This week:

Hypothesis Testing II:

***Chapter 8:* The Two-Sample Case**

Chapter 10: Hypothesis testing: Chi square

CHANGE IN SCHEDULE:
Problem solving Assignment # 4 due 11:30 a.m. sharp – 6%
Change from March 19th to March 26th (DUE)

REVIEWS THIS WEDNESDAY & FRIDAY DURING NORMAL TUTORIAL TIMES!

9-1

Chapter 7: Last week

Hypothesis Testing:

The One-Sample Case

Compare a sample statistic with a population parameter

We take a sample of Brock students; calculate a statistic (mean GPS),
& then ask: do they differ significantly from all students in Ontario (the population parameter)?

9-2

Chapter 7: before text

Hypothesis Testing:

The One-Sample Case

Compare a sample statistic with a population parameter

We take a sample of Brock students; calculate a statistic (mean GPS),
& then ask: do they differ significantly from all students in Ontario (the population parameter)?

TODAY: Chapter 8:

Hypothesis Testing II:

The Two-Sample Case

Compare a sample statistic with another sample statistic

Eg. We take a "sample of Brock students"... calculate their "mean GPA"
We take a "sample of Kings students"... calculate their "mean GPA"

Do the two sample differ significantly?

9-3

In this presentation you will learn about:

- The basic logic of the two sample case.
- Hypothesis Testing with
  - Sample Means (Large Samples),
  - Sample Means (Small Samples)
  - Sample Proportions (Large Samples)
- The difference between "statistical significance" and "importance"
- A few more words on setting "alpha"
- **Bivariate tables and Chi square (Chapter 10)**

9-4

Example:

- Do middle- and working-class persons differ in their use of email?
- The data below report the average number of times per day that people check their email in two random samples (one of middle class individuals and the other working class individuals):

E-mail Messages	
Sample 1 (Middle Class)	Sample 2 (Working Class)
$\bar{X}_1 = 8.7$	$\bar{X}_2 = 5.7$
$s_1 = 0.3$	$s_2 = 1.1$
$N_1 = 89$	$N_2 = 55$

- The middle class seem to check their email more than the working class, but is the difference *significant*?

9-5

Hypothesis Test for Two Samples: Basic Logic

We begin with a difference between sample statistics (means).

The question we test:
"Is the difference between the samples large enough to allow us to conclude (with a known probability of error) that the populations represented by the samples are different?"

The null hypothesis, H_0, is that the samples represent populations that are the same:
There is no difference between the parameters of the two populations. $H_0: \mu_1 = \mu_2$

If the difference between the sample statistics is large enough, or, if a difference of this size is *unlikely* assuming H_0 is true, we *reject* the H_0
Conclude that there is a significant difference between the populations.

$H_1: \mu_1 \neq \mu_2$ or $H_1: \mu_1 > \mu_2$ or $H_1: \mu_1 < \mu_2$

Changes from One- to Two-Sample Case

- **Step 1:** in addition to samples selected according to EPSEM principles, samples must be selected independently: **Independent random sampling**.
- **Step 2:** null hypothesis statement will say the two populations are not different.
- **Step 3:** sampling distribution refers to ***difference* between the sample statistics**.
- **Step 4:** In computing the test statistic, we use *Z*(obtained) or *t*(obtained) with slightly revised formula, depending on the size of our sample (forthcoming)
- **Step 5:** same as before: If the test statistic, *Z*(obtained) or *t*(obtained), falls into the critical region, as marked by *Z*(critical) or *t*(critical), reject the H_0.

9-7

NOTE: STEP 4 USES DIFFERENT FORMULA!!!

- **Step 4:** In computing the test statistic, we use *Z*(obtained) or *t*(obtained) with slightly revised formula, depending on the size of our sample (forthcoming)

We will work with 3 options & 3 sets of formulae

1. If comparing sample means (2 large samples)
   1a. With population standard deviations
   1b. With only sample standard deviations

2. If comparing sample means (small samples: n_1 and $n_2 < 100$)

3. If comparing sample proportions (large samples)

9-8

1a. If comparing sample means (2 large samples) with σ

$$Z(\text{obtained}) = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{x}-\bar{x}}} \quad \text{with} \quad \sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

1b. If comparing sample means (2 large samples) with s

$$Z(\text{obtained}) = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{x}-\bar{x}}} \quad \text{with} \quad \sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

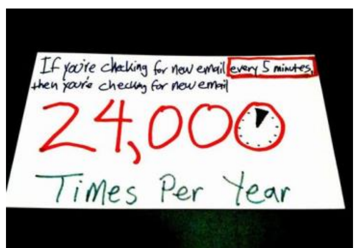
2. If sample means (small samples)

$$t(\text{obtained}) = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{x}-\bar{x}}} \quad \text{with} \quad \sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

3. If sample proportions (large samples)

$$Z(\text{obtained}) = \frac{(P_{s1} - P_{s2})}{\sigma_{p-p}} \qquad \sigma_{p-p} = \sqrt{P_u(1-P_u)}\sqrt{\frac{n_1+n_2}{n_1 n_2}} \qquad P_u = \frac{n_1 P_{s1} + n_2 P_{s2}}{n_1 + n_2}$$

9-9



"Psst! Buddy – you wanna check your email?"

