WELCOME TO NDACAN MONTHLY OFFICE HOURS!

NATIONAL DATA ARCHIVE ON CHILD ABUSE AND NEGLECT
DUKE UNIVERSITY, CORNELL UNIVERSITY, & UNIVERSITY OF CALIFORNIA: SAN FRANCISCO

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

- The session will begin at 11am EST
  - 11:00 - 11:30am – LeaRn with NDACAN (Introduction to R)
  - 11:30 - 12:00pm – Office hours breakout sessions
- Please submit LeaRn questions to the Q&A box
- This session is being recorded.
- See ZOOM Help Center for connection issues: https://support.zoom.us/hc/en-us
  - If issues persist and solutions cannot be found through Zoom, contact Andres Arroyo at aa17@cornell.edu.

LEARN WITH NDACAN

Presented by Frank Edwards

MATERIALS FOR THIS COURSE

- Course Box folder (https://cornell.box.com/v/LeaRn-with-R-NDACAN-2024-2025) contains
  - Data (will be released as used in the lessons)
    - Census state-level data, 2015-2019
    - AFCARS state-aggregate data, 2015-2019
    - AFCARS (FAKE) individual-level data, 2016-2019
    - NYTD (FAKE) individual-level data, 2017 Cohort
  - Documentation/codebooks for the provided datasets
  - Slides used in each week's lesson
  - Exercises as that correspond to each week's lesson
  - An .R file that will have example, usable R code for each lesson – will be updated and appended with code from each lesson

WEEK 5: DESCRIPTIVE STATISTICS

February 21, 2025

DATA USED IN THIS WEEK'S EXAMPLE CODE

- AFCARS fake aggregated ./Data/afcars_aggreg_suppressed.csv
  - Simulated aggregate data on children in foster care following the AFCARS structure
  - Can order full data from NDACAN:
    - https://www.ndacan.acf.hhs.gov/datasets/request-dataset.cfm

BASIC DESCRIPTIVE STATISTICS IN R

CENTRAL TENDENCY

- mean() computes the arithmetic mean of a vector $\frac{1}{n}\sum_{i=1}^{n} x_i$
  - Can be directly computed as sum(x) / length(x)
- median() returns the value at the 0.5 quantile of the data after arranging
  - Can also be computed as quantile(x, 0.5)

DISPERSION

- Standard deviation: sd(x)
- Variance: var(x)
- Interquartile range: quantile(x, c(0.25, 0.75))
- Minimum: min(x)
- Maximum: max(x)

CROSSTABS AND GROUPED SUMMARIES

- For univariate or bivariate crosstabs in a data.frame: table(dfx, dfy)
- For more advanced applications of grouped operations (beyond frequencies), use tidyverse group_by() %>% summarize()

WHY NOT ALL OF THEM?

- We can also just use summary() on a data.frame to obtain a good collection of descriptive statistics

OVER TO RSTUDIO

R CODE, PAGE 1 OF 3

```
#### Week 5: descriptive statistics
#### project: leaRn
#### Author: Frank Edwards
#### Email:  frank.edwards@rutgers.edu
#---------------------------------------

library(tidyverse)

afcars_demo<-read_csv("./data/afcars_aggreg_suppressed.csv")


# base R ------------------------------------------------------------------
## central tendency
# mean of exits
mean(afcars_demo$exited)

# wait what?! oh yeah, missing values!
mean(afcars_demo$exited, na.rm=T)

# how many missing values in exits?
table(is.na(afcars_demo$exited))

# how about the median?
median(afcars_demo$exited, na.rm=T)

## Variation, dispersion
# OK, mean > median, long right tail on this data
# how much variable are exits?
sd(afcars_demo$exited, na.rm=T)
var(afcars_demo$exited, na.rm=T)
```

R CODE, PAGE 2 OF 3

```
# what about the range of the data?
min(afcars_demo$exited, na.rm=T)
max(afcars_demo$exited, na.rm=T)

# How about the IQR?
quantile(afcars_demo$exited, c(0.25, 0.75), na.rm=T)

# How about the central 90% of the data
quantile(afcars_demo$exited, c(0.05, 0.95), na.rm=T)

## crosstabs
# what does our state and year coverage look like?
table(afcars_demo$fy)
table(afcars_demo$state)

# what about missing data on exits by year?
table(afcars_demo$fy, is.na(afcars_demo$exited))

# tidyverse ------------------------------------------------------------
#What if we want the mean exits for each year?
afcars_demo %>%
  group_by(fy) %>%
  summarize(exited = mean(exited, na.rm=T))
# and medians!
afcars_demo %>%
  group_by(fy) %>%
  summarize(exited_mn = mean(exited, na.rm=T),
            exited_med = median(exited, na.rm=T))
```

