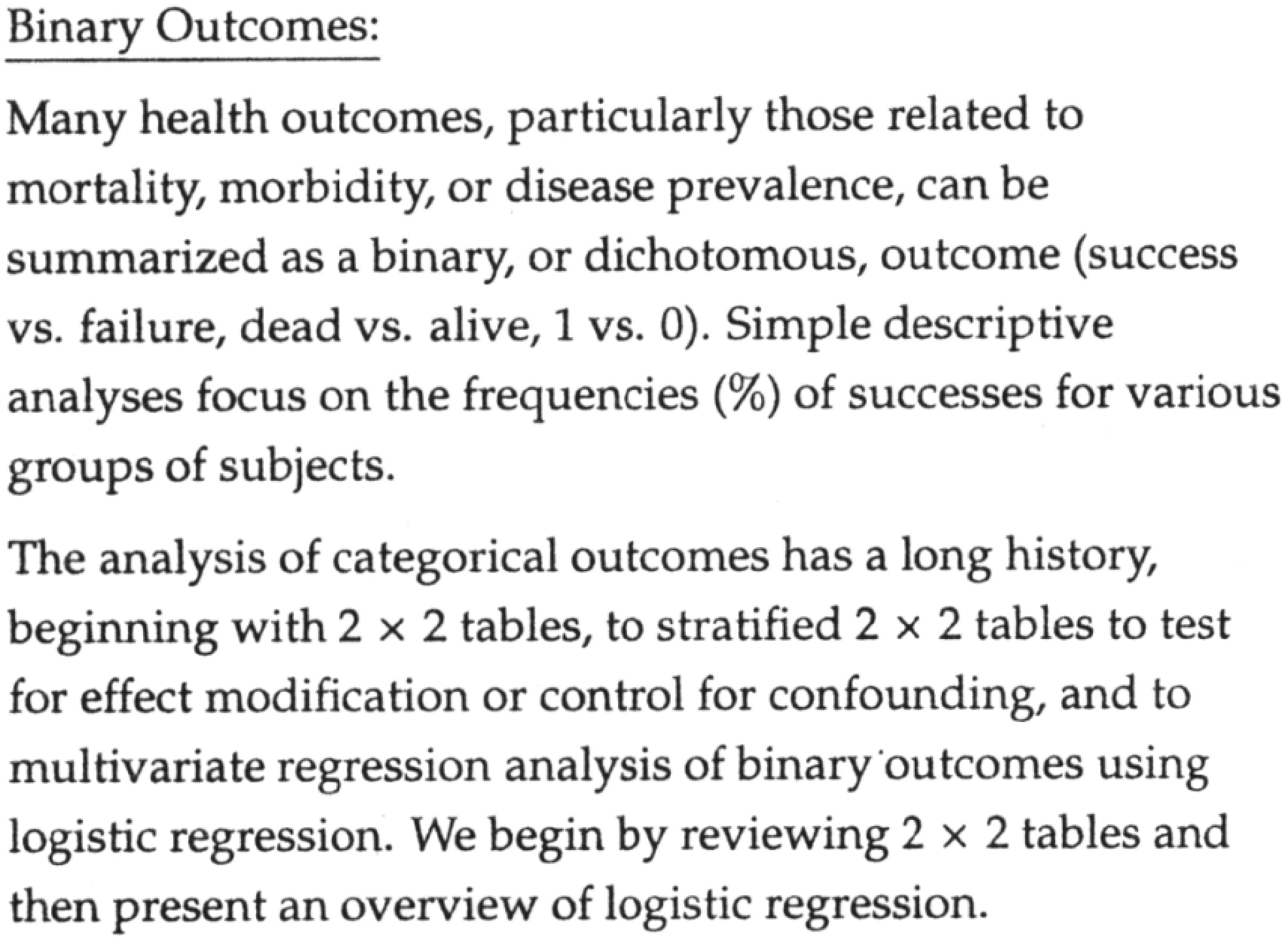


Binary Outcomes:

Many health outcomes, particularly those related to mortality, morbidity, or disease prevalence, can be summarized as a binary, or dichotomous, outcome (success vs. failure, dead vs. alive, 1 vs. 0). Simple descriptive analyses focus on the frequencies (%) of successes for various groups of subjects.

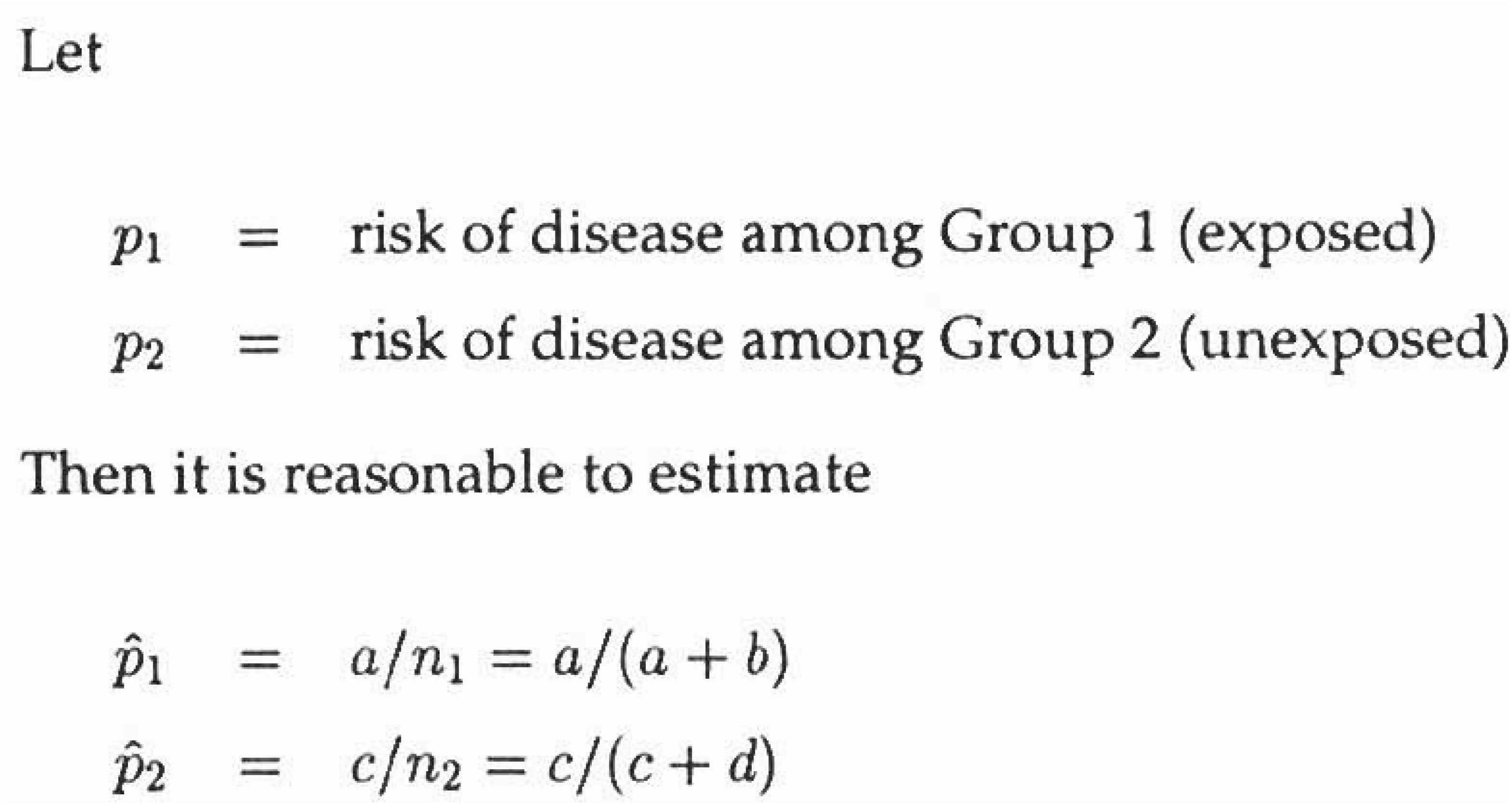
The analysis of categorical outcomes has a long history, beginning with 2 × 2 tables, to stratified 2 × 2 tables to test for effect modification or control for confounding, and to multivariate regression analysis of binary outcomes using logistic regression. We begin by reviewing 2 × 2 tables and then present an overview of logistic regression.

Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support

2 × 2 Tables:

Consider a sample of N individuals randomly selected from a population of interest. Cross-classify these individuals on the basis of their outcome status and their exposure of interest (treatment 1 vs. treatment 2, smoker vs. nonsmoker, comparison group vs. control, etc.):

		Disease		
		Yes	No	
Group	1	a	b	n_1
Group	2	c	d	n_2
		m_1	m_2	N



Let

$$p_1 \quad = \quad \text{risk of disease among Group 1 (exposed)}$$

$$p_2 \quad = \quad \text{risk of disease among Group 2 (unexposed)}$$

Then it is reasonable to estimate

$$\hat{p}_1 \quad = \quad a/n_1 = a/(a+b)$$

$$\hat{p}_2 \quad = \quad c/n_2 = c/(c+d)$$

Measures of Association: There are three commonly used methods for comparing risks between two groups:

- Risk difference (or attributable risk)

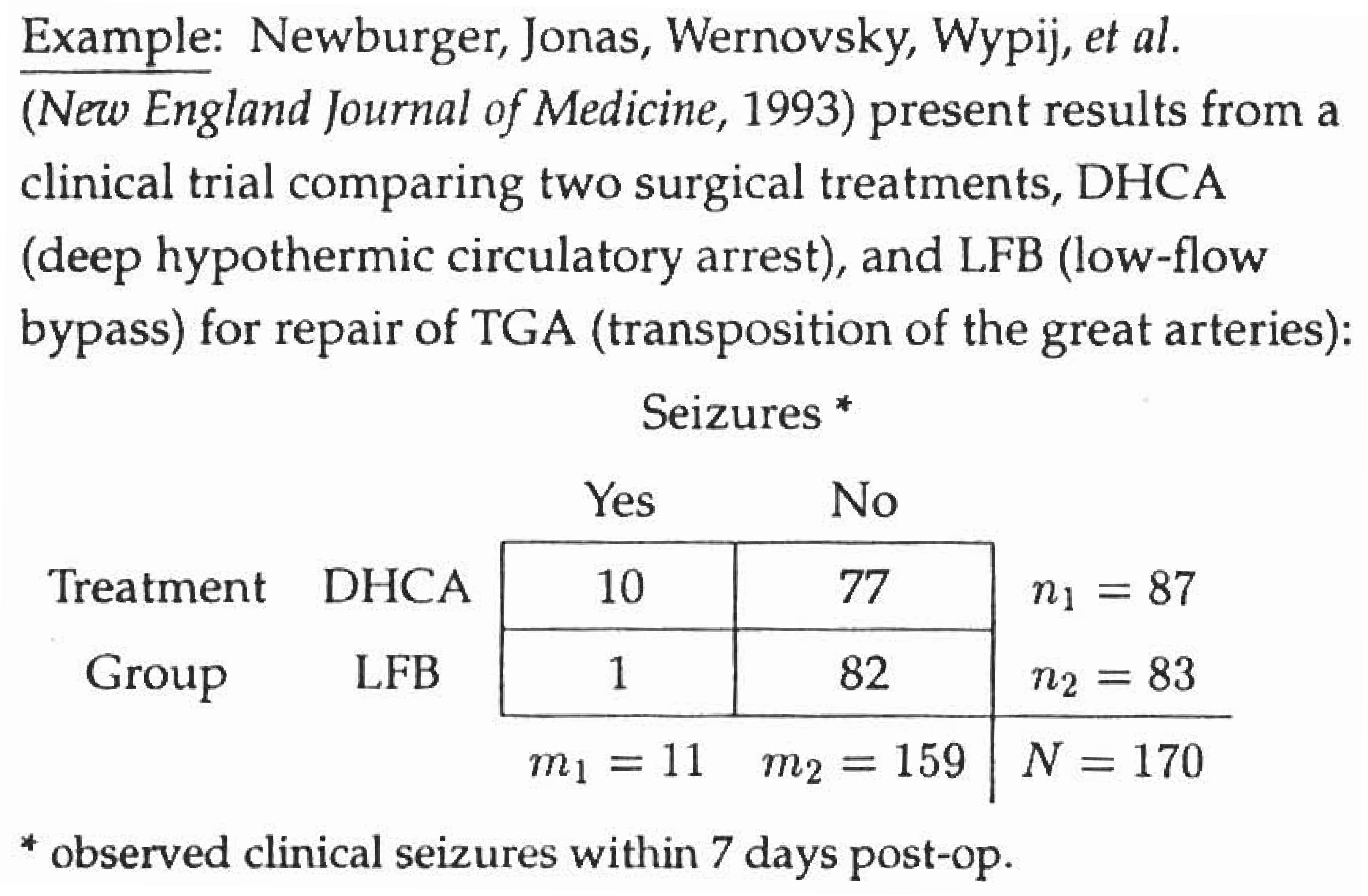
$$RD = p_1 - p_2$$

- Risk ratio (or relative risk)

$$RR = p_1/p_2$$

- Odds ratio (or relative odds)

$$\begin{aligned} OR &= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \\ &= \frac{\text{odds of disease in group exposed to risk factor}}{\text{odds of disease in group not exposed to risk factor}} \end{aligned}$$



Example: Newburger, Jonas, Wernovsky, Wypij, *et al.* (*New England Journal of Medicine*, 1993) present results from a clinical trial comparing two surgical treatments, DHCA (deep hypothermic circulatory arrest), and LFB (low-flow bypass) for repair of TGA (transposition of the great arteries):

		Seizures *		
		Yes	No	
Treatment	DHCA	10	77	$n_1 = 87$
Group	LFB	1	82	$n_2 = 83$
		$m_1 = 11$	$m_2 = 159$	$N = 170$

* observed clinical seizures within 7 days post-op.

