Stats for Clinical Trials, Math 150
Jo Hardin
Fall 2010
Simpson's Paradox & Retrospective Studies

Retrospective versus Prospective Studies

After much research (and asking many people who do not all agree!), I finally came across a definition of retrospective that I like. Note, however, that many many books *define* retrospective as synonymous with case-control. That is, they define a retrospective study to be one in which the observational units were chosen based on their status of the response variable. I disagree with that definition. As you see below, retrospective studies are defined based on the when the variables were *measured*. I've also given a quote from the Kuiper text where retrospective is defined as any study where historic data are collected (I like this definition less).

> Studies can be classified further as either prospective or retrospective. We define a prospective study as one in which exposure and covariate measurements are made before the cases of illness occur. In a retrospective study these measurements are made after the cases have already occurred... Early writers referred to cohort studies as prospective studies and to case-control studies as retrospective studies because cohort studies usually begin with identification of the exposure status and then measure disease occurrence, whereas case-control studies usually begin by identifying cases and controls and then measure exposure status. The terms prospective and retrospective, however, are more usefully employed to describe the timing of disease occurrence with respect to exposure measurement. For example, case-control studies can be either prospective or retrospective. A prospective case-control study uses exposure measurements taken before disease, whereas a retrospective case-control study uses measurements taken after disease. [**Modern Epidemiology**, 2nd edition, Rothman & Greenland, page 74]

> Retrospective cohort studies also exist. In these designs past (medical) records are often used to collect data. As with prospective cohort studies, the objective is still to first establish groups based on an explanatory variable. However since these are past records the response variable can be collected at the same time. [**Stat2Labs**, S. Kuiper, chapter 6, page 24]

Simpson's Paradox

Consider the example on smoking and 20-year mortality (case) from section 3.4 of *Regression Methods in Biostatistics*, pg 52-53. Because the data broken down by age was not available, I made it up using the original data as a base and the reported OR to guide me.

age		smoker	nonsmoker	prob smoke	odds smoke	empirical OR	book OR
all	case	139	230	0.377	0.604	0.685	
	control	443	502	0.469	0.882		
18-44	case	61	32	0.656	1.906	1.627	1.777
	control	375	320	0.540	1.172		
45-64	case	34	66	0.340	0.515	1.308	1.320
	control	50	127	0.282	0.394		
65+	case	44	132	0.250	0.333	1.019	1.018
	control	18	55	0.247	0.327		

What we see is that the vast majority of the controls were young, and they had a high rate of smoking. A good chunk of the cases were older, and the rate of smoking was substantially lower in the oldest group. However, within each group, the cases were more likely to smoke than the controls.

R code / logistic regression on Simpson's Paradox smoking data

```
death <- c(rep(1,93),rep(0,695), rep(1,100),rep(0,177), rep(1,176), rep(0,73))
smoke <- c(rep(1,61), rep(0,32), rep(1,375), rep(0,320), rep(1,34), rep(0,66),
          rep(1,50), rep(0,127), rep(1,44), rep(0,132), rep(1,18), rep(0,55))
age <- c(rep("young", 788), rep("middle", 277), rep("old", 249) )


> summary(glm( death ~ smoke, family="binomial"))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.78052    0.07962  -9.803  < 2e-16 ***
smoke       -0.37858    0.12566  -3.013  0.00259 **

    Null deviance: 1560.3  on 1313  degrees of freedom
Residual deviance: 1551.1  on 1312  degrees of freedom
AIC: 1555.1

> summary(glm( death ~ smoke + as.factor(age), family="binomial"))
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.6684     0.1347  -4.961 7.03e-07 ***
smoke                 0.3122     0.1539   2.028   0.0425 *
as.factor(age)old     1.4745     0.1881   7.838 4.59e-15 ***
as.factor(age)young  -1.5248     0.1731  -8.809  < 2e-16 ***

    Null deviance: 1560.3  on 1313  degrees of freedom
Residual deviance: 1231.5  on 1310  degrees of freedom
AIC: 1239.5

> summary(glm( death ~ smoke * as.factor(age), family="binomial"))
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.6545     0.1517  -4.313 1.61e-05 ***
smoke                       0.2689     0.2691   0.999    0.318
as.factor(age)old           1.5300     0.2209   6.927 4.29e-12 ***
as.factor(age)young        -1.6481     0.2396  -6.880 6.00e-12 ***
smoke:as.factor(age)old    -0.2505     0.4201  -0.596    0.551
smoke:as.factor(age)young   0.2177     0.3548   0.614    0.540

    Null deviance: 1560.3  on 1313  degrees of freedom
Residual deviance: 1230.0  on 1308  degrees of freedom
AIC: 1242.0
```