# Hypothesis Testing II

ANA 500 – Foundations of Data Analytics

Module 2 - week 4A

# Hypothesis testing with two samples

- We are often interested in **making comparison between groups**.
- Is there a difference in average salary between male and female lawyers?
- Is there a difference in the proportion of times students are late to class between public and private schools?
- Is there a difference in the average price of a 4-star hotel room between Washington DC and Baltimore?
- Is there a difference in the proportion of households with internet access between those living in the North versus those living in the South?

- We will almost always calculate a difference using random samples, **what we want to know** is whether this is a true difference or simply due to random chance.

# Hypothesis testing with two samples

- The two groups can be independent or matched pairs:
  - Independent groups consist of two samples from two independent populations (e.g. population 1 is female and population 2 is male)
  - Matched pairs are two samples that are dependent (e.g. completion time before training and completion time after training).
- All that is really changing compared to hypothesis testing with one mean is the type of question being asked. The approach to the test will be the same.
  - Set up hypothesis, determine distribution, calculate test statistic and p-value, make decision.

# Hypothesis testing with two samples

- We know, thanks to the CLT, that the distribution of a mean is normal. It is also true that the distribution of differences in means is normal.
- We use a standard normal distribution if we know the population standard deviations.
- we use a student's T distribution if the standard deviations need to be estimated.
- For differences in means the standard error is estimated by:

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

# Hypothesis testing with two samples

- The t-stat is still the difference between our estimate and the value being tested (i.e. the value specified in Ho) divided by the standard error.
  - If Ho is true (which is assumed) how many standard deviations is our estimate from the mean?

$$\text{t-stat} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The null hypothesis being a statement of no difference or no effect.

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

# Hypothesis testing with two samples

- Degrees of freedom = $$\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{s_1^2}{n_1}\right)^2+\left(\frac{1}{n_2-1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$$

# Hypothesis testing with two samples-example

- E.g. Is there a difference in the average price of a 4-star hotel room between Washington DC and Baltimore?
- Denote Washington DC group 1 and Baltimore group 2

- Suppose we have the following sample statistics:
  - $\overline{x_1}$ =290, $\overline{x_2}$ =270,
  - $n_1$ =30,$n_2$ =22,
  - $s_1$ =40,$s_2$ =32

$$H_0\text{: } \mu_1 - \mu_2 = 0$$

$$H_0\text{: } \mu_1 - \mu_2 \neq 0$$

# Hypothesis testing with two samples-example

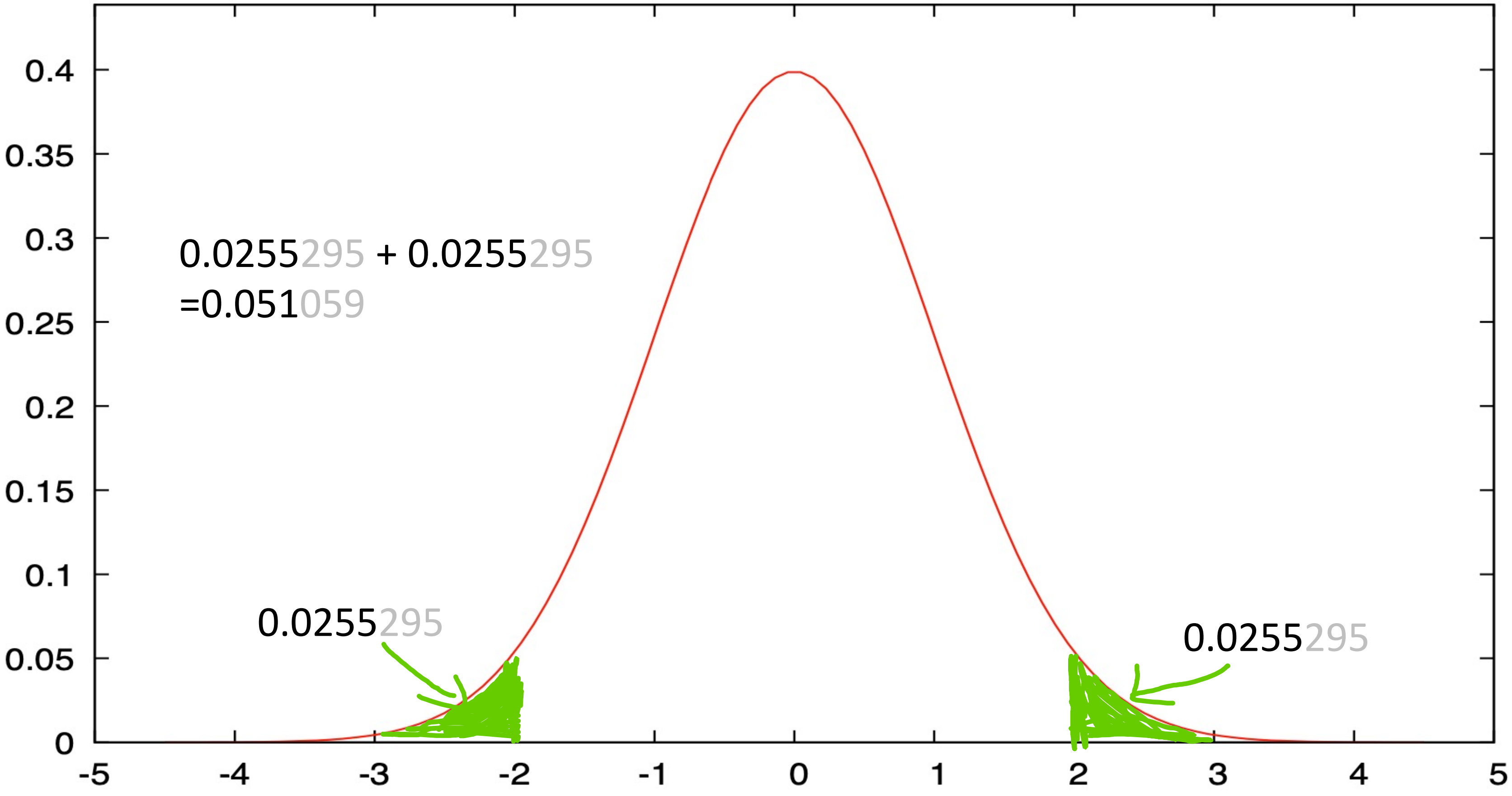
$$\text{t-stat} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(290-270)-(0)}{\sqrt{\frac{40^2}{30}} + \sqrt{\frac{32^2}{22}}} = \frac{20}{\sqrt{99.88}} = 2$$

