MATH 1070 - Chapter 21 Notes

Two Sample Problems

Recall: So far, we have dealt with inference (confidence intervals and hypothesis testing) pertaining to:

- Single sample of data.
- A matched pairs design for which the analysis resembles that of a single sample.

Goal of Two Sample Inference is to compare the responses to two treatments or to compare the characteristics of two **independent** populations. The important aspects of this analysis are:

- Each sample is **independent** of each other. What does this imply?

- The units are not matched. So, the samples can be of differing sizes say n_1 and n_2.

Example (Exercise and Pulse Rates): A study is performed to compare the mean resting pulse rate of adult subjects who regularly exercise to the mean resting pulse rate of those who do not regularly exercise.

	n	Sample Mean	Standard Error
Exercisers	29	66	8.6
Non-Exercisers	31	75	9.0

Conditions for Comparing Two Means:

(1) We have **two independent SRSs**, from two distinct populations.

- One sample has no influence on the other (**Note:** Matching would violate independence.)
- We measure the same variable for both samples.

(2) Both populations are **Normally distributed.**

- The means (μ_1 and μ_2) and standard deviations (σ_1 and σ_2) of the populations are all unknown.

Two-Sample t Procedures:

What are we trying to do inference about?

What is the statistic that we are using to estimate this parameter?

What is the distribution of this statistic if $\bar{X}_1 \sim N(\mu_1, \sigma_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma_2)$

If we do not know σ_1 and σ_2, then they are each estimated by:

The standard deviation of $\bar{X}_1 - \bar{X}_2$ is estimated by:

It is referred to as the **standard error** of the difference in sample means, or the **estimated standard deviation** of the difference in the sample means.

1. Two-Sample t Confidence Interval

Under the following conditions:

(1) SRS of size n_1 from first population.
(2) A second SRS of size n_2 from second population.
(3) Both samples independent of each other.
(4) Both populations Normal with unknown means μ_1 and μ_2.

A **confidence interval** for $\mu_1 - \mu_2$ is given by:

where t^* is the critical value for confidence level C from the t density curve where the degrees of freedom are equal to the smaller of $n_1 - 1$ and $n_2 - 1$. **The critical values of the t distribution are given in Table C (page 701) of your text book.**

Example (Exercise and Pulse Rates): Find a 95% confidence interval for the difference in population means (non-exercisers minus exercisers) and interpret your answer. (Assume the two populations are Normally distributed and the two samples are independent SRSs from the respective populations.)

2. Two-Sample t Significance Tests

Under the conditions:

(1) SRS of size n_1 from first population.
(2) A second SRS of size n_2 from second population.
(3) Both samples independent of each other.
(4) Both populations Normal with unknown means μ_1 and μ_2.

The **test statistic** to test the hypothesis: $H_0 : \mu_1 = \mu_2$ is given by:

The p-values associated with the test statistic are calculated in the usual manner (see page 4 of notes from Chapter 20) from the T distribution corresponding to degrees of freedom equal to the minimum of $n_1 - 1$ and $n_2 - 1$.

Example (Exercise and Pulse Rates): Is the mean resting pulse rate of adult subjects who regularly exercise different from the mean resting pulse rate of those who do not regularly exercise? Test this hypothesis using a significance level of 0.01.

3. Robustness of t procedures for two sample procedures.

Thumb rules for using the t procedure:

(1) Except in the case of small samples, the assumption that each sample is an independent SRS from the population of interest is more important than the assumption that the two population distributions are Normal.

(2) Small sample sizes ($n_1 + n_2 < 15$): Use t procedures if each data set appears close to Normal (symmetric, single peak, no outliers). If a data set is skewed or if outliers are present, do not use t.

(3) Medium sample sizes ($n_1 + n_2 \geq 15$): The t procedures can be used except in the presence of outliers or strong skewness in a data set.

(4) Large samples: The t procedures can be used even for clearly skewed distributions when the sample sizes are large, roughly $n_1 + n_2 \geq 40$.

Details of t Degrees of Freedom:

(1) Using degrees of freedom as the smallest of $n_1 - 1$ and $n_2 - 1$ is only a rough approximation to the actual degrees of freedom for the two-sample t procedures.

(2) A better approximation that is used by software uses a function of the sample sizes and sample standard deviations to compute degrees of freedom. Let us refer to this as *df* for now.

(3) Use of degrees of freedom calculated via the software (*df*) gives more accurate results than when simply using the smaller of $n_1 - 1$ and $n_2 - 1$.

The formula for *df* is given by:

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

This formula yields an approximation to the true t distribution in question while dealing with a two sample t test.

The approximation is accurate when both n_1 and n_2 are 5 or larger.

Example (Exercise and Pulse Rates): Compute the degrees of freedom *df* used by software to analyze these data using two-sample t procedures.

Chapter 21 Objectives:

- Motivation for two sample t test.
- Population parameter of interest.
- Two sample t confidence interval.
- Two sample t hypothesis test.
- Calculation of df.