Mendelian Genetics 2
Probability Theory and Statistics

Mathematicians distinguish two kinds of processes:

deterministic	**outcomes predicted exactly**	**flip coin with two heads-> H**
stochastic	**outcomes have probabilities**	**flip coin with H and T**

Models in science:

Deterministic	**Newtonian physics**	**Stochastic**	**quantum theory**

In everyday language, stochastic ≈ random; in probability theory, random sometimes restricted to cases where all outcomes are equally probable. I'll usually say "strictly random".

fair coin	**stochastic and (strictly) random**
weighted coin	**stochastic, not strictly random**

Computers generate pseudorandom numbers: start with number you give it or the time of day, go through long series of arithmetic operations. Resulting series of numbers could be predicted exactly and hence are determiistic IF you knew the starting number. If you don't, almost impossible to distinguish from strictly random.

How Mendel saw randomness as variation among progeny of different plants.

F_2 from a cross showed 336 round:102 wrinkled, very close to 3:1. Stochastic or deterministic? Mendel looked at individual plants, got the following among others:

round	45	43	14
wrinkled	11	2	15
	3:1	20:1	1:1

Can use Punnett squares and intuition to solve many problems in genetics, i.e. to predict kinds and frequencies of gametes and progeny, phenotypes and genotypes. But can be very complicated; e.g. 3-factor cross may require up to 64-block Punnet square. Better to learn to use a bit of basic probability theory.

Terminology:

roll die P(6) = 1/6 = 0.1667 % Don't use %!! Use fraction or decimal fraction.

probability of an event or outcome can range 0 (impossible) --> 1 (must happen)

P(r r --> R gamete) = 0
P(r r --> r gamete) = 1
P(R r --> R gamete) = 1/2

Two "Kinds" of Probabilities, or Two Ways of Thinking About Them

***1. a priori* probabilities are based on model or hypothesis**

E.g. toss coin. Whether lands H or T depends on details of how one flips it and where one catches it. Assume we could never control thumb and hand precisely enough to control outcome. Then either outcome equally likely, or P(H) = P(T) = 1/2. (I have read that chaos theory has been used to verify this.)

E.g. Mendel hypothesized that fusion of gametes is random with respect to the genes he studied in peas. Self *A a*, P(*A* pollen & *A* egg) = P(*A* pollen & *a* egg), etc.

2. *a posteriori* probabilities = observed frequencies of events

E.g. toss coin many times, frequency of heads = f(H) ≈ 1/2.

E.g. Mendel observed frequencies *A A* = *A a* = *a A* = *a a*

To do most kinds of genetics, need learn only two basic probability rules and how to apply them.

1. **Independent events**: Occurrence of one doesn't affect probability of the other.

e.g. toss 2 coins or 1 coin 2 times, H1 and T2 are independent
pick 1 egg and 1 pollen from *Rr* plant, *R* egg and *R* pollen are independent

If events M, N, O, ... are independent, P(M & N & O ...) = P(M) P(N)P(O) ...

e.g. toss 2 coins P(H1 & T2) = P(H1)P(T2) = (1/2)(1/2) = 1/4
P(R egg and r pollen from Rr) = (1/2)(1/2) = 1/4

2. **Mutually exclusive events**: Cannot occur together.

e.g. toss 1 coin, H and T are mutually exclusive, can get one or the other, not both
1 gamete from *Rr* plant --> *R* or *r*, not both

If events A, B, C ... are mutually exclusive, P(A or B or C ...) = P(A) + P(B) + P(C) ...

e.g. P(H or T) = (1/2) + (1/2) = 1
P(F2 from Rr X Rr is round) = P(RR or Rr) = P(RR) + P(Rr) = (1/4) + (1/2) = 3/4

Same result as Punnet square and intuition.

This is easy. Hard part:
Know which rule to use.
Know how to combine rules to solve problem.

Do simple cases, relate to intuition and Punnett square.

(1) Toss 2 coins. P(1 H & 1 T) = ? Order not specified, want any order.

P(T,T) = P(H1 & T2 or T1 & H2) = [P(H1)P(T2)] + [P(T1)P(H2)]
indep. indep.
mutually exclusive (compound events)
= (1/2)(1/2) + (1/2)(1/2) = (1/4) + (1/4) = 1/2

cf. Punnett square
1/4 H T + 1/4 H T = 1/2 H T

		Toss 2	
		1/2 H	1/2 T
Toss 1	1/2 H	1/4 H H	1/4 H T
	1/2 T	1/4 T H	1/4 T T

(2) Rr × Rr --> ?

