STATISTICS II

Bachelor's degrees in Economics, Finance and Management

2nd year/2nd Semester
2025/2026

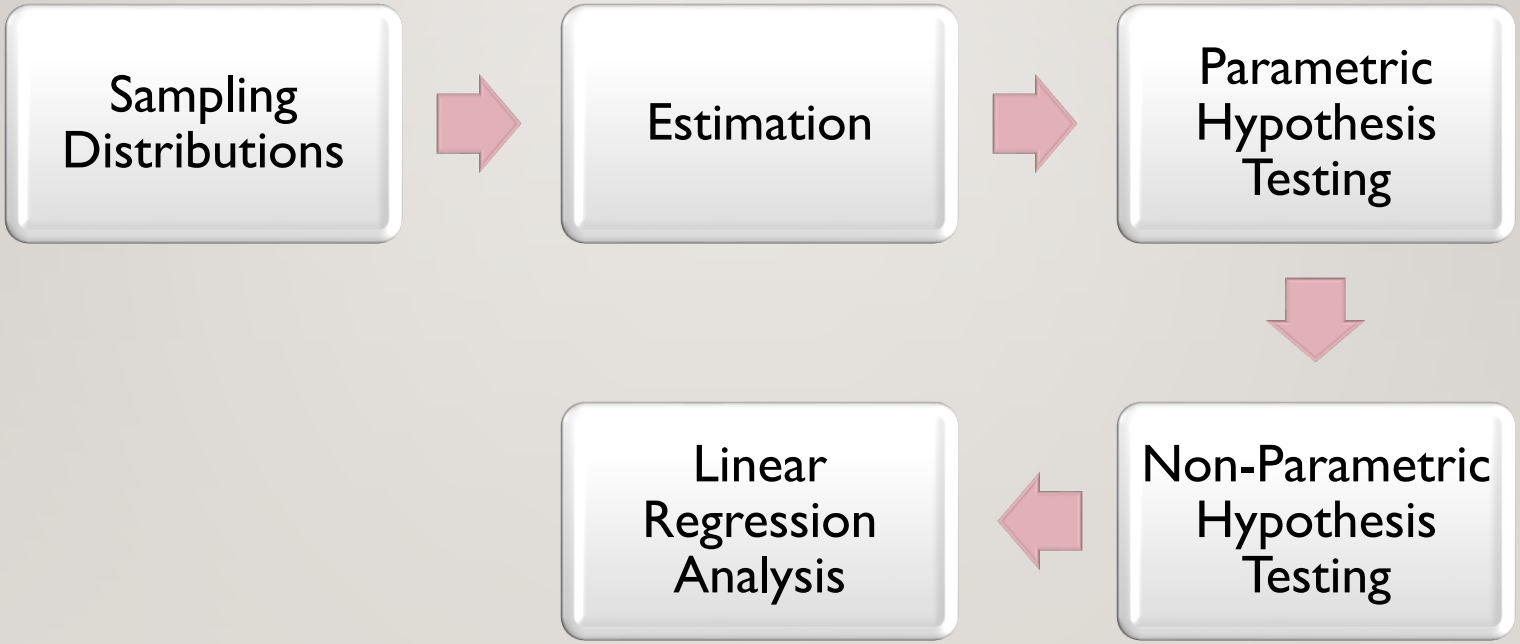
CONTACT

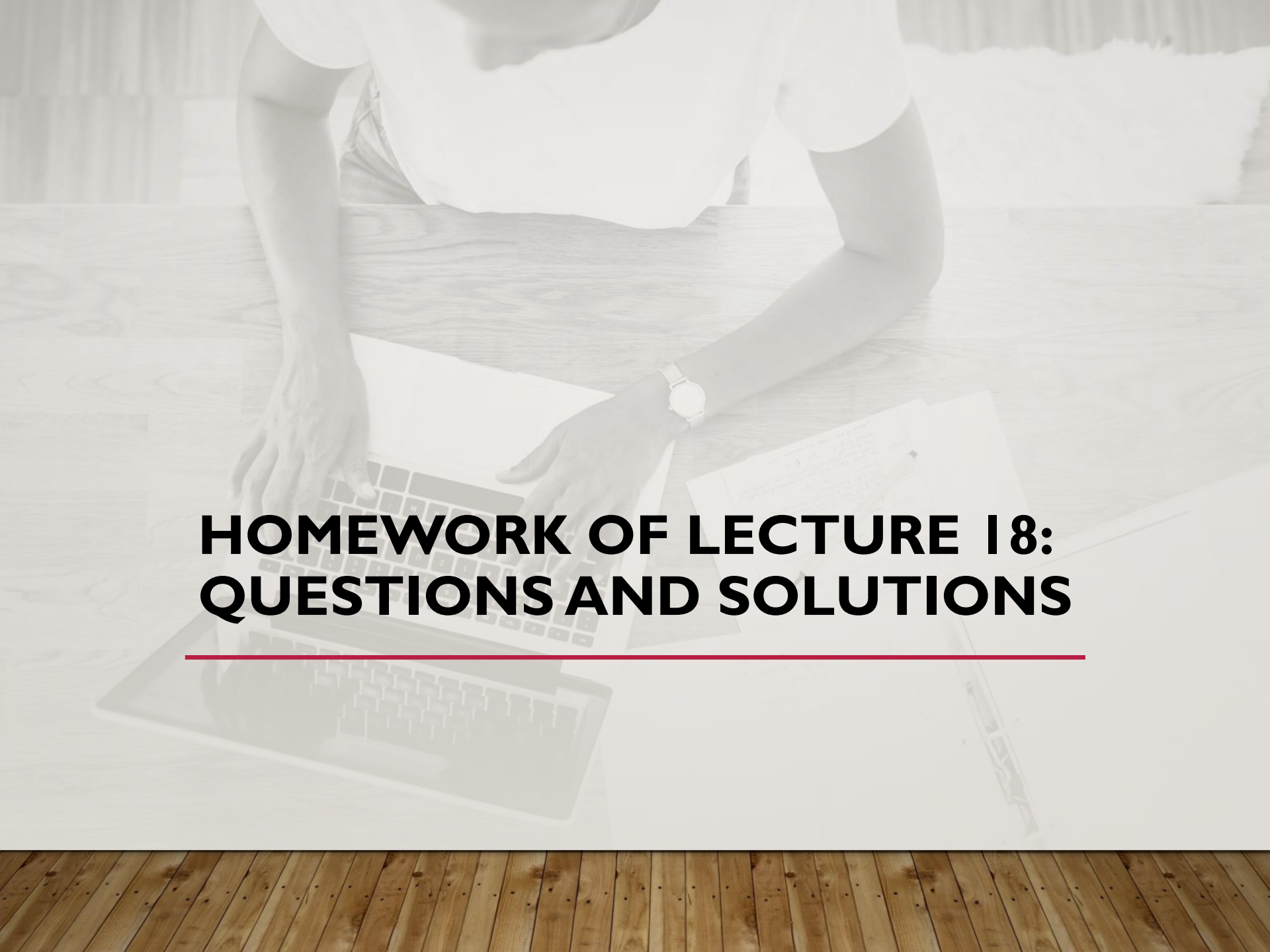
Professor: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

PROGRAM


Sampling Distributions
Estimation
Parametric Hypothesis Testing
Non-Parametric Hypothesis Testing
Linear Regression Analysis



HOMEWORK OF LECTURE 18: QUESTIONS AND SOLUTIONS

EXERCISE 10.15

10.15 Random samples of 900 people in the United States and in Great Britain indicated that 60% of the people in the United States were positive about the future economy, whereas 66% of the people in Great Britain were positive about the future economy. Does this provide strong evidence that the people in Great Britain are more optimistic about the economy?

Newbold et al (2013)

EXERCISE 10.15: SOLUTION

✅ Answer:

Left-Tailed Test

1. Hypotheses

Let p_{US} = proportion of positive respondents in the United States,
and p_{UK} = proportion in Great Britain.

$$H_0 : p_{\text{US}} \geq p_{\text{UK}} \quad \text{(UK is not more optimistic)}$$

$$H_1 : p_{\text{US}} < p_{\text{UK}} \quad \text{(UK is more optimistic)}$$

This is a **left-tailed test**.

2. Sample Data

Country	n	x	$\hat{p}$
US	900	540	0.60
UK	900	594	0.66

EXERCISE 10.15: SOLUTION

✅ Answer:

Step 1: Compute the pooled proportion

$$\hat{p}_0 = \frac{x_{US} + x_{UK}}{n_{US} + n_{UK}} = \frac{0.60 \cdot 900 + 0.66 \cdot 900}{1800} = 0.63$$

Step 2: Compute the standard error under H_0

$$SE = \sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_{US}} + \frac{1}{n_{UK}}\right)} = \sqrt{0.63 \cdot 0.37 \cdot \frac{2}{900}} \approx 0.0228$$

Step 3: Compute the z-statistic

$$z = \frac{\hat{p}_{US} - \hat{p}_{UK}}{SE} = \frac{0.60 - 0.66}{0.0228} = \frac{-0.06}{0.0228} \approx -2.63$$

EXERCISE 10.15: SOLUTION

✅ Answer:

Decision Using P-value

Step 4: Compute the p-value

- This is a **one-tailed test** ($H_1 : p_{US} < p_{UK}$)
- Look up $z = -2.63$ in the standard normal table:

$$p\text{-value} \approx 0.0043$$

Step 5: Conclusion

- Typical significance level: $\alpha = 0.05$
- $p\text{-value} = 0.0043 < 0.05 \rightarrow$ **reject** H_0

✅ **Conclusion:** There is strong evidence that people in the United States are **less optimistic** than people in Great Britain about the future economy.

EXERCISE 10.15: SOLUTION

✅ Answer:

Decision Using RR

5. Critical Value

For a **left-tailed test** at $\alpha = 0.05$:

$$z_{0.05} = -1.645$$

RR =] -∞; -1.645]

6. Decision

$$z = -2.64 < -1.645 \implies \text{reject } H_0$$

7. Conclusion

There is **strong statistical evidence** that a higher proportion of people in **Great Britain** are optimistic about the economy compared to the United States.

Using the **separate variance method** for the standard error confirms the result and is more appropriate when the proportions are different.

EXERCISE 10.16

10.16 A random sample of 1,556 people in country A were asked to respond to this statement: *Increased world trade can increase our per capita prosperity.* Of these sample members, 38.4% agreed with the statement. When the same statement was presented to a random sample of 1,108 people in country B, 52.0% agreed. Test the null hypothesis that the population proportions agreeing with this statement were the same in the two countries against the alternative that a higher proportion agreed in country B.

Newbold et al (2013)

EXERCISE 10.16: SOLUTION

Answer: Right-Tailed Test

1. Hypotheses

Let

$$p_A = \text{proportion agreeing in country A}, \quad p_B = \text{proportion agreeing in country B}$$

$$H_0 : p_B \leq p_A \quad \text{(country B is not more positive)}$$

$$H_1 : p_B > p_A \quad \text{(country B has a higher proportion agreeing)}$$

This is a **right-tailed test**.

