Lecture 6: Improving visualiztion quality using uncertainty

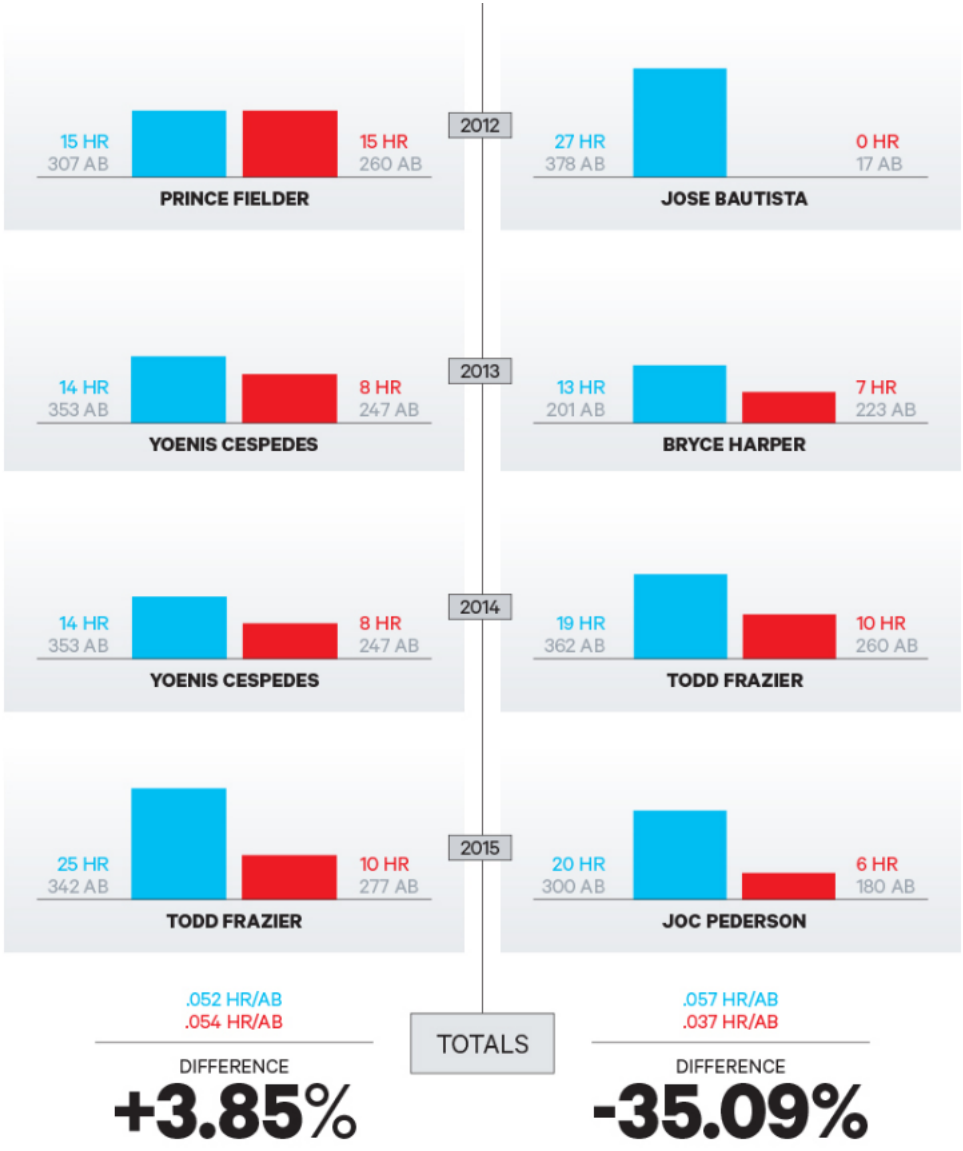
Michael Lopez, Skidmore College

Examples

1. What are the plots trying to show?
2. What are the plots actually showing?
3. Are the plots accurate?
4. Truth continuum
5. Model bugs: audience

Example plot 1

`MLB's real home run derby curse impacts second place`, (link)



Example plot 2

`The web is dead`, (Wired.com)

