Is It Normal?

by Peter Flanagan-Hyde
Phoenix Country Day School
Paradise Valley, Arizona

Making the Case for Normality

In a variety of settings, students in AP Statistics are asked to make a judgment about whether or not the assumption that a given sample has been drawn from a normal distribution is a reasonable one. For example, on the 2001 AP Statistics Exam (problem 5), students were asked to provide statistical evidence about the difference in the active ingredients in two brands of pills. To use the appropriate *t*-statistic as part of their solution, they needed to assess whether the distribution of differences appeared to be a normal distribution, given that the sample size was only 10. The best solutions presented a graphical display of the sample distribution along with a statement that it was reasonable to assume that the underlying distribution of differences is normal. In this article, I'll look at what students in an introductory course can do to make the case for the normality of the population and note the limitations of these procedures.

Students often have trouble with the idea of the assumptions for inference. These assumptions are generally about specific properties of the population that must be satisfied for the probability calculations to be mathematically justified. The common dilemma: We don't know enough about the population and so cannot certify that these properties are present. The best we can do is to argue from the sample in hand that it is reasonable to *act as if* these population properties are indeed satisfied.

Many of the inference procedures, both hypothesis testing and construction of confidence intervals, are based on probability calculations that are applied to a normal distribution. If the actual population is not normally distributed, inferences from these calculations may be questionable. The conundrum that students face is that the situations in which it is most urgent that the underlying distribution be normal -- small samples -- are just those cases when it is most difficult to determine if the assumption of normality is plausible.

This is a difficult task for many students because we are asking them to make a very sophisticated argument without providing them much experience with distributions, which are normal or are not normal. Imbedded in the problem on which they are working is another inference problem: If the population is normal, how likely is it that I would get a pattern in a sample that departs from perfect normality as much as this one? Without much experience with this kind of question, students' intuitive judgment must substitute for formal assessment. And, as pointed out above, small samples are particularly difficult to assess.

In light of this difficulty, let's keep our hopes very modest for what students can accomplish in determining whether it is plausible to assume a normal distribution in the population based on a sample. On the 2000 AP Statistics Exam, problem 2, students were asked to assess the reasonableness of the assumptions for inference, including the normality assumption. The grading rubric indicates that students could get full credit for saying either 1) it *is not* reasonable to assume a normal distribution, or 2) it *is* reasonable to assume a normal distribution -- based on the same information! So it's knowing the question should be asked as well as providing a reasonable response that are important, rather than the specific conclusion reached.

As a guiding principle in what follows, I hope to discuss procedures that are easily implemented and about which students' intuition about normal distributions supports the conclusion reached. Here is a situation, too, in which the steady drumbeat of skepticism that I maintain throughout the course needs to be suspended somewhat. Too many of my students are ready to condemn a distribution as not normal based on a display from a small sample, so our goal is to find a procedure that is relatively robust in the face of small samples.

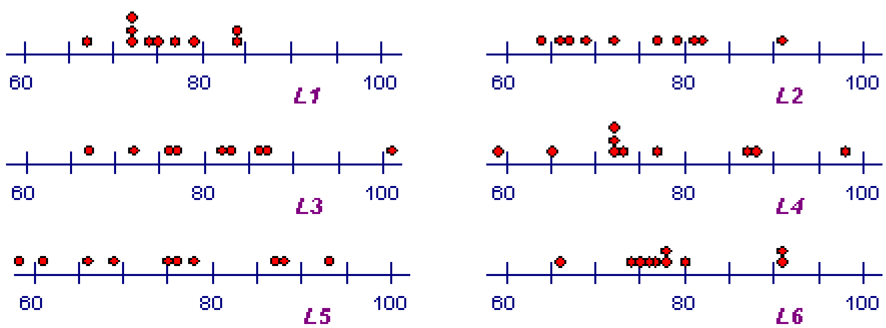
Most of us have watched enough television to be familiar with the police lineup. (Let's hope that's how you became familiar with this device!) One suspect is paraded in front of a witness along with five others; the question is whether the witness can spot the perpetrator among the innocents. I will take a similar approach: Six data sets are presented, five of which are samples from a normal distribution, and one of which is not. Can you spot the perpetrator?

Imagine that you give a test to 10 students who are randomly selected from a population that has a mean score of 75, a standard deviation of 10, and is normally distributed. In the table below are the results of five such samples, but not necessarily the first five, and a sixth sample drawn from a different population. This other population has approximately the same mean and standard deviation, but a different shape. (It's actually a uniform distribution.) You can generate lists like this on a TI-83 calculator with the command `round(randNorm(75,10,10),0)` for the normal distribution and `randInt(57,93,10)` for the uniform. Can you tell which one of the samples below is not from a normally distributed population?

