VOL. 81 | NO. 8
OCTOBER 2019

THE AMERICAN BIOLOGY TEACHER

NABT
National Association of Biology Teachers

VOL. 81 | NO. 8
OCTOBER 2019

THE AMERICAN BIOLOGY TEACHER

About Our Cover

This Blue Ridge red salamander (*Pseudotriton ruber nitidus*) was found under a rotting log in a secondary-growth pine forest along the Appalachian Trail in southwestern Virginia. These large, lungless salamanders are widely distributed throughout the eastern and southern United States. Their skin secretions contain pseudotritontoxin and are poisonous if ingested, making this an excellent example of aposematic coloration. In fact, skin secretions of the mud salamander (*P. montanus*), spring salamander (*Gyrinophilus porphyriticus*), and eastern newt (eft stage; *Notophthalmus viridescens*) are also toxic, forming a Müllerian mimicry complex in which all members use red coloration to warn predators of their unpalatability.

The photograph was taken by Bob Remedi with a Canon 7D Mark II and Canon EF 100mm f/2.8L macro lens set at F14, 1/100 second and ISO 200. Bob Remedi is a full-time faculty member at College of Lake County, Grayslake, IL 60030. Bob would like to thank Dr. James Organ, Dr. Kevin Hamid, and Jerry Hinkley for helping to nurture his fascination with salamanders and love for taking students on the Appalachian Trail.

Contents

Feature Article

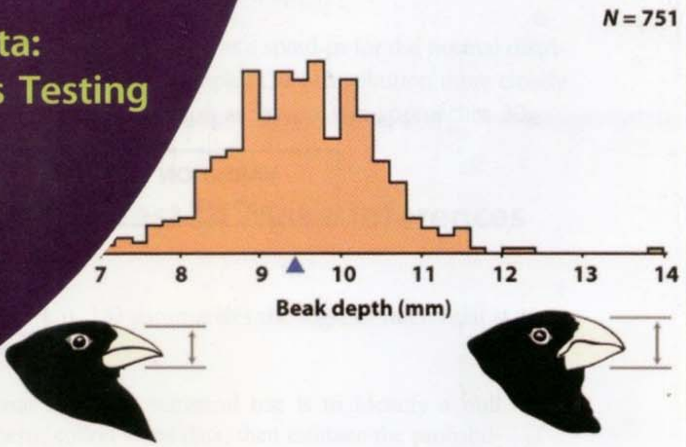
Research on Learning

Inquiry & Investigation

Tips, Tricks & Techniques

Departments

Making Decisions with Data: Understanding Hypothesis Testing & Statistical Significance



ROBERT A. COOPER

FREEMAN, SCOTT; HERRON, JON C., EVOLUTIONARY ANALYSIS, 4th, ©2007. Reprinted by permission of Pearson Education, Inc., New York, New York.

Abstract

Statistical methods are indispensable to the practice of science. But statistical hypothesis testing can seem daunting, with P-values, null hypotheses, and the concept of statistical significance. This article explains the concepts associated with statistical hypothesis testing using the story of "the lady tasting tea," then walks the reader through an application of the independent-samples t-test using data from Peter and Rosemary Grant's investigations of Darwin's finches. Understanding how scientists use statistics is an important component of scientific literacy, and students should have opportunities to use statistical methods like this in their science classes.

Key Words: *AP Biology; college science; data analysis; data interpretation; IB Biology; inquiry instruction; secondary school science; statistical analysis; statistics.*

○ Introduction

Statistical methods are indispensable to the practice of science, and understanding science includes understanding the role statistics play in its practice. Students must be given opportunities to analyze data in their science classes, using statistical methods that are suited to the data and age-appropriate. Middle school science students should be able to construct and interpret graphs, understand variation, calculate a mean, and understand what standard deviation tells us about a distribution. High school and college biology students should be able to construct and interpret error bars, and perform and interpret statistical hypothesis tests like chi-square and the independent-samples *t*-test. Here, I explain the meaning of *statistical significance* and related terms associated with hypothesis testing, using an application of the independent-samples *t*-test as an example.

Students must be given opportunities to analyze data in their science classes, using statistical methods that are suited to the data and age-appropriate.

○ Variation & Sampling

As we engage students with inquiry labs, situations arise where students must make decisions based on data. Statistics allow us to organize data for interpretation and deal with variation in the data. There are many sources of variation. Some variation, like the genotypic and phenotypic differences between organisms, is characteristic of the systems we study. But some of the variation we see is induced by data collection (Wild, 2006). Figure 1 distinguishes the sources of induced variation from the real variation in which we are interested. Measurement error can arise from mistakes made by the person making the measurements or from limitations or flaws in the measuring devices. Other errors can occur during the collection and processing of data. For example, a number could be entered in the wrong column on a data table or spreadsheet. Finally, there is always sampling error. Sampling error results when a sample that is intended to represent the entire population does not adequately do so.

Being meticulous in your data collection and sampling methods may reduce or eliminate many of these sources of induced variation. For example, careful attention to detail can reduce or eliminate the chance of measurement errors or accidents occurring during data collection and processing. But measuring devices will always have limitations resulting in some degree of variation and uncertainty, however small. And unless we only deal with cases where populations are very small and we can measure every individual, there will always be some sampling error. Statistical methods help us filter out any real variation in sample data from the surrounding noise caused by induced variation so that we can learn something about our population of interest.

