Association of Attributes

By

Dr. Moutushi Chatterjee

Definition:

A categorical variable has a measurement scale consisting of a set of categories.
For example, students' responses to an exam question, with the categories correct & incorrect.

Categorical data may even occur in highly quantitative fields, such as
i) How soft to touch a certain fabric is;
ii) How good a particular food tastes.

Categorical variable

Explanatory variable — Response

Remark:

Date / /

The way that a variable is measured, determines its classification. For example,

i) "education" is only nominal when measured as Govt. school or private school;

ii) It is ordinal when measured by highest-degree attained, using the categories of primary, high school, bachelors' & so on,

iii) It is interval when measured by no. of years of education, using the integers 1, 2, ...

Contingency Table

Let X & Y denote two categorical response variables, X & with I categories & Y with J categories. Classifications of subjects on both variables ⇒ have IJ possible combinations. The responses (X, Y) of a subject chosen randomly from some popⁿ have a probability distⁿ.

A rectangular table having I rows for categories of X & J col.s for categories of Y displays this distⁿ. The cells of the table represent the IJ possible outcomes. When the cells contain frequency counts of outcomes for a sample, the table is called a contingency table [Karl Pearson, 1904].

A contingency table with I rows & J col.s is called an I x J (or I by J) table.

Sensitivity & Specificity

Consider a survey on the effectiveness of a diagnosis procedure for a particular disease.

Let X denotes the true disease status (ie. whether a person suffers from the disease) & let Y = diagnosis (ie +ve or -ve), where a +ve outcome predicts that the person is having suffering from the disease.

With diagnostic tests for a disease, the two correct diagnoses are-

i) a +ve test outcome, when the subject actually is suffering from the disease.

ii) a -ve test outcome when the subject is not suffering.

	Diagnostic Test		Total
Disease.	+ve	-ve	
Yes			
NO			

a) Given that the subject has the disease, the conditional prob. that the diagnostic test is +ve is called the sensitivity.

b) Given that the subject does not have the disease, the conditional probability that the test is -ve. is called specificity.

Ideally, both should be high.

Independence & Association of attributes in the context of 2×2 contingency table:

		A	α	Total
B	B	f_{AB}	$f_{\alpha B}$	f_B
β	β	$f_{A\beta}$	$f_{\alpha\beta}$	f_β
	Total	f_A	f_α	n

In the context of 2×2 contingency table, suppose there are two attributes each having two forms & based on the presence or absense of the said characteristic. Thus for the 1st attribute, the presence of the characteristic may be denoted by A or the absence by α. Similarly, the presence of the second ch. may be denoted by B & absence by β.

Let us now consider the following two assumptions:

i) The individuals under consideration constitute the pop^{n} itself & not just a sample from the pop^{n}.

ii) None of the marginal frequencies is zero.

Cell frequencies: f_{AB}, $f_{A\beta}$, $f_{\alpha B}$, $f_{\alpha\beta}$

Marginal frequencies:

$f_A = f_{AB} + f_{A\beta}$; $\quad f_\alpha = f_{\alpha B} + f_{\alpha\beta}$;

$f_B = f_{AB} + f_{\alpha B}$, $\quad f_\beta = f_{A\beta} + f_{\alpha\beta}$.

f_{AB}/f_A : Proportions of the members of the pop^{n} having B among those having A.

$f_{\alpha B}/f_\alpha$: Proportions . . . having B among those having α (ie not having A).

Independence

Thus, $f_{AB}/f_A = f_{\alpha B}/f_\alpha$ implies that the presence or absence of the character A does not have any impact on the presence of the character B in an individual. In such a situation, A & B are said to be independent.

On the other hand, if A & B are associated, then $\frac{f_{AB}}{f_A} \neq \frac{f_{\alpha B}}{f_\alpha}$.

Association:

When the two characters A & B are not mutually independent, one may encounter either of the following two cases, viz.

i) $f_{AB} > \frac{f_A f_B}{n}$: the attributes are said to be positively associated. In this case, A & B jointly occur more frequently, than they would have occured indiridually, had they been independent.

ii) $f_{AB} < \frac{f_A f_B}{n}$: the attributes A & B are said to be negatively associated (or disassociated). In this case, A & B jointly occur less frequently than they would have occured indiridually, had they been mutually independent.

Measures of Association for 2×2 contingency table.

Properties of a good measure:

In the context of association of attributes, a good measure should possess the following properties

i) A good measure should be independent of the total frequency (n) - it should depend only on the relative frequencies in the cells - rather than the absolute ~~frequencies~~ cell frequencies. Otherwise, by merely increasing the total freq. one can manipulate the index value.

ii) A good measure should assume the value 0 for independence; negative for negative association & +ve for +ve association.

iii) It should increase from its lowest possible value through zero to its highest possible value as one proceeds from perfect -ve association through independence to perfect +ve association.

iv) It should preferably vary betⁿ two definite limits, like (-1) & (+1).

Yule's first measure of Association.

Coefficient of Association

Yule used the difference $\delta_{AB} = f_{AB} - \frac{f_A f_B}{n}$ while defining a measure of association, since when the two attributes are mutually independent, $\delta_{AB} = 0$ & it also possesses the 3rd desirable property discussed above..