9-10

Example:

- Do middle- and working-class persons differ in their use of email?
- The data below report the average number of times per day that people check their email in two random samples (one of middle class individuals and the other working class individuals):

E-mail Messages

Sample 1 (Middle Class)	Sample 2 (Working Class)
$\bar{X}_1 = 8.7$	$\bar{X}_2 = 5.7$
$s_1 = 0.3$	$s_2 = 1.1$
$N_1 = 89$	$N_2 = 55$

- Is the difference *significant*?

Testing Hypotheses: The Five Step Model

1. Make assumptions and meet test requirements.
2. State the H_0.
3. Select the Sampling Distribution and Determine the Critical Region.
4. Calculate the test statistic.
5. Make a Decision and Interpret Results.

Return to our example:

E-mail Messages	
Sample 1 (Middle Class)	Sample 2 (Working Class)
$\bar{X}_1 = 8.7$	$\bar{X}_2 = 5.7$
$s_1 = 0.3$	$s_2 = 1.1$
$N_1 = 89$	$N_2 = 55$

Step 1: Make Assumptions and Meet Test Requirements

- Model:
  - Independent Random Samples
    - The samples must be independent of each other (i.e. the selection of cases in the first sample has no bearing on the selection of cases in the second)
  - Level of Measurement is Interval-Ratio
    - Number of email messages -> can work with our means
  - Sampling Distribution's shape
    - *N* = (85+55 =144) cases which is > 100 so we can assume a normal shape.

Step 2: State the Null Hypothesis

- No direction for the difference has been predicted, so a two-tailed test is called for, as reflected in the research hypothesis:
  - H_0: $\mu_1 = \mu_2$
    - The Null asserts there is no significant difference between the populations (the two populations represented by our samples are equally likely to be using email)
  - H_1: $\mu_1 \neq \mu_2$
    - The research hypothesis contradicts the H_0 and asserts there is a significant difference between the populations.

9-15

Step 3: Select Sampling Distribution and Establish the Critical Region

- Sampling Distribution = *Z* distribution
- Alpha (α) = 0.05
- note: unless otherwise stated, use 0.05 in all significance tests (i.e. the default in most tests)
- With two tailed test: *Z* (critical) = ± 1.96

Step 4: Compute the Test Statistic

With two sample tests, use the appropriate formula (below) to compute the obtained *Z* score:

$$Z(\text{obtained}) = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}-\bar{X}}}$$

The denominator in this formula is the standard deviation of the sampling distribution (i.e. the standard error)

Step 4 *(continued)*

NOTE: How do we calculate this standard error that enters into the denominator of Z(obtained)? When the population standard deviations are known, we use the following formula:

$$\sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

but when we only have the sample standard deviations, we use the following:

$$\sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

i.e. we substitute *s* as an estimator of *σ,* suitably corrected for the bias (*n* is replaced by *n-1* to correct for the fact that *s* is a biased estimator of *σ*).

Again, the above formula only apply if the combined size of the two samples is at least N> 100

9-17

In this example: compute the Test Statistic

E-mail Messages

Sample 1 (Middle Class)	Sample 2 (Working Class)
$\bar{X}_1 = 8.7$	$\bar{X}_2 = 5.7$
$s_1 = 0.3$	$s_2 = 1.1$
$N_1 = 89$	$N_2 = 55$

We have the "sample standard deviations",..

So: calculate standard error (population standard deviations unknown):

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}} = \sqrt{\frac{.3^2}{89-1} + \frac{1.1^2}{55-1}} = \sqrt{.001 + .022} = .15$$

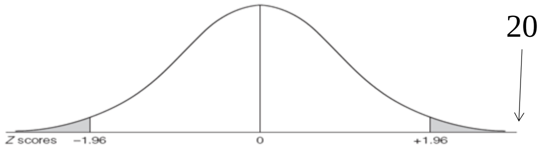
On this basis, you can calculate Z (obtained) with the standard error in the denominator

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}-\bar{X}}} = \frac{8.7 - 5.7}{.15} = 20$$

9-18

Step 5: Make Decision and Interpret Results

The obtained test statistic (*Z* = 20) falls in the Critical Region so *reject* the null hypothesis.



- The difference between the sample means is so large that we can conclude, at $\alpha = 0.05$, that a difference exists between the populations represented by the samples.
- The difference between email usage of middle- and working-class individuals *is significant.*

Hypothesis Test for Two-Sample Means: Student's *t* distribution

(Small Samples)

Hypothesis Test for Two-Sample Means: Student's *t* distribution

(Small Samples)

- For small samples (combined *N's<100), s* is too unreliable an estimator of *σ* so do not use standard normal distribution. Instead we use Student's *t* distribution.
- The formula for computing the test statistic, *t*(obtained), is:

FORMULA 8.6

$$t\,(\text{obtained}) = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{x}-\bar{x}}}$$

where $\sigma_{\bar{x}-\bar{x}}$ is defined as:

FORMULA 8.5

$$\sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

9-21

Hypothesis Test for Two-Sample Means: Student's *t* distribution

(continued)

- The logic of the five-step model for hypothesis testing is followed, using the *t* table, Appendix B, where the degrees of freedom (df) = $N_1 + N_2 - 2$.

