Supplementary Handout for Dis 12: Inference about Two Populations

1 Motivation

- Last discussion, we talked about how to conduct hypothesis testings on parameters obtained from one population.
- This week, we will look at how to conduct hypothesis for parameters obtained from two different populations.
- One immediate question: how to test something like

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

  where both sides of the equation contain something unknown (μ_1 on the left hand side, μ_2 on the right hand side)?

  – Solution: we can rewrite the above hypotheses as the following:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

    Now the left hand side is something unknown related to the population parameters & waiting to be tested, and the right hand side is a concrete number to test the unknown against.

- Similar to last discussion about the one population case, we will discuss how to test / compare the following three sets of population parameters:
  1. The population means ($\mu_1 - \mu_2$)
  2. The population variances ($\frac{\sigma_1^2}{\sigma_2^2}$)
  3. The population success proportions ($p_1 - p_2$, from two binomial experiments)

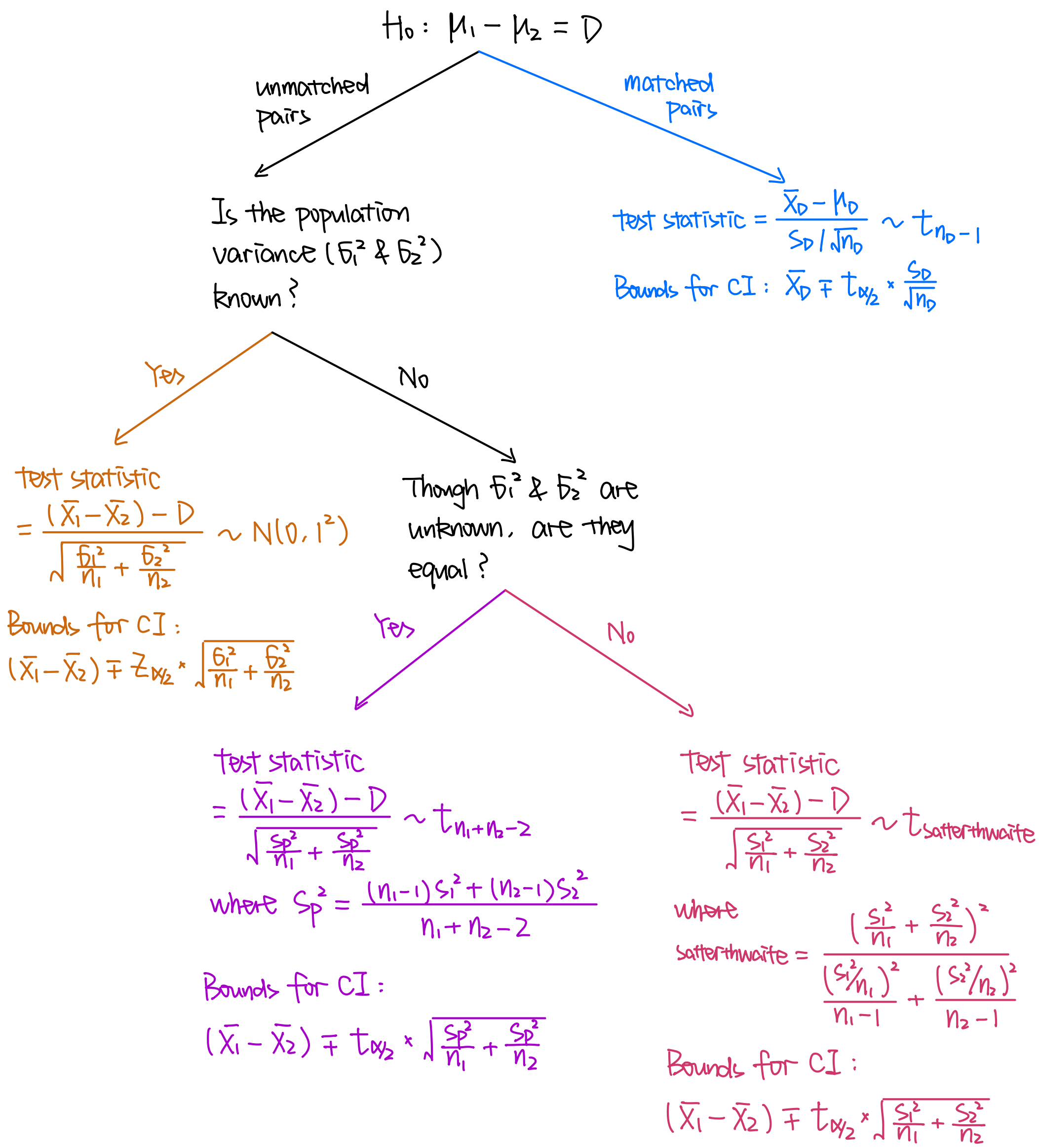
2 General Approach

- The general testing approach that we will take is very similar to what we talked about last discussion:
  – If using the rejection region & test statistic method:
    1. Find out what distribution the test statistic follows.
    2. Set up significance level & the appropriate sizes for the tail-ends of the distribution. Find cutoff values to construct rejection region.
    3. Calculate test statistic from the given sample, and see if it falls within the rejection region.
       * If test statistic falls within the rejection region, then we reject H_0 under the specified significance level.
       * If test statistic doesn't fall within the rejection region, then we fail to reject H_0 under the specified significance level.

