Published in final edited form as:
Dev Biol. 2010 December 1; 348(1): 3–11. doi:10.1016/j.ydbio.2010.08.018.

Categorical Data Analysis in Experimental Biology

Bo Xu[a], **Xuyan Feng**[a], and **Rebecca D. Burdine**[*],[a]

[a]Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544 U.S.A

Abstract

The categorical data set is an important data class in experimental biology and contains data separable into several mutually exclusive categories. Unlike measurement of a continuous variable, categorical data can not be analyzed with methods such as the student's t-test. Thus, these data require a different method of analysis to aid in interpretation. In this article, we will review issues related to categorical data, such as how to plot them in a graph, how to integrate results from different experiments, how to calculate the error bar/region, and how to perform significance tests. In addition, we illustrate analysis of categorical data using experimental results from developmental biology and virology studies.

Keywords

categorical data; ternary diagram; statistical significance; confidence interval; chi-square

Introduction

Experimental data in biology typically fall into three major classes. Class I and Class II data are quantitative in nature. Class I comes from the measurements of continuous variables; for example, measuring the tumor volume in subcutaneous mouse xenografts, which can yield numbers of any value such as 52.1 mm^3 or 67.2 mm^3. Class II comes from the measurement of discrete variables; for example, counting the number of cilia in kidney tubules, which yield whole numbers such as 52, 75, etc. For these kinds of data, one can calculate the arithmetic mean and the variance of the sample, which can be used to estimate the mean and variance of the population, and apply a student's t-test when comparing with data from another sample. This type of statistical analysis was recently reviewed by Cumming et al (Cumming et al., 2007). Class III, the categorical frequency data set, is quite different. This class deals with the qualitative attributes of objects. Categorical data represent the distribution of samples into several mutually exclusive categories, which usually involves counting how many objects are in each qualitative category. The sum of categorical data typically equals 100%. For example, the percentage of female and male students in a class could be 51% and 49%. Categorical data is commonly found in the study of genetics (Hartl and Jones, 2005; Klug and Cummings, 2006). In fact, the categorical data Mendel generated from his work on peas helped to define genetic inheritance. In developmental and cell biology, categorical data are often acquired when analyzing mutant phenotypes or biological situations. For example, oocytes are activated after in vitro injection of spermatozoa, and the

[*]Author for correspondence Rebecca D. Burdine, Phone: (609) 258-7515, Fax: (609) 258-6730, rburdine@princeton.edu.

proportion of activated oocytes can vary from 0 to 100%. Oocytes from different mutant backgrounds may have different activation percentages. For example, you could have two different alleles of a gene that affect oocyte activation. After analysis, you determine that allele 1 produces 89% activated oocytes, while allele 2 produces 80%. When presenting these results, you would like to know how accurately they reflect the true phenotype if you were able to measure all the oocytes ever produced by these alleles. This can be calculated and displayed with an error bar. Furthermore, the percentages of activated oocytes produced by each allele are not equal, but they are close. How can we determine if this reflects real differences between the two alleles, or just experimental variation? In other words, are the differences statistically significant? The statistical methods used to analyze categorical data are different than those typically employed for Class I or Class II data, although it should be noted that the experimentalist can easily convert discrete data into categorical data and vice versa depending on his/her interpretations and experimental questions. In this article we will review approaches for presenting categorical data, methods to determine the error range or confidence interval, and ways to compare two data sets and determine if they are statistically different. In addition, we provide a new method of graphing and visualizing categorical data that is particularly useful when analyzing categorical data with three or four categories.

1. How to analyze proportional data of two categories

Categorical data sets can be divided into two main classes. The first contains what is termed ordinal variables or data that can be put into a ranked order. For example, in evaluating customer service one could assign categories of excellent, good, fair and poor. These categories have a natural order to them with excellent being the best or highest rank, and poor being the least or lowest rank. In biology, categories to describe data that have a natural order to them belong in this class. For example, in performing an siRNA experiment in tissue culture cells, results could be grouped into categories based on the level of target protein knock-down. Category labels could include complete knockdown, partial knockdown, and no effect. The important characteristic to note is that the categories have a natural rank order to them as partial knockdown is logically a state somewhere between no knockdown and complete knockdown. Categories that can be ranked or ordered have a relationship with each other and this affects the types of statistics that should be used for analysis. Using incorrect statistical methods may prevent you from unveiling meaningful trends in rank order data since the categories are mutually exclusive, but yet related. For more information on statistics to use in analysis of rank order data, please see (Agresti, 1996).

In this paper, we focus on statistics to analyze the second class of categorical data which contains nominal variables, or data categories with no natural ordering. For example, the gender of people has no natural ordering as it doesn’t matter if you list the number of females in a class first or second in a table. The number of flies with red eyes compared to those with white eyes from the same parents is another example of this type of categorical data. We note that the difference between ranked and nominal categorical data can be a matter of interpretation. For example, you could count the number of cells in a given experiment that are in mitosis versus those in interphase. This would generate nominal data because it doesn’t matter which set of data you list first or second. However if you categorize the same cells in terms of their phase in the cell cycle (G1, S, G2, M), the categories are now ranked because there is a logical order to cell cycle progression and applying statistical methods for rank order data to this set may provide additional information as to the relationships between the categories. However, the experimentalist could still treat this data as nominal if he/she recognizes that some information could be lost.

First, we will explain how to calculate the mean, the standard error (SE) and confidence interval (CI) for nominal data with two categories.

1.1 Mean and Standard Error

The mean of your sample is the average and is used to estimate the population mean. Standard error measures the spread of data around the mean and tells you how much variation there is from the mean in your data set. A low standard error indicates that on average your data points are very close to the calculated mean.

First let's calculate the mean. Let's say you get N independent samples from a large population, and P out of N show phenotype I, while (N-P) show phenotype II. We expect the proportion of the whole population showing phenotype I is $p_0 = P/N$. The sampling distribution of proportion p has the following attributes:

$$\mathrm{Mean:}\mathrm{p}_0=\frac{P}{N}$$

To calculate the standard error of this data set, you use the following equation:

$$\mathrm{SE}=\sqrt{\frac{p0*(1-p0)}{N}}=\frac{1}{N}\sqrt{\frac{P*(N-P)}{N}}\ \text{(Agresti, 1996)}$$

From these equations one can tell that the SE decreases when N increases, indicating that more samples (a higher N) will yield a result with smaller variation.