9-22

Example: Research on Obesity,.. How to deal with the problem?

9-23

Example:

Studying "weight loss" strategies:

- 1st sample – combined cardio (30 minutes a day & weight training 30 minutes a day)

  Mean weight loss: 20 pounds
  s =5
  Sample size (n_1 = 29)

2nd sample – Solely cardio (45 minutes a day)

Mean weight loss: 18 pounds
s = 4
Sample size (n_2 = 33)

Is there a significant difference between the two?? 9-24

Step 1: Make Assumptions and Meet Test Requirements

- Model:
  - Independent Random Samples
  - Level of Measurement is Interval-Ratio
    - Weight loss-> can work with our means
  - Sampling Distribution’s shape
    - N = (29+33=62) cases which is less than 100 so we must work with t distribution

9-25

Step 2: State the Null Hypothesis

- No direction for the difference has been predicted, so a two-tailed test is called for, as reflected in the research hypothesis:
  - H_0: $\mu_1 = \mu_2$
    - The Null asserts there is no significant difference in the weight loss for the two populations
  - H_1: $\mu_1 \neq \mu_2$
    - The research hypothesis contradicts the H_0 and asserts there is a significant difference in weight loss

9-26

Step 3: Select Sampling Distribution and Establish the Critical Region

- Sampling Distribution = *t* distribution
- Alpha (α) = 0.05
- note: unless otherwise stated, use 0.05 in all significance tests (i.e. the default in most tests) $df = n_1 + n_2 - 2 = 60$
- With two tailed test: *t* (critical) = ? (from Appendix B)

Appendix B Distribution of *t*

Degrees of Freedom (df)	Level o[illegible]ificance for One-tailed Test					
	.10	.05	[illegible]	.01	.005	.0005
	Level of Significance for Two-tailed Test					
	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	[illegible]	2.771	3.690
28	1.313	1.701	2.048	[illegible]	2.763	3.674
29	1.311	1.699	2.045	[illegible]	2.756	3.659
30	1.310	1.697	2.042	[illegible]	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	~~1.671~~	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	~~1.645~~	1.960	2.326	2.576	3.291

Source: Table III of Fisher & Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd., London (1974), 6th edition (previously published by Oliver & Boyd Ltd.

9-28

Step 3: Select Sampling Distribution and Establish the Critical Region

- Sampling Distribution = *t* distribution
- Alpha (α) = 0.05
- note: unless otherwise stated, use 0.05 in all significance tests (i.e. the default in most tests) df = $N_1 + N_2 - 2 = 60$
- With two tailed test: *t* (critical) = ± 2.00 (from Appendix B)

Step 4: Compute the Test Statistic

With two sample tests, use the appropriate formula (below) to compute the obtained *t* score:

$$t(\text{obtained}) = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{x}-\bar{x}}}$$

$$\uparrow$$

BUT: must first calculate the denominator (SE)

Step 4 *(continued)*

NOTE: How do we calculate this standard error ?

When the population standard deviations are <u>unknown</u>, we use Formula 8.5 to calculate $\sigma_{\bar{x}-\bar{x}}$:

Again, the above formula only apply if the combined size of the two samples is less than 100

FORMULA 8.5

$$\sigma_{\bar{x}-\bar{x}} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = \sqrt{\frac{(29)(5)^2 + (33)(4)^2}{29 + 33 - 2}}\sqrt{\frac{29 + 33}{(29)(33)}} =$$

$$= 1.16$$

In this example: compute the Test Statistic

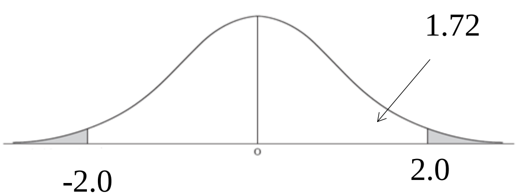
On this basis, you can calculate t (obtained) with the standard error in the denominator

$$t(obtained) = \frac{(\overline{X}_1 - \overline{X}_2)}{\sigma_{\overline{X}-\overline{X}}} = \frac{20 - 18}{1.16} = 1.72$$

9-31

Step 5: Make Decision and Interpret Results

The obtained test statistic (t = 1.72) does not fall in the Critical Region so we can not *reject* the null hypothesis.
Recall: t(critical) +/- 2.0



- The difference between the sample means is not large enough that we can

- conclude, at α = 0.05, that a difference exists between the populations represented by the samples.
-
- The difference between the two populations using the different exercise regimes *is NOT significant.*

9-32

TWO sample test with Proportions (or percentages)….

We conduct research on educational outcomes

AFN's National Chief, Perry Bellegarde has urged the Trudeau Government to act on "education"!!

