VERIFICATION OF MENDELIAN RATIO USING CHI-SQUARE (χ^2) TEST.
(PAPER: ZOOG-CC4-4-P)

==========

INTRODUCTION OF THE TOPIC:

As a student of Life Science you are free to do any sort of experiment. Suppose with this urge you have performed an experiment and got some numerical value. Now how can you conclude that the way you have performed the experiment is right and the result you got is correct as you have expected. Let me clear you again with an example.

Suppose like the simple figure given below you are performing a **genetic cross** in which you know the **genotypes of the parents**. In this situation, you might **hypothesize** that the cross will result in a **certain ratio of phenotypes** in the offspring.

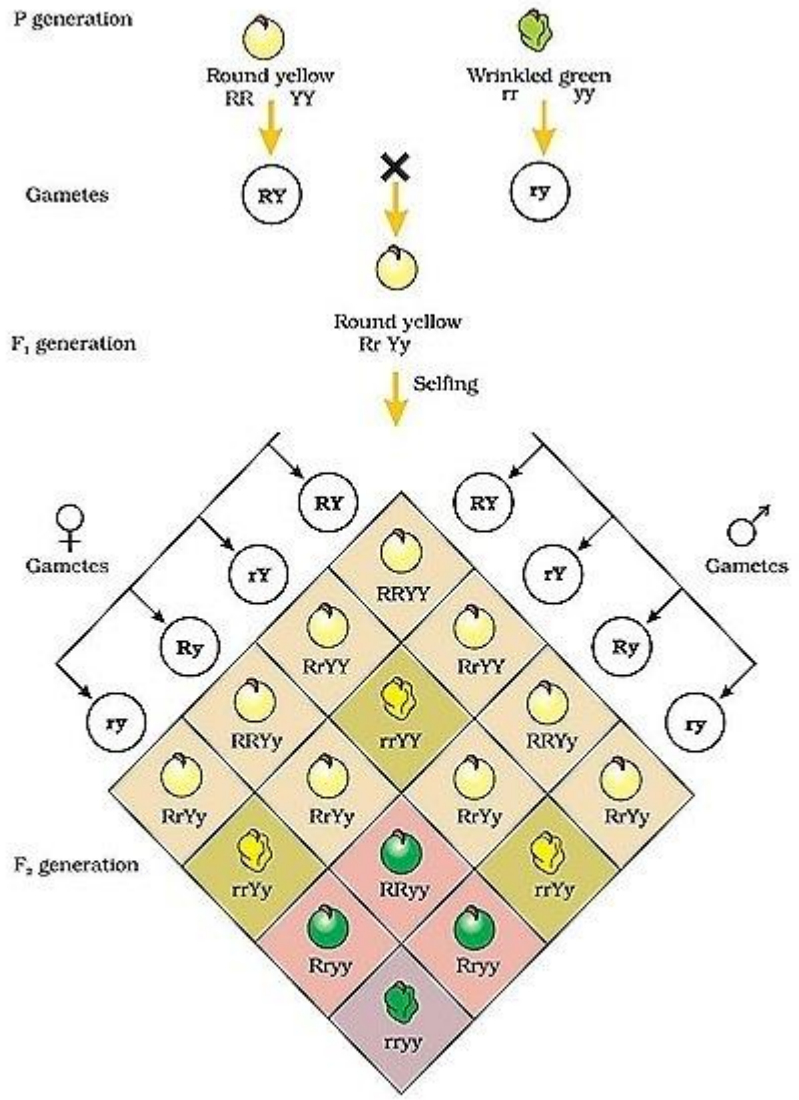

Figure 5.7 Results of a dihybrid cross where the two parents differed in two pairs of contrasting traits: seed colour and seed shape

Thus, you know the **genotypes of the parents** (i.e. both the parents are **RrYy genotype** in the **F_2 generation** and involved in a self pollination) and you have hypothesized that the cross will generate **9:3:3:1 phenotypic ratio** in the offspring.

Now if you got the **same** or **almost same** result as you expected or hypothesized then ok, but what will happen if your observed results do not exactly match your expectations or hypothesis? How can you tell whether this deviation was due to chance? The key to answering these questions is the use of **statistics**, which allows you to determine whether your data are consistent with your hypothesis.

In real life, the results obtained in samples do not always fit exactly with the theoretical results expected according to the rules of probability.

For example, in dihybrid cross of Mendelian genetics, it is expected that the peas of 4 shapes round and yellow; round and green; wrinkled and yellow; wrinkled and green will appear in the proportion 9:3:3:1 respectively. But in reality we will never get the exact ratio. This is also happen in case of a monohybrid 3:1 ratio.

So, it is very much important to see whether the experiment supports the exact ratio or not. This evokes the concept of observed and theoretical frequencies.

