ANALYZING SURVEY DATA IN R

Summarizing quantitative data

Kelly McConville
Assistant Professor of Statistics

Summary statistics

```
NHANESraw %>%
  filter(Age >= 12) %>%
  select(DaysPhysHlthBad)
```

```
# A tibble: 14,390 x 1
   DaysPhysHlthBad
             <int>
 1               0
 2               2
 3              20
 4               2
 5               0
 6               0
 7               0
 8              NA
 9               0
10               0
# ... with 14,380 more rows
```

Mean, total, and median

```
svymean(x = ~DaysPhysHlthBad, design = NHANES_design, na.rm = TRUE)

                   mean     SE
DaysPhysHlthBad 3.3315 0.1128

svytotal(x = ~DaysPhysHlthBad, design = NHANES_design, na.rm = TRUE)

                    total       SE
DaysPhysHlthBad 7.65e+08 35784824

svyquantile(x = ~DaysPhysHlthBad, design = NHANES_design, na.rm = TRUE,
            quantiles = 0.5)

                0.5
DaysPhysHlthBad   0
```

Summarizing by group

```
svyby(formula = ~DaysPhysHlthBad, by = ~SmokeNow,
      design = NHANES_design,
      FUN = svymean, na.rm = TRUE,
      row.names = FALSE)

  SmokeNow DaysPhysHlthBad        se
1       No        3.908984 0.1996290
2      Yes        4.951750 0.2346189
```

Summarizing by group

```
svyby(formula = ~Age, by = ~SmokeNow,
              design = NHANES_design,
              FUN = svymean, na.rm = TRUE,
              keep.names = FALSE)

  SmokeNow      Age        se
1       No 54.57933 0.6249442
2      Yes 42.76574 0.4087738
```

ANALYZING SURVEY DATA IN R

Let's practice!

ANALYZING SURVEY DATA IN R

Visualizing a quantitative variable

Kelly McConville
Assistant Professor of Statistics

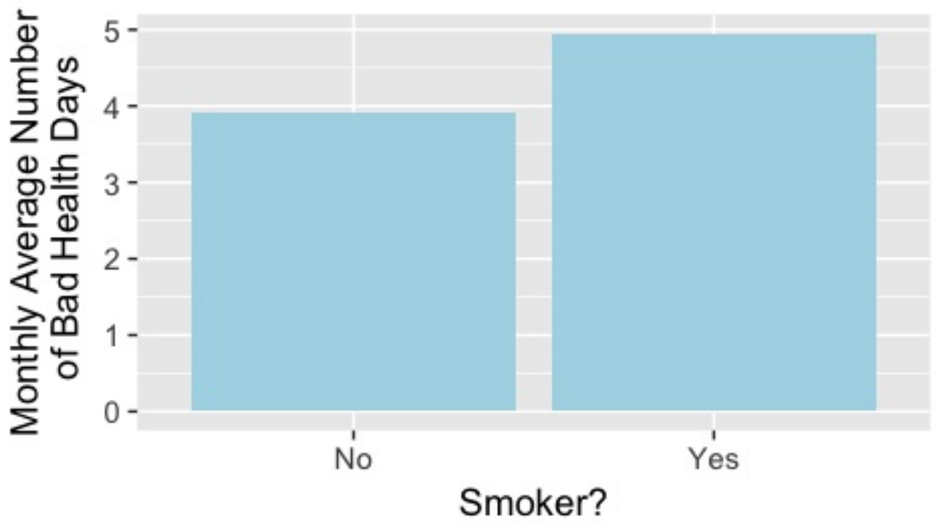
Table of means

```
out <- svyby(formula = ~DaysPhysHlthBad, by = ~SmokeNow,
             design = NHANES_design,
             FUN = svymean, na.rm = TRUE,
             keep.names = FALSE)
out
```

