CollegeBoard
Advanced Placement Program

Student Performance Q&A:

2006 AP® Statistics Free-Response Questions

The following comments on the 2006 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 primary goals of this question were to: (1) assess a student's ability to use simple graphical displays (dotplots in this case) to compare and contrast two distributions; and (2) evaluate a student's ability to recognize what statistical information is most useful in making different practical decisions.

How well did students perform on this question?

The mean score was 2.11 out of a possible 4 points. On the whole, students performed well on this question that required interpretation of comparative graphical displays. Most students included comments about shape, center, and spread (and even outliers) in their discussions of similarities and differences between the two distributions. In addition, a majority of students used the graphical information to make practical decisions based on their understanding of the properties of center and spread in the context of the question.

What were common student errors or omissions?

Part (a):

- Some students did not comment on any similarities and differences but rather provided a separate list of descriptors for shape, center, and spread for the two distributions, with no comparison between them. Quite a few did not address all three characteristics—shape, center, and spread—on this comparative data analysis question.

- Several students gave vacuous comments about similarities and differences in the distributions, such as "The shapes of the distributions are similar (or different)," with no statistical evidence to support the statement. They also gave weak comparisons (e.g., "the measures are different") without commenting on the nature of the difference.
- Students found it difficult to describe the shape of these dotplots. They seem to have acquired a very limited vocabulary for describing shapes of distributions. Many students used nonstandard and/or incorrect terminology (e.g., "evenly distributed") in describing the shapes of the two distributions.
- In some responses, students used the word "spread" as synonymous with "range." Range is one way to measure spread, as is standard deviation, or IQR. Moreover, some students misused "range" by giving an interval of numbers. Range is a single number that is calculated using Range = Maximum – Minimum.
- A number of students focused too heavily on the modes of the two distributions when commenting on center or shape. Generally speaking, the median or mean are better measures of the center of a distribution and should be used instead of the mode(s).
- Although students were not required to state specific numerical values for measures of center or spread, many students were penalized for giving an incorrect value of a chosen statistic, e.g., "Catapult A's distribution has a median of 135."

Part (b):

- Several students did not use appropriate statistical terminology in explaining the smaller variability in the distances traveled by balls launched with catapult B. Students said colloquial things about catapult B's distribution—"more consistent," "more reliable," "more accurate," or "less sporadic"—instead of the statistically preferable "less variable."

Part (c):

- Students who used the mean or median distance from the target line for placing catapult B found it difficult to explain why they had chosen this location based on a statistical property of the mean or median. Few responses addressed the fact that the mean (or median) would provide a good summary of the center of a roughly symmetric, somewhat mound-shaped distribution.
- Not enough students focused on the goal of maximizing the probability of having balls land in the shaded band, arguing instead that their positioning of the catapult resulted in a reasonably high (but not necessarily the highest) proportion of balls landing in the band.
- Some students thought that the width of the shaded band was 10 cm instead of 5 cm.
- A number of students gave the distance to the front or back of the shaded band rather than the distance to the target line as requested.
- Several students picked a distance of 137 cm from the target line because 137 is one of the modes of catapult B's distribution, without any consideration of the proportion of balls that might land in the target band.

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?

Many students' responses were longer than they needed to be. The more students write, the more likely it is that they will make a statistically inaccurate statement. Students should be encouraged to answer each question completely but succinctly and then to move on to the next question. In parts (b) and (c) the majority of students did not convey that they were using information from these *samples* of shots to draw conclusions about the *populations* of shots that could be fired from these catapults. Teachers should be sure that students understand the differences between a sample and a population. Finally, many students spent unnecessary time keying data into their calculators; students should read through the entire question first and not immediately begin keying in data until they have determined whether it is necessary.

Question 2

What was the intent of this question?

The primary goal of this question was to assess a student's ability to identify the estimated regression line and to identify and interpret important statistics from regression output provided by statistical software in the context of a practical problem.

How well did students perform on this question?

The mean score was 0.46 out of a possible 4 points. Most of the scores were in the 0–2 range. More students tended to earn points for part (a), while parts (b) and (c) of this question presented a challenge.

What were common student errors or omissions?

Part (a):

- A significant number of students could not read the correct values from the computer output.
- Of those students who earned at least partial credit on this part of the question, most presented the correct variables in the model and described them appropriately. Few defined only one.
- Many students did not use the standard fitted regression notation either by words (estimation, prediction) or by notation ($\hat{y}$) in their response.
- Some confused variables with the parameters.

