ONLINE MATERIAL THAT ACCOMPANIES CHAPTER 6

Box 1. What is a p-value and why do we not use it in this book?

Recall from Chapter 5 that we can estimate the probability of observing a certain value for a variable or values more extreme based on our sample once we standardize the variable so that it follows a z-distribution. In a similar fashion, if we have data on disease and exposure status, we can compute the probability of observing an association like the one we estimate from our data or an association that is more extreme under different hypotheses about the underlying association in the population.

Typically, we start with the hypothesis that there is no relation between exposure and disease in the population, often called the null hypothesis. If there is no association between exposure and disease, what would be the probability of observing the association that we observed, or something more extreme? We conduct tests that put two competing claims head to head: the claim that there is no association (null hypothesis) and the claim that there is an association (alternative hypothesis). To conduct the test, we compute a value called a test statistic, which is computed directly from the data. We then estimate the probability of observing a test statistic value like the one we got or a value more extreme if the claim that "there is no association" is really true. This probability is called the p-value. The smaller this probability, the less likely it is that we would observe these data under the assumption of no association between exposure and disease. With a small p-value, we have evidence that the claim "there is no association" is wrong. In other words, we reject the null hypothesis.

But how do we decide how small the p-value needs to be for us to reject a null hypothesis? In many fields, a standard is 5%, as suggested by Fisher, one of the principal architects of hypothesis testing in the medical and social sciences (Fisher 1950). If the probability, or the p-value, is less than 0.05, or 5%, we conclude that there is sufficient evidence to reject the hypothesis that there is no relation between exposure and outcome. Neyman and Pearson proposed to formalize Fisher's concept of an objective p-value cut-off that would provide sufficient evidence against a null hypothesis (Neyman and Pearson 1933), and the concept of statistical significance was born.

While the p-value as a useful tool, statistical significance as a concept is increasingly met with criticism (Sterne and Davey Smith 2001, Greenland and Poole 2013). The principal critiques are in the broad misinterpretation of the p-value and abuses of statistical significance in medical research (Meehl 1967, Gigerenzer 2004). The central critiques of the p-value when applied through a hypothesis test are that statistical significance (commonly a p-value less than 0.05 or 5%) is not good evidence that the null hypothesis is wrong, and a p-value above 0.05 is not good evidence that the null hypothesis is right. Further, p-values do not provide information about the magnitude or strength of the association between an exposure and an outcome. For example, a p-value less than 0.001 does not necessarily indicate that the association between exposure and outcome is strong or of public health relevance. Finally, p-values only assess the probability of observing the data or data more extreme assuming the null hypothesis is true due to chance, and do not take into account errors in design, measurement, or analysis of data, including potential non-comparability of exposed and unexposed persons. In observational epidemiology, non-comparability is often a greater threat to validity than a mistakenly significant result.

We do not encourage the use of statistical significance tests as measures of evidence in epidemiologic studies. Rather, we present confidence intervals to provide a range of plausible values for the parameter of interest based on non-comparability that can arise from the sampling process alone. Confidence intervals become more narrow as sample size increases, reflecting greater precision in the resulting measures of association. Moreover, confidence intervals allow the researcher to interpret the role of chance in the study results with more nuance than a binary indicator of significant or not significant.

Major epidemiologic journals now require confidence intervals or more advanced statistical techniques, and stipulate that p-values can be presented along with confidence intervals provided that they are not presented with arbitrary statistical significance cut-offs. Our book reflects this broader movement in epidemiologic science, seeking to appropriately and responsibly use the critical tool of probability to quantify chance in the sampling process without overly relying on these tools to guide our public health decision-making.

Citations

Fisher, R. A. (1950). Statistical methods for research workers. London, Oliver and Boyd.

Gigerenzer, G. (2004). "Mindless statistics." The Journal of Socio-Economics 33: 587-606.

Greenland, S. and C. Poole (2013). "Living with p values: resurrecting a Bayesian perspective on frequentist statistics." Epidemiology 24(1): 62-68.

Meehl, P. E. (1967). "Theory-testing in psychology and physics: A methodological paradox." Philosophy of Science 34(2): 103-115.

Neyman, J. and E. Pearson (1933). "On the problem of the most efficient tests of statistical hypotheses." Philos Trans Roy Soc A 231: 289-337.

