# Lecture 6

9/17/2020

DGP: $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\theta, \sigma^2)$

$$\underbrace{\hat{\theta} = \bar{X} \sim N(\theta, (\tfrac{\sigma}{\sqrt{n}})^2)}_{SE[\bar{X}]} \iff \underbrace{\frac{\bar{X}-\theta}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)}_{\text{standardize the estimator}} \quad \text{(exactly)}$$

What if $\sigma$ is unknown?

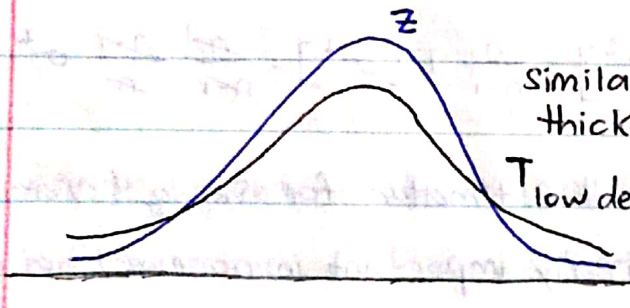
$S^2$ estimates $\sigma^2 \Rightarrow S = \sqrt{S^2}$ estimates $\sigma$.

Does $\dfrac{\bar{X}-\theta}{\frac{S}{\sqrt{n}}} \overset{?}{\sim} N(0,1)$ No! But close!

In 1907 Gosset proved:

$$\frac{\bar{X}-\theta}{\frac{S}{\sqrt{n}}} \sim T_{n-1} \quad \text{(exactly)}$$

Student's standard T distribution with $n-1$ "degrees of freedom" (the parameter for the standard T distr.).




Similar to normal but with thicker tails

T low degree of freedom < 20

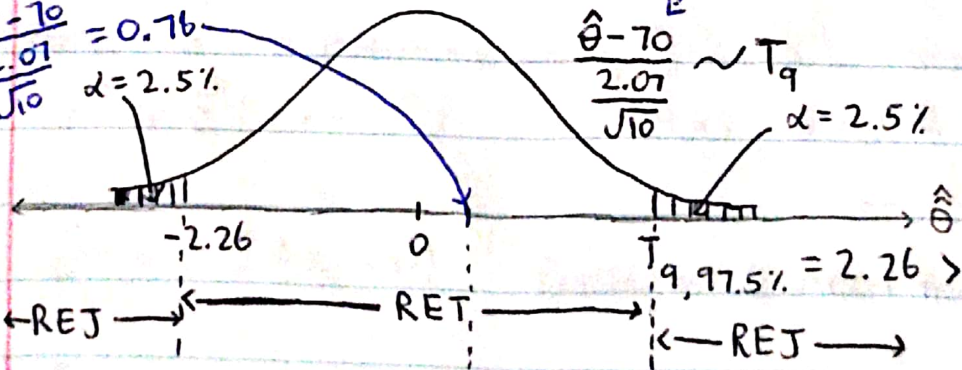
Data from $n=10$ male student heights: $\bar{X} = 70.5$, $S = 2.07$

$H_a: \theta \neq 70$, $H_0: \theta = 70$, $\alpha = 5\%$

$$\frac{\hat{\theta}-70}{\frac{S}{\sqrt{n}}} = \frac{70.5-70}{\frac{2.07}{\sqrt{10}}} = 0.76$$

