Population & Quantitative Genetics

Dr. Nabil Amor

2018/2019

DEPARTMENT OF ZOOLOGY
King Saud University

Brief Contents

A. Population Genetics

Objectives: This is an observational lab, wherein you will do simulations that help to understand the basic principals of population genetics. When you finish this lab, you should:

1. Understand the underlying genetic principals behind the Hardy-Weinberg equilibrium equation
2. Understand the role of assortative mating in the disruption of the Hardy-Weinberg equilibrium
3. Understand the role of large population size in the maintenance of the Hardy-Weinberg equilibrium

I. PROCEDURE

PART I. DEMONSTRATION OF HARDY-WEINBERG EQUATION

A. Allele frequencies in human populations. It is possible to use data from your class to determine the allele frequencies for common genetic traits. Several human traits are determined simply by one genetic locus with two alleles that interact in Mendelian fashion (i.e., one dominant and one recessive allele). You will determine your own phenotype for each of these characteristics, then collect the data from the entire class.

Traits:

i) **Ear lobes** may be free (detached) or attached to the head. The dominant allele is "detached".

ii) "**Hitchhikers" thumb** is a trait wherein the middle joint of the thumb is hyper-extensible, so that the thumb bends back at nearly right-angles to the plane of the hand. The alternative phenotype is a straight thumb that does not hyper extend. The dominant allele is "hitchhikers".

iii) **Widow's peak** is when the hair line across the forehead forms a distinct "V" in the center (note that this is distinct from pattern baldness in men). The alternative phenotype is a straight hairline, without the "V". The dominant allele is "widow's peak".

iv) **Dimpled chin** is a distinct depression in the middle of the chin. The alternative phenotype is a smooth chin. The dominant allele is "dimpled".

Your phenotypes:

Earlobes_____ Thumb_____

Hairline_____Chin_____

Calculating class allele frequencies:

Table 1. Phenotype and allele frequencies of four traits for the entire class

	# dominant phenotype	# recessive phenotype	frequency (dominant allele)	frequency (recessive allele)
Ear lobes				
Thumb				
Hairline				
Chin dimples				

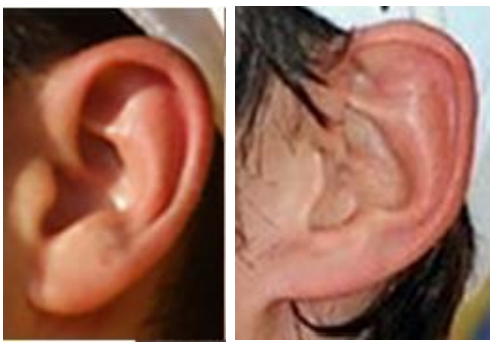
Sample calculation: We cannot know the frequency of the dominant allele (why?) and must therefore calculate it using the Hardy Weinberg equation.

If the number of students with attached earlobes (the recessive allele) is 9 out of 20, then we know that $q^2 = 9/20 = 0.45$ and $q = 0.67$. We can then calculate p (the frequency of the dominant allele) as $1-q = 0.33$.

Note that in some cases, the dominant alleles and their corresponding phenotype are not the most common phenotype in the population. Allelic "dominance" is *not* correlated with population frequency, although this is a common misconception.

Attached earlobe: The myth??

i) **Ear lobes** may be free (detached) or attached to the head. The dominant allele is "detached". \



- Two parents with attached earlobes could they have a child with a free earlobe?

……
……
……
……
……
……
……
……
……
……

Family studies

Lai and Walsh (1966) called earlobes in which the lowest point on the earlobe was the attachment point "attached," and they classified all other earlobes as "free." They recorded the following data on families in New Guinea:

Parents	F offspring	A offspring	Percent F
F x F	12	22	35%
F x A	72	114	39%
A x A	37	90	29%

There are slightly more A offspring from A x A matings, but the large numbers of F offspring from A x A matings and A offspring from F x F matings indicate that this is not a one-locus, two-allele trait.

Mohanraju and Mukherjee (1973) performed a similar study in India and found similar results:

Parents	F offspring	A offspring	Percent F
F x F	13	1	93%
F x A	7	7	50%
A x A	5	29	15%

They found a much stronger association between parents and offspring, but the five F offspring of A x A matings are inconsistent with the myth that this is a one-locus, two-allele trait.

Conclusion:

Earlobes do not fall into two categories, "free" and "attached"; there is continuous variation in attachment point, from up near the ear cartilage to well below the ear. While there is probably some genetic influence on earlobe attachment point, family studies show that it does not fit the simple one-locus, two-allele myth. You should not use earlobe attachment to demonstrate basic genetics.

PART II. HARDY-WEINBERG EQUILIBRIUM

Using colored beads, we will simulate a Hardy-Weinberg equilibrium for one locus with 2 alleles. Each bead represents an allele (r or y) at the "colored bead" gene locus. The imaginary species is **diploid**, so each individual in the population is represented by a *pair* of beads. The three possible genotypes will **be rr, ry, and yy**. Let ***p*** be the frequency of red alleles in the population and ***q*** frequency of yellow alleles.

Geneticists typically call the starting population **Generation 0 (zero)**. Once you have established your initial conditions, you will use **random "matings"** to generate subsequent generations, and then record the genotypic frequencies and allele frequencies for each of these descendent generations. If our simulated population is in Hardy-Weinberg equilibrium, what should happen with allele frequencies over time? Genotype frequencies?

a) **Generation 0 (zero).** We will start with an initial population of 80 individuals using 160 beads. Decide on the initial allele frequencies for your population selecting initial allele frequencies, ***p* & *q***, between 0.3 and 0.7. You can now calculate the initial population's genotype frequencies using the Hardy-Weinberg equation. Then calculate how many beads of each color you will need for your starting population (i.e. multiply *p* x 160 and *q* x 160). Count the right number of yellow and red beads and put them in a container. Record the initial frequencies of genotypes and alleles in the "Generation 0" row of Table 2.

b) The 160 beads in your container represent the gene pool, or the frequencies of the two gamete types in Generation 0. During reproduction, 80 *pairs* of gametes will be drawn from this gene pool to create the next (diploid) generation. We now need to simulate the process of reproduction among the 80 members of our population. In real life, each individual would produce perhaps hundreds or thousands of gametes and mate multiple times (human females produce relatively few gametes. Imagine these are corn or fruit flies). From your container, randomly select 2 beads. Record the genotype, and then return the 2 beads to the container (so that the "parents" can reproduce again). Repeat this process until you have tallied a total of 80 offspring, simulating random mating among your 80 individuals to produce an offspring generation with the same population size. In Table 2, generation 1, record the numbers of each genotype. The number of each allele must be added, and the frequency of each allele calculated.

c) Adjust the allele frequencies in the container to reflect the gene pool you recorded for generation 1. To do so, you need to adjust the number of alleles in your container to match the number of alleles you calculated for generation 1 in table 2. Then, repeat the mating process to produce 80 generation 2 genotypes.

Table 2. Genotypes and allele frequencies by generation for simulation 1: Hardy-Weinberg equilibrium

Generation	Numbers of:						
	Genotypes[1]			Alleles[2]		Frequencies[3]	
	rr	ry	yy	r	y	P	q
0	7	33	40				
1	9	32	39				
2	8	34	38				

1. Recorded during simulated mating.
2. Calculated from genotypes recorded for that population.
3. Calculated from the number of alleles out of total population (%)

Results:

Describe what happened to genotype frequencies as the initial population produced successive generations. Did allele frequencies change?

Since there will always be some variability in the results when you sample a population, we need to use a statistical test to determine if the results we find are statistically significant. In other words, is the variation that we see in our samples the result of random sample error, or the result of a real difference between what our model (H-W) leads us to predict, and reality? Does our model reflect reality well? This is the same process you used in the Mendilian genetics experiments with corn populations (cobs)!

To determine if any deviations you observed are significant, we will use a chi-square test, as in the Mendelian genetics lab, to test allele frequencies.

Table 3. Chi-square test for Hardy-Weinberg equilibrium with no selection

Allele	gen. 0 frequency	Expected freq., gen 2	observed freq. gen 2	O-E	$\frac{(O-E)^2}{E}$
p					
q					
				Sum = X^2 value	

Chi-square calculations:

$$X^2 = \sum \frac{\textbf{(observed count - expected count)}^2}{\textbf{expected count}}$$

Look up the probability for the calculated value of X^2 in table provided at the end of this lab. You should use the p=0.05 level as your critical value.

Did your X^2 value exceed the value at p=0.05 level?

If yes, or no, what does that mean?