Sterne, J. A. and G. Davey Smith (2001). "Sifting the evidence-what's wrong with significance tests?" BMJ 322(7280): 226-231.

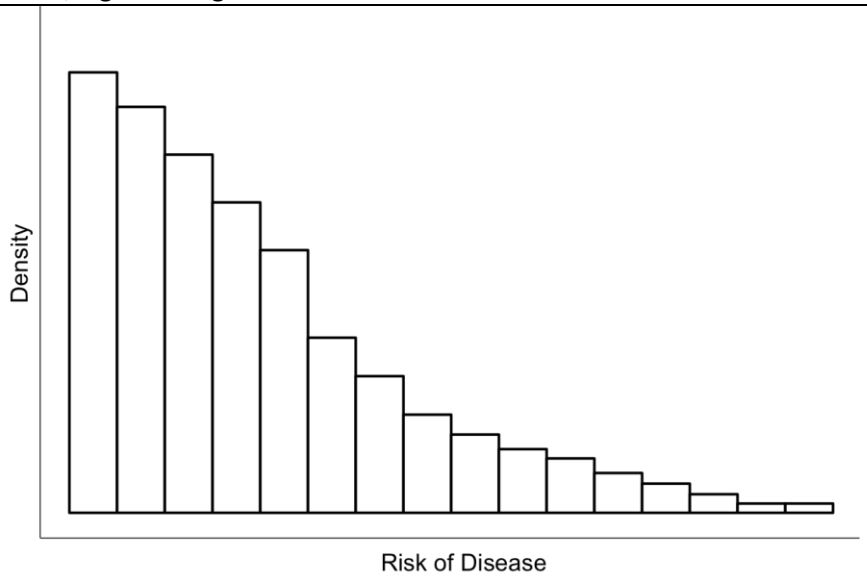
Box 2. What are a log and a natural log, and why do we need to know?

Recall from Chapter 5 that we can quantify and standardize variables to a z-distribution if they are normally distributed. Many variables, however, are not normally distributed. When a random variable is not normally distributed, one solution is to take each value and use some kind of mathematical operator on it in order to re-express the values so that they do correspond to a normal distribution or some other distribution that is easy to work with. Taking the logarithm of each value is one such common re-expression, but many others are also possible (e.g., squaring each number or taking it to an even higher power, taking the square root, etc.).

When would we want to take the logarithm of a random variable in order that the distribution

of the values is better approximated by a normal distribution? Typically, when a variable that is right skewed – that is when there is a pile-up of values close to 0, and a long tail of data to the right.

Box 2, Figure 1. Right-skewed distribution



When we take the log of each of the values in a distribution such as the above, the resulting distribution will appear much closer to a normal distribution.

Even when we discuss taking the logarithm of a variable, there are several different options, which depend on the base of the logarithm. It is common to use logs with respect to base 10, or even more common to use logs with respect to base e, also called natural logarithms. e is a number that has many useful mathematical properties, and is a constant of approximately 2.71828 that has many handy mathematical properties and is useful in determining rates of growth. This is useful for us because if we apply the log with respect to base e to some types of random variables, we can often get a non-normal distribution to be much closer to normal. Then we can standardize the distribution of log values to a Z-distribution in order to estimate confidence intervals and conduct hypothesis tests.

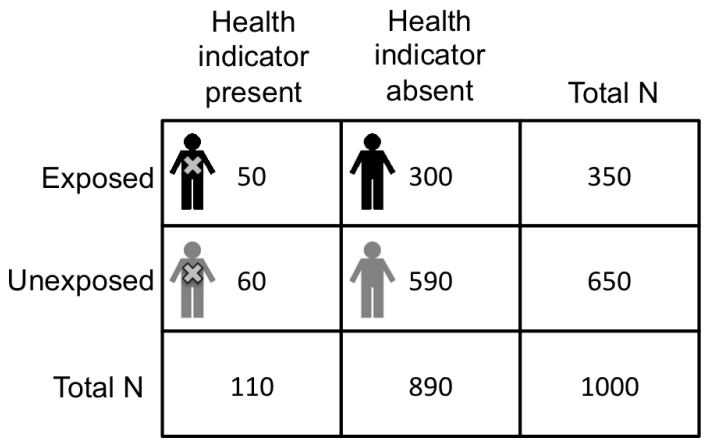
As it turns out, risk ratios, rate ratios, and odds ratios are not normally distributed. The smallest possible value of any of these variables is 0, but the largest is infinity. However, they are approximately normally distributed on the natural log scale. Thus, we take the natural log (log to base e) of ratio measures before estimating the standard error. We can then use a normal distribution to get an accurate approximation of the potential range of ratios measures accounting for sampling variability.