- If using the confidence interval method:
  1. Find out what distribution the test statistic follows.
  2. Construct $(1-\alpha)$ level of confidence interval based on the confidence level $(1-\alpha)$ and the associated cutoffs from the distribution.
  3. Check if the hypothesized null falls within the $(1-\alpha)$ confidence interval:
     * If the hypothesized null falls within the $(1-\alpha)$ confidence interval, then we fail to reject H_0 at α significance level.
     * If the hypothesized null doesn't fall within the $(1-\alpha)$ confidence interval, then we reject H_0 at α significance level.

3 Inference about Two Populations

3.1 About the difference between two population means $(\mu_1 - \mu_2)$



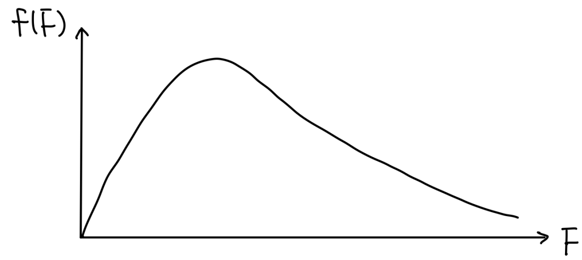
3.2 About two population variances ($\frac{\sigma_1^2}{\sigma_2^2}$)

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$\downarrow$$

$$\text{test statistic} = \frac{S_1^2}{S_2^2} \sim F_{\underbrace{n_1-1}_{\text{numerator DOF } (\nu_1)},\ \underbrace{n_2-1}_{\text{denominator DOF } (\nu_2)}}$$

- We are introduced to a new distribution: **F distribution**
  - F distribution usually looks like the following:

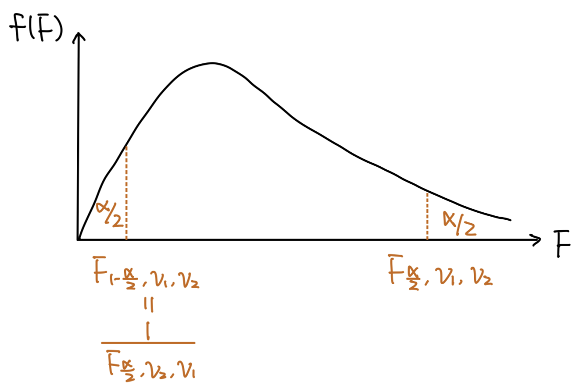


    It looks very similar to the Chi-squared distribution. (In fact, F distribution is actually obtained by dividing a Chi-squared distribution by a different Chi-squared distribution.)
  - A F distribution is denoted as the following:

$$F_{\nu_1,\nu_2}$$

    Here, ν_1 is the numerator degree of freedom, and ν_2 is the denominator degree of freedom.

- The F distribution table is usually given for a right tail area only – how does one find the appropriate left tail cutoff value, if we are performing a two-tail test?



Solution: the left tail cutoff value can be found as right tail cutoff value from a slightly different F distribution (switching the numerator and denominator degree of freedom)

$$F_{1-\frac{\alpha}{2},\nu_1,\nu_2} = \frac{1}{F_{\frac{\alpha}{2},\nu_2,\nu_1}}$$

3.3 About the difference between two success proportions ($p_1 - p_2$)

$H_0: p_1 - p_2 = D$

$\downarrow$

Does $\hat{p}_1 - \hat{p}_2$ follow a normal distribution?

i.e. Check if ALL of the following hold:

① $n_1 \hat{p}_1 \geq 5$

② $n_1(1-\hat{p}_1) \geq 5$

③ $n_2 \hat{p}_2 \geq 5$

④ $n_2(1-\hat{p}_2) \geq 5$

No (one/more fail(s))

Yes (ALL 4 hold)

Does $D = 0$?

Yes ($H_0: p_1 - p_2 = 0$)

test statistic

$$= \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{a}{\sim} N(0, 1^2)$$

where $\hat{p} = \dfrac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$

Bounds for CI:

$$(\hat{p}_1 - \hat{p}_2) \mp z_{\alpha/2} \times \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

No ($H_0: p_1 - p_2 = D$, where $D \neq 0$)

test statistic

$$= \frac{(\hat{p}_1 - \hat{p}_2) - D}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \overset{a}{\sim} N(0, 1^2)$$

Bounds for CI:

$$(\hat{p}_1 - \hat{p}_2) \mp z_{\alpha/2} \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$