# CEP932 and EAD946A Practice Questions

## 1 Introduction

These practice questions cover topics in introductory statistics and econometrics discussed in Michigan State University's graduate courses CEP932 (introductory statistics) and EAD946A (econometrics). These questions aren't meant to stand in for comprehensive exam practice, but they should serve as a practical review of concepts with which you should be familiar prior to taking causal inference (MSU's EAD946B).

This set includes 20 questions with 10 questions from each course. The questions generally follow the chronological topic ordering, and they are generally increasing in difficulty. If you have questions or want to review topics, feel free to return to the notes for either course. Also, note that questions 7 and 8 relate to the chi-square goodness of fit test and the chi-square test of association. These tests are somewhat specific to statistics and may not have been covered in your introductory statistics course, so it is fine to skip those questions if you aren't familiar with those topics. In addition, you can use the $t$-table provided at the end for questions as well.

## 2 Introductory Statistics Review

1. Consider the following data from a random variable $X$.

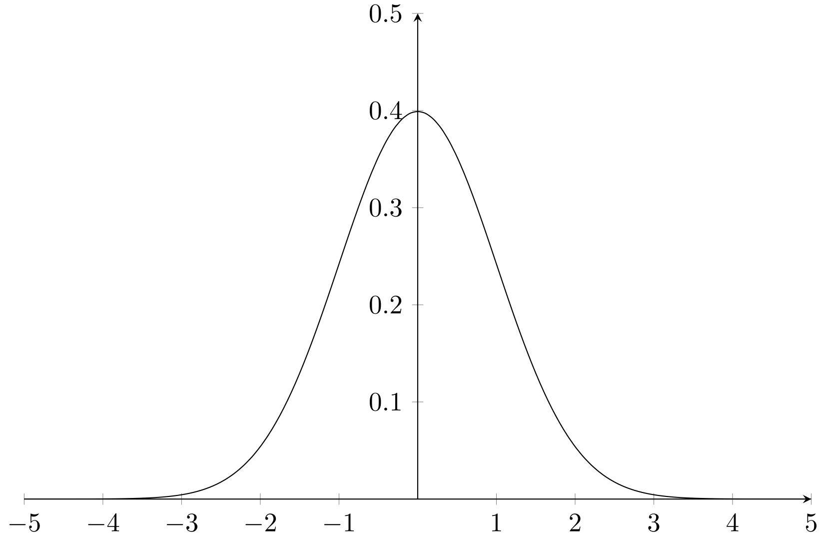
Table 1: Sample Data

| $i$ | $X$ |
|---|---|
| 1 | 2 |
| 2 | 3 |
| 3 | 7 |
| 4 | 5 |
| 5 | 3 |
| 6 | 6 |
| 7 | 4 |
| 8 | 5 |
| 9 | 3 |
| 10 | 5 |
| 11 | 1 |

For this question, conduct all computations using the population formulas.

(a) Compute the mean of $X$.
(b) Compare the mean of $X$ to the median of $X$.
(c) Compute the variance of $X$.
(d) Create a histogram of observed $X$ data and comment on whether the distribution of $X$ appears to be normal.