L1	L2	L3	L4	L5	L6
74	72	82	77	93	77
84	81	77	98	87	75
75	79	86	88	78	74
77	91	83	72	76	76
79	69	101	59	61	66
67	67	76	73	66	78
72	82	67	65	75	78
72	64	73	72	69	91
72	77	87	87	88	80
84	66	72	72	58	91

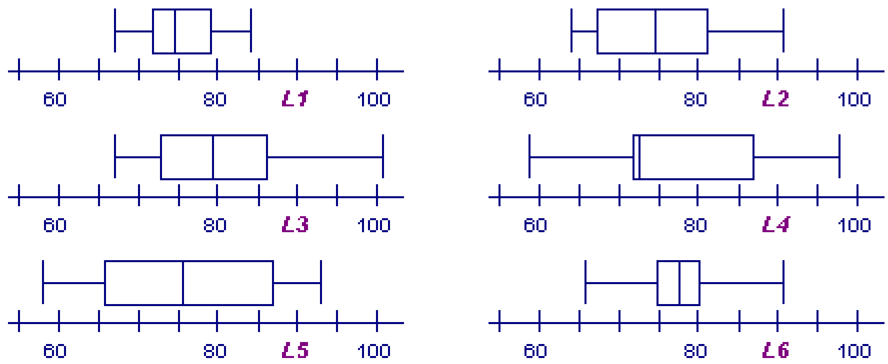
No? Don't feel bad -- it is virtually impossible to read the lists and make this assessment. In all cases, whether it is reasonable to assume normality will be based on a graphical display. Let us look at several options that are either easily made with a graphing calculator or available as hand-created displays and see which is most helpful.

First, how about a dotplot? This is extremely easy to construct for small samples and contains all the information in the sample. Below are dotplots for each of the lists. Which one is not a random sample from a normal distribution?



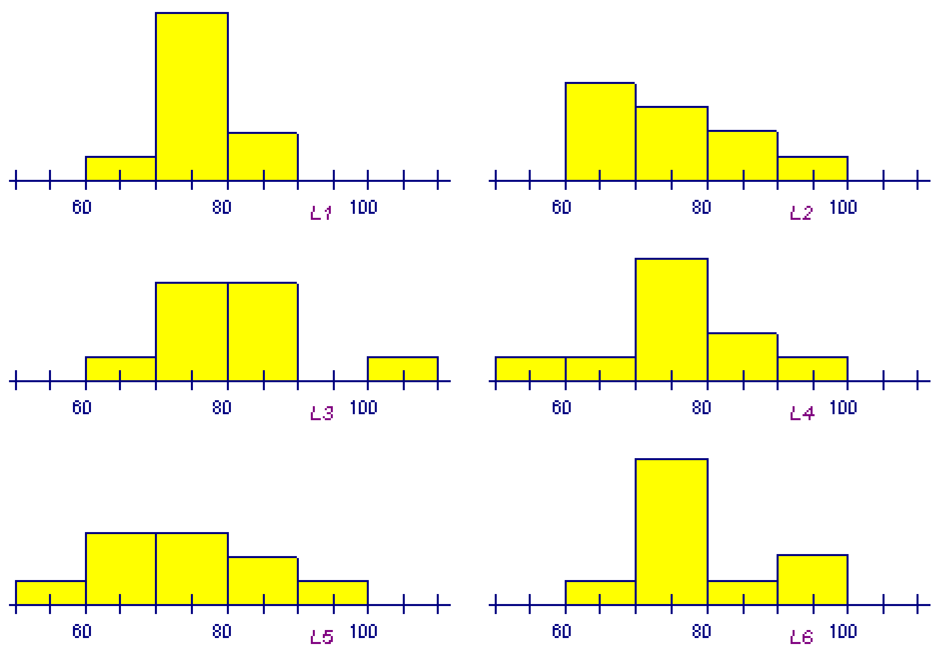
It's still not so easy to tell. For a mound-shaped normal distribution, you'd expect the dots to be symmetrically arranged. They should also be more bunched up in the middle of the distribution and more spread out on either end. Using this criterion, it appears that good candidates for the normal distribution are lists 3, 4, and 6. List 1 seems a little thin in the middle, with almost a "U" shape; list 2 looks like it's more heavily weighted to the left; and list 5 doesn't seem dense enough in the middle. I'm not sure we're at the climactic "That's the guy!" quite yet! But lists 1, 2, and 5 remain suspects.

Let's try another graphical display: boxplots. These help students see the overall pattern in the distribution without getting distracted by the location of each individual point. Below are boxplots for each of the lists.



