LSN1-GradSchoolBerkeley

Jim Starling

1/5/2022

Exploring the Berkeley Grad School Dataset

Let's change the data to categorical data:

```
grad2 <- grad %>% mutate(Program = as.factor(Program)) %>%
         mutate(Sex = as.factor(Sex)) %>%
         mutate(Accepted = as.factor(Accepted))
```

Explore the Berkeley Data set

Let's take a look at the data:

```
head(grad)
```

```
## # A tibble: 6 x 3
##   Program Sex   Accepted
##   <chr>   <chr> <chr>
## 1 A       Male  Yes
## 2 A       Male  Yes
## 3 A       Male  Yes
## 4 A       Male  Yes
## 5 A       Male  Yes
## 6 A       Male  Yes
```

```
glimpse(grad)
```

```
## Rows: 1,647
## Columns: 3
## $ Program  <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "~
## $ Sex      <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male~
## $ Accepted <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes"~
```

```
table(grad$Program)
```

```
##
##   A   F
## 933 714
```

```
table(grad$Program, grad$Sex)
```

```
##
##     Female Male
##   A    108  825
##   F    341  373
```

```
table(grad$Accepted,grad$Sex)
```

```
##
##       Female Male
##   No     336  665
##   Yes    113  533
```

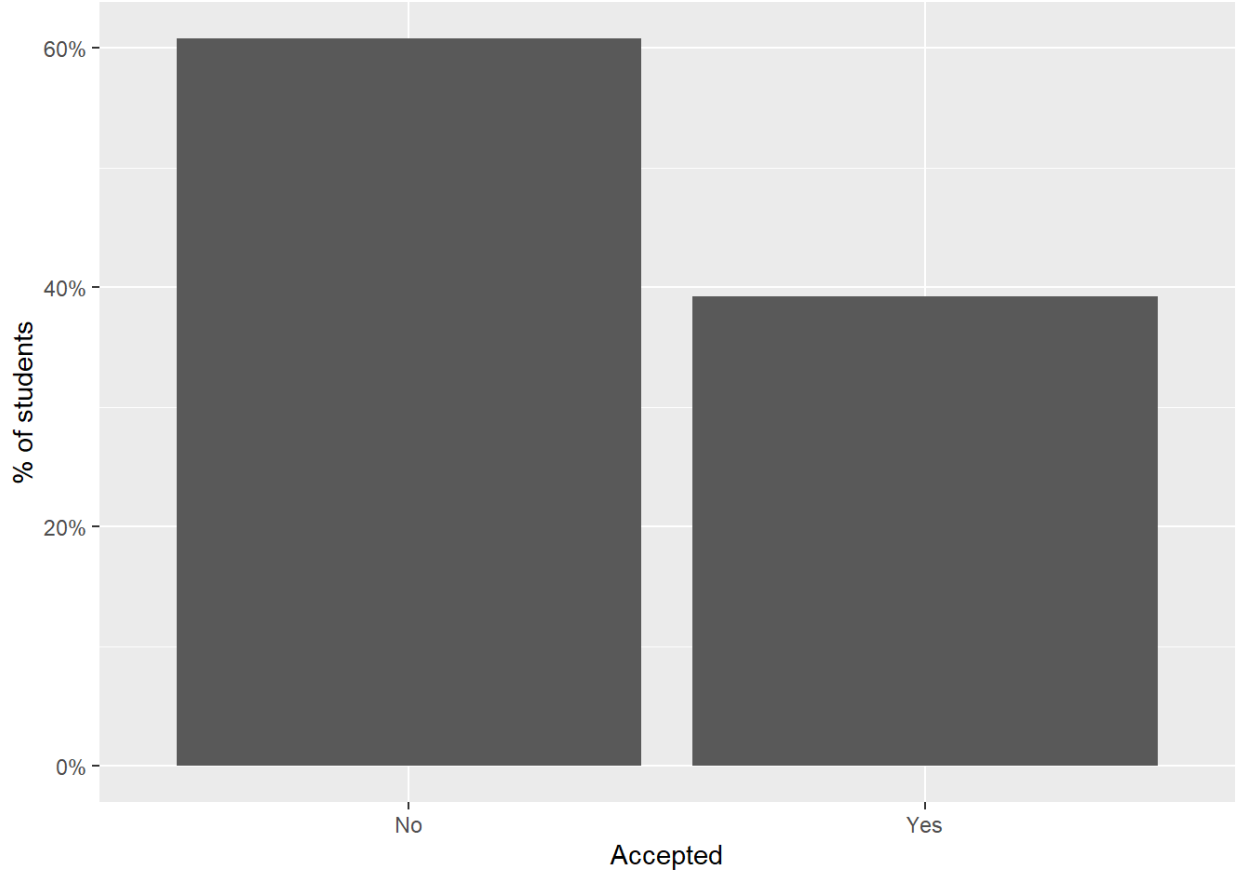
Observational units: *People who are applied to grad school at UC Berkeley.*

