4. A survey organization conducted telephone interviews in December 2008 in which 1,009 randomly selected adults in the United States responded to the following question.

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

Of the 1,009 adults surveyed, 676 responded "yes." In December 2007, 622 of 1,020 randomly selected adults in the United States had responded "yes" to the same question. Do the data provide convincing evidence that the proportion of adults in the United States who would respond "yes" to the question changed from December 2007 to December 2008 ?

5. A recent report stated that less than 35 percent of the adult residents in a certain city will be able to pass a physical fitness test. Consequently, the city's Recreation Department is trying to convince the City Council to fund more physical fitness programs. The council is facing budget constraints and is skeptical of the report. The council will fund more physical fitness programs only if the Recreation Department can provide convincing evidence that the report is true.

   The Recreation Department plans to collect data from a sample of 185 adult residents in the city. A test of significance will be conducted at a significance level of $\alpha = 0.05$ for the following hypotheses.

$$H_0 : p = 0.35$$
$$H_a : p < 0.35,$$

   where p is the proportion of adult residents in the city who are able to pass the physical fitness test.

   (a) Describe what a Type II error would be in the context of the study, and also describe a consequence of making this type of error.

   (b) The Recreation Department recruits 185 adult residents who volunteer to take the physical fitness test. The test is passed by 77 of the 185 volunteers, resulting in a p-value of 0.97 for the hypotheses stated above. If it was reasonable to conduct a test of significance for the hypotheses stated above using the data collected from the 185 volunteers, what would the p-value of 0.97 lead you to conclude?

   (c) Describe the primary flaw in the study described in part (b), and explain why it is a concern.

AP® STATISTICS
2012 SCORING GUIDELINES

Question 4

Intent of Question

The primary goal of this question was to assess students' ability to identify, set up, perform, and interpret the results of an appropriate hypothesis test to address a particular question. More specific goals were to assess students' ability to (1) state appropriate hypotheses; (2) identify the name of an appropriate statistical test and check appropriate assumptions/conditions; (3) calculate the appropriate test statistic and *p*-value; (4) draw an appropriate conclusion, with justification, in the context of the study.

Solution

Step 1: States a correct pair of hypotheses.

Let p_{07} represent the population proportion of adults in the United States who would have answered "yes" about the effectiveness of television commercials in December 2007. Let p_{08} represent the analogous population proportion in December 2008.

The hypotheses to be tested are $H_0: p_{07} = p_{08}$ versus $H_a: p_{07} \neq p_{08}$.

Step 2: Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a two-sample *z*-test for comparing proportions.

Because these are sample surveys, the first condition is that the data were gathered from independent random samples from the two populations. This condition is met because we are told that the subjects were randomly selected in the two different years. Although we are not told whether the samples were selected independently, this is a reasonable assumption given that they are samples of different sizes selected in different years.

The second condition is that the sample sizes are large, relative to the proportions involved. This condition is satisfied because all sample counts (622 "yes" in 2007; $1{,}020 - 622 = 398$ "no" in 2007; 676 "yes" in 2008; $1{,}009 - 676 = 333$ "no" in 2008) are all at least 10 (or, are all at least 5).

An additional condition may be checked: The population sizes (more than 200 million adults in the United States) are much larger than 10 (or, 20) times the sample sizes.

Step 3: Correct mechanics, including the value of the test statistic and *p*-value (or rejection region).

The sample proportions who answered "yes" are:

$$\hat{p}_{07} = \frac{622}{1{,}020} \approx 0.6098 \text{ and } \hat{p}_{08} = \frac{676}{1{,}009} \approx 0.6700.$$

The combined proportion, $\hat{p}_c$, who answered "yes" in these two years is:

$$\hat{p}_c = \frac{622 + 676}{1{,}020 + 1{,}009} = \frac{1{,}298}{2{,}029} \approx 0.6397.$$

Question 4 (continued)

The test statistic is:

$$z = \frac{\hat{p}_{07} - \hat{p}_{08}}{\sqrt{\hat{p}_c\left(1 - \hat{p}_c\right)\left(\frac{1}{n_{07}} + \frac{1}{n_{08}}\right)}} = \frac{0.6098 - 0.6700}{\sqrt{(0.6397)(1 - 0.6397)\left(\frac{1}{1{,}020} + \frac{1}{1{,}009}\right)}} \approx -2.82.$$

The p-value is $2P(Z \leq -2.82) \approx 0.0048$.

Step 4: State a correct conclusion in the context of the study, using the result of the statistical test.

Because this p-value is smaller than any common significance level such as $\alpha = 0.05$ or $\alpha = 0.01$ (or, because this p-value is so small), we reject H_0 and conclude that the data provide convincing (or, statistically significant) evidence that the proportion of all adults in the United States who would answer "yes" to the question about the effectiveness of television commercials changed from December 2007 to December 2008.

Scoring

Each of steps 1, 2, 3, and 4 are scored as essentially correct (E), partially correct (P), or incorrect (I).

Step 1 is scored as follows:

Essentially correct (E) if the response identifies correct parameters *AND* both hypotheses are labeled and state the correct relationship between the parameters.

Partially correct (P) if the response identifies correct parameters *OR* states correct relationships, but not both.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: Either defining the parameter symbols in context, or simply using common parameter notation, such as p_{07} and p_{08}, with subscripts clearly relevant to the context, is sufficient.

Step 2 is scored as follows:

Essentially correct (E) if the response correctly includes the following three components:
1. Identifies the correct test procedure (by name or by formula).
2. Checks for randomness.
3. Checks for normality.

Partially correct (P) if the response correctly includes two of the three components listed above.

Incorrect (I) if the response correctly includes one or none of the three components listed above.

Question 4 (continued)

Step 3 is scored as follows:

Essentially correct (E) if the response correctly calculates both the test statistic and a *p*-value that is consistent with the stated alternative hypothesis.

Partially correct (P) if the response correctly calculates the test statistic but not the *p*-value,
OR
if the response calculates the test statistic incorrectly but then calculates the correct *p*-value for the computed test statistic.

Incorrect (I) if the response fails to meet the criteria for E or P.

Step 4 is scored as follows:

Essentially correct (E) if the response provides a correct decision in context, also providing justification based on linkage between the *p*-value and conclusion.

Partially correct (P) if the response provides a correct decision, with linkage to the *p*-value, but not in context,
OR
if the response provides a correct decision in context, but without justification based on linkage to the *p*-value.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: If the decision is consistent with an incorrect *p*-value from step 3, and also in context with justification based on linkage to the *p*-value, then step 4 is scored as E.

Each essentially correct (E) step counts as 1 point. Each partially correct (P) step counts as ½ point.

4 **Complete Response**

3 **Substantial Response**

2 **Developing Response**

1 **Minimal Response**

If a response is between two scores (for example, 2½ points), use a holistic approach to decide whether to score up or down, depending on the overall strength of the response and communication.