The most important and popular method is the **chi-square (χ^2) test** [where the **χ is the Greek letter chi**] that determines whether the results obtained in samples supports exactly with the theoretical frequency or not.

FORMING AND TESTING A HYPOTHESIS

Before performing an experiment the primary work of a **researcher / scientist / experimenter (it may be you)** is to form a hypothesis about the experiment outcome. This often takes the form of a **null hypothesis (Ho)**, which is a statistical hypothesis that states there will be no difference between observed and expected data. The null hypothesis is proposed by a scientist before completing an experiment, and **it can be either supported by data** or **disproved** in favor of an **alternate hypothesis (H_1)**.

Pearson's Chi-Square Test for Goodness-of-Fit

One of **Karl Pearson's** most significant achievements occurred in **1900**, when he developed a statistical test called Pearson's chi-square (X^2) test, also known as the **chi-square test for goodness-of-fit** (Pearson, 1900). Pearson's chi-square test is used to examine the role of chance in producing deviations between observed and expected values. The test depends on an **extrinsic hypothesis**, because **it requires theoretical expected values to be calculated**. The test indicates the probability that chance alone produced the deviation between the expected and the observed values (Pierce, 2005). When the probability calculated from Pearson's chi-square test is high (i.e. Critical **chi-square value is high than the calculated chi-square value)**, it is assumed that **chance alone produced the difference**. Conversely, when the probability is low (i.e. Critical **chi-square value is low than the calculated chi-square value)**, it is assumed that a **significant factor other than chance** produced the deviation.

In **1912, J. Arthur Harris** applied Pearson's chi-square test to examine **Mendelian ratios** (Harris, 1912). It is important to note that when Gregor Mendel studied inheritance, he did not use statistics, and neither did Bateson, Saunders, Punnett, and Morgan during their experiments that discovered genetic linkage. Thus, until Pearson's statistical tests were applied to biological data, scientists judged the goodness of fit between theoretical and observed experimental results simply by inspecting the data and drawing conclusions (Harris, 1912). Although this method can

work perfectly if one's data exactly matches one's predictions, scientific experiments often have variability associated with them, and this makes statistical tests very useful.

Definition of chi-square (χ^2) test:

It may be defined as a statistical comparison of observed ratios with the theoretical ratios.

Simply, when sample subjects are distributed among discrete categories (e.g. tall and dwarf plants), the Chi-square distribution is frequently used. This statistical hypothesis test was invented by Karl Pearson in 1900.

Few important definition:

Testing of hypothesis:

Determine whether to support or reject a hypothesis by comparing the data to the predictions of the hypothesis.

Null hypothesis:

To test the hypotheses it is required to make a concise statement about the population mean (μ). This statement is called a null hypothesis is denoted by Ho, because it expresses the concept of "no difference ". Unless data provides convincing evidence that it is false, Ho is accepted.

Alternative hypothesis:

Alternative hypothesis is a statement that contradicts with the null hypothesis (Ho) and is denoted by Ha or some times H_1. If it is concluded that a null hypothesis is false, then an alternative hypothesis is assumed to be true.

Level of significance: the maximum probability with which a null hypothesis is rejected is called the level of significance of the statistical test. Generally level of significance is considered at **1% (0.01 level)** or **5% level** (**0.05 level)** or any other level depending upon the consequences of statistical decision.

Degrees of freedom (df): The number of degrees of freedom denotes the number of comparisons that can be made between any one observation and the rest of the observations, taking them in pairs. This is the values of a sample which are freely variable without affecting the mean

If there are 30 observations (n), the number of degrees of freedom will be (30−1) = 29, since only one single observation can be compared with each of the remaining 29 observations taking one at a time.

==========

GOODNESS OF FIT:

This common type of Chi square test is also known as test for goodness of fit" because it is used to compare an "observed" ratio with an "expected" ratio", and to determine how closely the former (i.e. observed) fits the latter (i.e. expected).

In genetics and breeding, tests for goodness of fit are widely used to compare an observed Mendelian ratio with a theoretical (expected) ratio.

This is determined by the following formula:

$$\chi^2 = \Sigma \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$(\chi^2) = \frac{[(o_1 - e_1)^2]}{e_1} + \frac{[(o_2 - e_2)^2]}{e_2} + \frac{[(o_3 - e_3)^2]}{e_3} + \ldots\ldots\ldots\ldots + \frac{[(o_n - e_n)^2]}{e_n}$$

Or simply,

$$(\chi^2) = \Sigma \frac{(O - E)^2}{E}$$

$$(\chi^2) = \frac{\Sigma [(O - E) - 0.5]^2}{E}$$

Where, O = Observed frequency

E = expected or theoretical frequency,

Σ = is the symbol denoting summation.

(O – E) is the deviation between each observed and expected class value.

