

The Trials of Scipio: An Investigative Task

Floyd Bullard

North Carolina School of Science and Mathematics

Durham, North Carolina

Introduction

I teach two sections of high school statistics at the North Carolina School of Science and Mathematics. Students here doing research projects in biology occasionally need help with the design of their experiments and the analysis of their data, and my statistics colleagues and I may be called upon to provide this help. Two years ago I worked with a student named Tommy Miller, who wanted to perform an experiment to show that mice can learn and transfer the concept of “middle.”

For a variety of reasons, Miller’s final study was quite different from the one he originally designed: mice required much care, mice didn’t behave as expected, mice were uncooperative, mice died, etc. In the end he was able to collect useful data on only *one* mouse he named Scipio. Obviously, extrapolation to any population was impossible. But he learned quite a bit about conducting research in biology, which was the main goal of his research class, and his mouse Scipio produced a set of data that provided me with a very good investigative task to share with my statistics students at the end of the school year, after they had learned all the hypothesis tests that are in the AP curriculum. (In fact, the data presented here are only from the second stage of Miller’s study. The complete study was more extensive.)

In this article, I will first share the investigative task, then provide commentary on some of the solutions my students came up with, and finally discuss a source of subjectivity inherent in hypothesis testing of which statistics students should be aware.

The purpose of this investigative task is threefold. First, it provides the students a chance to practice different types of hypothesis testing. Second, by allowing many possible approaches, it helps students to see that sometimes problems have no one right answer or one right approach. And third, because different valid approaches to the problem lead to different conclusions, it helps students see that the way one approaches a problem -- which is chosen subjectively by the researcher based on his expectations prior to doing the experiment -- has an effect on the conclusion.

The Problem As Presented to Students

A student named Tommy Miller conducted an experiment to see whether a mouse named Scipio was capable of learning to find the middle of seven markers. Miller filled a plastic wading pool with water and placed in it seven empty tin cans without lids. Attached to them were drinking straws that extended above the surface of the water so that the swimming mouse could see the markers. On top of the middle tin can was a clean glass dish, resting just underneath the surface of the water. It was invisible from the surface, so Scipio could not tell just by looking which, if any, of the drinking straw markers was attached to a platform on which he could gain solid footing.

Miller scheduled nine *sessions* with Scipio, separated by several hours or days. Each session consisted of five *trials*, separated by a few minutes. A trial consisted of Miller putting Scipio in the water and counting how many times the mouse tried to climb up one of the “wrong” drinking straws (and found no platform there) before finally climbing up the one in the middle. Once

Scipio found the middle one, Miller rewarded him by taking him out of the water and warming him in his hands for a few minutes. The table below shows the number of “errors” Scipio made (i.e., attempts to climb on a platform at a marker other than the middle one) before climbing on the platform at the middle marker.

		Trial number					
		1	2	3	4	5	Totals
Session number	1	1	6	6	6	3	22
	2	11	0	6	3	3	23
	3	3	2	0	0	1	6
	4	3	8	8	4	1	24
	5	10	3	6	6	8	33
	6	6	0	0	0	2	8
	7	1	1	2	5	2	11
	8	5	1	1	2	2	11
	9	0	0	4	2	0	6
						144	

Miller’s hypotheses were:

H_0 : Scipio is *not* learning where the platform is.

H_A : Scipio *is* learning where the platform is.

Consider Miller’s research question: *Is Scipio learning?* How strong is the evidence provided by his data that Scipio is, in fact, learning to find the middle platform? Choose an inference procedure that you think is appropriate and answer Miller’s question. Be prepared to share with the rest of the class what inference procedure you chose, why you chose it, and what conclusion it led you to.

Some Student Approaches

My students came up with many possible solutions to the problem. They are listed below, with a commentary after each one. Note that some approaches are completely inappropriate, some are appropriate but do not take full advantage of all the data, and others are appropriate and take greater advantage of all the data. *However, there is no single right way to approach the problem.* Each appropriate approach reflects a different model of the problem, and each has its own strengths and weaknesses.

