STAT 201

Week 8

Lecture goals:

By the end of this lecture, the students are expected to be able to:

- Describe a t-distribution and its relationship with the normal distribution.
- Use results from the assumption of normality or the Central Limit Theorem to perform estimation and hypothesis testing.
- Compare and contrast the parts of estimation and hypothesis testing that differ between simulation- and resampling-based approaches with the assumption of normality or the Central Limit Theorem-based approaches.
- Write a computer script to perform hypothesis testing based on results from the assumption of normality or the Central Limit Theorem.
- Discuss the potential limitations of these methods.

Central Limit Theorem

Central Limit Theorem

- Let $x_1, x_2, \dots, x_n$ be a random sample from a population.

- The CLT states that for large sample sizes (large n) the sampling distribution of the sample mean (or sample proportion) will converge to the Normal distribution.

- $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ $\qquad$ $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Central Limit Theorem: Assumptions

- The central limit theorem makes the following assumptions:
  - The sample is drawn in an independent fashion.

  - In general, if your sample size is greater than 10% of the population size, there will be a severe violation of independence.

  - The sample size must be large enough.
    - For the proportion, you can check if $np \geq$ and $n(1-p) \geq 10$.
    - For the sample mean, there is no universal guideline, and we might need a large sample size. Usually, however, sample sizes between 30 and 50 are enough to get a reasonable approximation (but it is not guaranteed).

Student’s t distribution family

Student's t

- Imagine a random variable X mean μ and standard deviation σ.

- Given that the CLT conditions are satisfied, the sample average, $\bar{X}$, is normally distributed with mean μ and standard error $\sigma/\sqrt{n}$.

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- We can standardize $\bar{X}$ by calculating the Z-score:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Student’s t

- We can standardize $\bar{X}$ by calculating the Z-score:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- However, in almost all cases in practice, σ is also unknown. Then, we’d need to replace σ with its estimator, s.

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- Problem solved!

Student's t

- We can standardize $\bar{X}$ by calculating the Z-score:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- However, in almost all cases in practice, σ is also unknown. Then, we'd need to replace σ with its estimator, s.

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- Problem solved! Is it?

Student’s t

- We can standardize $\bar{X}$ by calculating the Z-score:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- Note that we are replacing a constant (a fixed parameter), σ, with a random quantity (a statistic), s.

- As we know, s will also change from sample to sample. Therefore, by using

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

the formula is changing from sample to sample.

Student's t

- For large values of n, the standard error of s will go down, and s will be more stable and closer to σ. Therefore, we can expect T and Z to have similar distribution: $N(0,1)$.

- But for not so large values of n, s will be more unstable, with a higher standard error. In this case, besides the variation of $\bar{X}$, we need to account for the extra variation brought on by s.

- The Student's t distribution family accounts for this extra uncertainty.

Properties of Student’s t

- The Student’s t distribution family is indexed by one parameter only, ν, the *degrees of freedom*.

- The higher the *degrees of freedom*, the closer the t distribution is to the $N(0,1)$.

- Student’s t distribution:
  - is always centred around 0;
  - is symmetric;
  - has heavier tails than the Gaussian distribution (for low values of ν).

Hypothesis Testing with CLT

One-proportion z-test

- Remember the steps for hypothesis testing:
  1. Formulate the hypotheses;
  2. Define the test statistic;
  3. Calculate the null model (i.e., the sampling distribution of the test statistic if H_0 were true)
  4. Get a sample;
  5. Calculate the observed test statistic;
  6. Contrast the observed test statistic with the null model to assess if there is enough evidence to reject H_0.

- The only thing we are going to change from what you’ve done in week 6, is Step 3. This time we will use the CLT to obtain the null model, instead of simulation.

Hypothesis testing for one proportion

One-proportion z-test

- Suppose we want to check if the proportion of white and red balls in a box is the same.

