# Computing and reporting descriptive statistics

## Frequency tables with descr & tidyverse

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

# BRFSS data import & cleaning

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

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

# cleaning the TRNSGNDR variable
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'))
```

# More frequency table options

- The **descr** (short for descriptives) package has a good option for a basic table of frequencies and percentages with the `freq()` function.
- A graph is automatically printed with the `freq()` output and the graph is not always useful, so use the `plot = FALSE` option with `freq()` to stop the graph from printing with the output.
- Because this is just one command from descr, use the `::` format to access the `freq()` function.

```
descr::freq(x = brfss.2014.cleaned$TRNSGNDR, plot = FALSE)
```

```
## brfss.2014.cleaned$TRNSGNDR
##                          Frequency   Percent Valid Percent
## Male to female                 363   0.07812       0.23562
## Female to male                 212   0.04562       0.13761
## Gender non-conforming          116   0.02496       0.07529
## Not transgender             150765  32.44603      97.85995
## Not sure                      1138   0.24491       0.73866
## Refused                       1468   0.31593       0.95286
## NA's                        310602  66.84443
## Total                       464664 100.00000     100.00000
```

# Reading the freq() output

```
## brfss.2014.cleaned$TRNSGNDR
##                          Frequency   Percent Valid Percent
## Male to female                 363   0.07812       0.23562
## Female to male                 212   0.04562       0.13761
## Gender non-conforming          116   0.02496       0.07529
## Not transgender             150765  32.44603      97.85995
## Not sure                      1138   0.24491       0.73866
## Refused                       1468   0.31593       0.95286
## NA's                        310602  66.84443
## Total                       464664 100.00000     100.00000
```

- Notice there are two columns in the output that show percentages.
- Reviewed the columns to find that the *Percent* column includes the missing data (`NA`) in the calculation of the percentage of observations in each category.
- The *Valid Percent* column removes the `NA` values and calculates the percentage of observations that falls into each category *excluding the observations missing values on this variable*.

# Making a table with the tidyverse

- **tidyverse** can also be used to make tables.

```
# use tidyverse to make table of frequency and percent
brfss.2014.cleaned %>%
  group_by(TRNSGNDR) %>%
  summarize(freq.trans = n()) %>%
  mutate(perc.trans = 100*(freq.trans / sum(freq.trans)))
```

```
## # A tibble: 7 x 3
##   TRNSGNDR                freq.trans perc.trans
##   <fct>                        <int>      <dbl>
## 1 Male to female                 363     0.0781
## 2 Female to male                 212     0.0456
## 3 Gender non-conforming          116     0.0250
## 4 Not transgender             150765    32.4
## 5 Not sure                      1138     0.245
## 6 Refused                       1468     0.316
## 7 <NA>                        310602    66.8
```

# Summarizing the information in the tables

**The 2014 BRFSS had a total of 464,664 participants. Of these, 310,602 (66.84%) were not asked or were otherwise missing a response to the transgender status question. A few participants refused to answer (n = 1,468, .32%) and a small number were unsure of their status (n = 1,138, .24%). Most reported being not transgender (n = 150,765; 32.44%), 116 were gender non-conforming (.03%), 212 were female to male (.05%), and 363 were male to female (.08%).**

- This works for reporting frequencies and percentages, but is not the *valid* percentages including just the people who responded to the trans question.
- To add the valid percent to the table, use `[ ]` to omit the `NA` from `TRNSGNDR` and calculate the valid percentages.

```
# use tidyverse to make table of frequency and percent
brfss.2014.cleaned %>%
  group_by(TRNSGNDR) %>%
  summarize(freq.trans = n()) %>%
  mutate(perc.trans = 100*(freq.trans / sum(freq.trans))) %>%
  mutate(valid.perc = 100*(freq.trans / (sum(freq.trans[na.omit(TRNSGNDR
```

# Summarizing the results

```
## # A tibble: 7 x 4
##   TRNSGNDR              freq.trans perc.trans valid.perc
##   <fct>                      <int>      <dbl>      <dbl>
## 1 Male to female               363     0.0781     0.236
## 2 Female to male               212     0.0456     0.138
## 3 Gender non-conforming        116     0.0250     0.0753
## 4 Not transgender           150765    32.4       97.9
## 5 Not sure                    1138     0.245      0.739
## 6 Refused                     1468     0.316      0.953
## 7 <NA>                      310602    66.8      202.
```

- Ignore the 202. in the last row of `valid.perc` for now, it is a tricky data management problem to delete this value to have a perfect table.

# Write the summary including the valid percentages

- Summarize the descriptive statistics again, using the valid percentages:

**The 2014 BRFSS had a total of 464,664 participants. Of these, 310,602 (66.84%) were not asked or were otherwise missing a response to the transgender status question. Of the 33.16% who responded, some refused to answer (n = 1,468, .95%) and a small number were unsure of their status (n = 1,138, .74%). Most reported being not transgender (n = 150,765; 97.86%), 116 were gender non-conforming (.08%), 212 were female to male (.14%), and 363 were male to female (.24%).**

# Achievement 2: Check your understanding

Use one of the methods shown to create a table of the frequencies for the `HADMAM` variable, which indicates whether or not each survey participant had a mammogram. Review the question and response options in the codebook and recode to ensure that the correct category labels show up in the table before you begin.

# Answer (one of several possible)

```
brfss.2014.cleaned %>%
    mutate(HADMAM = recode_factor(.x = HADMAM,
                                  `1` = 'Yes',
                                  `2` = 'No',
                                  `7` = 'Not sure',
                                  `9` = 'Refused')) %>%
  group_by(HADMAM) %>%
  summarize(freq.trans = n())
```

```
## # A tibble: 5 x 2
##   HADMAM   freq.trans
##   <fct>         <int>
## 1 Yes          204705
## 2 No            51067
## 3 Not sure        253
## 4 Refused         317
## 5 <NA>         208322
```