The American Biology Teacher, Vol. 81, No. 8, pp. 535–542, ISSN 0002-7685, electronic ISSN 1938-4211. © 2019 National Association of Biology Teachers. All rights reserved. Please direct all requests for permission to photocopy or reproduce article content through the University of California Press's Reprints and Permissions web page, www.ucpress.edu/journals.php?p=reprints. DOI: https://doi.org/10.1525/abt.2019.81.8.535.

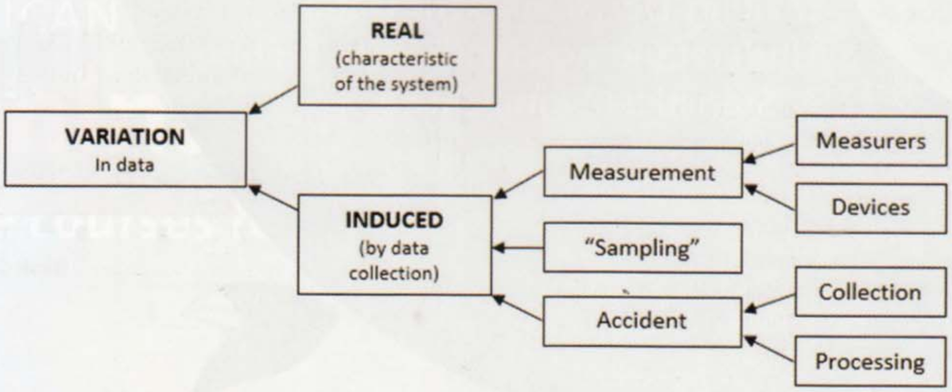


Figure 1. Sources of variation in data (Wild, 2006).

○ The Lady Tasting Tea & Statistical Significance

A common situation that arises in inquiry labs involves determining whether the difference between two groups, for example a treatment and a control group, is statistically significant. But statistical methods can seem daunting, with their *P*-values and null hypotheses. And for that matter, what does "statistical significance" actually mean? Many instructors and students struggle to understand these concepts. A story from the history of statistics about a lady tasting tea should make significance testing and related concepts more accessible.

The idea of a test of significance was conceived by Ronald Fisher (1890–1962). He played a major role in developing experimental designs and statistical methods that helped to revolutionize the practice of science in the twentieth century (Salsburg, 2001). In his book *The Design of Experiments* (1971; first published in 1935), Fisher introduced the concept of a test of significance by recounting the following story. One summer afternoon in the late 1920s, Fisher and several colleagues were having tea. When Fisher handed Lady Muriel Bristol a cup, she declined because Fisher had poured the tea into the cup first. Lady Bristol declared that tea tasted different depending on whether the milk was poured into the tea or the tea poured into the milk. Fisher and the other scientists were skeptical and began to discuss how they could test Lady Bristol's claim.

The scientists decided to arrange eight cups, four with the milk poured into the tea and four with the tea poured into the milk. The cups were presented to Lady Bristol for tasting one at a time in random order, and she was told that she had to identify the four milk-first cups. In his book, Fisher explains how to determine the probabilities associated with having the lady evaluate eight cups of tea. Figure 2 shows the probabilities he calculated for each number of milk-first cups the lady could potentially identify correctly (Fisher, 1971; Gorroochurn, 2012). With eight cups of tea presented in random order, the probability of the lady correctly identifying all four milk-first cups by guessing is 1 in 70, or 0.014. Assuming that she is unable to distinguish the cups, the most likely outcome will be that she guesses two out of four cups correctly. In repeated experiments, this will happen by chance 51.4% of the time. If she cannot really tell the difference, the probability of Fisher being fooled by a random occurrence where the lady happens to guess all four of the cups correctly is 0.014, or 1.4%. So if Fisher puts her to the test and she evaluates all the cups correctly – an outcome of the experiment that is unlikely to have occurred by chance – he can tentatively conclude that she can tell which liquid was poured first.

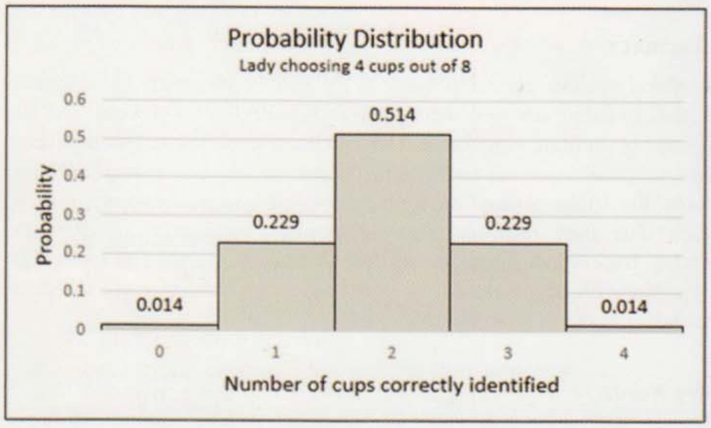


Figure 2. Possible outcomes of the "lady tasting tea" experiment.

While Fisher described the experimental design for the lady tasting tea, he did not tell us the outcome. However, Salsburg (2001) has it from a reliable source that the lady correctly identified all the cups. It is important to understand that by setting up this experiment and having her demonstrate her talent on one occasion, Fisher has not proved that the lady can make the necessary distinction. Even if she can't tell the difference, there is still a 1.4% chance that the outcome of the experiment was a random occurrence, albeit a very unlikely one. However, getting an unlikely result in an experiment like this is good evidence that Fisher's initial assumption, that she cannot tell the difference, may not be true. So, we can be reasonably safe in rejecting the idea that her success is just by chance and conclude, based on the result of this experiment, that she can tell the difference.

