

1. Suppose that you repeatedly take random samples from a population with slope coefficient β and that from each sample you calculate a 96% t-confidence interval for β .

- (a) Would all of the intervals have the same width (difference between endpoints)? Explain briefly.

No. A CI for the slope parameter is computed using: $b \pm t_{n-2}SE_b$. Because the SE_b is calculated from each sample, each CI will have a different SE_b and therefore a different width.

- (b) Is it true that in the long run 96% of your intervals will contain the sample slope coefficient, b ? Explain briefly.

No, every interval will contain b !!! (see CI equation above) However, we'd expect 96% of the intervals to contain β , the true population slope parameter.

2. Consider a population of 2328 people who smoke. The average age that they started smoking is 18.18 years and with a standard deviation of 5.39 years. Consider taking a random sample of 40 people from this population.

- (a) What is the variable of interest? Is it categorical or quantitative?
Age at which people started smoking. It is a quantitative variable.

- (b) What is the sampling distribution for the sample mean (of the 40 people)? Explain what the sampling distribution represents in words. Additionally, write out the explicit information of the sampling distribution. (Check any assumptions you need.)

The sampling distribution represents all possible values and associated probabilities of \bar{X} values when taking samples of size 40. Because we have a reasonably large sample size, we don't have to worry about the normality of the underlying population.

$$\bar{X} \sim N(18.18\text{yrs}, 5.39/\sqrt{40} = 0.852\text{yrs})$$

- (c) Given the characteristics of the population, find the probability that the average age of your sample of 40 is smaller than 17.5 years old.

$$\begin{aligned} P(\bar{X} < 17.5) &= P\left(\frac{\bar{X} - 18.18}{0.852} < \frac{\bar{X} - 18.18}{0.852}\right) \\ &= P(Z < -0.797) = 0.213 \end{aligned}$$

- (d) Now consider a sample of size 40 such that: $\bar{X} = 17.95$ years, $s = 6.98$ years. **Assume for parts (d) and (e) that you don't know anything about the**

population. Test the claim that the average age of first smoking is greater than 17.5 years old.

$$\begin{aligned} \mu &= \text{true average age of initial smoking} \\ H_0 &: \mu \leq 17.5 \text{years} \\ H_A &: \mu > 17.5 \text{years} \\ t^* &= \frac{17.95 - 17.5}{6.98/\sqrt{40}} = 0.41 \\ p\text{-value} &= P(t_{39} > 0.41) > 0.1 \end{aligned}$$

We cannot reject H_0 . We do not have enough evidence to claim that the average initial age of smoking is greater than 17.5 years old.

- (e) Is it possible that you made an error? If so, which type did you make? Specify the type of error you might have made in terms of the problem (i.e., smoking).

In hypothesis testing, it is always possible that an error has been made. Here, the error would be (and is!) that we didn't reject H_0 but we should have. This is a type II error. We did not have enough power to show that the average age of smoking is greater than 17.5 years old.

3. What does taking a larger sample size do to (short answers, no explanation necessary):

- (a) The size of the hypothesis test: nothing
- (b) The power of the hypothesis test: increases
- (c) The SE of the estimate of interest: decreases
- (d) The accuracy of the estimate of interest: increase only if we weren't taking random samples to start with. If we are taking random samples we expect to get the parameter value on average regardless of the sample size.
- (e) The width of the CI for the parameter of interest: decreases
- (f) The percent coverage of the parameter for a CI: nothing

4. An investigator collected data on heights and weights of college students. The correlation between height and weight for the men was about 0.6; for the women the correlation was also about 0.6. If you take the men and the women together, the correlation between height and weight would be _____

just about 0.6 somewhat lower somewhat higher

somewhat higher – think of the picture that has both sexes on it... the distance from the new average (\bar{x}, \bar{y}) will be bigger, and the correlation will be higher. Also, you can draw the lines at \bar{X} and \bar{Y} and see that you will have many more points in the “positive” quadrants when combining the data.

5. $X \sim N(1000, 50)$

- (a) If bricks are fired at a temperature above 1125° F, they will crack and must be disposed of. If the bricks are placed randomly throughout the kiln, what proportion of bricks will crack during the firing process?

$$P(X \geq 1125) = P\left(\frac{X-1000}{50} \geq \frac{1125-1000}{50}\right) = P(Z \geq 2.5) = 0.0062$$

- (b) If, after glazing many bricks, we notice that about 10% of them are discolored (due to low temperature), at what temperature are the bricks discoloring? (Again, assume the bricks are placed randomly throughout the kiln.)

$$\begin{aligned} P(X \leq ?) &= 0.1 \\ P(Z \leq \frac{? - 1000}{50}) &= 0.1 \\ \frac{? - 1000}{50} &= -1.28 \\ ? &= 936^\circ F \end{aligned}$$

6. Data were collected in an investigation of environmental causes of disease. One variable was the annual mortality rate per 100,000 for males, averaged over the years 1958-1964 (think average number of deaths per year) for a random sample of 61 large towns in England and Wales. Additionally, they recorded whether or not the town is north of Derby.

where	sample size	mean	std. dev.
North of Derby	35	1633.6	136.9
South of Derby	26	1376.8	140.3

Create a 98% confidence interval for the difference in average mortality rate per 100,000 males in towns north of Derby as compared to towns south of Derby. Remember to interpret the interval in terms of the problem.

μ_n = average mortality for the population north of Derby

μ_s = average mortality for the population south of Derby

$$s_p^2 = \frac{34 \cdot 136.9^2 + 25 \cdot 140.3^2}{35 + 26 - 2} = 138.35^2$$

$$SE(\bar{Y}_n - \bar{Y}_s) = s_p \sqrt{1/35 + 1/26} = 35.82$$

$$df = n_n + n_s - 2 = 35 + 26 - 2 = 59$$

$$t_{59,.01} = 2.39$$

$$CI: \bar{Y}_n - \bar{Y}_s \pm t_{59,.01} SE(\bar{Y}_n - \bar{Y}_s)$$

$$256.8 \pm 2.39 \cdot 35.82$$

$$(171.2, 342.4)$$

We are 98% confident that the true average difference in mortality rate between towns north of Derby compared with those south of Derby is between 171.2 and 342.4 men per 100,000 people.