IX. Two-Sample Hypothesis Testing

A. A Few Important Definitions

1. **Independent Simple Random Samples – two (or more) simple random samples taken in such a way that the observed values in one sample are unrelated to the observed values in the other sample(s).**
2. **Matched (Dependent) Simple Random Samples – two (or more) simple random samples for which for every element in one sample corresponds to a unique identical element in the other sample(s).**

The sampling approach taken will determine, to a large extent, how one can compare two populations.

B. Differences between Means of Two Populations – Independent Samples

Let μ_1 be the mean of population 1 and μ_2 be the mean of population 2

We can write the difference between the two population means as $\mu_1 - \mu_2$.

To make a point estimate of the difference between the two population means $\mu_1 - \mu_2$ we take independent samples of sizes n_1 and n_2 from the two populations and calculate

$$\bar{x}_1 - \bar{x}_2 = \frac{\sum_{i=1}^{n_1} x_{i,1}}{n_1} - \frac{\sum_{i=1}^{n_2} x_{i,2}}{n_2}$$

The point estimate $\bar{x}_1 - \bar{x}_2$ has a sampling distribution with the following characteristics:

$$\text{Expected Value}: E\left[\bar{x}_1 - \bar{x}_2\right] = \mu_1 - \mu_2$$

$$\text{Standard Deviation}: \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where

σ_1 is the standard deviation of population 1
σ_2 is the standard deviation of population 2

If both samples are large ($n_1 \geq 30$ *and* $n_2 \geq 30$) or a small sample is drawn from a normal parent population, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is *approximately* normal.

This sampling distribution can be used to build confidence intervals and test hypotheses for the difference between the two population means μ_1 - μ_2 :

1. **Interval Estimate of the difference between the two population means μ_1 - μ_2 – Large Sample Case ($n_1 \geq 30$ *and* $n_2 \geq 30$) or the populations are normal, and we know the population standard deviations σ_1 and σ_2:**

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2}$$

where 1 - α is the confidence coefficient.

Note we rarely know the population standard deviations σ_1 and σ_2, and so consequently we rarely know the standard deviation of the difference between the two population means $\sigma_{\bar{x}_1-\bar{x}_2}$.

Example: Suppose we wish to assess the difference in incomes earned by Whatsamatta University accounting and marketing majors ten years after they receive their bachelors degrees at the 95% level of confidence. From historical records, the Whatsamatta U. Alumni Relations Department knows the standard deviation of incomes earned ten years after graduation for their alumni with undergraduate degrees in accounting and marketing are 15,000 and 20,000, respectively. The Alumni Relations Department collects independent random samples of accounting and marketing majors who graduated ten years ago and records information about each selected alumnus. The data on current incomes collected by the Alumni Department are provided on the following slide.

Accounting			Marketing		
$65,510	$74,290	$56,870	$97,440	$89,060	$77,650
$92,810	$77,560	$106,600	$102,910	$88,840	$88,600
$70,100	$69,160	$65,530	$89,080	$55,100	$53,570
$82,240	$85,130	$94,280	$90,470	$76,920	$66,050
$113,390	$97,950	$66,820	$114,560	$81,920	$54,810
$109,880	$61,820	$60,210	$96,060	$82,150	$63,980
$98,090	$66,770	$65,460	$108,040	$112,330	$71,140
$70,780	$64,150	$44,720	$124,140	$100,840	
$112,270	$78,390	$88,140	$113,930	$59,700	
$85,020	$57,390	$60,960	$75,340	$93,690	
$56,100	$62,840	$77,920	$97,370	$116,750	
$100,900	$92,880	$95,210	$99,600	$159,800	
$78,080	$72,800	$91,490	$76,510	$97,180	

Since we are interested in assessing the difference in mean incomes of the two groups, we do know the population standard deviations σ_{ACT} and σ_{MKT}, and we have taken large ($n_{ACT} \geq 30$ *and* $n_{MKT} \geq 30$) independent random samples of the two groups, we can build a confidence interval for the difference between the two population means $\mu_{ACT} - \mu_{MKT}$:

$$\left(\bar{x}_{ACT} - \bar{x}_{MKT}\right) \pm z_{\alpha/2}\sigma_{\bar{x}_{ACT}-\bar{x}_{MKT}} \text{ where } \sigma_{\bar{x}_{ACT}-\bar{x}_{MKT}} = \sqrt{\frac{\sigma^2_{ACT}}{n_{ACT}} + \frac{\sigma^2_{MKT}}{n_{MKT}}}$$

Since we have

$$\bar{x}_{ACT} = \frac{\sum_{i=1}^{n_{ACT}} \bar{x}_{ACT}}{n_{ACT}} = \frac{3,070,509}{39} = 78,731.00, \sigma_{ACT} = 15,000$$

$$\bar{x}_{MKT} = \frac{\sum_{i=1}^{n_{MKT}} \bar{x}_{MKT}}{n_{MKT}} = \frac{2,975,522}{33} = 90,167.33, \sigma_{MKT} = 20,000$$

and

$$\sigma_{\bar{x}_{ACT}-\bar{x}_{MKT}} = \sqrt{\frac{\sigma^2_{ACT}}{n_{ACT}} + \frac{\sigma^2_{MKT}}{n_{MKT}}} = \sqrt{\frac{15,000^2}{39} + \frac{20,000^2}{33}} = 4,229.71$$

and

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

...so

$$\begin{aligned}&\left(\bar{x}_{ACT} - \bar{x}_{MKT}\right) \pm z_{\alpha/2}\sigma_{\bar{x}_{ACT}-\bar{x}_{MKT}} \\ &= \left(78,731.00 - 90,167.33\right) \pm 1.96\left(4,229.71\right) \\ &= -11,436.33 \pm 8,290.23 \\ &= \left(-19,726.6, -3,146.1\right)\end{aligned}$$

95% of all confidence intervals estimated in a similar manner on random samples of the same size taken from these two populations will contain the true value of $\mu_{ACT} - \mu_{MKT}$ (note that this interval does not contain 0).

2. Hypothesis Testing of the difference between the two population means $\mu_1 - \mu_2$ – Large Sample Case ($n_1 \geq 30$ *and* $n_2 \geq 30$) or the populations are normal, and we know the population standard deviations σ_1 and σ_2:

- Null and Alternative Hypotheses

H_o: $\mu_1 - \mu_2$ ($\leq = \geq$) D_o H_a: $\mu_1 - \mu_2$ ($< \neq >$) D_o

- Appropriate Test Statistic

$$z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - D_0}{\sigma_{\bar{x}_1-\bar{x}_2}} = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Example: Suppose we wish to test that there is no difference in incomes earned by Whatsamatta University accounting and marketing majors ten years after they receive their bachelors degrees at an $\alpha = 0.05$ level of significance. From historical records, the Whatsamatta U. Alumni Relations Department knows the standard deviation of incomes earned ten years after graduation for their alumni with undergraduate degrees in accounting and marketing are 15,000 and 20,000, respectively. The Alumni Relations Department collects independent random samples of accounting and marketing majors who graduated ten years ago and records information about each selected alumnus. The data on current incomes collected by the Alumni Department are provided on the following slide.

