Lecture 4

From the last lecture, we followed gene segregation in a cross of a true breeding *shibire* fly with a wild type fly.

<div align="center">

Shibire x wild type

↓

F$_1$: all not paralyzed

↓

F$_2$:   3 not paralyzed : 1 paralyzed

</div>
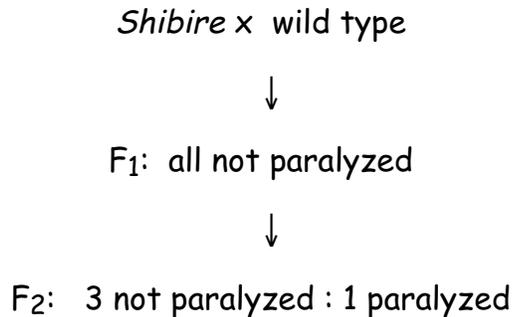
This is the segregation pattern expected for a single gene. But in an actual experiment how do we know that the phenotypic ratio is really 3 : 1 ?

There is no logical way to prove that we have a 3 :1 ratio. Nevertheless, we can think of an alternative hypothesis then show that the alternative hypothesis does not fit the data. Usually, we then adopt the simplest hypothesis that still fits the data.

A possible alternative hypothesis is that recessive mutations in two different genes are needed to get a paralyzed fly.

In this case a true breeding paralyzed fly would have genotype: a/a , b/b

Whereas wild type would have genotype: A/A , B/B

F$_1$: A/a B/b not paralyzed

F$_2$: $p(a/a$ and $b/b) = (1/4)^2 = 1/16$

$p(a/a$ and $B/-) = 1/4 \times 3/4 = 3/16$

$p(b/b$ and $A/-) = 3/16$

$p(A/-$ and $B/-) =$ the rest $= 9/16$

This is the classic ratio for two gene segregation   9 : 3 : 3 : 1
⌐─┘
paralyzed

For our hypothesis we should see a phenotypic ratio of 15 not paralyzed : 1 paralyzed.

Therefore, to distinguish one-gene segregation from two-gene segregation we need a statistical test to distinguish 3 : 1 from 15 : 1. Intuitively, we know that in order to get statistical significance, we need to look at a sufficient number of individuals.

For a **chi-square test** you start with a specific hypothesis that gives a precise expectation. The test is then applied to the actual experimental results and will give the probability of obtaining the results under the hypothesis. The test is useful for ruling out hypotheses that would be very unlikely to give the actual results.

Say we look at 16 flies in the F_2 and observe 14 not paralyzed and 2 paralyzed flies.

Under the hypothesis of two genes we expect 15 not paralyzed flies and 1 paralyzed fly.

We calculate the value χ^2 using the formula below. Where O is the number of individuals observed in each class and E is the number of individuals expected for each class.

$$\chi^2 = \sum_{\text{(all classes)}} \frac{(O-E)^2}{E} = \frac{1^2}{15} + \frac{1^2}{1} = 0.067 + 1 = 1.067$$

degrees of freedom (df) = number of classes – 1

From the table using 1 df, $0.05 < p < 0.5$

The convention we use is that $p \leq 0.05$ constitutes a deviation from expectation that is significant enough to reject the hypothesis. Therefore, on the basis of this sample of 16 flies we can't rule out the hypothesis that two genes are required.

Say we look at 64 F_2 flies and find that 12 are paralyzed. For the hypothesis of two genes the expectation is that 4 would be paralyzed. The χ^2 for this data:

$$\chi^2 = \frac{8^2}{60} + \frac{8^2}{4} = 1.07 + 16 = 17.1$$

From the table $p < 0.005$ so we reject the two-gene hypothesis.

Let's use this data to test the hypothesis of one gene segregation which would be expected to give 16 paralyzed flies from 64 F_2 flies,

$$\chi^2 = \frac{4^2}{48} + \frac{4^2}{16} = 0.33 + 1 = 1.33$$

From the table using 1 df, $0.5 < p < 0.5$. Thus the data still fits the hypothesis of one-gene segregation.

So far, the hypothesis that one gene is responsible for the paralyzed trait is the simplest explanation that fits the data.

The way to distinguish most easily between a heterozygote and a homozygote expressing a dominant trait is to cross to a homozygous recessive test strain.

Test cross: cross to homozygote recessive:

$A/A \times a/a$ gives all A/a .   i.e. all offspring will express the dominant trait.

$A/a \times a/a$ gives $1/2$ A/a and $1/2$ a/a. i.e. one half of the offspring will express the dominant trait.

Mendelian inheritance in humans

For humans we can't do test crosses, of course, but by following inheritance of a trait for several generations the modes of inheritance can usually be identified by applying basic principles of Mendel. The following are guidelines for identifying different modes of inheritance in pedigrees.