PART III. NON-RANDOM MATING

In most plants and animals, non-random mating is the rule rather than the exception. Males and females may choose their mates using a variety of criteria, such as color, body size, territory size, or display behavior. Even when mate choice is not occurring, differential survival of particular phenotypes in particular environments may increase the probability that organisms with similar genotypes mate. Using the beads as our model, we will simulate non random-mating by subdividing the gene pool according to parental genotype/phenotype, and allowing only individuals within each gene pool (those with the same phenotype) to mate with each other.

(a) Start with the same generation 0 population used for **experiment 1**. Assume that the **r** allele is dominant over the **y** allele, so that **rr** and **ry** individuals have the phenotype of **"red"**. Record the starting genotype frequencies and allele frequencies in **Table 4**, generation 0. We will assume that organisms prefer to mate with other organisms of their own phenotype, **something called positive assortative mating**. In nature, such a preference would not be 100 percent effective, but we can simulate perfect compliance by taking all the beads for rr and ry individuals and putting them in one container, and putting all the beans for yy in another. These two containers represent separate gene pools ***within*** the same population.

(b) We do not need to simulate mating in the *yy* container because all the offspring in this case must be *yy*. There will be 20 offspring from the *yy* subpopulation's mating.
The combined results of the rr/ry and yy matings represent the generation 1 offspring for the entire population (be sure to add the results of the two gene pools together). Record these genotype frequencies and allele frequencies in the table 4.

(c) Be sure that the total number of individuals in the population is 80. Repeat the mating simulation with these new gene pools to produce generation 2. Tally the genotypes of the second round of mating and record the combined second-generation results in Table 4.

Table 4. Genotype and allele frequencies by generation for simulation 2: Effects of non-random mating on genotype frequency and allele frequency.

Generation	Numbers of:						
	Genotypes[1]			Alleles[2]		Frequencies[3]	
	rr	ry	yy	r	y	p	q
0	7	33	40				
1	16	16	48				
2	19	9	52				

Results

What happened to the genotype frequencies? Did non-random mating affect the frequency of the alleles?

Was the change in allele frequencies only sampling error, or a real deviation? I.e. are the results significant? To answer this question, compare allele frequencies in generation 2 to those in generation 0 using a Chi-squared test:

……
……
……
……
……
……
……
……
……
……

Table 5. Chi-square test for Hardy-Weinberg equilibrium with non-random mating

Allele	gen. 0 frequency	expected freq., gen 2	observed freq., gen 2	O-E	$\frac{(O-E)^2}{E}$
p					
q					
				Totals = X^2 value	

How do your results relate to the Hardy-Weinberg conditions discussed in the beginning of the lab?

Can you think of any examples of violation of this condition in nature? (Hint: are examples to be found in our own population?)

……
……
……
……
……
……
……
……
……
……

PART IV. GENETIC DRIFT

In the previous exercises we have kept the size of the breeding population constant and fairly high. In nature, reproductive populations may go through periods of greatly reduced numbers. This can occur through natural events, such as dispersal of a few individuals to a new habitat like an island, or disease. These "bottlenecks" can also happen through man-made destruction of natural habitat. In any case, the population is reduced often without any difference among phenotypes in survival (i.e., we are assuming that the mortality is not biased towards certain members of the population). When only a small number of larger population can mate, there is potential for significant **genetic drift** - sharp changes in gene frequency caused by sampling error or the random survival of just a few, randomly selected individuals that likely do not have the same allele frequencies as the initial, large population. We will simulate genetic drift by taking a small subsample of individuals from your population.

Methods:

Return the beads to the starting population used in the first experiment. Pick out four **pairs** of beads *at random* and record their genotypes in Table 6. These pairs represent a pioneer population of four individuals. Replace the beads in the container, mix well, and repeat the process for four more pioneer populations (a total of five pioneer populations). Record the genotypes of the four pioneers in each population in the chart. Then calculate the genotype frequencies and allele frequencies in each of these populations.

Table 6. Genotype and allele frequencies by population for part IV: Effects of genetic drift.

Population	Numbers of:						
	Genotypes			Alleles		Frequencies	
	rr	ry	yy	r	y	*p*	*q*
Initial	7	33	40				
Pioneer 1	1	2	5				
Pioneer 2	6	1	1				
Pioneer 3	4	3	1				
Pioneer 4	1	1	6				
Pioneer 5	2	1	5				

Results:

Compare the frequencies of genotypes and alleles in the pioneer populations to those in the starting population. Genetic drift can cause the **fixation** or **extinction** of alleles in the pioneer population. Fixation occurs when a single allele achieves a frequency of one (100%); extinction occurs when the frequency of an allele drops to zero. Did you observe this in any of your pioneer populations?

How would you test whether significant genetic drift has occurred in each of your new populations?

Chi-Square Probability Distribution

ρ	0.99	0.95	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01
df													
1	0.00	0.004	0.02	0.06	0.15	0.27	0.45	0.71	1.07	1.64	2.71	3.84	6.63
2	0.02	0.10	0.21	0.45	0.71	1.02	1.39	1.83	2.41	3.22	4.61	5.99	9.21
3	0.11	0.35	0.58	1.01	1.42	1.87	2.37	2.95	3.66	4.64	6.25	7.81	11.34
4	0.30	0.71	1.06	1.65	2.19	2.75	3.36	4.04	4.88	5.99	7.78	9.49	13.28
5	0.55	1.15	1.61	2.34	3.00	3.66	4.35	5.13	6.06	7.29	9.24	11.07	15.09

……
……
……
……
……
……
……
……
……
……
……
……

PART V Migration

Migration into or out of a population is the fourth and final factor that can affect its genetic composition. Obviously, if immigrants are genetically different from the population they are entering, this will cause the population's genetic composition to be altered. The evolutionary importance of migration stems from the fact that many species are composed of a number of distinct subpopulations, largely isolated from each other but connected by occasional migration. (For an extreme example of population subdivision, think of ant colonies.) Migration between subpopulations gives rise to gene flow, which acts as a sort of 'glue', limiting the extent to which subpopulations can diverge from each other genetically.

The simplest model for analysing migration assumes that a given population receives a number of migrants each generation, but sends out no emigrants. Suppose the frequency of the A_1 allele in the resident population is p, and the frequency of the A_1 allele among the migrants arriving in the population is p_m. The proportion of migrants coming into the population each generation is m (i.e. as a proportion of the resident population.) So post-migration, the frequency of the A_1 allele in the population is:

$$p' = (1 - m)\, p + m p_m$$

The change in gene frequency across generations is therefore:

$$\Delta p = p' - p$$

$$= - m\,(p - p_m)$$

Therefore, migration will increase the frequency of the A_1 allele if $p_m > p$, decrease its frequency if $p > p_m$, and leave its frequency unchanged if $p = p_m$. It is then a straightforward matter to derive an equation giving the gene frequency in generation t as a function of its initial frequency and the rate of migration. The equation is:

$$p_t = p_m + (p_0 - p_m)(1 - m)^t$$

where p_0 is the initial frequency of the A_1 allele in the population, i.e. before any migration has taken place. Since the expression $(1 - m)^t$ tends towards zero as t grows large, it is easy to see that equilibrium is reached when $p_t = p_m$, i.e. when the gene frequency of the migrants equals the gene frequency of the resident population.

This simple model assumes that migration is the only factor affecting gene frequency at the locus, but this is unlikely to be the case. So it is necessary to consider how migration will interact with selection, drift and mutation. A balance between migration and selection can lead to the maintenance of a deleterious allele in a population, in a manner closely analogous to mutation-selection balance, discussed above. The interaction between migration and drift is especially interesting. We have seen that drift can lead the separate subpopulations of a species to diverge genetically. Migration opposes this trend—it is a homogenising force that tends to make subpopulations more alike. Mathematical models suggest that that even a fairly small rate of migration will be sufficient to prevent the subpopulations of a species from diverging genetically. Some theorists have used this to argue against the evolutionary importance of group selection, on the grounds that genetic differences between groups, which are essential for group selection to operate, are unlikely to persist in the face of migration.

PART VI Mutation

Consider an allele couple A_1 et A_2 of frequencies p_t and q_t respectively in generation t; consider u as the direct mutation rate of A_1 towards A_2 in each generation and v the reverse mutation rate of A_2towards A_1. We define the mutation rate as the probability for a mutation to appear per gamete and per generation.

For example, suppose a population only composed of individuals of the genotype A_1 A_1 which contribute to the production of 2N gametes in the next generation.

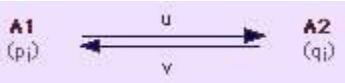
If u represents the mutation rate of A_1 locus towards all the other possible alleles in a diploid population of N individuals, as we have 2N concerned genes, the number of new mutants that we get are 2N x u, in a given generation. In the population, on 2N genes, only one gene among them (1/2N) will be fixed. Thus, the probability for a mutant allele to appear in the generation with a fixed gene, is 2N x u x 1/2N = **u**.

