Assignment 3

Andrew Shaw

2025-08-26

1. Come up with a mean from economics that you'd like to know more about (example: male to female wage gap). State your best guess for the null hypothesis.

What percentage of total energy consumption comes from renewable sources at data centers?

$H_0 : \mu = 0.40$

2. State the alternative hypothesis in 3 different ways.

$H_a : \mu < 0.40$

$H_a : \mu > 0.40$

$H_a : \mu \neq 0.40$

3. Can you accept your alternative hypothesis? Why or why not?

No, because are testing for H_0 and H_a cannot be proven true. If null is rejected, the absence of evidence for the null does not make the alternative automatically true.

4. You randomly draw a large sample of your desired mean and find that the sample mean is 2 times as large as you expected and suppose we know that the standard deviation is 3 times the expected mean. What is the probability that your expectation about the mean is correct?

$\frac{2\mu_{Y,0} - \mu_{Y,0}}{3\mu_{Y,0}} = 0.3333$

$2\phi(-|0.333|) = 0.7414$

5. What is your intuition for why dividing the difference between your sample and expected average by the standard deviation gives you a good approximation for how likely something is to occur?

The difference tells how far away the sample mean is from the hypothesized mean. The standard deviation standardizes this result in the z-score which can give a rough approximation of probability by seeing where the score falls on the curve and approximating how much area is in the tail or tails.

6. Describe the distinction between standard error and the standard deviation, and when can we use the standard error?

The standard deviation is the deviation from the sample mean of individual observations. The standard error is the deviation of the sample mean from the true population mean. Standard error can be used when determining the variability of the estimator whereas the standard deviation answers how variable the data is.

7. Suppose your sample size from 4 was 100. What is the 95% confidence interval for your sample in 4? Describe what this confidence interval tells you.

$SE(\bar{Y}) = \frac{3\mu_{Y,0}}{\sqrt{100}} = 0.3\mu_{Y,0}$ $\mu_{Y,0} \pm 1.96 * SE = [1.412\mu_{Y,0},\ 2.588\mu_{Y,0}]$

This states that with repeated samples, the value of the true population mean would fall within this interval 95% of the time.

8.Define the standard error in a comparison of a two-means sample group.

The standard error represents the combined variability of the difference of the two population means.

9. What happens to the standard error if one group remains small but the other grows large? Does it converge towards zero still?

No, it wouldn't converge toward 0 rather it would converge toward the standard error of the population group that is not growing.

10. Define some differences that you'd like to know more about in economics.

- Morning people are more productive
- Is Solar energy is cheaper than grid electricity?
- Do rural areas resist adopting renewable technologies?

11. How would you collect the sample to find the smallest standard error while holding the overall sample size constant?

Use precise tools and control for any influencing factors.

12. Why do you think a null hypothesis that the differences between two means is 0 is the most widely used hypothesis in science?

Two-sided tests have a wider range of probabilities and it's easier to determine whether populations are no different than try to test for some value of difference which has little reference.

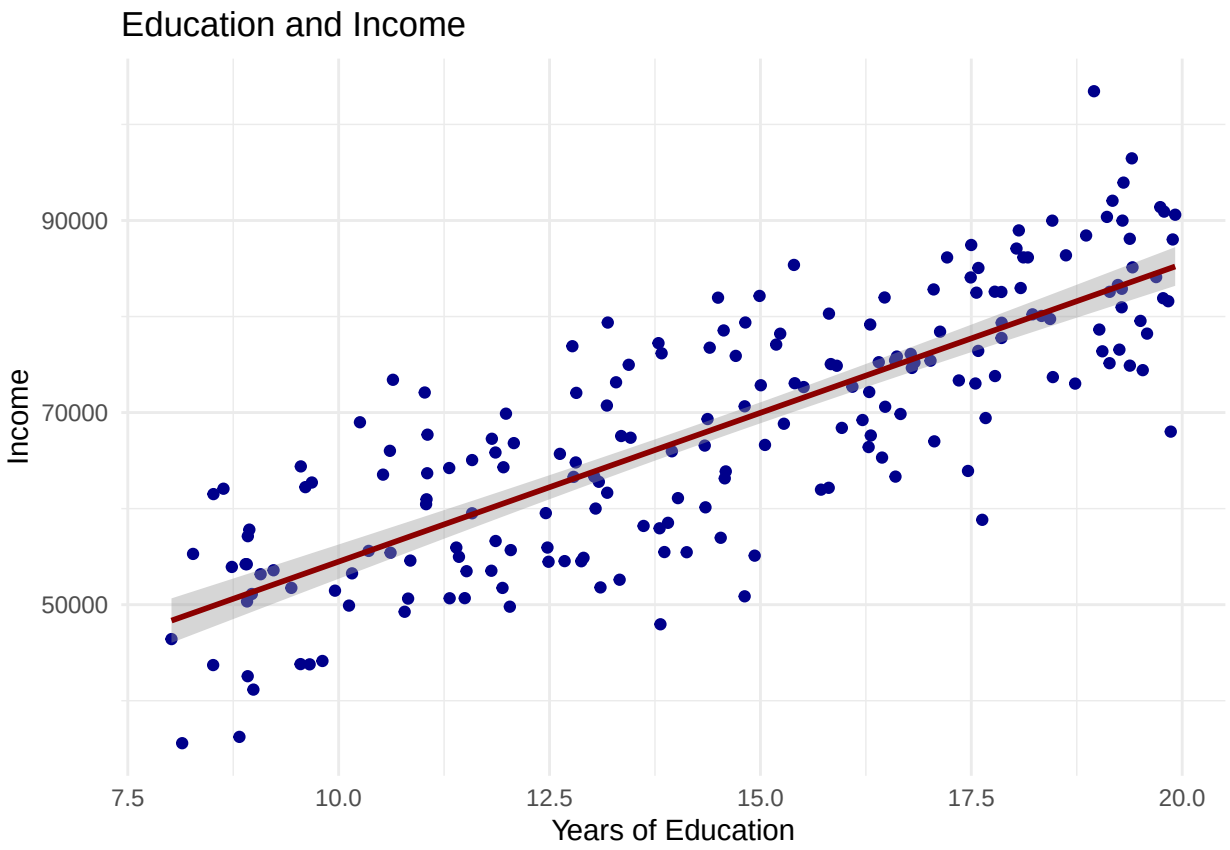
13. A scatterplot is a plot that shows what?

It shows the observations of a sample plotted by two of its dimensions on X and Y axis. Doing so gives a correlation of those dimensions of the data.

14. Draw a fictional scatterplot that relates earnings to years of education.

```
set.seed(456)

n <- 200

education <- runif(n, 8, 20)
# base salary of 25000 for minimal education + 3000 per year of education + variability
# to avoid uniformity
income <- 25000 + 3000 * education + rnorm(n,0, 8000)

df <- data.frame(education, income)

ggplot(df, aes(x=education, y = income)) +
  geom_point(color = 'darkblue') +
  geom_smooth(method="lm", se=TRUE, color = "darkred") +
  labs(title = 'Education and Income',
       x = "Years of Education",
       y = "Income") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



15. What does correlation measure in a scatterplot in linear samples? Is it the slope of the line?

Correlation is not the slope of the line but it does indicate the slope of the line. It measures the ratio of the sample covariance and the sample standard deviations which shows the relationship between the two variables

16. What will be the correlation coefficient of a variable that is quadratic over its sample?

0, correlation is a linear measurement.