P(RR) = P(Rf & Rm) = (1/2)(1/2) = 1/4

P(rr) = P(rf & rm) = (1/2)(1/2) = 1/4

P(Rr) = P(Rf & rm or rf & Rm) = (1/2)(1/2) + (1/2)(1/2) = 1/2

or P(Rr) = 1 – P(RR or rr) = 1 – [(1/4) + (1/4)] = 1/2

The last point is very important; in many cases it is easier to calculate the probability that something does not happen and subtract it from 1 than it is to calculate the probability that it does happen directly.

(3) Rr Yy Tt × Rr yy tt ---> 2,000 seeds How many do we expect to be round and green and produce tall plants?

Translate to genotypes: expect how many R– yy T– ?

Three pairs of alleles segregate independently, so start by doing each one separately.
Rr × Rr --> 3/4 R–
Yy × yy --> 1/2 yy
Tt × tt --> 1/2 Tt

P(R– yy T–) = (3/4)(1/2)(1/2) = 3/16 = expected frequency

expected number = (3/16)(2,000) = 375

Conditional Probabilities

Conditional probabilities show how our assignment of probabilities depend on our prior knowledge.

e.g. Rr X Rr --> 1/4 RR 1/2 Rr 1/4 rr What proportion of round peas are homozygous? Translate to probability language: what is the probability that a pea is homozygous, given that it is round?

There is a law of conditional probabilities:

P(A given B) = P(A and B)/P(B)

P(A|B) = P(A∩B)/P(B)

P(RR|R-) = P(RR∩R-)/P(R-) = (1/4)/(3/4) = 1/3

But you don't have to use it in any situation that we will consider.

Instead, note that when I specified that the peas must be round, I eliminated one possible outcome, wrinkled peas. This changes the probabilities:

RR 1/4	Rr
Rr	rr

4 equally likely outcomes

RR 1/3	Rr
Rr	

3 equally likely outcomes

I have 2 children. What is P(2 F)?

P(FF) = P(1stF & 2ndF) = P(1stF)P(2ndF) = (1/2)(1/2) = 1/4

FF 1/4	FM
MF	MM

I have 2 children. The first one is F. (A condition is put on it.) What is P(2F)? We have eliminated two possible outcomes, MF and MM.So the Punnett square is:

FF 1/2	FM

P(FF|F1) = P(FF∩F1)/P(F1) = (1/4) /(1/2) = 1/2

Punnett Squares With Unequal Probabilities

Punnett squares are ways of getting all possible combinations of things.
e.g. all combinations of gametes

In cases we have considered, the probabilities are all equal. But they don't have to be. Consider tossing a weighted coin which has probability of heads = 0.6 and tails 0.4. Toss it twice (or toss two such coins):

What are the probabilities?

P(HH) = (0.6)(0.6) = 0.36
P(HT) = P(TH) = (0.6)(0.4) = 0.24
P(TT) = (0.4)(0.4) = 0.16
Check: 0.36 + 2(0.24) + 0.16 = 1

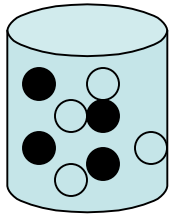
We could use a Punnett Square as follows:

		First toss	
		0.6 H	**0.4 T**
Second Toss	**0.6 H**	**0.36 H H**	**0.24 H T**
	0.4 T	**0.24 H T**	**0.16 T T**

Binomial Probability Distribution

Problem: Mendel observed approximately 3:1 ratio in the F2 of one-factor crosses, but when he looked at small samples from single pods, he often got ratios very different from 3:1. Suppose you repeat one of his crosses but only look at one pod and get ratio 4 round and 4 wrinkled peas. This could happen, but how likely is it?

Can get 4 r and 4 w in many different orders or permutations:
w w w w r r r r
r w w r r r w w
etc.



P(one order) = $(3/4)^4 (1/4)^4$
Orders mutually exclusive so if we knew how many there were, we could get the answer by multiplying the P(one order) X number orders.

n trials (experiments), each with same possible mutually exclusive outcomes which have same probabilities in each trial. Want probability that a particular outcome will happen w times, another happens x time, etc.

