Evaluating evidence of student understanding

USCOTS 2019

Camille Fairbourn, Senior Academic Specialist – Teaching, MSU | fairbour@msu.edu
John Keane, Academic Specialist - Teaching, MSU | keanejoh@msu.edu
Workshop materials available at https://msu.edu/~fairbour/Presentations.html

Fourteen exercises you might find in an introductory statistics course are listed on the following pages. Organize these questions into two groups, which you may title in whatever manner you find appropriate. The groups need not be equal in size. Use the following questions to guide your sorting:

1. *What would a successful student response look like for questions in this group?*

2. *What are the qualities of the questions in this group that promote these responses?*

Group 1 Title:	**Group 2 Title:**
List out the numbers of the exercises in this group.	List out the numbers of the exercises in this group.

Once you've sorted the exercises into two groups, jot down your group's responses to the second guiding question:

What are the qualities of the questions in this group that promote these responses?

1. **Nicotine Patches versus Zyban ~** Medical researchers are investigating whether the use of an antidepressant medication might be an effective aid to those attempting to give up smoking. A recent study compared the effectiveness of nicotine patches to the effectiveness of the antidepressant bupropion, which is marketed with the brand name Zyban.

   893 patients were randomly allocated to four treatment groups. To keep participants blind as to their treatment, they all used a patch (nicotine or placebo) and took a pill (Zyban or placebo). The proportion of each treatment group that was not smoking 6 months after the start of the treatment is recorded below.

Treatment	Subjects	Proportion not smoking
Placebo only	160	0.188
Nicotine patch	244	0.213
Zyban	264	0.348
Zyban and nicotine patch	245	0.388

   Suppose one of your peers is trying to quit smoking and is already using Nicotine patches. Would you recommend they also use Zyban? Use at least two statistical procedures to support your opinion.

---

2. **Vaccine for Malaria ~** For a vaccine to be effective, it should reduce a person's chance of acquiring a disease. Consider a hypothetical vaccine for malaria – a tropical disease that kills between 1.5 and 2.7 million people every year. Suppose this vaccine is tested with 500 volunteers in a village who are malaria free at the beginning of the trial. Two hundred of the volunteers will be randomly selected to receive the experimental vaccine and the rest will not be vaccinated. Suppose that the chance of contracting malaria is 10% for those who are not vaccinated.

   Below are the results of this study. Do they suggest the vaccine is effective at reducing the risk of contracting malaria? Provide at least two statistical procedures as justification for your decision.

	Contracts Malaria	Does not contract Malaria
Vaccinated	13	187
Not Vaccinated	37	263

---

3. **Changing standard deviation ~** Suppose an observed set of data has 100 observations with $\bar{x} = 10$ and $s = 4.6$. Suppose we are going to add two observations to the original data set of 100 observations.

   3 and 17 ***8 and 12*** ***0 and 20*** ***10 and 10***

   a. Which pair (or pairs) of numbers will decrease the set's standard deviation?

   b. Which pair (or pairs) of numbers will increase the set's standard deviation?

   c. Briefly explain your answers.

4. **Impatient Professor ~** The wait time X for an elevator in Wells Hall is normally distributed with mean 10 seconds and standard deviation 3 seconds [i.e., $X \sim N(10,3)$]. A professor uses this elevator 9 times each day and will choose to take the stairs anytime he waits more than 15 seconds. Assuming the wait times are IID, what is the probability he will take the stairs at least 3 times on a randomly-selected day?

---

5. **Chocolate bars ~** The price per ounce for each of 14 chocolate bars is shown in the table below.

0.68	0.72	0.92	1.14	1.42	0.94	0.77
0.57	1.51	0.57	0.55	0.86	1.41	0.90

Calculate the **sample variance** for the price per ounce

---

6. **Mechanics of test statistics ~** Suppose two researchers are testing the hypotheses $H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 > \mu_2$.

   a. Fill in the table of sample statistics below such that the results of this hypothesis test would produce a p-value of at least 0.50. Show that your values satisfy this condition.