1.2 Confidence Interval

We obtained the mean for our sample with the formula above. But how can we determine the percentage of the whole population that displays phenotype I? We can not know the exact number unless we test the entire population. But with the results from a limited sample, we can determine a range in which the "true" result would fall with a certain confidence. This range is termed the confidence interval. For example, we can use statistical methods to determine the interval around the sample mean where the actual mean of the whole population displaying phenotype I is likely to reside with 95% confidence.

Since intensive computation is involved when calculating the exact CI based on binomial distribution, several approximations have been developed to estimate this interval. Here we introduce two methods:

Wald Method—The more straightforward approach is the simple asymptotic method (Wald method), which uses normal approximation.

$$\mathrm{CI}_{(\mathrm{Wald})}:\mathrm{p}_0\mathrm{z}*\sigma(\mathrm{p}),\mathrm{p}_0+\mathrm{z}*\sigma(\mathrm{p})\quad \text{with } \mathrm{p}_0=\frac{P}{N},\sigma(\mathrm{p})=\mathrm{SE}$$

The critical value is a cutoff value in a statistical test used to decide whether or not to reject a null hypothesis. In the equations above, z is the critical value from the standard normal distribution with a given confidence level (CL): z is 1.64 for 90% CL; z is 1.96 for 95% CL; z is 2.58 for 99% CL. The interval is symmetric around p_0, with the range of $z*\sigma(p)$on each side. However, this method can result in an aberration as sometimes it gives an interval with a boundary lower than 0 or higher than 1. This method is not typically suggested for scientific literature, though it gives a rapid, rough estimate.

Wilson method (Wilson, 1927)—This method doesn't cause an aberration and gives a closer approximation of the exact CI compared with the Wald method (Newcombe, 1998), and thus is recommended (Brown et al., 2001).

$$\mathrm{CI}_{(\mathrm{Wilson})}:\frac{p0+t/2}{1+t}\frac{\sqrt{p0*(1-p0)*t+t*t/4}}{1+t},\frac{p0+t/2}{1+t}+\frac{\sqrt{p0*(1-p0)*t+t*t/4}}{1+t}$$

Here t = z*z/N, and z has the same value as in the Wald method above. The Wilson interval is not symmetric around p_0, and range is larger towards 50% while it is shorter away from 50% as shown in Fig.1. Additional recommended methods include the Clopper-Pearson interval (the "exact" interval) or the Jefferys interval (Brown et al., 2001).

To illustrate the analysis of categorical data with two classes, we provide analysis of actual experimental data from our laboratory. The zebrafish *seahorse* mutant causes cysts to form in the pronephric tubules (Serluca et al., 2009). We identified 25 embryos out of 36 mutant embryos with cysts at 2.5 days post fertilization, while 11 did not have detectable cysts. In this case, P is 25 and N is 36. The expected proportion of embryos with cysts in a larger set of mutant embryos is P/N or 69.44%; with a standard error of this estimate of 7.68%. The calculated Clopper-Pearson interval for this data is (51.89%, 83.65%); the Wald interval is (54.40%, 84.49%); and the Wilson interval is (53.14%, 82.00%). For each CI, the confidence level (CL) is 95%. These results are shown in Fig.1 A.

In Fig.1 B, we use a hypothetical data set, where 18 embryos out of 19 develop cysts, to better demonstrate the differences among these three methods and the aberration that can occur using the Wald Method. For this data set, the expected proportion of embryos with cysts in a larger set of mutant embryos is P/N or 94.74% (18/19); with a standard error of this estimate of 5.12%. The calculated Clopper-Pearson interval is (73.97%, 99.87%); the Wald interval is (84.70%, 104.78%); and the Wilson interval is (75.36%, 99.01%). For each CI, the CL is 95%. Note that with this data set, the Wald interval has an aberration since the upper range of values is above 100%.

It is worth mentioning that the SE of categorical data decreases dramatically as the sample size increases. For example, if we found 50 embryos out of 100 with cysts, then the result is 50%, with a SE of 5%. If we found 500 embryos out of 1000 with cysts, then the result is 50%, with a SE of 1.58%. The SE is related to the width of the CI. The larger N is, the smaller the SE is and the narrower the CI is.

To facilitate the computation of confidence intervals for proportional data with two categories, we have designed a website to perform the calculation and plot the Wilson CI in a graph: https://webscript.princeton.edu/~rburdine/stat/2categories. The Clopper-Pearson interval can be calculated through this website: http://statpages.org/confint.html

1.3 Statistical significance

In order to tell whether the difference between two sets of results is significant or not, one typically uses the student's t-test for continuous and discrete variables. However, categorical data requires an alternative method. Pearson's chi-square test (Pearson, 1900) for the goodness of fit has been widely used in classical genetics studies (Hartl and Jones, 2005; Klug and Cummings, 2006). Here we introduce this method to determine significance of categorical data. .

In the experiments mentioned above, we counted 25 out of 36 *seahorse* mutant embryos with cysts. We designed a morpholino antisense oligo against the *seahorse* gene, which binds to *seahorse* mRNA and blocks its splicing and thus protein production. In the morpholino injected embryos, 90 out of 127 showed cysts (Serluca et al., 2009). We assume that *seahorse* morpholino injected embryos display a similar phenotype when compared to *seahorse* mutant embryos. Thus, the null hypothesis is these two groups behaved similarly in terms of cyst formation. To determine whether the null hypothesis is true, we first place the data into the so-called 2 by 2 contingency table as in Table1.

We put the raw data (11, 25, 37, and 90) into the corresponding cells under observed results. Note here we must use the actual counts and not the percentages. Then we calculate the total sum of each row and column. The expected value is calculated by multiplying the sums in the row and column where the cell resides, and dividing by the total number of embryos from each category, which is 163 in this example (11+25+37+90=163). For example, the cell of *seahorse* mutants with cysts has number 25, and the total number of that row is 36 while the total number of that column is 115. Thus, the expected counts of *seahorse* mutants with cysts are 36*115/163 = 25.40. Once the expected results have been calculated, we can use these values to calculate the chi-square value of the sample, which is a measure of the likelihood that the two experimental data sets exhibit similar phenotypes. A higher chi-square value suggests that there is a lower likelihood that the two data sets exhibit similar phenotype.