2. Sample Data

Country	n	x	$\hat{p}$
A	1,556	0.384×1,556 ≈ 597	0.384
B	1,108	0.520×1,108 ≈ 576	0.520

EXERCISE 10.16: SOLUTION

✅ Answer:

3. Pooled Proportion

$$\hat{p} = \frac{x_A + x_B}{n_A + n_B} = \frac{597 + 576}{1,556 + 1,108} = \frac{1,173}{2,664} \approx 0.440$$

4. Standard Error

$$SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} = \sqrt{0.44 \cdot 0.56\left(\frac{1}{1,556} + \frac{1}{1,108}\right)}$$

$$\frac{1}{1,556} + \frac{1}{1,108} \approx 0.0006428 + 0.000902 = 0.0015448$$

$$SE = \sqrt{0.2464 \cdot 0.0015448} = \sqrt{0.0003804} \approx 0.0195$$

5. Test Statistic

$$z = \frac{\hat{p}_B - \hat{p}_A}{SE} = \frac{0.520 - 0.384}{0.0195} = \frac{0.136}{0.0195} \approx 6.97$$

EXERCISE 10.16: SOLUTION

✅ Answer:

Decision Using RR

6. Critical Value

Right-tailed test at $\alpha = 0.05$:

$$z_{0.95} \approx 1.645$$

RR = [1.645, +∞[

7. Decision

$$z = 6.97 > 1.645 \implies \text{reject } H_0$$

8. Conclusion

There is **very strong statistical evidence** that a **higher proportion of people in country B agree** that increased world trade can increase per capita prosperity, compared to country A.

The observed difference (52% vs 38.4%) is **highly significant** due to the large sample sizes.

EXERCISE 10.24

10.24 It is hypothesized that the total sales of a corporation should vary more in an industry with active price competition than in one with duopoly and tacit collusion. In a study of the merchant ship production industry it was found that in 4 years of active price competition, the variance of company A's total sales was 114.09. In the following 7 years, during which there was duopoly and tacit collusion, this variance was 16.08. Assume that the data can be regarded as an independent random sample from two normal distributions. Test, at the 5% level, the null hypothesis that the two population variances are equal against the alternative that the variance of total sales is higher in years of active price competition.

Newbold et al (2013)

EXERCISE 10.24: SOLUTION

✅ Answer:

Right-Tailed Test

1. Hypotheses

Let

$\sigma^2_{\text{competition}}$ = population variance of total sales in years of active price competition,
$\sigma^2_{\text{duopoly}}$ = population variance of total sales in years of duopoly with tacit collusion.

$$H_0 : \sigma^2_{\text{competition}} \leq \sigma^2_{\text{duopoly}} \quad \text{(variances are equal or lower)}$$

$$H_1 : \sigma^2_{\text{competition}} > \sigma^2_{\text{duopoly}} \quad \text{(variance is higher during competition)}$$

2. Sample Data

Period	n	s^2
Active price competition	4	114.09
Duopoly/collusion	7	16.08

EXERCISE 10.24: SOLUTION

✅ Answer:

3. Test Statistic

$$F = \frac{s^2_{\text{competition}}}{s^2_{\text{duopoly}}} = \frac{114.09}{16.08} \approx 7.10$$

Degrees of freedom:

$$df_1 = n_1 - 1 = 3, \quad df_2 = n_2 - 1 = 6$$

4. Critical Value

For a **right-tailed F-test** at $\alpha = 0.05$ with df1 = 3 and df2 = 6:

$$F_{0.95,3,6} \approx 4.76$$

RR = [4.76, +∞[

EXERCISE 10.24: SOLUTION

✅ Answer:

Decision Using RR

5. Decision

$$F = 7.10 > 4.76 \implies \text{reject } H_0$$

6. Conclusion

At the **5% significance level**, there is **strong evidence** that the variance of total sales is **higher during years of active price competition** compared to years of duopoly with tacit collusion.

This supports the hypothesis that competitive pressure increases variability in total sales.

EXERCISE 10.25

10.25 In light of a number of recent large-corporation bankruptcies, auditors are becoming increasingly concerned about the possibility of fraud. Auditors might be helped in determining the chances of fraud if they carefully measure cash flow. To evaluate this possibility, samples of midlevel auditors from CPA firms were presented with cash-flow information from a fraud case, and they were asked to indicate the chance of material fraud on a scale from 0 to 100. A random sample of 36 auditors used the cash-flow information. Their mean assessment was 36.21, and the sample standard deviation was 22.93. For an independent random sample of 36 auditors not using the cash-flow information, the sample mean and standard deviation were respectively 47.56 and 27.56.

Test the assumption that population variances for assessments of the chance of material fraud were the same for auditors using cash-flow information as for auditors not using cash-flow information against a two-sided alternative hypothesis.

Newbold et al (2013)

EXERCISE 10.25: SOLUTION

Two-Tailed Test

✅ Answer:

1. Hypotheses

Let

σ^2_{cash} = population variance for auditors using cash-flow information,
$\sigma^2_{\text{no-cash}}$ = population variance for auditors **not using** cash-flow information.

$$H_0 : \sigma^2_{\text{cash}} = \sigma^2_{\text{no-cash}} \quad \text{(variances are equal)}$$

$$H_1 : \sigma^2_{\text{cash}} \neq \sigma^2_{\text{no-cash}} \quad \text{(variances are different)}$$

2. Sample Data

Group	n	s
Cash-flow information	36	22.93
No cash-flow information	36	27.56

EXERCISE 10.25: SOLUTION

✅ Answer:

3. Test Statistic

$$F = \frac{s^2_{\text{larger}}}{s^2_{\text{smaller}}} = \frac{27.56^2}{22.93^2} = \frac{759.0}{525.9} \approx 1.44$$

Degrees of freedom:

$$df_1 = n_{\text{larger}} - 1 = 35, \quad df_2 = n_{\text{smaller}} - 1 = 35$$

*(For a two-sided F-test, we always put the **larger variance in the numerator**.)*

---

4. Critical Values

Two-sided test at $\alpha = 0.05$, df1 = 35, df2 = 35:

$$F_{0.975,35,35} \approx 1.77, \quad F_{0.025,35,35} \approx 1/1.77 \approx 0.565$$

RR = [0, 0.565]U[1.77; +∞[

EXERCISE 10.25: SOLUTION

✅ Answer:

Decision Using RR

RR = [0, 0.565]U[1.77; +∞[

5. Decision

$$F = 1.44 \in [0.565, 1.77] \implies \text{fail to reject } H_0$$

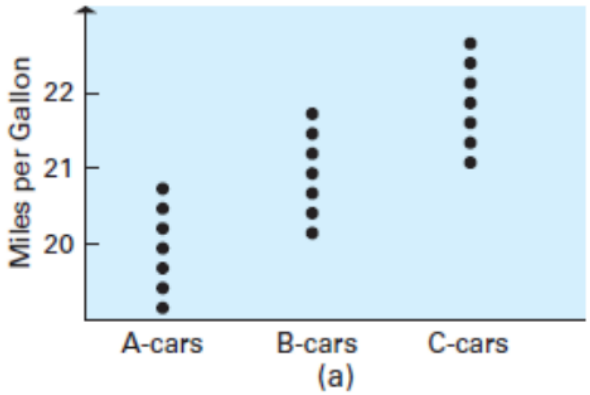
6. Conclusion

At the **5% significance level**, there is **insufficient evidence** to conclude that the population variances of assessments differ between auditors using and not using cash-flow information.

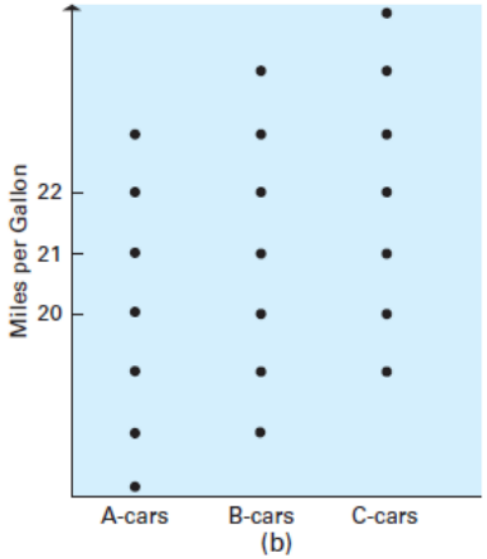
The variability in fraud assessments appears to be **statistically similar** for both groups.

LECTURE 19: ANALYSIS OF VARIANCE

COMPARISON OF SEVERAL POPULATION MEANS



- Small variation around the sample means compared to the variation among the sample means



- Large variation around the sample means compared to the variation among the sample means

Newbold et al (2013)

ONE-WAY ANALYSIS OF VARIANCE

Note: Although ANOVA is typically used to compare three or more groups, it can also be applied when comparing only two groups. In this case, the ANOVA test is equivalent to the independent samples t-test, and both methods lead to the same conclusion regarding the equality of means.

- Evaluate the difference among the means of three or more groups

  Examples: Average production for 1st, 2nd, and 3rd shifts
  Expected mileage for five brands of tires

- Assumptions
  - Populations are normally distributed
  - Populations have equal variances
  - Samples are randomly and independently drawn

Newbold et al (2013)

HYPOTHESES OF ONE-WAY ANOVA

Hypotheses

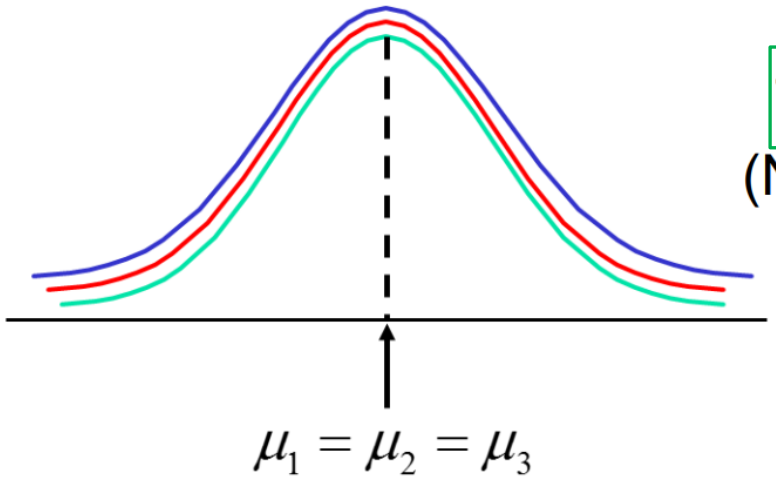
- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$
  - All population means are equal
  - i.e., no variation in means between groups
- $H_1 : \mu_i \neq \mu_j$ for at least one i , j pair.
  - At least one population mean is different
  - i.e., there is variation between groups
  - Does not mean that all population means are different (some pairs may be the same)

Newbold et al (2013)

ONE-WAY ANOVA

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

H_1 : Not all μ_i are the same



All Means are the same:

The Null Hypothesis is True

(No variation between groups)

Newbold et al (2013)

ONE-WAY ANOVA

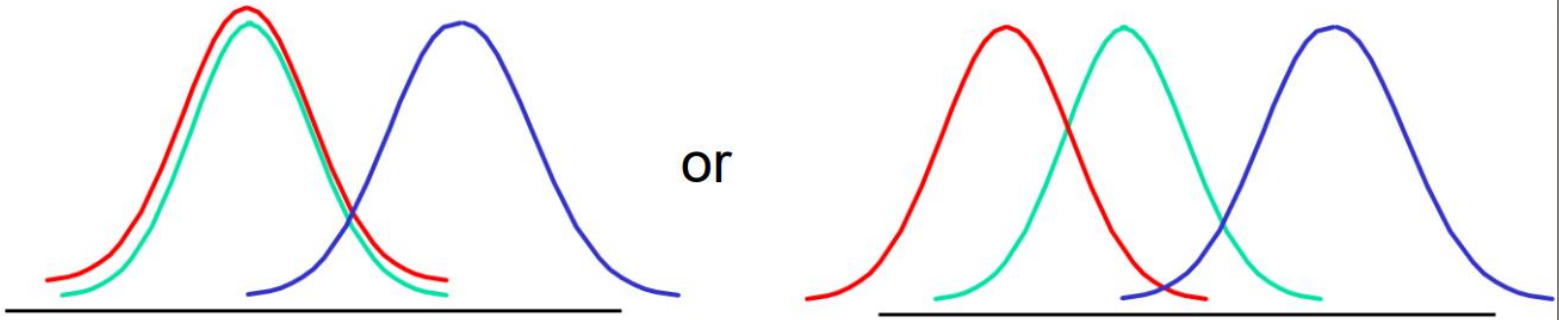
$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_K$

H_1 : Not all μ_i are the same.

