431 Class 10

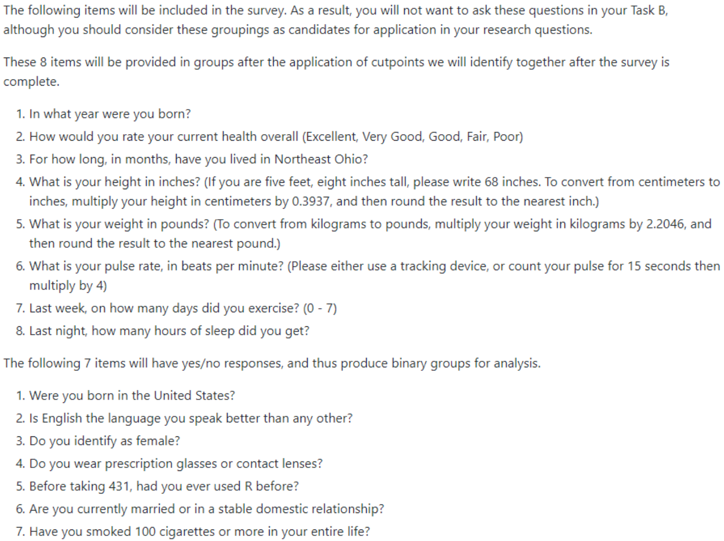
Thomas E. Love

2017-09-28

Today's Agenda

- Forming Project Task B Groups
- The Western Collaborative Group Study
    - Dealing with Factors
    - Building Tables Effectively
    - Dealing with Missingness
    - Scatterplot and Correlation Matrices

15 Questions Dr. Love plans to include in the Survey



The following items will be included in the survey. As a result, you will not want to ask these questions in your Task B, although you should consider these groupings as candidates for application in your research questions.

These 8 items will be provided in groups after the application of cutpoints we will identify together after the survey is complete.

1. In what year were you born?
2. How would you rate your current health overall (Excellent, Very Good, Good, Fair, Poor)
3. For how long, in months, have you lived in Northeast Ohio?
4. What is your height in inches? (If you are five feet, eight inches tall, please write 68 inches. To convert from centimeters to inches, multiply your height in centimeters by 0.3937, and then round the result to the nearest inch.)
5. What is your weight in pounds? (To convert from kilograms to pounds, multiply your weight in kilograms by 2.2046, and then round the result to the nearest pound.)
6. What is your pulse rate, in beats per minute? (Please either use a tracking device, or count your pulse for 15 seconds then multiply by 4)
7. Last week, on how many days did you exercise? (0 - 7)
8. Last night, how many hours of sleep did you get?

The following 7 items will have yes/no responses, and thus produce binary groups for analysis.

1. Were you born in the United States?
2. Is English the language you speak better than any other?
3. Do you identify as female?
4. Do you wear prescription glasses or contact lenses?
5. Before taking 431, had you ever used R before?
6. Are you currently married or in a stable domestic relationship?
7. Have you smoked 100 cigarettes or more in your entire life?

Project Task B groups

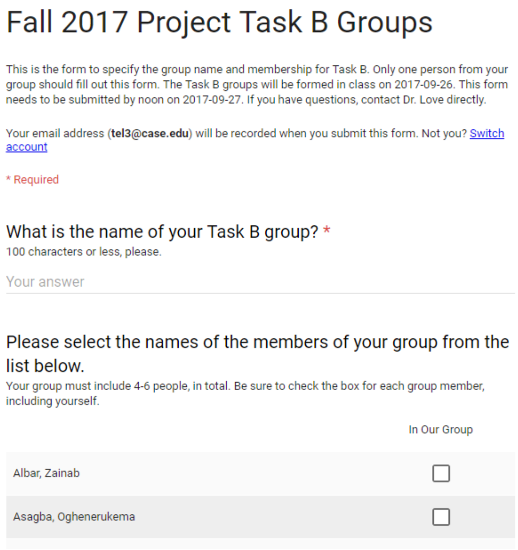
We need ten such groups, each with about 5 people involved.

Google Form is available at https://goo.gl/forms/WaQOdCEAW0wxdjJh2 and needs to be done by 5 PM today.

Task B meetings in class will be held next Tuesday, and also on 2017-10-12.

Details on Task B specified at https://github.com/thomaselove/431project

The Form's Questions. . .