The chi-square is defined as:

$$\chi^2 = \sum_i \frac{(Oi - Ei)^2}{Ei}$$

Σ is the sign of summary. i stands for each cell we have. *Oi* is the observed value in that cell, and in this case, is the number of samples falling into each of the two categories. *Ei* is the expected value in that cell. For example, for the cell of *seahorse* mutants with cysts, the observed value is 25, while the expected value is 25.40.

From our example above we can calculate the chi-square value as follows:

$$\chi^2 = \frac{(11 - 10.60)^2}{10.60} + \frac{(25 - 25.40)^2}{25.40} + \frac{(37 - 37.40)^2}{37.40} + \frac{(90 - 89.60)^2}{89.60} = 0.027$$

Similar to the t-test, we will compare the computed chi-square value for our data with the critical value under a given CL and degree of freedom (provided in Table 4). In the example above, we have two experimental conditions and two categories of phenotype, thus we have a 2 by 2 table of results. The degree of freedom is obtained by multiplying (number of rows minus 1) and (number of columns minus 1) of the table. In this case it is (2-1)*(2-1) = 1. If we have a result table of 3 rows and 4 columns, the degree of freedom would be (3-1)*(4-1) = 6. If we choose to determine the 95% CL, then the critical value is 3.841 (Table 4). Since the chi-square we calculated is 0.027, which is far smaller than 3.841, the two sets of data do ***not*** show a statistically significant difference. Thus, we cannot reject the hypothesis that the morpholino injected embryos displayed similar defects compared to *seahorse* mutants in terms of cyst formation. This result is consistent with the fact that the phenotypes of the two populations are similar.

As mentioned in section 1.2, the confidence interval is affected by the sample size. If we look at the calculation to determine the Wilson interval, one can see that given the same CL and percentage data, the larger N is, the smaller the confidence interval is. Accordingly, when we compare groups to look for differences, the sample size also matters. How big a sample size you need for a given experiment, however, is difficult to determine without prior knowledge of the results. In experiments where the phenotypic expression is variable, a higher N value can reveal the difference between groups while a smaller N might not. For example, we count 20 samples from mutant A and 6 out of the 20 show phenotype I. Then we count 20 samples from mutant B and 8 out of the 20 show phenotype I. Although 30% (mutant A) is different from 40% (mutant B), since the N value is low, the chi-square value is only 0.44, and thus, the p value is about 0.51. As a result, we can not reject the hypothesis that these two mutants are actually similar with regards to phenotype I. However, if we count 200 samples of A and 60 show phenotype I, and we count 200 samples of B and 80 show phenotype I, the conclusion will be different. Although samples from A and B still show 30% and 40% expression of phenotype I respectively, the chi-square value is 4.396 and the p value is less than 0.05. Thus, the difference between the two groups is significant.

The chi-square test is very useful in comparing categorical data, but it has some limitations. First, the data sets being compared must be independent, and each set must not affect the others. In other words, each category must be mutually exclusive, and data can only be placed in one category and not in another. Seahorse mutants either have kidney cysts, or they don't. Secondly, when using a 2 by 2 table, the counts in every cell should be at least 5 (Norman and Streiner, 2000). If a smaller number of samples is involved, the Yates' correction or Fisher's exact test is recommended instead (Norman and Streiner, 2000). Alternatively, with the emerging power of fast computing, one can use Fisher' s exact test to compute the p-value of a 2 by 2 table online (http://statpages.org/ctab2×2.html)

2. How to analyze proportional data of three categories

In biology, experiments can generate data that can be distributed into more than two categories. For example, you could determine the number of different eye colors that exist in a class of students. The categories might be blue, brown, green, and other. When analyzing categorical data with more than two categories, similar statistical methods are used (described above). However, determining how to best display data of multiple categories can be difficult. It is important to display your data in a way that is accessible to your readers and easy to interpret.

2.1 How to present three component categorical data in a bar graph or ternary diagram

Experimental biologists typically use some form of bar graph to show percentage categorical data. For example, in an experiment of 100 samples from a population, 53 show phenotype A, 16 show phenotype B, and 31 show phenotype C. Traditionally, the result could be shown as in Fig.2a.

However, when you need to compare multiple experiments across categories, bar graphs can become cluttered and complicated (Fig.2b, 2c). Alternatively, a ternary diagram can help to visualize the data in a more direct way. Fig.2d shows the same result as in Fig.2b, 2c, with ternary diagram. One can easily compare the results from many populations within the diagram. Ternary diagrams have been used in the current studies of genomics and bioinformatics (Raymond et al., 2003; Steinke et al., 2006; White et al., 2007). Here we introduce a similar diagram to developmental biologists in order to represent and analyze the relative percentage proportional data of three categorical components. It is required that the sum of the three percentages be a constant, in this case, 100%. Fig.3 shows ternary diagram plots and how to interpret them. The diagrams graphically depict the percentages by plotting a point (M) within an equilateral triangle. First we determine how much of M is due to the percentage of samples with phenotype A (Fig.3a). A line is drawn from M parallel to BC, which crosses line CA at point D (line MD). Then we calculate the length ratio of line segments CD/CA. This ratio equals the percentage of samples with phenotype A that point M represents. Using the percentage abundance scale (the group of lines parallel to BC, divide CA into several segments of the same length), we can estimate M represents 53% of the samples are of phenotype A. As we can see, every point along a line parallel to BC shares the same percentage of phenotype A, and the closer a points locates to A, the higher percentage of A it represents. Thus, all the points on line BC represent 0% of A, while point A represents 100% of samples with phenotype A. Similarly, in order to get the percentage of B that M represent, we draw line ME parallel to CA, which crosses CA at point E. Then the ratio of AE/AB is the percentage of B that M represents (Fig.3b). Similarly, points on line AC shows 0% B, while point B means 100% of B. Fig.3c shows the calculation for C% (which is the ratio of BF/BC) and Fig.3d shows all the percentages together. The sum of ratios CD/CA, AE/AB and BF/BC is 100%.

