Categorical Data Analysis

Data:

A set of information which is collected for a specific purpose.

Categorical Data:

As the name implies, is grouped into some sort of category or multiple categories.

OR

A categorical data or categorical variable is one for which the measurement scale consists of a set of categories.

For example:

- Gender (Male & female).
- Eye Color (Black, Brown and Blue).
- Blood type of a person (A, B, AB or O).
- The city that a person lives in (Sargodha, lahore, rawalpindi)
- The political party that a voter might vote for e.g PMLN, PTI, MQM & PPP.
- Economic Status (low, medium and high).
- Education Level (undergraduate, graduate, post graduate).
- Race, material status.

Categorical data is also data that is collected in an yes or no fashion.

For example:

- Do you have car (Yes or no)

Is age a categorical data?

Age is a numerical data - It is often more informative to categorize it into a relatively small number of groups.

So the age groups are categorical.

* age group (12-22, 23-33 ...)
* Income (< 30000, 30000, > 30000)
* Likert scale (SA, A, N, D, SD)

Applications of CDA

* Social Sciences (for measuring attitudes and opinions on various issues).
* Health Sciences (for measuring such responses as whether a patient survives an operation {yes or no}, severity of an injury {none, mild, moderate, severe} and the stage of a disease {intial, advanced}).
* Behavioral Sciences (for diagnosis of type of mental illness {schizophrenia, depression, neurosis}).
* Public health (for whether awareness of lungs cancer has led to increased smoking {Yes or No})
* Zoology (for alligator's primary food preference {fish, invertebrate or reptile})

* Education (for student's responses to an exam question {correct or incorrect})
* Marketing (Preference among leading brands of a product {Brand A, Brand B or Brand C})
* Engineering sciences and industrial quality control (to whether or not they conform to certain standards {Yes or No})

Types of Categorical data.

Categorical variables have two main types of measurement scales. "Nominal variables" & "Ordinal variables".

Nominal variables:

Categorical variables having unordered scales are called nominal variables.

Example:

→ Religious Affiliation (Hindu, Jewish, Muslim, other).

→ Mode of transportation (automobile, bicycle, bus, subway, walk).

Ordinal variables:

Categorical variables having ordered scales are called ordinal variables:

→ response to a medical treatment (excellent, good, fair, poor).

→ Methods designed for nominal variables give the same results no matter how the categories are listed. Methods designed for ordinal variables utilize the categorical ordering. Whether we list the categories from low to high or from high to low is irrelivent in terms of substantive conclusions, but results of ordinal analyses would change if the categories were recorded in any other way.

⇒ Method designed for ordinal variables cannot be used with nominal variables since nominal variables do not have ordered categories.

Method designed for nominal variables can be used with nominal or ordinal variables, since they only required a categorical scale.

Distributions for Categorical Data.

Inferential statistical analysis require assumptions about the probability distribution of the response variable.

The key distributions for categorical responses/data: Binomial, multinomial

Binomial Distribution:

An experiment is called a binomial probability experiment if it possesses the following four properties:

1- The outcomes of each trail may be classified into one of two categories.

2- The probability of success (π) remains constant for all trials.

3- The successive trails are all independent

4- The experiment is repeated a fixed number of times, say n.

Let $X_1, X_2, \dots X_n$ denote responses for n independent and identical trials such that $P(X=1) = \pi$ and $P(X=0) = 1-\pi$. we use the generic labels "success" and "failure" for outcomes 1 and 0. The total number of successes, $X = \sum_{i=1}^{n} X_i$ has the binomial distribution with independent n and parameter "π", denoted by $bin(n, \pi)$

The pmf for the possible outcomes y for Y is

$$f(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

$$x = 0, 1, 2 \dots, n$$

$$\mu = E(X) = n\pi \quad , \quad \sigma^2 = var(X) = n\pi(1-\pi)$$

The distribution converges to normality as n increases, for fixed π.

Multinomial Distribution:

A Multinomial experiment has the following properties:

1- The outcomes of each trial may be classified into one of k mutually exclusive categories.

2- The probability of the ith outcome is π_i which remain constant and $\sum \pi_i = 1$

3- The successive trails are all independent

4- The experiment is repeated a fixed number of times, say n.

Suppose that each of n independent, identical trails can have outcome in any of c categories. let $Y_{ij} = 1$ if trial i has outcome in category j and $Y_{ij} = 0$ otherwise. Then $Y_i = (Y_{i1}, Y_{i2}, \dots Y_{ic})$ represent a multinomial trail

When the trials are independent with the same category probabilities for each trial, the distribution of counts in the various categories is the multinomial

Let c denote the number of outcome categories. We denote their probabilities by $\{\pi_1, \pi_2, \dots \pi_c\}$ where $\sum \pi_j = 1$

For n independent observations, the multinomial probability that n_1 fall in category 1, n_2 fall in category 2 ... n_c fall in category c, where $\sum n_j = n$, equals.

$$f(x) = \frac{n!}{n_1! n_2! \cdots n_c!} \pi_1^{n_1} \cdot \pi_2^{n_2} \cdots \pi_c^{n_c}$$

$$\mu = E(X) = n\pi_i \qquad \sigma^2 = var(x) = n\pi_i(1 - \pi_i)$$

Most methods for categorical data assume the binomial distribution for a count in a single category. and the multinomial distribution for a set of counts in several categories.