BIOSTATS 540 – Introductory Biostatistics
Fall 2022

Introduction to R
03 – Numerical Summaries

Welcome!
In any analysis of data, an important first step is to inspect the dataset structure and obtain descriptive statistics on every variable. This R illustration focuses on numerical summarization.

Dataset Used in this Illustration (right click to download):
`lung_demo.xlsx`

Packages Used in this Illustration:
`{stargazer}, {summarytools}`

#1. Highlights of Introduction to R 02 – R Essentials II

Set Your Working Directory. At the start of every session, set your working directory using either: 1) menus: SESSION > SET WORKING DIRECTORY > Choose Directory; or 2) in the console, using `setwd( )`	R needs to know where to read from and where to write to.
Work with Packages. While there is much you can do with the packages that come pre-installed when you installed R and R Studio, very often you will be using packages that are external. For these there are 2 steps: Step 1: Installation (1 time) Step 2: Load to session (each session)	To install a package, do either 1) installation from the packages pane (usually lower right); or (2) in the console by issuing the command `install.packages("packagename")`. Quotes are required. To load a package, issue the command `library(packagename)` There must be NO QUOTES.
Import Data from Excel. Often, we will be working with data that is imported from Excel.	To import a dataset from Excel, use the menus: FILE > IMPORT DATASET > FROM EXCEL Tip. Be sure to review an make your selections at lower left under "Import Options:"

#2. Introduction to the UCLA Study of Chronic Respiratory Disease

Citations:

Detels R., Coulson A, Tashkin D and Rokaw S (1975). Reliability of plethysmography, the single breath test and spirometry in population studies. *Bulletin de Physiopathologie Respiratoire*, **11**, 9-30.

Tashkin DP, Clark VA, Simmons M, Reems C, Coulson AH, Bourque LB, Sayre JW, Detels R and Rokaw S (1984). The UCLA population studies of chronic obstructive respiratory disease. VII. Relationship between parents smoking and children's lung function. *American Review of Respiratory Disease*, **129**, 891-97.

Description. The data used in this illustration is a subset of the data from UCLA study of chronic obstructive respiratory disease (CORD). The original study followed over 15,000 persons and obtained measurements of lung function (FVC and FEV1, explained below) at two points in time so that they could investigate the change in lung function in relationship to location of residence, a proxy for exposure to air pollution.

lung_demo.xlsx The data in used for this illustration (**lung_demo.xlsx**) is for the first time period only. It is a sample of n=150 ***families.*** There are 6 variables.

Key:

FVC = Forced Vital Capacity (liters). It is the amount of air that can be forcibly exhaled after taking the deepest breath possible.

FEV1 = Forced Expiratory Volume in 1 Second (liters). It is the amount of air that can be forcibly exhaled during the first second of an FVC test.

`lung_demo.xlsx` Data Dictionary

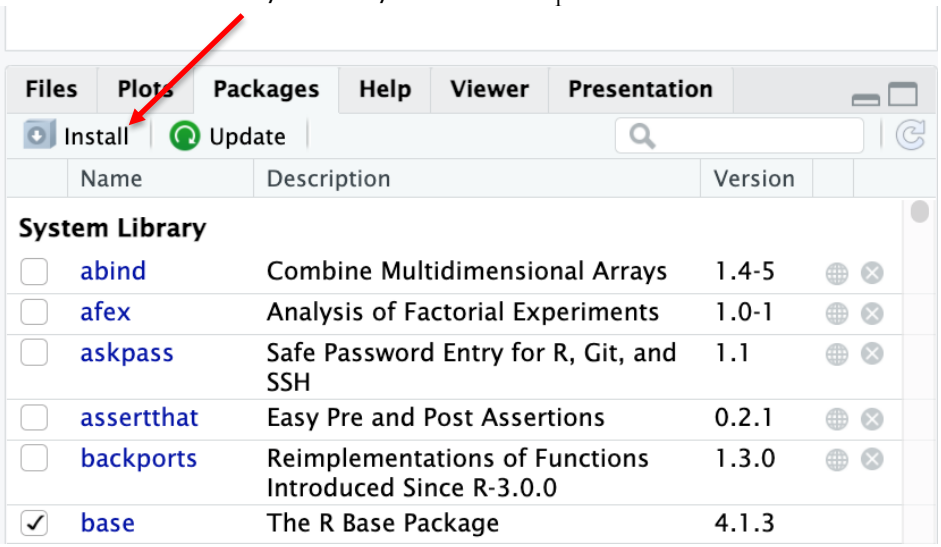
Variable	Variable Label	Units	Codes
id	Unique Study ID	-	1 to 150
area		-	1 = Burbank 2 = Lancaster 3 = Long Beach 4 = Glendora
mfvc	Forced vital capacity, mother	Continuous, liters	-
mfev1	Forced expiratory volume in 1 second, mother)	Continuous, liters	-
ocsex	Sex at Birth, oldest child		1 = Male 2 = Female
ocfvc	FVC, oldest child (liters	Continuous, liters	Continuous, liters