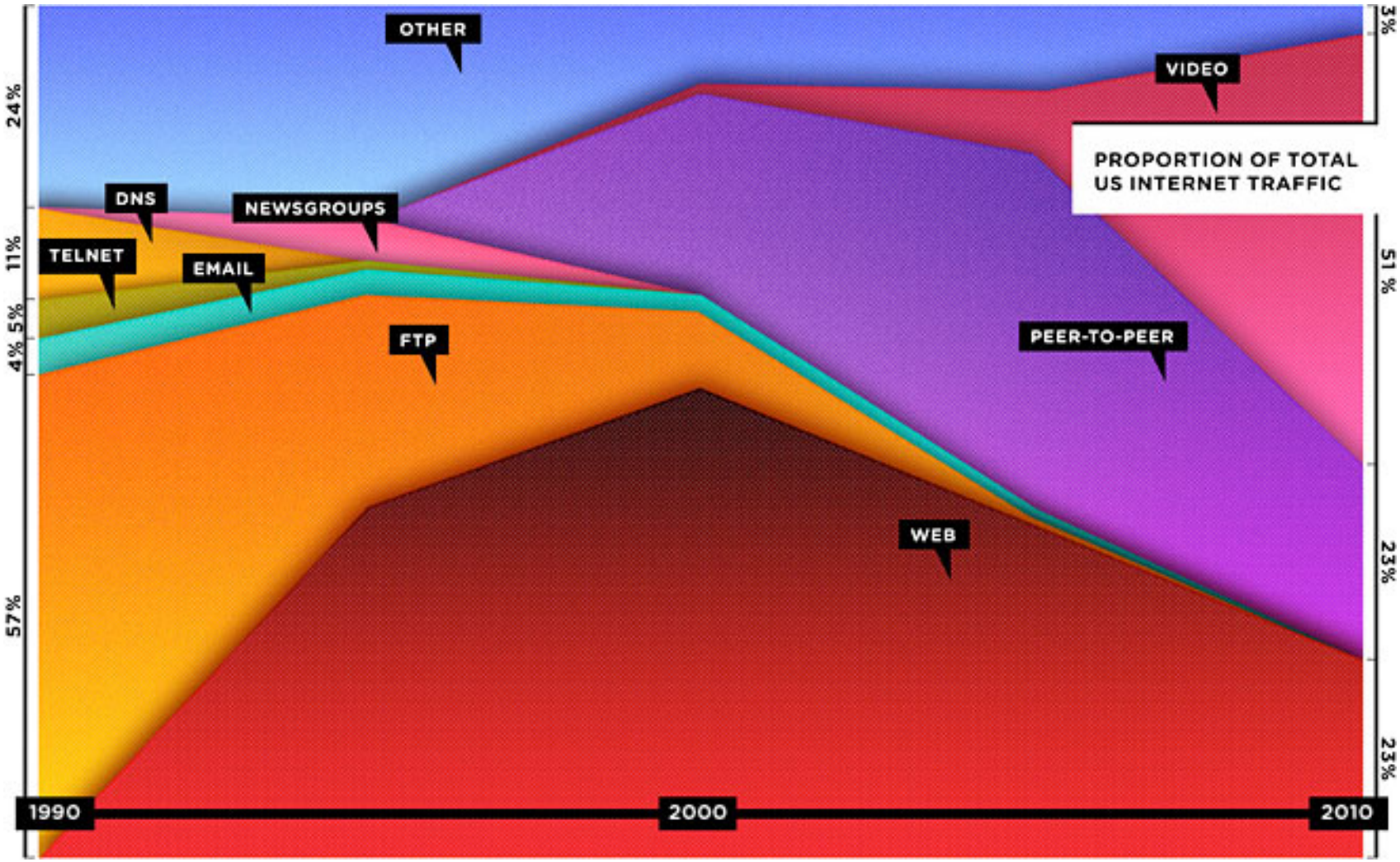
