Name _____ **Period** _____

AP Biology **Date** _____

LAB ____: THE CHI-SQUARE TEST

Probability, Random Chance, and Genetics

Why do we study random chance and probability at the beginning of a unit on genetics?

Genetics is the study of inheritance, but it is also a study of probability. Most eukaryotic organisms are diploid, meaning that each cell contains two copies of every chromosome, so there are two copies of each gene that controls a trait (alleles). In sexual reproduction, these two copies of each chromosome separate, and are randomly sorted into the reproductive cells (gametes). When gametes from two different parents combine in fertilization, new combinations of alleles are created. Thus chance plays a major role in determining which alleles, and therefore which combinations of traits end up in each new individual. The important point is that the inheritance of characteristics is the result of random chance. Therefore, it is important to understand the nature of chance and probability and the resulting implications for the science of genetics. In short, the genes that an individual organism inherits depends on the "luck of the draw," and the luck of the draw is dependent on the laws of probability.

The Laws of Probability

There are three Laws of Probability that are important in genetics and they can be easily demonstrated using simple models like flipping a coin or choosing cards from a deck:

- **The Rule of Independent Events**: Past events have no influence on future events.

  Question: If a coin is tossed 5 times, and each time a head appears, then what is the chance that the next toss will be heads?

  Answer: 1/2 (1 chance in 2), because coins have 2 sides.

- **The Rule of Multiplication**: The chance that two or more independent events will occur together is equal to the product of the probabilities of each individual event.

  Question: What are the chances of drawing a red nine from a standard deck of cards?

  Answer: 1/26 (1 chance in 26), because there is 1/2 chance of drawing a red card and 1 chance in 13 of drawing a nine. Therefore, 1/2 x 1/13 = 1/26 or 1 chance in 26 of drawing a red nine.

- **The Rule of Addition**: The chance of an event occurring when that event can occur two or more different ways is equal to the sum of the probabilities of each individual event

  Question: If 2 coins are tossed, what is the chance that the toss will yield 2 unmatched coins (1 head & 1 tail)?

  Answer: 1/2 (1 chance in 2) because the combination of 2 unmatched coins can come about in 2 ways: Result A (coin #1 heads, coin #2 tails) as well as Result B (coin #1 tails, coin #2 heads). Therefore (1/2 x 1/2) + (1/2 x 1/2) = 1/2, or the chance of Result A plus the chance of Result B.

Paired Coins and Genetics

Using paired coins, in fact, mimics genetics closely. Each coin can serve as the model for a gamete during fertilization, because it's the "luck of the draw" governing which sperm fertilizes which egg.

When you toss two coins, there are three possible outcomes:

- 2 heads
- 2 tails
- 1 head, 1 tail

The probability of each of these outcomes is based on the 3 Laws of Probability we just discussed:

- 2 heads: 1/4 chance

  1/2 heads on coin #1 x 1/2 heads on coin #2 = 1/4,

  which is generalized as $\mathbf{p^2}$ because [$\mathbf{p \times p = p^2}$]
- 2 tails: 1/4 chance

  1/2 tails on coin #1 x 1/2 tails on coin #2 = 1/4,

  which is generalized as $\mathbf{q^2}$ because [$\mathbf{q \times q = q^2}$]
- 1 head, 1 tail: 1/2 chance

  (1/2 heads on coin #1 x 1/2 tails on coin #2) + (1/2 tails on coin #1 x 1/2 heads on coin #2), which is generalized as **2pq** because [(**p x q**) + (**q x p**) **= 2pq**]

Therefore, all the expected results from tossing two coins can be summarized as follows:

$$\mathbf{p^2 + 2pq + q^2 = 1}$$

(double heads) + (heads/tails) + (double tails) = 100%

You will see this formula again when we learn about genetics of populations, so it would be good to become familiar with it now.

