Is That an Assumption or a Condition?

by Dave Bock
Ithaca High School
Ithaca, New York

The nemesis of many an AP student (and teacher) is the requirement to test assumptions and/or conditions before using a statistical procedure. Each year many students who write otherwise very nice solutions to free-response questions about inference don't receive credit for a "Complete" response because they fail to deal correctly with the assumptions and conditions. AP Central's "Notes on the 2002 AP Statistics Free-Response Questions" (available in "See also," below) observes that many students "failed to provide conditions or gave an incomplete set of conditions for using the selected statistical test" or "listed the conditions for using the selected statistical test, but did not check them." That year was not unique -- this is a perennial issue. How can we help our students understand and satisfy these requirements?

Many students struggle with these questions:

- What *are* assumptions and conditions?
- What, if anything, is the difference between them?
- Why bother checking them?

Out of some combination of inattention, confusion, or annoyance, students often forget to write anything at all, or they write things that are incorrect, or they write everything they can think of whether relevant or not. Even students who seem to know what the assumptions and conditions are often just list them and make little check marks nearby. Apparently these students believe that the check marks somehow add authority to their list, when in fact they really need to show some work that actually demonstrates that the appropriate condition has been met.

You can help your students understand which assumptions and conditions apply and how to check them and -- most important -- assist students in understanding why this is an important step in doing statistics. Here are some suggestions about how to avoid, ameliorate, and attack the misconceptions and mysteries about assumptions and conditions.

- Start early. Assumptions and conditions aren't just for inference.
- Clearly distinguish between assumptions and conditions. They are not the same thing.
- Clearly connect the assumptions to the procedure you hope to perform. There's a reason why each is important.
- Pair each assumption with a corresponding condition (if there is one).
- Show students how to actually *check* the conditions. There's more to it than making those little check marks!

Start Early
Distinguish Between Assumptions and Conditions
Our Next Encounter: The 68-95-99.7 Rule
Regression Models
Bernoulli Trials
Let's Take Stock...
Inference for a Proportion
Inference for Means
Inference for the Difference of Two Means
Inference for Matched Pairs
Inference for Chi-Square
Inference for Regression
And That's That

Start Early

Inference is a difficult topic for students. The vocabulary is new, the reasoning can be confusing, and the interpretations are tricky. There are many new ideas on the table at once, and adding the issue of assumptions and conditions to the mix just makes things worse. Inference will be less daunting if you discuss assumptions and conditions from the very beginning of the course. Make checking them a requirement for every statistical procedure you do. For example...

1. What kind of graphical display should we make -- a bar graph or a histogram? The key issue is whether the data are categorical or quantitative. Students should always think about that before they create any graph. If they decide on a pie chart or a bar graph, require that they write down the...

   Categorical Data Condition: These data are categorical. (Of course, in the event they decide to create a histogram or boxplot, there's a **Quantitative Data Condition** as well.)

2. While it's always okay to summarize quantitative data with the median and IQR or a five-number summary, we have to be careful not to use the mean and standard deviation if the data are skewed or there are outliers. Don't let students calculate or interpret the mean or the standard deviation without checking the...

   Not Skewed/No Outliers Condition: A histogram shows the data are reasonably symmetric and there are no outliers. Note that students must *check* this condition, not just state it; they need to show the graph upon which they base their decision.

Of course, these conditions are not earth-shaking, nor critical to inference or the course. They serve merely to establish early on the understanding that doing statistics requires clear thinking and communication about what procedures to apply and checking to be sure that those procedures are appropriate.

Distinguish Between Assumptions and Conditions

The plot thickens, and soon. All of mathematics is based on "If..., then..." statements. Students have seen such statements before. For example, they know (we hope) not to apply the Pythagorean theorem unless they have a right triangle. The fact that it's a right triangle is the assumption that guarantees the equation $a^2 + b^2 = c^2$ works, so we should always check to be sure we are working with a right triangle before proceeding.

The same is true in statistics -- we just don't see the theorems themselves very often. All those theorems, though, are in that same "If..., then..." form. The "If" part sets out the underlying assumptions used to prove that the statistical method works. If those assumptions are violated, the method may fail. The assumptions are about populations and models, things that are unknown and usually unknowable. And that presents us with a big problem, because *we will probably never know whether an assumption is true*.