Accounting			Marketing		
$65,510	$74,290	$56,870	$97,440	$89,060	$77,650
$92,810	$77,560	$106,600	$102,910	$88,840	$88,600
$70,100	$69,160	$65,530	$89,080	$55,100	$53,570
$82,240	$85,130	$94,280	$90,470	$76,920	$66,050
$113,390	$97,950	$66,820	$114,560	$81,920	$54,810
$109,880	$61,820	$60,210	$96,060	$82,150	$63,980
$98,090	$66,770	$65,460	$108,040	$112,330	$71,140
$70,780	$64,150	$44,720	$124,140	$100,840	
$112,270	$78,390	$88,140	$113,930	$59,700	
$85,020	$57,390	$60,960	$75,340	$93,690	
$56,100	$62,840	$77,920	$97,370	$116,750	
$100,900	$92,880	$95,210	$99,600	$159,800	
$78,080	$72,800	$91,490	$76,510	$97,180	

- **State the Null and Alternative Hypotheses**

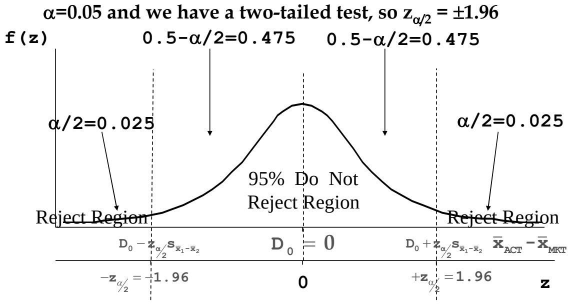
$$H_o: \mu_{ACT} - \mu_{MKT} = 0$$
$$H_a: \mu_{ACT} - \mu_{MKT} \neq 0$$

- **Select the Appropriate Test Statistic – $n_{ACT} = 39 \geq 30$ and $n_{MKT} = 33 \geq 30$, and we know the population standard deviations σ_{ACT} and σ_{MKT}, so use**

$$z = \frac{(\bar{x}_{ACT} - \bar{x}_{MKT}) - D_0}{\sigma_{\bar{x}_{ACT} - \bar{x}_{MKT}}} = \frac{(\bar{x}_{ACT} - \bar{x}_{MKT}) - D_0}{\sqrt{\frac{\sigma^2_{ACT}}{n_{ACT}} + \frac{\sigma^2_{MKT}}{n_{MKT}}}}$$

- **State the Desired Level of Significance α, Find the Critical Value(s) and State the Decision Rule**

α=0.05 and we have a two-tailed test, so $z_{\alpha/2} = \pm 1.96$



Decision rule – do not reject H_o if $-1.96 \leq z \leq 1.96$ otherwise reject H_o

- **Calculate the Test Statistic**

$$z = \frac{\left(\bar{x}_{ACT} - \bar{x}_{MKT}\right) - D_0}{\sqrt{\frac{\sigma^2_{ACT}}{n_{ACT}} + \frac{\sigma^2_{MKT}}{n_{MKT}}}} = \frac{(78,731.00 - 90,167.33) - 0}{\sqrt{\frac{15,000^2}{39} + \frac{20,000^2}{33}}}$$

$$= \frac{11,436.33}{4,229.71} = -2.704$$

- **Use the Decision Rule to Evaluate the Test Statistic and Decide Whether to Reject or Not Reject the Null Hypothesis**

$-z_{\alpha/2} \nleq z \leq z_{\alpha/2}$, i.e., $-1.96 \nleq -2.704 \leq 1.96$

so reject H_o - There is strong evidence of a difference in mean incomes earned by Whatsamatta University accounting and marketing majors ten years after they receive their bachelors degrees .

Again note we rarely know the population standard deviations σ_1 and σ_2 (why not?) - fortunately, sample standard deviation are very reliable estimators of population standard deviations – if the sample is reasonably large, we can directly substitute s for σ!

3. **Interval Estimate of the difference between the two population means $\mu_1 - \mu_2$ – Large Sample Case ($n_1 \geq 30$ *and* $n_2 \geq 30$) and the population standard deviations σ_1 and σ_2 are unknown:**

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2} \text{ where } s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and 1 - α is the confidence coefficient.

Since both samples are large, the sample standard deviations s_1 and s_2 are very good estimators of the population standard deviations σ_1 and σ_2, and so are directly substituted without any adjustment.

Example: Suppose we wish to assess the difference of incomes earned by Whatsamatta University accounting and marketing majors ten years after they receive their bachelors degrees at the 95% level of confidence. The Alumni Relations Department collects independent random samples of accounting and marketing majors who graduated ten years ago and records information about each selected alumnus. The data on current incomes collected by the Alumni Department are provided on the following slide.

Accounting			Marketing		
$65,510	$74,290	$56,870	$97,440	$89,060	$77,650
$92,810	$77,560	$106,600	$102,910	$88,840	$88,600
$70,100	$69,160	$65,530	$89,080	$55,100	$53,570
$82,240	$85,130	$94,280	$90,470	$76,920	$66,050
$113,390	$97,950	$66,820	$114,560	$81,920	$54,810
$109,880	$61,820	$60,210	$96,060	$82,150	$63,980
$98,090	$66,770	$65,460	$108,040	$112,330	$71,140
$70,780	$64,150	$44,720	$124,140	$100,840	
$112,270	$78,390	$88,140	$113,930	$59,700	
$85,020	$57,390	$60,960	$75,340	$93,690	
$56,100	$62,840	$77,920	$97,370	$116,750	
$100,900	$92,880	$95,210	$99,600	$159,800	
$78,080	$72,800	$91,490	$76,510	$97,180	

Since we are interested in assessing the difference in mean incomes of the two groups, we don't know the population standard deviations σ_{ACT} and σ_{MKT}, and we have taken large ($n_{ACT} \geq 30$ *and* $n_{MKT} \geq 30$) independent random samples of the two groups, we can build a confidence interval for the difference between the two population means $\mu_{ACT} - \mu_{MKT}$:

$$\left(\bar{x}_{ACT} - \bar{x}_{MKT}\right) \pm z_{\alpha/2} s_{\bar{x}_{ACT} - \bar{x}_{MKT}} \text{ where } s_{\bar{x}_{ACT} - \bar{x}_{MKT}} = \sqrt{\frac{s_{ACT}^2}{n_{ACT}} + \frac{s_{MKT}^2}{n_{MKT}}}$$

Since we have

$$\bar{x}_{ACT} = \frac{\sum_{i=1}^{n_{ACT}} x_{ACT}}{n_{ACT}} = \frac{3,070,509}{39} = 78,731.00$$

$$\bar{x}_{MKT} = \frac{\sum_{i=1}^{n_{MKT}} x_{MKT}}{n_{MKT}} = \frac{2,975,522}{33} = 90,167.33$$

and

$$s_{\bar{x}_{ACT}-\bar{x}_{MKT}} = \sqrt{\frac{s^2_{ACT}}{n_{ACT}} + \frac{s^2_{MKT}}{n_{MKT}}}$$

where

$$s^2_{ACT} = \frac{\sum_{i=1}^{n_{ACT}} \left(x_{ACT,i} - \bar{x}_{ACT}\right)^2}{n - 1}$$

$$= \frac{(65,510 - 78,731.00)^2 + (92,810 - 78,731.00)^2 + \cdots + (91,490 - 78,731.00)^2}{39 - 1}$$

$$= 17,527.9305$$

and

$$s^2_{MKT} = \frac{\sum_{i=1}^{n_{MKT}} \left(x_{MKT,i} - \bar{x}_{MKT}\right)^2}{n - 1}$$

$$= \frac{(97,440 - 90,167.33)^2 + (102,910 - 90,167.33)^2 + \cdots + (71,140 - 90,167.33)^2}{33 - 1}$$

$$= 22,708.1202$$

so

$$s_{\bar{x}_{ACT}-\bar{x}_{MKT}} = \sqrt{\frac{s^2_{ACT}}{n_{ACT}} + \frac{s^2_{MKT}}{n_{MKT}}} = \sqrt{\frac{17,527.9305^2}{39} + \frac{22,708.1202^2}{33}} = 4,848.06$$

and

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

…and so

$$\left(\bar{x}_{ACT} - \bar{x}_{MKT}\right) \pm z_{\alpha/2} s_{\bar{x}_{ACT}-\bar{x}_{MKT}}$$

$$= (78,731.00 - 90,167.33) \pm 1.96(4,848.06)$$

$$= -11,436.33 \pm 9,502.20$$

$$= (-20,938.5, -1,934.1)$$

Note the interval now wider – this is because our estimated standard deviations s_{ACT} and s_{MKT} are each larger than the population standard deviations σ_{ACT} and σ_{MKT} they are estimating (why?).

4. Hypothesis Testing of the difference between the two population means $\mu_1 - \mu_2$ – Large Sample Case ($n_1 \geq 30$ *and* $n_2 \geq 30$) or the populations are normal, and the population standard deviations σ_1 and σ_2 are unknown

- **Null and Alternative Hypotheses**

H_o: $\mu_1 - \mu_2$ ($\leq = \geq$) D_o $\qquad$ H_a: $\mu_1 - \mu_2$ ($< \neq >$) D_o

- **Appropriate Test Statistic**

$$z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - D_0}{s_{\bar{x}_1-\bar{x}_2}} = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Example: Suppose we wish to test there is no difference in incomes earned by Whatsamatta University accounting and marketing majors ten years after they receive their bachelors degrees at an $\alpha = 0.05$ level of significance. The Alumni Relations Department collects independent random samples of accounting and marketing majors who graduated ten years ago and records information about each selected alumnus. The data on current incomes collected by the Alumni Department are provided on the following slide.

