

## Student Performance Q&A: 2010 AP<sup>®</sup> Statistics Free-Response Questions

The following comments on the 2010 free-response questions for AP<sup>®</sup> Statistics were written by the Chief Reader, Allan Rossman of California Polytechnic State University, San Luis Obispo. They give an overview of each free-response question and of how students performed on the question, including typical student errors. General comments regarding the skills and content that students frequently have the most problems with are included. Some suggestions for improving student performance in these areas are also provided. Teachers are encouraged to attend a College Board workshop to learn strategies for improving student performance in specific areas.

### Question 1

#### ***What was the intent of this question?***

The primary goals of this question were to assess students' ability to (1) apply terminology related to designing experiments; (2) construct an appropriate plot that could be used to begin to investigate the fit of the data to a linear model; and (3) decide, from a graphical display, whether a linear regression model is appropriate for a set of data.

#### ***How well did students perform on this question?***

The mean score was 2.05 out of a possible 4 points, with a standard deviation of 1.25 points.

#### ***What were common student errors or omissions?***

Part (a), subpart i

- Many students claimed that the treatments were garlic oil or garlic oil concentration (note the singular), rather than the five different concentrations of garlic oil.
- Some students mistakenly reported the treatments to be the experimental units (the birds), as in "the 8 birds receiving 0% concentration, the 8 birds receiving 2% concentration, ...."
- Some students reported the experimental units, response variable, explanatory variable, and other things related to this experiment, rather than identifying the treatments.

Part (a), subpart ii

- Some students claimed that the food granules were the experimental units, but the food granules were not what the response was measured on.
- Some students mistakenly grouped the birds, for example, by saying that the five groups of eight birds each were the experimental units.

Part (a), subpart iii

- Many students wrote that the response variable was the *mean* number of granules consumed, apparently not realizing that the mean is a summary calculated about the response variable rather than the variable itself.
- Some students thought that the response was the amount of food consumed per treatment (concentration), rather than the amount per experimental unit (bird).
- Some students reported the purpose of the experiment as the response variable.

Part (b), subpart i

- Many students used a scale on the horizontal axis that presented the five concentrations as equally spaced.
- Some students did not label one or both axes.
- Some students provided a bar chart or histogram-like graph rather than the requested scatterplot.
- Some students reversed the two axes.

Part (b), subpart ii

- Many students attempted to answer this question without referring to the graph.
- Some students talked about correlation or slope as a justification for linearity.
- Some students calculated a regression line and used the line's existence as a justification for linearity.
- Some students who drew a bar graph discussed the graph as if it were a scatterplot.
- Some students referred to skew in the scatterplot or appealed to normality as justification.

***Based on your experience of student responses at the AP Reading, what message would you like to send to teachers that might help them to improve the performance of their students on the exam?***

Be very precise when applying the terminology of experiments. One particular difficulty for many students is confusing a variable with the values/levels/categories of the variable. In this study the explanatory variable was garlic oil concentration, but the treatments were the five particular values of garlic oil concentration that were applied to the granules fed to the birds.

Another important distinction that can be difficult for students is the difference between a *variable* (such as number of food granules consumed by the bird) and a *summary* statistic calculated for that variable (such as the average number of food capsules consumed). One strategy for helping students to recognize these distinctions is to ask them to identify the variables and observational/experimental units in every example that they encounter throughout the course.

With regard to graphing, encourage students always to provide labels in context for all axes and to ensure that scaling on axes is done consistently. Make clear that the best way to determine the appropriateness of a linear relationship for a set of data is to examine appropriate graphs, not to examine a correlation value.

Also, strive to help students see connections among various topics in the course. For example, this question combined elements of descriptive statistics and data analysis with aspects of data

collection and experimental design. Some students may have been surprised to see these two topic areas represented in the same question.

## Question 2

### ***What was the intent of this question?***

The primary goals of this question were to assess students' ability to (1) describe a sampling distribution of a sample mean and (2) set up and perform a normal probability calculation based on the sampling distribution.

