StatChat013019

```r
# Find raw disparities in preemptory strikes from 225 trials with race data
combo <- jurors %>%
  left_join(trials, by = c("trial__id" = "id")) %>%
  filter(race != "Unknown")

combo %>% summarise(num_trials = n_distinct(trial__id))   # 226
```

```
## # A tibble: 1 x 1
##   num_trials
##        <int>
## 1        226
```

```r
# Examine some categorical variables
combo %>% count(struck_by)
```

```
## # A tibble: 9 x 2
##   struck_by                         n
##   <chr>                         <int>
## 1 Juror chosen as alternate       240
## 2 Juror chosen to serve on jury  2462
## 3 Juror excused/absent             43
## 4 Juror not struck               2902
## 5 Struck by the defense          1467
## 6 Struck by the state            1286
## 7 Struck for cause               1342
## 8 Struck without notation         362
## 9 Unknown                          18
```

```r
combo %>% count(race)
```

```
## # A tibble: 4 x 2
##   race       n
##   <chr>  <int>
## 1 Asian      3
## 2 Black   3877
## 3 Latino     1
## 4 White   6241
```

```r
combo %>% count(defendant_race)
```

```
## # A tibble: 4 x 2
##   defendant_race     n
##   <chr>          <int>
## 1 Asian             54
## 2 Black           7819
## 3 Unknown          193
## 4 White           2056
```

```r
combo %>% count(verdict)
```

```
## # A tibble: 4 x 2
##   verdict                             n
##   <chr>                           <int>
## 1 Aquitted on all counts           1529
```

```
## 2 Guilty on at least one offense  8045
## 3 Mistrial                         447
## 4 Unknown                          101
```

```r
combo %>% count(strike_eligibility)
```

```
## # A tibble: 5 x 2
##   strike_eligibility         n
##   <chr>                  <int>
## 1 Both State and Defense  3504
## 2 Defense                  155
## 3 n/a                     4700
## 4 Neither                  136
## 5 State                   1627
```

```r
combo %>% count(def_attny_1)
```

```
## # A tibble: 92 x 2
##    def_attny_1                 n
##    <chr>                   <int>
##  1 A.E. (Rusty) Harlow, Jr.    36
##  2 Aelicia L. Thomas           58
##  3 Alison Steiner             125
##  4 Andre de Gruy               53
##  5 Antwayn Patrick            115
##  6 Austin Vollor              199
##  7 Azki Shah                   57
##  8 B. Leon Johnson             63
##  9 Bennie L. Jones, Jr.        70
## 10 Bernard C. Jones, Jr.       98
## # ... with 82 more rows
```

```r
combo %>% count(offense_title_1)
```

```