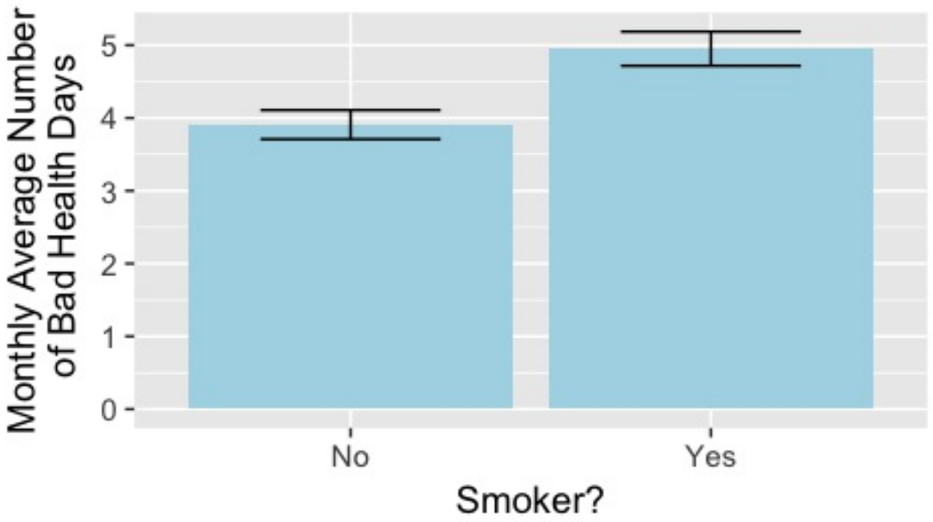
```
  SmokeNow DaysPhysHlthBad        se
1       No        3.908984 0.1996290
2      Yes        4.951750 0.2346189
```

Bar graphs

```
ggplot(data = out, mapping = aes(x = SmokeNow, y = DaysPhysHlthBad)) +
  geom_col() +
  labs(y = "Monthly Average Number\n of Bad Health Days", x = "Smoker?")
```



Bar graphs with error bars



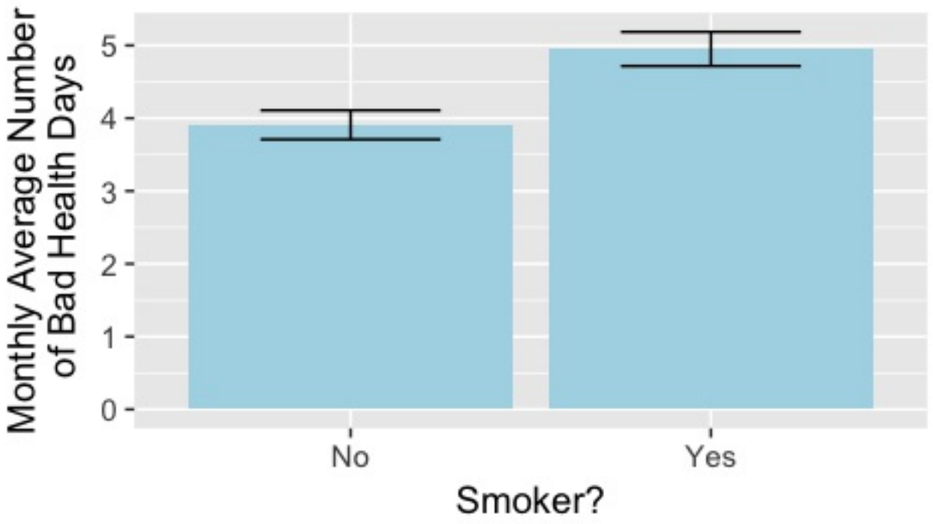
Bar graphs with error bars

```
out <- mutate(out, lower = DaysPhysHlthBad - se,
              upper = DaysPhysHlthBad + se)
out

  SmokeNow DaysPhysHlthBad        se    lower    upper
1       No        3.908984 0.1996290 3.709355 4.108613
2      Yes        4.951750 0.2346189 4.717131 5.186369
```