$P(w,x) = (n!/w!x!)p^w q^x$ = number of permutations (orders) × probability of one permutation

where n = number of trials
w = number of occurrences of outcome E1 with probability p
x = number of occurrences of outcome E2 with probability q
w + x = 1 P + q = 1

$P(4\ r, 4\ w) = (8!/4!4!)\ (3/4)^4\ (1/4)^4 = 70$ X 0.001236 = 0.0865

Reminders about factorials

$n! = 1 \times 2 \times \ldots n$

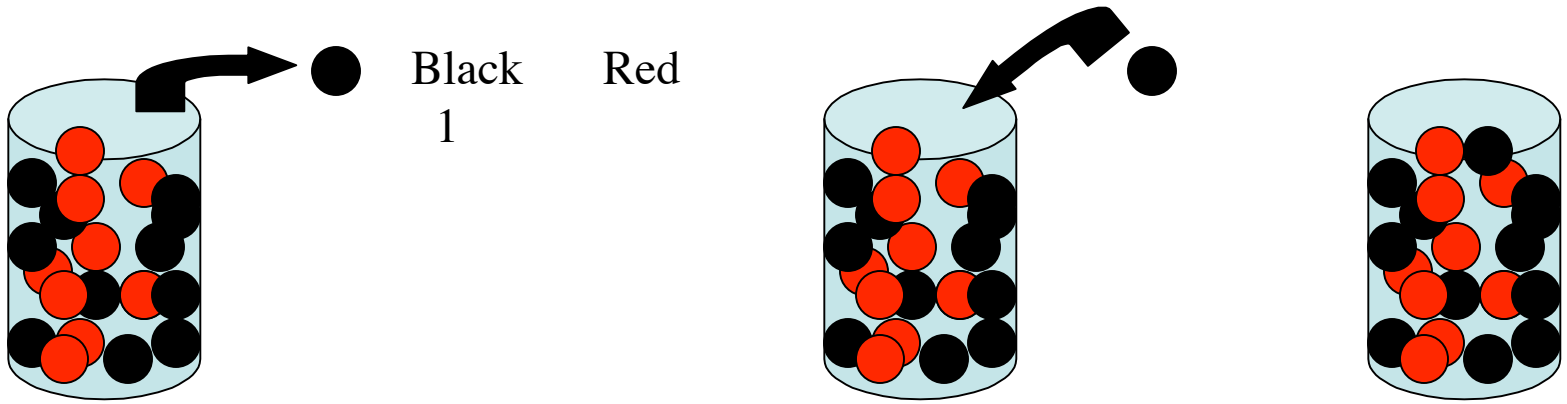
$4! = 4 \times 3 \times 2 \times 1 = 24$

$0! = 1$

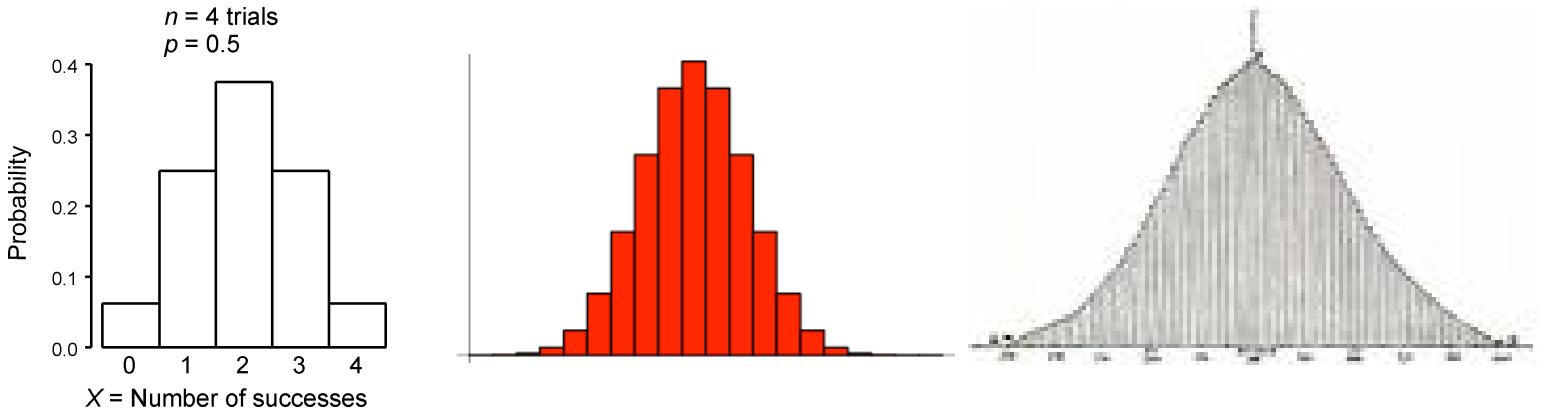
Factorials often cancel:

$$\frac{6!}{4!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1} = 6 \times 5$$

Binomial distribution is also used in an urn model called sampling with replacement. Imagine an urn with lots of balls, half of which are labeled female and half are labelled male. Draw a ball, look at the label and record it, then return the ball and draw again. Draw n times. Each time the probability of drawing a ball labeled female is 1/2 and the same for a ball labeled male.