In fact, there are three types of assumptions:

1. Unverifiable. We must simply accept these as reasonable -- after careful thought.
2. Plausible, based on evidence. We test a condition to see if it's reasonable to believe that the assumption is true.
3. False, but close enough. We know the assumption is not true, but some procedures can provide very reliable results even when an assumption is not fully met. In such cases a condition may offer a rule of thumb that indicates whether or not we can safely override the assumption and apply the procedure anyway.

A condition, then, is a testable criterion that supports or overrides an assumption.

Our Next Encounter: The 68-95-99.7 Rule

On a recent AP Exam, students were given summary statistics about a century of rainfall in Los Angeles and asked if a year with only 10 inches of rain should be considered unusual. Many students observed that this amount of rainfall was about one standard deviation below average and then called upon the 68-95-99.7 Rule or calculated a Normal probability to say that such a result was not really very strange. Those students received no credit for their responses. The other rainfall statistics that were reported -- mean, median, quartiles -- made it clear that the distribution was actually skewed. Students should have recognized that a Normal model did not apply. (The correct answer involved observing that 10 inches of rain was actually at about the first quartile, so 25 percent of all years were even drier than this one.)

Students will not make this mistake if they recognize that the 68-95-99.7 Rule, the z-tables, and the calculator's Normal percentile functions work only under the...

Normal Distribution Assumption: The population is Normally distributed.

That's a problem. We can never know whether the rainfall in Los Angeles, or anything else for that matter, is truly Normal. We never see populations; we can only see sets of data, and samples never are and cannot be Normal. However, if the data come from a population that is close enough to Normal, our methods can still be useful. We can plot our data and check the...

> **Nearly Normal Condition:** The data are roughly unimodal and symmetric.

Require that students always state the Normal Distribution Assumption. If the problem specifically tells them that a Normal model applies, fine. If not, they should check the nearly Normal Condition (by showing a histogram, for example) before appealing to the 68-95-99.7 Rule or using the table or the calculator functions. Not only will they successfully answer questions like the Los Angeles rainfall problem, but they'll be prepared for the battles of inference as well. With practice, checking assumptions and conditions will seem natural, reasonable, and necessary.

Regression Models

Least squares regression and correlation are based on the...

> **Linearity Assumption:** There is an underlying linear relationship between the variables.

Students should not calculate or talk about a correlation coefficient nor use a linear model when that's not true. As always, though, we cannot know whether the relationship really is linear. We can, however, check two conditions:

> **Straight Enough Condition:** The scatterplot of the data appears to follow a straight line.
> **Outlier Condition:** The scatterplot shows no outliers.

Bernoulli Trials

Tossing a coin repeatedly and looking for heads is a simple example of Bernoulli trials: there are two possible outcomes (success and failure) on each toss, the probability of success is constant, and the trials are independent. We can use binomial probability models to calculate probabilities of certain outcomes, but before applying such methods we must make the...

> **Independent Trials Assumption:** The trials are independent.

If we are tossing a coin, we assume that the probability of getting a head is always $p = 1/2$, and that the tosses are independent. This assumption seems quite reasonable, but it is unverifiable. There's no condition to test; we just have to think about the situation at hand.

Things get stickier when we apply the Bernoulli trials idea to drawing without replacement. We face that whenever we engage in one of the fundamental activities of statistics, drawing a random sample. Then the trials are no longer independent. Just as the probability of drawing an ace from a deck of cards changes with each card drawn, the probability of choosing a person who plans to vote for candidate X changes each time someone is chosen. We don't really care, though, provided that the sample is drawn randomly and is a very small part of the total population -- commonly less than 10 percent. Note that in this situation the Independent Trials Assumption is known to be false, but we can proceed anyway because it's close enough. We need only check two conditions that trump the false assumption:

> **Random Condition:** The sample was drawn randomly from the population.
> **10 Percent Condition:** The sample is less than 10 percent of the population.

When we are dealing with more than just a few Bernoulli trials, we stop calculating binomial probabilities and turn instead to the Normal model as a good approximation. A binomial model is not really Normal, of course. After all, binomial distributions are discrete and have a limited range of from 0 to n successes. Normal models are continuous and theoretically extend forever in both directions. Nonetheless, binomial distributions approach the Normal model as n increases; we just need to know how large an n it takes to make the approximation close enough for our purposes. We can trump the false **Normal Distribution Assumption** with the...

