Chapter 3 Part 1

04/26/2021

’‘

Distribution Theory - Learning Objectives

After finishing this chapter, you should be able to:

- Write definitions of non-normal random variables in the context of an application.
- Identify possible values for each random variable.
- Identify how changing values for a parameter affects the characteristics of the distribution.
- Recognize a form of the probability density function for each distribution.
- Identify the mean and variance for each distribution.
- Match the response for a study to a plausible random variable and provide reasons for ruling out other random variables.
- Match a histogram of sample data to plausible distributions.
- Create a mixture of distributions and evaluate the shape, mean, and variance.

Libraries Needed (none new?)

```r
# Packages required for Chapter 3
library(gridExtra)
library(knitr)
library(kableExtra)
library(tidyverse)
```

Introduction

- What if it is not plausible that a response is normally distributed?
- You may want to construct a model to predict whether a prospective student will enroll at a school or model the lifetimes of patients following a particular surgery.
- In the first case you have a binary response (enrolls (1) or does not enroll (0)), and in the second case you are likely to have very skewed data with many similar values and a few hardy souls with extremely long survival.
- These responses are not expected to be normally distributed; other distributions will be needed to describe and model binary or lifetime data.
- Non-normal responses are encountered in a large number of situations. Luckily, there are quite a few possibilities for models.

Discrete Random Variables

- A discrete random variable has a countable number of possible values
  - Ex: we may want to measure the number of people in a household
  - Ex: model the number of crimes committed on a college campus.
- With discrete random variables, the associated probabilities can be calculated for each possible value using a **probability mass function** (pmf).
- A pmf is a function that calculates $P(Y = y)$, given each variable's parameters.

Binary Random Variable

- Consider the event of flipping a (possibly unfair) coin.
- If the coin lands heads, let's consider this a success and record $Y = 1$.
- A series of these events is a **Bernoulli process**, independent trials that take on one of two values (e.g., 0 or 1).
- These values are often referred to as a failure and a success
    - the probability of success is identical for each trial.

Binary Random Variable

- Suppose we only flip the coin once, so we only have one parameter, the probability of flipping heads, p.
- If we know this value, we can express $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$.
- In general, if we have a Bernoulli process with only one trial, we have a **binary distribution** (also called a **Bernoulli distribution**) where

$$P(Y = y) = p^y(1 - p)^{1-y} \quad \text{for} \quad y = 0, 1.$$

- If $Y \sim \mathrm{Binary}(p)$, then Y has mean $\mathsf{E}(Y) = p$ and standard deviation $\mathsf{SD}(Y) = \sqrt{p(1 - p)}$.

Example 1

- Your playlist of 200 songs has 5 which you cannot stand. What is the probability that when you hit shuffle, a song you tolerate comes on?
- We want to understand the example and be able to translate that into notation/model it with a distribution

Example 1

- Assuming all songs have equal odds of playing, we can calculate $p = \frac{200-5}{200} = 0.975$
  - so there is a 97.5% chance of a song you tolerate playing, since $P(Y = 1) = .975^1 * (1 - .975)^0$.

Binomial Random Variable

- Does anybody know how this relates to a Bernoulli random variable?

Binomial Random Variable

- We can extend our knowledge of binary random variables.
- Suppose we flipped an *unfair* coin n times and recorded Y, the number of heads after n flips.
- If we consider a case where $p = 0.25$ and $n = 4$, then here $P(Y = 0)$ represents the probability of no successes in 4 trials
  - 4 consecutive failures.
  - The probability of 4 consecutive failures is $P(Y = 0) = P(TTTT) = (1 - p)^4 = 0.75^4$.

Binomial Random Variable

- Next we consider $P(Y = 1)$, and are interested in the probability of exactly 1 success *anywhere* among the 4 trials.
- How many different ways can we get exactly 1 success in the four trials?
- To find the probability of this what else do we need to count?
- What is the probability?

Binomial Random Variable

- There are $\binom{4}{1} = 4$ ways to have exactly 1 success in 4 trials, $P(Y = 1) = \binom{4}{1} p^1 (1-p)^{4-1} = (4)(0.25)(0.75)^3$.

Binomial Distribution

- In general, if we carry out a sequence of n Bernoulli trials (with probability of success p) and record Y, the total number of successes, then Y follows a **binomial distribution**, where

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} \quad \text{for} \quad y = 0, 1, \ldots, n. \tag{1}$$

- If $Y \sim \mathrm{Binomial}(n, p)$, then $\mathsf{E}(Y) = np$ and $\mathsf{SD}(Y) = \sqrt{np(1-p)}$. - $E(y)$ is the expected value of Y. If we repeated the experiment many times we would expect on average $n * p$ successes.

Binomial Distribution Graphs

- Typical shapes of a binomial distribution
- I will show you these in R

Binomial Distribution Graphs

- On the left side n remains constant.
- We see that as p increases, the center of the distribution ($\mathrm{E}(Y) = np$) shifts right.
- On the right, p is held constant.
- As n increases, the distribution becomes less skewed.

Binomial Distribution vs Bernoulli

- Note that if $n = 1$,

$$\begin{aligned} P(Y = y) &= \binom{1}{y} p^y (1-p)^{1-y} \\ &= p^y (1-p)^{1-y} \quad \text{for} \quad y = 0, 1, \end{aligned}$$

a Bernoulli distribution! - In fact, Bernoulli random variables are a special case of binomial random variables where $n = 1$.

Graphing/calling in R

- In R we can use the function `dbinom(y, n, p)`, which outputs the probability of y successes given n trials with probability p, i.e., $P(Y = y)$ for $Y \sim \mathrm{Binomial}(n, p)$.
- Lets try it

Example 2

- While taking a multiple choice test, a student encountered 10 problems where she ended up completely guessing, randomly selecting one of the four options.
- What is the chance that she got exactly 2 of the 10 correct?
- What assumption do we need to make about the questions?
- How would we do this without using a distribution?

Example 2

- Knowing that the student randomly selected her answers, we assume she has a 25% chance of a correct response.
- n *factorial* is denoted $n!$ and $n! = n \cdot (n-1) \cdot ... \cdot 3 \cdot 2 \cdot 1$
- Here we used a *combination* $\binom{n}{p} = \frac{n!}{p!(n-p)!}$ read *n choose p*
- Thus, $P(Y = 2) = \binom{10}{2}(.25)^2(.75)^8 = 0.282$.
- We can use R to verify this:

```
dbinom(2, size = 10, prob = .25)
```

```
## [1] 0.2815676
```

Therefore, there is a 28% chance of exactly 2 correct answers out of 10.