Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Vol. 32, 338–345, 2011

High- and Low-Tech Approaches to Teaching Statistical Skills in Introductory Biology

Gillian Gass

Department of Biology, Dalhousie University, Halifax NS B3H 4J1 CAN
(**Gillian.Gass@dal.ca**)

Students from a wide range of academic programs (science, arts, engineering, nursing, kinesiology, and so on) take Introductory Biology at Dalhousie University. We want to make sure that all of our students can benefit from the class, while also making sure that they are properly prepared for upper-year biology classes. Introducing students to basic statistical techniques important in biology, such as the chi-square test and comparing means using 95% confidence intervals, presents a challenge: how can we teach these skills to students from a broad range of backgrounds, while keeping the focus on biological (rather than purely statistical) concepts during class time? During this workshop, I'll share some hi-tech and low-tech solutions to this issue. Making use of technologies such as Camtasia Studio software and tablet PCs, I produced short pre-lab videos that students could watch on the course website to help them prepare for using statistical techniques in their lab exercises; during the lab, students could also consult a short stats-skills document included in the lab manual (please see the Supplemental Materials section for links to videos and to the Stats Primer document). We'll also look at how these videos, documents and exercises can be adapted for use in online courses. Workshop participants will have the opportunity to try out some of the tools used for making the videos, and to share their perspectives and ideas for successful integration of statistical techniques in biology labs and classrooms.

A basic guide to statistics begins on the next page. Addition-al supplemental materials for this workshop can be accessed through the following link:

Link to Supplemental Materials: www.ableweb.org/volumes/vol-32/gass/supplement.htm

Appendix: Statistics in Introductory Biology

In virtually every published primary research article in science, the Results section will contain the results of a number of statistical tests performed on the data collected by the researchers. Scientists use statistics to demonstrate mathematically that their results (for example, that plants treated with Fertilizer A grew larger than plants treated with Fertilizer B) are meaningful. For example, a biologist might weigh the plants in the two groups and find that the average weights calculated for each group were different values. The researcher would not stop there, but would then want to find out whether that the difference observed between the two groups is a legitimate or *significant* one (Fertilizer A really does promote plant growth better than Fertilizer B), rather than just an accident of chance (that is, the plants chosen for measurement in treatment A just happened to be heavier than the plants chosen for measurement in treatment B, even though there was no real difference in weights caused by the fertilizer used). Another biologist might have used a hypothesis to generate a prediction of the frequency of a particular phenotype in the offspring of a cross between two plants. When he or she actually performs that cross by breeding the plants together, do the frequencies match what was predicted? If they don't match exactly, are they close enough, or do the expected and the observed differences differ significantly?

In this Appendix, you will learn the basic statistical techniques that you will need to use in your laboratory activities. More advanced biology classes make use of more advanced statistical techniques, but many of these techniques are based on the same concepts you will use in your labs this year.

There are three sections to this Appendix:

I. Basic descriptive statistics: mean and standard deviation

II. Statistical tests: the chi-square test

III. Standard error and 95% confidence intervals

I. Basic Descriptive Statistics: Mean and Standard Deviation

When a group of measurements are taken, we often want to be able to characterize that group using descriptive statistics: for example, what was the middle or average weight of a plant in that group? How much did individual plants in that group tend to differ in weight from one another?

A common measure of the middle or average value used in biology is the **mean**. You have likely calculated means in secondary school math: the mean is found by adding up all of the observed values, then dividing by the number of observed values. The number of observations or data points is referred to as **n**. The Greek letter sigma (∑) indicates that you should sum up whatever comes immediately after the sigma. We can represent the procedure for finding the mean like this:

mean = ∑observed values / n.

In spreadsheet programs such as Microsoft Excel or Google Docs Spreadsheets you can calculate the mean using the "=AVERAGE" formula.