Lab Activity

1. Divide the class into 10 teams.

2. Each team of students will toss a pair of coins exactly 100 times and record the results on the data table labeled "Team Data." Each team must check their results to be certain that they have exactly 100 tosses.

3. Record your team results on the chalkboard and then record the summarized results on the data table labeled "Class Data."

4. Analyze both the team data and the class data separately using the Chi-square analysis explained below.

Chi-square Analysis

The Chi-square is a statistical test that makes a comparison between the data collected in an experiment versus the data you expected to find. It is used beyond genetics studies and can be used whenever you want to compare the differences between expected results and experimental data.

Variability is always present in the real world. If you toss a coin 10 times, you will often get a result different than 5 heads and 5 tails. The Chi-square test is a way to evaluate this variability to get an idea if the difference between real and expected results are due to normal random chance, or if there is some other factor involved (like an unbalanced coin).

Genetics uses the Chi-square to evaluate data from experimental crosses to determine if the assumed genetic explanation is supported by the data. In the case of genetics (and coin tosses) the expected results can be calculated using the Laws of Probability (and possibly the help of a Punnett square). The Chi-square test helps you to decide if the difference between your observed results and your expected results is probably due to random chance alone, or if there is some other factor influencing the results.

- Is the variance in your data probably due to random chance alone and therefore your hypothesis about the genetics of a trait is supported by the data?
- Are the differences between the observed and expected results probably not due to random chance alone, and your hypothesis about the genetics of a trait is thereby not supported by the data?
- Should you consider an alternative inheritance mechanism to explain the results?

The Chi-square test will not, in fact, prove or disprove if random chance is the only thing causing observed differences, but it will give an estimate of the likelihood that chance alone is at work.

Determining the Chi-square Value

Chi-square is calculated based on the formula below:

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

A. For your individual team results, complete column A of the Chi-square Analysis Data Table by entering your observed results in the coin toss exercise.

B. For your individual team results, complete column B of the Chi-square Analysis Data Table by entering your expected results in the coin toss exercise.

C. For your individual team results, complete column C of the Chi-square Analysis Data Table by calculating the difference between your observed and expected results.

D. For your individual team results, complete column D of the Chi-square Analysis Data Table by calculating the square of the difference between your observed and expected results. (This is done to force the result to be a positive number.)

E. For your individual team results, complete column E of the Chi-square Analysis Data Table by dividing the square in column D by the expected results.

F. Calculate the X^2 value by summing each of the answers in column E. The ∑ symbol means summation.

G. Repeat these calculations for the full class data and complete the Class Data Chi-square Analysis Table.

H. Enter the "Degrees of Freedom" based on the explanation below.

Interpreting the Chi Square Value

With the Chi-square calculation table completed, you would look up your Chi-square value on the Chi-square Distribution table at the back of this lab. But to know which column and row to use on that chart, you must now determine the degrees of freedom to be used and the acceptable probability that the Chi-square you obtained is caused by chance alone or by other factors. The following two steps will help you to determine the degrees of freedom and the probability.

Degrees of Freedom

Which row do we use in the Chi-square Distribution table?

The rows in the Chi-square Distribution table refer to degrees of freedom. The degrees of freedom are calculated as the one less than the number of possible results in your experiment.

In the double coin toss exercise, you have 3 possible results: two heads, two tails, or one of each. Therefore, there are two degrees of freedom for this experiment.

In a sense degrees of freedom is measuring how many classes of results can "freely" vary their numbers. In other words, if you have an accurate count of how many 2-heads, and 2-tails tosses were observed, then you already know how many of the 100 tosses ended up as mixed head-tails, so the third measurement provides no additional information.

Probability = p

Which column do we use in the Chi-square Distribution table?

The columns in the Chi-square Distribution table with the decimals from .99 through .50 to .01 refer to probability levels of the Chi-square.

