2021 Russell Sage Foundation Summer Institute in Social-Science Genomics

Problem Set 1

The purpose of this problem set is to teach some basic concepts that will be useful for the course, most of which we will start using right away in the early lectures. Note that the problem set looks longer than it is, because the problems themselves are teaching the material. Nonetheless, we expect that the problem set will require a serious time commitment, so make sure you give yourself enough time to complete the problem set before the Summer Institute begins.

Please email your completed problem set to rsf.genomics.school@gmail.com with the subject line, "Completed Problem Set 1", **before the opening dinner on Sunday, August 8.**

1. Hardy-Weinberg equilibrium

At a typical *locus* (meaning "location") in the genome, each individual in a population inherits one of two possible *alleles* from each parent. Each child receives one of the father's two alleles at random and one of the mother's two alleles at random. For concreteness, we denote one possible allele as "+" and the other as "−". An individual's *genotype* refers to the combination of alleles at the locus. Therefore, there are three possible genotypes at each locus: $++$, $-+$, and $--$.

(You may assume that the biological function of the locus does not depend on which allele was received by which parent. This is generally true, the exception being cases of *genomic imprinting*, in which the allele of only one of the parents is expressed.)

Denote the frequency of the "+" allele in the population by p, and denote the frequency of the "−" allele by $q = 1 - p$. Denote the population frequencies of the three genotypes by $P_0 \equiv Pr(--)$, $P_1 \equiv Pr(-+)$, and $P_2 \equiv Pr(++) = 1 - P_0 - P_1$.

Hardy-Weinberg equilibrium (HWE) refers to a particular relationship between the allele frequencies, p and q, and the genotype frequencies, P_0, P_1, and P_2. It serves as an important benchmark case in both theoretical and empirical analyses in genetics. This problem defines HWE, and the next problem develops some of its implications.

a. Explain why the following two equations must always hold, regardless of the values of P_0, P_1, and P_2:

$$p = P_2 + \frac{1}{2}P_1$$
$$q = P_0 + \frac{1}{2}P_1.$$

Hint: Consider the number of each type of allele (+ and −) in each genotype.

Thus, if we know P_0, P_1, and P_2, then the values of p and q are pinned down.

b. Now consider the reverse: If we know p and q, what can we conclude about P_0, P_1, and P_2? Show that the two triplets shown below are consistent with the same values of p and q.

(i) P_0: 0.5 P_1: 0.4 P_2: 0.1

(ii) P_0: 0.6 P_1: 0.2 P_2: 0.2

Therefore, knowing p and q, the population frequencies of + and − alleles, does not uniquely determine P_0, P_1, and P_2, the population frequencies of different genotypes, unless we make some assumptions. These assumptions and their implications are the subject of the remainder of this problem.

Specifically, we make five assumptions about the population:

i. The population is large.
ii. There are no mutations.
iii. There is no migration (i.e., no immigration or emigration).
iv. There is no selection (i.e., individuals of any genotype have the same number of offspring on average than individuals of any other genotype).
v. Mating is random (i.e., the probability of mating with a partner of any particular genotype is independent of one's own genotype).

c. Explain why Assumptions i-iv, taken together, imply that p and q are constant from one generation to the next.

The addition of Assumption v (random mating) takes the implications further. In particular, we will focus on how the distribution of genotypes, (P_0, P_1, P_2), changes from one generation to the next.

d. Explain why Assumption v implies that the alleles one inherits from one's father and mother are independent from each other.

e. Using this fact, conclude that

$$P_0 = q^2 \quad (1)$$
$$P_1 = 2pq \quad (2)$$
$$P_2 = p^2 \quad (3)$$

Equations (1)-(3) characterize the HWE. It is an equilibrium in the sense that once the genotypes, (P_0, P_1, P_2), are equal to their *Hardy-Weinberg frequencies*, they will be constant from one generation to the next. Note that regardless of the genotype frequencies in the parents' generation, the population will be in HWE in the offspring generation, provided that the assumptions underlying HWE have been met.

When working with individual-level genetic data, we observe both the allele frequencies and the genotype frequencies in the sample. Let $\hat{p}$ and $\hat{q}$ denote the allele frequencies in the sample, and let $\hat{P}_0$, $\hat{P}_1$, and $\hat{P}_2$ denote the genotype frequencies in the sample. Note that even if the sample is randomly drawn from a population that is in HWE, the sample analogs of equations 1-3 are unlikely to hold exactly due to random sampling error.

If, at a high level of statistical confidence, we can reject the null hypothesis that the sample analogs of equations 1-3 hold, it may be that the assumptions underlying HWE are violated. In practice, Assumptions i-v are rarely perfectly satisfied, and yet it turns out that the HWE equations are typically robust to deviations from these assumptions.

Some important additional notes:

Note 1: A standard test for HWE is Pearson's χ^2 test. The test statistic is a measure of the distance between the genotype frequencies expected under HWE and the genotype frequencies observed in the sample: $\left[\frac{(\hat{P}_0-\hat{q}^2)^2}{\hat{q}^2} + \frac{(\hat{P}_1-2\hat{p}\hat{q})^2}{2\hat{p}\hat{q}} + \frac{(\hat{P}_2-\hat{p}^2)^2}{\hat{p}^2}\right] N$, where N is the sample size. Under the null hypothesis of HWE, this test statistic follows a χ^2 statistic with 1 degree of freedom. For this distribution, the 5% significance threshold is 3.84. Therefore, at the 5% significance level, one would reject the null hypothesis of HWE when the value of the test statistic exceeds 3.84.