Therefore, the probability for one allele to be substitute by a muted allele is equal to the mutation rate, per gamete and per generation, in a given locus.

---

a. Recurrence law

In the presence of mutations, the evolution of allele frequencies depends on p_i and q_i frequencies in generation i, but also on the mutation rates of u and v.



The frequency of an allele A_1, one generation after mutation ($p_i + 1$), corresponds to the frequency of an allele A_1 in the preceding generation (frequency p_i) which would not have muted (probability equal to 1-u), or to an allele A_2 (frequency q_i) which would have muted in A_1 (probability v).

$$\mathbf{p_{i+1} = (1-u)\ p_i + v\ q_i}$$

$$\mathbf{q_{i+1} = (1-v)\ q_i + u\ p_i}$$

Example: The peppered moth (with the kind authorisation of Pr. Georges Périquet, University of Tours)

The peppered moth, *Biston betularia* (Order: Lepidoptera) is common in Northern Europe. The individuals fly during the night and in the day lay down against the light-coloured bark of trees. Since the 19th century, this species has been often studied; it is represented by two morphs, one **light**-coloured (*typica*) and another dark-coloured, or **melanic** (*carbonaria*).

- **Genetic determinism and allele frequencies**

The genetic determinism of this coloration is monogenetic and autosomic. The allele *carbonaria* (C) is dominant on the allele *typica* (c). The allele frequencies C and c will be named p and q respectively.

Phenotypes	**[Melanic form]**	**[Light form]**
Genotypes	CC ou Cc	cc

In the middle of the 19th century, among the British populations, the *typica* form was largely majority. It is only in 1848 that an individual *carbonaria* capture had been reported in Manchester area. This form frequency had strongly increased and other melanic individuals, afterwards, had been captured in others industrial areas of England. The frequency increase of the *carbonaria* form was very rapid in Manchester region; it reached 98 %, in 1895. In less than 50 years, this form had become the great majority into the region. As the variations always occurred in the way of a *carbonaria* form increase of frequency, it could not be due to a random phenomenon.

- **Two hypothesis**

Two hypothesis are possible; the effect of recurrent mutations, or the action by selection.
The hypothesis of recurrent mutations:

We could think that the *carbonaria* allele was obtained frequently enough by mutation and thus replace the *typica* allele. Therefore, we can calculate the mutation rate v from c towards C, which would corresponds to an evolution as rapid as the one observed in the nature. Under mutations effect alone, the recurrence relation for the allele frequencies c, between two successive generations, can be written as:

$$q_1 = q_0 - vq_0$$

The c allele frequency will be the same as the preceding generation (q_0) reduced by the muted allele frequency (vq_0). Here, We consider that the allele C does not mute in c, which correspond to the most rapid possible evolution under the mutation effect.
As the process is repeated with the course of the generations, we deduce:

$$q_t = q_0 (1-v)^n$$

and

$$v = 1 - (q_t/q_0)^{1/n}$$

Where n represents the generation number.

If we consider that the melanic form frequency, in the Manchester region, was 0,01 in 1848 and 0.98 in 1895, and knowing that the peppered moth gives one generation per year in this region, it is possible, therefore, to calculate the v value which better express the rapidity of the observed evolution.

Thus, in 47 generations, we obtain:

$$\mathbf{v = 1 - (0,141/0,995)^{1/47} = 0,041.}$$

Then, the calculate value obtained was about 4 %. That means, to obtain a rapid evolution, as fast as what we can observe in the nature, about 4 gametes per 100 at each generation should be muted from *typica* to *carbonaria*. In fact, it is an incompatible rate mutation in respect with the spontaneous mutation rate observed in reality (about 10^{-5} to 10^{-6}). Furthermore, we have to explain the brutal change of mutation rate since 1848!
If the rate of 4 5.10^{-2} really exists, it could be explained by the existence of Transposons (doubtful) or by the effect of other factors of evolution such as the selection.

- **Equilibrium state**

The equilibrium state in which performed direct and reverse mutations.

$$\mathbf{q_1 = (1-v)\, q_0 + u\, p_0}$$

$$\mathbf{q_1 = q_0 - v\, q_0 + u\, p_0}$$

Δ_q = the difference between 2 successive generations

$$\mathbf{\Delta_q = q_{n+1} - q_n}$$

Following the sign Δ_q we know if:

$$\mathbf{\Delta_q > 0 \text{ ----> } q_{n+1} > q_n \text{ ----> } A_2 \text{ increase}}$$

$$\mathbf{\Delta_q < 0 \text{ ----> } q_{n+1} < q_n \text{ ----> } A_2 \text{ decrease}}$$

$$\mathbf{\Delta_q = 0 \text{ ----> } q_{n+1} = q_n = q_e}$$ **Equilibrium state**

If we pose at this level of equilibrium state

$$\mathbf{q_e = q_e - v\, q_e + u\, p_e}$$

$\mathbf{u\ p_e = v\ q_e}$

$\mathbf{u\ p_e = v\ (1 - p_e)}$

$\mathbf{u\ p_e + v\ p_e = v}$

$\mathbf{p_e = v / (u + v)}$

and

$\mathbf{q_e = u / (u + v)}$

The equilibrium state is function of the mutation rate u and v.

$q_e \varepsilon]\ 0\ ,\ 1\ [$

Example:

If u = v ---> q_e = 1/2

If u = 10^{-5} and v = 10^{-6} ---> q_e = 0.91

What ever the mutation rates are, we can have any equilibrium state.

To pass of $q_0 = 0.51$ to $q_n = 0.71$ and $q_e = 0.91$

Mutations can produce equilibrium states, but they will be reached very slowly through the time. The mutation process does not have major effect on the genetic structure of populations; the variation of allele frequencies is very low through the time. This evolution of allele frequencies varies according the equation:

$\mathbf{p_{n+1} = p_e + (p_0 - p_e)(1-u-v)^n}$

$\mathbf{q_{n+1} = q_e + (q_0 - q_e)(1-u-v)^n}$

PART VII. Selection at One Locus

Ex: A population is examined for a trait that is codominantly expressed. It is found that there are 785 individuals who exhibit the genotype A_1A_1, 20 with the A_1A_2 genotype and 195 with the A_2A_2 genotype. Are these two alleles in Hardy-Weinberg equilibrium? **Show all work** including X^2 test (X^2 table on back of this sheet).

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……..

……

……

……

……

……

PART VIII Simulations

1. Simulating with PopG PopGen Fishpond

This program simulates the evolution of random-mating populations with two alleles, arbitrary fitnesses of the three genotypes, an arbitrary mutation rate, an arbitrary rate of migration between the replicate populations, and finite population size.

The programs simulate simultaneously evolving populations with you specifying the population size, the fitnesses of the three genotypes, the mutation rates in both directions (from *A* to *a* and from *a* to *A*), and the initial gene frequency. They also ask for a migration rate among all the populations, which will make their gene frequencies more similar to each other. Much of the time (but not always!) you will want to set this migration rate to zero. In most respects the program is self-explanatory.

Initially there are ten populations. You can set the number of simultaneously-evolving populations to any number from 0 to 1000. The population size for each population can be any number from 1 to 10000. Note that a larger population, a larger number of generations run, and a larger number of populations can lead to longer runs.

When you make a menu selection that causes the program to run, a graph of the gene frequencies of the *A* allele in each of the populations will be drawn in the window.

The program can simulate a wide variety of cases, and you should explore some of these. Here are some suggestions:

- Try cases with no mutation, no migration, and all fitnesses 1.0 so that there is no selection. Does genetic drift in a population of size 1000 accomplish roughly the same changes in 1000 generations as genetic drift in a population of size 100 does in 100 generations? By running a largish number of populations, can you check whether the probability that an allele is fixed by genetic drift is equal to its initial frequency in the populations?
-
- Try a case with no mutation or migration, with the *A* allele favored by natural selection (with fitness of the *AA* genotype set highest and fitness of the *aa*genotype set lowest). Start with a small frequency of *A*. Is it always fixed? If one starts with a single copy of the allele, how does the probability that *A* is fixed compare with the selection coefficient favoring it in the heterozygote (the fraction by which the *Aa* genotype is higher compared to the fitness of the *aa* genotype)? Is this fixation probability larger than the one you would get with the same initial frequency but with no selection?

2. Populus Genetic Drift Computer Simulation

The Monte Carlo model simulates genetic drift using a random number generator to sample genes from a small parental population and passes them on to offspring. Population size is assumed to be constant from generation to generation and gene frequency changes the result only from the random sampling process.

- **Exercise Instructions**

Click on the Populus "P" icon to start

In the upper left corner, click on "Model"

From the pull-down menu choose "Genetic Drift Model" and then from that menu choose "Genetic Drift"

- **Experiment 1. Population Size**

Goal: To investigate the effect of population size on the loss of polymorphism.

