8.6 Inference for two samples of binary data

NOTE: This supplement to the first edition is being released online as an extension to Chapter 8. Its pagination corresponds to the printed first edition text. This supplement includes a set of exercises and solutions to the odd-numbered exercises.

Methods for summarizing data from two samples of binary data were introduced in Sections 1.6.2 and 8.5, where relative risk and odds ratios were introduced. Methods for inference for the difference of two proportions were discussed in Sections 8.2 and 8.3. This supplement provides more details about those measures when used to compare two groups, adding information about terminology, interpretation, hypothesis tests, and confidence intervals for relative risk and odds ratios. For convenience, some of the earlier material on inference for the difference in two proportions is reviewed in this supplement.

8.6.1 Introduction and terminology

This section reviews concepts and terminology for summary measures of association in two groups (e.g., an intervention or exposure) when the outcome has two values (e.g., yes/no, success/failure, or 0/1). The ideas summarized here are discussed in more detail in later sections.

Risk ratio (relative risk) and risk difference

The term risk is often used in statistics and epidemiology as the likelihood of a condition or an outcome from a disease, such as the risk of diabetes among young adults or the risk of Covid-19 infection in a particular population. Since prevalence is a better term for the likelihood of a condition in a population, this section limits the term risk to settings in which the potential causal effect of an intervention or exposure is examined, such as a study examining whether a new treatment for Covid-19 reduces the risk of severe disease in a cohort of infected individuals. A causal effect can be estimated only in studies where an outcome is measured after a study participant has received an intervention or experienced an exposure. Randomized experiments are the best designs for estimating a potential causal effect since they reduce the chance of confounding, but risk can also be suggested by some observational studies, so the term risk is used here more widely than for just randomized experiments.

In the LEAP study discussed in the first section of Chapter 1 the study team investigated whether an intervention beginning in infancy would reduce the risk of a child developing a peanut allergy by age 5 years. Of the 263 children randomized to the peanut avoidance group, 36 showed signs of a peanut allergy at 5 years of age. The estimated risk of an allergy is $36/263 = 0.137$. Five of the 267 assigned to the peanut consumption group experienced signs of an allergy, for an estimated risk of $5/267 = 0.019$. The estimated risk ratio (also called relative risk)[23] of developing an allergy, comparing the avoidance to the consumption group, is $0.137/0.019 = 7.21$. Children in the avoidance group were more than 7 times as likely to develop an allergy compared to those in the consumption group.

The estimated risk difference in the LEAP study is $0.137 - 0.019 = 0.118$. On an additive scale, the risk of a peanut allergy increases by slightly more than 0.10 (10%) for children in the avoidance group. Risk ratios and differences provide important summary statistics when comparing groups, but in some settings one is more informative than the other. When overall risk is small, as is the case with peanut allergies, risk ratios are often more informative. A small risk difference of 0.118

[23] The term relative risk is more common than risk ratio, but the latter is more descriptive and is used in this section.

is associated with a large multiplicative increase in the probability of outcome. When overall risk is larger, a risk ratio may potentially obscure the magnitude of an effect. For example, suppose overall risk is 0.40 and an intervention under study is thought to reduce risk to 0.35. In a large population, this reduction in absolute risk of 0.05 may be clinically relevant; in a population of 1,000,000 a reduction in risk from 0.40 to 0.35 will reduce the occurrence of the condition from 400,000 to 350,000, affecting 50,000 individuals. The relative risk of 0.40/0.35 = 1.14 does not convey the same message as the risk difference. Whichever summary statistic is used as the primary measure of comparison, both should be provided in the interpretation of a study.

The calculation of risk ratio in the LEAP study used the peanut consumption group as the baseline. The risk in the peanut avoidance group could have been used for the baseline, yielding a relative risk of 0.019/0.137 = 0.139. This risk of allergy in the consumption group is approximately 0.14 times that of avoidance group. While there is no set convention for the choice of the baseline group, risk ratios greater than 1 are easier for most people to interpret so the baseline group is usually chosen to be the one with the smaller risk.

Prevalence ratio and prevalence difference

The calculations for prevalence ratios and differences mirror those for risk ratios and differences, but the different terminology reflects an important difference in interpretation. The prevalence of a disease is the proportion of a population experiencing the disease. Cross-sectional studies sample a population during a prespecified (usually short) time interval and can be used to estimate the prevalence of a disease and features of the population that may be associated with the disease. Since a cross-sectional study does not measure an outcome occurring subsequent to an exposure, it cannot estimate risk of an outcome from an exposure. Cross-sectional studies can, however, provide important information about the association between outcome and features of a population that might justify additional studies.

The US CDC estimates that approximately 14.9% of non-Hispanic Asian adults in the United States have Type 2 diabetes (T2D);[24] the prevalence of T2D in this population is 0.149. For non-Hispanic white adults, the prevalence of T2D is 0.119 (11.9%). The prevalence difference between the groups, comparing non-Hispanic Asian to non-Hispanic white adults, is 0.149 - 0.119 = 0.03. The prevalence ratio comparing Asian to white non-Hispanics is 0.149/0.119 = 1.252. The prevalence of T2D for Asian adults is 1.252 times as large as that for white adults.

Odds ratios

Odds ratios are used to estimate an association between an outcome and exposure when baseline risk or prevalence cannot be estimated, such as in a case-control study. In a dataset, the observed odds of an event is the number of times the event happens divided by the number of times it does not. The odds ratio (OR) is the odds of an event occurring in one group divided by the odds of an event occurring in the baseline group. Somewhat surprisingly, even when risk or prevalence ratio cannot be estimated, the OR comparing the odds of an outcome between exposed and unexposed groups can.

Figure 8.19 in Section 8.5 summarizes the results of a study examining the association of persistent pulmonary hypertension of a newborn (PPHN) with exposure to maternal use of a selective serotonin re-uptake inhibitor (SSRI) during pregnancy. For convenience, the figure is repeated here as Figure 8.22.

Participants in the PPHN study were sampled and grouped according to whether their infants did or did not suffer from PPHN; the study did not count the number of PPHN outcomes among women using an SSRI during pregnancy. Thus, the absolute risk of PPHN given SSRI use,

[24] Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020.

PPHN present	Yes	No	Total
SSRI exposed	14	6	20
SSRI unexposed	323	830	1153
Total	337	836	1173

Figure 8.22: SSRI exposure versus observed number of PPHN cases in newborns.

$P(\text{PPHN}|\text{SSRI})$, cannot be estimated from these data. Since it is not possible to compute the relative risk of PPHN comparing the SSRI groups, the OR is used instead.

Among the 337 cases, 14 were exposed to an SSRI and 323 were not, so the estimated odds of exposure among the cases are 14/323. Similarly, the estimated odds of SSRI exposure among the controls are 6/830. The estimated OR compares the odds of exposure among the cases to that among the controls:

$$\widehat{\text{OR}}_{\text{exposure, cases vs. controls}} = \frac{14/323}{6/830} = \frac{(14)(830)}{(323)(6)} = 6.00.$$

Infants exposed to SSRI during maternal pregnancy have 6 times the odds of PPHN than unexposed infants.

8.6.2 Inference for risk or prevalence differences

Confidence intervals and tests for risk or prevalence differences use the methods for comparing two binomial proportions outlined in Section 8.2. When the conditions described in Section 8.2.1 are met, a 95% confidence interval for the difference $p_1 - p_2$ of two proportions is given by

$$\hat{p}_1 - \hat{p}_2 \pm (1.96 \times \text{SE}_{\hat{p}_1 - \hat{p}_2}).$$

The estimates $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of the outcome of interest in the two groups, and the standard error SE of the difference in estimated proportions is given by

$$\text{SE}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

where n_1 and n_2 are the two group sizes.