Note 2: As an empirical matter, major deviations from HWE are often an indication of genotyping errors caused by factors such as "heterozygote dropout"; see, e.g., chapter 16 (p. 375) in Neale et al. (2007). For this reason, a test of HWE is often used as a quality-control check in genome-wide analyses, and genetic variants that depart substantially from HWE are dropped from the analysis.

<u>References</u>

Sullivan, PF, and S Purcell. 2007. "Analyzing genome-wide association study data: a tutorial

using PLINK." In *Statistical Genetics: Gene Mapping through Linkage and Association*, eds. BN Neale, MA Ferreira, SE Medland, and D Posthuma. Taylor & Francis Group, 355–394.

2. Additive and dominance variance components

A *phenotype* refers to any individual-level outcome that may be affected by genes. For example, height, schizophrenia risk, or educational attainment could each be a phenotype. We will denote the value of individual i's phenotype, say height in centimeters, by y_i.

In this problem, we will study the relationship between individuals' phenotypes and their genotype at a single locus. In particular, we will decompose the relationship into a linear component (usually called the "additive component") and a non-linear component (usually called the "dominance component"). This decomposition is useful in empirical work, and it is critical to understand conceptually in order to interpret the meaning of empirical findings in genomics research.

As in Problem 1, we denote the two alleles that can be found at a particular locus by "+" and "−", yielding three possible genotypes: $++$, $-+$, and $--$. Define an individual's *genotype score*, $x_i \in \{0,1,2\}$, as the number of + alleles he or she has:

$$x_i = [0 \; for \; genotype \; -- \; 1 \; for \; genotype \; -+ \; 2 \; for \; genotype \; ++]$$

In an abuse of terminology, the genotype score is often called the "genotype." Note that we have defined it with respect to + being the *reference allele*, but either allele could be used as the reference (this choice is completely arbitrary).

Suppose we know the joint distribution of (y_i, x_i) in the population. Our goal is to find the *best predictor function*, $BP(y_i|x_i)$: the function that gives the best prediction of y_i given any particular observed value of x_i.

To formalize this problem, we need to be precise about what we mean by "best prediction." The most common approach, which we will pursue here, is to minimize the expected squared prediction error. To be precise, we define the best predictor function for y_i as a function of x_i by

$$BP(y_i|x_i) \equiv argE\left[\left(y_i - g(x_i)\right)^2|x_i\right]. \tag{4}$$

That is, we find the function $g(x_i)$ so that, when we draw randomly from the population distribution of y_i, the expected value of the squared difference between y_i and $g(x_i)$ is minimized.

It can be shown that the best predictor function of y_i is the *conditional expectation function*:

$$BP(y_i|x_i) = E[y_i|x_i].$$

(Note: If you are interested, try reaching this solution by yourself by considering the problem for each fixed value of x_i separately, say starting with $x_i = 0$ and defining $z \equiv g(0)$. To find the value of $BP(y_i|x_i = 0)$, take the derivative of $E[(y_i - z)^2|x_i = 0]$ with respect to z and set it equal to 0.)

a. Explain why (by definition of the conditional expectation) we can always decompose the value of the phenotype into the sum of its conditional expectation and a residual that is uncorrelated with genotype:

$$y_i = E[y_i|x_i] + \epsilon_i,$$

where $E(\epsilon_i) = 0$ and $Cov(x_i, \epsilon_i) = 0$.

We will now decompose the conditional expectation function into the sum of a linear ("additive") part and a non-linear ("dominance") part. To simplify the expression we will obtain, we first re-center the phenotype measure (just subtracting a constant):

$$\check{y}_i \equiv y_i - \frac{E(y_i|x_i = 0) + E(y_i|x_i = 2)}{2}.$$

We define the *additive effect* as $a \equiv E\left(\check{y}_i|x_i = 2\right)$ and we define the *dominance deviation* as $d \equiv E\left(\check{y}_i|x_i = 1\right)$.

b. Show that

$$E\left(\check{y}_i|x_i = 0\right) = -a.$$

c. Show that

$$E\left(\check{y}_i|x_i\right) = (x_i - 1)a + I\{x_i = 1\}d, \tag{5}$$

where $I\{\cdot\}$ is an indicator function that takes the value 1 when the expression inside the curly brackets is true and takes the value 0 otherwise.

(Hint: Show that Equation 5 reduces to a when $x_i = 2$, to $-a$ when $x_i = 0$, and to d when $x_i = 1$.)

d. Given equation (5), explain why it makes sense to refer to a as the "additive effect" (of the "+" allele) and d as the "dominance deviation."

e. In one or more simple figures where the x-axis is x_i and the y-axis is $E(\check{y}_i|x_i)$, draw three cases: $d = 0$, $d = a$, and $d = -a$. Explain why $d = a$ corresponds to a situation of Mendelian inheritance where the phenotype is a "dominant trait" (i.e., inheriting a "+" allele from *either* the mother *or* the father is sufficient for enhanced phenotypic expression). Explain why $d = -a$ corresponds to a situation of Mendelian inheritance where the phenotype is a "recessive trait" (i.e., inheriting a "+" allele from *both* the mother *and* the father is needed for enhanced phenotypic expression). (Note that equation (5) is a generalization of these two cases, allowing for any value of d.)

The exercises above show that when $d \neq 0$, the best predictor function is non-linear. However, we will often want to rely on the best possible *linear* approximation to the relationship between y_i and x_i.