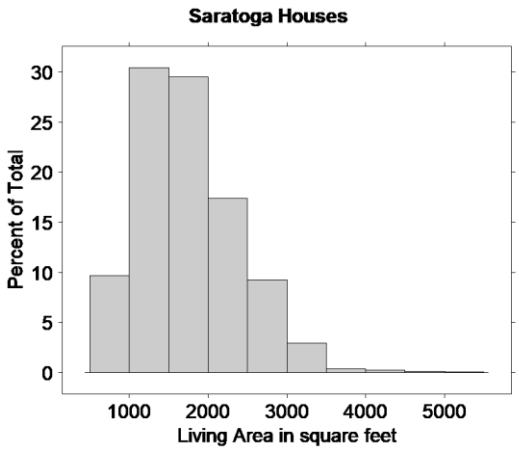
Sample size n_i	**Sample mean $\bar{x}_i$**	**Sample standard deviation s_i**
$n_1 =$	$\bar{x}_1 =$	$s_1 =$
$n_2 =$	$\bar{x}_2 =$	$s_2 =$

   b. Suppose the two values you chose for n_1 and n_2 were doubled. Show at least two ways you could make changes to the other four values in the table such that you still arrived at the same p-value.

---

7. **Saratoga homes ~** Use the histogram below to answer the following questions.

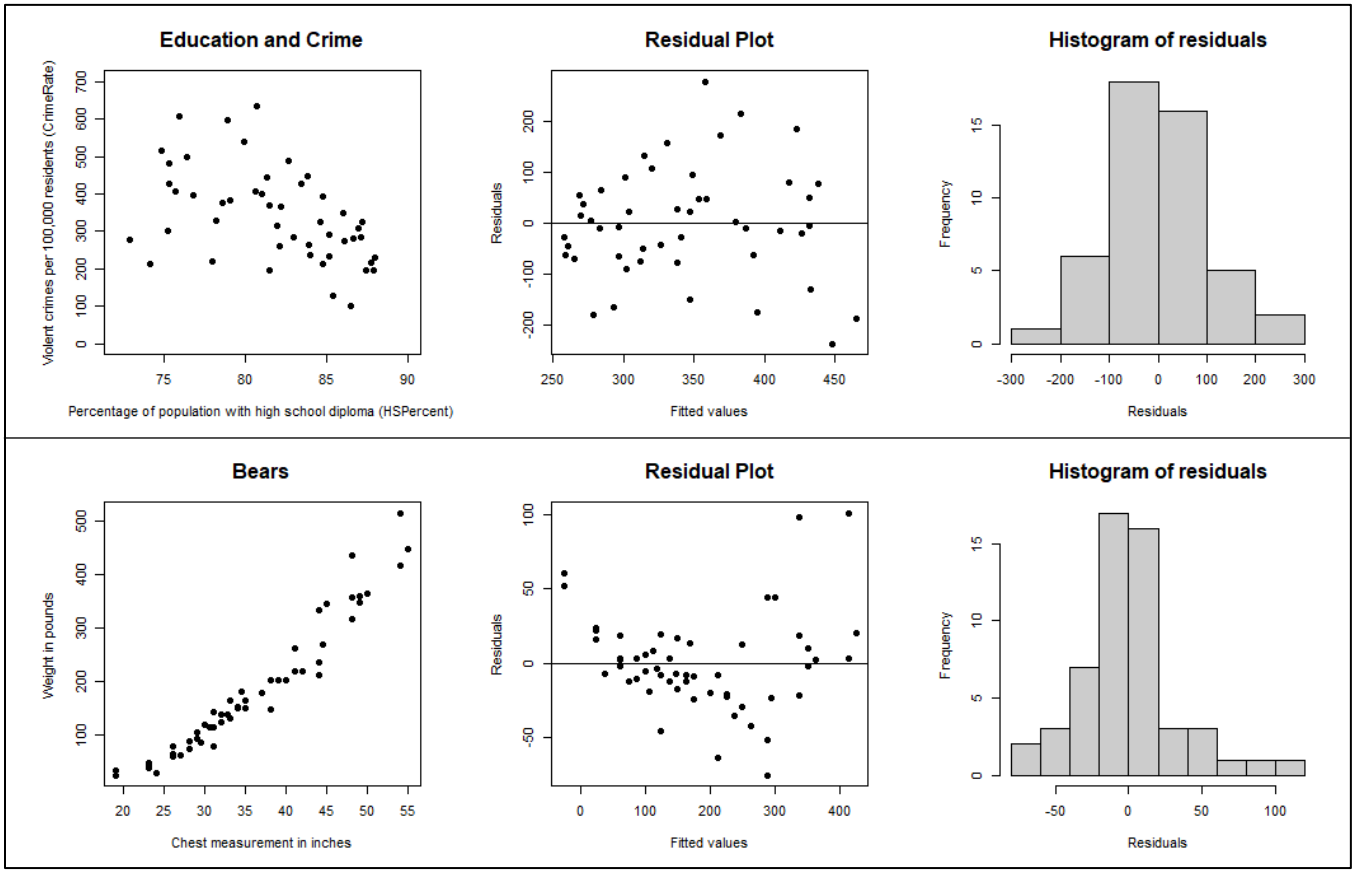
a. Approximately what percentage of houses have a living area between 1500 and 2500 square feet?



