CollegeBoard
Advanced Placement Program

Student Performance Q&A:

2007 AP® Statistics Free-Response Questions

The following comments on the 2007 free-response questions for AP® Statistics were written by the Chief Reader, Brad Hartlaub of Kenyon College in Gambier, Ohio. 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 goals of this question were to assess a student's ability to: (1) explain how a commonly used statistic measures variability; (2) use a graphical display to address the research question of interest in a simple comparative experiment; and (3) use a confidence interval to make an appropriate inference.

How well did students perform on this question?

The mean score was 1.1 out of a possible 4 points. The overall student performance on this exploratory analysis of data question was not as good as it has been in the past. Students had a considerable amount of trouble describing how the standard deviation summarizes variability in the discoloration ratings for the control group. Most students were able to use the dotplots to correctly comment on the effectiveness of the preservative in lowering the amount of discoloration in strawberries. Many students, however, could not interpret the confidence interval provided in the context of this study.

What were common student errors or omissions?

Part (a)

- Most students were not able to explain that standard deviation represents the "average" or "typical" distance of the points in the control group from the mean discoloration rating.
- Many students used the 68–95–99.7 rule in an attempt to summarize variability, without providing evidence that the distribution of discoloration ratings in the control group is approximately normal.

- Several students simply tried to comment on whether 2.141 represents a small or a large amount of variability.

Part (b)

- Students generally performed well on part (b), demonstrating a solid understanding of the connection between the effectiveness of the preservative and comparative measures of relative standing in the treatment and control groups.
- Some students compared shape, center, spread, and even outliers for the two groups, but they failed to link the relevant portion of their analysis (center) with their decision about the effectiveness of the preservative.

Part (c)

- Quite a few students noted that the fact that 0 was not included in the confidence interval was suggestive of a difference in population means.
- In this experimental setting, students found it very difficult to express their conclusions in terms of the population mean discoloration ratings for treated and untreated strawberries. Part of their difficulty may have stemmed from the fact that the populations are hypothetical.
- A number of students gave a generic interpretation of the 95 percent confidence interval, such as, "We are 95 percent confident that the difference in population means for treated and untreated strawberries is between 0.16 and 2.72." A better interpretation, one that conveys the *direction* of the difference, would be: "We are 95 percent confident that the actual difference in population mean discoloration ratings is between 0.16 and 2.72 units higher for treated than for untreated strawberries."
- We saw many incorrect interpretations of the confidence interval and the confidence level.

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?

Students do not have trouble calculating the standard deviation, but they have no idea how this statistic measures variability. Asking students to explain orally and in writing how the standard deviation, and other statistics, measure variability would be extremely beneficial. Students need to repeat these conceptual descriptions in a variety of different settings before they are able to gain a solid understanding of them. Similarly, interpreting confidence intervals in context takes practice. Asking students to comment on these important concepts once or twice is not enough. They will benefit from repeated applications in different settings.

Question 2

What was the intent of this question?

The three primary goals of this question were to evaluate a student's ability to: (1) clearly explain the importance of a control group in the context of an experiment; (2) describe the randomization process required for three groups; and (3) reduce variability by grouping experimental units as homogeneously as possible.

How well did students perform on this question?

The mean score was 1.9 out of a possible 4 points. Overall, student performance on this design question was very good. In fact, it was the highest scoring question this year, and one of the highest scoring design questions in the history of the program. Students were able to describe the advantage of a control group and a method for random assignment, though some students did not provide enough details for their method to be implemented. Many students correctly pointed out that the primary purpose of blocking is to reduce variability by forming groups of homogeneous experimental units, but they did not understand that the best blocking variable would have the strongest association with the response variable (a measure of joint and hip health).

What were common student errors or omissions?

Part (a)

- Most students answered this part correctly.
- Some students gave an advantage of a control group but not in context.
- Some students correctly talked about "reducing effects of confounding variables."
- Some students incorrectly discussed bias.