Fall 2017 Project Task B Groups

This is the form to specify the group name and membership for Task B. Only one person from your group should fill out this form. The Task B groups will be formed in class on 2017-09-26. This form needs to be submitted by noon on 2017-09-27. If you have questions, contact Dr. Love directly.

Your email address (**tel3@case.edu**) will be recorded when you submit this form. Not you? Switch account

* Required

What is the name of your Task B group? *

100 characters or less, please.

Your answer

Please select the names of the members of your group from the list below.

Your group must include 4-6 people, in total. Be sure to check the box for each group member, including yourself.

	In Our Group
Albar, Zainab	□
Asagba, Oghenerukema	□

Today's R Setup

```
library(Epi); library(viridis); library(broom)
library(GGally); library(mice)
library(forcats); library(tidyverse)

source("Love-boost.R")
```

Cleaning up Loose Ends in the VHL Study

```
VHL <- read.csv("vonHippel-Lindau.csv") %>% tbl_df
```

Von Hippel - Lindau study Codebook

- `p.ne` = plasma norepinephrine (pg/ml)
- `tumorvol` = tumor volume (ml)
- `disease` = 1 for patients with multiple endocrine neoplasia type 2
- `disease` = 0 for patients with von Hippel-Lindau disease

We want to add a new variable (factor) called `diagnosis`, which takes the values `von H-L` or `neoplasia`.

Creating a Factor to represent disease diagnosis

We want to add a new variable, specifically a factor, called `diagnosis`, which will take the values `von H-L` or `neoplasia`.

- Recall `disease` is a numeric 1/0 variable (0 = von H-L, 1 = neoplasia)
- Use `fct_recode` from the `forcats` package…

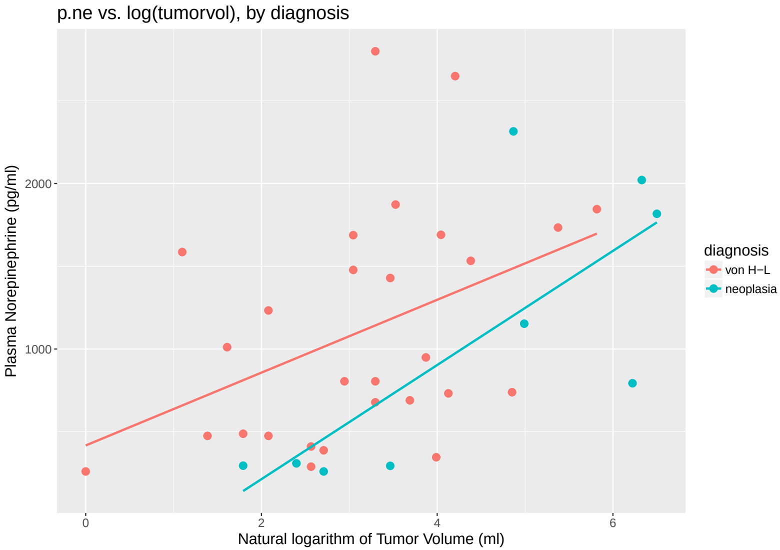
```
VHL <- VHL %>%
  mutate(diagnosis = fct_recode(factor(disease),
                                "neoplasia" = "1",
                                "von H-L" = "0")
  )
```

Now, what does VHL look like?

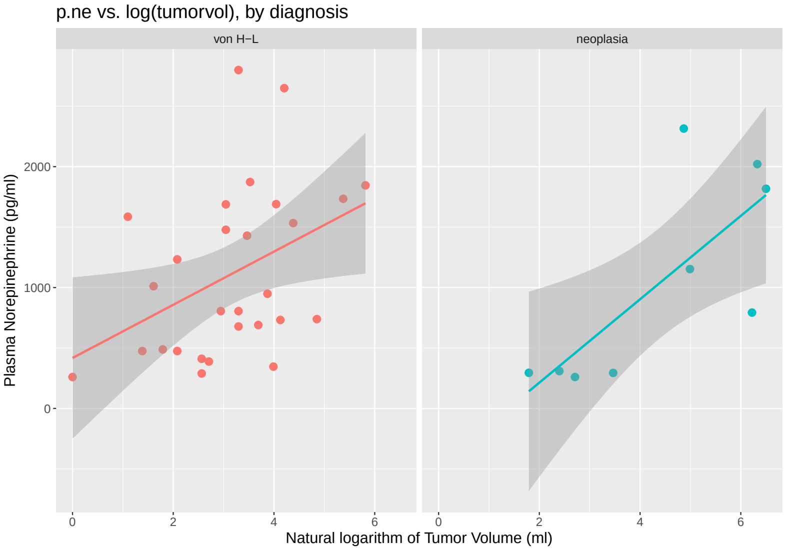
```
VHL
```

```
# A tibble: 37 x 5
      id disease  p.ne tumorvol diagnosis
   <int>   <int> <int>    <int>    <fctr>
 1   101       0   289       13   von H-L
 2   102       1   294       32 neoplasia
 3   103       0  2799       27   von H-L
 4   104       0  2649       67   von H-L
 5   105       0   346       54   von H-L
 6   106       0  1690       57   von H-L
 7   107       0   805       19   von H-L
 8   108       1  1153      147 neoplasia
 9   109       0   678       27   von H-L
10   110       1  1817      665 neoplasia
# ... with 27 more rows
```

