431 Homework A Sketch

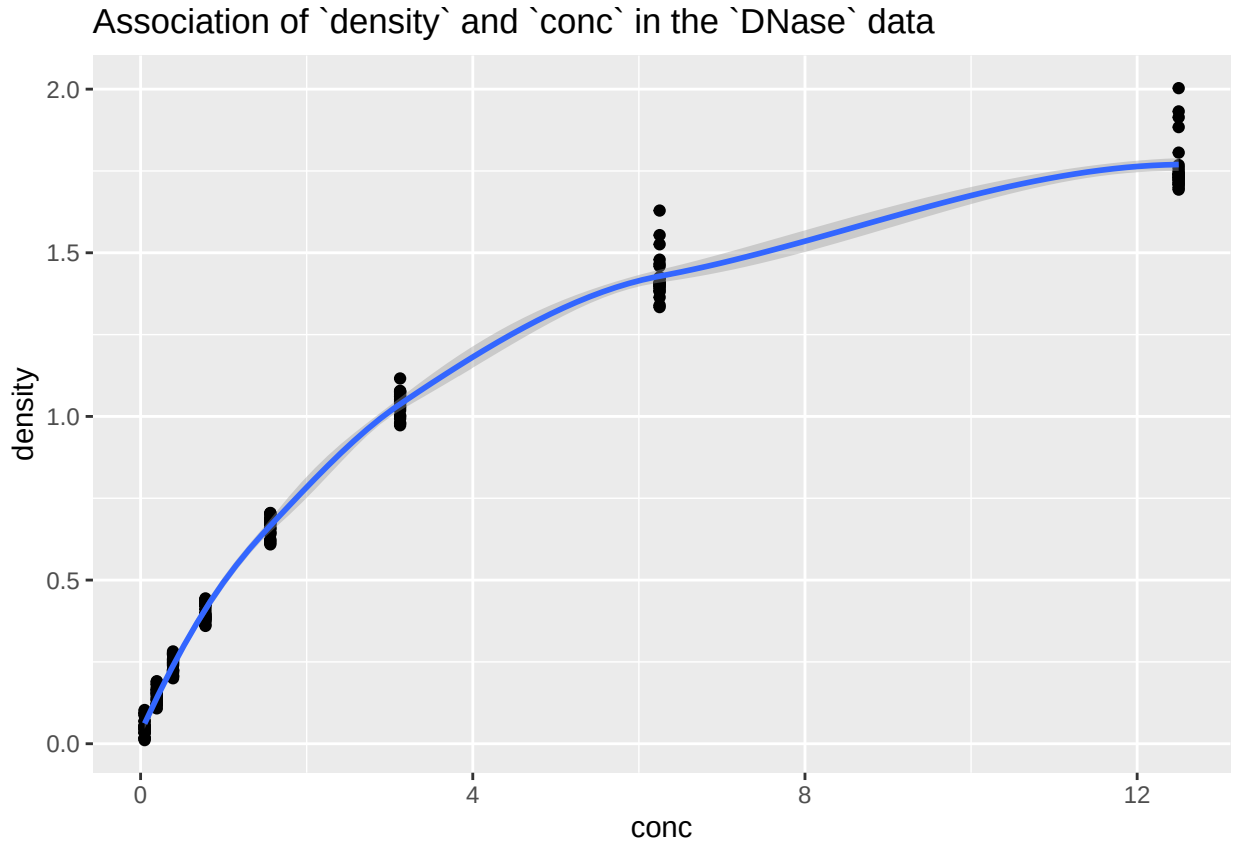
*Due **2018-08-30** at 2 PM. Last Edited 2019-09-02 21:15:23*