Variables

- **Response variable:**
  - *Accepted (binary) (yes/no)*
- **Explanatory variables:**
  - *Sex (categorical) (Male/Female)*
  - *Program (categorial) (A/F)*

Create a bar plot of accepted vs rejected.

```
acc2 <- grad2 %>% pull(Accepted) %>%
  fct_count() %>%
  mutate(perc = n / nrow(grad2))
# same as above: # acc2 <- fct_count(grad2$Accepted) %>% mutate(perc = n / nrow(grad2))
acc2 %>% ggplot(aes(x=f)) +
  geom_bar(aes(y=perc),stat = "identity") +
  labs(x="Accepted", y="% of students") +
  scale_y_continuous(labels=scales::percent)
```



What do we see here?

Most (60%) of the applicants were rejected.

Create a Contingency Table

We can ask the question, “Is the admissions process at Berkeley fair?”. Or a more focused question, “Does the admissions process at Berkeley disproportionately affect women applicants?”

```
table1(~Accepted | Sex, data=grad2)
```

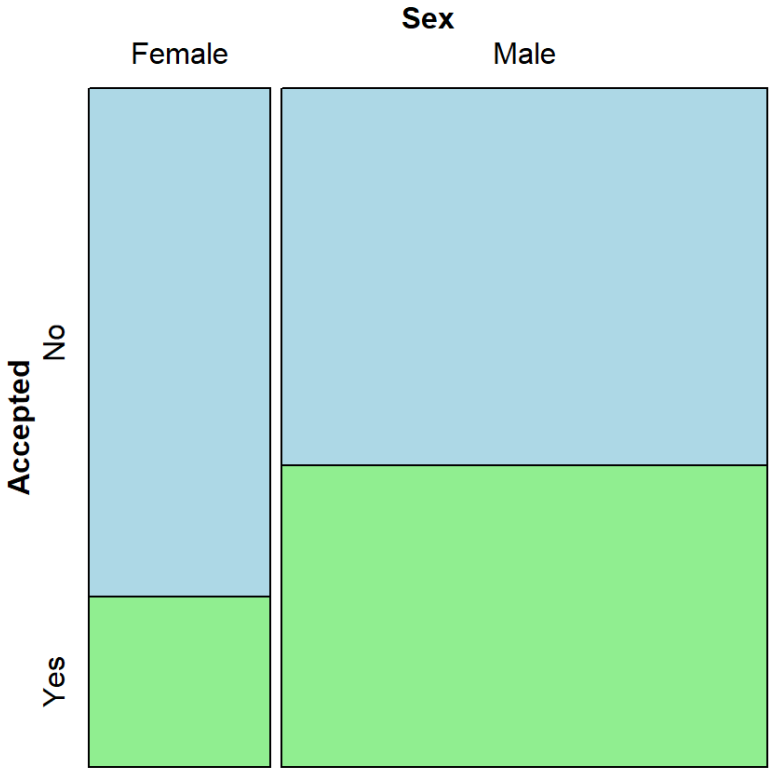
	Female (N=449)	Male (N=1198)	Overall (N=1647)
Accepted			
No	336 (74.8%)	665 (55.5%)	1001 (60.8%)
Yes	113 (25.2%)	533 (44.5%)	646 (39.2%)

What do we see here? Is the

Create Mosaic plot

Create a mosaic plot to see the data related to sex and acceptance:

```
# labs <- round(prop.table(table(grad2$Accepted, grad2$Sex)),2)
mosaic(~Sex+Accepted, data=grad2, highlighting="Accepted", highlighting_fill = c("lightblue", "lightgreen"), direction=c("v"
,"h"))
```