9-33

Example:

Sample from Non-Aboriginal Population (N=60)
Ps1 = .23 (23 % are university educated)

Sample from Aboriginal Population (N=72)
Ps2 = .10 (10% are university educated)

Are Non-Aboriginal Canadians significantly more likely than Aboriginal Canadians to have a university degree?

Problem here: can we infer from our samples, that are not that large?

Formula for Hypothesis Testing with Sample Proportions (Large Samples)

- Formula for proportions:

$$Z(obtained) = \frac{\mathrm{P}_{s1} - \mathrm{P}_{s2}}{\sigma_{p-p}}$$

Where P_{s1} is the proportion associated with the first sample, and P_{s2} is the proportion associated with the second.

- See next slide for how to calculate the denominator in this equation (*standard error*)* and the "pooled estimate of the population proportion"*….
- *Note that you need to calculate both these values in order to solve the denominator of the above equation!

To obtain standard error, most first calculate something called: P_u (the Pooled Estimate of the Population Proportion)

- To calculate P_u (the pooled estimate, see p. 255):

$$P_u = \frac{n_1 P_{s1} + n_2 P_{s2}}{n_1 + n_2}$$

- Which is then inserted into the following equation for the standard deviation of the sampling distribution (standard error):

$$\sigma_{p-p} = \sqrt{\mathrm{P}_u(1-\mathrm{P}_u)}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

Which then enters into the aforementioned formula for our test statistic Z(obtained)

Again, use the basic 5 step model in testing for significance…

9-37

Step 1.
Model has independent random samples,
Level of measurement is “nominal” -> work with proportions
Sampling distribution can be considered normal since N> 100

Step 2. State null hypothesis: direction? Yes, one tailed test

H_0: $P_{\mu 1} = P_{\mu 2}$
The Null asserts there is no significant difference in the proportion with a university degree for the two populations

H_1: $P_{\mu 1} > P_{\mu 2}$
The research hypothesis contradicts the H_0 and asserts there is a significant difference: Non-Aboriginal people have a higher education.. Than Aboriginal Canadians..

Step 3.

Select the sampling distribution and establish critical region

Sampling distribution is the Z distribution

Alpha is .05 one tailed

Appendix A table indicates Z(critical) = 1.65

Step 4. Calculate the test statistic

Start with "pooled estimate on the proportion"

$$P_u = \frac{n_1 P_{s1} + n_2 P_{s2}}{n_1 + n_2}$$

$$P_u = \frac{(60)(.23) + (72)(.10)}{60 + 72} = .159$$

Next: get our standard error

$$\sigma_{p-p} = \sqrt{\mathrm{P}_u(1 - \mathrm{P}_u)}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

$$\sigma_{p-p} = \sqrt{.159(1 - .159)}\sqrt{\frac{60 + 72}{(60)(72)}} = 0.064$$

Step 4 (continued)

Then obtain your test statistic:

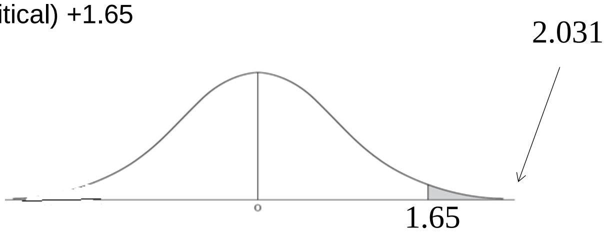
$$Z(obtained) = \frac{P_{s1} - P_{s2}}{\sigma_{p-p}}$$

$$Z(obtained) = \frac{.23 - .10}{.064} = 2.031$$

Step 5: Make Decision and Interpret Results

The obtained test statistic Z = 2.031) falls in the Critical Region so we can *reject* the null hypothesis.

Recall: Z(critical) +1.65



- The difference between the proportions is large enough to conclude, at α = 0.05, that Non-Aboriginal Canadians are significantly more likely to have a university education than "Aboriginal Canadians"
- The difference between the two populations *is significant.*

9-42

Some comments on Alpha Levels

- By assigning an alpha level, α, one defines an “unlikely” sample outcome.
- Alpha level is the probability that the decision to reject the null hypothesis, H_0, is incorrect.
- If we set our Alpha at .05, and we end up rejecting our null hypothesis,.. We are 95% certain that we are correct

If we set our Alpha at .10, and we end up rejecting our null hypothesis, we are 90% certain that we are correct..

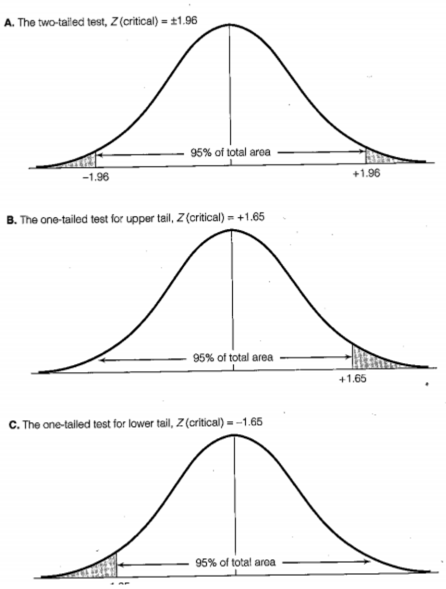
Etc…

Do note: that our sampling distribution tells us that sometimes we can be wrong!!

