.33 + 6.00 + 2.00 = \mathbf{9.33}$$



Accounts Receivable: Decision & Interpretation

Mathematical Decision:

- Computed $\chi^2 = 9.33$.
- Critical $\chi^2 = 5.991$.
- Since $9.33 > 5.991$, it falls in our rejection region.



Accounts Receivable: Decision & Interpretation

Mathematical Decision:

- Computed $\chi^2 = 9.33$.
- Critical $\chi^2 = 5.991$.
- Since $9.33 > 5.991$, it falls in our rejection region.

We **Reject** H_0 at the 5% significance level.



Accounts Receivable: Decision & Interpretation

Mathematical Decision:

- Computed $\chi^2 = 9.33$.
- Critical $\chi^2 = 5.991$.
- Since $9.33 > 5.991$, it falls in our rejection region.

We **Reject** H_0 at the 5% significance level.

Business Interpretation:

- The regional accounts receivable portfolio does not reflect the national distribution.
- Looking at the data, the region has significantly fewer late accounts than expected (120 vs 150) but slightly more uncollectible accounts (60 vs 50).
- The credit manager should investigate the collection policy to address uncollectible accounts!



Core Limitations of the Chi-Square Statistic

While the Chi-Square test is incredibly flexible, there are strict limits to prevent mathematical distortion.

The Small Frequency Warning:

If the expected frequency (f_e) in any category cell is too small, the calculated χ^2 statistic becomes artificially inflated!



Core Limitations of the Chi-Square Statistic

While the Chi-Square test is incredibly flexible, there are strict limits to prevent mathematical distortion.

The Small Frequency Warning:

If the expected frequency (f_e) in any category cell is too small, the calculated χ^2 statistic becomes artificially inflated!

Two Rules of Thumb to Prevent Distortion:

- 1 For goodness-of-fit with $df = 1$ (2 categories), each expected frequency f_e must be **at least 5**.
- 2 For goodness-of-fit with $df \geq 2$, no more than 20% of cells should have expected frequencies below 5, and **no cell should have an expected frequency of 0**.
- 3 **Solution:** Combine adjacent categories to increase expected frequencies!



Contingency Table Analysis

Categorical Associations: Bivariate Nominal Data

In simple correlation analysis, we measured relationships between two continuous variables. What if we want to study the relationship between **two categorical nominal-level variables**?

- Is there an association between a customer's gender and whether they order dessert?
- Is the choice of pension plan related to an employee's job classification?
- Does the type of bus engine (diesel vs. gasoline) relate to low or high maintenance costs?



Categorical Associations: Bivariate Nominal Data

In simple correlation analysis, we measured relationships between two continuous variables. What if we want to study the relationship between **two categorical nominal-level variables**?

- Is there an association between a customer's gender and whether they order dessert?
- Is the choice of pension plan related to an employee's job classification?
- Does the type of bus engine (diesel vs. gasoline) relate to low or high maintenance costs?

To solve this, we construct a **Contingency Table** and perform a **Chi-Square Test for Independence**.



Formulating the Hypotheses for Independence

In a test of independence, our hypotheses are always structured as follows:

Contingency Table Hypotheses:

- H_0 : The two nominal variables are completely independent of one another.
- H_1 : The two nominal variables are dependent (associated/related).



Formulating the Hypotheses for Independence

In a test of independence, our hypotheses are always structured as follows:

Contingency Table Hypotheses:

- H_0 : The two nominal variables are completely independent of one another.
- H_1 : The two nominal variables are dependent (associated/related).

If we reject H_0 , we conclude that a statistically significant relationship exists between the two categorical variables.



Calculating Expected Frequencies in a Grid

Unlike goodness-of-fit tests, our expected frequencies for independence are not specified by a historical ratio. Instead, we calculate them from the table row and column totals:

Expected Frequency for Cell (Row i , Column j):

$$f_e = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

Where:

- Row Total is the sum of observations in row i .
- Column Total is the sum of observations in column j .
- Grand Total (n) is the total number of observations in the entire sample.



