Name: _____ Date: _____

Human Variation and Disease: Discussion 1 Notes and Concept Checks

Answer Key

The goal of this discussion is to review the statistical concepts described in Appendix B described in Gillespie's "Population Genetics: A Concise Guide" that will be useful for this class. Questions to check conceptual understanding are provided, incorporating course material when applicable.

Complete this worksheet to the best of your ability and bring it to the first discussion section. Also, please email me (leblake@uchicago.edu) any concepts or sections that you found challenging by Thursday night at midnight.

0.1 What are random variables?

- A discrete random variable (e.g. $\mathbf{X}$), is a function that takes on certain values depending on the outcome of some event, trial, or experiment.
- An event with n outcomes and its associated random variable may be described as follows:

Outcome	Value of $\mathbf{X}$	Probability
1	x_1	p_1
2	x_2	p_2
3	x_3	p_3
.	.	.
.	.	.
i	x_i	p_i
.	.	.
n	x_n	n_i

- **Concept Check 1:** In the table above, what is Prob $\{\mathbf{X} = x_i\}$? p_i

- **Concept Check 2:** In the table above, what is Prob $\{\mathbf{X} = x_1 \cup \mathbf{X} = x_3\}$? $p_1 + p_3$

0.2 Moments of Random Variables

0.2.1 Mean/Expectation

- The mean is a weighted average of the values taken by the random variable.
- It is defined as $E\{\mathbf{X}\} = \sum_{i=1}^{n} p_i x_i$
- It is often denoted by μ.
- **Concept Check 3**: Population geneticists are often interested in the mean fitness of a population, $\bar{w}$. Here, the random variable takes on the fitnesses of genotypes and the probabilities are the frequencies of those genotypes. Using the table below, write the formula for the mean fitness of a population. (For now, don't worry about the biological meanings of h and s.)

Outcome	Value of $\mathbf{X}$	Probability
A_1A_1	1	p^2
A_1A_2	1 - hs	$2pq$
A_2A_2	1- s	q^2

$\bar{w} = (1)\ p^2 + (1\text{-}hs)\ 2pq + (1\text{-}s)\ q^2 = 1 - 2pqhs - q^2 s$

Note: this question is based off of the material covered in the Human Variation and Disease Lecture titled "Selection I"

- **Concept Check 4**: We have measured the beaks of 5 subpopulations of finches and recorded our findings in the table below. What is the expected value of beak size for the population of finches? (In other words, what is $E\{\mathbf{X}\}$? Assume only these subpopulations make up the population of finches that we are interested in and there is no variation in beak length within sub-populations.)

Subpopulation Number	Beak length (mm)	Subpopulation size
1	0.4	100
2	0.55	300
3	0.65	150
4	0.35	250
5	0.47	200

0.4*0.1 + 0.55*.3 + 0.65*.15 + 0.35*.25 + 0.47*0.2 = 0.484 mm

Note: this question is based off of the material covered in the Human Variation and Disease Lecture titled "Quantitative Genetics I"

0.2.2 Variance

- The variance of a random variable is the expectation of the squared deviations from the mean.
- It is defined by $\text{Var}\{\mathbf{X}\} = E\{(\mathbf{X} - \mu)^2\} = E\{\mathbf{X}^2\} - E\{\mathbf{X}\}^2$
- It is often denoted by σ^2.
- **Concept Check 5**: Using the table from the previous Concept Check, what is variance in beak size for this population of finches? (Again, assume only these subpopulations make up the population of finches that we are interested in.)

$E\{\mathbf{X}^2\} = 0.4^2 \times 0.1 + 0.55^2 \times 0.3 + 0.65^2 \times 0.15 + 0.35^2 \times 0.25 + 0.47^2 \times 0.2 = 0.24493$

$E\{\mathbf{X}\}^2 = (answer from above)^2 = 0.484^2 = 0.23426$

$\text{Var}\{\mathbf{X}\} = E\{\mathbf{X}^2\} - E\{\mathbf{X}\}^2 = 0.24493 - 0.23426 = 0.01067$

Note: this question is based off of the material covered in the Human Variation and Disease Lecture titled "Quantitative Genetics I"