Therefore, consider the problem of finding the *best linear predictor* for y_i as a function of x_i, $BLP(y_i|x_i)$, defined as the function $g(x_i)$ that solves the problem in equation (4) but restricted to have the form $g(x_i) = \alpha + \beta x_i$ for some constants α, β. This problem can be described formally as:

$$BLP(y_i|x_i) \equiv \left\{\alpha + \beta x_i | (\alpha, \beta) = argE\left[\left(y_i - (\alpha + \beta x_i)\right)^2\right]\right\}. \tag{6}$$

We solve this minimization problem by setting β equal to the *population regression coefficient* and setting α equal to the *population regression intercept*:

$$\beta = \frac{Cov(y_i, x_i)}{Var(x_i)},$$

$$\alpha = E[y_i] - \beta E[x_i].$$

$$BLP(y_i|x_i) = \alpha + \beta x_i.$$

(Note: If you are interested, try reaching this solution by yourself by taking the partial derivatives of $E\left[\left(y_i - (\alpha + \beta x_i)\right)^2\right]$ with respect to α and β, setting them both equal to 0, and solving the two equations simultaneously for α and β.)

We often refer to the best linear predictor function as the *population regression equation*.

As in Problem 1, denote the frequency of the "+" allele in the population by p, and denote the frequency of the "−" allele by $q = 1 - p$. From here onward in this problem, we will assume that the genotype frequencies are in HWE, meaning that the following equations apply:

$$P_0 = Pr\ Pr\ (x_i = 0) = q^2$$
$$P_1 = Pr\ Pr\ (x_i = 1) = 2pq$$
$$P_2 = Pr\ (x_i = 2) = p^2$$

f. Under this assumption, show that the mean and variance of x_i are given by:

$$E(x_i) = 2p,$$
$$Var(x_i) = 2p(1 - p).$$

Hint: Find expected values by expressing all allele frequencies in terms of p (where $q = (1 - p)$).

The *coefficient of determination*, or R^2, of a regression is defined as the proportion of variance explained by the predictor variables: $R^2 \equiv \frac{Var(\beta x_i)}{Var(y_i)}$.

g. Show that

$$R^2 = \frac{2p(1-p)\beta^2}{Var(y_i)}. \tag{7}$$

Hint: Use the expression for $Var(x_i)$ from part f.

Note that Equation (7) is useful empirically in genome-wide association study (GWAS) meta-analyses, in which results from regressions run in different datasets (typically with some differences in control variables across datasets) are combined to yield a single, overall estimate of β. In such meta-analyses, it is common to use the sample analog of equation (7)—i.e., a version of equation (7) calculated using the sample estimates of p, $Var(y_i)$, and β—to calculate an estimate of R^2. Equation (7) is also useful theoretically for conducting power calculations, as will be illustrated in Problem 3(j) below.

To simplify the calculations that follow, we will de-mean the phenotype and genotype variables: define $\tilde{y}_i \equiv y_i - E(y_i)$ and $\tilde{x}_i \equiv x_i - E(x_i)$.

h. Show that we can write the best linear predictor for $\tilde{y}_i$ as a function of $\tilde{x}_i$ as:

$$BLP(\tilde{y}_i|\tilde{x}_i) = \beta\tilde{x}_i,$$

where $\beta = \frac{Cov(y_i,x_i)}{Var(x_i)} = \frac{Cov(\tilde{y}_i,\tilde{x}_i)}{Var(\tilde{x}_i)}$.

Hint: Plug the de-meaned variables into the equations for α and β given above to show that when the variables are de-meaned, $\alpha = 0$ while β does not change.

i. Using the expression for $E(x_i)$ from part f, show that

$$E(\tilde{x}_i|x_i = 0) = (\tilde{x}_i|x_i = 0) = -2p,$$
$$E(\tilde{x}_i|x_i = 1) = (\tilde{x}_i|x_i = 1) = q - p,$$
$$E(\tilde{x}_i|x_i = 2) = (\tilde{x}_i|x_i = 2) = 2q.$$

j. In part i, you found that the expected values of the de-meaned genotype score depend solely on p and q. Now, we'll show how the overall and conditional expected values of the de-meaned phenotype depend on p and q, as well as the additive effect a and the dominance effect d. Using equation (5) together with the HWE genotype frequencies, show that

$$E(\check{y}_i) = a(p - q) + 2pqd.$$

k. Using the equation from part j, together with equation (5) and the fact that $\tilde{y}_i \equiv y_i - E(y_i) = \check{y}_i - E(\check{y}_i)$, show that

$$E(\tilde{y}_i|x_i = 0) = -2p(a + qd),$$
$$E(\tilde{y}_i|x_i = 1) = d(1 - 2pq) + a(q - p),$$
$$E(\tilde{y}_i|x_i = 2) = 2q(a - pd).$$

Next, we will prove that the slope of the best linear predictor function, β, is equal to

$$\beta = a + d(q - p). \tag{8}$$

We have provided the steps for you, since the math gets a bit messy.

We know that $Var(\tilde{x}_i) = Var(x_i)$ since the spread of a random variable does not change after subtracting a constant (which is all that has been done to de-mean the variable).

We also know that $Cov(\tilde{y}_i, \tilde{x}_i) = E(\tilde{y}_i\tilde{x}_i) - E(\tilde{y}_i)E(\tilde{x}_i) = E(\tilde{y}_i\tilde{x}_i) - (0 * 0) = E(\tilde{y}_i\tilde{x}_i)$. By the law of iterated expectations, $E(\tilde{y}_i\tilde{x}_i) = E(\tilde{x}_i E(\tilde{x}_i))$.