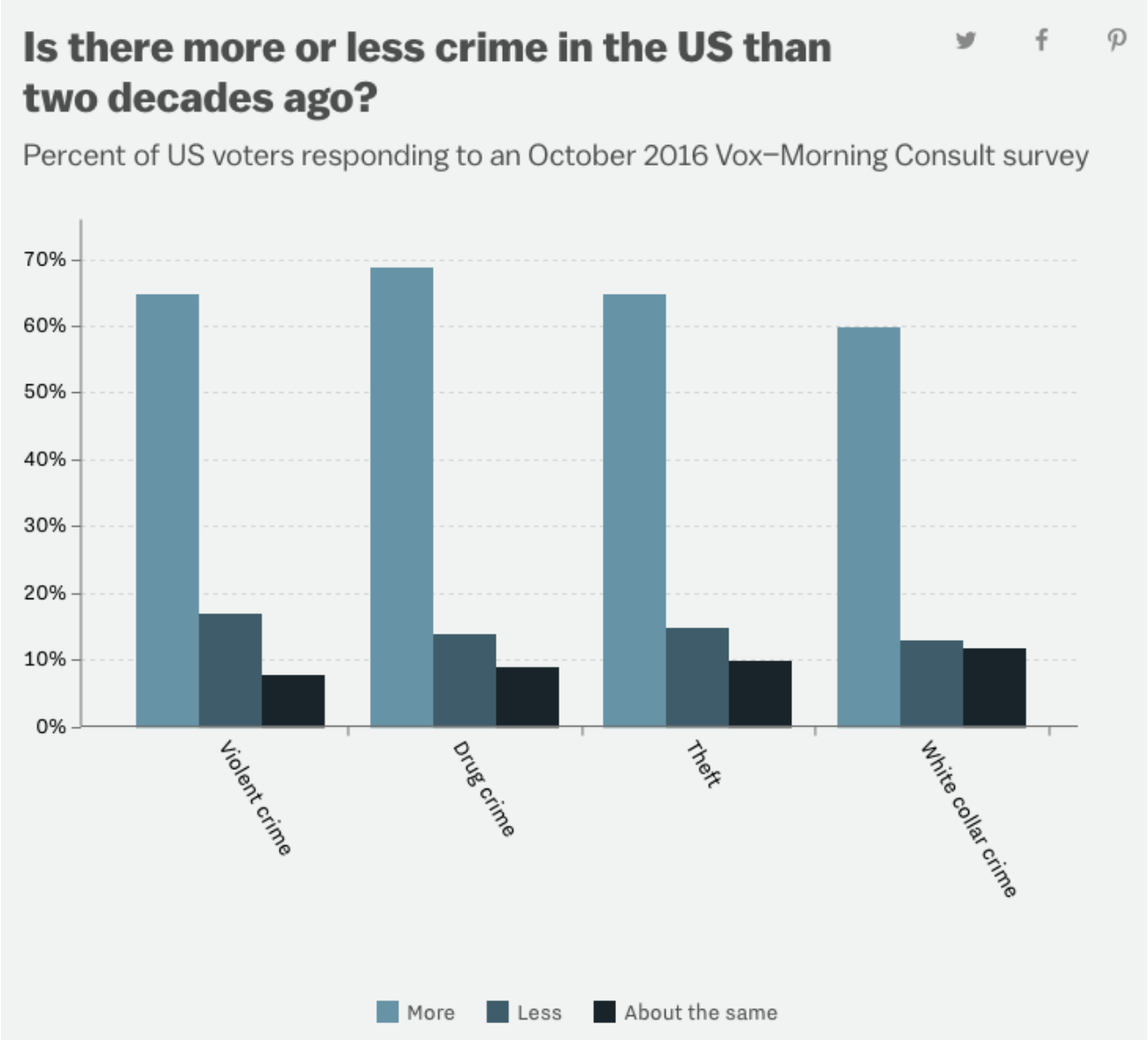