> **Success/Failure Condition:** If we expect at least 10 successes ($np \geq 10$) and 10 failures ($nq \geq 10$), then the binomial distribution can be considered approximately Normal. (Note that some texts require only five successes and failures.)

Remember, students need to *check* this condition using the information given in the problem. Simply saying "$np \geq 10$ and $nq \geq 10$" is not enough. If, for example, it is given that 242 of 305 people recovered from a disease, then students should point out that 242 and 63 (the "failures") are both greater than ten. Or if we expected a 3 percent response rate to 1,500 mailed requests

for donations, then np = 1,500(0.03) = 45 and nq = 1,500(0.97) = 1,455, both greater than ten.

Let's Take Stock...

What have we seen so far?

- The mathematics underlying statistical methods is based on important assumptions.
- We never know if those assumptions are true.
- Some assumptions are unverifiable; we have to decide whether we believe they are true.
- Other assumptions can be checked out; we can establish plausibility by checking a confirming condition.
- And some assumptions can be violated if a condition shows we are "close enough."

We've established all of this and have not done any inference yet! We can develop this understanding of sound statistical reasoning and practices long before we must confront the rest of the issues surrounding inference. By then, students will know that checking assumptions and conditions is a fundamental part of doing statistics, and they'll also already know many of the requirements they'll need to verify when doing statistical inference.

Inference for a Proportion

Inference for a proportion requires the use of a Normal model. Since proportions are essentially probabilities of success, we're trying to apply a Normal model to a binomial situation.

> **Independent Trials Assumption:** Sometimes we'll simply accept this. If we're flipping a coin or taking foul shots, we can assume the trials are independent. However, if we hope to make inferences about a population proportion based on a sample drawn without replacement, then this assumption is clearly false. We can proceed if the **Random Condition** and the **10 Percent Condition** are met. Close enough.
>
> The **Normal Distribution Assumption** is also false, but checking the **Success/Failure Condition** can confirm that the sample is large enough to make the sampling model close to Normal.

Inference for the Difference of Two Proportions

When we have proportions from two groups, the same assumptions and conditions apply to each. We need to have random samples of size less than 10 percent of their respective populations, or have randomly assigned subjects to treatment groups. In addition, we need to be able to find the standard error for the difference of two proportions. That's done by adding variances, and thus requires the...

> **Independent Groups Assumption:** The two groups (and hence the two sample proportions) are independent.

That's not verifiable; there's no condition to test. We just have to think about how the data were collected and decide whether it seems reasonable.

Inference for Means

Whenever samples are involved, we check the Random Sample Condition and the 10 Percent Condition. Beyond that, inference for means is based on *t*-models because we never can know the standard deviation of the population. The theorems proving that the sampling model for sample means follows a *t*-distribution are based on the...

> **Normal Population Assumption:** The data were drawn from a population that's Normal.

We can never know if this is true, but we can look for any warning signals. We've done that earlier in the course, so students should know how to check the...

> **Nearly Normal Condition:** A histogram of the data appears to be roughly unimodal, symmetric, and without outliers.

If so, it's okay to proceed with inference based on a t-model. But what does "nearly" Normal mean? If the sample is small, we must worry about outliers and skewness, but as the sample size increases, the t-procedures become more robust. By the time the sample gets to be 30-40 or more, we really need not be too concerned. Then our Nearly Normal Condition can be supplanted by the...

> **Large Sample Condition:** The sample size is at least 30 (or 40, depending on your text).

Note that understanding why we need these assumptions and how to check the corresponding conditions helps students know what to do. And it prevents the "memory dump" approach in which they list every condition they ever saw -- like np ≥ 10 for means, a clear indication that there's little if any comprehension there.

Inference for the Difference of Two Means

By now students know the basic issues. They check the **Random Condition** (a random sample or random allocation to treatment groups) and the **10 Percent Condition** (for samples) for both groups. They also must check the **Nearly Normal Condition** by showing two separate histograms or the Large Sample Condition for each group to be sure that it's okay to use *t*. And there's more.

As was the case for two proportions, determining the standard error for the difference between two group means requires adding variances, and that's legitimate only if we feel comfortable with the Independent Groups Assumption. Again there's no condition to check. We have to think about the way the data were collected.

