



Student Performance Q&A: 2012 AP[®] Statistics Free-Response Questions

The following comments on the 2012 free-response questions for AP[®] 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) describe a nonlinear association based on a scatterplot; (2) describe how an unusual observation may affect the appropriateness of using a linear model for bivariate data; and (3) implement a decision-making criterion on data presented in a scatterplot.

How well did students perform on this question?

The mean score was 1.31 out of a possible 4 points, with a standard deviation of 0.83.

What were common student errors or omissions?

Part (a):

- Many students did not describe three aspects of association, often describing only the direction without mentioning strength or form.
- Many students mistakenly described the form of the association as linear; few students attempted to describe the nonlinear form evident in the scatterplot.
- Some students neglected to refer to the context (price and quality of sewing machines) in their descriptions.

Part (b):

- Many students discussed the influence of the point without clearly explaining why the influential point made the linear regression model less appropriate for the data as a whole.
- Some students mentioned the lack of fit for the correct point without discussing the point's role in suggesting an alternative curved relationship for the data.

Part (c):

- Some students circled only one of the two points that satisfy the criteria.
- A few students circled points that did not satisfy the criteria.

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 recognize that just as there are three common features of a *distribution* to describe — center, variability, and shape — so too are there three aspects of *association* to describe: strength, direction, and form. Even better is to help students learn to write descriptions of association not in a rote manner but in a way that reflects on what the scatterplot reveals about the particular variables in the study.

Help students by providing several examples of scatterplots that reveal a nonlinear form of association between the variables.

Emphasize what the concept of an influential observation entails. An influential observation need not fall far from the least squares line; in fact, an influential observation often pulls the least squares line toward it, giving it a small residual value.

A challenging idea that warrants some attention is that of reasonableness of a linear model for a set of bivariate data. Help students realize that the issue of reasonableness applies to the dataset as a whole and concerns the question of whether a different (nonlinear) model might provide a better fit for the data as a whole.

Question 2

What was the intent of this question?

The primary goals of this question were to assess students' ability to (1) perform calculations and compute expected values related to a discrete probability distribution; and (2) implement a normal approximation based on the central limit theorem.

How well did students perform on this question?

The mean score was 2.13 out of a possible 4 points, with a standard deviation of 1.48.

What were common student errors or omissions?

Part (a):

- A few students reported counts instead of probabilities in the table.
- A few students wrote the probabilities in the wrong order.

Part (b):

- Some students reported the correct expected value but did not show their work in performing the calculation.
- Some students showed how to calculate the expected value correctly but made an arithmetic error.
- Some students mistakenly reported the most likely value (\$2) as the expected value.

Part (c):

- Some students used the most common value (\$2) rather than the expected value, obtaining a calculation of $500/2 = 250$ spins required.
- Some students reported the correct answer but did not show supporting work to justify their answer.
- Some students used a guess-and-check approach without further justifying their answer.

Part (d):

- Some students reversed the roles of the boundary value and mean in calculating the numerator of the z-score, calculating $(500 - 700)$ rather than the correct $(700 - 500)$.
- Some students indicated the wrong direction (e.g., ≤ 500 rather than ≥ 500) or showed no direction.
- Many students used calculator notation without clearly specifying the parameter and boundary values (e.g., saying “normalcdf(500, ∞ , 700, 92.79)” without clarifying that $\mu = 700$, $\sigma = 92.79$, and $P(X \geq 500)$ are being calculated).
- Some students divided by $\sqrt{1,000}$ in calculating the standard deviation.
- Some students reported z-scores and probabilities that were inconsistent with each other (e.g., $P(Z \geq 2.155) = 0.9844$).

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?

Encourage students to be extremely clear in communicating how they perform probability calculations. This includes specifying the probability distribution being used, the parameter values, and the interval whose probability is being calculated. Students should not be discouraged from using their calculator to perform the calculation, but they must be able to communicate what they are calculating without resorting solely to calculator notation. Provide ample opportunities to practice these skills and sufficient feedback on students' performance.