Using a ternary diagram, one can easily plot proportional data of three categories and visualize the difference between data sets. For example, we have analyzed two alleles of the zebrafish *pkd2* mutant *cup; cup*^{tc321} and *cup*^{ty30b}. These two alleles seem to behave differently in how they affect left-right patterning as visualized by the placement of organs. In these mutants, the positioning of the heart, liver and pancreas can be divided into 3 mutually exclusive categories: situs solitus (ss; correct pattern), situs inversus (si; reversed pattern), and heterotaxia (ht; any pattern not ss or si) (Schottenfeld et al., 2007). In 97 mutant embryos from *cup*^{tc321}, the percentages of ss, si and ht are 35.0%, 33.0% and 32.0% respectively. In 98 mutant embryos from *cup*^{ty30b}, the percentages of ss, si and ht are 37.8%, 45.9.0% and 16.3% respectively (Table 2). These results are displayed with a bar graph (Fig. 4a) or a ternary diagram (Fig.4b). There are two advantages that a ternary diagram has over a bar graph as seen in this example. First, the presentation of results is clearer when a lot of results are displayed together. Imagine we want to compare the left-right patterning phenotypes from 10 different mutant alleles. We would need draw to draw 30 bins in a bar graph, while we only need 10 dots in a ternary diagram. Secondly, we can present the error region of our results, which can only be done accurately with ternary diagrams. Although error bars can be used in a bar graph when presenting data of two categories, they cannot be used accurately to describe the error region associated with data of three categories in a bar graph.

2.2 How to draw the error bar/region on your categorical data

We previously discussed how to calculate the error bar for proportional data with two categories. However, with three categories plotted in a ternary diagram, we cannot draw an error bar because the display is two dimensional and the bar is one dimensional. Thus, we need a two dimensional error region. Similar to the confidence interval in the previous section, the error region defines where the true result resides at a given CL. If we use statistical methods to draw a region of 95% CL around the point representing our result in the diagram, this method will give us the region containing the true result with 95% probability. Although this region can be accurately calculated, intensive computational power is required especially when the sample is large. Watson and Nguyen have suggested using the chi-square method to approximate this value (Watson and Nguyen, 1985).

Here is an example: In Human cytomegalovirus infected cells, three types of enveloped particles can be seen in the cytoplasm: virions, non infectious enveloped particles (NIEPs) and dense bodies (DBs). Among 357 virus particles examined in BAD*wt* virus–infected cells, 120 were virions, 150 were NIEPs and 87 were DBs, or 33.5%, 42% and 24.5% respectively (Feng et al., 2006). How can we draw the error region for this result in a ternary diagram at a given CI (for example 95%)? In our plot we denote A as virions, B as NIEPs and C as DBs. First we plot point M in diagram (Fig.5a) to represent our result, 33.5% A, 42% B and 24.5% C. Then we look at every point in the diagram to see whether it is in the 95% confidence region of M. For example, point L represents 32% A, 45% B and 23% C in the diagram. If we assume the true result is L, then we can draw the region of which the experimental results would fall into with 95% probability if we count 357 samples from a population with distribution L. This can be done accurately by computing the trinomial distribution, but here we use chi-square test to approximate and improve the efficiency. Our results from this calculation define the region within the green boundary shown in Fig.5a. Since point M resides inside the green line, point L is in the error region of M at 95% CL. Now let's look at point K, which represents 25% A, 40% B and 35% C. The blue line is the boundary of the 95% confidence region of which the experimental results would fall into with 95% probability if we count 357 samples from a population with distribution K. Since point M is outside the blue line, point K is not in the error region of M at 95% CL. In general, for a given point X, representing xa of A, xb of B and (100%-xa-xb) of C in a population, we use the following formula to approximately determine whether M is inside the error region of X at 95% CL. Again, we are utilizing the chi-square method.

$$\chi^2=\sum_i\frac{(Oi-Ei)^2}{Ei}=\frac{(O_A-E_A)^2}{E_A}+\frac{(O_B-E_B)^2}{E_B}+\frac{(O_C-E_C)^2}{E_C}=\frac{(O_A-xa*N)^2}{xa*N}+\frac{(O_B-xb*N)^2}{xb*N}+\frac{(O_C-(1-xa-xb)*N)^2}{(1-xa-xb)*N}$$

N=357, O_A=120, O_B=150 and O_C=87. We are comparing two sets of proportional data, each with 3 categories. Thus, the degree of freedom is (2-1)*(3-1) = 2. The critical value of chi-square distribution with 2 degrees of freedom at 95% CL is 5.991 (Table 4). If χ^2 is less than 5.991, then the 95% error region of point X includes M, and thus X is inside the error region of M at 95% CL. In the case of M and L, with the sample size of 357, M (33.5% A, 42% B, 24.5% C) is the observed value, while L (32% A, 45% B, 23% C) is the expected value. So O_A= 120, O_B=150, O_C=87; E_A=114, E_B=161, E_C=82.

$$\chi^2(\mathrm{M,L})=\sum_i\frac{(Oi-Ei)^2}{Ei}=\frac{(120-114)^2}{144}+\frac{(150-161)^2}{161}+\frac{(87-82)^2}{82}=1.372$$

For the chi-square value of M and K (25% A, 40% B 35% C), we have O_A= 120, O_B=150, O_C=87; E_A=89, E_B=143, E_C=125.

$$\chi^2(\mathrm{M,K})=\sum_i\frac{(Oi-Ei)^2}{Ei}=\frac{(120-89)^2}{89}+\frac{(150-143)^2}{143}+\frac{(87-125)^2}{125}=22.69$$

Because χ^2 (M,L) < 5.991 while χ^2 (M,K) >5.991, L is inside the error region of M while K is not. In order to draw the boundary of the error region of point M, we need to draw the curve with (xa, xb, 100%-xa-xb) defined by χ^2 = 5.991. The analytical solution to this equation can be complicated. Therefore, we have designed a web-based program to calculate the error region for a ternary plot. We tested 20,000 points inside the diagram to evaluate the χ^2, and then draw the boundary along those points whose χ^2 values are less than 5.991 as shown in Fig.5b.

Using this algorithm of chi-square distribution, we constructed the following website to plot three component categorical data in a ternary diagram and automatically draw the error

region at CL of 90%, 95% or 99%:
https://webscript.princeton.edu/~rburdine/stat/three_categories

With this website, after drawing the error region, one can use the mouse cursor to track the boundary of the region and read out the corresponding percentage values in the table below the plotting graph.