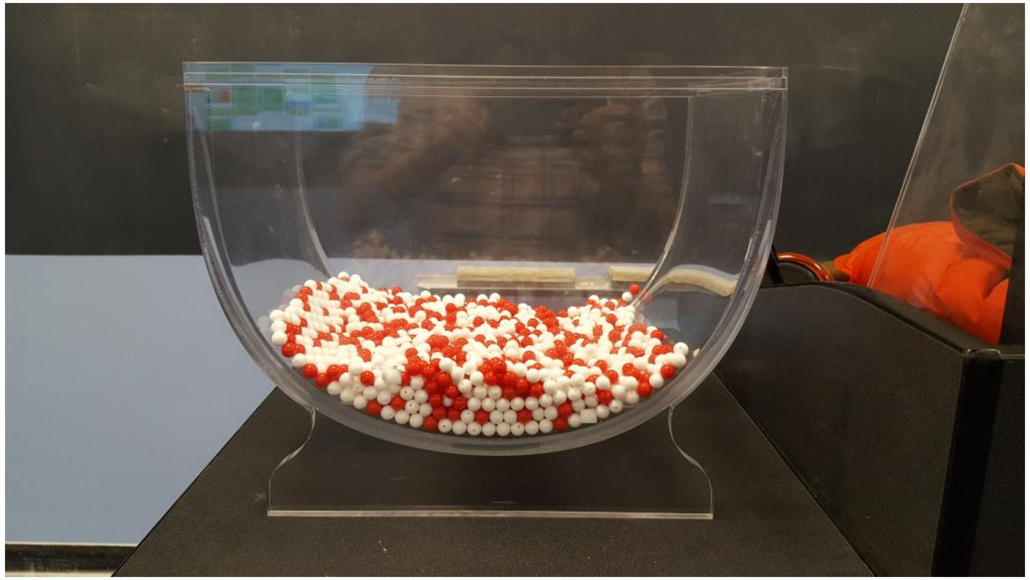


Image: Modern Dive

One-proportion z-test

Let p be the true but unknown proportion of red balls.

- Step 1: Formulate the hypothesis

$$H_0: p = 0.5 \quad vs \quad H_1: p \neq 0.5$$

- Step 2: Define the test statistic
  - Sample proportion!

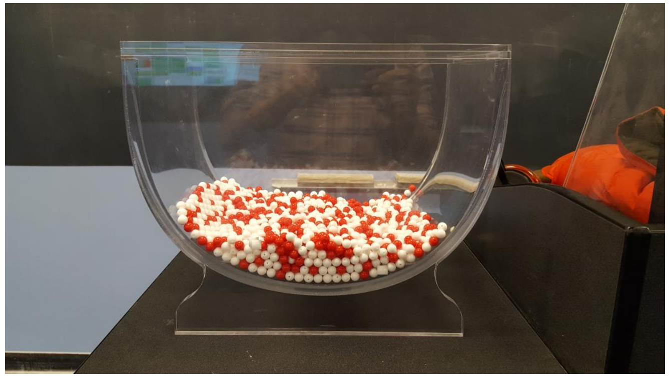


Image: Modern Dive

One-proportion z-test

- Step 3: Calculate the null model.
    - If the sample is: (1) taken at random; (2) independent; and (3) $np \geq 10$ and $n(1-p) \geq 10$, by the CLT we know that:

$$\hat{p} \sim N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

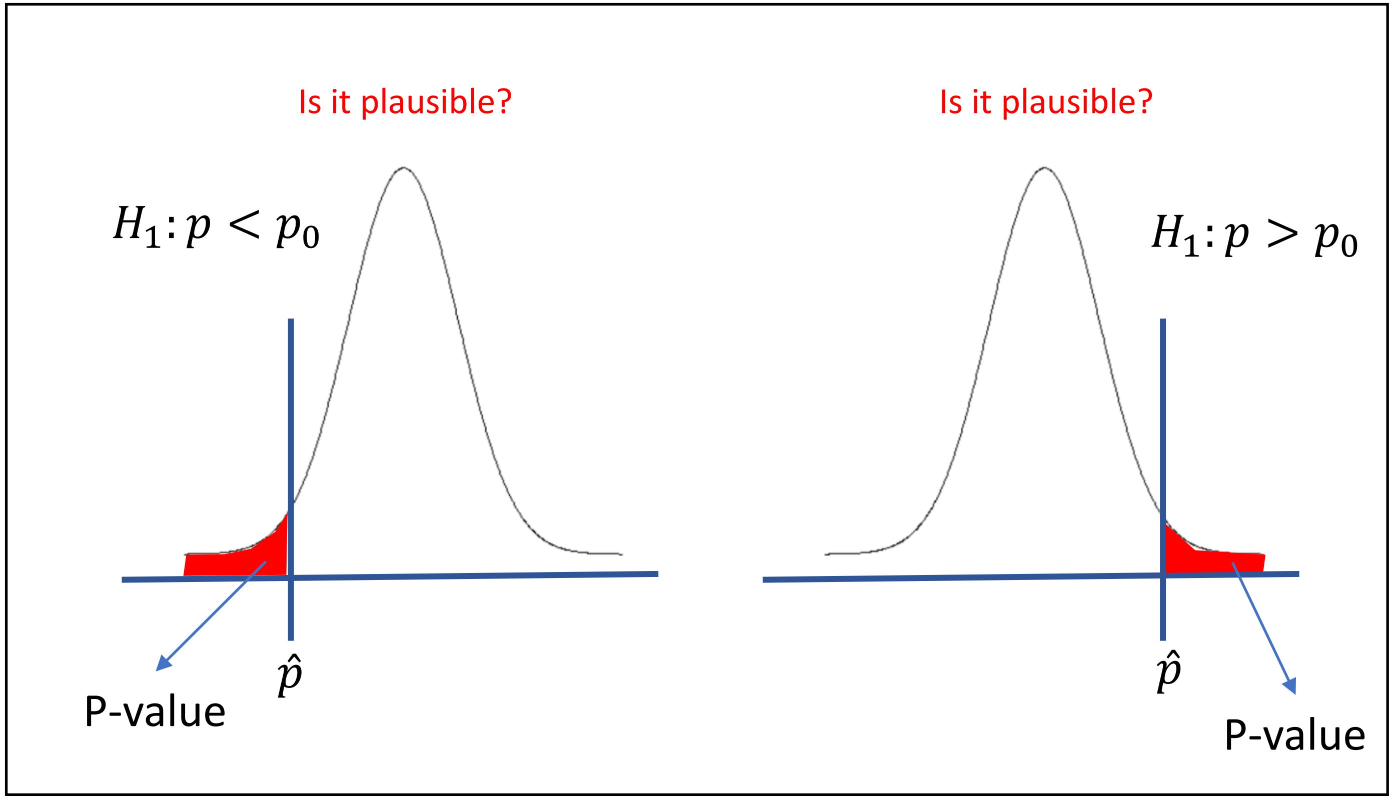
- Step 4: Get a sample.
    - Suppose we got a sample with 40 red balls and 60 white balls.

- Step 5: Calculate the observed test statistic: $\hat{p} = 0.40$.

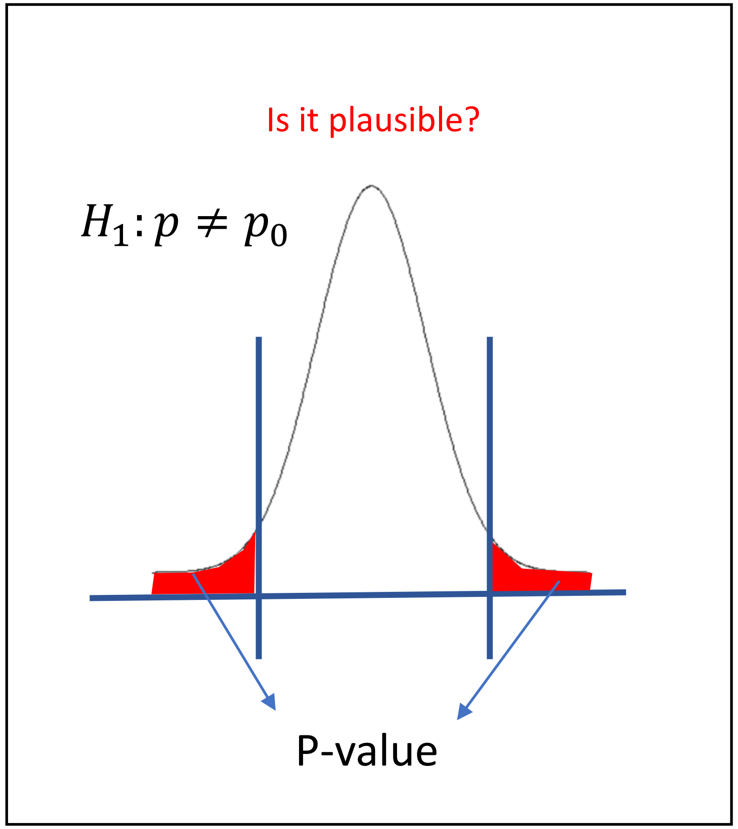
One-proportion z-test

- Step 6: Contrast the observed test statistic with the null model by calculating the p-value.