1. Chi-square test of independence.

Comment: This was by far the most common *inappropriate* approach to test the null hypothesis. The students who chose this method were reacting to the way the data were presented: a grid of whole numbers. If it were possible to demonstrate the dependence of row- and column-factors for these data, that would not indicate that the mouse was learning anything. It would only demonstrate that the mouse’s error count sequences were not proportional in all sessions, that its behavior in some of the sessions was qualitatively -- not just quantitatively -- different from some of the other sessions. This does not match the stated hypotheses of learning

to find the middle or not. Students should learn *not to decide upon a statistical approach based upon the format in which the data are presented*. They must *think* about the data and what the data mean!

2. Chi-square test of goodness-of-fit.

$$H_0 : p_1 = p_2 = \dots = p_9 = \frac{1}{9}$$

$$H_A : p_n \neq \frac{1}{9} \text{ for some } n = 1, 2, \dots, 9,$$

where p_n is the proportion of errors Scipio would make during session n .

Comment: A few students wanted to perform a chi-square test of goodness-of-fit, with the null hypothesis being that the errors would be uniformly distributed over each of the nine sessions. This makes more sense than the chi-square test of independence, but it has a problem: a deviation from the null hypothesis would not necessarily indicate a general reduction in error counts over time. Indeed, a dramatic increase in errors or a pronounced fluctuation in errors over the sessions would also cause the null to be rejected but would not indicate learning. So this approach is also inappropriate for answering the research question.

3. $H_0 : p = \frac{1}{9}$

$$H_A : p > \frac{1}{9},$$

where p represents the proportion of errors Scipio makes in the first session. *A priori*, you would expect him to make $\frac{1}{9}$ of his errors in the first session of nine if he were not learning and a greater proportion than that if he were learning. Is the observed proportion of errors that occurred during the first session sufficiently greater than $\frac{1}{9}$ to provide statistically significant evidence that Scipio was learning? The observation is $\hat{p} = \frac{22}{144}$, which provides a P-value of about 0.056: reasonably strong evidence that Scipio is learning. (Though not significant at the $\alpha=0.05$ level.)

Comment: This approach is reasonable, though probably not ideal. Students should know that a test of proportions has as a requirement that the number of observations be *fixed in advance*. That was not the case here. Many scientific studies in fact ignore this altogether. For example, in ecological research, a herpetologist may choose a sample size equal to however many snakes he happens to capture. This violation of the requirement is not a serious problem so long as the stopping rule is independent of the response variable being studied. In this case, we are not told why Miller chose to conclude his data collection after nine sessions. It may have been because he saw that the data were showing him what he wanted to see. In that case, this approach would *not* be reasonable (nor, indeed, would any approach students learn in the AP curriculum). But if we make the assumption that the decision to stop collecting after nine sessions was decided in advance or was arbitrary, then this approach is reasonable.

However, there is another weakness to this approach. It takes a table rich with numerical data and lumps all of it together into only one data value, erasing most of the information the table contained. This very likely has the effect of making this particular approach not very *powerful*. (Recall that *power* is the probability that a test of significance will reject the null hypothesis given that it is indeed false.) If the mouse is in fact learning to find the middle marker, then we want a test that will have a high probability of detecting that learning -- but by erasing information, we reduce that probability. We cannot really speak of the power of a test of significance without speaking of its power against a particular alternative hypothesis. This student approach would

only be optimally powerful if the erased information was in fact useless information; in other words, it would only be optimally powerful if the mouse did all of its learning during the first session. That would seem to be an unreasonable assumption to make *a priori*. Thus, while this approach is valid for answering the question of interest, there are better approaches.

$$4. \quad H_0 : p = \frac{2}{9}$$

$$H_A : p > \frac{2}{9},$$

where p represents the proportion of errors Scipio makes in the first *two* sessions. This is similar to approach 3, and the reasoning behind it is the same. The observation is $\hat{p} = \frac{22+23}{144}$, whose P-value is about 0.0046. This provides very good evidence that Scipio is learning.