Figure 2:Proportion of total internet traffic.

Example plot 3

`More or less crime?`, (Vox.com)



Today's goals

1. Communicating Variability
2. Variability by variable type
3. Graphing variability
4. Communicating change

Accounting for variability - communicating change

An example

Cal-Berkeley admissions, 1973

Table 1:Cal-Berkeley admissions, 1973

	Admitted	Rejected
Male	1198	1493
Female	557	1278

What does this look like? How can we visualize?

An example

```
library(dplyr); library(ggplot2)
cb.df <- data.frame(apply(UCBAdmissions, c(2, 1), sum))
cb.df
```

```
##        Admitted Rejected
## Male       1198     1493
## Female      557     1278
```

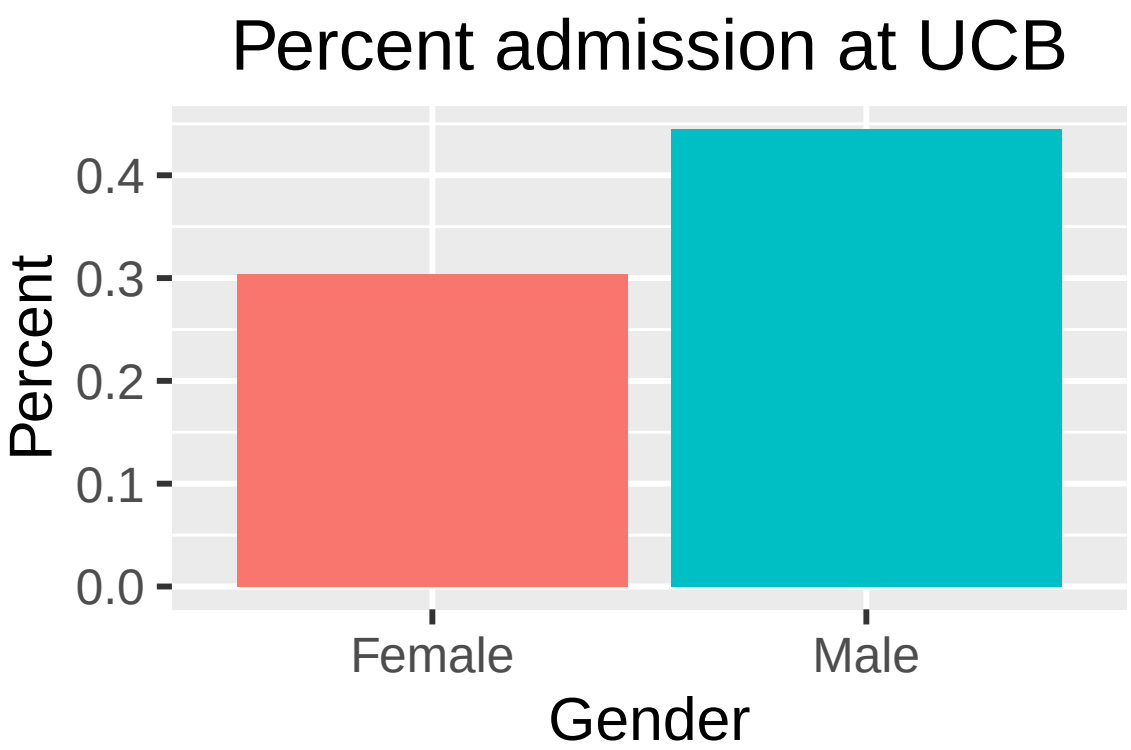
An example

```
cb.df <- cb.df %>%
  mutate(Gender = c("Male", "Female"),
         Percent = Admitted / (Admitted + Rejected),
         n = Admitted + Rejected)
cb.df
```

```
##   Admitted Rejected Gender   Percent    n
## 1     1198     1493   Male 0.4451877 2691
## 2      557     1278 Female 0.3035422 1835
```

