Previously on 4SSPP109…

- Estimating population mean & variance from a random sample
- Formulating a hypothesis about a population.
  - $H_0: E(Y) = \mu_{Y,0}$
  - $H_1: E(Y) \neq \mu_{Y,0}$

Do students dislike ice-cream?

- H_0: most students at King's dislike ice-cream
- H_1: the above is not true.
- Random sample of 1,000 King's students.
  - only 1 dislikes ice-cream.

> H_0 is almost surely false! (reject H_0)

Do average earnings of recent graduates equal 20£/hour?

- $H_0 : E(Y) = 20$
- $H_1 : E(Y) \neq 20$
- In your random sample (n=200), $\bar{Y} = 22.64$.
- How do you use the sample data to **reject** or **not reject** H_0 ?
  - *test statistics*
    - *Rejection (or critical) region*
  - *p-value*
    - *Level of significance*

P-value: formal definition

- p-value = $Pr_{H_0}\left[\left|\bar{Y} - \mu_{Y,0}\right| \geq \left|\bar{Y}^{act} - \mu_{Y,0}\right|\right]$

"Probability under the null hypothesis…

…that the difference between the sample mean and the null hypothesis…

…is at least as large as the one we obtained."

- Low p-value → null hypothesis is *probably* wrong.
- High p-value → *cannot reject* the null hypothesis.

How to calculate the p-value

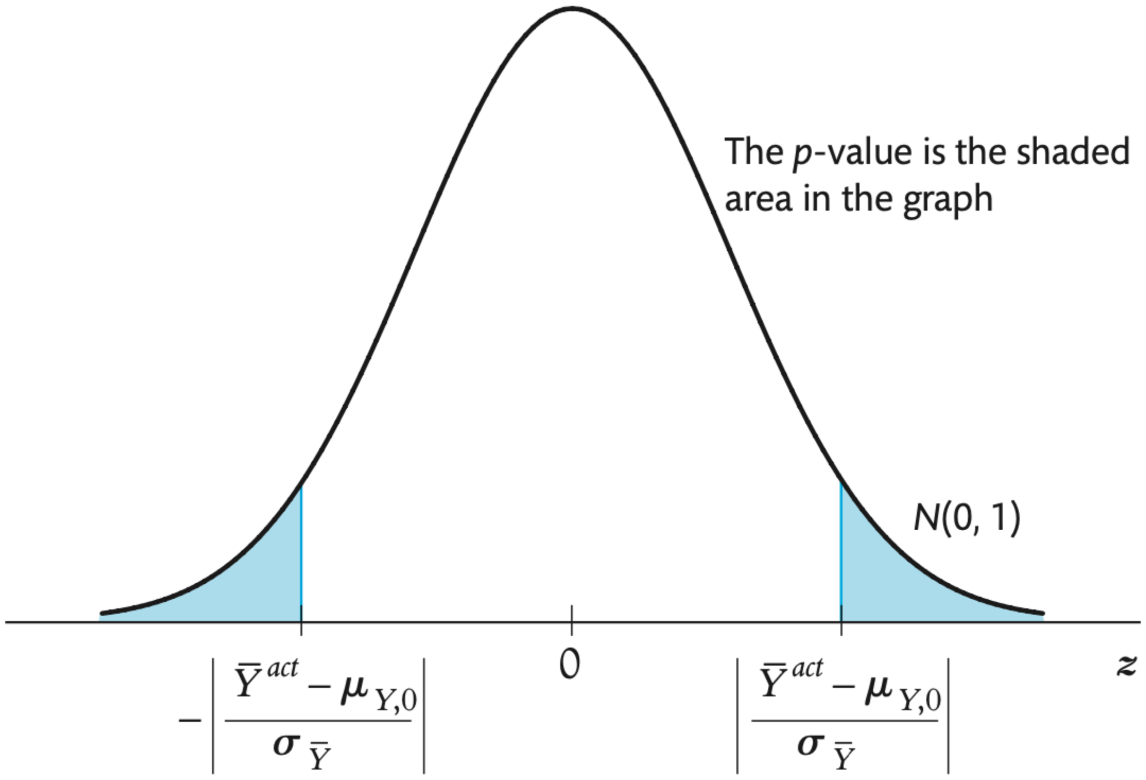
- With large *n*, assuming H_0 true:

$$\bar{Y} \sim N(\mu_{Y,0}, \sigma_{\bar{Y}}^2)$$

$$\rightarrow \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}} \sim N(0,1)$$

called t-statistics

- *p-value* = probability of obtaining a t-stat as far from 0 as you obtained
- *p-value*=probability that a N(0,1) RV falls as far as $|t|$ from zero
- *p-value*= $2\Phi(-|t|)$



The Standard Error of $\bar{Y}$

- We need $\sigma_{\bar{Y}}$ to compute t-stat & p-value.
- We know that $\sigma_{\bar{Y}} = \frac{1}{\sqrt{n}} \sigma_Y$
- We can estimate it using $\hat{\sigma} = \frac{1}{\sqrt{n}} s_Y$
- Called *standard error* of $\bar{Y}$: $SE(\bar{Y}) = \hat{\sigma} = \frac{1}{\sqrt{n}} s_Y$
- $SE(\bar{Y})$ measures the *precision* of $\bar{Y}$ as an estimate of μ_Y

Computing the p-value in practice

1. Compute sample mean ($\bar{Y}^{act}$) & sample SD (s_Y).
2. Compute $SE(\bar{Y}) = \frac{1}{\sqrt{n}} s_Y$
3. Compute t-stat $\mathrm{t} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}$
4. Write in STATA "display 2**normal(x)*" *where* $\mathrm{x} = -|t|$
   - because p-value = $2\Phi(-|t|)$
   - *(can also use Excel, or Table 6.1 in textbook)*

Calculating the p-value: an example

- We have wages for a sample of 200 recent graduates
- $H_o: \mu_Y = £20$
- In the sample, $\bar{Y}^{act} = £22.64; s_Y = £18.14$
- YOUR TURN - Calculate:
  1. $SE(\bar{Y})$,
  2. t-stat

  (we then compute p-value together)

Remember:

- $SE(\bar{Y}) = \hat{\sigma} = \frac{1}{\sqrt{n}} s_Y$
- t-stat = $\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}$
- p-value = $2\Phi\ (-|t|)$

Calculating the p-value: an example

- We have wages for a sample of 200 recent graduates
- $H_o: \mu_Y = £20$
- In the sample, $\bar{Y}^{act} = £22.64; s_Y = £18.1$

- $SE(\bar{Y}) = \hat{\sigma} = \frac{1}{\sqrt{n}} s_Y = \frac{18.14}{\sqrt{200}} = 1.28$
- t-stat = $\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})} = \frac{22.64-20}{1.28} = 2.06$
- p-value = $2\Phi\left(-|t|\right) = 2 * 0.0197 = 0.0394$

Accept or reject H_0?

Significance level

- How low should the p-value be, for us to reject the null hypothesis?

- Convention in social sciences: 0.05 (or 5%)

$$\text{Reject } H_0 \text{if } \mathrm{p} < 0.05$$

  - *0.05 (or 5%) significance level*
  - sometimes denoted as α

- max probability of a *type-I error* (= falsely rejecting the null) we are willing to accept

t is our test statistics!

- We reject the null based on the value of t.