Binomial distributions: frequency distributions approach normal distribution as number of trials increases.



The binomial equation can be extended to any number of events as a multinomial equation:

$$P(w,x,y...) = (n!/w!x!y!\ldots)p^w q^x r^y ...$$

R rYy X R rYy --> 5 progeny

$$P(3\ \text{R-Y-},\ 2\ \text{R-yy},\ 0\ \text{rrY-},0\ \text{rryy}) = \frac{5!}{3!\ 2!\ 0!\ 0!}(9/16)^3(3/16)^2(3/16)^0(1/16)^0$$

Or

$$\frac{5!}{3!\ 2!}(9/16)^3(3/16)^2$$

Expected ratios: how often do we expect to get them?

Toss coin 4X, expect 2 H & 2 T. Gamble with me: toss coin 4X. You bet on the expected result, 2 H & 2 T; I'll bet against it. Who will accept?

Expected ratios: how often do we expect to get them?

Toss coin 4X, expect 2 H & 2 T. Gamble with me: toss coin 4X. I will let you bet on the expected result, 2 H & 2 T; I'll bet against it. Who will accept?

I will win $10 for every $6 you win.

Probabilities calculated from binomial distribution:

4 H	**1/16**	
4 T	**1/16**	**sum 10/16 less likely to get 2 H & 2 T than something else**
3 H, 1 T	**4/16**	
1 H, 3 T	**4/16**	
2 H, 2 T	**6 /16**	**most likely single outcome ... that's what "expected" means**
	1	

Rr Yy Tt × Rr yy tt --> 3/16 round green tall
What is probability of getting exactly 375 round green tall out of 2,000 seeds?

$$\frac{2000!}{375!\ 1624!}\ (3/16)^{375}(13/16)^{1625} \approx 10^{-2.303} \approx 0.005$$

Used Stirling's approximation for large factorials; most computers can't handle these.

We don't expect to get exactly the expected frequencies. But if they are very different we might decide that our expectations are wrong ... i.e. that we used the wrong model or hypothesis or explanation. How much can our observed frequencies differ from the expected frequencies before we decide that our model is wrong?

Some biologists and other scientists think that one should never need to use statistics.They are using intuition to decide if their observations are significantly different from their expectations. How good is our intuition?

Four-O'Clock flowering plant *RR* = red *Rr* = pink *rr* = white

Nursery has a lot of seeds which are supposed to come from a cross *Rr* × *Rr*. The expected ratio of phenotypes is 1/4 RR red : 1/2 Rr pink : 1/4 rr white.

This is a case of incomplete dominance.

	observe				
red	**pink**	**white**	**expect**	χ^2	**P**
20	**44**	**36**	**25 50 25**		

Many experiments produce numerical data. If one thinks the results are obvious, this means one is doing statistics in one's head, i.e. one is doing bad statistics.

We don't expect to get exactly the expected frequencies. But if they are very different we might decide that our expectations are wrong ... i.e. that we used the wrong model or hypothesis or explanation. How much can our observed frequencies differ from the expected frequencies before we decide that our model is wrong?

Some biologists and other scientists think that one should never need to use statistics.They are using intuition to decide if their observations are significantly different from their expectations. How good is our intuition?

Four-O'Clock flowering plant *RR* = red *Rr* = pink *rr* = white

Nursery has a lot of seeds which are supposed to come from a cross *Rr* × *Rr*. The expected ratio of phenotypes is 1/4 RR red : 1/2 Rr pink : 1/4 rr white.

This is a case of incomplete dominance.

	observe						
red	**pink**	**white**	**expect**	**χ^2**	**P**		
20	**44**	**36**	**25 50 25**	**6.56**	**0.05 - 0.01**	**reject**	**N = 100**
8	**9**	**8**	**6 13 6**				

Many experiments produce numerical data. If one thinks the results are obvious, this means one is doing statistics in one's head, i.e. one is doing bad statistics.

We don't expect to get exactly the expected frequencies. But if they are very different we might decide that our expectations are wrong ... i.e. that we used the wrong model or hypothesis or explanation. How much can our observed frequencies differ from the expected frequencies before we decide that our model is wrong?

Some biologists and other scientists think that one should never need to use statistics.They are using intuition to decide if their observations are significantly different from their expectations. How good is our intuition?

Four-O'Clock flowering plant *RR* = red *Rr* = pink *rr* = white

Nursery has a lot of seeds which are supposed to come from a cross *Rr* × *Rr*. The expected ratio of phenotypes is 1/4 RR red : 1/2 Rr pink : 1/4 rr white.