This program allows you to change the population size and initial gene frequencies of six unlinked loci.

After you run the program you will see a graphical display of the change in frequency of allele A over time in generations. Each different colored line is a separate locus. Fixation of allele A occurs when its frequency reaches 1.0. Fixation for allele a occurs when the frequency of A is 0.

Run six unlinked loci simultaneously (collectively), each with initial gene frequencies of 0.5. To run the program and view the result graph, click "VIEW". Explore the effect of changing the population size by running this program for 3 different population sizes between 2 and 200 (you choose the three sizes). Run the simulations for 10x (the population size) generations (e.g. if your population size is 30, run the simulation for 300 generations). To change the number of generations click the circle marked "Other" under "Runtime" and you can specify an alternate value for the number of generations you would like to run the simulations.

- EACH TIME YOU RUN THE SIMULATION DO THE FOLLOWING

Record the number of fixed loci for A and for a as well as the number of loci which remain polymorphic.
Approximate the number of generations until fixation for each population you explore

REPEAT the simulation 5 times for a total of 30 replicates for each population size

---

- **Experiment 2. Initial Allele Frequency**
- **Goal: to explore the effect of initial allele frequency on fixation of alleles**

Run six unlinked loci simultaneously (collectively), each with an initial population size of 10. Run the simulations for 100 generations.

Explore the effect of changing the initial frequency by running this program for 4 different frequencies of allele A between 0 and 1. You will need to "SET FREQUENCIES COLLECTIVELY"
EACH TIME YOU RUN THE SIMULATION record the total number of loci fixed for allele A and allele a.

REPEAT the simulation 5 times for a total of 30 replicated for each initial allele frequency.

Calculate the proportion of loci fixed for allele A and a for each initial frequency.

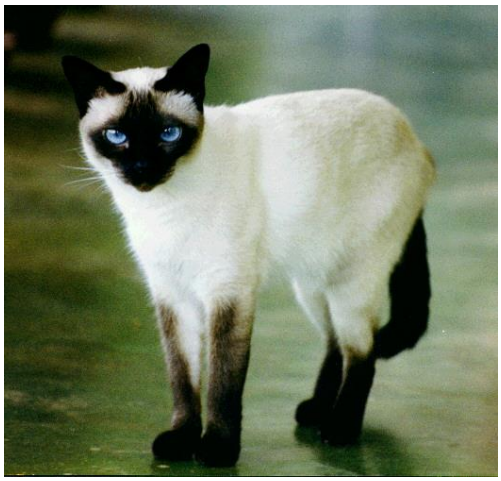
B. Quantitative Genetics (Genetics of complex traits)

PHENOTYPES ARE NOT ALWAYS A DIRECT REFLECTION OF GENOTYPES

Some alleles are only expressed in some environments, or have variable expression for other reasons.

Mendelian Traits that are also affected by the Environment

The effects of these mutations are usually only apparent at high temperatures. Siamese cats have a mutation in the C gene controlling dark pigment formation. The C^h allele of this gene is heat sensitive. The C^h allele can make dark pigment at low but not high temperatures. The permissive temperature for dark-color occurs at the extremities, and a restrictive temperature occurs in the body core.



Nutritional effects

Phenylketonuria is a human nutritional defect that can lead to severe physical and mental disorders in children, but only if they consume phenylalanine. The mutation prevents individuals from metabolizing this amino acid. The disease phenotype can be avoided by eliminating phenylalanine from the diet.

I. QUANTITATIVE TRAITS

Most phenotypic traits in plants and animals are affected by many genes (size, weight, shape, lifespan, physiological traits, and fecundity). Often, it is not feasible to determine the number of genes affecting a particular trait, and the individual effects of genes on the phenotype. Many of these traits can be measured on a quantitative, rather than a qualitative, scale. This is where the terms quantitative trait and quantitative genetics come from.

Quantitative traits:

1. Have continuous distributions, not discrete classes
2. Are usually affected by many genes (polygenic)
3. Are also affected by environmental factors

EXAMPLE: COLOR OF WHEAT KERNELS

This trait is determined by two genes that contribute "doses" of red pigment, and display partial dominance (heterozygotes intermediate). Each allele with the subscript "1" contributes 1 dose of red pigment. This trait demonstrates additive effects among different alleles at a single locus, and among alleles at different genetic loci. Genes with subscript "2 " don't contribute any red pigment.

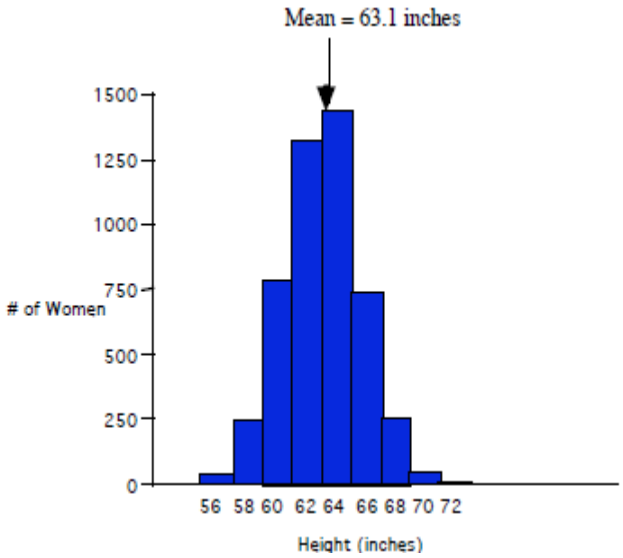
With two additive genes you have five phenotypic classes in the F2 offspring of true-breeding strains, instead of the three classes you would see if you had only one additive gene. Some classes are composed of several genotypes that are indistinguishable in phenotype. **The number of phenotypic classes expected when there are n diallelic additive loci is (2n+1)**. So, if have four additive genes, you should expect nine phenotypic classes in the F2 offspring of pure-breeding strains.

As the number of additive genes increases, the distribution of phenotypes becomes more continuous. In addition, as stated above, most quantitative traits are also affected by the environment. Environmental effects may obscure genetically-caused differences between phenotypic classes. For example, nutrition affects the adult size in many organisms. The distribution of phenotypes then becomes even more continuous. The distribution of quantitative traits often approximates a bell-shaped curve when you plot phenotypic value (height, for example) against the frequency of individuals in particular phenotypic classes. Such a plot is called a frequency histogram.

EXAMPLE: DISTRIBUTION OF HEIGHT IN 5000 BRITISH WOMEN:

In this graph, the column designated "62" includes all individuals with heights between 61 and 63 inches, "64" includes all individuals with heights between 63 and 65 inches, and so on.

This curve has two easily measured properties--the **MEAN (average)**, and the **VARIANCE** (variation



about the mean). Curves with the same mean may have very different variances.

If you were to measure the heights of a sample of about 5000 different British women, you would get a curve similar to the one shown above. To calculate the mean of a distribution, you need to sum up all the different heights, then divide by the number of individuals:

mean = average = sum of heights / number of women:

$x = \sum (xi)/N = 63.1$ inches. Another way of calculating the mean directly from a histogram (if you don't have available a list of the heights of all 5000 women, for example) is

$$\bar{x} = \sum_{i=1}^{n=\#classes} f_i x_i \Big/ \sum_{i}^{n} f_i$$

The variance is calculated from the sum of squared deviations of individuals from the mean:

$$Var = \Sigma (x_i - x)^2/(N-1) = 7.24\ in^2 \quad or$$

$$Var = \sum_{i=1}^{n} f_i (x_i - \bar{x})^2 / (N-1)$$

where N = total number of individuals (5000). We will use this concept of variance extensively in our discussion of quantitative genetics.

Quantitative traits are influenced by genetics and by the environment. Under some circumstances, we can partition the phenotypic variance in quantitative traits into variance that is associated with genetic effects, and variance that is associated with environmental effects:

VP = VG + VE

Where VP is the total phenotypic variance, VG is the genetic variance, and VE is the environmental variance.