ABOUT YATES CORRECTION

Pearson's chi-square test works well with genetic data as long as there are **enough expected values** in each group. In the case of small samples (less than 10 in any category) that have 1 degree of freedom, the test is not reliable. However, in such cases, the test can be corrected by using the Yates correction for continuity, which reduces the absolute value of each difference between observed and expected frequencies by 0.5 before squaring. Additionally, it is important to remember that the chi-square test can only be applied to numbers of progeny, not to proportions or percentages. The **Yates correction** is **usually recommended**, especially if the **expected cell frequencies** are below 10 (some authors put that figure at 5). All you really need to know is that if your expected cell frequencies are below 10, you *probably* should be using the Yates correction. Although some people recommend that you should use the correction only if your expected cell frequency is below 10 or even 5, others recommend that you **don't use it at all**.

0.5 or ½ = **Yates correction** [reduction of 0.5 from absolute difference between observed and expected frequencies, and is generally applicable in monohybrid crosses or similar problems when df = 1.

The Yates correction formula:

$$\chi^2_{Yates} = \sum^{k} \frac{(|f_0 - f_e| - 0.5)^2}{f_e}$$

Now if (χ^2) = 0, the **observed** and **expected frequencies** agree exactly. If the deviations of expected from observed events are small, (χ^2) approaches 0 and the fit is good. Whereas if (χ^2) > 0, that is the deviations are large, then χ^2 increases and hence, the fit is poor and in that case they will not agree exactly. The larger the value of (χ^2), the greater is the discrepancy between the observed and theoretical frequencies.

==========

STEPS FOR TESTING THE HYPOTHESIS AND CALCULATING CHI-SQUARE:

1. State the null hypothesis (Ho) and an alternative hypothesis (H_1).
2. Determine the expected numbers for each observational class. Remember to use numbers, not percentages.
3. Calculate (χ^2) using the formula. Complete all calculations to three significant digits. Round off your answer to two significant digits.
4. Use the chi-square distribution table to determine significance of the value.
5. State your conclusion in terms of your hypothesis.

If the calculated value of (χ^2) is < than the critical value of (χ^2) at a particular degrees of freedom then it denotes that the probability is greater than the level of significance specified (e.g. 0.05 or 0.01) and thus the deviation or difference between the observe and expected frequencies is not significant and therefore the null hypothesis (H_0) will not be rejected.

If the calculated value of (χ^2) is > than the critical value of (χ^2) at a particular degrees of freedom then it denotes that the probability is less than the level of significance specified (e.g. 0.05 or 0.01) and thus the deviation or difference between the observe and expected frequencies is significant and therefore the null hypothesis (H_0) will be rejected and alternative hypothesis (H_1) will be accepted.

PROBLEM 1

Expecting a Mendelian monohybrid cross ratio of 3:1, a geneticist crossed pure bred tall and dwarf pea plants, and out of 100 progeny he obtained 84 tall and 16 dwarf plants in F2 generation. Construct the null (Ho) and an alternative hypothesis (H_1) and use Chi square test for goodness of fit to conclude whether the geneticist can conclude as he expected or not. [$\chi^2_{0.05, (1)} = 3.841$]

SOLUTION:

Null hypothesis (Ho): 3:1

Alternative hypothesis (H_1): 1:1

You can construct the table like this:

	Events		
	Tall pea plants	Dwarf pea plants	Total
Observed number (O)	84	16	100
Expected ratio	3/4	1/4	
Expected number (E)	75	25	100
(O – E)	+9	−9	0
$(O – E)^2$	81	81	
$(O – E)^2 / E$	81/75 = 1.08	81/25 = 3.24	

OR like this:

		Observed number (O)	Expected ratio	Expected number (E)	(O – E)	$(O – E)^2$	$(O – E)^2 / E$
Events	Tall pea plants	84	3/4	75	+9	81	81/75 = 1.08
	Dwarf pea plants	16	1/4	25	−9	81	81/25 = 3.24
	Total	100		100	0		

According to the formula of chi-square test for goodness of fit for Menedian monohybrid cross:

$(\chi^2) = \Sigma (O - E)^2 / E = (1.08 + 3.24) = 4.32$

Here the degrees of freedom = [2 − 1] = 1

At 0.05 or 5% level of significance the critical value of (χ^2) is $\chi^2_{0.05, (1)} = 3.841$

Note:

Degrees of freedom represent the number of ways in which the observed outcome categories are free to vary. For Pearson's chi-square test, the degrees of freedom are equal to *n* - 1, where *n* represents the number of different expected phenotypes. In problem 1 there are two expected outcome phenotypes (tall and dwarf), so *n* = 2 categories, and the degrees of freedom equal 2 - 1 = 1. Thus, the calculated chi-square value (4.32) and the associated degrees of freedom (1), we can determine the probability by using a chi-square table (Table in the next page).