Plugging these into the formula $\beta = \frac{Cov(\tilde{y}_i,\tilde{x}_i)}{Var(\tilde{x}_i)}$, we get:

$$\beta = \frac{E\left(\tilde{x}_i E(\tilde{y}_i|\tilde{x}_i)\right)}{Var(x_i)}.$$

The numerator of this expression can be expanded as follows:

$$E\left(\tilde{x}_i E(\tilde{y}_i|\tilde{x}_i)\right) = \sum_{j=0}^{j=2} Pr(x_i = j)E(\tilde{x}_i|x_i = j)E(x_i = j)$$
$$= P_0 E(x_i = 0)E(x_i = 0) + P_1 E(x_i = 1)E(x_i = 1) + P_2 E(x_i = 2)E(x_i = 2).$$

Using the equations above for P_0, P_1 and P_2 (assuming we are in HWE), for $E(\tilde{x}_i|x_i = j)$, and for $E(\tilde{y}_i|x_i = j)$, we can first expand and then simplify the expression for the numerator.

$$= [q^2(-2p)(-2p(a + qd))] + [2pq(q - p)(d(1 - 2pq) + a(q - p))] + [p^2(2q)(2q(a - pd))]$$

$$= \left[-2pq^2\left(-2p(a + qd)\right)\right] + [2pq(q - p)\left(d(1 - 2pq) + a(q - p)\right)] + [2p^2q\left(2q(a - pd)\right)]$$

$$= [4p^2q^2a + 4p^2q^3d] + [(2pq^2 - 2p^2q)(d - 2pqd + aq - ap)] + [4p^2q^2a - 4p^3q^2d]$$

$$= [4p^2q^2a + 4p^2q^3d] + [2pq^2d - 4p^2q^3d + 2pq^3a - 2p^2q^2a - 2p^2qd + 4p^3q^2d - 2p^2q^2a + 2p^3qa] + [4p^2q^2a - 4p^3q^2d]$$

$$= 4p^2q^2a + 2pq^2d + 2pq^3a - 2p^2qd + 2p^3qa$$

Recall from part f that the denominator, $Var(x_i) = 2p(1 - p) = 2pq$. Then

$$\beta = \frac{E\left(\tilde{x}_i E(\tilde{y}_i|\tilde{x}_i)\right)}{Var(x_i)} = \frac{4p^2q^2a + 2pq^2d + 2pq^3a - 2p^2qd + 2p^3qa}{2pq}$$

$$= 2pqa + qd + q^2a - pd + p^2a.$$

Simplifying, we get the following equation for β:

$$\beta = a(p^2 + q^2 + 2pq) + d(q - p)$$

Since $p + q = 1$, it follows that $(p + q)^2 = 1$, and therefore $q^2 + p^2 + 2pq = 1$. Hence, our final formula for β:

$$\beta = a + d(q - p)$$

m. What is the substantive interpretation of β? In your answer, be sure to speak to the different genetic elements that this formula is composed of. (Hint: Can it be described simply as the additive effect a? Why does the dominance deviation d enter the formula?)

Earlier, we decomposed the best predictor function into an additive effect and a dominance deviation. We now have all the pieces we need to do a different decomposition, into what are called *variance components*.

We begin by defining the *genetic factor*,

$$G(x_i) \equiv BP(y_i|x_i) - E(y_i),$$

which is the best predictor of the phenotype given the genotype, de-meaned to make the math easier. We define the *additive variance component*, or the *additive component*, as the (de-meaned) best linear predictor of the phenotype given the genotype:

$$A(x_i) \equiv BLP(y_i|x_i) - E(y_i).$$

Finally, we define the *dominance variance component*, or the *dominance component*, as the improvement in prediction from using the best predictor relative to using the best linear predictor:

$$D(x_i) \equiv BP(y_i|x_i) - BLP(y_i|x_i).$$

Note that $G(x_i)$, $A(x_i)$, and $D(x_i)$ are constants when evaluated at a particular genotype x_i, but they are random variables when x_i is unobserved (as is the case in family studies or adoption studies).

n. Show that:

$$G(x_i) = A(x_i) + D(x_i).$$

o. Show that $A(x_i) = \beta\tilde{x}_i$.

Using the results from part k and the proof that follows, it can be shown that the following formulas hold (interested students can verify themselves):

x_i	$A(x_i)$	$D(x_i)$
0	$-2p\beta$	$-2dp^2$
1	$(q-p)\beta$	$2dpq$
2	$2q\beta$	$-2dq^2$

p. Using the results from part i and the formulas in the table above, show that:

$$E[A] = 0,$$
$$E[D] = 0.$$

It can also be shown that:

$$Var(A) = 2pq\beta^2,$$
$$Var(D) = (2pqd)^2.$$

q. Using either the results from part i together with the table from part o, and/or the properties of Ordinary Least Squares regression, show that

$$Cov(A, D) = 0.$$

Hint: Since A and D are both mean zero, $\mathrm{Cov}(A, D) = \mathrm{E}(AD)$.

This last equality is an important—and useful—property of the variance decomposition: it implies that the additive and dominance components are uncorrelated with each other.

r. Explain why it also implies that

$$Var(G) = Var(A) + Var(D).$$

Another important property of the variance decomposition is that, even if the dominance deviation is large, the additive variance component captures most of the variance in the genetic factor if one of the alleles is relatively rare in the population. Thus,

$$\frac{Var(A)}{Var(G)} = 1.$$

s. Explain intuitively why this is true.

