STA 580 — Spring 2009 — Dr. Charnigo

Lecture 13

Measures of effect: risk difference, relative risk, odds ratio

Introduction. In Lecture 11, I presented the correlation as a tool to describe the strength and direction of the (linear) relationship between two continuous variables. But what if we want to describe the strength and direction of the relationship between two dichotomous variables?

More specifically, suppose that we have a dichotomous response variable indicating whether a now-healthy individual develops a certain disease by a specified future date. Also, suppose that we have a dichotomous explanatory variable indicating whether the individual has a certain exposure suspected to affect the probability of developing the disease. Our goal is to ascertain whether the exposure does affect the probability of developing the disease and, if so, by how much.

Notation and terminology. Let p_1 denote the probability that a now-healthy exposed person develops the disease by a specified future date, and let p_2 denote the probability that a now-healthy non-exposed person develops the disease by a specified future date.[1]

We refer to each of p_1 and p_2 as a risk, and we refer to each of $p_1/(1-p_1)$ and $p_2/(1-p_2)$ as odds. Thus, the odds of developing the disease for an exposed individual equal the probability of developing the disease for an exposed individual divided by the probability of not developing the disease for an exposed individual, and similarly for a non-exposed individual.

[1] In fact, p_1 and p_2 are conditional probabilities. Let D be the event that a now-healthy individual develops the disease by a specified future date, and let E be the event that this individual has the exposure. Then $p_1 = P(D|E)$ and $p_2 = P(D|\bar{E})$.

Risk difference, relative risk, odds ratio. The risk difference is (Definition 13.4)

$$p_1 - p_2,$$

the relative risk or risk ratio is (Definition 13.4)

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

and the odds ratio is (Definition 13.6)

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

Example (risk difference, relative risk, odds ratio). Suppose that $p_1 = 0.01$ and $p_2 = 0.001$. Then the risk difference is

$$0.01 - 0.001 = 0.009,$$

the relative risk is

$$\frac{0.01}{0.001} = 10.0,$$

and the odds ratio is

$$\frac{0.01/0.99}{0.001/0.999} = 10.1.$$

Comments on risk differences. For the risk difference, the "neutral" value is 0. That is, $p_1 - p_2 = 0$ means that the exposure is neither harmful nor beneficial. While a positive risk difference means that the exposure is harmful, the risk difference may not accord with your intuition about <u>how</u> harmful the exposure is.

For instance, if I tell you that smoking increases your risk of developing lung cancer in the next five years from 0.001 to 0.01, you will most likely conclude that smoking is quite harmful. But if I tell you that eating fast food on Thursday nights increases your risk of developing hypertension in the next twenty years from 0.401 to 0.41, you will most likely conclude that eating fast food on Thursday nights is not too harmful.

Yet, the risk difference is the same in both scenarios: 0.009. In particular, for every 1000 people who stop smoking, we anticipate 9 fewer lung cancer cases in the next five years, and for every 1000 people who stop eating fast food on Thursday nights, we anticipate 9 fewer hypertension cases in the next twenty years.

Apart from the seriousness of the disease, then, why does our intuition tell us that these two scenarios are so different? Perhaps the main reason rests with our perception of how much control we have. In the first scenario, the 9 fewer lung cancer cases may be compared to the 1 lung cancer case that is still anticipated; a person who smokes and gets lung cancer cannot reasonably say, "I probably would have gotten lung cancer even if I had stopped smoking." In the second scenario, the 9 fewer hypertension cases may be compared to the 401 hypertension cases that are still anticipated; a person who eats fast food on Thursday nights and becomes hypertensive can reasonably say, "I probably would have become hypertensive even if I had stopped eating fast food on Thursday nights."

Comments on relative risks. For the relative risk, the neutral value is 1. A relative risk greater than 1 means that the exposure is harmful. The relative risk will usually accord with your intuition about how harmful the exposure is. Indeed, the relative risk is easy for the lay person to understand: if the relative risk is 10, then an exposed individual is 10 times as likely to develop the disease as a non-exposed individual. In particular, if there are equal numbers of exposed and non-exposed people, we anticipate about 10 times as many exposed people developing the disease as non-exposed people.

Comments on odds ratios. For the odds ratio, the neutral value is 1. An odds ratio greater than 1 means that the exposure is harmful.