Comment: The same comments made about approach 3 apply here. This method was one a student came up with after discussing approach 3 with his group. He found that it would produce a smaller P-value and thought it was therefore a better approach. At the end of this article, I will discuss one of the lessons students should take away from this activity, which is that you must choose a test statistic before, not after, data collection.

While approach 3 would only be optimally powerful if all the mouse's learning occurred during the first session, this one would only be optimally powerful if all the mouse's learning occurred during the first *two* sessions. It would be very difficult to argue in a research paper that the researcher expected *a priori* this type of learning to occur. Thus, this approach is not technically incorrect, but it would be difficult to justify using. Better approaches exist.

$$5. \quad H_0 : p = \frac{1}{9}$$

$$H_A : p < \frac{1}{9},$$

where p represents the proportion of errors Scipio makes in the *last* session. This is also similar to the previous two approaches and uses the same reasoning. The observation is $\hat{p} = \frac{6}{144}$, whose P-value is about 0.004, providing strong evidence of learning.

Comment: Again, the comments on approaches 3 and 4 apply here as well. This approach is about as appropriate as approach 3 but is somewhat less justifiable. One could perhaps make a case *a priori* that the mouse was expected to learn quickly and that all learning would occur during the first session. But how could one argue that all the learning was expected to suddenly occur after exactly *eight* sessions? Again, better approaches exist.

(Note that the requirement of a large sample size for the normal approximation to be used is not well met. Some texts teach that you should have at least 10 successes and at least 10 failures, others teach five or some other rule. The six successes here do make the normal approximation to the distribution of \hat{p} a bit weak, but the P-value of 0.004 is sufficiently small that we can be confident that even a more accurate calculation of the P-value would lead to rejecting the null hypothesis.)

$$6. \quad H_0 : p_E = p_L$$

$$H_A : p_E > p_L,$$

where p_E and p_L represent, respectively, the proportion of all the errors made in the “early” (first four) sessions and the proportion of all the errors made in the “late” (last four) sessions.

Comment: This approach is *incorrect*. When comparing population proportions using independent sample proportions, the latter need to be of samples taken from two distinct populations and must measure the proportion having a single common characteristic. But in this student approach, there is really only a single population (all the errors), and the two proportions represent two distinct characteristics (early or late occurrence).

Students occasionally make a similar error when they should be testing $p = 0.5$, and they instead attempt to test $p_1 = p_2$, where p_1 and p_2 are in fact complementary proportions.

$$7. \quad H_0 : \mu_E - \mu_L = 0 \\ H_A : \mu_E - \mu_L > 0,$$

where μ_E represents the mean number of errors per session that Scipio would make in his “early” sessions (before the fifth session), and μ_L represents the mean number of errors per session that Scipio would make in his “late” sessions (after the fifth session). With only four observations each of “early” and “late,” we must *assume* here that the distributions of errors in each period are approximately normal, despite some evidence to the contrary in the early sessions. (The 6 appears to be an outlier.) But if we perform inference anyway using two independent sample t-procedures, we observe a P-value of about 0.052: reasonably good evidence that Scipio is learning (though not significant at the $\alpha=0.05$ level).

