Example:

Fifth grade students from two neighboring counties took a placement exam.

- Group 1, from County A, consisted of 8 students. The sample mean score for these students was 77.2 and the sample variance is 15.3.
- Group 2, from County B, consisted of 10 students and had a sample mean score of 75.3 and the sample variance is 19.7.

Example:

From previous years of data, it is believed that the scores for both counties are normally distributed.

Derive a test to determine whether or not the two population means are the same.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

- Estimate $\mu_1 - \mu_2$ with $\overline{X}_1 - \overline{X}_2$.
- $\overline{X}_1 - \overline{X}_2$ is normally distributed

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}\right)$$

- **Estimate $\mu_1 - \mu_2$ with $\overline{\text{X}}_1 - \overline{\text{X}}_2$.**
- **$\overline{\text{X}}_1 - \overline{\text{X}}_2$ is normally distributed**

$$Z = \frac{\overline{\text{X}}_1 - \overline{\text{X}}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

- Estimate $\mu_1 - \mu_2$ with $\overline{X}_1 - \overline{X}_2$.
- $\overline{X}_1 - \overline{X}_2$ is normally distributed

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \; ?$$

If both sample sizes are large, the sample variances are decent approximations for the true variances, so do the test as in the last Lesson. (approximate Z-test)

- Estimate $\mu_1 - \mu_2$ with $\overline{X}_1 - \overline{X}_2$.
- $\overline{X}_1 - \overline{X}_2$ is normally distributed

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim ?$$

If at least one sample size is small, the sample variances are not great approximations for the true variances.

- Suppose that $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ is a random sample of size n_1 from the normal distribution with mean μ_1 and variance σ_1^2.

- Suppose that $X_{2,1}, X_{2,2}, \ldots, X_{2,n}$ is a random sample of size n_2 from the normal distribution with mean μ_2 and variance σ_2^2.

- Suppose that σ_1^2 and σ_2^2 are unknown and that the samples are independent.

- Suppose that σ_1^2 and σ_2^2 are equal!

- Since we are assuming that $\sigma_1^2 = \sigma_2^2$, there is no need for subscripts.

- Call the common value σ^2.

- We have two sample variances, S_1^2 and S_2^2 that we would like to combine into a single estimator for σ^2.

- Call the combined estimator a pooled variance and denote it by S_p^2.

Pooled Variance

How about

$$S_p^2 = \frac{S_1^2 + S_2^2}{2} \quad ?$$

We <u>won't</u> use this because:

- If one sample variance is from a larger sample, we'd like to give it more weight.

- The distribution…?

Pooled Variance

Define

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Note that

$$\chi^2(n_2 - 1)$$

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

$$\chi^2(n_1 - 1)$$

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim N(0, 1)$$

$$=$$

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2}}$$

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p^2}}$$

$$= \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2}} \cdot \sqrt{\frac{\sigma^2}{S_p^2}}$$

$$= \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2}} \Bigg/ \sqrt{\frac{S_p^2}{\sigma^2}}$$

N(0, 1)

$$= \frac{\dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2}}}{\sqrt{\dfrac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \Big/ (n_1 + n_2 - 2)}}$$

$\chi^2(n_1 + n_2 - 2)$

So,

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p^2}} \sim t(n_1 + n_2 - 2)$$

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

Step One:

Choose an estimator for $\theta = \mu_1 - \mu_2$.

$$\widehat{\theta} = \overline{X}_1 - \overline{X}_2$$

Step Two:

Give the "form" of the test.

Reject H_0, in favor of H_1 if either

$$\widehat{\theta} > c \qquad \text{or} \qquad \widehat{\theta} < -c$$

for some c to be determined.

$$H_0 : \mu_1 - \mu_2 = 0$$

Step Three:

$$H_1 : \mu_1 - \mu_2 \neq 0$$

$$\alpha = \text{P(Type I Error)}$$

$$= \text{P(Reject } H_0 \, ; \theta = 0 \,)$$

$$= \text{P}(\overline{X}_1 - \overline{X}_2 > c \text{ or } \overline{X}_1 - \overline{X}_2 < -c ; \theta = 0 \,)$$

$$= 1 - \text{P}(-c \leq \overline{X}_1 - \overline{X}_2 \leq c \, ; \underbrace{\theta = 0}_{\mu_1 - \mu_2 = 0} \,)$$

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Step Three:

$$= 1 - P(-c \leq \overline{X}_1 - \overline{X}_2 \leq c \; ; \theta = 0)$$

- **Subtract $\mu_1 - \mu_2$ (which is 0)**
- **Divide by**

$$\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}$$

Step Three:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

$$\alpha = 1 - P(\,-d \leq T \leq d\,)$$

where $T \sim t(n_1 + n_2 - 2)$

and

$$d = c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}$$

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Step Three:

$$P(-d \leq T \leq d) = 1 - \alpha$$

$$\Rightarrow \quad d = t_{\alpha/2, n_1 + n_2 - 2}$$

$$\Rightarrow \quad c = t_{\alpha/2, n_1 + n_2 - 2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}$$

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Step Four:

Conclusion:

Reject H_0, in favor of H_1, if

$$\overline{X}_1 - \overline{X}_2 > t_{\alpha/2, n_1+n_2-2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}$$

or

$$\overline{X}_1 - \overline{X}_2 < - t_{\alpha/2, n_1+n_2-2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}$$

Example:

Fifth grade students from two neighboring counties took a placement exam.

- Group 1, from County A, consisted of 8 students. The sample mean score for these students was 77.2 and the sample variance is 15.3.

- Group 2, from County B, consisted of 10 students and had a sample mean score of 75.3 and the sample variance is 19.7.

Example:

From previous years of data, it is believed that the scores for both counties are normally distributed.

Can we say that the true means for Counties A and B are different?

Test the relevant hypotheses at level 0.01.

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$

$$n_1 = 8 \qquad n_1 = 10$$

$$\bar{x}_1 = 77.2 \qquad \bar{x}_1 = 75.3$$

$$s_1^2 = 15.3 \qquad s_2^2 = 19.7$$

$$\alpha = 0.01 \qquad t_{0.005,16} = 2.92$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$= 17.775$$

$$\bar{x}_1 - \bar{x}_2 = 77.2 - 75.3 = 1.9$$

$$t_{\alpha/2, n_1+n_2-2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_P^2}$$

$$= 2.92 \sqrt{\left(\frac{1}{8} + \frac{1}{10}\right)(17.775)}$$

$$= 5.840$$

Since $\bar{x}_1 - \bar{x}_2 = 1.9$ is not

- above 5.840, or
- below -5.840

we fail to reject H_0, in favor of H_1 at 0.01 level of significance.

The data do not indicate that there is a significant difference between the true mean scores for counties A and B.