Accounting			Marketing		
$65,510	$74,290	$56,870	$97,440	$89,060	$77,650
$92,810	$77,560	$106,600	$102,910	$88,840	$88,600
$70,100	$69,160	$65,530	$89,080	$55,100	$53,570
$82,240	$85,130	$94,280	$90,470	$76,920	$66,050
$113,390	$97,950	$66,820	$114,560	$81,920	$54,810
$109,880	$61,820	$60,210	$96,060	$82,150	$63,980
$98,090	$66,770	$65,460	$108,040	$112,330	$71,140
$70,780	$64,150	$44,720	$124,140	$100,840	
$112,270	$78,390	$88,140	$113,930	$59,700	
$85,020	$57,390	$60,960	$75,340	$93,690	
$56,100	$62,840	$77,920	$97,370	$116,750	
$100,900	$92,880	$95,210	$99,600	$159,800	
$78,080	$72,800	$91,490	$76,510	$97,180	

- **State the Null and Alternative Hypotheses**

  $H_o: \mu_{ACT} - \mu_{MKT} = 0$

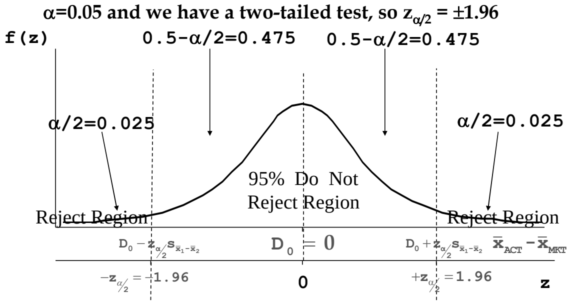
  $H_a: \mu_{ACT} - \mu_{MKT} \neq 0$

- **Select the Appropriate Test Statistic – $n_{ACT} = 39 \geq 30$ and $n_{MKT} = 33 \geq 30$, and we do not know the population standard deviations σ_{ACT} and σ_{MKT}, so use**

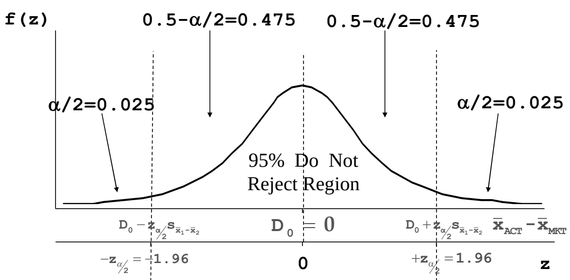
$$z = \frac{\left(\bar{x}_{ACT} - \bar{x}_{MKT}\right) - D_0}{s_{\bar{x}_{ACT} - \bar{x}_{MKT}}} = \frac{\left(\bar{x}_{ACT} - \bar{x}_{MKT}\right) - D_0}{\sqrt{\frac{s_{ACT}^2}{n_{ACT}} + \frac{s_{MKT}^2}{n_{MKT}}}}$$

- **State the Desired Level of Significance α, Find the Critical Value(s) and State the Decision Rule**

  α=0.05 and we have a two-tailed test, so $z_{\alpha/2} = \pm 1.96$



Decision rule – do not reject H_o if $-1.96 \le z \le 1.96$ otherwise reject H_o



- **Calculate the Test Statistic**

$$z = \frac{(\bar{x}_{ACT} - \bar{x}_{MKT}) - D_0}{\sqrt{\frac{s^2_{ACT}}{n_{ACT}} + \frac{s^2_{MKT}}{n_{MKT}}}} = \frac{(78,731.00 - 90,167.33) - 0}{\sqrt{\frac{17,527.93^2}{39} + \frac{22,708.12^2}{33}}}$$

$$= \frac{11,436.33}{4,848.061} = -2.359$$

- **Use the Decision Rule to Evaluate the Test Statistic and Decide Whether to Reject or Not Reject the Null Hypothesis**

  $-z_{\alpha/2} \nleq z \le z_{\alpha/2}$, i.e., $-1.96 \nleq -2.359 \le 1.96$

  so reject H_o - there is strong evidence of a difference in mean incomes earned by Whatsamatta University accounting and marketing majors ten years after they receive their bachelors degrees.

What if we are unable to obtain large samples? How do we then build confidence intervals for the difference between the two population means $\mu_1 - \mu_2$:

5. **Interval Estimate of the difference between the two population means $\mu_1 - \mu_2$ – Small Sample Case (n_1 < *and/or* n_2 < 30), the populations are normal, and population standard deviations σ_1 and σ_2 are known and equal**

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2} \text{ where } \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $1 - \alpha$ is the confidence coefficient.

Example: Rheal Toor works in the assessment office of Biendo Parish. In order to better understand the differences in property taxes collected in the two major cities in his parish (West Liberty and East Liberty), Mr. Toor wants to assess the differences in square footage of homes in the two cities at the 99% level of confidence. From historical records, Mr. Toor knows the square footage of homes in West Liberty and East Liberty are each normally distributed with a standard deviation of 2,100. He has collected independent random samples of square footages for homes in West Liberty and East Liberty and recorded this information. These data are provided on the following slide.

West Liberty	East Liberty
15520	16280
14200	13510
12950	12950
18100	17400
11550	18900
16500	18000
12050	16500
12075	12850
14900	15000
15650	17950
12500	16800
	17500

Since we are interested in assessing the difference in mean incomes of the two groups, we do know the population standard deviations σ_{WL} and σ_{EL} (and they are equal), but we have taken small (n_{WL} <30 *and* n_{EL} < 30) independent random samples of the two groups that are known to be normal, we can build a confidence interval for the difference between the two population means μ_{WL} - μ_{EL}:

$$\left(\bar{x}_{WL} - \bar{x}_{EL}\right) \pm z_{\alpha/2}\sigma_{\bar{x}_{WL}-\bar{x}_{EL}} \text{ where } \sigma_{\bar{x}_{WL}-\bar{x}_{EL}} = \sqrt{\sigma^2\left(\frac{1}{n_{WL}}+\frac{1}{n_{EL}}\right)}$$

Since we have

$$\bar{x}_{WL} = \frac{\sum_{i=1}^{n_{WL}} x_{WL}}{n_{WL}} = \frac{156,997}{11} = 14,272.45,\ \sigma_{WL} = 2,100$$

$$\bar{x}_{EL} = \frac{\sum_{i=1}^{n_{EL}} x_{EL}}{n_{EL}} = \frac{193,640}{12} = 16,136.67,\ \sigma_{EL} = 2,100$$

and

$$\sigma_{\bar{x}_{WL}-\bar{x}_{EL}} = \sqrt{\sigma^2\left(\frac{1}{n_{WL}}+\frac{1}{n_{EL}}\right)} = \sqrt{2100^2\frac{1}{11}+\frac{1}{12}} = 876.59$$

and

$$z_{\alpha/2} = z_{0.01/2} = z_{0.005} = 2.57$$

...so

$$\left(\bar{x}_{WL} - \bar{x}_{EL}\right) \pm z_{\alpha/2}\sigma_{\bar{x}_{WL}-\bar{x}_{EL}}$$
$$= \left(14,272.45 - 16,136.67\right) \pm 2.57\left(876.59\right)$$
$$= -1,864.21 \pm 2,252.84$$
$$= \left(-4,117.1,\ 388.6\right)$$

Again note we rarely know the population standard deviations σ_1 and σ_2 (why not?), and so consequently we rarely know the standard deviation of the difference between the two population means $\sigma_{\bar{x}_1-\bar{x}_2}$.

6. Hypothesis Testing of the difference between the two population means μ_1 - μ_2 – Small Sample Case (n_1 < 30 *and/or* n_2 < 30), the populations are normal, and the population standard deviations σ_1 and σ_2 are known and equal

- Null and Alternative Hypotheses

H_o: $\mu_1 - \mu_2$ (≤ = ≥) D_o $\qquad$ H_a: $\mu_1 - \mu_2$ (< ≠ >) D_o

- Appropriate Test Statistic

$$z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - D_0}{\sigma_{\bar{x}_1-\bar{x}_2}} = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - D_0}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Example: Rheal Toor works in the assessment office of Biendo Parish. In order to better understand the differences in property taxes collected in the two major cities in his parish (West Liberty and East Liberty), Mr. Toor wants to test his belief that homes in the East Liberty are at least 2000 square feet larger than homes in West Liberty at the $\alpha = 0.01$ level of significance. From historical records, Mr. Toor knows the square footage of homes in West Liberty and East Liberty are each normally distributed with a standard deviation of 2,100. He has collected independent random samples of square footages for homes in West Liberty and East Liberty and recorded this information. These data are provided on the following slide.