For instance, 3 events were observed in our coin toss exercise, so we already calculated we would use 2 degrees of freedom. If we calculate a Chi-square value of 1.386 from the experiment, then when we look this up on the Chi-square Distribution chart, we find that our Chi-square value places us in the "p=.50" column. This means that the variance between our observed results and our expected results would occur from random chance alone about 50% of the time. Therefore, we could conclude that chance alone could cause such a variance often enough that the data still supported our hypothesis, and probably another factor is not influencing our coin toss results.

However, if our calculated Chi-square value, yielded a sum of 5.991 or higher, then when we look this up on the Chi-square Distribution chart, we find that our Chi-square value places us in the "p=.05" column. This means that the variance between our observed results and our expected results would occur from random chance alone only about 5% of the time (only 1 out of every 20 times). Therefore, we would conclude that

chance factors alone are not likely to be the cause of this variance. Some other factor is causing some coin combinations to come up more than would be expected. Maybe our coins are not balanced and are weighted to one side more than another.

So what value of Probability (p) is acceptable in scientific research?

Biologists generally accept p=.05 as the cutoff for accepting or rejecting a hypothesis. If the difference between your observed data and your expected data would occur due to chance alone fewer than 1 time in 20 (p = 0.05, or 5%) then the acceptability of your hypothesis may be questioned. In other words, there's a 95% that the differences between your observed and your expected data are due to some other factor beyond chance. Biologists consider a p value of .05 or less to be a "statistically significant" difference.

A probability of more than 0.05 by no means proves that the hypothesis from which you worked is correct but merely tells you that from a statistical standpoint that it could be correct, and that the variation from your expected results is probably due to random chance alone. Furthermore, a probability of less than 0.05 does not prove that a hypothesis is incorrect; it merely suggests that you have reason to doubt the correctness or completeness of one or more of the assumptions on which your hypothesis is based. At that point, it would be wise as a researcher to explore alternative hypotheses.

Null hypothesis

So how is this directly applied to genetics research?

In classical genetics research where you are trying to determine the inheritance pattern of a phenotype, you establish your predicted genetic explanation and the expected phenotype ratios in the offspring as your hypothesis. For example, you think a mutant trait in fruit flies is a simple dominant inheritance. To test this you would set up a cross between 2 true-breeding flies:

mutant female x wild type male

You would then predict the ratios of phenotypes you would expect from this cross. This then establishes an hypothesis that any difference from these results will not be significant and will be due to random chance alone. This is referred to as your "null hypothesis". It, in essence, says that you propose that nothing else — no other factors — are creating the variation in your results except for random chance differences.

After the cross, you would then compare your observed results against your expected results and complete a Chi-square analysis. If the p value is determined to be greater than .05 then you would accept your null hypothesis (differences are due to random chance alone) and your genetic explanation for this trait is supported. If the p value is determined to be .05 or less then you would reject your null hypothesis — random chance alone can only explain this level of difference fewer than 1 time out of every 20 times — and your genetic explanation for this trait is unsupported. You therefore have to consider alternative factors influencing the inheritance of the mutant trait.

You would repeat this cycle of prediction-hypothesis-analysis for each of your crosses in your genetic research.

Stating conclusions

Once you have collected your data and analyzed them using the Chi-square test, you are ready to determine whether your original hypothesis is supported or not. If the p value in your Chi-square test is .05 or less (.05, .01, etc.) then the data do not support your null hypothesis that

nothing else but random chance is at work here. So, as a scientist, you would state your "acceptable" results from the Chi-square analysis in this way:

> "The differences observed in the data were not statistically significant at the .05 level." You could then add a statement like, "Therefore the data support the hypothesis that..."

And you will see that over and over again in the conclusions of research papers.

This is how a scientist would state "unacceptable" results from the Chi-square analysis:

> "The differences observed in the data were statistically significant at the .05 level." You could then add a statement like, "Therefore the data do not support the hypothesis that..."

N.B.: Do not forget in your writings that the word "data" is plural (datum is singular & rarely used).

