0604 Slides

MA 116

June 2025

Review: 0603 lecture

1. Set up background: Inference about two population proportions
2. Distribution of the difference between two proportions
    1. We want to gain information about p_1 vs. p_2, two population proportion from possibly different populations with possibly different characteristics.
3. Testing hypothesis regarding two population proportions
    1. New variable $\hat{p}_1 - \hat{p}_2$–does it have an approximately normal distribution?
    2. H_0 is always $p_1 = p_2$, H_1 is $p_1 >, <, \text{ or } = p_2$.
    3. Example of scenario: Investigate whether the percentage of Boston residents who drink coffee every day is about the same as the percentage of Boston residents who like coffee. Population is the same, so be careful that when obtaining the two sample data sets, both must be randomly obtained, and obtaining one data set must not depend on how we obtain the other sample data set.
4. Interval estimator for the difference between two population proportions

1. Determine H_0, H_1, test type.
2. Assume H_0 is true. Check if the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal. ($n_i < 0.05N_i$; $n_i\hat{p}_i(1-\hat{p}_i) \geq 10$ for $i = 1$ and 2.)
3. If normal, change to a standard normal variable z using the approximation formula of $\sigma_{\hat{p}_1-\hat{p}_2}$ (Keep assuming H_0 is true so that the $\sigma_{\hat{p}_1-\hat{p}_2}$ formula is

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \text{ where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

4. Determine the critical value(s) and the critical region on the standard normal curve diagram. (Keep assuming H_0 is true.)
5. Calculate the test statistic z_0 by plugging in $\hat{p}_1$ and $\hat{p}_2$ values of our particular sample into z formula. If z_0 falls into the critical region, reject H_0.

Interval estimator for the difference between two population proportions

If $\hat{p}_1$ and $\hat{p}_2$ are checked to have approximately normal distributions, then we may use the following formula

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

We say $(\hat{\text{p}}_1\text{-}\hat{\text{p}}_2) \pm E$ is an interval estimator of $(p_1 - p_2)$ to a confidence level of $(1-\alpha)100\%$.

06/03 last slide gives a wrong formula for E-missing the $z_{\alpha/2}$ factor.

Outline of today's new content

1. Two sample data sets that are dependent, matched-pairs
2. Inference about two means: matched-pairs design
   - distribution of μ_d
   - test hypothesis
3. Inference about two means: independent samples
   - population mean difference $\mu_1 - \mu_2$, distribution of variable $\overline{x_1} - \overline{x_2}$.
   - test hypothesis
   - Confidence interval/interval estimator

Independent data sets vs. Dependent data sets

Example: independent sample data sets, proportion

Investigate whether the proportion of Boston residents who drink coffee every day is approximately equal to the proportion who say they like coffee. Although both groups come from the same population, the two samples must be independently and randomly drawn. Be careful to ensure that selecting one group does not influence the selection of the other.

Example: independent sample data sets, mean

Investigate whether the average number of cups of coffee Boston residents consumed a day per person is about the same as the average number of cups of coffee New York residents consumed a day per person.

Example: dependent (matched-pair) sample data sets, mean

Investigate whether, on average, a Boston resident consumes more cups of coffee than cups of milk per day.

Matched-pair data

Require population to be quantitative.

Definition. (population mean difference μ_d)

This is a population parameter associated to the mean of difference. (Not the difference of the two means!)

In this example, we investigate whether, on average, a Boston resident consumes more cups of coffee than cups of milk per day. Our population parameter μ_d is **the population mean of (# of cups of coffee a Boston resident consumes per day-# of cups of milk a Boston resident consumes per day)**. μ_d is NOT defined to be μ_x(the population mean of # of cups of coffee a Boston resident consumes per day)-μ_y(the population mean of # of cups of milk a Boston resident consumes per day)! Even though, in this particular scenario we do have

$$\mu_d = \mu_x - \mu_y.$$

population parameter vs. sample statistic

When doing inference about two means in a matched-pair setting we only use population parameter μ_d, not μ_x or μ_y. We analyze the distribution of a new variable d, instead of x or y.

The sample statistic associated to this population parameter μ_d is $\overline{d}$.

$$\overline{d} = \frac{\sum_i (x_i - y_i)}{n} = \overline{x} - \overline{y}$$

We can talk about distributions of d and (wrt some fixed n) $\overline{d}$.

Sampling distribution of $\overline{d}$

Note. s_d is a sample statistic.

We say $t = \dfrac{\overline{d} - \mu_d}{s_d/\sqrt{n}}$ can be approximated by Student's t-distribution of $df = n - 1$ if either

1. d is approximately normally distributed, for example, when both x and y are normally distributed.
2. $n > 30$, so the distribution of t is approximately **standard** normal.

Hypothesis test of two means: matched-pairs design

$$H_0 : \mu_d = 0,\ H_1 :\ \mu_d >, <, \text{ or } \neq 0.$$

Assume H_0 is true, we may verify whether $\overline{d}$ follows Student's t-distribution. If it does, we may determine the critical value(s) and the critical region on the Student's t-distribution graph with the appropriate df. Finally we calculate the test statistic

$$t_0 = \frac{\overline{d} - \mu_d}{s_d/\sqrt{n}} = \frac{\overline{d} - 0}{s_d/\sqrt{n}}.$$

If the test statistic lies in the critical region, reject H_0.

example.

We want to construct an interval estimator for d. If the variable $t = \dfrac{\overline{d} - \mu_d}{s_d/\sqrt{n}}$ follows Student's t-distribution (there are 2 situations in which this happend), then we can use the following formula.

$$E = t_{\alpha/2} \frac{s_d}{\sqrt{n}}.$$