Thus, one can define

$$Q_{AB} = \frac{n\,\delta_{AB}}{f_{AB} f_{\alpha\beta} + f_{A\beta} f_{\alpha B}}$$

$$= \frac{f_{AB} f_{\alpha\beta} - f_{A\beta} f_{\alpha B}}{f_{AB} f_{\alpha\beta} + f_{A\beta} f_{\alpha B}}.$$

This is known as the Yule's coefficient of association bet^n two attributes.

This is labeled as in honor of the Belgian statistician Quetelet.

Properties of δ_{AB}:

1) $\delta_{AB} = 0$ iff $s_{AB} = 0$, ie iff A & B are independent
2) $-1 \leq \delta_{AB} \leq +1$.

3) $\delta_{AB} = -1$, when $f_{AB} f_{\alpha B} = 0$, ie when at least one of f_{AB} & $f_{\alpha B}$ is 0, ie when ∃ complete negative association bet^n the two attributes.

4) $\delta_{AB} = +1$, when $f_{AB} f_{\alpha B} = 0$, ie when at least one of f_{AB} & $f_{\alpha B}$ is 0, ie when ∃ complete +ve association bet^n the two ~~variables~~ attributes.

Thus, δ_{AB} satisfies all the desirable properties of a good measure of Association.

Yule's Second Measure of Association [Coefficient of Colligation]

$$\gamma_{AB} = \frac{\sqrt{f_{AB} f_{\alpha\beta}} - \sqrt{f_{A\beta} f_{\alpha B}}}{\sqrt{f_{AB} f_{\alpha\beta}} + \sqrt{f_{A\beta} f_{\alpha B}}}$$

It is easy to observe that γ_{AB}, similar to σ_{AB}, also satisfies all the desirable ~~con~~ properties of a good measure of association of attributes.

Numerical Example

- Some students of an Indian City, who were interviewed during a sample survey, are classified below according to their smoking and tea drinking habits. Calculate Yule's measures of association.

	Smoker	Non-smoker
Drinks tea	40	33
Does not drink tea	3	12

Odds Ratio

Date

odds

Odds ratio is one of the most important measures of association for 2×2 contingency tables.

If the probability of success is denoted by π, then the odds of success can be defined as $\pi/(1-\pi)$. Since π is probability & hence by definition, $0 \leq \pi \leq 1$, value of odds will always be non-negative.

Also, odds $= \pi/(1-\pi) \Rightarrow P(\text{success}) = \frac{\text{odds}}{(1+\text{odds})}$.

Thus, probability of success can be expressed in terms of the odds & vice-versa.

Interpretations:

odds > 1 $\Rightarrow$ A success is more likely than a failure.

This is due to the fact that,

$\pi/(1-\pi) > 1 \Rightarrow \pi > 1-\pi \Rightarrow 2\pi > 1 \Rightarrow \pi > \frac{1}{2}$ i.e. $\pi < \frac{1}{2}$.

i.e. $\Rightarrow$ π = P(success) $> \frac{1}{2}$ & $1-\pi$ = P(failure) $< \frac{1}{2}$.

odds = 4 $\Rightarrow$ A success is 4 times as likely as a failure.

Thus, here one can expect 4 successes for every single failure.

odds = $\frac{1}{4}$ $\Rightarrow$ A failure is 4 times as likely as a success.

Thus, here one can expect 4 failures for every single success.

Also, odds = 4 $\Rightarrow$ π = odds/(1+odds) = $\frac{4}{4+1} = 0.8$.

Thus, success probability is 0.8.

Thus, odds is the ratio of probability of success and the probability of failure.

Date ___/___

Odds and the elements of 2×2 Contingency Table

Suppose we are investigating the so-called association between Smoking & lung cancer. Here the (2×2) Contingency Table will be as follows:

	With lung Cancer (1)	Without lung Cancer (2)	
Smoker (1)	π_{11}	π_{12}	π_{10}
Non-Smoker (2)	π_{21}	π_{22}	π_{20}

Then, Ω_1 = odds for Smokers = $\frac{\text{Success prob. of having LC among Smokers}}{1 - \text{Success prob. of having LC among Smokers}}$

$$= \frac{\pi_{11}}{\pi_{12}} = \frac{(\pi_{11}/\pi_{10})}{(\pi_{12}/\pi_{10})}$$

$$= \frac{(\pi_{11}/\pi_{10})}{(\pi_{10} - \pi_{11})/\pi_{10}} = (\pi_{11}/\pi_{10}) / (1 - \pi_{11}/\pi_{10})$$

Similarly, Ω_2 = odds for non-smokers = π_{21} / π_{22}.

Hence odds ratio = $\Omega = \frac{\Omega_1}{\Omega_2} = (\pi_{11}\pi_{22}) / (\pi_{12}\pi_{21})$.

Note that here it does not matter whether we consider smoking as explanatory variable & LC as the response and vice versa. For this reason, odds ratio is more popular.

Formal definition of odds ratio:

The odds ratio can be defined as the estimated increase in the probability of success associated with a one-unit change in the value of the predictor variable.

Thus odds ratio is a relative measure of association, telling us how much more it is likely that same one who is exposed to the factor under study will develop the outcome as compared to same one who is not exposed.