

Supplemental Lecture on Analysis of Categorical Variables

1 Summarizing relationships between dichotomous variables

We have already discussed one method for comparing categorical data on 2 groups: compute the distribution of the response variable conditional on group membership (since there are two groups, there are two of these conditional distributions). If the response variable is *dichotomous* (i.e. only takes 2 values), then we compare the proportions of respondents who are classified as “yeses” in the 2 groups (so the 2 conditional distributions can be summarized by two proportions). One way to compare the 2 proportions would be to compute the difference between the two proportions, $\hat{p}_1 - \hat{p}_2$ (where \hat{p}_i is the sample proportion of “yeses” in group i).

1.1 Relative Risk

When we compare 2 proportions which are small, the difference between them will be small too even if one of the proportions is substantially larger than the other. For this reason, it often makes sense to look at the ratio of the 2 proportions $\frac{\hat{p}_1}{\hat{p}_2}$. When the explanatory variable is a risk factor for some disease (or other negative outcome, like getting hit by a car), and the response variable is the disease, we call this ratio the *relative risk*, and we will denote it RR or \hat{RR} is we estimate the quantity from data. If group 1 is the group which has the risk factor in question, we define the relative risk by

$$RR = \frac{p_1}{p_2},$$

and the sample quantity we use to estimate this number is

$$\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2}.$$

If the risk factor does increase the chances of developing the disease, \hat{RR} will be greater than 1. The relative risk tells you how many more times one is likely to develop the disease if one has the risk factor than if the risk factor is absent.

1.2 Odds Ratio

Although we have used probability as a method for quantifying uncertainty, one could also use *odds*. If some event happens with probability p , then the odds on the event are $\frac{p}{1-p}$ to 1. If p is high, then the odds on the event are high, while if the event is unlikely (so p is low) the odds on the event are low. Unlike probability, there is no upper bound on the odds on an event, but odds can never be negative (if an event can't happen, so that $p = 0$, then the odds on the event are zero).

The *odds ratio* is just like the relative risk, except we use odds instead of probabilities when we compute the ratio, that is, if we use OR for the odds ratio

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

The odds ratio is interpreted like the relative risk: it indicates how much your odds of developing the disease increase when you have an exposure to a risk factor. When the disease is rare (i.e. p_1 is almost zero), $1 - p_1$ is close to one, so that the odds $p_1/(1 - p_1)$ are very close to the probability p_1 , in which case the relative risk and the odds ratio are similar.

1.2.1 Estimation of the odds ratio and study design

A *prospective* study is a study in which the researchers get a collection of patients and observe what happens to the patients. In contrast, a *retrospective* study (also called a case-control study) collects data on certain types of patients and then finds out if these patients had some risk factor. These 2 types of studies provide very different information: the former provides information on the distribution of the response given values of the explanatory variables, while the latter provides information on the distribution of the explanatory variable given the response variable. To compute the relative risk, one needs the the probability

of developing the disease given the exposure to risk factors. Unfortunately, we can not obtain this sort of information from a retrospective study. On the other hand, we can estimate the odds ratio from either sort of study.

2 Inference for the Odds Ratio

While there are methods for conducting inference for the relative risk, we concentrate on inference for the odds ratio because of the wider applicability of this statistic. While the sampling distribution of the odds ratio is not normal, the normal distribution does provide a reasonable approximation to the natural logarithm (denoted \ln) of the odds ratio. The form for a confidence interval for $\ln(OR)$ is

$$\ln(\hat{OR}) \pm z^* \text{standard error of } \ln(\hat{OR}),$$

where z^* is a value from the normal table which determines the confidence level of the interval (e.g. for a 95% confidence interval $z^* = 1.96$). If a, b, c, d represent the 4 numbers in the 2 by 2 table, then

$$\text{standard error of } \ln(\hat{OR}) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

Once you have a confidence interval for the natural logarithm of the odds ratio, you can obtain a confidence interval for OR by exponentiating the endpoints of the interval for the natural logarithm.

To conduct a 2 sided test about the odds ratio at level α , you construct a $1 - \alpha$ level confidence interval and examine if the hypothesized value of OR lies in the interval. If you are interested in the hypothesis $H_0 : OR = 1$ it probably makes sense to just use Pearson's χ^2 test.

3 Evaluation of diagnostic tests

Suppose we have some test we use to determine if a patient has some illness. There are a variety of methods in use to characterize the quality of such tests. To simplify the presentation, let us use T^+ to denote a positive test result, T^- to denote a negative test result, D^+ to denote the person has the disease and D^- to indicate if the person does not have the disease.

3.1 Sensitivity and Specificity

The *sensitivity* and *specificity* relate to the test's ability to find ill people. These two quantities are defined by

$$\text{sensitivity} = P(T^+ | D^+)$$

$$\text{specificity} = P(T^- | D^-).$$

A good diagnostic test has high values for both of these.

3.2 Predictive value of the test

Another way to characterize a diagnostic test is to look at the test from the patient's prospective: given a positive test result, does it really mean the patient is sick. There are 2 commonly used measures here, the *positive predictive value* (PPV) and the *negative predictive value* (NPV). These are

$$\text{PPV} = P(D^+ | T^+)$$

$$\text{NPV} = P(D^- | T^-).$$

If we also know the *prevalence* (which is the probability that someone has the disease in the population who gets the test done, or $P(D^+)$), then we can relate the PPV and the NPV to the sensitivity and the specificity using Bayes theorem. We find

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})},$$

and

$$\text{NPV} = \frac{\text{specificity} \times (1 - \text{prevalence})}{\text{specificity} \times (1 - \text{prevalence}) + (1 - \text{sensitivity}) \times \text{prevalence}}.$$