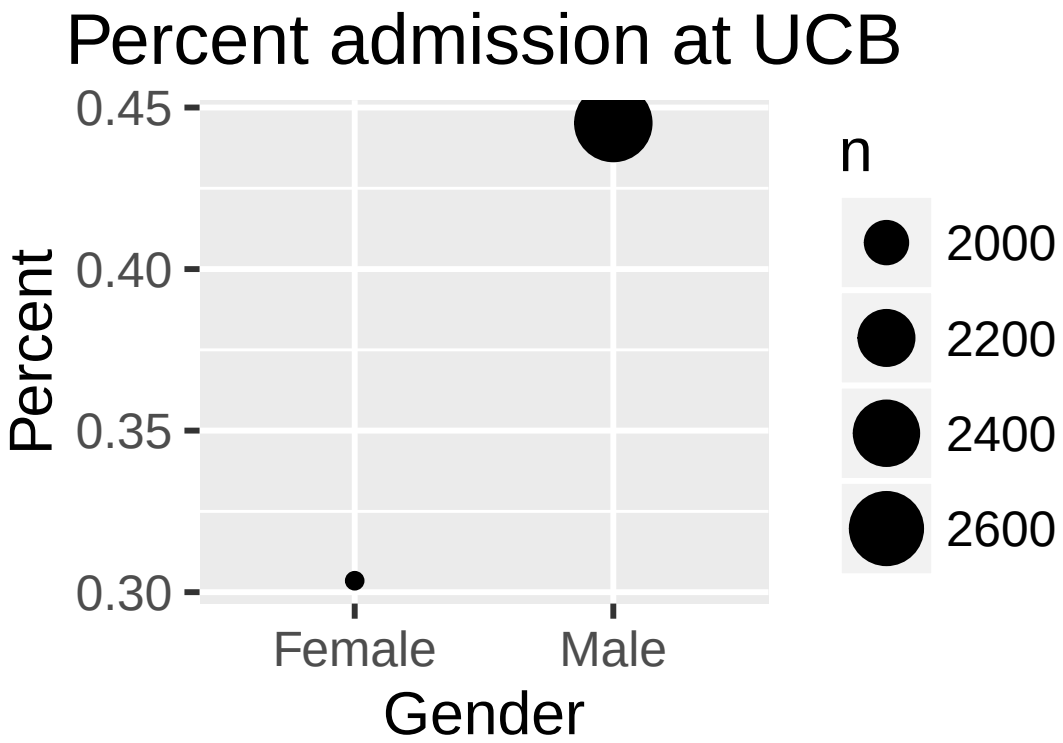
An example

```
ggplot(cb.df, aes(x = Gender, y = Percent, fill = Gender)) +
  geom_bar(stat = "identity") +
  ggtitle("Percent admission at UCB") +
  theme(legend.position = "none")
```



An example

```
ggplot(cb.df, aes(x = Gender, y = Percent, size = n)) +
  geom_point() +
  ggtitle("Percent admission at UCB")
```



Possible explanations?

Definitions in statistics

- Observation/unit
- Standard deviation
- Statistic
- Standard error
- Margin of error

Margins of error + assumptions

- Continuous data
- Proportions
- Trends

An aside

Margin of error for proportions: a proof for the rule of thumb

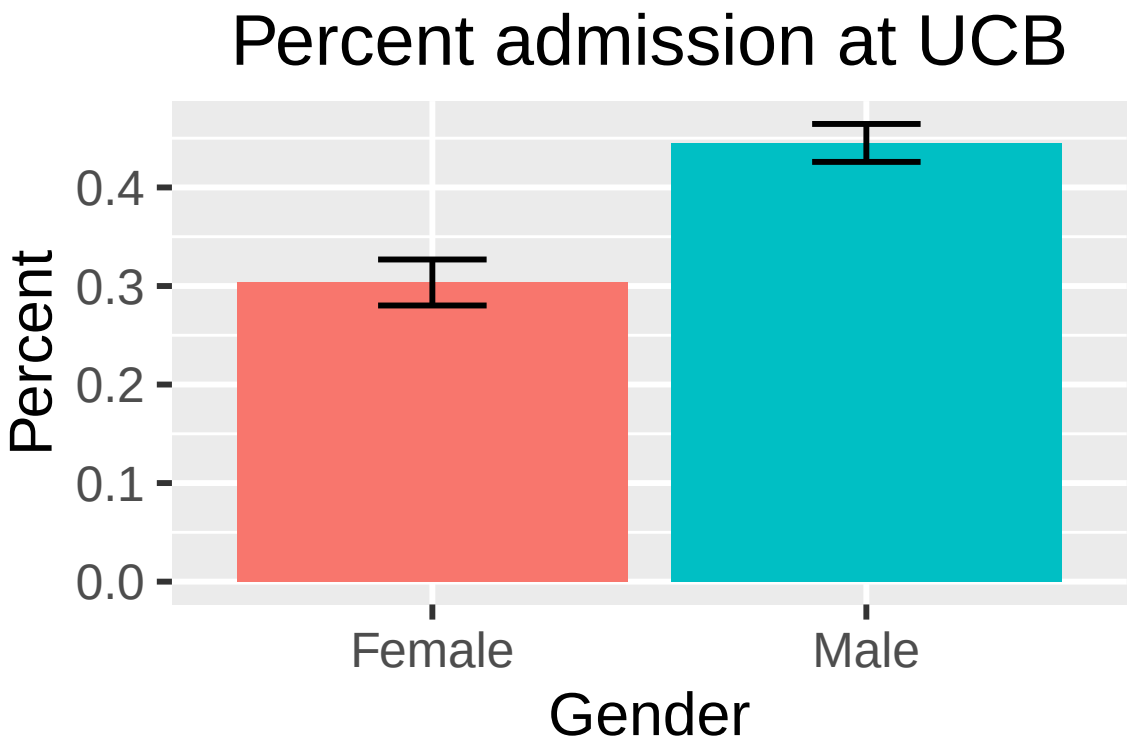
An example

```
cb.df <- mutate(cb.df,  MOE = 1/sqrt(n))
cb.df
```

```
##   Admitted Rejected Gender   Percent    n         MOE
## 1     1198     1493   Male 0.4451877 2691 0.01927716
## 2      557     1278 Female 0.3035422 1835 0.02334436
```