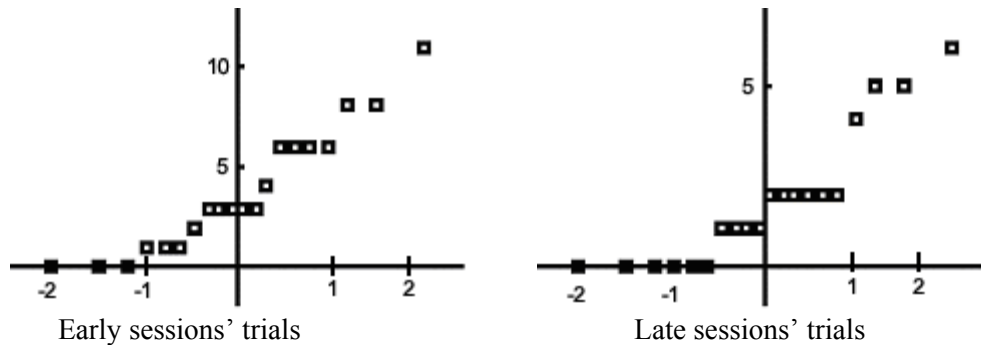
Comment: The student group who proposed this approach was aware of the problem of assuming normality. I was impressed by their grasp of the requirements for using this test comparing population means.

This approach sat poorly with many students because we were ignoring session 5. Why would we ignore data? It’s an excellent point, but it doesn’t invalidate the test. If this type of statistical analysis had been decided in advance, it might have made more sense for Miller to collect data from an even number of sessions so that half the sessions could be “early” and half “late” without ignoring any. It has been pointed out in the comments that approaches 3-5 above ignore information as well; they lump together all the errors made in many sessions without distinction of which session they belonged to. And so far, no proposal has paid attention to the separate trials. Yet ignoring information does not invalidate a conclusion. At worst, it will make a test of significance less powerful; i.e., it will make it less likely that you will pick up on learning that may be occurring. This approach is valid and somewhat justifiable, though the assumption of normality is probably unreasonable (as pointed out by the students). Better approaches exist still.

$$8. \quad H_0 : \mu_E - \mu_L = 0 \\ H_A : \mu_E - \mu_L > 0,$$

where μ_E represents the mean number of errors *per trial* that Scipio would make in his “early” sessions (before the fifth session) and μ_L represents the mean number of errors *per trial* that Scipio would make in his “late” sessions (after the fifth session).

This approach is similar to approach 7 above, only now we have samples of size $n = 20$ instead of $n = 4$, so instead of *assuming* normality of the populations, we can actually *check* the normality. Normal probability plots of the two samples are shown below. In both graphs, the data are on the vertical axis.



Normality is not grossly violated, although the skew in the second sample is sufficient to cause concern. (Whenever dealing with data containing a lot of zeros, one should be cautious.) If one chooses to perform a two-independent sample t-test using these data, the P-value is 0.01, which is pretty strong evidence that Scipio is learning.

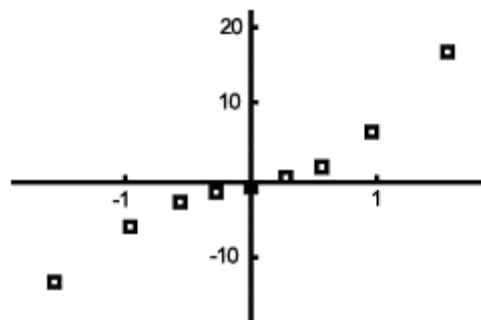
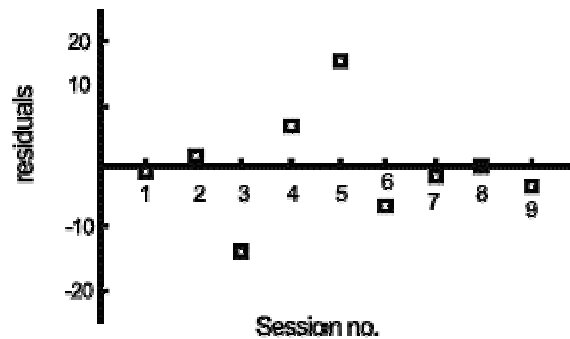
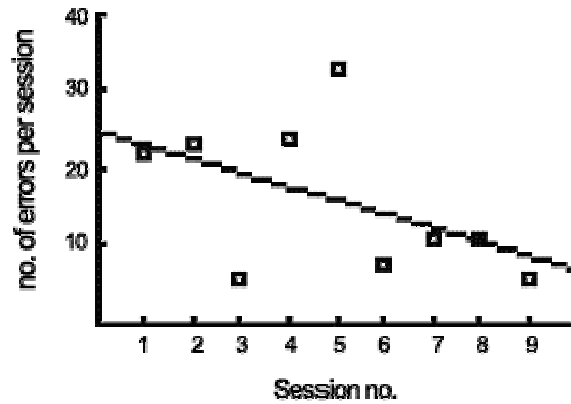
Comment: This approach is the first one so far to pay attention to the information in the individual trials. It is commendable that the students using this approach actually checked requirements instead of simply assuming the requirements were met. The approach is completely valid, and its conclusion quite justifiable.