Compare the patients by diagnosis



Facetted Scatterplots by diagnosis



Model accounting for different slopes and intercepts

```r
model2 <- lm(p.ne ~ log(tumorvol) * diagnosis, data = VHL)
model2
```

```

Call:
lm(formula = p.ne ~ log(tumorvol) * diagnosis, data = VHL)

Coefficients:
                     (Intercept)
                           417.2
                   log(tumorvol)
                           220.0
              diagnosisneoplasia
                          -893.3
log(tumorvol):diagnosisneoplasia
                           124.8
```

Model 2 results

`p.ne` = 417 + 220 log(`tumorvol`) - 893 (`diagnosis = neoplasia`) + 125 (`diagnosis = neoplasia`)*log(`tumorvol`)

where the indicator variable (`diagnosis = neoplasia`) = 1 for neoplasia subjects, and 0 for other subjects. . .

- Model for `p.ne` in von H-L patients:
    - 417 + 220 log(`tumorvol`)
- Model for `p.ne` in neoplasia patients:
    - (417 - 893) + (220 + 125) log(`tumorvol`)
    - -476 + 345 log(`tumorvol`)

Model 2 Predictions

What is the predicted `p.ne` for a single new subject with `tumorvol` = 55 ml (so log(tumorvol) = 4.01) in each diagnosis category?

```
predict(model2, newdata = data_frame(tumorvol = 55,
        diagnosis = "neoplasia"), interval = "prediction")
```

```
       fit       lwr      upr
1 905.7322 -456.1596 2267.624
```

```
predict(model2, newdata = data_frame(tumorvol = 55,
        diagnosis = "von H-L"), interval = "prediction")
```

```
       fit       lwr      upr
1 1299.003 -23.21001 2621.215
```

The broom package

```r
tidy(model2)
```

```
                              term  estimate std.error
1                      (Intercept)  417.2040 317.98858
2                    log(tumorvol)  220.0463  93.57335
3               diagnosisneoplasia -893.3017 658.56474
4 log(tumorvol):diagnosisneoplasia  124.7791 154.53279
   statistic    p.value
1  1.3120092 0.19857086
2  2.3515913 0.02481442
3 -1.3564372 0.18416834
4  0.8074602 0.42518351
```

The broom package

```
newdat <- augment(model2)

head(newdat, 2)
```

```
  p.ne log.tumorvol. diagnosis  .fitted  .se.fit
1  289      2.564949   von H-L 981.6115 131.6002
2  294      3.465736 neoplasia 718.9758 238.3433
     .resid        .hat   .sigma     .cooksd .std.resid
1 -692.6115 0.04312915 631.2164 0.01406830 -1.1173571
2 -424.9758 0.14146982 638.3783 0.02158137 -0.7237946
```

The broom package

```
glance(model2)
```

```
  r.squared adj.r.squared    sigma statistic
1 0.2903844     0.2258739 633.6815  4.501351
      p.value df    logLik      AIC      BIC deviance
1 0.009374592  4 -289.0914 588.1827 596.2373 13251224
  df.residual
1          33
```

The Western Collaborative Group Study

Original description: 3524 men aged 39-59 and employed in the San Francisco Bay or Los Angeles areas were enrolled in 1960 and 1961. In addition to determinations of behavior pattern, the initial examination included medical and parental history, socioeconomic factors, exercise, diet, smoking, alcohol consumption, diet, serum lipid and lipoprotein studies, blood coagulation studies, and cardiovascular examination.