0.3 Noteworthy discrete random variables

0.3.1 Bernoulli random variable

- The Bernoulli random variable is used when there is one trial of an event where there are only two possible outcomes: success (where $\mathbf{X} = 1$) and failure (where $\mathbf{X} = 0$).
- **Concept Check 6:** The mean of a Bernoulli random variable is $1 \times p + 0 \times q = p$ and the variance is pq

  $E\{\mathbf{X}^2\} = 1^2 \times p + 0^2 \times q = p$

  $E\{\mathbf{X}\}^2 = (mean)^2 = p^2$

  Var = $E\{\mathbf{X}^2\}$ - $E\{\mathbf{X}\}^2 = p - p^2 = p(1-p) = pq$
- **Concept Check 7**: What is a use of the Bernoulli distribution in genetics? In your example, describe what is a "success" and what is a '"failure".

One example is the proportion of haploid chromosomes carrying an allele for a particular Mendelian disease. A "success" is the chromosome carrying the disease allele and "failure" is the chromosome not carrying the disease allele. The probability of success is the frequency of the Mendelian disease-causing variant in the population and the probability of failure is the complement of the probability of success.

0.3.2 Binomial Random Variable

- These random variables represent the number of success in n independent trials when the probability of success for any one trial is p.
- The random variable can take o the values 0, 1, ..., n with probabilities

  Prob $\{\mathbf{X} = i\} = \binom{n}{x}\ p^i\ (1-p)^{n-i}$
- **Concept Check 8**: How are Bernoulli and Binomial random variables related?

  The binomial distribution allows for n number of independent Bernoulli trials.

- **Concept Check 9**: Keeping in mind your answer to Concept Check 2, what is the mean of a binomial random variable? The variance?

  Multiply the mean and then the variance of the Bernoulli distribution by n so that the mean is n and the variance is npq.

- **Challenge 1**: In the Wright-Fisher model that we will study in class, we want to know the proportion of offspring that carry allele A from some population. What factors will be important for us to consider given that the Wright-Fisher model is a binomial sampling distribution?

  Population size to find the number of gametes (2N) and the number of adult chromosomes that carry allele A will allow us to find the probability of j gametes with allele A in the offspring given i gametes with allele A in the adult population. Note: there are several assumptions that allow us to model this as a binomial distribution.

0.3.3 Poisson Random Variable

- Poisson Random Variables are obtained by taking the limit of binomial random variables as $n \to \infty$ and $p \to 0$.

- Poisson random variables can take values 0, 1, ..., ∞ with probabilities

  Prob $\{\mathbf{X} = i\} = \frac{e^{-\mu}\mu^i}{i!}$

- **Concept Check 10:** When can the Poisson distribution be used instead of the Binomial distribution?

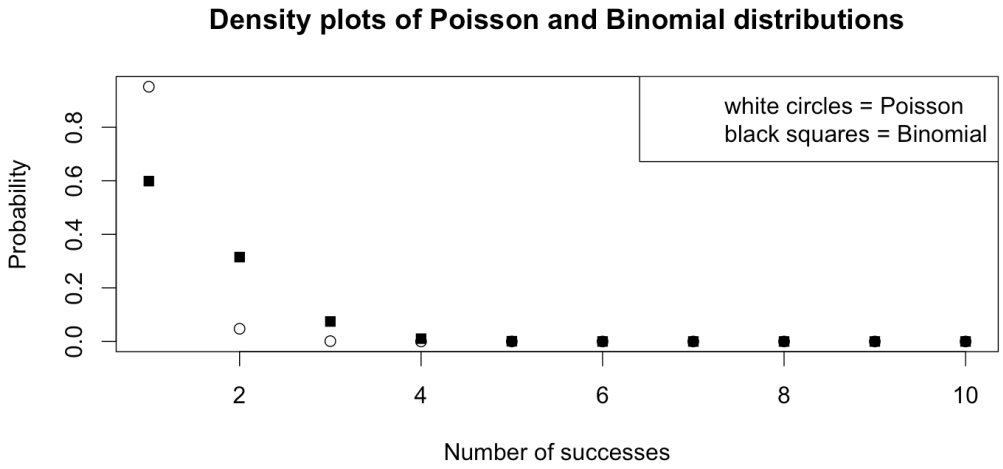
  When p (probability of a success) is much smaller than n (number of trialssample size).

- **Concept Check 11:** How does the Poisson distribution look different than the Binomial distribution?