> "Data are" is correct.
>
> "Data is" is **not** correct.

TEAM DATA

Toss	H/H	H/T	T/T
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			

Toss	H/H	H/T	T/T
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			
47			
48			
49			
50			
51			
52			
53			
54			
55			
56			
57			
58			
59			
60			
61			
62			
63			
64			
65			
66			
67			
68			

Toss	H/H	H/T	T/T
69			
70			
71			
72			
73			
74			
75			
76			
77			
78			
79			
80			
81			
82			
83			
84			
85			
86			
87			
88			
89			
90			
91			
92			
93			
94			
95			
96			
97			
98			
99			
100			
—	—	—	—
Total*			

* The sum of the total of each column must equal 100 tosses.

3 points

CLASS DATA

	1	2	3	4	5	6	7	8	9	10	Totals Obs	Totals Exp
H/H												
H/T												
T/T												
Total	100	100	100	100	100	100	100	100	100	100	1000	1000

TEAM DATA: CHI SQUARE ANALYSIS

	A	B	C	D	E
	Obs	Exp	Obs - Exp	$(Obs - Exp)^2$	$\frac{(Obs - Exp)^2}{Exp}$
H/H					
H/T					
T/T					
				X^2 Total	
				Degrees of Freedom	

CLASS DATA: CHI SQUARE ANALYSIS

	A	B	C	D	E
	Obs	Exp	Obs - Exp	$(Obs - Exp)^2$	$\frac{(Obs - Exp)^2}{Exp}$
H/H					
H/T					
T/T					
				X^2 Total	
				Degrees of Freedom	

3 points

CHI-SQUARE DISTRIBUTION TABLE

Degrees of freedom	Probability (p) value									
	← ACCEPT NULL HYPOTHESIS								REJECT →	
	0.99	0.95	0.80	0.70	0.50	0.30	0.20	0.10	**0.05**	0.01
1	0.001	0.004	0.06	0.15	0.46	1.07	1.64	2.71	**3.84**	6.64
2	0.02	0.10	0.45	0.71	1.30	2.41	3.22	4.60	**5.99**	9.21
3	0.12	0.35	1.00	1.42	2.37	3.67	4.64	6.25	**7.82**	11.34
4	0.30	0.71	1.65	2.20	3.36	4.88	5.99	7.78	**9.49**	13.28
5	0.55	1.14	2.34	3.00	4.35	6.06	7.29	9.24	**11.07**	15.09
6	0.87	1.64	3.07	3.38	5.35	7.23	8.56	10.65	**12.59**	16.81
7	1.24	2.17	3.84	4.67	6.35	8.38	9.80	12.02	**14.07**	18.48

In scientific research, the probability value of 0.05 is taken as the common cut off level of significance. A probability value (p-value) of .05 means that there is a 5% chance that the difference between the observed and the expected data is a random difference, and a 95% chance that the difference is real and repeatable — in other words, a significant difference. Therefore, if your p-value is greater than .05, you would accept the null hypothesis: "The difference between my observed results and my expected results are due to random chance alone and are not significant."

In genetics experiments (like your upcoming Fly Lab), accepting the null hypothesis would mean that your data are supporting your proposal for the genetics and inheritance scheme for the flies that you were breeding.

In medical research, the chi-square test is used in a similar — but interestingly different — way. When a scientist is testing a new drug, the experiment is set up so that the control group receives a placebo and the experimental group receives the new drug. Analysis of the data is trying to see if there is a difference between the two groups. The expected values would be that the same number of people get better in the two groups — which would mean that the drug has no effect. If the chi-square test yields a p-value greater than .05, then the scientist would accept the null hypothesis which would mean the drug has no significant effect. The differences between the expected and the observed data could be due to random chance alone. If the chi-square test yields a p-value ≤ .05, then the scientist would reject the null hypothesis which would mean the drug has a significant effect. The differences between the expected and the observed data could <u>not</u> be due to random chance alone and can be assumed to have come from the drug treatment.