Autosomal dominant

i) Affected individuals must have at least one affected parent

Exceptions to this rule will occur if a new mutation arises in one of the parents (in real life a more likely explanation is extramarital paternity). Another possibility is incomplete penetrance, where other genetic or environmental factors prevent the trait from being expressed in one of the parents.

Autosomal recessive

i) When both parents are carriers, on average $1/4$ of the children will be affected.

ii) When both parents are affected, then all of the children will be affected.

iii) If the trait is very rare then consanguinity is likely. That is, it is likely that parents of affected children are themselves related (e.g. cousins).

X-linked inheritance

$♀$ X^cX^+ $\times$ $♂$ X^+Y
(carrier)

↓

$♀$ X^cX^+, $♀$ X^+X^+, $♂$ X^cY, $♂$ X^+Y
(carrier)   (color blind)
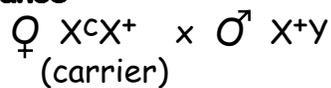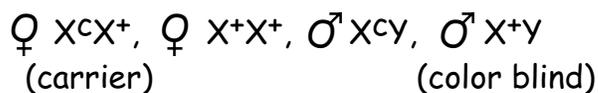
i) When parents are a carrier ♀ and an unaffected ♂, then on average, 1/2 of the daughters will be carriers and 1/2 of the sons will be affected.

If the trait is rare then the vast majority of affected individuals will be male which is the hallmark of X-linked traits.

ii) Affected sons inherit the allele from mother

- Maternal uncles often affected
- Since inherited only from mother, inbreeding doesn't increase the probability of an affected ♂.

Conditional probabilities

Consider the following pedigree of a recessive trait.
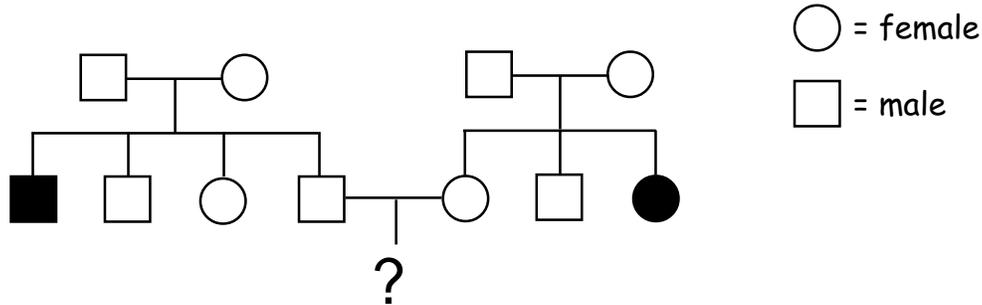
p(affected child) = p(mother carrier and father carrier and affected child)

$$= 2/3 \times 2/3 \times 1/4 = 1/9$$

However, if they have a child that is affected we must reassess the probability that their next child will be affected.

p(both parents carriers) = 1. So, p(next child affected) = $1/4$

This example shows how probability calculations are based on information. The probability changes not because the parents have changed but because our information about them has.

HANDBOOK for PROBABILITY CALCULATIONS

Many problems in diploid genetics rely on basic concepts of probability. This is because each individual inherits at random only one of two possible copies of a gene from each parent. Thus, breeding experiments or inheritance in human pedigrees have probabilistic rather than absolute outcomes. Everyone has an intuitive sense of probability but what we need is a precise definition that will allow probabilities to be manipulated quantitatively.

Probabilities are usually defined in terms of possible outcomes of a trial. A trial could be the toss of a coin, the roll of a die, or two parents having a child. If we define a specific event $\mathbf{a}$, p($\mathbf{a}$) or the probability of $\mathbf{a}$, can be defined as follows: after a very large number of trials, p($\mathbf{a}$) is simply the fraction of trials that give outcome $\mathbf{a}$. In principle, we could determine p($\mathbf{a}$) by actually performing a large number of trials and directly measuring the fraction of trials that produce event $\mathbf{a}$. This is sometimes called the "Monte Carlo method" named after a famous European casino and works well for computer simulations of complicated phenomena. However, in many cases there is a much simpler way to calculate probabilities. To directly calculate classical probabilities one must know enough about a process to break down the possible outcomes of a trial into some number of *equally probable* events. In these cases the probability of event $\mathbf{a}$ is:

$$p(\mathbf{a}) = \frac{n_a}{N}$$

where n_a is the number of outcomes that satisfy the criteria for $\mathbf{a}$ and N is the total number of *equally probable* outcomes. Note that since N includes all possible outcomes, $n_a \leq N$ and $0 \leq p(\mathbf{a}) \leq 1$.