EXAMPLE 8.33

Non-inferiority designs are used in clinical trials when an intervention may not be as good as the standard of care but has other advantages, such as having fewer side effects or being less expensive. In a 2021 study Bernard, et al.[a] reported the results of a randomized trial comparing 6 versus 12 weeks of antibiotic therapy for prosthetic joint infection. While twelve weeks of therapy was known to be effective, 6 weeks of treatment would be preferable as an intervention (due to lower cost and more convenience for patients) if the outcomes were not unacceptably worse. The study design specified that 6 weeks of therapy could be considered a viable alternative as long as the upper 95% confidence limit for the difference in risk of persistent infection did not exceed 10% (0.10 when measured as a proportion). In this prospective, randomized trial, risk difference can be estimated directly. In the 6-week treatment group, 35 of 193 evaluable participants had a persistent infection, while in the 12-week group 18 of 191 had a persistent infection. Did the trial establish that 6 weeks of therapy was acceptable?

(E)

The 95% confidence interval for the difference in the risk of persistent infection is

$$\left(\frac{35}{193} - \frac{18}{191}\right) \pm 1.96\sqrt{\frac{(\frac{35}{193})(1-\frac{35}{193})}{193} + \frac{(\frac{18}{191})(1-\frac{18}{191})}{191}} \quad \rightarrow \quad (0.018, 0.156).$$

The trial did not establish non-inferiority of the 6-week course because the upper bound of the 95% confidence interval for the risk difference exceeds 0.10. The data suggest that 6 weeks of therapy could lead to approximately 16% more persistent infections. In fact, since the confidence interval for the risk difference does not include 0, the data suggest that the 6-week therapy might be statistically significantly worse than the 12-week course of therapy.

[a] Louis Bernard et al. "Antibiotic Therapy for 6 or 12 Weeks for Prosthetic Joint Infection". In: *New England Journal of Medicine* 384.21 (2021), pp. 1991–2001. DOI: 10.1056/NEJMoa2020198. eprint: https://doi.org/10.1056/NEJMoa2020198. URL: https://doi.org/10.1056/NEJMoa2020198.

The use of a confidence interval in Example 8.33 is the proper method of inference, since the null hypothesis of no difference between the therapies was not relevant. There are many instances, however, in which a test of the hypothesis of no difference between groups is a central part of the analysis.

There are two widely used methods for testing the null hypothesis of no difference in risk or prevalence between two groups: 1), using a z test based on the approximate normal sampling distribution for the difference of two sample proportions (Section 8.2); and 2), using the χ^2 test for a 2×2 table (Section 8.3). The z statistic for testing $H_0 : p_1 = p_2$ (equivalently, $p_1 - p_2 = 0$) versus $H_A : p_1 \neq p_2$ (equivalently, $p_1 - p_2 \neq 0$) is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $\hat{p}_1$ and $\hat{p}_2$ are the two estimated proportions based on group sizes n_1 and n_2, and $\hat{p}$ is a pooled estimate of the outcome probability p under the null hypothesis of no difference,

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}.$$

The National Health and Nutrition Examination Survey (NHANES), introduced in Section 1.6, is a cross-sectional study conducted by the US CDC designed to assess the health and nutritional status of adults and children in the United States. The survey began in 1960 and was conducted approximately every 5 years until 1994, when the CDC began conducting the survey continuously.

A random sample of 500 adults from the dataset NHANES was used in Example 1.14 to illustrate scatterplots of height versus BMI and height versus weight.

The NHANES sample also contains responses to the questions of whether the participant has smoked at least 100 cigarettes in their lifetime and whether the participant uses marijuana regularly, with the responses shown in Figure 8.23. The data are used here to illustrate confidence intervals and tests for prevalence differences. These data come from a survey conducted between 2009 and 2012, and only 309 of the 500 participants responded to both questions, so the data provide no information about current marijuana use.

Reg. Marijuana Use	Yes	No	Total
Smoke $\geq$ 100 cig.	57	78	135
Smoke $<$ 100 cig.	23	151	174
Total	80	229	309

Figure 8.23: Smoking history versus regular marijuana use, observed counts.

EXAMPLE 8.34

Calculate a 95% confidence interval for the difference in the prevalence of regular marijuana use between individuals who have smoked at least 100 cigarettes in a lifetime versus have not. Use a z test to assess the evidence against the null hypothesis that the prevalence difference is 0.

The estimated prevalences of regular marijuana use are 57/135 = 0.422 and 23/174 = 0.132, respectively, for a prevalence difference of 0.422 - 0.132 = 0.29, and the 95% confidence interval for the prevalence of regular use is

(E)
$$\left(\frac{57}{135} - \frac{23}{174}\right) \pm 1.96\sqrt{\frac{(\frac{57}{135})(1-\frac{57}{135})}{135} + \frac{(\frac{23}{174})(1-\frac{23}{174})}{174}} \quad \rightarrow \quad (0.193, 0.387).$$

The pooled estimate of prevalence is $\hat{p} = (57+23)/(135+174) = 0.259$. The z–statistic is

$$\frac{0.29}{\sqrt{(0.259)(1-0.259)\left(\frac{1}{135}+\frac{1}{174}\right)}} = 5.15.$$

The z statistic has p-value < 0.001; the test and the confidence interval both support the conclusion that based on these data, there is a strong association between smoking and regular marijuana use, with smokers more likely to use marijuana regularly.

Even if these data were current and nearly all participants responded to the questions, this cross-sectional study can estimate only the association between smoking and marijuana use, not the risk that smokers will begin using marijuana.

The χ^2 test statistic can also be used with the data in Figure 8.23 to test the null hypothesis of no association between smoking and marijuana use. Under the null hypothesis of no association, smoking status provides no information about marijuana use, making the column variable (marijuana use) independent of the row variable (smoking). Under this hypothesis, the observed and expected counts within each cell should be approximately equal. The statistic is

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

The calculation of the expected counts is described in Section 8.3.1. The statistic has approximately a χ^2 distribution with one degree of freedom as long as the conditions outlined in Section 8.3.2 are

met.

EXAMPLE 8.35

Conduct a χ^2 test of the null hypothesis of no association between smoking and marijuana use.

Figure 8.24 shows the expected cell counts calculated under the hypothesis of no association (i.e., independence), using the formula outlined in Section 8.3.1. For example, the expected cell count in the upper left corner of the table is

(E)

$$\frac{\text{(row 1 total)(column 1 total)}}{\text{sample size}} = \frac{(135)(80)}{309} = 34.95.$$

The conditions for the test are easily met – the participants were sampled and surveyed independently and the expected cell counts are all at least 10. The χ^2 statistic has value 33.3 with 1 degree of freedom and, like the z test, has a p-value <0.001.

Regular Marijuana Use	Yes	No	Total
Smoke ≥ 100 cig.	34.95	100.05	135
Smoke < 100 cig.	45.05	128.95	174
Total	80	229	309

Figure 8.24: Smoking history versus regular marijuana use, expected counts.