### ***How well did students perform on this question?***

The mean score was 0.78 out of a possible 4 points, with a standard deviation of 1.02 points.

### ***What were common student errors or omissions?***

Part (a)

- Many students revealed a lack of understanding about what a sampling distribution is.
  - Some described a (presumably population) distribution of individual song lengths.
  - Some described a distribution of song lengths in a sample of 40 songs.
- Many students described the shape of the sampling distribution as normal, rather than approximately normal.
- Some students indicated that a symmetric distribution is the same as a normal distribution.
- Some students gave the correct parameter values for the sampling distribution but did not label them clearly.
- Some students reported the mean of the sampling distribution as “about 3.9 minutes” rather than “3.9 minutes.”
- Many students did not divide by  $\sqrt{n}$  when calculating the standard deviation of the sampling distribution.
- Some students used incorrect notation for the mean and standard deviation of the sampling distribution.
- Some students mistakenly reported the distribution for the total (combined) length of the 40 songs.

Part (b)

- Many students used the wrong standard deviation — often the standard deviation of the population rather than the standard deviation of the sample mean or the standard deviation of the total.
- Some students obtained the correct numerical answer but did not show how the probability was calculated.
- Some students reversed the order of subtraction in the numerator of the z-score calculation.
- Some students analyzed the question as a hypothesis test.

- Many students who took this approach did not use the given value of the population mean in the null hypothesis.
- If students used this hypothesis-testing approach correctly, the  $p$ -value would have been the requested probability. But many students went on to draw a conclusion about the population mean, which was not what was asked.
- Some students who attempted to perform the calculation in terms of total airtime rather than the sample mean failed to find the correct standard deviation of the total.
- Some students treated the random variables as discrete rather than continuous, for example, by finding  $\Pr(T \geq 161)$  rather than  $\Pr(T > 160)$ .

***Based on your experience of student responses at the AP Reading, what message would you like to send to teachers that might help them to improve the performance of their students on the exam?***

Do not underestimate the difficulty that students have with the concept of a sampling distribution. In particular, help students to recognize and appreciate the differences among three distributions:

- the population
- a sample of values from that population
- a sample statistic (in this question, the sample *mean*) when repeated random samples are selected from the population

Using simulations, including hands-on as well as technology-based simulations, is an effective way to help students see the differences among these three distributions. Exploring the impact of different population types (e.g., normal, symmetric but not normal, slightly skewed, strongly skewed) and sample sizes is also helpful.

Encourage students to be careful to use appropriate notation when working with sampling distributions. Another point is to realize that the Central Limit Theorem (CLT) speaks specifically to the *shape* of the sampling distribution of a sample mean; the mean and variance of the sample mean can be derived from properties of expected values and variances, without recourse to the CLT.

Show students the connection between calculations involving the sample *mean* and those involving the sample *total*, including how the means and standard deviations of these two random variables relate to each other.

Remind students always to show their work with probability calculations. Students should also be cautioned that providing calculator syntax is not a replacement for sound statistical communication. The distinction between a normal distribution and an *approximately* normal one should also be emphasized. Encourage students always to accompany such calculations with a well-labeled sketch, paying close attention to the variable and observational units of the graph, and shading the region of interest.

### **Question 3**

***What was the intent of this question?***

The primary goals of this question were to assess students' ability to (1) interpret the meaning of a confidence level; (2) use a confidence interval to test the plausibility of a claim about the value of a population parameter; and (3) perform a sample size calculation related to a confidence interval.

***How well did students perform on this question?***

The mean score for this question was 1.02 out of a possible 4 points, with a standard deviation of 0.97 point.

***What were common student errors or omissions?***

Part (a)

- A large number of students interpreted a particular confidence *interval* rather than the confidence *level*.
- Some students suggested that a particular (calculated) interval would contain the (true) population proportion 95 percent of the time.
- Some students were unclear about whether 95 percent referred to one particular interval or the collection of possible intervals.
- Some students omitted any mention of the idea of repeated sampling.
- Some students omitted any mention of an interval, for example, by saying that 95 percent of samples would contain the population proportion.

