ST 518: Data Analytics II

Types of Categorical Data

Types of Categorical Data

- Binary Data
- More than Two Categories

Some Examples

1. In the run up to US presidential elections, pollsters survey likely voters to ask their opinions about their preferred candidate and/or about candidate characteristics that are important to them.

2. In a clinical trial to compare a new treatment to an existing treatment, doctors are interested in whether the odds of survival is better on the new treatment than the standard treatment.

3. Risk analysts categorize credit card purchases to identify potential fraud.

Categorical Data

A **categorical random variable** can take only a discrete (distinct, disjoint) and finite set of values. For example:

- Political party affiliation might be one of Democrat, Independent or Republican
- Cancer diagnoses can fall into four categories: type 1, type 2, type 3, or type 4
- Genes have different forms, called alleles, responsible for variation
- A credit card purchase may be flagged as standard, unusual or highly unusual

Binary Variables

The simplest type of categorical random variable is the binary or Bernoulli (after Jacob Bernoulli the famous 17th century Swiss mathematician) random variable:

- The Bernoulli random variable can only take one of two possible values (for example, true/false, yes/no, 0/1).
- We typically recode the two categories as 0 or 1, and write, for a Bernoulli random variable X,

$$Pr(X = 1) = p,$$

  where $p \in [0, 1]$ is a probability between 0 and 1.
- This specifies the entire probability distribution of the Bernoulli random variable since the rules of probability give us:

$$Pr(X = 0) = 1 - Pr(X = 1) = 1 - p.$$

The Bernoulli Distribution

The Bernoulli distribution or probability mass function, written in it's general form for $Z \sim \text{Bernoulli}(p)$ is:

$$Pr(Z = z) = p^z(1-p)^{1-z}$$

for $z = 0, 1$ and $p \in [0, 1]$.

- Consider $z = 1$, then $Pr(Z = 1) = p^1(1-p)^{1-1} = p^1(1-p)^0 = p$.
- Consider $z = 0$, then $Pr(Z = 0) = p^0(1-p)^{1-0} = p^0(1-p)^1 - 1 - p$.

More than Two Categories

In a generalization of the binary variable, consider a situation where there are more than two categories for the response

- For example, suppose you have to identify your political affiliation as Democrat, Independent or Republican (3 categories).
- In this case, we could record the response as 1,2 or 3, according to Democrat, Independent, Republican, respectively.
- Mathematically, it turns out to be easier to record the response as one of three possible vectors:

$$\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{matrix}$$

or

$$\begin{matrix} 0 & 0 & 1 \end{matrix}$$

More than Two Categories

For the political party affiliation example, the categorical variable is actually defined as a vector of three values, say,

$$(W_1, W_2, W_3),$$

where one and only one of the values is 1 and the others are 0.

- Corresponding to each of W_1, W_2 and W_3 is a category probability,

$$p_j = Pr(W_j = 1) \text{ for } j = 1, 2, 3$$

- Since one and only one of the W_j's must be 1, we have

$$\begin{aligned} Pr(W_1 = 1 \text{ or } W_2 = 1 \text{ or } W_3 = 1) &= Pr(W_1 = 1) + Pr(W_2 = 1) + Pr(W_3 = 1) \\ &= p_1 + p_2 + p_3 \\ &= 1. \end{aligned}$$

More than Two Categories

For a categorical variable, (W_1, W_2, W_3), the probability mass function is

$$Pr(W_1 = w_1, W_2 = w_2, W_3 = w_3) = p_1^{w_1} p_2^{w_2} p_3^{w_3},$$

where

- $(w_1, w_2, w_3) \in \{(1,0,0), (0,1,0), (0,0,1)\}$
- $p_1, p_2, p_3 \in [0,1]$
- $p_1 + p_2 + p_3 = 1.$

Next Steps

The Bernoulli distribution and this simple categorical distribution serve as building blocks for more complicated distributions, called binomial distributions and multinomial distributions.

- A binomial random variable is defined as the sum of n independent Bernoulli random variables, all with probability p.
- A multinomial random variable is defined as the sum of N independent categorical random variables, all with identical category probabilities, $p_1, p_2, \ldots, p_k$.

We'll consider these types of random variables in the next module.