Box 3. Example of 95% confidence interval for a risk ratio

We conduct a study among 1,000 Farrlandians measuring the association between having a family history of Alzheimer's disease (AD) and the incidence of AD among those aged >70. We select a random sample of 1,000 individuals aged >70 with no symptoms of AD and follow them

for 20 years, measuring symptoms of AD every year. There are no losses to follow-up. We obtain the data shown in Box 3, Figure 1.

Box 3, Figure 1. Association between family history of Alzheimer's disease and the incidence of Alzheimer's disease



$$\text{Risk ratio} = \frac{\left(\frac{50}{350}\right)}{\left(\frac{60}{650}\right)} = 1.55$$

In our study sample, those with a family history of AD have approximately 1.55 times the risk of developing AD over 20 years compared with those who do not have such a family history. Now, let us construct a confidence interval around this risk ratio in order to understand the range of values that are plausible in these data given sampling variability.

Step 1: Take the natural log of the risk ratio

$$\text{Ln}(\text{Risk ratio}) = \text{Ln}(1.55) = 0.44$$

Step 2: Estimate the standard error of the log of the risk ratio

$$\text{SE}(\text{Ln}[\text{Risk ratio}]) = \sqrt{\left(\left(\frac{1}{50}\right) - \left(\frac{1}{350}\right) + \left(\frac{1}{60}\right) - \left(\frac{1}{650}\right)\right)} = 0.18$$

Step 3: Estimate upper and lower bound confidence intervals on the log scale

Upper bound:

$$0.44 + (1.96 * 0.18) = 0.79$$

Lower bound:

$$0.44 - (1.96 * 0.18) = 0.09$$

Step 4: Take the antilogarithm to obtain the upper and lower bounds of the confidence interval

$$e^{.79} = 2.20$$

$$e^{.09} = 1.09$$

Step 5: Report and interpret the estimate and the confidence interval

Individuals >70 in Farrlandia with a family history of AD had 1.55 times the risk of developing AD over 20 years, with a 95% confidence interval for the risk ratio of 1.09 to 2.20.

In summary**,** our study in Farrlandia indicates that, accounting for sampling variability, the association between family history of AD and risk of AD could be as low as 1.09 or as high as 2.20. Would we conclude from this that chance in the sampling process could explain our findings? Well, even at the lowest bound, there remains a small association between family history of AD and risk of AD; that is, the confidence interval does not include 1.0 (if 1.0 were in the interval, this would suggest that the absence of an association, or sometimes referred to as the null, is possible within sampling variability). Thus, after accounting for sampling variability alone, we would not rule out the existence of an association. However, note that this confidence interval does not take into account any systematic error in the measurement of our constructs, or any potential non-comparability between those with a family history of AD and those without. We will detail these systematic error processes in Chapters 8 through 10.

Box 4. What are odds, and what is the relationship between odds and risk?

An odds is the probability that an event occurs divided by the probability that the event does not occur. For example, suppose that we have a classroom of 400 introductory epidemiology students, and we want to know the odds that someone in the classroom has a cold. We have our graduate student teaching assistants count the number of individuals who either cough or sneeze during our lecture as a measure of having a cold (for the purpose of this example, we will assume that all coughs and sneezes perfectly indicate having a cold). A total of 10 students cough/sneeze. The odds of a cold would thus be as follows:

The proportion who have a cold divided by 1 minus the proportion who have a cold:

$$\frac{10}{400} = 0.03$$

$$1 - \left(\frac{10}{400}\right) = 0.97$$

$$\frac{0.025}{0.975} = 0.03$$

Therefore, the odds of having a cold in the introductory epidemiology class are 0.03.

We could also reduce the proportions above for a more intuitive interpretation:

$$\frac{0.03}{0.03} = 1$$

$$\frac{0.97}{0.03} = 32.33$$

The odds of having a cold are about 1 to 32, or alternatively, the odds that anyone in the classroom does not have a cold are about 32 to 1.

To summarize the odds, a general formula for calculating an odds is:

$$\frac{\mathrm{P(D)}}{1 - \mathrm{P(D)}}$$

Where P(D) is the proportion with the disease of interest.

