Summaries for Qualitative Variables

Data Science 101 Team

Qualitative variables

Let's think about variables that take on discrete qualitative values.

- Eye color
- Allele at different positions in the genome
- Animal species in a national park

How can we define summaries for these?

Specifically, how can we think of **center** and **spread**?

Center

We already have one notion of "center" that we argued works well for qualitative variables

Suppose we have data of the form

$$\begin{array}{cccc} v_1 & v_2 & \cdots & v_m \\ n_1 & n_2 & \cdots & n_m \end{array}$$

where n_j are the number of observations with value v_j

Recall that with $d(x, y) = \mathbf{1}\{x \neq y\}$, the minimizer of the average distance is the v_j for which n_j is maximum

This is called the **mode**, one sensible measure of "center" for qualitative variables

Diversity

When dealing with a qualitative variable, rather then "spread" we talk about diversity

Let's say that we have data in the following form

$$\begin{array}{cccc} v_1 & v_2 & \cdots & v_m \\ p_1 & p_2 & \cdots & p_m \end{array}$$

where $v_1, \ldots, v_m$ are the different outcomes of the variable (ex. "Brown eyes","Blue eyes","Green eyes",...) and $p_1, \ldots, p_m$ the relative frequencies with which they are observed in the data ($p_1 + p_2 + \cdots + p_m = 1$)

- Q: What would be the values of $p_1, \ldots, p_m$ that correspond to the least diversity?
- Q: How about the value of $p_1, \ldots, p_m$ that correspond to the most diversity?

Another comparison

Say that population 1 is such that

$$\begin{array}{cccc} v_1 & v_2 & \cdots & v_m \\ \frac{1}{m} & \frac{1}{m} & \cdots & \frac{1}{m} \end{array}$$

And population 2 is

$$\begin{array}{cccc} w_1 & w_2 & \cdots & w_k \\ \frac{1}{k} & \frac{1}{k} & \cdots & \frac{1}{k} \end{array}$$

Q: Which of the two populations is more diverse if $k \leq m$?

An index of diversity

Now that we have some practice in thinking about diversity, let's see if we can come up with an "index" for diversity

$$\begin{array}{cccc} v_1 & v_2 & \cdots & v_m \\ p_1 & p_2 & \cdots & p_m \end{array}$$

$$D = ?$$

Imagine you go fishing and you are going to get a sense of the species diversity of the fish population in the lake from the first two fish that you capture

Q: What are the possible outcomes in two "fish catching events," assuming that we care only about getting a sense of how diverse the fish population is?

One proposal

$$D = 1 - \sum_{i=1}^{m} p_i^2$$

Probability that if you capture two fish they are not the same species

Q: Why?

Let's work this out using reasoning. Assume that the species of the first fish I catch doesn't affect the species of the second fish (I practice catch and release, fish of a scale don't school together, etc. . .)

What's the chance that the first fish I catch is of species 1?

What's the chance that the second fish I catch is of species 1?

Now, what's the chance that **both** the first and second fish I catch are of species 1? Is it higher or lower than the chance that just the first fish is of species 1?

Verifying that the index does what we want

$$D = 1 - \sum_{i=1}^{m} p_i^2$$

- $0 \leq D \leq 1$
- $D = 0$ if one $p_i = 1$
- Let's calculate the value of the index when $p_1 = p_2 = \cdots = p_m = \frac{1}{m}$

$$\sum_{i=1}^{m} p_i^2 = \sum_{i=1}^{m} (\frac{1}{m})^2 = m\frac{1}{m^2} = \frac{1}{m}$$

So, the diversity $1 - \frac{1}{m}$ is larger if m is larger

- $D = 1$ when there are an infinite number of species

Notes

- One can verify that for a population with m outcomes, $D \leq 1 - \frac{1}{m}$
- This index is known as Gini (again!) or Simpson's diversity index (be careful that actually there are multiple versions of the Simpson index)
- There are other measures of diversity. Most importantly one known as Shannon's index that is based on **entropy**

$$H = -\sum_{i=1}^{m} p_i \log(p_i)$$

Gini index in genetics

- When analyzing data on the frequency of different alleles in genetics, the $D = 1 - \sum_{i=1}^{m} p_i^2$ is preferred
- This is because it has a very easy genetic interpretation: it represents the **probability of a heterozygous genotype**

Summarizing multiple qualitative variables

Suppose I have **two** qualitative variables (ex: eye color, hair color)

How could I compactly store them?

Idea: just make two tables like

v_1	v_2	$\cdots$	v_J
n_1	n_2	$\cdots$	n_J

u_1	u_2	$\cdots$	u_K
m_1	m_2	$\cdots$	m_K

Q: What do you think of this idea?

Summarizing multiple qualitative variables: contingency tables

So when we collapsed our data into two separate tables, we lost the ability to say things like "x% of people in the data have both brown hair and brown eyes"

In other words, we lost information on the **dependence** between the variables

Q: How could we still reduce data size, while **keeping** this info?

Contingency table

A **contingency table** stores the counts of the number of observations of every possible **combination** of the values of the two variables.

Example: hair and eye color

	u_1	u_2	$\cdots$	u_K
v_1	n_{11}	n_{12}	$\cdots$	n_{1K}
v_2	n_{21}	n_{22}	$\cdots$	n_{2K}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
v_J	n_{J1}	n_{J2}	$\cdots$	n_{JK}

Contingency tables and dependence

Recall that we lost information about **dependence** by making two separate tables

Consider three different cases:

1. There is no relationship between hair and eye color
2. There is a moderate relationship between hair and eye color
3. There is a very strong relationship between hair and eye color

Q: In which case would you lose the most information by making two separate tables?

Q: Now suppose you have a contingency table. Can you think of any ways to measure how strong the depencence is between the two variables just based on the table?