#3. Install Packages as Needed

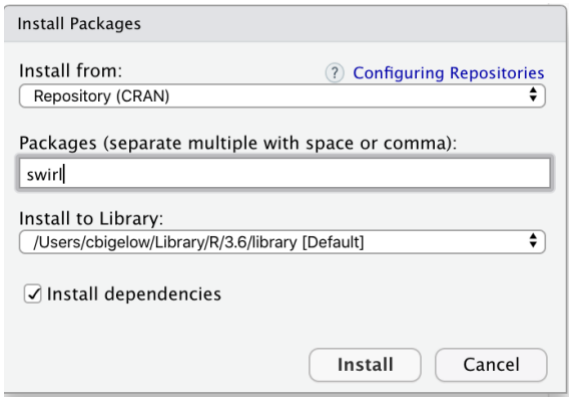
`{stargazer}`
`{summarytools}`

Recall. How to Install a Package Using R Studio Menus (recommended)
Example: `{ swirl }`

From the **FILES/PLOTS/PACKAGES** pane choose **PACKAGES** and click **Install**



Key:



In Install from: **default (Repository CRAN) is fine**

In Packages (separate multiple with space or comma:) **swirl**

In Install to Library: **leave as is**

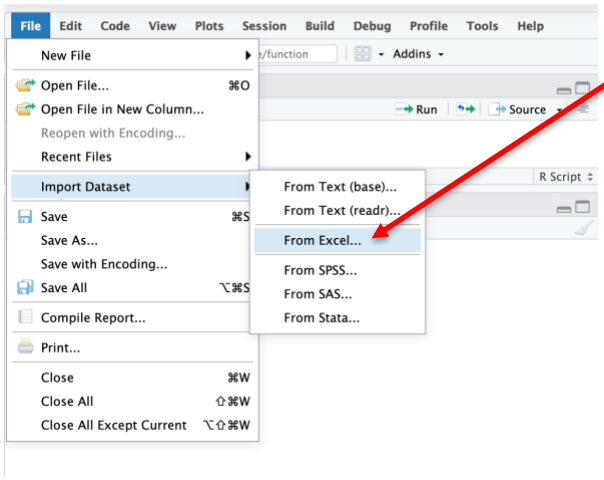
Check box for "Install dependencies": **check**

At bottom, click **Install**

#4. Import lung_demo.xlsx

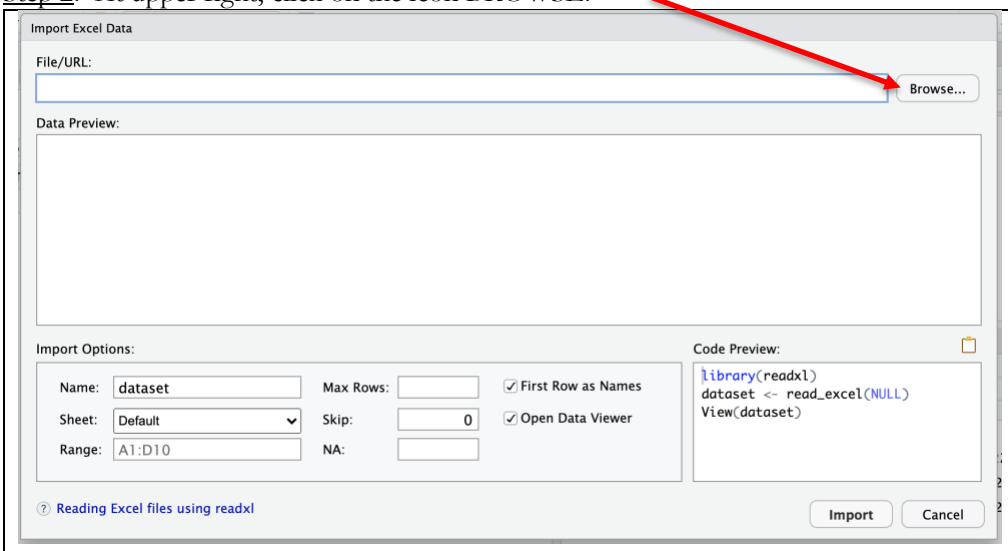
Recall. How to Import Excel Data Using R Studio Menus