Part (b)

- Many students correctly described a method for random assignment.
- Some students did not include enough details for the method to be (unambiguously) implemented.
- A few methods did not produce completely randomized designs. These included rolling a die and requiring 100 dogs in each group, and nested designs in which, for each clinic, the 30 dogs were randomly assigned to each group.

Part (c)

- Many students appeared to understand that blocking was to reduce variability, but most students did not relate it to the response variable.
- Many students related the blocking variable to those treatments that related more to interaction (interaction effect: different breeds react differently to the different treatments).
- Many students did not indicate a choice of the "stronger" relationship or the "larger" variability.
- Many students talked about confounding variables, some correctly and some incorrectly.
- A few students incorrectly discussed bias.

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?

When asking students to describe random assignment of the experimental units to the treatments, make sure they understand that someone reading their description should be able to implement their method. Simply saying that a coin, dice, random number generators, or tables of random numbers will be used is not enough. Students continue to have trouble with the concepts of blocking and confounding, especially when they are asked to apply or discuss these concepts in a specific context.

Question 3

What was the intent of this question?

This question was developed to assess a student's understanding of the sampling distribution of the sample mean. In particular, a student's ability to: (1) compare probabilities concerning sample means from different sample sizes; (2) compute an appropriate probability; and (3) recognize that an application of the Central Limit Theorem is being evaluated.

How well did students perform on this question?

The mean score was 1.16 out of a possible 4 points. Students had a considerable amount of trouble applying and explaining major properties of the sampling distribution of a sample mean. Justifications lacked clarity and contained ambiguous language. Poor communication resulted in lower than expected scores.

What were common student errors or omissions?

Part (a)

- A large majority of students chose the correct sample size.
- The most common error was ambiguity in expressing the justification. Many students said something to the effect of "larger samples have less variability." If they were talking about variability *within a sample*, this was not true at all. Students need to clearly express that they are referring to variability *between sample means*.
- Another version of this ambiguity error was to use the word "it" unclearly. Many students said something like, "it varies more with smaller sample size," without clearly specifying what "it" referred to. Again, if "it" was the sample, then the statement was wrong. But if "it" was the distribution of sample means, then the statement was correct.
- Many students commented that the distribution (of sample means) becomes more normal as the sample size increases. This is actually not a true statement for this problem: because the population (of fish lengths) is assumed to be normally distributed, the sample mean follows a normal distribution for all sample sizes. More importantly, the *shape* of the sampling distribution is not the relevant issue for answering this question; the *variability* in the sampling distribution is the key consideration.
- Another common error was to cite the Law of Large Numbers and assert that the sample mean $\bar{x}$ approaches the population mean μ as the sample size increases. This is (roughly) true, but it does not directly address the issue of how *variable* the sampling distribution of $\bar{x}$ is for a given sample size. Many students' wording of this assertion also made the mistake of suggesting that $\bar{x}$ has to be closer to μ when *n* is large than when *n* is small, apparently not recognizing that $\bar{x}$ is a random variable with a probability distribution.
- Many students used the word "accurate" ambiguously in their justification.
- Some students gave an incomplete argument that with smaller samples, a single observation can have more influence on the mean than with larger samples. This is the right idea behind understanding why averages vary more with small samples than with large samples. But most students who gave this "influence" argument did not go on to talk about the greater variability in sample means with smaller samples. Some students mistakenly argued that getting a single fish longer than 10 inches was more likely with a small sample than with a large sample.

Part (b)

- A common error was to interpret 0.3 as the population standard deviation σ and therefore divide by $\sqrt{50}$ to calculate the standard deviation of the sampling distribution of $\bar{x}$.
- Some students gave the correct answer and the calculator command without providing more support for their answer. That support could have come in any one of three forms: a sketch, a *z*-score calculation, or identifying the components of the calculator command.
- Some students looked in the wrong tail for the probability, or they exchanged $\bar{x}$ and μ, resulting in the negative of the correct *z*-score.
- Using a *t*-distribution rather than a normal (*z*-) distribution was another noticeable error.
- A few students raised the correct answer to the 50^{th} power, apparently trying to find the probability that all 50 fish in the sample would be longer than 10 inches.
- Many students could have strengthened their responses by providing a well-labeled sketch of the normal distribution with the region of interest shaded.
- Notational errors were prevalent but overlooked. These included expressing the probability as $P(x > 7.5)$ or $P(\mu > 7.5)$ rather than the correct $P(\bar{x} > 7.5)$ and equating the *z*-score and the probability (e.g., $z = -1.67 = .0475$).