Part (b):

- Many students described the standard deviation of the regression as a measure of variability in a single variable.
- Students did not recognize this variability as variability about the line or as variability in the response variable $y =$ height of soapsuds at a given amount of detergent, x.
- Students could not give a meaningful interpretation of the standard deviation of the regression. Many students tried to interpret the standard deviation by assessing its size.
- Students frequently did not give any context in their response, as they were instructed to do.

Part (c):

- Some students had difficulty identifying the correct number from the computer output.
- Students related the standard error of the slope to deviations from the line (residuals) rather than variability in slope estimates.
- Very few students successfully identified the source of variability in the slope, failing to recognize that the slope estimate is a statistic subject to variation in repeated sampling.

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?

Many students were able to correctly identify the requested values in computer output but were unable to interpret those values in a meaningful way in the context of the question. Teachers should be sure that students are able to connect the numbers in computer output to the context of the question and can correctly interpret the meaning of those numbers in context.

Question 3

What was the intent of this question?

The primary goals of this question were to assess a student's ability to: (1) recognize the random variable of interest, identify its probability distribution, and calculate a probability for a linear combination of a normal random variable and a constant; (2) use basic probability rules to find a different probability; and (3) use the sampling distribution of the sample mean to find a probability about the mean of three observations.

How well did students perform on this question?

The mean score was 0.64 out of a possible 4 points. Students did not seem to understand that they had to use the answer in part (a) to answer part (b), or they had no idea what to do in part (b). Students also seemed to have difficulty answering part (c). They did not know that the distribution for the mean of the three independent depth measurements was also normal or how to calculate $\sigma_{\bar{x}}$.

What were common student errors or omissions?

Parts (a) and (c):

- Many students calculated the correct probability but showed little or no support, or just calculator commands.
- Students misunderstood the question, thinking it was a hypothesis test, not a probability distribution problem.
- Students used statistical terminology incorrectly in the solution (e.g., *p*-value, *z*-test) and incorrect mathematical notation such as $-1.33 = z = 0.0918$.
- Many students seemed to think that the distribution was discrete. Instead of using $P(Z < -2)$, they used $P(Z < -2.1)$.

- A number of students wrote incorrect statements, such as $E > -2, E < 2, P(E < -2), P(E < -2.1),$ etc. A sketch of the distribution would have greatly helped these students.
- Students used the wrong tail in the calculation of the probability or some calculated negative probabilities, not recognizing that this was a problem.

Part (b):

- Some students recognized the need to use the result from part (a) but did not answer the question asked. They calculated the probability of 1 of 3, not P(at least 1).
- Students used probability rules but made errors such as omitting the binomial coefficients.

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 who used a drawing in their solution were more likely to understand what was being asked and made fewer mistakes. Encouraging students to represent the probability with a drawing may help them better internalize the meaning of that probability. Students should also be reminded that relying solely on calculator syntax as a way to justify an answer is not acceptable.

Question 4

What was the intent of this question?

The primary goals of this question were to evaluate a student's ability to: (1) identify and compute an appropriate confidence interval, after checking the necessary conditions; (2) interpret the interval in the context of the question; and (3) use that confidence interval to conduct an appropriate test of significance.

How well did students perform on this question?

The mean score was 1.04 out of a possible 4 points. Few students were able to state and assess all of their chosen method's assumptions. Most students showed a correct confidence interval; however, the supporting details were often spotty. Students' interpretative statements about their intervals sometimes went awry if they attempted to say more than necessary. For the test inference, it was encouraging to see how many students based their decision on the absence (or presence) of 0 in their interval. However, here again, errors were introduced by an attempt to say too much.

What were common student errors or omissions?

- Students failed to identify the method they were using or selected an incorrect method (either due to not having studied 2-sample T-methods or having chosen an incorrect method).
- Students failed to assess the normality of *each* sample mean's sampling distribution.
- Students confused statements about exactly which distribution is approximately normal.
- Students failed to mention the required independence of samples.

- Students omitted or presented incorrectly the interpretative statement for the confidence interval.
- Students tried to explain the meaning of confidence *level* (not requested by the problem) and erred in doing so.
- Students tried to do a direct test of hypotheses rather than basing an answer on the interval (as requested).
- Students failed to recognize that the 2-sided confidence interval that they generated in part (a) should only be used for a test of inference with the 2-sided alternative hypothesis

  $H_a : \mu_a - \mu_S \neq 0.$
- Most students failed to verify 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?

Generally, students should be sure to answer the question that is being presented and include supporting work that is consistent with their conclusions and/or final numerical results. For questions that require the application of a particular statistical test or procedure, students should understand which test or procedure is appropriate for the given situation and be able to justify its use by verifying appropriate assumptions or conditions.

