Homework 3

PPHA 31002 | Statistics for Data Analysis I

Jay Ballesteros

November 17, 2025

STATISTICAL EXERCISES

1. Suppose a friend collects data from a random sample of college-educated, full-time workers: weekly.expenditures indicates a person's weekly expenditures in dollars, and weekly_income indicates that person's weekly income.

1.1. Explain what the value of 710.0 is, and how to properly interpret it. Is it substantively meaningful in this context?

710 is the coefficient of the dependent variable `weekly.expenditures` (intercept) and represents the predicted weekly expenditure when the weekly income (independendent variable) is 0. In this context, it means that even if a worker has no income, they are expected to spend $710 per week, it is meaningful as it provides information that even without income, spending persists (at $710)

1.2. Explain what the value of 0.6 is, and how to properly interpret it.

0.6 is the coefficient for the independent variable `weekly.income`. This 0.6 result indicates the expected change in weekly expenditures for each additional unit of weekly income.

1.3. Interpret the R^2

The R^2 value of 0.13 indicates that the prediction errors are reduced by only 13%. In other words, only 13% of the variability in weekly expenditures (independent) can be explained by the weekly income (dependent). This suggests that while weekly income has some influence on expenditures, 87% of the variability is due to other factors not included in the model.

1.4. What is the predicted expenditure for a person with a weekly income of $200 based on the regression?

It is 830.0. To get that result, I used the following equation:

$$Y = \alpha + \beta * X$$

Where Y is the predicted expenditure, α is the intercept (710.0), β is the slope (0.6), and X is the weekly income (200).

$$Y = 710.0 + 0.6 * 200 = 830.00$$

1.5. Suppose a worker in the sample reports weekly expenditures of 1,000 and has a weekly income of 200. Find the size of the residual for this worker.

The residual for this worker is 170.00. To get that result, I used the following equation:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

Where $\hat{\epsilon}_i$ is the residual, Y_i is the actual expenditure, and $\hat{Y}i$ is the predicted expenditure. As a result, the calculation is:

$$\hat{\epsilon}_i = 1000 - 830.00 = 170.00$$

1.6. Briefly explain why we have residuals.

Residuals are the differences between the actual values and the predicted values from a regression model Considering the last question, that worker spends 170.00 more that then model predicts. We have them to assess the accuracy of our model's predictions (a sort of prediction error), because probably not all factors influencing expenditures are captured in the model.

1.7. Calculate the predicted expenditure for a worker with a weekly income of $10,000. What disclaimer should we make of this predicted value?

$$Y = 710.0 + 0.6 * 10000 = 6710.00$$

The disclaimer for this predicted value is that it may not be reliable since, for instance, the maximum value of the `weekly_income` variable is 5000. So the prediction for a weekly income of $10,000 is beyond the range of the data used to fit the model, which can lead to inaccurate or misleading results.

2. A government official claims that 60% of residents favor expanding public transit. Based on survey data collected from a simple random sample of n = 100 residents, 51 indicate they are in favor of expanding public transit, and 49 respondents indicate that they do not favor expanding public transit. You plan to assess whether this survey provides sufficiently strong evidence to support the claim that the true proportion of residents who favor expansion is different from 60% (either higher or lower). Based on your conversation with the offical, you set a = 0.05 for your hypothesis test. Use the information above to develop a hypothesis test.

2.1. Write down the null and alternative hypotheses (using statistical notation) for a test that fits the situation detailed above.

For the null hypothesis,
$$H_0 : p = 0.60$$

Whereas for the alternative hypothesis,
$$H_A : p \neq 0.60$$

2.2. What is the appropriate test statistic to use for this analysis? Provide the correct formula and the value of the test statistic. Round to two decimal places.

The test statistic for this analysis is the z-test for proportion. The formula for this z-test is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where $\hat{p}$ is the sample proportion, p_0 is the hypothesized population proportion, and n is the sample size. Calculating the sample proportion:

$$\hat{p} = \frac{51}{100} = 0.51$$

Now, substituting the values into the formula:

$$z = \frac{0.51 - 0.60}{\sqrt{\frac{0.60(1-0.60)}{100}}} = \frac{-0.09}{\sqrt{\frac{0.60(0.40)}{100}}} = \frac{-0.09}{\sqrt{0.0024}} = \frac{-0.09}{0.049} = -1.84$$

2.3. Convert the test statistic you just found into a p-value. Round to three decimal places.

To convert the test statistic z = -1.84 into a p-value, because of the test being two-sided, I need to calculate for both tails (by symmetry, both are the same). Thus,

$$P(Z \geq -1.84) \approx 0.033$$

Thus, the p-value is:
$$p - value = 2 * 0.0332 = 0.066$$

2.4. What is your conclusion (i.e., your decision) based on this test?

Based on the p-value of 0.066 and the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. This means that there is not enough evidence to support the claim that the true proportion of residents who favor expanding public transit is different from 60%.

