Introductory Applied Bio-Statistics

IMPORTANT: you will need a calculator for this section of the course and for exam questions over this section!

For the purposes of this section, a population is a group of inter-breeding people with/of equal species that live together.

Frequency (*f*)

- What is the frequency of a genotype in a population if there are three (3) genotypes, C_1C_1, C_2C_2, C_1C_2 (C_{11}, C_{22}, C_{12}, respectively) and there are C_T people in the population (T means total)?

Frequency (*f*)

- The frequencies of each would be determined as shown right:
- For the frequency of C_{11}, it is equal to the number of homozygous genotypes divided by the total number of the people in the population with the C genotype.

$$f_{11} = \frac{\#C_{11}}{\#C_T}$$

$$f_{12} = \frac{\#C_{12}}{\#C_T}$$

$$f_{22} = \frac{\#C_{22}}{\#C_T}$$

$$f_{11} + f_{12} + f_{22} = 1$$

- The sum of the frequencies of these genotypes equals 1, i.e., CLUE: think per cent: if a you have a pie and it is cut into a number of pieces of different sizes, then the sum of the per cent of each piece of the pie is equal to 100%; alternatively, if you only use fractions rather than the per cent values, the sum of the fraction of each piece of pie is equal to unity (1).

How does one calculate the frequency of a specific allele?

- IMPORTANT: when ALL alleles are assumed to be observed in heterozygotes is our ASSUMPTION.
- Allele = one of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome.

$f\,C_1$		
$f\,C_1$	$f\,C_{11} + 1/2(f\,C_{12}) = C_{11}/C_T + 1/2(C_{12}/C_T)$	$= (C_{11} + 1/2(C_{12}))/C_T$
$f\,C_2$		
$f\,C_2$	$f\,C_{22} + 1/2(f\,C_{12}) = C_{22}/C_T + 1/2(C_{12}/C_T)$	$= (C_{22} + 1/2(C_{12}))/C_T$
f "rule"		
$f\,C_1 + f\,C_2 = 1$	$\therefore f\,C_1 = 1 - f\,C_2$	

E.g.,

- In 500 people, there are 268 with C_1C_1 (Normal), 100 with C_1C_2 (Carriers) and 132 with C_2C_2 (Disease). Determine the frequencies of C_{11}, C_{12}, C_{22}, the frequency of C_1 and the frequency of C_2:

$\#C_{11}$	$\#C_{12}$	$\#C_{22}$
268	100	132
$f\,C_{11}$	$f\,C_{12}$	$f\,C_{22}$
268/500 = 0.536	100/500 = 0.2	132/500 = 0.264
(53.6%)	(20%)	(26.4%)

And $f\,C_1$ is:

$$(268 + 1/2(100))/500 = 318/500 = 0.636$$

And $f\,C_2$ is:

$$1 - f\,C_1 = f\,C_2 = 1 - 0.636 = 0.364$$

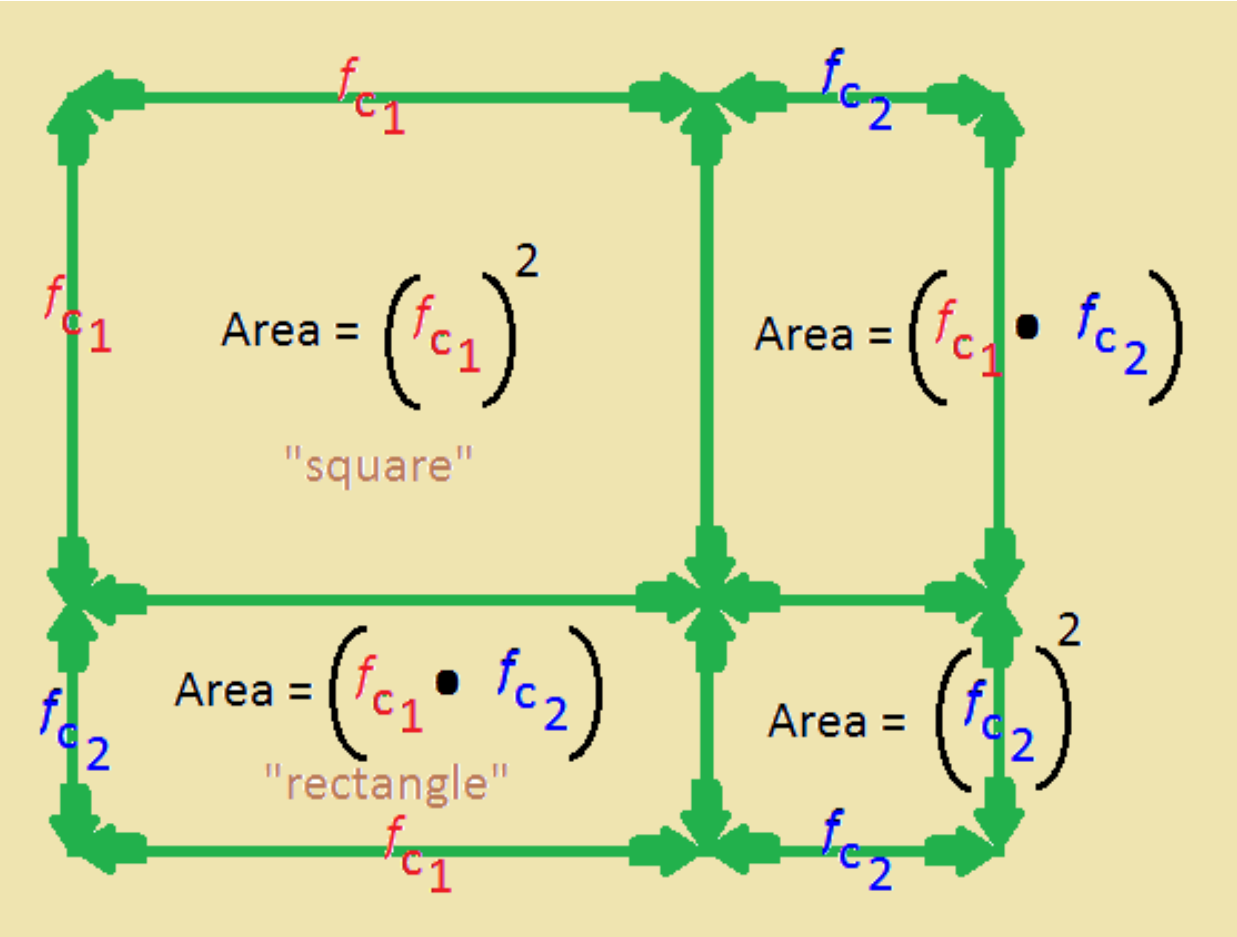
- How does one use this information?
- By definition, if the frequency value is calculated to be greater than or equal to 0.99, then this is a monomorphic gene. If the frequency value is calculated to be less than 0.99, then this is a polymorphic gene.
- A **monomorphic gene** is a gene with not more than one allele in high frequency;
- A **polymorphic gene** is a gene that has 2 or more alleles in substantial frequency.
- From our example, above, both C_1 and C_2 are polymorphic ($0.636 < 0.99$ and $0.364 < 0.99$, respectively). Although these alleles are not really frequent, both are polymorphic.

Hardy-Weinberg Principle

- The Hardy-Weinberg Principle shows a relationship between the frequency of the alleles and the frequency of the genotypes. The frequency of the genotypes for a gene with 2 different alleles are a binomial function of the frequency of the alleles.
- Four assumptions are inherent in this model:
  1. The population is "breeding" for 2 alleles, C_1 and C_2 with frequencies of C_1 and C_2 such that their sum is unity.
  2. A sperm and an ovum unite at random to produce a zygote.
  3. No other processes effect the variation of the genotype.
  4. The frequencies of all C's are identical in sperm and the ovum.

- Let's use our above example in understanding the Hardy-Weinberg Principle. We can examine this graphically with a sort of modification/expansion of the Punnett square and Mendel's work:

- If we make the long lengths of a square and rectangle equal to the $f\,C_1$ and the short sides of a rectangle equal to the $f\,C_2$, then we may construct the diagram:



- Remember that the area of a square is the product of two of its sides. Since a square has equal sides, the Area is equal to one side squared.
- The area of a rectangle is the product of its long side and its short side.
- In the case of the graphic, the area of a square is equal to $(f\,C_1)^2$, $(f\,C_2)^2$ and the area of a rectangle is $f\,C_1 \bullet f\,C_2$.

