# Hypothesis Testing for Two Groups

David Armstrong

UCI

## Assumptions for 2 Groups

Difference in 2 population proportions

- $n_1\hat{p}_1 \geq 10$ and $n_1\hat{q}_1 \geq 10$
- $n_2\hat{p}_2 \geq 10$ and $n_2\hat{q}_2 \geq 10$
- Sample size is less than 10% of the population size; if we are sampling is w/out replacement
- The sample can be regarded as a simple random sample from the population of interest.
- The data values are assumed to be independent of each other.

Population mean of paired differences (2 dependent samples)

- We need to have a large enough sample size of pairs ($n > 30$). For $n < 30$ with extreme skewness or outliers, you cannot use this method.
- Sample size is less than 10% of the population size; if we are sampling is w/out replacement
- The sample can be regarded as a simple random sample from the population of interest.
- The differences between the pairs are assumed to be independent of each other.
- Differences within pairs are dependent.

Difference in 2 population means for independent samples

- We need to have a large enough sample size for each group ($n_1 > 30$ and $n_2 > 30$). For $n_1 < 30$ or $n_2 < 30$ with extreme skewness or outliers, you cannot use this method.
- Sample size is less than 10% of the population size; if we are sampling is w/out replacement
- The sample can be regarded as a simple random sample from the population of interest.
- The data values are assumed to be independent of each other.
- The groups are independent as well.

# Hypothesis Testing - Proportions

| Parameter Description | Test Statistic | $H_o$: |
|---|---|---|
| 1 population proportion | $Z = \dfrac{\hat{p} - p_o}{\sqrt{\dfrac{p_o q_o}{n}}}$ | $p = p_o$ |
| Difference in 2 population proportions | $\boxed{Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \quad \text{where } \hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}}$ | $\boxed{p_1 - p_2 = 0}$ |

# Hypothesis Testing - Means

| Parameter Description | Test Statistic | $H_o$: |
|---|---|---|
| 1 population mean<br>($\sigma$ known) | $Z = \dfrac{\bar{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}}$ | $\mu = \mu_o$ |
| 1 population mean<br>($\sigma$ unknown) | $t = \dfrac{\bar{X} - \mu_o}{\frac{s}{\sqrt{n}}}$<br>where $df = n - 1$ | $\mu = \mu_o$ |
| Population Mean<br>of paired differences<br>($\sigma_d$ known) | $Z = \dfrac{\bar{X}_d - 0}{\frac{\sigma_d}{\sqrt{n_d}}}$ | $\mu_d = 0$ |
| Population Mean<br>of paired differences<br>($\sigma_d$ unknown) | $t = \dfrac{\bar{X}_d - 0}{\frac{s_d}{\sqrt{n_d}}}$<br>where $df = n_d - 1$ | $\mu_d = 0$ |
| Difference in 2 population<br>means for independent samples<br>($\sigma_1$ and $\sigma_2$ known) | $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ | $\mu_1 - \mu_2 = 0$ |
| Difference in 2 population<br>means for independent samples<br>($\sigma_1$ and $\sigma_2$ unknown)<br>($\sigma_1 \neq \sigma_2$) | $t = \dfrac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$<br>where $df = min(n_1 - 1, n_2 - 1)$ | $\mu_1 - \mu_2 = 0$ |
| Difference in 2 population<br>means for independent samples<br>($\sigma_1$ and $\sigma_2$ unknown)<br>($\sigma_1 = \sigma_2$) | $t = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$<br>where $df = n_1 + n_2 - 2$ | $\mu_1 - \mu_2 = 0$ |

| Family | Sports | |
| --- | --- | --- |
| $n_2 = 200$ | $n_1 = 150$ | $\hat{p}_1 - \hat{p}_2 = 0.3333 - 0.3$ |
| $x_2 = np = 200(0.30) = 60$ | $x_1 = 50$ | $= 0.0333$ |
| $\hat{p}_2 = 0.30$ | $\hat{p}_1 = \frac{x}{n} = \frac{50}{150} = 0.3333$ | |

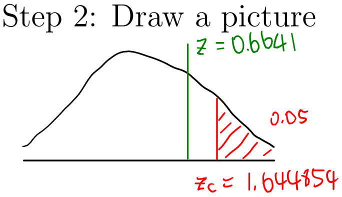
Ex: Sports car owners in a town complain that the state vehicle inspection station judges their cars differently from the family style cars. In a random sample of 200 family style cars, 30% failed the inspection the first time through. In a random sample of 150 sports cars, 50 failed the inspection on the first time through. Is there sufficient evidence to indicate that the proportion of first failures for sports cars is higher than the proportion for family style cars at the 0.05 significance level?

