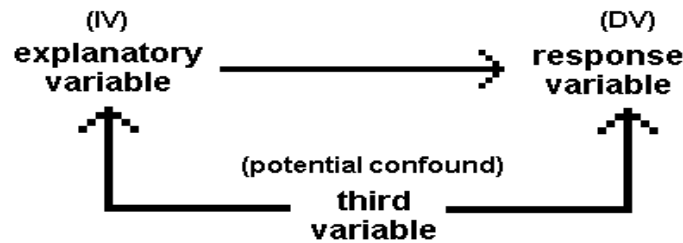


Analyses With Three Categorical Variables

- A. confounding variables
- B. collapsing over or across a variable
- C. what does it mean to control one variable for another variable?
- D. Simpson's paradox
- E. the correct analysis
- F. graphics



A. Confounding Variables (the case of all categorical variables)

A confound is not just "something that goes wrong with an experiment." A confounding variable is one that works as follows. If a third variable is related to BOTH the explanatory variable (IV) and the response variable (DV), then it is potentially a confounding variable. It becomes a confound when it creates an apparent relationship between these two variables that does not appear when the third variable is included in the data analysis, or when it wipes out or changes a relationship that would otherwise appear.

When data are derived from observational, correlational, or quasi-experimental research designs, i.e., when subjects are not randomly assigned to groups or conditions (a true randomized experiment), the potential for confounding is high. For example, researchers have found that people who regularly smoke marijuana have poorer short-term memory performance than people who don't smoke marijuana. Why? It could be an effect of marijuana smoking. In fact, that's a good possibility. However, there is a long list of things that can affect short-term memory (the DV). Is it possible that any of those things are also related to whether or not a person smokes marijuana (the IV)? How about alcohol use? How about use of other drugs? How do we know the memory deficit seen in the smoking group was actually due to marijuana use and not to some other behavior the smokers engage in but the nonsmokers do not? Is there some way we could do this study so as to avoid these potential confounds?

The R built-in dataset UCBAAdmissions is a classic dataset describing admissions to the top six grad programs at UC Berkeley in 1973, broken down by gender, department, and whether or not the applicant was admitted.

```
> data(UCBAAdmissions)           # copy the dataset to the workspace
> dimnames(UCBAAdmissions)      # see the names of the variables and their values
$Admit
[1] "Admitted" "Rejected"

$Gender
[1] "Male"    "Female"

$Dept
[1] "A" "B" "C" "D" "E" "F"
```

```

> margin.table(UCBAdmissions, margin=c(2,1)) -> admit.by.gender # collapse
> admit.by.gender
      Admit
Gender  Admitted Rejected
Male    1198     1493
Female   557     1278

> prop.table(admit.by.gender, margin=1) * 100
      Admit
Gender  Admitted Rejected
Male    44.51877 55.48123
Female  30.35422 69.64578
> 44.52/30.35 # likelihood ratio (calculate an odds ratio by hand)
[1] 1.466886

```

What is your conclusion concerning possible gender bias in graduate admissions to Berkeley in 1973?

B. Collapsing Over or Across a Variable

Because these data were added across departments, we say we collapsed across (or over) departments. I.e., the department variable was ignored, as if it didn't exist. This is dangerous when that variable could be a confound! Let's check to see if "Dept" might be a confound by checking to see if it is related to both "Gender" and "Admit."

```

> margin.table(UCBAdmissions, margin=c(3,1)) -> admit.by.dept
> prop.table(admit.by.dept, margin=1)
      Admit
Dept  Admitted Rejected
A 0.64415863 0.35584137
B 0.63247863 0.36752137
C 0.35076253 0.64923747
D 0.33964646 0.66035354
E 0.25171233 0.74828767
F 0.06442577 0.93557423
> prop.table(admit.by.dept, margin=1)
      Admit
Dept  Admitted Rejected
A 0.64415863 0.35584137
B 0.63247863 0.36752137
C 0.35076253 0.64923747
D 0.33964646 0.66035354
E 0.25171233 0.74828767
F 0.06442577 0.93557423
> margin.table(UCBAdmissions, c(3,2)) -> gender.by.dept
> prop.table(gender.by.dept, margin=1)
      Gender
Dept  Male  Female
A 0.88424437 0.11575563
B 0.95726496 0.04273504
C 0.35403050 0.64596950
D 0.52651515 0.47348485
E 0.32705479 0.67294521
F 0.52240896 0.47759104

```

Yes, the department variable is related to both admission (there is a much higher likelihood of being

admitted into departments A and B than into other departments) and to gender of applicant (the likelihood of an applicant to departments A and B being male is much higher than it is in other departments). So the department variable must be considered a potential confound.