the standardized distribution of $\hat{\theta} \mid H_0$

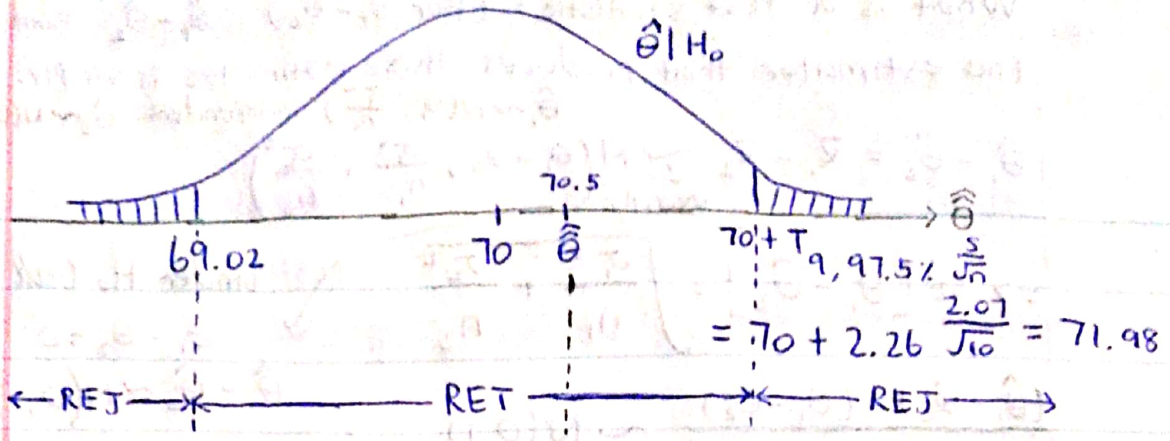
$$\frac{\hat{\theta}-70}{\frac{2.07}{\sqrt{10}}} \sim T_9$$




$T_{9,97.5\%} = 2.26 > 1.96 = Z_{97.5\%}$

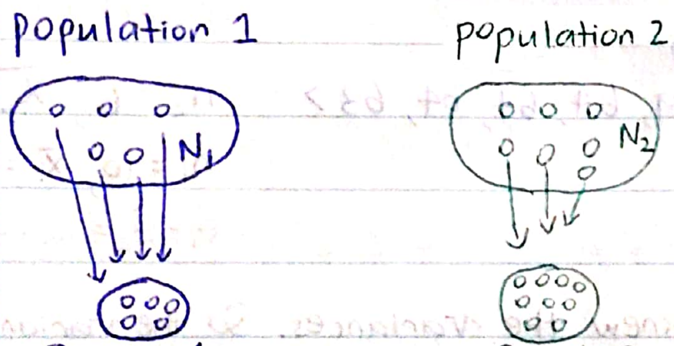
$\Rightarrow$ Retain $H_0$

Scanned with CamScanner




$\Rightarrow$ Retain $H_0$

We just did our first "one-sample two sided T test" (of a mean)




Sample 1 size $n_1$

$x_{11}, x_{12}, \ldots, x_{1n_1}$

Sample 2 size $n_2$

$x_{21}, x_{22}, \ldots, x_{2n_2}$

$N_1 \approx \infty$, $N_2 \approx \infty \Rightarrow$ i.i.d.

Assume $X_{11}, \ldots, X_{1n_1} \overset{i.i.d.}{\sim} N(\theta_1, \sigma_1^2)$

independent of

$X_{21}, \ldots, X_{2n_2} \overset{i.i.d.}{\sim} N(\theta_2, \sigma_2^2)$

$\sigma_1^2, \sigma_2^2$ are known but $\theta_1, \theta_2$ are unknown

There are 3 types of tests that are usually done.

i) $H_a: \theta_1 \neq \theta_2 \Rightarrow H_0: \theta_1 = \theta_2$ equivalent...

$H_a: \theta_1 - \theta_2 \neq 0 \Rightarrow H_0: \theta_1 - \theta_2 = 0$

ii) $H_a: \theta_1 < \theta_2 \Rightarrow H_0: \theta_1 \geq \theta_2$ equivalent...

$H_a: \theta_1 - \theta_2 < 0 \Rightarrow H_0: \theta_1 - \theta_2 \geq 0$

iii) $H_a: \theta_1 > \theta_2 \Rightarrow H_0: \theta_1 \leq \theta_2$ equivalent...

$H_a: \theta_1 - \theta_2 > 0 \Rightarrow H_0: \theta_1 - \theta_2 \leq 0$

What is a test statistic? (For $\theta_1 - \theta_2$) $\hat{\theta}_1 - \hat{\theta}_2$ Point estimate.