Part (b)

- Many students concluded that 0.39 *is* the value of the population proportion, tantamount to accepting a null hypothesis.
- Some students were confused about how the question was asked, answering “Yes, because 0.39 is in the interval,” rather than “No, because 0.39 is in the interval.”
- Some students based their conclusion on whether 0 (rather than 0.39) was in the interval.
- Some students attempted to conduct a hypothesis test without referring to the interval.

Part (c)

- Many students calculated the minimum sample size necessary for a margin of error of at most 0.119, rather than what this question asked about the sample size for this particular survey.
- Some students used 0.39 or 0.5 instead of the observed sample proportion 0.419 in the calculation.
- Some students did not recognize that the sample size needed to be an integer.
- Some students set up the correct expression to be solved but then did not correctly solve for  $n$ .
- Some students attempted to use the margin of error for a  $t$ -interval for a population mean.

***Based on your experience of student responses at the AP Reading, what message would you like to send to teachers that might help them to improve the performance of their students on the exam?***

Help students to recognize the difference between interpreting a confidence *interval* and interpreting a confidence *level*. An interpretation of confidence level should refer to something that conceptually happens many, many times (taking random samples and computing a confidence interval from each), which works about 95 percent of the time (meaning that about 95 percent of

the possible intervals would capture the population parameter). Provide frequent feedback to improve students' communication regarding difficult concepts such as confidence level.

Teach students to make sure that all necessary words are in their response. For example, a common incomplete response for part (a) was "In repeated sampling, 95% of all samples will contain the true proportion." Notice that the essential word "interval" was never mentioned in this response.

Make clear to students that any conclusion equivalent to "accepting the null hypothesis" is inappropriate. Rather, such a conclusion needs to be phrased in terms of "not finding sufficient evidence to reject the null hypothesis" or "not finding sufficient evidence in support of the alternative hypothesis."

Remind students to read carefully and answer only the question asked. Many students interpreted the interval (which was not even provided at that point) rather than the confidence level in part (a), and in part (c) many students determined the smallest sample size to achieve a certain margin of error, rather than the sample size used in this particular survey.

Present intervals to students in the form "estimate  $\pm$  margin of error," not always in the form "lower endpoint, upper endpoint."

Finally, remind students that sample sizes need to be integers.

## **Question 4**

### ***What was the intent of this question?***

The primary goals of this question were to assess students' ability to (1) calculate an expected value and a standard deviation; (2) recognize the applicability of a binomial distribution and perform a relevant binomial probability calculation (or recognize the applicability of a normal approximation and use it to perform a relevant probability calculation); and (3) suggest an appropriate sampling method to achieve a given goal.

### ***How well did students perform on this question?***

The mean score was 1.11 out of a possible 4 points, with a standard deviation of 1.18 points.

### ***What were common student errors or omissions?***

Part (a)

- Many students rounded the correct expected value of 15.62 to 16, apparently not realizing that an expected value, even with a random variable that only takes on integer values, does not need to be an integer.
- Many students could not calculate the standard deviations correctly.
  - Many did not recognize the relevance of the binomial distribution.
  - Some calculated the standard deviation of a sample proportion.
  - Some used 2,323 (the number of owners of type E) rather than the sample size of 2,000 in the standard deviation calculation.

Part (b)

- Some students who calculated a binomial probability mistakenly included  $\Pr(X = 12)$  or mistakenly excluded  $\Pr(X = 0)$  or used the wrong sample size of 2,323.
- Some students who calculated a normal probability mistakenly divided by  $\sqrt{n}$  in the standard deviation or mistakenly reversed the order of subtraction in the numerator of the  $z$ -score calculation.
- Some students only reported that 12 is below the expected value without going on to support their judgment of the unusualness of this outcome based on a numerical calculation such as a  $z$ -score.

Part (c)

- Many students omitted one of the three essential components in their description of the sampling method.
  - Some did not make clear that the total sample size was 2,000 owners.
  - Some did not make clear that their sampling method would ensure at least 12 owners in the sample for each of the five car models.
  - Some did not make clear how *randomness* would play a role in their sampling method.
- Some students mistakenly referred to stratified random sampling as cluster sampling or systematic sampling.
- Some students described a randomized experiment in which car models were randomly assigned to owners.

