Math 170S
Lecture Notes Section 8.2 *†
Tests about two means

Instructor: Swee Hong Chan

NOTE: Materials that appear in the textbook but do not appear in the lecture notes might still be tested.
Please send me an email if you find typos.

*Version date: Wednesday 18\textsuperscript{th} November, 2020, 13:23.
†This notes is based on Hanbaek Lyu’s and Liza Rebrova’s notes from the previous quarter, and I would like to thank them for their generosity. “Nanos gigantum humeris insidentes (I am but a dwarf standing on the shoulders of giants)”.
1 Setting: dependent $X$ and $Y$

Object: $X$ and $Y$ are (possibly dependent) random variables with unknown mean $\mu_X$ and $\mu_Y$, and $D$ is the difference $D := X - Y$.

Hypotheses:

- **Null Hypothesis** $H_0$: $\mu_X$ is equal to $\mu_Y$. Equivalently $\mu_D = 0$.

- **Alternative Hypothesis** $H_1$: The alternative hypothesis is one of these three forms:
  
  (a) $\mu_X$ is strictly greater than $\mu_Y$. Equivalently $\mu_D > 0$;

  (b) $\mu_X$ is strictly smaller than $\mu_Y$. Equivalently $\mu_D < 0$;

  (c) $\mu_X$ is not equal to $\mu_Y$. Equivalently $\mu_D \neq 0$. 

The strategy is to apply tests for one mean from Section 8.1 to $D$.

**Input:** Random samples $X_1, \ldots, X_n$ for $X$, random samples $Y_1, \ldots, Y_n$ for $Y$, and significance level $\alpha$.

**Methodology:**

- Compute $\overline{D} := \frac{(X_1 - Y_1) + \ldots + (X_n - Y_n)}{n}$
- Compute the critical region that depends on $\alpha$; or
- Compute the $p$-value that depends on $\overline{D}$.

**Output:**

- Reject the hypothesis if $\overline{D}$ is contained in the critical region. Equivalently, reject the hypothesis if the $p$-value is smaller than $\alpha$.
- Do not reject the hypothesis (i.e., test is inconclusive) otherwise.
2 Example: dependent $X$, $Y$

Twenty-four students were subjected to a brainwashing program to increase their midterm scores.

- The null hypothesis is that the brainwashing program does nothing to their midterm score;

- The alternative hypothesis is that the brainwashing program increases their midterm score.

Let $X$ be the midterm score before the program, and $Y$ be the midterm score after the program.

Let $D = X - Y$ be a normal random variable with unknown mean and unknown variance.

Suppose that sample mean $\bar{D}$ is $-0.079$ and sample standard deviation $s_D$ is $0.255$.

Should we reject the null hypothesis with $\alpha = 0.05$?
3 Answer: dependent $X$ and $Y$

This is the scenario of **normal** random variable with **unknown mean** and **unknown variance**.

From Section 8.1,

$$t_{\alpha}(n - 1)\frac{s}{\sqrt{n}} = t_{0.05}(23)\frac{0.255}{\sqrt{24}} = (1.714)(\frac{0.255}{\sqrt{24}}) \approx 0.09.$$

So the critical region is

$$\left(-\infty, 0 - t_{\alpha}(n - 1)\frac{s}{\sqrt{n}}\right] = \left(-\infty, -0.09\right].$$

Since the sample mean $\bar{D} = -0.079$ is not contained in the critical region, the test is inconclusive.
4 Setting: independent $X$ and $Y$ with known variances

Object: $X$ and $Y$ are independent random variables with unknown mean $\mu_X$ and $\mu_Y$ but with known variances $\sigma^2_X$ and $\sigma^2_Y$.

Hypotheses:

- Null Hypothesis $H_0$: $\mu_X$ is equal to $\mu_Y$.

- Alternative Hypothesis $H_1$: It takes one of these three forms:
  
  (a) $\mu_X$ is strictly greater than $\mu_Y$;

  (b) $\mu_X$ is strictly smaller than $\mu_Y$;

  (c) $\mu_X$ is not equal to $\mu_Y$.  

