2 GENETIC DATA ANALYSIS

2.1 Strategies for learning genetics

We will begin this lecture by discussing some strategies for learning genetics. Genetics is different from most other biology courses you have taken in that memorization is not very important. You are expected to learn vocabulary and some examples of genetic disorders, formulae, etc. But learning and applying concepts is much more important. In particular, you need to be able to think about and analyze **genetic data**. Almost all the homework and exam questions in this part of the course will require you to examine and analyze data, and to make some conclusion based upon your analysis. In this way, the work you do in this course is very similar to work done by real geneticists.

This is one reason why students who have always done well in biology sometimes do poorly in genetics. Learning how to approach a set of genetic data in a logical and consistent way takes patience and a good deal of practice.

I will introduce you to ways to think about problems during lectures. It is therefore very important that you attend lecture. I will use several techniques of "active learning", because these techniques have been proven to be more effective in increasing student learning than the traditional lecture-only format. I think these techniques also make class time more interesting for students.

2.2 Rules of Probability

Sum rule: The combined probability of two events that are mutually exclusive *is the sum of the individual probabilities.*

Genetic example of the sum rule:

Parental genotypes (monohybrid cross): GG x gg *'x' indicates a genetic cross*
F1 genotype: Gg
F2 (produced by mating F1 individuals): 1/4 GG: 1/2 Gg: 1/4 gg *(1:2:1 genotypic ratio)*

Q: What is the probability that an F2 offspring of a monohybrid cross has the dominant phenotype (is either GG *OR* Gg)?

A: P[dominant phenotype] = P[GG] + P[Gg] = ____ + ____ = ____
(Note: the P[x] stands for "Probability of x")

Product rule: The probability of both of two independent events *is the product of the individual probabilities.*

Genetic example using product rule and sum rule:

Parental genotypes (dihybrid cross): GGww x ggWW
F1: GgWw
F2: 9/16 G–W– : 3/16 G–ww : 3/16 ggW– : 1/16 ggww

(Note: the – indicates either the dominant or recessive allele G– indicates an individual with the dominant phenotype: either GG or Gg)

Q: What is the probability that an F2 offspring of the following dihybrid cross will have at least one dominant allele for each trait (i.e. that it will be G–W–)?

Solution: First note that segregation of the G locus is independent of segregation at the W locus (Mendel's law of independent assortment) So that

P[G–W–] = P[G–] * P[W–]

As above, GG and Gg are mutually exclusive (an individual can be GG or Gg, but not both). WW and Ww are also mutually exclusive.

P[G-] = P[Gg] *or* P[GG] = P[Gg] + P[GG]

P[Gg] = 1/2 P[GG] = 1/4

So: P[G-] = 1/2 + 1/4 = 3/4

P[W-]= P[Ww] + P[WW] = 1/2 + 1/4 =3/4

P[G-W-] = P[G-] * P[W-] = 3/4 * 3/4

Answer:

iii. Complex genetic problem--6 independent genes:

Parental genotypes:	AA bb CC DD ee ff	x	aa BB cc dd EE FF
F1:	Aa Bb Cc Dd Ee Ff	x	Aa Bb Cc Dd Ee Ff

Q: What proportion of F2 progeny will be AA bb Cc DD ee Ff ?
A:

iv. Genetic example using conditional probability

Q: In the F2 progeny from monohybrid cross, what is the proportion of heterozygotes among dominant progeny?

A: P[heterozygous] / P[dominant] =
P [Gg] / P[GG or Gg] =
1/2 / [1/4 + 1/2] =
1/2 / 3/4 = 2/3

2.3 Genetic Data Analysis: Goodness of Fit

What if the F2 generation of a monohybrid cross has a phenotypic ratio close to 3:1, but not exactly 3:1?

Q: How do we tell if deviations from expected proportions in a genetic experiment are due to chance or due to the fact that our genetic hypothesis is wrong?