Part (c)

- A common error was to answer "no" with a justification that a normal calculation cannot be done without a normal distribution. This response failed to recognize that the probability in (b) is based on the sampling distribution of the sample mean rather than on the distribution of fish lengths.
- Many students answered correctly ("yes") but provided an incomplete justification. Students should have noted the large sample size in this case ($n = 50$) and referred to the Central Limit Theorem, either by name or description, as the justification for the sampling distribution of $\bar{x}$ being approximately normal even when the population distribution is nonnormal. Many students mentioned the sample size without referring to the Central Limit Theorem, and some mentioned the Central Limit Theorem without noting the sample size.
- Again, many students used the word "it" ambiguously (e.g., "it is normal because *n* is large"). This response does not make clear whether the student is (correctly) referring to the distribution of sample means or (incorrectly) referring to the distribution of fish lengths in either the sample or the population.
- Some students mistakenly referred to the Central Limit Theorem as the Law of Large Numbers. The Law of Large Numbers does not pertain to a normal distribution, so it is not relevant for this problem.

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?

The sampling distribution of a sample mean is a major topic in almost every introductory statistics course. However, students have a considerable amount of trouble understanding and applying sampling distributions. They tend to rely on generalizations like "larger samples result in less variability" and "if the sample size is large, then the distribution is approximately normal." Students need to focus on what variability is decreased and what distribution becomes approximately normal. In short, memorizing brief and ambiguous generalizations and using them on this type of question will result in a low score. Asking students to support their probability calculations with clearly labeled sketches of the appropriate distribution would be extremely beneficial.

Question 4

What was the intent of this question?

This statistical inference question was designed to assess a student's ability to distinguish paired-data procedures from two-sample procedures and to execute the selected procedure. Providing a complete statistical justification is an important skill that was being evaluated with this standard inference problem.

How well did students perform on this question?

The mean score was 1.07 out of a possible 4 points. The overall student performance on this standard inference question was not as good as it has been in previous years. Many students made serious errors setting up the hypotheses, stating and checking conditions, calculating the appropriate test statistic and *p*-value, and/or making a conclusion in the context of this study.

What were common student errors or omissions?

Part 1

- Students used nonstandard notation.
- Students failed to define terms used in the context of the problem.
- Students used a one-sided alternative.

Part 2

- Many students identified an incorrect test, such as two-sample *t*, chi-square, regression test, *z*-diff, or two-sample *z*.
- Many students did not check conditions.
- Some students who dealt with conditions simply listed what they had memorized (e.g., np >10).
- Some students discussed their conclusion from a graph without actually providing a sketch of the graph.
- Many students did not provide linkage between the graph and their conclusion.
- Some students graphed the data instead of the differences, even when conducting a paired *t*-test.

Part 3

- Students used incorrect mechanics for the identified test.
- Students used a confidence interval approach without reporting the confidence level.
- Students computed one-sided *p*-value instead of two-sided *p*-value, forgetting to double the area obtained from the table.
- Students provided an incorrect formula for the *t*-statistic.

Part 4

- Many students did not provide linkage between the *p*-value and α .
- Many students interpreted the *p*-value without referring to the conditional part of the probability.
- Some students accepted the null hypothesis.
- Many students provided a conclusion without context.

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?

Distinguishing paired data from two sample data requires practice. The more examples and exercises students see, the more comfortable they will become with making the appropriate choice. Students still have trouble realizing that certain conditions must be satisfied in order to use inference procedures. These conditions must be checked whenever statistical inferences are made. Asking students to provide a thorough rationale for their decision and to explain their conclusion in the context of the study would be extremely beneficial for exercises and assignments dealing with statistical inference.