At least one mean is different:

The Null Hypothesis is Not true

(Variation is present between groups)



or

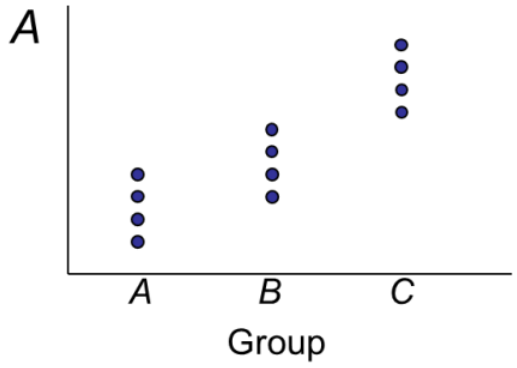
$\mu_1 = \mu_2 \neq \mu_3$

$\mu_1 \neq \mu_2 \neq \mu_3$

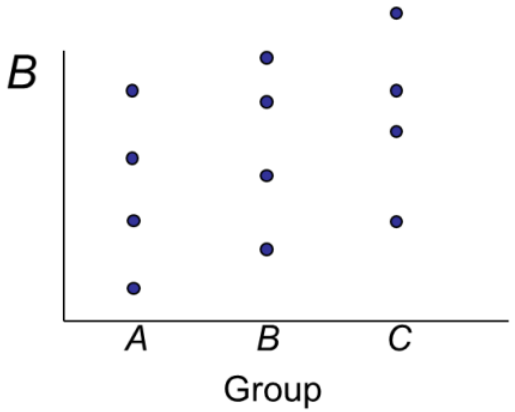
Newbold et al (2013)

VARIABILITY

- The variability of the data is key factor to test the equality of means
- In each case below, the means may look different, but a large variation within groups in *B* makes the evidence that the means are different weak



Small variation within groups



Large variation within groups

Newbold et al (2013)

SUM OF SQUARES DECOMPOSITION

- Total variation can be split into two parts:

$$SST = SSW + SSB$$

SST = Total Sum of Squares

Total Variation = the aggregate dispersion of the individual data values across the various groups

SSW = Sum of Squares Within Groups

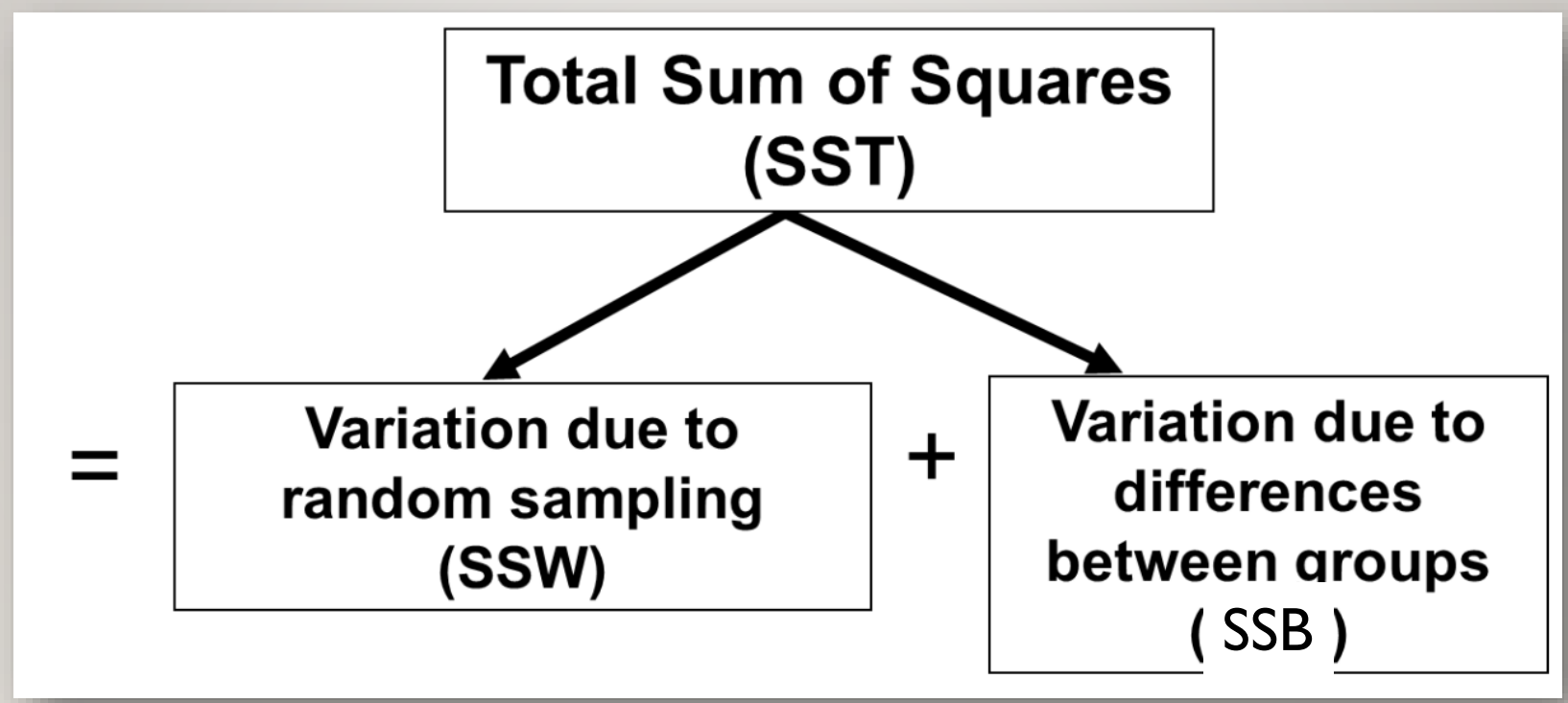
Within-Group Variation = dispersion that exists among the data values within a particular group

SSB = Sum of Squares Between Groups

Between-Group Variation = dispersion between the group sample means

Newbold et al (2013)

SUM OF SQUARES DECOMPOSITION



Newbold et al (2013)

TOTAL SUM OF SQUARES

$$SST = SSW + SSB$$

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{\bar{x}} \right)^2$$

Where:

SST = Total sum of squares

K = number of groups (levels or treatments)

n_i = number of observations in group i

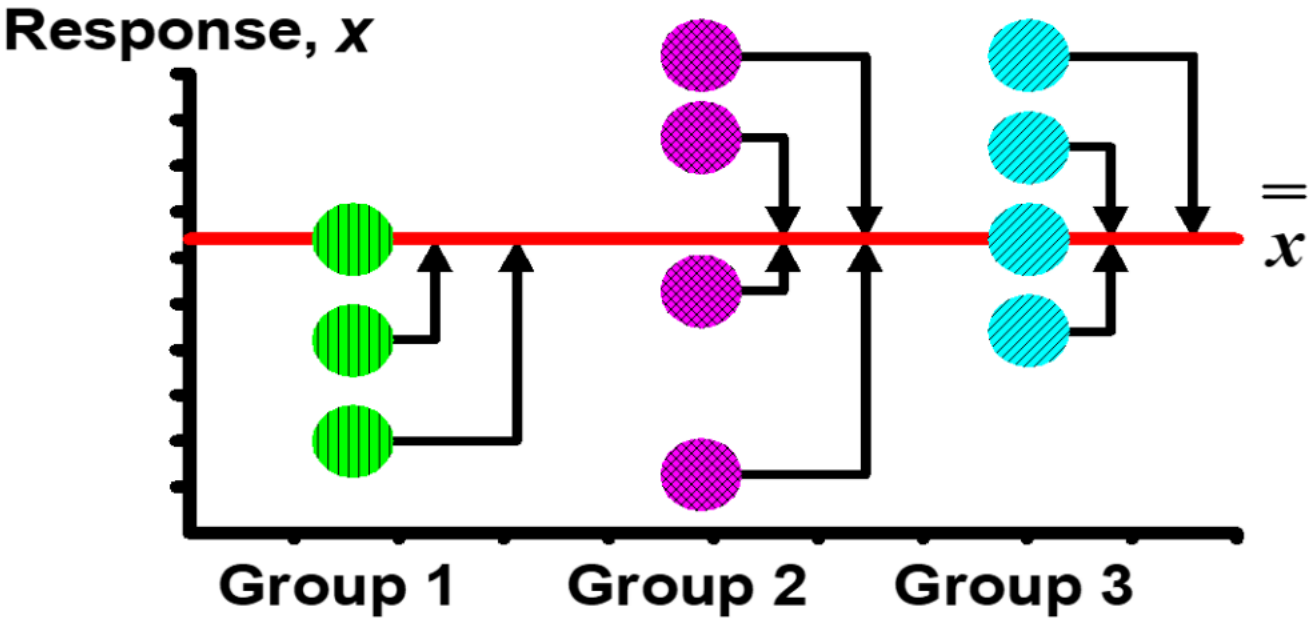
x_{ij} = j^{th} observation from group i

$\bar{\bar{x}}$ = overall sample mean

Newbold et al (2013)

TOTAL SUM OF SQUARES

$$SST=\left(x_{11}-\bar{\bar{x}}\right)^2+\left(x_{12}-\bar{\bar{x}}\right)^2+\cdots+\left(x_{Kn_K}-\bar{\bar{x}}\right)^2$$



Newbold et al (2013)

WITHIN-GROUP VARIATION

$$SST = SSW + SSG$$

$$SSW = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i \right)^2$$

Where:

SSW = Sum of squares within groups

K = number of groups

n_i = sample size from group i

$\bar{x}_i$ = sample mean from group i

$x_{ij} = j^{\text{th}}$ observation in group i

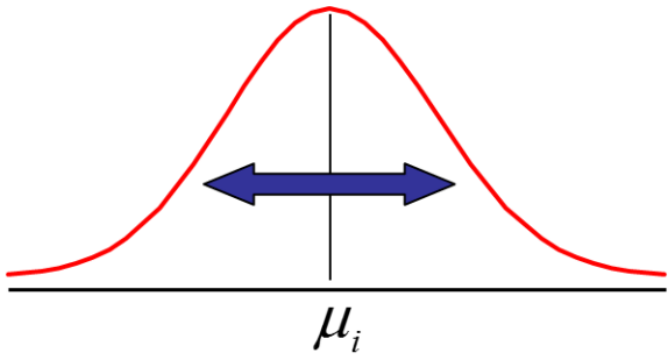
Newbold et al (2013)

WITHIN-GROUP VARIATION

$$\mathrm{SSW} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i \right)^2$$

Summing the variation within each group and then adding over all groups

$$\mathrm{MSW} = \frac{\mathrm{SSW}}{n - K}$$

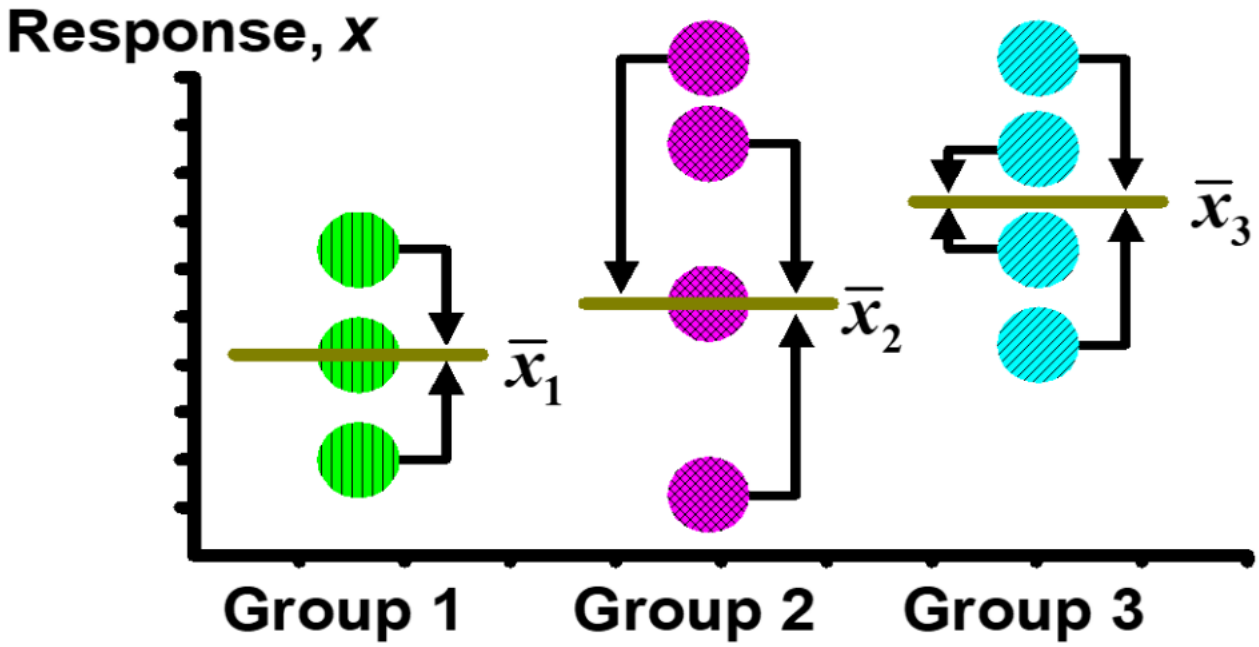


$$\text{Mean Square Within} = \frac{\mathrm{SSW}}{\text{degrees of freedom}}$$

Newbold et al (2013)

WITHIN-GROUP VARIATION

$$\mathrm{SSW} = \left(x_{11} - \bar{x}_1\right)^2 + \left(x_{12} - \bar{x}_1\right)^2 + \ldots + \left(x_{Kn_K} - \bar{x}_K\right)^2$$



Newbold et al (2013)

BETWEEN-GROUP VARIATION

$$SST = SSW + SSB$$

$$SSB = \sum_{i=1}^{K} n_i \left(\bar{x}_i - \bar{\bar{x}} \right)^2$$

Where:

SSB = Sum of squares between groups

K = number of groups

n_i = sample size from group i

$\bar{x}_i$ = sample mean from group i

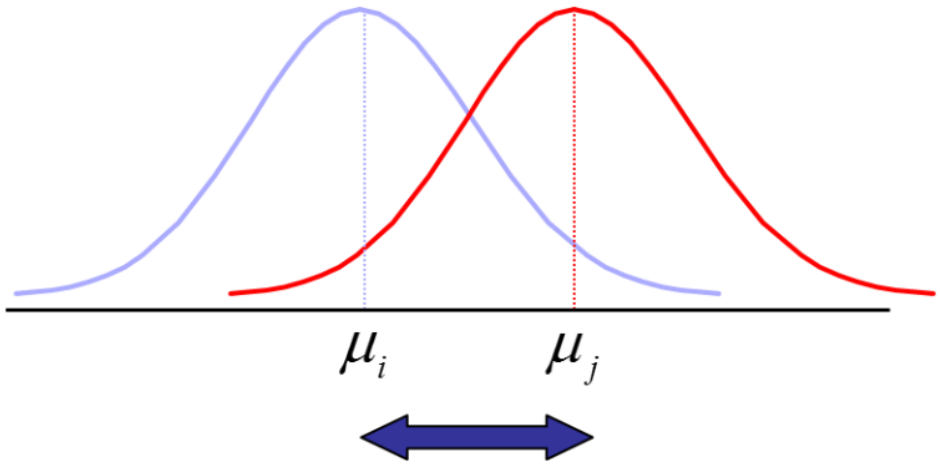
$\bar{\bar{x}}$ = grand mean (mean of all data values)

Newbold et al (2013)

BETWEEN-GROUP VARIATION

$$SSB = \sum_{i=1}^{K} n_i \left(\bar{x}_i - \bar{\bar{x}} \right)^2$$

Variation Due to Differences Between Groups



$$MSB = \frac{SSB}{K-1}$$

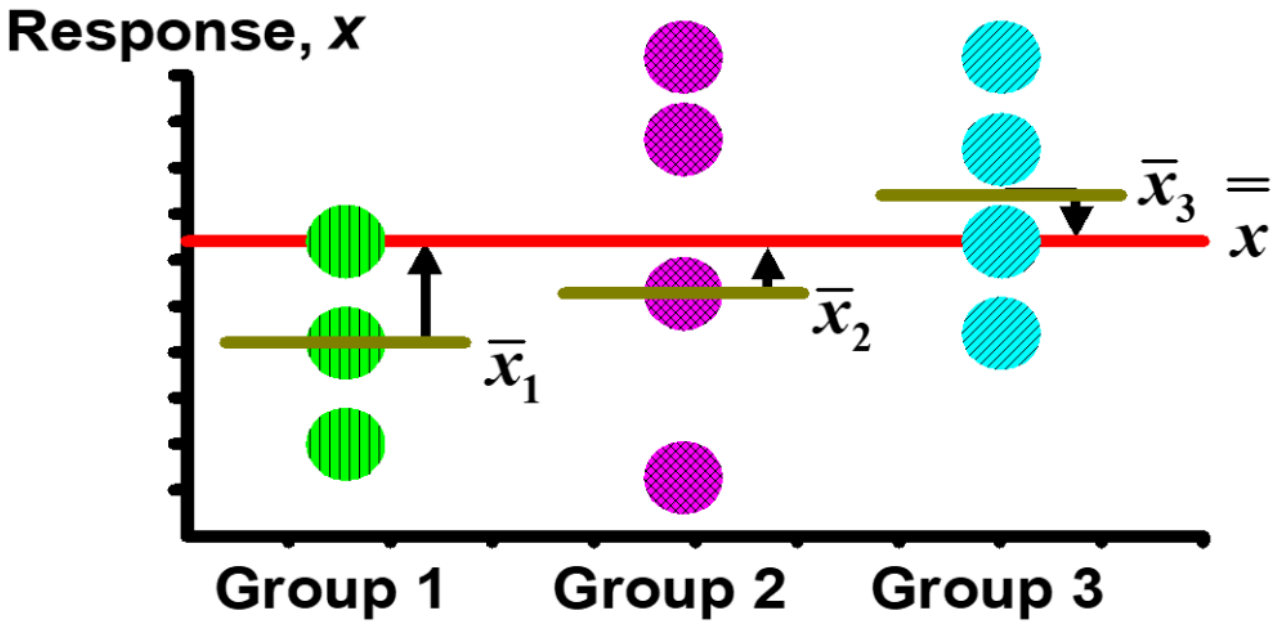
Mean Square Between Groups

$$= \frac{SSB}{\text{degrees of freedom}}$$

Newbold et al (2013)

BETWEEN-GROUP VARIATION

$$\text{SSB} = n_1\left(\bar{x}_1 - \bar{\bar{x}}\right)^2 + n_2\left(\bar{x}_2 - \bar{\bar{x}}\right)^2 + \ldots + n_K\left(\bar{x}_K - \bar{\bar{x}}\right)^2$$



Newbold et al (2013)

OBTAINING THE MEAN SQUARES

$$\text{MST} = \frac{\text{SST}}{n-1}$$

$$\text{MSW} = \frac{\text{SSW}}{n-K}$$

$$\text{MSB} = \frac{\text{SSB}}{K-1}$$

Where n = sum of the sample sizes from all groups

K = number of populations

Newbold et al (2013)

ONE-WAY ANOVA TABLE

Source of Variation	SS	df	MS (Variance)	*F* ratio
Between Groups	SSB	$K - 1$	$\text{MSB} = \frac{\text{SSB}}{K - 1}$	$F = \frac{\text{MSB}}{\text{MSW}}$
Within Groups	SSW	$n - K$	$\text{MSW} = \frac{\text{SSW}}{n - K}$	
Total	SST = SSB + SSW	$n - 1$		

K = number of groups

n = sum of the sample sizes from all groups

df = degrees of freedom

Newbold et al (2013)

ONE-FACTOR ANOVA F TEST STATISTIC

Hypotheses

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$$

H_1 : At least two population means are different

- **Test Stastistic**

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

MSG is mean squares between variances

MSW is mean squares within variances

- Degrees of freedom
  - $\text{df}_1 = K - 1$ (K = number of groups)
  - $\text{df}_2 = n - K$ (n = sum of sample sizes from all groups)

Newbold et al (2013)

INTERPRETING THE F STATISTIC

- The *F* statistic is the ratio of the between estimate of variance and the within estimate of variance
  - The ratio must always be positive
  - $df_1 = K - 1$ will typically be small
  - $df_2 = n - K$ will typically be large

Decision Using P-Value

P-value = $P(F > f_0)$
f_0 is the value of test statistic

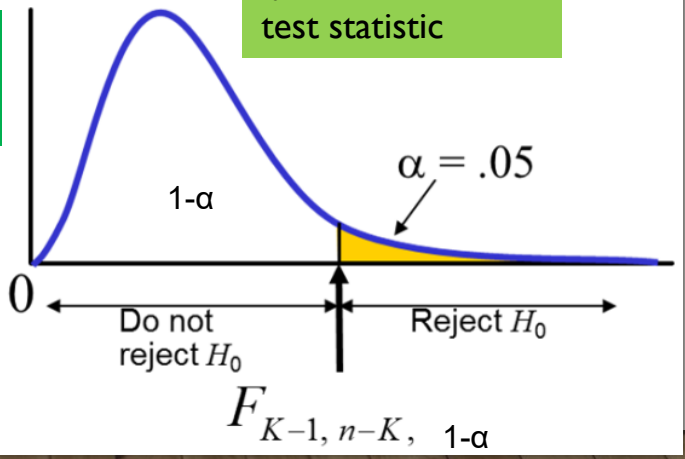
Decision Rule:

- Reject H_0 if

$f_0 > F_{K-1,\, n-K,\ 1-\alpha}$

Decision Using Rejection Region

$RR = [F_{1-\alpha;\, k-1,\, n-k}; +\infty[$



$F_{k-1,\, n-k;\, 1-\alpha} = F_{1-\alpha;\, k-1,\, n-k}$

Newbold et al (2013)

ONE-FACTOR ANOVA F TEST EXAMPLE

You want to see if three different golf clubs yield different distances. You randomly select five measurements from trials on an automated driving machine for each club. At the .05 significance level, is there a difference in mean distance?