The variability of the data set (how much the values tended to differ from the mean) is described using the **standard deviation**. Together, the mean and the standard deviation tell you about the distribution of your observations: what value they cluster around, and how narrow or wide that cluster is. The larger the differences between each observation and the mean, the larger the standard deviation. The procedure for finding the standard deviation of a sample is more complex than the procedure for finding the mean. Some values will fall below the mean (resulting in a negative number), while some values will fall above the mean (resulting in a positive number), so the values need first to be squared so that all of the differences will be positive, then a square root taken. Here is the formula describing this procedure:

standard deviation = √ (∑ (observed value – sample mean)^2 / n-1)

You can use the "=STDEV" formula in spreadsheet software to calculate the standard deviation. You will need to know what the mean and standard deviation tell you about your set of data. We will build on this knowledge: you will learn how to use **n** and standard deviation to calculate a related value called standard error, so that when graphing your data you can quickly assess whether the means of two groups are likely to be significantly different, as in the case of the Fertilizer A and B treatments described above.

II. Statistical Tests: The Chi-Square Test

You will carry out and interpret a **statistical test** called the chi-square (χ^2) test of goodness of fit. This test will allow you to test hypotheses by comparing your predictions to your observations, as in the plant cross example described above. On the next page, you will find complete instructions for carrying out and interpreting the chi-square test. This test is just one of a very wide range of statistical tests used in science, and if you take upper-year courses in biology you will likely encounter many different statistical tests. However, these tests tend to share some common features:

- The purpose of the test is to help you decide whether or not to reject some hypothesis. The hypothesis itself will differ depending on the study being performed and the statistical test being used, but at the end of the test you should be able to say whether the hypothesis should be rejected or not. Notice that we do not say that the hypothesis is "supported" or "proven" simply that we fail to reject it.
- At the end of the mathematical operations involved in the test, you have computed what is called a **test statistic**. In the chi-square test, the test statistic is the χ^2 value that you calculate by adding up the squared differences between observed and expected values divided by the expected value; other types of tests (the Student's t-test or the Mann-Whitney U test, for example) have their own test statistics arrived at by their own procedures.
- Each test also requires that you find the number of **degrees of freedom**, which is related to the number of different categories being studied. Together, the test statistic and the degrees of freedom value will allow you to interpret the results of your test. When the test statistic and degrees of freedom have been calculated, you use these values to consult **statistical tables** (on paper or in computer databases) specific to each statistical test. In your chi-square test, the degrees of freedom value tells you which row of the table to look in. When you find your test statistic (or the nearest values to your test statistic) in this row, you then look to the top row of the table to find the corresponding p-value.
- The **p-value** is what helps you decide whether or not to reject your hypothesis. It is a probability with a value of between 0 and 1: the probability that the difference you are looking at is due to chance alone. In the chi-square test, the p-value tells you how likely it is that the differences between your expected and observed values are due to chance alone, rather than these differences being due to a faulty hypothesis that should be rejected. The p-value indicates whether the difference between the two things you are comparing is likely *significant*, or if the distance is instead *non-significant* – that is, due only to chance.
- You must decide what p-value will function as the cut-off line or threshold between significant or non-significant differences. We will use $p < 0.05$ is used: if your chi-square value corresponds to a p-value of less than 0.05, then we interpret that to mean that the probability that the differences are due to chance alone is too small to worry about, and as such we say that the differences are *significant* and reject our hypothesis. However, 0.05 is not the only possible threshold to use: some people might decide that $p = 0.1$ is sufficient, while others might decide that $p = 0.05$ isn't strict enough, and choose a smaller value such as $p = 0.01$. For this reason, the p-value is always stated explicitly when the results of a statistical test are given.

How to Perform the Chi-Square Test

The chi-square test is a test of goodness of fit: how well do the observations or measurements that you made fit with what you predicted that you would observe? For example, in a laboratory exercise you will be using Mendelian ratios to predict the number of corn kernels of each colour that you will observe, then actually making the observations by counting corn kernels. Even if the hypothesis is okay, the fit won't be perfect: observed numbers will almost always differ from predicted numbers at least a little, due to the effects of chance. For example: you counted seven rows of kernels to get your results. What if you'd counted a different seven rows? Your observed results would likely be a little bit different. The chi-square test is a way to determine whether these differences between what you observed and what you expected are *significant* – that is, due to a problem with your hypothesis – or *not significant* -- that is, just due to chance. If this difference between what was observed and what was expected is likely due to an incorrect hypothesis, then we will want to know that so we can reject our hypothesis. If the difference is likely due to chance alone, then we will not reject our hypothesis.