When both risks are small, the odds ratio is numerically similar to the relative risk because we can approximate $1 - p_1$ by 1 and $1 - p_2$ by 1 to obtain

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

When the risks are not small, the odds ratio can be much further away from 1 than the relative risk.

If the odds ratio is 10.1, then an exposed individual has odds of developing the disease 10.1 times as large as a non-exposed individual. Unfortunately, many researchers will state that an exposed individual is 10.1 times as likely to develop the disease as a non-exposed individual. Such a statement is at best casual and at worst seriously misleading.

How can we explain an odds ratio of 10.1 to the lay person unfamiliar with the concept of odds? This is difficult. One attempt is as follows: "Suppose that there are equal numbers of exposed people who do not develop the disease and non-exposed people who do not develop the disease. Then, for every 10 non-exposed people who do develop the disease, we anticipate that 101 exposed people will develop the disease."

Types of studies

Introduction. Since we want to ascertain whether exposure affects the probability of developing the disease, we will want to obtain point and interval estimates for at least one of the three measures of effect (risk difference, relative risk, odds ratio). The type of study determines the measures of effect for which we can obtain such estimates. The type of study may also have implications for the speed with which results can be obtained and for how seriously the results may be affected by various biases.

Prospective and retrospective studies. In a prospective study (or cohort study), we recruit groups of exposed and non-exposed individuals who are currently healthy. We then monitor the subjects in both groups for development of the disease over a specific time period, so that we can relate disease status at the end of the study to exposure status at the beginning of the study (Definition 13.1).

Prospective studies are usually observational, in that the investigators do not randomly assign the subjects to be exposed or non-exposed; the subjects assign themselves, as it were. For example, randomly assigning half of one's subjects to smoke would be unethical; however, given that a person has decided to smoke, observing what happens to that person is not unethical. Exceptions may occur when the "exposure" is really a treatment intended to reduce the probability of developing the disease; in such instances, the study is more of a clinical trial than an epidemiologic investigation.

In a retrospective study (or case-control study), we recruit groups of diseased and non-diseased individuals. We then find out which individuals were exposed in the past, so that we can relate past exposure status to current disease status (Definition 13.2). Retrospective studies are necessarily observational.

Advantages of retrospective studies. Key advantages of retrospective studies are as follows.

• Savings of time: With a retrospective study, we know almost immediately a subject's current disease status and past exposure status. With a prospective study, we must wait (say) ten years to know each subject's disease status ten years from now. A long delay in reaching a conclusion may impede public health officials from addressing an important public health problem in a timely fashion.

• Affordability: With a retrospective study, we can oversample cases (i.e., recruit disproportionately many individuals with the disease) and thereby ensure that we have enough cases to reach a firm conclusion. With a prospective study, a massive number of subjects will be required if the disease is rare. This is because the precision of our estimates will be limited by the smallest entry in a table such as Table 13.1. If we conduct a prospective study and do not have enough subjects to ensure that a and c are reasonably large, then we may not be able to reach a firm conclusion.

Advantages of prospective studies. Key advantages of prospective studies are as follows.

• Estimation capabilities: With a prospective study, we can estimate p_1 by a/n_1 and p_2 by c/n_2 (notation as in Table 13.1). Hence, any function of p_1 and p_2 can be estimated using solely the information from a two-by-two contingency table. With a retrospective study, we cannot estimate either p_1 or p_2 using solely the information from a two-by-two contingency table.

However, we can still estimate the odds ratio.[2]

• Avoidance of recall bias: With a retrospective study, we depend on the subjects to recall and report their past exposure status. With a prospective study, we require only that the subjects report their current exposure status, so there is less potential for inaccurate reporting.

• Avoidance of selection bias: With a retrospective study, the diseased individuals we recruit may not be representative of diseased individuals in general. Our subjects may tend to be less ill, as they are still alive and willing to participate in our study.