Activity 1. Interpreting a Visualization Built in R

Professor Love used R and the `tidyverse` to build the plot below using the `DNase` data set from the `datasets` package automatically loaded by R. Here's the plot again, and the code I used to build it.

```
ggplot(DNase, aes(x = conc, y = density)) +
    geom_point() +
    geom_smooth() +
    labs(title = "Association of `density` and `conc` in the `DNase` data")
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Use the Help window in R to learn about the `DNase` data set, and in particular, about the two variables displayed in the plot and their scientific context. Then write a paragraph (no more than 100 words) which explains what the plot indicates about the relationship between the two variables, and (more generally) what you have learned about the data (or science) from the plot.

The `DNase` data set: Help file

- *Description*: The `DNase` data frame has 176 rows and 3 columns of data obtained during development of an ELISA assay for the recombinant protein DNase in rat serum.

- The variable `conc` is a numeric vector giving the known concentration of the protein.
- The variable `density` is a numeric vector giving the measured optical density (dimensionless) in the assay. Duplicate optical density measurements were obtained.

What Were We Looking For?

We don't write answer sketches for essay questions, and that's sort of what this is. We'll likely share some excerpts written by students in the class (anonymously) later, but we can tell you what we were hoping to see.

1. We want you to write in complete, grammatically correct English sentences. We want you to make your points as clearly as possible, in your own words, not, for instance, just copy-and-pasting what's in the help file.
2. We want you to accurately describe what the graph indicates about the relationship between `conc` and `density` as shown by the data, specifically that higher concentrations of the DNase protein are associated with higher values of measured optical density in the assay.
3. We wanted you also to describe the shape of the relationship, specifically that it appeared somewhat non-linear. It appears that the impact of changing the `conc` level is a bit more substantial on `density` at lower `conc` levels than at higher levels.
4. We wanted you to remark on the nature of the experiment, that several `density` measures were taken at each `conc` level, and perhaps to suggest that the blue smooth curve follows fairly closely to the average of those `density` measures at each observed `conc` level. This explains why the points in the plot fall in vertical lines at certain `conc` levels, and do not appear at other levels - that is the design of this study.