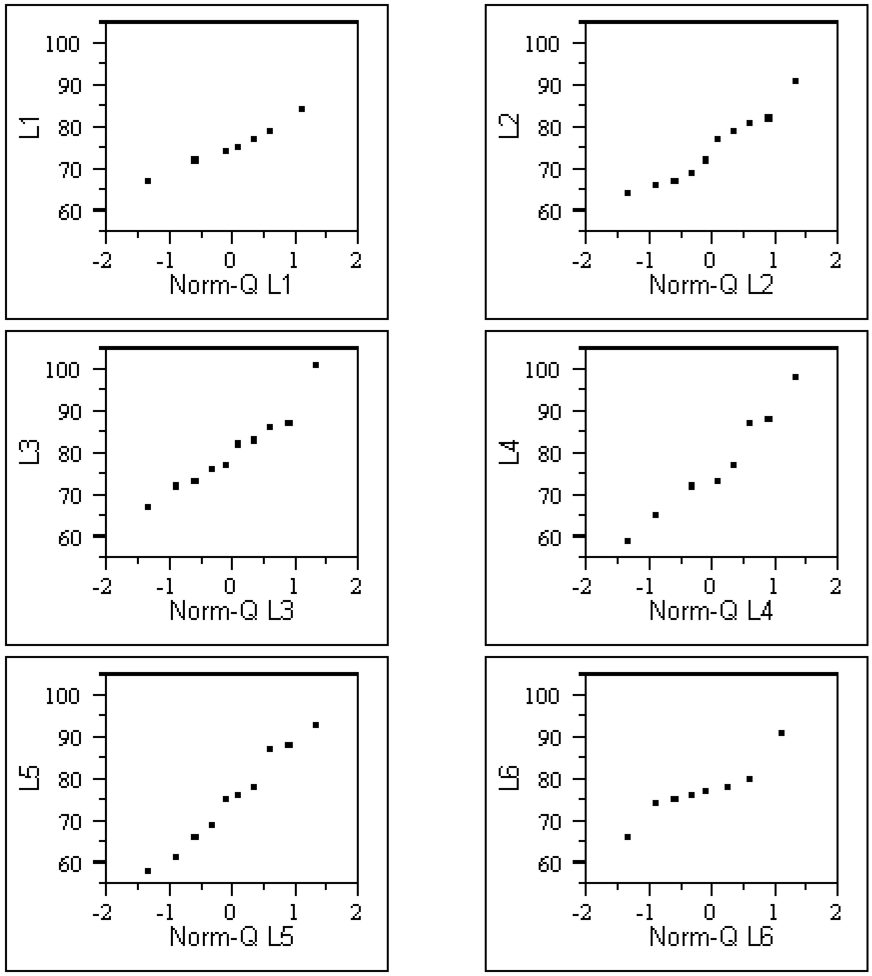
You would expect a normal distribution to show a symmetrical display, present most strongly in lists 1, 5, and 6. Even in the others, some aspects of the plot are symmetrical, either the box area or the "whiskers." Given the small samples, the location of any of the five vertical bars could move quite a bit with just one change in the data. The symmetries of these displays are not too bad, with perhaps list 4 being the most suspicious. Let's consider another aspect of the boxplot that should be present if it comes from a normal distribution -- the proportion of the plot that is in the "box."

In a normal distribution, the lower and upper quartiles are located about two-thirds of a standard deviation from the center. This provides an expectation for the width of the "box." Also, in a sample of size 10, it would be surprising if the values did not span at least 85 percent of the area under the curve. This corresponds to about 1.5 standard deviations on either side of the center and provides an expectation about the total length of the plot. Most typical, then, for samples of size 10 or so from a normal distribution, is that the range of values (total width of the plot) is more than twice the interquartile range (width of the box). In the graphs below, it appears that this is clearly present in lists 1, 3, 4, and 6; marginally present in list 2; and not present in list 5, where the box is too wide. How about a histogram of the data? Like the boxplot, this allows you to focus on the overall pattern, rather than the individual values. In a normal distribution, you expect to see the symmetry of the distribution and its mound shape to be higher in the middle and lower on either end. Here are histograms of our lists, with class width 10.



The expected properties are present in all lists except list 2, which appears to be skewed. Thinking more about how high you might expect the peak to be suggests that list 1 may have an unusually tall peak, and list 5 seems perhaps too broad-shouldered. Another issue with histograms for small data sets is that the appearance of the plot varies quite a lot with different class intervals. Ideally, we would like something less fickle.

Here's one last display: a normal probability plot (or normal quantile plot). This special scatterplot compares the actual data values with expected "normal quantile" values. For a large population that is normally distributed, the points in a normal probability plot will fall on a straight line. For samples from a normal population, points should fall roughly on a line. Like the dotplot, all the information in the sample is present. Below are normal probability plots for the six lists.