For proportional data with four categories, the data should be plotted in a three dimensional space. Instead of in a unilateral triangle, we use a pyramid plot and error region will be a 3 dimensional cloud. However, in order to illustrate the result in a two dimensional webpage, we project the data to the four sides of a pyramid, respectively. For each set of four categorical data, we have four ternary diagrams to illustrate the error region. A data point is considered to reside inside the confidence region in a pyramid plot if and only if this point is within the confidence region in all of the 4 triangle projections from the pyramid plot.

https://webscript.princeton.edu/~rburdine/stat/four_categories

2.3 How to determine whether one set of data is significantly different from another

Statistical significance is very useful in comparing two sets of data, to judge whether or not they are different from each other. As mentioned above, the error regions drawn in a ternary diagram can be used to visualize differences between data sets; alternatively, the chi-square test can also be used to judge the significance.

In the example of Human cytomegalovirus infected cells, 3 types of particles can be seen in the cytoplasm, virions, NIEPs and DBs. Among 357 virus particles examined in BAD*wt* virus–infected cells, 120 were virions, 150 were NIEPs and 87 were DBs. Among 320 virus particles examined in BAD*in*US24 virus–infected cells, 91 were virions, 154 were NIEPs and 75 were DBs (Feng et al., 2006). Figure 6a illustrates the two data sets plotted in the ternary diagram with their error region of 95% CL, according to the procedures mentioned above. Since the two data points (red dot and green dot) reside within each other's error region of 95% CL, we can not reject the hypothesis that the two different viruses affect virion particle formation in a similar way. Thus, the particle phenotypes of these two viruses are similar. On the other hand, if we analyze another virus mutant, and find out among 210 particles, 30 are virions, 40 are NIEPs and 140 are DBs. This data point (Fig.6a blue dot) does not reside within the wildtype virus error region of 95% CL, and there is no overlap between the two error regions (blue circle and red circle), we can reject the hypothesis that the two viruses affect virion formation in a similar way with 95% CL and state that they have different effects.

However, when the error regions overlap, the conclusion is less clear. In the example of *pkd2* mutant embryos we used before, there are two alleles, *tc321* and *ty30b*. Given the data presented (Schottenfeld et al., 2007), we ask whether or not these two alleles affect left-right organ patterning in different ways (Table 2). Similarly, we plot the result in Figure 6b. However, the pattern is different from Figure 6a. In this case, neither of the points resides within the other's error region of 95% CL, but the error regions overlap. The location of the data points seems to indicate that these alleles affect organ patterning differently, but since there is some overlap in their error regions, we should return to the chi-square test to be sure.

First we assume they are not different, in other words, they are just two sets of samples from the same population. This is our null hypothesis (H_0). Now we calculate the chi-square value of our data and compare it with 5.991 (the critical value of chi-square distribution of 2 degrees of freedom at 95% CI) as mentioned above. If the value is greater than 5.991, this suggests the likelihood that the two sets of experimental data are chosen from the populations with the same distribution is less than 5%. Thus, we would conclude that these two alleles behave differently in affecting left-right patterning with statistical significance. On the other hand, if the chi-square value of the data is less than 5.991, we can not reject H_0, that is to say, we can not claim they are different in affecting left-right patterning.

First we calculate the χ^2 value. Since we start with the null hypothesis, we need to calculate the percentages of the whole population. As discussed in part 2, the percentages of ss, si and ht in the whole population (our best estimation from the two data sets) would be:

$$\begin{gathered}\text{ss}:(34+37)/(97+98)=36.4\%;\\ \text{si}:(32+45)/(97+98)=39.5\%;\\ \text{ht}=(31+16)/(97+98)=24.1\%\end{gathered}$$

These are our expected values for the whole population under the hypothesis that the two data sets are drawn from the populations with the same left-right patterning distribution. Then we calculate the expected results in each experiment given these percentages. In other words, given the percentages of 36.4% ss, 39.5% si and 24.1% ht, what do we expect to observe in an experiment of 97 samples and what do we expect with 98 samples in another experiment?

Expected data in *tc321* (97 samples):

$$\begin{gathered}\text{ss}:97*36.4\%=35.3;\\ \text{si}:97*39.5\%=38.3;\\ \text{ht}:97*24.1\%=23.4\end{gathered}$$

Expected data in *ty30b* (98 samples):

$$\begin{gathered}\text{ss}:98*36.4\%=35.7;\\ \text{si}:98*39.5\%=38.7;\\ \text{ht}:98*24.1\%=23.6\end{gathered}$$

Now we summarize the chi-square values from all 6 cells in the table. Note, this method is mathematically equivalent to the method we used for two categories above. The chi-square value is:

$$\chi^2=\sum_i\frac{(Oi-Ei)^2}{Ei}=\frac{(34-35.3)^2}{35.3}+\frac{(32-38.3)^2}{38.3}+\frac{(31-23.4)^2}{23.4}+\frac{(37-35.3)^2}{35.3}+\frac{(45-38.3)^2}{38.3}+\frac{(16-23.4)^2}{23.4}=7.14>5.991$$

Since the chi-square value is larger than the critical value, it is not likely that two data sets are drawn from the population of the same left-right patterning distribution with 95% CL. Thus we reject the null hypothesis, and claim that the two alleles probably affect left-right patterning in different ways. To further illustrate this point, you can calculate the chi-square value in the virus example. In that case, the degree of freedom is 2, so the critical value of 95% CL is 5.991. The chi-square value between two virus infections is 2.91, and it is less than the critical value. So the chi-square test gives the same result as the ternary diagram, though since the ternary diagram is clear, one does not have to compute the chi-square value for this set.

Using Microsoft Excel, we can calculate the p_value for our data with the function CHIDIST(chi-square value, degrees of freedom). In a table of j rows and k columns, the degree of freedom is (j-1)*(k-1). So Table 2 has 2 degrees of freedom. We can calculate p = CHIDIST(7.14, 2)= 0.028. Since 0.028 is less than 0.05, we consider this to be statistically significant. By convention, a p_value less than 0.05 is considered statistically significant. Thus, if the two alleles of *cup* do have the same left-right patterning distribution, the chance that such an event could occur is less than 1 in 20.