**Input:** Significance level $\alpha$, random samples $X_1, \ldots, X_n$ for $X$, random samples $Y_1, \ldots, Y_m$ for $Y$. Note that $n$ is not necessarily equal to $m$.

**Methodology:**

- Compute the critical region that depends on $\alpha$; or
- Compute the $p$-value that depends on $\bar{X}$ and $\bar{Y}$.

**Output:**

- Reject the hypothesis if $\bar{X} - \bar{Y}$ is contained in the critical region. Equivalently, reject the hypothesis if the $p$-value is smaller than $\alpha$.
- Do not reject the hypothesis (i.e., test is inconclusive) otherwise.
5 Theorem: independent $X$ and $Y$ with known variances

Theorem 1.  

• For the case $\mu_X > \mu_Y$,

\[
\text{critical region} = \left[ z_{\alpha} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, \infty \right),
\]

\[
p-value = 1 - \Phi \left( \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \right).
\]

• For the case $\mu_X < \mu_Y$,

\[
\text{critical region} = \left( -\infty, -z_{\alpha} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right),
\]

\[
p-value = \Phi \left( \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \right).
\]
• For the case $\mu_X \neq \mu_Y$,

\[
\text{critical region} = \left( -\infty , -z_{\alpha/2} \sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}} \right) \cup \left[ z_{\alpha/2} \sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}}, \infty \right),
\]

\[
p-value = 2 \left[ 1 - \Phi \left( \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}}} \right) \right].
\]
6 Example: independent $X, Y$ with known variances

Let $X$ be the midterm grade of a random student from Section 1, and let $Y$ be the midterm grade of a random student from Section 2.

- The null hypothesis is that $\mu_X = \mu_Y$;

- The alternative hypothesis is that $\mu_X < \mu_Y$.

Let $X$ be a normal random variable with standard deviation $\sigma_X = 1.08$, and let $Y$ be a normal random variable with standard deviation $\sigma_Y = 1.55$.

Suppose that $X$ has sample mean $\bar{X} = 67.01$ with $n = 50$, and $Y$ has sample mean $68.41$ with $m = 40$ students. Should we reject the null hypothesis at an $\alpha = 0.01$ significance level?
7 Answer: independent $X, Y$ with known variances

This is case (b), so we have

$$z_\alpha \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} = z_{0.01} \sqrt{\frac{(1.08)^2}{50} + \frac{(1.55)^2}{40}} = (2.326) \sqrt{\frac{(1.08)^2}{50} + \frac{(1.55)^2}{40}} \approx 0.672$$

So the critical region is

$$\left( -\infty, 0 - z_\alpha \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right] = \left( -\infty, -0.672 \right].$$

Since $\overline{X} - \overline{Y} = 67.01 - 68.41 = -1.4$ is contained in the critical region, we reject the null hypothesis.
The case of independent $X$ and $Y$ with unknown variances

Object: $X$ and $Y$ are independent random variables with unknown mean $\mu_X$ and $\mu_Y$ and unknown but equal variances $\sigma^2$.

Hypotheses:

- **Null Hypothesis** $H_0$: $\mu_X$ is equal to $\mu_Y$.

- **Alternative Hypothesis** $H_1$: The alternative hypothesis can take one of these three forms:
  
  (a) $\mu_X$ is strictly greater than $\mu_Y$;
  
  (b) $\mu_X$ is strictly smaller than $\mu_Y$;
  
  (c) $\mu_X$ is not equal to $\mu_Y$. 

**Input:** Significance level $\alpha$, random samples $X_1, \ldots, X_n$ for $X$, random samples $Y_1, \ldots, Y_m$ for $Y$. Note that $n$ is not necessarily equal to $m$.