The linear pattern seems strongest in lists 1 and 5. List 2 shows a somewhat curved pattern, list 3 has an apparent outlier, list 4 is a little ragged, and the plot for list 6 is "S-shaped." The result is somewhat inconclusive, and it is a judgment call as to how straight the plot has to be. Certainly, though, the suspicions are focused on lists 2 and 6.

So, who is the culprit in this caper? List 5 is not a sample from a normal distribution, but is sampled from a uniform distribution instead. Let's examine how our displays did in picking out list 5 as a suspect.

The dotplots had list 5 as a mild suspect, but there were others in the running as well. The boxplots probably had 5 as the prime suspect, but list 4 was in the running, too. The histograms were pointing most strongly to list 2, with list 5 as another candidate. Finally, normal probability plots, despite their name, had pretty well exonerated list 5 and had us snooping about the private lives of lists 2 and 6, looking for more clues to their guilt.

My personal favorite display for assessing whether a distribution is plausibly normal is the boxplot, if the sample size is at least 10. I have seen students focus too much on individual dots in the dotplot, thinking, "It can't be normal; look at that dot there." Histograms can show very different results with small changes in class intervals. And finally, for small data sets, students have a very hard time separating the typical raggedness present in a normal probability plot from a true pattern of departure for these to be of much use. Boxplots are reasonably effective; easy to create, either by hand or with a calculator; and match students' intuition better than the other displays.

This is the procedure that I recommend to my students, when n is at least 10:

1. Make a boxplot, either by hand or with the calculator.

2. Look for symmetry. It is sufficient to have *either* the box or the whiskers symmetric, as long as the other isn't too far off. Insisting on both is too stringent a criterion for small samples. If neither aspect shows symmetry, the distribution is probably skewed.

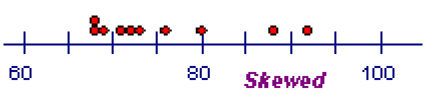
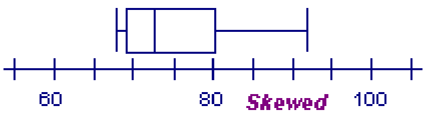
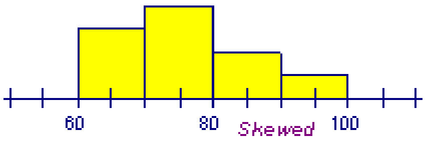
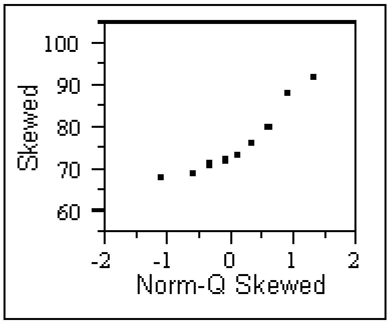
3. Compare the width of the box (IQR) to the width of the plot (range). The box should be less than or equal to half the width of the plot. If the box is too big, reject the distribution on grounds of shape.

4. Look for too-long whiskers. In a small data set, outliers (more than 1.5 IQR above or below the nearest quartile) should cause an expression of concern, and extreme outliers (more than 3 IQR above or below the nearest quartile) are cause for rejection.

For very small data sets *n* less than 10), a dotplot is probably the best bet, but for samples this small, it is hard to make any kind of a convincing argument.

Let's look at one more example and see how this procedure works. Since all the distributions that we've looked at thus far are symmetric, we'll try one that is clearly skewed. The list below was created with `round(8*randNorm(0,1,10)2+67,0)`.

73	68	92	72	71	76	68	80	69	88

<table>
<tr><td>Dotplot: Does not appear to be a normal distribution (symmetrical, points bunched in the middle). It shows the skewed distribution effectively.</td><td></td></tr>
<tr><td>Boxplot: Does not appear to be a normal distribution. Neither the "box" nor the "whiskers" show the expected symmetry, so it appears skewed.</td><td></td></tr>
<tr><td>Histogram: Plausibly a normal distribution, certainly more promising than the histogram for list 2 above.</td><td></td></tr>
<tr><td>Normal probability plot: Shows some curvature, but is this too curved or not? It's hard to tell, and it is not apparently much different from the plot for list 2 above.</td><td></td></tr>
</table>

A boxplot or a dotplot is most effective at reaching the conclusion that this is not a normal distribution. Of course, we present this example with the special knowledge that it wasn't drawn from a normal distribution, and there certainly *are* random samples taken from normal distributions that are as skewed as this one. Our goal can only be to give students some tools that are typically effective and encourage them to look at a variety of samples so that their understanding of the variability of different samples can be improved.

See also...

- AP Statistics Course Home Page
- AP Statistics EDG
- Teachers' Resources
- Institutes & Workshops
- CB Store: Products for AP Statistics