Comparing Two Proportions

Categorical Response Variable

- For a *categorical response variable*, inferences compare groups in terms of their population proportions in a particular category.
- Let p_1 represent the population proportion for the first group and p_2 the population proportion for the second group.
- We can compare the groups by their difference, $(p_1 - p_2)$. This is estimated by the difference of the sample proportions, $(\widehat{p_1} - \widehat{p_2})$.

Step 1: Assumptions

- Check assumptions

  - Population proportions are defined for each of the two groups

  - n_1 and n_2 are large enough, $(\mathrm{n}_1 + \mathrm{n}_2 > 30)$

Example:

Table 10.1 Whether Subject Died of Cancer, for Placebo and Aspirin Treatment Groups

	Death from Cancer		
Group	Yes	No	Total
Placebo	347	11,188	11,535
Aspirin	327	13,708	14,035

Is there a significant difference between the two groups? Use a 95% confidence level.

Step 1: Assumptions

- Both sample sizes are large enough

Step 2: State the Hypotheses

Null Hypothesis

$$H_0: p_1 = p_2 \; (p_1 - p_2 = 0)$$

Alternative Hypothesis

$$H_a: p_1 \neq p_2$$

Step 3: Compute the Test Statistic

$$z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{se_0}$$

where $se_0 = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$ and

$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$, the pooled estimate

Step 3: Compute the Test Statistic

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{347 + 327}{11535 + 14035} = 0.0264$$

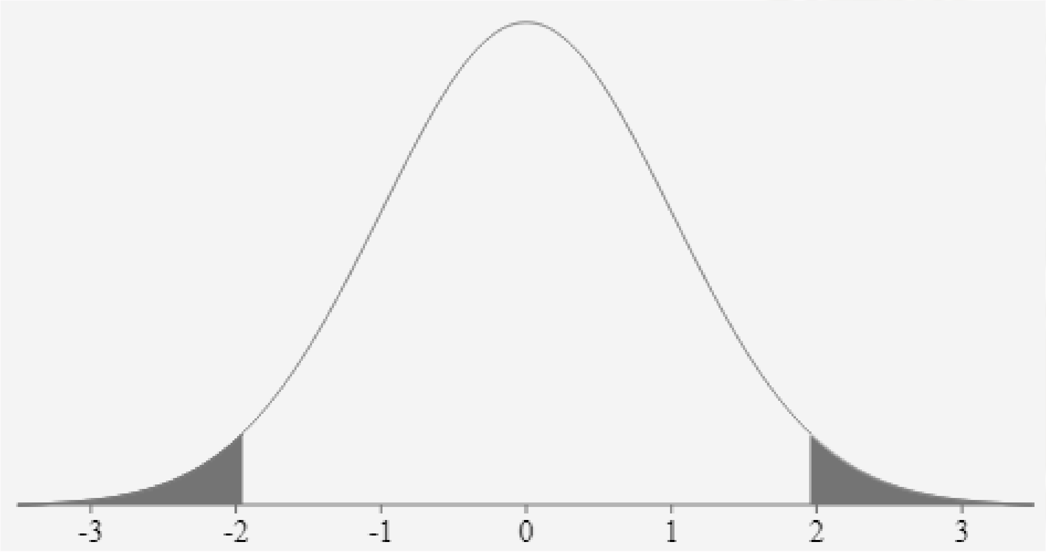
$$se_0 = \sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

$$= \sqrt{0.0264\,(1 - 0.0264)(\frac{1}{11535} + \frac{1}{14035})} = 0.002013$$

$$z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{se_0} = \frac{(347/11535 - 327/14035) - 0}{0.002013} = 3.37$$

Step 4: Interpret the Test Statistic (Using Rejection Region)

- 95% confidence level, $\alpha = 0.05$
- $z_c = \pm 1.96$



Step 4: Interpret the Test Statistic (Using p-value)

- 95% confidence level, $\alpha = 0.05$
- $z = 3.37$
- p-value = 2(0.5 - .4996) = 2(0.0004) = 0.0008

Step 5: Make a Conclusion

- Since the test statistic lies in the RR or
- since the p-value is less than α,

- then we reject the null hypothesis.

- Therefore, there is a difference between the proportions of the two groups.

Comparing Two Means

Independent Samples
Paired Samples

Samples

- Most comparisons of groups use **independent samples** from the groups. The observations in one sample are *independent* of those in the other sample.
- Significance test for comparing two means with **paired samples**, that is, sample observations are *naturally paired*. This is a result of a group being tested twice. Usually, this is used in determining the difference of means of before and after observations.

Independent Samples

Most comparisons of groups use **independent samples** from the groups. The observations in one sample are *independent* of those in the other sample.