There remains still some lumping together of information, however, that potentially lowers the power of the test against the real phenomenon. This test is only optimally powerful if all the mouse's learning occurs after session 4 or 5. It makes no distinction between sessions 1-4 or sessions 6-9 but treats all 20 trials in each group as the same. A method yet remains that would take into account the order of the sessions: linear regression. Approach 8 is definitely commendable -- but still probably not the best one available to students.

$$9. \quad H_0 : \beta = 0$$

$$H_A : \beta < 0,$$

where β is the slope of the hypothetical line modeling Scipio's error counts per session as a function of session number. A t-test on the slope requires that the relationship be linear and that the residuals be normally distributed with the same standard deviation. The graphs below show respectively the linear regression line, the residuals plotted against the session number, and a normal probability plot of the residuals with the data on the vertical axis.



The t-test on the slope gives a P-value of 0.086, very weak evidence that Scipio is learning.

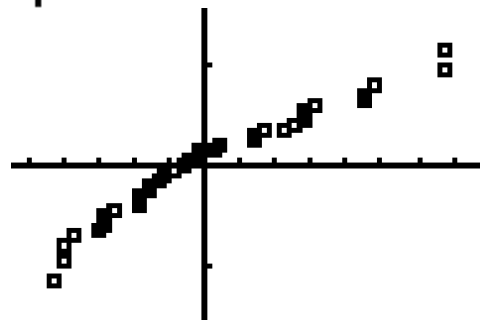
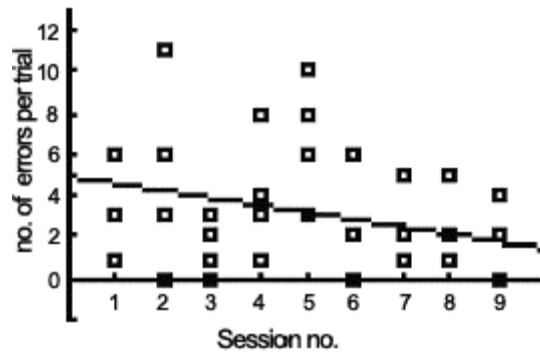
Comment: This student solution shows some good thinking in the attempt to use a regression line and incorporate the time-ordered nature of the data but an unfortunate lack of understanding of the underlying requirements for the t-test on the slope of a regression line. First of all, the first graph does not really suggest a linear model for the data. Second, the normal probability plot of the residuals has a curvature that suggests “heavy-tails” -- i.e., more weight in the low and high tails of the residual distribution than a normal distribution would have. (With the data on the vertical axis, this plot shows high data values that are higher than z-scores would predict and low data values that are lower than z-scores would predict.) This is particularly troublesome when performing inference with only nine data values: the Central Limit Theorem requires a large number of data values to overcome skewed or heavy-tailed distributions. This approach might have been reasonable had the data been more cooperative. But with the given data, this approach is questionable.