Question 5

What was the intent of this question?

The primary goals of this statistical interference question were to assess a student's ability to: (1) distinguish an observational study from an experiment; (2) state the appropriate hypotheses for a research problem; (3) check the appropriate conditions for an inference procedure; and (4) interpret standard results for an inference procedure that is unfamiliar to students.

How well did students perform on this question?

The mean score was 1.1 out of a possible 4 points. The overall student performance on this inference question was similar to the performance on question 4, but not as good as it has been on inference questions in previous years. Many students had trouble stating the hypotheses, recognizing that one of the conditions for inference was not met, and making a conclusion in the context of this study based on the *p*-value provided. Many students did not realize this was an experiment.

What were common student errors or omissions?

Part (a)

- Many students indicated incorrectly that this was not an experiment because there was no control group.
- Some students used an incorrect definition of an experiment to justify their answer (e.g., stating experimenters imposed the simulation or task).
- A few students discussed experiment and observational study in such a way that it was not clear which they were selecting.

Part (b)

- Some students stated the hypotheses in words that parroted the question (e.g., they did not reference population proportion [parameters]).
- Some students used a two-tailed alternative instead of a one-sided alternative.
- Some students used "more distracted" in H_0.
- Some students stated hypotheses in terms of the sample instead of a population.
- Some students used "significant difference" in the hypotheses.

Part (c)

- Many students discussed SRS but not random assignment to treatment groups.
- Many students did not show calculations for normality check (np and n(1-p)) or compare to a value (5 or 10) .
- Some students used "x" to mean that the condition was not met; this is poor communication.
- Many students addressed the two-sample *z* procedure for means, not proportions.

Part (d)

- Many students did not write answers in context.
- Many students did not include the condition, "Assuming the null is true … ," in their interpretation of *p*-value.
- Some students interpreted the *p*-value using "as extreme as . . ." instead of "as extreme *or more extreme* as . . ."
- Some students are still "accepting H_0" instead of "failing to reject H_0."
- A few students are "rejecting H_a" instead of "failing to reject H_0."
- Students were unclear on their decision, writing, for example, "*p*-value is less than .05, so fail to reject" (fail to reject what?).
- Students were unclear on their conclusion (e.g., "fail to reject, so phones *are* more distracting").

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?

Control groups are extremely beneficial in many situations, but not all experiments have control groups. Students must recognize that the lack of a control group does not necessarily change an experiment to an observational study. The primary distinction between an observational study and an experiment is that the researchers impose treatments in an experiment. One of the striking results for many introductory students is that *p*-values from very different settings have similar interpretations. Asking students to provide thorough explanations of *p*-values for different procedures in a variety of settings with real data will help them recognize the similarities and differences.

Question 6

What was the intent of this question?

This question was designed to evaluate a student's ability to make inferences for simple linear regression models. Interpreting model parameters and comparing and contrasting different models are important skills that were also being assessed. Finally, a multiple regression model with a special variable, an indicator variable, was introduced to investigate whether the relationship between the predictor and the response variable differs for two different groups of people. Students were asked to sketch the estimated line for both groups and interpret the estimated parameters in the multiple regression model.

How well did students perform on this question?

The mean score was 0.79 out of a possible 4 points. This was the lowest scoring question on the exam and one of the lowest scoring investigative tasks over the last few years. Students could not interpret the estimated slope in the context of this study. Many students correctly picked the appropriate model but did

not provide a justification, or the justification they provided was not based on sound statistical reasoning. In part (c) many students did not realize they were being asked to conduct a statistical test. Finally, students did not label the estimated regression lines, and they could not interpret the estimated coefficients.

What were common student errors or omissions?

Part (a)

- Students made a deterministic statement, such as, "For every increase of 1 foot in the actual distance, a person will perceive an additional 1.08 feet," rather than a statement like, "For every increase of 1 foot in the actual distance, the model predicts that a person will perceive an additional 1.08 feet," or "For every increase of 1 foot in the actual distance, on average a person will perceive an additional 1.08 feet."
- Students interpreted the estimated slope of 1.08 in such a way as to imply that to get the perceived distance, one can multiply the actual distance by 1.08, thereby ignoring the *y*-intercept (e.g., "For every foot apart the objects are placed, the subject tends to perceive that the objects are 1.080 feet apart").