In fact, chi-square analysis tables can go to much lower p-values than the one above — they could have p-values of .001 (1 in 1000 chance), .0001 (1 in 10,000 chance), and so forth. For example, a p-value of .0001 would mean that there would only be a 1 in 10,000 chance that the differences between the expected and the observed data were due to random chance alone, whereas there is a 99.99% chance that the difference is really caused by the treatment. These results would be considered highly significant.

QUESTIONS

1. What is the Chi-square test used for? _____

The Chi square is a statistical test that makes a comparison between the data collected in an experiment versus the data you expected to find.

2. Why is probability important in genetics? _____

In genetics

3. Briefly describe how the Chi-square analysis may be used in genetics.

Genetics uses the Chi square to evaluate data from experimental crosses to determine if the assumed genetic explanation is supported by the data.

4. Suppose you were to obtain a Chi-square value of 7.82 or greater in your data analysis (with 2 degrees of freedom). What would this indicate?

5. Suppose you were to obtain a Chi-square value of 4.60 or lower in your data analysis (with 2 degrees of freedom). What would this indicate?

6. A heterozygous white-fruited squash plant is crossed with a yellow-fruited plant, yielding 200 seeds. Of these, 110 produce white-fruited plants while only 90 produce yellow-fruited plants. Are these results statistically significant? Explain using Chi-square analysis.

7. What if there were 2000 seeds and 1100 produced white-fruited plants & 900 yellow-fruited? Are these results statistically significant? Explain using Chi-square analysis.

_____

_____

_____

8. **TEAM DATA:** What was your hypothesis (expected values) for your individual team coin toss?

_____

_____

What was your calculated Chi-square value for your individual team data? _____

What p value does this Chi-square correspond to? _____

Was your hypothesis supported by your results? Explain using your Chi-square analysis.

_____

_____

_____

_____

9. **CLASS DATA:** What was your hypothesis (expected values) for the class coin toss?

_____

_____

What was your calculated Chi-square value for the class data? _____

What p value does this Chi-square correspond to? _____

Was your hypothesis supported by the results? Explain using your Chi-square analysis.

_____

_____

10. When scientists design research studies they purposely choose large sample sizes. Work through these scenarios to see why:

    a. Just as in your experiment, you flipped 2 coins, but you only did it 10 times. You collected these data below. Use the chart to calculate the Chi-square value:

	Obs	Exp	Obs - Exp	$(Obs - Exp)^2$	$\frac{(Obs - Exp)^2}{Exp}$
H/H	1				
H/T	8				
T/T	1				
				X^2 Total	

    Would you accept or reject the null hypothesis? Explain using your Chi-square analysis.

    _____

    _____

    _____

    _____

    b. Now you flipped your 2 coins again, but you did it 100 times. You collected these data below. Use the chart to calculate the Chi-square value:

	Obs	Exp	Obs - Exp	$(Obs - Exp)^2$	$\frac{(Obs - Exp)^2}{Exp}$
H/H	10				
H/T	80				
T/T	10				
				X^2 Total	

    Would you accept or reject the null hypothesis? Explain using your Chi-square analysis.

    _____

    _____

    _____

    _____

c. Now you flipped your 2 coins again, but you did it 1000 times. You collected these data below. Use the chart to calculate the Chi-square value:

	Obs	Exp	Obs - Exp	$(Obs – Exp)^2$	$\frac{(Obs – Exp)^2}{Exp}$
H/H	100				
H/T	800				
T/T	100				
				X^2 Total	

Would you accept or reject the null hypothesis? Explain using your Chi-square analysis.

_____

_____

_____

_____

d. Now, using your understanding of the Chi-square test, explain why scientists purposely choose large sample sizes when they design research studies.

_____

_____

_____

_____

_____

_____

_____

_____