It should also be pointed out here that performing linear regression on the total number of errors per *session* versus session number is different only by a multiplicative factor of 5 from performing regression on the *mean* number of errors per *trial* versus session number. In general, regression on means can give drastically different results from regression on the raw data and should be avoided whenever possible. In this case, it is of course possible to avoid regression on the means: the raw data are the numbers of errors per trial, which are given.

$$10. H_0 : \beta = 0$$

$$H_A : \beta < 0,$$

where β is the slope of the hypothetical line modeling Scipio's error counts *per trial* as a function of session number. A t-test on the slope requires that the relationship be linear and that the residuals be normally distributed with the same standard deviation. The graphs below show respectively the linear regression line, the residuals plotted against the session number, and a normal probability plot of the residuals with the data on the vertical axis.



It appears from the first graph that a reasonable model for these data may be a linear one. The second graph shows a troubling decrease in variability as the session numbers increase, which violates one of the requirements of inference, and the third graph shows a slight amount of right-skew in the distribution of residuals. Fortunately, with 45 data values, this small amount of right-skew will be overcome by the Central Limit Theorem, so inference may be performed. The P-value of the test against zero slope is 0.017, providing good evidence that Scipio is learning.

Comment: This is a very thorough and well-thought-out approach. The data are used in all their richness, and appropriate checks are made on the requirements of inference. The skew in the residuals may be greater than the student thinks, but her caution is noted along with her appropriate rationale for continuing with the inference, viz., the large sample size. The student notes the decrease in the spread of the residuals over the sessions (this violates the *homoscedasticity* requirement) but doesn't seem to know what to do about it. Fortunately, there is a way around this problem that is a part of the AP Statistics curriculum: a transformation of the data.

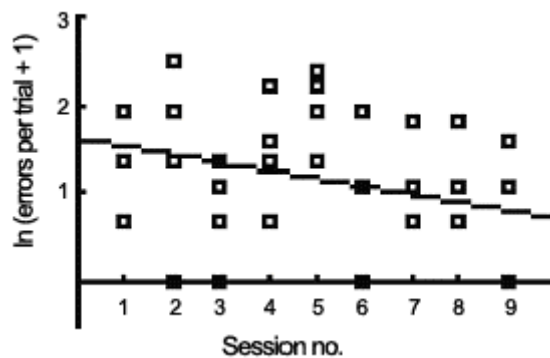
It should be mentioned before continuing that inference on the slope of a regression line has another requirement that is violated here: the independence of the observations. These data are actually time-series data rather than independently collected observations, since the x -variable is session number, a measure of time in this case. Like the problem with a fixed number of trials mentioned in approach 3, this problem is often ignored if the residuals do not show any obvious trend over time, which they do not appear to do here; the inference is roughly correct.

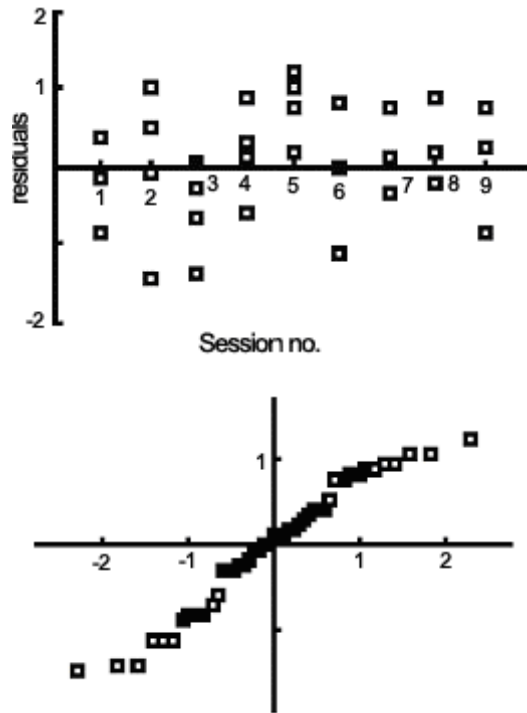
$$11. H_0 : \beta = 0$$

$$H_A : \beta < 0,$$

where β is the slope of the hypothetical line modeling $\ln(\text{error counts per trial} + 1)$ as a function of session number. The log transformation was chosen because the log function tends to lessen differences in variability, and the "+ 1" was added because some of the error counts were zero.