b. The summary statistics for the variable **price** are shown in the table below. One house in the data set had a price of $399,000. Calculate the z-score for the price of this house and determine if it should be considered an outlier.

Mean	SD	n
211966.7	98411.4	1727

8. **Regression Assumptions ~** Below are the scatter plots and plots for residuals for the **Education and Crime** regression model and the **Bears** chest and weight regression model from the previous two questions.



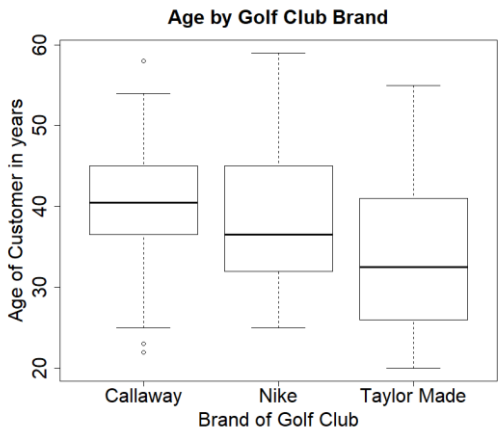
In which scenario is it more appropriate to use the linear regression model?

(Circle one) **Education and Crime** **Bears**

and ***briefly explain*** the reason for your choice in the space below.

---

9. **Geezer Golf ~** A sporting goods store sells three different brands of golf clubs, Nike, Callaway, and Taylor Made. For the last 100 sales of golf clubs, employees recorded both the brand purchased and the age of the customer. The side-by-side boxplot shows the age distribution for each brand.



Which golf club brand has a customer age distribution with the smallest interquartile range (IQR)?

10. **The Median Value ~** The median of the data set (x, y, 8, 11) is 19. Camille believes that, for this to be true, x must be greater than 23. John, on the other hand, believes x can be at most 23. Who is correct and why?

---

11. **Creating data sets ~** A data set contains eight numbers, four of which are 10, 10, 11, and 12. Identify four remaining numbers so that the standard deviation of the entire data set is 0. Explain why the standard deviation is 0 or argue that it is not possible to identify four such numbers.

---

12. **IPOisson ~** The number of companies making their initial public offering of stocks (IPO) can be modeled by a Poisson distribution with a mean of 15 per month. What is the probability of exactly 16 companies making their IPO next month?