Below you'll see the equation used in the chi-square test. The tables provided in your lab manual work you through this equation in a series of steps to get a chi-square value at the end. The chi-square symbol is χ^2 (χ is the Greek letter chi), and the Greek letter sigma (**Σ) indicates that you take the sum.**

$$\chi^2 = \Sigma \; \frac{(\text{Observed - Expected })^2}{\text{Expected}}$$

To find the expected numbers for each class, take the proportion that you predicted that you would see, and multiply the total number observed by this proportion. For example, if you predicted a ratio of 2 pink kernels : 1 white kernel, then your prediction tells you to expect 2/3 of the kernels to be pink and 1/3 of the kernels to be white. If you counted a total of 2003 kernels, you find how many kernels of each colour you expect to see by multiplying the total number that you observed by the fraction expected to be that colour:

2/3 * 2003 = 1335 pink kernels expected (round to a whole number of kernels), and
1/3 * 2003 = 668 white kernels expected.

Notice that the number of pink kernels (1335) plus the number of white kernels (668) is equal to the total number of kernels observed (2003).

The chi-square test is a way of adding up all of the differences between what you observed and what you expected. The larger the difference between the observed and expected values, the greater the magnitude of the χ^2 test statistic. A large χ^2 value means that the differences between observed and expected were probably not due to chance; these differences are considered *significant*. But how large does the χ^2 value have to be for us to decide that the differences are significant? To interpret the χ^2 value, there are four more steps: 1) determine the degrees of freedom, 2) look up your χ^2 test statistic in the Table of Critical Chi-square Values, 3) find the p-value that corresponds to your χ^2 test statistic, and 4) interpret your p-value. These steps are explained below.

1) **Determine the degrees of freedom.** This step is simple: the degrees of freedom is equal to the number of classes (phenotypes) minus 1. When using the chi-square test to compare expected and observed genotype frequencies in a population (for example, when testing whether a population is in Hardy-Weinberg equilibrium), the degrees of freedom is equal to the number of different genotypes minus the number of alleles.
2) **Look up your χ^2 value in the "Critical χ^2 values" table.** This table will be provided on your lab benches, or you can look up a copy online. Find the row that matches your calculated degrees of freedom, then ignore the other rows. Scan across the numbers in that row horizontally to find the χ^2 value that is closest to your calculated value. Your χ^2 value will likely not match exactly any of the values in that row, but will instead fall between two values. In this case, you will focus on the two columns that your χ^2 value falls between.
3) **Find the p-value that corresponds to your χ^2 value.** Scan vertically up the column(s) to find the probability or range of probabilities that your data fits into. These are the decimal numbers in bold along the top row of the table, and they are called "p-values". If your χ^2 value fell between two values on the χ^2 table, then your p-value will also fall between two values at the top of the table; you can represent this by stating that "[lower p-value] < p < [higher p-value]" (for example, $0.05 < p < 0.2$).
4) **Interpret the p-value.** The p-value is the probability that the difference between what you observed and what you expected is due to chance, so a low p-value (0.01 or less) means that there is a very small probability that the differences are due to chance. If your p-value is 0.05 or smaller, then the differences between what you observed and what you expected to observe are considered *significant*, and you should reject your hypothesis. If your p-value is greater than 0.05, then the differences between what you observed and what you expected to observe are considered *not significant*, and you do not reject your hypothesis.

You may have noticed that in this test, our two choices are to reject or fail to reject the hypothesis: *when using the chi-square test, it is not appropriate to claim that you have "supported" or "proven" the hypothesis*. The hypothesis can only be disproven, never proven – because the very next run of this experiment, and the next chi-square test you perform, might find that the hypothesis should be rejected. It's not easy being a hypothesis!

III. Standard Error And 95% Confidence Intervals

It is very common in biology to compare groups of measurements to determine whether the groups differ. For example, for a research article published in 2006, Dalhousie biologist Dr. Jeffrey Hutchings and his colleagues measured the length of the pelvic fins of a large number of Atlantic cod, and then compared the measurements made on male cod to the measurements made on female cod. The researchers were able to determine that the pelvic fins of male cod are larger than the pelvic fins of female cod, and in their article they discussed these findings in connection with the evolutionary significance of sexual dimorphism in cod (Skjaeraasen, Rowe, & Hutchings, 2006).