Inference for Matched Pairs

Whenever the two sets of data are not independent, we cannot add variances, and hence the independent sample procedures won't work. Such situations appear often. We might collect data from husbands and their wives, or before and after someone has taken a training course, or from individuals performing tasks with both their left and right hands. Matching is a powerful design because it controls many sources of variability, but we cannot treat the data as though they came from two independent groups. Instead we have the...

> **Paired Data Assumption:** The data come from matched pairs.

There's no condition to be tested. Instead students must think carefully about the design. This helps them understand that there is no "choice" between two-sample procedures and matched pairs procedures. Either the data were from groups that were independent or they were paired. The design dictates the procedure we must use. Looking at the paired differences gives us just one set of data, so we apply our one-sample t-procedures. We already know the appropriate assumptions and conditions.

> **Independence Assumption:** The individuals are independent of each other. We base plausibility on the **Random Condition.**
> **Normal Distribution Assumption:** The population of all such differences can be described by a Normal model. We verify this assumption by checking the...
> **Nearly Normal Condition:** The histogram of the differences looks roughly unimodal and symmetric.

Note that there's just one histogram for students to show here. We don't care about the two groups separately as we did when they were independent. As before, the Large Sample Condition may apply instead.

Inference for Chi-Square

Although there are three different tests that use the chi-square statistic, the assumptions and conditions are always the same:

> **Counted Data Condition:** The data are counts for a categorical variable. This prevents students from trying to apply chi-square to percentages or, worse, quantitative data.
>
> **Large Sample Assumption:** The sample is large enough to use a chi-square model. But how large is that? We confirm that our group is large enough by checking the...
> **Expected Counts Condition:** In every cell the expected count is at least five.

Inference for Regression

We close our tour of inference by looking at regression models. The slope of the regression line that fits the data in our sample is an estimate of the slope of the line that models the relationship between the two variables across the entire population. Sample-to-sample variation in slopes can be described by a *t*-model, provided several assumptions are met. Each can be checked with a corresponding condition.

> **Linearity Assumption:** The underling association in the population is linear. By this we mean that the means of the *y*-values for each *x* lie along a straight line. Check the...
> **Straight Enough Condition:** The pattern in the scatterplot looks fairly straight.
>
> **Independence Assumption:** The errors are independent. By this we mean that

there's no connection between how far any two points lie from the population line. Check the...
Random Residuals Condition: The residuals plot seems randomly scattered.

Normality Assumption: Errors around the population line follow Normal models. By this we mean that at each value of *x* the various *y* values are normally distributed around the mean. Check the...
Nearly Normal Residuals Condition: A histogram of the residuals looks roughly unimodal and symmetric.

Equal Variance Assumption: The variability in *y* is the same everywhere. By this we mean that all the Normal models of errors (at the different values of *x*) have the same standard deviation. Check the...
Does the Plot Thicken? Condition: The residuals plot shows consistent spread everywhere. No fan shapes, in other words!

And That's That

Let's summarize the strategy that helps students understand, use, and recognize the importance of assumptions and conditions in doing statistics.

- Start early: ***Assumptions and Conditions aren't just for inference.***
- Distinguish assumptions (unknowable) from conditions (testable).
- Note that conditions may *verify* that an assumption is plausible, or *override* an assumption that is violated.
- Insist that students *always* check conditions before proceeding.

All the assumptions and their corresponding conditions can be nicely summarized on a single page, a good way to end this article. Your students may find this page helpful (it's available as a Web page or in PDF format below), but to answer the inevitable question: No, they cannot take it into the AP Exam with them! If only...

- Assumptions for Inference (Web page)
- Assumptions for Inference (.pdf/81KB)

Dave Bock has been a high school math teacher since 1969. He holds a BA in mathematics and an MS in statistics from the University at Albany. He has taught statistics at Ithaca High School, Cornell University, Ithaca College, and Tompkins-Cortland Community College, and has been teaching AP Statistics since its inception. He has served as an AP Statistics Reader and consultant since 1999, and leads the St. Johnsbury Summer Institute for teachers of AP Statistics. Dave is co-author of Barron's AP Calculus review book and of Stats:Modeling the World*, an AP Statistics text published by Addison-Wesley.*

See also...

- Notes on the 2002 AP Statistics Free-Response Questions
- Common Errors on the 2003 AP Statistics Exam
- AP Statistics Course Home Page
- Teachers' Resources

CONTACT | CAREERS | ABOUT US | PRIVACY | TERMS OF USE | PRESS

Copyright © 2005 collegeboard.com, Inc.