Step 1: Assumptions

- Check assumptions
  - A quantitative response variable for the two groups
  - Independent random samples
  - Approximately normal population distribution for each group
  - For large sample sizes, use the z-table
  - For small sample sizes, $n_1, n_2 < 30$, use the t-table

Example:

Records of 40 used passenger cars and 40 used pickup trucks (none used commercially) were randomly selected to investigate whether there was any difference in the mean time in years that they were kept by the original owner before being sold. For cars the mean was 5.3 years with standard deviation 2.2 years. For pickup trucks the mean was 7.1 years with standard deviation 3.0 years.

Test whether there is a difference between the means. Use a 10% level of significance.

Step 1: Assumptions

- Both sample sizes are large enough (use the z-table)

Step 2: State the Hypotheses

Null Hypothesis

$$H_0: \mu_1 = \mu_2 \; (\mu_1 - \mu_2 = 0)$$

where μ_1 is the mean for the first group and μ_2 is the mean for the second group

Alternative Hypothesis

$$H_a: \mu_1 \neq \mu_2$$

Step 3: Compute the Test Statistic

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{se}$$

where $se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$\overline{x_1}$: mean of the first group

$\overline{x_2}$: mean of the second group

s_1: standard deviation of the first group

s_2: standard deviation of the second group

n_1: sample size of the first group

n_2: sample size of the second group

Step 3: Compute the Test Statistic

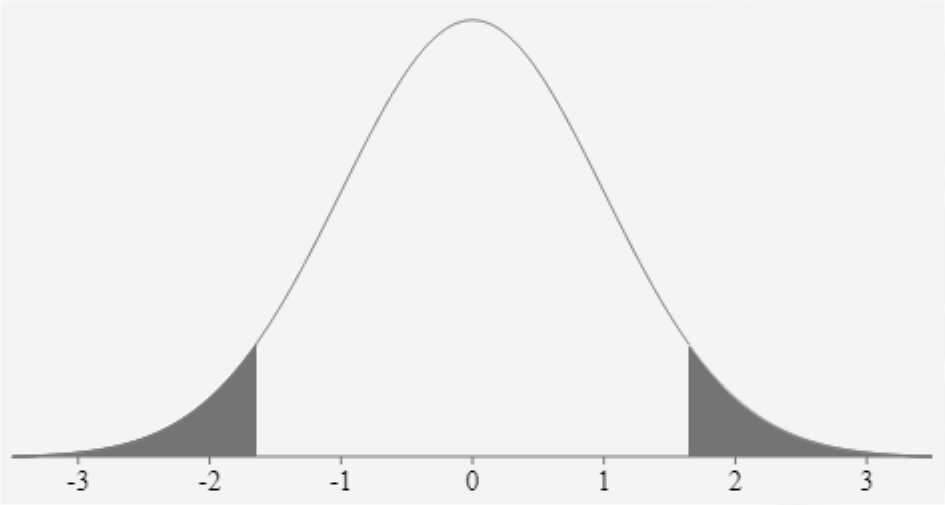
$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{2.2^2}{40} + \frac{3.0^2}{40}} = 0.5882$$

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{se} = \frac{(5.3 - 7.1) - 0}{0.5882} = -3.06$$

$$\overline{x_1} = 5.3$$
$$\overline{x_2} = 7.1$$
$$s_1 = 2.2$$
$$s_2 = 3.0$$
$$n_1 = 40$$
$$n_2 = 40$$

Step 4: Interpret the Test Statistic (Using Rejection Region)

- $\alpha = 0.1$
- $z_c = \pm 1.645$



Step 5: Make a Conclusion

- Since the test statistic lies in the RR
- then we reject the null hypothesis.
- Therefore, there is a difference between the means of the two groups.

Example:

A gardener sets up a flower stand in a busy business district and sells bouquets of assorted fresh flowers on weekdays. To find a more profitable pricing, she sells bouquets for Php15 each for ten days, then for Php10 each for five days. Her average daily profit for the two different prices are given below.

	n	$\overline{x}$	s
Php 15	10	171	26
Php 10	5	198	29

Test whether there is a difference between the means. Use a 10% level of significance.