  The Poisson distribution will have a higher density closer to 0 than the Binomial distribution. We can see that in the plot below.

  Overlay an example of each of the distributions on a plot below. Each distribution should have $\mu = 0.05$ and n= 10:

To create this plot, I used the following commands:
plot(dpois(x=0:9, lambda=0.05), pch=1, main = "Density plots of Poisson and Binomial distributions", xlab = "Number of successes", ylab = "Probability")
points(dbinom(x=0:9, 10, 0.05), pch = 15)
leg.txt ¡- c("white circles = Poisson", "black squares = Binomial")
legend('topright', leg.txt, lty = 1, col = c('white', 'white'))



- **Concept Check 12:** Would you use the Binomial or Poisson distribution to describe the distribution of crossovers along a chromosome in meiosis? Justify your answer

The Poisson distribution because, in general, there is a small number of crossovers compared to the relatively large number of possible crossovers given the entire length of the chromosomes.

0.3.4 Geometric Random Variable

- The geometric random variable can take on values 1, 2, ..., ∞ describes the time until the first success in a sequence of independent trials with the probability of success, p, and the probability of failure, q = *1-p*.
- Prob $\{\mathbf{X} = i\} = q^{i-1}p$

- **Concept Check 13:** Why would you use the geometric distribution instead of the binomial distribution? (Hint: How are the questions answered by each different?)

The geometric distribution is used to get the probability of a given number of failures until a success. This means that as soon as there is a success, we stop recording trials. Alternatively, the binomial distribution allows us to have more than one success and the success(es) can be scattered throughout the independent trials rather than at the end.

0.4 Correlated random variables

- **Concept Check 14:** If Prob$\{\mathbf{X} = x_i\} = p_i$ and Prob$\{\mathbf{Y} = x_i\} = p_j$, then the random variables X and Y are independent if $p_{ij} = p_i \times p_j$
- Covariance is defined as Cov $\{\mathbf{X},\mathbf{Y}\} = E\{(\mathbf{X} - \mu_x)\ (\mathbf{Y} - \mu_y)\}$
- The covariance is a measure of the tendency of two random variables to vary together.

- **Concept Check 15:** If X and Y tend to be large together and small together, then their covariance will be positive
- If two random variables are independent, then their covariance is 0
- The correlation coefficient of X and Y =

$$\frac{Cov(X,Y)}{\sqrt{(Var(X))}\sqrt{(Var(Y))}}$$

- **Concept Check 16:** In the definition of the correlation coefficient, the covariance is scaled by the variance. Why is this "normalization" step important? What additional information can we get from this step? We are comparing the relative size of the covariance to the variances. If the covariance is big, we do not know if this is due to a truly strong correlation between X and Y or if it is being driven by large

variances of X and Y. The correlation coefficient helps us to answer this question. Additionally, after normalization, we can compare correlation coefficients across different studies. While before the covariance may be big, this can be due to

- **Challenge 2:** Suppose we have two sites on a chromosome, A and B, each with two alleles in a haploid individual. If we were to plug in the probability of A_1 and B_1 into the formula for correlation coefficient, describe the information in words that we will get from this.

  This question introduces the concept of linkage disequilibrium. This means that we are looking for how often the alleles A_1 and B_1 "travel together" a.k.a. how often they appear together compared to what would be expected by chance.

  Note: These questions are based on material found in the Human Variation and Disease lectures titled "LD" and "Quantative Genetics I".

0.5 Operations on random variables

- If $\mathbf{Y}$ is a transformed random variable, $\mathbf{Y} = \mathrm{a}\mathbf{X} + \mathrm{b}$,
  then $E\{\mathbf{Y}\} = aE\{\mathbf{X}\} + b$ and the $\mathrm{Var}\{\mathbf{Y}\} = a^2\mathrm{Var}\{\mathbf{Y}\}$.

- $E\{\mathbf{X} + \mathbf{Y}\} = E\{\mathbf{X}\} + E\{\mathbf{Y}\}$ regardless of whether X and Y and independent or dependent.

- $\mathrm{Var}\{\mathbf{X} + \mathbf{Y}\} = \mathrm{Var}\{\mathbf{X}\} + \mathrm{Var}\{\mathbf{Y}\} + 2\mathrm{Cov}\{\mathbf{X},\mathbf{Y}\}$.