Question 5

What was the intent of this question?

The primary goals of this question were to evaluate a student's ability to: (1) identify the treatments in a biological experiment; (2) present a *completely randomized design* to address the research question of interest; (3) describe the benefit of limiting sources of variability; and (4) describe the limitations to the scope of inference for the biologist.

How well did students perform on this question?

The mean score was 1.00 out of a possible 4 points. For each part, large numbers of students had correct responses, but it was unusual for one student to respond correctly on all four parts. Although part (c) was the most challenging for students, significant numbers of students missed each part. Overall, students performed best on part (d).

What were common student errors or omissions?

Part (a):

- Students often listed the three nutrients and two salinity levels, giving 5 treatments. Even though these 5 treatments could have been listed in part (a), students sometimes used a tree diagram to illustrate the 6 treatments in part (b), indicating a lack of understanding of what constitutes a treatment when more than two factors are present.

- Some students introduced a "no nutrient" level and a "no salinity" level, leading them to have 12 instead of 6 treatments.

Part (b):

- Students frequently did not recognize that tanks were the experimental units and that treatments had to be randomized to tanks (not shrimp). They often gave a detailed description of the randomization of shrimp to tanks. Because students had been told that the shrimp were randomly assigned to the tanks, this information was considered extraneous.
- The process of randomization of treatments to tanks was often omitted or not presented in enough detail. When detail was given, some randomization processes did not ensure that exactly two tanks would be assigned to each treatment.
- Sometimes students incorrectly referred to a two-stage randomization process (e.g., random assignment of salinity levels to tanks followed by a random assignment of nutrients within salinity levels) as blocking.

Part (c):

- The most common error was the improper use of "confounding variable" or "lurking variable." When students identified the advantage of reduced variability, they often did not express why this was an advantage.

Part (d):

- Students did not always explain that the inability to generalize was because other shrimp species may have responded differently to the treatments.

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 answering statistical design questions, students should not use terminology that is not appropriate for the situation. For example, in this question many students referred to 'confounding'; however, in a completely randomized design, confounding is not possible. While students seemed to know that some type of randomization was necessary, the level of their understanding of that process was often very minimal. They either were unable to provide additional details or described randomizations of treatments that were incorrect or not possible. Helping students to understand exactly what the treatments are in an experiment and the reasons (and details) of the related randomization(s) may help them to improve their performance on design questions.

Question 6

What was the intent of this question?

The primary goals of this question were to evaluate a student's ability to apply the concepts of significance testing to a new setting; in particular to: (1) state hypotheses for a parameter of interest, given a research question; (2) evaluate a new test statistic and use the probability distribution associated with that statistic to test the hypotheses of interest; (3) identify the values of the test statistic that would lead to rejection of the null hypothesis on a graph; and (4) interpret simulated sampling distributions for different populations.

How well did students perform on this question?

The mean score was 0.83 out of a possible 4 points. Although there were not many blank responses to this question, student performance was disappointing. Many students did not seem to realize they needed to use the earlier parts of the question to help them answer the later parts, instead viewing parts (a) through (f) as independent. Also, some students seemed to forget that question 6, the investigative task, is one in which they will not only use their knowledge of statistical concepts and principles (e.g., significance testing) but one where they should be prepared to integrate concepts in new ways.

What were common student errors or omissions?

Part (a):

- There were two population variances in the question—the variance of the readings of the population of thermostats in the past (known to be 1.52 degrees Fahrenheit squared) and the variance of the readings of recently manufactured thermostats. Students often wrote the null hypothesis, for example, in one of these forms: "The variance of the "true" population is $\sigma^2 = 1.52$ (or $\sigma^2 = 1.52$)" (with no definition of σ^2). In such cases, it was not clear that the student knew that σ^2 stands for the variance of the readings of recently manufactured thermostats. Symbols used in hypotheses should be appropriate (σ^2 for population variance, for example, and not μ or s^2) and should always be precisely defined. Hypotheses that used the symbol *s* made it appear that the hypotheses referred (incorrectly) to the sample.
- Occasionally, the hypotheses were incorrectly written as if this were a two-sample test $\sigma_1^2 = \sigma_2^2$ or a two-sided test.

Part (b):

- Some students incorrectly used their calculator to compute a variance with n as the divisor rather than the correct $(n - 1)$ for a sample variance.

Part (c):

- Some students computed an incorrect test statistic using a formula from a discrete χ^2 test.
- Many students omitted either one or both of the following—conclusions for significance tests with linkage to the *p*-value (or to the test statistic and critical value), or conclusions in terms of the context of the situation.