---

13. According to a Gallup poll conducted in August 2015, 35% of Americans report spending more on gasoline than they did in 2014. This figure is based on a random sample of 3,000 adults living in all 50 U.S. states.

    a. Use the above information to calculate a 90% confidence interval for the proportion of U.S. adults who are spending more on gasoline this year than last year.
    b. If you wanted to replicate this survey in your state but didn't have the funds to contact as many people as Gallup did, what size sample do you need in order to create an interval with a margin of error of 0.04 at 90% confidence? Use a planning value based on the Gallup poll result.
    c. If you want a margin of error of 0.04 but only have enough money to sample 300 people, what is the maximum confidence level you could get? Use a planning value of 0.35.

---

14. **Living Area ~** A random sample of one-story homes and a random sample of two-story homes were selected from recent home sales in the Midwest.

Type of home	**95% confidence interval for mean living area in ft²**
One story	(1296.0, 1336.5)
Two story	(1821.6, 1889.3)

    a. Suppose the standard deviation for the two types of homes sampled were equal but the sample sizes were different. Provide plausible values for n_1 and n_2.

$n_1 =$ _____ $n_2 =$ _____

    b. The interval (1871.6, 1839.3) is either a 90% or a 99% confidence interval for the mean living area of two story homes. John believes it is a 99% interval because it provides a more precise estimate of the parameter. Camille thinks that John is wrong, but cannot explain to him why. Please help Camille.

Sample Introductory Statistics Learning Objectives

Chapter 0 – Informal Probability

1. Define the terms: outcome, event, random, independent, disjoint, complement, law of large numbers
2. Express chance as a ratio of desired outcomes/total outcomes.
3. Given a list of real-life events, rank them from impossible to certain on a likelihood scale, and translate these likelihood rankings to and from numeric probability values between 0 and 1.
4. Given 2 events, be able to explain both numerically and in words why they could or could not be considered independent.
5. Given a naturalistic scenario, identify a research question, construct a contingency table, calculate conditioned proportions, and choose which calculated values answer the research question.

Chapter 1 – Gathering & Summarizing Data

1. Given a described research question and/or results, be able to
    a. Identify the population of interest and the sample.
    b. Identify the experimental unit and the variables.
    c. Classify variables as categorical (ordinal/not ordinal) or numerical (discrete/continuous).
2. Given the description of conducted research, be able to
    a. Classify the research as a designed experiment or an observational study.
    b. Identify explanatory and response variables.
3. For Designed Experiments, given a description of the research, be able to
    a. Identify the treatment and control groups.
    b. Identify and evaluate the use of randomization.
    c. Explain why randomization in the assignment of subjects to treatment and control groups is important.
    d. Identify the use/non-use of a placebo and explain its role in the experiment.
    e. Distinguish between blind and double-blind experiments and explain the reasons for each.
    f. Explain when and why randomized, controlled, double-blind experiments allow the cautious inference of causation.
4. For Observational Studies, given a description of the research, be able to
    a. Identify and evaluate the sampling methods used
        i. Simple random sampling
        ii. Cluster and multi-stage cluster sampling
        iii. Stratified sampling
        iv. Systematic sampling
        v. Convenience sampling
    b. Identify and evaluate the use of randomization.
    c. Identify and explain types of bias that may be introduced based on the sampling method.
        i. Nonresponse
        ii. Response
        iii. Selection
    d. Identify and explain possible confounding factors.
    e. Explain why observational studies do not allow the inference of causation.
5. Calculate basic descriptive statistics for small data sets using statistical functions of a calculator. (mean, median, variance, SD, quantiles, percentiles, range, IQR)
6. Interpret data visualizations (dot plot, histogram, box plot, scatter plot, bar chart, mosaic plot).
7. Estimate descriptive statistics from visualizations.
8. Use histograms to identify the shape of a data distribution (symmetric, right/left skew).