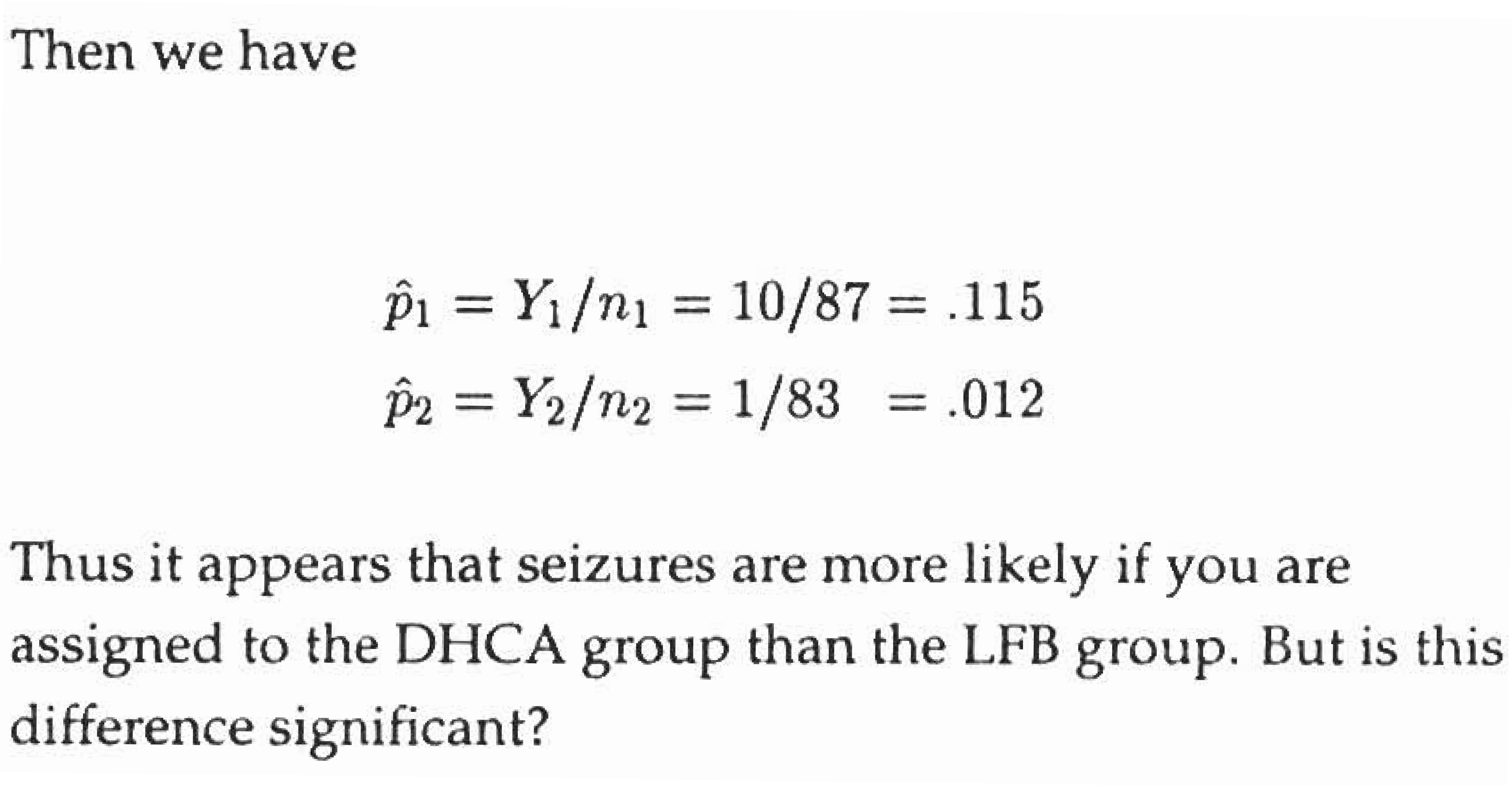
Specifying a little more notation:

		Seizures		
		Yes	No	
Treatment	DHCA	$Y_1 = 10$	77	$n_1 = 87$
Group	LFB	$Y_2 = 1$	82	$n_2 = 83$

$$\begin{aligned} Y_1 &\sim Bin(n_1, p_1) \\ Y_2 &\sim Bin(n_2, p_2) \end{aligned}$$

where

$$\begin{aligned} p_1 &= P(\text{seizures in DHCA group}) \\ p_2 &= P(\text{seizures in LFB group}) \end{aligned}$$



Then we have

$$\begin{aligned} \hat{p}_1 &= Y_1/n_1 = 10/87 = .115 \\ \hat{p}_2 &= Y_2/n_2 = 1/83 \quad = .012 \end{aligned}$$

Thus it appears that seizures are more likely if you are assigned to the DHCA group than the LFB group. But is this difference significant?

Let's estimate our three <u>measures of association:</u>

<u>Risk difference:</u>

$$\begin{aligned}\hat{RD} &= \hat{p}_1 - \hat{p}_2 \\ &= \frac{10}{87} - \frac{1}{83} = .103\end{aligned}$$

Thus the difference in the risk of seizures in the DHCA group compared to the LFB group is .103, or 10.3%.

Risk ratio:

$$\begin{aligned} \hat{RR} &= \hat{p}_1/\hat{p}_2 \\ &= (10/87)\,/\,(1/83) = 9.54 \end{aligned}$$

Thus the risk of seizures in the DHCA group is 9.54 times higher than the risk of seizures in the LFB group.

Odds ratio:

$$\hat{O}R = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}$$

$$\begin{aligned}
\text{O}\hat{\text{d}}\text{ds in DHCA group} &= \frac{\hat{p}_1}{1-\hat{p}_1} = \frac{10/87}{77/87} = \frac{10}{77} = .130 \\
\text{O}\hat{\text{d}}\text{ds in LFB group} &= \frac{\hat{p}_2}{1-\hat{p}_2} = \frac{1/83}{82/83} = \frac{1}{82} = .012
\end{aligned}$$

Thus, for the DHCA group, the odds are .130 to 1 (or 1 to 7.7) that an infant will develop seizures. (In general, an infant will not develop seizures).

Similarly, for the LFB group, the odds are .012 to 1 (or 1 to 82) that an infant will develop seizures.

Then

$$\hat{OR} = \frac{10/77}{1/82} = \frac{10 \cdot 82}{1 \cdot 77} = 10.65$$

Thus the odds of developing seizures is 10.65 times higher for the DHCA group than the LFB group.

Note: The odds ratio is the only measure of association that can generally be estimated with retrospective case-control studies. But all three measures of association can be estimated in cohort studies, clinical trials, or cross-sectional designs.

- The estimate of the odds ratio can be put in the form

$$\hat{OR} = \frac{a \cdot d}{b \cdot c},$$

leading to this estimator being called the cross-product ratio.

- More formal comparisons demand either hypothesis testing or confidence intervals.

Pearson Chi-square test: A method of testing the hypothesis of no association in a 2 × 2 table is with the Pearson χ^2 test.

Recall the observed counts:

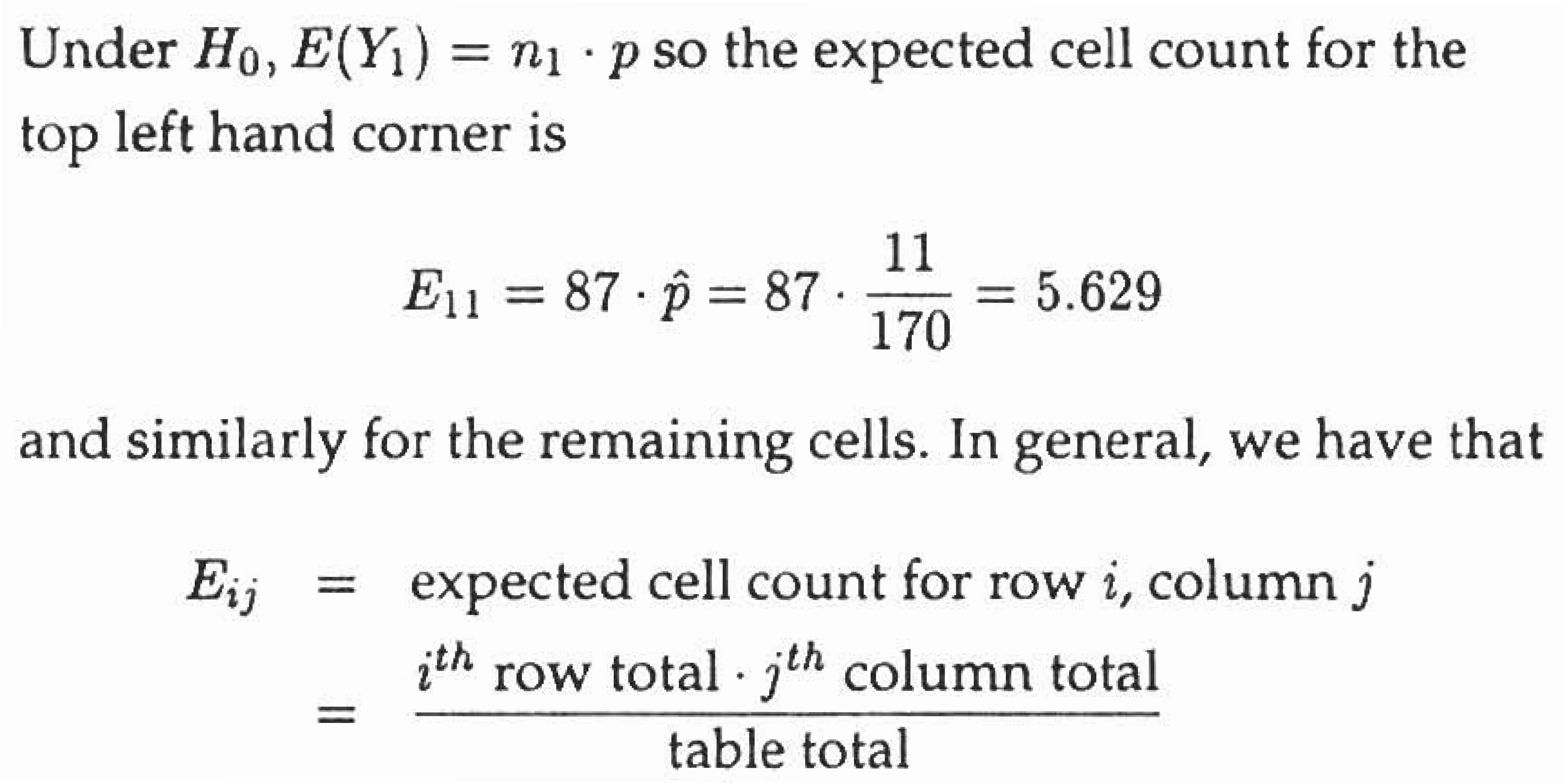
	Seizures		
	Yes	No	
DHCA	10	77	$n_1 = 87$
LFB	1	82	$n_2 = 83$
	$m_1 = 11$	$m_2 = 159$	$N = 170$

What would be the expected counts if $H_0 : p_1 = p_2$ was true?

If H_0 was true, we would estimate

$$\hat{p} = (10 + 1)/(87 + 83) = 11/170 = .065$$

from the combined sample.

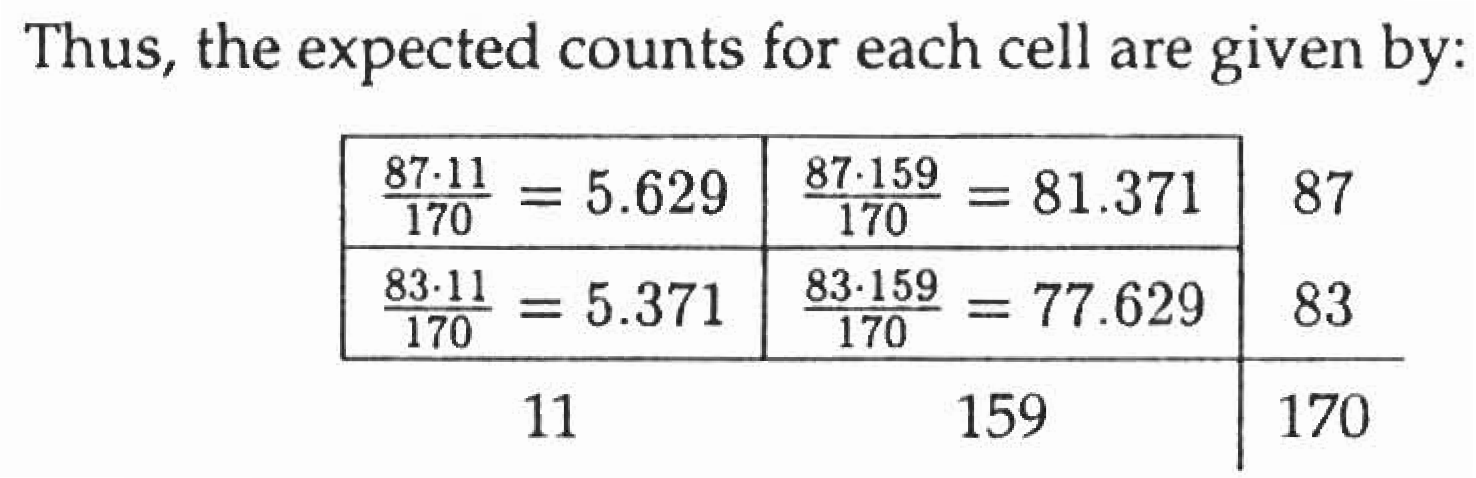


Under $H_0, E(Y_1) = n_1 \cdot p$ so the expected cell count for the top left hand corner is

$$E_{11} = 87 \cdot \hat{p} = 87 \cdot \frac{11}{170} = 5.629$$

and similarly for the remaining cells. In general, we have that

$$\begin{aligned} E_{ij} &= \text{expected cell count for row } i\text{, column } j \\ &= \frac{i^{th} \text{ row total} \cdot j^{th} \text{ column total}}{\text{table total}} \end{aligned}$$



Thus, the expected counts for each cell are given by:

$\frac{87\cdot 11}{170} = 5.629$	$\frac{87\cdot 159}{170} = 81.371$	87
$\frac{83\cdot 11}{170} = 5.371$	$\frac{83\cdot 159}{170} = 77.629$	83
11	159	170