***Based on your experience of student responses at the AP Reading, what message would you like to send to teachers that might help them to improve the performance of their students on the exam?***

Ensure that students realize that expected values do not have to be integers, even when the random variable of interest can only take on integer values. One way to achieve this goal is to frequently ask students to interpret expected value as the long-run average value that a random variable would approach after a large number of repetitions.

Give students practice identifying when binomial and normal distributions are appropriate and also with determining expected values and variances of such distributions.

Emphasize to students the difference between random *sampling* and random *assignment*. The goal of random sampling is to select a representative sample from a population so that findings about the sample can be generalized to the larger population. The goal of random assignment is to create treatment groups in a randomized experiment that are as similar as possible in all respects except for the explanatory variable, so a significant difference in the responses between the groups can be taken as evidence of a cause-and-effect relationship between the explanatory and response variables. Students should be discouraged from using the word “randomly” in a casual, nontechnical manner.

This question provides another illustration of the importance of making connections among different topics in the course. Parts (a) and (b) concerned probability distributions, while part (c) addressed issues of sampling and data collection.

## Question 5

### ***What was the intent of this question?***

The primary goal of this question was to assess students' ability to set up, perform and interpret the results of a significance test. 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) compute the appropriate test statistic and  $p$ -value; and (4) draw an appropriate conclusion, with justification, in the context of the study.

### ***How well did students perform on this question?***

The mean score was 1.59 out of a possible 4 points, with a standard deviation of 1.33 points.

### ***What were common student errors or omissions?***

Some students failed to realize that this question called for a significance test, and they based their response solely on a descriptive analysis of the sample data.

#### Step 1

- Many students used standard symbols for parameters but then defined the symbols incorrectly.
  - Some omitted any mention of population.
  - Some omitted any mention of mean.
- Some students used nonstandard notation for parameter symbols without defining them as means of populations.
- Some students reversed the null and alternative hypotheses.
- Some students used 1 and 2 as subscripts but did not identify them with specific suppliers.
- Some students attempted a confidence interval approach but almost none adjusted for the one-sided alternative (e.g., making a conclusion at a 5 percent significance level using a 90 percent confidence interval).

#### Step 2

- Some students named an incorrect procedure, either a  $z$ -test or a matched-pairs test, or a pooled  $t$ -test without addressing the equal variances condition.
- Many students simply said "SRS" or "random sample" or "random" without specifying that *two* independent random samples were taken.
- Many students did not address the randomness condition at all.
- Many students did not include graphs to assess normality.
- Many students did not distinguish between stating the condition (the two populations of fish lengths follow normal distributions) and checking the condition (because the two graphs of sample fish lengths are roughly symmetric with no outliers, it is reasonable to assume that the population distributions are normal).
- Many students did not address the normality condition at all.



- Some students mistakenly believed that the sample sizes of the two groups must be the same.

### Step 3

- Some students had the correct test statistic but an incorrect  $p$ -value.
  - Some found the probability of the complement and obtained a  $p$ -value = 0.6.
  - Some calculated a two-sided  $p$ -value = 0.8.
- Some students reported the correct  $p$ -value but did not report the value of the test statistic.
- Some students forgot to square the standard deviations when using the  $t$ -test statistic formula.

### Step 4

- Many students did not include the idea of *mean* fish lengths in their conclusion.
- Some students stated their conclusion as if the alternative hypothesis had been two-sided (e.g., “we cannot conclude that the population means are different”).
- Many students phrased their conclusion as accepting the null hypothesis and concluding that the population mean lengths were the same.
- Some students phrased their conclusion as “retaining the null hypothesis” without clarifying what this meant in this context.
- Some students did not explain the linkage between the  $p$ -value and the conclusion they reached.