## # A tibble: 81 x 2
##    offense_title_1                              n
##    <chr>                                    <int>
##  1 Accessory after the fact of murder          51
##  2 Aggravated assault                        1047
##  3 Aggravated Assault                          51
##  4 Aggravated assault (attempt)                27
##  5 Aggravated driving under the influence     104
##  6 armed robbery                               46
##  7 Armed robbery                              817
##  8 Arson of state supported school building    70
##  9 Attempted building burglary                 35
## 10 Attempted burglary of a dwelling            63
## # ... with 71 more rows
```

```r
combo %>% count(cause_number)
```

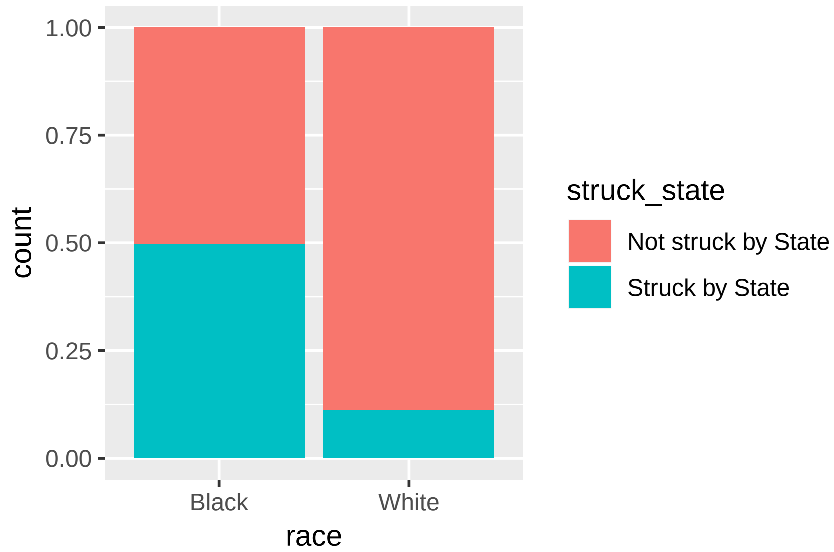
```
## # A tibble: 218 x 2
##    cause_number     n
##    <chr>        <int>
##  1 1992-2061       21
##  2 1992-2087       21
##  3 1992-4399       14
```

```
## 4 1992-9708        6
## 5 1993-2114       34
## 6 1993-2156       40
## 7 1993-2165       70
## 8 1993-3106       33
## 9 1993-4141       12
## 10 1993-4516      19
## # ... with 208 more rows
```

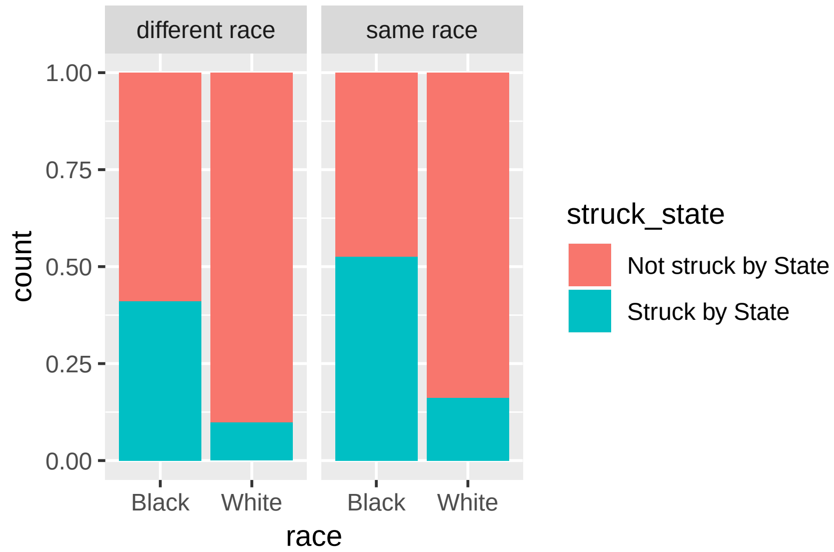
```r
# Smaller version of combo to look at a few variables that might be associated
#   with being struck by the state (assuming eligible to be struck by state)
combo_small <- combo %>%
  select(id, trial__id, struck_by, race, gender, defendant_race,
         strike_eligibility, cause_number) %>%
  filter(strike_eligibility == "Both State and Defense" |
           strike_eligibility == "State") %>%
  filter(race == "Black" | race == "White") %>%
  mutate(same_race = ifelse(race == defendant_race, "same race",
                            "different race"),
         struck_state = ifelse(struck_by == "Struck by the state",
                               "Struck by State", "Not struck by State"),
         year = parse_number(str_sub(cause_number)))

# Ratio of black prob to white prob is 4.45 = .498 / .112 (matches report)
combo_small %>%
  group_by(race) %>%
  summarise(prop_struck = mean(struck_state == "Struck by State"),
            num_struck = sum(struck_state == "Struck by State"),
            total = n())
```

```
## # A tibble: 2 x 4
##   race  prop_struck num_struck total
##   <chr>       <dbl>      <int> <int>
## 1 Black       0.498        902  1811
## 2 White       0.112        372  3318
```

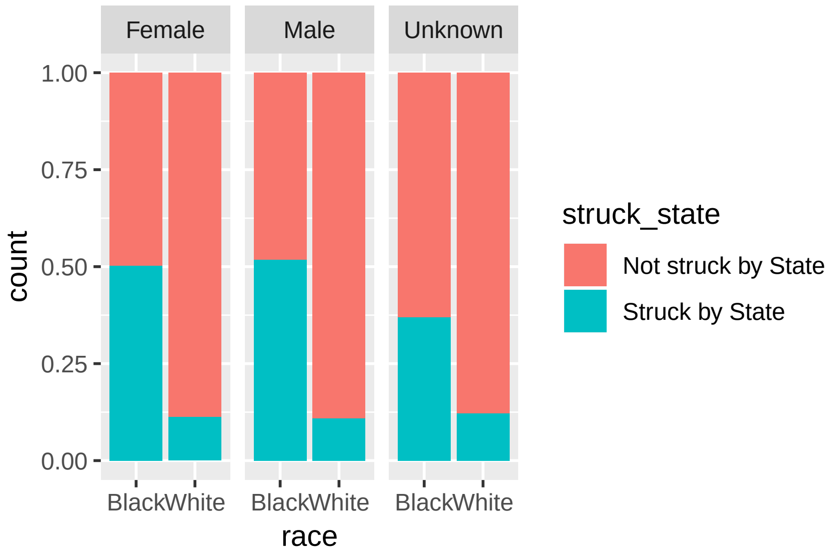
```r
ggplot(combo_small) +
  geom_bar(aes(x = race, fill = struck_state), position = "fill")
```




```
ggplot(combo_small) +
  geom_bar(aes(x = race, fill = struck_state), position = "fill") +
  facet_grid(. ~ same_race)
```