Degrees of Freedom for Independence

Because we are working in a two-dimensional grid of rows and columns, our degrees of freedom are defined by the number of independent cells we can freely fill:

Contingency Table Degrees of Freedom:

$$df = (r - 1)(c - 1)$$

Where:

- r is the total number of rows in our contingency table.
- c is the total number of columns in our contingency table.
- *Note: Row and column totals are excluded from row/column counts.*



Step-by-Step Example 5: Gender vs. Dessert Orders

Textbook Example: Restaurant Marketing

A restaurant director wishes to study the relationship between a guest's gender ($r = 2$) and whether they order dessert ($c = 2$).

A random sample of $n = 200$ recent customers is summarized in the contingency table:

Dessert Ordered	Male	Female	Total
Yes	32	15	47
No	68	85	153
Total	100	100	200

Task: At the **0.05 significance level**, can we conclude that ordering dessert is dependent on the guest's gender?

Gender vs. Dessert: Step 1 to 4 Setup

Step 1: State the Hypotheses.

- H_0 : Ordering dessert is independent of the customer's gender.
- H_1 : Ordering dessert is dependent (associated) on gender.

Step 2: Level of Significance. $\alpha = 0.05$.



Gender vs. Dessert: Step 1 to 4 Setup

Step 1: State the Hypotheses.

- H_0 : Ordering dessert is independent of the customer's gender.
- H_1 : Ordering dessert is dependent (associated) on gender.

Step 2: Level of Significance. $\alpha = 0.05$. **Step 3: Select Test Statistic.** Chi-Square
($\chi^2 = \sum (f_o - f_e)^2 / f_e$).



Gender vs. Dessert: Step 1 to 4 Setup

Step 1: State the Hypotheses.

- H_0 : Ordering dessert is independent of the customer's gender.
- H_1 : Ordering dessert is dependent (associated) on gender.

Step 2: Level of Significance. $\alpha = 0.05$. **Step 3: Select Test Statistic.** Chi-Square ($\chi^2 = \sum (f_o - f_e)^2 / f_e$). **Step 4: Formulate the Decision Rule.**

- Calculate degrees of freedom:

$$df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1 \times 1 = 1$$

- Look up $df = 1, \alpha = 0.05$ in Appendix B.7 $\rightarrow$ **Critical** $\chi^2 = 3.841$.
- **Decision Rule:** Reject H_0 if computed $\chi^2 > 3.841$.



Gender vs. Dessert: Step 5 (Calculations Part 1)

Calculate the expected frequencies (f_e) for each of our four cells:



Gender vs. Dessert: Step 5 (Calculations Part 1)

Calculate the expected frequencies (f_e) for each of our four cells:

- **Male & Yes Dessert:**

$$f_e = \frac{47 \times 100}{200} = \mathbf{23.5}$$



Gender vs. Dessert: Step 5 (Calculations Part 1)

Calculate the expected frequencies (f_e) for each of our four cells:

- **Male & Yes Dessert:**

$$f_e = \frac{47 \times 100}{200} = 23.5$$

- **Female & Yes Dessert:**

$$f_e = \frac{47 \times 100}{200} = 23.5$$



Gender vs. Dessert: Step 5 (Calculations Part 1)

Calculate the expected frequencies (f_e) for each of our four cells:

- **Male & Yes Dessert:**

$$f_e = \frac{47 \times 100}{200} = 23.5$$

- **Female & Yes Dessert:**

$$f_e = \frac{47 \times 100}{200} = 23.5$$

- **Male & No Dessert:**

$$f_e = \frac{153 \times 100}{200} = 76.5$$



Gender vs. Dessert: Step 5 (Calculations Part 1)

Calculate the expected frequencies (f_e) for each of our four cells:

- **Male & Yes Dessert:**

$$f_e = \frac{47 \times 100}{200} = 23.5$$

- **Female & Yes Dessert:**

$$f_e = \frac{47 \times 100}{200} = 23.5$$

- **Male & No Dessert:**

$$f_e = \frac{153 \times 100}{200} = 76.5$$

- **Female & No Dessert:**

$$f_e = \frac{153 \times 100}{200} = 76.5$$



Gender vs. Dessert: Step 5 (Calculations Part 2)