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

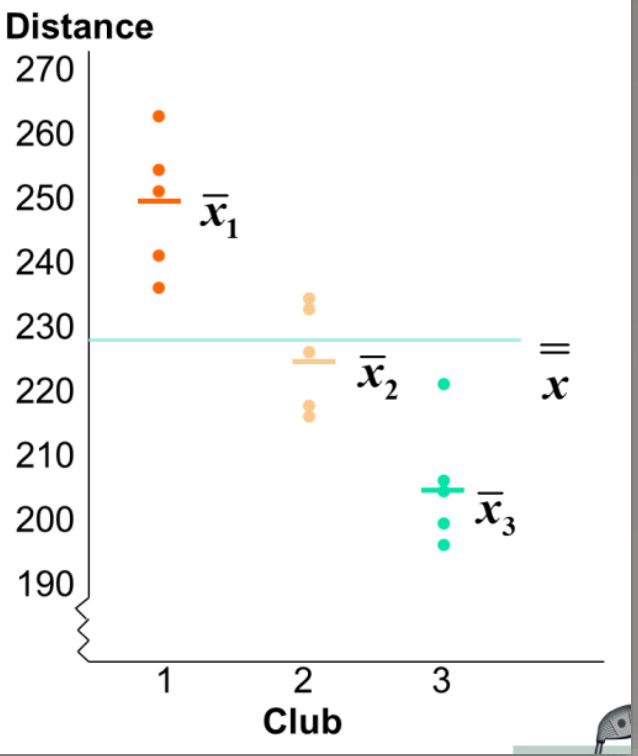
Newbold et al (2013)

ONE-FACTOR ANOVA EXAMPLE: SCATTER DIAGRAM

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2 \quad \bar{x}_2 = 226.0 \quad \bar{x}_3 = 205.8$

$\bar{\bar{x}} = 227.0$



Newbold et al (2013)

COMPUTATIONS

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$$\bar{x}_1 = 249.2 \qquad n_1 = 5$$
$$\bar{x}_2 = 226.0 \qquad n_2 = 5$$
$$\bar{x}_3 = 205.8 \qquad n_3 = 5$$
$$\bar{\bar{x}} = 227.0 \qquad n = 15$$
$$k = 3$$

$$\text{SSB} = 5(249.2-227)^2 + 5(226-227)^2 + 5(205.8-227)^2 = 4716.4$$

$$\text{SSW} = (254-249.2)^2 + (263-249.2)^2 + \ldots + (204-205.8)^2 = 1119.6$$

$$\text{MSB} = \frac{4716.4}{(3-1)} = 2358.2$$

$$\text{MSW} = \frac{1119.6}{(15-3)} = 93.3$$

$$f_0 = \frac{2358.2}{93.3} = 25.275$$

Newbold et al (2013)

ONE-FACTOR ANOVA EXAMPLE SOLUTION

Since the value of the test statistic (f_0 = 25.275) lies in the rejection region (RR), we reject H_0.

Hypotheses

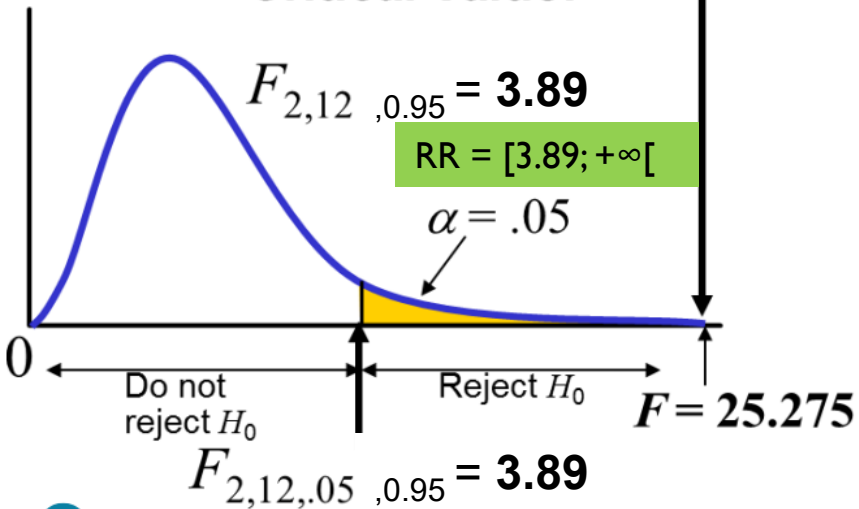
$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : \mu_i$ not all equal

$\alpha = .05$

$df_1 = 2 \qquad df_2 = 12$

Critical Value:



Test Stastistic:

$$f_0 = \frac{MSB}{MSW} = \frac{2358.2}{93.3} = 25.275$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

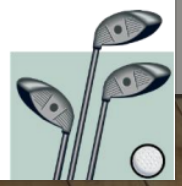
There is evidence that at least one μ_i differs from the rest

ANOVA - SINGLE FACTOR: EXCEL OUTPUT

Excel: data | data analysis | ANOVA: single factor

Summary						
Groups	Count	Sum	Average	Variance		
Club 1	5	1246	249.2	108.2		
Club 2	5	1130	226	77.5		
Club 3	5	1029	205.8	94.2		
ANOVA						
Source of Variation	SS	df	MS	*F*	*P*-value	*F* crit
Between Groups	**4716.4**	**2**	**2358.2**	**25.275**	4.99E-05	**3.89**
Within Groups	**1119.6**	**12**	**93.3**			
Total	5836.0	14				

P-value < 0.05



Newbold et al (2013)

EXERCISE 15.1

15.1 Given the following analysis of variance table, compute mean squares for between groups and within groups. Compute the F ratio and test the hypothesis that the group means are equal.

Source of Variation	Sum of Squares	Degrees of Freedom
Between groups	1,000	4
Within groups	750	15
Total	1,750	19

Newbold et al (2013)

EXERCISE 15.1: SOLUTION

✅ Answer:

Source of Variation	Sum of Squares	Degrees of Freedom
Between groups	1,000	4
Within groups	750	15
Total	1,750	19

1. Compute the Mean Squares

Mean Square Between Groups (MSB)

$$MS_B = \frac{SS_B}{df_B} = \frac{1000}{4} = 250$$

Mean Square Within Groups (MSW)

$$MS_W = \frac{SS_W}{df_W} = \frac{750}{15} = 50$$

2. Compute the F ratio

$$f_0 = \frac{MS_B}{MS_W} = \frac{250}{50} = 5$$

EXERCISE 15.1: SOLUTION

✅ Answer:

Hypotheses

Hypotheses

- H_0: All group means are equal
- H_1: At least one group mean is different

Test Stastistic

$$F = 5$$

with:

- numerator degrees of freedom: $df_1 = 4$
- denominator degrees of freedom: $df_2 = 15$

Decision Rule

At significance level α, reject H_0 if:

$f_0 \quad F_{1-\alpha;\,4,15}$

For example, at $\alpha = 0.05$,

$F_{0.95;\,4,15} \approx 3.06$

Conclusion

RR = [3.06; +∞[

Since:

$$5 > 3.06,$$

we **reject** H_0 at the 5% significance level.

4. Final Conclusion

There is **statistically significant evidence** to conclude that **not all group means are equal**.

LECTURE 20: NON-PARAMETRIC HYPOTHESIS TESTS

TWO NON-PARAMETRIC TESTS TO BE CONSIDERED

- Use the chi-square goodness-of-fit test to determine whether data fits specified probabilities
- Set up a contingency analysis table and perform a chi-square test of association

Newbold et al (2013)

NON-PARAMETRIC STATISTICS

- Nonparametric Statistics
  - Fewer restrictive assumptions about data levels and underlying probability distributions
    - Population distributions may be skewed
    - The level of data measurement may only be ordinal or nominal

Newbold et al (2013)

CHI-SQUARE GOODNESS-OF-FIT TEST

- Are technical support calls equal across all days of the week? (i.e., do calls follow a uniform distribution?)
  - Sample data for 10 days per day of week:

	Sum of calls for this day:
Monday	290
Tuesday	250
Wednesday	238
Thursday	257
Friday	265
Saturday	230
Sunday	192
	$\sum = 1722$

Newbold et al (2013)

LOGIC OF GOODNESS-OF-FIT TEST

- If calls **are** uniformly distributed, the 1722 calls would be expected to be equally divided across the 7 days:

$$\frac{1722}{7} = 246 \text{ expected calls per day if uniform}$$

- Chi-Square Goodness-of-Fit Test: test to see if the sample results are consistent with the expected results

Newbold et al (2013)

OBSERVED VS. EXPECTED FREQUENCIES

	Observed O_i	Expected $E_i = n \times p_i = 1722 \times 1/7$
Monday	290	246
Tuesday	250	246
Wednesday	238	246
Thursday	257	246
Friday	265	246
Saturday	230	246
Sunday	192	246
Total	1722	1722

Newbold et al (2013)

CHI-SQUARE TEST STATISTIC

Hypotheses

H_0 : The distribution of calls is uniform over days of the week

H_1 : The distribution of calls is not uniform

H_0: $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 1/7$
Vs
H_1: $\exists\ j: p_j \neq 1/7, j = 1, \ldots, 7$

- The t

Test Stastistic

$$Q = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \quad (\text{where d.f.} = K - 1)$$

where:

K = number of categories

O_i = observed frequency for category i

E_i = expected frequency for category i

Newbold et al (2013)