The chi-square test is not limited to a 2 by 3 table as mentioned here. However, the critical chi-square value is affected by the CI depending on different degrees of freedom (so it is not always 5.991). One can look for the critical value in Table 4. For a quick guide, at 95% CI, the critical value is 3.842 for one degree of freedom, 7.815 for three degrees of freedom and 9.489 for four degrees of freedom. One thing worth mentioning is that the chi-square test also has its limitations as mentioned in part I. If more than 20% of the expected values in the cells of a data table are less than 5, instead of chi-square, Fisher's exact test should be used (Norman and Streiner, 2000).

The excel file which will automatically perform the chi-square test can be downloaded at: http://www.princeton.edu/~rburdine/stat/chi_square_test.xls

3. How to combine data from different experiments

Finally, we wish to emphasize how to properly combine data from different experiments. Let's say you perform two experiments with two different samples from the same population and each experiment gave a set of proportional data. Based on these two results, what is our best estimate about the proportional distribution of the whole population if we could measure all of them? For example, we discovered a new mutant with left-right patterning defects and analyzed it as described above for *cup*. We analyzed 20 (N1) mutant embryos in the first experiment, and found the number of ss, si and ht were 3 (A1), 8 (B1), and 9 (C1) respectively. Therefore, the percentage of ss, si and ht are 15%, 40% and 45%. In the next experiment, we analyzed another 80 (N2) embryos from the same parents, and found the number of ss, si and ht were 40 (A2), 22 (B2), and 18 (C2), respectively. Thus, the percentage of ss, si and ht are 50%, 27.5% and 22.5%. What should we report as the overall ratio of ss, si and ht in the whole population of this mutant based on our observations? Can we calculate the percentages in each experiment and then average them? No! Taking the average of the percentages eliminates the difference of sample size between the two experiments. To be correct, the percentages of ss, si and ht in the whole population should be reported as:

$$\begin{gathered}\mathrm{ss}:(A1+A2)/(N1+N2)=(3+40)/(20+80)=43\%,\\ \mathrm{si}:(B1+B2)/(N1+N2)=(8+22)/(20+80)=30\%,\\ \mathrm{ht}:(C1+C2)/(N1+N2)=(9+18)/(20+80)=27\%.\end{gathered}$$

The results are different from the arithmetic average of the percentages from the two experiments as you can see by using the incorrect method:

$$\begin{gathered}\mathrm{ss}'=(15\%+50\%)/2=32.5\%,\\ \mathrm{si}'=(40\%+27.5\%)/2=33.75\%,\\ \mathrm{ht}'=(45\%+22.5\%)/2=33.75\%\end{gathered}$$

It can be proven that, in general, if we have done the experiments m times, and each time we have N_i sample, of which we observed A_i of phenotype P, B_i of phenotype Q, C_i of phenotype R, …(Ai, Bi, Ci… are the numbers of counts, not the percentages.) Then the best estimate of the whole population should be:

$$p=(A1+A2+\ldots+Am)/(N1+N2+\ldots+Nm);$$
$$q=(B1+B2+\ldots+Bm)/(N1+N2+\ldots+Nm);$$
$$r=(C1+C2+\ldots+Cm)/(N1+N2+\ldots+Nm);\ \ldots\ldots$$

However, to be cautious one should first apply a chi-square test to the individual results from different experimental repeats before combining all of the data together. If some repeats show results which are statistically significant from others, the researcher may need to carefully look at the conditions of each experimental repeat, and/or reconsider the hypothesis.

4. Websites to facilitate the analysis of categorical data

Plot data with two categories and draw the Wilson CI:

https://webscript.princeton.edu/~rburdine/stat/2categories

Calculate the Clopper-Pearson interval for data with two categories:

http://statpages.org/confint.html

Calculate the p-value of a 2 by 2 table with Fisher' s exact test:

http://statpages.org/ctab2×2.html

Plot data with three categories and draw the Wilson CI:

https://webscript.princeton.edu/~rburdine/stat/three_categories

Plot data with four categories and draw the Wilson CI:

https://webscript.princeton.edu/~rburdine/stat/four_categories

Excel file to calculate chi-square value:

http://www.princeton.edu/~rburdine/stat/chi_square_test.xls

5. R packages for Websites

R is a programming language for statistical computing (http://www.r-project.org/). Many calculations mentioned here can also be executed by installing R and some relative packages, for example, to compute the Wald interval, Wilson interval, etc. For readers who are familiar with R/Bioconductor, please refer to this online file http://cran.r-project.org/web/packages/MKmisc/MKmisc.pdf for more details.

Acknowledgments

We thank Prof. Ned Wingreen, Prof. Carlos Brody, Prof. John Story, Dr. Monica Skoge, Liang Li and Andrew Miri for helpful discussions; Andrew Miri, Prof. Eric Wieschaus, and members of the Burdine lab for critical reading of the manuscript; and Drs. Arcady Mushegian and Hua Li for their helpful critique of the manuscript and suggestions related to R/Bioconductor.

References

Agresti, A. An introduction to categorical data analysis. Wiley; New York: 1996.

Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. Statistical Science 2001;16:101–133.

Cumming G, Fidler F, Vaux DL. Error bars in experimental biology. J Cell Biol 2007;177:7–11. [PubMed: 17420288]

Feng X, Schroer J, Yu D, Shenk T. Human cytomegalovirus pUS24 is a virion protein that functions very early in the replication cycle. J Virol 2006;80:8371–8. [PubMed: 16912288]

Hartl, DL.; Jones, EW. Genetics : analysis of genes and genomes. Jones and Bartlett Publishers; Sudbury, Mass: 2005.

Klug, WS.; Cummings, MR. Concepts of genetics. Pearson/Prentice Hall; Upper Saddle River, NJ: 2006.

Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998;17:857–72. [PubMed: 9595616]

Norman, GR.; Streiner, DL. Biostatistics : the bare essentials. B.C. Decker; Hamilton: 2000.

Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London: 1900.