An example

```
limits <- aes(ymin = Percent - MOE, ymax = Percent + MOE)
ggplot(cb.df, aes(x = Gender, y = Percent, fill = Gender)) +
  geom_bar(stat = "identity") +
  geom_errorbar(limits, width=0.25) +
  ggtitle("Percent admission at UCB") +
  theme(legend.position = "none")
```



Example, continuous data

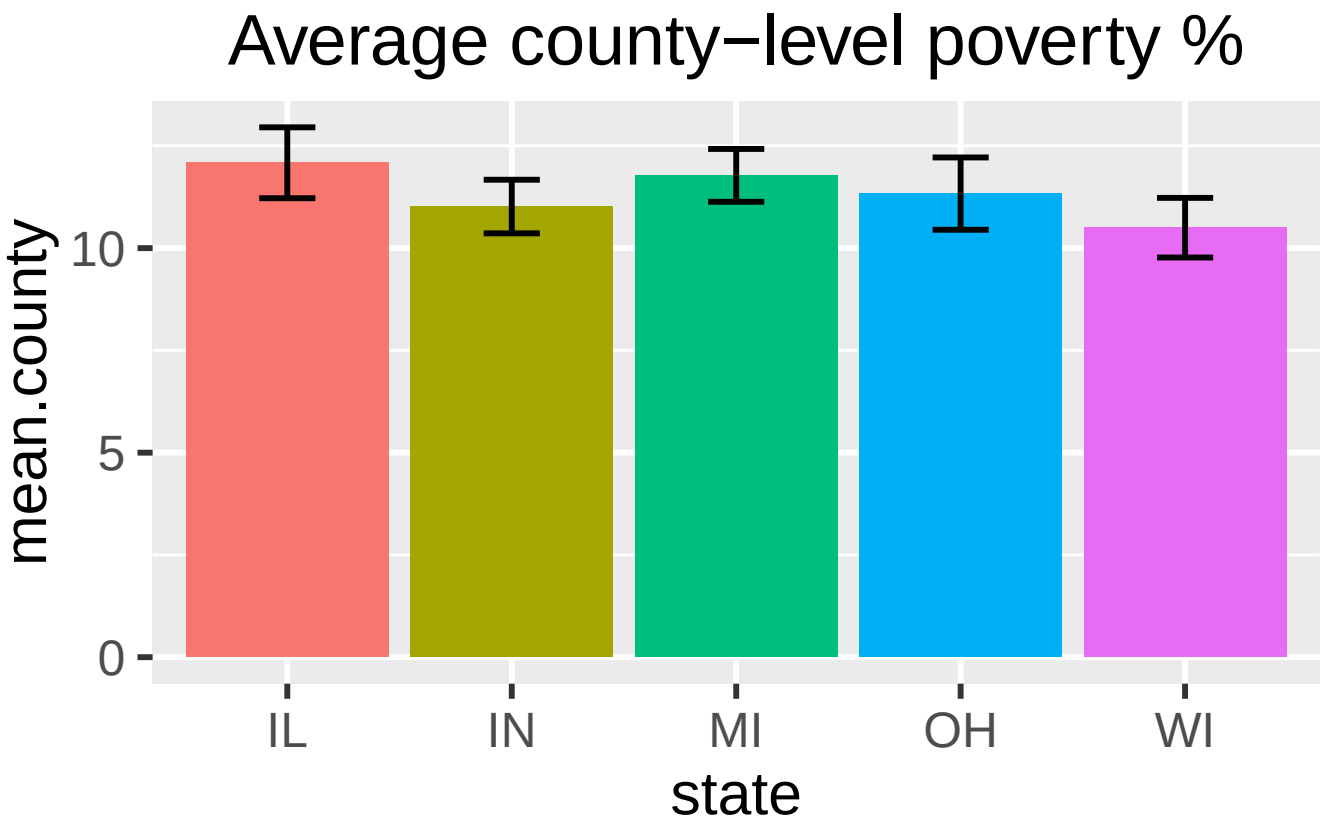
Midwest data - county-level percentages of elderly in poverty, by state.

```
midwest1 <- midwest %>%
  group_by(state) %>%
  summarise(mean.county = mean(percelderlypoverty),
            sd.county = sd(percelderlypoverty),
            n.county = n()) %>%
  mutate(moe.county = 2*sd.county/sqrt(n.county))
midwest1
```

```
## Source: local data frame [5 x 5]
##
##   state mean.county sd.county n.county moe.county
##   (chr)       (dbl)     (dbl)    (int)      (dbl)
## 1    IL    12.08606  4.372512      102  0.8658863
## 2    IN    11.01617  3.141212       92  0.6549881
## 3    MI    11.77796  2.940704       83  0.6455683
## 4    OH    11.33227  4.145811       88  0.8838898
## 5    WI    10.49910  3.092291       72  0.7288600
```