R CODE, PAGE 3 OF 3

```
# and standard deviations
afcars_demo %>%
  group_by(fy) %>%
  summarize(exited_mn = mean(exited, na.rm=T),
            exited_med = median(exited, na.rm=T),
            exited_sd = sd(exited, na.rm=T))

# summary will bundle many of these for us
summary(afcars_demo)

#Thanks!
```

R CODE WITH DEMO OUTPUT, PAGE 1 OF 7

```
R version 4.4.1 (2024-06-14) -- "Race for Your Life"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

 Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to
help.
Type 'q()' to quit R.

> library(tidyverse)
── Attaching core tidyverse packages ──────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts ────────────────────────── tidyverse_conflicts() ── ✖ dplyr::filter() masks
stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package to force all conflicts to become errors
> afcars_demo<-read_csv("./data/afcars_aggreg_suppressed.csv")
Rows: 4100 Columns: 10
```

R CODE WITH DEMO OUTPUT, PAGE 2 OF 7

```
── Column specification ────────────────────────────────────────────────
Delimiter: ","
chr (1): sex
dbl (9): fy, state, raceethn, numchild, phyabuse, sexabuse, negl...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> # base R
> ----------------------------------------------------------------
> ## central tendency
> # mean of exits
> mean(afcars_demo$exited)
[1] NA
> head(afcars_demo)
# A tibble: 6 × 10
     fy state sex   raceethn numchild phyabuse sexabuse neglect
  <dbl> <dbl> <chr>    <dbl>    <dbl>    <dbl>    <dbl>   <dbl>
1  2015     1 1            1     2180      352       88     568
2  2015     1 1            2     1245      198       46     331
3  2015     1 1            4       10       NA        0      NA
4  2015     1 1            5       NA        0        0      NA
5  2015     1 1            6      245       30       NA      71
6  2015     1 1            7      204       56       22      60
# ℹ 2 more variables: entered <dbl>, exited <dbl>
```

R CODE WITH DEMO OUTPUT, PAGE 3 OF 7

```
> glimpse(afcars_demo)
Rows: 4,100
Columns: 10
$ fy       <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2…
$ state    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2…
$ sex      <chr> "1", "1", "1", "1", "1", "1", "1", "2", "2", "2",…
$ raceethn <dbl> 1, 2, 4, 5, 6, 7, 99, 1, 2, 3, 4, 5, 6, 7, 99, 1,… $ numchild <dbl> 2180, 1245, 10, NA,
245, 204, NA, 2085, 1120, NA,… $ phyabuse <dbl> 352, 198, NA, 0, 30, 56, NA, 305, 217, 0, NA, 0, … $
sexabuse <dbl> 88, 46, 0, 0, NA, 22, 0, 156, 73, 0, 0, 0, 12, 40… $ neglect  <dbl> 568, 331, NA, NA, 71,
60, NA, 572, 251, NA, NA, N… $ entered  <dbl> 1073, 565, NA, NA, 87, 92, NA, 1035, 519, NA, NA,…
$ exited   <dbl> 864, 452, NA, 0, 99, 91, NA, 866, 385, 0, NA, NA,…
> # wait what?! oh yeah, missing values!
> mean(afcars_demo$exited, na.rm=T)
[1] 365.7624
> # how many missing values in exits?
> table(is.na(afcars_demo$exited))

FALSE  TRUE
3077  1023
> # how about the median?
> median(afcars_demo$exited, na.rm=T)
[1] 106
> ## Variation, dispersion
> # OK, mean > median, long right tail on this data # how much variable
> are exits?
> sd(afcars_demo$exited, na.rm=T)
[1] 686.2612
> var(afcars_demo$exited, na.rm=T)
[1] 470954.5
```