Fisher termed outcomes of well-designed experiments that are unlikely to have occurred by chance *statistically significant* outcomes, and a test designed to demonstrate this is a *test of significance*. Methods of inference commonly used in science to support or reject claims based on data, like the chi-square test or the independent-samples *t*-test, are also tests of significance (Moore et al., 2009).

One key point to remember from the story of Fisher and the lady is that before performing an experiment, we should consider all possible outcomes and how to interpret them. Fisher used the distribution in Figure 2 to forecast all possible results of the lady tasting tea because that was an appropriate distribution to model an experiment in which a series of random events occurred, with each event having one of two possible outcomes (Gorroochurn, 2012). In *The Design of Experiments*, he wrote:

> In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed on each one of them. (Fisher, 1971, p. 12)

Fisher called the distribution of all possible random outcomes the *null distribution*. In evaluating the outcome of an experiment, a scientist is testing whether or not the actual outcome conforms to one of the most likely outcomes predicted by the null distribution. In other words, he is testing the assumption that the result of the experiment is a highly probable outcome and is therefore likely governed mainly by chance. This assumption is called the *null hypothesis*. If the outcome of an experiment deviates significantly from the most likely outcomes predicted by the null distribution, then the result is statistically significant.

The null distribution we use depends on the design of our experiment and the type of data we collect. For example, biology teachers are familiar with chi-square distributions as they are used in classical genetics. Chi-square distributions (Figure 3) provide models of discrete data (that is, data derived from counting) where each datum randomly falls into one of two or more categories. Another important family of distributions are normal distributions that model continuous data where the measurements, like the heights of a group of people, can be meaningfully subdivided into smaller and smaller increments. Data that fit a normal distribution fall randomly around an average value, the mean, but tend to cluster near the mean and tail off to either side. To use the normal distribution, we need a large sample size and we need to know the standard deviation of the population, but this is usually not the case. The *t*-distribution, which we will use here in a case related to Darwin's finches, is a null distribution used as a stand-in for the normal distribution when we have a small sample. The *t*-distribution more closely resembles the normal distribution as sample size approaches 30.

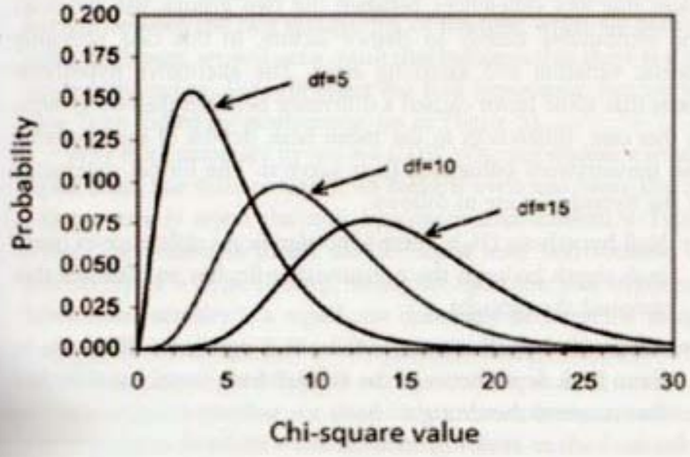


Figure 3. Distribution of chi-square values with different degrees of freedom. Modified from Engineering 360 (https://www.globalspec.com/reference/69568/203279/11-8-the-chi-square-distribution).

○ Using the *t*-Test to Make Inferences from Data

McDonald (2014, p. 14) summarizes the logic of inferential statistical procedures as follows:

> The basic idea of a statistical test is to identify a null hypothesis, collect some data, then estimate the probability of getting the observed data if the null hypothesis were true. If the probability of getting a result like the observed one is low under the null hypothesis, you conclude that the null hypothesis is probably not true.

In the following, I use data on the morphological characteristics of Darwin's finches to illustrate the logic of inferential statistics described by McDonald. The appropriate statistical test to analyze the finch data is the independent-samples *t*-test (referred to hereafter as the *t*-test). In the concluding section, I explain the parallels between Fisher's tea experiment and the *t*-test.

The activity provided by HHMI BioInteractive (2014) includes a spreadsheet with data from a randomized sample of 100 medium ground finches collected and measured by Peter and Rosemary Grant in the Galápagos; 50 of the birds survived the 1977 drought on Daphne Major, and the other 50 died as a result of the drought. The spreadsheet records data on a number of the birds' traits, among them beak depth. We want to know whether the birds' beak depth made a difference in their survival during the drought. We will use the *t*-test to compare the mean beak depths of survivors and nonsurvivors, to see if there is a statistically significant difference between them.

The assumption we begin with is our null hypothesis, that beak depth made no difference in the survival of the birds. If that is the case, then we expect that the group of survivors will have a very similar mean beak depth to that of the nonsurvivors, and that any difference we see between the survivor and the nonsurvivor group is due to sampling error. (To understand this expectation, it helps to look at HHMI BioInteractive [2017]. Readers are encouraged to experiment with this interactive activity. The related student worksheet provides guided instruction in some foundational statistics concepts.)