Step 1: Set up the null and alternative hypothesis

$H_0: P_1 = P_2 \longrightarrow P_1 - P_2 = 0$

$H_A: P_1 > P_2 \longrightarrow P_1 - P_2 > 0 \longrightarrow$ The failure rate for sports car owners is greater than that of family style cars

Step 2: Draw a picture




Assuming $H_0$ is true

$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{50 + 60}{150 + 200} = 0.3143$

Step 3: Calculate the test statistic Assuming $H_0$ is true

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.0333 - 0}{(0.3143)(1 - 0.3143)\left(\frac{1}{150} + \frac{1}{200}\right)} = 0.6640971$$

Step 4: Calculate the critical value

$z_c$ = qnorm(0.95, 0, 1) = 1.644854

Step 5: Make and justify a statistical decision at 0.05 level, and state your conclusions in context of the problem.

Since the test statistic is not more extreme than the critical value, we fail to reject the null hypothesis at the $\alpha = 0.05$ level. There is not statistically significant evidence the failure rate is greater than that of family style cars.

$\sigma_d = ?$

$\mu_d = 0$ $\quad \bar{x}_d = -3.7$ $\quad s_d = 6.73$ $\quad n_d = 10$ $\quad \alpha = 0.10$

Ex: A researcher wanted to show that after taking a full time job, the amount of time spent on leisure activities decreases. She took a random sample of 10 adults and asked them about the time they spend per week on leisure activities. After 1 year at their new full-time job, they were asked the same question. Their responses (in hours) are listed below with a mean difference of -3.7 and standard deviation 6.73. Assume that the times spent on leisure activities by all adults follows a bell shaped curve. Test the researcher's claim at the 0.10 significance level

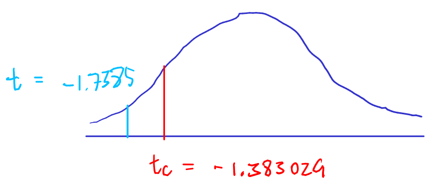
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pre | 14 | 25 | 22 | 38 | 16 | 26 | 19 | 23 | 41 | 33 |
| Post | 10 | 25 | 24 | 32 | 10 | 27 | 20 | 22 | 20 | 30 |

Step 1: Set up the null and alternative hypothesis

$H_0: \mu_d = 0$

$H_A: \mu_d < 0$ The time spent on leisure activities decreases after taking a full time job.

Step 2: Draw a picture




Assuming $H_0$ is true

Step 3: Calculate the test statistic

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n_d}}} = \frac{-3.7 - 0}{\frac{6.73}{\sqrt{10}}} = -1.7385$$

Step 4: Calculate the critical value

$t_c = qt(\alpha, df) = qt(0.10, n-1) = qt(0.10, 9) = -1.383029$

Step 5: Make and justify a statistical decision at 0.10 level, and state your conclusions in context of the problem.

Since the test statistic is more extreme than the critical value, we reject the null hypothesis at the $\alpha = 0.10$ level.
There is statistically significant evidence that the amount of time spent on leisure activities decreases

Drug | Control

$\bar{x}_1 = 1.05$ | $\bar{x}_2 = 1.2$ | $\bar{x}_1 - \bar{x}_2 = -0.15$

$s_1 = 0.5$ | $s_2 = 0.3$

$n_1 = 50$ | $n_2 = 75$ | $\alpha = 0.05$

Ex: Suppose a research neurologist is interested in testing the effect of a drug on response time by injecting 50 rats with a unit dose of the drug, subjecting each rat to a neurological stimulus, and recording the response time. The mean response time for this sample is 1.05 seconds. The mean response time is 1.2 seconds for a sample of 75 rats that are not injected with the drug. Suppose that the drug injected rats have a sample standard deviation of 0.5 and the control rats have a sample standard deviation of 0.3. Does the mean response time from drug-injected rats differ from those that are not injected? Test the researcher's claim at the $\alpha = 0.05$ level.