Let's construct our differences table:

Cell (Row, Col)	Observed (f_o)	Expected (f_e)	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
Male, Yes	32	23.5	+8.5	72.25	3.074
Female, Yes	15	23.5	-8.5	72.25	3.074
Male, No	68	76.5	-8.5	72.25	0.944
Female, No	85	76.5	+8.5	72.25	0.944



Gender vs. Dessert: Step 5 (Calculations Part 2)

Let's construct our differences table:

Cell (Row, Col)	Observed (f_o)	Expected (f_e)	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
Male, Yes	32	23.5	+8.5	72.25	3.074
Female, Yes	15	23.5	-8.5	72.25	3.074
Male, No	68	76.5	-8.5	72.25	0.944
Female, No	85	76.5	+8.5	72.25	0.944

Summing the last column: $\chi^2 = 3.074 + 3.074 + 0.944 + 0.944 = \mathbf{8.036}$.



Gender vs. Dessert: Step 6 (Decision & Interpretation)

Mathematical Decision:

- Computed $\chi^2 = 8.036$.
- Critical $\chi^2 = 3.841$.
- Since $8.036 > 3.841$, it falls in our right-tail rejection region.



Gender vs. Dessert: Step 6 (Decision & Interpretation)

Mathematical Decision:

- Computed $\chi^2 = 8.036$.
- Critical $\chi^2 = 3.841$.
- Since $8.036 > 3.841$, it falls in our right-tail rejection region.

We **Reject** H_0 at the 5% significance level.



Gender vs. Dessert: Step 6 (Decision & Interpretation)

Mathematical Decision:

- Computed $\chi^2 = 8.036$.
- Critical $\chi^2 = 3.841$.
- Since $8.036 > 3.841$, it falls in our right-tail rejection region.

We **Reject** H_0 at the 5% significance level.

Business Interpretation:

- Ordering dessert is statistically dependent on the customer's gender.
- Looking at our percentages, Male customers are twice as likely to order dessert compared to Female customers (32% of males ordered dessert vs. 15% of females).
- **Marketing Action:** The restaurant should target dessert advertisements or promotions toward male diners!



Pop-up Question: Contingency Tables

Discussion

Pop-Up Question!

An HR team analyzes sessional promotions across 3 regional branches (Branch A, B, C) and 2 promotion statuses (Promoted vs. Not Promoted).

The table has $r = 2$ rows and $c = 3$ columns.

1. What is the degrees of freedom (df) for this test of independence?
2. If the computed $\chi^2 = 5.23$, and the critical $\chi^2_{2,0.05} = 5.991$, what is your decision?



Pop-up Question: Contingency Tables

Discussion

Pop-Up Question!

An HR team analyzes sessional promotions across 3 regional branches (Branch A, B, C) and 2 promotion statuses (Promoted vs. Not Promoted).

The table has $r = 2$ rows and $c = 3$ columns.

1. What is the degrees of freedom (df) for this test of independence?
2. If the computed $\chi^2 = 5.23$, and the critical $\chi^2_{2,0.05} = 5.991$, what is your decision?

Answers:

- **1:** $df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 1 \times 2 = 2$.
- **2:** Since $5.23 < 5.991$, we **Fail to Reject** H_0 . Promotion rate is independent of the branch location.



Software Lab Application

MegaStat for Excel: Two-Proportions Test

We will now execute these nonparametric analyses using software. On your Windows laptops, open Microsoft Excel with the MegaStat add-in enabled:

- 1 Select the **Add-Ins** tab on the top menu ribbon.
- 2 Select **MegaStat**, choose **Hypothesis Tests**, and click **Compare Two Independent Proportions**.



MegaStat for Excel: Two-Proportions Test

We will now execute these nonparametric analyses using software. On your Windows laptops, open Microsoft Excel with the MegaStat add-in enabled:

- ❶ Select the **Add-Ins** tab on the top menu ribbon.
- ❷ Select **MegaStat**, choose **Hypothesis Tests**, and click **Compare Two Independent Proportions**.
- ❸ In the dialog window:
 - ▶ Group 1: Enter $x = 19$ and $n = 100$.
 - ▶ Group 2: Enter $x = 62$ and $n = 200$.



