Course: Fundamentals of Genetics

Class: - Ist Year, IInd Semester

Lecture No. VI

Title of topic: - Probability and Chi-square

Prepared by- Vinod Kumar, Assistant Professor, (PB & G)
College of Agriculture, Powarkheda

Probability:-

Probability expresses the likelihood of the occurrence of a particular event. It is the number of times that a particular event occurs, divided by the number of all possible outcomes.

Example-

I. A deck of 52 cards contains only one king of hearts. The probability of drawing one card from the deck at random and obtaining the king of hearts is 1/52, because there is only one card that is the king of hearts (one event) and there are 52 cards that can be drawn from the deck (52 possible outcomes). The probability of drawing a card and obtaining an ace is 4/52, because there are four cards that are aces (four events) and 52 cards (possible outcomes). Probability can be expressed either as a fraction (4/52 in this case) or as a decimal number (0.077 in this case).

In other cases, we determine the probability of an event by making a large number of observations.

II. When a weather forecaster says that there is a 40% chance of rain on a particular day, this probability was obtained by observing a large number of days with similar atmospheric conditions and finding that it rains on 40% of those days. In this case, the probability has been determined empirically (by observation).

Prediction of probability :- Two rules of probability are useful for predicting the ratios of offspring produced in genetic crosses.

1. The first is the **multiplication rule**, which states that the probability of two or more independent events occurring together is calculated by multiplying their independent probabilities. To illustrate the use of the multiplication rule, let's again consider the roll of a die. The probability of rolling one die and obtaining a four is 1/6. To calculate the probability of rolling a die twice and obtaining 2 fours, we can apply the multiplication rule. The probability of obtaining a four on the first roll is 1/6 and the probability of obtaining a four on the second roll is 1/6; so the probability of rolling a four on both is 1/6 _ 1/6 _ 1/36

The key indicator for applying the multiplication rule is the word *and*; in the example just considered, we wanted to know the probability of obtaining a four on the first roll *and* a four on the second roll. For the multiplication rule to be valid the events whose joint probability is being calculated must be independent— the outcome of one event must not influence the outcome of the other. For example, the number that comes up on one roll of the die has no influence on the number that comes up on the other roll; so these events are independent.

However, if we wanted to know the probability of being hit on the head with a hammer and going to the hospital on the same day, we could not simply multiply the probability of being hit on the head with a hammer by the probability of going to the hospital. The multiplication rule cannot be applied here, because the two events are not independent—being hit on the head with a hammer certainly influences the probability of going to the hospital.

2. **The addition rule** The second rule of probability frequently used in genetics is the **addition rule**, which states that the probability of any one of two or more mutually exclusive events is calculated by adding the probabilities of these events. Let's look at this rule in concrete terms.

To obtain the probability of throwing a die once and rolling *either* a three *or* a four, we would use the addition rule, adding the probability of obtaining a three (1/6) to the probability of obtaining a four (again, 1/6), or 1/6 _ 1/6 _ 2/6 _ 1/3.

The key indicators for applying the addition rule are the words *either* and *or.* For the addition rule to be valid, the events whose probability is being calculated must be mutually exclusive, meaning that one event excludes the possibility of the occurrence of the other event. For example, you cannot throw a single die just once and obtain both a three and a four, because only one side of the die can be on top. These events are mutually exclusive.

The application of probability to genetic crosses:-

- The multiplication and addition rules of probability can be used in place of the Punnett square to predict the ratios of progeny expected from a genetic cross. Let's first consider a cross between two pea plants heterozygous for the locus that determines height, *Tt_Tt.* Half of the gametes produced by each plant have a *T* allele, and the other half have a *t* allele; so the probability for each type of gamete is 1/2.
- The gametes from the two parents can combine in four different ways to produce offspring. Using the multiplication rule, we can determine the probability of each possible type. To calculate the probability of obtaining *TT* progeny, for example, we multiply the probability of receiving a *T* allele from the first parent (1/2) times the probability of receiving a *T* allele from the second parent (1/2). The multiplication rule should be used here because we need the probability of receiving a *T* allele from the first parent *and* a *T* allele from the second parent—two independent events. The four types of progeny from this cross and their associated probabilities are:

TT (*T* gamete and *T* gamete) 1/2 _ 1/2 _ 1/4 tall
Tt (*T* gamete and *t* gamete) 1/2 _ 1/2 _ 1/4 tall
tT (*t* gamete and *T* gamete) 1/2 _ 1/2 _ 1/4 tall
tt (*t* gamete and *t* gamete) 1/2 _ 1/2 _ 1/4 short

- Notice that there are two ways for heterozygous progeny to be produced: a heterozygote can either receive a *T* allele from the first parent and a *t* allele from the second or receive a *t* allele from the first parent and a *T* allele from the second. After determining the probabilities of obtaining each type of progeny, we can use the addition rule to determine the overall phenotypic ratios. Because of dominance, a tall plant can have genotype *TT*, *Tt*, or *tT*; so, using the addition rule, we find the probability of tall progeny to be 1/4 _ 1/4 _ 1/4 _ 3/4. Because only one genotype encodes short (*tt*), the probability of short progeny is simply 1/4.