West Liberty	East Liberty
15520	16280
14200	13510
12950	12950
18100	17400
11550	18900
16500	18000
12050	16500
12075	12850
14900	15000
15650	17950
12500	16800
	17500

- State the Null and Alternative Hypotheses

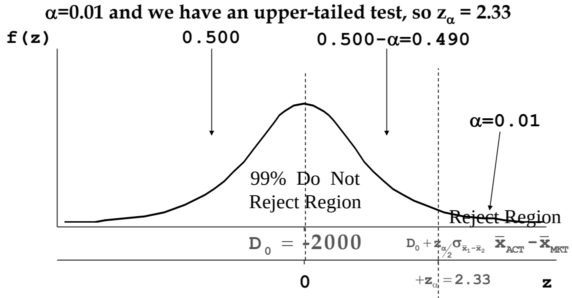
$$H_o: \mu_{WL} - \mu_{EL} \le -2000$$
$$H_a: \mu_{WL} - \mu_{EL} > -2000$$

- Select the Appropriate Test Statistic – $n_{WL} = 11 < 30$ and $n_{EL} = 12 < 30$, the populations are normal, and we know the population standard deviations σ_{WL} and σ_{EL} (and they are equal), so use

$$z = \frac{(\bar{x}_{WL} - \bar{x}_{EL}) - D_0}{\sigma_{\bar{x}_{WL}-\bar{x}_{EL}}} = \frac{(\bar{x}_{WL} - \bar{x}_{EL}) - D_0}{\sqrt{\sigma^2\left(\frac{1}{n_{WL}} + \frac{1}{n_{EL}}\right)}}$$

- **State the Desired Level of Significance α, Find the Critical Value(s) and State the Decision Rule**

α=0.01 and we have an upper-tailed test, so z_α = 2.33



Decision rule – do not reject H_o if $z \le 2.33$ otherwise reject H_o

- **Calculate the Test Statistic**

$$z = \frac{\left(\bar{x}_{WL} - \bar{x}_{EL}\right) - D_0}{\sqrt{\sigma^2\left(\frac{1}{n_{WL}} + \frac{1}{n_{EL}}\right)}} = \frac{\left(14,272.45 - 16,136.67\right) - (-2000)}{\sqrt{2100^2\left(\frac{1}{11} + \frac{1}{12}\right)}}$$

$$= \frac{135.79}{876.59} = 0.15$$

- **Use the Decision Rule to Evaluate the Test Statistic and Decide Whether to Reject or Not Reject the Null Hypothesis**

$z \le z_\alpha$, i.e., $0.15 \le 2.33$

so do not reject H_o - there is insufficient evidence to conclude the homes in East Liberty are at least 2000 square feet larger than homes in West liberty.

What if we are unable to obtain large samples and we don't know the population standard deviations σ_1 and σ_2? How do we then build confidence intervals for the difference between the two population means $\mu_1 - \mu_2$:

7. **Interval Estimate of the difference between the two population means $\mu_1 - \mu_2$ – Small Sample Case ($n_1 <$ *and/or* $n_2 < 30$), the populations are normal, and population standard deviations σ_1 and σ_2 are unknown but known to be equal**

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{\alpha/2,df}\, s_{\bar{x}_1 - \bar{x}_2} \text{ where } s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{and } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where $1 - \alpha$ is the confidence coefficient and $df = n_1 + n_2 - 2$ are the degrees of freedom.

Example: Rheal Toor works in the assessment office of Biendo Parish. In order to better understand the differences in property taxes collected in the two major cities in his parish (West Liberty and East Liberty), Mr. Toor wants to assess the differences in square footage of homes in the two cities at the 99% level of confidence. From historical records, Mr. Toor knows the square footage of homes in West Liberty and East Liberty are each normally distributed with similar variation. He has collected independent random samples of square footages for homes in West Liberty and East Liberty and recorded this information. These data are provided on the following slide.

West Liberty	East Liberty
15520	16280
14200	13510
12950	12950
18100	17400
11550	18900
16500	18000
12050	16500
12075	12850
14900	15000
15650	17950
12500	16800
	17500

Since we are interested in assessing the difference in mean incomes of the two groups, we do not know the population standard deviations σ_{WL} and σ_{EL} (but have an indication they are equal), and we have taken small (n_{WL} <30 *and* n_{EL} < 30) independent random samples of the two groups that are known to be normal, we can build a confidence interval for the difference between the two population means μ_{WL} - μ_{EL}:

$$\left(\bar{x}_{WL} - \bar{x}_{EL}\right) \pm t_{\alpha/2,df} s_{\bar{x}_{WL}-\bar{x}_{EL}} \text{ where } s_{\bar{x}_{WL}-\bar{x}_{EL}} = \sqrt{s_p^2\left(\frac{1}{n_{WL}} + \frac{1}{n_{EL}}\right)}$$

Since we have

$$\bar{x}_{WL} = \frac{\sum_{i=1}^{n_{WL}} \bar{x}_{WL}}{n_{WL}} = \frac{156,997}{11} = 14,272.45$$

$$\bar{x}_{EL} = \frac{\sum_{i=1}^{n_{EL}} \bar{x}_{EL}}{n_{EL}} = \frac{193,640}{12} = 16,136.67$$

where

$$s_{WL} = \sqrt{\frac{\sum_{i=1}^{n_{WL}}\left(x_{WL,i} - \bar{x}_L\right)^2}{n_{WL} - 1}}$$

$$= \sqrt{\frac{(15,520 - 14,272.45)^2 + (14,200 - 14,272.45)^2 + \cdots + (12,500 - 14,272.45)^2}{11 - 1}}$$

$$= 2,052.94$$

and

$$s_{EL} = \sqrt{\frac{\sum_{i=1}^{n_{EL}}\left(x_{EL,i} - \bar{x}_{EL}\right)^2}{n_{EL} - 1}}$$

$$= \sqrt{\frac{(16,280 - 16,136.67)^2 + (13,510 - 16,136.67)^2 + \cdots + (17,500 - 16,136.67)^2}{12 - 1}}$$

$$= 2,080.03$$

…so

$$s_p^2 = \frac{(n_{WL} - 1)s_{WL}^2 + (n_{EL} - 1)s_{EL}^2}{n_{WL} + n_{EL} - 2}$$

$$= \frac{(11 - 1)2057.94^2 + (12 - 1)2080.03^2}{11 + 12 - 2} = 2067.13$$

and

$$s_{\bar{x}_{WL} - \bar{x}_{EL}} = \sqrt{s_p^2\left(\frac{1}{n_{WL}} + \frac{1}{n_{EL}}\right)} = \sqrt{2067^2\left(\frac{1}{11} + \frac{1}{12}\right)} = 862.87$$

and

$$t_{\alpha/2,df} = t_{0.01/2,11+12-2} = t_{0.005,21} = 2.831$$

…so

$$\left(\bar{x}_{WL} - \bar{x}_{EL}\right) \pm t_{\alpha/2,df}\, s_{\bar{x}_{WL} - \bar{x}_{EL}}$$

$$= (14,272.45 - 16,136.67) \pm 2.831(862.87)$$

$$= -1,864.21 \pm 2,481.63$$

$$= (-4,345.8, 617.4)$$

Note the interval now wider – this is because our estimated standard deviations s_{WL} and S_{EL} are larger than the population standard deviations σ_{WL} and σ_{EL} they are estimating (why?).

