AMS 315
Data Analysis
Chapter Four Study Guide
Probability and Probability Distributions
Spring 2023

Context

Random variables are the principal tool of statistics. Probability theory, as developed in AMS 310 and 311, documents the properties of random variables and is the foundation of statistics. My expectation of you is that you will review this material from your prerequisite class. If the material in this chapter is new to you, then you should read the relevant sections of the text carefully. I will review the contents of each section in this handout to help you prioritize your studying.

Chapter Four

4.1. Introduction

This section is introductory. You should review the discussion of classic equiprobable models, relative frequency, and subjective interpretations of probability.

4.2. Finding the probability of an event

You should review the use of statistical packages to generate random data.

4.3. Basic event relations and probability laws

This section reviews the high points of axiomatic probability theory. A key result that is of importance here is the theorem that the probability of the union of two events, *A* and *B*, is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This leads immediately to Boole's inequality, which states that $P(A \cup B) \leq P(A) + P(B)$ for any two events *A* and *B*. This will be used extensively in lecture when we discuss multiple comparisons and in your second computing project.

4.4. Conditional probability and independence

This section contains definitions of fundamental importance.

4.5. Bayes' formula

The general form of Bayes' theorem is the basis of an increasing popular set of statistical procedures (called Bayesian statistics). The definition of prior and posterior probabilities is important. Most problems on your examinations that use Bayes' theorem will have at least three conditioning possibilities. Problems 1 and 2 below are examples.

4.6. Variables: discrete and continuous

This section was covered in lecture.

4.7. Probability distributions for discrete random variables

You should review this section

4.8. Two discrete random variables: the binomial and the Poisson

You should review this section carefully. When the random variable X has the binomial distribution on n trials with probability of success p with $0 < p < 1$, $E(X) = np$, $\text{var}(X) = np(1-p)$, and $\text{var}(X) < E(X)$. When the random variable Y has the Poisson distribution with mean μ, $E(Y) = \mu$, $\text{var}(Y) = \mu$, and $E(Y) = \text{var}(Y)$. In lecture, I mentioned "over-dispersed" discrete distributions. For example, if W has a negative binomial distribution, $\text{var}(W) > E(W)$.

4.9. Probability distributions for continuous random variables

This section discusses how to calculate probabilities from probability density functions.

4.10. A continuous probability distribution: the normal distribution

This section gives detailed instructions on calculating normal probabilities. This is a question type that will appear on your final examination. Review it carefully.

4.11. Random sampling

This section discusses how to obtain a random sample operationally. It is valuable background information. I will not test this material directly.

4.12. Sampling Distributions

This section discusses the distribution of the arithmetic average of a random sample of size n and the Central Limit Theorem. This material was covered in lecture. It is of fundamental importance and will be the basis of about 40% of your examination questions.

4.13. Normal Approximation to the binomial

Review this section for your own information.

4.14. Evaluating whether or nor a population distribution is normal

This section gives you procedures for testing for normality. You should consider them in your computing assignments.

4.15. Research study: inferences about performance-enhancing drugs among athletes

Read this section. It will not be tested directly.

4.16. Statistical package instructions

Review this section quickly. Each statistical package performs similar computations. Make sure that you understand how to do these tasks in the package that you have choses.

4.17. Summary and key formulas

Each formula listed is worth knowing.

Some Past Examination Problems from Chapter 4:

1. An individual has one of three genotypes called *A*, *B*, and *C*, respectively, for a gene associated with disease *X*. The probability that an individual has genotype *A* is 0.64; the probability that an individual has genotype *B* is 0.32; and the probability that an individual has genotype *C* is 0.04. The probability that an individual with the *A* genotype is affected with disease *X* is 0.05. The probability that an individual with the *B* genotype is affected with disease *X* is 0.80. The probability that an individual with the *C* genotype is affected with disease *X* is 0.99.
    a. What is the probability that an individual is affected with disease *X*?
    b. Given that an individual has disease *X*, what is the probability that the individual is genotype *B*?