THE REJECTION REGION

Hypotheses

H_0 : The distribution of calls is uniform over days of the week

H_1 : The distribution of calls is not uniform

Test Stastistic

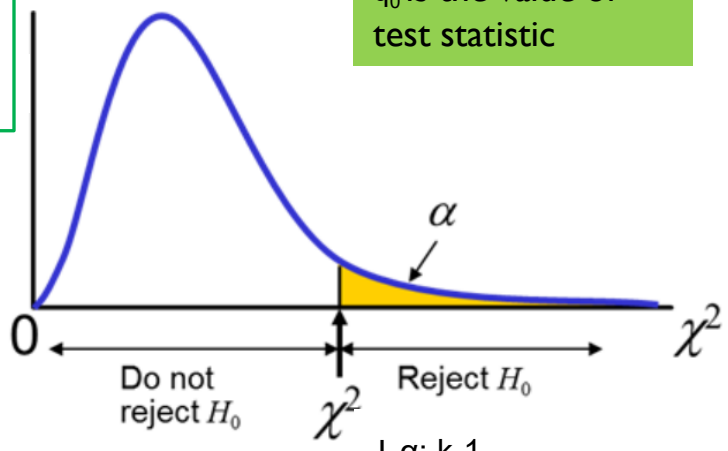
$$Q = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

Decision Using P-Value

P-value = $P(Q > q_0)$
q_0 is the value of test statistic

- Reject H_0 if $q_0 > \chi^2_{1-\alpha;\, k-1}$

(with $k - 1$ degrees of freedom)



Decision Using Rejection Region

RR = $[\chi^2_{1-\alpha;k-1}; +\infty[$

Newbold et al (2013)

OBSERVED VS. EXPECTED FREQUENCIES

	Observed O_i	Expected E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Monday	290	246	44	1936	7.870
Tuesday	250	246	4	16	0.065
Wednesday	238	246	−8	64	0.260
Thursday	257	246	11	121	0.492
Friday	265	246	19	361	1.467
Saturday	230	246	−16	256	1.041
Sunday	192	246	−54	2916	11.854
Total	1722	1722			$q_0 = 23.049$

Newbold et al (2013)

CHI-SQUARE TEST STATISTIC

H_0 : The distribution of calls is uniform over days of the week

H_1 : The distribution of calls is not uniform

$$q_0 = \frac{(290-246)^2}{246} + \frac{(250-246)^2}{246} + \ldots + \frac{(192-246)^2}{246} = 23.049$$

$K - 1 = 6$ (7 days of the week) so use 6 degrees of freedom:

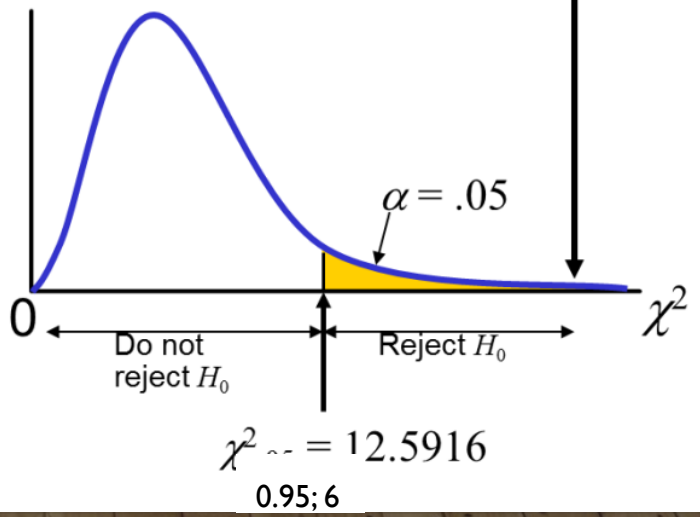
$$\chi^2_{0.95} = 12.5916$$

Conclusion:

RR = [12.5916; +∞[

$$q_0 = 23.05 > \chi^2_{1-\alpha} = 12.5916$$

so **reject** H_0 and conclude that the distribution is not uniform



Newbold et al (2013)

CHI-SQUARE GOODNESS-OF-FIT TEST: SUMMARY

Step 1: Hypotheses

- **Null hypothesis** (H_0): The observed frequencies follow the expected distribution.
- **Alternative hypothesis** (H_1): The observed frequencies do not follow the expected distribution.

Note:
The Chi-square goodness-of-fit test is used when we have a **categorical variable** (nominal or ordinal) and we want to test whether the **observed frequencies follow a specified theoretical distribution**.

Step 2: Test Statistic

$$Q = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-1}$$

- O_i = observed frequency
- $E_i = np_i$ = expected frequency
- n = total sample size
- p_i = theoretical probability for class i
- K = number of classes

Note:
The test statistic follows a Chi-square distribution with $K - 1$degrees of freedom.

Note:
Parametric tests rely on assumptions about the population distribution (typically normality), while non-parametric **tests do not require such assumptions**. Chi-square tests are non-parametric.

CHI-SQUARE GOODNESS-OF-FIT TEST: SUMMARY

Step 3: Rejection Region

- Degrees of freedom: $df = K - 1$ (or $K - 1 - p$ if parameters are estimated)
- Significance level: α
- **Reject H_0 if**

$$q_0 \geq \chi^2_{1-\alpha,\, K-1}$$

RR = $[\chi^2_{1-\alpha;k-1}; +\infty[$

α = significance level
1 – α = confidence level
$\chi^2_{1-\alpha;k-1}$ = represents the 1 – α quantile of the Chi-square distribution with k-1 degrees of freedom (see chi-square table)

Step 4: p-value

$$p\text{-value} = P(Q \geq q_0)$$

- Where q_0 is the observed value of the test statistic.
- The p-value indicates the probability of observing a test statistic as extreme as q_0, assuming H_0 is true.

Step 5: Decision

- **If q_0 falls in the rejection region** or **p-value < α** → reject H_0
- Otherwise → fail to reject H_0

EXERCISE 14.1

14.1 A random sample of 150 residents in one community was asked to indicate their first preference for one of three television stations that air the 5 p.m. news. The results obtained are shown in the following table. Test the null hypothesis that for this population their first preferences are evenly distributed over the three stations.

Station	A	B	C
Number of first preferences	47	42	61

Newbold et al (2013)

EXERCISE 14.1: SOLUTION

✅ Answer:

Data:

Station	A	B	C
Observed O_i	47	42	61

- Sample size: $n = 150$

Goodness-of-Fit Test

Step 1: Hypotheses + Observed vs. Expected Frequencies

$$H_0 : p_A = p_B = p_C = \frac{1}{3} \quad vs. \quad H_a : \text{not all } p_i \text{ are equal}$$

Station	Observed O_i	Expected $E_i = 150/3$
A	47	50
B	42	50
C	61	50

$$E_i = n \cdot p_i = 150 \cdot \frac{1}{3} = 50$$

EXERCISE 14.1: SOLUTION

Answer:

Step 2: Test statistic (Chi-square)

$$q_0 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(47-50)^2}{50} + \frac{(42-50)^2}{50} + \frac{(61-50)^2}{50}$$

Compute each term:

$$\frac{(47-50)^2}{50} = \frac{(-3)^2}{50} = \frac{9}{50} = 0.18$$

$$\frac{(42-50)^2}{50} = \frac{(-8)^2}{50} = \frac{64}{50} = 1.28$$

$$\frac{(61-50)^2}{50} = \frac{(11)^2}{50} = \frac{121}{50} = 2.42$$

$$q_0 = 0.18 + 1.28 + 2.42 = 3.88$$

The value of the test statistic is $q_0 = 3.88$.

Step 3: Degrees of freedom

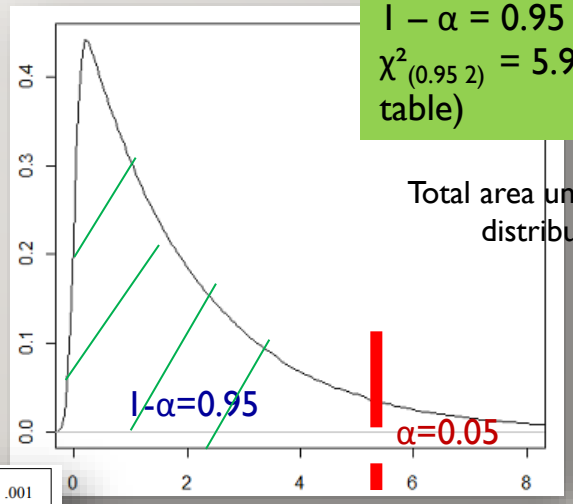
$$df = k - 1 = 3 - 1 = 2$$

CRITICAL VALUE $\chi^2_{1-\alpha;k-1}$: CALCULATION

Confidence Level (1-α = 0.95)

Significance Level (α=0.05)

α = 0.05 (significance level)
1 – α = 0.95 (confidence level)
$\chi^2_{(0.95\ 2)}$ = 5.991 (see chi-square table)



Total area under the chi-square distribution pdf = 1.

$\chi^2_{0.95;2}$ = 5.991

Given a significance level of α = 0.05, we have 1 − α = 0.95.

$\chi^2_{0.95;2}$ = 5.991 (see chi-square table)

$\chi^2_{n,\varepsilon} : P(X > \chi^2_{n,\varepsilon}) = \varepsilon$

n \ ε	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005	.001
1	.000	.000	.001	.004	.016	.102	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	.010	.020	.051	.103	.211	.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	.072	.115	.216	.352	.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	.207	.297	.484	.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	.412	.554	.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	.676	.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9							8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10							9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588

RR = [5.991; +∞[

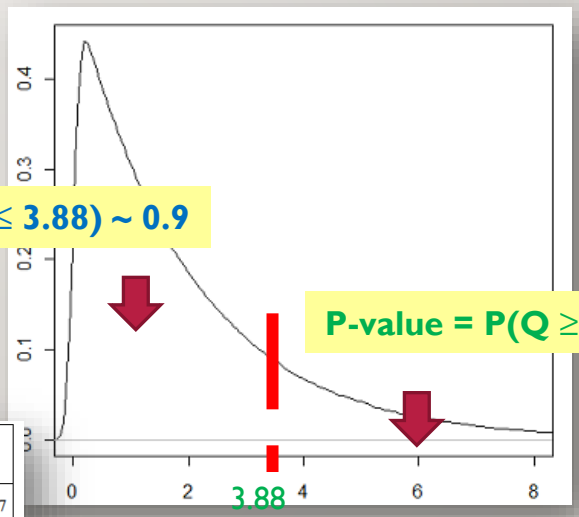
Note:
The chi-square table reports right-tail probabilities: $P(Q \geq q_0)$.

P-VALUE FOR A CHI-SQUARE STATISTIC: CALCULATION

The value of the test statistic is $q_0 = 3.88$

Total area under the chi-square distribution pdf = 1.

P-value = P(Q ≥ 3.88) ~ P(Q ≥ 4.605) = 0.1



$\chi^2_{n,\varepsilon} : P(X > \chi^2_{n,\varepsilon}) = \varepsilon$

n \ ε	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005	.001
1	.000	.000	.001	.004	.016	.102	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	.010	.020	.051	.103	.211	.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	.072	.115	.216	.352	.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	.207	.297	.484	.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	.412	.554	.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	.676	.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.5							15.987	18.307	20.483	23.209	25.188	29.588

Note:
The chi-square table reports right-tail probabilities: $P(Q \geq q_0)$.

Decision rule based on the p-value:
P-value = $P(Q \geq q_0) < \alpha \Rightarrow$ Reject H_0 for α

EXERCISE 14.1: SOLUTION

✅ Answer:

Step 4: Rejection region

- Significance level: $\alpha = 0.05$
- Critical value: $\chi^2_{0.95;2} \approx 5.991$
- **Reject** H_0 **if** $q > 5.991$

α = 0.05 (significance level)
1 – α = 0.95 (confidence level)
$\chi^2_{(0.95\ 2)}$ = 5.991 (see chi-square table)

Note:
The rejection region and the p-value were calculated in the two previous slides, respectively.