3. **The organizers of a public health campaign claim that, after the implemen-
tation of the campaign, daily sugar intake among residents is now, on average,
only 30 grams. Your friend is interested in investigating this claim, as she has a
suspicion that the true average daily sugar consumption is greater than 30 grams
for residents. You help her collect a simple random sample of n = 49 residents.
Based on the sample, the mean daily sugar intake is 34 grams, and the standard
deviation is 14 grams. She asks you to formally conduct a hypothesis test. Based
on her tolerance for type I error, you set the significance level to be a = 0.05.**

**3.1. Write down the null and alternative hypotheses, using statistical notation, based on the
situation detailed above.**

For the null hypothesis,
$$H_0 : \mu = 30$$

Whereas for the alternative hypothesis,
$$H_A : \mu > 30$$

**3.2. What is the appropriate test statistic in this exercise? Provide the correct formula and
the value of the test statistic. Round to two decimal places.**

The appropriate test statistic for this exercise is the t-test for a single mean. The formula for this t-test is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Where $\bar{x}$ is the sample mean, μ_0 is the hypothesized population mean, s is the sample standard deviation,
and n is the sample size. Substituting the values into the formula:

$$t = \frac{34 - 30}{14/\sqrt{49}} = \frac{4}{14/7} = \frac{4}{2} = 2.00$$

3.3. Convert the test statistic you just found into a p-value. Round to two decimal places.

To convert the test statistic t = 2.00 into a p-value, we need to find the area to the right of t = 2.00 in the
t-distribution with $n - 1 = 49 - 1 = 48$ degrees of freedom. I found that:

$$P(T \geq 2.00) \approx 0.03$$

**3.4. What is your conclusion (i.e., your decision) based on this test? Would your conclusion
be different if you had specified a two-sided alternative hypothesis instead? Briefly explain.**

Based on the p-value of 0.03 and the significance level of $\alpha = 0.05$, we reject the null hypothesis. This means
that there is enough evidence to support the claim that the true average daily sugar consumption among
residents is greater than 30 grams.

4. A survey conducted in June 2022 asked a random sample of U.S. adults about their preferences for electric vehicles versus gasoline-powered vehicles. The question stated: "If both options were available at a similar price, which type of car would you prefer to buy?" Among respondents under age 40, 100 out of 180 said they would prefer an electric vehicle, while among respondents age 40 and older, 155 out of 320 said they would prefer an electric vehicle. You are interested in testing whether the proportion of younger adults who state they prefer electric vehicles is greater than the proportion of older adults who state they prefer electric vehicles. The total sample size is n = 500.

4.1. Write down the null and alternative hypotheses, using statistical notation, for the situation detailed above.

For the null hypothesis,

$$H_0 : p_1 = p_2$$

Whereas for the alternative hypothesis,

$$H_A : p_1 > p_2$$

4.2. What is the appropriate test statistic in this exercise? Provide the correct formula and the value of the test statistic. Round to two decimal places.

The appropriate test statistic for this exercise is the z-test for two proportions. The formula for this z-test is:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions, p is the pooled proportion, and n_1 and n_2 are the sample sizes for the two groups. Calculating the sample proportions:

$$\hat{p}_1 = \frac{100}{180} \approx 0.5556$$

$$\hat{p}_2 = \frac{155}{320} \approx 0.4844$$

Calculating the pooled proportion:

$$p = \frac{100 + 155}{180 + 320} = \frac{255}{500} = 0.51$$

Now, substituting the values into the formula:

$$z = \frac{0.5556 - 0.4844}{\sqrt{0.51(1-0.51)(\frac{1}{180} + \frac{1}{320})}} = \frac{0.0712}{\sqrt{0.51(0.49)(0.00556 + 0.003125)}} = \frac{0.0712}{\sqrt{0.51(0.49)(0.008685)}} = \frac{0.0712}{\sqrt{0.002173}} = \frac{0.0712}{0.0466} \approx 1.5$$

So,

$$z \approx 1.53$$

4.3. Convert the test statistic you just found into a p-value.

To convert the test statistic z = 1.53 into a p-value, we need to find the area to the right of z = 1.53 in the standard normal distribution. I found that:

$$P(Z \geq 1.53) \approx 0.063$$

4.4. What is your conclusion (i.e., your decision) based on this test? Would your conclusion change if you had chosen a significance level of a = 0.01 instead? Briefly explain.