    Answer: The probability of being affected is 0.3276. The probability that an affected individual has genotype *B* is 0.7814.

2. An individual may have 0, 1, or 2 alleles that affect susceptibility to a disease *D*. The probability that an individual has 0 susceptibility alleles is $(1-p)^2$, $0 < p < 1$; the probability of 1 susceptibility allele is $2p(1-p)$; and the probability of 2 susceptibility alleles is p^2. The probability that an individual with 0 susceptibility alleles has the disease *D* is β; the probability that an individual with 1 susceptibility allele has the disease *D* is βr; and the probability that an individual with 2 susceptibility allele has the disease *D* is βr^2, where $0 < \beta < 1$, $r > 1$, and $\beta r^2 < 1$. Find the probability that an individual with the disease has 0 susceptibility alleles; that is, find the probability of 0 susceptibility alleles given that the individual has the disease. This problem is worth 40 points.

    Answer: $P(0 \text{ susceptibility alleles} \mid D) = \dfrac{(1-p)^2}{(1-p+rp)^2}$

3. Let *E*, *F*, and *G* be three events, each with positive probability. Prove or disprove that $P(E \mid F) = P(E \mid FG)P(G \mid F) + P(E \mid FG^C)P(G^C \mid F)$. Answer: The result is true.

4. Let *A* and *B* be two events, each with positive probability. Prove or disprove each of the following statements:

a. $P(A \mid B) + P(A^c \mid B) = 1$

b. $P(A \mid B) + P(A \mid B^c) = 1$

Answer: a. is true, and b. is not.

5. The random variable *X* has expected value μ_X and variance $\sigma_X^2 < \infty$. The random variable *Y* has expected value μ_Y and variance $\sigma_Y^2 < \infty$. The random variable *W* has expected value μ_W and variance $\sigma_W^2 < \infty$.These random variables are not necessarily independent. That is, the covariances of the pairs of random variables are given by $\text{cov}(X,Y) = \sigma_{XY}$, $\text{cov}(X,W) = \sigma_{XW}$, and $\text{cov}(Y,W) = \sigma_{YW}$. Let $V = aX + bY + cW$. Find $E(V)$ and $\text{var}(V)$.

Answer: $E(V) = a\mu_X + b\mu_Y + c\mu_W$, and

$\text{var}(V) = a^2\sigma_X^2 + b^2\sigma_Y^2 + c^2\sigma_W^2 + 2ab\sigma_{XY} + 2ac\sigma_{XW} + 2bc\sigma_{YW}$.

6. A research team has observed *n* values, $x_1, x_2, \ldots, x_n$. As usual, $\bar{x}_n = \dfrac{\sum_{i=1}^{n} x_i}{n}$. Prove or disprove that $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = (\sum_{i=1}^{n} x_i^2) - n(\bar{x}_n)^2$.

Answer: The result is true.

7. The random variables W_1 and W_2 are a random sample of 2 drawn from the random variable *W* which has expected value μ_W and variance $\sigma_W^2 < \infty$. Find $E(W_1 - W_2)$ and $E[(W_1 - W_2)^2]$. Prove your result.

Answer: $E(W_1 - W_2) = 0$; $\text{var}(W_1 - W_2) = 2\sigma_W^2$.

8. The random variable *Y* has $E(Y) = \mu_Y$ and $\text{var}(Y) = \sigma_Y^2 < \infty$, and θ is a constant. Prove or disprove $E[(Y - \theta)^2] = \text{var}(Y) + (\mu_Y - \theta)^2$.

Answer: The result is correct.

9. The random variables $Y_1, Y_2, \ldots, Y_n$ are independently distributed with expected value μ and finite variance σ^2. What is $E(\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2)$, where $\bar{Y}_n = \dfrac{\sum_{i=1}^{n} Y_i}{n}$? Prove your answer. Note that this is a hard problem that would not be asked in an examination.

Answer: $(n-1)\sigma^2$.