8. **Hypothesis Testing of the difference between the two population means μ_1 - μ_2 – Small Sample Case (n_1 < 30 *and/or* n_2 < 30), the populations are normal, and the population standard deviations σ_1 and σ_2 are unknown but known to be equal**

- **Null and Alternative Hypotheses**

H_o: $\mu_1 - \mu_2$ ($\leq = \geq$) D_o H_a: $\mu_1 - \mu_2$ ($< \neq >$) D_o

- **Appropriate Test Statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_{\bar{x}_1 - \bar{x}_2}} \text{ where } s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{and } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Example: Rheal Toor works in the assessment office of Biendo Parish. In order to better understand the differences in property taxes collected in the two major cities in his parish (West Liberty and East Liberty), Mr. Toor wants to test his belief that homes in East Liberty are more than 2000 square feet larger than homes in West Liberty at the α=0.01 level of significance. From historical records, Mr. Toor knows the square footage of homes in West Liberty and East Liberty are each normally distributed with similar variation. He has collected independent random samples of square footages for homes in West Liberty and East Liberty and recorded this information. These data are provided on the following slide.

West Liberty	East Liberty
15520	16280
14200	13510
12950	12950
18100	17400
11550	18900
16500	18000
12050	16500
12075	12850
14900	15000
15650	17950
12500	16800
	17500

- **State the Null and Alternative Hypotheses**

  H_o: $\mu_{WL} - \mu_{EL} \geq -2000$

  H_a: $\mu_{WL} - \mu_{EL} < -2000$

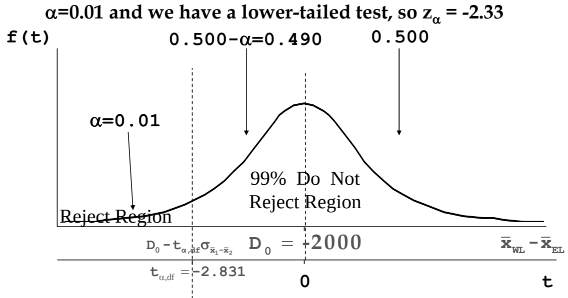
- **Select the Appropriate Test Statistic – n_{WL} = 11 < 30 and n_{EL} =12 < 30, the populations are normal, and we do not know the population standard deviations σ_{WL} and σ_{EL} (but we know they are equal), so use**

$$t = \frac{(\bar{x}_{WL} - \bar{x}_{EL}) - D_0}{s_{\bar{x}_{WL}-\bar{x}_{EL}}} \text{ where } s_{\bar{x}_1-\bar{x}_2} = \sqrt{s_p^2\left(\frac{1}{n_{WL}} + \frac{1}{n_{EL}}\right)}$$

$$\text{and } s_p^2 = \frac{(n_{WL} - 1)s_{WL}^2 + (n_{EL} - 1)s_{EL}^2}{n_{WL} + n_{EL} - 2}$$

- **State the Desired Level of Significance α, Find the Critical Value(s) and State the Decision Rule**

  α=0.01 and we have a lower-tailed test, so z_α = -2.33



Decision rule – do not reject H_o if $z \geq -2.33$ otherwise reject H_o

- **Calculate the Test Statistic**

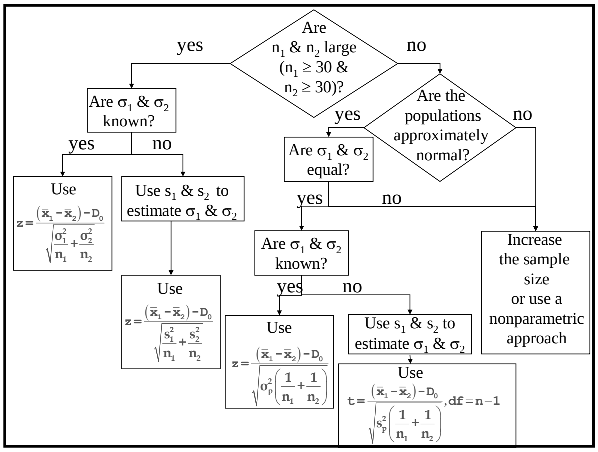
$$z = \frac{(\bar{x}_{WL} - \bar{x}_{EL}) - D_0}{\sqrt{s^2\left(\frac{1}{n_{WL}} + \frac{1}{n_{EL}}\right)}} = \frac{(14,272.45 - 16,136.67) - (-2000)}{\sqrt{2067.13^2\left(\frac{1}{11} + \frac{1}{12}\right)}}$$

$$= \frac{135.79}{862.87} = 0.16$$

- **Use the Decision Rule to Evaluate the Test Statistic and Decide Whether to Reject or Not Reject the Null Hypothesis**

  $z \geq -z_\alpha$, i.e., $0.16 \geq -2.33$

  so do not reject H_o - there is insufficient evidence to conclude homes in East Liberty are at least 2000 square feet larger than homes in West Liberty.



C. Differences between Means of Two Populations – Matched Samples

Define a new variable *d* where $d_i = x_{i,1} - x_{i,2}$ and let μ_d be the population mean of *d*, i.e.,

$$\mu_d = \frac{\sum_{i=1}^{N} d_i}{N} = \frac{\sum_{i=1}^{N} \left(x_{i,1} - x_{i,2}\right)}{N}$$

To make a point estimate of the mean difference between matched observations $d_1 - d_2$ we take a sample of size n matched observations and calculate

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{n} \left(x_{i,1} - x_{i,2}\right)}{n}$$

The point estimate $\bar{d}$ has a sampling distribution with the following characteristics:

$$\text{Expected Value}: E\left[\bar{d}\right] = \mu_d$$

$$\text{Standard Error}: \sigma_{\bar{d}} = \sigma_d / \sqrt{n} = \sqrt{\frac{\sum_{i=1}^{N} \left(d_i - \mu_d\right)^2}{N}} \Bigg/ \sqrt{n}$$

If the sample of matched observations is large ($n \geq 30$) or a small sample is drawn from a normal parent population, the sampling distribution of $\bar{d}$ is *approximately* normal.

This sampling distribution can be used to build confidence intervals and test hypotheses for the mean difference between the matched observations μ_d:

1. Interval Estimate of the population mean difference between matched observations μ_d – Large Sample Case (n ≥ 30) or the population is normal and the population standard deviation σ_d is known:

$$\bar{d} \pm z_{\alpha/2}\sigma_{\bar{d}} = \bar{d} \pm z_{\alpha/2}\, \sigma_d / \sqrt{n}$$

where 1 - α is the confidence coefficient.

Note we rarely know the population standard deviation σ_d and so consequently we rarely know the standard deviation of the mean difference between the matched observations $\sigma_{\bar{d}}$.

Example: Suppose we wish to assess the difference of actual diameters of holes drilled with a 3 cm. bit into 4 inch thick disks of stainless steel with machines supplied by two manufacturers, Driller Inc. and Hole-Maker Ltd, at the 90% level of confidence. We plan to allow each of 36 employees to drill one hole with a 3 cm. bit into a 4 inch thick disk of stainless steel with one machine supplied by each manufacturer. From historical records, we know the standard deviation of the differences in holes drilled with a 3 cm. bit into 4-inch thick disks of stainless steel by an employee with machines provided by the two suppliers is 0.100. The actual diameters of the thirty-six holes drilled by the employees with each machine are provided on the following slide.

Driller Inc.							Hole-Maker Ltd					
Emp Num	Diam (cm.)	Emp Num	Diam (cm.)	Emp Num	Diam (cm.)		Emp Num	Diam (cm.)	Emp Num	Diam (cm.)	Emp Num	Diam (cm.)
42	2.91	59	3.02	63	2.98		42	2.98	59	3.11	63	3.14
18	3.02	163	2.94	144	2.84		18	2.93	163	2.90	144	3.11
37	2.92	299	2.88	177	3.00		37	3.02	299	3.00	177	3.08
125	2.95	178	2.94	260	3.04		125	3.04	178	3.05	260	2.96
98	3.01	105	2.89	204	2.94		98	2.97	105	3.06	204	2.99
211	3.04	88	3.01	161	2.89		211	3.01	88	3.20	161	3.09
57	3.00	123	2.93	194	3.04		57	3.07	123	3.00	194	2.97
186	2.89	136	2.98	26	3.06		186	3.02	136	3.07	26	2.98
142	2.97	112	3.04	133	3.02		142	3.09	112	2.97	133	3.14
8	2.97	237	2.99	94	2.95		8	2.98	237	2.97	94	2.99
245	2.94	256	2.92	145	2.95		245	2.96	256	3.14	145	3.11
139	2.91	40	2.98	203	3.04		139	3.07	40	2.95	203	3.08

Since we have matched observations (an individual employee drills a hole with each of the two machines supplied) we first must calculate $d_i = x_{i,DI} - x_{i,HM}$ for each employee:

Employee	Diameter (cm.)	Employee	Diameter (cm.)	Employee	Diameter (cm.)
42	-0.07	59	-0.09	63	-0.16
18	0.09	163	0.04	144	-0.27
37	-0.10	299	-0.12	177	-0.08
125	-0.09	178	-0.11	260	0.08
98	0.04	105	-0.17	204	-0.05
211	0.03	88	-0.19	161	-0.20
57	-0.07	123	-0.07	194	0.07
186	-0.13	136	-0.09	26	0.08
142	-0.12	112	0.07	133	-0.12
8	-0.01	237	0.02	94	-0.04
245	-0.02	256	-0.22	145	-0.16
139	-0.16	40	0.03	203	-0.04

Since we are interested in assessing the mean difference in holes drilled with the two machines by the employees, we do know the population standard deviation σ_d, and we have taken a large ($n_d = 36 \geq 30$) random sample of matched observations, we can build a confidence interval for the mean difference in the holes drilled with the two machines by the employees μ_d:

$$\bar{d} \pm z_{\alpha/2}\sigma_{\bar{d}}$$

Since we have

$$\bar{d} = \frac{\sum_{i=1}^{n}\bar{d}}{n} = \frac{-2.40}{36} = -0.0667$$

and

$$\sigma_{\bar{d}} = \sigma_d/\sqrt{n} = 0.100/\sqrt{36} = 0.01667$$

and

$$z_{\alpha/2} = z_{0.10/2} = z_{0.05} = 1.65$$

…so

$$\begin{aligned}&\bar{d} \pm z_{\alpha/2}\sigma_{\bar{d}}\\ &= -0.0667 \pm 1.65(0.01667)\\ &= -0.0667 \pm 0.0260\\ &= (-0.09417, -0.03917)\end{aligned}$$

90% of all confidence intervals estimated in a similar manner on random samples of the same size taken from this population will contain the true value of μ_d (note that this interval does not contain 0).

2. **Hypothesis Testing of the mean difference between matched observations μ_d – Large Sample Case (n ≥ 30) or the population is normal and the population standard deviation σ_d is known:**

- **Null and Alternative Hypotheses**

  H_o: μ_d (≤ = ≥) D_o H_a: μ_1 (< ≠ >) D_o

- **Appropriate Test Statistic**

$$z = \frac{\bar{d} - D_0}{\sigma_{\bar{d}}} = \frac{\bar{d} - D_0}{\sigma_d / \sqrt{n}}$$

Example: Suppose we wish to test that there is no difference in actual diameters of holes drilled with a 3 cm. bit into 4 inch thick disks of stainless steel with machines supplied by two manufacturers, Driller Inc. and Hole-Maker Ltd, at the α=0.10 level of confidence. We plan to allow each of 36 employees to drill one hole with a 3 cm. bit into a 4 inch thick disk of stainless steel with one machine supplied by each manufacturer. From historical records, we know the standard deviation of the differences holes drilled with a 3 cm. bit into 4-inch thick disks of stainless steel by an employee with machines provided by the two suppliers is 0.100. The actual diameters of the thirty-six holes drilled by the employees with each machine are provided on the following slide.

Driller Inc.							Hole-Maker Ltd					
Emp Num	Diam (cm.)	Emp Num	Diam (cm.)	Emp Num	Diam (cm.)		Emp Num	Diam (cm.)	Emp Num	Diam (cm.)	Emp Num	Diam (cm.)
42	2.91	59	3.02	63	2.98		42	2.98	59	3.11	63	3.14
18	3.02	163	2.94	144	2.84		18	2.93	163	2.90	144	3.11
37	2.92	299	2.88	177	3.00		37	3.02	299	3.00	177	3.08
125	2.95	178	2.94	260	3.04		125	3.04	178	3.05	260	2.96
98	3.01	105	2.89	204	2.94		98	2.97	105	3.06	204	2.99
211	3.04	88	3.01	161	2.89		211	3.01	88	3.20	161	3.09
57	3.00	123	2.93	194	3.04		57	3.07	123	3.00	194	2.97
186	2.89	136	2.98	26	3.06		186	3.02	136	3.07	26	2.98
142	2.97	112	3.04	133	3.02		142	3.09	112	2.97	133	3.14
8	2.97	237	2.99	94	2.95		8	2.98	237	2.97	94	2.99
245	2.94	256	2.92	145	2.95		245	2.96	256	3.14	145	3.11
139	2.91	40	2.98	203	3.04		139	3.07	40	2.95	203	3.08

Since we have matched observations (an individual employee drills a hole with each of the two machines supplied) we first must calculate $d_i = x_{i,DI} - x_{i,HM}$ for each employee:

Employee	Diameter (cm.)	Employee	Diameter (cm.)	Employee	Diameter (cm.)
42	-0.07	59	-0.09	63	-0.16
18	0.09	163	0.04	144	-0.27
37	-0.10	299	-0.12	177	-0.08
125	-0.09	178	-0.11	260	0.08
98	0.04	105	-0.17	204	-0.05
211	0.03	88	-0.19	161	-0.20
57	-0.07	123	-0.07	194	0.07
186	-0.13	136	-0.09	26	0.08
142	-0.12	112	0.07	133	-0.12
8	-0.01	237	0.02	94	-0.04
245	-0.02	256	-0.22	145	-0.16
139	-0.16	40	0.03	203	-0.04

- **State the Null and Alternative Hypotheses**

  H_o: $\mu_d = 0$

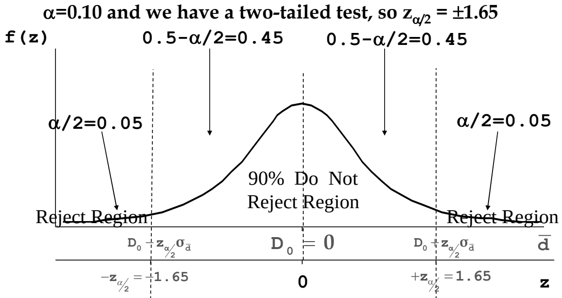
  H_a: $\mu_d \neq 0$

- **Select the Appropriate Test Statistic – n = 36 ≥ 30 and we know the population standard deviation σ_d, so use**

$$z = \frac{\bar{d} - D_0}{\sigma_{\bar{d}}} = \frac{\bar{d} - D_0}{\sigma_d / \sqrt{n}}$$

- **State the Desired Level of Significance α, Find the Critical Value(s) and State the Decision Rule**

  α=0.10 and we have a two-tailed test, so $z_{\alpha/2} = \pm 1.65$



Decision rule – do not reject H_o if $-1.65 \leq z \leq 1.65$ otherwise reject H_o

- **Calculate the Test Statistic**

$$z = \frac{\bar{d} - D_0}{\sigma_d / \sqrt{n}} = \frac{-0.0667 - 0.00}{0.10 / \sqrt{36}}$$

$$= \frac{-0.0667}{0.01667} = -4.00$$

- **Use the Decision Rule to Evaluate the Test Statistic and Decide Whether to Reject or Not Reject the Null Hypothesis**

$-z_{\alpha/2} \nleq z \leq z_{\alpha/2}$, i.e., $-1.96 \nleq -4.000 \leq 1.96$

so reject H_o - There is strong evidence of a difference in the sizes of holes drilled by an employee on the two machines.

What if the sample of matched differences is small (n < 30) and the standard deviation σ_d is unknown?

3. **Interval Estimate of the population mean difference between matched observations μ_d – Small Sample Case (n < 30) when the population is normal and population standard deviation σ_d is unknown.**

$$\bar{d} \pm t_{\alpha/2,df} s_{\bar{d}} = \bar{d} \pm t_{\alpha/2,df} \, s_d / \sqrt{n}$$

where 1 - α is the confidence coefficient and df = n - 1 is the degrees of freedom.

Example: Suppose we wish to assess the difference of actual diameters of holes drilled with a 3 cm. bit into 4 inch thick disks of stainless steel with machines supplied by two manufacturers, Driller Inc. and Hole-Maker Ltd, at the 90% level of confidence. We plan to allow each of 16 employees to drill one hole with a 3 cm. bit into a 4 inch thick disk of stainless steel with one machine supplied by each manufacturer. From historical records, we know the differences in holes drilled with a 3 cm. bit into 4-inch thick disks of stainless steel by an employee with machines provided by the two suppliers is normally distributed. The actual diameters of the sixteen holes drilled by the employees with each machine are provided on the following slide.