Step 1: Assumptions

- Both sample sizes are small, $n_1, n_2 < 30$, use the t-table

- Equal variance: $df = n_1 + n_2 - 2$

- Unequal variance: $df = \dfrac{\left[\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right]^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$

Step 2: State the Hypotheses

Null Hypothesis

$$H_0: \mu_1 = \mu_2 \; (\mu_1 - \mu_2 = 0)$$

where μ_1 is the mean for the first group and μ_2 is the mean for the second group

Alternative Hypothesis

$$H_a: \mu_1 \neq \mu_2$$

Step 3: Compute the Test Statistic

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{26^2}{10} + \frac{29^2}{5}} = 15.3558$$

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{se} = \frac{(171 - 198) - 0}{15.3558}$$

$$= -1.76$$

Step 4: Interpret the Test Statistic (Using p-values)

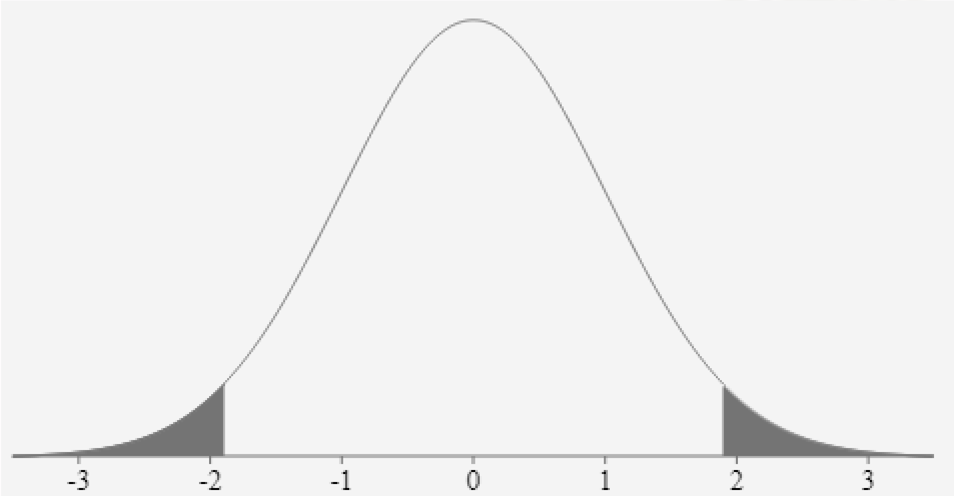
- $\alpha = 0.1$
- Use unequal variance,

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left[\frac{26^2}{10} + \frac{29^2}{5}\right]^2}{\frac{1}{10 - 1}\left(\frac{26^2}{10}\right)^2 + \frac{1}{5 - 1}\left(\frac{29^2}{5}\right)^2} = 7.33 \approx 7$$

Step 4: Interpret the Test Statistic (Using rejection region)

- $\alpha = 0.1$ (two-tailed), $\frac{\alpha}{2} = 0.05$
- df=7
- $t_c = \pm 1.894579$



Step 5: Make a Conclusion

- Since the test statistic lies in the FTR
- then we do not reject the null hypothesis.
- Therefore, there is no difference between the means of the two groups.

Paired Samples

a significance test for comparing two means with paired samples, that is, sample observations are naturally paired. This is a result of a group being tested twice. Usually, this is used in determining the difference of means of before and after observations.

Paired Samples

1. Compute the difference between the observations for each sample. Denote it by x_d
2. Compute for the sample mean difference, $\bar{x}_d$ of the difference scores, x_d
3. Hypotheses
- To test the hypothesis $H_0: \mu_1 = \mu_2$ of equal means, conduct a one-sample test of $H_0: \mu_d = 0$ with the difference scores.
- $H_0: \mu_d = 0 (or\ \mu_d \geq 0, or\ \mu_d \leq 0)$, where $\mu_d = \mu_2 - \mu_1$
- $H_a: \mu_d \neq 0 (or\ \mu_d < 0, or\ \mu_d > 0)$

4. Test Statistic
- $t = \frac{\bar{x}_d - 0}{se}$ where $se = \frac{s_d}{\sqrt{n}}$ where s_d is the standard deviation of the samples x_d

5. Compute for the p – value. The degree of freedom (*df*)= n – 1. Or use the rejection method.

6. Based on the p-value or according to the rejection region, make a decision about H_0. Relate the conclusion to the context of the study.

Example:

- Eight golfers were asked to submit their latest scores on their favorite golf courses. These golfers were each given a set of newly designed clubs. After playing with the new clubs for a few months, the golfers were again asked to submit their latest scores on the same golf courses. The results are summarized below.

Golfer	1	2	3	4	5	6	7	8
Own Clubs	77	80	69	73	73	72	75	77
New Clubs	72	81	68	73	75	70	73	75

- Test, at the 1% level of significance, the hypothesis that on average golf scores are the same with the new clubs.

Example:

Golfer	1	2	3	4	5	6	7	8
Own Clubs	77	80	69	73	73	72	75	77
New Clubs	72	81	68	73	75	70	73	75
x_d	5	-1	1	0	-2	2	2	2

$$\overline{x_d} = 1.125$$
$$sd = 2.167124$$
$$se = 0.766194$$
$$t = 1.468296 \approx 1.47$$

Example:

- $df = 7;\ \alpha = .01;\ \frac{\alpha}{2} = .005$
- $t_c = 3.49948$
- Hence, the test statistic lies in the FTR. So, we fail to reject the null hypothesis, $\mu_d = 0, \mu_1 = \mu_2$.
- Therefore, there is no significant difference between the before and after or the old club and the new club.