A t-test on the slope requires that the relationship be linear and that the residuals be normally distributed with the same standard deviation. The graphs below show respectively the linear regression line, the residuals plotted against the session number, and a normal probability plot of the residuals with the data on the vertical axis.





This model does a reasonably good job of linearizing the data (the residual plot shows no trend) and of creating equal variability at different treatment levels (the residual plot also shows no obvious change in magnitude of residuals). The normal probability plot of the residuals shows a small amount of deviation from normality, but happily, it is in the direction of *light-tailedness*; i.e., the distribution of residuals appears to be slightly less spread out than a normal distribution. (This can be seen by imagining a line through the normal probability plot. The upper data values are a little lower than the z-scores would predict, and the lower data values are a little higher than the z-scores would predict.) The Central Limit Theorem will eliminate most of this, but still, the light-tailedness will make our P-value more conservative, not less, than it would if the residuals were normally distributed.

Performing a t-test on the slope of the regression line gives a P-value of 0.02, which provides strong evidence that Scipio is learning.

Comment: Although all the 10 preceding approaches were proposed by my students, this one was not. It is included here to show an approach accessible to students who have completed the AP curriculum that puts together several different tools they've learned so as to take full advantage of the richness of the data and to overcome the problems with a linear model on the untransformed data. It is the best of the approaches proposed in this article because it is the most powerful against many alternative hypotheses (by not grouping the data excessively), and it does well at meeting its inference requirements.

It would be unusual for all but the most competent and talented AP Statistics students to come up with this or a similar approach entirely on their own, yet teachers should not consider it too lofty a goal to strive for. This is the kind of solution we would like from our students after they complete the AP Statistics course. We would like our students not only to have tools (such as normal probability plots and function transformations) "under their belts" but also to be able to discern when they are needed. We would like our students not only to recognize when inference

requirements are not met (e.g., “these residuals do not have equal variance”) but also to have some idea of what to do about it. We would like our students to be comfortable creating appropriate mathematical models.

Students may be very unhappy with this model because it seems rather contrived, what with the “+1” inserted in the transformation just to deal with the problem of zero-valued data. They should learn that statisticians work to create mathematical models with the hope that they are useful (by linearizing the data, reducing differences in variability, etc.) but with the understanding that they are imperfect.

Two Further Observations

Students may well come up with approaches other than the ones I described above. Indeed, once we had discussed in my class some of the above approaches, the students were inspired to think of similar ones (such as approach 4 inspired by approach 3). All of the approaches that students of AP Statistics are likely to propose require a certain assumption that not every statistician would readily accept. We are assuming that the error counts observed per trial behave like a random sample of error counts per trial drawn from the phantom population of error counts per trial that Scipio would exhibit were it possible to go back in time and repeat the experiment with Scipio again and again under “similar” conditions. Having data from more than one mouse would eliminate this requirement, for then our population would be all the lab mice that we drew from. As it is, with one mouse only, not only is extrapolation to more mice inherently impossible, but even inference about Scipio himself (“is Scipio learning”) requires us to accept that under slightly different conditions, we might have collected slightly different data, so our observations are, in a sense, a random sample from a population of hypothetical observations. Some statisticians are fine with this, while others are not.

It should be mentioned that the data here are actually waiting times until the first success, and should have a geometric distribution if the mouse’s visits to markers are assumed random and independent. If 1 is added to all values, the variable is the number of the trial on which the first success occurs and has mean equal to the reciprocal of the probability of success. As there are seven straws, this probability would be $1/7$ under pure guessing. Note that the adjusted session means are 5.4, 5.6, 2.2, 5.8, 7.6, 2.6, 3.2, 3.2, 2.2; the first few (except for the third) are not far from the expected 7. (That they are under 7 is unsurprising in view of the fact that the correct choice is always in the middle.)

