SAS for Data Management, Analysis, and Reporting

Lecture 1
June 16, 2004

Kate Cowles
374 SH
kcowles@stat.uiowa.edu

Critical features of the work of applied statisticians

• goal is to contribute to knowledge in the subject area – to get right answers and contribute to right decisions
• all aspects of experimental design and statistical analysis must be correct
  – appropriate sampling scheme
  – data values must be correct
  – appropriate statistical analysis
  – correct computer code
  – clear reporting and interpretation of statistical results
• most important question always is: Does this make sense?

Goals of the course

• to provide experience with skills required by
  – M.S.-level statisticians
  – Ph.D. statisticians in industry, government, and consulting
• to provide experience with aspects of applied statistics that may not be covered in other courses

• SAS is the statistical software package most commonly used in business, government, and many other applied research settings
  – has the best data management capabilities
  – can handle huge datasets
  – not as user-friendly or well-designed as R or S-PLUS in some other ways
  – R or S-PLUS are often preferred for research in theoretical and methodological statistics
• preparing data for analysis usually takes more time and work than doing the actual analysis
Aspects emphasized in this course

- data preparation
  - reading data from other sources into SAS
  - checking and validating data
  - merging data from multiple sources
  - reorganizing and reformatting data
- professional-looking graphical and numeric summaries of data and statistical results
- writing and documenting SAS code
- some written interpretation of results
- commonly-used statistical methods that you may not have seen elsewhere

Usual weekly schedule for class

- Mon. and Wed: lecture in classroom
- Tues. and Thurs.: instructed lab in computer lab
- Fri: open lab time for you to work on homework and projects with me present to help and answer questions
  - some weeks, including this week, half of the class period on Fri. will be lecture and the other half open lab
- exception: next week only, open lab will be on Thurs. 6/24 and instructed lab will be on Fri. 6/25
- special guest lecturer next Tues. 6/22: Kellie Poullin

Data checking and validation ("cleaning")

- making sure that raw data were accurately entered into a computer-readable file
- checking that character variables contain only valid values
- checking that numeric values are within predetermined ranges
- checking whether there are missing values for variables where complete data are necessary
- checking for and eliminating duplicate records
- checking for uniqueness of certain values, such as patient IDs
- checking for invalid date values
- checking that an ID number is present in each of several related files
- verifying that more complex multi-file rules have been followed

---

example: if an adverse event of type X occurs in one dataset, you expect an observation with the same ID number in another data set. In addition, the date of this observation must be after the adverse event and before the end of the study.

from Cody’s *Data Cleaning Techniques Using SAS Software* by Ron Cody, SAS Institute, 1999.
Example dataset: Patients.dat

<table>
<thead>
<tr>
<th>Date</th>
<th>Visit</th>
<th>Patho</th>
<th>Gender</th>
<th>HR</th>
<th>SBP</th>
<th>DBP</th>
<th>DX</th>
<th>AE</th>
</tr>
</thead>
<tbody>
<tr>
<td>00311/11/1998</td>
<td>88140</td>
<td>80</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>61</td>
<td>0</td>
</tr>
<tr>
<td>00311/13/1998</td>
<td>84120</td>
<td>78</td>
<td>10</td>
<td>31</td>
<td>05011/1998</td>
<td>40120</td>
<td>10</td>
<td>41</td>
</tr>
<tr>
<td>00311/13/1998</td>
<td>86300</td>
<td>20</td>
<td>41</td>
<td>01210/12/98</td>
<td>60122</td>
<td>74</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>01210/12/98</td>
<td>48144</td>
<td>02</td>
<td>21</td>
<td>02411/01/1999</td>
<td>74100</td>
<td>23</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>02411/01/1999</td>
<td>22130</td>
<td>90</td>
<td>1</td>
<td>02411/13/1998</td>
<td>84120</td>
<td>78</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>00311/12/1999</td>
<td>58112</td>
<td>74</td>
<td>0</td>
<td>0151</td>
<td>82148</td>
<td>88</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>11719/06/1999</td>
<td>58118</td>
<td>70</td>
<td>0</td>
<td>12315/12/1999</td>
<td>60</td>
<td>10</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>00311/09/1999</td>
<td>90400020</td>
<td>90</td>
<td>51</td>
<td>00311/09/1999</td>
<td>10</td>
<td>20</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>02410/10/1999</td>
<td>48141</td>
<td>22</td>
<td>8</td>
<td>02411/01/1999</td>
<td>223478</td>
<td>10</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>02411/01/1999</td>
<td>74102</td>
<td>68</td>
<td>51</td>
<td>02711/08/1997</td>
<td>85100</td>
<td>28</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>02711/08/1997</td>
<td>86190</td>
<td>30</td>
<td>3</td>
<td>02611/15/1999</td>
<td>10</td>
<td>8</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>00607/07/1999</td>
<td>82148</td>
<td>84</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SAS Code to read in the data

```sas
/* Program Name: Patients.sas in C:\Cleaning */
/* Purpose: To create a SAS data set called PATIENTS */
/* Date: May 29, 1998 */
*------------------------------------------------------------------*
OPTIONS P(stop='|----------|----------|----------|----------|Bunch=75, UPDATE;

* LIBNAME CLEAN "C:\Cleaning";

DATA CLEAN.PATIENTS;
DATA PATIENTS;
INFILE "C:temp\patients.dat" PAD; INFILE "ftp://ftp/pub\acowles\datasets\patients.dat" PAD;
INPUT 01 PATHO $3. 04 GENDER $1. 05 VISIT MMDDYY10. 06 HR $3. 07 SBP $3. 08 DBP $3. 09 DX $3. 10 AE $1. ;
LABEL PATHO = "Patient Number" GENDER = "Gender" VISIT = "Visit Date" HR = "Heart Rate" SBP = "Systolic Blood Pressure" DBP = "Diastolic Blood Pressure" DX = "Diagnosis Code" AE = "Adverse Event?";

FORMAT VISIT MMDDYY10.;
RUN;
```

Files on course web page

- data file: patients.dat
- SAS program: patients172.sas
New aspects of this data step

- PAD option on `infile` statement
  - adds blanks to the end of short records to the default logical record length or a length specified by another `infile` option, `lrecl`
  - prevents skipping to the next record of data when a shorter line is encountered

- `%` in `input` statement
  - tell SAS at which numeric column to begin reading each variable
  - needed when there are no delimiters between variable values in data file

- formats after each variable name
  - how many digits or characters in each value
  - identify character variables with `$`
  - `MMDDYY10` is built-in SAS format for reading dates

---

`Proc contents: Getting SAS to describe the contents of a dataset`

```
/*-----------------------------
Extra code: getting a description of the dataset
--------------------------------------------------------------*/
PROC CONTENTS DATA = PATIENTS ;
MOD ;

The SAS System 1

The CONTENTS Procedure

Data Set Name: WERF.PATIENTS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Len</th>
<th>Pos</th>
<th>Format</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 AE</td>
<td>Char</td>
<td>1</td>
<td>39</td>
<td></td>
<td>Adverse Event</td>
</tr>
<tr>
<td>6 DBP</td>
<td>Num</td>
<td>8</td>
<td>24</td>
<td></td>
<td>Diastolic Blood Pressure</td>
</tr>
<tr>
<td>7 DX</td>
<td>Char</td>
<td>3</td>
<td>36</td>
<td></td>
<td>Diagnosis Code</td>
</tr>
<tr>
<td>2 GENDER</td>
<td>Char</td>
<td>1</td>
<td>35</td>
<td></td>
<td>Gender</td>
</tr>
<tr>
<td>4 HR</td>
<td>Num</td>
<td>8</td>
<td>8</td>
<td></td>
<td>Heart Rate</td>
</tr>
<tr>
<td>3 VISIT</td>
<td>Num</td>
<td>8</td>
<td>0</td>
<td>MMDDYY10</td>
<td>Visit Date</td>
</tr>
<tr>
<td>1 PATNO</td>
<td>Char</td>
<td>3</td>
<td>32</td>
<td></td>
<td>Patient Number</td>
</tr>
<tr>
<td>5 SBP</td>
<td>Num</td>
<td>8</td>
<td>16</td>
<td></td>
<td>Systolic Blood Pressure</td>
</tr>
</tbody>
</table>

---Engine/Host Dependent Information----

```

Data Set Page Size: 8192
Number of Data Set Pages: 1
First Data Page: 1
Max Dbs per Page: 203
Dbs in First Data Page: 31
Number of Data Set Repairs: 0
File Name: /usr/tmp/SAS vapkEBBSK00003B05_ mouse/patients.sas7dat
```

---
Validity checks on character variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Valid values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>F, M</td>
</tr>
<tr>
<td>DX</td>
<td>numerals 1 through 999</td>
</tr>
<tr>
<td>AE</td>
<td>0, 1</td>
</tr>
</tbody>
</table>

- Are there other (invalid) values for these variables in the dataset?
- Are there missing values?
- Which observations in the dataset contain invalid or missing values?

Using proc freq to list all distinct values of a character variable that appear in the dataset

```plaintext
/*------------------------------------------*/
Program 1-2  Using PROC FREQ to list all the unique values for character variables

PROC FREQ DATA=CLEAN.PATIENTS;
TITLE "Frequency Counts for Selected Character Variables";
TABLES GENDER DX AE / NODIM NODPERCENT;
RUN;

Frequency Counts for Selected Character Variables

The FREQ Procedure

Gender

<table>
<thead>
<tr>
<th>GENDER</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>12</td>
</tr>
<tr>
<td>M</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>f</td>
<td>2</td>
</tr>
</tbody>
</table>

Frequency Missing = 1
```

Using proc print to list invalid character values and identify the observations

- `where` statement in many procedures will exclude observations that don’t meet a given logical condition
- Simple logical conditions involve comparing the value in a variable to some specified value
- Example: `where hr > 150`
- Example: `where gender not in (’M’ ’F’ )`
The Verify Function

- Verify(character_variable, verify_string)
- returns the first position in the character_variable that is not in the verify_string
- returns 0 if character_variable does not contain any invalid values

Formatting and documenting SAS code

- Make your SAS programs readable and understandable
  - to yourself (now and if you go back to the program 5 years from now!)
  - to your instructor or supervisor
  - to someone who takes over your job when you quit
- Use meaningful variable names
  - gender rather than x1
- Use white space
  - Skip lines between logical blocks of code
  - Indent for understandability
Methods you may not have studied: the Chi-square test for comparing population proportions

Comparing two proportions

Recall: In a two-independent sample problem, we want to compare two populations or the responses to two different treatments using data from two independent samples.

When we are interested in comparing the proportions of successes in two groups, the notation is:

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>proportion</td>
<td>size</td>
<td>proportion</td>
</tr>
<tr>
<td>1</td>
<td>( p_1 )</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( p_2 )</td>
<td>( n_2 )</td>
</tr>
</tbody>
</table>

Example: Do seatbelts affect whether children will survive car accidents?

- study of deaths among children involved in car accidents during an 18-month period
- two simple random samples
  - one sample from population of children who were wearing seatbelts at the time of car accident
  - one sample from population of children who were not wearing seatbelts at the time of car accident
- parameters of interest: proportions of children who die in car accidents from each of these populations

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>proportion</td>
<td>size</td>
<td>proportion</td>
</tr>
<tr>
<td>seatbelts</td>
<td>( p_1 )</td>
<td>123</td>
</tr>
<tr>
<td>no seatbelts</td>
<td>( p_2 )</td>
<td>290</td>
</tr>
</tbody>
</table>

To determine whether the study provides significant evidence that seatbelts affect the proportion of kids who die if they are involved in a car accident, we test the hypotheses:

\[
H_0 : p_1 - p_2 = 0 \quad \text{or} \quad H_0 : p_1 = p_2
\]

\[
H_a : p_1 - p_2 \neq 0 \quad \text{or} \quad H_a : p_1 \neq p_2
\]

To estimate how large the difference is, we compute a confidence interval for the difference \( p_1 - p_2 \).
Contingency Tables and the Chi-square test

A way of comparing two population proportions that generalizes to more than two populations.

Begin by presenting the data as a two-way table, with rows representing levels of one variable and columns representing levels of the other.

Seatbelt example:

<table>
<thead>
<tr>
<th>Seatbelts</th>
<th>Died</th>
<th>Did not die</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>3</td>
<td>120</td>
<td>123</td>
</tr>
<tr>
<td>No</td>
<td>13</td>
<td>277</td>
<td>290</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>397</td>
<td>413</td>
</tr>
</tbody>
</table>

To test the hypotheses

\[ H_0 : p_1 - p_2 = 0 \quad \text{or} \quad H_0 : p_1 = p_2 \]
\[ H_a : p_1 - p_2 \neq 0 \quad \text{or} \quad H_a : p_1 \neq p_2 \]

using the two-way table, we must compute the expected counts. These are the counts we would expect (except for random variation) if \( H_0 \) were true.

\[ \text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}} \]

If \( H_0 \) were true, there would be just one \( p \) shared by both populations.

Our best estimate is the pooled sample proportion

\[ \hat{p}_{\text{pooled}} = \frac{\text{total count of successes in both samples}}{n_1 + n_2} \]

\[ \hat{p}_{\text{pooled}} = 0.039 \]

The Chi-Square Test

- Recall that the expected counts were computed under the assumption that the null hypothesis was true.

- We can test the null hypothesis by determining whether the differences between the observed and expected counts are too large to be likely to be due to chance.

- Notation
  - \( O_i \) is the observed count in cell \( i \)
  - \( E_i \) is the expected count in cell \( i \)
The Chi-square statistic

The statistic that we use for this test is the sum over all the cells in the table of 
\[
(\text{observed count} - \text{expected count})^2
\]
expected count
The formula in mathematical notation is
\[
X^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}
\]
where \( rc \) is the total number of cells in the table,
- \( r \) is the number of rows
- \( c \) is the number of columns

- Think of \( X^2 \) as a measure of the distance of the observed counts from the expected counts.
- Like any distance, \( X^2 \)
  - is always zero or positive
  - is zero only when the observed counts are exactly equal to the expected counts
- Large values of \( X^2 \) are evidence against \( H_0 \).
  - indicate that observed counts are far away from what we would expect if \( H_0 \) were true.
- The Chi-square statistic \( X^2 \) follows a Chi-square distribution (\( X^2 \) distribution) with \((r - 1)(c - 1)\) degrees of freedom.

The Chi-Square test for the car-accidents example

\[
X^2 = \frac{(3 - 4.8)^2}{4.8} + \frac{(120 - 118.2)^2}{118.2} + \frac{(13 - 11.3)^2}{11.3} + \frac{(277 - 278.7)^2}{278.7} = 0.969
\]

Since we have 2 rows and 2 columns in our table, the degrees of freedom is 
\((r - 1)(c - 1) = 1(1) = 1\)

We will carry out our hypothesis test at \( \alpha = .05 \).
According to a Chi-square table, the .05 cutoff under the Chi-square distribution with 1 degree of freedom is 3.84.

The Chi-square test is always 2-sided. For the Chi-square test, we always reject if the test statistic is larger than the cutoff value.

Our computed value, 0.969, is smaller than this cutoff. Therefore we cannot reject \( H_0 \).
When does Chi-square test give accurate enough inference?

- rule of thumb: when expected counts in all cells are $\geq 5$
- not quite satisfied in this example

If these rules of thumb are not satisfied, use Fisher’s exact test instead.

Important point: Chi square test is applicable when the the population parameters of interest are proportions and the two samples are independent – that is, there is no matching or pairing between individual members of one sample and individual members of the other sample.

It would be possible but wrong to display the results as a regular $2 \times 2$ contingency table,

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Survive 5 yrs</th>
<th>Die within 5 yrs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>106</td>
<td>515</td>
<td>621</td>
</tr>
<tr>
<td>B</td>
<td>95</td>
<td>521</td>
<td>621</td>
</tr>
<tr>
<td></td>
<td>201</td>
<td>1041</td>
<td>1242</td>
</tr>
</tbody>
</table>

This is incorrect because it ignores the pairing.

McNemar’s Test

Example problem involving paired data and population proportions

- For a rare form of cancer, we want to compare treatment with chemotherapy vs. treatment with surgery
- We will pair up the subjects in the study
  - Each pair will be matched on age (within 5 years), sex, and clinical condition.
  - A randomly selected member of each pair will be assigned to chemotherapy; the other to surgery.
- Patients will be followed for 5 years.
- The outcome variable is whether they are alive at the end of 5 years (yes/no)
- A total of 1242 patients (621 pairs) are enrolled.

The appropriate table has the pair as the unit counted,

<table>
<thead>
<tr>
<th>Outcome in Trt B</th>
<th>Trt A pt Survive</th>
<th>Die</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survive</td>
<td>90</td>
<td>16</td>
<td>106</td>
</tr>
<tr>
<td>Die</td>
<td>5</td>
<td>510</td>
<td>515</td>
</tr>
<tr>
<td>Total</td>
<td>95</td>
<td>526</td>
<td>621</td>
</tr>
</tbody>
</table>

For this type of table, the research question should not be phrased as whether the proportion of 5-year survivors is the same under both treatments.

We will restate the null hypothesis as

$H_0$: There is no association between treatment type (chemo vs. surgery) and 5-year survival.
McNemar’s test

Concordant and discordant pairs

- concordant pairs: both members of pair had the same outcome
  - 600 of these in example
- discordant pairs: one member had one outcome; other member had other outcome
  - 21 of these in example

McNemar’s test, continued

Then the test statistic for McNemar’s test is

$$X^2 = \frac{(|r - s| - 1)^2}{r + s}$$

In the example, this is

$$X^2 = \frac{(|16 - 5| - 1)^2}{16 + 5}$$

$$= \frac{100}{21}$$

$$= 4.762$$

This test statistic has an approximate Chi-square ($\chi^2$) distribution with 1 degree of freedom.

McNemar’s test, continued

- Concordant pairs are not used in calculating the test statistic for McNemar’s test.
- Further breaking down the discordant pairs, let:
  - $r$ = number of pairs in which A patient survives and B patient dies
    $= 16$
  - $s$ = number of pairs in which B patient survives dies and A patient dies
    $= 5$

The test itself

- To conduct our test at $\alpha = .05$ significance level, we use a Chi-square table to find the .05 cutoff.
  - This is 3.84.
- Since our test statistic value 4.762 > 3.84, we reject $H_0$ at the .05 level.
  - (Recall that for a Chi-square test, we reject if the test statistic value is too large. We would never reject because a Chi-square statistic value was too small.)
  - Since 3.84 < 4.762 < 5.02, we can see that the p-value for the test is
    $$0.025 < p < 0.05.$$ 
- We conclude that there is an association between which drug is used and 5-year survival. 
  Note that SAS uses a different form of the
test statistic for McNemar’s test. SAS does not subtract the 1 in the numerator. This makes little difference when the sample size is very large. For small sample sizes, proc freq can compute the exact p-value if you specify the mcnem option on the exact statement.

Measures of relative probabilities of disease:

The relative risk and the odds ratio

- The relative risk is the ratio of the probabilities of disease (or other health outcome) in two different groups
  - groups defined by whether members have been exposed to a particular risk factor (or protective factor)
  
  \[
  RR = \frac{P(disease \mid exposed)}{P(disease \mid unexposed)}
  \]

- Example: Based on a large study of blood pressure and death from cardiovascular disease in men, it is estimated that the following probabilities hold for deaths during a two-year period
  - \( P(\text{death from c v d} \mid \text{low bp}) = 0.0078 \)
  - \( P(\text{death from c v d} \mid \text{high bp}) = 0.0165 \)

  Then the relative risk of death from c v d for men with high blood pressure compared to men with low blood pressure is

  \[ RR = \frac{0.0165}{0.0078} = 2.1 \]

- The relative risk does not depend on the sizes of probabilities in the two groups
  - makes it useful in studying rare events

- The relative risk is easily interpretable, but can not be estimated from certain kinds of studies,
  - case-control studies of rare diseases
  - so another way of comparing probabilities is needed
  - brings us to odds and odds ratios
Odds at the greyhound track

Odds on a dog are odds in favor of his losing
Odds favoring losing \( \frac{\Pr(\text{lose})}{\Pr(\text{win})} \)

Example: If \( \Pr(\text{OKey Donny loses}) = .6 \)
then \( \text{odds}_{OD} = \frac{.6}{.4} = 1.5 \) (or 3 to 2)

Example: If \( \Pr(\text{My Mickey D loses}) = .95 \)
then \( \text{odds}_{MMD} = \frac{.95}{.05} = 19 \)

Odds ratios

Odds ratio - the ratio of odds in 2 different groups

Example:
Suppose \( \Pr(\text{heart attack in next 12 mos.}) \) is .01 for male smokers

Then odds ratio for heart attacks in next 12 mos. in nonsmoker vs. smoker males is

\[
O.R. = \frac{\Pr(\text{heart attack|non-smoke})}{1 - \Pr(\text{heart attack|non-smoke})} \div \frac{\Pr(\text{heart attack|smoke})}{1 - \Pr(\text{heart attack|smoke})}
\]

\[
= \frac{.005}{.995} \div \frac{.005}{.995} = 0.497
\]

For any binary variable \( Y \),
odds in favor of \( Y = 1 = \frac{\Pr(Y = 1)}{\Pr(Y = 0)} = \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} \)

Example: If \( \Pr(\text{heart attack in next 12 mos.}) \) for male nonsmoker is .005, then odds for heart attack in next 12 mos. for male nonsmoker is \( \frac{.005}{.995} = .00503 \)

Interpretation of odds ratio

- \( OR \geq 0 \)
- if \( OR = 1.0 \), then \( \Pr(Y = 1) \) is the same in both samples
- if \( OR < 1.0 \), then \( \Pr(Y = 1) \) is less in numerator group than in denominator group
- \( OR = 0 \) if and only if \( \Pr(Y = 1) = 0 \) in numerator sample
Odds ratios: measuring the strength of the association between two nominal variables

- Recall that, if the probability of a success is $p$ then
  
  the odds in favor of success = $\frac{p}{1-p}$.

- Consider two populations defined by:
  - members of population 1 have been exposed to a particular risk factor for a disease
  - members of population 2 have not been exposed to the risk factor

- Define $p_1$ and $p_2$ as the probabilities of getting the disease in the two respective populations

- If we compute the odds in favor of success (disease!) for each population, then the ratio of odds or **odds ratio** becomes a useful measure
  - for relating the two probabilities
  - for measuring the strength of the association between exposure and disease

- The odds ratio
  
  $$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

Computing sample odds ratios

- As usual, we will use sample data to estimate the unknown true odds ratio in the populations.

- Laying out the data as a $2 \times 2$ contingency table simplifies computation of sample odds ratios,

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>No disease</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>n=a+b+c+d</td>
</tr>
</tbody>
</table>

- From this table we would estimate:

  $P(disease|exposed) = \frac{a}{a+c}$

  $P(disease|unexposed) = \frac{b}{b+d}$

- Then we estimate the odds in favor of the disease in the respective groups as

  $\hat{O}_{exposed} = \frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \frac{a}{c}$

  $\hat{O}_{unexposed} = \frac{\frac{b}{b+d}}{\frac{d}{b+d}} = \frac{b}{d}$

- So the estimated odds ratio is

  $$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$
Example: prophylactic antibiotics during breast surgery

Platt et al. (New England Journal of Medicine, 1990) observed patients who did receive the antibiotic cefonicid during breast surgery and patients who did not. The outcome of interest was whether or not patients required antibiotic treatment after surgery for any reason.

- The risk factor is not receiving prophylactic antibiotic during surgery.
- The “disease” is requiring postoperative antibiotic treatment.

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>43</td>
<td>26</td>
<td>69</td>
</tr>
<tr>
<td>No disease</td>
<td>260</td>
<td>277</td>
<td>537</td>
</tr>
<tr>
<td>Total</td>
<td>303</td>
<td>303</td>
<td>n = 606</td>
</tr>
</tbody>
</table>

Confidence intervals for odds ratios

- Now we have a point estimate, but we need a confidence interval to indicate the uncertainty in this estimate.
- As usual, this requires knowing something about the sampling distribution of the statistic.
- The sampling distribution of sample odds ratios is skewed to the right, so we cannot use a normal approximation for it.
- However, the natural log of the sample odds ratio has approximately a normal distribution
  - sample size requirement: expected count in each cell of contingency table \( \geq 5 \)

The estimated odds ratio for exposed vs. unexposed patients is

\[
OR = \frac{43 \times 277}{26 \times 260} = 1.76
\]

- The formula for a 95% confidence interval for the natural log of the odds ratio is

\[
\log(OR) \pm 1.96 \text{se}[\log(OR)]
\]

where “se” is an abbreviation for “standard error”

- The standard error of \( \log(OR) \) is estimated as

\[
\text{se}[\log(OR)] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}
\]

- If any one of the cell counts is 0, this will blow up. In this case, we add 0.5 to each of the cell counts before computing the standard error.

\[
\hat{\text{se}}[\log(OR)] = \sqrt{\frac{1}{a + 0.5} + \frac{1}{b + 0.5} + \frac{1}{c + 0.5} + \frac{1}{d + 0.5}}
\]
**Example**

None of the cell counts is 0, so we estimate the standard error as

\[
s \varepsilon [\log(OR)] = \sqrt{\frac{1}{43} + \frac{1}{26} + \frac{1}{260} + \frac{1}{277}}
\]

\[= 0.263\]

So our confidence interval for the log odds ratio is 95%

\[
\log(1.76) \pm 1.96 \times 0.263
\]

\[0.565 \pm 0.515\]

\[(0.05, 1.08)\]

**Conclusions from this confidence interval**

- We are 95% confident that the true ratio of the odds in the exposed population over the odds in the unexposed population lies between 1.05 and 2.94.
- An odds ratio of 1 would mean that the probability of disease was the same in both the exposed and the unexposed populations.
- Our confidence interval does not include 1.0. Thus we are 95% confident that the probabilities are not equal.

To get the confidence interval for the odds ratio itself, we exponentiate both endpoints.

\[\left(e^{0.05}, e^{1.08}\right)\]

\[(1.05, 2.94)\]

Note that this confidence interval is *not* symmetric.

- Since our confidence interval lies entirely to the right of 1.0, we are 95% confident that the true odds ratio > 1.0.
  - That is, we are 95% confident that there is higher probability of requiring postoperative antibiotics among patients who do not receive prophylactic antibiotics during breast surgery than among patients who do receive prophylactic antibiotics during breast surgery.