II. MODELS OF QUANTITATIVE INHERITANCE

SOURCES OF VARIANCE

In this case, all the phenotypic variance is due to differences in genotypes, so it is all genetic variance. However, some of the difference between the genotypes is due to additive effects of alleles, and some are due to dominance effects of alleles. For example, the difference in phenotype between aabb, aaBb, and aaBB (16, 17, and 18 respectively) are only due to the additive interactions between different alleles at the B locus. However, the difference in phenotype between AABB, AaBB, and aaBB (20, 20, and 18, respectively) is due to both additive effects (difference between aa and AA is 2 "units" of the A allele and the difference in height is 2 inches--so the average effect of a "unit" of A is one inch) and dominance effects at the A locus. Therefore, some of the genetic variation is due to additive effects and some is due to dominance effects. Therefore, some of the genetic variance in the F2 is additive variance, and some is dominance variance. In fact, although the derivation of the equation is beyond the scope of this course, **the amount of additive (VA) vs. dominance variance (VD)** is easily calculated in this example: **VA**= 1.0 and **VD** = 0.333; **($VA=2\Sigma p_i q_i a_i$ 2; $VD=\Sigma(2pqd^2)$, where the sum is over each locus contributing to the trait)**. In this simplified example, there are no environmental effects, so the environmental variance (VE) is zero. If there were environmental effects, the phenotypic variance would be the sum of the genetic and environmental variance. The additive and non-additive genetic variance (VA and VD) together are known as the genotypic or genetic variance, which is abbreviated VG. In the bird-wing example, VG = 1.333, which is also the total phenotypic variance is abbreviated **VP = VG + VE.**

These considerations (and the mathematical fact that variances due to independent sources of variation can be summed) lead to the following equations:

VP = VG + VE

VP = VA + VD + VE

In the bird-wing example, all the phenotypic variance was due to genetic variance (there were no environmental effects on the trait). We can also imagine a trait that has no genetic variance, so that the differences between individuals are due only to environmental effects. For example, assume that you grow a clonal strain of corn in a field, so that every individual has exactly the same genotype. At the end of the growing season, there will be differences in height (VP) among different plants, but all the differences will be due to local soil, moisture, temperature, and light conditions. Thus, all the height differences are due to environmental variance (VE).

A useful thing for plant and animal breeders to know is, for any trait of interest, how much of the phenotypic variability of that trait is due to genetic variance, and how much is due to non-genetic environmental factors.

This is the broad-sense heritability: $\mathbf{H^2 = VG/VP}$

It is even more useful to know what proportion of the phenotypic variation is due to additive genetic effects.

The heritability (narrow-sense) of a trait is defined as the proportion of the total phenotypic variation that is due to heritable (additive genetic) effects:

$\mathbf{VA/VP = h^2}$

h2 is the proportion of variability that can be passed on from parent to offspring, and this is why this quantity is of interest to animal and plant breeders. They want to know whether selection on the parents will produce inherited changes in the offspring. The degree to which selection on parents will produce inherited changes in the offspring is determined by the narrow sense heritability of the trait.

When h2 = 0, none of the phenotypic variance among individuals is due to additive genetic differences (VA=0) and offspring will not closely resemble their parents for the trait of interest for genetic reasons.

Can you think of any traits in humans, other animals, or plants that probably have zero heritability?……

……

……

……

……

When h2 = 1, all the variation among individuals is due to heritable genetic differences (VP=VA) and offspring will resemble their parents very closely.

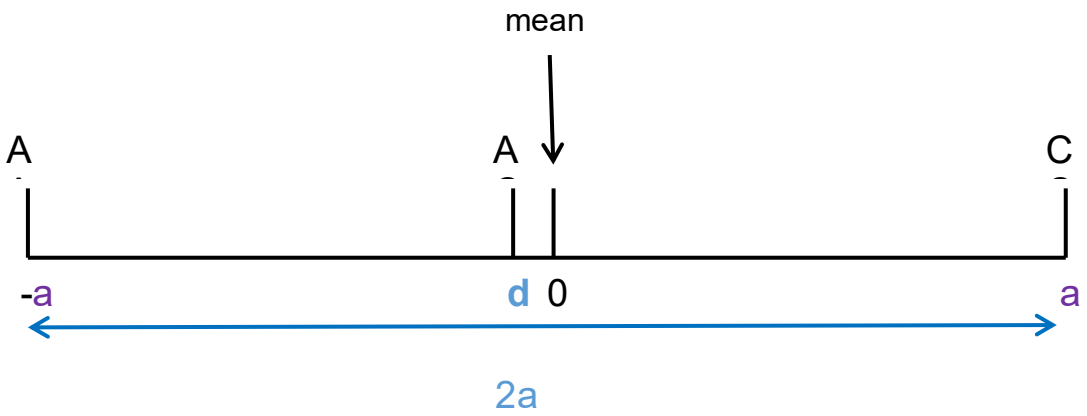
Can you think of any traits that probably have high heritability?

……

……

……

……



1. Calculate genotypic values (a and d)
2. Calculate the average effect of the alleles
3. Calculate the genotype frequencies
4. Calculate the mean IL6-R concentration in the population
5. Calculate how much of the variance is explained by this SNP
   (*Variance= Sum of squared deviations from the mean*)
6. Calculate heritability

.....

.....

.....

.....

.....

.....

.....

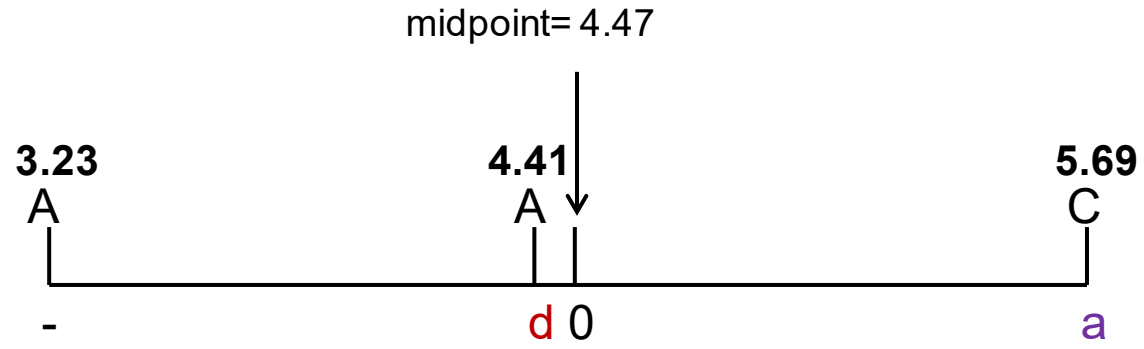
.....

.....

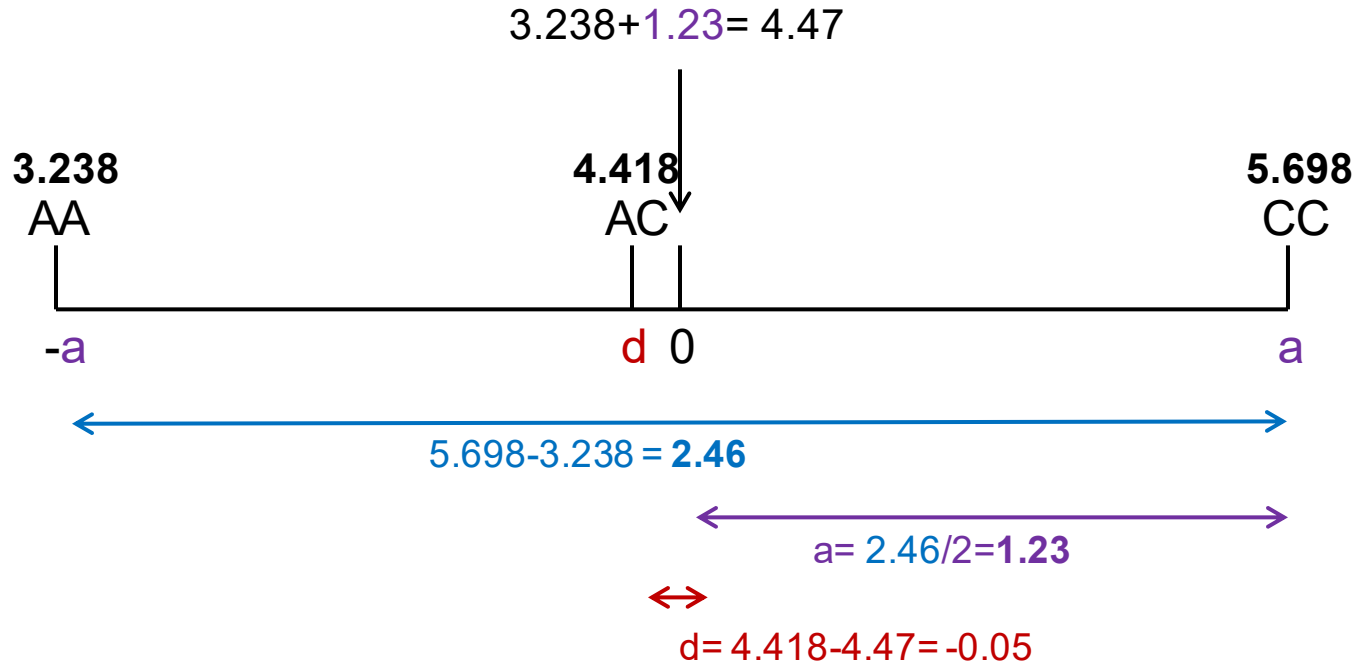
.....