Example, continuous data

```
limits <- aes(ymin = mean.county - moe.county, ymax = mean.county + moe.county)
ggplot(midwest1, aes(x = state, y = mean.county, fill = state)) +
  geom_bar(stat = "identity") +
  geom_errorbar(limits, width=0.25) +
  ggtitle("Average county-level poverty %") +
  theme(legend.position = "none")
```



Back to the UCB data

```
admitted.males <- UCBAdmissions[1,,][1,]
applicants.males <- UCBAdmissions[1,,][1,] + UCBAdmissions[2,,][1,]
admitted.females <- UCBAdmissions[1,,][2,]
applicants.females <- UCBAdmissions[1,,][2,] + UCBAdmissions[2,,][2,]
df.UCB <- data.frame(Department = letters[1:6],
                     Admitted = c(admitted.males, admitted.females),
                     Applied = c(applicants.males, applicants.females),
                     Gender = c(rep("Male", 6), rep("Female", 6)))
df.UCB
```

```
##    Department Admitted Applied Gender
## 1           a      512     825   Male
## 2           b      353     560   Male
## 3           c      120     325   Male
## 4           d      138     417   Male
## 5           e       53     191   Male
## 6           f       22     373   Male
## 7           a       89     108 Female
## 8           b       17      25 Female
## 9           c      202     593 Female
## 10          d      131     375 Female
## 11          e       94     393 Female
## 12          f       24     341 Female
```