Figure 4, an image captured from HHMI BioInteractive (2017), shows two graphs plotting means of random samples drawn from a large population. In each case, 500 samples were selected at random, their means were calculated, and the means were plotted on the graphs to produce the distributions shown. Each of the 500 sample means provides an estimate of the actual mean of the population, which is 50 kg. The key point here is that we can expect any two random samples taken from the same population to have similar means. If you compare distribution A, consisting of 500 samples of 25 individuals each, with distribution B, consisting of 500 samples of 100 individuals each, you can see that the estimates cluster around the population mean of 50 kg in each case, but more tightly so in B. So, if we could take 500 random samples of 50 birds from our finch population and plot them, what would

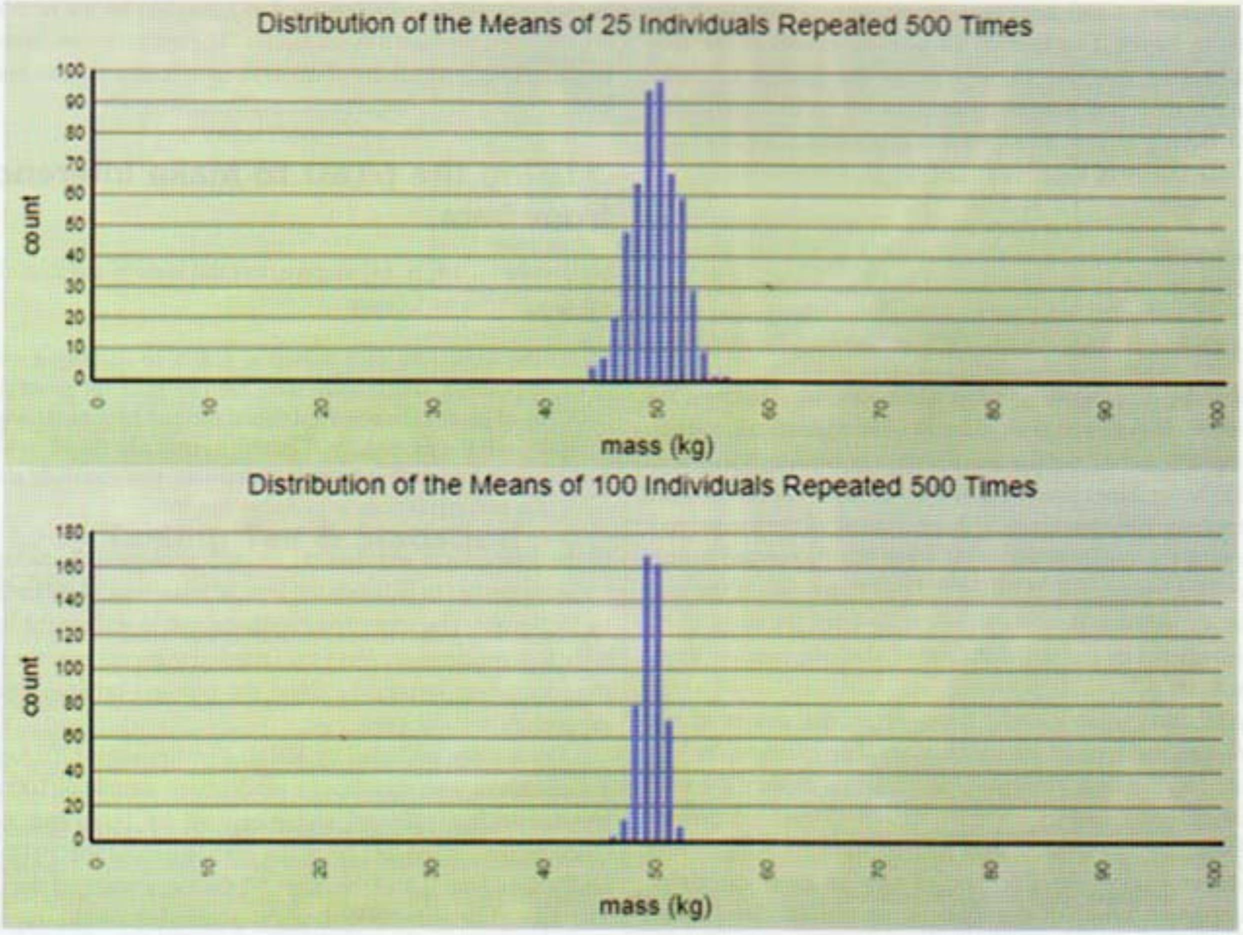


Figure 4. Means of random samples drawn from a large population. Image captured from HHMI BioInteractive (2017; copyright 2015 Howard Hughes Medical Institute, used with permission; https://www.BioInteractive.org).

we expect to see? We should see a similar normal distribution as in A and B, but with a range intermediate between A and B. So, if beak depth had no bearing on survival, our two samples should have very similar means, just as any two random samples drawn from the same population should have.

It is possible to get two random samples that have a statistically significant difference just by chance, but it is unlikely. Figure 5 shows two groups randomly selected from a large population that, by chance, have a statistically significant difference between their means as determined by a *t*-test. If these two samples were selected for an experiment, one as a control and the other as a treatment group, and the independent variable actually had no effect, a *t*-test would incorrectly result in the rejection of the null hypothesis. Statisticians call this a Type I error.