Emphasize proper interpretations of probabilities and expected values. This particular question did not ask for interpreting the expected value, but students who understood expected values as long-run average values might have been less tempted to make common errors. The issue of rounding up to the next largest integer when calculating a minimum sample size is also one to be emphasized, and teachers can help students to realize that this step (giving the next largest integer as the final answer) needs to be explained in the student's response.

Give students opportunities to practice applying normal approximations in situations other than approximating the sampling distribution of a sample mean.

Question 3

What was the intent of this question?

The primary goals of this question were to assess students' ability to (1) compare two distributions presented with histograms; and (2) comment on the appropriateness of using a two-sample t -procedure in a given setting.

How well did students perform on this question?

The mean score was 1.93 out of a possible 4 points, with a standard deviation of 1.15.

What were common student errors or omissions?

Part (a):

- Many students did not use comparative language, describing the two distributions without comparing them.
- Some students omitted one or more of the three features — center, variability, shape — that were expected.
- Some students neglected to relate their comments to the context of comparing household sizes between 1950 and 2000.
- Some students made contradictory statements, especially with regard to shape (e.g., saying that the distributions were skewed to the right and also normal).

Part (b):

- Many students mistakenly believed that the sample had to be normally distributed in order for a t -procedure to be valid.
- Many students considered the normality and sample size conditions to be separate issues, not realizing that a large sample size allows for a t -procedure to be valid even with a population that is not normally distributed.
- Many students did not clearly specify that both samples needed to be randomly selected from their populations.
- Many students did not clearly distinguish between *stating* and *checking* conditions for inference.
- Some students tried to implement completely inappropriate checks, such as $np > 10$.
- Some students attempted to check the condition that the population size is at least 10 times larger than the sample size, but they often seemed to be unaware of why this condition matters and how it relates to other conditions.

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?

Provide considerable opportunities for practice with comparing distributions of data, based on a variety of types of graphs. Model good responses and insist that students provide comparisons with complete sentences, not bullet lists of descriptions. Comparisons of center and variability should involve statements of which group is larger (with respect to center or variability) or that the groups have similar centers/variability. When possible, such statements should be supported with specific numerical evidence, such as means/median and standard deviations/IQRs. With regard to comparing shapes of distributions, caution students against using multiple descriptors (such as “skewed right” and “normal”) for the same distribution.

Expect students to clearly state and check conditions for inference often. In addition, help students understand the reasons behind the conditions. For example, the t -distribution is not a close approximation to the sampling distribution of a sample mean when the sample size is small and the population distribution is nonnormal, so a 95 percent confidence based on the t -distribution might not successfully capture the actual parameter value in 95 percent of all random samples.

Encourage students to read questions carefully and make sure that their responses address the question asked. Some students might have been surprised to see a question about checking conditions for inference paired with a question about comparing distributions presented in histograms.

Question 4

What was the intent of this 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; 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 1.56 out of a possible 4 points, with a standard deviation of 1.28.

What were common student errors or omissions?

- Many students did not include all four components of a significance test.
- Some students included written descriptions of hypotheses that pertained to samples rather than populations, even though few students used symbols for sample statistics in their hypothesis statements. One common way to do this was to mistakenly say that one of the parameters is the proportion of adults who answered “yes” in 2008; the use of past tense (“answered” rather than “would have answered”) made the description about the sample rather than the population.
- Many students seem to have learned that conditions need to be checked for inference without understanding what purpose the conditions serve.
- Some students reported only a p -value without also providing the value of the test statistic.
- Some students provided the correct values for the test statistic and p -value but also included additional information that was mistaken, for example, by writing a formula for the test statistic that was based on population parameters rather than sample statistics.
- Many students did not explicitly justify their conclusion by comparing the p -value with an assumed significance level α or by commenting generally on the size of the p -value.
- Many students presented their conclusion in terms of the samples rather than the populations, for example, by concluding that the proportion who answered “yes” in 2007 was different from the proportion who answered “yes” in 2008.
- Some students attempted to interpret what the p -value meant, often not doing so correctly.
- A few students did not express their conclusion in the context of this study.

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?