2. Consider the following graph of a normal distribution:

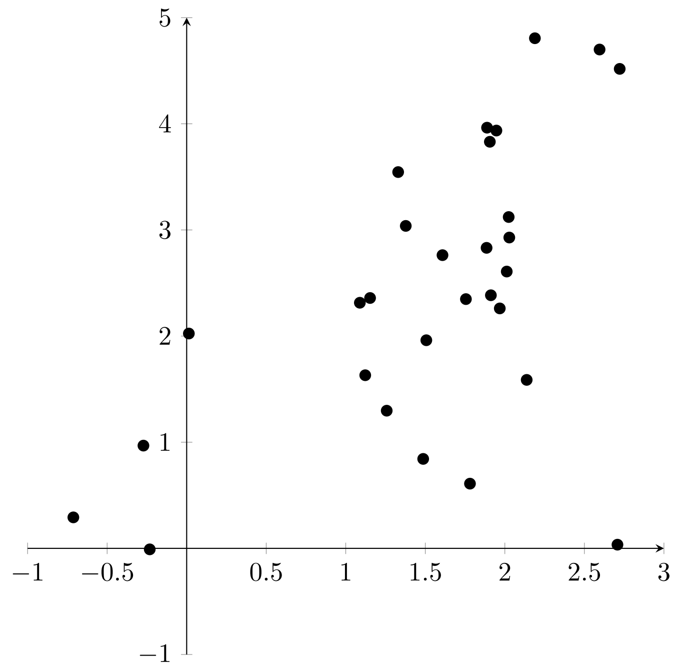



(a) What is the sum of the area under the line of in the graph above? Explain why in words.
(b) Approximately what percentage of values fall within one standard deviation of the mean of the standard normal distribution?
(c) Suppose the normal distribution above is a $N(0, 2)$. What is the z-score of a value $\bar{x} = 3$?

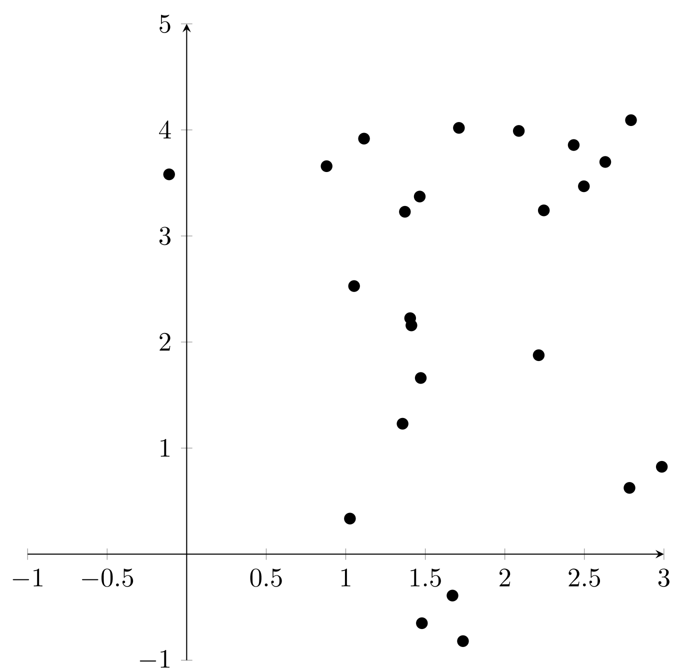
3. In one or two sentences, define a Type I error and a Type II error. In addition, discuss how you could decrease the likelihood of making one of these errors. Which of these do researchers typically aim to minimize?
4. Provide a hypothetical example of an intervention in which a one-tailed test would be an appropriate analysis.
5. Suppose that researchers try a new online curriculum for students. To determine whether the online curriculum is more effective than the old curriculum, the resaerchers randomly sample students and assign some to be in an online class and some to be in an in-person class. Denote the mean of test scores for online students $\mu_O$ and the mean of test scores for in-person students $\mu_I$.
   (a) In notation, write the null and alternative hypotheses that you would use to determine whether the online class had an effect on student outcomes.
   (b) We don't know whether test scores follow a normal distribution, but we are able to randomize 100 students to be in each group. Can we use a t-test for these data? Why or why not?
   (c) Suppose that test scores for students in online classrooms have a mean of 89 test scores for students in in-person classes have a mean of 82. In addition, suppose that the standard deviation of test scores in both groups is 10. Regardless of your answer to the previous question, compute a $t$-statistic that you can use to test the null hypothesis.
   (d) Using the value of your test statistic, would you reject the null hypothesis at the 5% level? Hint: The critical value for the $t$-distribution with more than 120 degrees of freedom is 1.96.
6. Compute the specified confidence interval given the information in each of the following cases. You may need to look up particular critical values in z-tables or $t$-tables.
   (a) Compute the 95% confidence interval for the mean given that $\bar{x} = 2$, $s = 1$, there are 51 observations and 50 degrees of freedom, and the distribution of $x$ is approximately normal.
   (b) Compute the 99% confidence interval for the mean given that $\bar{x} = 5$, $s^2 = 16$, and there are 1001 observations and 1000 degrees of freedom. Note that we do not necessarily know what the distribution of the $x$ is.
   (c) Suppose that we estimate a model via OLS using 1500 observations and find that $\hat{\beta}_1 = 2.5$ and the standard error of $\hat{\beta}_1$ is 0.75. Find the 95% confidence interval for this estimate. Without calculating a $t$-statistic, can you say whether this estimate is significant at the 5% level?
7. Suppose that we know from historical data that 40% of MSU students will major in STEM fields, 40% of students will major in humanities fields, and 20% will double major (for clarity, these three categories are mutually exclusive). Now, suppose that we poll 100 students in the honors college at MSU and find that 65 are STEM majors, 23 are humanities majors, and 12 are double majors. Compute a chi-square goodness of fit test to determine whether there is evidence that students in the honors college follow different major trends relative to other MSU students.
8. Describe in words what the chi-square test of association tells you.
9. Guess the correlations in each of the following graphs. In addition, given that variance of $x$ in each dataset is