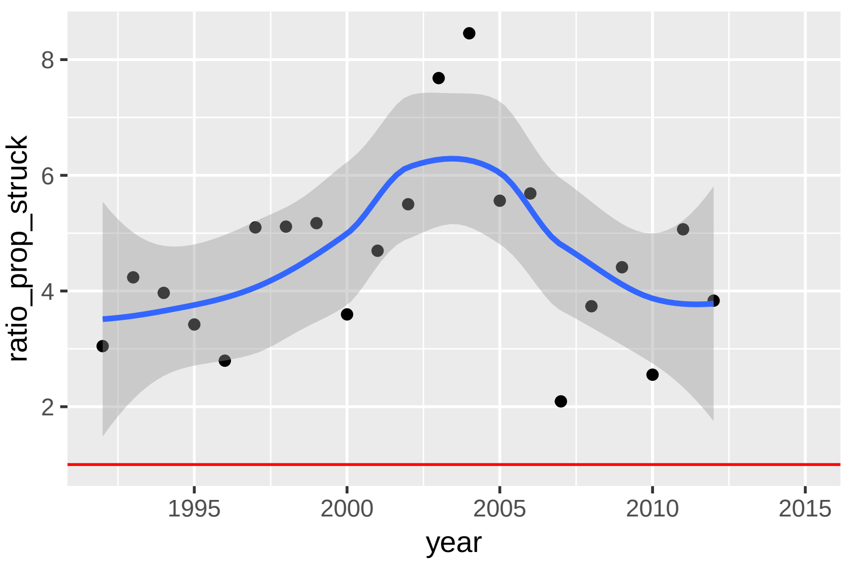



```
ggplot(combo_small) +
  geom_bar(aes(x = race, fill = struck_state), position = "fill") +
  facet_grid(. ~ gender)
```

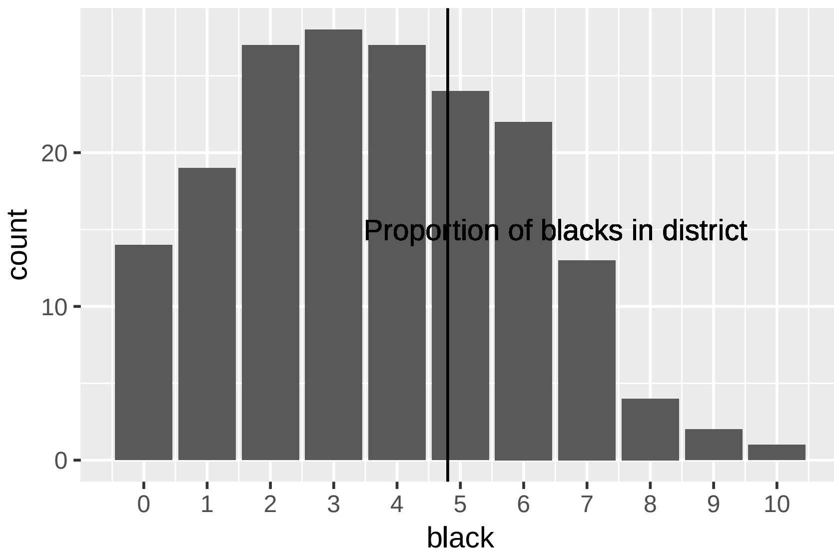



```
# Look at patterns over time in strike ratio
combo_small %>%
  group_by(year, race) %>%
  summarise(prop_struck = mean(struck_state == "Struck by State")) %>%
  spread(key = "race", value = "prop_struck") %>%
  mutate(ratio_prop_struck = Black / White) %>%
  ggplot(aes(x = year, y = ratio_prop_struck)) +
    geom_point() +
    geom_smooth() +
    geom_hline(yintercept = 1, color = "red")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing missing values (geom_point).
```



```r
# Histogram of jury composition by race
combo %>%
  filter(struck_by == "Juror chosen to serve on jury") %>%
  group_by(trial__id) %>%
  summarise(jurors = n(),
            black = sum(race == "Black"),
            white = sum(race == "White")) %>%
  filter(jurors == 12) %>%
  ggplot(aes(x = black)) +
    geom_bar() +
    geom_vline(xintercept = 4.8) +
    scale_x_continuous(breaks = seq(0, 12, 1)) +
    geom_text(x = 6.5, y = 15, label = "Proportion of blacks in district")
```




```r
# Logistic regression analysis with voirdire data added

# 89 trials represented in voirdire (not counting 1 with trial_id = NA),
#   although 2 have n=1
print(voirdire %>% count(juror_id__trial__id) %>% arrange(n), n = Inf)
```

```
## # A tibble: 90 x 2
##    juror_id__trial__id     n
##                  <dbl> <int>
##  1                  78     1
##  2                 270     1
##  3                  NA     1
##  4                  18    21
##  5                  72    22
##  6                 153    22
##  7                   1    24
##  8                  13    25
##  9                  73    25
## 10                   3    26
## 11                  31    26
## 12                 101    26
```

```
## 13          24   27
## 14          30   27
## 15          21   28
## 16          27   28
## 17          57   28
## 18          96   28
## 19          19   29
## 20          53   30
## 21         109   30
## 22         134   30
## 23         164   30
## 24         201   30
## 25           4   31
## 26           5   31
## 27          15   31
## 28          16   31
## 29          39   31
## 30          60   31
## 31          64   31
## 32          68   31
## 33          93   31
## 34         107   31
## 35         132   31
## 36         177   31
## 37          62   32
## 38          71   32
## 39          91   32
## 40         133   32
## 41         169   32
## 42           9   33
## 43         168   34