The χ^2 and z tests are equivalent in that both will either reject or fail to reject the null hypothesis together—both provide the same p-value; in fact, the square of the z statistic equals the χ^2 statistic. The details provided by the two approaches are different, however. The χ^2 test is based on a data summary that is compact and easily understood (i.e., a 2×2 table). The approach based on inference for a difference of two proportions (z test) provides a confidence interval for the difference that is not available from the χ^2 test. Sometimes a confidence interval is the main tool for inference. In Example 8.33, there was no null hypothesis of equality of risk of persistent infection; this is a case in which the χ^2 test would not have answered the scientific question of interest. The χ^2 test is inherently two-sided, since it assesses evidence that the rows and columns are not independent, while the z statistic can be used for either one- or two-sided tests.

The standard normal and χ^2 distributions for these test statistics are continuous distributions that only approximate the sampling distributions of the statistic; there are alternative versions of these test statistics that either attempt to improve the approximation with small adjustments (i.e., continuity corrections), or avoid the approximation altogether by using theoretically exact sampling distribution (exact methods). These alternatives are discussed in Section 8.6.5.

8.6.3 Inference for risk and prevalence ratios

Prevalence and risk ratios can also be used to summarize the differences between two groups in cross-sectional studies and studies in which an exposure or intervention precedes an outcome. In the NHANES data in Figure 8.23, the prevalence ratio for regular marijuana use, comparing smokers to non-smokers, is

$$\widehat{\text{PR}} = \frac{(57/135)}{(23/174)} = 3.19.$$

The null hypothesis $H_0 : \text{PR} = 1$ is equivalent to a prevalence difference of 0, so the χ^2 statistic calculated in Example 8.35 supports the conclusion of a prevalence ratio different from 1.

In Example 8.33 the risk of persistent infection in the group treated for 6 weeks was 35/193 = 0.181; the risk in the 12 week group was 18/191 = 0.094, for a risk ratio of 1.93. The 6-week group is almost 2 times more likely to experience persistent infection.

Since confidence intervals for a RR or PR use the same calculation, the steps for computing a confidence interval are phrased in terms of risk ratios. A confidence interval for a risk ratio is a two-step process, starting with a confidence interval for the natural log of the RR, then exponentiating the upper and lower bounds to obtain upper and lower bounds for the RR.

CONFIDENCE INTERVAL FOR LOG RISK RATIO

Suppose E is an event that has two possible outcomes, labeled *yes*, *no*. Let y_1 and y_2 be the observed counts of the value *yes* in two groups of size n_1 and n_2, let the estimated proportions of *yes* outcomes be $\hat{p}_1 = y_1/n_1$ and $\hat{p}_2 = y_2/n_2$, and let $\widehat{\text{RR}} = \hat{p}_1/\hat{p}_2$ be the estimated population RR comparing group 1 to group 2. If the two groups can be viewed as random samples from a larger population and the conditions described in Section 8.2.1 are met, $\log(\widehat{\text{RR}})$ is approximately normally distributed with mean log(RR) and standard error (SE)

$$\text{SE}_{\log(\widehat{\text{RR}})} = \sqrt{\frac{1-\hat{p}_1}{y_1} + \frac{1-\hat{p}_2}{y_2}}.$$

A $100(1-\alpha)\%$ confidence interval for log(RR) is given by

$$\log(\widehat{\text{RR}}) \pm (z^\star \times \text{SE}), \tag{8.36}$$

where $z^\star$ is the point on a z distribution with area $(1-\alpha/2)$ in the left tail.

EXAMPLE 8.37

Calculate a 95% confidence interval for RR in the joint infection trial, comparing 6 weeks to 12 weeks of treatment.

The estimated probabilities are $\hat{p}_{6\text{wk}} = 0.181$ and $\hat{p}_{12\text{wk}} = 0.094$, so the standard error of log(RR) is

$$\text{SE} = \sqrt{\frac{1-0.181}{193} + \frac{1-0.094}{191}} = 0.095.$$

(E)

The 95% confidence interval for the log(RR) is

$$\log\left(\frac{0.181}{0.094}\right) \pm (1.96)(0.095) = 0.655 \pm 0.186$$
$$\rightarrow (0.468, 0.841).$$

The 95% interval for RR is $(e^{0.468}, e^{0.841}) = (1.597, 2.318)$. With 95% confidence, the risk of persistent infection on the 6-week therapy is between 1.6 and 2.3 times the risk on the 12-week therapy.

Examples 8.33 and 8.37 illustrate the use of both risk difference and ratio for the same study. The study was designed to estimate the treatment effect on risk difference, so the investigators presented that as their primary analysis. The upper bound of the confidence interval for the risk difference showed that the difference in the risk of persistent infection could be as much as 16%. Since the overall risk is relatively low (approximately 9% on the 12 week treatment), the upper bound of the risk difference of approximately 16% translates into a upper bound for the RR of approximately 2.3. The 6 week therapy could lead to more than 2 times the risk of persistent infection compared to the 12 week therapy.

8.6.4 Inference for odds ratios

The odds of an event E are $P(E)/(1-P(E))$; odds are the proportion of times an event occurs divided by the proportion it does not. Figure 8.25 shows that probabilities and odds increase or decrease together; thus, a factor that is associated with a change in the probability of an event will also be associated with a change in the odds of the event and vice versa. The figure also demonstrates that while probabilities always have values between 0 and 1, odds can be much larger than 1; odds should not be interpreted as probabilities.

	Probability	Odds
1	0.05	0.05
2	0.10	0.11
3	0.20	0.25
4	0.30	0.43
5	0.40	0.67
6	0.50	1.00
7	0.60	1.50
8	0.70	2.33
9	0.80	4.00
10	0.90	9.00
11	0.95	19.00

Figure 8.25: Probability versus odds for selected values.

In a 2019 paper in the journal *Headache*, Togha et al.[25] report a case-control study examining the association between migraine headaches and vitamin D levels. The investigators enrolled 70 healthy individuals (the controls) and 70 age- and sex-matched individuals with either chronic or episodic migraine headaches (the cases), and measured vitamin D levels (the exposure) in both cases and controls. Figure 8.26 shows the number of participants classified by vitamin D levels and the presence of migraines. In this table participants were categorized as having low vitamin D if they were either vitamin D deficient or insufficient using standard definitions given in the paper.

Migraine	Yes	No	Total
Vitamin D low	36	18	54
Vitamin D normal	34	52	86
Total	70	70	140

Figure 8.26: Vitamin D level versus presence of migraine.

EXAMPLE 8.38

Calculate the OR of having low vitamin D level in patients with migraines compared to those without.

(E)

The odds of low vitamin D levels among the participants suffering from migraines are 36/34 = 1.06. The corresponding odds among participants not suffering from migraines are 18/52 = 0.35, so the OR = 1.06/0.35 = 3.06.

Since the migraine data are from an outcome-based sampling design, the relative risk of a migraine comparing participants with low versus normal vitamin D levels cannot be estimated from these data. However, the OR comparing the odds of a migraine given low vitamin D levels to the odds of a migraine given normal levels can be calculated – it is, in fact, identical to the OR

[25] M. Togha et al. "Serum Vitamin B12 and Methylmalonic Acid Status in Migraineurs: A Case-Control Study". In: *Headache* 59.9 (Oct. 2019), pp. 1492–1503.

calculated in Example 8.38, 3.06. The data from the migraine study suggest that low vitamin D levels may be associated with a tripling of the odds of chronic or episodic migraines; however, it is important to note that these data are from a small study and that the observed association may be a result of unmeasured confounding.

The general structure of a 2×2 table can be used to show that the OR for outcome given exposure is identical to the OR for exposure given outcome. Typically, 2×2 tables are organized with the exposure as the row variable and the column variable as the outcome, as shown in Figure 8.27.

