Genetics Society of America
PREP
Peer-Reviewed Education Portal

In-Class Exercise: Basic Probability and Chi-Squared Tests

Margaret A. Bakewell[1] and Patricia J. Wittkopp*[1,2]

[1]Department of Ecology and Evolutionary Biology
[2]Department of Molecular, Cellular, and Developmental Biology
University of Michigan, Ann Arbor, MI

** Corresponding Author: wittkopp@umich.edu*

Accepted for publication November 1, 2013

Citation:

> Bakewell, Margaret A. and Wittkopp, Patricia J. (2013). Basic Probability and Chi-Squared Tests. *Genetics Society of America Peer-Reviewed Education Portal (GSA PREP)*: 2013.005; doi: 10.1534/gsaprep.2013.005

Synopsis:

This resource describes an inquiry-based, in-class exercise designed for students working in small groups. We have used it in discussion sections of 20-30 students each from a class of over 400 total students, but it could also be adapted for use with the full class in a lecture hall setting. It is designed to review and enrich student understanding of probability, how probabilities of individual events can be combined to make predictions about more complex outcomes, and how observed data can be compared to a null model based on probabilities using a chi-squared test. These skills are used extensively for classical genetic analysis. We assume that a more complete presentation of these topics has been provided to students during lecture or in assigned readings prior to using this exercise. Throughout the activity, peers and instructors guide students through the process of developing and solving problems using probabilities and chi-squared tests in small groups.

Introduction:

Understanding probability, combining probabilities to make predictions about outcomes, and evaluating the fit of observed data to predictions based on a null model are critical skills for classical genetic analysis. The goal of this activity is to give students practice working with these skills. This is an inquiry-based activity. Students are guided by the instructors through the process of developing and solving probability-based problems in small groups.

Estimated time:

50 minutes

Group formation:

We recommend dividing students into groups of 3 or 4, ideally with a diversity of backgrounds in each group. One way to do this is to divide up the students based on other classes they have taken or by class year (freshman/sophomore/junior/senior). If it seems that the groups will be very unequal, you can simply have the students count off to form random groups. It is not suggested to allow them to form their own groups.

Part 1:

Probability: Intro (7 minutes)
Distribute the handout to students.

Provide a brief introduction/review of probability, including the multiplication “AND” rule and addition “OR” rule as well as defining *independent* and *mutually exclusive* events. These are also written in the student handout. For the multiplication rule, emphasize that it only applies to independent events. For the addition rule, explain the meaning of mutually exclusive (e.g., you could draw a Venn-type diagram with circles that do not overlap to represent mutually exclusive events). Give the students a chance to ask questions as you go. Also ask the students to supply some extra examples for the different scenarios, preferably examples not related to cards or the phenotypes being used in the handout. This will probably be an easy question for them, which should help them feel comfortable speaking up in class.

Get a count of the males and females in the class by asking students to raise their hands (allowing them to self-identify as male or female) and write the counts on the board. Next, ask students to raise their hand if they were born in January and write that count on the board. Repeat for February

through December. Convert all count data into proportions and write those proportions on the board. Students will use these data as described below.

Probability – group work (10 minutes)
After the introduction, briefly review what the students will be doing in their groups – basically cover what the handout says. Tell them that they will have about 10 minutes to complete their questions and solve another group's questions.

Once they get started, circulate and answer questions. Check that students are arriving at the correct answers. When the first two groups finish writing their problems, have them exchange problem sets and solve these new problems. Continue until all groups have swapped question sets. If an odd number of groups, shuffle among 3 groups. During this problem solving, continue to circulate and answer questions. As groups finish, they can compare their answers with the authoring group to determine if they are right or wrong and to discuss any inconsistencies.

When all groups have at least finished answering their questions, (but maybe not finished comparing with the authors depending on how much time is left), announce that you'd like to have them share some examples of the questions they solved. Take 2 or 3 examples, having a student explain each one. The student might write the solution on the board to help other students follow along.