**Methodology:**

- Compute the sample variance $s^2_X$ and $s^2_Y$.
- Compute the pulled estimator
  $$s_P := \sqrt{\frac{(n - 1)s^2_X + (m - 1)s^2_Y}{n + m - 2}}.$$
- Compute the critical region that depends on $\alpha$.

**Output:**

- Reject the hypothesis if $\bar{X} - \bar{Y}$ is contained in the critical region. Equivalently, reject the hypothesis if the $p$-value is smaller than $\alpha$.
- Do not reject the hypothesis otherwise.
Theorem 2. If \( n + m - 2 > 30 \), then use the following formula:

- For the case \( \mu_X > \mu_Y \),

\[
\text{critical region} = \left[ z_\alpha \, s_P \sqrt{\frac{1}{n} + \frac{1}{m}} , \, \infty \right),
\]

- For the case \( \mu_X < \mu_Y \),

\[
\text{critical region} = \left( -\infty , \, -z_\alpha \, s_P \sqrt{\frac{1}{n} + \frac{1}{m}} \right],
\]

- For the case \( \mu_X \neq \mu_Y \),

\[
\text{critical region} = \left( -\infty , \, -z_{\alpha/2} \, s_P \sqrt{\frac{1}{n} + \frac{1}{m}} \right] \cup \left[ z_{\alpha/2} \, s_P \sqrt{\frac{1}{n} + \frac{1}{m}} , \, \infty \right),
\]
If \( n + m - 2 \leq 30 \), use the following formula:

- **For the case** \( \mu_X > \mu_Y \),

  \[
  \text{critical region} = \left[ t_\alpha(n + m - 2)s_P \sqrt{\frac{1}{n} + \frac{1}{m}}, \infty \right).
  \]

- **For the case** \( \mu_X < \mu_Y \),

  \[
  \text{critical region} = \left( -\infty, -t_\alpha(n + m - 2)s_P \sqrt{\frac{1}{n} + \frac{1}{m}} \right].
  \]

- **For the case** \( \mu_X \neq \mu_Y \),

  \[
  \text{critical region} = \left( -\infty, -t_{\alpha/2}(n + m - 2)s_P \sqrt{\frac{1}{n} + \frac{1}{m}} \right]\]
  \[
  \left[ t_{\alpha/2}(n + m - 2)s_P \sqrt{\frac{1}{n} + \frac{1}{m}}, \infty \right).
  \]
9 Example: independent $X$ and $Y$ with unknown variances

Let $X$ be the net worth of a random citizen in Atlantis, and let $Y$ be the net worth of a random citizen in Shangrila.

- The null hypothesis is that $\mu_X = \mu_Y$;
- The alternative hypothesis is that $\mu_X \neq \mu_Y$.

Suppose that $X$ and $Y$ are two **normal** random variable with **same unknown variance**. Suppose that

- $X$ has sample mean $\overline{X} = 1076.75$, sample variance $s_X^2 = 29.30$ with $n = 12$ citizens;

- $Y$ has sample mean $\overline{Y} = 1072.33$, sample variance $s_Y^2 = 26.24$ with $m = 12$ citizens.

Should we reject the null hypothesis at an $\alpha = 0.1$ significance level?
10 Answer: independent $X$ and $Y$ with unknown variances

This is case (c), so we have

$$s_P = \sqrt{\frac{(n - 1)s_X^2 + (m - 1)s_Y^2}{n + m - 2}} = \sqrt{\frac{(12 - 1)(29.30) + (12 - 1)(26.24)}{12 + 12 - 2}} = 5.267,$$

which gives us

$$t_{\alpha/2}(n + m - 2)s_P \sqrt{\frac{1}{n} + \frac{1}{m}} = (1.717)(5.267)\sqrt{\frac{1}{12} + \frac{1}{12}} = 3.69.$$

So the critical region is

$$\left( -\infty, -3.69 \right] \cup \left[ 3.69, \infty \right)$$

Since $\bar{X} - \bar{Y} = 1076.75 - 1072.33 = 4.42$ is contained in the critical region, we reject the null hypothesis.