2021 Russell Sage Foundation Summer Institute in Social-Science Genomics

Formula Sheet and Glossary of Useful Terms

This document contains formulas that may prove useful as you work through Problem Sets 1 and 2. Though it is not exhaustive, and you may want or need to use additional formulas (or alternative specifications of the formulas provided here), you may think of it as a quick statistics refresher. We will add to this document for later problem sets.

Formula sheet

Expected value: The expected value of a sum of random variables is just the sum of the expected values:

$$\mathrm{E}(X+Y)=\mathrm{E}(X)+\mathrm{E}(Y)$$

This implies that known, fixed quantities can be removed from the expectation statement. For example, if a and b are constants and X and Y are random variables:

$$\mathrm{E}(aX)=a\mathrm{E}(X)$$

$$\mathrm{E}(aX+bY)=a\mathrm{E}(X)+b\mathrm{E}(Y)$$

$$\mathrm{E}(aX+b)=a\mathrm{E}(X)+b$$

Variance: The variance of a random variable, X, is:

$$\mathrm{Var}(X)\equiv \mathrm{E}\left[\left(X-\mathrm{E}(X)\right)^2\right]=\mathrm{E}(X^2)-[\mathrm{E}(X)]^2$$

Note that, unlike expected values, variance is not a linear operator. Assume that a and b are constants and X and Y are random variables:

$$\mathrm{Var}(aX)=a^2\mathrm{Var}(X)$$

$$\mathrm{Var}(aX+bY)=a^2\mathrm{Var}(X)+b^2\mathrm{Var}(Y)+2ab\mathrm{Cov}(X,Y)$$

For example, note carefully in the following derivation how variance does not behave in the same way as expectation:

Whereas $\mathrm{E}(aX+b)=a\mathrm{E}(X)+b$, instead:

$$\mathrm{Var}(aX+b) = \mathrm{E}[(aX+b)^2] - (\mathrm{E}[aX+b])^2$$

$$= \mathrm{E}[a^2X^2 + 2abX + b^2] - (a\mathrm{E}(X) + b)^2$$

$$= a^2\mathrm{E}(X^2) + 2ab\mathrm{E}(X) + \ b^2 - a^2[\mathrm{E}(X)]^2 - 2ab\mathrm{E}(X) - b^2$$

$$= a^2\mathrm{E}(X^2) - a^2[\mathrm{E}(X)]^2$$

$$= a^2\mathrm{Var}(X)$$

"De-meaned" (centered) variables: "De-meaning" a random variable, X, means subtracting the expected value from every observation:

$$\tilde{X} \equiv X - \mathrm{E}(X)$$

The expected value of a de-meaned variable, $\tilde{X}$, is 0:

$$\mathrm{E}(\tilde{X}) = \mathrm{E}(X - \mathrm{E}(X)) = 0$$

Standardized variables: Centered and transformed to have a standard deviation (and variance) of 1:

$$\check{X} \equiv \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}$$

Covariance: Consider two random variables, X and Y. Their covariance is:

$$\mathrm{Cov}(X,Y) \equiv \mathrm{E}\big[\big(X - \mathrm{E}(X)\big)\big(Y - \mathrm{E}(Y)\big)\big]$$

$$= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$$

Suppose now that X, Y, and Z are random variables, and a is a constant. Then the following rules apply:

$$\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$$

$$\mathrm{Cov}(X+Y,aZ) = \mathrm{Cov}(X,aZ) + \mathrm{Cov}(Y,aZ) = a\mathrm{Cov}(X,Z) + a\mathrm{Cov}(Y,Z)$$

Covariance of two de-meaned random variables, $\tilde{X}$ and $\tilde{Y}$:

$$\mathrm{Cov}\big(\tilde{X},\tilde{Y}\big) =$$
$$= \mathrm{E}\big(\tilde{X}\tilde{Y}\big) - 0$$

$$= \mathrm{E}(\tilde{X}\tilde{Y})$$

Correlation: The correlation between two random variables, *X* and *Y*, is:

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

Correlation of two standardized random variables, $\check{X}$ and $\check{Y}$:

$$\mathrm{Corr}\left(\check{X}, \check{Y}\right) = \frac{\mathrm{Cov}(\check{X}, \check{Y})}{\sqrt{\mathrm{Var}\left(\check{X}\right)\mathrm{Var}(\check{Y})}}$$

$$= \frac{\mathrm{Cov}(\check{X}, \check{Y})}{\sqrt{1 * 1}}$$

$$= \mathrm{Cov}(\check{X}, \check{Y})$$

Discrete random variables: Can take on a finite number of possible values. (Genotype score is an example of a discrete random variable.)