RR = [5.991; +∞[

Step 5: P-value

$$p = P(Q > 3.88) \approx 0.144$$

Exact value

P-value = P(Q > 3.88) ~ 0.1
Approximate value (from the chi-table)

Step 6: Conclusion

- $\chi^2 = 3.88 < 5.991 \Rightarrow$ not in rejection region
- $p = 0.144 > 0.05$

Decision: Fail to reject H_0

Interpretation (slide-ready): There is **no statistically significant evidence** that first preferences are not evenly distributed among the three TV stations.

CHI-SQUARE GOODNESS-OF-FIT TEST: POPULATION PARAMETERS UNKNOWN

Goodness-of-Fit Tests When Population Parameters Are Estimated

Suppose that a null hypothesis specifies category probabilities that depend on the estimation (from the data) of m unknown population parameters. The appropriate **goodness-of-fit test with estimated population parameters** is precisely as in Section 14.1, except that the number of degrees of freedom for the chi-square random variable is

$$\text{degrees of freedom} = (K - m - 1) \tag{14.3}$$

where K is the number of categories and m is the number of unknown population parameters.

Chi-Square Random Variable

A random sample of n observations, each of which can be classified into exactly one of K categories, is selected. Suppose the observed numbers in each category are $O_1, O_2, \ldots, O_K$. If a null hypothesis (H_0) specifies probabilities $P_1, P_2, \ldots, P_K$ for an observation falling into each of these categories, the expected numbers in the categories, under H_0, would be as follows:

$$E_i = nP_i \quad \text{for } i = 1, 2, \ldots, K \tag{14.1}$$

If the null hypothesis is true and the sample size is large enough that the expected values are at least 5, then the random variable associated with

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \tag{14.2}$$

is known as a **chi-square random variable**, and has, to a good approximation, a chi-square distribution with $(K - 1)$ degrees of freedom.

Newbold et al (2013)

EXERCISE 14.11

14.11 The number of times a machine broke down each week was observed over a period of 100 weeks and recorded in the accompanying table. It was found that the average number of breakdowns per week over this period was 2.1. Test the null hypothesis that the population distribution of breakdown is Poisson.

Number of breakdowns	0	1	2	3	4	5 or more
Number of weeks	10	24	32	23	6	5

Newbold et al (2013)

EXERCISE 14.11: SOLUTION

✅ Answer:

Chi-square goodness-of-fit test for a Poisson distribution

✔ Step 1: Hypotheses

- H_0: The number of breakdowns per week follows a Poisson distribution
- H_1: It does not follow a Poisson distribution

✔ Step 2: Parameter estimate

Given mean over 100 weeks:

$$\hat{\lambda} = 2.1$$

So under H_0:

$$X \sim \text{Poisson}(2.1)$$

✔ Step 3: Observed frequencies

x	0	1	2	3	4	≥5
O	10	24	32	23	6	5

Total = 100

EXERCISE 14.11: SOLUTION

Answer:

✔ Step 4: Compute probabilities

$$P(X = k) = e^{-2.1}\frac{2.1^k}{k!}$$

First:

$$e^{-2.1} \approx 0.1225$$

Probabilities p_i

- $P(0) = 0.1225$
- $P(1) = 0.1225 \cdot 2.1 = 0.2573$
- $P(2) = 0.1225 \cdot \frac{2.1^2}{2} = 0.2702$
- $P(3) = 0.1225 \cdot \frac{2.1^3}{6} = 0.1891$
- $P(4) = 0.1225 \cdot \frac{2.1^4}{24} = 0.0993$

Now:

$$P(\geq 5) = 1 - \sum_{k=0}^{4} P(k)$$

$$P(\geq 5) = 1 - (0.1225 + 0.2573 + 0.2702 + 0.1891 + 0.0993) = 0.0616$$

EXERCISE 14.11: SOLUTION

✅ Answer:

✔ Step 5: Expected frequencies (n = 100)

x	O_i	p_i	$E_i = n \times p_i$
0	10	0.1225	12.25
1	24	0.2573	25.73
2	32	0.2702	27.02
3	23	0.1891	18.91
4	6	0.0993	9.93
≥5	5	0.0616	6.16

EXERCISE 14.11: SOLUTION

✅ Answer:

✔ Step 6: Chi-square statistic

$$Q = \sum \frac{(O-E)^2}{E}$$

Compute each term:

- 0: $(10 - 12.25)^2/12.25 = 0.41$
- 1: $(24 - 25.73)^2/25.73 = 0.12$
- 2: $(32 - 27.02)^2/27.02 = 0.92$
- 3: $(23 - 18.91)^2/18.91 = 0.88$
- 4: $(6 - 9.93)^2/9.93 = 1.56$
- ≥5: $(5 - 6.16)^2/6.16 = 0.22$

Total:

$$q_0 \approx 4.11$$

EXERCISE 14.11: SOLUTION

✅ Answer:

✔ Step 7: Degrees of freedom

- Classes = 6
- Estimated parameters = 1 (λ)

$$df = 6 - 1 - 1 = 4$$

✔ Step 8: Critical value

At **10% significance level**:

$$\chi^2_{0.90,4} = 7.779$$

EXERCISE 14.11: SOLUTION

✅ Answer:

✔ Step 9: Decision

- Test statistic = 4.11
- Critical value = 7.779

$$4.11 < 7.779$$

👉 Do not reject H_0

RR = [7.779; +∞[

✔ Final conclusion

At the 10% significance level, there is not enough evidence to reject the hypothesis that the number of machine breakdowns per week follows a Poisson distribution.

CONTINGENCY TABLES

Contingency Tables

- Used to classify sample observations according to a pair of attributes
- Also called a cross-classification or cross-tabulation table
- Assume *r* categories for attribute A and *c* categories for attribute B
  - Then there are $(r \times c)$ possible cross-classifications

Newbold et al (2013)

R TIMES C CONTINGENCY TABLE

	Attribute B				
Attribute A	**1**	**2**	**. . .**	***c***	**Totals**
1	O_{11}	O_{12}	...	O_{1c}	R_1
2	O_{21}	O_{22}	...	O_{2c}	R_2
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Totals	C_1	C_2	...	C_c	n

Newbold et al (2013)

TEST FOR ASSOCIATION

- Consider n observations tabulated in an $r \times c$ contingency table
- Denote by O_{ij} the number of observations in the cell that is in the i^{th} row and the j^{th} column
- The null hypothesis is

Hypotheses

H_0 : No association exists between the two attributes in the population

- The appropriate test is a chi-square test with $(r-1)(c-1)$ degrees of freedom

Newbold et al (2013)

TEST FOR ASSOCIATION

Decision Using P-Value

P-value = $P(Q > q_0)$
q_0 is the value of test statistic

- Let R_i and C_j be the row and column totals
- The expected number of observations in cell row *i* and column *j*, given that H_0 is true, is

$$E_{ij} = \frac{R_i C_j}{n}$$

- A test of association at a significance level α is based on the chi-square distribution and the following decision rule

Test Stastistic

Reject H_0 if $Q = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} > \chi^2_{(r-1)(c-1),\ 1-\alpha}$

Decision Using Rejection Region

RR = $[\chi^2_{1-\alpha;(r-1)(c-1)}; +\infty[$

Newbold et al (2013)

TEST FOR ASSOCIATION: EXAMPLE

Left-Handed vs. Gender

- Dominant Hand: Left vs. Right
- Gender: Male vs. Female

Hypotheses

H_0 : There is no association between hand preference and gender

H_1 : Hand preference is not independent of gender

Newbold et al (2013)

TEST FOR ASSOCIATION: EXAMPLE

Sample results organized in a contingency table:

sample size = n = 300:

120 Females, 12 were left handed

180 Males, 24 were left handed

Gender	Hand Preference		
	Left	Right	
Female	12	108	120
Male	24	156	180
	36	264	300

Newbold et al (2013)

LOGIC OF THE TEST

H_0 : There is no association between hand preference and gender

H_1 : Hand preference is not independent of gender

- If H_0 is true, then the proportion of left-handed females should be the same as the proportion of left-handed males
- The two proportions above should be the same as the proportion of left-handed people overall

Newbold et al (2013)

FINDING EXPECTED FREQUENCIES

120 Females, 12 were left handed

180 Males, 24 were left handed

Overall:

$$P(\text{Left Handed}) = \frac{36}{300} = .12$$

If no association, then

$$P(\text{Left Handed} \mid \text{Female}) = P(\text{Left Handed} \mid \text{Male}) = .12$$

So we would expect 12% of the 120 females and 12% of the 180 males to be left handed…

i.e., we would expect $(120)(.12) = 14.4$ **females to be left handed**

$(180)(.12) = 21.6$ **males to be left handed**

Newbold et al (2013)

EXPECTED CELL FREQUENCIES

- Expected cell frequencies:

$$E_{ij} = \frac{R_i C_j}{n} = \frac{\left(i^{\text{th}}\ \text{Row total}\right)\left(j^{\text{th}}\ \text{Column total}\right)}{\text{Total sample size}}$$

Example:

$$E_{11} = \frac{(120)(36)}{300} = 14.4$$

Newbold et al (2013)

OBSERVED VS. EXPECTED FREQUENCIES

Observed frequencies vs. expected frequencies:

Gender	Hand Preference		
	Left	Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

Newbold et al (2013)

THE CHI-SQUARE TEST STATISTIC

The Chi-square test statistic is:

Test Statistic

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \quad \text{with} \quad \text{d.f.} = (r-1)(c-1)$$

- where:

O_{ij} = observed frequency in cell (i, j)

E_{ij} = expected frequency in cell (i, j)

r = number of rows

c = number of columns

Newbold et al (2013)

OBSERVED VS. EXPECTED FREQUENCIES

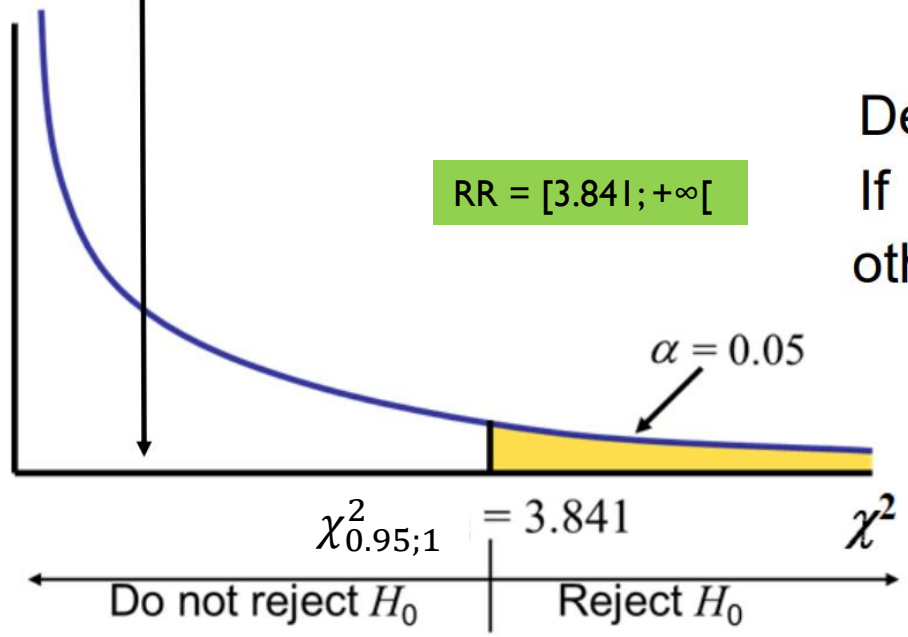
Gender	Hand Preference		
	Left	Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

$$q_0 = \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = 0.7576$$

Newbold et al (2013)

CONTINGENCY ANALYSIS

$q_0 = \boxed{0.7576}$ with $\text{d.f.} = (r-1)(c-1) = (1)(1) = 1$



RR = [3.841; +∞[

Decision Rule:

If $q_0 > 3.841$, reject H_0,

otherwise, do not reject H_0

Here, $q_0 = 0.7576 < 3.841$,

so we do not reject H_0 and conclude that gender and hand preference are not associated

Newbold et al (2013)

CHI-SQUARE TEST OF INDEPENDENCE: SUMMARY

1. Hypotheses

- **Null hypothesis (H_0):**
  The two categorical variables are independent.
  (There is no association between the variables.)
- **Alternative hypothesis (H_1):**
  The two categorical variables are not independent.
  (There is an association between the variables.)