1, guess the magnitude and direction of the covariance in each graph. Estimates are fine in this problem, but you should be able to make a good guess based on the graph and the variance.

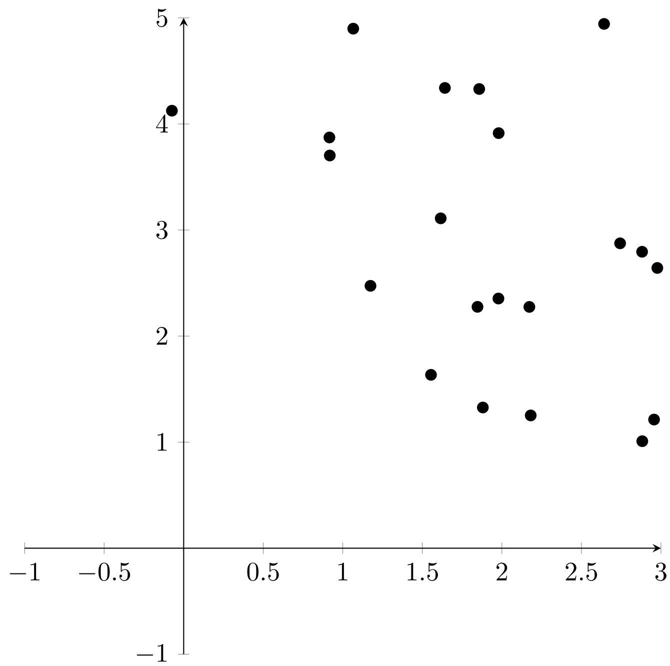
(a) Guess the correlation.



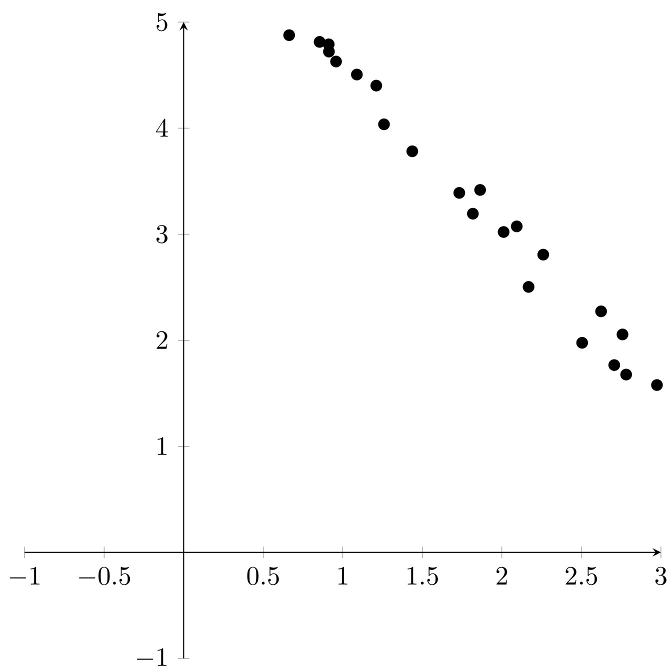

(b) Guess the correlation.




(c) Guess the correlation.




(d) Guess the correlation.