	Driller Inc.					Hole-Maker Ltd		
Emp Num	Diam (cm.)	Emp Num	Diam (cm.)		Emp Num	Diam (cm.)	Emp Num	Diam (cm.)
42	2.91	59	3.02		42	2.98	59	3.11
18	3.02	163	2.94		18	2.93	163	2.90
37	2.92	299	2.88		37	3.02	299	3.00
125	2.95	178	2.94		125	3.04	178	3.05
98	3.01	105	2.89		98	2.97	105	3.06
211	3.04	88	3.01		211	3.01	88	3.20
57	3.00	123	2.93		57	3.07	123	3.00
186	2.89	136	2.98		186	3.02	136	3.07

Since we have matched observations (an individual employee drills a hole with each of the two machines supplied) we first must calculate $d_i = x_{i,DI} - x_{i,HM}$ for each employee:

Employee	Diameter (cm.)	Employee	Diameter (cm.)
42	-0.07	59	-0.09
18	0.09	163	0.04
37	-0.10	299	-0.12
125	-0.09	178	-0.11
98	0.04	105	-0.17
211	0.03	88	-0.19
57	-0.07	123	-0.07
186	-0.13	136	-0.09

Since we are interested in assessing the mean difference in holes drilled with the two machines by the employees, we do not know the population standard deviation σ_d, and we have taken a small (n = 16 < 30) random sample of matched observations from a normal population, we can build a confidence interval for the mean difference in the holes drilled with the two machines by the employees μ_d:

$$\bar{d} \pm t_{\alpha/2,df} s_{\bar{d}}$$

Since we have

$$\bar{d} = \frac{\sum_{i=1}^{n} \bar{d}}{n} = \frac{-1.10}{16} = -0.06875$$

and

$$s_d = \frac{\sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2}{n - 1} = \frac{1.18728}{15} = 0.079152$$

and

$$s_{\bar{d}} = s_d \big/ \sqrt{n} = 0.07928 \big/ \sqrt{16} = 0.019788$$

and

$$t_{\alpha/2,df} = z_{0.10/2,16\text{-}1} = z_{0.05,15} = 1.753$$

…so

$$\begin{aligned} &\bar{d} \pm t_{\alpha/2,df} s_{\bar{d}} \\ &= \text{-}0.06875 \pm 1.753(0.019788) \\ &= \text{-}0.06875 \pm 0.034688 \\ &= (\text{-}0.10344, \text{-}0.03406) \end{aligned}$$

90% of all confidence intervals estimated in a similar manner on random samples of the same size taken from these two populations will contain the true value of μ_d (note that this interval does not contain 0).

2. Hypothesis Testing of the mean difference between matched observations μ_d – Small Sample Case (n ≥ 30) when the population is normal and the population standard deviation σ_d is unknown:

- **Null and Alternative Hypotheses**

$$H_o: \mu_d\ (\leq = \geq)\ D_o \qquad H_a: \mu_1\ (< \neq >)\ D_o$$

- **Appropriate Test Statistic**

$$t = \frac{\bar{d} - D_0}{s_{\bar{d}}} = \frac{\bar{d} - D_0}{s_d \big/ \sqrt{n}}$$

Example: Suppose we wish to test that there is no difference in actual diameters of holes drilled with a 3 cm. bit into 4 inch thick disks of stainless steel with machines supplied by two manufacturers, Driller Inc. and Hole-Maker Ltd, at the α=0.10 level of confidence. We plan to allow each of 16 employees to drill one hole with a 3 cm bit into a 4 inch thick disk of stainless steel with one machine supplied by each manufacturer. From historical records, we know the differences in holes drilled with a 3 cm. bit into 4-inch thick disks of stainless steel by an employee with machines provided by the two suppliers is normally distributed. The actual diameters of the sixteen holes drilled by the employees with each machine are provided on the following slide.

Driller Inc.				Hole-Maker Ltd			
Emp Num	Diam (cm.)	Emp Num	Diam (cm.)	Emp Num	Diam (cm.)	Emp Num	Diam (cm.)
42	2.91	59	3.02	42	2.98	59	3.11
18	3.02	163	2.94	18	2.93	163	2.90
37	2.92	299	2.88	37	3.02	299	3.00
125	2.95	178	2.94	125	3.04	178	3.05
98	3.01	105	2.89	98	2.97	105	3.06
211	3.04	88	3.01	211	3.01	88	3.20
57	3.00	123	2.93	57	3.07	123	3.00
186	2.89	136	2.98	186	3.02	136	3.07

Since we have matched observations (an individual employee drills a hole with each of the two machines supplied) we first must calculate $d_i = x_{i,DI} - x_{i,HM}$ for each employee:

Employee	Diameter (cm.)	Employee	Diameter (cm.)
42	-0.07	59	-0.09
18	0.09	163	0.04
37	-0.10	299	-0.12
125	-0.09	178	-0.11
98	0.04	105	-0.17
211	0.03	88	-0.19
57	-0.07	123	-0.07
186	-0.13	136	-0.09

- **State the Null and Alternative Hypotheses**

$$H_o: \mu_d = 0$$

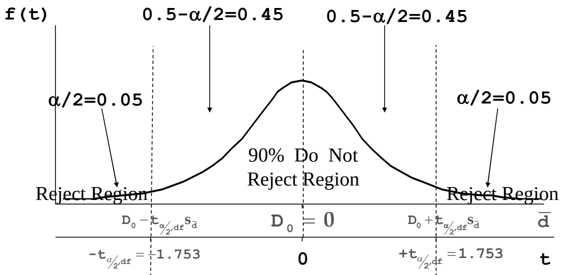
$$H_a: \mu_d \neq 0$$

- **Select the Appropriate Test Statistic – n = 16 < 30 and we know the population is normal, but we do not know the population standard deviation σ_d, so use**

$$t = \frac{\bar{d} - D_0}{s_{\bar{d}}} = \frac{\bar{d} - D_0}{s_d / \sqrt{n}}$$

- **State the Desired Level of Significance α, Find the Critical Value(s) and State the Decision Rule**

  α=0.10 and we have a two-tailed test with n-1=15 degrees of freedom, so $t_{\alpha/2,15} = \pm 1.753$



Decision rule – do not reject H_o if $-1.753 \le t \le 1.753$ otherwise reject H_o

- **Calculate the Test Statistic**

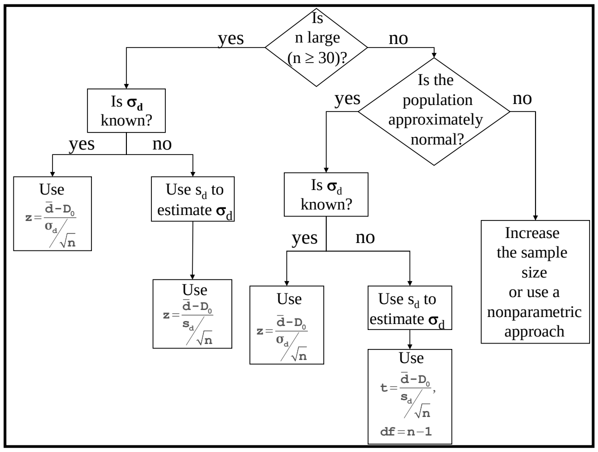
$$t = \frac{\bar{d} - D_0}{s_d / \sqrt{n}} = \frac{-0.06875 - 0.00}{0.079152 / \sqrt{16}} = \frac{-0.06875}{0.01979} = -3.474$$

- **Use the Decision Rule to Evaluate the Test Statistic and Decide Whether to Reject or Not Reject the Null Hypothesis**

  $-t_{\alpha/2,df} \nleq t \le t_{\alpha/2,df}$, i.e., $-1.753 \nleq -3.474 \le 1.753$

  so reject H_o - There is strong evidence of a difference in the sizes of holes drilled by an employee on the two machines.

Note that all we have actually done with matched observations is to define a new variable *d* where $d_i = x_{i,1} - x_{i,2}$ and treat this variable as we did other single quantitative variables when building confidence intervals (Chapter 8 of ASW) and testing hypotheses (Chapter 9 of ASW)!



D. Differences between Proportions of Two Populations – Independent Samples

Let p_1 be the proportion of population 1 and p_2 be the proportion of population 2

We can write the difference between the two population proportions as $p_1 - p_2$.