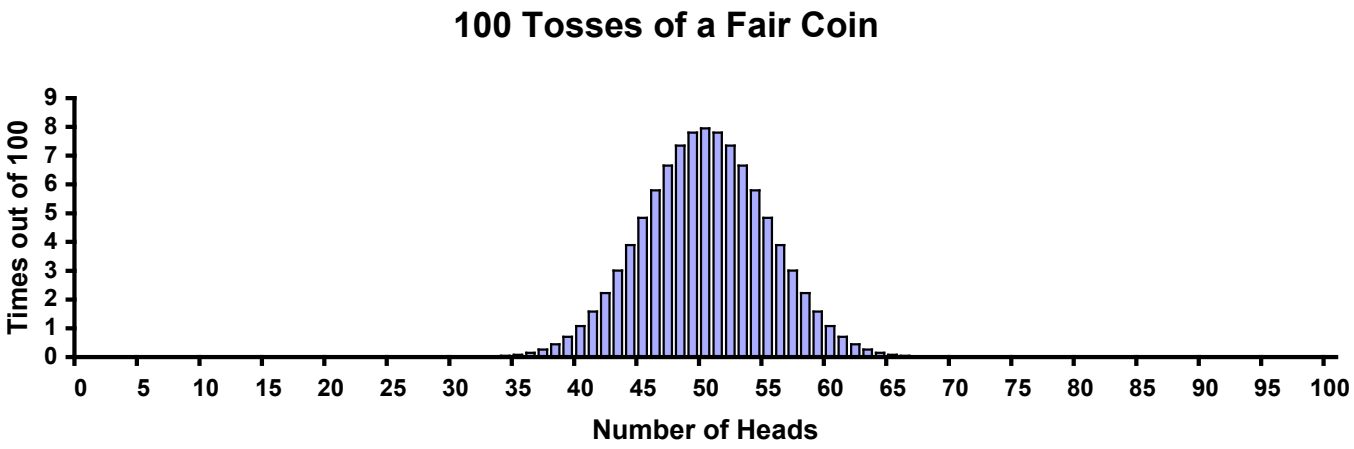
A: Repeat an experiment many times. E.g, the experiment is: toss a fair coin 100 times and count how many heads are throw. Repeat this experiment 100 times.

Q: What is the most likely outcome of a single experiment?

A:

Q: Will all 100 experiments have the same outcome if the coin is fair?

A:



Getting **65 H and 35 T** is an **unlikely result, but not impossible.** How unlikely is it? The probability is actually 0.0018. In other words, this result is expected to occur 18 times out of every thousand times the experiment is repeated (or 1.8 times out of every one hundred trials).

In same way, consider a cross of *Aa* to an individual of unknown genotype that yields 100 progeny. We hypothesize that the unknown individual has genotype *aa*.

Q: If the progeny contain 65 of the dominant phenotype, and 35 of the recessive phenotype, can we reject our hypothesis that the unknown parent was *aa* and that the cross was therefore *Aa* x *aa*? Another way to phrase the question is, "how likely are we to get a result this different from the expected result if our hypothesis is true?"

First, what is the expected result? What is the observed result? How could we measure the difference between the expected and observed results?

A:

A test commonly used to measure the goodness of fit is the **CHI-SQUARE TEST** (pronounced with a hard 'C' and to rhyme with 'pie').

The **observed number** (O) in each category is compared to the **expected number** (E) under your hypothesis. We use the square of the difference between observed and expected numbers, and standardize by dividing by the expected number in a category. (O-E) is referred to as the "difference" (d). the ∑ symbol means to sum the values over each category or "class" of the data

$$\chi^2 = \sum [(O-E)^2]/E = \sum (d^2/E)$$

So for our testcross example:

class	O	E	(O-E)	$(O-E)^2 = d^2$	$(O-E)^2/E$
Aa	65	50	15	225	225/50 = 4.5
aa	35	50	-15	225	225/50 = 4.5
					$\chi^2 = 9$
					d.f. = (#classes -1) = 1