	Outcome A	Outcome B	Sum
Exposure 1	a	b	$a+b$
Exposure 2	c	d	$c+d$
Sum	$a+c$	$b+d$	$a+b+c+d=n$

Figure 8.27: A general 2×2 table of outcome by exposure.

The odds of Exposure 1 vs. Exposure 2 among participants with Outcome A are a/c, while the odds of Exposure 1 vs 2 with Outcome B are b/d, so the OR for Exposure 1 vs 2, given outcome, is

$$OR_{\text{Exposure 1 versus 2, comparing Outcome A to Outcome B}} = \frac{a/c}{b/d} = \frac{ad}{bc}.$$

The odds of Outcome A versus B with Exposure 1 are a/b, while the odds of Outcome A versus B with Exposure 2 are c/d, so the OR for Outcome A versus B, given exposure, is

$$OR_{\text{Outcome A versus B, comparing Exposure 1 to Exposure 2}} = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

Because of this algebraic identity, it is numerically correct to calculate directly the OR for outcome given exposure regardless of the outcome-based sampling design.

The estimated OR from a table depends on the organization of the rows of the table. Figure 8.28 interchanges the rows in Figure 8.27. The new cross product ratio (cb/da) is the reciprocal of the cross-product ratio from the original table and estimates the OR for outcome A versus outcome B where Exposure 1 is the baseline group instead of Exposure 2. To ensure that the computed OR matches the intent of the analysis, it is advisable to calculate ORs from the definitions of odds and odds ratio rather than automatically compute the cross-product ratio from a 2×2 table.

	Outcome A	Outcome B	Sum
Exposure 2	c	d	$c+d$
Exposure 1	a	b	$a+b$
Sum	$a+c$	$b+d$	$a+b+c+d=n$

Figure 8.28: A general 2×2 table of outcome by exposure, with rows interchanged.

When there is no association between an outcome and an exposure (i.e., the outcome and exposure are independent) the population odds ratio is 1. The null hypothesis of no association H_0 : OR = 1 can be tested against the two-sided alternative H_A : OR $\neq$ 1 with the χ^2 test for the independence of row and column variables in a 2×2 table (provided that the conditions for using the χ^2 test are satisfied). A test of the null hypothesis of no association between vitamin D exposure and migraine headaches uses the same approach described in Example 8.35. The conditions for χ^2 test are met in this example (calculations not shown), and the value of the statistic is 9.77 with one degree of freedom; the p-value for the test is 0.002. This case-control study suggests a significant association between vitamin D and migraine headaches, but it is a small observational study and definitive evidence of a causal relationship would require a prospective randomized trial.

As with relative risks, two steps are used to calculate a confidence interval for an OR: begin with a confidence interval for log(OR), then exponentiate the upper and lower bounds to obtain a

confidence interval for OR.

CONFIDENCE INTERVALS FOR LOG ODDS RATIO

Suppose the data from a study are summarized in a 2×2 table and the estimate of the population OR is computed. When the data are a random sample from a population and the conditions for the validity of the χ^2 test are met, the estimated log(OR) is approximately normally distributed with mean equal to the population OR and standard error (SE) given by

$$\text{SE}_{\log(\widehat{\text{OR}})} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}},$$

where a, b, c, and d are the four cell counts.

A $100(1-\alpha)\%$ confidence interval for log(OR) is given by

$$\log(\widehat{\text{OR}}) \pm (z^\star \times \text{SE}), \tag{8.39}$$

where $z^\star$ is the point on a z distribution with area $(1-\alpha/2)$ in the left tail.

EXAMPLE 8.40

Calculate a 95% confidence interval for the estimated OR in the migraine study.

The $\log(\widehat{\text{OR}}) = \log(3.06) = 1.118$. Its standard error is

(E)

$$\text{SE} = \sqrt{\frac{1}{36} + \frac{1}{18} + \frac{1}{34} + \frac{1}{52}} = 0.363.$$

A 95% confidence interval for the log(OR) is

$$1.118 \pm 1.96(0.363) \quad \rightarrow \quad (0.406, 1.830).$$

The corresponding confidence interval for the OR is $(e^{0.406}, e^{1.830}) = (1.501, 6.234)$.

8.6.5 Alternative versions of statistics

Continuity corrections

Confidence intervals and p-values usually rely on a result from theoretical statistics that the sampling distributions of nearly all test statistics are approximately normal in moderate to large sample sizes. The result follows from the Central Limit Theorem, introduced in Section 4.1.1 for the special case of the sample mean. Perhaps surprisingly, the use of the χ^2 distribution in the test for independence in a 2×2 table also relies on the Central Limit Theorem.

The normal distribution is a continuous distribution – all values are possible. In contrast, the possible values of the χ^2 statistic correspond to all the ways the counts in 4 cells of a table can be rearranged for a fixed sample size; there are a finite number of such rearrangements, and each rearrangement yields a value of the statistic. Continuity corrections have been proposed to improve the accuracy of using a continuous distribution to approximate a discrete one when sample size is small, and many texts discuss these modifications. The continuity correction for the χ^2 statistic, for instance, reduces the absolute value of each difference between observed and expected counts

by 0.5 before squaring the difference:

$$\chi^2_{cc} = \sum_{\text{all cells}} \frac{(|\text{observed} - \text{expected}| - 0.5)^2}{\text{expected}}. \tag{8.41}$$

Similar modifications can be made to the test for the difference of two proportions.

It is important to be aware of these continuity corrections but they are not highlighted in this supplement or in the main text. The exact methods outlined in the next section provide a better alternative in small samples, where, for instance, some of the expected cell counts in a 2×2 table are close to or less than 10. Exact methods are now implemented in all major software packages and are easy to use. The general consensus is that using the continuity correction in small samples tends to result in p-values that are too large, so that a test being conducted using $\alpha = 0.05$ fails to reject the null hypothesis as often as it should.

Exact Methods

Exact methods are based on the actual sampling distribution of a summary statistic (under some assumptions) instead of the normal approximation. Exact sampling distributions are often complicated, so these methods are almost always performed with software rather than by hand. Example 8.7 illustrates an exact calculation of a p-value in a single sample of binomial observations, using data considered by the US FDA when giving approval for the use of Avastin when treating a form of brain cancer; Section 8.3.5 describes the use of Fisher's exact test in a small study examining the use of fecal infusion to treat an infection.

Fisher's exact test is used when row and column totals in the table can be considered known in advance. Under that assumption and the null hypothesis of independence of the row and column variables, the conditional distribution of the count in any particular cell (of the 4 table cells) has the hypergeometric distribution discussed in Section 3.5.3. In case-control studies it is used to test the null hypothesis $H_0 : \text{OR} = 1$. In exposure based sampling designs, such as randomized trials, Fisher's test is used for the equivalent null hypothesis $H_0 : \text{RR} = 1$ or $H_0 : p_1 = p_2$, where p_1 and p_2 are outcome probabilities for two groups. In the fecal infusion study discussed in Section 8.3.5, participants were randomized either to the infusion or standard antibiotic therapy (vancomycin). Fisher's exact test was used to test the null hypothesis that the probability of cure did not differ between the two randomized groups. The dataset in this study was small enough that the p-value corresponding to Fisher's test could be calculated by hand, but the calculation in Section 8.3.5 is shown primarily for instructional purposes. In important analyses, Fisher's test should always be calculated using software. The second lab for Chapter 8 illustrates how to use the R function `fisher.test` to calculate both one- and two-sided p-values.

