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 $n\left(1 - \hat{p}\right) > 10$. 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.

2000 AP Statistics Free-Response Question 2

2. Anthropologists have discovered a prehistoric cave dwelling that contains a large number of adult human footprints. To study the size of the adults who used the cave dwelling, they randomly selected 20 of the footprints from the population of all footprints in the cave and measured the length of those footprints. Some statistics resulting from this random sample are as follows.

Sample size	20	Minimum	15.2 cm
Mean	24.8 cm	First quartile	18.7 cm
Standard deviation	7.5 cm	Median	21.5 cm
		Third quartile	30.0 cm
		Maximum	37.0 cm

The anthropologists would like to construct a 95 percent confidence interval for the mean foot length of the adults who used the cave dwelling.

(a) What assumptions are necessary in order for this confidence interval to be appropriate?

(b) Discuss whether each of the assumptions listed in your response to (a) appears to be satisfied in this situation.

Question 3(d) on the 2004 exam was similar. Students were asked whether conclusions from a sample of dinosaur bones could be extrapolated to all dinosaurs. They were to realize that the sample was not random and that the assumption that it was representative of all dinosaurs might well be unreasonable.

2004 AP Statistics Free-Response Question 3

3. At an archaeological site that was an ancient swamp, the bones from 20 brontosaur skeletons have been unearthed. The bones do not show any sign of disease or malformation. It is thought that these animals wandered into a deep area of the swamp and became trapped in the swamp bottom. The 20 left femur bones (thigh bones) were located and 4 of these left femurs are to be randomly selected without replacement for DNA testing to determine gender.

(a) Let X be the number out of the 4 selected left femurs that are from males. Based on how these bones were sampled, explain why the probability distribution of X is not binomial.

(b) Suppose that the group of 20 brontosaurs whose remains were found in the swamp had been made up of 10 males and 10 females. What is the probability that all 4 in the sample to be tested are male?

(c) The DNA testing revealed that all 4 femurs tested were from males. Based on this result and your answer from part (b), do you think that males and females were equally represented in the group of 20 brontosaurs stuck in the swamp? Explain.

(d) Is it reasonable to generalize your conclusion in part (c) pertaining to the group of 20 brontosaurs to the population of all brontosaurs? Explain why or why not.

If a free-response question does not appear to be focusing primarily on the assumptions (as in the cave problem) but is instead asking students to perform a hypothesis test or construct a confidence interval, then the students need not write paragraphs about the assumptions, but they should quickly and clearly state them and, if possible, check that that the assumptions are reasonable. (Some students erroneously think that "checking assumptions" means drawing a little check mark next to them. It does not! Checking assumptions means writing a sentence or an inequality or a graph or calculation that communicates in what way the data are consistent with the assumptions.)

Suppose the student is asked to construct a confidence interval estimate of the mean weight of apples from an orchard given the weights of eight apples randomly sampled from the orchard. Here is an example of a poor check of assumptions.

$np > 10$

$n(1-p) > 10$

$n < 15$, assume normal

Obviously the first two statements do not apply here in the context of means. Yet AP Statistics Readers will confirm that these show up like a talisman on a surprising number of responses to questions having nothing to do with proportions. Their presence indicates that the student doesn't know what these formulas are for; they may be trying a "shotgun approach," assuming (again, incorrectly) that the correct formulas will be found and the incorrect ignored. But even if that poor start is ignored, the third line still does not constitute a proper check of assumptions.

Writing "$n < 15$, assume normal" does not communicate much. A sentence stating *what* was being assumed to have a normal distribution (the population of apple weights) would have been an improvement. But with the eight data values given in the problem, the student would be expected to perform some kind of check to see whether the assumption that they came from a normal distribution was at least plausible.

Here is a much better check of assumptions:

"I will construct a *t*-interval estimate of the mean weight of the apples in the orchard. This requires that the sample be random, which the problem states. It also requires, with such a small sample size, that the population of apple weights is approximately normal. Here's a normal probability plot of the eight data values: [shows picture]. It's not grossly nonlinear, and the data set has no outliers, so the normality of the population is not contradicted by the data, so we'll assume it."

While delightfully complete, the response may seem awfully wordy, especially in the context of a timed exam. The following much shorter check communicates the same ideas:

"Sample is random (given). Assume apple weights in orchard are normally distributed. Check: [boxplot here], no skew or outliers. So assumption is reasonable."

Conclusion

Students can easily end up thinking that a check of assumptions is merely a pointless routine on the road to "doing the problem." This mindset can get them into trouble both on AP Exam questions and in their thinking about real statistical analyses they perform or read about. Statistics is largely about creating mathematical models to explain observations in the world. These models are never exactly right. Mother Nature is not so simplistic! However, the models can still be useful for understanding the world, so long as they are reasonably good. The "goodness" of the models depends largely upon the reasonableness of the assumptions behind them. Assumption checking isn't something that students do in AP Statistics only and from which statisticians have "graduated," leaving it behind. Rather, checking assumptions is an important part of every statistical endeavor: it is the assumptions that drive the model. Statisticians don't let statisticians drive without the mathematical support provided by assumptions!

AP Statistics teachers should encourage in their students the habit of always checking the assumptions behind the inference procedures they're using. One reason so many students think of checking assumptions as just a pointless routine is that they are rarely exposed to problems in which the assumptions are not met. I would encourage teachers to present problematic data to their students on a fairly regular schedule to keep them on their toes and help them see the point of checking assumptions. Sometimes that will mean concluding that the inference techniques of the AP Statistics curriculum are insufficient for a particular problem, and so it must be left unresolved. That's perfectly fine.

Also, students are far more likely to encounter assumptions that are not met when they design and conduct their own studies than when they read about studies in their textbooks. The examples presented in this section are not true stories but are motivated by actual classroom experience, describing the sorts of problems that students have encountered when they conduct their own studies.

When students conduct their own studies, the most common difficulty is devising a sampling method such that the assumption of a random sample, or "like" a random sample, is reasonable. Following that distantly is the collection of a sufficient amount of data for the limiting distribution assumption (normal, *t*, or chi-square) to be reasonable. Sometimes students come up with interesting studies for which the data are relatively easy to collect in large numbers. But other times the students realize that although their study's design is good, it will require far more time in data collection than they are able to spend.

Finally, the habit of thinking about and checking assumptions must not stop as a mental exercise or an oral discussion but must be something students are in the habit of putting in complete sentences on paper. This is part of the broader and very important goal of AP Statistics, that students be able to communicate their ideas clearly in writing. The study of *theoretical* statistics is a rich and beautiful endeavor, and like theoretical mathematics no context is required. However, AP Statistics is applied statistics, applied statistics does not exist without context, and statistical solutions to problems do not exist if they are not clearly communicated.