Part (b)

- Students stated, without any justification, that Model 2 is better because a simpler model is easier to deal with. One correct justification was to say that if the two objects are placed side by side (so that the actual distance is zero), then we would expect the subjects to perceive that the distance between the objects is zero. A second justification was to use the given standard error of 0.260 to argue that the *y*-intercept in Model 1 is not statistically significant, so the researcher is justified in going with the simpler model.
- Students misinterpreted the meaning of standard error.

Part (c)

- Students failed to understand that a test of significance is needed.
- Students tested the null hypothesis that $\beta = 0$. No response that tested $\beta = 0$ in part (c) received a score of 4.
- Students stated the null hypothesis, $\beta = 1$, in words that seemed to imply that the null hypothesis is that each person will perceive the distance exactly (e.g., "There is no difference between the actual distance and the perceived distance." Better wording would have been, "For every increase of 1 foot in actual distance, on average people perceive an increase of 1 foot.").
- Students divided the standard error by $\sqrt{40}$ in the denominator of the test statistic.
- Students used incorrect degrees of freedom. Here, only one parameter is being tested, so there are 40 – 1 = 39 degrees of freedom.
- Students omitted the formula, either in symbols or with numbers substituting the variables.
- Students "accepted" the null hypothesis.
- Students used a two-sided alternative hypothesis.
- Students failed to define the notation they used, or they used nonstandard notation such as μ for the slope.
- Students provided a conclusion that was not in context.

Part (d)

- Students failed to label which line was for contact wearers and which was for non-contact wearers.
- Students graphed lines inaccurately.

Part (e)

- Students correctly interpreted the coefficients of the two equations graphed in part (d), 1.05 and 1.17, but failed to interpret 0.12 separately.
- Students interpreted the slope in a way that does not reflect an understanding that it is a prediction, estimate, or average and does not exactly determine the predicted distance for each subject.

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?

Interpreting estimated regression coefficients is a very important skill. Exposing students to different regression models and asking for detailed explanations of the estimated model will help reinforce this skill. Investigative tasks require a different level of understanding. Students must think about their responses to each part and try to pull the information together with the statistical knowledge they have gained in the course. Practice with case studies and real data is essential.

General Comments on Exam Performance

Overall student performance on the multiple-choice questions was up slightly from a very low point in 2006. The average score on the free-response questions was also up from 2006. While both averages for 2007 are above the corresponding averages for 2006, they are still lower than the corresponding averages for 2002 through 2005. The investigative task clearly has been one of the most challenging questions in the last few years. However, students did not perform well on the exploratory data analysis question, the probability question, and the inference questions. This is cause for concern as the program continues to grow.

General Recommendations for Teachers

One general recommendation is clear from this exam: students need more practice explaining major statistical concepts in the context of a study. Students were asked to interpret the standard deviation, identify the appropriate blocking variable, apply the sampling distribution for a sample mean, interpret a *p*-value, and interpret estimated coefficients. The overall scores indicate that students had trouble with these interpretations, explanations, and applications of major statistical concepts. Future exams will focus on different statistics or different applications of these concepts, and students must be able to communicate their knowledge of statistics to the Exam Readers. As teachers, we must try to help our students avoid ambiguous language and provide clear, thorough responses.

Another major observation from this exam is that students are relying too heavily on generalizations that they may or may not understand. Students who say “larger samples reduce variability” may understand exactly what is happening in the context of the question, but they will be graded on the clarity and completeness of their response. The Reader cannot add words to a student’s response; the Reader must score what has been written. Students must be precise when responding to specific questions about statistical concepts.

Finally, students need more practice with all aspects of statistical inference. The scoring guidelines for statistical inference questions have remained the same for years, but student performance on these questions is not improving.