Misc practice exercises

1. The distribution of annual income in a geographic region is skewed right, and has two modes. The overall mean and standard deviation (in thousands of dollars) are 45.8 and 36.2, respectively. What shape and summary statistics (i.e., mean and standard deviation) would we expect for the following distributions:

   a. Incomes in a single random sample of size 1025 drawn from this population.

   b. The sampling distribution of mean incomes in random samples of size 5.

   c. Same as previous question, except random samples are of size 1025.

2. Two quantitative variables V1 and V2 are approximately linearly associated, with a correlation of $r = -0.9$. Assume both variables contain only positive values.
   (a) Sketch a qualititatively reasonable scatter plot showing the association.
   (b) Sketch another qualititatively reasonable scatter plot showing the association after standardizing V1 and V2.

3. Give a complete and precise statement of the central limit theorem for sample means.

4. An ecologist has gathered data on the trunk diameter and age of a species of trees, and found there is a positive linear association between them. The data satisfies all the conditions for linear regression, and the line of best fit for predicting the age (in years) from the trunk diameter (in inches) has slope of 1.18 with y-intercept of 9.95. Identify the explanatory and response variables, and interpret the slope and intercept in context (with correct units).

   a. Identify the explanatory and response variables.

   b. Interpret the slope and intercept in context (with correct units).

   c. A regression model for the sampled data is nice! But what relationship is the ecologist, most likely, really interested in studying?

   d. Write the general form of that relationship using standard notation.

e. Write the hypotheses for testing whether there is a statistically significant linear relationship between the variables of interest.

f. Suppose the sample size is 37, and the standard error of the slope is 0.24, find the P-value and infer an appropriate conclusion.

5. According to a study by the U.S. Public Interest Research Group (USPIRG), a typical student spends an average of about \$898 per year on textbooks, with a standard deviation of \$268. A smart statistics student at Earlham College is interested in studying the implications of this for her class of 30 students.

   a. Describe (with numbers, sketch and a few words) the sampling distribution of mean textbook costs per year for that sample of 30 students. Clearly state any assumptions you make.

   b. Find the probability that the total textbook expenditure for that class will be more than \$30,000 this year. Show all steps in your solution.

   c. There is only a 5% chance that the total textbook costs for the class will be less than what value?

6. A public-service organization carried out a survey of the tuition and fees for the 2017-18 school year at a sample of 44 private colleges in the midwestern U.S. The mean and standard deviation (in thousands of dollars) were found to be 32.4 and 7.2, respectively.

   a. Carry out a hypothesis test to determine whether these data suggest the mean costs in the midwest differ significantly from the national average of \$34,800 published by the U.S. Department of Education.

   b. Infer an appropriate conclusion to the same question by constructing and interpreting a matching confidence interval.

   c. What sample size should we use if we want to estimate the true mean costs at midwestern colleges to within a \$1000 margin of error?

7. The primary goal of this question is to practice using the T-tables in both directions.

   a. Find the t^* value for a 90% confidence interval with sample size 20.

   b. Find the t^* value for a 98% confidence interval with sample size 73.

c. Find the P-value for a 1-tailed hypothesis test with $n = 20$ and $t = 2.46$.

d. Same as previous Q, except the hypothesis test is 2-tailed.

e. Find the t-score that cuts off the highest 2.5% of the area when $df = 17$.

8. Each of the following questions requires a word, phrase, or numerical value as the answer. No reasoning or justification is needed.

   i) A survey organization conducted telephone interviews in which 1,248 randomly selected adults in the United States were asked to respond to the question:

   > "At the present time do you think television commercials are an effective way to promote a new product?"

   Identify the following as precisely as possible

   * The population: ____

   * The parameter(s): ____

   ii) A survey of employee job satisfaction at a large corporation reported the correlations shown in the table. The variables are: YS=years of service; SL=salary; PR=promotion rate; and JS=job satisfaction.

	YS	**SL**	**PR**	**JS**
YS	1			
SL	0.23	1		
PR	0.58	0.74	1	
JS	−0.79	0.82	0.88	1

   Assuming the conditions necessary for interpreting correlations are met, are the following true or false:

   * Higher promotion rates are associated with longer years of service:

   ____

   * Longer years of service are associated with greater job satisfaction:

   ____

   iii) Every statement about a confidence interval contains two parts – the level of confidence, and the interval. An insurance agent estimates the mean loss claimed by auto accident clients is given by the 90% confidence interval ($4650, $7650).

* Find the margin of error: ________________

* If the sample size is 22, what is the critical t^* value: ________________

9. A team of scientists, researching the consequences of vitamin B_{12} deficiency, tracked a group of 57 adults with B_{12} deficiency for 7 years. At the end of this period they found 14 people in their sample exhibited symptoms of major depression. Use a confidence interval to estimate the true rate of depression among those with B_{12} deficiency, based on data from this sample. Be sure to include all steps and state an appropriate conclusion.

10. Now that cigarette smoking has been clearly tied to lung cancer, researchers are exploring possible links to other diseases. An article in the *American Journal of Public health* gives data on smoking rates and coronary heart disease (CHD) in 21 countries. The mean cigarette consumption in these countries was 2148 cigarettes per adult per year, with a standard deviation of 809 cigarettes per adult per year. The mean CHD rate was 144.9 deaths per 100,000 citizens, with standard deviation of 66.5 deaths per 100,000. The association between cigarette consumption and CHD rates was found to be approximately linear, positive, and with correlation $r = 0.73$.
(a) Construct a linear regression model to predict CHD death rates from cigarette consumption rate.
(b) Interpret its slope in this particular application context.

11. Give brief answers to each of the following questions as instructed:
    i) True or false: If a population distribution is highly skewed, then the central limit theorem will never apply to the sampling distribution of sample means.
    ii) Suppose a researcher conducts a hypothesis test at a significance level of 5% and gets a P-value of 0.034. Can the researcher conclude the null hypothesis is: False? True? Neither?

iii) A study on seat-belt usage among drivers and passengers based on a random sample failed to find evidence of a significant change compared to 3 years back. Explain what the study's P-value of 0.19 means in the context of this application.

iv) True or false: For a given confidence level, cutting the margin of error in half requires doubling the sample size.

v) True or false: For a given sample size, higher confidence means a smaller margin of error.