Now, what is the relation between the odds and the risk? Let us go back to our cold example. The proportion of individuals with a cold, also known as the risk of a cold, was 10/400, or 0.03. The odds of having a cold were 0.03. Do these seem similar? Yes! To see why, let us look again at the calculation of the odds:

$$\frac{0.03}{0.97}$$

0.97 is very close to 1.0; if we were to divide 0.03 by 1.0, we would get the same number as the risk of having a cold (0.03). As the denominator of the odds is closer and closer to 1.0, the odds will be closer and closer to the calculation of the risk. Because the denominator of the odds is the complement of the numerator, we can also say that as the risk of disease becomes more rare, the odds will increasingly be closer to the risk because the denominator will be closer and closer to 1.0.

Here are three examples that show this principle more clearly. Suppose we have three diseases, with the following incidence proportions: 1%, 20% and 45%. Let us see how closely the odds will approximate the risk in the three examples:

Disease 1:

$$\mathrm{Risk} = 0.01$$

$$\mathrm{Odds} = \frac{0.01}{1 - 0.01} = \frac{0.01}{0.99} = 0.0101$$

Disease 2:

$$\text{Risk} = 0.20$$

$$\text{Odds} = \frac{0.20}{1 - 0.20} = \frac{0.20}{0.80} = 0.25$$

Disease 3:

$$\text{Risk} = 0.45$$

$$\text{Odds} = \frac{0.45}{1 - 0.45} = \frac{0.45}{0.55} = 0.82$$

We see that the odds of Disease 1 are very similar to the risk of disease 1: 0.01 versus 0.0101. The odds of Disease 2 begin to diverge from the risk, but they are still relatively close: 0.20 versus 0.25. For disease 3, the odds of disease are much different than the risk of disease: 0.45 versus 0.82.

In summary, the odds will approximate the risk when the outcome is rare. This has been known in epidemiology as the rare disease assumption for interpreting odds ratios from case control studies, however this is somewhat of a misnomer as the extent to which the odds ratio will approximate the risk ratio depends on exactly how the controls were sampled for the case control study (see Box 3 of the online material that accompanies for Chapter 6).

Box 5. Why is the exposure odds ratio equivalent to the disease odds ratio?

Below we provide the mathematical notation that shows why the exposure odds ratio is equivalent to the disease odds ratio. We begin with a basic 2x2 table cross-classifying exposure and disease.

Box 5, Figure 1. General 2x2 notation with person years

	Health indicator present	Health indicator absent	**Total**	Person years
Exposed	a	b	a+b	person years exposed
Unexposed	c	d	c+d	person years exposed
Total	a+c	b+d	total population	total person years

Disease odds in the exposed:

$$\frac{\frac{a}{a+b}}{1-\left(\frac{a}{a+b}\right)} = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \left(\frac{a}{a+b}\right) * \left(\frac{a+b}{b}\right) = \frac{a}{b}$$

Disease odds in the unexposed:

$$\frac{\frac{c}{c+d}}{1-\left(\frac{c}{c+d}\right)} = \frac{\frac{c}{c+d}}{\frac{d}{c+d}} = \left(\frac{c}{c+d}\right) * \left(\frac{c+d}{d}\right) = \frac{c}{d}$$

Disease odds ratio:

$$\frac{a \div b}{c \div d} = \frac{a * d}{b * c}$$

The disease odds ratio reduces to a*d/b*c.

Exposure odds ratio in the diseased:

$$\frac{\frac{a}{a+c}}{1-\left(\frac{a}{a+c}\right)} = \frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \left(\frac{a}{a+c}\right) * \left(\frac{a+c}{c}\right) = \frac{a}{c}$$

Exposure odds ratio in the nondiseased:

$$\frac{\frac{b}{b+d}}{1-\left(\frac{b}{b+d}\right)} = \frac{\frac{b}{b+d}}{\frac{d}{b+d}} = \left(\frac{b}{b+d}\right) * \left(\frac{b+d}{d}\right) = \frac{b}{d}$$

Exposure odds ratio:

$$\frac{a \div c}{b \div d} = \frac{a * d}{b * c}$$

The exposure odds ratio reduces to a*d/b*c.

Exposure odds ratio = Disease odds ratio

$$\frac{a * d}{b * c} = \frac{a * d}{b * c}$$