- **Concept Check 17:** As we can see, the variance of $\{\mathbf{X} + \mathbf{Y}\}$ relies on the variance of $\{\mathbf{X}\}$, the variance of $\{\mathbf{Y}\}$, and the covariance of $\{\mathbf{X},\mathbf{Y}\}$. Why do you think that is?

  This implies that the variance of the mean increases with the average of the correlations. In other words, additional correlated observations are not as effective as additional independent observations at reducing

the uncertainty of the mean. (Answer from the Wikipedia article titled "Variance".)

- **Concept Check 18:** What is the Var{**X** + **Y**} when **X** and **Y** are independent? Var{**X**} + Var{**Y**} because 2Cov{**X**,**Y**} = 0.

0.6 Noteworthy continuous random variables

0.6.1 Overview

- **Concept Check 19:** Continuous random variables can have values from -∞ to ∞. Therefore, the probability that a continuous random variable takes on 1 particular value is negligible (0)
- As a result, we will write the probability of a continuous random variable in a particular interval:

  Prob $\{a < \mathbf{X} < b\} = \int_a^b f(x)\ dx$

  where $f(x)$ is the probability density function.
- **Concept Check 20:** Prob $\{-\infty < \mathbf{X} < \infty\} = \int_{-\infty}^{\infty} f(x)\ dx$ equals 1

0.6.2 Normal random variable

- **Concept Check 21:** How can you tell if a dataset is well described by the normal distribution?

  We want to use the 68-95-99.7 rule for normal distributions. This means that we should check if approx. 68% of the data falls within 1 standard deviation of the mean, 95% fall within 2 standard deviations of the mean and 99.7% fall within 3 standard deviations of the mean.

- **Concept Check 22:** List some phenotypes that we might model as normally distributed:

  Human height, blood pressure

- The probability density function of the normal distribution is

$$f(x) = \frac{1}{\sqrt{(2\pi)\sigma^2}} exp(-\frac{(x_i - \mu)^2}{2\sigma^2}) \tag{1}$$

- We will often standardize a normal distribution, e.g. N(3, 10), using the transformation Z = $\frac{X-\mu}{\sigma}$

- **Concept Check 23:** What are the advantages of "standardizing" a normal distribution with parameters μ and σ^2 so that it is N(0,1)?

The biggest advantage is the ability to compare across studies, particularly if the original data in each of the studies are modeled by normal distributions with different means and variances.