- We reject the null if

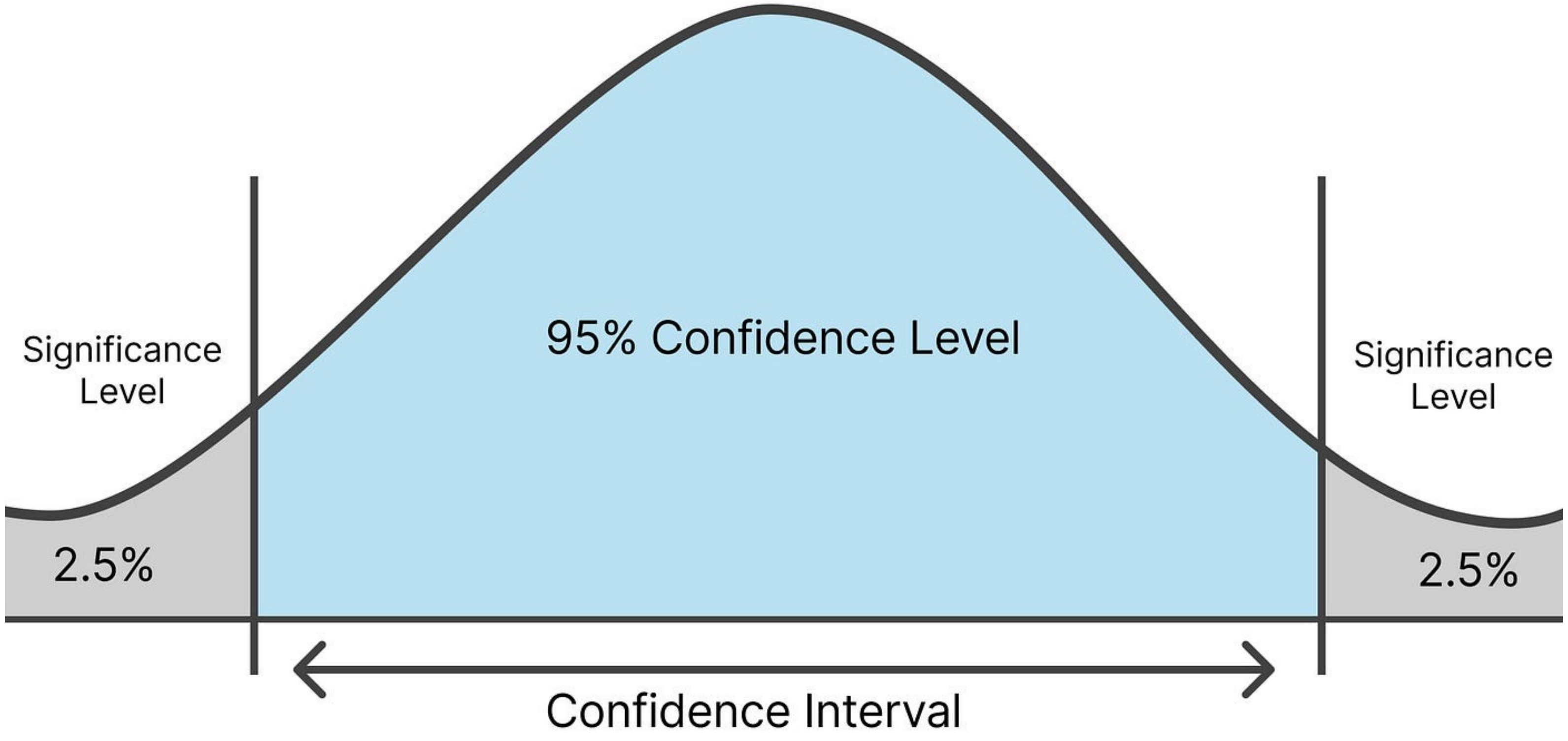
$$p = 2\Phi\left(-|t|\right) < \alpha$$

- With $\alpha = 0.05$, this means we reject the null if

$$|t| > 1.96$$

  - this is our *rejection region*!

3. Confidence intervals



Confidence intervals

- 95% confidence interval: a range of values that is 95% likely to include the population mean.

- The set of all values for μ_Y that we *cannot* reject at the 5% significance level.