## 44          14   35
## 45          48   35
## 46         175   35
## 47         191   35
## 48          55   36
## 49         197   36
## 50          70   37
## 51         100   37
## 52         139   37
## 53         166   37
## 54          65   38
## 55          85   38
## 56         108   39
## 57         110   39
## 58         104   40
## 59          28   41
## 60          33   41
## 61          92   41
## 62         162   41
## 63         140   42
## 64          90   43
## 65         128   43
## 66          42   44
```

```
## 67                   135    44
## 68                   186    44
## 69                     6    45
## 70                    12    45
## 71                     8    46
## 72                    22    46
## 73                    77    46
## 74                   189    47
## 75                   182    49
## 76                   152    51
## 77                   141    52
## 78                    26    55
## 79                    98    55
## 80                    88    57
## 81                    10    60
## 82                   144    62
## 83                   301    62
## 84                   163    64
## 85                    47    69
## 86                     7    73
## 87                   102    88
## 88                    95    97
## 89                   255   105
## 90                   268   130
```

```r
# all data for jurors in 89 trials with complete voir dire transcript
master <- voirdire %>%
  left_join(trials, by = c("juror_id__trial__id" = "id")) %>%
  filter(!is.na(juror_id__trial__id)) %>%
  left_join(jurors, by = c("juror_id" = "id"))

# Confirm 89 trials
master %>% summarise(num_trials = n_distinct(trial__id))
```

```
## # A tibble: 1 x 1
##   num_trials
##        <int>
## 1         89
```

```r
# Examine some categorical variables
master %>% count(struck_by)
```

```
## # A tibble: 9 x 2
##   struck_by                         n
##   <chr>                         <int>
## 1 Juror chosen as alternate       112
## 2 Juror chosen to serve on jury  1037
## 3 Juror excused/absent             53
## 4 Juror not struck                141
## 5 Struck by the defense           696
## 6 Struck by the state             573
## 7 Struck for cause                912
## 8 Struck without notation          20
## 9 Unknown                           1
```

```r
master %>% count(race)
```

```
## # A tibble: 3 x 2
##   race        n
##   <chr>   <int>
## 1 Black    1290
## 2 Unknown    87
## 3 White    2168
```

```r
master %>% count(defendant_race)
```

```