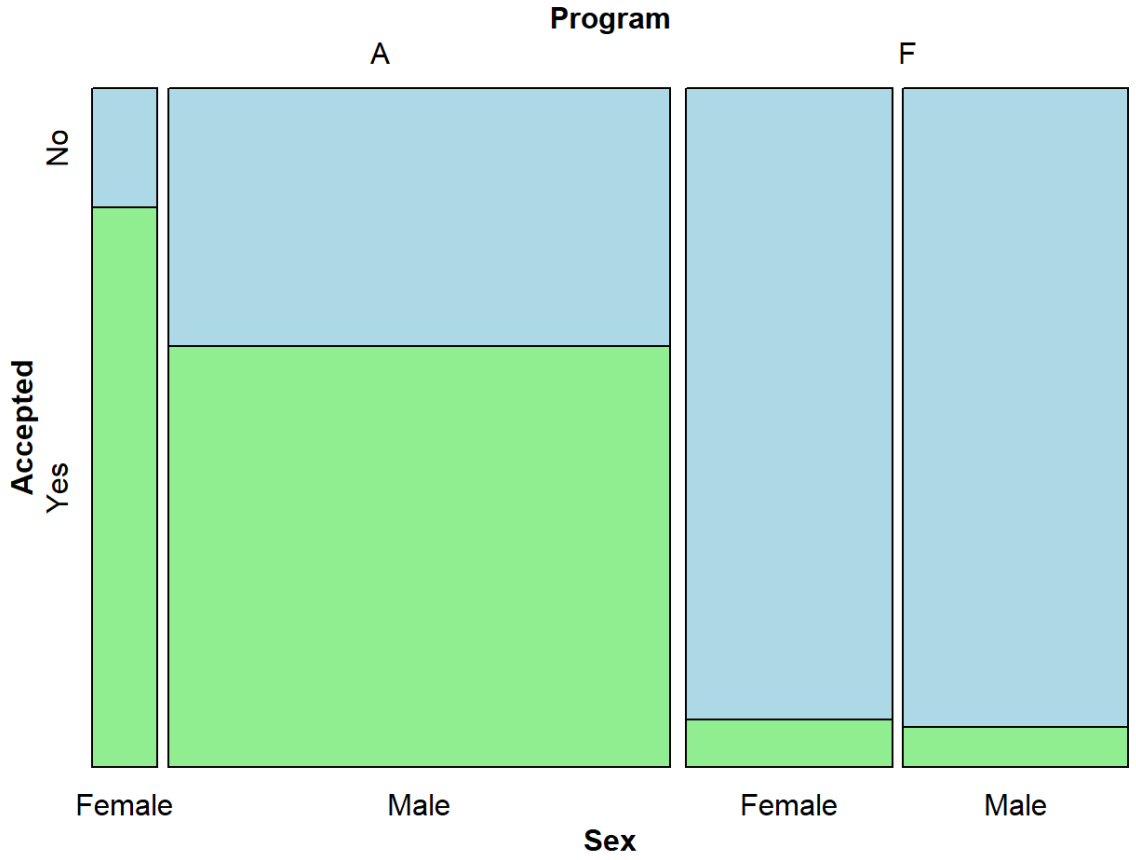


What does this plot tell us? Are sex and acceptance *associated*?

It looks like there are less females accepted. Also, more males applied to grad school. Sex and acceptance are associated.

Does the program a person applied to make a difference? To explore this, we will condition on the program applied to:

```
mosaic(~Program+Accepted+Sex, data=grad2, highlighting="Accepted", highlighting_fill = c("lightblue", "lightgreen"), directi
on=c("v","h","v"))
```



