• We will use data from a study of risk factors for low infant birth weight (i.e. <2500 grams) to motivate our discussion.

• The data come from 189 births at Baystate Medical Center in Springfield, Mass. during 1986 (Hosmer and Lemeshow, 1989 Appendix I).

• Our goal will be to answer the following questions:
  1. Are mother’s who smoke during pregnancy at higher risk for having a low weight baby?
  2. Are teenage mother’s or older mother’s (≥35 years of age) at higher risk for having a low weight baby than other mothers?
  3. Are the number of physician visits during the first trimester associated with an increased risk for having a low weight baby?

---

### Variables in the Low Birth Weight Data Set

Data are Listed in Appendix I  
Hosmer and Lemeshow (1989)  
Applied Logistic Regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Code</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification code</td>
<td>4-226</td>
<td>ID</td>
</tr>
<tr>
<td>Low birth weight</td>
<td>0 if ≥2500g, 1 if &lt;2500g</td>
<td>LOW</td>
</tr>
<tr>
<td>Age</td>
<td>Years</td>
<td>AGE</td>
</tr>
<tr>
<td>Weight at last menstrual period</td>
<td>Pounds</td>
<td>LWT</td>
</tr>
<tr>
<td>Race</td>
<td>1=White, 2=Black, 3=Other</td>
<td>RACE</td>
</tr>
</tbody>
</table>
Variables in the Low Birth Weight Data Set
Data are Listed in Appendix I
Hosmer and Lemeshow (1989)
Applied Logistic Regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Code</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking status during pregnancy</td>
<td>1=Yes</td>
<td>SMOKE</td>
</tr>
<tr>
<td></td>
<td>0=No</td>
<td></td>
</tr>
<tr>
<td>History of Premature Labor</td>
<td>0=None,</td>
<td>PTL</td>
</tr>
<tr>
<td></td>
<td>1=One,</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of Hypertension</td>
<td>1=Yes,</td>
<td>HT</td>
</tr>
<tr>
<td></td>
<td>0=No</td>
<td></td>
</tr>
<tr>
<td>Presence of Uterine Irritability</td>
<td>1=Yes,</td>
<td>UI</td>
</tr>
<tr>
<td></td>
<td>0=No</td>
<td></td>
</tr>
<tr>
<td>Number of physician visits during the</td>
<td>0=None</td>
<td>FTV</td>
</tr>
<tr>
<td>first trimester</td>
<td>1=One,</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Which of the following study designs best describes the Low Birth Weight Study?

- Cross-sectional
- Case-control
- Prospective cohort
- Retrospective cohort
- Randomized trial

What effect, if any, does the choice of study design have on data analysis and/or on the interpretation of the results?

We will begin our analysis by considering the proportion of women who have had a low weight baby. 

Cost of dichotomization (Cohen [Applied Psychological Measurement 1983])?
data x1; infile "c:\Chl5407h\data\bwt\bwt.dat";
input id low age lwt race smoke ptl ht ui f tv bwt;
;
proc freq; tables low;

<table>
<thead>
<tr>
<th>LOW</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>130</td>
<td>68.8</td>
<td>130</td>
<td>68.8</td>
</tr>
<tr>
<td>1</td>
<td>59</td>
<td>31.2</td>
<td>189</td>
<td>100.0</td>
</tr>
</tbody>
</table>

This is considerably higher than the 1986 U.S. national rate of 7% (Health United States 1995).

Is this difference statistically significant?

What is the 95% confidence interval for the risk of having a low weight baby according to our data?

What sampling distribution best describes the number of low weight babies from the 189 births?

The distribution of the random variable $Y$ (=number of low weight babies) follows the binomial distribution with parameters $n$ and $\pi$ when...

1. There are a fixed number, $n$, of observations.
2. The $n$ observations are all independent.
3. Each observation falls into one of two categories (e.g. low birth weight vs. normal birth weight).
4. The probability of having a low weight baby is constant and equal to $\pi$.

Are all of these assumptions likely to hold?
Binomial Distribution  
(Agresti, 1996 Section 1.2.2) 

Let the random variable $Y$ be binomially distributed. Then...

$$Pr(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

where $y=0,...,n$.

The term

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

$(n! = n(n-1)...2(1))$ denotes the number of ways of partitioning $n$ objects into two groups; one group of size $y$ and the other of size $(n-y)$ when order is unimportant.

Heuristic Justification

Suppose that $y=0$...
Suppose that $y=1$...
Suppose that $y=2$...

How can we derive the mean and the variance for $Y$, a binomially distributed random variable?

Let $Y = \sum_{i=1}^{n} Y_i$ where $Y_i = 1$ for low weight babies and $Y_i = 0$ otherwise.

Then each of the "iid" random variables $Y_i$ come from a Bernoulli distribution with probability given by...

$$Pr(Y_i = y_i) = \pi^{y_i}(1-\pi)^{1-y_i}$$ so that

$$Pr(Y_i = 0) = 1 - \pi, \hspace{1em} \text{and}$$

$$Pr(Y_i = 1) = \pi.$$
Mean

\[
E(Y_i) = \sum_{y_i=0}^{1} y_i \, Pr(Y_i = y_i)
\]
\[
= 0 \times Pr(Y_i = 0) + 1 \times Pr(Y_i = 1)
\]
\[
= 0 \times (1 - \pi) + 1 \times \pi
\]
\[
= \pi
\]

\[
E(Y) = E(Y_1 + ... + Y_n)
\]
\[
= E(Y_1) + ... + E(Y_n)
\]
\[
= n\pi
\]

Variance

\[
Var(Y_i) = E(Y_i - E(Y_i))^2
\]
\[
= (0 - \pi)^2 \times Pr(Y_i = 0) + (1 - \pi)^2 \times Pr(Y_i = 1)
\]
\[
= \pi^2(1 - \pi) + (1 - \pi)^2\pi
\]
\[
= \pi(1 - \pi)
\]

\[
Var(Y) = Var(Y_1 + ... + Y_n)
\]
\[
= Var(Y_1) + ... + Var(Y_n)
\]
\[
= n\pi(1 - \pi)
\]

Note that here we make use of the independence assumption.
Notes

- The expectation of a sum is the sum of the expectations.

- The variance of a sum is the sum of the variances, if the observations are independent.

- Suppose that the independent random variables

  \[ X \sim \text{Bin}(n, \pi) \quad \text{and} \quad Z \sim \text{Bin}(m, \pi) \]

  What is the distribution of \( W = X + Z \)?

- What does a binomial distribution look like?

\[ Y \sim \text{Bin}(n = 10, \pi = 0.1) \]
\[ Y \sim \text{Bin}(n = 10, \pi = 0.5) \]

\[ Y \sim \text{Bin}(n = 189, \pi = 0.312) \]
Central Limit Theorem

The central limit theorem states that standardized sums or averages of independent random variables are asymptotically standard normal so long as the random variables come from a distribution with a finite variance.

Thus asymptotically...
\[
\frac{Y - E(Y)}{SE(Y)} \sim N(0, 1)
\]

Similarly for large \( n \) and \( \hat{\pi} = Y/n \)...
\[
\frac{\hat{\pi} - E[\hat{\pi}]}{SE(\hat{\pi})} \sim N(0, 1)
\]

Large Sample Inference: A Single Proportion

\( H_O : \pi = 0.07 \) vs. \( H_A : \pi \neq 0.07 \)

\[
Z = \frac{\hat{\pi} - 0.07}{\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}}
\]

\[
= \frac{0.3122 - 0.07}{\sqrt{0.3122(1 - 0.3122)/189}}
\]

\[
= \frac{0.3122 - 0.07}{0.0337} = 7.18
\]

The associated two-tailed P-value is given by...

\[
P = 2 \times Pr(Z \geq 7.18) < 0.001
\]

where \( Z \) is a standard normal random variable.

What can we conclude?
Notes

• Was it reasonable to perform a two-sided hypothesis test?

• This statistical inference is approximate. Which means?

• Exact vs. approximate statistical inferences

• The approximate one sample hypothesis test on proportions is generally considered valid so long as both \( n\pi_0 > 5 \) and \( n(1 - \pi_0) > 5 \) (Agresti, 1996 page 12).

• The test statistic we derived is identical to a Wald test constructed according to likelihood theory.

• What might be the consequences of replacing the estimated variance of \( \hat{\pi} \) by the \( \text{Var}(\hat{\pi}) \) obtained assuming the null hypothesis to be true?

Confidence Interval Construction

The 95% confidence interval for \( \pi \), the true risk of having a low weight baby is given by...

\[
(\hat{\pi} - 1.96\hat{SE}(\hat{\pi}), \hat{\pi} + 1.96\hat{SE}(\hat{\pi}))
\]

\[
(0.3122 - 1.96(0.0337), 0.3122 + 1.96(0.0337))
\]

\[
(0.25, 0.38)
\]

where \( \hat{SE}(\hat{\pi}) = \sqrt{\hat{\pi}(1 - \hat{\pi})}/n \).

• This often appears in the medical literature in statements such as “We are therefore 95% confident that the true risk of having a low weight baby is between 25% and 38%.”

• By which we mean that the confidence limits were calculated by a method which includes \( \pi \), 95% of the time.

• However, we have data from only a single sample.
A 95% confidence interval for \( \pi \) is given by

\[
(\hat{\pi} - 1.96\sqrt{\hat{\pi}(1 - \hat{\pi})/n}, \hat{\pi} + 1.96\sqrt{\hat{\pi}(1 - \hat{\pi})/n})
\]

Questions

- What is the effect of increasing sample size on the width of a confidence interval?
- Are 95% confidence intervals wider or narrower than 99% confidence intervals?
- What effect does \( \hat{\pi} \) have on the width of a confidence interval for \( \pi \)?

Confidence Intervals and Test Statistics

- Confidence intervals may be constructed by ‘inverting’ a test statistic.
- The associated confidence interval will consequently provide identical statistical inferences to the test statistic from which it was derived.
- An approximate 95% confidence interval for \( \pi \) is found by solving the equation given by...

\[
0.95 = Pr\left(-1.96 \leq \frac{\hat{\pi} - \pi}{SE(\hat{\pi})} \leq 1.96\right)
\]

\[
= P(\hat{\pi} - 1.96SE(\hat{\pi}) \leq \pi \leq \hat{\pi} + 1.96SE(\hat{\pi}))
\]

where \( SE(\hat{\pi}) = \sqrt{\hat{\pi}(1 - \hat{\pi})/n} \).
Is maternal smoking associated with low birth weight?

```sas
options ls=80;
data x1; infile "c:\Chl5407h\data\bwt\bwt.dat";
input id low age lwt race smoke ptl ht ui ftv bwt;
;
proc format; value lwfmt 1='low'
    0='normal';

proc sort; by smoke;

proc freq order=formatted;
tables low / binomial(p=0.07);
format low lwfmt.; by smoke; run;

<table>
<thead>
<tr>
<th>SMOKE</th>
<th>N</th>
<th>Y</th>
<th>PIHAT</th>
<th>ASE</th>
<th>L95</th>
<th>U95</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>115</td>
<td>29</td>
<td>0.25217</td>
<td>0.040495</td>
<td>0.17280</td>
<td>0.33154</td>
</tr>
<tr>
<td>1</td>
<td>74</td>
<td>30</td>
<td>0.40541</td>
<td>0.057074</td>
<td>0.29354</td>
<td>0.51727</td>
</tr>
</tbody>
</table>

Note that the 95% confidence intervals overlap.

---------------------------- smoke=0 ----------------------------

<table>
<thead>
<tr>
<th>low</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>29</td>
<td>25.22</td>
<td>29</td>
<td>25.22</td>
</tr>
<tr>
<td>normal</td>
<td>86</td>
<td>74.78</td>
<td>115</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Binomial Proportion
for low = low

Proportion 0.2522
ASE 0.0405
95% Lower Conf Limit 0.1728
95% Upper Conf Limit 0.3315

Exact Conf Limits
95% Lower Conf Limit 0.1758
95% Upper Conf Limit 0.3417

Test of H0: Proportion = 0.07
ASE under H0 0.0238
Z 7.6568
One-sided Pr > Z <.0001
Two-sided Pr > |Z| <.0001

```

Note that the 95% confidence intervals overlap.
Large Sample Inference:  
Two Independent Proportions

Let...

\[ Y_0 = 29 \] denote the number of low weight babies born to mothers who did not smoke while they were pregnant,

\[ Y_1 = 30 \] denote the number of low weight babies born to mothers who smoked while they were pregnant,

\[ n_0 = 115 \] denote the number of mothers who did not smoke while they were pregnant,

\[ n_1 = 74 \] denote the number of mothers who smoked while they were pregnant,

\[ \hat{\pi}_0 = Y_0/n_0 = 29/115 = 0.25 \] denote the observed risk of having a low weight baby for mothers who did not smoke while they were pregnant.

\[ \hat{\pi}_1 = Y_1/n_1 = 30/74 = 0.41 \] denote the observed risk of having a low weight baby for mothers who smoked while they were pregnant.

\[ Y = \sum_{i=0}^{1} Y_i = 59 \]

\[ n = \sum_{i=0}^{1} n_i = 189 \]

Two × Two Contingency Tables

<table>
<thead>
<tr>
<th>Smoke</th>
<th>Birth Weight</th>
<th>No</th>
<th>Yes</th>
<th>n – Y = 130</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( n_0 - Y_0 = 86 )</td>
<td>( n_1 - Y_1 = 44 )</td>
<td>( n - Y = 130 )</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>( Y_0 = 29 )</td>
<td>( Y_1 = 30 )</td>
<td>( Y = 59 )</td>
<td></td>
</tr>
<tr>
<td>( n_0 = 115 )</td>
<td>( n_1 = 74 )</td>
<td>( n = 189 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
options ls=76;
data x1; infile "c:\Chl5407h\data\bwt\bwt.dat";
input id low age lwt race smoke pttl ht ui ftv bwt;
;
proc freq; tables low*smoke;

TABLE OF LOW BY SMOKER

<table>
<thead>
<tr>
<th>LOW</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>86</td>
<td>44</td>
<td>130</td>
</tr>
<tr>
<td>1</td>
<td>29</td>
<td>30</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>115</td>
<td>74</td>
<td>189</td>
</tr>
</tbody>
</table>

Hypothesis Test for the Risk Difference

\[ H_O : \delta = \pi_1 - \pi_0 = 0 \quad H_A : \delta = \pi_1 - \pi_0 \neq 0 \]

(Agresti 1996 Section 2.2)

Which of the two test statistics are appropriate?

\[ Z = \frac{\hat{\pi}_1 - \hat{\pi}_0}{\sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0} + \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}}} \]

\[ Z = \frac{\hat{\pi}_1 - \hat{\pi}_0}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}} \]

where \( \hat{\pi} = \frac{Y}{n} \)

What distinguishes these two test statistics?
What if we wanted to test the null hypothesis that \( \delta = 0.2 \)?
Hypothesis Test for the Risk Difference

\( H_O : \delta = \pi_1 - \pi_0 = 0 \quad H_A : \delta = \pi_1 - \pi_0 \neq 0 \)

\[
Z = \frac{\hat{\pi}_1 - \hat{\pi}_0}{\sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0} + \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}}}
\]

\[
= \frac{0.41 - 0.25}{\sqrt{[0.25(1-0.25)/115] + [0.41(1-0.41)/74]}}
\]

\[= 2.19 \quad (P = 0.0285)\]

Confidence Interval for the Risk Difference

Confidence intervals for the risk difference are usually given by...

\[
\left( (\hat{\pi}_1 - \hat{\pi}_0) - z_\alpha \sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0} + \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}} \right)
\]

\[
(\hat{\pi}_1 - \hat{\pi}_0) + z_\alpha \sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0} + \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}}
\]

A 95% confidence interval for the risk difference is given by...

\[
\left( (0.41 - 0.25) - 1.96\sqrt{0.0048973}, \right)
\]

\[
(0.41 - 0.25) + 1.96\sqrt{0.0048973} \quad \}
\]

\[
(0.02, 0.29)
\]

What can we conclude?
Determination of Power and Sample Size:
Inferences on the Risk Difference
for Two Independent Samples
Two Sided Test

Power = \( Pr(\text{reject } H_0 \mid H_A \text{ true}) \)
= \( 1 - \beta \) Error

If under \( H_A : \delta = \pi_1 - \pi_0 > 0 \) ...

Power = \( Pr \left( \left| \frac{\hat{\pi}_1 - \hat{\pi}_0}{SE(\hat{\pi}_1 - \hat{\pi}_0)} \right| \geq z_{\alpha/2} \right) \)

\[ \approx \frac{Pr \left( Z \leq -z_{\alpha/2} + \frac{\delta}{SE(\hat{\pi}_1 - \hat{\pi}_0)} \right)}{\sqrt{\pi_0(1-\pi_0) + \pi_1(1-\pi_1)/n}} \]
\[ + Pr \left( Z \leq -z_{\alpha/2} - \frac{\delta}{SE(\hat{\pi}_1 - \hat{\pi}_0)} \right) \]

\[ \approx Pr \left( Z \leq -z_{\alpha/2} + \frac{\delta}{SE(\hat{\pi}_1 - \hat{\pi}_0)} \right) \]

= \( Pr(Z \leq z_\beta) \) where

\[ SE(\hat{\pi}_1 - \hat{\pi}_0) = \sqrt{\frac{\pi_0(1-\pi_0) + \pi_1(1-\pi_1)}{n}} \]

This formula is easily extended to allow for unequal \( n \) per sample.

A standard sample size formula (Agresti, 1996 page 130) can be obtained by solving the following equation for \( n \), the number of subjects per sample...

\[ z_\beta = -z_{\alpha/2} + \frac{\delta}{\sqrt{\pi_0(1-\pi_0) + \pi_1(1-\pi_1)/n}} \]
\[ z_{\alpha/2} + z_\beta = \frac{\delta}{\sqrt{\pi_0(1-\pi_0) + \pi_1(1-\pi_1)/n}} \]

\[ n = \frac{(z_\beta + z_\beta)^2 \left[ \pi_0(1-\pi_0) + \pi_1(1-\pi_1) \right]}{(\pi_1 - \pi_0)^2} \]

Comparisons between this sample size formula and alternative approaches are provided by...

Example: Prospective Cohort Study of Second Hand Smoking on Infant Birth Weight
You are asked to design a prospective cohort study comparing the risk that nonsmoking mothers will have a low weight baby depending on the father’s smoking status.
How many families need to be enrolled in this study?

Suppose that...

1. The risk of having a low weight baby is approximately equal to 10% for families in which neither parent smokes.
2. We would like to be able to detect a risk difference of $\pi_1 - \pi_0 = 0.15$.
3. We plan on using a two-tailed hypothesis test with a Type I error rate of 5% and 80% power.

$$n = \frac{(z_{0.025} + z_{0.2})^2[\pi_0(1 - \pi_0) + \pi_1(1 - \pi_1)]}{(\pi_1 - \pi_0)^2}$$
$$= \frac{(1.96 + 0.841)^2[0.10(0.90) + 0.25(0.75)]}{(0.25 - 0.10)^2} = 96.8$$

Therefore we need to enroll at least 100 families in which neither parent smokes and 100 families in which only the father smokes.

It is always a good idea to consider the effect of a range of plausible alternative smoking prevalence values on the required sample size as part of a sensitivity analysis.

Further adjustments to the sample size may be needed allowing for missing data, and planned multivariable analyses.
What is the effect on the required sample size of...

1. increasing the specified treatment effect (i.e. increasing $\pi_1 - \pi_0$)?

2. increasing the Type I ($\alpha$) error rate?

3. increasing the Type II ($\beta$) rate?

4. studying a rare vs. a common disease?

A Review of Statistical Power in Randomized Clinical Trials
Moher et al. [JAMA 1994;272:122-124]

- All too often clinical trials are underpowered.

- Moher et al. (1994), for example, reviewed 37 reports of randomized trials having negative results published in JAMA, Lancet, or NEJM in 1990.

- Only 43% (16/37) of these trials justified the sample size.

- Only 42% (10/24) had sufficient power to detect a relative difference of 50%.
Why are sample sizes for clinical trials often too small?

1. Cost.

2. Overestimate effectiveness of new interventions.

3. Omit consideration for an adequate sensitivity analysis.

4. Control groups do better than expected.
   - Inclusion/exclusion criteria exclude high risk subjects.
   - Control subjects receive more attention than might normally be given thus improving prognosis.
   - Temporal trends.

5. Underestimate loss to follow-up and non-compliance.

6. Overestimate instrument precision.

Inferences on the Odds Ratio for Two Independent Samples
(Agresti 1996 Section 2.3)

The odds of having a low weight baby is equal to...

\[
\frac{Y_1/(n_1 - Y_1)}{Y_0/(n_0 - Y_0)} = 30/44 \text{ for smokers.}
\]

\[
\frac{Y_0/(n_0 - Y_0)}{Y_0/(n_0 - Y_0)} = 29/86 \text{ for nonsmokers.}
\]

Therefore the odds ratio is given by...

\[
\hat{\psi} = \frac{Y_1/(n_1 - Y_1)}{Y_0/(n_0 - Y_0)}
\]

\[
= \frac{[Y_1/n_1]/[(n_1 - Y_1)/n_1]}{[Y_0/n_0]/[(n_0 - Y_0)/n_0]}
\]

\[
= \frac{\hat{\pi}_1/(1 - \hat{\pi}_1)}{\hat{\pi}_0/(1 - \hat{\pi}_0)}
\]

\[
= \frac{30/44}{29/86} = 2.02
\]

Thus the estimated odds ratio of having a low weight baby is 2.02 comparing mothers who smoke to mothers who do not smoke.
Odds Ratio and Risk Difference

- If the odds ratio is less than one the risk difference is...
- If the odds ratio is equal to one the risk difference is...
- If the odds ratio is greater than one the risk difference is...

Sampling Distribution of the Odds Ratio and of the log(Odds Ratio)

- The central limit theorem ensures that the sampling distribution of the estimated odds ratio is approximately normally distributed for a sufficiently large sample size.
- In practice sample sizes are not usually sufficiently large to satisfy central limit theorem requirements ensuring that the sampling distribution of the estimated odds ratio is approximately normal.
- Rather the sampling distribution of the odds ratio tends to be positively skewed.
- The sampling distribution of the log(odds ratio) is more likely to be normally distributed for samples of moderate size.
The log(Odds Ratio)

\[
\log \hat{\psi} = \log \left( \frac{\hat{\pi}_1/(1 - \hat{\pi}_1)}{\hat{\pi}_0/(1 - \hat{\pi}_0)} \right)
\]

\[
= \log \left( \frac{\hat{\pi}_1}{1 - \hat{\pi}_1} \right) - \log \left( \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} \right)
\]

\[
= \logit(\hat{\pi}_1) - \logit(\hat{\pi}_0)
\]

Therefore the log(odds ratio) can be expressed as a difference in “logits.”

Since all observations are assumed to be independent...

\[
\text{Var}(\log \hat{\psi}) = \text{Var} [\logit(\hat{\pi}_1) - \logit(\hat{\pi}_0)]
\]

\[
= \text{Var} [\logit(\hat{\pi}_1)] + \text{Var} [\logit(\hat{\pi}_0)]
\]

\[
= \text{Var} [\logit(\hat{\pi}_0)] + \text{Var} [\logit(\hat{\pi}_1)]
\]

\[
\approx \frac{1}{n_0 \hat{\pi}_0 (1 - \hat{\pi}_0)} + \frac{1}{n_1 \hat{\pi}_1 (1 - \hat{\pi}_1)}
\]

\[
= \frac{1}{n_0 \hat{\pi}_0} + \frac{1}{n_0 (1 - \hat{\pi}_0)} + \frac{1}{n_1 \hat{\pi}_1} + \frac{1}{n_1 (1 - \hat{\pi}_1)}
\]

The estimated variance of \(\log(\hat{\psi})\) is then given by...

\[
\text{Var}[\log(\hat{\psi})] = \frac{1}{n_0 \hat{\pi}_0 (1 - \hat{\pi}_0)} + \frac{1}{n_1 \hat{\pi}_1 (1 - \hat{\pi}_1)}
\]

\[
= \frac{1}{n_0 \hat{\pi}_0} + \frac{1}{n_0 (1 - \hat{\pi}_0)} + \frac{1}{n_1 \hat{\pi}_1} + \frac{1}{n_1 (1 - \hat{\pi}_1)}
\]

\[
= \frac{1}{Y_0} + \frac{1}{n_0 - Y_0} + \frac{1}{n_1} + \frac{1}{Y_1 - n_1 Y_1}
\]

\[
= \frac{1}{29} + \frac{1}{86} + \frac{1}{30} + \frac{1}{44} = 0.1022
\]

Sampling Distribution

for the Estimated Log Odds Ratio

\[
\log(\hat{\psi}) = \logit(\hat{\pi}_1) - \logit(\hat{\pi}_0)
\]

\[
\sim N \left( \log(\psi), \text{Var}[\log(\hat{\psi})] \right)
\]

\[
\text{Var}[\log(\hat{\psi})] = \frac{1}{n_0 \hat{\pi}_0 (1 - \hat{\pi}_0)} + \frac{1}{n_1 \hat{\pi}_1 (1 - \hat{\pi}_1)}
\]
Two-Tailed Hypothesis Test

\[ H_O : \psi = 1 \quad H_A : \psi \neq 1 \]
\[ H_O : \log(\psi) = 0 \quad H_A : \log(\psi) \neq 0 \]

\[ Z = \frac{\log(\hat{\psi}) - 0}{SE(\log(\hat{\psi}))} \]
\[ = \frac{\log(2.0219)}{\sqrt{0.1022}} \]
\[ = 2.2027 \quad (P = 0.0276) \]

Conclusion?

Confidence Interval for an Odds Ratio

A 95% confidence interval for the log odds ratio is given by...

\[ \left( \log(\hat{\psi}) - 1.96SE(\log(\hat{\psi})) , \log(\hat{\psi}) + 1.96SE(\log(\hat{\psi})) \right) \]

\[ \left( \log(2.0219) - 1.96\sqrt{0.1022} , \log(2.0219) + 1.96\sqrt{0.1022} \right) \]

\[ (0.077560 , 1.33056) \]

Transforming back to the odds ratio scale...

\[ (\exp(0.077560) , \exp(1.33056)) \]

\[ (1.08 , 3.78) \]

Conclusion?

Will inferences on odds ratios and risk differences always be identical?
data x1; infile "c:\nklar\Ch15407h\data\bwt\bwt.dat";
   input id low age lwt race smoke ptyl ht ui ftv bwt;
;
proc freq; tables low*smoke / relrisk;
run;

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>86</td>
<td>44</td>
<td>130</td>
</tr>
<tr>
<td>1</td>
<td>29</td>
<td>30</td>
<td>59</td>
</tr>
<tr>
<td>Total</td>
<td>115</td>
<td>74</td>
<td>189</td>
</tr>
</tbody>
</table>

(Asymptotic) 95% Confidence Limits
Case-Cntrl 2.0219 1.0807 3.7831
(Odds Ratio)

Does SAS provide point estimates and statistical inferences for risk differences or relative risks?