From the data provided in HHMI BioInteractive (2017), the nonsurviving finches have a mean beak depth of 9.11 mm, with a standard deviation of 0.88 mm. The surviving finches have a mean beak depth of 9.67 mm, with a standard deviation of 0.84 mm. Is this difference between the groups large enough to be considered statistically significant, or are they just random samples from the same population that, due to sampling error, have a difference in their means? To answer this question, we perform a *t*-test on our data. Inferential statistical procedures like the *t*-test have five basic steps. The steps in the *t*-test are applied to the finch data as follows.

Steps 1 & 2: Choose the Appropriate Statistical Test & State the Hypotheses

The *t*-test is appropriate for comparing the mean beak depths of two small samples of continuous data points. The null hypothesis states that any differences between the two groups will be small and attributable mainly to chance factors, in this case primarily genetic variation and sampling error. The alternative hypothesis states that some factor caused a difference between the two groups; in this case, differences in the mean beak depths of the survivors and nonsurvivors influenced their survival. The formal statements of the hypotheses are as follows.

- Null hypothesis (H_0): There is no significant difference in mean beak depth between the nonsurviving finches and finches that survived the drought.
- Alternative hypothesis (H_1): There is a significant difference in mean beak depth between the nonsurviving finches and finches that survived the drought.

Step 3: Choose the Decision Criterion

Next we determine what criterion to use in deciding whether the difference between means is large enough to be statistically significant. As Fisher wrote, we must "forecast all possible results of the

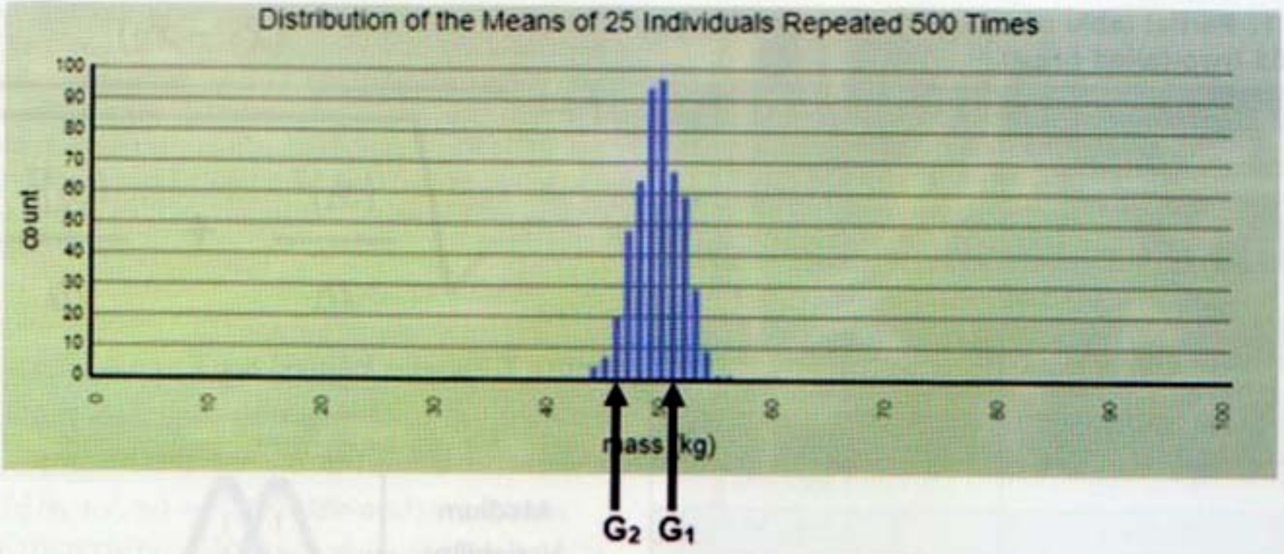


Figure 5. Two groups, $G_1(n_1 = 25, \bar{x}_1 = 51$ kg, $s_1 = 8$ kg) and $G_2(n_2 = 25, \bar{x}_2 = 46$ kg, $s_2 = 8$ kg), were randomly selected from a large population, yet they still have a statistically significant difference in their means. This is a Type I error. The probability of getting a difference this large between these two random samples by chance is $P = 0.032$ ($P < 0.05$). Image captured from HHMI BioInteractive (2017; copyright 2015 Howard Hughes Medical Institute, used with permission; https://www.BioInteractive.org).

experiment, and [decide] what interpretation shall be placed on each one of them" (1971, p. 12). The selection of the t-distribution as our null distribution forecasts all possible results. In order to interpret each of them, we need two numbers – a significance level and the degrees of freedom. The significance level, designated by the Greek letter alpha (α), is a choice you make as the scientist conducting the research. It represents the probability of incorrectly rejecting the null hypothesis, so it plays a similar role as Fisher's 1.4% probability that the lady guesses all the cups correctly by chance.

In other words, we are deciding how tolerant we will be of getting two random samples with a large difference just by chance and incorrectly rejecting the null hypothesis – a Type I error. The null assumes that any differences between the two groups will be small and attributable primarily to sampling error. It is customary to use $\alpha = 0.05$. This means that even if there really is no significant difference between the two groups, 5% of the time when we perform this experiment, we will get a result that indicates that there is a significant difference. We will reject the null hypothesis, but we will have been fooled by randomness (as in Figure 5).

Why is it customary to use $\alpha = 0.05$? Why not choose a smaller alpha level, like 0.01 or 0.001, to make it even less likely that we will incorrectly reject the null hypothesis and commit a Type I error? The reason is that a smaller alpha level will increase the probability of a Type II error, failing to reject the null hypothesis when there actually is a significant difference between the sample means. Choosing $\alpha = 0.05$ allows us to strike a balance between the risks of Type I and Type II errors.