One-tailed tests



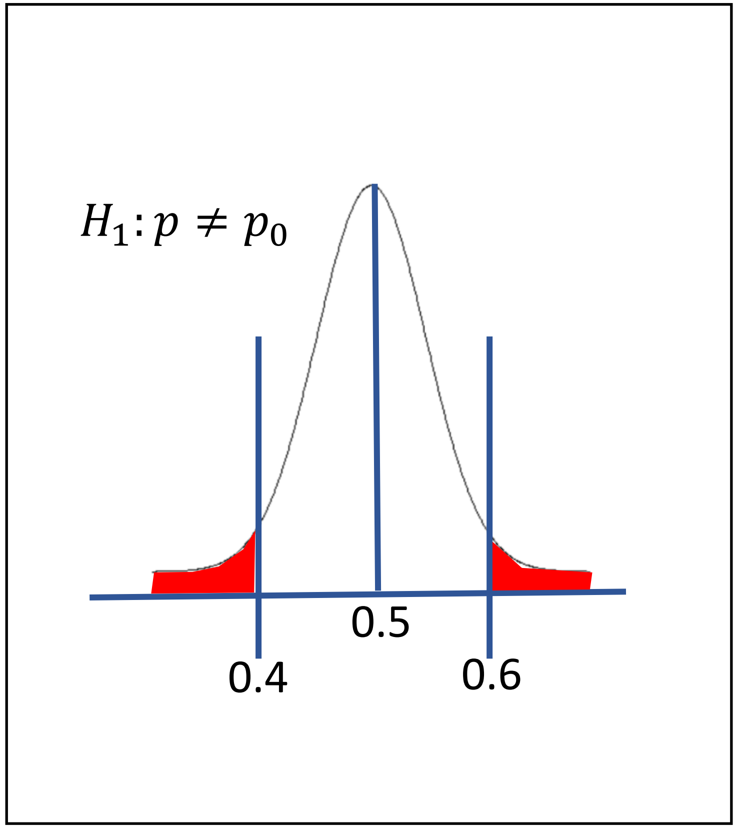
Two-tailed test



One-proportion z-test

- Step 6: Contrast the observed test statistic with the null model by calculating the p-value.

Two-tailed test



$$p_0 = 0.5$$
$$\hat{p} = 0.4$$

Null model: $N\left(0.5, \sqrt{\frac{0.25}{100}}\right)$

p-value $= 2P(\hat{p} < 0.4 | H_0 \text{ is true}) = 0.0455$

Hypothesis testing for difference in proportions

Two-proportion z-test

- Suppose we have two bins and want to check if the proportion of red balls in both bins are the same.

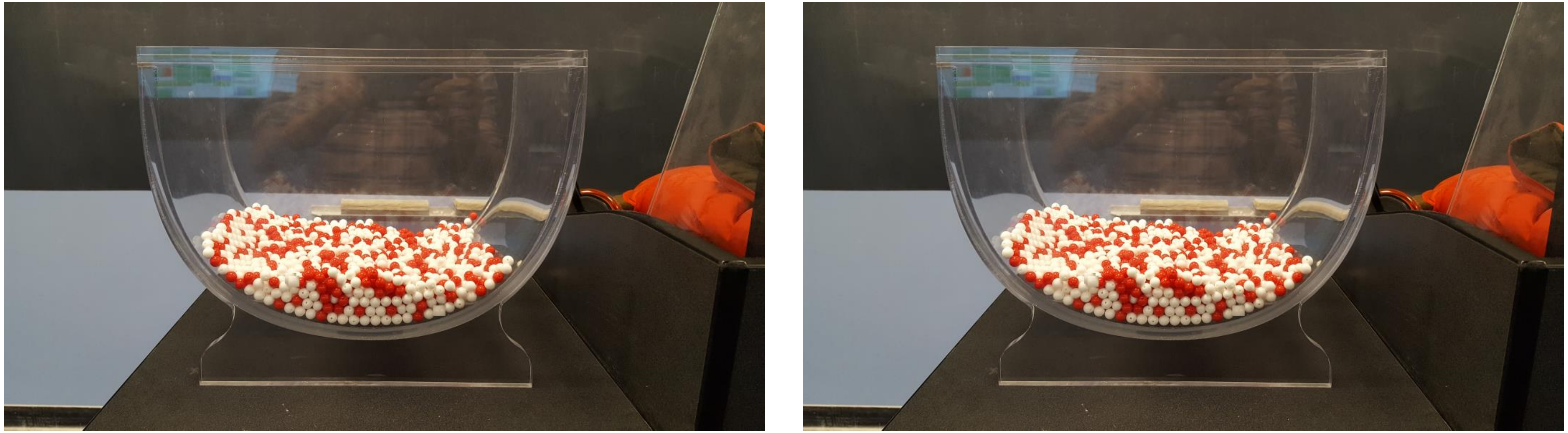


Image: Modern Dive

Two-proportion z-test

Let p_1 and p_2 be the true but unknown proportions of red balls in bin 1 and 2, respectively.