10. Consider the following generic regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Suppose that we get some data and estimate this model, and we find the following regression estimates:

$$\begin{array}{rccccc} y = & 2.876 & + & 2.034x_1 & + & 1.02x_2 \\ & (1.101) & & (0.054) & & (0.751) \end{array}$$

This regression is based on 233 observations, the sum of squared residuals is 358.02, and the total sum of squares is 571.10. In addition, standard errors are printed beneath the regression estimates. Based on this information, answer the following questions.

(a) In notation, write the null and alternative hypotheses to test whether $\beta_1 = 0$.

(b) Compute the test statistic and formally test the hypothesis from part (a) for $\alpha = 0.05$.

(c) Compute the coefficient of determination, or $R^2$, for this regression model. Comment on how well this model predicts the outcome and how much we should consider this when evaluating the model overall.

(d) Write the five statistical assumptions for bivariate regression. As a bonus, comment on how these assumptions relate to the Gauss-Markov assumptions.

# 3 Econometrics Review

11. For each of the following PMFs or PDFs, find the expected value.
    (a) Consider a fair, 6-sided die. What is the expected value of a roll of this die?
    (b) Consider a fair, 6-sided die and a fair coin. Suppose that flipping a heads is equal to a value of 1 and flipping a tails is equal to a value of 0. What is the expected value of the rolling the die and flipping the coin?
    (c) What is the expected value of a standard normal distribution?
12. In words, describe the intuition for ordinary least squares estimation. In particular, what are some properties of this estimator that make it useful for econometric analysis?
13. There are five assumptions under which OLS is the best linear unbiased estimator, or BLUE. List these assumptions and discuss the relative difficulty of satisfying each of them.
14. Consider the following multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Suppose that we get some data and estimate this model, and we find the following regression estimates:

$$\begin{array}{lllll} y_i = & \hat{\beta}_0 & + \quad \hat{\beta}_1 x_{1i} & + & \hat{\beta}_2 x_{2i} + \hat{u}_i \\ & 1.52 & 3.5 & & 6.82 \\ & (0.75) & (1.76) & & (9.81) \end{array}$$

Here, the residual sum of squares is equal to 177.2, the explained sum of squares is equal to 363.4, and there are 50 observations in the data.

    (a) Compute the $R^2$ from this regression.
    (b) Test the null hypothesis that $\beta_2 = 0$ in this model assuming that the errors are normally distributed at the 5% level.
    (c) Test the null hypothesis that $\beta_2 = 0$ in this model without assuming that the errors are normally distributed at the 5% level.
    (d) Given that this model has 48 degrees of freedom, would you report the results of part (b) or (c) in a policy brief? Why?
15. Answer each of the following questions with TRUE or FALSE. Explain your answers.
    (a) When two random variables are standardized, their covariance equals their correlation.
    (b) If you have enough data, omitted variable bias is not an issue.
    (c) If two variables are not statistically significant on their own, then they are also not jointly significant.
    (d) Heteroskedastic-robust standard errors will solve issues of heteroskedasticity and serially correlated errors.
16. Consider a bivariate regression model relating number flowers in each county $c$ to rainfall measured in inches. Provide the interpretation of $\beta_1$ in each of the following models.
    (a) $flowers_c = \beta_0 + \beta_1 rain_c + u$
    (b) $flowers_c = \beta_0 + \beta_1 \ln(rain_c) + u$
    (c) $\ln(flowers_c) = \beta_0 + \beta_1 rain_c + u$
    (d) $\ln(flowers_c) = \beta_0 + \beta_1 \ln(rain_c) + u$
17. Consider the following bivariate regression model:

$$y = \beta_0 + \beta_1 x_1 + u$$

The interpretation of $\beta_1$ in this model is that a one unit increase in $x_1$ is associated with a $\beta_1$ unit change in $y$. Answer the following questions about how data transformations would affect the OLS estimators.

    (a) How would the interpretation of $\beta_1$ change if we added 2 to each observation of $x_1$?
    (b) How would the interpretation of the OLS estimators change if we multiplied each observation of $x_1$ by 2?
    (c) How would the interpretation of the OLS estimators change if we multiplied each variable by ln(2)?

(d) How would the interpretation of the OLS estimators change if we squared $x_1$?