The second number we need to set the decision criterion is the degrees of freedom – the number of values in the final calculation of a statistic that are free to vary. For the independent-samples t-test, the degrees of freedom equals the sum of the number of measurements in the two groups minus 2. There are 50 birds in each of the two groups, so the degrees of freedom (df) = 50 + 50 – 2 = 98.

Now we take $\alpha = 0.05$ and df = 98 to a table of critical t-values like Table 1. Notice that df = 98 is not on our table. When this is the case, we choose the critical t-value associated with the next lowest level on the table, in this case df = 80, with a critical t-value of 1.990. If you examine the table, you will see that the lower the degrees of freedom, the higher the critical t-value. The higher the critical value, the harder it is to achieve statistical significance. So, by choosing the next lower value for degrees of freedom, we make it less likely that we will commit a Type I error and reject the null hypothesis when we should not do so. We want to avoid being fooled by a random occurrence if possible.

Ultimately, the t-test comes down to comparing two numbers, the critical t-value that we just determined (1.990) and the observed t-value we will soon calculate using the finch data. Figure 6 shows a t-distribution centered on $t = 0$; this is the most likely outcome when we compare the means of any two random samples taken from the same population. (Recall the example distributions from HHMI BioInteractive [2017] in Figure 4.) To calculate a t-value, we subtract the means ($\bar{x}_1 - \bar{x}_2$ in Figure 7), so the t-value will equal zero when the two means are identical, and the value will be close to zero when their difference is small. The greater the difference between the two means, the farther to the left or right of center the observed t-value will fall. If $\bar{x}_1$ in the formula shown in Figure 7 is significantly greater than $\bar{x}_2$, then the calculated t-value will be greater than the critical value of 1.990. If $\bar{x}_1$ is significantly less than $\bar{x}_2$, then the calculated t-value will be less than –1.990.

The critical values for the finch study are marked as vertical lines on the graph in Figure 6. If the observed t-value is less than –1.990 or greater than 1.990, we reject the null hypothesis and conclude that the difference is statistically significant – that is, that the difference is large enough that it is unlikely to have occurred by chance. We used $\alpha = 0.05$, so 95% of the area under the curve falls in the center of the distribution between the two lines marking the critical values, and 5% of the area falls to the left and right of those lines. We have effectively divided the distribution into

Table 1. Partial table of critical *t*-values for $\alpha = 0.05$ (two-tailed *t*-test).

df	t_{crit}
1	12.706
2	4.303
3	3.182
4	2.776
5	2.571
6	2.447
7	2.365
8	2.306
9	2.262
10	2.228
•	•
•	•
•	•
50	2.009
60	2.000
70	1.994
80	1.990
100	1.984
Infinity	1.960

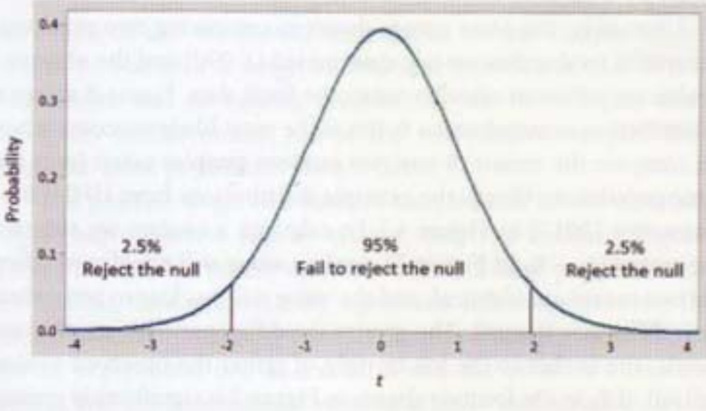


Figure 6. The *t*-distribution, which assumes that the null is true. Modified from Statistics By Jim (https://statisticsbyjim.com/hypothesis-testing/t-tests-t-values-t-distributions-probabilities/).

two regions: (1) the region under the center of the *t*-distribution, representing a small difference between the sample means that is compatible with the null hypothesis; and (2) the region under the extreme tails of the *t*-distribution, representing large differences between the sample means that are very unlikely to occur if the null hypothesis is true.

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

Figure 7. Equation for calculating a *t*-value.

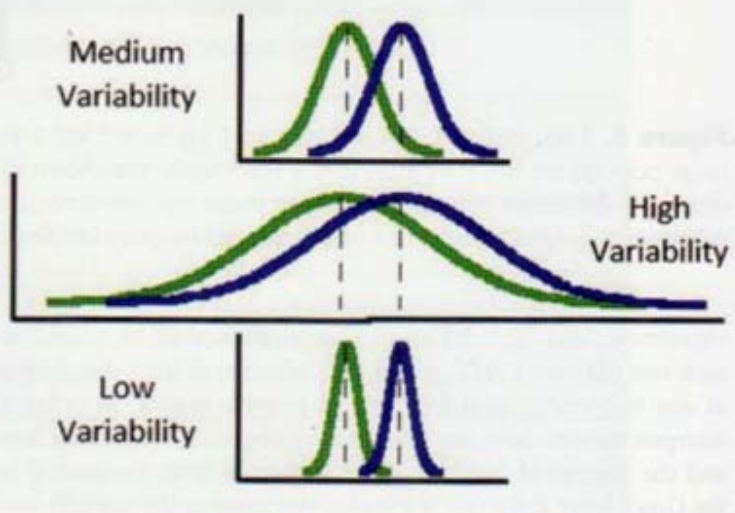


Figure 8. Examples of sample distributions with differing degrees of variability (from Web Center for Social Research Methods, http://www.socialresearchmethods.net/kb/stat_t.php).