Chapter 2 – Introduction to Inference

1. Given a description of realistic research, be able to
    a. Identify the research question.
    b. Construct a null hypothesis.
    c. Explain how to construct a randomization test for one- and two-sample proportions.
    d. Use an applet to run a randomization test for one- and two-sample proportions.
    e. Identify and/or calculate an approximate p-value from the applet and explain its meaning.
    f. Identify whether or not the null hypothesis should be rejected and interpret the conclusion in the context of the research question.
2. Given the average and standard deviation of a normally distributed data set, be able to
    a. calculate and interpret z-scores.
    b. calculate the percentage of the data that falls within a specified range.
    c. calculate the percentile rank for a given data point.
    d. calculate the raw score for a given percentile rank.
3. Given a real-world scenario, be able to identify the mean and standard deviation and use a normal curve approximation to calculate probabilities.
4. Given a histogram and/or a QQ-plot of data, be able to evaluate the appropriateness of the normal model.
5. Given a sampling situation, be able to distinguish between a population, a parameter, a sample, and a statistic.
6. Given a sample point estimate and a standard error, be able to calculate a confidence interval for the population proportion at a given confidence level.
7. Be able to correctly interpret a confidence interval.

Chapter 3 – Inference for Categorical Data

1. Given a real-life decision situation, identify when and be able to conduct a hypothesis test for
    a. a single population proportion
    b. a difference in two population proportions
    c. goodness-of-fit for multiple proportions
    d. independence in two-way tables

    by

    a. formulating a null and alternative hypothesis appropriate to the research question
    b. identifying and checking necessary conditions for the test
    c. calculating a test statistic and p-value (including degrees of freedom, where applicable)
    d. interpreting the p-value and drawing a conclusion regarding the hypotheses in the context of the research question

2. Given a real-life estimation procedure, be able to construct and interpret confidence intervals for
    a. a single population proportion.
    b. a difference in two population proportions.
3. Given a real-life estimation procedure for a single proportion and a confidence level, be able to calculate the sample size needed to obtain a given margin of error.
4. Given a description of a hypothesis testing situation for proportions, be able to identify and contextually interpret Type I and Type II errors.

Chapter 4 – Inference for Quantitative Data

1. Given a real-life decision situation, identify when and be able to conduct a hypothesis test for
    a. a single population mean
    b. paired data
    c. a difference in two population means
    d. a difference in more than two population means (ANOVA)

    by

    a. formulating a null and alternative hypothesis appropriate to the research question
    b. identifying and checking necessary conditions for the test
    c. calculating a test statistic and p-value (including degrees of freedom, where applicable)
    d. interpreting the p-value and drawing a conclusion regarding the hypotheses in the context of the research question
2. Given a real-life estimation procedure, be able to construct and interpret confidence intervals for
    a. a single population mean
    b. paired data
    c. a difference in two population means
3. Given a real-life estimation procedure for a single mean and a confidence level, be able to calculate the sample size needed to obtain a given margin of error.
4. Given a description of a hypothesis testing situation for means, be able to identify and contextually interpret Type I and Type II errors.
5. Given a hypothesis test scenario for a single population mean, be able to calculate the chance of making a Type II error.

Chapter 5 – Introduction to Regression

1. Given a scatter plot, be able to roughly estimate the correlation coefficient.
2. Given the description of a linear association of two variables, be able to interpret the sign and strength of the correlation coefficient.
3. Given a scatter plot, be able to distinguish between a
    a. positive and negative association
    b. linear and non-linear association
    c. strong and weak association (linear or non-linear).
4. Given a five-number summary for a set of data and/or computer output, be able to calculate and/or find the equation of the regression line.
5. Given a linear regression equation, be able to
    a. interpret the slope and intercept of the estimated regression line.
    b. use the regression line to compute point estimates.
6. Given a linear regression equation and information about an observed data point, be able to calculate and interpret the residual for that point.
7. Given a scatter plot and residual plots, be able to
    a. evaluate the whether the conditions have been met for creating a linear regression model for the data.
    b. identify outliers and how they influence the least squares line.
8. Given computer output of a linear regression model, be able to
    a. use a t-test to test for the significance of the slope.
    b. construct and interpret a confidence interval for the slope of the regression line.
9. Given a scenario involving a test-retest situation, be able to recognize and explain the regression effect and the regression fallacy.