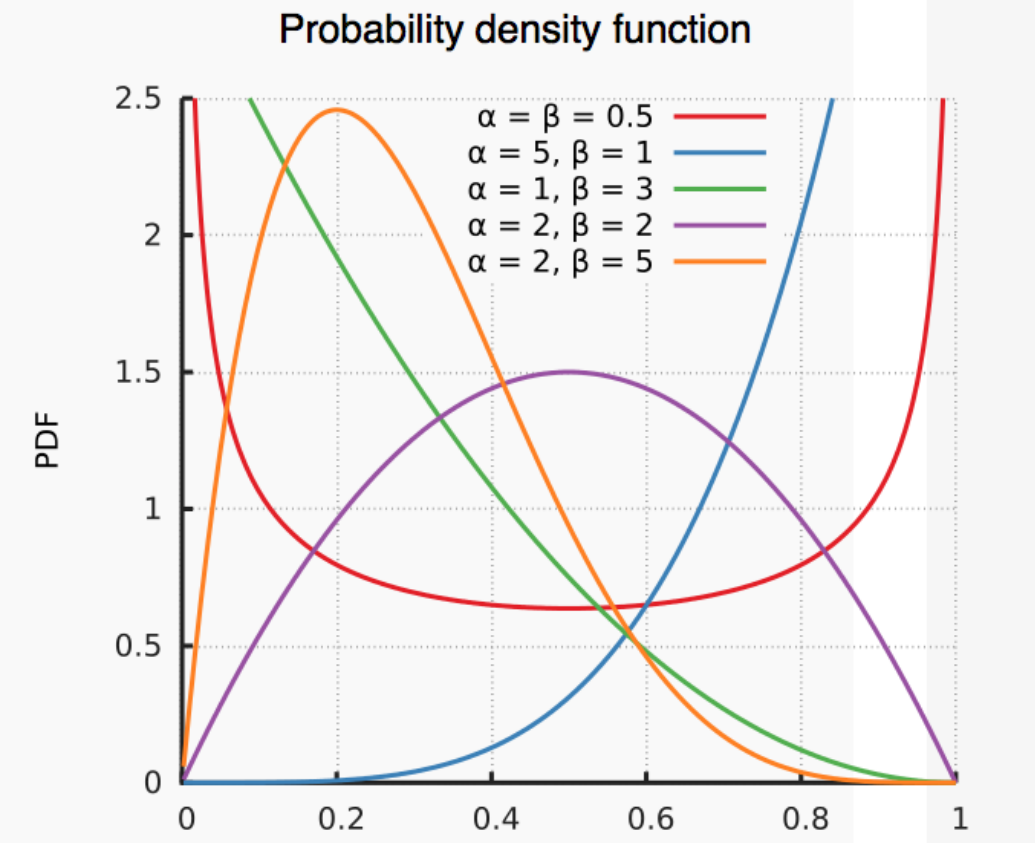
0.6.3 Beta random variable

- The Beta distribution represents a distribution of probabilities. Therefore, 0 <x <1.

- If n is a positive integer, $\Gamma(n) = (n - 1)!$

- The beta random variable is distributed according to the density:

$$f(x) = \frac{\Gamma(\alpha + \beta)x^{\alpha-1}(1 - x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \tag{2}$$



- **Concept Check 24:** Above are PDFs of beta distributions with different alpha and beta values. You will see that on the x-axis are probabilities. How could having a distribution of probabilities be useful in a genetics context?

  We frequently are unsure of the true population probability and we want to find the best estimates of probabilities (e.g. allele frequencies, disease prevalence, etc.)

0.6.4 Bivariate normal random variable (Challenging)

- We can represent two correlated random variables that each follow a normal distribution with the bivariate normal distribution. We can represent these correlated variables with random vector Z. This vector Z will have the vector mean μ and covarariance matrix Σ.

- The probability density for the bivariate normal distribution Z ~ (X,Y) incorporates the means of each of the random variables (X and Y), the variances of the two random variables, and the correlation coefficient, ρ.

$$f(x,y) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

where (μ_x, μ_y) is the mean vector and the variance-covariance matrix is

$$\begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

- The bivariate normal can be generalized to include more than two univariate random variables, called multivariate normal distributions.
- A handful of the journal articles that we read in this class will use the bivariate or multivariate normal distribution (e.g. Pritchard et al. 2000 Genetics 155:945).
- We will also use the moments and properties of the bivariate normal distribution, for example, in linear regression.
- **Concept Check 25:** What is the goal of linear regression? To determine if variable X is a good predictor of variable Y. (Here, a "good predictor" is defined by being better than by chance. Also, there must be a linear relationship between X and Y.)
- The regression of **Y** on **X** is

  $E\ \{\mathbf{Y} \mid \mathbf{X} = \mathrm{x}\ \} = \mu_y + \beta\ (\ x - \mu_x\ )$

  where

  $\beta = \frac{Cov\{\mathbf{X},\mathbf{Y}\}}{\sigma_x^2}$

Concept Check 26: Use concepts that we have covered in this discussion and what you know about linear regression to write each of the parts of this equation in words and how it is related to linear regression.

$E\ \{\mathbf{Y} \mid \mathbf{X} = \mathrm{x}\ \}$ is

The expectation of Y given a particular value of X. This coincides nicely with the goal of regression because we want to be able to give our model a value of X and use that to come up with our best prediction of Y.

μ_y is

The average value of y. This is used because the mean is the best statistic to describe a variable in linear regression.

β is

This is how strong the relationship between X and Y is relative to the variance (spread) of X. In genetics, this is often referred to as the "effect size".

x - μ_x is

This is the difference between the random variable X and the mean of X. This is useful because when you know how far the given value of X deviates from the mean so that you can use this information to figure out how the given value of Y deviates from the mean.

Note: This worksheet was created by Lauren E. Blake using the material from "Population Genetics, A Concise Guide" by John Gillespie 2nd edition and "Introduction to Probability" by D. P. Bertsekas and J. N. Tsitsiklis, 1st edition. Kindly edited by course instructor John Novembre.