The degrees of freedom is equal to 49.57 It‘s always safer to round down, so we would use degrees of freedom equal to 49.

P-value = 0.051

- 0.051 > 0.05 --- fail to reject Ho, there is not a statistically significant difference between Washington DC and Baltimore in the average price of a hotel room.

# Hypothesis testing with two samples-example



# Hypothesis testing with two samples-proportions

- A similar approach is used for testing differences in population proportions.
- Assume: two independent random samples with at least 5 “successes” and 5 “failures” in each sample.
- If two sample proportions are different, it could be because they really are in the population or because of random chance--this is what we want to know.
- Differences in proportions follow a normal distribution.

# Hypothesis testing with two samples-proportions

- A “pooled proportion” is used to conduct the test.

$$p_c = \frac{x_a + x_b}{n_a + n_b}$$

Where

- $x_a$ is the number of successes in the first group.
- $x_b$ is the number of successes in the second group.
- $n_a$ is the number of observations or trials in the first group.
- $n_b$ is the number of observations or trials in the second group.

# Hypothesis testing with two samples-proportions

$$p'_a - p'_b \sim N\left(0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_a}+\frac{1}{n_b}\right)}\right.$$

- Our test statistic, which again will measure how many standard deviations are estimated Difference is from the mean of zero

$$\text{z-stat} = \frac{(p'_a - p'_b) - (p_a - p_b)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_a}+\frac{1}{n_b}\right)}}$$

# Hypothesis testing with two samples-proportions

- E.g. Is there a difference in the proportion of households with internet access between those living in the North versus the South?
- Denote North group a and South group b

- Suppose we have the following sample statistics:
- $p'_a = 0.74$, $p'_b = 0.68$,
- $n_a = 42$, $n_b = 38$,
- $p'_c = \frac{31+26}{42+38} = 0.71$