Compare χ^2 to a **theoretical value** from a chi-square table that has the same **degrees of freedom.** Theoretical value based on: Assuming the hypothesis (null hypothesis) we proposed is correct, if we repeated our genetic experiment, many times (counting 100 progeny each time) and calculated a χ^2 for each, we could obtain a distribution of chi-square values. In a certain number of these repeated experiments, we would get 65 or more Aa offspring out of 100, and the χ^2 value associated with these cases would be greater than or equal to the χ^2 value we obtained above (9). The proportion of times that this occurs, is the probability of getting a value of 9 or greater, when you have 1 degree of freedom. In this case the probability associated with a χ^2 value of 9 is less than 1%.

χ^2 Table

	Probabilities					
df	0.90	0.50	0.20	0.05	0.01	0.001
1	0.02	0.46	1.64	3.84	6.64	10.83
2	0.21	1.39	3.22	5.99	9.21	13.82
3	0.58	2.37	4.64	7.82	11.35	16.27
4	1.06	3.36	5.99	9.49	13.28	18.47

0.001<P<0.01 < 1% chance

Degrees of freedom refers to integer number that is generally equal to the number of categories in the data, minus 1. In our coin-toss problem, the degrees of freedom is the number of phenotypic classes that are independent of each other. For example, if there are two categories, there is only one degree of freedom. This **constraint** occurs because, given the total number, **100 progeny**, and the number in one class (65), the number in the other class (35) is fixed. Only one class can vary **freely**, the other is constrained.

To reiterate, due to chance, χ^2 will be **large** for a few samples and small in most other samples. The **Critical Values** of the χ^2 distribution are exceeded by chance only rarely, and that rarity is indicated by the **probability value.** The common practice in genetics is to set the critical value at **5%**. This means that an event that should occur by chance less than 5% of the time, is considered **statistically significant**, and justifies the rejection of the **null hypothesis** (in this example the null hypothesis is that the coin is fair, so that we expect 50% heads and 50% tails). So, we say that the standard for rejecting the null hypothesis is **$P < 0.05$**.

Data from a genetic experiment:

Assume you have crossed two pea plants with purple flowers. Your hypothesis is that both plants are heterozygous for a dominant allele at a single locus controlling flower color. You could write this in symbols as:

H: P: (Ww X Ww)
F1: 3/4 W- (purple) and 1/4 ww (white). *H stands for 'genetic Hypothesis'*
Or a 3:1 phenotypic ratio

Observed Results: Of 166 progeny, 110 had purple flowers and 56 had white flowers.

Expected Results: If your genetic hypothesis is true, then you expect 3/4 purple and 1/4 white out of **166 total offspring. That is, you expect 124.5 purple and 41.5 white.**

Note: If you have learned to think about statistical tests in terms of a "null hypothesis" and an "alternative hypothesis", then the your null hypothesis is the F1 offspring numbers are consistent with a 3:1 phenotypic ratio. The alternative hypothesis is that the F1 offspring numbers are not consistent with a 3:1 ratio.

Q: Is the difference between observed and expected too big to be due to chance? If it is, then you reject your proposed genetic hypothesis (you reject the null hypothesis in favor of the alternative hypothesis).

χ^2 TEST

Class	O	E	(O-E)	$(O-E)^2 = d^2$	$(O-E)^2/E$
Purple	110	124.5	-14.5	210.25	210.25/124.5 = 1.69
White	56	41.5	14.5	210.25	210.25/41.5 = 5.07
					$\chi^2 = 6.76$
					1 degree of freedom

This χ^2 value is a relatively large number indicating that the observed numbers are not very close to the expected, but we need an objective measure of how close to the expected it really is. Since 6.76 is larger than the value in the table for P=0.01 with one degree of freedom, we can be fairly confident that our hypothesis is **not correct**. We would expect this large a deviation from our expected results **due to chance alone** fewer than once out of every 100 repeated experiments.