Point and interval estimation

Estimation in prospective studies. Suppose that the data are organized as in Table 13.1. Let $\hat{p}_1 := a/n_1$ and $\hat{p}_2 := c/n_2$. Note that $\hat{p}_1$ and $\hat{p}_2$ are the proportions of subjects in the exposed and non-exposed groups who have developed the disease within the specified time period; that is, $\hat{p}_1$ and $\hat{p}_2$ are sample proportions. A $100(1-\alpha)\%$ confidence interval for the risk difference is (Equation 13.1, omitting the continuity correction)

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

A $100(1-\alpha)\%$ confidence interval for the relative risk is (Equation 13.6)

$$\frac{\hat{p}_1}{\hat{p}_2}\exp\left[\pm z_{1-\alpha/2}\sqrt{b/(a \cdot n_1) + d/(c \cdot n_2)}\right].$$

---

[2] This is because

$$\frac{P(D|E)/P(\bar{D}|E)}{P(D|\bar{E})/P(\bar{D}|\bar{E})} = \frac{P(E|D)/P(\bar{E}|D)}{P(E|\bar{D})/P(\bar{E}|\bar{D})}.$$

While $P(D|E)$ and $P(D|\bar{E})$ cannot be estimated from a contingency table in a retrospective study, $P(E|D)$ and $P(E|\bar{D})$ can be estimated from a contingency table in a retrospective study.

A $100(1-\alpha)\%$ confidence interval for the odds ratio is (Equation 13.11)

$$\frac{\hat{p}_1(1-\hat{p}_2)}{\hat{p}_2(1-\hat{p}_1)} \exp\left[\pm z_{1-\alpha/2}\sqrt{1/a+1/b+1/c+1/d}\right].$$

Example (estimation in prospective studies). Consider the data in Table 13.3. The disease is lung cancer, and the exposure is heavy consumption of alcoholic beverages. Note that $\hat{p}_1 = 33/1700 = .0194$ and that $\hat{p}_2 = 27/2300 = .0117$. The 95% confidence interval for the risk difference is

$$.0194 - .0117 \pm 1.96\sqrt{\frac{.0194(.9806)}{1700} + \frac{.0117(.9883)}{2300}} = -.0002 \text{ to } .0156.$$

The 95% confidence interval for the relative risk is

$$\frac{.0194}{.0117} \exp\left[\pm 1.96\sqrt{1667/(33 \cdot 1700) + 2273/(27 \cdot 2300)}\right] = 0.998 \text{ to } 2.74,$$

while the 95% confidence interval for the odds ratio is

$$\frac{.0194(.9883)}{.0117(.9806)} \exp\left[\pm 1.96\sqrt{1/33 + 1/1667 + 1/27 + 1/2273}\right] = 0.998 \text{ to } 2.78.$$

Remarks on confidence intervals. Each confidence interval starts with a point estimate. The point estimate for the risk difference is adjusted additively to yield a confidence interval, while the point estimates for the relative risk and the odds ratio are adjusted multiplicatively. The sample sizes must be reasonably large for these confidence intervals to be valid. A common rule of thumb is to require $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$.

Estimation in retrospective studies. The $100(1-\alpha)\%$ confidence interval for the odds ratio is

$$\frac{ad}{bc}\exp\left[\pm z_{1-\alpha/2}\sqrt{1/a+1/b+1/c+1/d}\right].$$

This is equivalent to the formula used for prospective studies; however, we must remember that $\hat{p}_1$ and $\hat{p}_2$ have no meaning in retrospective studies. If the disease is rare, then this confidence interval may be viewed as an approximate confidence interval for the relative risk.

Reading the SAS output. Refer to {Lungcancer.pdf}. The contingency table at the top of page 1 provides frequencies, overall percentages, percentages within row (i.e., within exposure group), and percentages within column (i.e., within disease group). In the "Column 1 Risk Estimates" box, the first row pertains to p_1, the second row to p_2, and the last row to $p_1 - p_2$. The entry in the "Risk" column is a point estimate, and the entry in the "Asymptotic 95% Confidence Limits" column is a large-sample 95% confidence interval. These entries would not be meaningful with a retrospective study.

Continue to the "Estimates of the Relative Risk" box on page 2. The first row pertains to OR, and the second row pertains to RR. The second row would not be meaningful with a retrospective study.

Remarks on confounding. The data analysis in {Lungcancer.pdf} is problematic in that smoking confounds the relationship between lung cancer and heavy alcohol consumption: smokers are more likely both to drink heavily and to develop lung cancer. A consequence is that the relative risk and odds ratio estimates are misleadingly high. Methods to "adjust" or "control" for confounding variables like smoking are discussed in CPH 630 and CPH 930.