- http://www.epi.umn.edu/cvdepi/study-synopsis/western-collaborative-group-study/

The WCGS data describe 3,154 subjects and 22 variables. For now, let's examine a few interesting variables, and a sample of 500 of the observations.

```
wcgs.full <- read.csv("wcgs.csv")
wcgs.full <- tbl_df(wcgs.full)
```

Select the variables of interest, and sample 500 subjects

```r
set.seed(43101)
wcgs1 <-
  wcgs.full %>%
  select(id, age, chol, arcus, dibpat, bmi,
         wghtcat, smoke, ncigs, chd69) %>%
  sample_n(500, replace = FALSE)
```

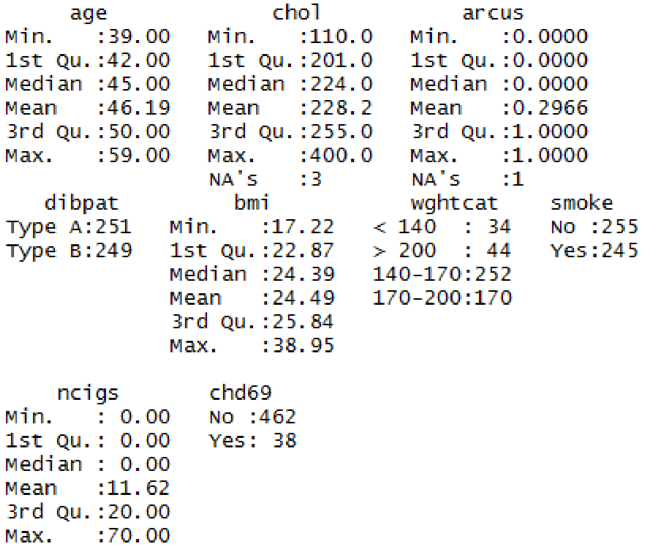
Resulting tibble

```
# A tibble: 500 x 10
      id   age  chol arcus dibpat       bmi wghtcat
   <int> <int> <int> <int> <fctr>     <dbl>  <fctr>
 1  6039    42   218     0 Type A 25.05680 140-170
 2  3713    44   207     0 Type B 22.87093 170-200
 3  3465    50   249     0 Type A 24.38980 140-170
 4  3923    42   236     0 Type B 25.10714 170-200
 5  3503    46   247     1 Type A 21.47629 140-170
 6 13167    40   206     0 Type B 26.45372 170-200
 7 11325    39   267     1 Type B 23.56510 140-170
 8  3800    51   214     0 Type A 23.67245 140-170
 9 12582    43   235     1 Type A 19.25676 140-170
10 19096    48   234     0 Type B 28.12608 170-200
# ... with 490 more rows, and 3 more variables:
#   smoke <fctr>, ncigs <int>, chd69 <fctr>
```

Codebook

Name	Stored As	Type	Details (units, levels, etc.)
`id`	integer	(nominal)	ID #, nominal and uninteresting
`age`	integer	quantitative	age, in years - no decimal places
`chol`	integer	quantitative	total cholesterol, mg/dL
`arcus`	integer	(nominal)	arcus senilis present (1) or absent (0)
`dibpat`	factor (2)	(binary)	behavioral pattern: A or B
`bmi`	number	quantitative	body-mass index
`wghtcat`	factor (4)	(ordinal)	wt: $<$ 140, 140-170, 170-200, $>$ 200
`smoke`	factor (2)	(binary)	cigarette smoker: Yes or No
`ncigs`	integer	quantitative	number of cigarettes smoked per day
`chd69`	factor (2)	(binary)	CHD event: Yes or No

Summary of this sample (without id)



```
      age              chol           arcus       
 Min.   :39.00   Min.   :110.0   Min.   :0.0000  
 1st Qu.:42.00   1st Qu.:201.0   1st Qu.:0.0000  
 Median :45.00   Median :224.0   Median :0.0000  
 Mean   :46.19   Mean   :228.2   Mean   :0.2966  
 3rd Qu.:50.00   3rd Qu.:255.0   3rd Qu.:1.0000  
 Max.   :59.00   Max.   :400.0   Max.   :1.0000  
                 NA's   :3       NA's   :1       
    dibpat         bmi            wghtcat     smoke    
 Type A:251   Min.   :17.22   < 140  : 34   No :255  
 Type B:249   1st Qu.:22.87   > 200  : 44   Yes:245  
              Median :24.39   140-170:252            
              Mean   :24.49   170-200:170            
              3rd Qu.:25.84                          
              Max.   :38.95                          
                                                     
     ncigs        chd69    
 Min.   : 0.00   No :462  
 1st Qu.: 0.00   Yes: 38  
 Median : 0.00            
 Mean   :11.62            
 3rd Qu.:20.00            
 Max.   :70.00            
```

```
summary(wcgs1)[,-1]
```