5. It also would have been helpful to avoid suggesting any sort of causal relationship. We don't know enough about the study to even suggest that higher `conc` *caused* larger values of `density` or anything like that. Among other things, we don't know what else might influence this relationship, and we don't know what else might have been controlled for in this study.

The TAs will provide a few comments, centered around these ideas, in reaction to your paragraph. We hope this is helpful to you, as you think through future work.

Activity 2. Completing a Survey - Google Form

No real need here for an answer sketch. We're not looking for particular answers, just trying to understand where your attitudes are at the start of the class.

Attitudes toward Statistics items

Several of the items were drawn from the Attitudes Toward Statistics scale. See Wise SL (1985) The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement*, 45, 401-405.

- SA = Strongly Agree (5 points in standard coding)
- A = Agree (4 points)
- N = Neutral (3 points)
- D = Disagree (2 points)
- SD = Stongly Disagree (1 point)

There were n = 62 responses.

Standard Coded Items

Item	SA	A	N	D	SD	Mean Score
I feel that statistics will be useful to me in my profession.	51	10	0	0	1	4.77
Most people would benefit from taking a statistics course.	36	23	2	1	0	4.52
Statistics is an inseparable aspect of scientific research.	44	14	2	0	2	4.58
I am excited at the prospect of using statistics in my work.	42	16	4	0	0	4.61
One becomes a more effective "consumer" of research findings if one has some training in statistics.	36	24	2	0	0	4.55
Statistical training is relevant to my performance in my field of study.	43	17	2	0	0	4.66
Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.	8	19	18	14	3	3.24

Reverse Coded Items

The Mean Score for these items is 6 - score for the standard coded items.

- So for these items, SA = 1, and not 5, etc.

Item	SA	A	N	D	SD	Mean Score
I have difficulty seeing how statistics relates to my field of study.	2	0	0	10	50	4.71 **R**
Dealing with numbers makes me uneasy.	2	9	9	27	14	3.71 **R**
Statistical analysis is best left to the "experts" and should not be part of a typical scientist's job.	2	2	5	27	26	4.18 **R**

By Respondent

If we add up the scores (standard scoring for 7 items and reverse scoring for the other 3) and divide by 10, we get an index for each student. Results for the 62 respondents are tabulated below.

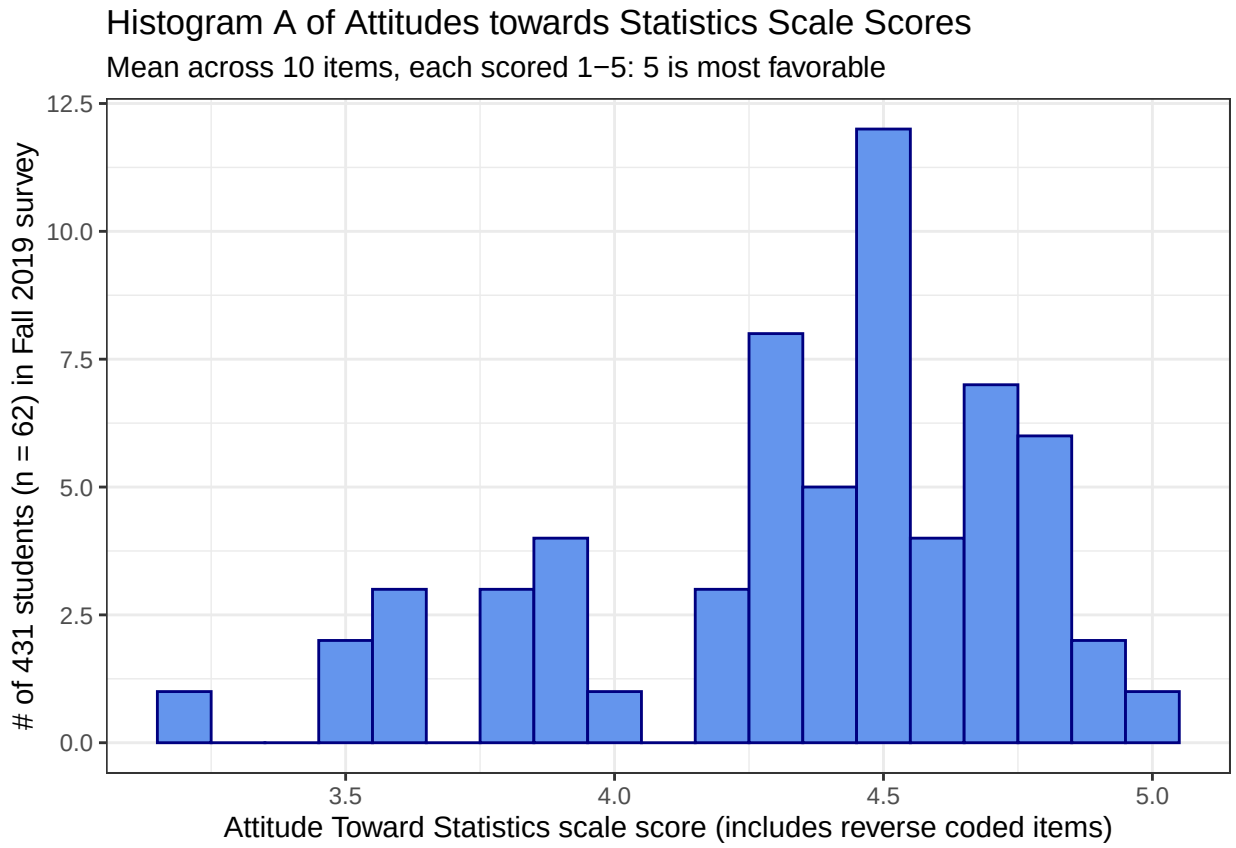
```
scores <- c(5, 4.9, 4.9, 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4.7, 4.7, 4.7, 4.7, 4.7, 4.7, 4.7,
            4.6, 4.6, 4.6, 4.6, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5,
            4.4, 4.4, 4.4, 4.4, 4.4, 4.3, 4.3, 4.3, 4.3, 4.3, 4.3, 4.3, 4.3, 4.2, 4.2, 4.2,
            4.0, 3.9, 3.9, 3.9, 3.9, 3.8, 3.8, 3.8, 3.6, 3.6, 3.6, 3.5, 3.5, 3.2)

