Summaries for Qualitative Variables

Data Science 101 Team
Qualitative variables

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

- Race/Ethnicity information in census data
- 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?
Diversity

When dealing with a qualitative variable, rather than “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 modalities of the variable (ex. “Asian”, “Black”, “Hispanic”, “White”) and \(p_1, \ldots, p_m\) the relative frequencies with which they are observed in the data \((p_1 + p_2 + \cdots + p_m = 1)\)

▶ What would be the value of \(p_1, \ldots, p_m\) that correspond to the least diversity?

▶ 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

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

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.

\[ v_1 \quad v_2 \quad \cdots \quad v_m \]
\[ p_1 \quad p_2 \quad \cdots \quad p_m \]
\[ D = ? \]

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

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

- Probability that if you capture two fishes they are not the same.
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 \( p_1 = p_2 = \cdots = p_m = \frac{1}{m} \)

\[ \sum_{i=1}^{m} p_i^2 = \sum_{i=1}^{m} \left( \frac{1}{m} \right)^2 = m \frac{1}{m^2} = \frac{1}{m} \]

So, that the diversity \( 1 - \frac{1}{m} \) is larger the larger is \( m \)

- \( D = 1 \) when there are an infinite number of species
We really should verify that for a population with \( m \) modalities, 
\[
D \leq 1 - \frac{1}{m}
\]

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 \( \text{entropy} \)

\[
H = - \sum_{i=1}^{m} p_i \log(p_i)
\]
When analyzing data relative to 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 an heterozygous genotype.

Genetic diversity is connected to the size/age of the population: over time, more mutations can appear and can become reasonably frequent.
The Rosenberg Lab Website makes available a lot of data collected over the years on human genetic diversity.

In one study published in 2002, the scientific team collected data on

- 377 genetic markers
- in 52 human populations
- for a total of 1056 individuals

We downloaded information on the allele frequencies of each of these markers and each of these populations and studied their diversity.
Average marker diversity for distinct populations

- BiakaPygmy
- MbutiPygmy
- BantuKenya
- San
- Yoruba
- Mandenka
- Karitiana
- Surui
- Maya
- Pima
- Uygur
- Balochi
- Brunnio
- Hazara
- Makrani
- Pathan
- Sindh
- Cambodian
- Han
- Han−NChina
- Dai
- Daur
- Daur
- Hezhen
- Lahu
- Miao
- Oroqen
- She
- Tujia
- Xibo
- Yi
- Mongola
- Naxi
- Japanese
- Yakut
- Basque
- French
- Sardinian
- Orcadian
- Italian
- Russian
- Adygei
- Mozabite
- Palestinian
- Bedouin
- Melanesian
- Papuan
Average marker diversity for distinct continents

- Africa
- Middle East
- Central South Asia
- Europe
- East Asia
- Oceania
- America

Continental Diversity

- America (0.6)
- Oceania (0.7)
- East Asia (0.7)
- Europe
- Central South Asia (0.7)
- Middle East (0.7)
- Africa (0.7)