A Key Research Question

Were the men with Type A behavioral patterns more likely to suffer a CHD event?

```r
table(wcgs1$dibpat, wcgs1$chd69)
```

```
        No Yes
Type A 222  29
Type B 240   9
```

What's not so great about this table?

Re-specify the factor to re-order of the levels

```r
wcgs1$chdevent <-
  factor(wcgs1$chd69, levels = c("Yes", "No"))
tab1 <- table(wcgs1$dibpat, wcgs1$chdevent)
tab1
```

```
         Yes  No
  Type A  29 222
  Type B   9 240
```

and this is **standard epidemiological format**.

Aha! A Two-by-Two Table!

```
twoby2(tab1) ## twoby2 is part of the Epi package
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Yes
Comparing : Type A vs. Type B

       Yes  No    P(Yes) 95% conf. interval
Type A  29 222    0.1155    0.0815   0.1613
Type B   9 240    0.0361    0.0189   0.0680

                                   95% conf. interval
             Relative Risk: 3.1965    1.5450   6.6134
         Sample Odds Ratio: 3.4835    1.6132   7.5221
Conditional MLE Odds Ratio: 3.4754    1.5589   8.5398
    Probability difference: 0.0794    0.0334   0.1279
```

A 4 by 2 table

```r
table(wcgs1$wghtcat, wcgs1$chdevent)
```

```
          Yes  No
  < 140     1  33
  > 200     3  41
  140-170  20 232
  170-200  14 156
```

Why is this a poor table?

An Improved 4 x 2 table

```r
wcgs1$wghtcat <- factor(wcgs1$wghtcat,
       levels = c("< 140", "140-170", "170-200", "> 200"))
tab2 <- table(wcgs1$wghtcat, wcgs1$chdevent)
knitr::kable(tab2)
```

	Yes	No
< 140	1	33
140-170	20	232
170-200	14	156
> 200	3	41

A Three-Way Table (Not Flattened)

```
table(wcgs1$dibpat, wcgs1$wghtcat, wcgs1$chdevent)
```

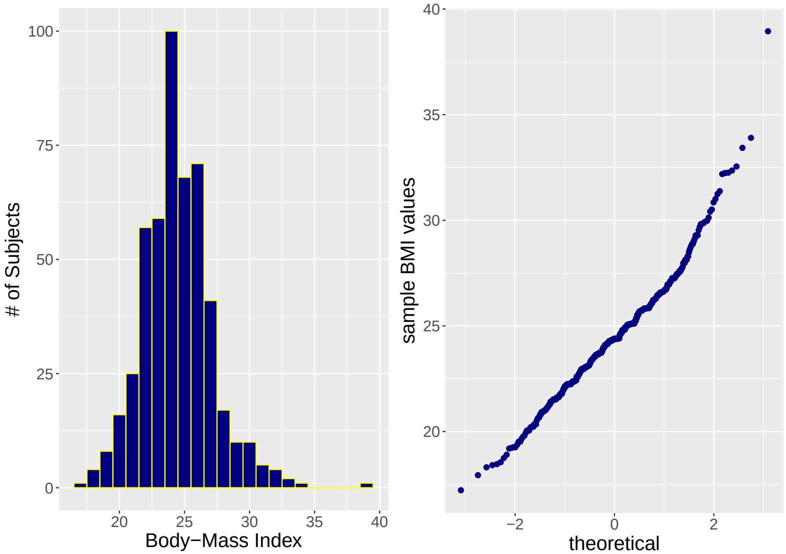
```
, ,  = Yes

         < 140 140-170 170-200 > 200
  Type A     0      15      13     1
  Type B     1       5       1     2

, ,  = No

         < 140 140-170 170-200 > 200
  Type A    15     104      82    21
  Type B    18     128      74    20
```

A Three-Way Table (Flattened)

```
ftable(wcgs1$dibpat, wcgs1$wghtcat, wcgs1$chdevent)
```

```
                  Yes  No
                         
Type A < 140        0  15
       140-170     15 104
       170-200     13  82
       > 200        1  21
Type B < 140        1  18
       140-170      5 128
       170-200      1  74
       > 200        2  20
```

How can we describe the distribution of BMI?