Strive to help students not only learn the steps involved with a hypothesis test but also understand the overall reasoning process and how those steps relate to each other. Provide detailed feedback on student performance with this task, even at the level of looking at the tense of verbs, which indicates whether the student is interpreting a conclusion in terms of the population or sample. Remind students frequently that hypotheses are about population parameters rather than sample statistics.

Identifying parameters clearly is also a big challenge for many students, so they should receive ample practice with that skill. Continue to make students aware of the importance of always checking conditions for inference, based on the specific details of the study at hand, rather than merely stating assumptions for inference, when conducting a significance test or producing a confidence interval. Make students aware that these checks require examination of the sample data and consideration of data-collection procedures.

Provide considerable practice and feedback with summarizing conclusions from significance tests. Encourage students to be very clear in stating how their conclusion follows from the p -value. Remind students frequently about the need to express conclusions in the context of the research question presented.

Question 5

What was the intent of this question?

The primary goals of this question were to assess students' ability to (1) describe a Type II error and its consequence in a particular study; (2) draw an appropriate conclusion from a p -value; and (3) describe a flaw in a study and its effect on inference from a sample to a population.

How well did students perform on this question?

The mean score was 1.31 out of a possible 4 points, with a standard deviation of 1.08.

What were common student errors or omissions?

Part (a):

- Some students described Type I error rather than Type II error.
- Many students did not describe the error in terms of the parameter of interest (proportion of adults in this city who are able to pass the physical fitness exam).
- Some students referred to “accepting the null hypothesis” or “rejecting the alternative hypothesis” when they should have referred to “failing to reject the null hypothesis.”
- Some students described only part of the error (e.g., “we fail to reject the null hypothesis”) without specifying the condition (e.g., not going on to say “when the null hypothesis is actually false”).
- Some students gave a textbook definition without relating it to this context or describing a consequence in this context.
- Some students described a consequence that was inconsistent with the error described.
- Some students made multiple attempts, at least one of which was incorrect.

Part (b):

- Many students did not provide an explicit connection (linkage) between their test decision/conclusion and how the p -value related to the given significance level.
- Many students did not clearly refer to the parameter of interest in stating their conclusion.
- Some students attempted to provide an interpretation of the p -value and did so incorrectly.
- Some students rejected the null hypothesis, despite the very large p -value.
- Some students stated a conclusion that was equivalent to accepting the null hypothesis, for example, concluding that the population proportion able to pass the physical fitness exam is equal to 0.35.

- Some students did not refer to the one-sided nature of the alternative hypothesis in stating their conclusion.
- Some students phrased their conclusion in terms of the sample rather than the population, for example, by drawing a conclusion about the proportion who passed the exam.

Part (c):

- Many students identified the small sample size as a flaw, without commenting on the nonrandom nature of how the sample was selected.
- Many students correctly commented that volunteers are likely to be different from, or nonrepresentative of, other people without specifying *how* they are likely to be different or nonrepresentative (e.g., arguing that volunteers are more likely to be physically fit than the population as a whole).
- Many students stopped short of describing a problem with making an inference, for example, by saying only that healthy people would be overrepresented in the sample without further commenting that the inference drawn about the proportion would could pass the exam in the population would be invalid.
- Some students used the term “bias” without clearly explaining what it means and how it pertains to this study.
- Some students used statistical terminology incorrectly (e.g., saying that the results are “skewed”).

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?

Emphasize that students understand concepts related to statistical inference, not simply the ability to apply specific procedures of statistical inference. Closely related to this is the need to ask, and provide feedback on, frequent questions about concepts of inference that are not presented in the standard four-part (hypothesis, conditions, calculations, conclusion) manner.

Model for students that conclusions must always be presented in the context of the study at hand. For inference questions this almost always means writing the conclusion in terms of the relevant parameter of interest.

Repeatedly make the point that significance tests assess the strength of evidence provided by the sample data *against* the null hypothesis. It is not appropriate to draw conclusions about the strength of evidence for the null hypothesis or against the alternative hypothesis. One way to achieve this goal is by requiring students to address whether the sample data provide compelling evidence for the alternative hypothesis.

Be vigilant in making sure that students use statistical vocabulary (e.g., bias, confounding, skew) correctly at all times. In fact, students might be well advised to avoid using statistical vocabulary on the exam unless they are quite sure that they are using it correctly. Advise students to describe the principle rather than simply use the term.