- Step 1: Formulate the hypothesis

$$H_0: p_1 = p_2 \quad vs \quad H_1: p_1 \neq p_2$$

- Step 2: Define the test statistic
  - Difference of sample proportions!

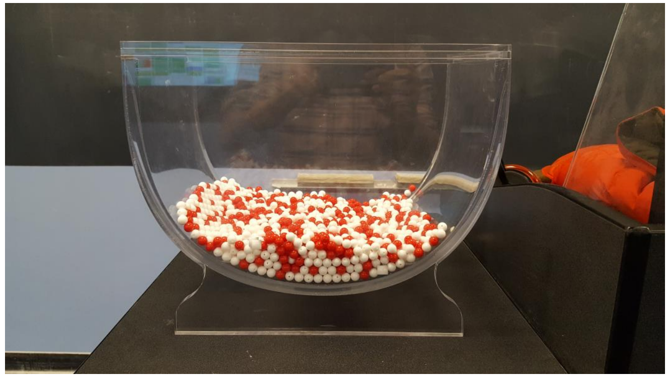


Image: Modern Dive

Two-proportion z-test

- Step 3: Calculate the null model.
  - If the sample is: (1) taken at random; (2) independent; and (3) $np \geq 10$ and $n(1-p) \geq 10$, by the CLT we know that:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(0, \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

  where p is the overall proportion (i.e., the total number of red balls in your sample divided by the total size of your samples)

- Step 4: Get a sample.
  - Suppose we got a sample with 40 red balls and 60 white balls from bin 1 and 20 red balls and 40 white balls from bin 2.

- Step 5: Calculate the observed test statistic: $\hat{p} = 0.40$.

Two-proportion z-test

- Step 4: Get a sample.
    - Suppose we got a sample with 40 red balls and 60 white balls from bin 1 and 20 red balls and 40 white balls from bin 2.

- Step 5: Calculate the observed test statistic: $\hat{p}_1 = 0.40$ and $\hat{p}_2 = 0.5$:

$$\hat{p}_1 - \hat{p}_2 = -0.1$$

Two-proportion z-test

- Step 6: Contrast the observed test statistic with the null model by calculating the p-value.

$$p = \frac{40 + 20}{100 + 60} = \frac{60}{160} = 0.375$$

$$\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \approx 0.0791$$

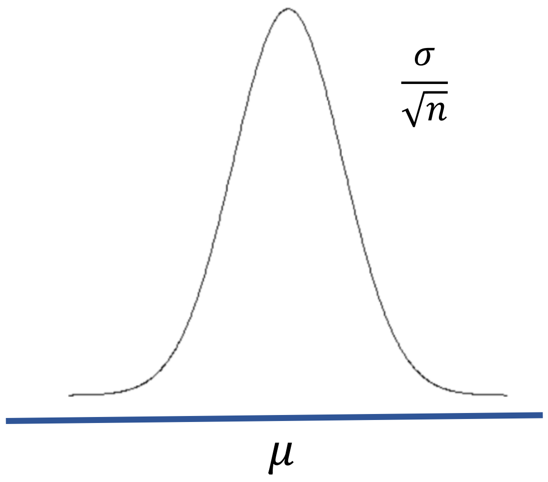
Null model: $\hat{p}_1 - \hat{p}_2 \sim N(0, 0.0791)$

p-value $= 2 \times pnorm(-0.1, 0, 0.0791) = 0.2062$

Testing hypotheses for the mean

Confidence intervals based on CLT: Mean

- The process for the mean is the same! We just need to figure out our test statistic and null model.



To test

- $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$
- $H_0: \mu = \mu_0$ vs $H_1: \mu < \mu_0$
- $H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$

- Assuming the CLT conditions are satisfied, we have that:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Testing hypotheses for difference in means

Confidence intervals based on CLT: Mean

To test

- $H_0: \mu_1 - \mu_2 = \delta_0$ vs $H_1: \mu_1 - \mu_2 \neq \delta_0$
- $H_0: \mu_1 - \mu_2 = \delta_0$ vs $H_1: \mu_1 - \mu_2 < \delta_0$
- $H_0: \mu_1 - \mu_2 = \delta_0$ vs $H_1: \mu_1 - \mu_2 > \delta_0$

- Assuming the CLT conditions are satisfied, we have that:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$$

where the degrees of freedom ν is $\min(n_1, n_2)$. Note, in fact, we are going to see a much better approximation of ν in the Worksheet.