Step 1: At upper left; FILE > IMPORT DATASET > FROM EXCEL



Note: R may return a message saying that you need to install readxl. Click YES. Then wait until you get a prompt.

Step 2: At upper right, click on the icon BROWSE.



Step 3: Navigate to choose lung_demos.xlsx.

Step 4: At lower right click IMPORT.

#5. Preliminary. Inspect Your Dataset Structure

```
lung_demo <- as.data.frame(lung_demo)
str(lung_demo)
'data.frame':        150 obs. of  6 variables:
 $ id   : num  1 2 3 4 5 6 7 8 9 10 ...
 $ area : num  1 1 1 1 1 1 1 1 1 1 ...
 $ mfvc : num  370 411 309 265 245 349 492 342 357 364 ...
 $ mfev1: num  331 347 265 206 233 306 425 271 313 345 ...
 $ ocsex: num  2 1 1 2 1 1 2 1 2 2 ...
 $ ocfvc: num  296 323 114 256 260 389 218 460 289 192 ...
```

#6. Introduction to Factors: How to Create a Labelled Categorical Variable

Exploration and analysis of categorical data in R involves working with R factor objects.
To work with categorical data in R rn our work (e.g., BIOSTATS 540), we think about continuous variables and categorical variables. Categorical variables are discrete and their values can be nominal or ordinal. When we want to work with what we call categorical data, we must create what R calls a factor object.

How to create a labelled categorical variable from a numeric variable

```
lung_demo$ocsexf <- factor(lung_demo$ocsex,
                           levels = c(1,2),
                           labels= c("Male","Female"))
```

Important: If you want to create a labelled categorical variable from a character (string) variable, the levels must be enclosed in quotes, as in the following example.

```
cityf <- factor(city,
                levels = c(“boston”, “seattle”),
                labels = c("Boston","Seattle"))
```

#7. Quick Summary Statistics for Every Variable

Using summary() in {base}

```
summary(lung_demo)
```

```
       id              area           mfvc           mfev1           ocsex         ocfvc      
 Min.   :  1.00   Min.   :1.00   Min.   :206.0   Min.   :175.0   Min.   :1.0   Min.   :107.0  
 1st Qu.: 38.25   1st Qu.:2.00   1st Qu.:306.0   1st Qu.:263.2   1st Qu.:1.0   1st Qu.:201.0  
 Median : 75.50   Median :3.00   Median :349.5   Median :299.0   Median :1.5   Median :291.0  
 Mean   : 75.50   Mean   :2.74   Mean   :350.2   Mean   :297.3   Mean   :1.5   Mean   :304.3  
 3rd Qu.:112.75   3rd Qu.:4.00   3rd Qu.:396.2   3rd Qu.:328.0   3rd Qu.:2.0   3rd Qu.:376.0  
 Max.   :150.00   Max.   :4.00   Max.   :567.0   Max.   :460.0   Max.   :2.0   Max.   :689.0  
    ocsexf          areaf     
 Male  :75   Burbank   :24  
 Female:75   Lancaster :49  
             Long Beach:19  
             Glendora  :58  
```