There are a variety of different statistical methods to use when comparing two groups of measurements. You will see how the mean, standard deviation, and *n* can be used to construct a 95% confidence interval. In order to generate this interval, you will need to use the standard deviation and *n* to calculate another value, known as the **standard error**. This value is simple to calculate, but a bit tricky to understand. Here's the general idea: if you are making measurements of, for example, fin length in male Atlantic cod, you will not be measuring the fins of the entire male Atlantic cod population of the world. You will just be measuring the fins of a sample of male cod (say, 400 individuals), and then using that data to draw conclusions about the overall male cod population. Your set of data will have a certain mean value that you can calculate, and we call this mean the *sample mean*. Similarly, if it were possible for you to measure the fin length of every male Atlantic cod alive today, you could calculate the mean value for that group, and we would call this mean the *population mean*.

A measurable characteristic like pelvic fin length is likely to show small differences between all individuals, due to many different causes: genetic differences across multiple genes, environmental differences during development and adulthood, measurement error, and so on. For this reason, the lengths of cod pelvic fins are likely to be **normally distributed**: that is, if you were able to measure the entire male Atlantic cod population for the pelvic fin length trait, you would find that the measurements would vary around the mean in a normal distribution: most of the measurements would be close to the mean, with fewer and fewer measurements at values farther away from the mean, distributed symmetrically across the mean and peaking at the mean. A statistical rule called the "68-95-99.7 Rule" describes this phenomenon: in a normally distributed population, 68% of the individual measurements fall within one standard deviation of the mean, 95% fall within two standard deviations of the mean, and 99.7% fall within three standard deviations of the mean. (DeVeaux, Velleman, & Bock, 2008)

Since your sample only consists of some of the male cod in the population, your *sample mean* will likely be a bit different from the *population mean*. If you then measured another sample of 400 male cod, you would likely get yet another slightly different sample mean; and if you caught yet another sample of 400 male cod and measured their fins, you would get yet another sample mean. The sample means will vary around the population mean (Fig. 1), just as your individual observations will vary around your sample mean.

