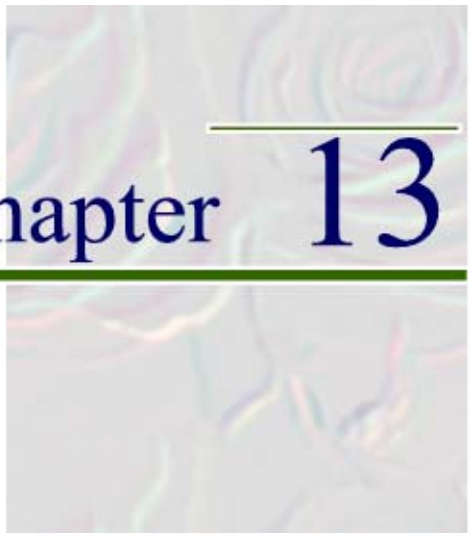


DESIGN AND ANALYSIS TECHNIQUES FOR EPIDEMIOLOGIC STUDIES

REVIEW OF KEY CONCEPTS

SECTION 13.1 Study Design

The following are the principal study designs used in epidemiologic research.

Prospective Study—A study in which a group of disease-free individuals at baseline are followed over time for the development of disease. The study population at baseline is sometimes referred to as a cohort. The goal of the study is usually to estimate the incidence rate of disease over time and to relate it to other risk factors defined at baseline (and also occasionally to risk factors which themselves change over time).

This type of study may be observational where subjects are simply observed over time or may be experimental (i.e., a clinical trial) , where different therapies are randomly assigned to patients at baseline and incidence rates are compared between treatment groups.

Retrospective Study—a study in which a group of patients with disease (the cases) and without disease (the controls) are defined in advance and differences in prior health habits (called exposures) are compared between groups. This type of study is called a case/control study.

Cross-Sectional Study—a study in which disease status is ascertained at one point in time and is related to either current or previous health habits.

Prospective studies are considered the gold standard of epidemiologic research because of the avoidance of *recall bias* (i.e., bias in recalling previous exposure history) on the part of subjects with disease, or *selection bias* (e.g., the selection of a control group in a case/control study that is not comparable to the case group or the selection of a case group that is not representative of the total population of cases).

SECTION 13.2 Measures of Effect for Epidemiological Data

In some instances, we are interested in comparing binomial proportions between two groups, but the groups are heterogeneous with respect to risk and it is desirable to subdivide each group into strata that are more homogeneous. Consider the following study.

Special wards exist for the care of babies of very low birthweight. One of the leading causes of death during the 1st year of life in this group is the respiratory distress syndrome. During the period 1985–1990, surfactant was introduced to widespread clinical use. Surfactant is administered intratracheally (the trachea is more commonly called the windpipe) to infants with the respiratory distress syndrome. The data on the following page were reported regarding in-hospital mortality for the period before and after surfactant use

on the wards in 14 hospitals [1]. Since mortality is related to birthweight, it was important to account for possible birthweight differences between the two time periods.

One approach is to do separate analyses within each birthweight stratum. The authors used this approach and reported their results in terms of an *odds ratio* for each birthweight stratum.

Birthweight (g)	Before surfactant			After surfactant			OR
	n	No. died	(%)	*n*	No. died	(%)	(95% CI)
500–749	769	479	(62)	330	177	(54)	0.7 (0.5–0.9)
750–999	903	257	(28)	414	78	(19)	0.6 (0.4–0.8)
1000–1249	1027	142	(14)	415	40	(10)	0.7 (0.5–1.0)
1250–1500	1223	82	(7)	548	40	(7)	1.1 (0.7–1.6)

What is an odds ratio? The odds in favor of a disease $= p/(1-p)$, where $p =$ probability of disease. The *odds ratio* (OR) = odds in favor of disease for an exposed group/odds in favor of disease for an unexposed group. For example, for the stratum with birthweight 500–749 g, the odds ratio relating in-hospital mortality to surfactant use is obtained from

$$\text{Odds of in-hospital mortality after surfactant use} = \frac{177}{330-177} = 1.157$$

$$\text{Odds of in-hospital mortality before surfactant use} = \frac{479}{769-479} = 1.652$$

$$\text{Estimated OR relating in-hospital mortality to surfactant use} = \hat{\text{OR}} = \frac{1.157}{1.652} = 0.700$$

Thus, the odds in favor of in-hospital mortality was 30% lower after surfactant was introduced than before. An OR can also be written in the form $ad/(bc)$, where *a*, *b*, *c*, and *d* are the cells of the 2×2 table relating disease to exposure. This table is given below for the 500–749-g subgroup, where $a = 177$, $b = 153$, $c = 479$, and $d = 290$.

		In-hospital mortality +	In-hospital mortality -	
Surfactant use	+	177	153	330
	-	479	290	769
		656	443	1099

To obtain a CI for the underlying OR, we use the Woolf approach. We first find a CI for ln(OR) $= (c_1, c_2)$, where

$$c_1 = \ln\left(\hat{\text{OR}}\right) - z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$c_1 = \ln\left(\hat{\text{OR}}\right) + z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

A $100\% \times (1-\alpha)$ CI for OR is then given by $[\exp(c_1), \exp(c_2)]$.

In this example, a 95% CI for ln(OR) is (c_1, c_2), where

$$c_1 = \ln(0.700) - 1.96\sqrt{\frac{1}{177}+\frac{1}{153}+\frac{1}{479}+\frac{1}{290}} = -0.356 - 0.261 = -0.617$$

$$c_2 = -0.356 + 0.261 = -0.095$$

The 95% CI for OR $= [\exp(-0.617),\ \exp(-0.095)] = (0.54,\ 0.91)$.

There are other comparative measures of effect for epidemiologic data, namely, the *risk difference* and the *risk ratio* defined by

$$\text{Risk difference} = p_1 - p_2$$

$$\text{Risk ratio} = \frac{p_1}{p_2}$$

The odds ratio, risk ratio and risk difference can all be estimated from prospective studies; only the odds ratio can be estimated from case/control studies.

13.2.1 Stratified Data

Since individual strata are often small, it is usually worthwhile to obtain an overall estimate of the odds ratio over all strata (in this example, over all birthweight groups). This can be accomplished with the *Mantel-Haenszel* (*MH*) *odds-ratio estimator* given by

$$\text{OR}_{\text{MH}} = \frac{\sum_{i=1}^{k} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i c_i}{n_i}}$$

where a_i, b_i, c_i, d_i, and n_i are the individual cells and grand total for the ith stratum and there are a total of k strata. In this example, the MH odds-ratio estimator can be interpreted as an odds ratio controlling for the confounding effects of birthweight. Note that this is different from simply pooling the data over all strata, where the birthweight distribution might be different before and after the use of surfactant.

In this case, the four 2×2 tables relating in-hospital mortality to surfactant use in individual birthweight strata are given as follows:

		Birthweight											
		500–749			750–999			1000–1249			1250–1500		
		Mortality			Mortality			Mortality			Mortality		
		+	–		+	–		+	–		+	–	
Surfactant	+	177	153	330	78	336	414	40	375	415	40	508	548
	–	479	290	769	257	646	903	142	885	1027	82	1141	1223
		656	443	1099	335	982	1317	182	1260	1442	122	1649	1771

The MH odds-ratio estimator is given by

$$\text{OR}_{\text{MH}} = \frac{\frac{177(290)}{1099} + \ldots + \frac{40(1141)}{1771}}{\frac{479(153)}{1099} + \ldots + \frac{82(508)}{1771}} = \frac{135.29}{192.70} = 0.702$$

Thus there is an overall 30% reduction in the odds of in-hospital mortality after surfactant was introduced. It is also possible to obtain a 95% CI for the overall OR—see Equation 13.16, in Chapter 13 of the text.

A commonly used approach to assessing whether surfactant use has an effect on in-hospital mortality, while controlling for birthweight is to perform the *Mantel-Haenszel test*. This is a test of the hypothesis $H_0: \text{OR} = 1$ versus $H_1: \text{OR} \neq 1$, where OR represents the underlying odds ratio relating in-hospital mortality to surfactant use within specific birthweight strata. To implement this test, use the test statistic

$$X^2_{\text{MH}} = \frac{(|O - E| - .5)^2}{V} \sim \chi^2_1 \text{ under } H_0$$

where

$O = \sum_{i=1}^{k} O_i$ = total observed number of in-hospital deaths after surfactant was introduced over all birthweight strata

$E = \sum_{i=1}^{k} E_i$ = total expected number of in-hospital deaths after surfactant was introduced over all birthweight strata

V = variance of $O = \sum_{i=1}^{k} V_i$,

where

$$E_i = \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$

$$V_i = \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)}$$

The p-value $= Pr(\chi_1^2 > X_{\text{MH}}^2)$

In this example,

$$O = 177 + \ldots + 40 = 335$$

$$E = \frac{330(656)}{1099} + \ldots + \frac{548(122)}{1771} = 392.42$$

$$V = \frac{330(769)(656)(443)}{1099^2(1098)} + \ldots + \frac{548(1223)(122)(1649)}{1771^2(1770)} = 166.39$$

$$X_{\text{MH}}^2 = \frac{(|335 - 392.42| - .5)^2}{166.39} = \frac{3239.39}{166.39} = 19.47 \sim \chi_1^2 \text{ under } H_0$$

Since $\chi_{1,\,.999}^2 = 10.83 < X_{\text{MH}}^2$, it follows that $p < .001$. Thus, there was a significant reduction of in-hospital mortality after surfactant was introduced, even after controlling for possible birthweight differences between the two periods.

The Mantel-Haenszel test can also be extended to study the relationship between a dichotomous disease variable such as the prevalence of bronchitis in the first year of life and an ordered categorical variable such as the number of cigarettes smoked per day by the mother, where one wants to stratify by other potential confounding variables (such as the presence of other siblings in the household). A separate estimate of slope relating the prevalence of bronchitis to the number of cigarettes smoked can be obtained in households where there are 1+ siblings and in households where there are no siblings. A weighted average of these slope estimates can then be obtained and assessed for statistical significance. This technique is referred to as the *Mantel-Extension test*—see Equation 13.19, in Chapter 13 of the text.

SECTION 13.3 Multiple Logistic Regression

In Section 13.2, we described a study of the effect of surfactant on in-hospital mortality among babies with very low birthweights. Since birthweight itself is a powerful predictor of in-hospital mortality, the data were stratified by birthweight and the Mantel-Haenszel test was used to compare in-hospital mortality before and after the use of surfactant, while controlling for birthweight. The authors were also concerned with other potential risk factors for in-hospital mortality, including gender and race. How can we look at

the effect of surfactant while controlling for several potential confounding variables? The authors used *multiple logistic regression* to accomplish this task.

13.3.1 Multiple Logistic-Regression Model

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \ldots + \beta_4 x_4$$

where

p = probability of in-hospital mortality
$x_1 = 1$ if male sex, 0 otherwise
$x_2 = 1$ if white race, 0 otherwise
x_3 = birthweight (g)
$x_4 = 1$ after surfactant became available
$= 0$ before surfactant became available

13.3.2 Interpretation of Regression Parameters

For dichotomous predictor variables, such as x_4 = surfactant availability, the regression parameter has the following interpretation: $\exp(\beta_4)$ = odds ratio relating in-hospital mortality to surfactant availability among infants that are the same for all other risk factors (the same sex, race, and birthweight). For continuous predictor variables, such as x_3 = birthweight, the regression parameter has the following interpretation: $\exp(\beta_3)$ = odds ratio relating in hospital mortality to a 1-g difference in birthweight (i.e., the odds in favor of in-hospital mortality for the heavier infant versus the lighter infant), among babies that are the same for all other risk factors (sex, race, and surfactant availability). Because 1 g is a small increment, the authors chose a different increment for the purpose of reporting results (viz., 100 g). In this case, the odds ratio relating in-hospital mortality to a 100-g difference in birthweight= $\exp(100\beta_3)$. Thus, for both dichotomous and continuous predictor variables, the regression coefficients play the same role as do partial-regression coefficients in multiple linear regression.

13.3.3 Hypothesis Testing

We wish to test the hypothesis $H_0: \beta_j = 0$, all other $\beta_l \neq 0$, versus H_1: all $\beta_j \neq 0$. This is a test of the specific effect of the *j*th independent variable after controlling for the effect of all other variables in the model. We use the test statistic

$$z = \frac{\hat{\beta}_j}{se\left(\hat{\beta}_j\right)} \sim N(0,1) \text{ under } H_0$$

The p-value $= 2 \times \Phi(z)$ if $z < 0$, or $= 2 \times [1 - \Phi(z)]$, if $z \geq 0$. The standard error of $\hat{\beta}_j$ is difficult to compute and is always obtained by computer. The results of the study were as follows:

Variable	Odds ratio	95% CI	*p*–value
Male sex	1.6	(1.4–1.8)	< .001
White race	1.0	(0.9–1.2)	> .05
Surfactant availability	0.7	(0.6–0.8)	< .001
Birthweight	0.6[a]	(0.6–0.7)	< .001

[a] For a 100-g increase in birthweight.

The results indicate that the odds in favor of in-hospital mortality were 30% lower after surfactant became available than before, after controlling for the effects of sex, race, and birthweight ($p < .001$). This is similar to the results in Section 13.2 using the Mantel-Haenszel test, where only birthweight was controlled for. Also, the odds in favor of in-hospital mortality are 60% higher for boys than girls ($p < .001$). Finally, the odds in favor of in-hospital mortality are 40% lower for the heavier infant than the lighter infant of the same race, sex, and time period, where the heavier infant is 100 g heavier than the lighter infant. No significant effect of race was seen.

Similar goodness-of-fit procedures can be performed with multiple logistic regression as with linear regression, including residual analyses that can help identify outlying and influential observations.

SECTION 13.4 Meta Analysis

In all previous work in this text, we have only considered methods of design and analysis for a single study. In some cases, there isn't enough power to test a hypothesis in a single study. In other cases, several different studies have been performed to test a specific hypothesis which seem to provide contradictory results. In both situations, it is desirable to combine results over more than one study. The technique for doing this is called *meta-analysis*.

Example: For many years, there have been studies comparing blood pressure levels between black and white children. These studies have often showed small differences despite the fact that large differences exist between black and white adult blood pressure levels. To improve precision and differentiate between early childhood vs adolescent bp levels, a meta analysis was performed combining data on black vs white bp differences in 9 large epidemiologic studies [2]. Because few children have high enough blood pressure to be called hypertensive by adult standards, a child was considered to have high blood pressure if he/she was above the 90th percentile for systolic blood pressure for his/her 1-year age-sex group after correcting for height. We present data from 3 of the largest of these studies for adolescent (age 13–17) girls as follows:

	Black			White							$\ln(OR_i)$	
Study	+	total	prev*	+	total	prev*	OR_i	s_i^2	Δ^2	w_i^*	(y_i)	$w_i^* y_i$
Bogalusa	121	911	.133	105	1342	.078	1.80	0.0199	0.0326	19.06	0.590	11.25
Dallas	416	5142	.081	293	4447	.066	1.25	0.0063		25.73	0.222	5.70
S. Carolina	91	536	.170	86	567	.152	1.14	0.0269		16.79	0.134	2.25
Total										61.58		19.21

* prevalence of high blood pressure

To combine results over more than 1 study, we calculate a weighted average of the study-specific ln odds ratios (y_i) given by

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i^* y_i}{\sum_{i=1}^{k} w_i^*}$$

The weight w_i^* ; is inversely proportional to the study-specific variance of the log-odds ratio given by $s_i^2 + \Delta^2$. The within-study variation is given by s_i^2 , where

$$s_i^2 = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

and a_i, b_i, c_i, and d_i, are the cell counts in the 2×2 table in the ith study. The between-study variation is estimated by Δ^2 (see Equation 13.39, Chapter 13, text).

If the between-study variation is large relative to the within-study variation, then different studies get about equal weight. If the within-study variation predominates, then larger studies will tend to get more weight. In the above example, we have

$$\hat{\mu} = \frac{19.21}{61.58} = 0.312$$

The standard error of $\hat{\mu} = se(\hat{\mu}) = \left(1 \Big/ \sum_{i=1}^{3} w_i^*\right)^{1/2} = (1/61.58)^{1/2} = 0.127$. The test statistic is

$$z = \frac{\hat{\mu}}{se(\hat{\mu})} = 2.447 \sim N(0,1) \text{ under } H_0.$$

The p-value $= 2 \times \Phi(z)$ if $z < 0$ or $2 \times [1 - \Phi(z)]$ if $z \geq 0$. In this case, the p-value $= 2 \times [1 - \Phi(z)]$ $= 2 \times [1 - \Phi(2.447)] = .014$. Thus, there is a significant difference between the proportion of black and white adolescent girls with high blood pressure. The point estimate of the odds ratio

$$\exp(\hat{\mu}) = \exp(0.312) = 1.37.$$

A 95% CI for the OR $= \exp[\hat{\mu} \pm 1.96 se(\hat{\mu})] = \exp[0.312 \pm 1.96(0.127)] = (1.06,\ 1.75)$. Thus, black adolescent girls have a significantly higher rate of high systolic blood pressure than white adolescent girls.

SECTION 13.5 The Cross-Over Design

A *cross-over design* is a type of clinical trial where the same patient receives both types of treatment in different periods. Consider the following example. Erythropoietic protoporphyria (EPP) is a rare dermatologic disease where patients are very sensitive to sunlight. A study was performed comparing cysteine versus placebo for the treatment of this condition. Patients were randomized to either group 1 or group 2. Patients in group 1 received cysteine for an 8-week period, went off drug for 1 week (during a "washout period") and then received placebo for the next 8 weeks. Patients in group 2 received placebo for 8 weeks, went off drug for 1 week and then received cysteine for the next 8 weeks. There were 8 patients in group 1 and 7 patients in group 2. Patients were assigned to groups using a random-number table. Patients were phototested at the end of each active drug period (i.e., subjected to a controlled intensity of artificial light) and the time to develop erythema (excessive redness) was noted. How can we assess the relative effectiveness of cysteine versus placebo with this type of study design?

To assess the significance of the results, for each patient, we compute

$x = x_C - x_P$ = time to develop erythema while on cysteine – time to develop erythema while on placebo

$\bar{d}_1$ = mean of x for all subjects in group 1

$\bar{d}_2$ = mean of x for all subjects in group 2

$s_{d_1}^2$ = variance of x for all subjects in group 1

$s_{d_2}^2$ = variance of x for all subjects in group 2

We use the test statistic

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{4}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \text{ under } H_0$$

where

$$\bar{d} = .5\left(\bar{d}_1 + \bar{d}_2\right)$$
$$s_d^2 = \frac{(n_1-1)s_{d_1}^2 + (n_2-1)s_{d_2}^2}{n_1+n_2-2}$$

The p-value $= 2 \times Pr\left(t_{n_1+n_2-2} > t\right)$ if $t > 0$, or $= 2 \times Pr\left(t_{n_1+n_2-2} < t\right)$ if $t \le 0$.

The following results were obtained from the EPP study:

	Group 1	Group 2
$\bar{d}_i$ [a]	6.75	21.71
s_{d_i} [a]	7.78	9.86
n_i	8	7

[a] Minutes.

Thus,

$$\bar{d} = \frac{6.75 + 21.71}{2} = 14.23$$
$$s_d^2 = \frac{7(7.78)^2 + 6(9.86)^2}{13} = 77.46$$
$$t = \frac{14.23}{\sqrt{\frac{77.46}{4}\left(\frac{1}{8}+\frac{1}{7}\right)}} = \frac{14.23}{2.28} = 6.25 \sim t_{13} \text{ under } H_0$$
$$p\text{-value} = 2 \times Pr(t_{13} > 6.25) < .001$$

Thus, cysteine significantly increases the time until symptoms develop relative to placebo.

One issue that often arises in cross-over studies is the possible presence of a carry-over effect. A *carry-over effect* is defined to be present when the effect of the treatment received in the first period is present also in the second treatment period. For example, in the preceding study, for the patients in group 1 the effect of cysteine might extend into the second period, making it hard to identify treatment effects for subjects in group 1. In this case, if a carry-over effect is present, the average time before symptoms occur over the two periods for a patient in group 1 will be higher than for patients in group 2. To test for carry-over effect, we compute

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \text{ under } H_0$$

where

y = average time to develop erythema for a patient over the two time periods
$\bar{y}_1$ = mean of y over all patients in group 1
$\bar{y}_2$ = mean of y over all patients in group 2
s_1^2 = variance of y over all patients in group 1
s_2^2 = variance of y over all patients in group 2

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The p-value $= 2 \times Pr\left(t_{n_1+n_2-2} > t\right)$ if $t > 0$, or $= 2 \times Pr\left(t_{n_1+n_2-2} < t\right)$ if $t \le 0$.

In this case, the following results were obtained from the EPP study:

	Group 1	Group 2
$\bar{y}_i$	16.63	22.00
s_i	11.26	7.39
n_i	8	7

The pooled-variance estimate is

$$s^2 = \frac{7(11.26)^2 + 6(7.39)^2}{13} = 93.48$$

The test statistic is

$$t = \frac{16.63 - 22.00}{\sqrt{93.48\left(\frac{1}{8} + \frac{1}{7}\right)}} = \frac{-5.37}{5.00} = -1.07 \sim t_{13} \text{ under } H_0$$

The p-value $= 2 \times Pr\left(t_{13} < -1.07\right) = .30$. Thus, there is no significant carry-over effect in this study. An important assumption of the overall test for treatment efficacy is that no significant carry-over effect is present as is the case with the EPP data.

SECTION 13.6 Clustered Binary Data

In the two sample test for binomial proportions, we have assumed that each sample point is independent. In some applications, we have replicate observations for each subject . In general, the replicate observations cannot be considered statistically independent.

Example: Retinitis pigmentosa is a genetic eye disease which often leads to reduced vision and sometimes to legal blindness. There are several genetic types including dominant (DOM), recessive (AR), sex-linked (XL) and isolate (ISO). Patients with different genetic types seem to have a different severity. A cross-sectional study was performed [3] to more clearly differentiate the clinical course of patients with different genetic types. One way to define reduced vision is by a visual acuity of 20/50 or worse (the threshold for not being able to drive in many states). An issue is that a patient may have reduced vision in both eyes, one eye, or neither eye and the outcomes for individual eyes are not independent. For example, in the above study the following results were obtained from DOM and XL patients:

	Number of eyes with reduced vision			Total patients	Total affected eyes	Total eyes	Proportion of affected eyes
	2	1	0				
DOM	7	6	15	28	20	56	.357
XL	14	2	3	19	30	38	.789

If the eye is used as the unit of analysis, then we need to account for the dependence between outcomes for individual eyes of the same person before comparing the proportion of eyes with reduced vision between genetic types.

In general, we will use the two sample test for binomial proportions, but will adjust the variance for dependence between replicate observations for the same subject. If there are m replicates per subject, then the test statistic is as follows:

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{[1+(m-1)\hat{\rho}]}{2} \times \left(\frac{1}{M_1} + \frac{1}{M_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{M_1} + \frac{1}{M_2}\right)[1+(m-1)\hat{\rho}]}} \sim N(0,\ 1) \text{ under } H_0$$

where $\hat{\rho}$ is the estimated intraclass correlation between replicate responses for the same subject

M_i = total number of replicates in the i th group

$\hat{p}_i$ = proportion of successes among replicates in the i th group

$$\hat{p} = \frac{\sum_{i=1}^{2} M_i \hat{p}_i}{\sum_{i=1}^{2} M_i}$$

The p-value $= 2 \times [1 - \Phi(z)]$. We can use the same formula for computing intraclass correlation as for normally distributed data (see Equation 12.36, Chapter 12, text) except here the response is binary, which is coded as 1 if a success (i.e. affected) and 0 if a failure (i.e., not affected).

For the RP data, we use MINITAB to obtain the variance components by first creating separate columns for outcomes for the right (OD) and left (OS) eyes and then running a 1-way ANOVA. The 21 patients with two affected eyes are coded as $\text{OD} = 1$ and $\text{OS} = 1$. There are 8 patients with 1 affected eye of whom 4 are arbitrarily coded as $\text{OD} = 1$ and $\text{OS} = 0$ and 4 are coded as OD = 0 and OS = 1. The 18 patients with no affected eyes are coded as $\text{OD} = 0$ and $\text{OS} = 0$. The results of the ANOVA are as follows:

Between MS $= 0.4218$, Within MS $= 0.0851$. Thus, since $m = 2$

$$\hat{\sigma}^2 = 0.0851$$

$$\hat{\sigma}_A^2 = \frac{0.4218 - 0.0851}{2} = 0.1684$$

$$\hat{\rho}_{\text{ICC}} = \frac{0.1684}{0.1684 + 0.0851} = .664$$

The overall proportion of eyes with reduced vision $= \hat{p} = 50/94 = .532$ and

$$\hat{q} = 1 - .532 = .468.$$

Therefore,

$$z = \frac{|.357 - .789| - \frac{1.664}{2}\left(\frac{1}{56} + \frac{1}{38}\right)}{\sqrt{.532(.468)\left(\frac{1}{56} + \frac{1}{38}\right)(1.664)}} = \frac{.432 - .037}{\sqrt{.2490(.0442)(1.664)}} = \frac{.396}{.135} = 2.924 \sim N(0,1) \text{ under } H_0$$

Thus, the p-value $= 2 \times Pr[1 - \Phi(2.924)] = .003$. Therefore, the percent of patients with reduced vision is significantly greater in the sex-linked group than the dominant group.

The test statistic given above can be generalized to accommodate an unequal number of replicates (see Equation 13.51, Chapter 13, text, for details).

SECTION 13.7 Special Topics in Epidemiological Data Analysis

13.7.1 Establishing Clinical Equivalence

In many previous examples considered in this study guide, the goal has been to establish superiority of an active treatment vs. a control treatment (usually placebo). However, in some instances, an effective treatment already exists and it is unethical to withhold active treatment from patients. Instead, if a new treatment is introduced, the goal is to establish approximate clinical equivalence between the new treatment and the standard treatment. For example, effective treatments exist for the prevention of mortality in patients with an evolving myocardial infarction (e.g. streptokinase, TPA). If a new drug is introduced, it may be reasonable to establish that its ability to reduce mortality is roughly similar to standard therapy. Such a study is called an *equivalence study*. The analytic approach generally taken is to establish a threshold δ, such that if $p_1 =$ survival probability for the standard therapy and $p_2 =$ survival probability for the new treatment, then with high confidence e.g., 95% confidence, we can say that $p_1 - p_2$ is less than δ. Thus, in words, if δ is for example 5%, then we will be 95% confident that the new therapy has a survival probability that is no more than 5% lower than standard therapy. To achieve this goal, we construct a 1-sided $100\% \times (1 - \alpha)$ CI for $p_1 - p_2$ and consider the treatments equivalent if the upper bound of this CI is $< \delta$. Sample sizes needed for equivalence studies are generally larger than for superiority studies, particularly, if the threshold δ is close to zero. (see Section 13.9, text for more details about equivalence studies).

13.7.2 Accounting for Measurement Error

In many epidemiologic studies, exposure variables are measured with error. For example, one hypothesis of interest is that end-stage renal disease (ESRD) among diabetic patients is more common among diabetics who consume a high protein diet. To test this hypothesis, it might be reasonable to fit a logistic regression of the form:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

where $p =$ probability of developing ESRD and $X =$ dietary protein as measured by a dietary questionnaire such as a 24-hour recall. The problem is that X is measured with error both because a 24-hour recall is subject to recall bias and because intake is estimated only for 1 day. Even if we averaged many 24 hour recalls we still could not eliminate the recall bias. Instead, what is really desired is the "true" protein intake as measured by average intake over many days which we denote by X^*. Thus, if we knew X^* then we could fit the logistic regression

$$\ln\left(\frac{p}{1-p}\right) = \alpha^* + \beta^* X^*$$

The goal of measurement error correction is to estimate β^*.To accomplish this, we need some additional data that defines the amount of measurement error. There are two distinct situations. In one case, we can actually estimate X^* (or at least obtain an unbiased estimate of it). For example, a diet record (DR) is an instrument where a subject records what they eat on a real-time basis. If DR intake is averaged over many weeks, this might provide a reasonable estimate of X^*. We can then run a *validation study* where both

dietary protein from a single 24-hour recall (X) and the gold standard = average protein intake from many weeks of DR (X^*) are obtained on the same patients. By combining information from the validation study where we can relate X to X^* and the main study where we can relate p to X, we can indirectly estimate β^* (see Section 13.12.2 for details).

In another instance, we might be interested in relating urinary protein (X) to ESRD by logistic regression. There is no gold standard for urinary protein. We might consider the average of an infinite number of urinary protein values (X^*) as a gold standard, but we cannot directly measure it. However, if we conduct a reproducibility study, and obtain several replicate measures of urinary protein, then we can use the estimated intraclass correlation among the replicates and the results from the main study to indirectly estimate β^* (see Section 13.12.3 for details).

PROBLEMS

Cardiovascular Disease
In a 1985 study of the effectiveness of streptokinase in the treatment of patients who have been hospitalized after myocardial infarction, 9 of 199 males receiving streptokinase and 13 of 97 males in the control group died within 12 months [4].

13.1 Compute a 95% CI for the difference in 12-month mortality rates between the streptokinase and control groups.

13.2 Compute the odds ratio in favor of death within 12 months for streptokinase therapy versus control therapy.

13.3 Provide a 95% CI for the true odds ratio corresponding to your answer to Problem 13.2.

In the streptokinase study in Problem 13.1, 2 of 15 females receiving streptokinase and 4 of 19 females in the control group died within 12 months.

13.4 After stratifying the data by sex, perform a significance test for association between treatment group and 12-month mortality.

13.5 After stratifying the data by sex, estimate the odds ratio in favor of 12-month mortality in the streptokinase group versus the control group.

13.6 Provide a 95% CI for the true odds ratio corresponding to your answer to Problem 13.5.

Obstetrics
Suppose there are 500 pairs of pregnant women who participate in a prematurity study and are paired in such a way that the body weights of the 2 women in a pair are within 5 lb of each other. One of the 2 women is given a placebo and the other drug A to see if drug A has an effect in preventing prematurity. Suppose that in 30 pairs of women, *both* women in a pair have a premature child; in 420 pairs of women, *both* women have a normal child; in 35 pairs of women, the woman taking drug A has a normal child and the woman taking the placebo has a premature child; in 15 pairs of women, the woman taking drug A has a premature child and the woman taking the placebo has a normal child.

13.7 Estimate the odds ratio for premature birth for women assigned to drug A vs. woman assigned to placebo.

13.8 Provide a 95% CI for the estimate in Problem 13.7.

Cardiovascular Disease
A hypothesis has been suggested that a principal benefit of physical activity is to prevent sudden death from heart attack. The following study was designed to test this hypothesis: 100 men who died from a first heart attack and 100 men who survived a first heart attack in the age group 50–59 were identified and their wives were each given a detailed questionnaire concerning their husbands' physical activity in the year preceding their heart attacks. The men were then classified as active or inactive. Suppose that 30 of the 100 who survived and 10 of the 100 who died were physically active. If we wish to test the hypothesis, then

13.9 Compute the odds ratio in favor of survival after an MI for physically active versus physically inactive men.

13.10 Compute a 95% CI for the odds ratio referred to in Problem 13.9.

Obstetrics
An issue of current interest is the effect of delayed childbearing on pregnancy outcome. In a recent paper a population of first deliveries was assessed for low-birthweight deliveries (< 2500 g) according to the woman's age and prior pregnancy history [5]. The data in Table 13.1 were presented.

Table 13.1 Relationship of age and pregnancy history to low-birthweight deliveries

Age	History[a]	n	Percentage low birthweight
≥ 30	No	225	3.56
≥ 30	Yes	88	6.82
< 30	No	906	3.31
< 30	Yes	153	1.31

[a] History = yes if a woman had a prior history of spontaneous abortion or infertility; = no otherwise

Source: Reprinted with permission of the *American Journal of Epidemiology*, **125**(l),101–109,1987.

13.11 Estimate the odds ratio relating age to the prevalence of low-birthweight deliveries, while controlling for the possible confounding effect of pregnancy history. Provide a 95% CI for the odds ratio.

13.12 Estimate the odds ratio relating pregnancy history to the prevalence of low-birthweight deliveries, while controlling for the possible confounding effect of age. Provide a 95% CI for the odds ratio.

13.13 Is there evidence of effect modification between age and pregnancy history regarding their effect on low-birthweight deliveries?

Cancer

A prospective study was conducted among 55–59-year-old postmenopausal women looking at the association between the use of postmenopausal hormones and the incidence of breast cancer. Women were categorized as never users, current users or past users by mail questionnaire in 1976 and the incidence of breast cancer was assessed in the 10-year interval following the questionnaire (1976–1986) by follow-up mail questionnaires. The results are given in Table 13.2.

Table 13.2 Association between postmenopausal hormone use and the incidence of breast cancer

Status	Number of cases	Number of women
Never users	129	7175
Current users	51	2495
Past users	46	2526

13.14 What test can be used to test the overall hypothesis that there is (or is not) a difference in incidence of breast cancer among the three groups?

13.15 Implement the test in Problem 13.14 and report a *p*-value.

13.16 Estimate the odds ratio relating breast-cancer risk for current users versus never users. Provide a 95% CI for the odds ratio.

13.17 Estimate the odds ratio relating breast-cancer risk for past users versus never users. Provide a 95% CI for the odds ratio.

13.18 Compute risk ratios and 95% CI's corresponding to the odds ratios in Problems 13.16 and 13.17. How do the risk ratios and odds ratios compare? How do the CI's compare?

Obstetrics

Interest has increased in possible relationships between the health habits of the mother during pregnancy and adverse delivery outcomes. An issue that is receiving attention is the effect of marijuana usage on the occurrence of congenital malformations. One potential problem is that many other maternal factors are associated with adverse pregnancy outcomes, including age, race, socioeconomic status (SES), smoking, and so forth. Since these factors may also be related to marijuana usage, they need to be simultaneously controlled for in the analysis. Therefore, data from the Delivery Interview Program (DIP) were used, and a multiple logistic regression was performed relating the occurrence of major congenital malformations to several maternal risk factors [6]. The results are given in Table 13.3.

13.19 Perform a test to assess the significance of each of the risk factors after controlling for the other risk factors.

13.20 Compute an odds ratio and an associated 95% CI relating the presence or absence of each independent variable to the risk of malformations.

Table 13.3 Multiple logistic-regression analysis relating the occurrence of major congenital malformations to several risk factors, using 12,424 women in the DIP study

Variable[a]	Regression coefficient	Standard error
Marijuana usage (any frequency)	0.307	0.173
Any previous miscarriage	0.239	0.141
White race	0.191	0.161
Alcohol use in pregnancy	0.174	0.145
Age 35+	0.178	0.186
Any previous stillbirth	0.049	0.328
On welfare	0.030	0.182
Smoking 3+ cigarettes per day at delivery	–0.174	0.144
Any previous induced abortion	–0.198	0.168
Parity > l[b]	–0.301	0.113

[a] All variables are coded as 1 if yes and 0 if no.
[b] Wornen with at least one previous pregnancy.

Source: Reprinted with permission of the *American Journal of Public Health*, **73**(10), 1161–1164, 1983.

13.21 How do you interpret the effect of marijuana usage on congenital malformations, given your findings in Problems 13.19 and 13.20?

Dermatology
In the dermatology study referred to in Section 13.5 (Study Guide) [7], in addition to phototesting, the patients were instructed to go out in the sun and determine the number of minutes until erythema developed. The data in Table 13.4 were obtained (not all subjects completed this part of the study).

13.22 Assess if there is a difference between the mean time of exposure to sunlight until the development of erythema for subjects while on cysteine versus placebo.

13.23 Assess if there are any carry-over effects. What does a carry-over effect mean in this context?

Table 13.4 Mean exposure times (minutes) to symptom development: daylight exposures

Group 1			Group 2		
Patient number	Period 1	Period 2	Patient number	Period 1	Period 2
2	27.70	27.80	1	28.86	45.00
3	187.35	156.86	4	38.63	42.41
7	55.88	20.00	6	12.17	38.79
11	19.61	22.51	8	33.31	48.75
12	155.13	164.55	13	84.48	104.40
14	40.23	45.00			

Hypertension
Refer to the adolescent blood pressure study described in Section 13.4. Data were also obtained for adolescent boys and are presented as follows:

	Black			White		
Study	+*	Total	%	+	Total	%
Bogalusa	73	923	.079	95	1510	.063
Dallas	424	5101	.083	450	4524	.099
S. Carolina	66	492	.134	88	552	.159

* High blood pressure = SBP above the 90th percentile after correcting for age and height

13.24 Obtain an odds ratio relating the prevalence of high blood pressure for black vs. white boys that accounts for both between and within study variation. Test the hypothesis that black boys have a significantly different prevalence of high blood pressure than white boys.

13.25 Provide a 95% CI for the odds ratio estimate obtained in Problem 13.24.

13.26 Compare the results obtained for adolescent boys with similar results obtained in Section 13.4 for adolescent girls.

Opthalmology
Refer to the RP study described in Section 13.6. Data were also obtained for the number of eyes with reduced vision in patients with recessive (AR) and isolate (ISO) genetic types. The results for all genetic types are given as follows:

Genetic Type	Number of eyes with reduced vision*			Total Patients	Total Affected Eyes	Total Eyes	% of Affected Eyes
	2	1	0				
DOM	7	6	15	28	20	56	0.357
AR	9	5	7	21	23	42	0.548
XL	14	2	3	19	30	38	0.789
ISO	57	24	67	148	138	296	0.466

* Reduced vision = $20/50$ or worse

13.27 Compare the percent of eyes with reduced vision in recessive (AR) vs sex-linked (XL) patients. Provide a p-value accounting for the correlation between two eyes of the same patient.

13.28 Answer the question in Problem 13.27 comparing the percent of eyes with reduced vision in isolate (ISO) (i.e., no other known affected relatives) patients vs. sex-linked (XL) patients.

13.29 Summarize the results obtained for Problems 13.27, 13.28, and Section 13.6 regarding the comparative prevalence of reduced vision among different genetic type of RP.

SOLUTIONS

13.1 A 95% CI for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$
$$= 0.0452 - 0.1340 \pm 1.96\sqrt{\frac{.0452(.9548)}{199} + \frac{.1340(.8660)}{97}}$$
$$= -.0888 \pm 1.96(0.0376) = -.0888 \pm .0737$$
$$= (-.162,\ -.015)$$

13.2 The odds ratio is given by OR

$$\frac{\frac{9}{190}}{\frac{13}{84}} = \frac{0.0474}{0.1548} = 0.306.$$

13.3 A 95% CI for ln(OR) is given by $(c_1,\ c_2)$, where

$$c_1 = \ln(0.306) - 1.96\sqrt{\frac{1}{9} + \frac{1}{190} + \frac{1}{13} + \frac{1}{84}}$$
$$= -1.184 - 0.888 = -2.072$$
$$c_2 = -1.184 + 0.888 = -0.296$$

The corresponding 95% CI for OR is

$$(e^{-2.072},\ e^{-0.296}) = (0.13,\ 0.74).$$

13.4 We have the following observed and expected values for the 2×2 tables for males and females respectively (with expected values in parentheses):

Males-12-month mortality status

	Dead	Alive	
Streptokinase	9 (14.79)	190 (184.21)	199
Control	13 (7.21)	84 (89.79)	97
	22	274	296

Females-12-month mortality status

	Dead	Alive	
Streptokinase	2 (2.65)	13 (12.35)	15
Control	4 (3.35)	15 (15.65)	19
	6	28	34

We compute the test statistic

$$X^2_{\text{MH}} = \frac{(|O - E| - .5)^2}{V}$$

where

$$O = 9 + 2 = 11$$
$$E = 14.79 + 2.65 = 17.44$$
$$V = \frac{199 \times 97 \times 22 \times 274}{296^2 \times 295} + \frac{15 \times 19 \times 6 \times 28}{34^2 \times 33}$$
$$= 4.502 + 1.255 = 5.757$$

Thus,

$$X^2_{\text{MH}} = \frac{(|11 - 17.44| - .5)^2}{5.757} = \frac{5.94^2}{5.757} = 6.12 \sim \chi^2_1 \text{ under } H_0$$

Since $\chi^2_{1,\ .975} = 5.02$, $\chi^2_{1,\ .99} = 6.63$ and

$$5.02 < 6.12 < 6.63,$$

it follows that $1 - .99 < p < 1 - .975$ or

$$.01 < p < .025.$$

Thus, there is a significant association between 12-month mortality status and treatment group, after adjusting for gender.

13.5 We have

$$\text{OR}_{\text{MH}} = \frac{\sum_{i=1}^{k} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i c_i}{n_i}} = \frac{\frac{9\times84}{296} + \frac{2\times15}{34}}{\frac{13\times190}{296} + \frac{4\times13}{34}} = \frac{2.554 + 0.882}{8.345 + 1.529} = \frac{3.436}{9.874} = 0.35.$$

13.6 We use the approach in Equation 13.16 (text, Chapter 13) as follows:

$\text{Var}\left(\ln \hat{\text{OR}}\right) = A + B + C$ where

$$A = \frac{\frac{93(9)(84)}{296^2} + \frac{17(2)(15)}{34^2}}{2\left[\frac{9(84)}{296} + \frac{2(15)}{34}\right]^2} = \frac{1.2436}{23.6178} = 0.0527$$

$$B = \frac{\left[\frac{93(13)(190)}{296^2} + \frac{17(4)(13)}{34^2} + \frac{203(9)(84)}{296^2} + \frac{17(2)(15)}{34^2}\right]}{2\left[\frac{9(84)}{296} + \frac{2(15)}{34}\right]\left[\frac{13(190)}{296} + \frac{4(13)}{34}\right]}$$
$$= \frac{5.5793}{67.8622} = 0.0822$$

$$C = \frac{\frac{203(13)(190)}{296^2} + \frac{17(4)(13)}{34^2}}{2\left[\frac{13(190)}{296} + \frac{4(13)}{34}\right]^2} = \frac{6.4875}{194.9920} = 0.0333$$

Thus, $\text{Var}\left(\ln \hat{\text{OR}}\right) = 0.1681$ and a 95% CI for $\ln\left(\hat{\text{OR}}\right)$ is

$$\ln\left(\hat{\text{OR}}\right) \pm 1.96 se\left[\ln\left(\hat{\text{OR}}\right)\right]$$
$$= \ln(0.35) \pm 1.96\sqrt{0.1681} = -1.0555 \pm 0.8037$$
$$= (-1.8592,\ -0.2518)$$

The corresponding 95% CI for OR is

$$(e^{-1.8592},\ e^{-0.2518}) = (0.16,\ 0.78).$$

13.7 The estimated odds ratio $= 35/15 = 2.33$.

13.8 We refer to Equation 13.18, Chapter 13, text. A 95% CI for $\ln(\text{OR})$ is

$$\ln(2.33) \pm 1.96\sqrt{\frac{1}{50(\frac{35}{50})(\frac{15}{50})}} = 0.847 \pm 0.605$$
$$= (0.242,\ 1.452)$$

The corresponding 95% CI for OR is

$$(e^{0.242},\ e^{1.452}) = (1.27,\ 4.27).$$

13.9 The observed 2×2 table is shown as follows:

Observed table relating sudden death from a first heart attack and previous physical activity

		Physical activity: Active	Physical activity: Inactive	
Mortality status	Survived	30	70	100
	Died	10	90	100
		40	160	200

The odds ratio in favor of survival after an MI for physically active versus physically inactive men is given by

$$\hat{\text{OR}} = \frac{30 \times 90}{70 \times 10} = \frac{2700}{700} = 3.86$$

13.10 We use the Woolf method. We first obtain a 95% CI for ln(OR) given by

$$\ln\hat{\text{OR}} \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$
$$= \ln(3.86) \pm 1.96\sqrt{\frac{1}{30}+\frac{1}{70}+\frac{1}{10}+\frac{1}{90}}$$
$$= 1.350 \pm 0.781 = (0.569,\ 2.131)$$

The corresponding 95% CI for OR is

$$(e^{0.569},\ e^{2.131}) = (1.77,\ 8.42).$$

13.11 Form the following 2×2 table to assess age effects among women with a negative history:

Women with a negative history

		Low birthweight: Yes	Low birthweight: No	
Age	≥ 30	8	217	225
	< 30	30	876	906
		38	1093	1131

Form the following 2×2 contingency table among women with a positive history:

Women with a positive history

		Low birthweight: Yes	Low birthweight: No	
Age	≥ 30	6	82	88
	< 30	2	151	153
		8	233	241

We use the Mantel-Haenszel odds-ratio estimator as follows:

$$\hat{\text{OR}}_{\text{MH}} = \frac{\sum_{i=1}^{2}\frac{a_i d_i}{n_i}}{\sum_{i=1}^{2}\frac{b_i c_i}{n_i}} = \frac{\frac{8(876)}{1131}+\frac{6(151)}{241}}{\frac{30(217)}{1131}+\frac{2(82)}{241}} = \frac{9.956}{6.436} = 1.55$$

To obtain a 95% CI we use the Robins method given in Equation 13.16 (text, Chapter 13). We have the following summary statistics:

History	i	P_i	Q_i	R_i	S_i
Negative	1	0.7816	0.2184	6.1963	5.7560
Positive	2	0.6515	0.3485	3.7593	0.6805
Total				9.9556	6.4365

History	i	$P_i R_i$	$P_i S_i$	$Q_i R_i$	$Q_i S_i$
Negative	1	4.8431	4.4989	1.3532	1.2571
Positive	2	2.4490	0.4433	1.3103	0.2372
Total		7.2921	4.9422	2.6635	1.4942

Therefore,

$$\text{Var}\left(\ln\hat{\text{OR}}_{\text{MH}}\right) = \frac{\sum_{i=1}^{2}P_i R_i}{2\left(\sum_{i=1}^{2}R_i\right)^2} + \frac{\sum_{i=1}^{2}(P_i S_i + Q_i R_i)}{2\left(\sum_{i=1}^{2}R_i\right)\left(\sum_{i=1}^{2}S_i\right)} + \frac{\sum_{i=1}^{2}Q_i S_i}{2\left(\sum_{i=1}^{2}S_i\right)^2}$$
$$= \frac{7.2921}{2(9.9556)^2} + \frac{4.9422 + 2.6635}{2(9.9556)(6.4365)} + \frac{1.4942}{2(6.4365)^2}$$
$$= 0.0368 + 0.0593 + 0.0180 = 0.1142$$

Therefore, a 95% CI for $\ln(\hat{\text{OR}})$ is

$$\ln(1.55) \pm 1.96\sqrt{0.1142} = 0.4362 \pm 0.6623$$
$$= (-0.2261,\ 1.0984).$$

A 95% CI for OR $= (e^{-0.2261},\ e^{1.0984}) = (0.80,\ 3.00)$.

13.12 We construct 2×2 tables relating low birthweight to pregnancy history within specific age groups as follows:

Age < 30

		Low birthweight +	Low birthweight –	
Pregnancy	+	2	151	153
history	–	30	876	906
		32	1027	1059

Age ≥ 30

		Low birthweight +	Low birthweight -	
Pregnancy	+	6	82	88
history	–	8	217	225
		14	299	313

We use the Mantel-Haenszel odds-ratio estimator as follows:

$$\hat{\text{OR}}_{\text{MH}} = \frac{\frac{2(876)}{1059} + \frac{6(217)}{313}}{\frac{30(151)}{1059} + \frac{8(82)}{313}} = \frac{5.8141}{6.3735} = 0.91$$

To obtain a 95% CI, we use the Robins method as follows:

Age	i	P_i	Q_i	R_i	S_i
< 30	1	0.8291	0.1709	1.6544	4.2776
≥ 30	2	0.7125	0.2875	4.1597	2.0958
Total				5.8141	6.3735

Age	i	P_iR_i	P_iS_i	Q_iR_i	Q_iS_i
< 30	1	1.3716	3.5465	0.2828	0.7311
≥ 30	2	2.9637	1.4932	1.1961	0.6026
Total		4.3353	5.0397	1.4789	1.3338

Therefore,

$$\text{Var}\left(\ln \hat{\text{OR}}_{\text{MH}}\right) = \frac{4.3353}{2(5.8141)^2} + \frac{5.0397 + 1.4789}{2(5.8141)(6.3735)} + \frac{1.3338}{2(6.3735)^2}$$
$$= 0.0641 + 0.0880 + 0.0164 = 0.1685$$

A 95% CI for $\ln(\hat{\text{OR}})$ is

$$\ln(0.91) \pm 1.96\sqrt{0.1685} = -0.0919 \pm 0.8045$$
$$= (-0.8964,\ 0.7127)$$

The corresponding 95% CI for OR

$$(e^{-0.8964},\ e^{0.7127}) = (0.41,\ 2.04).$$

13.13 To test for effect modification, we use the Woolf test for homogeneity of odds ratios over different strata given in Equation 13.17 (text, Chapter 13). We have the test statistic

$$X^2_{\text{HOM}} = \sum_{i=1}^{2} w_i \left[\ln(\hat{\text{OR}}_i)\right]^2 - \frac{\left[\sum_{i=1}^{2} w_i \ln(\hat{\text{OR}}_i)\right]^2}{\sum_{i=1}^{2} w_i} \sim \chi^2_1 \text{ under } H_0$$

where we test H_0: $\text{OR}_1 = \text{OR}_2$ versus H_1: $\text{OR}_1 \neq \text{OR}_2$ where $\text{OR}_1 =$ odds ratio relating low birthweight to age for women with a negative pregnancy history and $\text{OR}_2 =$ the corresponding odds ratio for women with a positive pregnancy history and $w_i = (1/a_i + 1/b_i + 1/c_i + 1/d_i)^{-1}$. We have the following summary statistics:

History	i	w_i	$\hat{\text{OR}}_i$	$\ln(\hat{\text{OR}}_i)$	$w_i \ln(\hat{\text{OR}}_i)$	$w_i[\ln(\hat{\text{OR}}_i)]^2$
Negative	1	6.0945	1.0765	0.0737	0.4492	0.0331
Positive	2	1.4588	5.5244	1.7092	2.4934	4.2616
Total		7.5533			2.9426	4.2947

Thus, we have

$$X^2_{\text{HOM}} = 4.2947 - \frac{2.9426^2}{7.5533} = 3.15 \sim \chi^2_1 \text{ under } H_0$$

Since $3.15 < 3.84 = \chi^2_{1,\,.95}$ it follows that $p > .05$ and there is no significant heterogeneity among the two odds ratios. Thus, there is no significant effect modification.

13.14 We can set up the data in the form of a 3×2 contingency table as follows:

PMH status	Breast Cancer +	Breast Cancer –	Total	Percentage of breast cancer
Never use	129	7046	7175	(1.8)
Current use	51	2444	2495	(2.0)
Past use	46	2480	2526	(1.8)
	226	11,970	12,196	

The smallest expected value

$$226 \times \frac{2495}{12{,}196} = 46.2 \geq 5.$$

Thus, we can use the chi-square test for $r \times c$ tables.

13.15 We compute the expected table corresponding to the observed table in Problem 13.14. We have

$$E_{11} = \frac{226 \times 7175}{12{,}196} = 133.0$$
$$E_{12} = 7175 - 133.0 = 7042.0$$
$$E_{21} = \frac{226 \times 2495}{12{,}196} = 46.2$$
$$E_{22} = 2495 - 46.2 = 2448.8$$
$$E_{31} = \frac{226 \times 2526}{12{,}196} = 46.8$$
$$E_{32} = 2526 - 46.8 = 2479.2$$

The chi-square statistic is given by

$$X^2 = \frac{(129 - 133.0)^2}{133} + \frac{(7046 - 7042.0)^2}{7042.0} + \frac{(51 - 46.2)^2}{46.2} + \frac{(2444 - 2448.8)^2}{2448.8} + \frac{(46 - 46.8)^2}{46.8} + \frac{(2480 - 2479.2)^2}{2479.2}$$
$$= 0.118 + 0.002 + 0.491 + 0.009 + +0.014 + 0.000$$
$$= 0.635 \sim \chi^2_2 \text{ under } H_0$$

Since $\chi^2_{2,\,.50} = 1.39 > 0.635$, and

$$\chi^2_{2,\,.25} = 0.58 < 0.635,$$

it follows that $1 - .5 < p < 1 - .25$ or $.5 < p < .75$, and we accept H_0 that the incidence of breast cancer is the same in the three groups.

13.16 The estimated odds ratio

$$\frac{\frac{51}{2444}}{\frac{129}{7046}} = 1.14.$$

A 95% CI for $\ln(\text{OR})$ is given

$$\ln(1.14) \pm 1.96 \times \sqrt{\frac{1}{51} + \frac{1}{2444} + \frac{1}{129} + \frac{1}{7046}}$$
$$= 0.1308 \pm 0.3274 = (-0.1966,\ 0.4583)$$

The corresponding 95% CI for OR is

$$[\exp(-0.1966),\ \exp(0.4583)] = (0.82,\ 1.58).$$

13.17 The estimated odds ratio

$$\frac{\frac{46}{2480}}{\frac{129}{7046}} = 1.01.$$

A 95% CI for ln(OR) is given by

$$\ln(1.01) \pm 1.96 \times \sqrt{\frac{1}{46} + \frac{1}{2480} + \frac{1}{129} + \frac{1}{7046}}$$
$$= 0.0130 \pm 0.3397 = (-0.3267,\ 0.3527).$$

The corresponding 95% CI for OR

$$[\exp(-0.3267),\ \exp(0.3527)] = (0.72,\ 1.42).$$

13.18 For current users versus never users,

$$\text{RR} = \frac{\frac{51}{2495}}{\frac{129}{7175}} = 1.14.$$

To obtain confidence limits, we use the formula in Equation 13.6 (text, Chapter 13). A 95% CI for $\ln \text{RR}$ is

$$\ln(1.14) \pm 1.96\sqrt{\frac{2444}{51(2495)} + \frac{7046}{129(7175)}}$$
$$0.1283 \pm 0.3210$$
$$= (-0.1927,\ 0.4493)$$

The corresponding 95% CI for RR

$$[\exp(-0.1927),\ \exp(0.4493)] = (0.82,\ 1.57).$$

For past users versus never users,

$$\text{RR} = \frac{\frac{46}{2526}}{\frac{129}{7175}} = 1.01.$$

A 95% CI for ln RR is

$$\ln(1.01) \pm 1.96\sqrt{\frac{2480}{46(2526)} + \frac{7046}{129(7175)}}$$
$$0.0128 \pm 0.3335 = (-0.3207,\ 0.3463).$$

The corresponding 95% CI for RR

$$[\exp(-.3207),\ \exp(.3463)] = (0.73,\ 1.41).$$

Both the point estimates and the 95% Cl's for OR and RR are very similar for both current and past users respectively, versus never users. Larger differences would be expected if either OR or RR were further away from 1.0. Since this is a prospective study, the risk ratio is the more appropriate measure.

13.19 We wish to test $H_0: \beta_i = 0$ versus $H_1: \beta_i \neq 0$ for each of the risk factors. We will use the test statistic $z_i = \hat{\beta}_i \big/ se(\hat{\beta}_i) \sim N(0, 1)$ under H_0 to test for the effect of each variable after simultaneously controlling for the effects of all other variables. We have the following table:

Variable	$\hat{\beta}_i \big/ se(\hat{\beta}_i)$	Two-tail p-value
Marijuana usage	1.775	.076
Any previous miscarriage	1.695	.090
White race	1.186	.24
Alcohol use in pregnancy	1.200	.23
Age 35+	0.957	.34
Any previous stillbirth	0.149	.88
On welfare	0.165	.87
Smoking 3+ cigarettes per day at delivery	–1.208	.23
Any previous induced abortion	–1.179	.24
Parity > 1	–2.664	.008

The only risk factor significantly affecting the risk of congenital malformations is parity, with children born from women with no previous pregnancies at higher risk than children born from women who have had at least 1 previous pregnancy. In addition, marijuana usage at the current pregnancy and previous miscarriages showed a trend toward increasing the risk of congenital malformations, but were not statistically significant after controlling for the other risk factors.

13.20 The estimated odds ratios are given by $\exp(\hat{\beta}_i)$ and the associated 95% Cl's by $\exp[\hat{\beta}_i \pm 1.96 se(\hat{\beta}_i)]$. These are given in Table 13.5.

Table 13.5 Odds ratios and 95% CI of risk factors for congenital malformations from the DIP study

Variable	Odds ratio	95% CI
Marijuana usage	$\exp(0.307) = 1.36$	$\exp[0.307 \pm 1.96(0.173)] = (0.97,\ 1.91)$
Any previous miscarriage	$\exp(0.239) = 1.27$	$\exp[0.239 \pm 1.96(0.141)] = (0.96,\ 1.67)$
White race	$\exp(0.191) = 1.21$	$\exp[0.191 \pm 1.96(0.161)] = (0.88,\ 1.66)$
Alcohol use in pregnancy	$\exp(0.174) = 1.19$	$\exp[0.174 \pm 1.96(0.145)] = (0.90,\ 1.58)$
Age 35+	$\exp(0.178) = 1.19$	$\exp[0.178 \pm 1.96(0.186)] = (0.83,\ 1.72)$
Any previous stillbirth	$\exp(0.049) = 1.05$	$\exp[0.049 \pm 1.96(0.328)] = (0.55,\ 2.00)$
On welfare	$\exp(0.030) = 1.03$	$\exp[0.030 \pm 1.96(0.182)] = (0.72,\ 1.47)$
Smoking 3+ cigarettes per day at delivery	$\exp(-0.174) = 0.84$	$\exp[-0.174 \pm 1.96(0.144)] = (0.63,\ 1.11)$
Any previous induced abortion	$\exp(-0.198) = 0.82$	$\exp[-0.198 \pm 1.96(0.168)] = (0.59,\ 1.14)$
Parity > 1	$\exp(-0.301) = 0.74$	$\exp[-0.301 \pm 1.96(0.113)] = (0.59,\ 0.92)$

13.21 Marijuana usage in the current pregnancy shows a trend toward statistical significance with odds ratio $= 1.36$ after controlling for all other risk factors. This means that the odds in favor of a congenital malformation are 36% higher for offspring of users of marijuana compared with offspring of nonusers of marijuana, all other factors held constant. However, we should be cautious in interpreting these findings, due to the borderline significance of the results.

13.22 The mean and sd of the difference in the time to develop erythema between the cysteine and placebo treatment periods by group are given as follows:

	Group 1	Group 2
d_i [a]	8.20	16.38
s_{d_i} [a]	19.67	8.32
n_i	6	5

[a] Minutes.

The overall mean is $\bar{d} = (8.20 + 16.38)/2 = 12.29$ min. We have the pooled-variance estimate

$$s_d^2 = \frac{5(19.67)^2 + 4(8.32)^2}{9} = \frac{2210.902}{9} = 245.66$$

The test statistic is

$$t = \frac{12.29}{\sqrt{\frac{245.66}{4}\left(\frac{1}{6} + \frac{1}{5}\right)}} = \frac{12.29}{4.745}$$
$$= 2.59 \sim t_9 \text{ under } H_0$$
$$\text{The } p\text{-value} = 2 \times Pr(t_9 > 2.59) = .029$$

Thus, there is a significant difference between the time to developing erythema after exposure to sunlight while on cysteine vs. placebo as measured by patient diary; with the time being longer while on cysteine.

13.23 We compute the average time to symptom development over the two periods for each patient and compute the mean and sd of the average time within each group as follows:

	Group 1	Group 2
y_i [a]	76.89	47.68
s_i [a]	69.53	26.88
n_i	6	5

[a] Minutes.

The pooled-variance estimate is

$$s^2 = \frac{5(69.53)^2 + 4(26.88)^2}{9} = 3006.73$$

The test statistic is

$$t = \frac{76.89 - 47.68}{\sqrt{3006.73(\frac{1}{6} + \frac{1}{5})}} = \frac{29.21}{33.20} = 0.88 \sim t_9 \text{ under } H_0.$$

The p-value $= 2 \times Pr(t_9 > 0.88) = .40$. Thus, there is no significant carry-over effect in these data.

A carry-over effect means that cysteine when taken in the 1st period still has an effect in the 2nd period. This would bias the estimate of treatment efficacy in Group 1.

13.24 We use the meta analysis random effects model (Equation 13.39, Chapter 13, text). We obtain the $\log_e$ odds ratio (y_i) and within-study variance (s_i^2) for each study as shown in the MINITAB output. To compute the between study variance, we compute

$$Q_w = \sum_{i=1}^{3} w_i (y_i - y_w)^2 = 6.489$$

as shown in the spreadsheet below. The between study variance is then given by

$$\hat{\Delta}^2 = \frac{Q_w - (k-1)}{\sum_{i=1}^{3} w_i - \frac{\sum_{i=1}^{3} w_i^2}{\sum_{i=1}^{3} w_i}} = 0.0396$$

The estimated $\log_e$ odds ratio is then

$$\hat{\mu} = \frac{\sum_{i=1}^{3} w_i^* y_i}{\sum_{i=1}^{3} w_i^*} = \frac{-3.53926}{51.7524} = -0.0683$$

where $w_i^* = \left(s_i^2 + \hat{\Delta}^2\right)^{-1}$. The corresponding odds ratio estimate $= \exp(-0.0683) = 0.93$. This odds ratio is not significantly different from 1 (z statistic $= -0.49$, p-value $= .62$).

ROW	study	a_i	b_i	c_i	d_i	OR_i	y_i	s2_i
1	bogalusa	73	850	95	1415	1.27920	0.246231	0.0261081
2	dallas	424	4677	450	4074	0.82074	–0.197546	0.0050400
3	s. carolina	66	426	88	464	0.81690	–0.202237	0.0310177

ROW	w_i	wi*yi^2	sumwiyi2	wi*yi	sumwiyi	sumwi	Q_w
1	38.302	2.32225	11.3838	9.4312	–36.2845	268.955	6.48867
2	198.413	7.74294		–39.1957			
3	32.240	1.31859		–6.5200			

ROW	sumwi2	delta_2	wi*	sumwi*	sumwi*yi	mu	OR
1	41874.3	0.0396306	15.2117	51.7524	–3.53926	–0.0683883	0.933898
2			22.3861				
3			14.1546				

ROW	se(mu)	z	cdf(z)	p-value	mu_1	mu_2
1	0.139006	–0.49198	0.311367	0.622734	–0.340841	0.204064

ROW	OR_1	OR_2
1	0.711172	1.22638

13.25 A 95% CI for ln OR is given by

$$\hat{\mu} \pm 1.96 \times se(\hat{\mu}) = (-0.341,\ 0.204).$$

The corresponding 95% CI for OR

$$[\exp(-0.341),\ \exp(0.204)] = (0.71,\ 1.23).$$

13.26 The OR for black vs white girls was significantly greater than 1 indicating that black girls are more likely to have high blood pressure than white girls. However, there is no significant difference between the percent of black vs. white boys with high blood pressure. In other analyses of this dataset it was shown that black adolescent girls are heavier than white adolescent girls while no difference in weight was observed between black and white adolescent boys. Since high weight is a strong risk factor for high blood pressure, this provides a possible explanation for the different results for boys and girls.

13.27 We refer to Equation 13.52, Chapter 13, text. We have:

$$z = \frac{|.548 - .789| - \left[\frac{1+\hat{\rho}}{2}\right]\left(\frac{1}{42} + \frac{1}{38}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{42} + \frac{1}{38}\right)(1+\hat{\rho})}}$$

In this case, $\hat{p} = 53/80 = .6625$, $\hat{q} = .3375$. We run a one-way ANOVA to determine the intraclass correlation. For this purpose, we create a data set with 23 subjects with $\text{OD} = 1$ and $\text{OS} = 1$, 7 subjects with $\text{OD} = 1$ and $\text{OS} = 0$ and 10 subjects with $\text{OD} = 0$ and $\text{OS} = 0$. In essence, we create 2 columns with the 40 ID numbers repeated twice in one column and the individual OD and OS values in another column. We have: Between $\text{MS} = 0.3689$, Within $\text{MS} = 0.0875$. Thus,

$$\hat{\sigma}_A^2 = \frac{0.3689 - 0.0875}{2} = 0.1407$$

$$\hat{\sigma}^2 = 0.0875$$

$$\hat{\rho} = \frac{0.1407}{0.1407 + 0.0875} = \frac{0.1407}{0.2282} = 0.617$$

Therefore,

$$z = \frac{.2419 - \left(\frac{1.617}{2}\right)(0.0501)}{\sqrt{.6625(.3375)(0.0501)(1.617)}} = \frac{.2013}{.1346}$$

$$= 1.496 \sim N(0,\ 1) \text{ under } H_0$$

$$p\text{-value} = 2 \times [1 - \Phi(1.496)] = .135.$$

Thus, there is no significant difference between the % of eyes with reduced vision between AR and XL RP patients.

13.28 We use the same approach as in Problem 13.27. We have:

$$z = \frac{|.466 - .789| - \left(\frac{1+\hat{\rho}}{2}\right)\left(\frac{1}{296} + \frac{1}{38}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{296} + \frac{1}{38}\right)(1+\hat{\rho})}}$$

In this case, $\hat{p} = 168/334 = .503$, $\hat{q} = .497$. We run a one-way ANOVA to determine the intraclass correlation between 2 eyes of the same person. For this purpose, we create a dataset with 71 subjects with 2 affected eyes, 26 subjects with 1 affected eye and 70 subjects with 0 affected eyes. The result is as follows: Between $\text{MS} = 0.4247$, Within $\text{MS} = 0.0778$,

$$\hat{\sigma}_A^2 = \frac{0.4247 - 0.0778}{2} = 0.1735$$
$$\hat{\sigma}^2 = 0.0778$$
$$\hat{\rho} = \frac{0.1735}{0.1735 + 0.0778} = 0.690$$

Therefore,

$$z = \frac{.3233 - \left(\frac{1.690}{2}\right)(0.0297)}{\sqrt{.503(.497)(0.0297)(1.690)}}$$
$$= \frac{.3233 - .0251}{.1120} = \frac{.2982}{.1120}$$
$$= 2.662 \sim N(0, 1) \text{ under } H_0$$
$$p\text{- value} = 2 \times [1 - \Phi(2.662)] = .008$$

Thus, isolate patients have a significantly smaller proportion of eyes with reduced vision than sex-linked patients.

13.29 We found that there were significant differences between the sex-linked group and both the dominant and isolate groups, respectively, with sex-linked patients having a greater prevalence of eyes with reduced vision. No significant differences were found between the AR and XL groups, although there was a trend in the direction of sex-linked patients having a greater prevalence of eyes with reduced vision. Although this was not a longitudinal study, these results are consistent with the clinical impression that sex-linked patients progress the fastest.

REFERENCES

[1] Schwartz, R. M., Luby, A. M., Scanlon, J. W., & Kellogg, R. J. (1991). Effect of surfactant on morbidity, mortality and resource use in newborn infants weighing 500 to 1500 g. *New England Journal of Medicine*, **330**(21), 1476–1480.

[2] Rosner, B., Prineas, R., Daniels, S. R. and Loggie, J. Black/white blood pressure differences in relationship to body size among U.S. children and adolescents. In press, *American Journal of Epidemiology*, 2000.

[3] Berson, E. L., Rosner, B., and Simonoff, E. (1980). An outpatient population of retinitis pigmentosa and their normal relatives: risk factors for genetic typing and detection derived from their ocular examinations. *American Journal of Ophthalmology*, **89**, 763–775.

[4] Kennedy, J. W., Ritchie, J. L., Davis, K. B., Stadius, M. L., Maynard, C., & Fritz, J. K. (1985). The western Washington randomized trial of intracoronary streptokinase in acute myocardial infarction: A 12-month follow-up report. *New England Journal of Medicine*, **312**(17), 1073–1078.

[5] Barkan, S. E., & Bracken, M. (1987). Delayed childbearing: No evidence for increased risk of low birth weight and preterm delivery. *American Journal of Epidemiology*, **125**(l), 101–109.

[6] Linn, S., Schoenbaum, S. C., Monson, R. R., Rosner, B., Stubblefield, R C., & Ryan, K. J. (1983). The association of marijuana use with outcome of pregnancy. *American Journal of Public Health*, **73**(10), 1161–1164.

[7] Mathews-Roth, M. M., Rosner, B., Benfell, K., & Roberts, J. E. (1994). A double-blind study of cysteine photoprotection in erythropoietic protoporphyria. *Photodermatology, Photoimmunology & Photo-medicine*; **10**, 244–248.