Two-Sample Problem

Two independent populations are to be compared for a characteristic, either a quantitative measurement or a categorical one. A random sample is taken from each population to study. The samples are independent. In a controlled study, the subjects are selected with uniformity with respect to known source(s) of variation (similar weight, similar health condition, etc).
Examples

1. Is there “grade inflation” in WMU? How does the average GPA of WMU students today compare with 10 years ago? Suppose a random sample of 100 student records from 10 years ago yields a sample average GPA of 2.90 with a standard deviation of .40. A random sample of 100 current students today yields a sample average of 2.98 with a standard deviation of .45. The difference between the two sample means is $2.98 - 2.90 = .08$. Is this proof that GPA’s are higher today than 10 years ago? Or the chance variability is at work?

2. How does the reduction of BMI of the Atkins diet compare to that of Zone diet at 12 months?

Notation

- Population 1 has mean $\mu_1$ and standard deviation $\sigma_1$ (usually unknown).
- Population 2 has mean $\mu_2$ and standard deviation $\sigma_2$ (usually unknown).
- Sample of size $n_1$ is taken from population 1: the sample mean is $\bar{X}_1$, the sample standard deviation is $SD_1$, and the standard error is $SE_1 = SD_1 / \sqrt{n_1}$.
- Sample of size $n_2$ is taken from population 2: the sample mean is $\bar{X}_2$, the sample standard deviation is $SD_2$, and the standard error is $SE_2 = SD_2 / \sqrt{n_2}$.
- The two samples are independent.
Estimating the Difference in Means 
of two independent populations: $\mu_1 - \mu_2$

- **point estimate**: $\overline{X}_1 - \overline{X}_2$
- **standard error**:

$$SE = \sqrt{[SE(\overline{X}_1)]^2 + [SE(\overline{X}_2)]^2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

- **95% confidence interval for $\mu_1 - \mu_2$**:

$$(\overline{X}_1 - \overline{X}_2) \pm 1.96 \times \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

**Statistical Significance**

- If the 95% confidence interval for $\mu_1 - \mu_2$ **excludes zero**, then we say that the difference is **statistically significant** or that the mean for one group differs significantly than that for the other group.

- If the 95% confidence interval for $\mu_1 - \mu_2$ **includes zero**, then we say that the difference is **insignificant** or that there is insignificant difference between the two population means.
Difference in GPA Averages Example
Let the GPAs today be of group 1 and that of 10 years ago be of group 2.

- \( \bar{X}_1 - \bar{X}_2 = 2.98 - 2.90 = 0.08 \).
- Standard error:
  \[
  SE = \sqrt{\frac{(.45)^2}{100} + \frac{(.40)^2}{100}} = 0.06
  \]
- Margin of error:
  \[
  ME = 1.96 \times SE = 1.96 \times .06 = 0.118
  \]
- 95% c.i. for \( \mu_1 - \mu_2 \):
  \[
  (0.08 - 0.118, 0.08 + 0.118) = (-0.038, 1.198)
  \]
- The interval includes zero and hence, the difference is insignificant. Simple chance variability can be a viable explanation for the observed difference.

Diet Comparison Example
Consider Atkins and Zone diets at 12 months. Denote \( \mu_1 \) the mean change in BMI for Zone diet group and \( \mu_2 \) the mean change in BMI for Atkins diet group.

- \( \bar{X}_1 - \bar{X}_2 = (-.53) - (-1.65) = 1.12 \).
- Standard error:
  \[
  SE = \sqrt{\frac{(2.00)^2}{79} + \frac{(2.54)^2}{77}} = 0.367
  \]
- Margin of error:
  \[
  ME = 1.96 \times 0.367 = 0.72
  \]
- 95% c.i. for \( \mu_1 - \mu_2 \):
  \[
  (1.12 - 0.72, 1.12 + 0.72) = (0.40, 1.84)
  \]
which excludes zero and hence the difference is statistically significant.
iClicker Question 19.1

A 95% confidence interval for the difference in the means of a numerical measurement on two independent populations was calculated from two independent samples. The result is (1.2, 10.5). Which of the following is true?

A. the confidence interval excludes 0 hence the difference is insignificant
B. the confidence interval includes 0 hence the difference is insignificant
C. the confidence interval includes 0 hence the difference is significant
D. the confidence interval excludes 0 hence the difference is significant
E. none of the previous

iClicker Question 19.2

A 95% confidence interval for the difference in the means of a numerical measurement on two independent populations was calculated from two independent samples. The result is (1.2, 10.5). Which of the following is a correct interpretation of the confidence interval?

A. we are 95% confident that the difference in sample means is between 1.2 and 10.5
B. we are 95% confident that the difference in the true means is between 1.2 and 10.5
C. there is 95% chance that the difference in the true means is between 1.2 and 10.5
D. there is 95% chance that the difference in sample means is between 1.2 and 10.5
E. none of the previous
iClicker Question 19.3

What effect there is on the standard error of the difference in means if the sample sizes are each quadrupled?

A. the standard error is likely to decrease
B. the standard error is likely to increase
C. there is no effect

Diet Comparison Example, continued

Is it viable to explain that the difference in changes in BMI of 1.12 is due to chance variability?

Note that 1.12 is $1.12 / 0.367 = 3.05$SE. If the true difference were 0, the chance to observe a value as far as $\pm 3.05$ or more SE’s is

$$P(Z < -3.05 \text{ or } Z > 3.05) = 2P(Z > 3.05)$$
$$= 2[1 - P(Z \leq 3.05)] = 0.0022$$

a very small chance making it hard to believe that the true difference is 0. Hence, we conclude that statistically, the two means are different. Or we can say that the means are significantly different.
Computing the P-value for the Difference in means of two independent populations

\[ P\text{-value} = 2 \times \left[ 1 - P\left( Z \leq \frac{|\bar{X}_1 - \bar{X}_2|}{SE} \right) \right] \]