Response

1. **QUESTION Calculate genotypic values (a and d)**



Midpoint = (3.238 + 5.698)/2 = 4.47



$a = 0.5*(sIL6R_{CC} - sIL6R_{AA}) = 0.5*(5.698-3.238) = 1.23$

$d = sIL6R_{AC} - (sIL6R_{AA} + a) = 4.418 - (3.238 + 1.23) = -0.05$

---

2. **Calculate the average effect of the two alleles**

Minor allele: C, frequency: p= 0.39

Major Allele: A, frequency: q=0.61

(Allele frequencies from Sib-pair (http://genepi.qimr.edu.au/staff/davidD/#sib-pair) using the best linear unbiased estimator (BLUE) option, which accounts for familial relatedness)

a=1.23, d= -0.05

Average effect C= q[a+d(q-p)] = 0.61[1.23-0.05(0.61-0.39)] **= 0.74**

Average effect A= -p[a+d(q-p)]= -0.39[1.23-0.05(0.61-0.39)] **= -0.48**

3. **Calculate the genotype frequencies**
4. **Calculate the mean IL6-R concentration in the population**
5. **Calculate how much of the variance is explained by this SNP**

Minor allele: C, frequency: p=0.39

Major Allele: A, frequency: q =0.61

	CC		**CA**		**AA**
Genotype frequencies	$= p^2$	+	2pq	+	q^2
	$= 0.39^2$	+	2*0.39*0.61	+	0.61^2
	= **0.15**	+	**0.48**	+	**0.37**
mean IL6R	5.698		4.418		3.238

Population mean $p^2 * 5.698$ + 2pq * 4.418 + q^2 * 3.238

0.15 * 5.698 + 0.48 * 4.418 + 0.37 * 3.238 **= 4.17**

Deviation from mean	= 5.698-4.17	4.418 - 4.17	3.238 - 4.17
	= 1.52	0.24	-0.94
Squared deviation	$= 1.52^2 = 2.32$	$0.24^2 = 0.06$	$-0.94^2 = 0.88$

	CC	**CA**	**AA**
Squared deviation	$= 1.52^2 = 2.32$	$0.24^2 = 0.06$	$-0.94^2 = 0.88$

Variance (SNP) $= p^2 * 2.32$ + $2pq * 0.06$ + $q^2 * 0.88$

= 0.15* 2.32 + 0.48* 0.06 + 0.37*0.88 = **0.71** (SD = 0.84)

Alternative way of estimating the mean: **Mean = a(p-q) + 2pqd**

a = 1.23 and d= -0.05

a(p-q) = 1.23(0.39 - 0.61) = 1.23 *(-0.22) = -0.27

2pqd = 2*0.39*0.61*-0.05 = -0.012

This value of the mean, however, is measured from the mid-homozygote point, which was 4.47 for these data.

Thus: (-0.27 – 0.012) + 4.47 = 4.19

Alternative way of estimating genetic variance (due to this SNP)

$a = 0.5*(sIL6R_{CC} - sIL6R_{AA}) = 1.23$

$d = sIL6R_{AC} - (sIL\text{-}6R_{AA} + a) = -0.05$

$V_A = 2pq\,[a + d\,(q-p)]^2 =$

$2*0.39*0.61[1.23-0.05*(0.61-0.39)]^2 = 0.71$

$V_D = (2pqd)^2 =$

$(2*0.39*0.61*-0.05)^2 = 5.66 \times 10^{-4}$

Calculate how much of the total variance is explained by this SNP

Total variance of IL-6R concentration= 1.35

% variance explained by SNP= 0.71/1.35 = 53%

Measures of central tendency and dispersion

Biological phenomena can often be described by a **measure of central tendency** (the average, or arithmetic **mean**), and a **measure of dispersion** (**variance**). To express dispersion in terms of **magnitude** without regard to sign, the difference from the mean is **squared**. To express dispersion in the **same units** as the mean, the square root of the variance is the **standard deviation**.

Mean = sum of ***i*** individual values of variable **X**, divided by number of individuals ***N***

$= \Sigma (x_i) / N = \bar{x}$ [read as, "***X bar***"]

$= (x_1 + x_2 + x_3 + \ldots x_n) / N$

Variance = average squared deviation of individuals from the mean

$= (1 / N) \Sigma (x_i - \bar{x})^2 = \sigma^2$ [read as, "***sigma squared***"]

computationally, this is more easily calculated as

$= (1 / N) \Sigma (x_i^2) - \bar{x}^2$

which formula can be remembered as

= "***mean of squares***" minus "***square of means***" [MOSSOM]

Standard deviation = square root of variance

$s = (\sigma^2)^{1/2}$

or more strictly, for a finite sample size **N**

$s = [(N) / (N-1)] (\sigma)^{1/2}$

III. QUANTITATIVE GENETICS PROBLEMS

1. Suppose that in a population of Peacocks the phenotypic variance for tail length is 2.5 and the slope of the father – offspring regression for this trait is 0.2. From a long-term captive population you also have data from a line of completely inbred individuals. In this line the phenotypic variance among individuals is 0.50. Assume that there are no shared environmental effects (Ves) and no epistatic variance (VI) for this trait. (Note that these questions are not given in the order that you need to solve them)

a) What is the total genetic variance for tail length?

……

……

……

……

b) What is the additive genetic variance?

……

……

……

……

c) What is the dominance genetic variance?

……

……

……

……

d) What is the environmental variance?

……

……

……

……

……

……

---

e) What is the narrow-sense heritability (h2)?

……

……

……

……

f) What is the expected phenotypic covariance among full-sibs?

……

……

……

……

2. wo inbred lines of beans are intercrossed. In the F_1, the variance in bean weight is measured at 1.5. The F_1 is selfed; in the F_2, the variance in bean weight is 6.1. Estimate the broad heritability of bean weight in the F_2 population of this experiment.

……

……

……

……

……

……

……

……

♦ Solution ♦. The key here is to recognize that all the variance in the F_1 population must be environmental because all individuals must be of identical genotype. Furthermore, the F_2 variance must be a combination of environmental and genetic components, because all the genes that are heterozygous in the F_1 will segregate in the F_2 to give an array of different genotypes that relate to bean weight. Hence, we can estimate

$$s_e^2 = 1.5$$
$$s_e^2 + s_g^2 = 6.1$$

Therefore

$$s_g^2 = 6.1 - 1.5 = 4.6$$

and broad heritability is

$$H^2 = \frac{4.6}{6.1} = 0.75\ (75\%)$$

3. In an experimental population of *Tribolium* (flour beetles), the body length shows a continuous distribution with a mean of 6 mm. A group of males and females with body lengths of 9 mm are removed and interbred. The body lengths of their offspring average 7.2 mm. From these data, calculate the heritability in the narrow sense for body length in this population.

……
……
……
……
……
……
……
……

♦ Solution ♦. The selection differential is 9 − 6 = 3 mm, and the selection response is 7.2 − 6 = 1.2 mm. Therefore, the heritability in the narrow sense is:

$$h^2 = \frac{1.2}{3} = 0.4\ (40\%)$$

Questions:

1. A botanist studying a plant with geographically disjunct and morphologically distinct populations wanted to know the degree to which the differences between populations was attributable to genetic factors. Outline a simple experiment she can do to address this question.

2. If you measured the narrow-sense heritability of a crop plant trait at 0.67 , what would that value mean in terms of the ease of selecting the trait in a selective breeding program. Explain your answer.

3. Broad-sense heritability of a trait is estimated at 0.8. If the total phenotypic variance is estimated at 32.20, what is the genetic variance?

4. In a controlled cross for a quantitative trait, 600 F2 offspring are obtained, 10 of which show the phenotype of one of the parentals. How many genes are estimated to control this trait based on these results?

5. The incidence of obesity in the United States has doubled over the past couple of decades. Is this due more to genetic or environmental factors? Cite a reason for your answer.

6. Pigmentation in the imaginary river-bottom dweller *Mucus yuccas* is a quantitative character controlled by a set of five independently segregating polygenes with two alleles each: *A/a, B/b, C/c, D.d,* and *E/e*. Pigment is deposited at three different levels, depending on the threshold of gene products produced by a capital-letter alleles. Greyish brown pigmentation is seen if at least four capital-letters alleles are present, light tan pigmentation is seen if two or three capital-letter alleles are present, and whitish-blue pigmentation is seen if these thresholds are not met. If an *AA, BB, CC, DD, EE* animal is crossed to an *aa, bb, cc, dd, ee* animal and the progeny are intercrossed, what kinds of phenotypes are expected in the F1 and F2?

7. Two inbred lines of beans are intercrossed. In the F1, the variance in bean weight is measured at 1.5. The F1 is selfed; in the F2, the variance in bean wight is 6.1. Estimate the broad sense heritability of bean weight in the F2 population of this experiment.