Are you starting to see a story emerge here? Notice that the departments that are easiest to get into are also the departments that have the highest percentage of male applicants.

C. What Does It Mean to Control One Variable For Another Variable?

Let's break the data down by departments. It already is broken down this way in the original contingency table, but I want to rearrange the variables just a bit to put the explanatory variable of interest (Gender) in the rows.

```
> margin.table(UCBAdmissions, margin=c(2,1,3)) -> UCBAadm
> UCBAadm
, , Dept = A
```

Gender	Admit	
	Admitted	Rejected
Male	512	313
Female	89	19

```
, , Dept = B
```

Gender	Admit	
	Admitted	Rejected
Male	353	207
Female	17	8

```
, , Dept = C
```

Gender	Admit	
	Admitted	Rejected
Male	120	205
Female	202	391

```
, , Dept = D
```

Gender	Admit	
	Admitted	Rejected
Male	138	279
Female	131	244

```
, , Dept = E
```

Gender	Admit	
	Admitted	Rejected
Male	53	138
Female	94	299

```
, , Dept = F
```

Gender	Admit	
	Admitted	Rejected
Male	22	351
Female	24	317

In which departments do the odds of admission favor men, in which do the odds of admission favor women, and in which are the odds about even? (Yes, we can have R do this for us, but it doesn't hurt to fire up the calculator every once in awhile so that it doesn't feel left out! We will use odds because odds are quicker to calculate than likelihoods and give the same information). How can this be?

In this analysis we have controlled for department. To say this in its most complete form, we have controlled the effect of gender on admission for department applied to. When you control one variable for another (admission for department, in this case), that means you look at the effect of the primary variable of interest (the IV) at each level (or value) of the controlled variable, rather than ignoring (or collapsing over) the controlled variable. This is a form of statistical control that can be used when experimental control (e.g., randomization) cannot be used. Experimental control is very much to be preferred, but sometimes statistical control is the best we can do.

In this case, we would say that "gender is confounded with department." Explain how that is true.

D. Simpson's Paradox

When an effect is produced by collapsing data over a third variable, this is called Simpson's paradox. In other words, in Simpson's paradox an effect appears when the frequencies are added across a third variable that does not appear when the frequencies are analyzed at each level of the third variable.

E. The Correct Analysis

If we are looking for possible gender bias, we must take into account the department variable. There are two statistical techniques that would allow us to do that in this case: log-linear analysis and logistic regression. We'll look at those later. For the moment, we will depend upon a graphical analysis.

F. Graphics

Seeing relationships among 3 variables on a 2D piece of paper can be tricky. One way to do it is to reduce one of the variables to a single summary statistic, such as likelihood or odds. That is the strategy we will use here.

```
> margin.table(UCBAdmissions, margin=c(2,3,1)) -> UCBAadm2
> UCBAadm2
, , Admit = Admitted

      Dept
Gender  A   B   C   D   E   F
Male   512 353 120 138  53  22
Female  89  17 202 131  94  24

, , Admit = Rejected

      Dept
Gender  A   B   C   D   E   F
Male   313 207 205 279 138 351
Female  19   8 391 244 299 317
```

What happens now if we take the first layer of this table (Admit = Admitted) and divide it by the second layer (Admit = Rejected)? What would the result be?

