Lecture 14:
Lesson and Activity Packet

MATH 232: Introduction to Statistics

November 16, 2016

Homework and Announcements

- Homework 13 due in class Monday (check Canvas)
- Post on book discussion forum on Canvas before 11:59 p.m. Wednesday next week
- Submit book summary on Canvas before 11:59 p.m. Wednesday next week
- Submit exam corrections in class on Friday (optional)
- Submit election make-up on Canvas before 11:59 p.m. on Friday (optional)

Today

- Descriptive vs. Inferential Statistics
- Random Variables (Discrete and Continuous)
- Probability Distributions
- Mean, Variance, and Standard Deviation for a Probability Distribution

Question 1

How would you compute the mean roll of a die? The standard deviation?

In the first part of the course, we would have rolled the die a number of times, and put together a data set—perhaps representing it using a frequency table. For example, you might have the following results after 40 rolls of the die:

x	1	2	3	4	5	6
f	8	10	9	12	11	10

In this case, the mean and standard deviation can be computed using the formulas from Chapter 3 of the text. The mean is $\bar{x} = 3.5$ and the standard deviation is $s = 1.7$.

But instead of constructing a frequency distribution and finding a mean and standard deviation based on actual observed data, we can develop a theoretical model of the distribution of results, and then find the mean and standard deviation of this theoretical model:

x	1	2	3	4	5	6
$P(x)$	1/6	1/6	1/6	1/6	1/6	1/6

In essence, the **probability distribution** above describes the relative frequency table for a die rolled an *infinite number of times*. With this knowledge of the population of outcomes, we are able to find important characteristics, such as the mean and the standard deviation. The remainder of this class and the core concepts of **inferential statistics** are based on probability distributions.

For the particular probability distribution corresponding to the rolling of a die, the mean is $\mu = 3.5$ and the standard deviation is $\sigma = 1.7$. We will learn how to compute these by the end of this class, but first, we begin with some key concepts.

Definition 1 (*Random Variable*)

A random variable is a variable (typically represented by x that has a single numerical value, determined by chance, for each outcome of a procedure.

Example 1

- *The age of a rock sample randomly chosen from an archaeological site;*
- *The score a randomly chosen student earns on an exam;*
- *The age or weight of a randomly chosen respondent of a survey*

Definition 2 (*Probability Distribution*)

A probability distribution is a description that gives the probability for each value of the random variable. It is often expressed in the format of a graph, table, or formula.

Example 2

For the example of tossing a die, the random variable is ____________________, and the probability distribution is

x	1	2	3	4	5	6
$P(x)$	1/6	1/6	1/6	1/6	1/6	1/6

Note that random variables are typically not perfectly random. Strictly speaking, neither are they variables. But some people still refer to avocados as “alligator pears”, and avocados are neither alligators nor pears. That is to say, be cautious of the nomenclature.

Random variables come in two categories: discrete and continuous.

Definition 3 (*Discrete Random Variable*)

A discrete random variable can take on a finite number of values or a countably infinite number of values (as many values as there are whole "counting" numbers). For nearly all discrete random variables, the possible values form a subset of the integers.

Definition 4 (*Continuous Random Variable*)

A continuous random variable arises when we deal with quantities measured on a continuous scale without gaps or interruptions, and can take on uncountably many values.

Example 3

- ***Discrete***: *Let x be the number of eggs that a bird lays in a day. This is a discrete random variable, because its only possible values are 0, 1, 2, 3,.... No bird can lay 2.8546584 eggs, for example, which would have had to have been possible had the variable been continuous.*
- ***Discrete***: *The head count of students in class today is a whole number, and is therefore a discrete variable.*
- ***Continuous***: *Let x be the time it takes for a runner to finish a race. This is a continuous random variable, because it can have any value over a continuous span; a runner might finish a hundred-meter dash in a time between 10 and 30 seconds. It would be possible to get 12.354452 seconds, because the time is not restricted to discrete numbers of seconds like 10, 11, 12,.... [Finite precision of measurement devices may limit the practical ability to observe any continuous random variable in real life. But pretend we have infinitely precise measuring equipment if you must.]*
- ***Continuous***: *The measurement of voltage for a particular smoke detector battery can be any value between 0 and 9 volts. It is therefore a continuous random variable.*

Group Exercise 1

Suppose you buy one lottery ticket every week over a year (for a total of 52 tickets). Over those weeks, you count the number of times you win something. In this context, what is the random variable, and what are its possible values?

Group Exercise 2

Determine whether each random variable is discrete or continuous.