Leung, et al.[26] report a series of experiments testing the effectiveness of surgical masks in reducing viral shedding, i.e., the addition of viral particles to the nearby environment from the breathing of infected individuals. Individuals infected with one of seasonal coronavirus, influenza, or rhinovirus were randomly assigned either to wear or not wear a surgical mask. The study team measured viral shedding through the presence of droplet particles larger than 5 micrometers ($> 5\mu m$) and aerosol particles $\leq 5\mu m$; figure 8.29 contains the data from the experiment measuring shedding of aerosol particles by participants infected with with influenza.

[26] Nancy HL Leung et al. "Respiratory virus shedding in exhaled breath and efficacy of face masks". In: *Nature medicine* 26.5 (2020), pp. 676–680. URL: https://doi.org/10.1038/s41591-020-0843-2.

Particles Present	Yes	No	Total
No mask	8	15	23
Mask	6	21	27
Total	14	36	50

Figure 8.29: Surgical mask wearing versus aerosol viral shedding, influenza.

EXAMPLE 8.42

Do the data in Figure 8.29 support a claim that wearing a surgical mask reduces the chance that an individual with influenza will shed viral particles?

The relative risk of viral shedding, comparing no mask to mask use is (8/23)/(6/27) = 1.57. Individuals with seasonal influenza not wearing a mask are estimated to be 1.57 times more likely to shed particles containing the virus. The expected count for the cell corresponding to no mask and particles detected (the upper left cell) is less than 10 ((14)(23)/50 = 6.44), so the usual χ^2 statistic is not appropriate. The function `fisher.test` in R reports a p-value of 0.36 for the null hypothesis OR = 1 (equivalent to RR = 1), so despite the estimated increase in risk of viral shedding, the table does not support a claim that surgical masks reduce viral shedding from seasonal influenza. The function `riskratio` in the R package `epitools` provides the 95% confidence interval (0.59, 4.70) for the RR after adjusting for the small sample. The confidence interval shows that because of the small sample size, there is considerable uncertainty in the estimate of RR.

Fisher's test was originally proposed by Ronald Fisher in 1934 and its use was initially limited to very small experiments where p-values could be calculated by hand. Data analysts often used the continuity correction to the χ^2 statistic to try to produce analyses similar to exact methods, since exact methods were computationally unavailable. As software for the test became available, Fisher's test was used more often in studies like the fecal infusion study. With current computational power, the test is now used in all but the largest datasets.

Despite its widespread use, Fisher's exact test does have drawbacks, some theoretical, some practical. Some statisticians have questioned the validity of conditioning on the row and column totals, i.e, treating them as if they were known in advance. In randomized experiments like the viral shedding study, the numbers of participants in each group (the row totals) are known once the randomization has been conducted. The numbers of outcomes in the columns will not be known in advance, but research has shown this has little practical impact.

More practically, the discreteness of the hypergeometric distribution may make it impossible to achieve a pre-specified value of α for the test, such as 0.05. An artificial but instructive example illustrates this aspect of the exact test.

EXAMPLE 8.43

Suppose the data in the following table summarize the results of a small randomized trial with 10 participants, in which half are assigned to control and half to treatment. Of those in the treatment group, 3 respond to treatment; only 1 patient in the control group responds to treatment.

	Response	No Response	Total
Treatment	3	2	5
Control	1	4	5
Total	4	6	10

Suppose researchers are interested in understanding whether treatment is superior to control. Enumerate all possible sets of results that favor treatment over control and identify the sets of results that reject the null hypothesis of no association at $\alpha = 0.05$.

The table above favors treatment over control, since the risk ratio comparing treatment to control is (3/5)/(1/5) = 3.00. The only other possible table that favors treatment is the one with a 4 in the upper left table cell. It is not possible to have a table with 5 in the upper left, since the total of individuals who have a response is fixed at 4.

E

	Response	No Response	Total
Treatment	4	1	5
Control	0	5	5
Total	4	6	10

In a table with small cell counts, it is possible to calculate exact test p-values directly instead of relying on the hypergeometric distribution. First, calculate the probability of the observed table under the null hypothesis of independence between treatment and outcome. There are $\binom{10}{5} = 252$ ways to draw 5 individuals from the 10 total individuals; i.e., 252 ways to select 5 individuals out of 10 to be in the treatment group. Given the marginal totals, there are $\binom{4}{3}\binom{6}{2} = 60$ ways to observe 5 individuals of which 3 individuals show a response and 2 individuals do not. Thus, the probability of the observed table is 60/252 = 0.238.

Using similar logic, the probability of the table with 4 in the upper left cell is $\frac{\binom{4}{4}\binom{6}{1}}{252} = 0.024$.

The one-sided p-value for the observed set of results equals 0.238 + 0.024 = 0.262, which is greater than 0.05 and fails to reject the null hypothesis.

A table with 4 in the upper left cell would result in a significant p-value, $p = 0.024$. There is no outcome that produces a p-value between 0.262 and 0.024, due to the discrete nature of the data. Since 0.024 is the largest p-value smaller than 0.05 that can occur for these data, Fisher's test is actually testing at the 0.024 significance level rather than 0.05.

A test that does not reject often enough under the null hypothesis will not reject often enough under the alternative; its power will be less than intended. Fisher's test is widely used, however, because despite the reduction in power, its significance level is guaranteed to be less than 0.05 (or any chosen value of α). When a test has this property, statisticians call it conservative.

There are two reasons for the non-significant result even with a risk ratio of 3.00. Because of the small sample size there is considerable uncertainty in the estimated risk ratio, and the discreteness of the distribution of the test statistic is such that only the most extreme result favoring treatment would have been significant. An increase in the sample size helps with both issues, but the discreteness of the distribution for the exact test continues to have an effect, even as that effect diminishes with increasing sample size.

8.6.6 Design versus the method of analysis

Students of statistics are often surprised by the variety of methods applied to data in simple 2×2 tables. Even experienced statisticians are sometimes uncertain about how to proceed. There are, however, a few guidelines which help in starting an analysis.

Exposure-based sampling

If the study randomized participants to one of two interventions or sampled participants according to exposure to a risk factor, a risk ratio or risk difference is the preferred summary statistic. There is no widespread agreement on the choice between risk ratio and risk difference, so in many instances it is appropriate to calculate both and provide carefully worded interpretations. When absolute risk is small, a small risk difference may imply that the difference between groups is unimportant. In these settings, a risk ratio may show that a member of a group is substantially more likely to experience an outcome when compared to a member of the other group, such as in the LEAP study discussed in Section 8.6.1, where a risk difference of 0.118 for a peanut allergy at age 5 years corresponded to a risk ratio of 7.21. Whenever a risk ratio is reported, however, it is important to give the baseline risk. If the absolute risk in both groups is small, the risk ratio without the baseline risk can be misleading.

An odds ratio is always a valid measure of association in a 2×2 table, but in randomized experiments or exposure-based sampling, the odds ratio should not be used as the primary summary statistic. People unfamiliar with statistics tend to mistakenly interpret the OR as the ratio of probabilities and so think of it as a risk ratio. Odds ratios also tend to be larger than risk ratios, sometimes strikingly so. In the viral shedding experiment, the OR should not be used as a primary summary statistic. The estimated RR is 1.56; the odds of viral shedding among those without masks are 8/15 and are 6/21 for those wearing masks, so the estimated the OR is 1.87, substantially larger than 1.56. Reporting the OR instead of the RR could lead to a news account claiming that not wearing a mask led to viral shedding at almost twice the rate of viral shedding when wearing a mask.

Cross-sectional studies

Cross-sectional studies such as NHANES measure a potential exposure and outcome at the same time; these studies estimate the prevalence of exposure and outcome and the association between them. If the participants are a random sample from a population, prevalence differences and ratios can be estimated from the data but cannot support the conclusion that the potential exposure leads to a change in the risk of the outcome, especially when the outcome might change behavior. In the NHANES example on smoking and marijuana use moderate to heavy smokers may use marijuana more often, but the reverse may also be true – regular users of marijuana may tend to begin smoking cigarettes more often.

An OR is often used as a measure of association in cross-sectional studies, but as with randomized trials, it should not be interpreted as a prevalence ratio. In Example 8.35 the OR for regular marijuana use, comparing individuals who have smoked more than 100 cigarettes lifetime to those who have not is

$$\frac{57/23}{23/174} = 4.80,$$

a value that is substantially larger than the prevalence ratio of 3.19. Odds ratios from cross-sectional studies are sometimes called prevalence odds ratios (POR).

Case-control studies

In case-control studies, an OR should be used as the primary summary statistic; risk ratios and differences should not be calculated. Case-control studies enroll participants according to outcome and can estimate the probability of exposure given observed outcome but not the probability of outcome given exposure – neither absolute nor relative risk can be estimated. It might be tempting to use the data in Figure 8.26 to calculate the risk ratio for migraine headaches comparing participants with low versus normal vitamin D levels, but the design does not support that calculation.

When outcomes are rare, the estimated OR in a case-control study can be a useful approximation to an RR. In Figure 8.27 suppose Outcome A is the outcome of interest (perhaps the presence of a disease), and let

$$p_1 = P(\text{Outcome A}|\text{Exposure 1})$$

and

$$p_2 = P(\text{Outcome A}|\text{Exposure 2}).$$

The odds ratio for the table is

$$\begin{aligned}\text{OR} &= \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} \\ &= \frac{(p_1)(1-p_2)}{(p_2)(1-p_1)} \\ &= \frac{p_1}{p_2}\frac{1-p_2}{1-p_1} \\ &= \text{RR}\frac{1-p_2}{1-p_1}.\end{aligned}$$

The rare disease assumption is satisfied when the probability of Outcome A is small for both exposures; i.e., both p_1 and p_2 are near 0. When this is true, the OR and RR are approximately equal. Since neither p_1 nor p_2 can be estimated from a case-control study, the rare disease assumption must be justified from additional data external to the study.

Persistent pulmonary hypertension (PPHN) discussed in Section 8.5 is a rare condition; CDC data on birth outcomes show that it occurs in approximately 1.9 per 1,000 live births, so the risk of PPHN in a live birth has probability 0.0019. Section 8.6.1 shows that the OR for the condition is 6.00, comparing women who used an SSRI during pregnancy to those who did not. Because it is a rare condition, the RR for PPHN, comparing women who did with those who did not use an SSRI, is also approximately 6. The 95% confidence interval for the OR returned by `fisher.test`, (2.14, 19.17), can be viewed as an approximate 95% confidence interval for RR.

The potential for confounding

Observational studies should never be used to draw causal conclusions. Even when estimating an association, the potential for confounding is substantial; if all the data from a study can be summarized in a simple 2×2 table, nothing can be done to adjust for confounders. Students should keep in mind that even when the proper method of analysis is applied to a 2×2 table and all the calculations are done correctly, an observed association (either in the form of a statistically significant result or confidence interval for OR/RR that does not include 1) may well be due to unmeasured confounders.

The potential for confounding should be kept in mind even for results that may not contradict intuition. For example, the OR of 6 in the PPHN study is striking and may seem to confirm suspicions that SSRI use during pregnancy is potentially risky, since statistically, there is a strong association between SSRI use and PPHN when the only data available are those summarized in Figure 8.22. However, it is important to consider the scientific context. There may be underlying

conditions in a pregnancy that are associated with both depression and higher risk of PPHN in a newborn.

With additional data, logistic regression can be used to estimate odds ratios after adjustment for other predictors. Links to labs, lecture slides and a supplement on logistic regression can be found on the website for this text (https://www.openintro.org/book/biostat/).

8.7 Notes

Two-way tables are often used to summarize data from medical research and public health studies, and entire texts have been written about methods of analysis for these tables. This chapter covers only the most basic of those methods.

Until recently, Fisher's exact test could only be calculated for 2×2 tables with small cell counts. Research has produced faster algorithms for enumerating tables and calculating p-values, and the computational power of recent desktop and laptop computers now make it possible to calculate the Fisher test on nearly any 2×2 table. There are also versions of the test that can be calculated on tables with more than 2 rows and/or columns. The practical result for data analysts is that the sample size condition for the validity of the χ^2 test can be made more restrictive. Some guidelines still recommend that expected cell counts should be at least 5. This chapter recommends using the χ^2 test only when cell counts in a 2×2 table are greater than 10 since there are no computational barriers to Fisher's test when the smallest expected counts are between 5 and 10.

When cell counts are small, some websites and texts recommend using the modified version of the χ^2 statistic shown in Equation 8.41. This version, called the Fisher-Yates statistic, is no longer used as often as it once was because of the widespread availability of Fisher's exact test.

The Fisher test is not without controversy, at least in the theoretical literature. Conditioning on the row and column totals allows the calculation of a p-value from the hypergeometric distribution, but in principle restricts inference to the set of tables with the same row and column values. In practice, this is less serious than it may seem. For tables of moderate size, the p-values from the χ^2 and Fisher tests are nearly identical and for tables with small counts, the Fisher test guarantees that the Type I error will be no larger than the specified value of α.

The two labs for this chapter examine methods of inference for the success probability in binomial data then generalizes inference for binomial proportions to two-way contingency tables. Lab 2 also discusses measures of association in two-by-two tables. The datasets in the labs are similar to datasets that arise frequently in medical statistics. Lab 1 assesses the evidence for a treatment effect in a single uncontrolled trial of a new drug for melanoma and whether outcomes in stage 1 lung cancer are different among patients treated at Dana-Farber Cancer Institute compared to population based statistics. In Lab 2, students analyze a dataset from a published clinical trial examining the benefit of using a more expensive but potentially more effective drug to treat HIV-positive infants.

8.43 CNS disorder. Suppose an investigator has studied the possible association between the use of a weight loss drug and a rare central nervous system (CNS) disorder. He samples from a group of volunteers with and without the disorder, and records whether they have used the weight loss drug. The data are summarized in the following table:

	Drug Use	
CNS disorder	**Yes**	**No**
Yes	10	2000
No	7	4000

(a) Can these data be used to estimate the probability of a CNS disorder for someone taking the weight loss drug?

(b) For this study, what is an appropriate measure of association between the weight-loss drug and the presence of CNS disorder?

(c) Calculate the measure of association specified in part (b).

(d) Interpret the calculation from part (c).

(e) What test of significance is the best choice for analyzing the hypothesis of no association for these data?

8.44 Asthma risk. Asthma is a chronic lung disease characterized as hypersensitivity to a variety of stimuli, such as tobacco smoke, mold, and pollen. The prevalence of asthma has been increasing in recent decades, especially in children. Some studies suggest that children who either live in a farm environment or have pets become less likely to develop asthma later in life, due to early exposure to elevated amounts of microorganisms. A large study was conducted in Norway to investigate the association between early exposure to animals and subsequent risk for asthma.