THE CHI-SQUARE TEST

With DeVries's data, and with other genetic data as well, we need an objective procedure to compare the results of the experiment with the predictions of the underlying hypothesis. This procedure has to take into account how chance might affect the outcome of the experiment. Even if the hypothesis is correct, we do not anticipate that the results of the experiment will exactly match the predictions of the hypothesis. If they deviate a bit, as Mendel's data did, we would ascribe the deviations to chance variation in the outcome of the experiment. However, if they deviate grossly, we would suspect that something was amiss. The experiment might have been executed poorly—for example, the crosses might have been improperly carried out, or the data might have been incorrectly recorded—or, perhaps, the hypothesis is simply wrong. The possible discrepancies between observations and expectations obviously lie on a continuum from small to large, and we must decide how large they need to be for us to entertain doubts about the execution of the experiment or the acceptability of the hypothesis.

One procedure for assessing these discrepancies uses a statistic called **chi-square** (χ^2). A *statistic* is a number calculated from data—for example, the mean of a set of examination scores. The χ^2 statistic allows a researcher to compare data, such as the numbers we get from a breeding experiment, with their predicted values. If the data are not in line with the predicted values, the χ^2 statistic will exceed a critical number and we will decide either to revaluate the experiment—that is, look for a mistake in technique—or reject the underlying hypothesis. If the χ^2 statistic is below this number, we tentatively conclude that the results of the experiment are consistent with the predictions of the hypothesis. The χ^2 statistic therefore reduces hypothesis testing to a simple, objective procedure.

As an example, let's consider the data from the experiments of Mendel and DeVries. Mendel's F_2 data seemed to be consistent with the underlying hypothesis, whereas DeVries's F_2 data showed some troubling discrepancies.

Table-1

F_2 Phenotype	**Observed Number**	**Expected Number**	**$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$**
Mendel's dihybrid cross			
Yellow, round	315	313	0.01
Green, round	108	104	0.15
Yellow, wrinkled	101	104	0.09
Green, wrinkled	32	35	0.26
Total	**556**	**556**	**0.51 = χ^2**
DeVries's dihybrid cross			
Red, hairy	70	88.9	4.02
White, hairy	23	29.6	1.47
Red, smooth	46	29.6	9.09
White, smooth	19	9.9	8.36
Total	**158**	**158**	**22.94 = χ^2**

Now imagine carrying out the experiment—carefully and correctly—many times, and each time, calculating a χ^2 statistic. Fortunately, the χ^2 frequency distribution is known from statistical theory, so we don't actually need to carry out many replications of the experiment to get it. The critical value is the point that cuts off the upper 5 percent of the distribution. By chance alone, the χ^2 statistic will exceed this value 5 percent of the time. Thus, if we perform an experiment once, compute a χ^2 statistic, and find that the statistic is greater than the critical value, we have either observed a rather unlikely set of results—something that happens less than 5 percent of the time—or there is a problem with the way the experiment was executed or with the appropriateness of the hypothesis. Assuming that the experiment was done properly, we are inclined to reject the hypothesis. Of course we must realize that with this procedure we will reject a true hypothesis 5 percent of the time. Thus, as long as we know the critical value, the χ^2 testing procedure leads us to a decision about the fate of the hypothesis. However, this critical value—and the shape of the associated frequency distribution—depends on the number of phenotypic classes in the experiment. Statisticians have tabulated critical values according to the **degrees of freedom** associated with the χ^2 statistic (**Table 2**). This index to the set of χ^2 distributions is determined by subtracting one from the number of phenotypic classes.

Table 2: Table of Chi-Square (χ^2) 5% Critical Values[2]

Degrees of Freedom	5% Critical Value
1	3.841
2	5.991
3	7.815
4	9.488
5	11.070
6	12.592
7	14.067
8	15.507
9	16.919
10	18.307
15	24.996
20	31.410
25	37.652
30	43.773

Selected entries from R. A. Fisher and Yates, 1943, *Statistical Tables for Biological, Agricultural, and Medical Research*. Oliver and Boyd, London

In each of our examples, there are 4 -1= 3 degrees of freedom. The critical value for the χ^2 distribution with 3 degrees of freedom is 7.815. For Mendel's data, the calculated χ^2 statistic is 0.51, much less than the critical value and therefore no threat to the hypothesis being tested. However, for DeVries's data the calculated χ^2 statistic is 22.94, very much greater than the critical value. Thus, the observed data do not fit with the genetic hypothesis. Ironically, when DeVries presented these data in 1905, he judged them to be consistent with the genetic hypothesis. Unfortunately, he did not perform a χ^2 test. DeVries also argued that his data provided further evidence for the correctness and widespread applicability of Mendel's ideas—not the only time that a scientist has come to the right conclusion for the wrong reason. To solidify your understanding of the χ^2 procedure, answer the question posed in Solve It: Using the Chi-Square Test.

KEY POINTS

- The chi-square statistic is $\chi^2=\sum$ (observed number - expected number)^2/ expected number, with the sum computed over all categories comprising the data.
- Each chi-square statistic is associated with an index, the degrees of freedom, which is equal to the number of data categories minus one.