Which Approach Is Correct?

It should be seen by now that there is no single, correct approach to studying these data. It is not uncommon in scientific studies for multiple inferential approaches or models to be possible. The approaches presented in this article differ largely in how they define “learning,” which is vague in the worded statements of the hypotheses.

If multiple approaches are possible, shouldn’t they still all give the same P-value? From the example of Scipio, clearly not. Although many of these approaches provide some evidence that Scipio is learning, the P-values vary quite a bit. One can think of reasonable approaches that do not produce significance evidence of Scipio learning at all. (For example, try testing the null hypothesis that one-third of Scipio’s errors would occur in one of the first three sessions against the alternate hypothesis that more than one-third would occur there.) In hypothesis testing, different methods of inference, even applied to the same data, do not necessarily provide the same strength of evidence against the null.

Given that the choice of approach affects the conclusion, how do you choose an approach? There are several answers to this question. The first is, you do NOT choose an approach after seeing the data. In that respect, this investigative task is a good example of what NOT to do in practice! If you do this activity with students, it is *crucial* that you do not stop before discussing this final aspect of the investigative task with them! Part of the point of the activity is for them to practice different hypothesis tests, yes, but a more important part is for them to think about *how a method of inference should be chosen in the first place*, and one thing that they absolutely must understand is that you cannot choose a method *after* you see the data. Then you could just pick one that produces a P-value you like!

(Indeed, if you wish to share this activity with your students in a way more similar to what should be done in practice, you should give your students the data *grid*, without any actual data in the cells, and have them determine how they will analyze the data first. Then give them the data and have them apply their methods. The main reason the activity is not presented that way here is that the data themselves often spark students' creativity more than an empty data table.)

So you have to decide before collecting the data how you will analyze them. The question remains: how do you choose one inference procedure over another? You choose *subjectively*, based on what you expect to happen. Consider the case of Scipio. Do you expect him to do most of his learning in the first session? Then approach 3 would be a reasonable method. Do you expect him to do most of his learning near the end, to "finally get it"? Then approach 4 above would be reasonable. Do you expect a marked decrease in errors around the fifth session? Then approach 8 might be a good choice. Do you expect a gradual decrease in errors to occur continually over all the sessions? Then inference based on the slope of some regression line might be appropriate. There is no single right answer. The "best" method is the most powerful one against the real phenomenon -- i.e., the one that is most likely to detect an effect that is present -- and the probability of detecting an effect depends upon how the effect manifests itself, which can only be guessed at in advance. (Preliminary studies can guide researchers.)

This is one way that subjectivity plays a role in statistics. We make a choice and hope it is a good one, but we can't tell for sure in advance. If we choose a method A and then find that method A does not reject the null while method B would do so, it is unethical to switch to method B after seeing the data.

Conclusion

The data presented in this article are actual data collected by a student performing a study on mouse behavior. They provide a good context for discussing many different hypothesis tests that students of AP Statistics study during the year. Arguably, the best model among those presented here is a rich one incorporating inference on a regression line as well as data transformation. When presented as a class activity in which students come up with their own ways of analyzing the data, these data also help students understand the subjectivity inherent in selecting a test statistic from among many choices. This should help them see that there is a certain art to scientific inquiry, and that experience can guide researchers to making their decisions wisely.

Floyd Bullard received his bachelor's degree in mathematical sciences from the Johns Hopkins University in 1991 and his master's degree in statistics from the University of North Carolina at Chapel Hill in 1999. He has taught high school math as a Peace Corps volunteer in Bénin, West Africa, at the Horace Mann School in New York, and most recently at the North Carolina School of Science and Mathematics (NCSSM) in

Durham, North Carolina. He is now on a leave of absence from NCSSM to study in a doctoral program in statistics at Duke University. Floyd is a Reader for the AP Statistics Exam.