STAT 109 Lecture Notes

4 Probability Distributions

4.1 Introduction to Probability Distributions

Any variable may take on a certain number of values:

- If the possible values a variable can take on can be written in a list, that variable is discrete.
- If the possible values for that variable are uncountably infinite (say, if they could be any decimal between 0 and 5), we say that variable is continuous.

Definition 1: A variable whose value is based on the outcome of a random event is called a random variable. We usually use capital letters, like X, to denote random variables, and we use lower case letters, like x, to denote specific values that the variable X can have.

Definition 2: A random variable whose outcomes can be listed is called a discrete random variable. A random variable whose outcomes cannot be listed is called a continuous random variable. We will discuss continuous random variables later in this chapter, but for now, think about a variable that can take on ANY value between 5 and 10 (not just 5, 6, 7, 8, 9, and 10 - I mean it could take on any decimal value too, including non-repeating, non-terminating decimals; there's no way we can possibly list all possible decimal values between 5 and 10 - it is literally impossible).

Example 4.1: Let's play a game. It costs \$5 to play. I will shuffle a deck of cards. You will draw a card from the shuffled deck. If you draw the Ace of Hearts, you win \$100. If you draw any of the other aces, you win \$10. If you draw any other heart, you win \$5. If you draw any other card, you do not win any money. What are the possible values for the amount of money you can win (take into account the fact that you have to pay to play)? What is the probability of winning each amount?

Definition 3: The collection of outcomes and their probabilities for a random variable is called the probability distribution of the random variable.

Definition 4: The expected value of a discrete random variable measures, on average, what you can expect the outcome of the random phenomenon to be. This measures the *center* of the discrete random variable. If X is the name of the random variable and x are the values that X can be, we compute the expected value $E(X)$ as follows:

$$\mu = E(X) = \sum x \cdot P(x)$$

Example 4.2: Using the probability model we constructed for the game in Example 1, calculate the expected value of the random variable. What does "expected value" mean in this context?

Definition 5: The standard deviation of a discrete random variable measures how much the outcome of the random phenomenon tends to vary from the expected value. This measures the *spread* of the discrete random variable. Just like with quantitative variables, standard deviation is denoted by σ (or sometimes $\text{SD}(X)$ if X is the name of our random variable), and it is equal to the square root of the variance: $\text{SD}(X) = \sqrt{\text{Var}(X)}$. Remember that with quantitative variables, we calculated the variance: $\sigma^2 = \dfrac{\sum(x-\overline{x})^2}{n-1}$. It is similar for discrete random variables:

$$\sigma^2 = \text{Var}(X) = \sum \left[(x - E(X))^2 \cdot P(x)\right]$$

Example 4.3: Calculate the standard deviation of the random variable in Example 1.

Example 4.4: Compute the expected value, variance, and standard deviation for the following game based on random outcomes: We roll a die. You lose 5 points for rolling a 1. You earn zero points for rolling a 2 or 3. You earn 5 points for rolling a 4 or 5. You earn 50 points for rolling a 6.

Example 4.5: Find the expected value and the standard deviation of the probability distribution given below for the discrete random variable X.

x	$P(X = x)$
11	0.14
-9	0.42
24	0.06
-60	0.38

5.2 The Normal Model as a Probability Distribution

Definition 6: For *continuous* random variables, the concept of "probability distribution" from discrete random variables does not translate (we can no longer write a list of the values in a table, because there are infinitely many possible values). Instead, we consider probability density functions (pdf). In order for a function to be a pdf, it must:

1. have total area under the graph of the function (between the graph of the function and the x-axis) equal to 1, and
2. the function must never be negative (the graph is always above the x-axis).

Do we know of any functions like this?

Probability density functions give information about probability based on the area. For any pdf $f(x)$ of a random variable X,

1. $P(X < x)$ is equal to the area under $f(x)$ to the left of x
2. $P(X > x)$ is equal to the area under $f(x)$ to the right of x
3. $P(x_1 < X < x_2)$ is equal to the area under $f(x)$ between x_1 and x_2
4. $P(X = x)$ is equal to ZERO (for an individual value, this corresponds to the area of a line segment, which is 0)

Example 5.6: Suppose Z is a random variable with pdf given by the graph of $N(0, 1)$.

1. What is $P(Z < 0.5)$?

2. What is $P(1.2 < Z < 1.83)$?

3. What is $P(Z \geq -0.62)$?

Example 5.7: I have found that my average monthly spending on groceries is $217, with a standard deviation of $90. If we assume that my spending on groceries is random from month to month and is normally distributed, answer the following questions.

1. What is the probability that I will spend more than $200 on groceries this month?

2. What is the probability that I will spend less than $300 on groceries this month?

3. What is the probability that I will spend between $200 and $250 on groceries this month?

4. What is the probability that I will spend more than $400 on groceries this month?

Example 5.8: The manufacturer of a box of cookies states on the outside of each box that the product contents weigh 16 ounces. The population of boxes of cookies has a mean of 16.25 ounces and a standard deviation of 0.30 ounces.

The manufacturer would like to ensure that the contents of each box weigh between 15.90 and 16.90 ounces. Less than 15.90 ounces may lead to customer satisfaction issues, while more than 16.90 ounces results in too much product being given away in each box. What is the probability of selecting a box of cookies with product content weighing between 15.90 and 16.90 ounces?