Using data from national registers, researchers identified 11,585 children known to have asthma at age 6 years out of the 276,281 children born in Norway between January 1, 2006 and December 31, 2009. Children whose parents were registered as "animal producers and related workers" during the child's first year of life were defined as being exposed to farm animals. Of the 958 children exposed to farm animals, 19 had an asthma diagnosis at age 6.

(a) Do these data support previous findings that living in a farm environment is associated with lower risk of childhood asthma? Conduct a formal analysis and summarize your findings. Be sure to check any necessary assumptions.

(b) Is the relative risk an appropriate measure of association for these data? Briefly explain your answer.

(c) In language accessible to someone who has not taken a statistics course, explain whether these results represent evidence that exposure to farm animals reduces the risk of developing asthma. Limit your answer to no more than seven sentences.

8.45 Tea consumption and carcinoma. In a study examining the association between green tea consumption and esophageal carcinoma, researchers recruited 300 patients with carcinoma and 571 without carcinoma and administered a questionnaire about tea drinking habits. Out of the 47 individuals who reported that they regularly drink green tea, 17 had carcinoma. Out of the 824 individuals who reported they never drink green tea, 283 had carcinoma.

(a) Analyze the data to assess evidence for an association between green tea consumption and esophageal carcinoma from these data. Summarize your results.

(b) Report and interpret an appropriate measure of association.

8.8.6 Inference for two samples of binary data

8.46 Prevalence difference versus prevalence ratio, I.

(a) Assume that the prevalence for a particular disease in two groups is 40.4% and 42%. Calculate the prevalence difference and ratio for the disease, comparing the group with the higher prevalence to the one with the lower prevalence. For each summary measure, provide an interpretation that a non-statistician would understand.

(b) Now assume that the prevalence for a particular disease in two groups is 1.2% and 2.8%. Calculate the prevalence difference and ratio for the disease, comparing the group with the higher prevalence to the one with the lower prevalence. For each summary measure, provide an interpretation that a non-statistician would understand.

8.47 Prevalence difference versus prevalence ratio, II.

(a) Assume that the prevalence for a particular disease in two groups is 10% and 15%. Calculate the prevalence difference and ratio for the disease, comparing the group with the higher prevalence to the one with the lower prevalence. For each summary measure, provide an interpretation that a non-statistician would understand.

(b) Now assume that the prevalence for a particular disease in two groups is 40% and 45%. Calculate the prevalence difference and ratio for the disease, comparing the group with the higher prevalence to the one with the lower prevalence. For each summary measure, provide an interpretation that a non-statistician would understand.

8.48 Birth defects and paternal alcohol consumption. A 2021 study by Zhou et. al[48] in JAMA Pediatrics discussed the possible association of congenital heart defects in a newborn and paternal alcohol consumption. The study was described as a prospective study in which the study team recruited more than 529,090 couples who were planning to become pregnant in the next 6 months, then recorded alcohol consumption and birth defects. Of the participating couples, 364,939 fathers did not drink alcohol before conception (defined as at least drink per week) and 164,151 did. Among the fathers who consumed alcohol, there were 363 birth defects. Among the fathers who did not consume alcohol, there were 246 birth defects.

(a) Should the analysis of this study use risk or odds ratio as summary statistic for the association of paternal alcohol consumption and fetal birth defects? Why?

(b) Calculate the statistic you have recommended in part (a).

(c) Calculate a 95% confidence interval for the measure of association in part (a).

8.49 High salt diet and cardiovascular disease related death. Suppose a retrospective study is done in a specific county of Massachusetts; data are collected on men ages 50-54 who died over a 1-month period. Of 35 men who died from CVD, 5 had a diet with high salt intake before they died, while of the 25 men who died from other causes, 2 had a diet with high salt intake. These data are summarized in the following table.

	CVD Death	Non-CVD Death	Total
High Salt Diet	5	2	7
Low Salt Diet	30	23	53
Total	35	25	60

(a) In this study sample, what are the estimated odds that a male had a high salt diet? A low salt diet?

(b) Among the men where the recorded death was due to CVD, what are the odds that the male had a high salt diet? What are the odds of a low salt diet in the same group?

(c) What is the OR for a CVD related death, comparing a high to a low salt diet?

(d) What is the OR for a death not related to CVD, comparing a high to a low salt diet?

8.50 Diabetes. In the United States, approximately 9% of the population have diabetes.

(a) What are the odds that a randomly selected member of the US population has diabetes?

(b) Suppose that in a primary care clinic, the prevalence of diabetes among the patients seen in the clinic is 12%. What is the probability that a randomly selected patient in the clinic has diabetes? What are the odds of diabetes for that patient?

(c) If in a particular population the probability of diabetes is twice what it is in the general population, does the odds of diabetes double?

[48] Qiongjie Zhou et al. "Association of Preconception Paternal Alcohol Consumption With Increased Fetal Birth Defect Risk". In: *JAMA Pediatrics* 175.7 (July 2021), pp. 742–743. ISSN: 2168-6203. DOI: 10.1001/jamapediatrics.2021.0291. eprint: https://jamanetwork.com/journals/jamapediatrics/articlepdf/2778779/jamapediatrics_zhou_2021_ld_210003_1625081160.23765.pdf. URL: https://doi.org/10.1001/jamapediatrics.2021.0291.

8.51 Treatment for Covid-19, I. Guimarães, et al.[49] reported the results of a randomized trial comparing tofacitinib to placebo in patients in Brazil hospitalized with Covid-19 pneumonia. Since there were no known effective treatments for Covid-19 pneumonia when the trial was conducted, a placebo control group was considered ethical. Of the 145 participants assigned to placebo, 42 experienced the outcome of interest, death or respiratory failure during the 28 day follow-up period; 26 out of the 144 assigned to tofacitinib experienced the outcome.

(a) Calculate a 95% confidence interval for the between group difference in the risk of death or respiratory failure.

(b) Conduct a test of the hypothesis of no difference between the groups.

(c) Calculate a 95% confidence interval for the risk ratio, comparing tofacitinib to placebo.

8.52 Treatment for Covid-19, II. Using the data in Problem 8.51:

(a) Construct a 2×2 table summarizing the data, with the treatment variable in the rows and outcome in the columns.

(b) Calculate the expected cell counts under the null hypothesis of no treatment effect. Are the conditions for the χ^2 test met?

(c) Verify that the χ^2 statistic has value 4.77.

8.53 Fisher's exact test, I. Suppose the partial data in the following table summarize the results of a small randomized trial with 11 participants, in which 6 are assigned to control and 5 to treatment. Of those in the treatment group, 3 respond to treatment.

	Response	No Response	Total
Treatment	4		5
Control			6
Total	5	6	11

(a) Show that the 4 in the upper left cell determines the counts in the rest of the table.

(b) What is the relative risk for a response, comparing treatment to control?

(c) What are the tables that are as or more extreme whose results favor treatment?

(d) Calculate the Fisher's exact test one-sided p-value for a test of the null hypothesis of no treatment effect on response.

8.54 Fisher's exact test, II. Suppose the partial data in the following table summarize the results of a small randomized trial with 11 participants, in which 6 are assigned to control and 5 to treatment. Of those in the treatment group, 3 respond to treatment.

	Response	No Response	Total
Treatment	3		5
Control			6
Total	5	6	11

(a) Show that the 3 in the upper left cell determines the counts in the rest of the table.