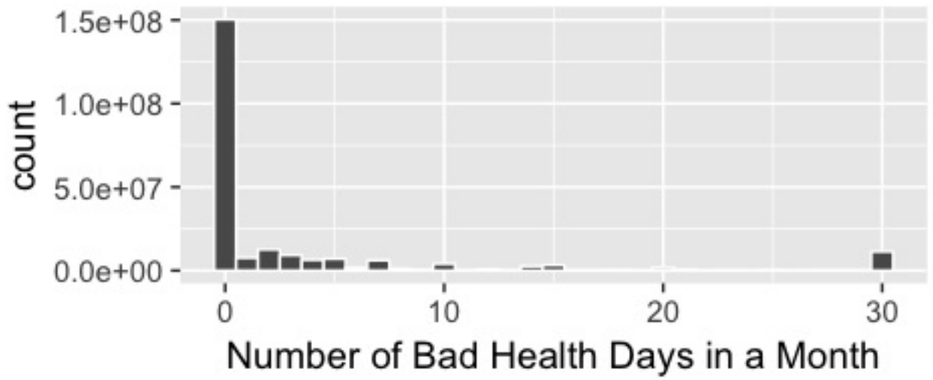
Bar graphs with error bars

```
ggplot(data = out, mapping = aes(x = SmokeNow, y = DaysPhysHlthBad,
                                 ymin = lower, ymax = upper)) +
  geom_col(fill = "lightblue") + geom_errorbar(width = .5) +
  labs(y = "Monthly Average Number\n of Bad Health Days", x = "Smoker?")
```



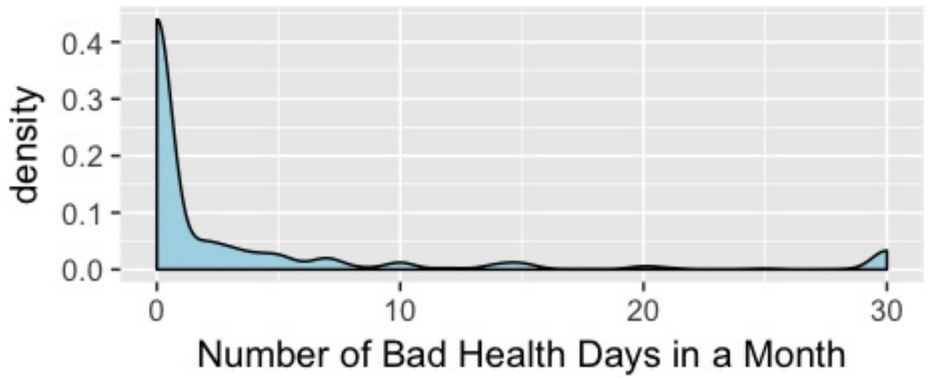
Histogram

```
ggplot(data = NHANESraw, mapping = aes(x = DaysPhysHlthBad,
                                       weight = WTMEC4YR)) +
  geom_histogram(binwidth = 1, color = "white") +
  labs(x = "Number of Bad Health Days in a Month")
```



Density plot

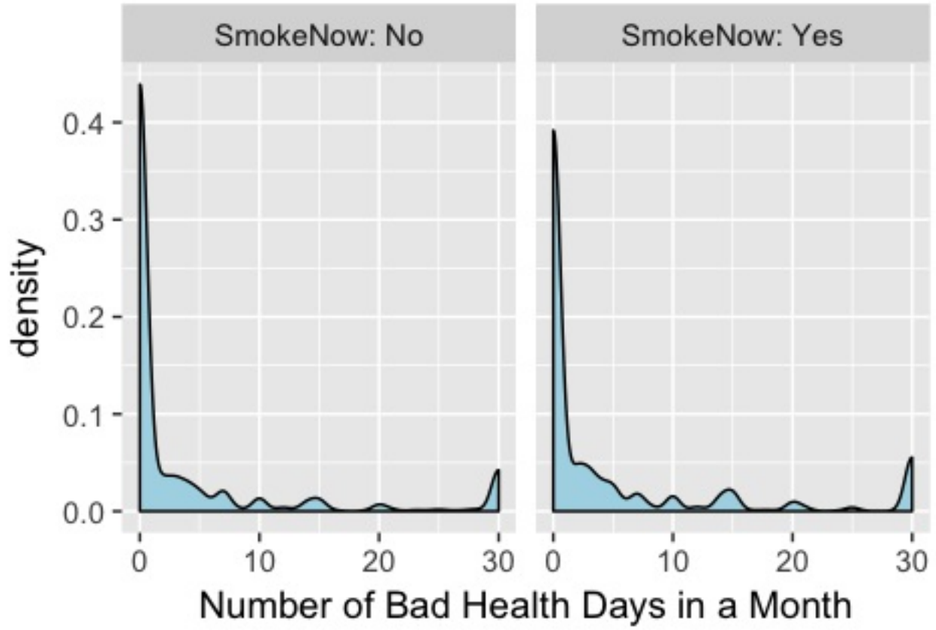
```
NHANESraw %>%
filter(!is.na(DaysPhysHlthBad)) %>%
  mutate(WTMEC4YR_std =
         WTMEC4YR/sum(WTMEC4YR)) %>%
ggplot(mapping =
         aes(x = DaysPhysHlthBad,
        weight = WTMEC4YR_std)) +
  geom_density(bw = .6,
               fill = "lightblue") +
  labs(x = "Number of Bad Health
       Days in a Month")
```