- 95% confidence interval for μ_Y:

$$\bar{Y} - 1.96 * SE(\bar{Y}) \leq \mu_Y \leq \bar{Y} + 1.96 * SE(\bar{Y})$$

Confidence intervals

YOUR TURN: *Compute 95% confidence interval for hourly earnings*

- In the sample, $\bar{Y}^{act} = \$22.64;\ \mathrm{SE}(\bar{Y}) = 1.28$

Reminder: a 95% confidence interval for μ_Y is:

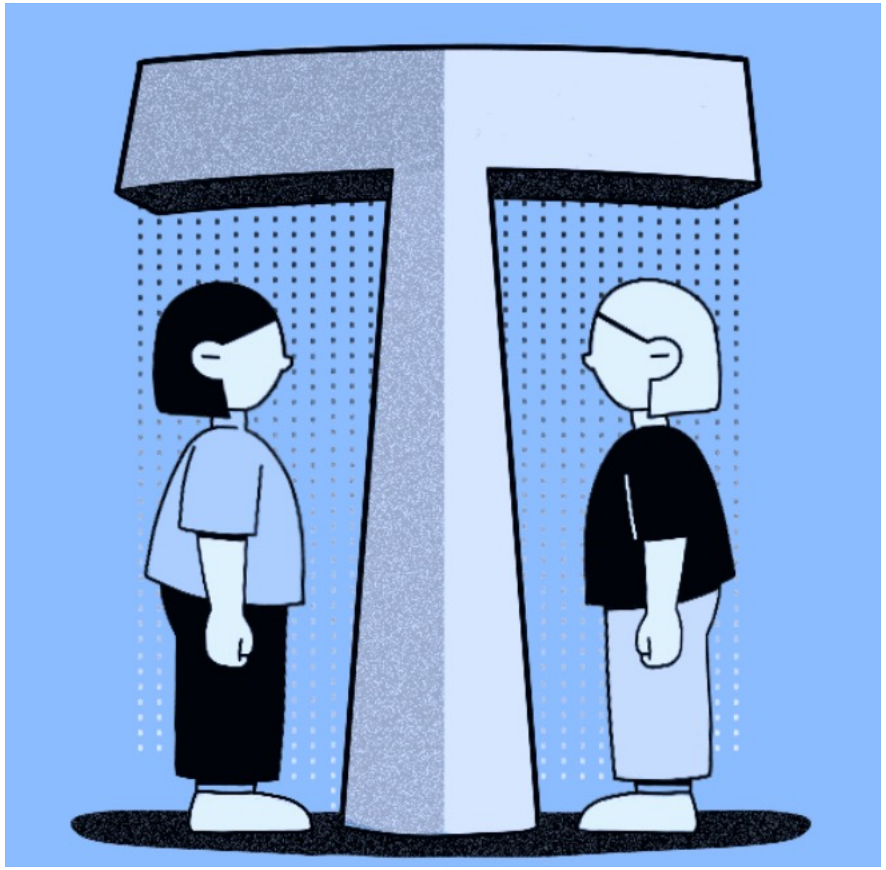
$$\bar{Y} - 1.96 * SE(\bar{Y}) \leq \mu_Y \leq \bar{Y} + 1.96 * SE(\bar{Y})$$

Confidence intervals

YOUR TURN: Calculate a 95% confidence interval for hourly earnings

- In the sample, $\bar{Y}^{act} = \$22.64;\ \mathrm{SE}(\bar{Y}) = 1.28$
- Upper bound: $\bar{Y} + 1.96 * SE(\bar{Y})$ = 22.64 + 1.96∗1.28 = 25.15
- Lower bound: $\bar{Y} - 1.96 * SE(\bar{Y})$ = 22.64 − 1.96∗1.28 = 20.13
- $20.13 \leq \mu_Y \leq 25.15$

4. Testing differences in means



TESTING DIFFERENCES BETWEEN MEANS

- $H_0 : \mu_m - \mu_w = d_0$ vs. $H_1 : \mu_m - \mu_w \neq d_0$
- $E(\bar{Y}_m - \bar{Y}_w) = \mu_m - \mu_w$
- $(\bar{Y}_m - \bar{Y}_w) \sim N(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w})$
- $SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$
- $t = \frac{(\bar{Y}_m - \bar{Y}_w) - d_0}{SE(\bar{Y}_m - \bar{Y}_w)} \rightarrow p - value = 2\Phi(-|t^{act}|)$

Thank you for your attention