- Some students did not understand that the *p*-value refers to a tail of the distribution and is a probability computed by assuming that the null hypothesis is true. That is, the following interpretation of the *p*-value is incomplete:

  "The *p*-value of 0.21 indicates that it is not unlikely, just by chance, to get a sample variance such as ours."

  A more complete interpretation would be:

  "The *p*-value of 0.21 indicates that it is not unlikely, just by chance (or, better, just by variability in sampling), to get a sample variance as large as or even larger than ours, given that the variance of recently manufactured thermostats remains at 1.52."

- Most students were able to find the correct *p*-value from the test statistic computed in part (b), but often the *p*-value was not linked to the conclusion. Linkage could have been achieved by appealing either to a rejection region or to the strength of the evidence against the null hypothesis.

- Some students wrote "Accept H_0" or the equivalent, such as stating that the variance of the recently manufactured thermostats was still 1.52. Such a statement is too strong and the conclusion was scored as incorrect.

Part (d):

- Many students who could not find the critical value of 16.92 from the table or estimate it from the χ^2 CDF function of their calculator realized that the value requested must be the cutoff point for the upper 5 percent of the distribution and so marked a reasonable estimate on the graph. This estimate could then be used for full credit in parts (e) and (f). This is a good example of how students who understand the "flow" of an investigative task can perform very well overall, even if they are unable to complete every part of the question perfectly.

Part (e):

- Some students failed to understand that the simulated sampling distributions were from populations where the variance was larger than 1.52 (even though that was clearly stated) and hence could not make the connection that the null hypothesis was false, so test statistics that do not fall to the right of 16.92 would result in a Type II error.

- Many students marked only the critical value of 16.92 and failed to identify a region in the right tail by shading or circling it.

Part (f):

- Almost all students were able to select Histograms III and II as the ones representing the populations with the largest and smallest variance, respectively. However, most justifications were weak, revealing little understanding that these histograms are approximate sampling distributions or that the regions represent likelihood of rejecting the (incorrect) null hypothesis. Typically, the justification for selecting Histograms III and II referred only to the spread of those histograms themselves: "Histogram III represents the population with the largest variance because it has the largest spread." There was rarely a connection to why the population with the variance farthest

above 1.52 would result in a sampling distribution of this test statistic with the largest region to the right of 16.92. A complete answer to part (f) should refer to the regions identified in part (e) and make it clear what the sizes of the regions represent—that the further the population variance is above 1.52, the larger the test statistic tends to be, resulting in more values above 16.92 and so a greater probability of (correctly) rejecting the null hypothesis (the concept of power).

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?

As is true with all investigative tasks, students should answer each part completely, realizing that they may need to use information from earlier parts of the question to respond correctly to the later parts; students should also understand that they will need to use concepts in new ways in responding to the later parts of such questions.

General Comments on Exam Performance

Overall performance on the multiple-choice questions was down from 2005, and in fact it was the lowest in the past five years. Scores on the free-response questions were similarly down (significantly) from 2005, and they also were the lowest in the past five years. While there were some challenges in parts of the questions, more than the usual number of students (as compared with past years) tended to earn lower scores on even the more straightforward parts of the exam, such as presenting correct mechanics in a hypothesis testing situation, or stating conclusions and findings in the context of the question.

General Recommendations for Teachers

Whether the student is answering questions that focus on comparing and contrasting distributions, conducting a test of significance, interpreting statistical results, or providing information about an experiment, some of the same recommendations apply.

- Students should always read each question completely first, think about what is being asked, and respond using statistical justifications. This means not only using correct statistical content but also using correct statistical vocabulary.
- Questions with more than one part (i.e., parts a, b, c, etc.) are often structured to familiarize the student with the question's setting in the early parts so that they can build on that information for the later parts. That is, students are expected to use their experience from answering the earlier parts to help them to answer the later parts of the question.
- Providing an interpretation of one's results and/or findings is always expected in every question, and that interpretation should always be presented in the context of the question. Numerical results that are not tied to a relevant context are meaningless.
- As students are progressing through the AP Statistics course, they need to realize (and be reminded regularly) that their ever-expanding understanding of statistics will require them to think critically when they are faced with new statistics problems. They should not, for example, blindly attempt to recall a process or procedure when faced with a question but instead should determine whether the question is asking them to analyze distributions, comment on a sampling or design plan, conduct an inference procedure, or do something else. Helping students to develop this skill throughout the course will make them more ready and able to face unfamiliar and less routine questions in an exam situation.