Based on the p-value of 0.063 and a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. There is not enough evidence to support the claim that the proportion of under 40 adults who prefer electric vehicles is greater than that of 40+. If we had chosen a significance level of $\alpha = 0.01$ instead, our conclusion would remain the same, as the p-value of 0.063 is still greater than 0.01.

5. A local government implemented a job training program to help unemployed residents increase their earnings potential. Ten participants were randomly selected, and their monthly incomes (in dollars) were recorded before and after completing the program:

5.1. Formulate the appropriate null and alternative hypotheses, using statistical notation, based on the information provided above.

For the null hypothesis,
$$H_0 : \mu\Delta = 0$$

Whereas for the alternative hypothesis,
$$H_A : \mu\Delta > 0$$

5.2. Calculate and report the sample mean and the sample variance of the differences in the participants monthly incomes. If needed, round to two decimal places.

The total difference in monthly incomes is 2930. Thus, the sample mean of the differences is:

$$\bar{x} = \frac{2930}{10} = 293.00$$

Whereas the sample variance of the differences is:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{2062400}{9} \approx 229156.67$$

5.3. What is the appropriate test statistic in this exercise? Provide the correct formula and the value of the test statistic. Round to two decimal places.

The appropriate test statistic for this exercise is the t-test for paired samples. The formula for this t-test is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Where $\bar{x}$ is the sample mean of the differences, μ_0 is the hypothesized population mean of the differences, s is the sample standard deviation of the differences, and n is the sample size. Calculating the sample standard deviation:

$$s = \sqrt{229156.67} \approx 478.73$$

Now, substituting the values into the formula:

$$t = \frac{293.00 - 0}{478.73/\sqrt{10}} = \frac{293.00}{478.73/3.162} = \frac{293.00}{151.36} \approx 1.94$$

5.4. Convert the test statistic you just found into a p-value. Round to two decimal places.

Considering the $t = 1.94$ and $n - 1 = 10 - 1 = 9$ degrees of freedom,

$$P(T \geq 1.94) \approx 0.04$$

5.5. Based on the data, can we conclude that participants' monthly incomes increased on average?

Based on the p-value of 0.04 and the significance level of $\alpha = 0.05$, we reject the null hypothesis. There is evidence to conclude that participants' monthly incomes increased on average after completing the program.

6. Now, treat the data as two independent samples (Sample A and Sample B), instead of repeated observations on the same participants. Specifically, 10 randomly selected participants in Sample A did not complete the job training program (control group), whereas another 10 randomly selected participants in Sample B completed the job training program (treatment group). Results for these two independent samples are given in the table below (you can think of these as the monthly incomes for the participants in groups A and B). For the questions below, you should again use a significance level of a = 0.05.

6.1. Which hypothesis test should you use to investigate whether the group of participants who completed the job training program, on average, had higher incomes?

A two-sample t-test. This test is appropriate because we are comparing the means of two groups.