   Hint: Suppose p is small, say $p = 0.1$. What are the HWE genotype frequencies in the population? If the best linear predictor of y_i as a function of x_i (almost) completely ignores the least common genotype, how large will the difference be between the best predictor function and the best linear predictor function?

t. Consider a population in Hardy-Weinberg equilibrium. For a given child, let x_m and x_f denote the mother's and father's genotype, respectively. As in Problem 1, let P_j denote the probability that child i has genotype $x_i = j$ for $j \in \{0,1,2\}$. To save you time, the table below shows the probability of each mother-father combination and the resulting probability of each genotype for the child.

x_m	x_f	$Prob(x_m, x_f)$	P_0	P_1	P_2
0	0	$q^2 \times q^2$	1	0	0
0	1	$q^2 \times 2pq$	½	½	0
0	2	$p^2 \times q^2$	0	1	0
1	0	$2pq \times q^2$	½	½	0
1	1	$2pq \times 2pq$	¼	½	¼
1	2	$2pq \times p^2$	0	½	½
2	0	$q^2 \times p^2$	0	1	0
2	1	$p^2 \times 2pq$	0	½	½
2	2	$p^2 \times p^2$	0	0	1

Classical studies of heritability (called *behavior genetics* studies) compare observed correlations between family members with theoretically expected correlations. Here, we will focus on deriving some theoretically expected parent-child correlations—for concreteness, the mother-child correlation.

Show that the mother-child correlation in the additive genetic component is:

$$Corr(A_m, A_i) \equiv \frac{Cov(A_m, A_i)}{\sqrt{Var(A_m)Var(A_i)}} = \frac{1}{2}.$$

Hint: the algebra is again tedious, so the following roadmap may be helpful:

1. From the information in the table, we can deduce:

$$Prob(x_i = 0|x_m = 0) = q^2 \times 1 + \ 2pq \times ½ + p^2 \times 0 = q,$$
$$Prob(x_i = 1|x_m = 0) = q^2 \times 0 + \ 2pq \times ½ + p^2 \times 1 = p,$$
$$Prob(x_i = 2|x_m = 0) = q^2 \times 0 + \ 2pq \times 0 + p^2 \times 0 = 0.$$

(Alternatively, the algebra can be skipped by reasoning directly from the assumption of random mating.) Use the same strategy to solve for $P(x_i = j|x_m = 1)$ and $P(x_i = j|x_m = 2)$ for $j \in \{0,1,2\}$.

2. Note that the probability that both the mother and the child have the value of $A(x_i)$ corresponding to genotype score $x_i = 0$ is:

$$\begin{aligned} &Prob\big(A_m = A(0), A_i = A(0)\big) \\ &= Prob(x_m = 0, x_i = 0) \\ &= Prob(x_i = 0|x_m = 0)Prob(x_m = 0) \\ &= q \times q^2 = q^3, \end{aligned}$$

   where $Prob(x_m = 0)$ is equal to the HWE genotype frequency of genotype score 0. Similarly calculate $Prob\big(A_m = A(j), A_i = A(k)\big)$ for all of the (other 8) combinations of genotype scores for the mother $j \in \{0,1,2\}$ and the child $k \in \{0,1,2\}$.

3. By definition, the covariance is:

$$Cov(A_m, A_i) = \sum_{j=0}^{2} \sum_{k=0}^{2} Prob\big(A_m = A(j), A_i = A(k)\big) \times \big[A(j) - E[A]\big] \times \big[A(k) - E[A]\big],$$

where $A(j)$ and $A(k)$ are the values from the table in part o. Show that this equation can be rearranged to get:

$$Cov(A_m, A_i) = q^2[qA(0)A(0) + pA(0)A(1)] +$$
$$2pq\left[\frac{1}{2}qA(1)A(0) + \frac{1}{2}A(1)A(1) + \frac{1}{2}pA(1)A(2)\right] +$$

$$p^2[qA(2)A(1) + pA(2)A(2)].$$

4. Simplify the equation in part 3 and divide by $\sqrt{Var(A_m)Var(A_i)}$ (where each of these two variances is given by the expression in part p) to get the desired result.

   Show that the mother-child correlation in the dominance genetic component is:

$$Corr(D_m, D_i) \equiv \frac{Cov(D_m, D_i)}{\sqrt{Var(D_m)Var(D_i)}} = 0.$$

Some important additional notes

Note 1: Calculations like those in part t can be used to help draw inferences about the relative importance of the additive and dominance genetic components for a given phenotype. For example, using similar calculations, it can be shown that between siblings, the correlation in the additive genetic component is $\frac{1}{2}$, and the correlation in their dominance genetic component is $\frac{1}{4}$. The fact that the parent-child correlation in height is similar to the sibling correlation therefore constitutes suggestive evidence that the dominance component is small, and hence that genetic effects on height can be well approximated by a linear model. We will discuss this kind of inference in more detail during the lectures.

Note 2: The fact that the parent-child correlation of the dominance component is zero is important in its own right, for the purposes of studying evolutionary dynamics.

<u>Background Reading</u>

- The fundamental statistical concepts used in the problem set question—properties of the covariance operator, the conditional expectation function, the law of iterated expectations, etc.—are covered in standard statistical texts. A good source if you want to brush up on these concepts is chapter 5 in Goldberger (1991).
- The single-locus model is covered in standard texts, see for example chapter 6-9 in Falconer and Mackay (1996), Kwan, Purcell and Sham (2007) and Lynch and Walsh (1998). See also Goldberger's (2005) treatment (especially the appendices), which emphasize the BLP interpretation of additive genetic variance.

<u>References</u>

Falconer, D, and T Mackay. 1996. Introduction to Quantitative Genetics *Introduction to Quantitative Genetics*. 4th ed. Harlow, Essex: Longman Group Ltd.