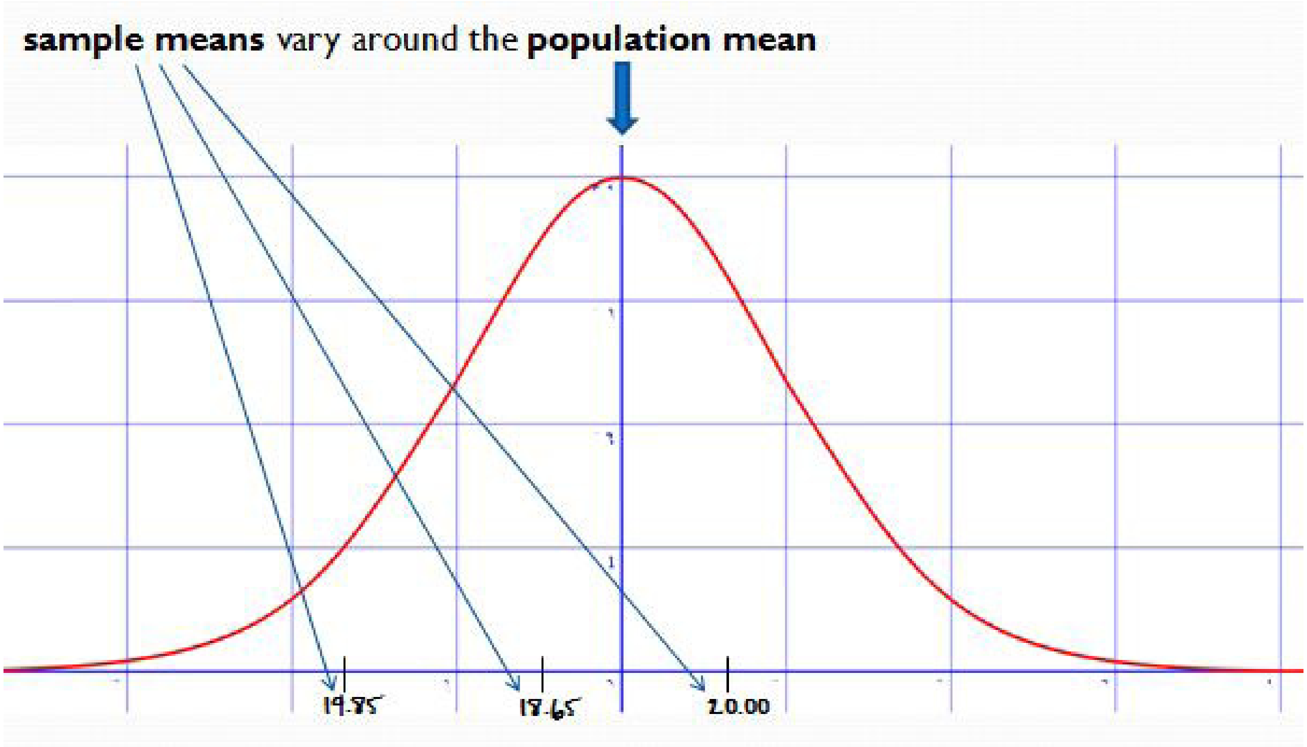

Figure 1. Sample means vary around the population mean.

Just as we use standard deviation as an indication of how variable the measurements in your sample are around the sample mean, we use the **standard error** as an indication of how variable these sample means are around the population mean.

To calculate standard error, you divide your sample's standard deviation by the square root of the number of observations in the sample, **n**:

$$\text{standard error} = \frac{\text{standard deviation}}{\sqrt{\mathbf{n}}}$$

Now that you know what standard error is and how to calculate it, what can you do with it? One method to compare two means that is often used by biologists is to calculate a **95% confidence interval** for each mean, consisting of the mean plus and minus two times the standard error. This interval is often represented visually by graphing the mean, plus error bars extending above and below the mean that correspond to two times the standard error of this mean.

How does this work? Remember that according to the 68-95-99.7 Rule, when considering the values in a normally-distributed population, 95% of these values will fall within two standard errors of the mean. Similarly, if samples are being taken from a population, 95% of the sample means will fall within two standard errors of the population mean. A 95% confidence interval indicates that you are 95% confident that the population mean can be found somewhere within this interval. (DeVeaux, Velleman, & Bock 2008).

If we construct 95% confidence intervals for two means, then by comparing the two intervals to see if they overlap, we can get a fairly good idea of whether the population means of the two groups are likely to be different.[1] (See Figure 2 at the end of this Appendix for an example.)

This is the same as saying that the difference between your two groups of measurements is likely to be significant: there is a difference between the two groups that is probably not due to chance alone. Confidence intervals are a quick way for us to test hypotheses: we begin with the hypothesis that there is no difference between the two groups, then calculate standard error and construct our confidence intervals. If the intervals overlap, then we do not reject our hypothesis that the groups do not differ. If the intervals do not overlap, then we reject our hypothesis that the groups do not differ.

Let's look at an example. Consider two large groups of aquatic newts, one group using blood worms as its primary food source and the other group using mantis shrimp as its primary food source. We want to know roughly whether the two groups of newts have different mean body weights. If we collect a sample of 40 newts from each group, weigh each newt, and find that the sample of newts eating blood worms has a mean body weight of 20.6 g and a standard deviation of 4.7 g, and the sample of newts eating mantis shrimp has a mean body weight of 25.2 g and a standard deviation of 5.2 g, how can we determine quickly whether the two groups' mean weights are likely different? First, we calculate the standard error for each sample. Then, we can make the comparison between the two groups graphically, by plotting each sample's mean plus and minus two standard errors as error bars. For each interval, we are 95% confident that it contains the population mean (that is, the mean weight of each large group of newts). If the bars do not overlap, then it is likely that the two groups' mean body weights are different.

The data is summarized in Table 1 on the next page, and the confidence intervals are graphed in Figure 2.