R CODE WITH DEMO OUTPUT, PAGE 4 OF 7

```
> # what about the range of the data?
> min(afcars_demo$exited, na.rm=T)
[1] 0
> max(afcars_demo$exited, na.rm=T)
[1] 7722
> # How about the IQR?
> quantile(afcars_demo$exited, c(0.25, 0.75), na.rm=T)
25% 75%
22 403
> # How about the central 90% of the data quantile(afcars_demo$exited,
> c(0.05, 0.95), na.rm=T)
   5%    95%
  0.0 1583.6
> ## crosstabs
> # what does our state and year coverage look like?
> table(afcars_demo$fy)

2015 2016 2017 2018 2019
819  817  817  824  823
> table(afcars_demo$state)

1  2  4  5  6  8  9 10 11 12 13 15 16 17 18 19 20 21 22 23 24 25
79 78 94 79 89 80 74 46 58 93 80 89 79 83 84 93 73 85 80 76 80 87
26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 44 45 46 47 48
75 80 80 80 92 78 81 72 77 70 77 80 81 90 71 80 83 74 85 67 80 99
49 50 51 53 54 55 56 72
80 65 79 85 76 81 72 51
> # what about missing data on exits by year?
> table(afcars_demo$fy, is.na(afcars_demo$exited))
```

R CODE WITH DEMO OUTPUT, PAGE 5 OF 7

```
 FALSE TRUE
 2015   626  193
 2016   608  209
 2017   624  193
 2018   612  212
 2019   607  216
> # tidyverse ----------------------------------------------------------
> # What if we want the mean exits for each year?
> afcars_demo %>%
+   group_by(fy) %>%
+   summarize(exited = mean(exited, na.rm=T))
# A tibble: 5 × 2
    fy exited
 <dbl>  <dbl>
1  2015   347.
2  2016   372.
3  2017   361.
4  2018   376.
5  2019   373.
> # and medians!
> afcars_demo %>%
+   group_by(fy) %>%
+   summarize(exited_mn = mean(exited, na.rm=T),
+             exited_med = median(exited, na.rm=T))
```

R CODE WITH DEMO OUTPUT, PAGE 6 OF 7

```
# A tibble: 5 × 3
    fy exited_mn exited_med
 <dbl>     <dbl>      <dbl>
1  2015      347.         98
2  2016      372.        109
3  2017      361.        104
4  2018      376.        116.
5  2019      373.        107
> # and standard deviations
> afcars_demo %>%
+   group_by(fy) %>%
+   summarize(exited_mn = mean(exited, na.rm=T),
+             exited_med = median(exited, na.rm=T),
+             exited_sd = sd(exited, na.rm=T))
# A tibble: 5 × 4
    fy exited_mn exited_med exited_sd
 <dbl>     <dbl>      <dbl>     <dbl>
1  2015      347.         98       679.
2  2016      372.        109       706.
3  2017      361.        104       678.
4  2018      376.        116.      686.
5  2019      373.        107       683.
```

R CODE WITH DEMO OUTPUT, PAGE 7 OF 7

```
> # summary will bundle many of these for us
> summary(afcars_demo)
       fy            state            sex              raceethn
 Min.   :2015   Min.   : 1.00   Length:4100        Min.   : 1.00
 1st Qu.:2016   1st Qu.:17.00   Class :character   1st Qu.: 2.00
 Median :2017   Median :29.00   Mode  :character   Median : 4.00
 Mean   :2017   Mean   :29.44                      Mean   :15.93
 3rd Qu.:2018   3rd Qu.:42.00                      3rd Qu.: 7.00
 Max.   :2019   Max.   :72.00                      Max.   :99.00

    numchild        phyabuse        sexabuse         neglect
 Min.   :   10   Min.   :   0.0   Min.   :   0.00   Min.   :    0
 1st Qu.:   53   1st Qu.:  11.0   1st Qu.:   0.00   1st Qu.:   31
 Median :  279   Median :  40.0   Median :  13.00   Median :  161
 Mean   : 1016   Mean   : 141.6   Mean   :  44.36   Mean   :  664
 3rd Qu.: 1104   3rd Qu.: 147.0   3rd Qu.:  46.00   3rd Qu.:  670
 Max.   :22237   Max.   :3267.0   Max.   :1290.00   Max.   :17204
 NA's   :929     NA's   :1166     NA's   :1340      NA's   :951
    entered          exited
 Min.   :   0.0   Min.   :   0.0
 1st Qu.:  30.0   1st Qu.:  22.0
 Median : 132.0   Median : 106.0
 Mean   : 426.4   Mean   : 365.8
 3rd Qu.: 459.5   3rd Qu.: 403.0
 Max.   :8603.0   Max.   :7722.0
 NA's   :1053     NA's   :1023
```