Student Handout

Basic Probability and *Chi*-Squared Tests

The goal of this activity is to improve your familiarity and confidence with basic probability and *chi*-squared tests. These skills are used extensively for genetic analysis.

In the following exercises, we will be working with the following phenotypes:
Gender (male or female)
Month of birth (Jan - Dec)

Part 1 – Probability

The multiplication "AND" rule:

If you want to know the probability of two independent events BOTH happening, then multiply the individual probabilities together.

Note that gender and month of birth are independent events.

Example: The probability of drawing a heart from a well-shuffled deck is ¼. The probability of drawing a 10 is 1/13. The probability of drawing a 10 AND drawing a heart (i.e., the 10 of hearts) is 1/4 * 1/13 = 1/52.

The addition "OR" rule:

If you want to know the probability of ONE OR THE OTHER of two mutually exclusive events happening, then add the individual probabilities together.

Note that the two traits of interest, gender or month of birth ARE NOT mutually exclusive (e.g., you can be a female born in January), but that the possible phenotypes within each trait (e.g., born in January or born in February) ARE mutually exclusive events.

Example: The probability of drawing a heart from a well-shuffled deck is ¼. The probability of drawing a spade is 1/4. Hearts and spades are mutually exclusive "phenotypes" because it's impossible for a card to be both hearts and spades. The probability of drawing a heart or a spade is ¼ + ¼ = ½.

In your group

Use the class data (from all groups, compiled by your instructor) to calculate the frequency of students in each group (male or female, month of birth) and write it down. For example, p(male) = X or p(January) = Y. You should have a frequency for each gender and each month of the year, for a total of 14 different frequencies.

Write three probability questions about these data as follows:

1. The first question must use ONLY the multiplication (or product) rule and ask about independent events.
2. The second question should use ONLY the addition (or sum) rule and should ask about mutually exclusive events.

3. The third question must use BOTH the multiplication and addition rules and ask only about independent and mutually exclusive events.

On a separate piece of paper, write an answer key for each of the questions. Let the instructor know when you are finished. You will be asked to exchange questions with another group and solve the questions written by the other group.

Part 2 – *Chi*-squared tests

Chi-squared tests are used to evaluate whether data are consistent with a null model. You will use the data collected about gender and birth month phenotypes to evaluate null hypotheses about enrollment in this class.

The *null hypothesis* is defined by the experimenter and can differ from test to test. It usually reflects the simplest or most common assumption(s). For example, the null hypothesis for a series of coin flips is that heads and tails will appear with equal frequency. In genetic analysis, the null hypothesis is often used to predict the number and kinds of offspring expected if certain conditions (for example, Mendelian inheritance of alleles) are true.

A *chi-squared test* is used to determine how likely the observed data are if the null hypothesis is true. For example, in the coin flip example, the null hypothesis predicts that heads will appear 50% of the time and tails will appear 50% of the time. So if a coin is flipped 10 times, we “expect” to see 5 heads and 5 tails. But, what if we observe 6 heads and 4 tails? Is this consistent with the null hypothesis? What if we observe 7 heads and 3 tails, or 8 heads and 2 tails? A chi-squared test allows us to answer these questions.

The chi-squared test statistic is calculated by comparing the observed data (O) to the data expected (E) under the null hypothesis. Briefly, for each group (e.g., heads or tails), we calculate $(O-E)^2/E$ and sum these values for all groups. We then determine the degrees of freedom (*df*) for the test, which is often simply the number of groups minus 1, and use these values with a chi-square table to determine the probability (p-value) of the observed data occurring by chance if the null hypothesis were true.

In your group

Use the class data (from all groups, compiled by your instructor) and *chi*-squared tests to evaluate the following two null hypotheses:

1. an equal number of males and females are enrolled in this class.

2. an equal number of people in this class were born in each quarter of the year.

In each case, you must (1) calculate the *chi*-squared value, (2) determine the proper degrees of freedom for the test, (3) use the table below to determine the approximate p-value, and (4) decide whether or not the class data is consistent with the null hypotheses shown (i.e. should you accept or reject the null hypothesis in each case). The instructor is there to provide help and guidance when you need it.

When you finish with the tests, discuss the following questions:

- Why did (or could) the class data cause you to reject the null hypothesis used in each case?
- How could we have made more accurate null models to test the idea that students in the class have the expected distribution by gender and month of birth?

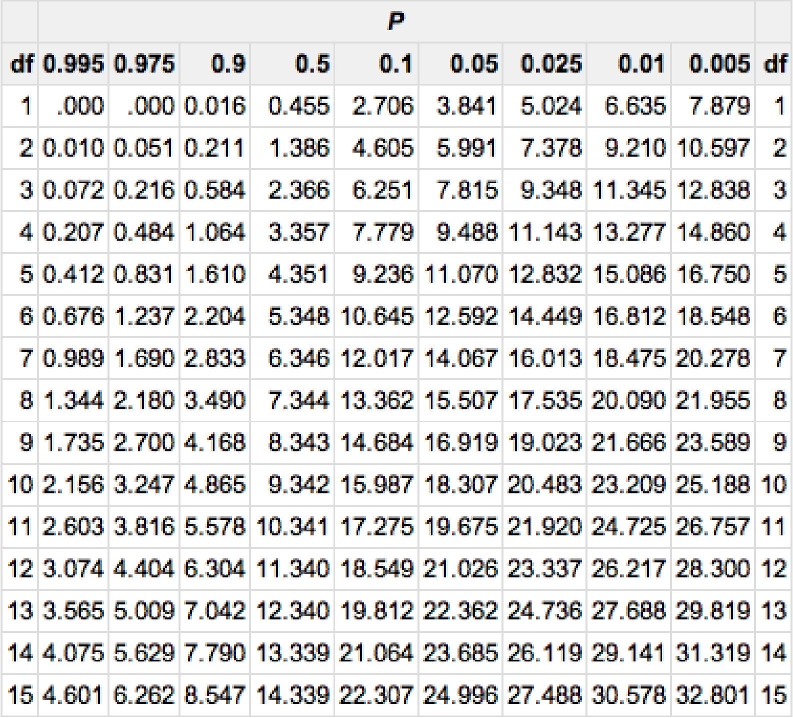
Part 3 – Genetics application

After crossing true-breeding yellow and green pea plants, Mendel allowed the F_1 plants to self. He observed 6022 yellow and 2001 green pea plants resulting from this F_1 self-cross. He used these data to develop his law of segregation.

Write the genotypes for the true-breeding yellow and green plants, the F1 hybrids, and the green and yellow progeny from the F_1 self-cross. Be sure to indicate which allele is dominant with your notation.

Using a chi-squared test, determine if the 6022 yellow and 2001 green pea plants Mendel observed are consistent with his law of equal segregation. Be sure to set up a table of observed and expected data and record the chi-squared value, degrees of freedom, approximate p-value (use the table above), and indicate whether the null hypothesis should be rejected.

***Chi*-square table:**



					P					
df	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15