- If we plug in the numbers from our original example, we'll find that $(fC_1)^2$ equals 0.4045 (40.45% C_1C_1) and $(fC_2)^2$ equals 0.132 (13.2% C_2C_2) and $2(fC_1\ fC_2)$ equals 0.464 (46.4% C_1C_2).
- If we add up all the areas of the squares and rectangles, we obtain the following equation for the complete square:
-

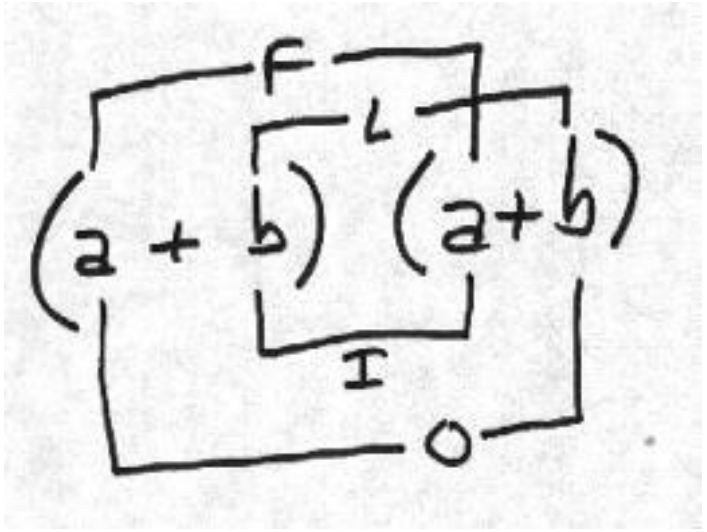
$$(fC_1)^2 + 2(fC_1\ fC_2) + (fC_2)^2 = 1$$

Which ALSO equals:

$$(fC_1 + fC_2)^2$$

ASIDE MATH Primer Reminder:

- Equations of the form $(a + b)^2$ are called BINOMIAL (has exactly 2 terms); using the FOIL method, $(a + b)^2$ expands to that found below:



F means to multiply the **F**irst terms together in the two parenthetic phrases; O means to multiply the **O**uter terms; **I** means to multiply the inner terms together; **L** means to multiply the last two terms together.

We then obtain:

$a^2 + ab + ba + b^2$

which is equal to:

$a^2 + 2ab + b^2$

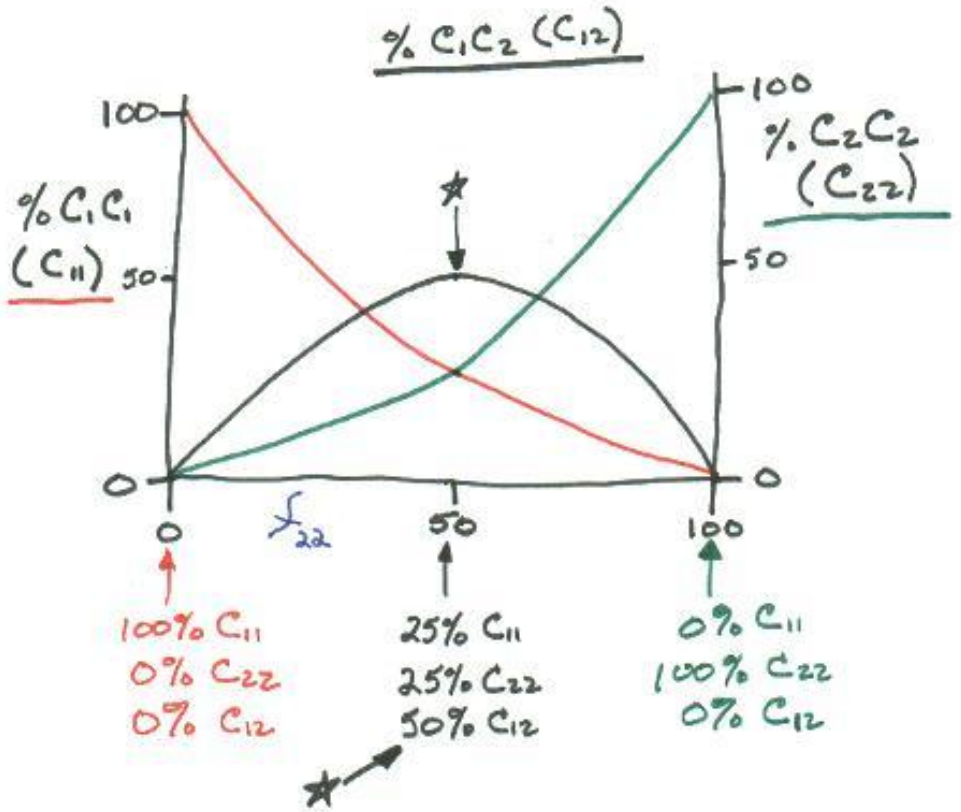
which LOOKS just like:

$(fC_1)^2 + 2(fC_1\, fC_2) + (fC_2)^2$

Hence, the proportion of progeny are a function of binomial arithmetic

END of ASIDE

- This binomial relationship allows us to examine the relationships between the three alleles we've been working with:
  1. **The maximum frequency of C_{11} occurs when the frequency of C_1 = 1 and the frequency of C_2 = 0**.
  2. **The maximum frequency of C_{22} occurs when the frequency of C_2 =1 and the frequency of C_1 = 0.**
  3. **The maximum frequency of C_{12} occurs when the frequency of C_1 is identical to the frequency of C_2; this value is 0.5, and marked with a star**.
- The bottom line, here, is that as C_1C_1 decreases, both C_1C_2 and C_2C_2 increase; conversely, as C_2C_2 decreases, C_1C_1 and C_1C_2 increase.



- Not withstanding the numerical evaluation, the critical concept of the Hardy-Weinberg Principle is that genotypic frequencies stay in the EXACT proportions of binomial arithmetic, *ad infinitum*!

As an extension of the Hardy-Weinberg Principle, one may estimate the proportions of a carrier in a given, defined population, e.g.:

$$(fC_2)^2 = \#C_{22}/\#C_T$$

and

$$\sqrt{(fC_2)^2} = \sqrt{(\#C_{22}/\#C_T)} = \text{estimation of the frequency of } C_2$$

ASSUMPTION: all alleles are NOT observed in the heterozygote, i.e., one dominant trait coded for and one recessive trait coded for: this is more REALISTIC!

The Estimation of a Carrier in A Population ($\equiv E_P$)

This value for a given, defined population is:

$$E_{PC_2} = 2\sqrt{(\#C_{22}/\#C_T)} * [1 - \sqrt{(\#C_{22}/\#C_T)}]$$

OR

$$E_{PC_2} = 2\,(fC_2)\,(1 - fC_2)$$

Example

Clubfoot appears in one of every thousand births. Determine what per cent of the population are carriers for this allele. From the previous equations:

$$f_{\text{clubfoot}} = \sqrt{(1/1000)} = 0.0316$$

and

$$1 - f_{\text{clubfoot}} = 1 - 0.0316 = 0.9684$$

and $E_{p_{\text{clubfoot}}} = (2)(0.0316)(0.9684) = 0.0612$

% = 0.0612 * 100 = 6.12% of the population (1 in 16 people) are carriers of the clubfoot allele. Surprising?!

The Chi Squared (χ^2) Test

- In some cases, it is necessary to use statistics to determine if the observed frequency of genotypes is close enough to Hardy-Weinberg expectations. The

  "big" test is the Chi squared (χ^2) test. This statistical manipulation is equal to the following generic equation:

$$\chi^2 = \Sigma\left[(\text{Observed \#'s} - \text{Expected \#'s})^2/\text{Expected \#'s}\right]$$

NOTE: <u>NUMBERS</u> NOT frequencies!!!!!!

- χ^2 is dependent on the degrees of freedom (dof);
- GENERALLY this equal to the number of characteristics less one.

One may use χ^2 2 ways:

Confidence			Error		
"Goodness of Fit"			"Sloppy Sampling"		
Confidence Level			p Level		
dof	99%	95%	dof	0.05+	0.01*
1	6.63	3.84	1	0.0039	0.0002
2	9.21	5.99	2	0.103	0.0201
3	11.3	7.81	3	0.352	0.115
4	13.3	9.49	4	0.711	0.297
8	20.1	15.5	8	2.73	1.65
10	23.2	18.3	10	3.94	2.56
For confidence of "fit of observed data" to that expected			+5%, *1% error; for error in data sampling -- examples coming up shortly		

APPLICATION for both uses:

χ^2 calculated < χ^2 table means:

Confidence: GOOD

Error: < 5%+ or 1%*

- If p (probability) > 0.05, you probably do not have a correct hypothesis;
- If p < 0.05, this suggests < 5%+ error in your hypothesis;
- If p < 0.01 suggests that there is less than 1%* error in your hypothesis.