MegaStat for Excel: Two-Proportions Test

We will now execute these nonparametric analyses using software. On your Windows laptops, open Microsoft Excel with the MegaStat add-in enabled:

- 1 Select the **Add-Ins** tab on the top menu ribbon.
- 2 Select **MegaStat**, choose **Hypothesis Tests**, and click **Compare Two Independent Proportions**.
- 3 In the dialog window:
 - ▶ Group 1: Enter $x = 19$ and $n = 100$.
 - ▶ Group 2: Enter $x = 62$ and $n = 200$.
 - ▶ Select **not equal** as the Alternative Hypothesis, and click **OK**.

MegaStat outputs the pooled proportion (0.270) and computed z-statistic (-2.21), matching our manual calculations perfectly!



MegaStat for Excel: Goodness-of-Fit Tests

To run a Goodness-of-Fit test in MegaStat (such as our Accounts Receivable study):

- 1 Enter your observed frequencies in Column B ('B2:B4') and expected frequencies in Column C ('C2:C4').
- 2 Select **MegaStat** → **Chi-Square/Crosstabs** → **Goodness of Fit Test**.



MegaStat for Excel: Goodness-of-Fit Tests

To run a Goodness-of-Fit test in MegaStat (such as our Accounts Receivable study):

- ❶ Enter your observed frequencies in Column B ('B2:B4') and expected frequencies in Column C ('C2:C4').
- ❷ Select **MegaStat** → **Chi-Square/Crosstabs** → **Goodness of Fit Test**.
- ❸ In the dialog box:
 - ▶ For Observed values, select 'B2:B4'.
 - ▶ For Expected values, select 'C2:C4'.
 - ▶ Enter **0** as the Number of parameters estimated from the data.



MegaStat for Excel: Goodness-of-Fit Tests

To run a Goodness-of-Fit test in MegaStat (such as our Accounts Receivable study):

- ❶ Enter your observed frequencies in Column B ('B2:B4') and expected frequencies in Column C ('C2:C4').
- ❷ Select **MegaStat** → **Chi-Square/Crosstabs** → **Goodness of Fit Test**.
- ❸ In the dialog box:
 - ▶ For Observed values, select 'B2:B4'.
 - ▶ For Expected values, select 'C2:C4'.
 - ▶ Enter **0** as the Number of parameters estimated from the data.
 - ▶ Click **OK**.

MegaStat instantly displays the completed χ^2 table, the computed value (9.33), and the exact P-value (0.0094). Since $0.0094 < 0.05$, we reject the null hypothesis.



MegaStat for Excel: Contingency Tables

To run a test of independence on a contingency table:

- ➊ Input your contingency matrix into cells 'A1' to 'C3'. Include row and column headings, but **do NOT select row or column totals!**
- ➋ Select **MegaStat** → **Chi-Square/Crosstabs** → **Contingency Table**.



MegaStat for Excel: Contingency Tables

To run a test of independence on a contingency table:

- ➊ Input your contingency matrix into cells 'A1' to 'C3'. Include row and column headings, but **do NOT select row or column totals!**
- ➋ Select **MegaStat** → **Chi-Square/Crosstabs** → **Contingency Table**.
- ➌ In the dialog window:
 - ▶ For Input Range, select cells 'A1:C3'.
 - ▶ Check the **chi-square** and **Expected values** options.



MegaStat for Excel: Contingency Tables

To run a test of independence on a contingency table:

- ➊ Input your contingency matrix into cells 'A1' to 'C3'. Include row and column headings, but **do NOT select row or column totals!**
- ➋ Select **MegaStat** → **Chi-Square/Crosstabs** → **Contingency Table**.
- ➌ In the dialog window:
 - ▶ For Input Range, select cells 'A1:C3'.
 - ▶ Check the **chi-square** and **Expected values** options.
 - ▶ Click **OK**.

MegaStat outputs a table displaying your observed counts, calculated expected counts in brackets, the computed χ^2 statistic (8.04), and the P-value (0.0046).

