AP STAT DEBRIEF – Question 2

Billy Esra, AP Statistics Teacher
Thomasville High School Scholars Academy

2016 QUESTION 2: Chi Square Test and Effectiveness of Treatments
Please view the questions here: https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap16_frq_statistics.pdf as well as additional question resources at http://apcentral.collegeboard.com/apc/members/exam/exam_information/8357.html

Question 2: INTENT OF THE QUESTION

"The primary goals of this question were to assess a student's ability to (1) identify, set up, perform, and interpret the results of an appropriate hypothesis test to address a particular question and (2) assess the effectiveness of treatments in a controlled experiment" (https://secure-media.collegeboard.org/digitalServices/pdf/ap/apcentral/ap16_statistics_q2.pdf).

SAMPLE SOLUTION:

This question referenced fake snacks called Apple-Zuties and Choco-Zuties, which made the question very memorable to students. It had the next to lowest mean score (1.22) but the highest standard deviation (1.21). Students had a hard time deciding which hypothesis test to use (or whether to use one at all). They had a hard time understanding that Group C was a control group that allowed them to detect advertisement effectiveness by comparing Group A and Group B to Group C. So, it was not unusual to read a response that took the student a lot of time to construct that earned them no points, or little to no points. Alternatively, it was quite easy for students who had a good grasp on chi-square procedures to earn points.

a. The student had to choose to carry out the appropriate test (chi-square test of homogeneity).
    - Students could identify the test as chi-square, chi-square test of homogeneity, or chi-square test of association/independence to name the correct test. (They could also get credit for naming if they had the chi-square symbol somewhere in their response.) Goodness of Fit was not accepted, though.
    - At least one of the hypothesis needed to be in context and the hypothesis had to represent the population, not the sample.
    - For conditions of inference:
        - Random Assignment explicitly stated so it did not have to be mentioned.
        - Expected counts were to be reported (but not as integers) and checked.
    - Student reported correct chi-square value, degrees of freedom, and p-value.

- Students linked the reported p-value to a correct conclusion about the alternative hypothesis in context.

b. The student had to correctly describe the effect of ad Type A and ad Type B (in context), compared to the control group.

NOTES/COMMON MISTAKES

1. Deciding which test to use was easy for most students that recognized that chi-square was the best choice, but some students spent a lot of time on non-conventional methods. Students could sometimes earn a score of 2 on the question if they used multiple 2-proportion z-tests, multiple 2-proportion confidence intervals, or three 1-proportion confidence intervals. Unfortunately, students who did this made mistakes in their execution of these alternative methods or elsewhere in the question and ended up with a score of 0 or 1 for the question. My guess is that some teachers did not make it to chi-square topic since it is often one of the last topics in most AP Statistics textbooks.
2. Students did not have to explicitly say chi-square for homogeneity or chi-square test for association/independence. In fact, they did not have to name the test at all. If they just typed $\chi^2 = 10.291$, that was enough to get credit for naming the test. They only lost points if they called it a chi-square goodness of fit test, which did not happen too very often.
3. A surprising number of students switched the null and alternative hypothesis.
4. Many students incorrectly found the expected counts by assuming that expected number would be the same for each cell in the table (dividing the total number in study by 6), instead of correctly using the formula for expected counts or expected value matrix in calculator. Also students were not allowed to round their expected counts to integers.
5. Many students lost points for conditions by stating something like “Simple Random Sample was stated.” In fact, there was not SRS, though the subjects were randomly assigned.
6. Students lost points by referring to the “independence of groups or ads” as being a condition that was met. Since the students were asked if there was an association between type of ad and choice, then they could not assume independence. Students could mention that children are independent with no penalty.
7. Students lost points by mentioning the normality condition, which is not appropriate for chi-square.
8. Students forgot to mention the degrees of freedom.
9. Linking the p-value and significance level to the decision in context proved difficult for many students. Students could give the correct interpretation of the p-value or explain how the conclusion follows from the p-value and not mention alpha levels (or significance levels) at all.

10. Very few students seemed to get full credit for part b. They had to realize that since the Choco-Zuties only distribution was very similar to the control group, the advertisement had little impact. And that since the Apple-Zuties only distribution was much different from the control group (since about half of the kids chose Apple-Zuties), the advertisement seemed to have an impact on the choice of snack. Few students were able to do this successfully. They wanted to say that both ads were ineffective (since predominantly chose the chocolate snack) or that they were both effective (since more chose chocolate in Group A and more chose apple in Group B than they did in the other groups). The biggest issue seemed to be that they did not understand that Group C was a control group (since it was not explicitly stated) and therefore the treatments should be compared to Group C. They just wanted to compare all three groups with one another.
11. Part b is all about making arguments with numbers. Using the counts or proportions to compare the effectiveness of the treatments. Students lost points for not including numerical justification, even if their description of the effectiveness of the ads was spot-on.

TEACHER SUGGESTIONS

When I first realized that I would be reading this questions, I was excited because I thought that students would be very successful with a chi-square question. Chi-square is often one of the last topics covered and students generally seem to do well with that topic in my classroom. I was very, very wrong (as I usually am about how students will respond to questions on the test). The problem with this question seemed to be that the stem made sense to students. They understood what they were being asked to do, they just didn't use the correct tools to craft their answers. They felt that they could put together some sort of semblance of a correct answer that would win them some points. And write they did. Many students filled all available space on both pages, but, unfortunately, much of what they wrote did not earn them points. They wrote paragraphs and paragraphs about choco-zuties and apple-zuties and probably spent a lot of time crafting their responses but …

As an AP Statistics teacher, I plan to do a better job at working with students on choosing the appropriate test after finishing up the inference unit. Perhaps I'm too quick to begin review, at large, and could help them with choosing the right test for the right situation. Students should have recognized the categorical variables and immediately thought chi-square when asked "provide convincing statistical evidence." But if they did not work through a chi-square test, they could, at most, get one point on this question (unless they successfully used one the other strategies mentioned earlier – but that was rare).

Perhaps students need a more formulaic method for organizing inference questions. I've started using a "four-corners" approach where students draw two intersecting lines to organize what goes where. I heard about this at the Best Practices night at the reading a few years ago,

and I've started using it with students. So far, anecdotally, it appears to be helping them remember all of the components of an inference procedure answer.

To much consternation of statistics teachers, the rubrics for inference procedure questions have not required students to use formulas. Just reporting the necessary statistics has been enough to earn credit for calculations. But if students do use formulas, they must use the correct formula and fill it in correctly.

Part b was difficult for students. The "Neither" group was not spelled out in the prompt as a control group, and I'm not sure it would have helped if it were spelled out. One thing is certain, I'll be including this type of discussion in the experimental design unit in my classes. We tell students that part of the experimental design process is comparing treatments, including comparing to a treatment group. But I cannot off-hand think of a time where I've actually asked them to explicitly do this.

Again, I loved this question. It was fun and approachable, but I really wanted students to perform better than they did, overall.