If we repeatedly take two random samples from the same population that should have no differences between them and compare their means with a *t*-test, 95% of the time the *t*-value we calculate will fall in the 95% area of the distribution. Five percent of the time, we will get a *t*-value that falls to the left or right of the critical value indicating statistical significance (as in Figure 5), even though both samples were chosen randomly from the same population and there is no actual difference between them.

Step 4: Calculate the *t*-Statistic

The *t*-statistic can be thought of as a ratio of "signal to noise." The expression in the numerator, $(\bar{x}_1 - \bar{x}_2)$, the difference between the two means, is the signal. The greater the difference, the stronger the signal, the larger the *t*-value will be, and the more likely we will achieve statistical significance.

But statistical significance also depends on the noise: the variability in the two sample datasets (Figure 8). The smaller the variability in the sample data, the more likely we are to find statistical significance. The sample variances, the squares of the standard deviations (s_1^2 and s_2^2), represent the variability in the beak depth data: $s_1^2 = (0.84\,\text{mm})^2 = 0.71\,\text{mm}^2$ and $s_2^2 = (0.88\,\text{mm})^2 = 0.77\,\text{mm}^2$. Smaller variances make a smaller denominator, making the *t*-value larger and making it more likely that we will achieve statistical significance. Sample size also influences the outcome of the test. Smaller sample sizes will typically result in a larger denominator and more noise, and make it less likely that we find a statistically significant result. Larger sample sizes will increase the likelihood of achieving statistical significance.

Using the formula for calculating the observed t-value in our finch case, we get $t = 3.26$. The calculations are as follows (SQRT = square root):

$$(\bar{x}_1 - \bar{x}_2) = 9.67\,\text{mm} - 9.11\,\text{mm} = 0.56\,\text{mm}$$
$$s_1^2/n_1 = 0.71\,\text{mm}^2/50 = 0.0142\,\text{mm}^2$$
$$s_2^2/n_2 = 0.77\,\text{mm}^2/50 = 0.0154\,\text{mm}^2$$
$$s_1^2/n_1 + s_2^2/n_2 = 0.0142\,\text{mm}^2 + 0.0154\,\text{mm}^2 = 0.0296\,\text{mm}^2$$
$$\text{SQRT}(s_1^2/n_1 + s_2^2/n_2) = \text{SQRT}(0.0296\,\text{mm}^2) = 0.172\,\text{mm}$$
$$(\bar{x}_1 - \bar{x}_2)/\text{SQRT}(s_1^2/n_1 + s_2^2/n_2) = 0.56\,\text{mm}/0.172\,\text{mm} = 3.26$$

Step 5: Evaluate the Result

In the final step, we compare the calculated t-value of 3.26 with the critical t-value of 1.990. If the calculated t-value falls in the 5% region we marked in Figure 6, we reject the null hypothesis because it indicates that getting a difference between the two means as large as we observed is unlikely to have occurred by chance, assuming under the null that our two groups of finches actually are just two random samples from the same population.

This is where the P-value comes into play. The P-value tells us how much uncertainty we have when we conclude that two samples have a significant difference between them. In other words, with a significance level of $\alpha = 0.05$, there is a 5% or lower probability that the difference we found is attributable to sampling error.

Using a spreadsheet or statistical analysis software, we can get the actual P-value (just as Fisher was able to calculate for the tea experiment). For the finch data, $P = 0.0015$. As a reminder, the null hypothesis is that beak depth had no bearing on survival of the birds during the drought. So, if the null hypothesis is true, and we repeatedly drew two random samples from the same finch population to compare their means with a t-test, for every 1000 times we did this experiment we would get a significant difference only about one or two times by chance – not a very likely outcome. We conclude that the difference we observed between the survivors and nonsurvivors is statistically significant. It is unlikely to have occurred by chance and so, like Fisher in the tea experiment, we reject the null hypothesis and tentatively conclude that birds with larger beaks were, on average, better able to survive the drought.

How do we know that this isn't just a rare chance occurrence, like winning a lottery? How can we establish our conclusion with greater certainty? We repeat the experiment. If the significant difference we found was just a chance occurrence, it is unlikely to be repeated in subsequent experiments. The Grants have repeated the experiment. They have observed similar fluctuations in beak depth as droughts have alternated with rainy periods over multiple years. Figure 9 shows the fluctuations in mean beak depth of medium ground finches correlated with fluctuations in weather over a period of eight years.