Step 1: Set up the null and alternative hypothesis

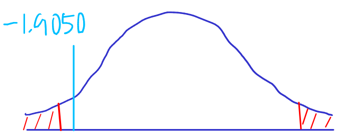
$H_0: \mu_1 = \mu_2 \longrightarrow \mu_1 - \mu_2 = 0$

$H_A: \mu_1 \neq \mu_2 \longrightarrow \mu_1 - \mu_2 \neq 0$ The average response time for the two groups is not the same.

Assume $H_0$ is true.

Step 2: Draw a picture

$t = -1.9050$



$t_c = -2.0096$ $t_c = 2.0096$

$$\text{p-value} = 2(pt(t, df)) = 2(pt(-1.9050, 49)) =$$

No mistake = specificity

Mistake = Type II Error

Step 3: Calculate the test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{-0.15 - 0}{\sqrt{\frac{0.5^2}{50} + \frac{0.3^2}{75}}} = -1.9050$$

Step 4: Calculate the critical value

$$t_c = qt\left(1 - \frac{\alpha}{2}, df = n - 1\right) = qt(0.975, 49) = \pm 2.0096$$

Step 5: Make and justify a statistical decision at 0.05 level, and state your conclusions in context of the problem.

Since the test statistic is not more extreme than the critical value, we fail to reject the null hypothesis at the $\alpha = 0.05$ level. There is not statistically significant evidence that the average response time for the two groups is not the same.

## $\alpha$ levels

Don't believe too strongly in an arbitrary alpha level. It is used as a guideline!

## Practical Versus Statistical Significance

At some point in the analysis of data from a study, researchers face the question, "How important are these results?" When a large sample size is used in hypothesis testing, the results may be statistically significant even though the difference between the statistic and the null value may have no practical significance (meaning it is not large enough to be considered important in the real world).

## Errors in Hypothesis Testing

Because a statistician must make inferences (or conclusions) based on random data that is subject to sampling errors, we can make mistakes in hypothesis testing. In fact, there are two types of errors that can be made:

- Type I error = the null hypothesis is true, but we mistakenly reject it.

  The probability of a Type I error is denoted by $\alpha$ and is called the level of significance of the test. This threshold is where we reject the null hypothesis. If that is a mistake, than the probability of making that error is $\alpha$.

- Type II error = the null hypothesis is false, but we fail to reject it.

  The probability of a Type II error is denoted by $\beta$ .

<table>
<tr><td></td><td colspan="3">Statistical Decision</td></tr>
<tr><td rowspan="3">Truth</td><td></td><td>REJECT $H_o$</td><td>FAIL TO REJECT $H_o$</td></tr>
<tr><td>$H_o$ TRUE</td><td>$\alpha$<br>(Type I Error)</td><td>correct decision<br>(specificity of the test)</td></tr>
<tr><td>$H_a$ TRUE</td><td>power<br>(sensitivity of the test)</td><td>$\beta$<br>(Type II error)</td></tr>
</table>

- Reducing Type I and Type II Errors:

  The only way to reduce both errors is to increase sample size.

  We have more information about our population parameter

  and a better estimate of reality.

# Statistical Power

- Power of the test

  1- Type II error

  The test's ability to detect a false hypothesis.

- The power depends on effect size.

  This is the distance between the null value and the true parameter.

  The larger the effect size, the greater the power of the test.

EX:

Let $p$ be the proportion of parachute failures for a skydiving company.
$H_o$: $p = 0.0001$ $H_a$: $p > 0.0001$
Draw a truth table for this scenario:

| | Reject $H_0$ $P > 0.0001$ | Fail to Reject $H_0$ $P = 0.0001$ |
|---|---|---|
| $H_0$ is True $P = 0.0001$ | Type I Error | Specificity |
| $H_0$ is False $H_a$ is True $P > 0.0001$ | Power Sensitivity | Type II Error |

Describe in words a Type I Error, Type II Error, specificity, and Power:

Type I Error : we conclude the proportion of parachute failures is greater than 0.0001. However, we were wrong

Type II Error : we conclude the proportion of parachute failures is 0.0001. However, we were wrong.

Specificity : we conclude the proportion of parachute failures is 0.0001 and we were correct.

Power Sensitivity : we conclude the proportion of parachute failures is greater than 0.0001 and we were correct.