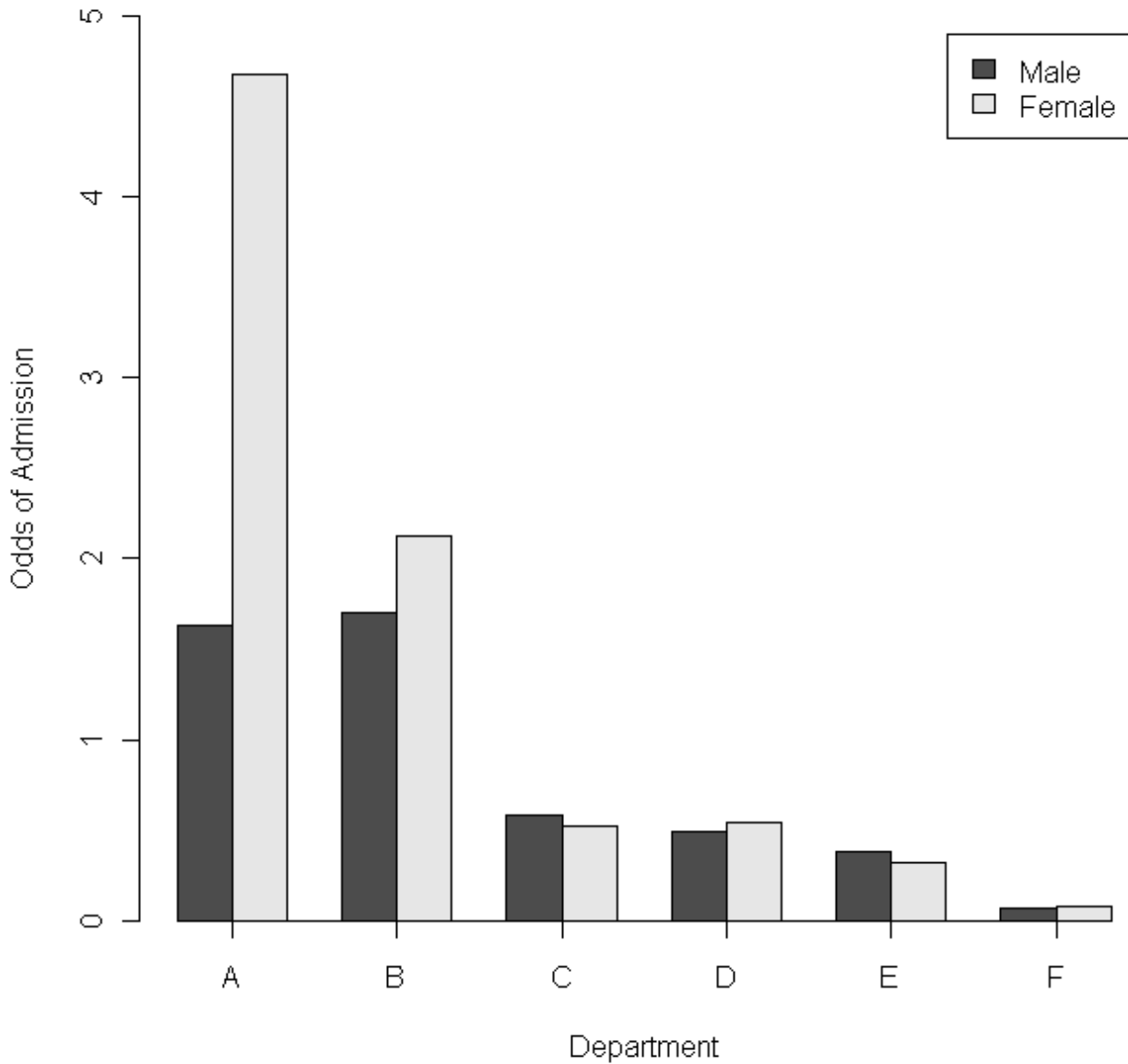
```

> UCBAadm2[1:2,1:6,1] / UCBAadm2[1:2,1:6,2] -> odds
> odds
      Dept
Gender  A      B      C      D      E      F
Male   1.635783 1.705314 0.5853659 0.4946237 0.3840580 0.06267806
Female 4.684211 2.125000 0.5166240 0.5368852 0.3143813 0.07570978

> barplot(odds, beside=T, legend=T, ylim=c(0,5), axis.lty=1)
> title(xlab="Department", ylab="Odds of Admission")
> title(main="Odds of Admission into UC Berkeley Grad Programs by Gender")

```

Odds of Admission into UC Berkeley Grad Programs by Gender



Moral of the Story: When you have three variables, you should analyze three variables, especially when that third variable might be a confound. Now here's a hard question! Can you see any reason why the odds of admission might favor men in some departments and women in other departments?