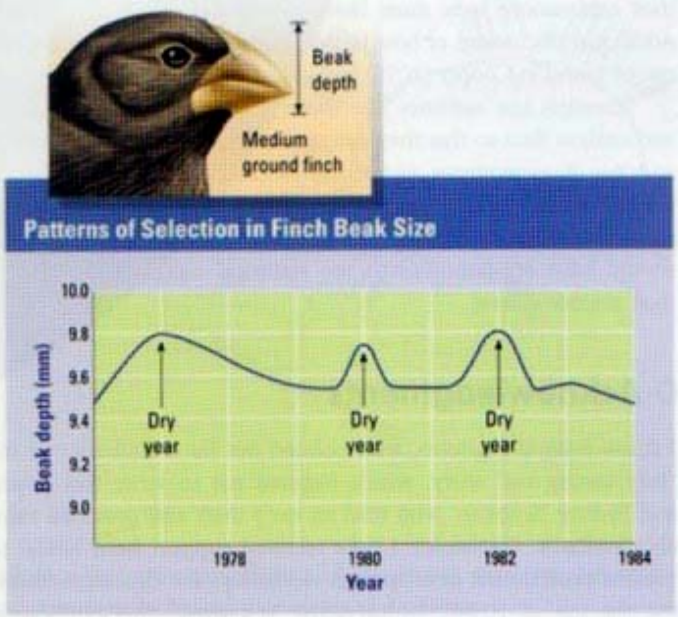


Figure 9. Fluctuations in mean beak depth of medium ground finches correlated with fluctuations in weather over a period of eight years. FREEMAN, SCOTT; HERRON, JON C., EVOLUTIONARY ANALYSIS, 4th, ©2007. Reprinted by permission of Pearson Education, Inc., New York, New York. (http://bodell.mtchs.org/OnlineBio/BIOCD/text/chapter14/concept14.4.html).

○ Conclusion

What do Fisher's test of the lady tasting tea and the t-test on the Grants' finch data have in common? In both cases, we started by assuming that there was no effect. The lady cannot tell the difference between milk-first and tea-first; beak depth has no bearing on survival. These are null hypotheses. Given these assumptions, we selected an appropriate null distribution to predict the outcomes of random events we expect during the experiment: the distribution in Figure 2 for the tea experiment; t-distribution for the finches. Experiments were carried out and data were collected and analyzed. In each case, the outcome deviated from what the null distribution predicted, by an amount that we previously determined in the design of the experiment. Fisher arranged matters so that the lady had a 1.4% probability of getting all cups correct by guessing, an outcome that was very unlikely to occur by chance. This was the P-value for Fisher's experiment. With the finches, we chose an alpha level of 0.05 as a decision point, so that if the difference between the means, given the variation in the samples, was great enough to produce a calculated t-value with a probability less than or equal to 0.05, we could say that a difference that large was unlikely to have occurred by chance. The actual P-value was 0.0015. In each case the outcomes were very unlikely, so we rejected the null hypothesis and concluded that there likely was a causal relationship.

Students in my AP Biology class performed the activity described here as one part of a unit introducing evolution as a major theme of biology. Emphasis in the unit was placed on understanding fundamental concepts, and constructing evidence-based arguments and explanations. Student understanding of statistics was not directly assessed; however, students did show improvement in their explanations of natural selection as the cause of adaptive changes in populations, and their explanations were more likely to reference empirical evidence. Additional discussions of how HHMI resources were used in the unit can be found in Cooper (2016) and in Lucci and Cooper (2019).

Scientists use statistics like those illustrated here to organize and analyze data so that they can make inferences from the dataset and use it as evidence (AAAS, 2011; NGSS Lead States, 2013; College Board, 2019). Understanding how scientists use statistics is an important component of biological literacy, and students should have opportunities to use statistical methods like this in their science classes.

○ Acknowledgments

I thank Brad Williamson, who pointed out the significance of the "lady tasting tea" story, which inspired me to write this paper, and Sydney Bergman, who read an early draft and provided valuable feedback. Disclaimer: I have received support from HHMI to present professional development workshops for educators featuring the use of HHMI BioInteractive resources. This publication was prepared and submitted independent of any HHMI support.

References

AAAS (2011). *Vision and Change: A Call to Action*. Washington, DC: American Association for the Advancement of Science.

College Board (2019). *AP Biology Course and Exam Description, rev. ed.* New York, NY: College Board.

Cooper, R. (2016, April 26). Need help teaching natural selection? Try this! [Web log post]. http://ncse.com/blog/2016/04/need-help-teaching-natural-selection-try-this-0017027.

Fisher, R.A. (1971). *The Design of Experiments, 8th ed.* (reprint). New York, NY: Hafner.

Gorroochurn, P. (2012). *Classic Problems of Probability*. Hoboken, NJ: Wiley.

HHMI BioInteractive (2014). Evolution in action: data analysis: finches dataset [Excel worksheet]. https://www.biointeractive.org/classroom-resources/evolution-action-data-analysis.

HHMI BioInteractive (2017). Sampling and normal distribution [interactive]. https://www.biointeractive.org/classroom-resources/sampling-and-normal-distribution.

Lucci, K. & Cooper, R.A. (2019). Using the I^2 strategy to help students think like biologists about natural selection. *American Biology Teacher, 81*, 88–95.

McDonald, J.H. (2014). *Handbook of Biological Statistics, 3rd ed.* Baltimore, MD: Sparky House. http://www.biostathandbook.com/HandbookBioStatThird.pdf.

Moore, D.S., McCabe, G.P. & Craig, B.A. (2009). *Introduction to the Practice of Statistics, 6th ed.* New York, NY: W.H. Freeman.

NGSS Lead States (2013). *Next Generation Science Standards: For States, by States*. Washington, DC: National Academies Press.

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York, NY: Holt.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal, 5*(2), 10–26.

ROBERT A. COOPER recently retired from Pennsbury High School, Fairless Hills, PA 19030, where he taught biology (general, honors, and AP); e-mail: bcooper721@gmail.com.