## # A tibble: 4 x 2
##   defendant_race     n
##   <chr>          <int>
## 1 Asian             31
## 2 Black           2792
## 3 Unknown           77
## 4 White            645
```

```r
master %>% count(strike_eligibility)
```

```
## # A tibble: 5 x 2
##   strike_eligibility         n
##   <chr>                  <int>
## 1 Both State and Defense  1569
## 2 Defense                   61
## 3 n/a                     1122
## 4 Neither                   67
## 5 State                    726
```

```r
master %>% count(accused)
```

```
## # A tibble: 2 x 2
##   accused     n
##   <lgl>   <int>
## 1 FALSE    3495
## 2 TRUE       50
```

```r
master %>% count(fam_accused)
```

```
## # A tibble: 2 x 2
##   fam_accused     n
##   <lgl>       <int>
## 1 FALSE        3078
## 2 TRUE          467
```

```r
master %>% count(death_hesitation)
```

```
## # A tibble: 2 x 2
##   death_hesitation     n
##   <lgl>            <int>
## 1 FALSE             3514
## 2 TRUE                31
```

```r
master %>% count(know_def)
```

```
## # A tibble: 2 x 2
##   know_def     n
```

```
##   <lgl>    <int>
## 1 FALSE     3079
## 2 TRUE       466
```

```r
master %>% count(fam_law_enforcement)
```

```
## # A tibble: 2 x 2
##   fam_law_enforcement     n
##   <lgl>               <int>
## 1 FALSE                2822
## 2 TRUE                  723
```

```r
# Logistic regression data (note they combine White and Unknown races)
master_logistic <- master %>%
  select(juror_id, trial__id, struck_by, race, defendant_race, accused,
         fam_accused, know_def, fam_law_enforcement, death_hesitation,
         strike_eligibility) %>%
  filter(strike_eligibility == "Both State and Defense" |
           strike_eligibility == "State") %>%
  mutate(same_race = ifelse(race == defendant_race, TRUE, FALSE),
         struck_state = ifelse(struck_by == "Struck by the state", 1, 0),
         is_black = ifelse(race == "Black", TRUE, FALSE))

# Ratio of black prob to white prob is 4.68 = .534 / .114 (matches report)
master_logistic %>%
  group_by(is_black) %>%
  summarise(prop_struck = mean(struck_state == 1),
            num_struck = sum(struck_state == 1),
            total = n())
```

```
## # A tibble: 2 x 4
##   is_black prop_struck num_struck total
##   <lgl>          <dbl>      <int> <int>
## 1 FALSE          0.114        177  1554
## 2 TRUE           0.534        396   741
```

```r
# logistic regression model
model1 <- glm(struck_state ~ accused + is_black + fam_accused +
  death_hesitation + know_def + same_race + fam_law_enforcement,
  family = binomial, data = master_logistic)
summary(model1)
```

```