5.4 Sampling Distributions

Suppose we have a variable y. It may be qualitative or quantitative.

If y is qualitative, say that we know that p percent of the population has a certain value. (For example, our variable is "housing status," our population is "Mercyhurst undergraduate students," and we know that 64% of the population has the value "on-campus.") If y is quantitative, say that we know that among the entire population, y has mean μ. (For example, our variable is "number of mobile subscriptions," our population is "households in the United States," and we know that the mean for the total population, as of 2013, is 2.8.)

Now, suppose we start taking random samples of our population of size n. Each of these samples will have a proportion with a certain value if y is qualitative (we'll call this value $\widehat{p}$) and its own mean if y is quantitative (we'll call this value $\overline{y}$).

If we plot the individual values of a single sample, we have the *distribution of the sample.*

Definition 7: If we graph the values of $\widehat{p}$ for a qualitative variable (or $\overline{y}$ for a quantitative variable) that come from all (or a LOT of) possible random samples from the population, this is called a sampling distribution.

Example 5.9: Consider the Excel file posted to the course website with sampling distributions for some of the variables in our first day of class survey.

The Big Idea: If we know the shape, center, and spread of the sampling distribution of a variable, we will be able to determine when a random sample is representative and when it is very different from the rest of the population. To do this, we will find an appropriate model for the sampling distribution, called a sampling distribution model.

The Central Limit Theorem. The mean of a random sample has a sampling distribution whose shape can be approximated by the Normal model.

It turns out that the Normal model is a great model for sampling distributions (and the bigger your sample size n, the closer the distribution is to the model). We only need to know the following information to apply the Normal model to a sampling distribution:

- n is large, and
- the samples are independent of each other

Unfortunately, sometimes these assumptions are difficult to check, so we'll settle for these easier-to-check conditions when testing whether or not the Normal model is appropriate:

- *Randomization:* Are the samples truly chosen randomly?
- *10%:* Is the sample size n less than 10% of the total population? (If the population is large, this is pretty easy...)
- *Success/Failures:* If y is qualitative, are np and nq both greater than or equal to 10?

When these assumptions and conditions are met, the Central Limit Theorem tells us the sampling distribution is approximately Normal. But to use the Normal model, we need a mean and a standard deviation. What do we do? **It is a fact that the mean of the sampling distribution is equal to the mean of the population.** This means the means of samples will be close to the mean of the actual population, and the proportion of samples will be close to the proportion of the actual population.

For qualitative variables with population proportion p, if the above assumptions/conditions are met, then the sampling distribution is close to the Normal model with mean p and standard deviation $\sqrt{\frac{pq}{n}}$.

For quantitative variables with population mean μ and population standard deviation σ, if the above assumptions/conditions are met, then the sampling distribution is close to the Normal model with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Examples

Example 5.10: We don't know it, but 52% of voters plan to vote "Yes" on an upcoming school budget. We poll a random sample of 300 voters.

1. What might the percentage of Yes-votes appear to be in our poll?

2. What is the standard deviation for this percentage?

3. What is the probability less than 50% of our sample plans to vote "Yes"?

Example 5.11: Math SAT scores should have mean 500 and standard deviation 100.

1. What do you expect to be the mean score for a random sample of 20 students?

2. What is the standard deviation for this expected score?

3. What is the probability a sample of 20 students has an average SAT math score between 550 and 660?

Example 5.12: It is generally believed that electrical problems affect about 14% of new cars. An automobile mechanic conducts diagnostic tests on 128 new cars on the lot.

1. How many successes (electrical failures) do you expect?

2. What is the probability that more than 30 of the new cars have electrical problems?

Example 5.13: One study found that the average age difference between married couples in the UK is approximately 2.24 years with a standard deviation of 4.1 years. Answer the following questions, assuming that the normal model is appropriate for the sampling distribution.

1. What percentage of marriages have an age difference larger than 5 years?

2. Suppose we will be considering random samples with $n = 50$. What percentage of random samples will have a mean age difference larger than 4 years?

3. What is the probability that a sample ($n = 50$) has a mean age difference between 1 and 2 years?

Example 5.14: Based on tests of the Chevy Cobalt, the fuel efficiency of the car is normally distributed with an average of 32 miles per gallon and a standard deviation of 3.5 miles per gallon.

1. What is the probability that a randomly selected Cobalt gets more than 34 miles per gallon on average?

2. Ten Cobalts are randomly selected. What is the probability that the average fuel efficiency of these cars is more than 34 miles per gallon?

3. Twenty Cobalts are randomly selected. What is the probability that the average fuel efficiency of these cars is more than 34 miles per gallon?

4. Fifty Cobalts are randomly selected.

   (a) What is the probability that the average fuel efficiency of these cars is more than 34 miles per gallon?

   (b) What is the probability that the average fuel efficiency of these cars is between 25 and 30 miles per gallon?

   (c) What is the probability that the average fuel efficiency of these cars is less than 29 miles per gallon?

   (d) What is the probability that the average fuel efficiency of these cars is between 30 and 34 miles per gallon?