# Inference for Difference of Means

What if we want to know $\mu_1 - \mu_2$

mean of some variable for 1 population

The mean for different pop^n

If we have matched pairs, $X_1$ and $X_2$ for a sample come from same individual

We can often consider the difference between values for a single individual as one observation.

We don't know $\mu$, but we have $\bar{X}$ from samples

$\bar{X}_1 \quad \bar{X}_2$

$\bar{X} \quad \bar{Y}$

$\bar{X}_1 - \bar{X}_2$ — This is random, subject to sampling variability

If $n_1$ and $n_2$ large enough

CLT

$\bar{X}_1 \sim N(\mu_1, \sigma_1/\sqrt{n_1}) \quad \bar{X}_2 \sim N(\mu_2, \sigma_2/\sqrt{n_2})$

Resembles $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ $\quad \left(\sqrt{p_1(1-p_1)}\right)^2$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Problem we don't know $\sigma_1$ or $\sigma_2$.

Estimate SE with $s_1$ and $s_2$

We do have $s_1$ and $s_2$

Sample SD from each sample

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim T \text{ distribution}$$

defined by degrees of freedom

$$df = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Big/ \left\{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}\right\}$$

Conserative estimate

$df = \min(n_1 - 1, n_2 - 1)$

We're going to use this in R....

Barbados Malnutrition Study

$df = \min(n_1 - 1, n_2 - 1)$ Use this in R....

# Barbados Malnutrition Study

52 children, about half of which experienced adolescent malnutrition, and wanted to know if scores on vocabulary tests differed for these two halves.

**Aside**

Similar to the case where $p_1 = p_2$ we calculate $p_{pooled} = \frac{\hat{p}_1 \cdot n_1 + \hat{p}_2 \cdot n_2}{n_1 + n_2}$

$\hookrightarrow SE_{pooled} = \sqrt{\frac{p_p(1-p_p)}{n_1} + \frac{p_p(1-p_p)}{n_2}}$

Similarly, if we believe $\sigma_1 = \sigma_2$

$$S_{pooled}^{x,y} = \frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{(n_x - 1) + (n_y - 1)}$$

$$SE_{pooled} = \sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}}$$

## Vocab Scores

Hospitalized at age < 1 with grade II or III protein energy malnutrition

| Malnutrition | Control |
|---|---|
| $n_m = 25$ | $n_c = 27$ |
| $\bar{X}_m = 38.03$ | $\bar{X}_c = 48.81$ |
| $S_m = 11.62$ | $S_c = 11.12$ |

$$SE = \sqrt{\frac{S_m^2}{n_m} + \frac{S_c^2}{n_c}} = \sqrt{\frac{11.62^2}{25} + \frac{11.12^2}{27}} \approx 3.159$$

## 95% CI

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{SE_{est}} \sim T_{df = \min(n_1 - 1, n_2 - 2)}$$

$$\mu_1 - \mu_2 : (\bar{X}_1 - \bar{X}_2) \pm T_{df=\dots} \cdot SE$$

$$(\bar{X}_c - \bar{X}_m) \pm \overset{2.064}{T_{df=24}} \cdot SE$$

$$(48.81 - 38.03) \pm 2.064 \cdot 3.159$$

$$(4.260, 17.300)$$

We are 95% confident that the difference $\mu_c - \mu_m$ is between $(4.26, 17.3)$

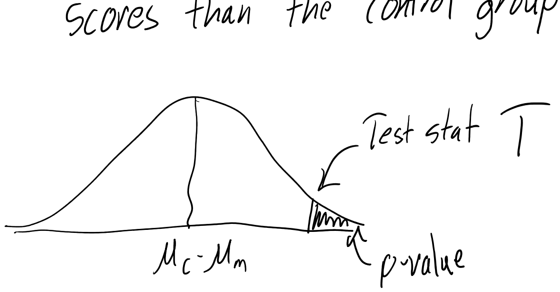
## Hypothesis Test

$H_0: \mu_c - \mu = 0 ; \mu_c = \mu_m$

There is no difference in average test scores for these two groups.

$H_a: \mu_c - \mu_m > 0 ; \mu_c > \mu_m$

The malnutrition group has lower average scores than the control group.




$$T_{df=24} = \frac{(\bar{X}_c - \bar{X}_m) - \cancel{(\mu_c - \mu_m)}^{\,0}}{\dots} = \frac{48.81 - 38.03}{3.159} = 3.412$$

$$T_{df=24} = \frac{(X_c - X_m) - \cancel{(\mu_c - \mu_m)}}{SE} = \frac{48.81 - 58.05}{3.159} = 3.412$$

P-value = 

$\rightarrow$ From R $\approx 0.001145$
or $.1145\%$

If $H_0$ is true, the observed data (or something more extreme) would be observed in $0.1145\%$ of samples. This is much less than standard $\alpha = 5\%$.

We reject $H_0$ and conclude $H_a$.

There is evidence that the malnutrition group scores lower on average than the control group.