Raymond J, Zhaxybayeva O, Gogarten JP, Blankenship RE. Evolution of photosynthetic prokaryotes: a maximum-likelihood mapping approach. Philos Trans R Soc Lond B Biol Sci 2003;358:223–30. [PubMed: 12594930]

Schottenfeld J, Sullivan-Brown J, Burdine RD. Zebrafish curly up encodes a Pkd2 ortholog that restricts left-side-specific expression of southpaw. Development 2007;134:1605–15. [PubMed: 17360770]

Serluca FC, Xu B, Okabe N, Baker K, Lin SY, Sullivan-Brown J, Konieczkowski DJ, Jaffe KM, Bradner JM, Fishman MC, Burdine RD. Mutations in zebrafish leucine-rich repeat-containing six-like affect cilia motility and result in pronephric cysts, but have variable effects on left-right patterning. Development 2009;136:1621–31. [PubMed: 19395640]

Steinke D, Salzburger W, Braasch I, Meyer A. Many genes in fish have species-specific asymmetric rates of molecular evolution. BMC Genomics 2006;7:20. [PubMed: 16466575]

Watson GS, Nguyen H. A confidence region in a ternary diagram from point counts. Mathematical Geology 1985;17:209–213.

White WT, Hills SF, Gaddam R, Holland BR, Penny D. Treeness triangles: visualizing the loss of phylogenetic signal. Mol Biol Evol 2007;24:2029–39. [PubMed: 17630280]

Wilson EB. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 1927;22:209–212.

Abbreviations List

CI	confidence interval
CL	confidence level
SE	standard error

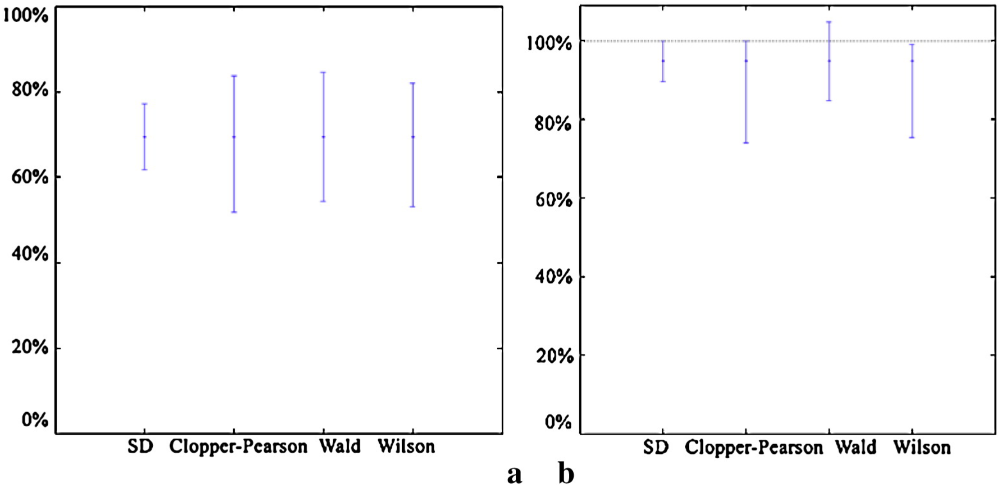

Fig.1. Comparison of Confidence Interval (CI) with three different method

a. Comparison of 95% confidence interval estimates calculated using Clopper-Pearson interval, Wald interval and Wilson interval for the *seahorse* mutant data set. Standard error (SE) for this data set is also shown. The y-axis denotes the proportion of mutant embryos with kidney cysts at 2.5 days. Our experimental result is 69.44%, indicated by the blue dots. On the x-axis, column 1 displays the SE, column 2 displays the calculated Clopper-Pearson interval, column 3 displays the calculated Wald interval and column 4 displays the calculated Wilson interval. Note that the upper parts of the error bars in column 2 and 4 are shorter than the lower parts, while they are of equal length in 1 and 3.

b. A hypothetical data set with the proportion of embryos having cysts as 94.74%. The difference between the lengths of the upper and lower parts of the error bars in column 2 and 4 is more obvious than in A. Note the aberration in the Wald interval which states the result could be greater than 100%.

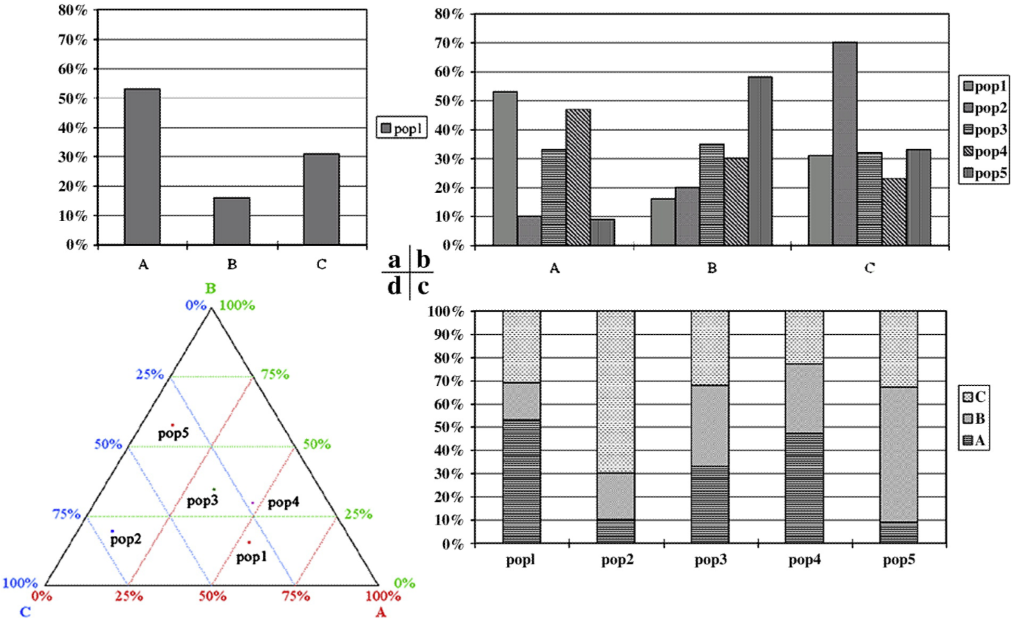

Fig.2. Presenting categorical data with traditional bar graphs and a ternary diagram

a. Clustered bar graph showing the distribution of phenotype A, B and C in samples from population 1(pop1).

b. Clustered bar graph to show the distributions of phenotype A, B and C of samples from 5 different populations, named pop1 through pop5.

c. Stacked bar graph to display the same data sets as in b.

d. Ternary diagram to display the same data sets as in b and c. Each dot represents the unique distribution of phenotypes from each population, marked as pop1 through pop5.