The estimator that produces these estimates is simply:

$\hat{\theta}_1 \sim N(\theta_1, \frac{\sigma_1^2}{n_1})$ independent $\hat{\theta}_2 \sim N(\theta_2, \frac{\sigma_2^2}{n_2})$

$$\hat{\theta}_1 - \hat{\theta}_2 = \bar{X}_1 - \bar{X}_2 \underset{\text{exactly}}{\sim} N\left(\theta_1 - \theta_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$\Rightarrow SE[\hat{\theta}_1 - \hat{\theta}_2] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

under $H_0$ (all 3)

$\theta_1 - \theta_2 = 0$

$$\hat{\theta}_1 - \hat{\theta}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \longrightarrow \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Standardized distr under $H_0$

Let's test if male mean height is different than female mean height.

$\bar{X} = \langle 60, 59, 64, 64, 64, 63 \rangle$ $\quad n_2 = 6, \ \bar{X}_2 = 62.3$

$n_1 = 10, \ \bar{X}_1 = 70.5$

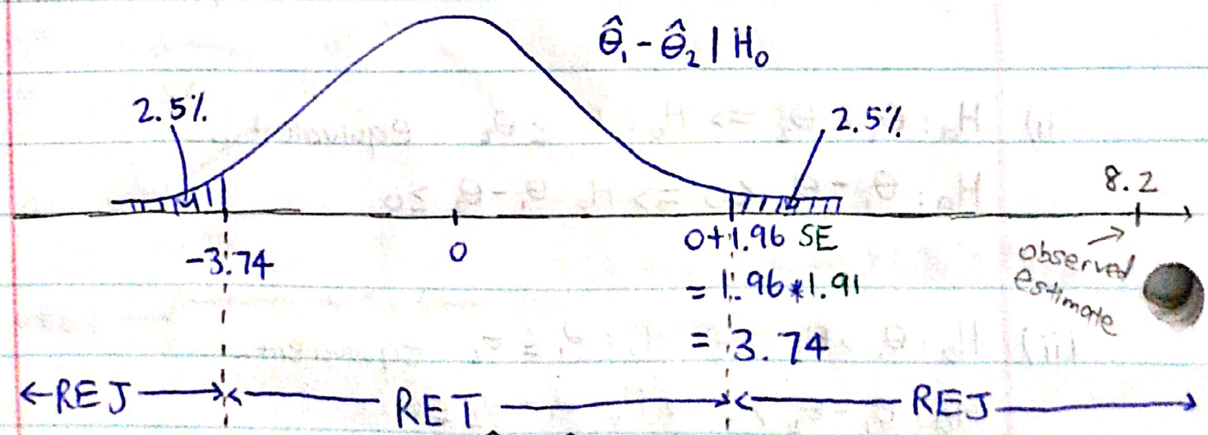
$\hat{\theta}_1 - \hat{\theta}_2 = 70.5 - 62.3 = 8.2$

We assumed we knew the variances. So the variance for the men was assumed to be $4^2$ and now the variance for the women is assumed to be $3.5^2$.

$\sigma_1^2 = 4^2, \ \sigma_2^2 = 3.5^2, \ \alpha = 5\%$ $\quad SE = \sqrt{\frac{4^2}{10} + \frac{3.5^2}{6}}$

$= 1.91$

We can now do our 2 sample 2-sided z test.