Summary of BMI data

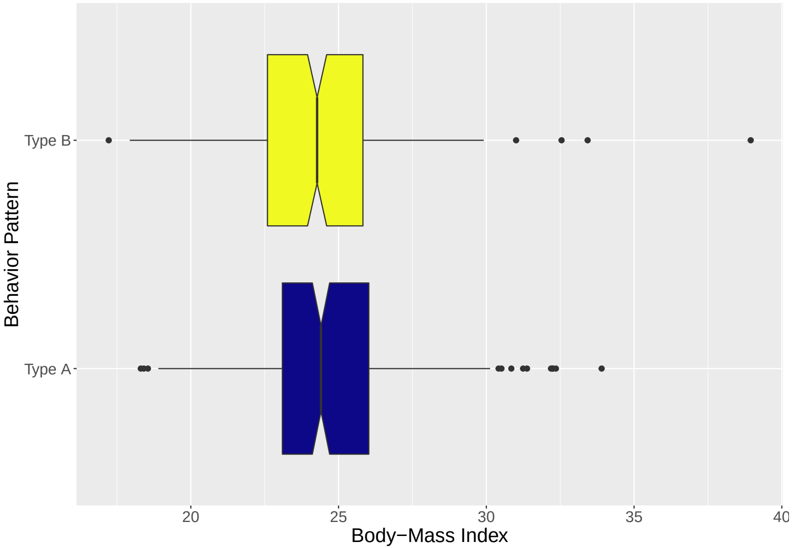
```r
mosaic::favstats(wcgs1$bmi)
```

```
      min       Q1   median       Q3      max     mean
 17.22242 22.87091 24.3898 25.84016 38.94737 24.48819
       sd   n missing
 2.680243 500       0
```

```r
psych::describe(wcgs1$bmi)
```

```
   vars   n  mean   sd median trimmed  mad   min   max
X1    1 500 24.49 2.68  24.39   24.39 2.15 17.22 38.95
   range skew kurtosis   se
X1 21.72 0.65     2.11 0.12
```

Comparing BMI by Behavior Pattern



Numerical BMI summaries, by behavior pattern

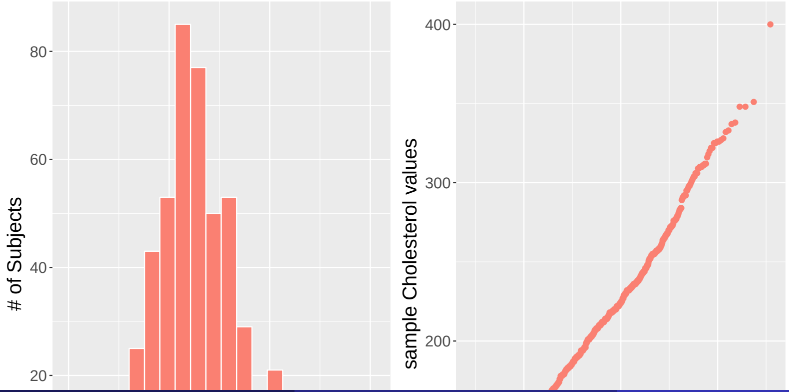
```r
by(wcgs1$bmi, wcgs1$dibpat, mosaic::favstats)
```

```
wcgs1$dibpat: Type A
      min       Q1   median       Q3      max     mean
 18.30729 23.09703 24.40514 26.02431 33.90239 24.64689
       sd   n missing
 2.700341 251       0
-----------------------------------------
wcgs1$dibpat: Type B
      min       Q1   median       Q3      max     mean
 17.22242 22.59412 24.27378 25.82449 38.94737 24.32823
      sd   n missing
 2.65565 249       0
```

How about Total Cholesterol, instead?

```
Warning: Removed 3 rows containing non-finite values
(stat_bin).
```

```
Warning: Removed 3 rows containing non-finite values
(stat_qq).
```



