Special Focus: Inference

Review of the Assumptions

Let's review the assumptions behind each of the inference procedures in the AP Statistics curriculum and how they might be checked for their reasonableness. For all studies in which conclusions are to be generalized to the population from which the sample was drawn:

Assumption: The sample is a random sample from the population.
Check: This cannot be checked from the data. The reasonableness of the assumption must be assessed based on how the sample was collected. Since true random samples are difficult at best to collect, the reasonableness of the assumption often reduces to whether the sample was collected in such a way that it does not appear to bias the responses by over- or underrepresenting certain responses as compared to the whole population. Reasonable people may disagree about whether a sampling method produces a sample that is sufficiently "like" a random sample to allow generalization.

For studies involving proportions:
Assumption: The sample size(s) is/are large enough to reasonably ensure that the sampling distribution(s) of the sample proportion(s) involved is/are approximately normal.
Check: There are a number of different checks people may use. A common one is $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$. Note that in the case of a hypothesis test of a single proportion, the hypothesized population proportion should be used in the check; for the construction of a confidence interval, the sample proportion should be used.

For studies involving means:
Assumption: The sample size(s) is/are large enough to reasonably ensure that the sampling distribution(s) of the sample mean(s) involved is/are approximately normal.
Check: In AP Statistics, different graphic checks are possible. (Analytic checks exist but are not in the syllabus.) A fairly sophisticated one is the normal probability plot, in which linearity in the plot corresponds to normality in the data. Histograms and boxplots are more crude but may be sufficient so long as students can recognize deviations from normality, such as skew or heavy tails. If a sample size is quite small (say, less than 15), then indications of a deviation from normality, such as skew or outliers, are quite troublesome and may invalidate the inference procedures. If the sample size is a bit larger (say, between 15 and 40) then skew, outliers, or heavy tails are less of a problem unless they are fairly severe. And if the sample size is quite large (say, greater than 40), only exceptionally severe deviations from normality will cause problems. If two samples are involved, each of them must be checked.

Note that a normally distributed population is still the assumption even when using "*t*" procedures.

For chi-square tests:
Assumption: The sample is large enough that the test statistic has approximately a chi-square distribution.
Check: A common check is that all of the expected cell counts must be at least 1, and no more than 20 percent of the expected cell counts may be less than 5.

For linear regression:
Assumption 1: The underlying relationship between x and y is linear.
Check: The residual plot shows no pattern, particularly no clear curvature.

Assumption 2: The errors have the same standard deviation for all values of x.
Check: "Eyeballing" the residual plot is sufficient for AP Statistics students. Be sure the residuals are of roughly the same magnitude across all values of x. In particular, be sure that they do not tend to grow as the response variable grows. If they do, then a transformation of the data may be appropriate.

Assumption 3: The errors are normally distributed.
Check: Look at the distribution of the residuals, either using a normal probability plot (better), a histogram, or a boxplot (more crude, but adequate). Be sure the residuals do not display obvious deviations from normality.

What Are Students Expected to Write?

At this point, you may be thinking, "This section about assumptions seems awfully long. Sometimes the discussion about whether an assumption was reasonable went on for a paragraph or more. Surely on the AP Exam itself our students aren't expected to write *that* much. So what must they write about assumptions in the free-response section?"

Students should know and state what the assumptions are behind the models they are using. If it is possible to check the reasonableness of assumptions using the data, they are expected to do that as well.

Sometimes assumptions will be stated explicitly in a problem for the students. For example, a problem may state explicitly that the given data represent a random sample from some population. That may be done because the assumption should not be the

focus of students' energies. It doesn't hurt for the students to repeat the assumption if they are performing inference that requires it, but it would not be absolutely necessary if the assumption were explicitly stated in the problem.

If a student thinks one of the assumptions required for inference is violated, but the question appears to demand inference nevertheless, the student would be wise to write something indicating his or her dilemma, such as, "I would ordinarily not want to use a *t* distribution when the data are so grossly nonnormal, but since this question seems to require a confidence interval calculation, I don't know what else to do, so I'll do that." That at least indicates that the student understands the connection between assumptions and inference. Some teachers suggest to their students that if they find themselves in that situation, they write, "Because the assumptions are not met, I will proceed with caution." That also indicates that the student is aware that something is wrong. Contrast that with a student who sees a small set of data and writes, "Check for normality," then sketches a boxplot showing skew and two outliers and goes on with inference anyway. Such a student response may not get any credit for checking assumptions at all, even with the supposed check for normality, since that student didn't seem to know what the purpose was for the check nor recognize that the assumption was not reasonable.

In situations such as described in the preceding paragraph, i.e., necessary assumptions are not met, the student has in all likelihood made an error somewhere. The free-response questions will not demand the use of unjustifiable inference procedures from students when the procedures in the AP syllabus are invalid. Even the best of students can still stumble, but students should recognize something is wrong and communicate this awareness as a part of their response.

The entire focus of a free-response question may be the validity of the assumptions in a particular situation. Students who make a very regular habit of beginning every inference question with a thoughtful check of assumptions will know what to do. Consider, for example, question 2 of the 2000 AP Statistics Exam in which students were told of a cave containing footprints of prehistoric humans. The question gave sample statistics and asked students what assumptions were required to construct a confidence interval. It also asked whether the assumptions were reasonable. In that question, students should have thoughtfully considered whether the sample was a random sample or anything like one. They should have realized that it was not, that indeed, many of the footprints may have been from the same person, or perhaps that some footprints (from heavier people, perhaps?) may have been more likely to appear in fossil form than others.