8. In a large herd of cattle, three different characters showing continuous distribution are measured and the variance in the table are calculated:

	Characters		
Variance	**Shank length**	**Neck length**	**Fat content**
Phenotypic variance	310.2	730.4	106.0
Environmental variance	248.1	292.2	53.0
Additive generic variance	46.5	73.0	42.4
Dominance genetic variance	15.6	365.2	10.6

a. Calculate the broad- and narrow sense heritabilities for each character.

b. In the population of animals studies, which character would respond best to selection? Why?

c. A project is undertaken to decrease mean fat content in the herd. The mean fat content is currently 10.5 percent. Animals of 6.5 % fat content are interbred as parents in the next generation. What mean fat content can be expected in the descendents of these animals?

Answers to above questions

1. Expose the different populations to similar environment. Any differences between the populations will be genetic.

2. Trait is easlily selected for because it is polygenic and encoded by additive genes. During selection, aim is to maximize the number of additive genes.

3. Genetic variance is 25.76.

4. $1/(4^n)$ = fraction of either parental phenotype in F2 = 10/600; looks like 1/64, therefore n=3,

5. It is environmental. Cannot be genetic because it would imply changes in genetic structure of a large population whiare are too fast.

6. All of F1 will be grey; Most F2 will be grey, many will be light tam, and very few white-blue.

7. Variance in F1 is all environmental; variance in F2 is environmental and genetic. Therefore Ve = 1.5, Vp = 6.1, therefore Vg = 6.1-1.5=4.6).

8. (a) shank length H^2= 0.2, h^2= 0.15; Neck length H^2=0.6, h^2+0.1; Fat content H^2=0.5, h^2+0.4 (b) Selection for back fat would be most promisisng. Most variation comming from additive effects of polygenes. (c) 8.9%.

MCQ

1. Both Hinor's mother and father are homozygous for a recessive gene associated with obesity. There is

a. a 75% chance that Hinor will be obese.
b. a 100% chance that Hinor will be obese.
c. a chance that Hinor will be obese.
d. no chance that Hinor will be obese.

2. Two parents have a child that has sickle-cell anemia. Which of the following is true?

a. Both parents must be homozygous for the sickle-cell allele.
b. One parent must be homozygous for the sickle-cell allele, but the other can be homozygous for the normal allele.
c. Both parents must have at least one copy of the sickle-cell allele.
d. One parent must be homozygous for the sickle-cell allele, but the other can be heterozygous.

3. Why are blood cells different than skin cells?

A. Blood cells have totally different DNA than skin cells.
B. Blood and skin cells have mostly the same DNA, with a few differing genes.
C. The same genes are active in blood and skin cells, but they make different proteins.
D. Blood and skin cells have the same DNA, but different genes are active.
E. Any of the above could be true, depending on the specific kind of cell.

4. The inheritance of all quantitative traits

A. Can be easily predicted given the parents' genotypes.
B. Is impossible to predict, because such traits are totally environmentally determined.
C. Can be accurately predicted given both genetic and environmental information.
D. May be very difficult to predict, because they contain random, genetic, and environmental components.

5. Which of the following is true of studies comparing identical to non-identical twins?

A. They attempt to reveal the role of genetic similarity when 2 people are raised in the same environments.
B. They attempt to reveal the role of environmental differences between 2 people that are genetically identical.
C. The environment that 2 non-identical twins are raised in is as similar as that which 2 identical twins are raised in.
D. All of the above.

6. Which of the following statements is FALSE?

a. Different types of cells in the body contain completely different combinations of genes.
b. Genes may be active in some cells of the body, but not in others.
c. Cells may have different genes active at different times.
d. Two very different cells in the body contain the same genes.

7. C is dominant for curly hair, and c is recessive for straight hair. What is the genotype of a person with straight hair?

a. CC
b. c
c. cc
d. Cc
e. C

8. True/False: Individuals who are carriers of genetic diseases usually have homozygous genotypes for the trait.

9. A woman who is heterozygous for sickle-cell anemia marries a man with the same genotype. Use a Punnett square to predict the possible genotypes and associated percentages of their children.

a. 25% homozygous unaffected, 50% heterozygous, 25% homozygous for anemia
b. 50% homozygous unaffected, 50% homozygous for anemia
c. 100% heterozygous
d. 50% heterozygous, 50% homozygous for anemia
e. none of the above

10. True/False: Between identical twins, any measurable differences in IQ are due to environmental as opposed to genetic factors.

11. A nutritional researcher finds a high similarity in total body fat between parents and children. Why is this not conclusive evidence of high heritability?

a. Parents and children also share the same environments.
b. Children are at different stages of their life.
c. Similarity has nothing to do with determining heritability.
d. Children are never genetically identical to their parents.

12. A heterozygous genotype is different than homozygous in that it contains

a. 2 different genes
b. 2 different alleles
c. 2 homologous chromosomes
d. 2 sister chromatids

13. Albinism is a recessive trait. If a carrier for albinism has kids with someone who is homozygous dominant, what are their chances of having an albino child?

a. 0
b. ¼
c. ½
d. ¾

14. What are the chances that 2 people who are carriers of the sickle cell allele will have a child with no sickle cell allele?
a. 0
b. ¼
c. ¾
d.100%

15. Two parents have a child that has sickle-cell anemia, a disease caused by a recessive mutation. Which of the following is true?
a. Both parents must be homozygous for the sickle-cell allele.
b. One parent must be homozygous for the sickle-cell allele, but the other can be homozygous for the normal allele.
c. Both parents must have at least one copy of the sickle-cell allele.
d. One parent must be homozygous for the sickle-cell allele, but the other can be heterozygous.

16. A study shows a high correlation in intelligence between adoptive parents and their adopted children. This supports the idea that intelligence
a. is highly heritable.
b. is not highly heritable.
c. has a strong environmental component.
d. has little environmental component.

17. Phenylketonuria (PKU) is a recessive genetic disease that leads to mental retardation when homozygous recessive individuals consume phenylalanine (look on a diet soda can). What is the genotype of a normal woman whose father had PKU?
a. PP
b. Pp
c. pp
d. Impossible to say.

18. True/False: A solid-colored cat is homozygous. Its spotted brother is heterozygous. The solid color is completely dominant to the spotted.

19. Alleles for petal color in a certain flower exhibit codominance. If allele A' codes for red coloration and allele A^+ codes for white coloration, what do the petals of a heterozygous flower look like?
a. Red
b. White
c. Red and white spotted
d. Chartreuse (yellow-green)

20. What are the chances that two carriers of cystic fibrosis will have child that inherits NO cystic fibrosis allele?

a. 0
b. 1/4
c. 1/2
d. 3/4

21. Which of the following is most likely to be a qualitative (non-quantitative) trait in humans?

a. Height
b. Skin color
c. The presence or absence of a genetic disease, like cystic fibrosis
d. Weight

22. True/False: If a quantitative trait has high heritability, it means that the trait is totally unaffected by the environment.

23. True/False: A (fictional) study shows a high correlation in intelligence between adoptive parents and their adopted children. This supports the idea that intelligence has high heritability.

24. True/False: If a trait was more similar between dizygotic than monozygotic twins, one could conclude that the trait was probably not highly heritable.

Answers

1. C. Because obesity is highly environmentally influenced, we cannot predict the exact chance that Hinor will be obese.
2. C. This is the only answer that MUST be true.
3. D. Different genes get turned on and off in different cells.
4. D. Even if you know the genetics and environment, you cannot predict their interaction!
5. A
6. A
7. C
8. False – they are by definition heterozygous
9. A
10. True – they share 100% of their genes
11. A. Correlation does not equal causation.
12. B
13. A – Aa x AA – looking for aa – Do the Punnett square!
14. B – S'S x S'S – looking for SS -- Do the Punnett square!
15. C
16. C – because adoptive parents and children have no more genes in common than any 2 humans.
17. B – If the woman's father had PKU, he had a **pp** genotype. He had to give his daughter a **p** in his sperm. If she is normal, she had a genotype of **Pp**, not PP.
18. False – if the heterozygote (Aa) is spotted, then spotted must be dominant.
19. C – codominance means that both alleles are expressed in the heterozygote.
20. B
21. C – the trait is not continuously variable – it is present or absent.
22. False
23. False – adopted children typically share 0% of their genes with adoptive parents.
24. True – dizygotic twins only share 50% of their genes.

IV. Appendix

Chi-Square Goodness of Fit Test

This lesson explains how to conduct a **chi-square goodness of fit test**. The test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

For example, suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60% were veterans but not All-Stars; and 10% were veteran All-Stars. We could gather a random sample of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company. The sample problem at the end of the lesson considers this example.

When to Use the Chi-Square Goodness of Fit Test

The chi-square goodness of fit test is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The variable under study is categorical.
- The expected value of the number of sample observations in each level of the variable is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis (H_o) and an alternative hypothesis (H_a). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

For a chi-square goodness of fit test, the hypotheses take the following form.