Examples of appropriate problems:

*1. What is the probability of being a female that is born in January? p(female)*p(January)*
2. What is the probability of being born in January, February, or March? p(Jan)+p(Feb)+p(Mar)
3. What is the probability of being a female born in May or a male born in July?
p(female)p(May) + p(male)p(July)

Note: asking for a specific gender or a specific month of birth is more complicated because these are not mutually exclusive events. (We cover this later in our course.) For example, if you ask for the probability of a female OR a person born in January and calculate it with the "normal" sum rule, p(female) + p(January), you have counted females born in January twice. To get the correct probability, you need to subtract the number of females born in Jan:
The probability of females OR students born in January is:
*p(female) + p(January) - [p(female)*p(January)].*

Part 2:

***chi*-squared - Introduction (3 minutes)**
Provide some introductory remarks about the *chi*-squared test, explaining how it is used to evaluate whether data are consistent with a null model. Emphasize that the null model is something that they will need to determine and is different for different problems/tests. Students often have questions about the interpretation of *p*-values, but you can defer these questions, saying that it will be discussed after they have worked some sample problems.

***chi*-squared – Group work (10 minutes)**
After introducing the *chi*-squared test, let the students know they will have about 5-7 minutes to work in groups on the 2 problems from the worksheet. To get them started, you can draw a sample chi-squared table (do not fill it out) on the board showing the columns for observed, expected, and (O-E)^2/E.

Circulate and answer questions while they are working. Check that students are arriving at the correct answers. Make sure they are using the correct df (1 for gender and 3 for quarter of year for births) and interpreting the *chi*-squared table correctly.
When most groups are finished, ask each group to provide one of the answers (*chi*-squared value, *df*, p-value, interpretation for each test) for the two tests, this will be a total of 8 items. After a group provides an answer, ask the rest of the class to show whether they agree or disagree by a show of hands. If they disagree, discuss other answers and why a group may have gotten a question wrong. Emphasize the interpretation of the *p*-value at this point: a *p*-value of 0.05 means that 5% of the time, data produced from the null model will deviate as much or more as the observed data from the expected values. This should go quickly if their answers are correct.

Discussion: (5 minutes) Two discussion questions are provided on the student handouts. Use these to lead discussion, as follows:

1. If the null hypothesis for one or both tests was rejected, discuss possible reasons why. If the null hypothesis was not rejected in any case, discuss in hypothetical terms why it could have been. For example, maybe there are more female than male biology majors, or more males prefer to have Friday off (if your class is meeting on Friday), or maybe more children are born during the summer months, or random chance. For this last possibility, point out that if 20 classes did this exercise and used a *p*-value cutoff of 0.05, one class is expected to have distributions of gender and birth months that would reject the null hypothesis even if it is true. **Emphasize that the *chi*-squared test can be rejected for many reasons and that it does not provide "proof" for any hypothesis. It should only be used to determine whether or not the data is *consistent with* the null model.**

2. How could we have made more accurate null models to test the ideas that students in the class have the expected distribution by gender and month of birth? One possible answer is to determine the percentage of males and females (or birth months) in the entire student population and use these frequencies to predict the percentages of different types of students in the class.

Part 3 (15 minutes):

This section is designed to help students see how the chi-squared test can be applied to analyze genetic data. After reading the first paragraph from this section, ask a student to describe Mendel's law of segregation (sample answer: an equal probability of inheriting either allele from a parent during gamete formation).

Ask them to develop their notation: "Write the genotypes for the yellow and green parents, the F1 hybrids, and the progeny from the F1 self-cross. Be sure to indicate which allele is dominant with your notation." Ask one or more students for the notation they used to write the parental and F1 genotypes and discuss how they chose the symbols. For example, if Yy is used to represent the F1 genotype, discuss which allele is dominant, which is recessive, and why the letter Y was chosen. Ask what color they expect the F1 plant to be (answer: yellow). Give them time to complete the chi-squared test. Ask what the null model was for this test. (answer: Mendel's law of equal segregation)

Ask a student to write their chi-square table on the board. This table should use either yellow and green or equivalent genotypic notation (e.g., Y_ = yellow and yy = green) as the categories. In this case, the three genotypic classes (e.g., YY, Yy, yy) cannot be used because the number of observed yellow plants that are homozygous and heterozygous is unknown. Ask other students for their chi-squared value, degrees of freedom, approximate p-value (based on the table) and whether or not the null hypothesis should be rejected. Discuss what this p-value (answer: >0.9) means: if this experiment were repeated an infinite number of times, more than 90% of these experiments would show a difference between the observed and expected data as large or larger than the difference observed in this experiment.

	Observed	Expected	$(O-E)^2/E$
Yellow	6022	6017.25	0.00375
Green	2001	2005.75	0.01125

Chi-squared = 0.00375 + 0.01125 = 0.015; df = 1; p-value > 0.9 and < 0.975