This is a case of incomplete dominance.

	observe						
red	pink	white	expect	χ^2	P		
20	44	36	25 50 25	6.56	0.05 - 0.01	reject	N = 100
8	9	8	6 13 6	1.96	0.5 - 0.3	accept	N = 25
16	18	16	12 25 13				

Many experiments produce numerical data. If one thinks the results are obvious, this means one is doing statistics in one's head, i.e. one is doing bad statistics.

We don't expect to get exactly the expected frequencies. But if they are very different we might decide that our expectations are wrong ... i.e. that we used the wrong model or hypothesis or explanation. How much can our observed frequencies differ from the expected frequencies before we decide that our model is wrong?

Some biologists and other scientists think that one should never need to use statistics.They are using intuition to decide if their observations are significantly different from their expectations. How good is our intuition?

Four-O'Clock flowering plant *RR* = red *Rr* = pink *rr* = white

Nursery has a lot of seeds which are supposed to come from a cross *Rr* × *Rr*. The expected ratio of phenotypes is 1/4 RR red : 1/2 Rr pink : 1/4 rr white.

This is a case of incomplete dominance.

observe							
red	pink	white	expect	χ^2	P		
20	44	36	25 50 25	6.56	0.05 - 0.01	reject	N = 100
8	9	8	6 13 6	1.96	0.5 - 0.3	accept	N = 25
16	18	16	12 25 13	4.38	0.2 - 0.1	accept	N = 50

Many experiments produce numerical data. If one thinks the results are obvious, this means one is doing statistics in one's head, i.e. one is doing bad statistics.

Statistical analysis is a way of determining how much confidence we can have in an interpretation of data with a stochastic component.

We can use the binomial distribution to calculate the exact probability of getting a particular result. But often we want to know only whether the observed results are significantly different from expectation. The Fisher exact probability test uses the binomial distribution to do this, but it is very computer-intensive for large samples.

Statistics =
The science of analyzing data
Data (better just to call them data)
Descriptive statistics = measures of central tendency (average = mean, mode, etc.) and of dispersion (variance, standard deviation, etc.).
Hypothesis-testing statistics = testing the validity of a model.

Recall that an *a priori* probability is defined by a model or hypothesis, while an *a posteriori* probability is defined by measuring the frequency of an event.

The validity of a model may be tested by comparing the *a priori* probabilities or expected frequencies defined by the model with the observed data. The comparison is made using a statistical test.

Appropriate statistical test for many kinds of genetic data is chi-square test. Read about it in text starting on p. 162. Will do an example in Discussion.

Mendel didn't cheat ... deliberately.

Mendel's results tended to be very close to expected. R. A. Fisher (1936) calculated pooled Chi-square for all Mendel's experiments. Chi-square = 41.6 84 d.f. P ≈ 0.99993
P (such good results by chance) ≈ 0.00007

Mendel was scrupulously honest. He communicated with Carl Nageli, most eminent student of heredity at the time. Nageli wasn 't interested in pea data, didn't understand Mendel's results. Urged Mendel to work with *Heiracium*. = hawkweed. Mendel did, but didn't get same results. Now know is because *Heiracium* reproduces asexually sometimes. Nevertheless, he described his results in a letter to Nageli. If Mendel was inclined to cheat, he should have done so here, cooked results to make *Heiracium* obey Mendel's laws, and maybe he could have got Nageli on his side. But he didn't.

Most likely explanation: Mendel cheated unconsciously.
e.g.:
- **Count yellow and green peas from huge bowl, get tired, stop before all done ...tend to stop when ratios near expected. Must decide in advance how many to count!**
- **Unconsciously pick peas so agree with expected ratio. Sample blind, or use table of random numbers, etc.**
- **Repeat or check experiments which give results that disagree with expectations, but not those that agree. Common practice, but wrong ... biases results in favor of expectations.**

Another possibility: Weiling noted three later geneticists got too good agreement with results when used peas. Suggested gamete sampling not strictly random: maybe the 4 pollen grains produced by one meiosis tend to stick together during pollination (like tetrad analysis), so gamete genotypes closer to equal frequencies than if strictly random.

What happened to Mendel's theory? Ignored until rediscovered in 1900.Why?

Mendel ahead of his time. Took other biologists > 2 decades to catch up.

- **Mathematical models became more popular in biology.**
- **Idea of a particulate gene proposed.**
- **Discovery of chromosomes, mitosis, and meiosis provided a plausible place for the genes and a physical basis for his laws.**

Time Line of Revolutions in Genetics