Example: A couple has two children, what is the probability that they are both girls? Assuming that the chances of having a boy or a girl are equal, there are 4 equally probable ways of having two children (boy, boy; girl, boy; boy, girl; girl, girl) and the probability of two girls is 1/4 or 0.25.

For classical probability problems you will always be able to arrive at the correct answer by writing out *all* of the possible outcomes of a trial and counting the fraction of outcomes that satisfy the criteria for a given event. Often, enumerating all of the outcomes for a trial is time-consuming and error-prone. It is usually faster and easier to break a problem down into simple parts and then to combine the probabilities for the individual parts. The following are useful ways that probabilities can be combined to speed probability calculations.

PRODUCT RULE

   p($\mathbf{a}$ and $\mathbf{b}$) = p($\mathbf{a}$) x p($\mathbf{b}$) if $\mathbf{a}$ and $\mathbf{b}$ are independent.

Two events are considered independent if they do not influence one another. The criterion of independence is very important — application of the product rule for events that are not independent will give an incorrect answer.

Examples: To find the probability that a couple with three children have three boys we first note that the sex of one child has no influence on the sex of another and therefore constitute independent events. For each child, p(boy) = 1/2 and by the product rule p(3 boys) = 1/2 x 1/2 x 1/2 = 1/8.

First, for a recessive trait to be expressed the progeny must inherit the recessive allele from both the mother *and* the father. Since the probability of inheriting a given allele from a heterozygote is 1/2, p(mutant from mother and mutant from father) = 1/2 x 1/2 = 1/4. Second, since unlinked genes are inherited independently, we can use the product rule again to calculate p(recessives at gene A and recessives at gene B) = 1/4 x 1/4 = 1/16.

SUM RULE

The probability that either **a** or **b** will occur can be written as p(**a** or **b**). If two events **a** and **b** cannot both occur they are *mutually exclusive* and the number of events that satisfy **a** or **b** is $n_a + n_b$. It should be apparent from our definition of probability that:

$$p(\mathbf{a} \text{ or } \mathbf{b}) = \frac{n_a + n_b}{N} = \frac{n_a}{N} + \frac{n_b}{N} = p(\mathbf{a}) + p(\mathbf{b})$$

A useful special case of the sum rule arises when we consider p(**not a**). By definition p(**a**) and p(**not a**) are mutually exclusive and they encompass all possible outcomes. Thus:

$$p(\mathbf{a} \text{ or } \mathbf{not\ a}) = 1 = p(\mathbf{a}) + p(\mathbf{not\ a}) \quad \text{and} \quad p(\mathbf{not\ a}) = 1 - p(\mathbf{a})$$

Examples: Find the probability that a family with three children has at least one girl. We begin by noting that instead of trying to count all possible families with at least one girl it is easier to realize that p(at least one girl) is the same as p(not all boys). Since p(all boys) = 1/8, p(not all boys) = 1– 1/8 = 7/8 = p(at least one girl).

In a cross where both parents are heterozygous for recessive mutations in two unlinked genes, what is the probability that one of their progeny will express at least one of the dominant traits? p(at least one dominant) = 1 – p(both recessive), and from above, p(both recessive) = 1/16. Therefore p(at least one dominant) = 1 – 1/16 = 15/16.

In cases where two events **a** and **b** are independent but *not mutually exclusive*, we can still calculate p(**a** or **b**). In this case we note that the two events **a** and (**b** and **not a**) are mutually exclusive and encompass all outcomes that satisfy **a** or **b** or both. For these mutually exclusive events we can apply the sum rule. Thus,

$$p(\mathbf{a} \text{ or } \mathbf{b}) = p(\mathbf{a} \text{ or } [\mathbf{b} \text{ and } \mathbf{not\ a}]) = p(\mathbf{a}) + p(\mathbf{b} \text{ and } \mathbf{not\ a})$$

Since **b** and **not a** are independent:

$$p(\mathbf{a}) + p(\mathbf{b} \text{ and } \mathbf{not\ a}) = p(\mathbf{a}) + p(\mathbf{b}) \text{ x } p(\mathbf{not\ a}) = p(\mathbf{a}) + p(\mathbf{b}) \text{ x } [1 - p(\mathbf{a})] =$$

$$p(\mathbf{a}) + p(\mathbf{b}) - [p(\mathbf{a}) \text{ x } p(\mathbf{b})]$$

Note that in the case where **a** and **b** are mutually exclusive, p(**a**) x p(**b**) = 0 giving the same formula as for the sum rule.

Example: We can use this formula as another way to solve the last example, which is a case in which the two events are independent but not mutually exclusive. p(at least one dominant) = p(dominant at gene A or dominant at gene B) = p(dominant at gene A)+p(dominant at gene B) – [p(dominant at gene A) x p(dominant at gene B)] = 3/4 + 3/4 – [3/4 x 3/4] =6/4 – 9/16 = 15/16.