***Based on your experience of student responses at the AP Reading, what message would you like to send to teachers that might help them to improve the performance of their students on the exam?***

First and foremost, lead students to realize that a question about whether sample data provide convincing evidence regarding a claim calls for a significance test. Emphasize the four components of such a test: stating a hypothesis, identifying the procedure and checking its conditions, performing the mechanics of calculations, and drawing a conclusion in context with linkage to  $p$ -value. Questions from previous AP Exams and their scoring guidelines can be effectively used to make students aware of the importance of providing these components and communicating them clearly. When using these scoring guidelines, however, note that beginning with this year’s exam, these steps of an open-ended inference question could be scored as partially correct as well as essentially correct and incorrect.

Recognize that stating parameters clearly is a challenge for many students. Remind students often that parameters concern populations and are a numerical summary of some sort (e.g., mean, proportion, correlation). For this question mentioning both “population” and “mean” were important for students who defined their parameter symbols.

Emphasize the importance of *checking* conditions and assumptions when conducting an inference procedure. Many students did not even acknowledge the existence of this step. Also remind students to check conditions involving random sampling or random assignment, as well as conditions about shapes of distributions or sample sizes. Students should reproduce (in rough form) any graphs that are used in checking conditions, and these graphs should be commented upon.

When using a calculator to carry out calculations, students should report both the test statistic value and the  $p$ -value (and, ideally, degrees of freedom when conducting a  $t$ - or chi-square test).

Finally, help students get in the habit of drawing conclusions in context and with clear linkage between the  $p$ -value and the conclusion drawn. Also note that in this question, the conclusion needed to refer to *mean* lengths of fish from the two suppliers.

## Question 6

### ***What was the intent of this question?***

The primary goals of this investigative task were to assess students' ability to (1) produce and comment on a graphical display; (2) calculate a test statistic based on rank data; and (3) use simulation results to draw an appropriate conclusion.

### ***How well did students perform on this question?***

The mean score was 1.92 out of a possible 4 points, with a standard deviation of 0.98 point. Student performance was higher on this investigative task than on any other investigative task for the past several years. Many students produced reasonable graphs in part (a), most students completed the calculations in parts (c) and (d) correctly, and students' conclusions from the simulation analysis were stronger than with similar questions from previous exams.

### ***What were common student errors or omissions?***

Part (a)

- Some students produced graphs that were essentially stacked bar graphs, treating the damage amounts as if they could be meaningfully added or averaged.
- Some students did not label the axes of their graphs in context.

Part (b)

- Many students did not communicate their descriptions clearly.
  - Some mistakenly claimed that Florida had the greatest damage amounts for all distance categories, rather than for four of five distance categories.
  - Many used vague phrases such as “in general” or “overall” when presenting information that should have been described more precisely.
  - Some failed to see overall patterns and resorted to a category-by-category description.
  - Some mistakenly referred to damage totals, or averages, across distances.
- Many students gave correct descriptions of a similarity or a difference but not both.

Parts (c) and (d)

- Some students calculated the median rank across distance categories for each region, rather than the average (mean) rank.

Part (e)

- Many students left this part blank.

- Many students did not determine an approximate  $p$ -value from the simulated distribution.
- Among students who did determine an approximate  $p$ -value, some did not relate the  $p$ -value to the rarity (or nonrarity) under the null hypothesis of the observed value of  $Q$ .
- Some students failed to use the observed value of  $Q$  and based their answer only on the graph of the simulated distribution.
- Some students based their conclusion on the *shape* of the simulated distribution.
- Some students ignored the simulated distribution and based their conclusion solely on the observed value of  $Q$  and its proximity to the expected value under the null hypothesis.

***Based on your experience of student responses at the AP Reading, what message would you like to send to teachers that might help them to improve the performance of their students on the exam?***

Give students even more experience with the reasoning process behind tests of significance and with interpreting the results of simulation analyses to assess statistical significance. In particular, students struggle to realize that such simulation analyses are conducted under the assumption that the null hypothesis is true, so an essential step is to use the simulated distribution to assess the rarity of the observed value of the test statistic, assuming the null hypothesis to be true.