MATH 205: Statistical methods

Lecture 33: Comparing the mean of two populations (cont.)

Announcements

- Homework 5 due on 12/02
- Wednesday 11/30: Quiz 5 (Hypothesis testing, one sample)
- Final exam:

  12/15/2022, Thursday
  3:30PM - 5:30PM
  Gore Hall Room 304

- Plan for the rest of the semester
  - Linear regression
  - Review of materials + Practice exam
- **There will be no lab next week**
- Course evaluations will be available 12/01 through 12/08

Chapter 7: Significance of evidence

Settings

Assumption

1. *$X_1, X_2, \ldots, X_m$ is a random sample from a population with mean μ_1 and variance σ_1^2.*
2. *$Y_1, Y_2, \ldots, Y_n$ is a random sample from a population with mean μ_2 and variance σ_2^2.*
3. *The X and Y samples are independent of each other.*

Proposition

$$E[\bar{X} - \bar{Y}] = \mu_1 - \mu_2$$

and

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Confidence intervals

When both population distributions are normal, standardizing $\overline{X} - \overline{Y}$ gives a random variable Z with a standard normal distribution. Since the area under the z curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$, it follows that

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

Manipulation of the inequalities inside the parentheses to isolate $\mu_1 - \mu_2$ yields the equivalent probability statement

$$P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right) = 1 - \alpha$$

Testing the difference between two population means

- Setting: independent normal random samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ with known values of σ_1 and σ_2. Constant Δ_0.
- Null hypothesis:

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

- Alternative hypothesis:
  (a) $H_a : \mu_1 - \mu_2 > \Delta_0$
  (b) $H_a : \mu_1 - \mu_2 < \Delta_0$
  (c) $H_a : \mu_1 - \mu_2 \neq \Delta_0$
- When $\Delta = 0$, the test (c) becomes

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

Testing the difference between two population means

Assume that we want to test the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ against each of the following alternative hypothesis

(a) $H_a : \mu_1 - \mu_2 > \Delta_0$

(b) $H_a : \mu_1 - \mu_2 < \Delta_0$

(c) $H_a : \mu_1 - \mu_2 \neq \Delta_0$

We use the test statistic:

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}.$$

and derive the p-value in the same way as the one-sample tests.

Practice problem

Each student in a class of 21 responded to a questionnaire that requested their GPA and the number of hours each week that they studied. For those who studied less than 10 h/week the GPAs were

$$2.80, 3.40, 4.00, 3.60, 2.00, 3.00, 3.47, 2.80, 2.60, 2.00$$

and for those who studied at least 10 h/week the GPAs were

$$3.00, 3.00, 2.20, 2.40, 4.00, 2.96, 3.41, 3.27, 3.80, 3.10, 2.50$$

Assume that the distribution of GPA for each group is normal and both distributions have standard deviation $\sigma_1 = \sigma_2 = 0.6$. Treating the two samples as random, is there evidence that true average GPA differs for the two study times? Carry out a test of significance at level .05.

Large-sample tests/confidence intervals

- Central Limit Theorem: $\bar{X}$ and $\bar{Y}$ are approximately normal when $n > 30 \to$ so is $\bar{X} - \bar{Y}$. Thus

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

  is approximately standard normal
- When n is sufficiently large $S_1 \approx \sigma_1$ and $S_2 \approx \sigma_2$
- Conclusion:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

  is approximately standard normal when n is sufficiently large

If $m, n > 40$, we can ignore the normal assumption and replace σ by S

Large-sample tests

We use the test statistic:

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}.$$

and derive the p-value in the same way as the one-sample tests.

Large-sample CIs

Proposition

Provided that m and n are both large, a CI for $\mu_1 - \mu_2$ *with a confidence level of approximately* $100(1-\alpha)\%$ *is*

$$\bar{x} - \bar{y} \ \pm \ z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where $-$*gives the lower limit and* $+$ *the upper limit of the interval. An upper or lower confidence bound can also be calculated by retaining the appropriate sign and replacing* $z_{\alpha/2}$ *by* z_α.

Example

Example

Let μ_1 and μ_2 denote true average tread lives for two competing brands of size P205/65R15 radial tires.

(a) Test

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

at level 0.05 using the following data: $m = 45$, $\bar{x} = 42,500$, $s_1 = 2200$, $n = 45$, $\bar{y} = 40,400$, and $s_2 = 1900$.

(b) Construct a 95% CI for $\mu_1 - \mu_2$.