Test for Independence

Analyzing the Association Between Variables

Chi-Squared Test

Test for Independence

- explain the association between variables by doing a chi-squared test

- The test statistic for the test of independence measures how close the observed cell counts fall to the expected cell counts.

Test for Independence

- This test utilizes a contingency table to analyze the data. A contingency table (also known as a *cross-tabulation*, *crosstab*, or *two-way table*) is an arrangement in which data is classified according to two categorical variables.
- The categories for one variable appear in the rows, and the categories for the other variable appear in columns. Each variable must have two or more categories. Each cell reflects the total count of cases for a specific pair of categories.

Example:

In a study conducted by a pharmaceutical company, 605 out of 790 smokers and 122 out of 434 nonsmokers were diagnosed with lung cancer. Is smoking and lung cancer independent? Use $\alpha = 0.05$

Contingency Table (Observed Count)

	LUNG CANCER		
SMOKING	present	absent	Total
Smoker	605	185	790
Non-smoker	122	312	434
Total	727	497	1224

Step 1: Assumptions

- Data are from random sampling
- Expected Count is greater that or equal to 5 in all cells

Step 2: State the Hypotheses

Null Hypothesis

H_0: The variables are independent

Variables are ***independent*** if there is no association between the variables

Alternative Hypothesis

H_a: The variables are dependent (associated)

Variables are ***dependent*** if there is an association between the variables

Step 3: Compute the Test Statistic

$$X^2 = \Sigma \frac{(observed\ count - expected\ count)^2}{expected\ count}$$

where

$$expected\ count = \frac{row\ total\ \times column\ total}{total\ sample\ size}$$

(if not given in the problem)

Step 3: Compute the Test Statistic

Contingency Table (Observed Count)

	LUNG CANCER		
SMOKING	present	absent	Total
Smoker	$E_{1,1}$	$E_{1,2}$	790
Non-smoker	$E_{2,1}$	$E_{2,2}$	434
Total	727	497	1224

Expected Count

$$E_{1,1} = \frac{(727)(790)}{1224} = 469.22 \approx 469$$

$$E_{1,2} = \frac{(497)(790)}{1224} = 320.78 \approx 321$$

$$E_{2,1} = \frac{(727)(434)}{1224} = 257.78 \approx 258$$

$$E_{2,2} = \frac{(497)(434)}{1224} = 176.22 \approx 176$$

Step 3: Compute the Test Statistic

Contingency Table (Observed Count)

	LUNG CANCER		
SMOKING	present	absent	Total
Smoker	605	185	790
Non-smoker	122	312	434
Total	727	497	1224

Expected Count

	LUNG CANCER		
SMOKING	present	absent	Total
Smoker	469	321	790
Non-smoker	258	176	434
Total	727	497	1224

Step 3: Compute the Test Statistic

$$X^2 = \frac{(605 - 469)^2}{469} + \frac{(185 - 321)^2}{321} + \frac{(122 - 258)^2}{258} + \frac{(312 - 176)^2}{176}$$

$$= 273.84$$

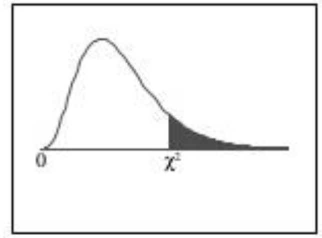
Step 4: Interpret the Test Statistic (Using Rejection Region)

- Compute for the critical value with the degree of freedom,
$$df = (r-1) \times (c-1)$$

- Use the chi-squared table

Step 4: Interpret the Test Statistic

Chi-Square Distribution Table

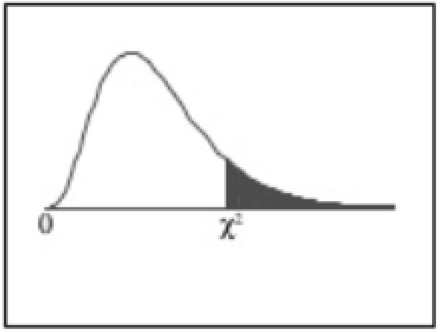


The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Step 4: Interpret the Test Statistic (Using Rejection Region)

- $\alpha = 0.05$
- $df = (2 - 1)(2 - 1) = 1$
- $X_c^2 = 3.841$



Step 5: Make a Conclusion

- Since the test statistic lies in the RR
- then we reject the null hypothesis.
- Therefore, the variables are dependent.