Goldberger, AS. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Goldberger, AS. 2005. "Structural Equation Models in Human Behavior Genetics." In *Identification and Inference for Econometric Models Essays in Honor of Thomas Rothenberg*, eds. DW Andrews and JS Stock. Cambridge: Cambridge University Press, 11–26.

Kwan, JS, S Purcell, and PC Sham. 2007. "Introduction to Biometrical Genetics." In *Statistical Genetics: Gene Mapping through Linkage and Association*, eds. BN Neale, MA Ferreira, SE Medland, and D Posthuma. Taylor & Francis Group, 17–42.

Lynch, M, and B Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. MA, USA: Sinauer Associates.

3. Statistical power

Suppose we want to estimate the effect of some genetic variable (e.g., the genotype score at a particular locus, or the value of a polygenic index that aggregates across many loci), x, on a phenotype, y. Denote the best linear predictor function of y given x—which we will call the *population regression equation*—by

$$y_i = \psi + \beta x_i + \epsilon_i, \tag{9}$$

where i indexes individuals, ψ is a constant (we reserve the notation α for a different variable, defined below), and ϵ_i is an error term that has mean 0 and is uncorrelated with x_i (where both are true by definition of the best linear predictor function, as you showed in Problem 2(a)). We denote the variance of x_i by σ_x^2. The parameter we want to estimate is β.

We will assume that ϵ_i is independent across individuals and has variance σ_ϵ^2 (that, again, does not depend on x_i). (Remember that each of these is a substantive assumption that does not follow from the fact that equation (9) is the best linear predictor function.)

Now, suppose we draw a random sample of N individuals from the population, and we estimate the regression equation (9) in our sample using Ordinary Least Squares. Let $\hat{\beta}$ denote the resulting estimate of β. We will assume that the distribution of $(\hat{\beta}\,|\beta)$ is

$$\hat{\beta}\,|\beta \sim Normal\left(\beta, \frac{\sigma_\epsilon^2}{\sigma_x^2 N}\right).$$

This equation is exactly true when ϵ_i is normally distributed, and the Central Limit Theorem implies that it is a very good approximation when N is large.

Suppose σ_ϵ^2 and σ_x^2 are known, so that we can calculate the true standard error of $\hat{\beta}$, which is defined as $\sqrt{Var(\hat{\beta}\,|\beta)}$. The standard error is thus $\frac{\sigma_\epsilon}{\sigma_x\sqrt{N}}$.

The observed t-statistic, which is defined as the estimated parameter divided by its standard error, is therefore

$$t_{obs} \equiv \frac{\hat{\beta}\sigma_x}{\frac{\sigma_\epsilon}{\sqrt{N}}}.$$

a. Show that the distribution of the observed t-statistic is

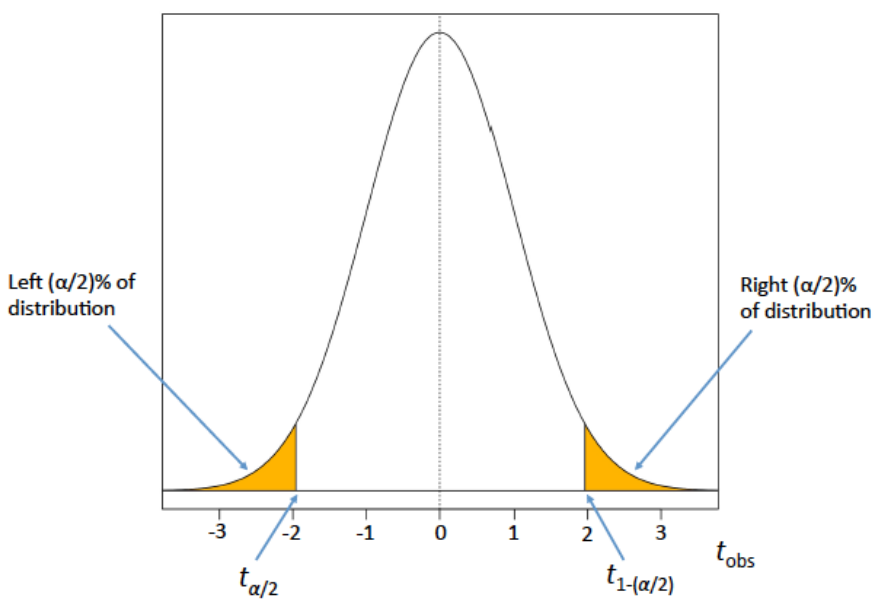
$$t_{obs} | \beta \sim Normal\left(\frac{\beta\sigma_x}{\frac{\sigma_\epsilon}{\sqrt{N}}}, 1\right).$$

Note: In practice, σ_ϵ^2 and σ_x^2 are usually not known and must be estimated. In that case, the t-statistic is defined as $\hat{\beta}$ divided by the *estimated* standard error. As a result, the t-statistic follows a t-distribution rather than a normal distribution, and hence the distribution has fatter tails. We assume that σ_ϵ^2 and σ_x^2 are known in order to keep the calculations simpler. Moreover, the normal distribution is a good approximation to the t-distribution when N is large.

Suppose the null hypothesis, H_0, is that there is no relationship between the genetic variable and the phenotype: $\beta = 0$. Note that under the null, the distribution of the t-statistic follows a standard normal distribution:

$$t_{obs} | (\beta = 0) \sim Normal(0,1).$$

Denote our *threshold for statistical significance* by α. We will reject the null hypothesis—and declare our estimated $\hat{\beta}$ to be *statistically significant*—whenever the observed t-statistic turns out to be in a region with less than α% probability under the null. Specifically, we will define the rejection region by two thresholds: $t_{\frac{\alpha}{2}}$ is the left $\frac{\alpha}{2}$% tail of the distribution, and $t_{1-\left(\frac{\alpha}{2}\right)}$ is the right $\frac{\alpha}{2}$% tail of the distribution, as illustrated below.