The expected value of a discrete random variable, *X,* is the sum of its possible values weighted by their probabilities of occurring (*p(x)*):

$$\mathrm{E}(X) = \sum_{all\ x} xp(x)$$

The variance of *X* can be found in a similar manner:

$$\mathrm{Var}(X) = \sum_{all\ x} \left(x - \mathrm{E}(X)\right)^2 p(x)$$

Uncorrelated variables: Two random variables, *X* and *Y*, are said to be uncorrelated when:

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y).$$

If *X* and *Y* are uncorrelated, then their covariance and correlation are both zero.

Independent variables: Two random variables, X and Y, are independent if their realizations do not depend on each other, i.e., the probability that $X = x$ is the same regardless of the realization of Y.

If X and Y are independent, then they are uncorrelated; but if they are uncorrelated, they are not necessarily independent.

For two independent random variables X and Y, joint and conditional probabilities are simple to determine:

Intersection: $\mathrm{P}(X \,\&\, Y) = \mathrm{P}(X) * \mathrm{P}(Y)$

Conditional probabilities: $\mathrm{P}(X \mid Y) = \mathrm{P}(X)$

$\mathrm{P}(Y \mid X) = \mathrm{P}(Y)$

Matrices: We'll use the following matrices as examples:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 10 & 9 \\ 8 & 7 \end{bmatrix}$$

- Notation: Matrices are typically denoted with capitalized, bolded Roman letters. Elements of a matrix are referenced with a lowercase letter and subscripted row and column identifiers.
    - Examples:
        - For matrix **A**, element $a_{1,2} = 2$
        - For matrix **B,** element $b_{2,1} = 3$

- Dimension: Matrices have dimension $i \times j$, where i is the number of rows and j is the number of columns.
    - Examples:
        - Matrix **A** has dimension 2×3, $\dim(\mathbf{A}) = 2 \times 3$
        - Matrix **B** has dimension 2×2, $\dim(\mathbf{B}) = 2 \times 2$

- Vector: Special type of matrix for which i or j equals 1. Column vectors (for which j equals 1) are more conventionally used than row vectors.
    - Vector length: Say we have a (column) vector $\boldsymbol{v} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$. Length of $\boldsymbol{v} = \sqrt{3^2 + 2^2} = \sqrt{13}$. We often standardize vectors to have a length of 1. To do so, we divide each element by the vector's original length.

- Transpose (denoted with '): Reverse rows and columns (e.g., $a_{i,j} = a'_{j,i}$).
    - Examples:

$$\mathbf{A}' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \quad \mathbf{B}' = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

- Matrix addition and subtraction: Adding or subtracting matrices proceeds element by element and results in a new matrix of the same dimension as the originals. Note that the matrices to be added or subtracted must be of equal dimension; if they are not, the matrix operation is undefined.
    - Examples:

$$\mathbf{A} + \mathbf{C} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 9 \\ 12 & 15 & 18 \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \text{Undefined, dim}\ (\mathbf{A}) \neq \text{dim}\ (\mathbf{B})$$

$$\mathbf{B} - \mathbf{D} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} 10 & 9 \\ 8 & 7 \end{bmatrix} = \begin{bmatrix} -9 & -7 \\ -5 & -3 \end{bmatrix}$$

- Scalar multiplication: Multiplying a matrix by a scalar proceeds element by element and results in a new matrix of the same dimension as the original.
    - Examples:

$$2\mathbf{A} = 2\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}$$

$$\left(\frac{1}{2}\right)\mathbf{B} = \left(\frac{1}{2}\right)\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1/2 & 2/2 \\ 3/2 & 4/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1 \\ 3/2 & 2 \end{bmatrix}$$

