Chapter 18

Two-Sample Problems
Case Study

Exercise and Pulse Rates

A study 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.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>mean</th>
<th>std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exercisers</td>
<td>29</td>
<td>66</td>
<td>8.6</td>
</tr>
<tr>
<td>Nonexercisers</td>
<td>31</td>
<td>75</td>
<td>9.0</td>
</tr>
</tbody>
</table>

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?
Two-Sample Problems

◆ The goal of inference is to compare the responses to two treatments or to compare the characteristics of two populations.

◆ We have a separate sample from each treatment or each population.
  – Each sample is separate. The units are not matched, and the samples can be of differing sizes.
Conditions for Comparing Two Means

- We have **two independent SRSs**, from two distinct populations
  - that is, one sample has no influence on the other—matching violates independence
  - we measure the same variable for both samples.
- Both populations are **Normally distributed**
  - the means and standard deviations of the populations are unknown
  - in practice, it is enough that the distributions have similar shapes and that the data have no strong outliers.
Two-Sample $t$ Procedures

- In order to perform inference on the difference of two means ($\mu_1 - \mu_2$), we’ll need the standard deviation of the observed difference $\bar{X}_1 - \bar{X}_2$:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
Two-Sample $t$ Procedures

- **Problem:** We don’t know the population standard deviations $\sigma_1$ and $\sigma_2$.

- **Solution:** Estimate them with $s_1$ and $s_2$. The result is called the standard error, or estimated standard deviation, of the difference in the sample means.

$$ SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} $$
Two-Sample $t$ Confidence Interval

- Draw an SRS of size $n_1$ form a Normal population with unknown mean $\mu_1$, and draw an independent SRS of size $n_2$ form another Normal population with unknown mean $\mu_2$.

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

$$\left( \bar{x}_1 - \bar{x}_2 \right) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  - here $t^*$ is the critical value for confidence level $C$ for the $t$ density curve. The **degrees of freedom** are equal to the smaller of $n_1 - 1$ and $n_2 - 1$. 


Case Study

Exercise and Pulse Rates

Find a 95% confidence interval for the difference in population means (nonexercisers minus exercisers).

\[
\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 75 - 66 \pm 2.048 \sqrt{\frac{(9.0)^2}{31} + \frac{(8.6)^2}{29}}
\]

\[
= 9 \pm 4.65
\]

\[
= 4.35 \text{ to } 13.65
\]

“We are 95% confident that the difference in mean resting pulse rates (nonexercisers minus exercisers) is between 4.35 and 13.65 beats per minute.”
Two-Sample $t$ Significance Tests

- Draw an SRS of size $n_1$ from a Normal population with unknown mean $\mu_1$, and draw an independent SRS of size $n_2$ from another Normal population with unknown mean $\mu_2$.

- To **test the hypothesis** $H_0: \mu_1 = \mu_2$, the test statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Use $P$-values for the $t$ density curve. The **degrees of freedom** are equal to the smaller of $n_1 - 1$ and $n_2 - 1$. 
**P-value for Testing Two Means**

- **Hₐ**: $\mu_1 > \mu_2$
  - $P$-value is the probability of getting a value as large or larger than the observed test statistic ($t$) value.

- **Hₐ**: $\mu_1 < \mu_2$
  - $P$-value is the probability of getting a value as small or smaller than the observed test statistic ($t$) value.

- **Hₐ**: $\mu_1 \neq \mu_2$
  - $P$-value is *two times* the probability of getting a value as large or larger than the absolute value of the observed test statistic ($t$) value.
Case Study

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?

- **Null**: The mean resting pulse rate of adult subjects who regularly exercise is the *same* as the mean resting pulse rate of those who do not regularly exercise? \( H_0 : \mu_1 = \mu_2 \)

- **Alt**: The mean resting pulse rate of adult subjects who regularly exercise is *different* from the mean resting pulse rate of those who do not regularly exercise? \( H_a : \mu_1 \neq \mu_2 \)

Degrees of freedom = 28 (smaller of 31 – 1 and 29 – 1).
1. **Hypotheses:**
   \[ H_0: \mu_1 = \mu_2 \quad \text{H}_a: \mu_1 \neq \mu_2 \]

2. **Test Statistic:**
   \[
   t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{75 - 66}{\sqrt{\frac{(9.0)^2}{31} + \frac{(8.6)^2}{29}}} \approx 3.961
   \]

3. **P-value:**
   \[ P\text{-value} = 2P(T > 3.961) = 0.000207 \quad \text{(using a computer)} \]
   P-value is smaller than 2(0.0005) = 0.0010 since \( t = 3.961 \) is greater than \( t^* = 3.674 \) (upper tail area = 0.0005)  
   \text{(Table C)}

4. **Conclusion:**
   Since the P-value is smaller than \( \alpha = 0.001 \), there is very strong evidence that the mean resting pulse rates are different for the two populations (nonexercisers and exercisers).
Robustness of $t$ Procedures

- The two-sample $t$ procedures are more robust than the one-sample $t$ methods, particularly when the distributions are not symmetric.

- When the two populations have similar distribution shapes, the probability values from the $t$ table are quite accurate, even when the sample sizes are as small as $n_1 = n_2 = 5$.

- When the two populations have different distribution shapes, larger samples are needed.

- In planning a two-sample study, it is best to choose equal sample sizes. In this case, the probability values are most accurate.
Using the $t$ Procedures

- 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.

- **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 $t$.

- **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.

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