Let Φ denote the cumulative distribution function (CDF) of the standard normal distribution, and let Φ^{-1} denote its inverse function. (Reminder: The CDF determines the probability that a randomly drawn *t*-statistic, *t*, is less than some specific value of *t*.)

b. Explain why the left threshold is given by $t_{\frac{\alpha}{2}} = \Phi^{-1}\left(\frac{\alpha}{2}\right)$, and explain why the right threshold is given by $t_{1-\left(\frac{\alpha}{2}\right)} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

Suppose the *alternative hypothesis*, H_1, corresponds to an *anticipated effect size*, β_1, for some specific value $|\beta_1| > 0$.

c. Show that the distribution of the observed *t*-statistic under the alternative hypothesis is

$$t_{obs} | (\beta = \beta_1) \sim Normal(NCP, 1),$$

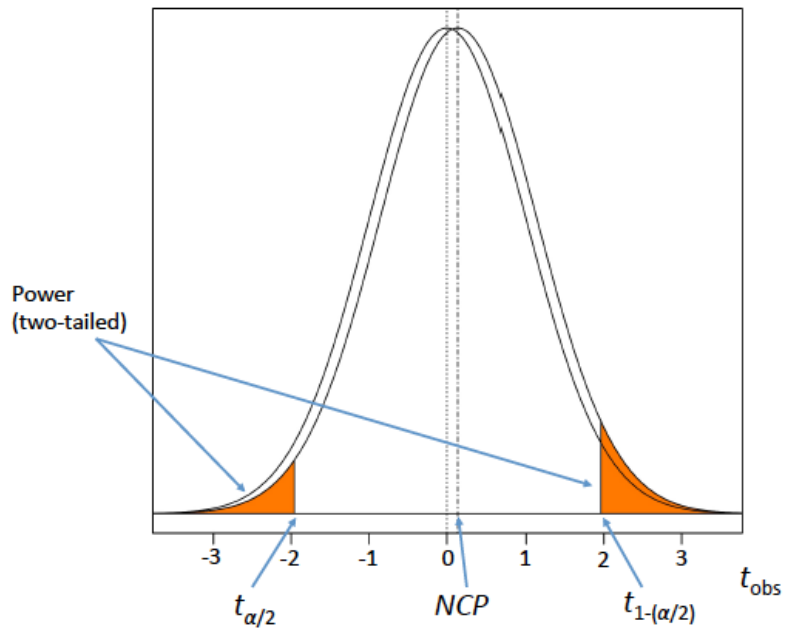
where $NCP \equiv \frac{\beta_1 \sigma_x}{\frac{\sigma_\epsilon}{\sqrt{N}}}$ is called the *non-centrality parameter*.

Conclude that

$$(t_{obs} - NCP) | (\beta = \beta_1) \sim Normal(0,1).$$

Power is the probability of rejecting the null hypothesis when the alternative hypothesis is true.

d. Explain why the area of the shaded region in the figure below is equal to the level of power. Further, explain what happens as the *NCP* grows (i.e., approaches infinity), holding all else constant. What happens as the *NCP* approaches 0?



e. We will now solve for power. Justify each of these steps of algebra to solve for the probability that, under the alternative hypothesis, t_{obs} falls in the rejection region in the left tail of the distribution:

$$\begin{aligned} Pr\left(t_{obs} \leq t_{\frac{\alpha}{2}}\right) &= Pr\left(t_{obs} - NCP \leq t_{\frac{\alpha}{2}} - NCP\right) \\ &= \Phi\left(t_{\frac{\alpha}{2}} - NCP\right) \\ &= \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - NCP\right). \end{aligned}$$

Using an analogous argument, show that the probability that t_{obs} falls in the rejection region in the right tail is

$$Pr\left(t_{obs} \geq t_{1-\left(\frac{\alpha}{2}\right)}\right) = 1 - \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - NCP\right).$$

Conclude that the level of power is given by

$$Power = \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - NCP\right) + 1 - \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - NCP\right). \tag{10}$$

Notice that calculating power using Equation 10 requires calculating NCP using

$$\frac{\beta_1 \sigma_x}{\sigma_\epsilon}.$$

f. If the phenotype y is educational attainment, then what are the units of $\frac{\beta_1 \sigma_x}{\sigma_\epsilon}$? What if the phenotype is height?

Hint: Consider the units of β_1, σ_x, and σ_ϵ. How do they cancel out (or not) in the equation for the NCP?

We will now develop an approximate equation for *NCP* that is also unitless and therefore comparable across phenotypes, but that is often more convenient to use.

g. Recall from Problem 2g that the R^2 of a regression is defined as the proportion of variance explained by the predictor variables. Under the alternative hypothesis, consider the population regression equation defined by Equation 9. Justify the following steps of algebra:

$$R^2 \equiv \frac{Var(\beta_1 x_i)}{Var(y_i)}$$

$$= \frac{Var(\beta_1 x_i)}{Var(\beta_1 x_i + \epsilon_i)}$$

$$= \frac{Var(\beta_1 x_i)}{Var(\beta_1 x_i) + Var(\epsilon_i)}$$

$$= \frac{(\beta_1)^2 \sigma_x^2}{(\beta_1)^2 \sigma_x^2 + \sigma_\epsilon^2}.$$

Next, show that

$$\frac{R^2}{1 - R^2} = \frac{(\beta_1)^2 \sigma_x^2}{\sigma_\epsilon^2}.$$

Using the fact that $\frac{R^2}{1-R^2} \approx R^2$ when R^2 is small, show that when the genetic variable explains a small amount of variance in the phenotype, we can approximate *NCP* by

$$NCP = \sqrt{N \frac{(\beta_1)^2 \sigma_x^2}{\sigma_\epsilon^2}} \approx \sqrt{NR^2}. \tag{11}$$

h. Write a program (in R, Matlab, Stata, etc.), or create a spreadsheet in Excel, to calculate power as a function of α, N, and R^2 using Equations 10 and 11. Fill in the entries of the table below with the level of power corresponding to each scenario. Holding the other variables constant, how does power change as sample size increases? How does it change as the explanatory power (R^2) of the genetic variant increases? And, is it higher or lower when using a more stringent significance threshold?