- *The number of people who are, at this moment, driving a car in Finland;*
- *The weight of all the pistachios in Turkey;*
- *The height of the last airplane that departed from Harriman & West Airport;*
- *The number of cans of soft drinks that you consumed in the last year;*
- *The cost of making a randomly selected movie*
- *The running time of a randomly selected movie;*

For now, we will consider only discrete random variables, and will pick up with continuous random variables after the end of Chapter 5.

There are various ways of representing a probability distribution of a discrete random variable. We saw a way of representing the distribution of the numbers rolled on a die by a table of values.

Another way of representing the probability distribution is with the **probability histogram**. This is best illustrated with an example.

Example 4

Consider the offspring of peas from parents both having the green/yellow combination of pod genes. Under these conditions, the probability that the offspring has a green pod is 3/4, or 0.75 (remember Punnett squares?). That is, $P(\text{green}) = 0.75$. *If five such offspring are obtained, and if we let* x *be the number of peas with green pods among 5 offspring peas, then* x *is a random variable because it depends on chance.*

In the table is a probability distribution giving the probability for each value of the random variable x. *[We will see how to use a binomial distribution to compute these values before the end of Chapter 5.]*

x	0	1	2	3	4	5
$P(x)$	0.001	0.015	0.088	0.264	0.396	0.237

This probability distribution can also be represented by the probability histogram:

Theorem 1 (*Requirements for a probability distribution*)

1. *The sum of all probabilities in the distribution of a discrete random variable must be 1. [That is,* $\sum_i P(x_i) = 1$*, where* x_1*,* x_2*, . . . are the values the variable can take on.]*
2. *Each probability in the distribution of a discrete random variable must be between 0 and 1 inclusive. [That is,* $0 \leq P(x_i) \leq 1$ *for all values* x_1*,* x_2*, . . . that the variable can take on.]*

- Requirement 1 arises from the fact that the random variable x represents **all** possible events in the entire sample space, so we are certain (with probability 1) that one of the events is bound to occur.

- Sometimes rounding errors force a particular distribution to violate Requirement 1 by a small amount. For example, in the given distribution corresponding to green/yellow pea pods, the probabilities sum to 1.001. This is fine.

Group Exercise 3

Based on a survey conducted by a high-school student you're tutoring, the probabilities for the number of cell phones in use per household is shown in the following table:

x	0	1	2	3
$P(x)$	0.19	0.26	0.33	0.13

Does this table constitute a probability distribution? Why or why not?

Group Exercise 4

Does $P(x) = \frac{x}{10}$*, where* $x \in \{0, 1, 2, 3, 4\}$*, constitute a probability distribution? Why or why not?*

The mean, variance, and standard deviation for a probability distribution can be computed using the following formulas:

Theorem 2

- *Mean:* $\mu = \sum_i [x_i \cdot P(x_i)]$
- *Variance:* $\sigma^2 = \sum_i \left[(x_i - \mu)^2 \cdot P(x_i)\right]$
- *Standard Deviation:* $\sigma = \sqrt{\sum_i [(x_i - \mu)^2 \cdot P(x_i)]}$

Example 5

Using the data in the distribution of the random variable predicting the number of peas with green pods in a sample of five randomly selected peas...

x	$P(x)$	$x \cdot P(x)$	$(x - \mu) \cdot P(x)$
0	*0.001*	$0 \cdot 0.001 = 0.000$	$(0 - 3.752)^2 \cdot 0.001 = 0.014078$
1	*0.015*	$1 \cdot 0.015 = 0.000$	$(1 - 3.752)^2 \cdot 0.015 = 0.113603$
2	*0.088*	$2 \cdot 0.088 = 0.000$	$(2 - 3.752)^2 \cdot 0.088 = 0.270116$
3	*0.264*	$3 \cdot 0.264 = 0.000$	$(3 - 3.752)^2 \cdot 0.264 = 0.149293$
4	*0.396*	$4 \cdot 0.396 = 0.000$	$(4 - 3.752)^2 \cdot 0.396 = 0.024356$
5	*0.237*	$5 \cdot 0.237 = 0.000$	$(5 - 3.752)^2 \cdot 0.237 = 0.369128$
Total		$\mu = 3.752$	$\sigma^2 = 0.940574$

Group Exercise 5

Based on past results found in the Information Please Almanac, *there is a 0.1919 probability that a baseball World Series contest will last four games, a 0.2121 probability that it will last five games, a 0.2222 probability that it will last six games, and a 0.3737 probability that it will last seven games.*

- *Does the given information describe a probability distribution?*
- *If it does describe a probability distribution, find the mean and standard deviation for the numbers of games in World Series contests.*