$$H_0: p_a - p_b = 0$$

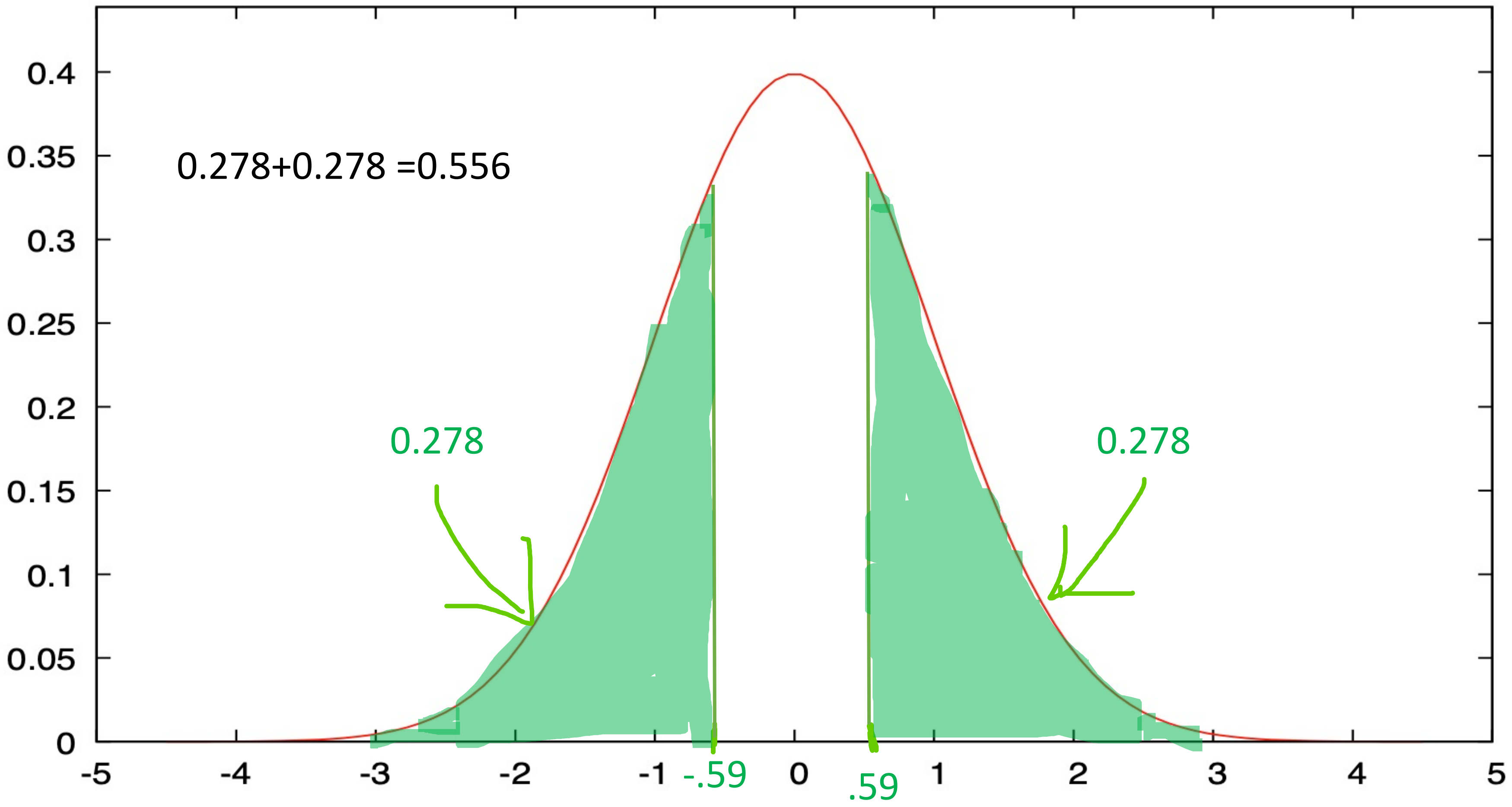
$$H_0: p_a - p_b \neq 0$$

# Hypothesis testing with two samples-proportions

$$\text{z-stat} = \frac{(p'_a - p'_b) - (p_a - p_b)}{\sqrt{p_c(1-p_c)(\frac{1}{n_a}+\frac{1}{n_b})}} = \frac{(0.74-0.68)-(0)}{\sqrt{(0.71)(1-0.71)\ (\frac{1}{42}+\frac{1}{38})}} = \frac{0.06}{\sqrt{0.0103}} = 0.59$$

- p-value = 0.555
- 0.555 > 0.05--fail to reject Ho, there is not a statistically significant difference between the proportion of households with internet access in the North versus the South.

# Hypothesis testing with two samples-proportions



# Hypothesis testing with two samples

- Matched samples

  - We have two measurements or two samples from the same entity.

  - Differences between the two samples are calculated. The differences are then used as a sample of data to calculate statistics of interest. We assume matched pairs have differences that come from a normal population, or the sample is large enough that the distribution is approximately normal.

# Hypothesis testing with two samples

- Matched samples
  - E.g. Does a new production process increase output produced, on average?

| Employee | Output using old production process | Output using new production process |
|---|---|---|
| A | 54 | 58 |
| B | 56 | 57 |
| C | 62 | 62 |
| D | 58 | 60 |
| E | 48 | 49 |
| F | 53 | 52 |
| G | 63 | 65 |
| H | 66 | 66 |

# Hypothesis testing with two samples

| Employee | Output using old production process | Output using new production process | Difference |
|---|---|---|---|
| A | 54 | 58 | 4 |
| B | 56 | 57 | 1 |
| C | 62 | 62 | 0 |
| D | 58 | 60 | 2 |
| E | 48 | 49 | 1 |
| F | 53 | 52 | -1 |
| G | 63 | 65 | 2 |
| H | 66 | 66 | 0 |

- The average difference = $\bar{x}$, = 1.125
- The standard deviation of the difference = $s$, = 1.55
- This test is conducted using a Student's t-distribution with $(n - 1)\ df$

# Hypothesis testing with two samples

One tailed test:

- Ho: $\mu_d = 0$
- Ha: $\mu_d > 0$

- t-stat = $\frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{1.125 - 0}{\frac{1.55}{\sqrt{8}}} = 2.05$

- p-value = 0.0398 < 0.05---reject Ho, the new production process increases output produced.

# Hypothesis testing with two samples