ats <- tibble(student = 1:62, ats_score = scores)
```

Graphical Summaries

A Histogram

```
ggplot(data = ats, aes(x = ats_score)) +
    geom_histogram(binwidth = 0.1,
                   fill = "cornflowerblue",
                   col = "navy") +
    theme_bw() +
    labs(title = "Histogram A of Attitudes towards Statistics Scale Scores",
         subtitle = "Mean across 10 items, each scored 1-5: 5 is most favorable",
         x = "Attitude Toward Statistics scale score (includes reverse coded items)",
         y = "# of 431 students (n = 62) in Fall 2019 survey")
```



Should we perhaps consider smoothing out some of the granularity here, perhaps by reducing the number of bins (or increasing the width of each bin)?

As it is, the histogram is basically just this stem-and-leaf display.

A Stem-and-Leaf Display

```
stem(ats$ats_score, scale = 2)
```

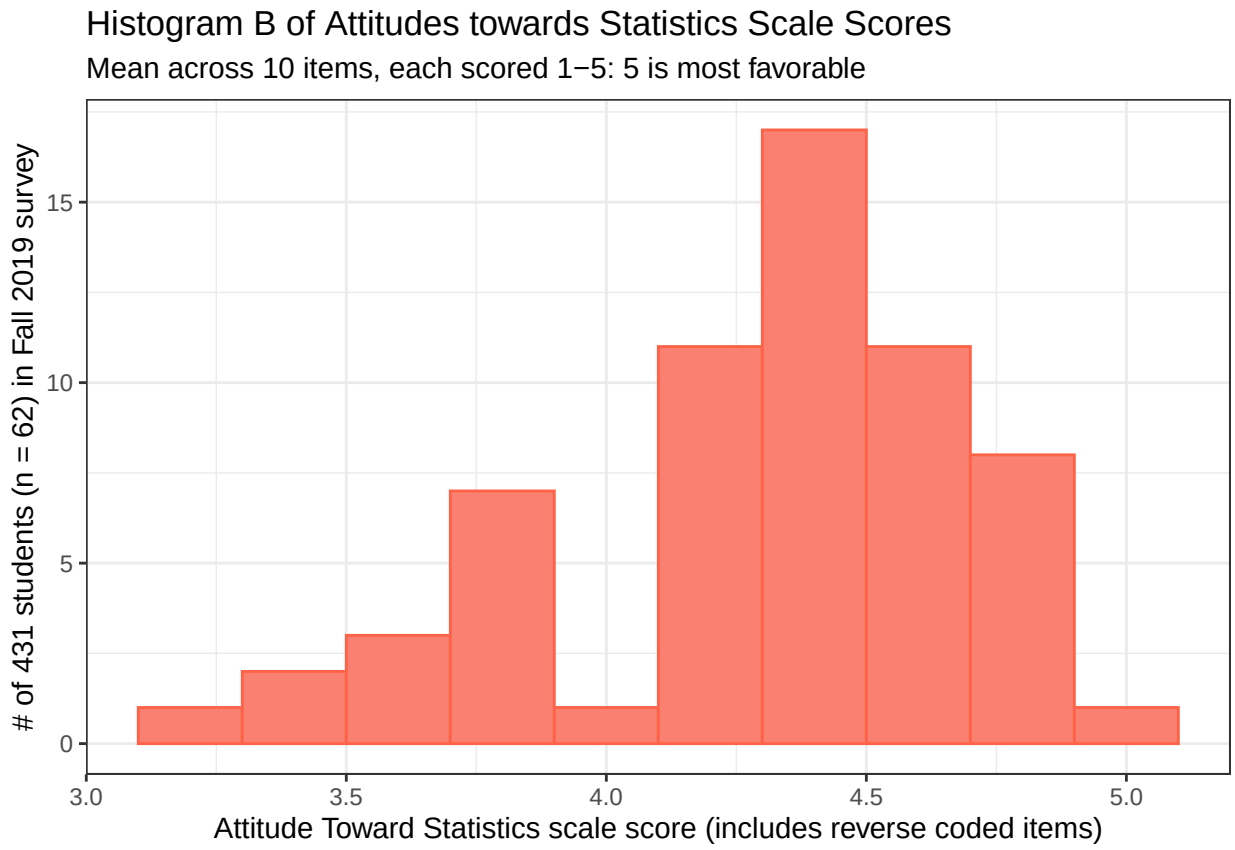
```
  The decimal point is 1 digit(s) to the left of the |