Why did we get error warnings?

```
mosaic::favstats(wcgs1$chol)
```

```
 min   Q1 median   Q3 max     mean       sd   n missing
 110 201     224 255 400 228.167 44.20081 497       3
```

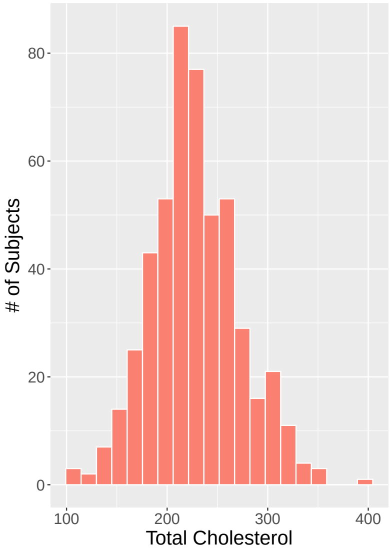
```
psych::describe(wcgs1$chol)
```

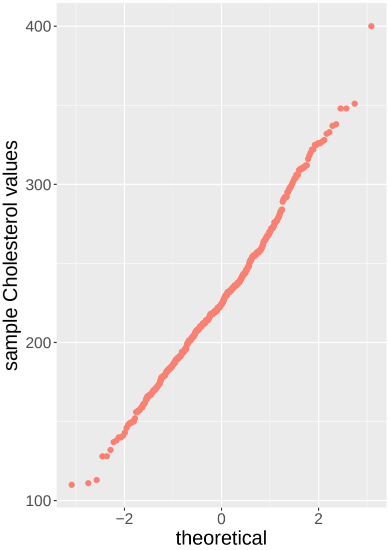
```
   vars   n   mean   sd median trimmed   mad min max
X1    1 497 228.17 44.2    224  226.64 41.51 110 400
   range skew kurtosis   se
X1   290 0.33     0.36 1.98
```

What should we do about missing values?

1. `ggplot2` doesn't include missing values in the plot, but it does warn that they've been removed.
   - We could suppress that warning by setting `na.rm = TRUE` in the call to a geom like `geom_histogram` or `geom_qq`.

Plot with `na.rm = TRUE`





Classification of Missing Data

There are three classifications we will think about in 431. Subtleties abound, but these three will suffice for most practical work.

- MCAR: Missing *Completely at Random*. This is the desirable scenario for us. MCAR means that there is no relationship between the probability of a data point being missing, and any values in the data set, either missing or observed.
    - The missing data are just a random subset of the data.
    - This is one kind of "ignorable" missingness.

Classification of Missing Data

- MAR: Missing *at Random*, which is definitely an unfortunate name. Less desirable, but there is still some hope. MAR means that the probability of a data point being missing has nothing to do with the missing value that would have been observed, but does have something to do with the values of some other variable that you did observe.
    - The idea is that if we can control for this other variable in our analysis, then we can treat this missingness as just a random subset of the data after that adjustment, which will eventually be pretty straightforward.
    - This is another kind of "ignorable" missingness.

Classification of Missing Data

- MNAR: Missing *not at Random* is a more serious issue. Now, additional thought and some special methods may well be required. Here, there is a relationship between the probability that a value is missing and what the actual (missing) value is.
  - This is what we mean by "non-ignorable" missingness.
  - Multiple Imputation methods (and Maximum Likelihood approaches) assume the data are MAR or MCAR, so the important distinction, generally, is between "at random" vs. "not at random."
- Some of these explanations come from `this link`

What should we do about missing values?

2. At times you will want to try to understand what makes observations with missing values different from observations with meaningful recorded values, especially if we're thinking that the missing mechanism is MCAR or MAR.
    - We might, for instance, compare the BMI values or perhaps the smoking status for those with and without missing cholesterol values, using the `is.na()` function to make a new variable to indicate those subjects without a cholesterol level.

- Some of this material is drawn from `R for Data Science`

For Tuesday: Working with Tables

For Class 11, I'd like you to read a 1981 paper by A.S.C. Ehrenberg that you'll find linked on the Class 10 README.

The paper is called *The Problem of Numeracy* and it provides some very helpful tips for working with tables, in particular.

There are three key tips related to the development of tables, in practice, as described by Ehrenberg, and also by Howard Wainer[1] who concisely states them as:

1. Order the rows and columns in a way that makes sense.
2. Round - a lot!
3. ALL is different and important.

[1] Visual Revelations (1997), Chapter 10.