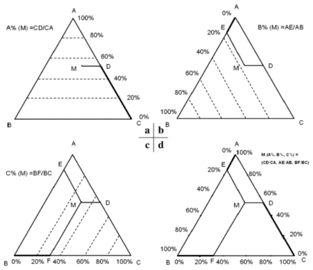

Fig.3. Interpretation of a ternary diagram

a. Line MD is parallel to BC, and crosses line CA at point D. The ratio of line segments CD/CA equals the percentage of samples with phenotype A that point M represents. Every point along a line parallel to BC shares the same percentage of phenotype A, and the closer a points locates to A, the higher percentage of A it represents. Thus, all the points on line BC represent 0% of A, while point A represents 100% of samples with phenotype A.

b. Line ME is parallel to CA, and crosses CA at point E. Then the ratio of AE/AB is the percentage of B that M represents. Similarly, points on line AC shows 0% B, while point B means 100% of B.

c. Line MF is parallel to AB, and crosses BC at point F. Then the ratio of BF/BC is the percentage of C that M represents. Similarly, points on line AC shows 0% B, while point B means 100% of B.

d. All the percentages are shown together. The sum of ratios CD/CA, AE/AB and BF/BC is 100%.

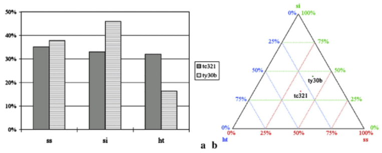

Fig.4. Comparison of bar graph and ternary diagram

a. Bar graph

b. Ternary diagram

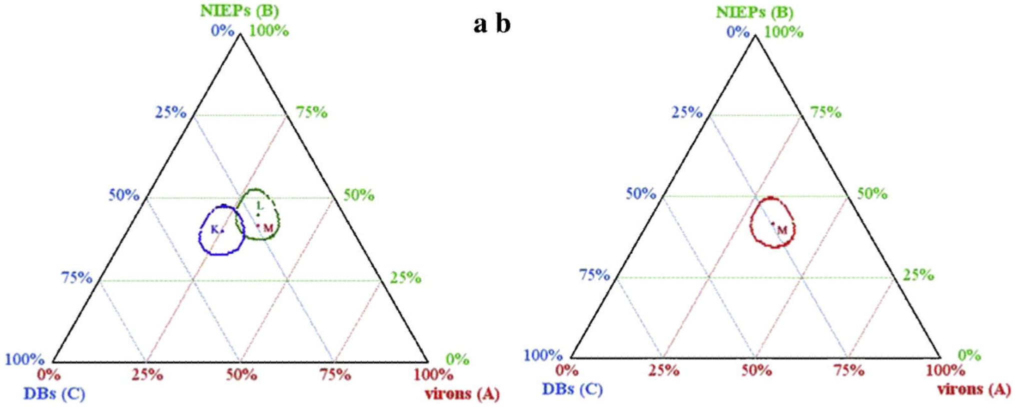

Fig.5. Calculation of error region from a given data set with 95% CL

M, the red dot, is the point of the result, representing 33.5% A, 42% B, and 24.5% C from 357 samples.

a. Point L represents 32% A, 45% B and 23% C, and green line is the boundary of 95% error region centered in L with 357 samples. Point K represents 25% A, 40% B and 35% C, and blue line is the boundary of 95% error region centered in K with 357 samples. Since blue line does not surround point M, point K is not in the error region of M at 95% CL.

b. With the method illustrated in Fig.5a, the error region of M at 95% CL is calculated and drawn with a red boundary. Thus any data we obtain from similar experiments counting virions should fall within this boundary with 95% possibility.

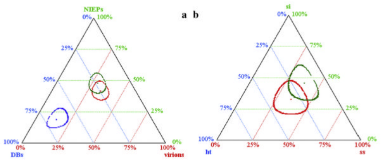

Fig.6. Using ternary diagram to determine the difference between data set

a. Comparison of wild type virus (red point represents the categorical data, while red circle represents the error region of 95% CL) with mutant virus (green point represents the categorical data, while green circle represents the error region of 95% CL), and a hypothetical data set (blue point represents the categorical data, while blue circle represents the error region of 95% CL). Note: the red point resides within the green circle and the green point resides with the red circle. The red point does not reside within the blue circle and the blue point does not reside with the red circle

b. Comparison of two alleles of *pkd2* mutants. *tc321* is in red, and *ty30b* is in green. Error region is drawn of 95% CL. Note: neither of the two points resides within the other one's error region, but there is overlap between the two circles.

Table 1

How to compute the expected value from experimental result

Observed results	**no cysts**	**cysts formed**	**Total**
seahorse mutant	11	25	36
morpholino injected	37	90	127
Total[*]	48	115	163
Expected results	**no cysts**	**cysts formed**	**Total**
seahorse mutant	10.60	25.40	36
morpholino injected	37.40	89.60	127
Total[*]	48	115	163

[*] Row 2,3,4 are experimental data, with row 4 the sum of row 2 and 3. Row 6,7,8 are expected values calculated from the experimental data, with row 8 the sum of row 6 and7.

Table 2

The phenotypes of two *cup* alleles

allele	n	ss (%)	si (%)	ht (%)
tc321	97	34 (35.0%)	32 (33.0%)	31 (32.0%)
ty30b	98	37 (37.8%)	45 (45.9%)	16 (16.3%)

Table 3

Observed and expected left-right patterning distributions of two *cup* alleles

		Observed (experimental data)			Expected (from null hypothesis)		
allele	n	ss	si	ht	ss	si	ht
tc321	97	34	32	31	35.3	38.3	23.4
ty30b	98	37	45	16	35.7	38.7	23.6

Table 4

Critical values of chi-square test

Critical value[*]	p_value					
degrees of freedom	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	2.706	**3.841**	5.024	6.635	7.879	10.827
2	4.605	**5.991**	7.378	9.210	10.597	13.815
3	6.251	**7.815**	9.348	11.385	12.838	16.266
4	7.779	**9.488**	11.143	13.277	14.860	18.466
5	9.236	**11.070**	12.832	15.086	16.750	20.515

[*] This table is computed with Microsoft Excel, values of 95% CI are shown in bold.