8-43

Alpha levels affect Critical Region in Step 3:

ESTABLISHING THE CRITICAL REGION, ONE-TAILED TESTS VERSUS TWO-TAILED TESTS, WITH REJECTION REGION FOR ALPHA = 0.05 IN SHADE



FINDING CRITICAL Z SCORES FOR ONE-TAILED TESTS (Single Sample Means)

		One-Tailed Value	
Alpha	Two-Tailed Value	Upper Tail	Lower Tail
0.10	±1.65	+1.29	−1.29
0.05	±1.96	+1.65	−1.65
0.01	±2.58	+2.33	−2.33
0.001	±3.29	+3.10	−3.10

8-44

Significance vs. Importance

- The probability of rejecting the null hypothesis in comparing statistics is a function of four independent factors:

1. The size of the difference (e.g., means of 8.7 and 5.7 for the example above).

2. The value of alpha (the higher the alpha, the more likely we are to reject the H_0).

3. The use of one- vs. two-tailed tests (we are more likely to reject with a one-tailed test).

4. The size of the sample (N) (the larger the sample the more likely we are to reject the H_0).

9-45

Significance vs. Importance

(continued)

- As long as we work with random samples, we must conduct a test of significance. However, **significance** is not the same thing as **importance**.
- Differences that are otherwise trivial or uninteresting may be significant, which is a major limitation of hypothesis testing.
  - When working with large samples, even small differences may be significant.
  - The value of the standard error is always an inverse function of N (i.e. the larger the N, the smaller the standard error)
  - The larger the N, the greater the value of the test statistic (standard error is always in the denominator), the more likely it will fall in the Critical Region and be declared significant.

9-46

Significance vs. Importance

(continued)

- In conclusion, significance is a necessary but not sufficient condition for importance.

- A sample outcome could be:
  - significant and important
  - significant but unimportant (e.g. with a very large N)
  - not significant but important (yikes: hazard of small N)
  - not significant and unimportant

9-47

Next Chapter: Chapter 10

Hypothesis Testing IV:

Chi Square

11-48

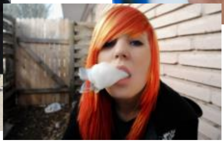
In this presentation you will learn about:

- Bivariate (Cross tabulation) Tables
- The basic logic of Chi Square
- If time:
- Perform the Chi Square test using the five-step model

11-49

Why examine a "bivariate table"?
Example: We are conducting research on smoking & education..

Small sample (N=600), is there a significant association??



11-50

Bivariate Tables

- Bivariate tables: display the scores of cases on two different variables at the same time.

Cell Counts

Level of Education

INDEPENDENT VARIABLE

DEPENDENT VARIABLE

		< H.S.	H. School Grad	Some Post Sec	
Smoking Behavior	No	60	100	300	460
	Yes	40	40	60	140
		100	140	360	600

Row marginals

Cell count for < HS and Non-smoker

column marginals

Total # of Cases (N)

More on Bivariate Tables

Cells are intersections of columns and rows.

- There will be as many cells as there are scores on the two variables combined.
- E.g. If 3 categories on dependent variable, and 5 categories on the indpendent, we have 3*5 = 15 cells

Marginals are the subtotals (either row or column)

N is the total number of cases in our cross tab..

- Crosstabs (or bivariate tables) provide evidence on potential "associations", i.e. two variables are said to be associated if the distribution of one variable changes for various categories of the other variable

11-52

For this course, we are following this convention:

- **Columns** will reflect different scores on the independent variable.
  - There will be as many columns as there are scores on the independent variable.
- **Rows** will reflect scores of the dependent variable.
  - There will be as many rows as there are scores on the dependent variable.

11-53

- Can calculate "column percentages".

Cell Counts and Column % **Level of Education**

		< H.S.	H. School Grad	Some Post Sec	
Smoking Behavior	**No**	60 60.00	100 71.43	300 83.33	460
	Yes	40 40.00	40 28.57	60 16.67	140
		100	140	360	600

100/140*100 60/360*100

Interpretation:
40% of < HS smoke, in contrast to 28.57% among HS graduates
And 16.67% among those with some college

Note: When working with a bivariate table!!!

If dependent variable is in your rows.. USE column % in interpretation.. The row %'s can potentially be very misleading..

If dependent variable happened to be in your columns, you would have to use the "row %" in interpretation!!

What if?

Sample of 690 clerical workers (1980)

		Independent		
		Women	Men	total
Dependent				
	smokers	65	45	110
	non-smokers	500	80	580
	Total	565	125	690

Row % or Column %???

What if?

Sample of 690 clerical workers (1980)

		Independent		
		Women	Men	total
Dependent				
	smokers	65	45	110
	non-smokers	500	80	580
	Total	565	125	690

Row %

		Independent		
		Women	Men	total
Dependent				
	smokers	59.1%	40.9%	100.0%
	non-smokers	86.2%	13.8%	100.0%
	Total			

OR?