Note:
The Chi-square test of independence is used when we have **two categorical variables** (nominal or ordinal) and want to determine whether there is a **statistically significant association** between them in a contingency table.

CHI-SQUARE TEST OF INDEPENDENCE: SUMMARY

2. Test Statistic

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(R-1)(C-1)}$$

- O_{ij} = observed frequency in cell (i, j)
- $E_{ij} = \dfrac{(\text{row total})_i \times (\text{column total})_j}{n}$ = expected frequency
- R = number of rows
- C = number of columns
- **Degrees of freedom:**

$$df = (R-1)(C-1)$$

Note:
The test statistic follows a Chi-square distribution with $(R-1)(C-1)$ degrees of freedom.

CHI-SQUARE TEST OF INDEPENDENCE: SUMMARY

3. Decision Rule

- **Rejection region:**

$$q_0 \geq \chi^2_{1-\alpha,\,(R-1)(C-1)}$$

RR = $[\chi^2_{1-\alpha;(R-1)(C-1)}; +\infty[$

- **p-value:**

$$p\text{-value} = P(Q \geq q_0)$$

(where q_0 is the observed value of the test statistic)

- **Decision:**
  - If q_0 falls in the rejection region **or** if p-value < α → **Reject** H_0
  - Otherwise → **Fail to reject** H_0

α = significance level
1 – α = confidence level
$\chi^2_{1-\alpha;(R-1)(C-1)}$ = represents the 1 – α quantile of the Chi-square distribution with (R-1)(C-1) degrees of freedom (see chi-square table)

EXERCISE 14.21

14.21 Following a presidential debate, people were asked how they might vote in the forthcoming election. Is there any association between one's gender and choice of presidential candidate?

	Gender	
Candidate Preference	Male	Female
Candidate A	150	130
Candidate B	100	120

Newbold et al (2013)

EXERCISE 14.21: SOLUTION

✅ Answer:

Chi-Square Test of Independence

Step 1: Hypotheses + Observed vs. Expected Frequencies

H_0 : Candidate preference is independent of gender $vs.$ H_1: Candidate preference is dependente of gender

Observed table O_{ij}:

Gender	Candidate A	Candidate B	Row Total
Male	150	100	250
Female	130	120	250
Total	280	220	500

EXERCISE 14.21: SOLUTION

✅ Answer:

Expected frequencies $E_{ij} = \frac{(\text{row total})(\text{column total})}{n}$:

$$E_{Male,A} = \frac{250 \cdot 280}{500} = 140$$

$$E_{Male,B} = \frac{250 \cdot 220}{500} = 110$$

$$E_{Female,A} = \frac{250 \cdot 280}{500} = 140$$

$$E_{Female,B} = \frac{250 \cdot 220}{500} = 110$$

Gender	Observed O_{ij}	Expected E_{ij}
Male	150 (A), 100 (B)	140, 110
Female	130 (A), 120 (B)	140, 110

EXERCISE 14.21: SOLUTION

Answer:

Step 2: Test Statistic

$$q_0 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(150-140)^2}{140} + \frac{(100-110)^2}{110} + \frac{(130-140)^2}{140} + \frac{(120-110)^2}{110}$$

Compute each term:

$$\frac{(150-140)^2}{140} = \frac{10^2}{140} \approx 0.714$$

$$\frac{(100-110)^2}{110} = \frac{(-10)^2}{110} \approx 0.909$$

$$\frac{(130-140)^2}{140} = \frac{(-10)^2}{140} \approx 0.714$$

$$\frac{(120-110)^2}{110} = \frac{10^2}{110} \approx 0.909$$

$$q_0 \approx 0.714 + 0.909 + 0.714 + 0.909 = 3.246$$

The value of the test statistic is $q = 3.246$.

EXERCISE 14.21: SOLUTION

Answer:

Step 3: Rejection Region

- Degrees of freedom: $df = (r-1)(c-1) = (2-1)(2-1) = 1$
- Significance level: $\alpha = 0.05$
- Critical value: $\chi^2_{0.05,1} \approx 3.841$
- **Reject** H_0 **if** $\chi^2 > 3.841$

α = 0.05 (significance level)
1 – α = 0.95 (confidence level)
$\chi^2_{(0.95\ 1)}$ = 3.841 (see chi-square table)

RR = [3.841; +∞[

Step 4: P-value

$$p = P(Q > 3.246) \approx 0.072$$

Exact value

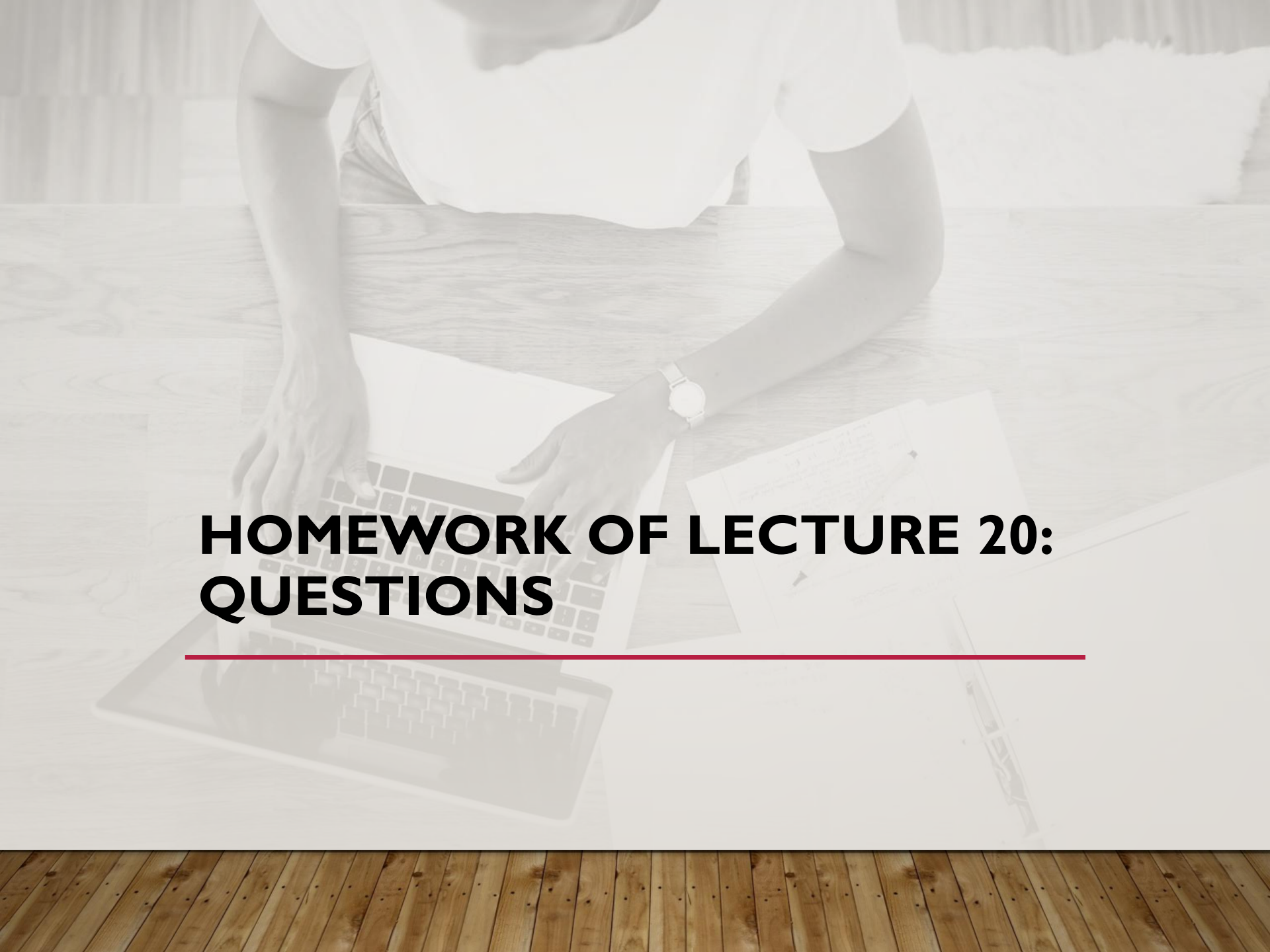
P-value = P(Q > 3.246) ~ 0.05
Approximate value (from the chi-square table)

Step 5: Conclusion

- $\chi^2 = 3.246 < 3.841 \Rightarrow$ not in rejection region
- P-value $p = 0.072 > 0.05$

Decision: Fail to reject H_0

Interpretation: There is **no statistically significant evidence** of an association between gender and candidate preference in this survey.



HOMEWORK OF LECTURE 20: QUESTIONS

EXERCISE 15.2

15.2 Given the following analysis of variance table, compute mean squares for between groups and within groups. Compute the F ratio and test the hypothesis that the group means are equal.

Source of Variation	Sum of Squares	Degrees of Freedom
Between groups	879	3
Within groups	798	16
Total	1,677	19

Newbold et al (2013)

EXERCISE 14.2

14.2 A 2008 survey investigated favorite water sports in Australia, and it found out that 45% of the interviewees voted for surfing, 40% voted for scuba diving, and the rest voted for other water sports. In 2011, a similar survey was conducted; out of a sample of 200 respondents, 102 declared they prefer surfing, 82 chose scuba diving, and the remaining 16 selected other water sports. Is it possible to conclude at the 5% level that in 2011 these preferences remained the same?

Newbold et al (2013)

EXERCISE 14.18

14.18 University administrators have collected the following information concerning student grade point average and the school of the student's major.

Determine if there is any association between GPA and major.

School	GPA < 3.0	GPA 3.0 or Higher
Arts and Sciences	50	35
Business	45	30
Music	15	25

Newbold et al (2013)

THANKS!

Questions?