Observed estimate $\hat{\theta}_1 - \hat{\theta}_2 \notin RET \Rightarrow$ REJECT $H_0$

$$\text{P-value} = 2P(\hat{\theta}_1 - \hat{\theta}_2 > 8.2) = 2P\left(\frac{\hat{\theta}_1 - \hat{\theta}_2}{SE} > \frac{8.2}{1.91}\right) = 2P(Z > 4.29)$$
$$= 1.8 \times 10^{-5} < \alpha$$
$$\Rightarrow \text{Reject}$$

Let's sample from two populations again however, this time we have the same variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$ which we still assume known.

$$X_{11}, \ldots, X_{1n_1} \overset{\text{i.i.d.}}{\sim} N(\theta_1, \sigma^2) \text{ independent of } X_{21}, \ldots, X_{2n_2} \overset{\text{i.i.d.}}{\sim} N(\theta_2, \sigma^2)$$

Under $H_0$, $\hat{\theta}_1 - \hat{\theta}_2 \sim N\left(0, \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}^{\,2}\right)$

also $\dfrac{\hat{\theta}_1 - \hat{\theta}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$

The test can be run again, you can probably assume $\sigma = 3.75$.

Same as above but $\sigma$ is unknown. How can we estimate the standard error?

$S_1^2$, $S_2^2$ are the sample variances in both samples 1 and 2.

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1,i} - \bar{X}_1)^2, \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2,i} - \bar{X}_2)^2$$

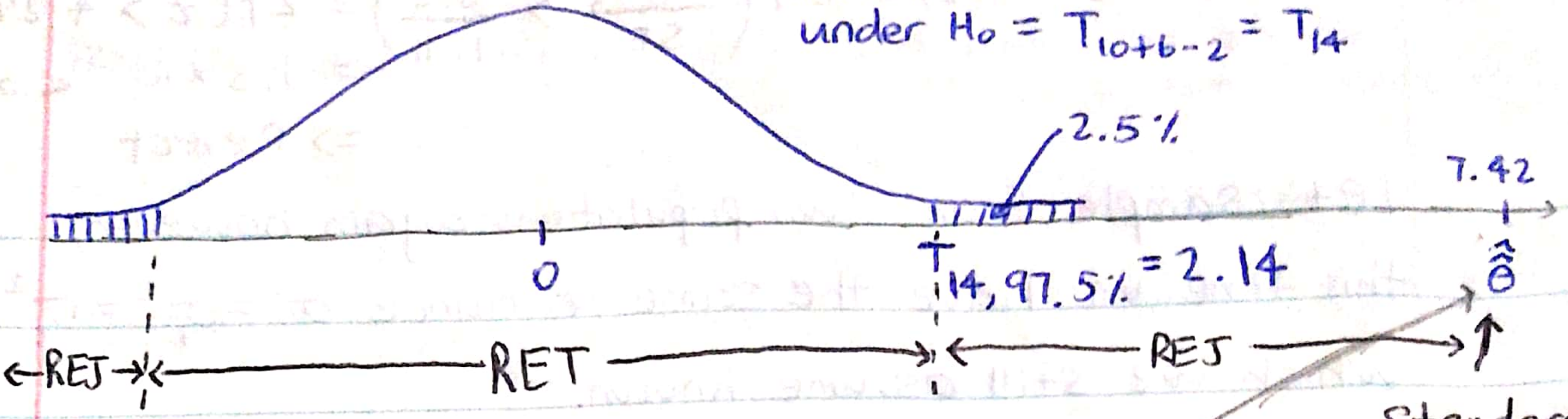
$$S^2_{\text{pooled}} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad \text{weighted average}$$

You can prove that

$$\frac{\hat{\theta}_1 - \hat{\theta}_2}{S_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T_{n_1 + n_2 - 2}$$

this allows you to do the "2-Sample T test of equal variance"

Standard estimator under $H_0 = T_{10+6-2} = T_{14}$




Standardized observed estimate

$$\hat{\theta} = \frac{70.5 - 62.3}{S_{pooled}\sqrt{\frac{1}{10} + \frac{1}{6}}} = \frac{8.2}{(2.14)(0.51)} = 7.42$$

$$S_{pooled}^2 = \frac{(10-1)(2.07)^2 + (6-1)(2.25)^2}{10+6-2} = \frac{9(2.07)^2 + 5(2.25)^2}{14}$$

$$= 4.56$$

$\Rightarrow$ Reject $H_0$

Scanned with CamScanner