Name: _____ Date: _____

Human Variation and Disease: Discussion 1 Notes and Concept Checks

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.

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 $\{\mathbf{X} = x_1\}$? _____
- **Concept Check 2:** In the table above, what is Prob $\{\mathbf{X} = x_i\}$? _____

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 1**: 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

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

- **Concept Check 2**: 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.)

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

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 $\text{Var}\{\mathbf{X}\} = E\{\big(\mathbf{X} - \mu\big)^2\} = E\{\mathbf{X}^2\} - E\{\mathbf{X}\}^2$
- It is often denoted by σ^2.
- **Concept Check 3**: 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.)

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 $\mathbf{X} = 0$).
- The mean of a Bernoulli random variable is _____ and the variance is _____.
- **Concept Check 1**: What is a use of the Bernoulli distribution in genetics? In your example, describe what is a "success" and what is a '"failure".

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

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

- **Challenge**: 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?

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 $\{\mathbf{X} = i\} = \frac{e^i \mu^i}{i!}$
- **Concept Check 1:** When should the Poisson distribution be used instead of the Binomial distribution?

- **Concept Check 2:** How does the Poisson distribution look different than the Binomial distribution?

  Overlay an example of each of the distributions on a plot below:

- **Concept Check 3:** Would you use the Binomial or Poisson distribution to describe the distribution of crossovers along a chromosome in meiosis? Justify your answer

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

0.4 Correlated random variables

- The random variables X and Y are independent if $p_{ij} =$ _____
- Covariance is defined as Cov $\{\mathbf{X,Y}\} = E\{(\mathbf{X} - \mu_x)\ (\mathbf{Y} - \mu_y)\}$
- The covariance is a measure of the tendency of two random variables to vary together.
- **Concept Check 1:** If X and Y tend to be large together and small together, then their covariance will be _____.
- If two random variables are independent, then their covariance is ___.
- The correlation coefficient of X and $Y =$

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

- **Concept Check 2:** 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?

- **Challenge:** 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.

Note: This question is based on material found in the Human Variation and Disease lecture titled "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 1:** As we can see, the variance of 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?

- **Concept Check 2:** What is the $\mathrm{Var}\{\mathbf{X} + \mathbf{Y}\}$ when $\mathbf{X}$ and $\mathbf{Y}$ are independent?

0.6 Noteworthy continuous random variables

0.6.1 Overview

- Continuous random variables can have values from -∞ to ∞. Therefore, the probability that a continuous random variable takes on 1 particular value is _____.
- 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 1:** Prob $\{-\infty < \mathbf{X} < \infty\} = \int_{-\infty}^{\infty} f(x)\ dx$ equals ___.

0.6.2 Normal random variable

- **Concept Check 1:** How can you tell if a dataset is well described by the normal distribution?
- **Concept Check 2:** List some phenotypes that are normally distributed:
- The probability density function of the normal distribution is

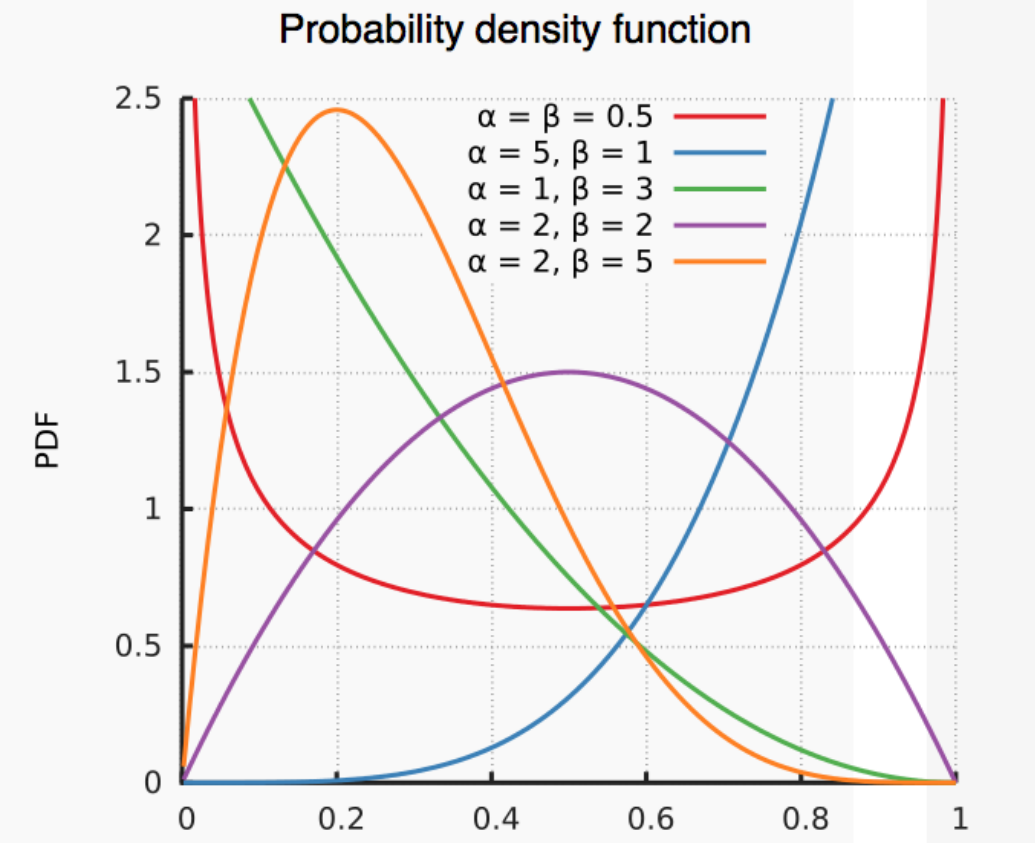
$$f(x) = \frac{1}{\sqrt{(2\pi)\sigma^2}} exp(-\frac{(x_i - \mu)^2}{2\sigma^2}) \quad (1)$$

- We will often standardize a normal distribution, e.g. N(3, 10), using the transformation Z = $\frac{X-\mu}{\sigma}$
- **Concept Check 3:** What are the advantages of "standardizing" a normal distribution with parameters μ and σ^2 so that it is N(0,1)?

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 1:** 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?

0.6.4 Bivariate normal random variable (Challenging)

1. We start with two independent normals, $U \sim \mathrm{N}(\mu, \sigma^2)$ and $V \sim \mathrm{N}(\mu, \sigma^2)$

2. We use of these variables in linear transformations with scalars a, b, c, and d.

   $X = aU + bV,$
   $Y = cU + dV$

3. Now, X and Y each are normally distributed (because they are linear functions of independent normal random variables).

4. Since X and Y are linear functions of the **same** two independent normal random variables, their joint probability distribution is the bivariate normal distribution.

- In the bivariate normal distribution, we need the means of each of the random variables, the variances of the two random variables, and the covariance of the correlation coefficient, ρ.

- A handful of the journal articles that we read in this class will use the bivariate normal distribution (e.g. Pritchard et al. 2000 Genetics 155:945).

- We will more frequently use the moments and properties of the bivariate normal distribution, for example, in linear regression.

- What is the goal of linear regression?

- The regression of $\mathbf{Y}$ on $\mathbf{X}$ is

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

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

Concept Check 1: 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

μ_y is

β is

x - μ_x is

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.