6.2. State your null and alternative hypotheses using statistical notation.
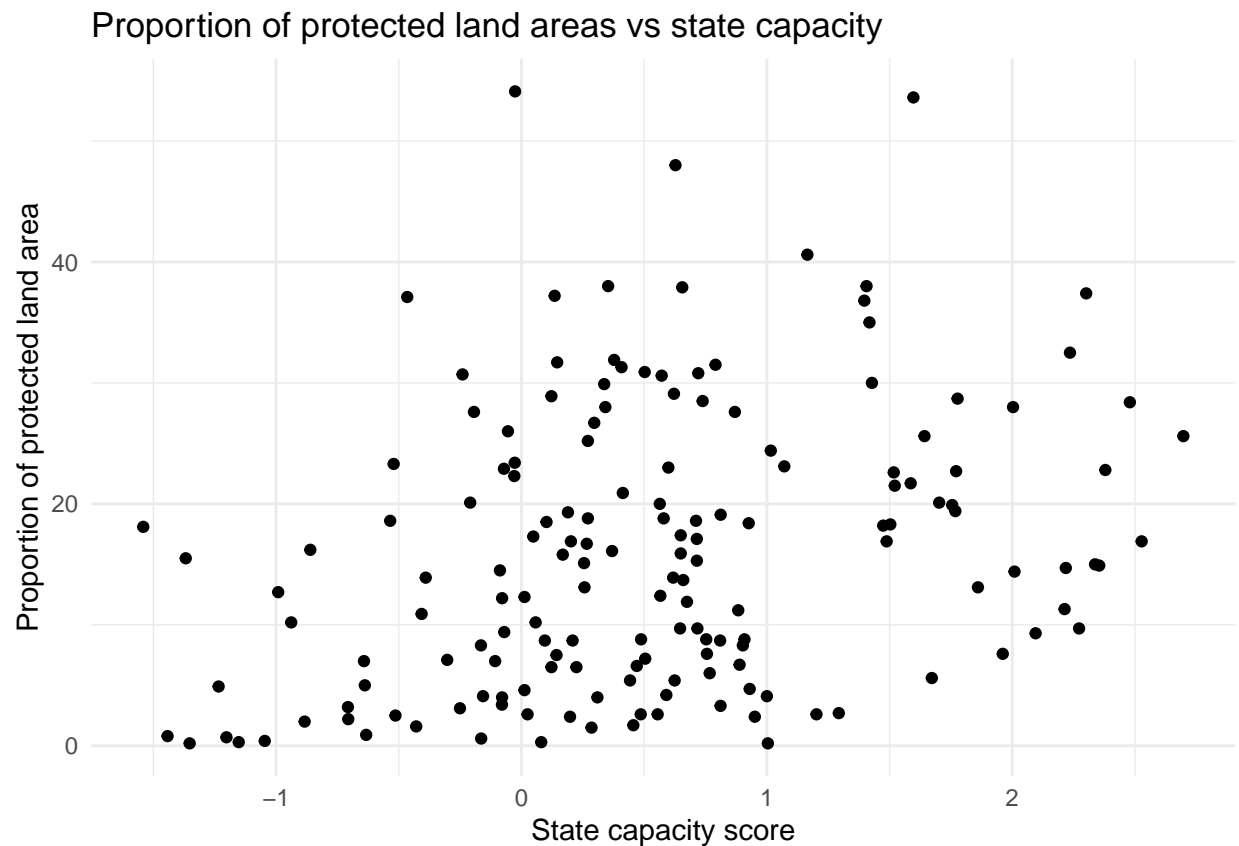
For the null hypothesis,
$$H_0 : \mu_A = \mu_B$$

Whereas for the alternative hypothesis,
$$H_A : \mu_A < \mu_B$$

DATA EXERCISE

7. Construct a scatterplot showing the relationship between the proportion of protected land areas and state capacity score. The proportion of protected land areas should be measured on the vertical axis (y-axis), and state capacity on the horizontal axis (x-axis). Make sure to label both axes and include a title as well. Provide a brief description of the relationship between these two variables based on a visual inspection.

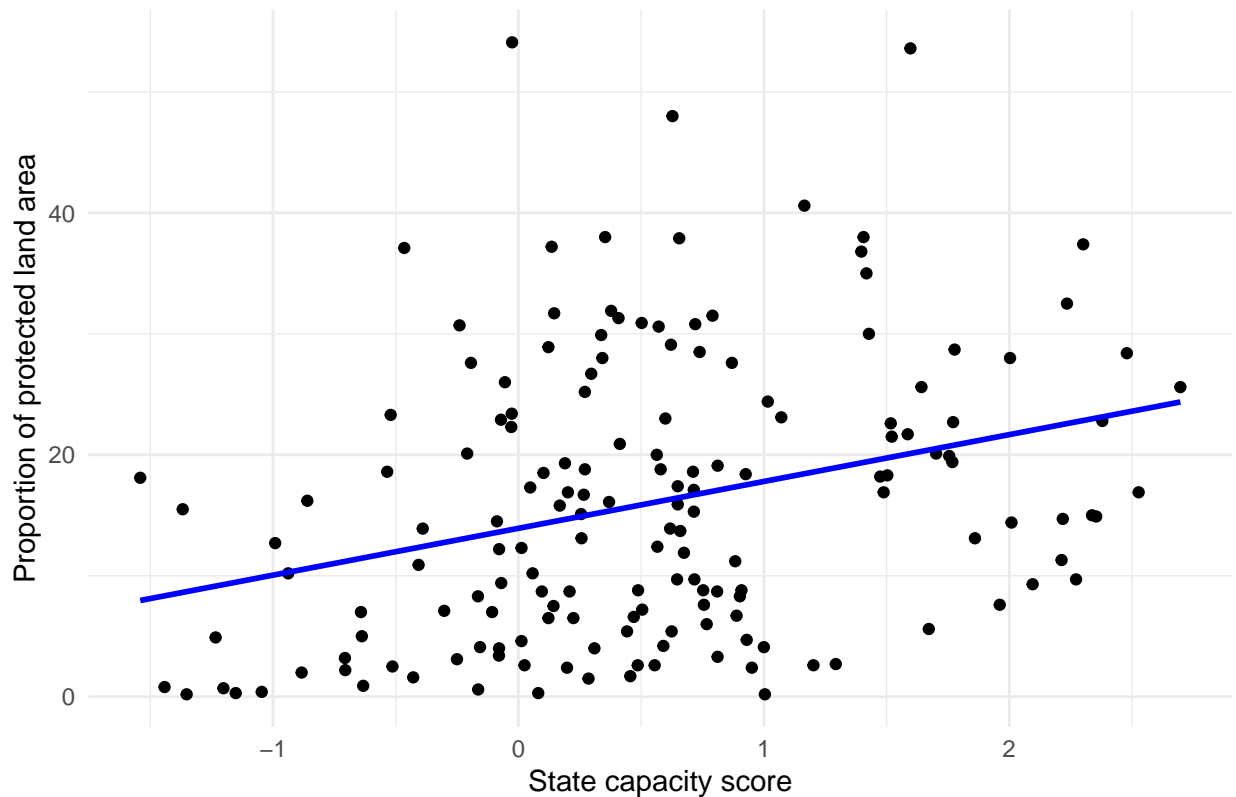Proportion of protected land areas vs state capacity