Using stargazer( ) in {stargazer}

```
library(stargazer)
stargazer(data=lung_demo, median=TRUE, type="text")
```

```
===============================================================
Statistic  N   Mean   St. Dev. Min Pctl(25) Median Pctl(75) Max
---------------------------------------------------------------
id        150 75.500   43.445   1    38.2    75.5   112.8   150
area      150  2.740   1.138    1     2       3       4      4
mfvc      150 350.233  60.427  206   306    349.5   396.2   567
mfev1     150 297.313  48.741  175  263.2    299     328    460
ocsex     150  1.500   0.502    1     1      1.5      2      2
ocfvc     150 304.253 126.393  107   201     291     376    689
---------------------------------------------------------------
```

#8. Single Variable Descriptives

Single Variable - Discrete

Using freq( ) in package {summarytools}

```
library(summarytools)
freq(lung_demo$ocsexf)
Frequencies
lung_demo$ocsexf
Type: Factor

               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
------------ ------ --------- -------------- --------- --------------
        Male     75     50.00          50.00     50.00          50.00
      Female     75     50.00         100.00     50.00         100.00
        <NA>      0                               0.00         100.00
       Total    150    100.00         100.00    100.00         100.00
```

Single Variable - Continuous

Using descr( ) in package {summarytools}

```
library(summarytools)
descr(lung_demo$mfvc)
Descriptive Statistics
lung_demo$mfvc
N: 150

                     mfvc
---------------- --------
            Mean   350.23
         Std.Dev    60.43
             Min   206.00
              Q1   306.00
          Median   349.50
              Q3   397.00
             Max   567.00
             MAD    65.98
             IQR    90.25
              CV     0.17
        Skewness     0.27
     SE.Skewness     0.20
        Kurtosis     0.45
         N.Valid   150.00
       Pct.Valid   100.00
```

Single Variable - Continuous

Using descr( ) in package {summarytools}: user chooses statistics, output is transposed

```
descr(lung_demo$mfvc,
      stats=c("n.valid","mean","sd","med","iqr"),
      transpose=TRUE)
```

```
Descriptive Statistics
lung_demo$mfvc
N: 150

             N.Valid     Mean   Std.Dev   Median     IQR
---------- --------- -------- --------- -------- -------
      mfvc    150.00   350.23     60.43   349.50   90.25
```

Nifty Trick: You can identify a set of variables of interest

```
myvars <- c("mfvc","mfev1", "ocsex")
```

Multiple Single Variables - Continuous

Using descr( ) in package {summarytools}: User chooses statistics to report and output is transposed

```
descr(lung_demo[myvars],
      stats=c("n.valid","mean","sd","med","iqr"),
      transpose=TRUE)
```

```
Descriptive Statistics
lung_demo
N: 150

              N.Valid     Mean   Std.Dev   Median     IQR
----------- --------- -------- --------- -------- -------
      mfev1    150.00   297.31     48.74   299.00   64.75
       mfvc    150.00   350.23     60.43   349.50   90.25
      ocsex    150.00     1.50      0.50     1.50    1.00
```

#9. 2 x 2 Table

2 x 2 Table

Using ctable( ) in package {summarytools}:

```
library(summarytools)
ctable(lung_demo$areaf,lung_demo$ocsexf,
                      prop="r")
```

```
Cross-Tabulation, Row Proportions
areaf * ocsexf
Data Frame: lung_demo
```

areaf	ocsexf	Male	Female	Total
Burbank		11 (45.8%)	13 (54.2%)	24 (100.0%)
Lancaster		28 (57.1%)	21 (42.9%)	49 (100.0%)
Long Beach		9 (47.4%)	10 (52.6%)	19 (100.0%)
Glendora		27 (46.6%)	31 (53.4%)	58 (100.0%)
Total		75 (50.0%)	75 (50.0%)	150 (100.0%)

#10. Single Continuous, by Group

Single Continuous Variable, by Group

Using descr( ) in package {summarytools}: User chooses statistics to report and output is transposed

```
with(lung_demo,
     stby(data=mfvc,
          INDICES=ocsexf,
          FUN=descr,
          stats=c("n.valid","min","mean","sd","max")))
```

```
Descriptive Statistics
mfvc by ocsexf
Data Frame: lung_demo
N: 75
```

	Male	Female
N.Valid	75.00	75.00
Min	206.00	215.00
Mean	350.63	349.84
Std.Dev	57.03	64.02
Max	496.00	567.00