Column %

		Independent		
		Women	Men	total
Dependent				
	smokers	11.5%	36.0%	
	non-smokers	88.5%	64.0%	
	Total	100.0%	100.0%	

What if?

Sample of 690 clerical workers (1980)

		Independent		
		Women	Men	total
Dependent				
	smokers	65	45	110
	non-smokers	500	80	580
	Total	565	125	690

Row %

		Independent		
		Women	Men	total
Dependent				
	smokers	X	40.9%	100.0%
	non-smokers		13.8%	100.0%
	Total			

OR?

Column %

		Independent		
		Women	Men	total
Dependent				
	smokers	11.5%	36.0%	
	non-smokers	88.5%	64.0%	
	Total	100.0%	100.0%	

Cell Counts and Column % Level of Education

		< H.S.	H. School Grad	Some Post Sec	
Smoking Behavior	**No**	60 60.00	100 71.43	300 83.33	460
	Yes	40 40.00	40 28.57	60 16.67	140
		100	140	360	600

Column %

OR (the exact same data) – both are okay, right?:

Row %

Level of education	Smoking	No	Yes	Total
	<H.S	60 60.0	40 40.0	100
	H. School Grad	100 71.4	40 28.6	140
	Some Post Sec.	300 83.3	60 16.7	360
	Total	460	140	600

- Interpret this table:

Independent variable

Incidence and % of Obesity by Province, 2008

Dependent variable		Nfld	PEI	NS	NB	Quebec
	Obese	173,298	36,998	230,913	229,299	1,739,628
	Not Obese	336,402	105,302	711,588	522,501	6,167,772
	Total	509,700	142,300	942,500	751,800	7,907,400

Interpretation

Not obvious with counts..

Can calculate column percentages to aid in interpretation since dependent variable is in the rows

Also: formal test of significance is possible… (chi square)

Interpretation?

Incidence and % of Obesity by Province, 2008

	Nfld	PEI	NS	NB	Quebec
Obese	173,298	36,998	230,913	229,299	1,739,628
	34.00%	26.00%	24.50%	30.50%	22.00%
Not Obese	336,402	105,302	711,588	522,501	6,167,772
	66.00%	74.00%	75.50%	69.50%	78.00%
Total	509,700	142,300	942,500	751,800	7,907,400
	100.00%	100.00%	100.00%	100.00%	100.00%

An association "appears to exist" between province of residence and obesity; the distribution of obese and non-obese vary across provinces e.g. 34% of Nfld are obese, as apposed to only 22% of Quebec residents NOTE: VERY LARGE #s here: LIKELY REAL!!!

What if we are working with relatively small numbers?

- Can we be sure an association (relationship) really exists for the larger population even if the %'s differ ???

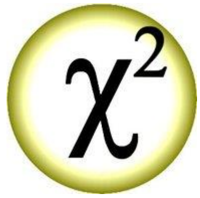
Incidence and % of Obesity by Province, 2008

	Nfld	PEI	NS	NB	Quebec
Obese	17	4	23	23	17
	33.33%	26.67%	24.47%	30.67%	21.52%
Not Obese	34	11	71	52	62
	66.67%	73.33%	75.53%	69.33%	78.48%
Total	51	15	94	75	79

- Numbers here are quite small.. Might the variation merely be the by-product of sampling error?
- There is a formal test to see whether the differences are significant or not -> chi square test..

11-62

Our Chi Square test is also called, the Chi Square test of ***"Independence"***….



What do we mean by "Independence" in this context?

The opposite of having an "association between two variables"… i.e. an absence of any type of association or relationship

11-63

- With this table? Is there a relationship between the two variables??

TABLE 11.2 THE CELL FREQUENCIES THAT WOULD BE EXPECTED IF RATES OF PARTICIPATION AND SEX WERE INDEPENDENT

	Sex					
Participation Rates	Male		Female			
High	50	66.67%	50	66.67%	**100**	**66.7**
Low	25	33.33%	25	33.33%	**50**	**33.3**
	75		75		**150**	**100**

Males are no more likely to participate than Females
NO RELATIONSHIP

"Independence"

- Two variables are **independent** if the classification of a case into a particular category of one variable has no effect on the probability that the case will fall into any particular category of the second variable.

- Let us return to our example with education and smoking…

Cell Counts and Column % Level of Education

		< H.S.	H. School Grad	Some Post Sec		
Smoking Behavior	No	60 60.00	100 71.43	300 83.33	460	77%
	Yes	40 40.00	40 28.57	60 16.67	140	23%
		100	140	360	600	100%

- Complete "Independence" would look like:

Expected frequencies, if we had independence..

Smoking behavior	< HS	H.School Grad	Some Post sec	
No	**77** **77%**	**107** **77%**	**276** **77%**	460 77%
Yes	**23** **23%**	**33** **23%**	**84** **23%**	140 23%
	100	**140**	**360**	600

Basic Logic of Chi Square TEST

- Again, a fundamental 5 step model!!!
- Question to answer:
- Does an "association" really exist? (given N)
- Or do we have "independence"?