- Matrix multiplication: Multiplying two matrices results in a new matrix. Examples of matrix multiplication and its requirements/properties are provided below.
    - Order matters
        - Example:

$$\mathbf{BD} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 10 & 9 \\ 8 & 7 \end{bmatrix} = \begin{bmatrix} (1*10)+(2*8) & (1*9)+(2*7) \\ (3*10)+(4*8) & (3*9)+(4*7) \end{bmatrix} = \begin{bmatrix} 26 & 23 \\ 62 & 55 \end{bmatrix}$$

$$\mathbf{DB} = \begin{bmatrix} 10 & 9 \\ 8 & 7 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} (10*1)+(9*3) & (10*2)+(9*4) \\ (8*1)+(7*3) & (8*2)+(7*4) \end{bmatrix} = \begin{bmatrix} 37 & 56 \\ 29 & 44 \end{bmatrix}$$

$$\mathbf{BD} \neq \mathbf{DB}$$

-
    - Need # columns of the first matrix to equal # rows of the second matrix). For example, if $\dim(\mathbf{M}_1) = i \times j$ and $\dim(\mathbf{M}_2) = m \times n$, and

$j = m$, then the matrix $\mathbf{M}_1\mathbf{M}_2$ has dimension $i \times n$. However, if $j \neq m$, then the matrix-multiplication operation is undefined.

- Examples:

$$\mathbf{AB} = \text{undefined}: \dim(\mathbf{A}) = 2 \times 3, \dim(\mathbf{B}) = 2 \times 2, 3 \neq 2$$

$$\mathbf{A}'\mathbf{B} = \text{defined}: \dim(\mathbf{A}') = 3 \times 2, \dim(\mathbf{B}) = 2 \times 2, \dim(\mathbf{A}'\mathbf{B}) = 3 \times 2$$

$$\mathbf{A}'\mathbf{B} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} (1*1)+(4*3) & (1*2)+(4*4) \\ (2*1)+(5*3) & (2*2)+(5*4) \\ (3*1)+(6*3) & (3*2)+(6*4) \end{bmatrix} = \begin{bmatrix} 13 & 18 \\ 17 & 24 \\ 21 & 30 \end{bmatrix}$$

- Determinant of a 2x2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
  - Example: $\det(\mathbf{B}) = (1)(4) - (2)(3) = 4 - 6 = -2$

- Inverse of a 2x2 matrix: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$
  - Example:

$$\mathbf{B}^{-1} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^{-1} = \frac{1}{(1)(4)-(2)(3)}\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} \frac{4}{-2} & -\frac{2}{-2} \\ -\frac{3}{-2} & \frac{1}{-2} \end{bmatrix}$$

$$= \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix}$$

Bayes' Rule:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

Glossary

Allele: The various sequences of DNA nucleotides for a locus or gene. E.g., the ABO blood system locus has the A, B, and O alleles and the Rhesus (Rh) blood group locus has positive (+) and negative (-) alleles. A gene or locus that has more than one common allele is called a polymorphism.

Locus: A continuous section (i.e., location) of DNA on a chromosome, sometimes a single SNP.

Genome: The total genetic information for an individual organism or for a species. The human genome consists of about 3 billion nucleotides.

Genotype: Genotype at a particular locus is determined by the combination of alleles at that locus--e.g., the AO genotype at the ABO locus.

Heritability: Proportion of observed individual differences in a trait attributable to genetic individual differences; equivalently, the proportion of phenotypic variance explained by genotypic variance. Broad sense heritability includes all types of gene action in the genetic variance. Narrow sense heritability includes only the additive effects of genes in the genetic variance.

Methylation: In genetics, the attachment of methyl groups to nucleotides, especially cytosine, resulting in a reduced transcription of genes.

Mutation: An irregular change in the DNA or a "spelling error" in the nucleotide sequence during cell division. Germinal mutations occur in the production of sperm or egg and are transmitted to the next generation. Somatic mutation, far more common than germinal mutations, influence all other cells of the organism. Mutations may affect only a single nucleotide (point mutation) or large sections of DNA up to a whole chromosome (e.g., trisomy 21 that causes Down's syndrome).

Polygenic: More than a single locus influences the phenotype. Sometimes used to denote the possibility that a large number of loci influence the phenotype.

Phenotype: An observable trait, i.e., an outcome that may be affected by genes.

Minor allele frequency (MAF): For variants with just two types of alleles (biallelic variants), the MAF is the frequency of the less common allele in a given population.