Example

Peas in Mendel's experiments exhibited 2 of 4 characteristics (NOTE: dof = 4 - 1 = 3):

Round	Yellow	Wrinkled	Green
R	Y	W	G

Mendel's theory of heredity predicts peas in the following proportions:

9 RY: 3 RG: 3 WY: 1 WG

In practice, he obtained as follows:

315 RY: 108 RG: 101 WY: 32 WG

1. Does the data (observed) fit with the expected data? A) @ 95% and B) @ 99% confidence?
2. Are the results subject to > 5% error?
3. Are the results subject to > 1% error?

FIRST: sum the number of peas:

$$315 + 108 + +101 + 32 = 556$$

SECOND: What is EXPECTED?

The sum of 9 + 3 + 3 + 1 = 16

∴ 9/16 are RY, 3/16 are RG, 3/16 are WY and 1/16 are WG

We'd expect the following, then:

9/16 * 556 = 313 RY	3/16 * 556 = 104 RG
3/16 * 556 = 104 WY	1/16 * 556 = 35 WG

THIRD: calculate χ^2:

$$\chi^2 = [(315-313)^2/313] + [(108-104)^2/104] + [(101-104)^2/104] + [(32-35)^2/35] =$$

$$4/313 + 16/104 + 9/104 + 9/35 = 0.510$$

FOURTH: calculate dof

4 characteristics, so 4-1 = 3 dof.

FIFTH: compare to table value: Step 1

	Confidence level	Calculated c^2		Table c^2	FIT
A	95%	0.51	vs.	7.81	Good
B	95%	0.51	Vs	11.3	Good
3 dof					

FIFTH: compare to table value: Step 2

	Calculated c^2		Table c^2
< 5% error	0.51	Vs.	0.352
YES			

FIFTH: compare to table value: Step 3

	Calculated c^2		Table c^2
< 1% error	0.51	Vs.	0.352
YES: Fit and error are good.			

Student's Two-Tailed T-test for Significance

Sometimes, though, you just want to compare values between two groups to determine if there is a significant difference. Student's Two-Tailed T-test for Significance does just that. The necessities for this test are listed and defined, below:

Group 1 Data	Group 2 Date
$Mean_1$	$Mean_2$
s_1	s_2
N_1	N_1

Mean = the average value in each group

s = the standard deviation in each group and equals a measure of the difference between a value and the average value "corrected" for the number of samples; determined on a calculator on the "sigma" (σ) key.

N = the number of samples

How to calculate p value:

$$Step\ 1: \sqrt{\left[\frac{(s_1)^2}{N_1}+\frac{(s_2)^2}{N_2}\right]}=Q$$

$$Step\ 2: \frac{(mean_1 - mean_2)}{Q}=Z\ (\pm)$$

$$Step\ 3: Look\ up\ Z\ in\ table\ , below$$

Z Values

p =	0.1	0.05	0.01	0.005	0.002
Z for 2-tail test	± 1.645	± 1.96	± 2.58	± 2.81	± 3.08
At infinite dof					

EXAMPLE

- In an experiment designed to study the effects of a DNA binding anti-cancer drug, ACS-19685, the drug was used to treat cancerous cells in culture. All cultures began with 50 cells per culture; 10 cultures were treated with ACS-19685 and ten were treated with the inert carrier solvent. The following data was obtained:

Culture #	ACS-19685 treated: cells remaining	Control: cells remaining
1	4	45
2	5	40
3	10	42
4	7	47
5	12	40
6	8	39
7	2	41
8	20	43
9	1	44
10	6	48

Are these results statistically significantly different by the 2-tailed t-test?

SOLUTION:

	ACS-19685 treated	Control
AVG	7.5	42.9
s	5.26	2.91
N	10	10

$$Q = \sqrt{\left[\frac{(5.26)^2}{10} + \frac{(2.91)^2}{10}\right]} = 4.303$$

$$Z = \frac{(7.5 - 42.9)}{4.303} = -8.227$$

$$\therefore\ p << 0.002$$

YES , they are

- This means that about 99.8% of the time, there was no error, therefore, % error possible is less than 0.2% and the difference is real between the two groups.