Question 6

What was the intent of this question?

The primary goals of this question were to assess students’ ability to (1) implement simple random sampling; (2) calculate an estimated standard deviation for a sample mean; (3) use properties of variances

to determine the estimated standard deviation for an estimator; and (4) explain why stratification reduces a standard error in a particular study.

How well did students perform on this question?

The mean score was 1.50 out of a possible 4 points, with a standard deviation of 1.04.

What were common student errors or omissions?

Part (a):

- Many students provided an incomplete description of how to implement their random sampling method.
- Some students did not specify to use four digits when using a random-digit table.
- Some students did not describe how to deal with numbers beyond the population size when using a random-digit table.
- Some students did not specify how to deal with repeat numbers when using a random-digit table or calculator or software.
- When describing a pull-names-from-a-hat approach, some students did not describe a process for mixing/randomizing the names prior to selecting them.
- Some students described a sampling process other than simple random sampling, such as selecting 60 girls and 40 boys.
- Some students described a systematic sampling method, such as selecting every 20th name from the list.

Part (b):

- Some students reported the correct answer but did not indicate how it was obtained, with either a formula or a calculation.
- Some students used $\frac{\sigma}{\sqrt{n}}$ rather than $\frac{s}{\sqrt{n}}$.
- Some students did not successfully simplify the expression $\frac{4.13}{\sqrt{100}}$.

Part (c):

- Some students did not properly use the 0.6 and 0.4 weights, for example, by calculating $\sqrt{\frac{1.80^2}{60} + \frac{2.22^2}{40}}$.
- Some students used the weights improperly, for example, by calculating $\sqrt{(0.6)\frac{1.80^2}{60} + (0.4)\frac{2.22^2}{40}}$ without squaring the weights.
- Some students attempted to combine standard deviations rather than variances, for example, by calculating $(0.6)\frac{1.80}{\sqrt{60}} + (0.4)\frac{2.22}{\sqrt{40}}$.
- Some students did not consider the sample sizes, for example, by calculating $\sqrt{0.6(1.80)^2 + 0.4(2.22)^2}$.
- Some students obtained an estimate for the standard deviation of the population and then divided by $\sqrt{100}$, for example, by calculating $\frac{0.6(1.80) + 0.4(2.22)}{\sqrt{100}}$.

Part (d):

- Many students correctly commented on relevant features of the dotplots, such as the smaller variability in soft-drink numbers within males and females separately for Rania's sample as compared with Peter's overall sample, but did not explain how this affects the standard deviation of Rania's estimator.
- Some students correctly commented on the differences in centers between males and females in Rania's dotplots but did not mention the smaller variability in soft-drink numbers within the two genders.
- Some students noted that when Rania's two dotplots are combined, the resulting data has less variability than Peter's. Though this observation is true, it misses the connection to reduced variability in Rania's estimator due to stratification.
- Some students compared only the variability in the distributions of data without referring to variability in estimators.
- Some students supplied only a general description of the benefits of stratification without referring to the information contained in the dotplots.
- Some students based their explanation only on sample sizes.
- Some students based their explanation on the apparent normality of Rania's distributions as compared with the skewed distribution of Peter's data.

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?

Make abundantly clear to students that descriptions of sampling methods must contain enough detail that the methods could be implemented solely based on the description. This includes dealing with issues such as repeats and out-of-bounds numbers when using a random-digit table, and accounting for mixing of names when pulling names from a hat. Give students detailed feedback on such descriptions to prepare them for the level of specificity expected on the exam.

Help students understand how to work with variances of random variables, including multiples and sums of random variables. Another important point that can be difficult for students to grasp is the general idea of using a random variable as a point estimator of a parameter. Constantly remind students of the importance of showing work when performing calculations.

Help students understand not only how to implement stratified random sampling but also what the benefits of stratification are. Encourage students to go beyond a superficial understanding that "stratification reduces variability" to understand variability of what, and under what circumstances. Providing visual explanations based on dotplots or histograms and giving examples where stratification does not help much, in addition to examples where it helps considerably, can be worthwhile.