# Computing and reporting descriptive statistics

## Messy data and mean, median, and mode

**Jenine Harris**
**Brown School**

# Messy data and computing measures of central tendency

- The `PHYSHLTH` variable is the number of physically unhealthy days a survey participant has had in the last 30 days.

- On page 11 of the BRFSS codebook, the `PHYSHLTH` values of `77` and `99` are `Don't know/Not sure` and `Refused`, so could be coded as missing before examining the variable.

- It also looks like `88` is `None` for the number of unhealthy days and should be coded as zero.

```
# import brfss data
brfss.trans.2014 <- read.csv(file = "data/transgender_hc_ch2.csv")

# open tidyverse for data management
library(package = "tidyverse")
```

# Data cleaning for PHYSHLTH

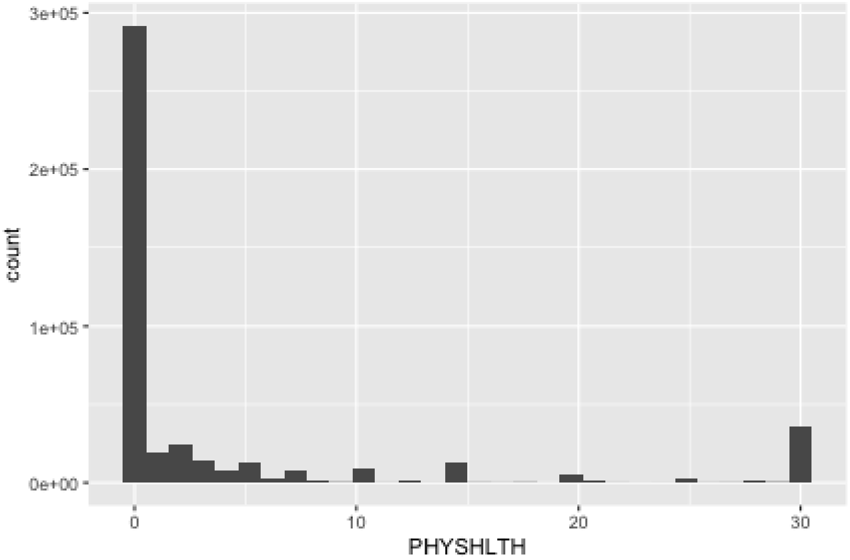
```r
# pipe in the original data frame
# recode the TRNSGNDR factor so it's easy to read
# recode 77, 88, 99 on PHYSHLTH
brfss.2014.cleaned <- brfss.trans.2014 %>%
  mutate(TRNSGNDR = recode_factor(.x = TRNSGNDR,
                                  `1` = 'Male to female',
                                  `2` = 'Female to male',
                                  `3` = 'Gender non-conforming',
                                  `4` = 'Not transgender',
                                  `7` = 'Not sure',
                                  `9` = 'Refused')) %>%
  mutate(PHYSHLTH = na_if(PHYSHLTH, 77)) %>%
  mutate(PHYSHLTH = na_if(PHYSHLTH, 99)) %>%
  mutate(PHYSHLTH = as.numeric(recode(PHYSHLTH, `88` = 0L)))

# examine PHYSHLTH to check data management
summary(object = brfss.2014.cleaned$PHYSHLTH)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.000   0.000   0.000   4.224   3.000  30.000   10303
```

# Make a histogram of PHYSHLTH

```
# make a histogram
brfss.2014.cleaned %>%
  ggplot(aes(x = PHYSHLTH)) +
  geom_histogram()
```



# Compute central tendency for PHYSHLTH

```r
# get mean, median, mode
mean(x = brfss.2014.cleaned$PHYSHLTH)
```

```
## [1] NA
```

```r
median(x = brfss.2014.cleaned$PHYSHLTH)
```

```
## [1] NA
```

```r
names(x = sort(x = table(brfss.2014.cleaned$PHYSHLTH), decreasing = TRUE
```

```
## [1] "0"
```

# Dealing with missing values

```
# get mean, median, mode
mean(x = brfss.2014.cleaned$PHYSHLTH, na.rm = TRUE)
```

```
## [1] 4.224106
```

```
median(x = brfss.2014.cleaned$PHYSHLTH, na.rm = TRUE)
```

```
## [1] 0
```

```
names(x = sort(table(brfss.2014.cleaned$PHYSHLTH), decreasing = TRUE))[1
```

```
## [1] "0"
```

# Using the tidyverse to examine central tendency

```
# get mean, median, mode
brfss.2014.cleaned %>%
  summarize(mean.days = mean(x = PHYSHLTH,
                             na.rm = TRUE),
            med.days = median(x = PHYSHLTH,
                              na.rm = TRUE),
            mode.days = names(x = sort(table(PHYSHLTH),
                                           decreasing = TRUE))[1])
```

```
##   mean.days med.days mode.days
## 1  4.224106        0         0
```

# Examine skewness

```
# skewness for PHYSHLTH
semTools::skew(object = brfss.2014.cleaned$PHYSHLTH)
```

```
##    skew (g1)           se            z            p
## 2.209078e+00 3.633918e-03 6.079054e+02 0.000000e+00
```

- `PHYSHLTH` has a skewness of 2.209078.
- After moving the decimal point 2 places to the right, z is 607.9054, which is much higher than seven.
- The graph showed a clear right skew, so there is plenty of evidence that this variable is not normally distributed.
- The **median** would be the best central tendency metric to report for PHYSHLTH.