##
## Call:
## glm(formula = struck_state ~ accused + is_black + fam_accused +
##     death_hesitation + know_def + same_race + fam_law_enforcement,
##     family = binomial, data = master_logistic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4693  -0.4874  -0.4107  -0.3127   2.4667
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.4307     0.1012 -24.017  < 2e-16 ***
## accusedTRUE               2.5128     0.5455   4.606 4.10e-06 ***
```

```
## is_blackTRUE              1.8972     0.1411  13.443  < 2e-16 ***
## fam_accusedTRUE           1.8476     0.1620  11.402  < 2e-16 ***
## death_hesitationTRUE      1.8243     0.5916   3.084 0.002044 **
## know_defTRUE              1.3257     0.2233   5.937 2.91e-09 ***
## same_raceTRUE             0.3603     0.1399   2.575 0.010036 *
## fam_law_enforcementTRUE  -0.5627     0.1622  -3.468 0.000524 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2579.5  on 2294  degrees of freedom
## Residual deviance: 1887.6  on 2287  degrees of freedom
## AIC: 1903.6
##
## Number of Fisher Scoring iterations: 5
```

```r
exp(coef(model1))
```

```
##             (Intercept)             accusedTRUE            is_blackTRUE
##              0.08797419             12.33917920              6.66695515
##         fam_accusedTRUE    death_hesitationTRUE            know_defTRUE
##              6.34456105              6.19872633              3.76480682
##           same_raceTRUE fam_law_enforcementTRUE
##              1.43369659              0.56968121
```

```r
exp(confint(model1))
```

```
## Waiting for profiling to be done...
```

```
##                              2.5 %    97.5 %
## (Intercept)             0.07182462  0.106826
## accusedTRUE             4.57474511 40.090813
## is_blackTRUE            5.06788006  8.815129
## fam_accusedTRUE         4.62782800  8.738251
## death_hesitationTRUE    2.01793799 20.917259
## know_defTRUE            2.43708908  5.854130
## same_raceTRUE           1.08819022  1.883933
## fam_law_enforcementTRUE 0.41209408  0.778917
```

```r
# get Wald CIs to match report
SE = summary(model1)$coefficients[,2]
beta = summary(model1)$coefficients[,1]
lower = beta - 1.96*SE
upper = beta + 1.96*SE
exp(cbind(lower, upper))
```

```
##                              lower      upper
## (Intercept)             0.07214457  0.1072771
## accusedTRUE             4.23601835 35.9430321
## is_blackTRUE            5.05587652  8.7914115
## fam_accusedTRUE         4.61821801  8.7162310
## death_hesitationTRUE    1.94411943 19.7643249
## know_defTRUE            2.43032142  5.8320559
## same_raceTRUE           1.08980275  1.8861082
## fam_law_enforcementTRUE 0.41451088  0.7829389
```