(Note that the effect size $R^2 = 0.0002$ corresponds to single genetic variant associated with educational attainment, and the effect size $R^2 = 0.07$ corresponds to a current polygenic index for educational attainment.)

	Standard significance threshold ($\alpha = 0.05$)				
	$N = 100$	$N = 1{,}000$	$N = 10{,}000$	$N = 100{,}000$	$N = 1{,}000{,}000$
$R^2 = 0.0002$					
$R^2 = 0.07$					

	Genome-wide significance threshold ($\alpha = 5\times10^{-8}$)				
	$N = 100$	$N = 1{,}000$	$N = 10{,}000$	$N = 100{,}000$	$N = 1{,}000{,}000$
$R^2 = 0.0002$					

Suppose that x_i is the genotype score at a single locus. Recall from Problem 2g that, under HWE, the proportion of variance in the phenotype explained by the locus is given by Equation 7:

$$R^2 = \frac{2p(1-p)\beta^2}{Var(y_i)},$$

where p is the frequency of the reference (+) allele. A common convention, which we will adopt in the remainder of this problem, is to choose the *minor allele* (the one that is less common in the population) as the reference allele. Loci where p is small—often defined as $p < 0.01$—are called *rare variants*, and other loci (where $0.01 \leq p \leq 0.5$) are called *common variants*.

i. Explain why equation (7) implies the following two facts:

(i) Holding constant N and α, a common variant that has population regression coefficient $\beta \neq 0$ will explain a greater proportion of variance than a rare variant that has the same population regression coefficient β.

(ii) Holding constant N and α, we have more power to detect a common variant that has population regression coefficient $\beta \neq 0$ than a rare variant that has the same population regression coefficient β.

j. Now consider two different loci: one has minor allele frequency p_1 and regression coefficient β_1, and the other has minor allele frequency p_2 and regression coefficient β_2. Show that, holding constant N and α, the power to detect these two loci is the same if

$$p_1(1-p_1)\beta_1^2 = p_2(1-p_2)\beta_2^2.$$

k. Supposing $p_1 = 0.5$ and $p_2 = 0.01$, show that the rare variant's regression coefficient must be roughly 5 times larger than that of the common variant in order for the study to have equal power to detect it.

4. Standardized regression

For some purposes, it is easier to work with variables that are *standardized*, meaning that they are transformed to have mean 0 and variance 1. For that reason, throughout this course, we will sometimes work with standardized variables. In this problem, we will derive some of the useful properties of regression with standardized variables.

Consider the population regression equation

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

where i indexes individuals, and ϵ_i is an error term that has mean 0 and variance σ_ϵ^2.

Define the standardized variables:

$$\tilde{y}_i \equiv \frac{y_i - \mu_y}{\sigma_y} \quad and \quad \tilde{x}_i \equiv \frac{x_i - \mu_x}{\sigma_x},$$

where μ_y is the mean of y_i, σ_y is the standard deviation of y_i, μ_x is the mean of x_i, and σ_x is the standard deviation of x_i.

a. Show that

$$\tilde{y}_i = \tilde{\beta}\tilde{x}_i + \tilde{\epsilon}_i, \tag{12}$$

where $\tilde{\beta} \equiv \frac{\beta\sigma_x}{\sigma_y}$, and $\tilde{\epsilon}_i$ is an error term that has mean 0 and variance $\frac{\sigma_\epsilon^2}{\sigma_y^2}$.

b. Suppose the units of y are years of education, and the units of x are the number of minor alleles at a particular locus. What are the units of $\tilde{y}$ and $\tilde{x}$? What are the units of β? Of $\tilde{\beta}$?

c. One useful property of standardized regression is that the regression coefficient, $\tilde{\beta}$, is equal to the correlation coefficient, r_{xy}. Prove that property.

   Hint: Use the following two facts: (i) by the properties of Ordinary Least Squares regression, $\beta = \frac{\sigma_{xy}}{\sigma_x^2}$, where σ_{xy} is the covariance between x and y; and (ii) the correlation coefficient between x and y is defined as $r_{xy} \equiv \frac{\sigma_{xy}}{\sigma_x\sigma_y}$.

d. Prove the following two claims:

(i) The R^2 from the standardized regression equation (12) is equal to the R^2 from the original regression.

(ii) The R^2 from the standardized regression equation (12) is equal to the squared correlation coefficient, r_{xy}^2.

Putting these claims together with part c gives us another useful property of standardized regression: $R^2 = \tilde{\beta}^2$.

Note: When the population regression equation is multivariate rather than univariate, the above properties still hold, but the correlation coefficient is replaced by the *partial correlation coefficient*. The partial correlation coefficient between x and y can be calculated by (i) running a regression of y on the other regressors (besides x); (ii) running a regression of x on the other regressors; and finally (iii) taking the correlation between the residuals from those two regressions.