- Chi Square, χ^2, is a test of significance based on bivariate, cross tabulation tables.
- Chi Square is a test for **independence**.

- Specifically, we are looking for significant differences between the *observed* cell frequencies in a table (f_o) and those that would be *expected* by random chance or if cell frequencies were **independent** (f_e):

11-66

Formulas for Chi Square

$$f_e = \frac{\text{Row marginal} \times \text{Column marginal}}{N}$$

.. Gives us our "expected frequencies" under assumption of "independence"

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

Formal test statistic
Step 4!

where f_o = the cell frequencies observed in the bivariate table
f_e = the cell frequencies that would be expected if the variables were independent

11-67

Computation of Chi Square: An Example

- Is there a relationship between support for privatization of healthcare and political ideology? Are liberals significantly different from conservatives on this variable?
  - The table below reports the relationship between these two variables for a random sample of 78 adult Canadians.

	Political Ideology		
Support	Conservative	Liberal	Total
No	14	29	43
Yes	24	11	35
Total	38	40	78

How do we calculate our "test statistic" in our chi squared test of independence?

Must first use:

$$f_e = \frac{\text{Row marginal} \times \text{Column marginal}}{N}$$

And then calculate:

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

where f_o = the cell frequencies observed in the bivariate table

f_e = the cell frequencies that would be expected if the variables were independent

11-69

An Example *(continued)*

Observed Frequencies (fo)

	Conservative	Liberal	Total
No	14	29	43
Yes	24	11	35
Total	38	40	78

$$f_e = \frac{\text{Row marginal} \times \text{Column marginal}}{N}$$

Use Formula 10.2 to find f_e.

– To obtain fe multiply column and row marginals for each cell and divide by *N*.

- (38*43)/78 = 1634 /78 = 20.9
- (40*43)/78 = 1720 /78 = 22.1
- (38*35)/78 = 1330 /78 = 17.1
- (40*35)/78 = 1400 /78 = 17.9

Expected frequencies (f_e)

Political Ideology

Support	Conservative	Liberal	Total
No	20.9	22.1	43
Yes	17.1	17.9	35
Total	38	40	78

11-70

Example:

Observed: (f_0)

	Political Ideology		
Support	Conservative	Liberal	Total
No	14	29	43
Yes	24	11	35
Total	38	40	78

Expected frequencies (f_e)

	Political Ideology		
Support	Conservative	Liberal	Total
No	20.9	22.1	43
Yes	17.1	17.9	35
Total	38	40	78

OUR test statistic tells us whether these are Significantly different!!

Example *(continued)*

- A computational table helps organize the computations.

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

	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
	14	20.9			
	29	22.1			
	24	17.1			
	11	17.9			
TOTAL	78	78			

- Subtract each f_e from each f_o. The total of this column *must* be zero.

	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
	14	20.9	-6.9		
	29	22.1	6.9		
	24	17.1	6.9		
	11	17.9	-6.9		
TOTAL	78	78	0		

- Square each of these values

	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
	14	20.9	-6.9	47.61	
	29	22.1	6.9	47.61	
	24	17.1	6.9	47.61	
	11	17.9	-6.9	47.61	
TOTAL	78	78	0		

Computation of Chi Square: An Example

(continued)

- Divide each of the squared values by the f_e for that cell. The sum of this column is chi square

	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
	14	20.9	-6.9	47.61	2.28
	29	22.1	6.9	47.61	2.15
	24	17.1	6.9	47.61	2.78
	11	17.9	-6.9	47.61	2.66
TOTAL	78	78	0		$\chi^2 = 9.87$

What to do with this chi square? 9.87?
The larger the chi square, the more likely the association is significant
We need a formal test…

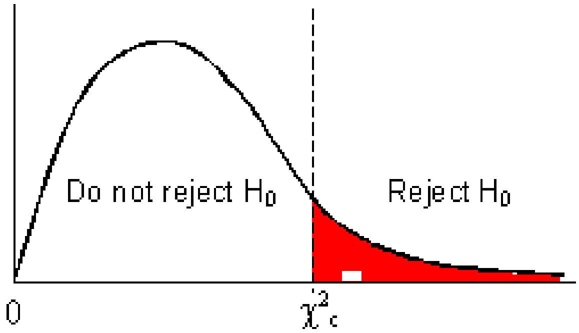
What about our "sampling distribution" and "critical score" in our Formal test?

Here, we use a sampling distribution called the

CHI square sampling distribution….

The Chi Square Distribution

- Type of sampling distribution
- The chi square distribution is asymmetric and its values are always positive (Appendix C).
- Its shape varies by the degrees of freedom involved in the test , which in turn is determined by the number of columns and rows in the table



Working with the chi square distribution

- χ^2 can be calculated for any bivariate table
- The shape of the χ^2 distribution is influenced by the number of rows and columns in the table df=(r-1)(c-1)

- The sampling distribution we are working with in this case (TABLE C) relates to all possible χ^2 under a hypothetical situation whereby we have independence with a table of given size (# of columns, # of rows)

- With our significance test, we work with this χ^2 distribution (with the null hypothesis that we have "independence"), and determine whether our test statistic χ^2 is likely or not,.. under this assumption

- If highly unlikely (we set our alpha at .05), we reject our null hypothesis, and conclude significance

- 95% confident that there is a relationship,.. If we set our alpha value at .05 and our test score falls within the critical area..