- H_o: The data are consistent with a specified distribution.
- H_a: The data are *not* consistent with a specified distribution.

Typically, the null hypothesis (H_o) specifies the proportion of observations at each level of the categorical variable. The alternative hypothesis (H_a) is that *at least* one of the specified proportions is not true.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the chi-square goodness of fit test to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. The chi-square goodness of fit test is described in the next section, and demonstrated in the sample problem at the end of this lesson.

Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequency counts, test statistic, and the P-value associated with the test statistic.

- Degrees of freedom. The degrees of freedom (DF) is equal to the number of levels (k) of the categorical variable minus 1.

  $DF = k - 1$

- Expected frequency counts. The expected frequency counts at each level of the categorical variable are equal to the sample size times the hypothesized proportion from the null hypothesis

  $E_i = np_i$

  where E_i is the expected frequency count for the *i*th level of the categorical variable, n is the total sample size, and p_i is the hypothesized proportion of observations in level *i*.

- Test statistic. The test statistic is a chi-square random variable (X^2) defined by the following equation.

  $X^2 = \Sigma [(O_i - E_i)^2 / E_i]$

  where O_i is the observed frequency count for the *i*th level of the categorical variable, and E_i is the expected frequency count for the *i*th level of the categorical variable.

- P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

Problem

Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans but not All-Stars, and 10% are veteran All-Stars.

Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  - Null hypothesis: The proportion of rookies, veterans, and All-Stars is 30%, 60% and 10%, respectively.
  - Alternative hypothesis: At least one of the proportions in the null hypothesis is false.
- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square goodness of fit test of the null hypothesis.
- **Analyze sample data**. Applying the chi-square goodness of fit test to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

DF = k - 1 = 3 - 1 = 2 $(E_i) = n * p_i$

$(E_1) = 100 * 0.30 = 30$

$(E_2) = 100 * 0.60 = 60$

$(E_3) = 100 * 0.10 = 10$ $X^2 = \Sigma [(O_i - E_i)^2 / E_i]$

$X^2 = [(50 - 30)^2 / 30] + [(45 - 60)^2 / 60] + [(5 - 10)^2 / 10]$

$X^2 = (400 / 30) + (225 / 60) + (25 / 10) = 13.33 + 3.75 + 2.50 = 19.58$

where DF is the degrees of freedom, k is the number of levels of the categorical variable, n is the number of observations in the sample, E_i is the expected frequency count for level i, O_i is the observed frequency count for level i, and X^2 is the chi-square test statistic.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.58.

We use the Chi-Square Distribution Calculator to find $P(X^2 > 19.58) = 0.0001$.

- **Interpret results**. Since the P-value (0.0001) is less than the significance level (0.05), we cannot accept the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the variable under study was categorical, and each level of the categorical variable had an expected frequency count of at least 5.

Chi-Square Test of Homogeneity

This lesson explains how to conduct a **chi-square test of homogeneity**. The test is applied to a single categorical variable from two or more different populations. It is used to determine whether frequency counts are distributed identically across different populations.

For example, in a survey of TV viewing preferences, we might ask respondents to identify their favorite program. We might ask the same question of two different populations, such as males and females. We could use a chi-square test for homogeneity to determine whether male viewing preferences differed significantly from female viewing preferences. The sample problem at the end of the lesson considers this example.

When to Use Chi-Square Test for Homogeneity

The test procedure described in this lesson is appropriate when the following conditions are met:

- For each population, the sampling method is simple random sampling.
- The variable under study is categorical.
- If sample data are displayed in a contingency table (Populations x Category levels), the expected frequency count for each cell of the table is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

Suppose that data were sampled from r populations, and assume that the categorical variable had c levels. At any specified level of the categorical variable, the null hypothesis states that each population has the same proportion of observations. Thus,

H_o: $P_{\text{level 1 of pop 1}} = P_{\text{level 1 of pop 2}} = \ldots = P_{\text{level 1 of pop r}}$

H_o: $P_{\text{level 2 of pop 1}} = P_{\text{level 2 of pop 2}} = \ldots = P_{\text{level 2 of pop r}}$

. . .

H_o: $P_{\text{level c of pop 1}} = P_{\text{level c of pop 2}} = \ldots = P_{\text{level c of pop r}}$

The alternative hypothesis (H_a) is that *at least* one of the null hypothesis statements is false.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the chi-square test for homogeneity to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. The chi-square test for homogeneity is described in the next section.

Analyze Sample Data

Using sample data from the contingency tables, find the degrees of freedom, expected frequency counts, test statistic, and the P-value associated with the test statistic. The analysis described in this section is illustrated in the sample problem at the end of this lesson.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to:

  $DF = (r - 1) * (c - 1)$

  where r is the number of populations, and c is the number of levels for the categorical variable.

- **Expected frequency counts.** The expected frequency counts are computed separately for each population at each level of the categorical variable, according to the following formula.

  $E_{r,c} = (n_r * n_c) / n$

where $E_{r,c}$ is the expected frequency count for population r at level *c* of the categorical variable, n_r is the total number of observations from population r, n_c is the total number of observations at treatment level *c*, and n is the total sample size.

- **Test statistic.** The test statistic is a chi-square random variable (X^2) defined by the following equation.

  $X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$

  where $O_{r,c}$ is the observed frequency count in population *r* for level *c* of the categorical variable, and $E_{r,c}$ is the expected frequency count in population *r* for level *c* of the categorical variable.

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

Problem

In a study of the television viewing habits of children, a developmental psychologist selects a random sample of 300 first graders - 100 boys and 200 girls. Each child is asked which of the following TV programs they like best: The Lone Ranger, Sesame Street, or The Simpsons. Results are shown in thecontingency table below.

	Viewing Preferences			Total
	Lone Ranger	Sesame Street	The Simpsons	
Boys	50	30	20	100
Girls	50	80	70	200
Total	100	110	90	300

Do the boys' preferences for these TV programs differ significantly from the girls' preferences? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  - Null hypothesis: The null hypothesis states that the proportion of boys who prefer the Lone Ranger is identical to the proportion of girls. Similarly, for the other programs. Thus,

H_o: $P_{\text{boys like Lone Ranger}} = P_{\text{girls like Lone Ranger}}$

H_o: $P_{\text{boys like Sesame Street}} = P_{\text{girls like Sesame Street}}$

H_o: $P_{\text{boys like Simpsons}} = P_{\text{girls like Simpsons}}$

  - Alternative hypothesis: At least one of the null hypothesis statements is false.
- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square test for homogeneity.
- **Analyze sample data**. Applying the chi-square test for homogeneity to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

DF = (r - 1) * (c - 1)
DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2

$E_{r,c} = (n_r * n_c) / n$
$E_{1,1} = (100 * 100) / 300 = 10000/300 = 33.3$
$E_{1,2} = (100 * 110) / 300 = 11000/300 = 36.7$
$E_{1,3} = (100 * 90) / 300 = 9000/300 = 30.0$
$E_{2,1} = (200 * 100) / 300 = 20000/300 = 66.7$
$E_{2,2} = (200 * 110) / 300 = 22000/300 = 73.3$
$E_{2,3} = (200 * 90) / 300 = 18000/300 = 60.0$

$X^2 = \Sigma[(O_{r,c} - E_{r,c})^2 / E_{r,c}]$
$X^2 = (50 - 33.3)^2/33.3 + (30 - 36.7)^2/36.7 +$
$(20 - 30)^2/30 + (50 - 66.7)^2/66.7 +$
$(80 - 73.3)^2/73.3 + (70 - 60)^2/60$
$X^2 = (16.7)^2/33.3 + (-6.7)^2/36.7 +$
$(-10.0)^2/30 + (-16.7)^2/66.7 +$
$(3.3)^2/73.3 + (10)^2/60$
$X^2 = 8.38 + 1.22 + 3.33 + 4.18 + 0.61 + 1.67 = 19.39$

where DF is the degrees of freedom, r is the number of populations, c is the number of levels of the categorical variable, n_r is the number of observations from population *r*, n_c is the number of observations from level *c* of the categorical variable, n is the number of observations in the sample, $E_{r,c}$ is the expected frequency count in population *r* for level *c*, and $O_{r,c}$ is the observed frequency count in population *r* for level *c*.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.39.

We use the Chi-Square Distribution Calculator to find $P(X^2 > 19.39) = 0.0000$. (The actual P-value, of course, is not exactly zero. If the Chi-Square Distribution Calculator reported more than four decimal places, we would find that the actual P-value is a very small number that is less than 0.00005 and greater than zero.)

- **Interpret results**. Since the P-value (0.0000) is less than the significance level (0.05), we reject the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the variable under study was categorical, and the expected frequency count was at least 5 in each population at each level of the categorical variable.