Faceted density plots

```
NHANESraw %>%
    filter(!is.na(DaysPhysHlthBad),
           !is.na(SmokeNow)) %>%
    group_by(SmokeNow) %>%
    mutate(
       WTMEC4YR_std =
       WTMEC4YR/sum(WTMEC4YR)
       ) %>%
    ggplot(mapping =
           aes(x = DaysPhysHlthBad,
           weight = WTMEC4YR_std)
           ) +
  geom_density(bw = .6,
               fill = "lightblue") +
  labs(x = "Number of Bad Health
       Days in a Month") +
  facet_wrap(~SmokeNow,
        labeller = "label_both")
```

Faceted density plots



ANALYZING SURVEY DATA IN R

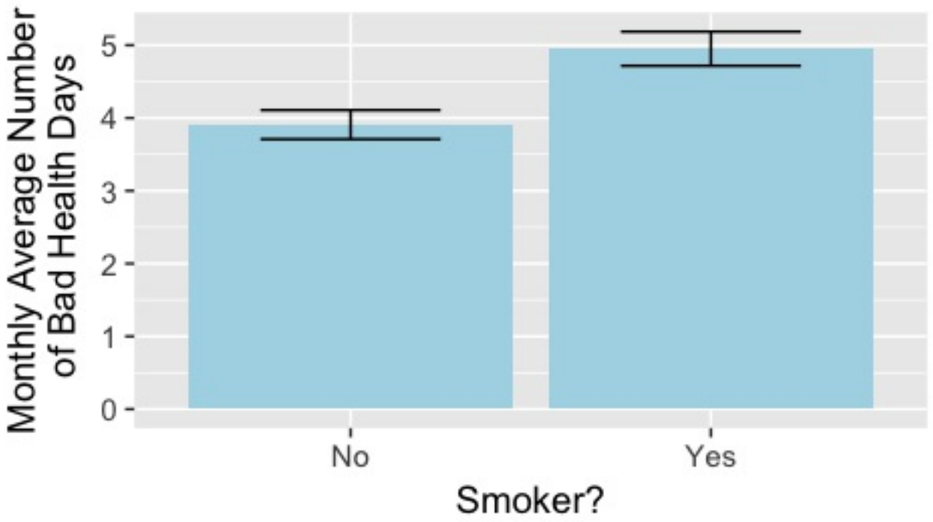
Let's practice!

ANALYZING SURVEY DATA IN R

Inference for quantitative data

Kelly McConville
Assistant Professor of Statistics

Inference for quantitative data



Survey-weighted t-test

Null Hypothesis: The monthly average number of poor health days is the same for smokers and non-smokers.

Alternative Hypothesis: The monthly average number of poor health days is different for smokers and non-smokers.

Test statistic: $t = \frac{\bar{y}_s - \bar{y}_n}{SE}$

Survey-weighted t-test

```
svyttest(formula = DaysPhysHlthBad ~ SmokeNow,
         design = NHANES_design)

Design-based t-test
data:  DaysPhysHlthBad ~ SmokeNow
t = 3.8208, df = 32, p-value = 0.0005778
alternative hypothesis: true difference in mean is not equal to 0
sample estimates:
difference in mean
          1.042766
```

ANALYZING SURVEY DATA IN R

Let's practice!