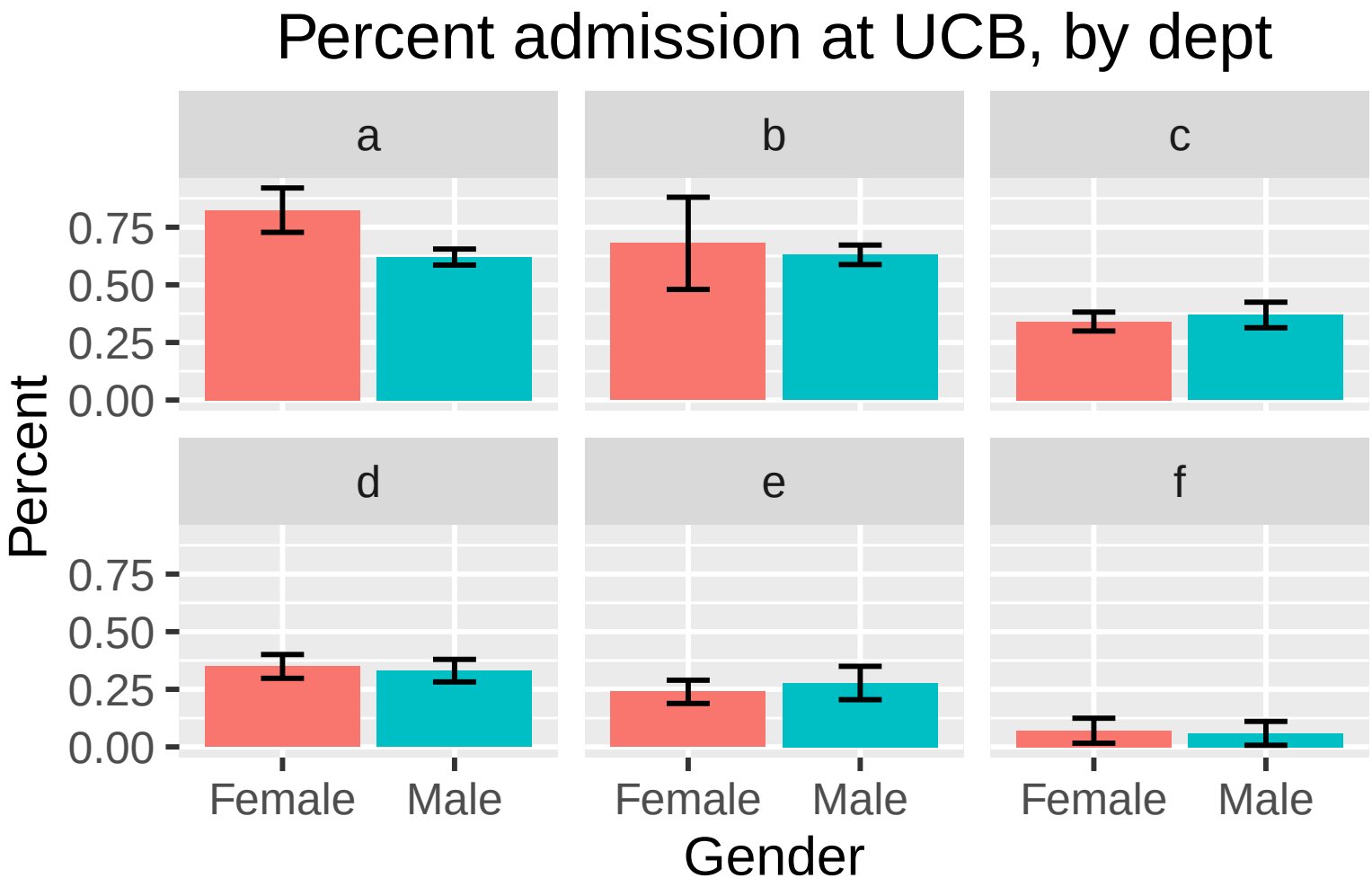
Back to the UCB data

```
df.UCB <- df.UCB %>%
  mutate(Percent = Admitted/Applied, MOE = 1/sqrt(Applied)) %>%
  arrange(Department)
df.UCB
```

```
##    Department Admitted Applied Gender    Percent        MOE
## 1           a      512     825   Male 0.62060606 0.03481553
## 2           a       89     108 Female 0.82407407 0.09622504
## 3           b      353     560   Male 0.63035714 0.04225771
## 4           b       17      25 Female 0.68000000 0.20000000
## 5           c      120     325   Male 0.36923077 0.05547002
## 6           c      202     593 Female 0.34064081 0.04106508
## 7           d      138     417   Male 0.33093525 0.04897021
## 8           d      131     375 Female 0.34933333 0.05163978
## 9           e       53     191   Male 0.27748691 0.07235746
## 10          e       94     393 Female 0.23918575 0.05044333
## 11          f       22     373   Male 0.05898123 0.05177804
## 12          f       24     341 Female 0.07038123 0.05415304
```

Back to the UCB data

```
limits <- aes(ymin = Percent - MOE, ymax = Percent + MOE)
ggplot(df.UCB, aes(x = Gender, y = Percent, fill = Gender)) +
  geom_bar(stat = "identity") +
  geom_errorbar(limits, width=0.25) +
  ggtitle("Percent admission at UCB, by dept") +
  facet_wrap(~Department)+
  theme(legend.position = "none")
```



Conclusions