Based on the graph, the relationship between this two variables is unclear and even a little noisier. So no strong or clear visual relationship between state capacity and the proportion of protected land area.

8. Run a regression of percent of protected land area (dependent variable) on the state capacity score (independent variable), and plot the regression line atop scatterplot you just created.

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

Proportion of protected land areas vs state capacity with regression line

9. Report the estimated intercept and coefficient, interpret them, and explain if the estimated intercept is substantively meaningful in this context.

The estimated intercept for the model is 13.923514 and the coefficient for the state capacity score is 3.872412. In this context, the results can be interpreted as follows:

- The intercept (13.923514) represents the predicted percentage of protected land area when the state capacity score is 0.

- The coefficient (3.872412) indicates that for each one-unit increase in the state capacity score, the percentage of protected land area is expected to increase by approximately 3.87 percentage points.

10. Calculate the predicted percent of land under protection in a country with a state capacity of 3. To what extent should we assign

substantive meaning to this predicted value?

Using the regression equation:

$$Y = \alpha + \beta * X$$

Where Y is the predicted percentage of protected land area, α is the intercept (13.923514), β is the coefficient (3.872412), and X is the state capacity score (3).

Calculating the predicted value:

$$Y = 13.923514 + 3.872412 * 3 = 25.54075$$

A state capacity score of 3 corresponds to a country not only above the mean, but also outside the maximum values observed in the data.

11. One of your friends suggests that assuming a linear relationship between the percentage of protected land area and the state capacity score might be imposing a very strong and unrealistic assumption about the structure of the data. They suggest you split the state capacity score into deciles (i.e. 10 equal-sized groups), and regress the outcome on indicators for a country's status in the 2nd, 3rd,..., 10th decile.

Observations	163
Dependent variable	percent_of_protected_land_area
Type	OLS linear regression

$F_{(9,153)}$	2.85
R^2	0.14
Adj. R^2	0.09

	Est.	S.E.	t val.	p
(Intercept)	5.90	2.68	2.20	0.03
capacity_deciles2	7.69	3.80	2.03	0.04
capacity_deciles3	8.94	3.80	2.35	0.02
capacity_deciles4	10.99	3.86	2.85	0.00
capacity_deciles5	11.87	3.86	3.08	0.00
capacity_deciles6	12.61	3.86	3.27	0.00
capacity_deciles7	10.24	3.86	2.66	0.01
capacity_deciles8	8.66	3.86	2.25	0.03
capacity_deciles9	17.84	3.86	4.63	0.00
capacity_deciles10	12.95	3.86	3.36	0.00

Standard errors: OLS

12. Which regression—the one based on the single variable for state capacity or its transformation to deciles—seems to provide better predictions of annual earnings within the sample? Explain your answer.

To assess which model provides better predictions, we can compare the R-squared values of the two regressions. The first linear model explains about 9% of the variation in protected land area ($R^2 \approx 0.09$). The decile model increases this to about 14% ($R^2 \approx 0.14$). Although the improvement is modest, the decile specification explains more of the variance and therefore provides better predictions within the sample (comparing the models).