18. Consider a regression between per-pupil spending and whether or not a student graduates on time. Suppose that you have a student-level dataset in which per-pupil spending is coded as a continuous variable, and the graduation outcome is coded as a binary variable. In addition, suppose that your dataset identifies parental income, whether a student is a student with a disability, whether a student is an English Learner, and student race as a categorical variable including five categories: White, African American, Hispanic, Asian, and Native American or Pacific Islander. How would you include parental income, disability status, EL status, and race in a regression? Describe in words how you would do this and write the regression equation that you would use. If you prefer, feel free to use matrix notation in your equation.
19. What is heteroskedasticity? Why is this a problem? Provide three ways to solve a heteroskedasticity issue.
20. Suppose that you have administrative data on student test scores and you want to determine whether a nutrition program affected them. For each of the following scenarios, describe an analysis you might use to determine whether the program had an effect on test scores. In each case, specify whether or not your suggested analysis would provide a causal estimate. Each scenario specifies certain information that you cannot assume that you have access to, but you may assume things that aren't specified if it is helpful in your explanation.

    *Note: answers will vary for this question and multiple analyses could suffice for each scenario. In each case, explain why your suggested analysis makes sense and what you would learn from it. Hint: regression is probably not the best answer for every scenario.*

    (a) Suppose that you have 20 student test scores and information on whether students participated in the nutrition program or not. You can assume that at least 5 students participated in the program and at least 5 students did not participate, but you cannot assume the exact number or how these students were chosen. You do not have any other information on students. How would you determine the effect of the program?

    (b) Suppose that you have 100 student test scores, information on whether students had the program or not, and basic student demographic information including grade, gender, and race. You can assume that approximately half of the students participated in the program, but you cannot assume how these students were chosen. How would you determine the effect of the program?

    (c) Suppose that you have 3000 student test scores, information on whether students had the program or not, and basic student demographic information. As in part (b), you can assume that approximately half of the students participated in the program, but you cannot assume how these students were chosen. How would you determine the effect of the program? Would you be more or less confident in this result than that from part (b)? Why?

    (d) **Bonus**: Suppose that you have average test scores from 500 schools and basic administrative data on these schools. In addition, suppose that these schools were randomly chosen by researchers, and all students in selected schools were offered access to the nutrition program. How would you determine the effect of the program?

# 4 Helpful Tables

Table 2: OLS Assumptions

| Assumption | Description |
| --- | --- |
| Linearity | The model must be linear in terms. I.e., it needs to be able to be written in the form $y = \beta_0 + \beta_1 x_1 + u$ |
| Random Sampling | The data must be randomly sampled from the population. |
| No Perfect Collinearity | None of the covariates can have exact linear relationships with another covariate. |
| Zero Conditional Mean | The error term $u$ has an expected value of 0 given any values of the independent variables. |
| Homoskedasticity | The error $u$ has the same variance given any value of the explanatory variables. |

Table 3: Normal Distribution Reference Values

| Standard Deviations | Percent of Data in Range |
| --- | --- |
| 1 | 68.2% |
| 2 | 95.4% |
| 3 | 99.7% |

Table 4: Critical Values of the $t$-Distribution

| | | Significance Level | | | | |
|---|---|---|---|---|---|---|
| 1-Tailed: | | .10 | .05 | .025 | .01 | .005 |
| 2-Tailed: | | .20 | .10 | .05 | .02 | .01 |
| | 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| | 2 | 1.886 | 2.920 | 4.303 | 6.965 | 6.625 |
| | 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| | 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| | 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| | 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| | 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| | 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| | 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| | 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| | 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| **D** | 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| **e** | 13 | 1.350 | 1,771 | 2.160 | 2.650 | 3.012 |
| **g** | 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| **r** | 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| **e** | 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| **e** | 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| **s** | 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| | 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| **o** | 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| **f** | 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| | 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| **F** | 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| **r** | 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| **e** | 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| **e** | 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| **d** | 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| **o** | 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| **m** | 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| | 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| | 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
| | 90 | 1.291 | 1.662 | 1.987 | 2.368 | 2.362 |
| | 120 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 |
| | $\infty$ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |