Introduction to Biostatistics (171:161)
Breheny

Lab #11

In lab #10, we learned how to use SAS/R to carry out χ^2-tests and Fisher's exact test. In lab #11, we will learn how to calculate and obtain confidence intervals for odds ratios and relative risks.

1 Refresher from lab #10

Cancer	No	Yes
Before age 25	4475	65
25 or older	1597	31

SAS:

```
PROC FREQ DATA = data.bc;
  TABLES FirstLabor*BreastCancer;
  EXACT OR;
  WEIGHT N;
RUN;
```

R:

```
bc <- matrix(c(4475, 65, 1597, 31),
    nrow = 2, ncol = 2, byrow = T)
dimnames(bc) <- list(c("Before age 25",
    "25 or older"), c("No", "Yes"))
fisher.test(bc)
```

2 Which odds ratio?

The odds ratio R/SAS gives you depends on what order you put the rows into. Clearly, this is arbitrary. For example, with the breast cancer data, R gives us an odds ratio of 1.34: the odds of breast cancer were about 34% higher for women who delivered their first child after age 25; alternatively, the odds of breast cancer for women who delivered their first child after age 25 are approximately 1.34 times the odds of breast cancer for women who delivered their first child before the age of 25. For the same data, SAS gives us an odds ratio of 0.75: the odds of breast cancer were cut by 25% if a woman gave birth before age 25; alternatively, the odds of breast cancer for women who gave birth before age 25 are approximately 0.75 times the odds of breast cancer for women who gave birth after 25. How do we get this odds ratio?

In SAS, we can specify `ORDER = DATA` to calculate the odds of breast cancer for women who delivered their first child after age 25 relative to women who delivered their first child before age 25. Other options for `ORDER =` include `FREQ`, `FORMATTED`, and `INTERNAL`. The table below summarizes these options.

PROC FREQ ORDER = Options	
Value of ORDER =	Levels Ordered By
DATA	Order of appearance in the input data set
FORMATTED	External formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value
FREQ	Descending frequency count; levels with the most observations come first in the order
INTERNAL	Unformatted value

In R, we could go back and re-enter the data, or just switch the columns around.

SAS:

```
PROC FREQ DATA = data.bc ORDER = DATA;
  TABLES FirstLabor*BreastCancer;
  EXACT OR;
  WEIGHT N;
RUN;
```

R:

```
fisher.test(bc[,2:1])
```

Note that that the hypothesis tests are unaffected, but the odds ratios and confidence intervals are "flipped". Another approach would simply be to manually invert the odds ratios and confidence intervals:

$$.75^{-1} = 1.33$$
$$.48^{-1} = 2.09$$
$$1.19^{-1} = 0.84$$

In this class, you are free to report whichever odds ratio you like, so long as you describe it correctly.

3 Power

Finally, let's look at the power of the CDC's breast cancer study in comparison with a case-control study using `PROC POWER` in SAS and `bpower` in `R`.

SAS

The overall syntax is fairly similar to the previous lab, only now `TWOSAMPLEFREQ` will replace `ONESAMPLEMEAN`. We still have to specify sample size, variability, and effect size, although these concepts manifest themselves in a different way for two samples instead of one, and for categorical data instead of continuous.

R

In `R` we need to install the package `Hmisc` and load the library. Within the `Hmisc` package is a function called `bpower` that we will use to calculate two-sample categorical power. All arguments supplied in this function can be vectors, which means multiple scenarios can be looked at with one function call. There are eight important arguments for `bpower`: `p1`, `p2`, `odds.ratio`, `percent.reduction`, `n`, `n1`, `n2`, and `alpha`. To use this function, `p1` must be specified and `p2`, `odds.ratio`, or `percent.reduction` must also be specified. You can either specify `n`, which will assume two equal samples, or you can specify `n1` and `n2`. The default value of `alpha` is 0.05, but can be changed.

Power calculations

First, let's get an idea of the power of the CDC's prospective breast cancer study. We have to specify the sizes of both samples, which were about 4500 and 1500. Next, we have to specify the frequency with which the event occurs. Breast cancer is relatively uncommon, occurring in only about 1% of the women in the study. Finally, we have to specify an effect size. Of course, this is unknown, but let's investigate the power of the CDC's study to detect a effect size such that the true odds ratio is 1.5.

SAS:

```
PROC POWER;
  TWOSAMPLEFREQ
    GROUPNS = (4500 1500)
    REFP = .01
    ODDSRATIO = 1.5
    POWER = .;
RUN;
```

R:

```
bpower(.01, odds.ratio = 1.5, n1 = 4500, n2 = 1500)
```

This tells us that the power is 36%. Now, let's consider an alternative, case-control design in which we include 150 cases (women with breast cancer) and 150 controls. Now, the event we are measuring is exposure to the risk factor – whether or not a woman had children after age 25 or not. This occurs about 35% of the time. We will keep the effect size the same, so as to get a sense of the comparative power of the studies:

SAS:

```
PROC POWER;
  TWOSAMPLEFREQ
    GROUPNS = (150 150)
    REFP = .35
    ODDSRATIO = 1.5
    POWER = .;
RUN;
```

R:

```
bpower(.35, odds.ratio = 1.5, n1 = 150, n2 = 150)
# or
bpower(.35, odds.ratio = 1.5, n = 300)
```

For this design, the power is 40%. Note that we obtain greater power with a 300-person retrospective study than we did with a 6,000-person prospective study (20 times fewer people). This is why researchers are often willing to risk the biases of retrospective studies – the increase in power (equivalently, decrease in sample size) can be dramatic.