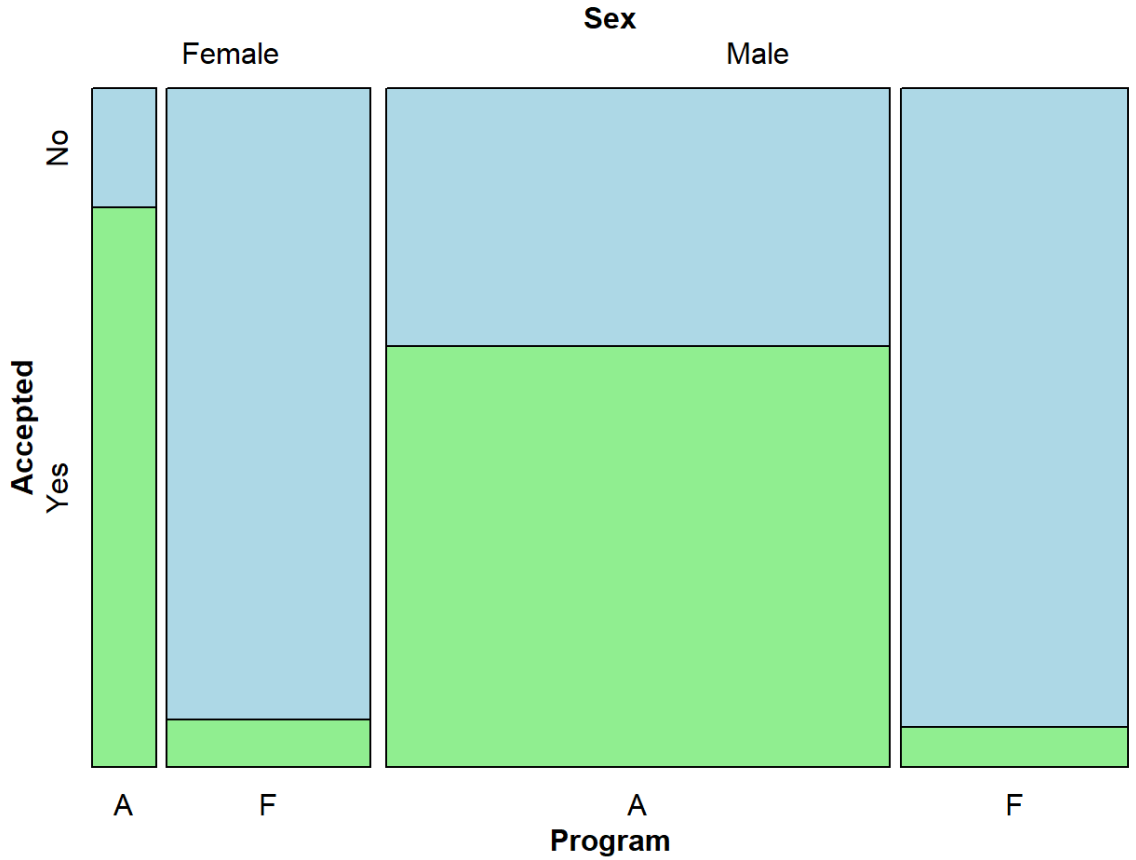
What does this plot tell us? Is program associated with acceptance? How is this different than the previous plot?

For program A, it appears that less females applied, but they were accepted at a higher rate than males. For program F, it appears that roughly an equal number of females and males applied and were accepted at roughly the same (lower) rate. We also see that program AND sex are associated with acceptance.

This is called Simpson's paradox, where the direction of the association between sex and acceptance reverses direction.

Here we look at the association between acceptance and program, conditioned on sex:

```
mosaic(~Sex+Accepted+Program, data=grad2, highlighting="Accepted", highlighting_fill = c("lightblue", "lightgreen"), directi
on=c("v","h","v"))
```



What do we notice here compared to our first mosaic plot above?

We notice that must consider the program and the sex when considering the acceptance rates for males and females.

What can we say about our original question, “Does the admissions process at Berkeley disproportionately affect women applicants?”? What do we call the variable *program*?

When looking only at acceptance rates and sex, we may incorrectly say that females have a disproportionate admission rate if we do not consider which program they applied to. ‘Program’ is called a confounding variable. (draw diagram if necessary)

Sources of Variation Diagram

On p. 8, the author provides the Sources of Variation diagram shown here.

Observed Variation in: Acceptance (Yes or No)	**Sources of explained variation**	**Sources of unexplained variation**
Inclusion criteria • Class (graduate students) • School (Berkeley)	• Sex (male or female) • Program	• Quality of application • Numerical data (e.g., test scores, grade point average) • Unknown ...

Are there other sources to consider?