[1] This is not quite the same as doing a statistical test to compare two means at p=0.05, though: if the standard errors of the two samples are roughly equal, comparing 84% confidence intervals actually gives a closer approximation of this kind of test. (See Payton, Greenstone, & Schenker, 2003.) In this exercise, we will stick to comparing 95% confidence intervals.

Table 1. Body weight of newts eating blood worms or mantis shrimp as the primary food source: mean weight, standard deviation, sample size (n), standard error, and 95% confidence intervals are given.

Food Source	Mean Weight (g)	Standard Deviation (g)	n	Standard Error (g)	95% Confidence Interval
Blood worms	20.6	4.7	40	0.74	19.12 - 22.08
Mantis shrimp	25.2	5.2	40	0.82	23.56 - 26.84

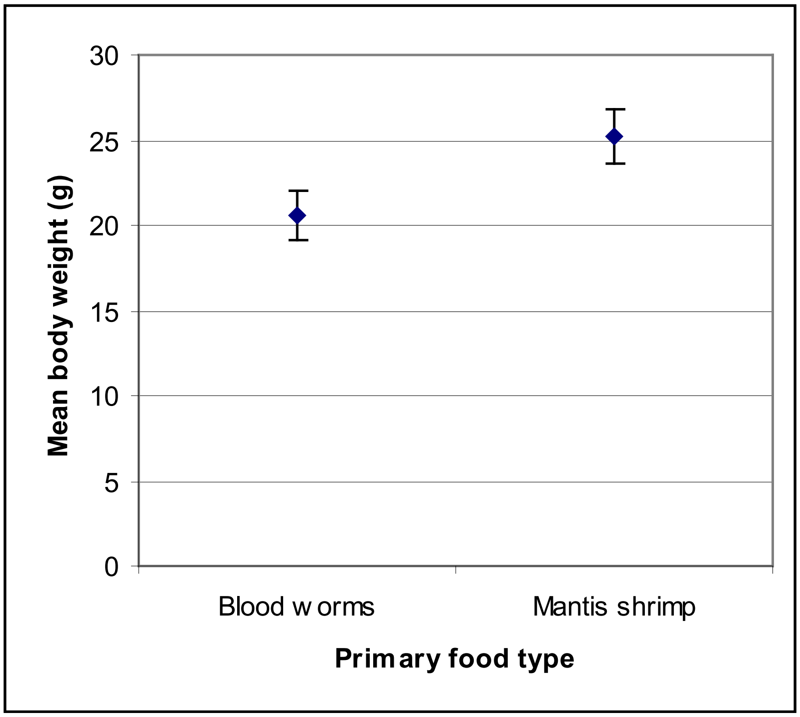

Figure 2. 95% confidence intervals for body weight of newts eating blood worms or mantis shrimp as the primary food type.

We can see from the table and figure above that the 95% confidence intervals for the mean body weight of newts in the two groups do not overlap, so we will **reject** our hypothesis that the two groups do not differ.

References

De Veaux, R. D., Velleman, P.F., & Bock, D.E. 2008. *Stats:Data and Models, 2nd Ed.* Boston: Pearson Addison-Wesley.

Parker, R.E. 1973. Introductory Statistics for Biology. *Studies in Biology, 43:* 1-122.

Payton, M.E., Greenstone, M.H., & Schenker, N. 2003. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science, 3 (34):* 6 pp.

Skjaeraasen, J.E., Rowe, S. & Hutchings, J.A. 2006. Sexual dimorphism in pelvic fin length of Atlantic cod. *Canadian Journal of Zoology, 84:* 865-870.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/**

Papers published in *Tested Studies for Laboratory Teaching: Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Although the laboratory exercises in this proceedings volume have been tested and due consideration has been given to safety, individuals performing these exercises must assume all responsibilities for risk. ABLE disclaims any liability with regards to safety in connection with the use of the exercises in this volume.

Citing This Article

Gass, G., M. 2011. High- and Low- Tech Approaches to Teaching Statistical Skills in Introductory Biology. Pages 338-345, in *Tested Studies for Laboratory Teaching,* Volume 32 (K. McMahon, Editor). Proceedings of the 32nd Conference of the Association for Biology Laboratory Education (ABLE), 445 pages. **http://www.ableweb.org/volumes/vol-32/?art=33**