Inference:

Since the calculated (χ^2) value is 4.32 which is greater than the critical value of (χ^2) i.e. $\chi^2_{0.05, (1)} = 3.841$, therefore, difference between the observed and expected frequencies are significant. So, the null hypothesis is (Ho) is rejected.

Finally, we can conclude that the data has not good fit to the Mendelian monohybrid cross ratio of 3:1.

HOW TO READ THE (χ^2) TABLE

Degrees of Freedom (df)	Probability (P)									
	0.995 (99.5%)	0.99 (99%)	0.975 (97.5%)	0.95 (95%)	0.90 (90%)	0.10 (10%)	0.05 (5%)	0.025 (2.5%)	0.01 (1%)	0.005 (0.5%)
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

(χ^2) table is organized with **degrees of freedom (df) in the left column** and **probabilities (P) at the top**. The (χ^2) values associated with the probabilities are in the center of the table. To determine the probability, first locate the row for the degrees of freedom for your experiment, then determine where the calculated chi-square value would be placed among the theoretical values in the corresponding row.

In problem 1 the question was asked in such a way that if the probability is less than 0.05, then null hypothesis will be rejected as the deviation would be significant and not due to chance. Now, looking at the row that corresponds to 1 degree of freedom, we can see that the calculated chi-square value of 4.32 falls between 3.841, which is associated with a probability of 0.05, and 5.024, which is associated with a probability of 0.025. Therefore, there is between a 2.5% and 5% probability that the deviation observed between expected and the observed numbers of tall and short plants is due to chance. In other words, the probability associated with the chi-square value is much less than the critical value of 0.05. This means that we will reject our null hypothesis, and the deviation between the observed and expected results is significant.

LEVEL OF SIGNIFICANCE

Determining whether to accept or reject a hypothesis is decided by the experimenter, who is the person who chooses the "level of significance" or confidence. Scientists commonly use the 0.05, 0.01, or 0.001 probability levels as cut-off values. For instance, in problem 1 used the 0.05 probability. Thus, $P \geq 0.05$ can be interpreted to mean that chance likely caused the deviation between the observed and the expected values (i.e. there is a greater than 5% probability that chance explains the data). If instead we had observed that $P \leq 0.05$, this would mean that there is less than a 5% probability that our data can be explained by chance. There is a significant difference between our expected and observed results, so the deviation must be caused by something other than chance.

PROBLEM 2 BELOW

PROBLEM 2

Expecting a Mendelian monohybrid cross ratio of 3:1, a geneticist crossed pure bred tall and dwarf pea plants, and out of 100 progeny he obtained 305 tall and 95 dwarf plants in F_2 generation. Construct the null (Ho) and an alternative hypothesis (H_1) and use Chi square test for goodness of fit at 0.01 significance level to conclude whether the geneticist can conclude as he expected or not.

SOLUTION:

Null hypothesis (Ho): 3:1

Alternative hypothesis (H_1): 1:1

You can construct the table like this:

	Events		
	Tall pea plants	Dwarf pea plants	Total
Observed number (O)	305	95	400
Expected ratio	3/4	1/4	
Expected number (E)	300	100	400
(O – E)	+5	−5	0
$(O – E)^2$	25	25	
$(O – E)^2 / E$	25/300 = 0.08	25/100 = 0.25	

According to the formula of chi-square test for goodness of fit for Menedian monohybrid cross:

$(\chi^2) = \Sigma (O – E)^2 / E = (0.08 + 0.25) = 0.33$

Here the degrees of freedom = [2 − 1] = 1

At 0.05 or 5% level of significance the critical value of (χ^2) is $\chi^2_{0.01, (1)} = 6.635$

Inference:

In problem 2 the question was asked in such a way that if the probability is less than 0.01, then null hypothesis will be rejected as the deviation would be significant and not due to chance. Now, looking at the row that corresponds to 1 degree of freedom, we can see that the calculated chi-square value of 0.33 falls between 0.016, which is associated with a probability of 0.9, and 2.706, which is associated with a probability of 0.10. Therefore, there is between a 10% and 90% probability that the deviation observed between expected and the observed numbers of tall and short plants is due to chance. In other words, the probability associated with the chi-square value is much greater than the

critical value of 0.01. This means that we will not reject our null hypothesis, and the deviation between the observed and expected results is not significant.

Since the calculated (χ^2) value is 0.33 which is less than the critical value of (χ^2) i.e. $\chi^2_{0.01,(1)}$ = 6.635, therefore, difference between the observed and expected frequencies are not significant. So, the null hypothesis is (Ho) is accepted.

Finally, we can conclude that the data has good fit to the Mendelian monohybrid cross ratio of 3:1.