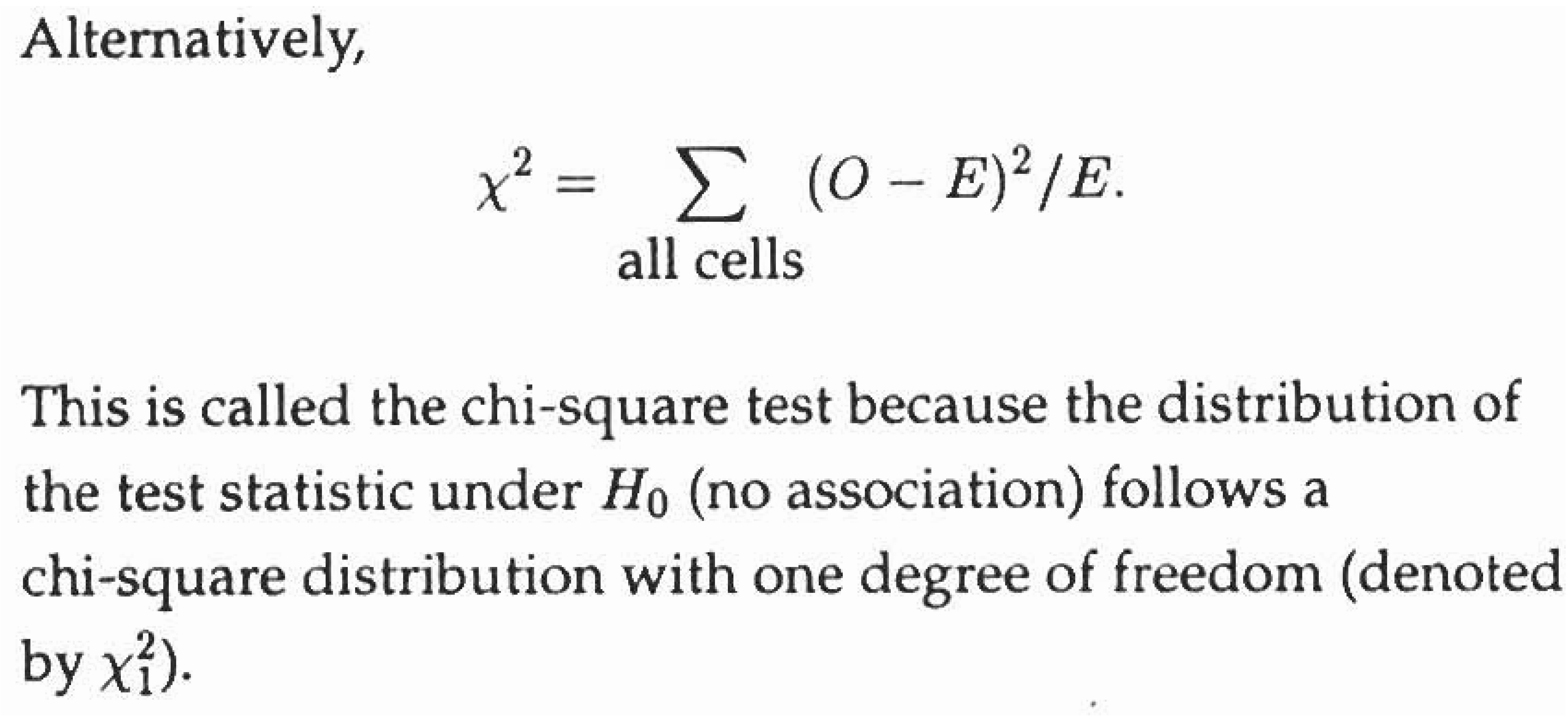
The Pearson Chi-Square test assesses how well the observed counts compare with the expected counts under the null hypothesis. If the discrepancies are small, we would have no evidence to reject H_0, while if the discrepancies are large, we would reject H_0. The test statistic is given by:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

$O_{ij} =$ observed cell count for row i, column j

$E_{ij} =$ expected cell count for row i, column j under H_0.



Alternatively,

$$\chi^2 = \sum_{\text{all cells}} (O - E)^2/E.$$

This is called the chi-square test because the distribution of the test statistic under H_0 (no association) follows a chi-square distribution with one degree of freedom (denoted by χ_1^2).

Seizure example, continued:

$$\begin{aligned}\chi^2 &= \frac{(10-5.629)^2}{5.629}+\frac{(77-81.371)^2}{81.371} \\ &\quad +\frac{(1-5.371)^2}{5.371}+\frac{(82-77.629)^2}{77.629} \\ &= 7.43\end{aligned}$$

Using a computer package like STATA or SAS we find that the *p*-value is given by

$$P_{H_0}(\chi^2 \geq 7.43) = 0.006$$

Fisher’s exact test:

Tests based upon χ^2 distributions work well for large sample sizes. Fisher suggested a more complicated test that is valid even for small samples.

Using a computer package like STATA or SAS we find that Fisher’s exact p-value is equal to 0.010. Since some of the cell counts are small in the seizures example, it may be preferable to use Fisher’s exact test for inferences. One rule of thumb is to always use Fisher’s exact test if any of the expected cell counts is less than 5. For large samples, Fisher’s and Pearson’s tests should give similar results.

Interval Estimation of our Measures of Association: In most applications we are more interested in the point and interval estimation of the risk difference, relative risk, or odds ratio than in hypothesis testing.

CI for the Risk Difference or Attributable Risk:

	D	$\bar{D}$	
E	Y_1		n_1
$\bar{E}$	Y_2		n_2

or

	D	$\bar{D}$	
E	a	b	n_1
$\bar{E}$	c	d	n_2

$$\begin{aligned} RD &= p_1 - p_2 \\ \hat{RD} &= \hat{p}_1 - \hat{p}_2 = \frac{Y_1}{n_1} - \frac{Y_2}{n_2} = \frac{a}{n_1} - \frac{c}{n_2} \end{aligned}$$

One can show that an approximate two-sided $100(1-\alpha)\%$ confidence interval for RD is given by:

$$\hat{RD} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$$

or

$$\hat{RD} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{Y_1(n_1-Y_1)}{n_1^3}+\frac{Y_2(n_2-Y_2)}{n_2^3}}$$

or

$$\hat{RD} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{a\cdot b}{n_1^3}+\frac{c\cdot d}{n_2^3}} .$$

Seizures example, continued:

$$\begin{aligned}\hat{p}_1 &= 10/87 = .115\\ \hat{p}_2 &= 1/83 = .012\\ \hat{RD} &= \hat{p}_1 - \hat{p}_2 = .103\end{aligned}$$

A 95% CI for the RD is given by:

$$.103 \pm 1.96\sqrt{\frac{(.115)(.885)}{87} + \frac{(.012)(.988)}{83}} = (.032, .174)$$

Since 0 is not included in the CI for the RD, we conclude that we have significant differences (at the two-sided 0.05 level).

Confidence Interval for the Relative Risk or Risk Ratio:

$$RR = p_1/p_2$$

$RR = 1$ no association between disease and exposure

$RR > 1$ positive association between disease and exposure

$RR < 1$ negative association between disease and exposure

$$\hat{RR} = \hat{p}_1/\hat{p}_2 = \left(\frac{Y_1}{n_1}\right) / \left(\frac{Y_2}{n_2}\right) = \left(\frac{a}{n_1}\right) / \left(\frac{c}{n_2}\right)$$

Using a natural logarithm transformation yields an approximate two-sided $100(1-\alpha)\%$ confidence interval for $\log RR$ as follows:

$$\log \hat{RR} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{q}_1}{n_1\hat{p}_1} + \frac{\hat{q}_2}{n_2\hat{p}_2}}.$$

or

$$\log \hat{RR} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{n_1 - Y_1}{n_1 \cdot Y_1} + \frac{n_2 - Y_2}{n_2 \cdot Y_2}}$$

or

$$\log \hat{RR} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{b}{a \cdot n_1} + \frac{d}{c \cdot n_2}}$$

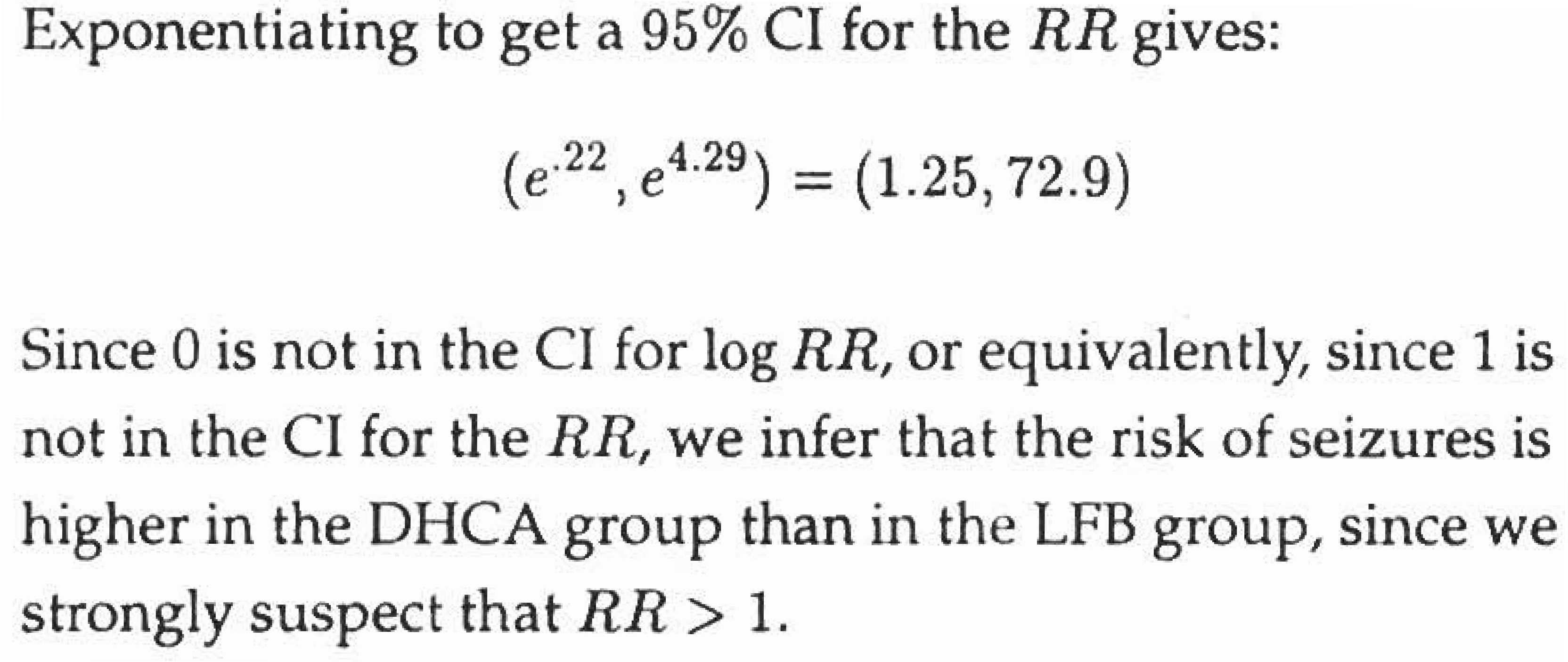
We exponentiate the CI for $\log RR$ to construct a CI for the RR.

Seizures example, continued:

$$\begin{aligned}
\hat{p}_1 &= 10/87 = .115 \\
\hat{p}_2 &= 1/83 = .012 \\
\hat{RR} &= \hat{p}_1/\hat{p}_2 = 9.54 \\
\log \hat{RR} &= 2.26
\end{aligned}$$

A 95% CI for the $\log(RR)$ is given by:

$$2.26 \pm 1.96\sqrt{\frac{.885}{(87)(.115)} + \frac{.988}{(83)(.012)}} = (.22, 4.29)$$



Exponentiating to get a 95% CI for the RR gives:

$$(e^{.22}, e^{4.29}) = (1.25, 72.9)$$

Since 0 is not in the CI for log RR, or equivalently, since 1 is not in the CI for the RR, we infer that the risk of seizures is higher in the DHCA group than in the LFB group, since we strongly suspect that $RR > 1$.

Confidence Interval for the Odds Ratio or Relative Odds:

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

$$\begin{aligned} OR &= 1 \text{ no association between disease and exposure} \\ OR &> 1 \text{ positive association between disease and exposure} \\ OR &< 1 \text{ negative association between disease and exposure} \end{aligned}$$

$$\begin{aligned} \hat{OR} &= \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{(Y_1/n_1)/((n_1-Y_1)/n_1)}{(Y_2/n_2)/((n_2-Y_2)/n_2)} \\ &= \frac{Y_1/(n_1-Y_1)}{Y_2/(n_2-Y_2)} = \frac{a/b}{c/d} = \frac{ad}{bc} \end{aligned}$$

Again using a natural logarithm transformation, an approximate two-sided $100(1-\alpha)\%$ confidence interval for $\log OR$ is given by:

$$\log \hat{OR} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_1 \hat{p}_1 \hat{q}_1} + \frac{1}{n_2 \hat{p}_2 \hat{q}_2}}$$

or

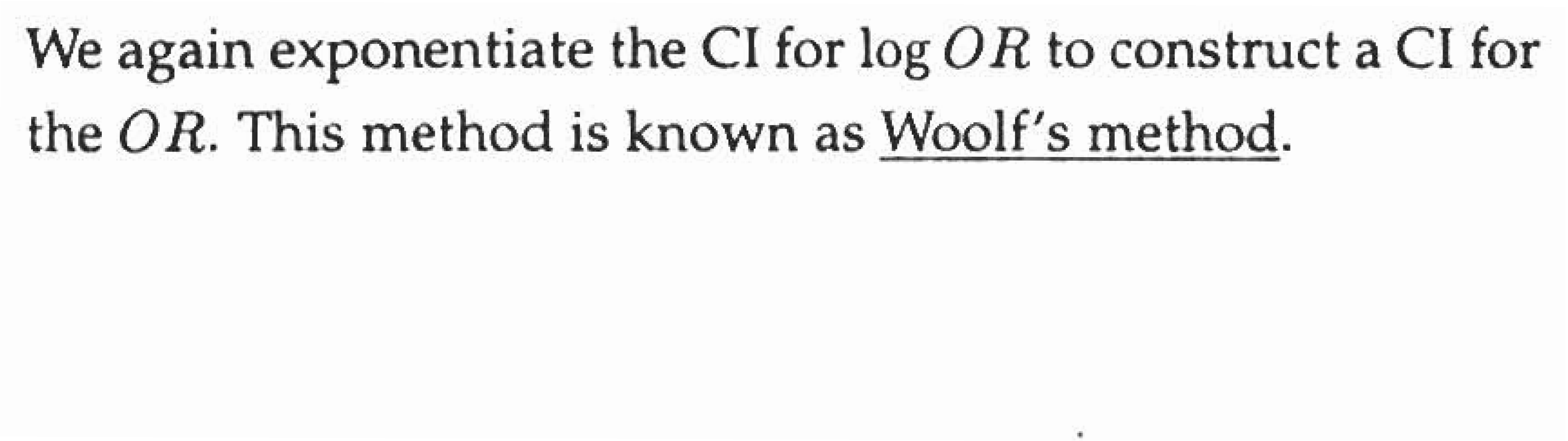
$$\log \hat{OR} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{n_1}{Y_1 \cdot (n_1 - Y_1)} + \frac{n_2}{Y_2 \cdot (n_2 - Y_2)}}$$

or

$$\log \hat{OR} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{n_1}{a \cdot b} + \frac{n_2}{c \cdot d}}$$

or

$$\log \hat{OR} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$



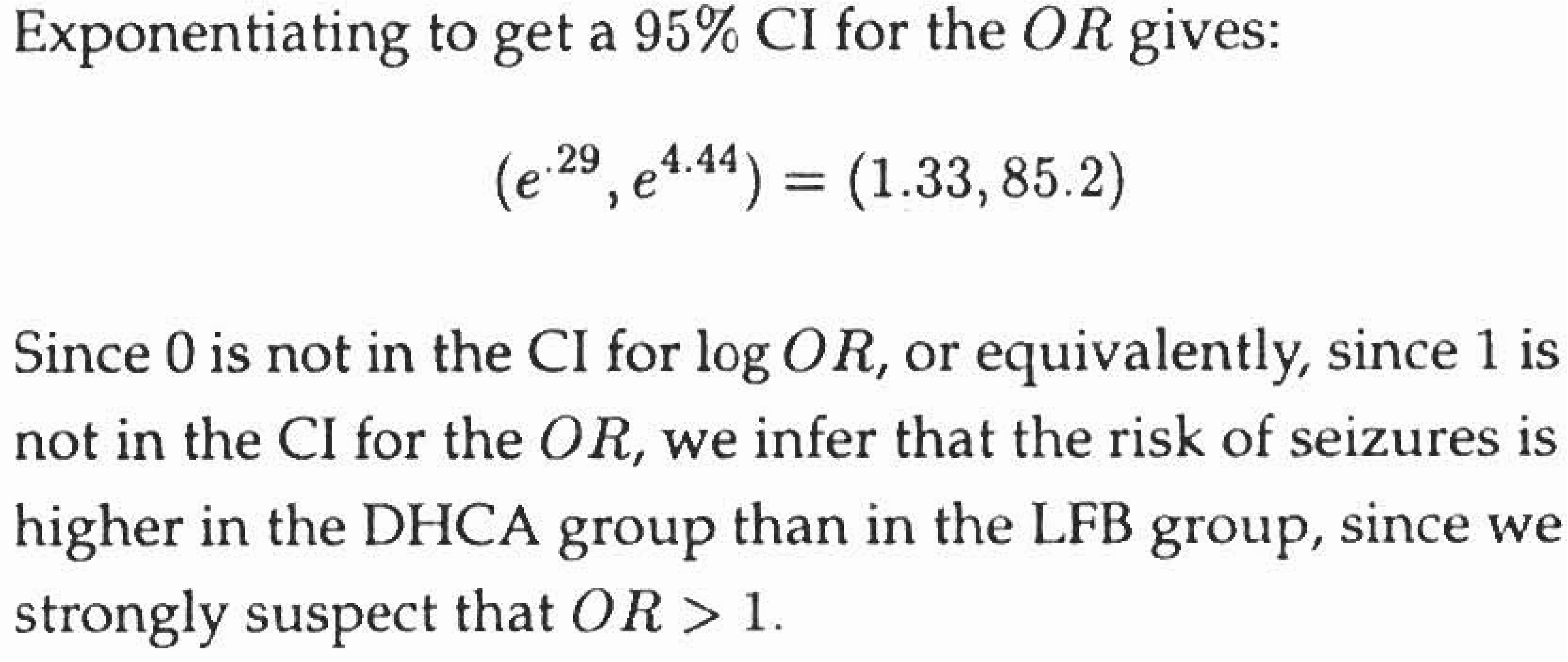
We again exponentiate the CI for $\log OR$ to construct a CI for the OR. This method is known as Woolf's method.

Seizure example, continued:

$$\begin{aligned}\hat{p}_1 &= 10/87 = .115\\ \hat{p}_2 &= 1/83 = .012\\ \hat{OR} &= \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{ad}{bc} = 10.65\\ \log \hat{OR} &= 2.37\end{aligned}$$

A 95% CI for $\log(OR)$ is given by:

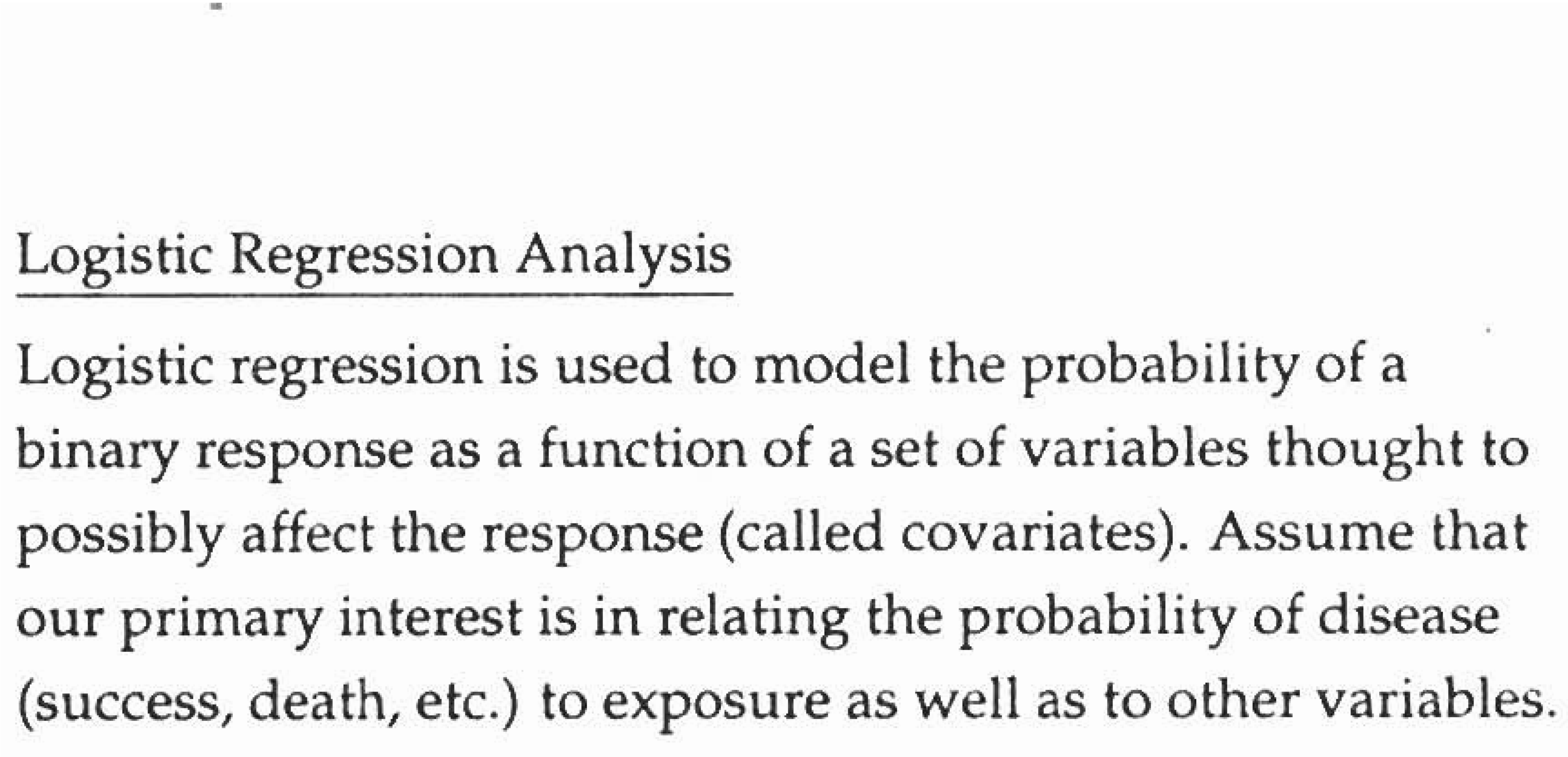
$$2.37 \pm 1.96\sqrt{\frac{1}{10}+\frac{1}{77}+\frac{1}{1}+\frac{1}{82}} = (.29, 4.44)$$



Exponentiating to get a 95% CI for the OR gives:

$$(e^{.29}, e^{4.44}) = (1.33, 85.2)$$

Since 0 is not in the CI for $\log OR$, or equivalently, since 1 is not in the CI for the OR, we infer that the risk of seizures is higher in the DHCA group than in the LFB group, since we strongly suspect that $OR > 1$.



Logistic Regression Analysis

Logistic regression is used to model the probability of a binary response as a function of a set of variables thought to possibly affect the response (called covariates). Assume that our primary interest is in relating the probability of disease (success, death, etc.) to exposure as well as to other variables.

Example: Consider again the clinical trial comparing deep hypothermic circulatory arrest (DHCA) vs. low-flow bypass (LFB) in infants with transposition of the great arteries (TGA). In this study, the binary response variable is given by:

$$Y = \begin{cases} 1 & \text{if infant develops seizures} \\ 0 & \text{if not} \end{cases}$$

.

The primary "exposure" variable in the clinical trial is given by: $x_1 = \begin{cases} 1 & \text{if the treatment group is DHCA} \\ 0 & \text{if the treatment group if LFB} \end{cases}$

But the other covariates may also be of interest, including:

$$x_2 = \begin{cases} 1 & \text{if the preop neurologic exam was normal} \\ 2 & \text{if the preop neurologic exam was possibly abnormal} \\ 3 & \text{if the preop neurologic exam was definitely abnormal} \end{cases}$$

$$x_3 = \text{duration of circulatory arrest (in minutes)}$$

Thus covariates can be qualitative (discrete), ordinal (ordered categorical), or quantitative (continuous).

Simple logistic regression (with a continuous covariate):

The modelled probability of seizures (p_x = P(Y=1)) as a function of minutes of circulatory arrest (x) is given by:

$$p_x = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

Sometimes analyses with a single covariate are called crude analyses, since there is no adjustment for any other factors.

The logistic regression parameters are estimated using maximum likelihood.

From a logistic regression package (e.g., STATA or SAS) we find that:

$\hat{\beta} = 0.063$	$\widehat{s.e.}(\hat{\beta}) = 0.022$	minutes or slope term
$\hat{\alpha} = -5.62$	$\widehat{s.e.}(\hat{\alpha}) = 1.24$	intercept or constant term

With large samples, $\hat{\beta}$ follows an approximately normal distribution with true mean β and an estimated variance of $\widehat{\text{Var}}(\hat{\beta}) = (\widehat{s.e.}(\hat{\beta}))^2$. Thus a 95% CI for β is given by:

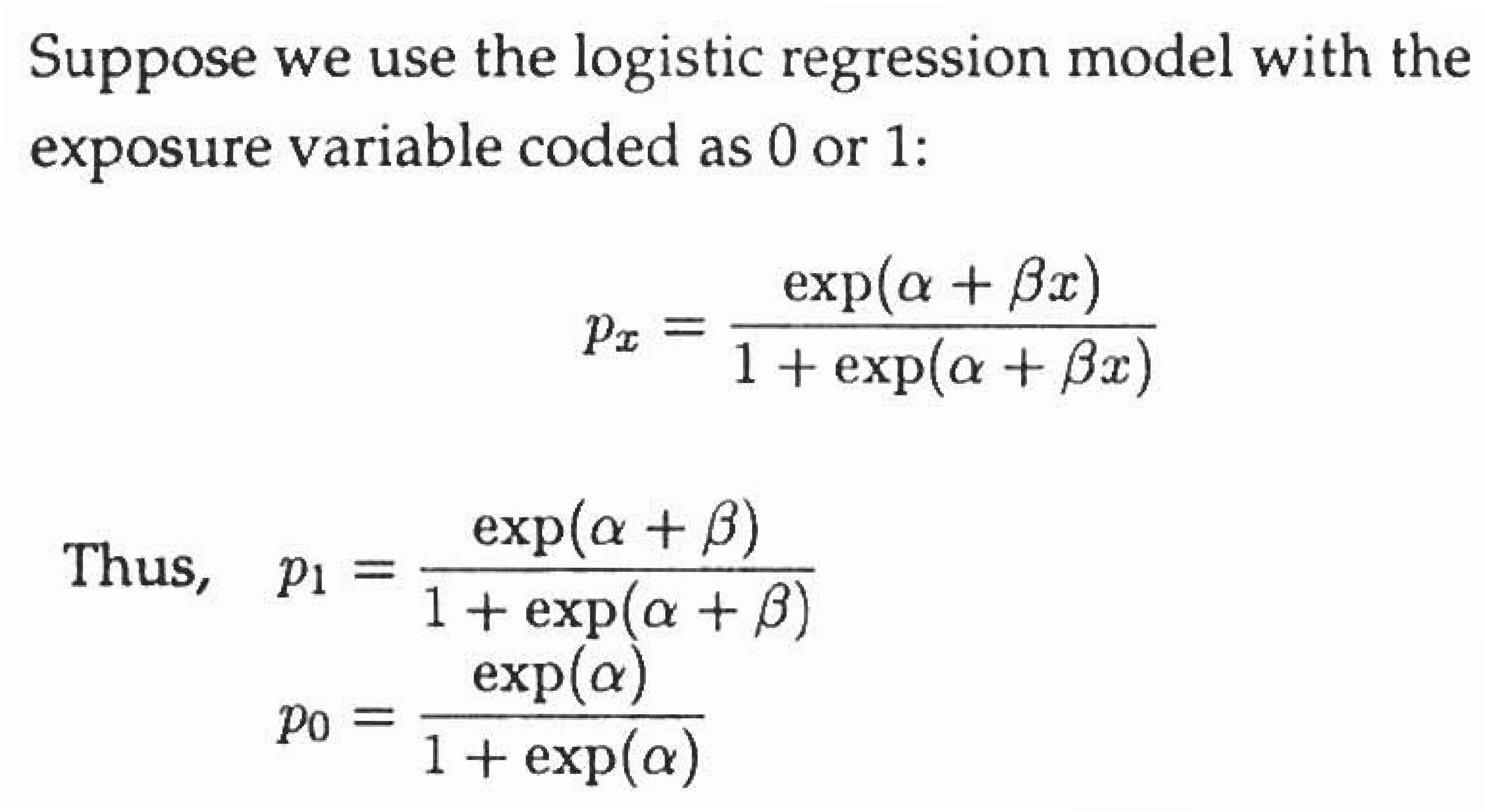
$$\begin{aligned} \hat{\beta} \pm 1.96\widehat{s.e.}(\hat{\beta}) &= 0.063 \pm 1.96(0.022) \\ &= (0.020, 0.107) \end{aligned}$$

Simple logistic regression (with a dichotomous covariate):
Suppose we are considering a study where the outcome variable is disease/non-disease and the predictor variable is exposed/ non-exposed, which we "code" as an indicator variable, or dummy variable. Let

$$Y = \begin{cases} 1 & D \\ 0 & \bar{D} \end{cases} \qquad x = \begin{cases} 1 & E \\ 0 & \bar{E} \end{cases}$$

$$\begin{aligned} \text{and } p_x &= \text{Prob(Disease given exposure } x) \\ &= P(Y = 1|x) \qquad x = 0, 1 \\ \text{Thus, } p_1 &= \text{probability of disease among exposed} \\ p_0 &= \text{probability of disease among non-exposed} \end{aligned}$$

Note: This is nothing more than the standard 2 × 2 table!



Suppose we use the logistic regression model with the exposure variable coded as 0 or 1:

$$p_x = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

Thus, $$p_1 = \frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)}$$
$$p_0 = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$$

The β coefficient corresponds to the logs odds ratio comparing the exposed to the unexposed populations:

$$\beta = \log(OR)$$

and hence

$$OR = \exp(\beta)$$

Thus, if we estimate β from our data, we can estimate

$$\log(\widehat{OR}) = \hat{\beta} \quad \text{or} \quad \widehat{OR} = \exp(\hat{\beta})$$

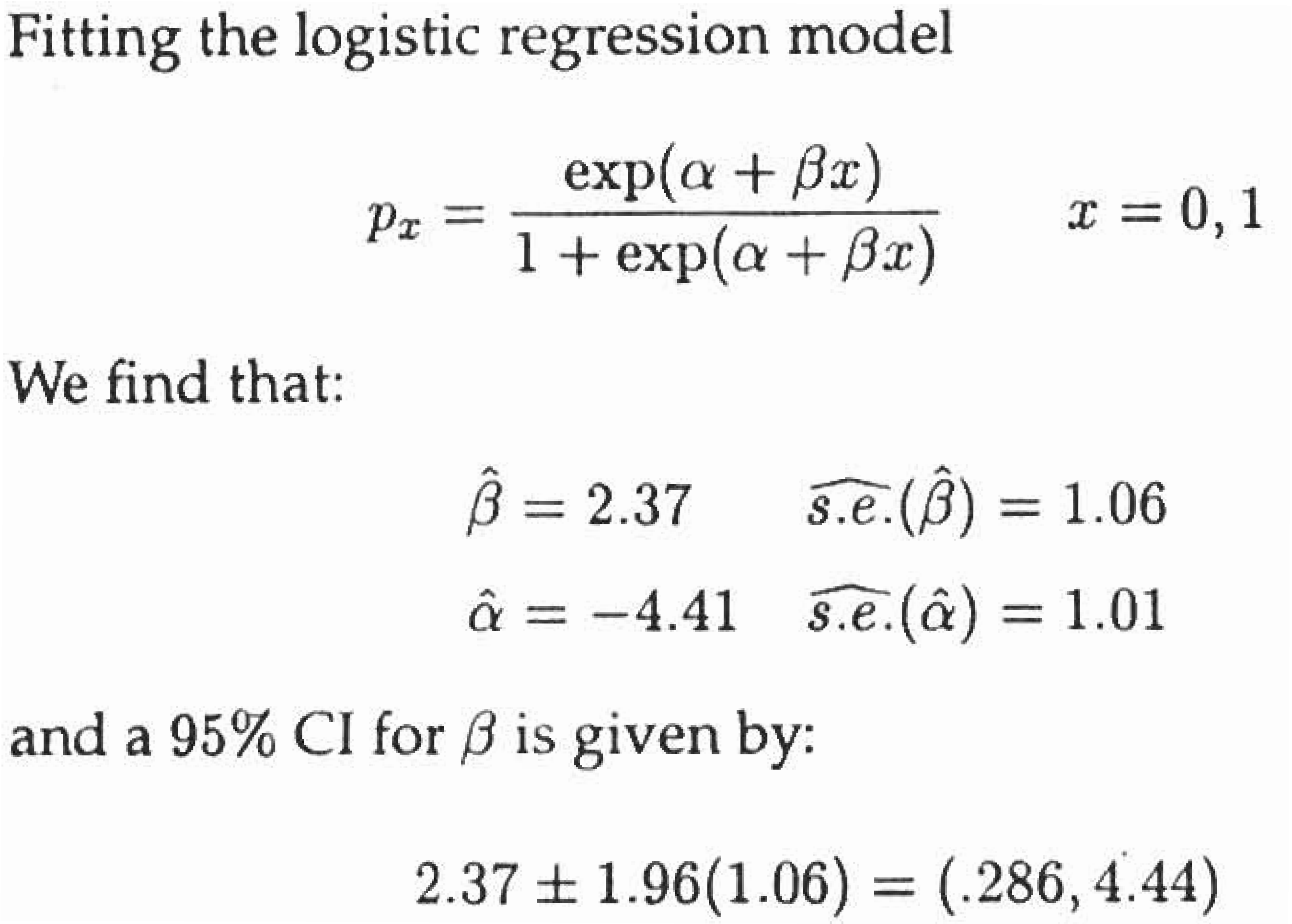
So far, we have examined single covariate models. We can also adjust for other covariates or possible confounding factors with multiple covariate models.

Example: Consider again the clinical trial on infants undergoing cardiac surgery for repair of TGA:

		Observed Clinical Seizures Within 7 Postop Days		
		Yes	No	
Treatment	DHCA	10	77	$n_1 = 87$
Group	LFB	1	82	$n_2 = 83$

If we code treatment $= \begin{cases} 1 & \text{for DHCA} \\ 0 & \text{for LFB} \end{cases}$

Then $\hat{p}_1 = 10/87 = .115$ and $\hat{p}_0 = 1/83 = .012$



Fitting the logistic regression model

$$p_x = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad x = 0, 1$$

We find that:

$$\hat{\beta} = 2.37 \qquad \widehat{s.e.}(\hat{\beta}) = 1.06$$

$$\hat{\alpha} = -4.41 \quad \widehat{s.e.}(\hat{\alpha}) = 1.01$$

and a 95% CI for β is given by:

$$2.37 \pm 1.96(1.06) = (.286, 4.44)$$

Also using the logistic regression results we find that

$$\begin{aligned}\widehat{OR} &= \exp(\hat{\beta}) \\ &= \exp(2.37) = 10.65\end{aligned}$$

is an estimate of the odds ratio measuring the association between treatment and seizures. Because of the coding, this odds ratio estimate compares DHCA (treatment = 1) to LFB (treatment = 0), a one unit change in the covariate.

Similarly a 95% CI for the OR can be obtained by exponentiating the the 95% CI for β, namely:

$$(\exp(.286)), \exp(4.44)) = (1.33, 85.16),$$

giving results very similar to those obtained previously using Woolf's CI for the OR.

Odds ratios can also be estimated from logistic regression models with continuous covariates. Let:

$$p_x = P(Y = 1|\text{covariate } x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

and

$$p_{x+1} = P(Y = 1|\text{covariate } x + 1) = \frac{\exp(\alpha + \beta(x + 1))}{1 + \exp(\alpha + \beta(x + 1))}.$$

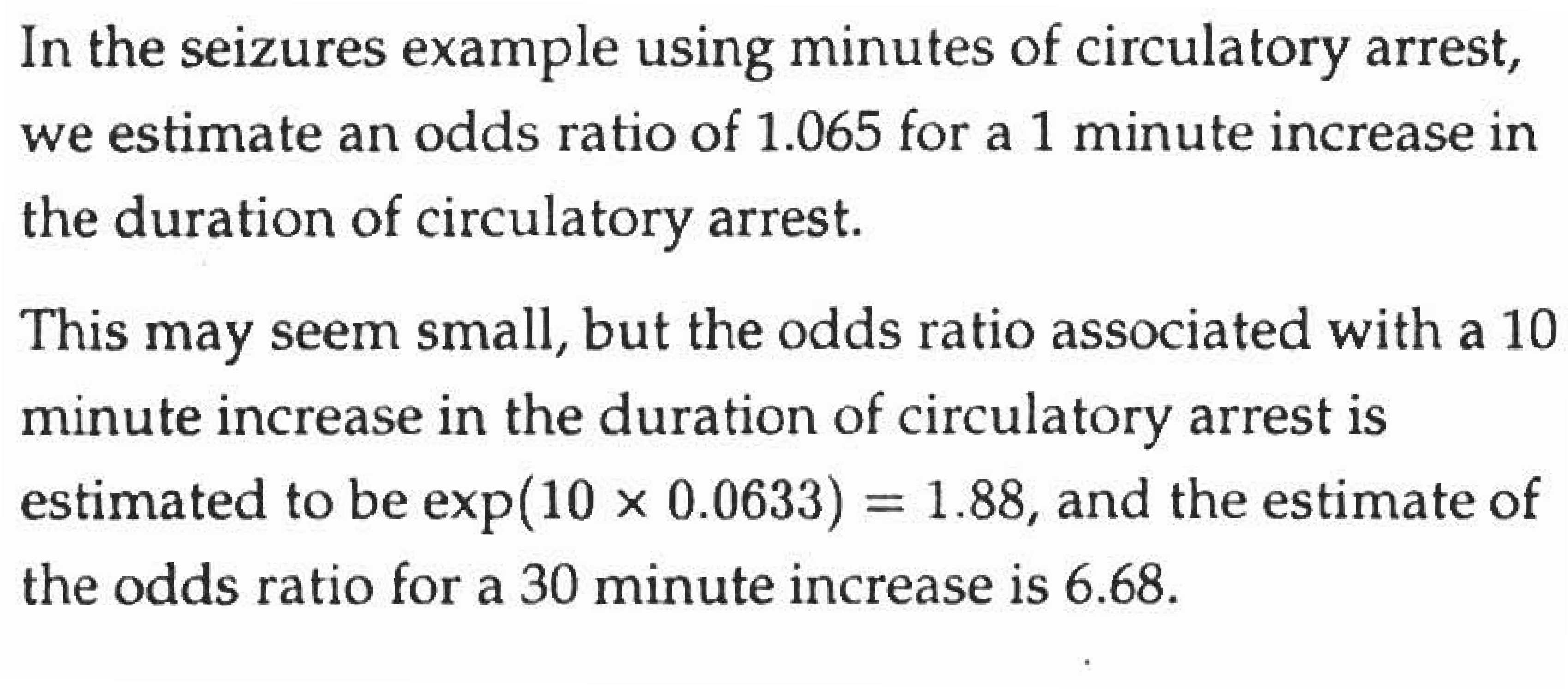
Then the OR associated with a 1 unit increase in x is given by:

$$\begin{aligned} OR &= \frac{p_{x+1}/(1 - p_{x+1})}{p_x/(1 - p_x)} \\ &\overset{algebra}{=} \exp(\beta \cdot 1) = e^{\beta} \end{aligned}$$

Similarly the OR associated with a 10 unit increase in x is given by:

$$OR = \exp(\beta \cdot 10) = e^{10\beta},$$

etc. Point estimates and CI's for β can be transformed to get point estimates and CI's for the OR.



In the seizures example using minutes of circulatory arrest, we estimate an odds ratio of 1.065 for a 1 minute increase in the duration of circulatory arrest.

This may seem small, but the odds ratio associated with a 10 minute increase in the duration of circulatory arrest is estimated to be $\exp(10 \times 0.0633) = 1.88$, and the estimate of the odds ratio for a 30 minute increase is 6.68.

Instead of seeing whether the CI for the OR contains 1, the Wald test offers a hypothesis testing approach. To test $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$, the test statistic given by:

$$Z = \frac{\hat{\beta}}{\widehat{s.e.}(\hat{\beta})}$$

follows a $N(0, 1)$ distribution under H_0 for large samples. We reject H_0 when Z is sufficiently large.

STATA gives $p = .026$ for the effect of minutes of circulatory arrest by the Wald test in the seizures example. Since $\hat{\beta} > 0$, we infer that an increase in circulatory arrest leads to an increase in the risk of seizures.

Some advantages of logistic regression:

- You can adjust for many covariates simultaneously.
- Handles qualitative and quantitative covariates.
- Allows direct tests of interaction (effect modification).
- You can assess potential confounders (fitting models with and without confounders, and comparing "crude" and "adjusted" estimates).
- You can obtain estimates and confidence intervals for odds ratios (crude or adjusted).
- Mathematically convenient; easy software availability.
- High tech.

Some disadvantages of logistic regression:

- Is abstract and mathematical.
- May be a barrier between the investigator and the data; you might get a better feel for what is going on with an analysis using classical methods (e.g., Mantel-Haenszel approach).
- Has implicit assumptions, which might be difficult or awkward to check.
- Several models may appear to fit well, and it can be difficult to select between these.
- Overconfidence in your final model; "I did an extensive computer analysis; therefore, my conclusions must be correct."
- High tech.

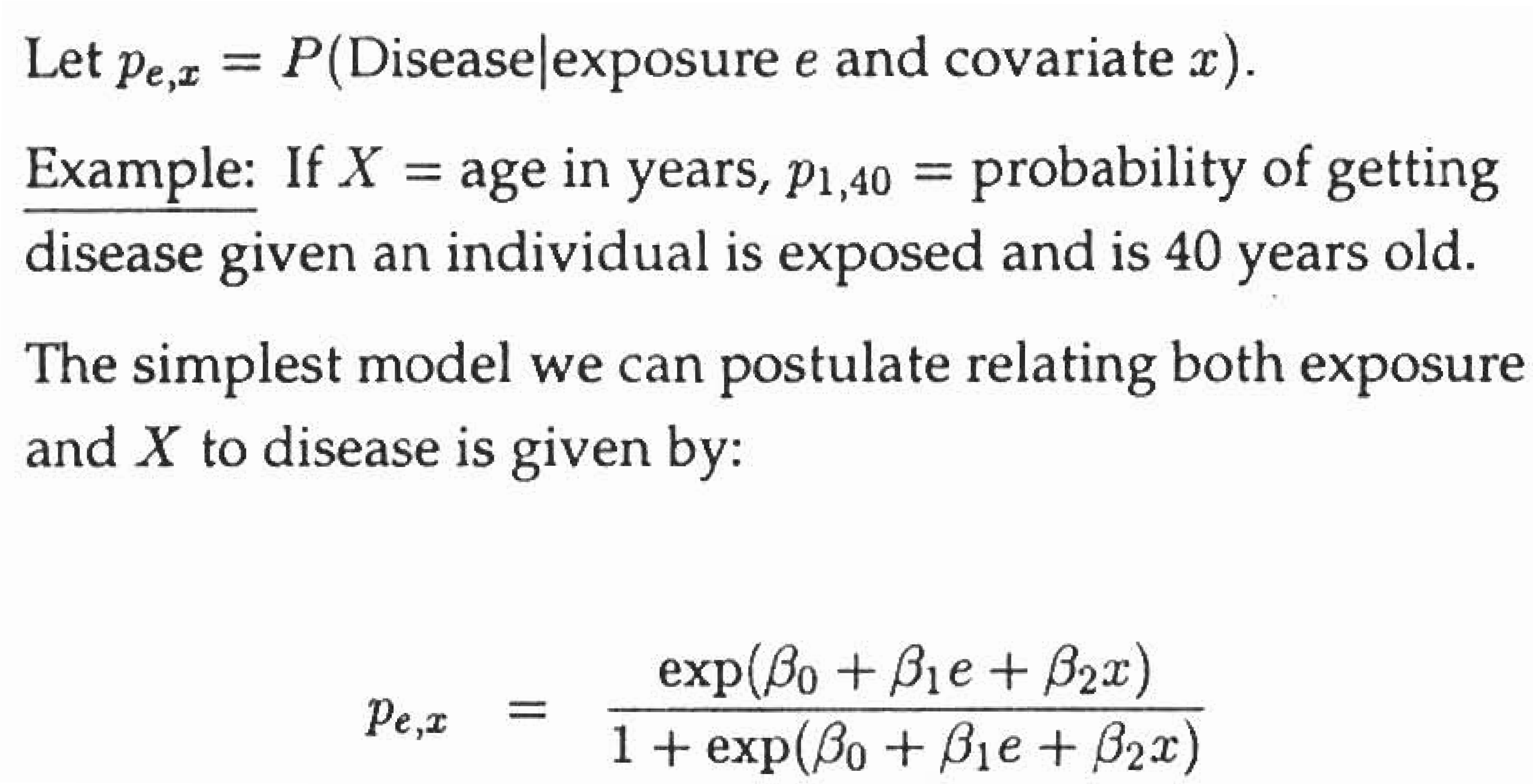
Multivariate Analysis and Control for Confounding

Logistic regression can be used to model the probability of response to many variables simultaneously. This can be used to control for confounding of other variables on the relationship between exposure and disease. By accounting for the effects of the confounding variables as well as exposure, we can separate out the effects that each of these have on the probability of disease.

For now, consider a single binary exposure variable,

$$E = \begin{cases} 1 & \text{exposed} \\ 0 & \text{no exposed} \end{cases}$$

and one confounding variable, X. This X could be quantitative (continuous) or qualitative (discrete).



Let $p_{e,x} = P(\text{Disease}|\text{exposure } e \text{ and covariate } x)$.

Example: If X = age in years, $p_{1,40}$ = probability of getting disease given an individual is exposed and is 40 years old.

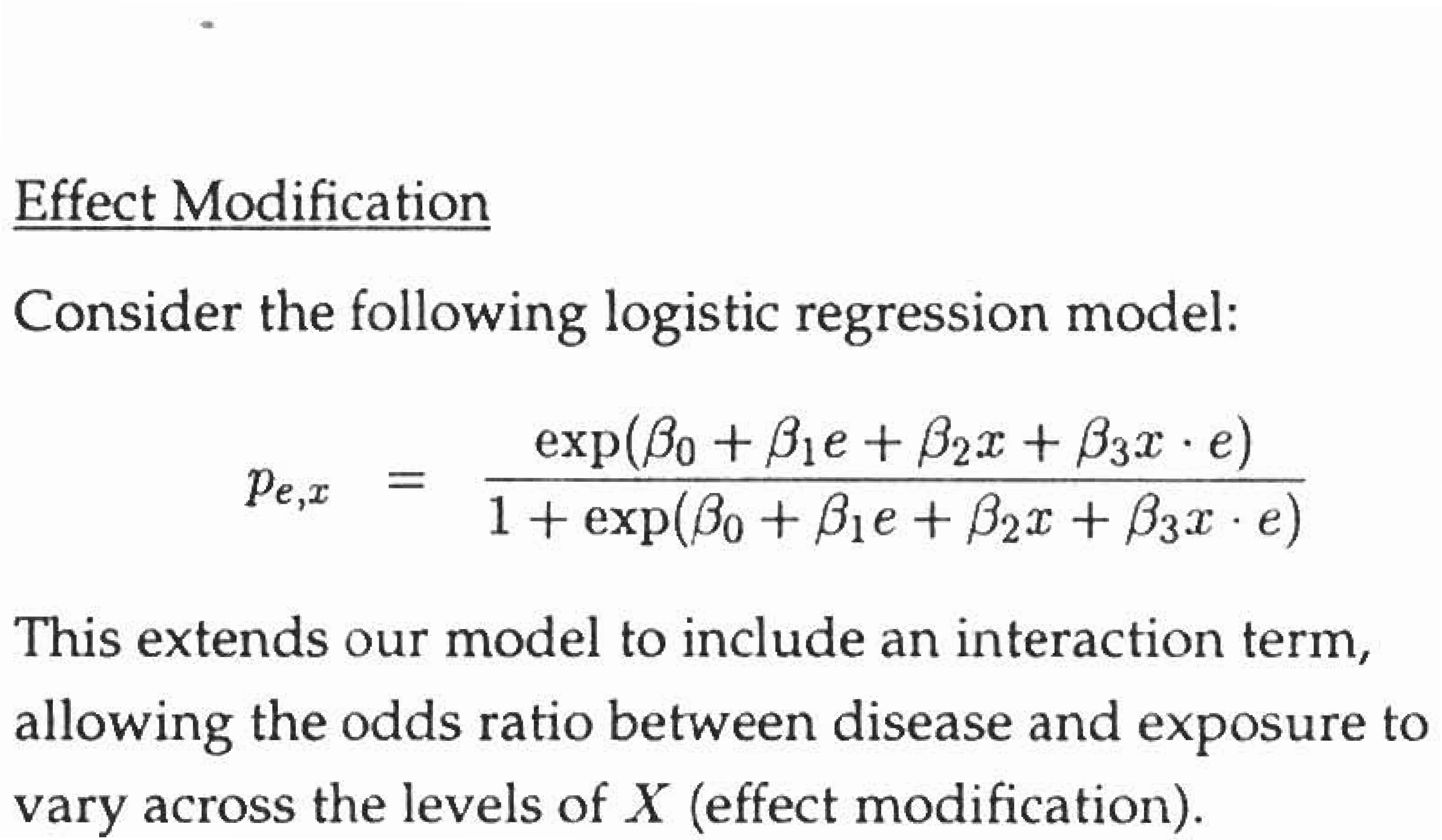
The simplest model we can postulate relating both exposure and X to disease is given by:

$$p_{e,x} = \frac{\exp(\beta_0 + \beta_1 e + \beta_2 x)}{1 + \exp(\beta_0 + \beta_1 e + \beta_2 x)}$$

Interpretation of β coefficients:

This model implies a common odds ratio e^{β_1} between disease and exposure for all possible values of X. By including both X and E in the model, we have separated out the effects of X and E on disease from each other, and thus we are controlling for (or adjusting for) the possible confounding effects of X on the relationship between disease and exposure.

Similarly, e^{β_2} is the common odds ratio between disease and a one unit increase in X, adjusting for E.

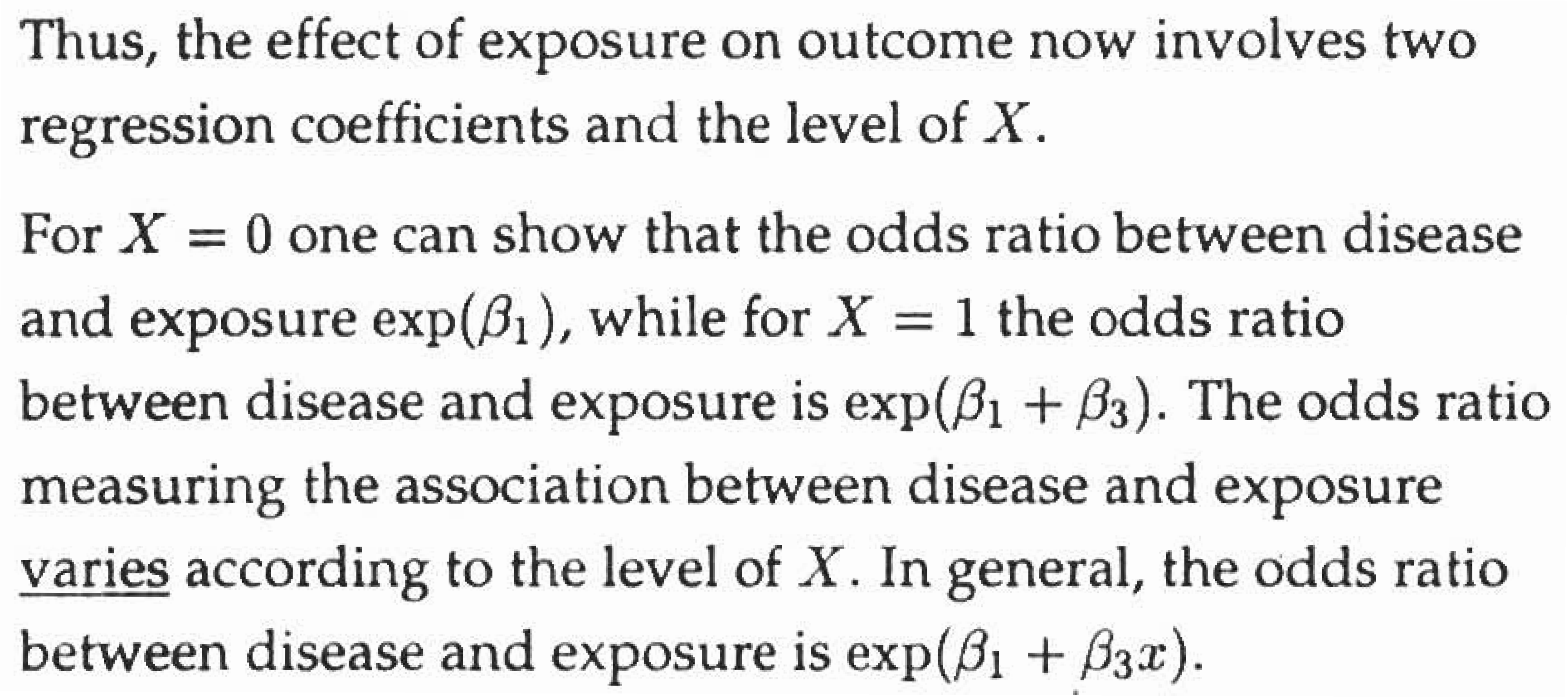


Effect Modification

Consider the following logistic regression model:

$$p_{e,x} \quad = \quad \frac{\exp(\beta_0 + \beta_1 e + \beta_2 x + \beta_3 x \cdot e)}{1 + \exp(\beta_0 + \beta_1 e + \beta_2 x + \beta_3 x \cdot e)}$$

This extends our model to include an interaction term, allowing the odds ratio between disease and exposure to vary across the levels of X (effect modification).



Thus, the effect of exposure on outcome now involves two regression coefficients and the level of X.

For $X = 0$ one can show that the odds ratio between disease and exposure $\exp(\beta_1)$, while for $X = 1$ the odds ratio between disease and exposure is $\exp(\beta_1 + \beta_3)$. The odds ratio measuring the association between disease and exposure varies according to the level of X. In general, the odds ratio between disease and exposure is $\exp(\beta_1 + \beta_3 x)$.

Hence, we can use this model to test for effect modification (interaction). If there was no effect modification, then β_3 should equal zero, otherwise β_3 would be different from zero.

A test for the homogeneity (equality) of the odds ratio across levels of X could then be posed as testing the null hypothesis:

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_A : \beta_3 \neq 0.$$

We could use Wald's test for this.

Categorical Predictor Variables with more than 2 Categories

Suppose we wish to model the probability of response, p, as a function of a categorical variable having $K \geq 2$ categories, labeled $0, 1, 2, \ldots, K - 1$. Here we consider the categories to have no specific order (e.g., eye color, race, state of residence).

A useful way to model this is again through the use of dummy variables (or design variables or indicator variables).

For example:

$$D_1 = I(\text{Cat} = 1) \quad = \quad \begin{cases} 1 & \text{if subject falls in category 1} \\ 0 & \text{if not} \end{cases}$$

$$D_2 = I(\text{Cat} = 2) \quad = \quad \begin{cases} 1 & \text{if subject falls in category 2} \\ 0 & \text{if not} \end{cases}$$

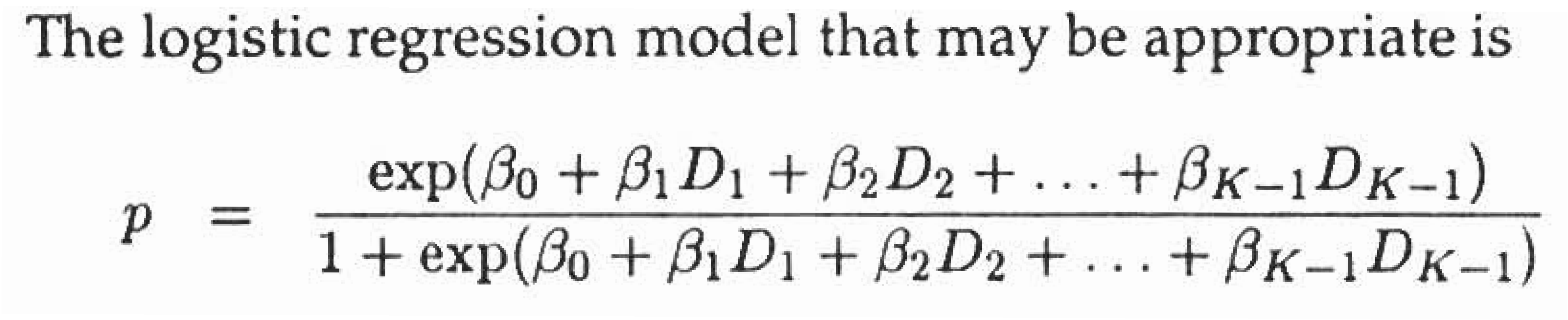
$$\vdots$$

$$D_{K-1} = I(\text{Cat} = K-1) \quad = \quad \begin{cases} 1 & \text{if subject falls in category } K-1 \\ 0 & \text{if not} \end{cases}$$

We refer to the category that is left out as the baseline or reference group (or the 0^{th} category).

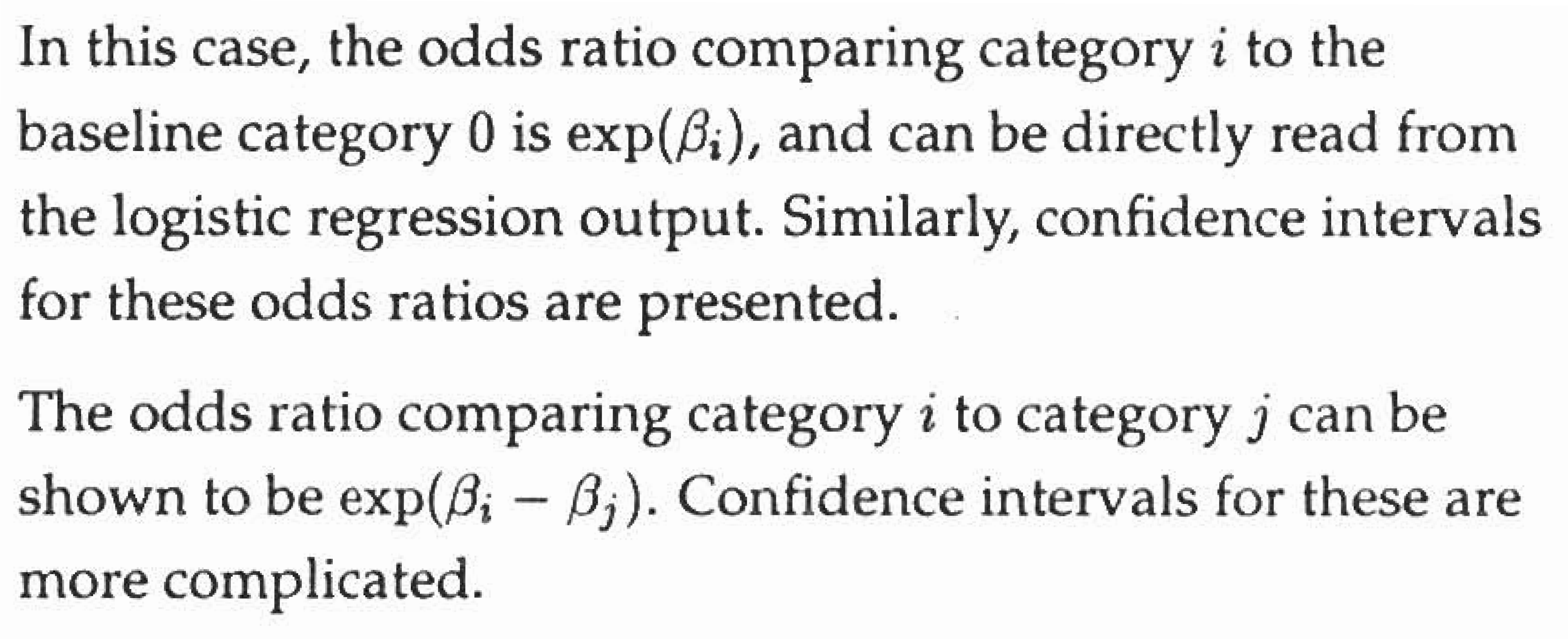
Example of coding:

		Dummy Variables			
Subject	Category	D_1	D_2	$\cdots$	D_{K-1}
1	2	0	1	$\cdots$	0
2	1	1	0	$\cdots$	0
3	0	0	0	$\cdots$	0
4	$K-1$	0	0	$\cdots$	1
$\vdots$	$\vdots$			$\vdots$	



The logistic regression model that may be appropriate is

$$p \quad = \quad \frac{\exp(\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_{K-1} D_{K-1})}{1 + \exp(\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_{K-1} D_{K-1})}$$



In this case, the odds ratio comparing category i to the baseline category 0 is $\exp(\beta_i)$, and can be directly read from the logistic regression output. Similarly, confidence intervals for these odds ratios are presented.

The odds ratio comparing category i to category j can be shown to be $\exp(\beta_i - \beta_j)$. Confidence intervals for these are more complicated.

Testing the null hypothesis of no association of the probability of response among different categories:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{K-1} = 0$$

$$H_A : \text{One or more of } \beta_1, \ldots, \beta_{K-1} \text{ are nonzero}$$

Under H_0, the probability of response is the same in the K categories. To test this hypothesis, we could use the likelihood ratio test, comparing:

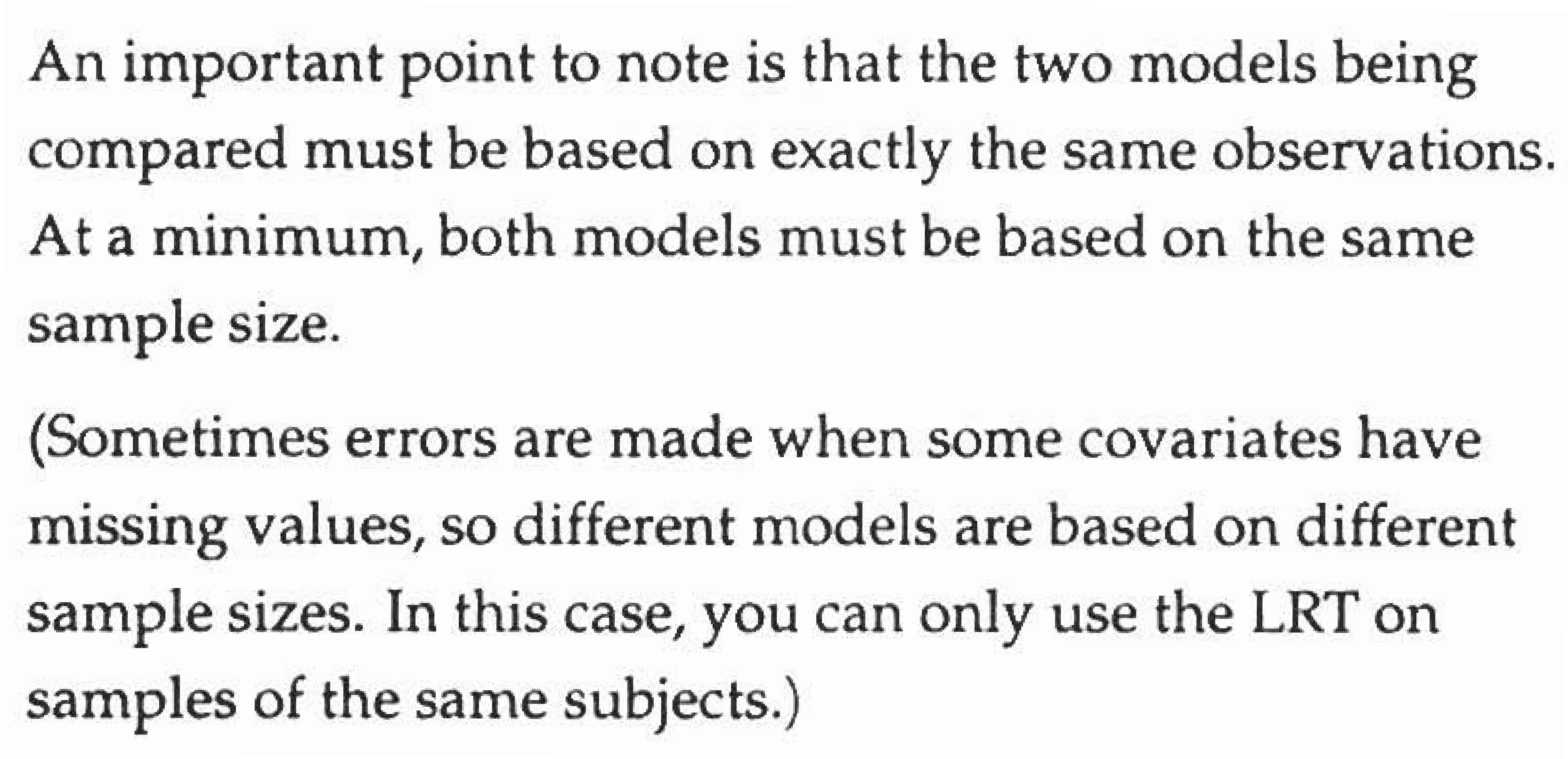
$$H_0 : p = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

$$H_A : p = \frac{\exp(\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_{K-1} D_{K-1})}{1 + \exp(\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_{K-1} D_{K-1})}$$

Likelihood ratio tests (LRT):

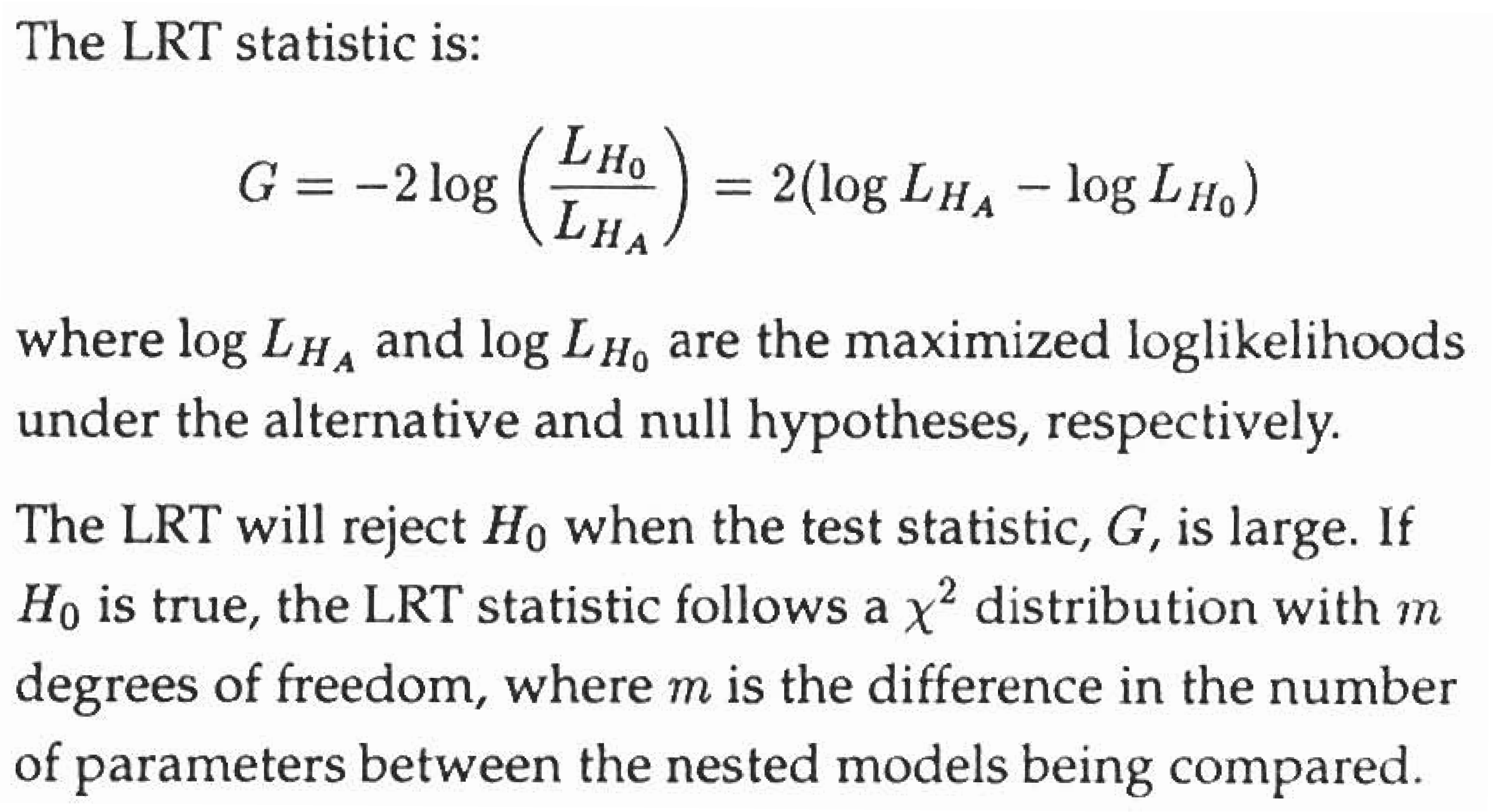
An alternate way (compared to Wald tests) of testing hypotheses in logistic regression models is with the use of a likelihood ratio test. This test is specifically designed to test between nested hypotheses, i.e., when a simpler hypothesis (H_0) is nested under a more complex hypothesis (H_A). Basically, the more complex model must include all of the parameters of the simpler model, plus one or more additional parameters.

A particular advantage of the LRT compared to the Wald test is that the LRT can be used to test more than one parameter at a time.



An important point to note is that the two models being compared must be based on exactly the same observations. At a minimum, both models must be based on the same sample size.

(Sometimes errors are made when some covariates have missing values, so different models are based on different sample sizes. In this case, you can only use the LRT on samples of the same subjects.)



The LRT statistic is:

$$G = -2\log\left(\frac{L_{H_0}}{L_{H_A}}\right) = 2(\log L_{H_A} - \log L_{H_0})$$

where $\log L_{H_A}$ and $\log L_{H_0}$ are the maximized loglikelihoods under the alternative and null hypotheses, respectively.

The LRT will reject H_0 when the test statistic, G, is large. If H_0 is true, the LRT statistic follows a χ^2 distribution with m degrees of freedom, where m is the difference in the number of parameters between the nested models being compared.

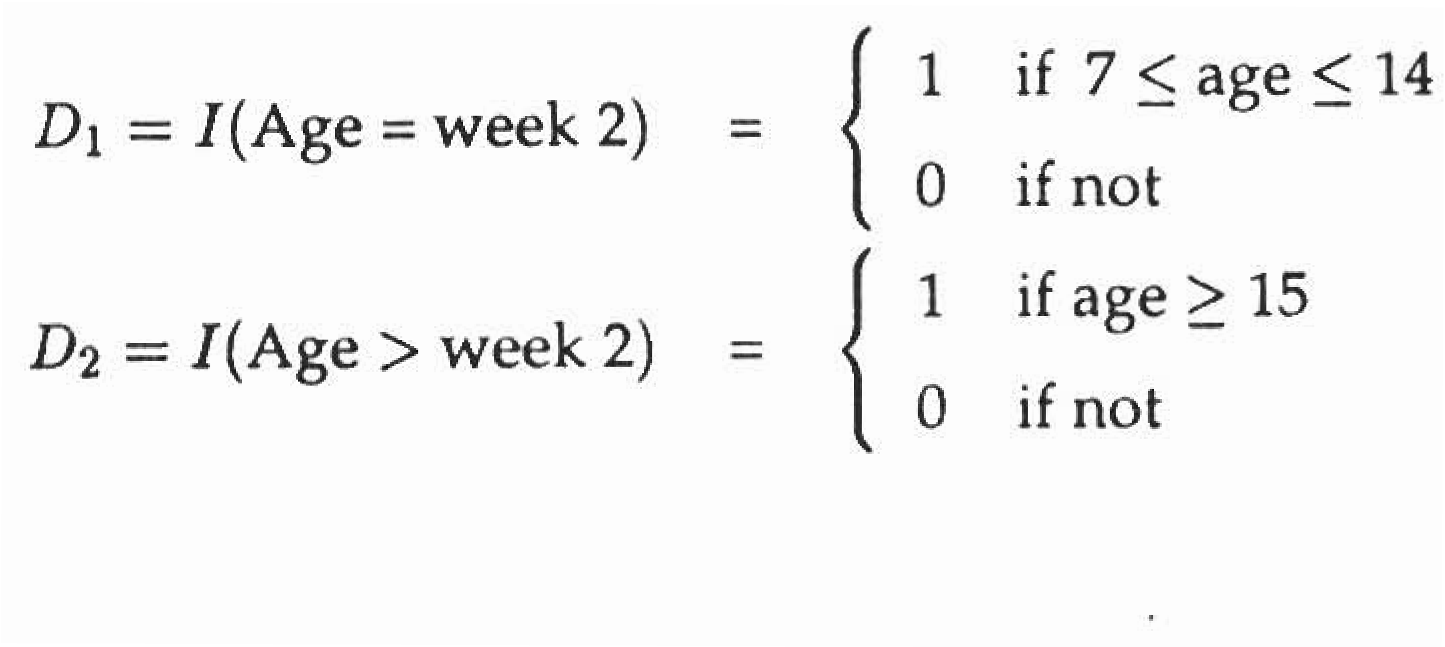
Example: 170 infants undergo cardiac surgery to repair transposition of the great arteries.

$$Y = \begin{cases} 1 & \text{if infant has definite seizures} \\ & \text{within 7 days postop} \\ 0 & \text{if not} \end{cases}$$

Covariate of interest: Age at surgery (in days), which ranges from 1 to 67 days in the study.

Suppose we break age at surgery into three categories:

Category	Age Range	
0	1–6	(baseline group)
1	7–14	
2	≥ 15	



$$D_1 = I(\text{Age} = \text{week } 2) \quad = \quad \begin{cases} 1 & \text{if } 7 \leq \text{age} \leq 14 \\ 0 & \text{if not} \end{cases}$$

$$D_2 = I(\text{Age} > \text{week } 2) \quad = \quad \begin{cases} 1 & \text{if age} \geq 15 \\ 0 & \text{if not} \end{cases}$$

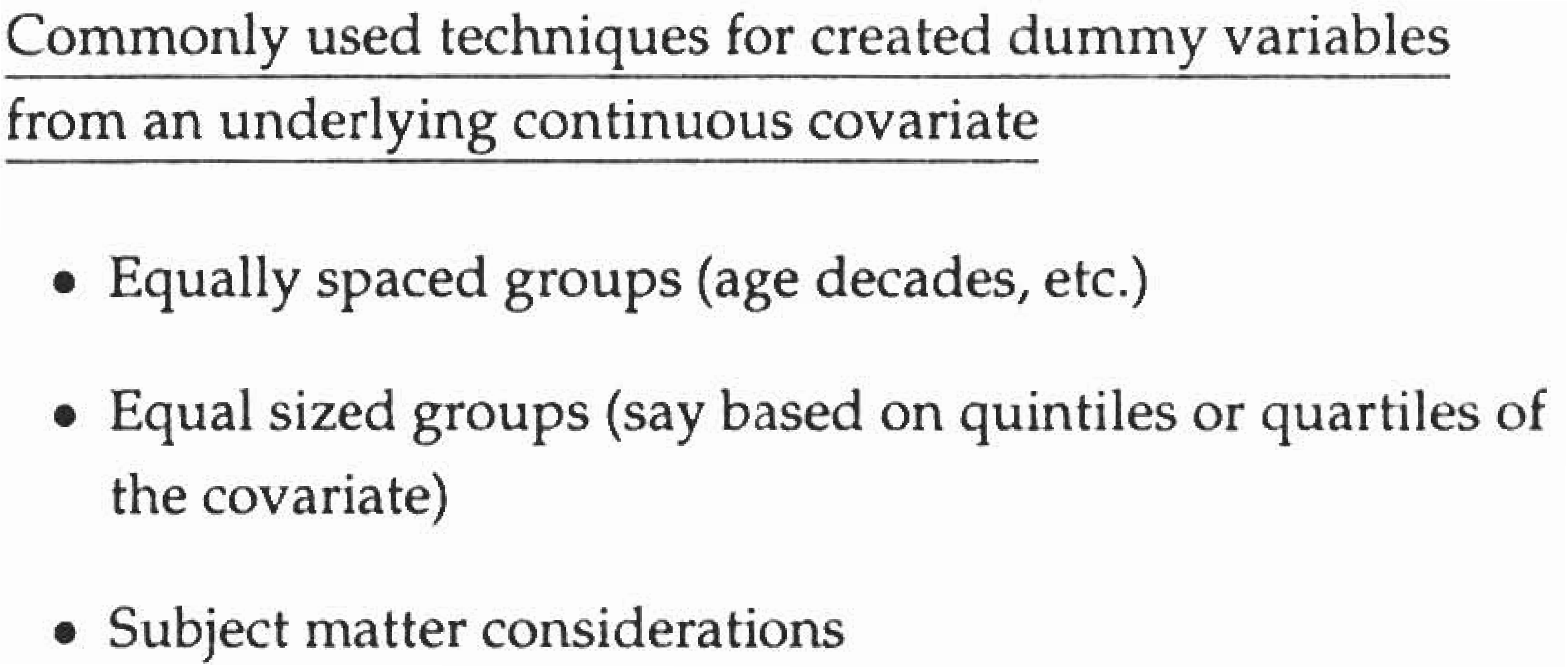
From the STATA output, the estimated odds ratio of seizures comparing high age to low age (the baseline group) is:

$$\exp(2.497) = 12.14 \left(= \frac{.2222/(1 - .2222)}{.0229/(1 - .0229)}\right)$$

The estimated odds ratio of seizures comparing high age to medium age is:

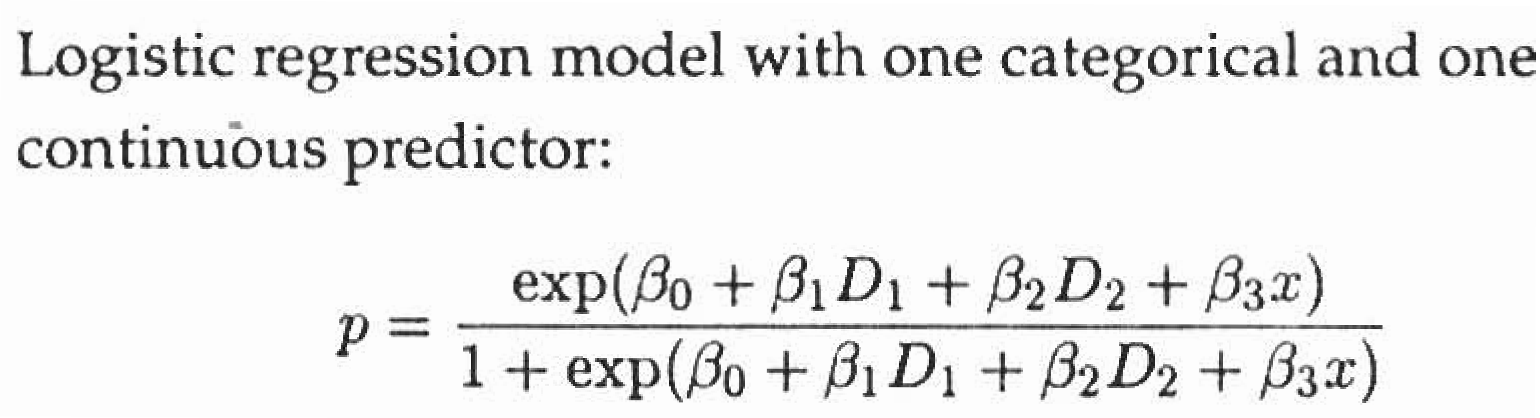
$$\exp(2.497 - 1.265) = 3.43 \left(= \frac{.2222/(1 - .2222)}{.0769/(1 - .0769)}\right)$$

The LRT comparing the model with the three age groups (having one intercept and two indicator variables) to the model with no covariates (having one intercept, thus one parameter in total) is given by $\chi_2^2 = 8.14$ (2 degrees of freedom) giving $p = 0.017$.



Commonly used techniques for created dummy variables from an underlying continuous covariate

- Equally spaced groups (age decades, etc.)
- Equal sized groups (say based on quintiles or quartiles of the covariate)
- Subject matter considerations



Logistic regression model with one categorical and one continuous predictor:

$$p = \frac{\exp(\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 x)}{1 + \exp(\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 x)}$$

If we add interaction terms:

$$p = \frac{\exp(\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 x + \beta_4 \cdot D_1 \cdot x + \beta_5 \cdot D_2 \cdot x)}{1 + \exp(\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 x + \beta_4 \cdot D_1 \cdot x + \beta_5 \cdot D_2 \cdot x)}$$

Model Building Strategies

Building multivariate logistic regression models can become a bewildering experience. There can be many covariates and interactions to consider, categorical vs. continuous covariates, data transformations, possible confounding or effect modifying factors, etc. I find it useful to work up hierarchically, looking at increasingly more complex structures of nested models. Subject matter knowledge, common sense, and statistical hypothesis tests help decide which variables are important in predicting response.

Some steps that may be helpful

- Begin with a careful univariate analysis of each variable as a screening tool.

- Run a multivariate model including all variables thought to possibly be important from the univariate analyses (Hosmer and Lemeshow recommend including any variable having $p < .25$ in a univariate analysis) and other variables known to be important (treatment or exposure variables, possible or known confounding variables, etc.)

- Consider the importance of each variable in the multivariate model, and look to see whether some variables could be deleted or if others need to be added.

- Once you feel "close" to a final model, look more carefully for possible interactions, recoding of variables that might be helpful, addition of "quadratic terms" or other transformations, etc.

- Use caution, logic, good sense, and biologic plausibility when using model building techniques. Also, have fun and be creative! There is as much an "art" to model building as "science."

Note: There may be more than one "final model." In complex data sets we may present the results of several related models, or select between models using subject matter considerations.

Note: "Statistical significance" is not the only reason to keep a variable in the model. Always include explanatory variables considered "essential" by subject matter considerations. If a variable is thought to be a confounding factor for the association between exposure and disease, leave it in!

A Step-Up Procedure (Forward Selection)

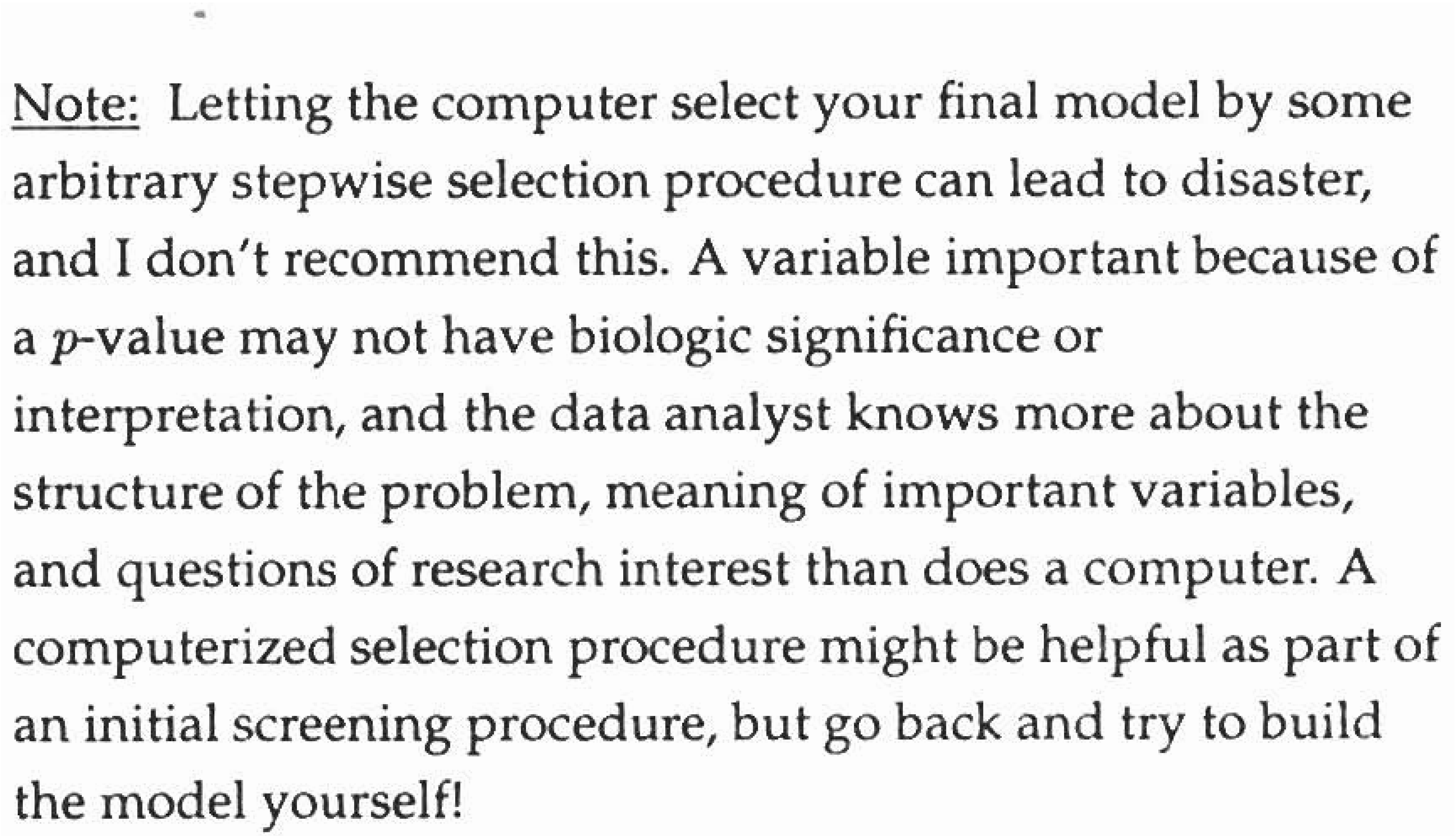
1. Fit the intercept only model, or some other relatively simple model that includes important covariates.

2. Fit all models that add an additional variable (or interaction) to the model. Choose the model with the best fit out of these. If this new model fits significantly better, keep this variable in. If not, you might stop with the current model.

3. Iterate (repeat) these steps until you want to stop.

Note: The specifics of determining how much improvement a new variable adds can be changed depending on circumstances, and can depend on statistical significance, possible confounding bias, or subject matter considerations.

Note: A step-down procedure (backward elimination) starts with a very complex model and then tries to delete variables that help the least. Often people use a hybrid method, trying to step up or down in tandem at each step.



Note: Letting the computer select your final model by some arbitrary stepwise selection procedure can lead to disaster, and I don't recommend this. A variable important because of a *p*-value may not have biologic significance or interpretation, and the data analyst knows more about the structure of the problem, meaning of important variables, and questions of research interest than does a computer. A computerized selection procedure might be helpful as part of an initial screening procedure, but go back and try to build the model yourself!

Goodness-of-Fit Testing

It can be helpful to consider the goodness-of-fit of a particular model you have selected. Goodness-of-fit tests cannot choose between two competing models, and it is possible for several different models to have reasonable goodness-of-fit. In such situations, one must use logic and subject matter considerations to choose between models.

Hosmer and Lemeshow developed a goodness-of-fit test that generally works well when there are a lot of different covariate patterns (e.g., as when you have continuous covariates, etc.). The test is based on grouping observations together that have similar $\hat{p}$ estimates, and is implemented in STATA.

Logistic Regression for Retrospective (Case-control) Studies:

So far we have been modeling the probability of "success" (disease, death, outcome) as a function of other variables (the covariates). In a case-control study, the researcher chooses the number of subjects with disease (cases) and the number without disease (controls), so strictly speaking we cannot model the probability of disease because we selected our sample differentially by disease status (oversampling cases).

However, the relationship between disease and other variables in a logistic regression model remain unaltered except for the "constant" term when we use a case-control design. Thus, logistic regression can still be used to estimate the odds ratios between disease and exposures.

Extensions:

- Conditional logistic regression is an extension that can be used when the data are "matched" on an individual basis or in small groups, when it would not be feasible to add enough parameters to adjust for the matching.

- Polychotomous logistic regression is an extension used to model polychotomous outcome variables that take on more than two categories (e.g., 0, 1, 2, or 0, 1, 2, 3, etc.). There are models appropriate for both unordered or ordered outcomes (e.g., health status coded as excellent, good, fair, or poor).