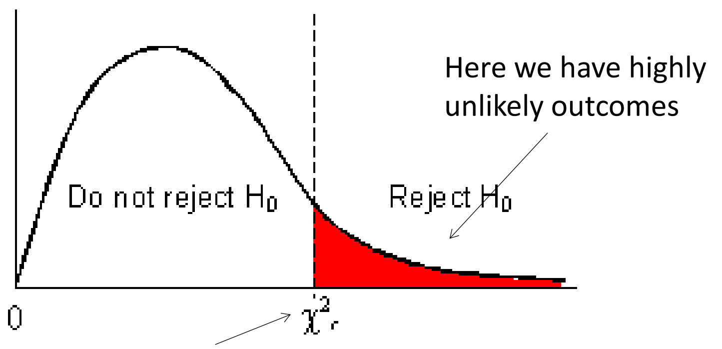
Appendix C Distribution of Chi Square

Critical values at alpha =.05

df	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.000	.001	.004	.016	.064	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341	16.268
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179

The Chi Square Distribution

- The chi square distribution is asymmetric and its values are always positive (Appendix C).
- Its shape varies by the degrees of freedom involved in the test



Appendix provides us with critical values for our test
We use an alpha of .05 unless otherwise specified

Back to our example

- Is there a relationship between support for privatization of healthcare and political ideology? Are liberals significantly different from conservatives on this variable?
  - The table below reports the relationship between these two variables for a random sample of 78 adult Canadians.

	Political Ideology		
Support	Conservative	Liberal	Total
No	14	29	43
Yes	24	11	35
Total	38	40	78

Performing the Chi Square Test Using the Five-Step Model

Step 1: Make Assumptions and Meet Test Requirements

- Independent random samples
- e.g. independent samples of conservatives & liberals
- Level of measurement is nominal
- e.g. support for privatization

11-82

Step 2: State the Null Hypothesis

- H_0: The variables are independent
  - Another way to state the H_0, more consistently with previous tests:
    - $H_0: f_o = f_e$
- H_1: The variables are dependent
  - Another way to state the H_1:
    - $H_1: f_o \neq f_e$

Step 3: Select Sampling Distribution and Establish the Critical Region

- Sampling Distribution = χ^2
- Alpha = .05
- df = (r-1)(c-1) = 1
- χ^2 (critical) = ?

2 rows and 2 columns, hence: df = 1

	Political Ideology		
Support	Conservative	Liberal	Total
No	14	29	43
Yes	24	11	35
Total	38	40	78

Appendix C Distribution of Chi Square

Critical values at alpha =.05

df	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.000	.001	.004	.016	.064	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341	16.268
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179

Step 3: Select Sampling Distribution and Establish the Critical Region

- Sampling Distribution = χ^2
- Alpha = .05
- df = (r-1)(c-1) = 1
- χ^2 (critical) = 3.841

Using Table C (page 510) in our appendix, we can indentify the χ^2 (critical) for alpha = .05
This χ^2 (critical) varies by the size of the table (# of rows/columns)

In this case, χ^2 (critical) allows us to identify in our sampling distribution a value of χ^2 which is quite unlikely, i.e. less than a 5% chance of getting it if our null hypothesis is true

11-86

Step 4. Get our test statisitc

Observed Frequencies (fo)

	Conservative	Liberal	Total
No	14	29	43
Yes	24	11	35
Total	38	40	78

$$f_e = \frac{\text{Row marginal} \times \text{Column marginal}}{N}$$

Use Formula 10.2 to find f_e.

- To obtain fe multiply column and row marginals for each cell and divide by *N*.
  - (38*43)/78 = 1634 /78 = 20.9
  - (40*43)/78 = 1720 /78 = 22.1
  - (38*35)/78 = 1330 /78 = 17.1
  - (40*35)/78 = 1400 /78 = 17.9

Expected frequencies (f_e)

Political Ideology

Support	Conservative	Liberal	Total
No	20.9	22.1	43
Yes	17.1	17.9	35
Total	38	40	78

Step 4: Calculate the Test Statistic

As demonstrated earlier:

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
14	20.9	-6.9	47.61	2.28
29	22.1	6.9	47.61	2.15
24	17.1	6.9	47.61	2.78
11	17.9	-6.9	47.61	2.66
78	78	0		$\chi^2 = 9.87$

Step 4: Calculate the Test Statistic

- χ^2 (obtained) = 9.87

Step 5: Make Decision and Interpret Results

- χ^2 (critical) = 3.841
- χ^2 (obtained) = 9.87
- The test statistic is in the Critical (shaded) Region:
  - We reject the null hypothesis of independence.
  - Opinion on healthcare privatization is associated with political ideology.