To make a point estimate of the difference between the two population proportions $p_1 - p_2$ we take independent samples of sizes n_1 and n_2 from the two populations and calculate

$$\bar{p}_1 - \bar{p}_2 = \frac{\sum_{i=1}^{n_1} x_{1,i}}{n_1} - \frac{\sum_{i=1}^{n_2} x_{2,i}}{n_2}$$

The standard error of the sampling distribution of $\bar{p}_1 - \bar{p}_2$ is

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Of course, we again encounter the problem of trying to use p_1 and p_2 to estimate p_1 and p_2!

Fortunately, if the sample sizes are relatively large – np_1, $n(1 - p_1)$, np_2, and $n(1 - p_2)$ all are at least 5 – we can use

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

1. **Interval Estimate of the difference between the two population means p_1 - p_2 – Large Sample Case ($np_1 \geq 5$, $n(1 - p_1) \geq 5$, $np_2 \geq 5$, *and* $n(1 - p_2) \geq 5$) :**

$$\left(\bar{p}_1 - \bar{p}_2\right) \pm z_{\alpha/2} s_{\bar{p}_1 - \bar{p}_2} \text{ where } s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1\left(1-\bar{p}_1\right)}{n_1} + \frac{\bar{p}_2\left(1-\bar{p}_2\right)}{n_2}}$$

and 1 - α is the confidence coefficient.

Example: Suppose we wish to assess the difference in employment rates for Whatsamatta University accounting and marketing majors ten weeks after they receive their bachelors degrees at the 95% level of confidence. The Alumni Relations Department collects independent random samples of 50 accounting majors and 45 marketing majors who graduated ten weeks ago and records information about each selected alumnus. Twenty-five of these accounting majors and thirty of these marketing majors were employed ten weeks after their graduation.

Since we are interested in assessing the difference in proportions of the two groups and we have taken large ($np_{ACT} \geq 5$, $n(1 - p_{ACT}) \geq 5$, $np_{MKT} \geq 5$, *and* $n(1 - p_{MKT}) \geq 5$) independent random samples of the two groups, we can build a confidence interval for the difference between the two population proportions p_{ACT} - p_{MKT}:

$$\left(\bar{p}_{ACT} - \bar{p}_{MKT}\right) \pm z_{\alpha/2} s_{\bar{p}_{ACT} - \bar{p}_{MKT}}$$

$$\text{where } s_{\bar{p}_{ACT} - \bar{p}_{MKT}} = \sqrt{\frac{\bar{p}_{ACT}\left(1-\bar{p}_{ACT}\right)}{n_{ACT}} + \frac{\bar{p}_{MKT}\left(1-\bar{p}_{MKT}\right)}{n_{MKT}}}$$

Since we have

$$\bar{p}_{ACT} = \frac{\sum_{i=1}^{n_{ACT}} x_{ACT}}{n_{ACT}} = \frac{25}{50} = 0.50 \text{ and } \bar{p}_{MKT} = \frac{\sum_{i=1}^{n_{MKT}} x_{MKT}}{n_{MKT}} = \frac{30}{45} = 0.667$$

and

$$s_{\bar{p}_{ACT}-\bar{p}_{MKT}} = \sqrt{\frac{\bar{p}_{ACT}\left(1-\bar{p}_{ACT}\right)}{n_{ACT}} + \frac{\bar{p}_{MKT}\left(1-\bar{p}_{MKT}\right)}{n_{MKT}}}$$
$$= \sqrt{\frac{0.50\left(1-0.50\right)}{50} + \frac{0.667\left(1-0.667\right)}{45}}$$
$$= \sqrt{0.004938 + 0.00500}$$
$$= 0.099691$$

and

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$

…so

$$\left(\bar{p}_{ACT} - \bar{p}_{MKT}\right) \pm z_{\alpha/2} s_{\bar{p}_{ACT}-\bar{p}_{MKT}}$$
$$= \left(0.500 - 0.667\right) \pm 1.96\left(0.099691\right)$$
$$= -0.1667 \pm 0.195394$$
$$= \left(-0.36206, 0.028727\right)$$

If we took independent random samples repeatedly, 95 percent of the confidence intervals we would construct from these samples would contain the true difference in the population proportion of accounting students and population proportion of marketing students employed after ten weeks - note the interval does contain zero (barely).

2. **Hypothesis Testing of the difference between proportions $p_1 - p_2$ – Large Sample Case ($np_1 \geq 5$, $n(1 - p_1) \geq 5$, $np_2 \geq 5$, *and* $n(1 - p_2) \geq 5$):**

- **Null and Alternative Hypotheses**

$H_o: p_1 - p_2 \ (\leq = \geq)\ D_o$ $\qquad$ $H_a: p_1 - p_2 \ (< \neq >)\ D_o$

- **Appropriate Test Statistic**

$$z = \frac{\left(\bar{p}_1 - \bar{p}_2\right) - D_0}{s_{\bar{p}_1-\bar{p}_2}} = \frac{\left(\bar{p}_1 - \bar{p}_2\right) - D_0}{\sqrt{\frac{\bar{p}_1\left(1-\bar{p}_1\right)}{n_1} + \frac{\bar{p}_2\left(1-\bar{p}_2\right)}{n_2}}}$$

Example: Suppose we wish to test that there is no difference in employment rates for Whatsamatta University accounting and marketing majors ten weeks after they receive their bachelors degrees at the α=0.05 level of significance. The Alumni Relations Department collects independent random samples of 50 accounting majors and 45 marketing majors who graduated ten weeks ago and records information about each selected alumnus. Twenty-five of these accounting majors and thirty of these marketing majors were employed ten weeks after their graduation.

- **State the Null and Alternative Hypotheses**

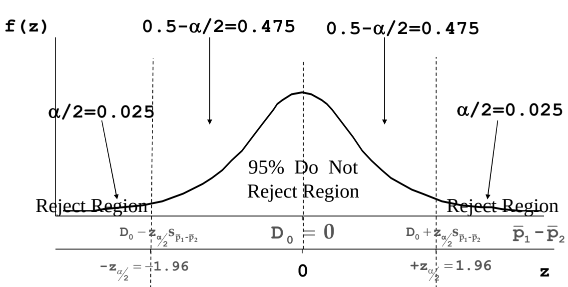
$$H_o: p_{ACT} - p_{MKT} = 0$$
$$H_a: \mu_{ACT} - P_{MKT} \neq 0$$

- **Select the Appropriate Test Statistic – $np_{ACT} \geq 5$, $n(1 - p_{ACT}) \geq 5$, $np_{MKT} \geq 5$, $n(1 - p_{MKT}) \geq 5$, so use**

$$z = \frac{(\bar{p}_{ACT} - \bar{p}_{MKT}) - D_0}{s_{\bar{p}_{ACT} - \bar{p}_{MKT}}}$$

- **State the Desired Level of Significance α, Find the Critical Value(s) and State the Decision Rule**

α=0.05 and we have a two-tailed test, so $z_{\alpha/2} = \pm 1.96$



Decision rule – do not reject H_o if $-1.96 \leq z \leq 1.96$ otherwise reject H_o

- **Calculate the Test Statistic**

$$z = \frac{(\bar{p}_{ACT} - \bar{p}_{MKT}) - D_0}{s_{\bar{p}_{ACT}-\bar{p}_{MKT}}} = \frac{(\bar{p}_{ACT} - \bar{p}_{MKT}) - D_0}{\sqrt{\frac{\bar{p}_{ACT}(1 - \bar{p}_{ACT})}{n_{ACT}} + \frac{\bar{p}_{MKT}(1 - \bar{p}_{MKT})}{n_{MKT}}}}$$

$$= \frac{(0.500 - 0.667) - 0.000}{\sqrt{\frac{0.500(1 - 0.500)}{50} + \frac{0.667(1 - 0.667)}{45}}} = \frac{-0.167}{0.099691} = -1.68$$

- **Use the Decision Rule to Evaluate the Test Statistic and Decide Whether to Reject or Not Reject the Null Hypothesis**

  $-z_{\alpha/2} \le z \le z_{\alpha/2}$, i.e., $-1.96 \le -1.68 \le 1.96$

  so do not reject H_o - there is insufficient evidence of a difference in the proportion of Accounting and Marketing majors who are employed ten weeks after graduation - note that at α=0.10 we would reach a different conclusion (barely).