- 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 one or both sample sizes are small. ($n_1 \leq 30$ and/or $n_2 \leq 30$)

$\boxed{\text{Goal:}}$ Find a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

Huge Assumption: $\sigma_1^2 = \sigma_2^2$

Step One:

An estimator: $\overline{X}_1 - \overline{X}_2$

Step Two:

Distribution of the estimator:

- $\overline{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$
- $\overline{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$

Step Two:

Distribution of the estimator:

- $\overline{X}_1 - \overline{X}_2$ is normally distributed

Mean:

$$E\left[\overline{X}_1 - \overline{X}_2\right] = E[\overline{X}_1] - E[\overline{X}_2]$$

$$= \mu_1 - \mu_2$$

Step Two:

Distribution of the estimator:

- $\overline{X}_1 - \overline{X}_2$ is normally distributed

Variance:

$$\mathrm{Var}\left[\overline{X}_1 - \overline{X}_2\right] = \mathrm{Var}[\overline{X}_1 + (-1)\overline{X}_2]$$

$$= \mathrm{Var}[\overline{X}_1] + \mathrm{Var}[(-1)\overline{X}_2]$$

$$= \mathrm{Var}[\overline{X}_1] + (-1)^2\mathrm{Var}[\overline{X}_2]$$

$$= \mathrm{Var}[\overline{X}_1] + \mathrm{Var}[\overline{X}_2]$$

Step Two:

Distribution of the estimator:

- $\overline{X}_1 - \overline{X}_2$ is normally distributed

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

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

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

We have S_1^2 and S_2^2 which are two independent estimators of the common variance σ^2.

How can we combine them?

Average them? No.

Sample Info:

- Sample of size n_1 from $N(\mu_1, \sigma_1^2)$ with $\overline{X}_1$ and S_1^2 reported.

- Sample of size n_2 from $N(\mu_1, \sigma_1^2)$ with $\overline{X}_2$ and S_2^2 reported.

Assume that $\sigma_1^2 = \sigma_2^2$.

Call the common value σ^2.

Use a weighted average that gives more weight to the one from the larger sample.

Pooled Variance:

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

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

- **We know that**

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1)$$

and

$$\frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$$

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

independent

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

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

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

$$\Downarrow$$

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

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

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

A N(0,1) divided by the square root of a χ^2 divided by its degrees of freedom!

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

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

Sample Info:

- Sample of size n_1 from $N(\mu_1, \sigma_1^2)$ with $\overline{X}_1$ and S_1^2 reported.

- Sample of size n_2 from $N(\mu_1, \sigma_1^2)$ with $\overline{X}_2$ and S_2^2 reported.

Assume that $\sigma_1^2 = \sigma_2^2$.

Call the common value σ^2.

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

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

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

Example: 90% CI for $\mu_1 - \mu_2$

$$n_1 = 9,\ \ \bar{x}_1 = 23.2,\ \ s_1^2 = 4.3$$

$$n_2 = 8,\ \ \bar{x}_2 = 24.7,\ s_2^2 = 5.2$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 4.72$$

$$t_{0.05,15} = 1.753$$

In R: qt(0.095,15)

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

$$23.2 - 24.7 \pm 1.753 \sqrt{4.72 \left(\frac{1}{9} + \frac{1}{8} \right)}$$

$$(-3.351, 0.351)$$

What if we can't say that σ_1^2 is equal to σ_2^2?

This is hard and is known as the **Behrens-Fisher problem**.

Welch's Approximation:

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\text{approx}}{\sim} t(\nu)$$

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

In R:

```
x<-rnorm(10)
y<-rnorm(14)
t.test(x,y,conf.level=0.90)
```

```
> x<-rnorm(14)
> x<-rnorm(10)
> y<-rnorm(14)
> t.test(x,y,conf.lev=0.90)

        Welch Two Sample t-test

data:  x and y
t = 0.030583, df = 16.916, p-value = 0.976
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -0.6289463  0.6514495
sample estimates:
 mean of x  mean of y
-0.2765715 -0.2878232

>
```