(b) What is the relative risk for a response, comparing treatment to control?

(c) What are the tables that are as or more extreme whose results favor treatment?

(d) Calculate the Fisher's exact test one-sided p-value for a test of the null hypothesis of no treatment effect on response.

[49] Patrícia O. Guimarães et al. "Tofacitinib in Patients Hospitalized with Covid-19 Pneumonia". In: *New England Journal of Medicine* 385.5 (2021). PMID: 34133856, pp. 406–415. DOI: 10.1056/NEJMoa2101643. eprint: https://doi.org/10.1056/NEJMoa2101643. URL: https://doi.org/10.1056/NEJMoa2101643.

8.39 (a) H_0: The distribution of the format of the book used by the students follows the professor's predictions. H_A: The distribution of the format of the book used by the students does not follow the professor's predictions. (b) $E_{hard\ copy} = 126 \times 0.60 = 75.6$. $E_{print} = 126 \times 0.25 = 31.5$. $E_{online} = 126 \times 0.15 = 18.9$. (c) Independence: The sample is not random. However, if the professor has reason to believe that the proportions are stable from one term to the next and students are not affecting each other's study habits, independence is probably reasonable. Sample size: All expected counts are at least 5. (d) $\chi^2 = 2.32$, $df = 2$, p-value = 0.313. (e) Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.

8.41 (a)

	CVD	No CVD
Age Onset $\leq$ 50 Years	15	25
Age Onset $>$ 50 Years	5	55

(b) The odds of CVD for patients older than 50 years when diagnosed with diabetes is 5/55 = 0.09. The odds of CVD for the patients younger than 50 years at diabetes onset is 15/25 = 0.60. The relative odds (or odds ratio, OR) is 0.09/0.60 = 0.15.

(c) The odds of CVD for someone with late onset diabetes is less than 1/5 that of people with earlier onset diabetes. This can be explained by the fact that people with diabetes tend to build up plaque in their arteries; with early onset diabetes, plaque has longer time to accumulate, eventually causing CVD.

(d) $H_0 : OR = 1$.

(e) The chi-square test can be used to test H_0 as long as the conditions for the test have been met. The observations are likely independent; knowing one person's age of diabetes onset and CVD status is unlikely to provide information about another person's age of diabetes onset and CVD status. Under H_0, the expected cell count for the lower left cell is (60)(20)/100 = 12, which is bigger than 5; all other expected cell counts will be larger.

(f) Since the study is not a randomized experiment, it cannot demonstrate causality. It may be the case, for example, that CVD presence causes earlier onset of diabetes. The study only demonstrates an association between cardiovascular disease and diabetes.

8.43 (a) No. This is an example of outcome dependent sampling. Subjects were first identified according to presence or absence of the CNS disorder, then queried about use of the drug. It is only possible to estimate the probability that someone had used the drug, given they either did or did not have a CNS disorder.

(b) The appropriate measure of association is the odds ratio.

(c) The easiest way of calculating the OR for the table is the cross-product of the diagonal elements of the table: [(10)(4000)]/[(2000)(7)] = 2.86. Using the definition, it can be calculated as:

$$\hat{OR} = \frac{\frac{\hat{P}(\text{CNS}|\text{ Usage})}{1-\hat{P}(\text{CNS}|\text{ Usage})}}{\frac{\hat{P}(\text{CNS}|\text{ No Usage})}{1-\hat{P}(\text{CNS}|\text{ No Usage})}} = \frac{ad}{bc} = \frac{(10)(4000)}{(2000)(7)} = 2.86$$

(d) The odds ratio has the interpretation of the relative odds of presence of a CNS disorder, comparing people who have used the weight loss drug to those who have not. People who have used the weight loss drug have odds of CNS that are almost three times as large as those for people who have not used the drug.

(e) Fisher's exact test is better than the chi-square test. The independence assumption is met, but the expected cell count corresponding the presence of a CNS disorder and the use of the drug is 5.68, so not all the expected cell counts are less than 10.

8.45 (a) The p-value is 0.92; there is insufficient evidence to reject the null hypothesis of no association. These data are plausible with the null hypothesis that green tea consumption is independent of esophageal carcinoma. (b) Since the study uses outcome-dependent sampling, the odds ratio should be used as a measure of association rather than relative risk. The odds ratio of esophageal carcinoma, comparing green tea drinkers to non-drinkers, is 1.08; the odds of carcinoma for those who regularly drink green tea are 8% larger than the odds for those who never drink green tea.

8.47 (a) The prevalence difference is 0.15 – 0.10 = 0.05 and the prevalence ratio is 0.15/0.10 = 1.50. The absolute prevalence of disease in one group is 0.05 higher than in the other group. For instance, in a population

of 100,000 one would expect 10,000 cases in the first group 15,000 in the second group, and increase of 5,000 cases. If the prevalence is 1.50 times as large as that in the other group, the difference of 10,000 vs 15,000 cases in the hypothetical example represents 50% more cases. (b) The prevalence difference is 0.45 – 0.40 = 0.05 and the prevalence ratio is 0.45/0.40 = 1.125. In a population of 100,000, one would expect 40,000 cases in the lower prevalence group and 45,000 cases in the higher prevalence group, a difference of 5,000 cases. The difference of 5,000 cases is a 12.5% incease.

8.49 (a) The estimated odds that a male had a high salt diet are 7/53 = 0.132 and the estimated odds that a male had a low salt diet are 53/7 = 7.58. (b) Among the men where the recorded death was due to CVD, the odds of high salt diet are 5/30 = 0.167. The odds of low salt diet in the same group are 30/5 = 6. (c) The OR for a CVD related death, comparing a high to a low salt diet are (5/2)/(30/23) = 1.92. (d) The OR for a non CVD related death, comparing a high to a low salt diet are (2/5)/(23/30) = 0.522.

8.51 (a) Let $\hat{p}_1$ represent the observed proportion who experience the outcome of interest among those assigned to placebo and $\hat{p}_2$ the observed proportion who experience the outcome of interest among those assigned to tofacitinib; $\hat{p}_1 = 42/145 = 0.290$ and $\hat{p}_2 = 26/144 = 0.181$. The 95% CI is (−0.0527, −.2709). (b) Test $H_0 : p_1 = p_2$ against $H_A : p_1 \neq p_2$. Let $\alpha = 0.05$. With the z-test method, the z-statistic is 2.186. The two-sided p-value is $P|Z| \geq 2.186 = 0.0289$, which is smaller than 0.05. There is sufficient evidence to reject the null hypothesis; the evidence suggests that tofacitinib is an effective treatment compared to placebo. (c) The 95% CI for the risk ratio is (1.0422, 2.469). There is a larger risk of death or respiratory failure during the follow-up period for individuals on the placebo group than for individuals on tofactinib that could be as high as over twice the risk or as low as 1.04 times the risk.

8.53 (a) Given that the upper left cell has value 4 and that the margins are fixed, the other values in the table (going clockwise) are 1, 5, and 1. (b) The relative risk for response, comparing treatment to control, is (4/5)/(1/6) = 4.8. (c) There is only one table more extreme whose results favor treatment; the table in which all 5 individuals in the treatment group show a response. (d) The one-sided p-value consists of the probability of the observed table plus the probability of the table with a 5 in the upper left cell. Thus, the p-value is $\frac{\binom{5}{4}\binom{6}{1}}{\binom{11}{5}} + \frac{\binom{5}{5}\binom{6}{0}}{\binom{11}{5}} = 0.067$.