  32 | 0
  33 |
  34 |
  35 | 00
  36 | 000
  37 |
  38 | 000
  39 | 0000
  40 | 0
  41 |
  42 | 000
  43 | 00000000
  44 | 00000
  45 | 000000000000
  46 | 0000
  47 | 0000000
  48 | 000000
  49 | 00
  50 | 0
```

Revised Histogram

Here's a histogram with a larger bin width. . .

```
ggplot(data = ats, aes(x = ats_score)) +
    geom_histogram(binwidth = 0.2,
                   fill = "salmon",
                   col = "tomato") +
    theme_bw() +
    labs(title = "Histogram B of Attitudes towards Statistics Scale Scores",
         subtitle = "Mean across 10 items, each scored 1-5: 5 is most favorable",
         x = "Attitude Toward Statistics scale score (includes reverse coded items)",
         y = "# of 431 students (n = 62) in Fall 2019 survey")
```



Numerical Summaries

Using `summary()`

```
ats %>%
    select(ats_score) %>%
    summary()
```

```
   ats_score
 Min.   :3.200
 1st Qu.:4.200
 Median :4.500
 Mean   :4.353
 3rd Qu.:4.675
 Max.   :5.000
```

Using `mosaic::favstats()`

```
mosaic::favstats(~ ats_score, data = ats)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2
```

```
 min  Q1 median    Q3 max     mean        sd  n missing
 3.2 4.2    4.5 4.675   5 4.353226 0.4051787 62       0
```

Using `Hmisc::describe()`

```
ats %>%
    select(ats_score) %>%
    Hmisc::describe()
```

```
.

 1  Variables      62  Observations
---------------------------------------------------------------------------
ats_score
       n  missing distinct     Info     Mean      Gmd      .05      .10
      62        0       15    0.987    4.353   0.4458    3.600    3.800
     .25      .50      .75      .90      .95
   4.200    4.500    4.675    4.800    4.800

Value        3.2   3.5   3.6   3.8   3.9   4.0   4.2   4.3   4.4   4.5
Frequency      1     2     3     3     4     1     3     8     5    12
Proportion 0.016 0.032 0.048 0.048 0.065 0.016 0.048 0.129 0.081 0.194

Value        4.6   4.7   4.8   4.9   5.0
Frequency      4     7     6     2     1
Proportion 0.065 0.113 0.097 0.032 0.016
---------------------------------------------------------------------------
```

Using `psych::describe()`

```
psych::describe(ats$ats_score)
```

```
   vars  n mean   sd median trimmed mad min max range  skew kurtosis   se
X1    1 62 4.35 0.41    4.5    4.39 0.3 3.2   5   1.8 -0.84    -0.02 0.05
```

Using `skimr::skim()`

```
skimr::skim_with(numeric = list(hist = NULL),
                 integer = list(hist = NULL))

ats %>%
    skimr::skim()
```

```
Skim summary statistics
 n obs: 62
 n variables: 2

-- Variable type:integer -------------------------------------------------------------------------
 variable missing complete  n mean    sd p0   p25  p50   p75 p100
  student       0       62 62 31.5 18.04  1 16.25 31.5 46.75   62

-- Variable type:numeric -------------------------------------------------------------------------
  variable missing complete  n mean   sd  p0 p25 p50  p75 p100
 ats_score       0       62 62 4.35 0.41 3.2 4.2 4.5 4.68    5
```