431 Lab 01 Sketch and Grading Rubric

Last Edited 2020-09-08 12:25:40

Part 1. Video

To grade Part 1, the TAs need to be able to save the video and watch it. The videos need to meet three criteria for full credit on part 1.

- They clearly state their given name and family name, so that the viewer can learn to pronounce it correctly.
- They tell us something about themselves.
- We asked them to keep this to a maximum of 30 seconds, but as long as they're less than 45 seconds that's fine.

If you can save and watch the video and they accomplish all three of these things, 25 points. I expect that almost every student will get the full 25 points here.

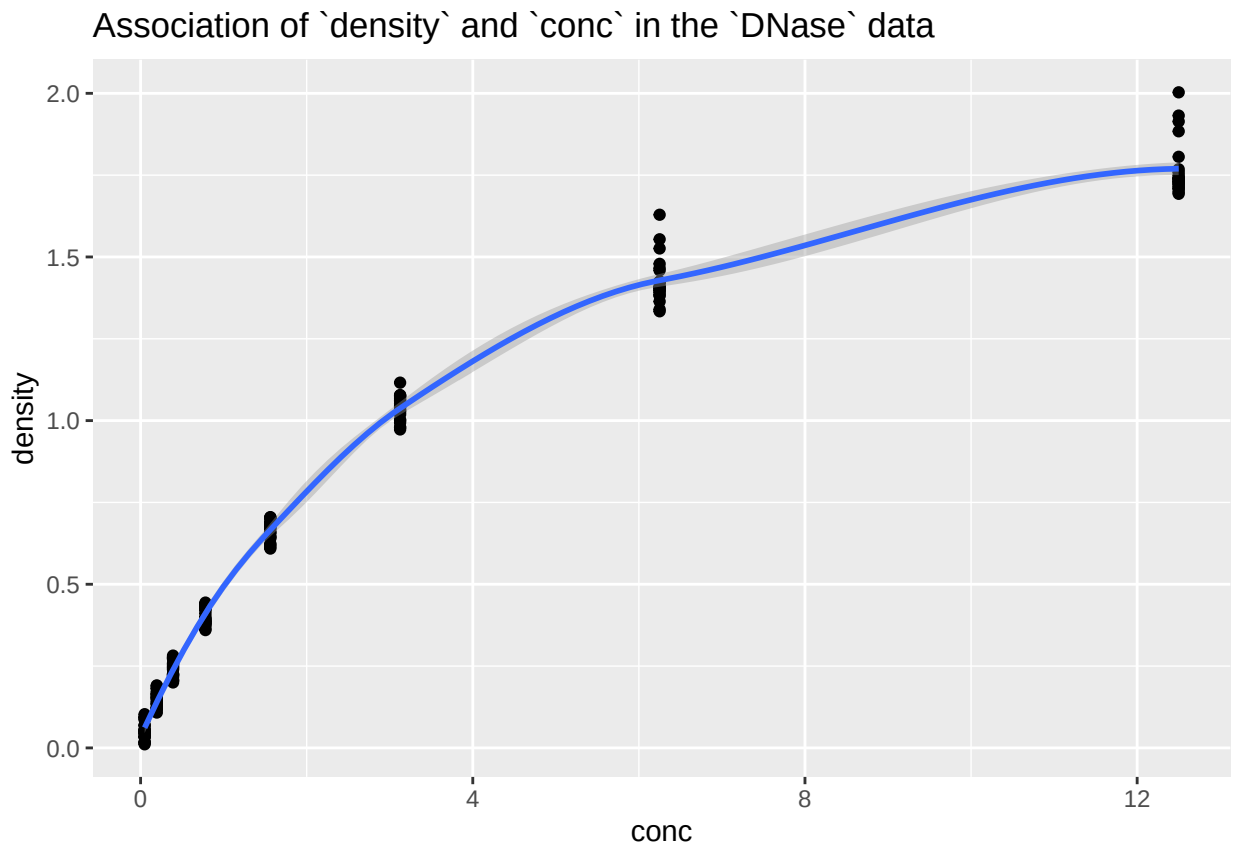
- If they don't manage to do any one of these things, they should receive 20 points
- If they don't manage to do two of these things, they should receive 15 points.
- If they don't do any of these things, they should receive 0 points.

Part 2. 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.

Grading Details

Students will receive 25 points if they've written a paragraph which explains what they learned from the plot and puts it in the context of what the two variables mean without factual inaccuracies. Again, I expect most students to get the full 25 points here.

- If they don't define what density and conc mean in context from looking at the DNase information, then they should lose 5 points for that, and should be told about it in TA comments.
- There are lots of ways in which they could describe the association in the plot. If they write something down which seems vaguely reasonable, then OK for now. If they use causal language (like writing "rising conc causes increased density") point out in TA comments that that's a stronger conclusion that can be justified here, but they shouldn't lose any points. A more tempered response like "higher levels of conc were associated with higher density" is sort of what we're aiming for. - It would be fine if they described the problems of interpolating when only a few "conc" values were studied, but if they don't talk about that, don't worry about it.
- If they write something (or more than one thing) that is definitely inaccurate, then they should be told about that in TA comments, and lose 5 points for factual inaccuracy (even if they have more than one such problem.)
- Comments should be provided about English grammar and syntax (I recognize that some TAs may be better at this than others) and in reaction to what they've discussed, but in this Lab, we won't drop any points for that.
- The paragraph was meant to be 100 words or less, but we won't count. If they've written something that looks to have exceeded that by quite a bit (200 words or more, perhaps) then they should be told about this in TA comments.

Part 3. Reaction to Spiegelhalter Introduction

We don't write answer sketches for essay questions.

Students will receive 25 points if they've written a paragraph which (a) describes a problem they are interested in solving and (b) shows some indication that they're thinking about how PPDAC (Problem-Plan-Data-Analysis-Conclusion) might be useful in the context of solving that problem. I expect, again, most students to receive 25 points.

- If they write an essay, but don't describe their problem at all, they should lose 10 points for that.
- If they write an essay, but don't indicate how PPDAC might help at all, they should lose 10 points. They certainly don't need to indicate how each piece of PPDAC might help, but at least some of the PPDAC stuff should be indicated for them to get credit.

- Detailed comments should be provided about English grammar and syntax (I recognize that some TAs may be better at this than others) and in reaction to what they've discussed, but in this Lab, we won't drop any points for that.
- The paragraph was meant to be 100 - 150 words, but we won't count. If they've written something that seems to substantially exceed this (250 words or more, perhaps) then they should be told about this in TA comments.

Activity 4. 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. People who get their response in on time will receive the full 25 points.

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)

Standard Coded Items

Item	SA	A	N	D	SD	Mean Score
I feel that statistics will be useful to me in my profession.	53	17	0	0	0	4.76
Most people would benefit from taking a statistics course.	43	19	8	0	0	4.50
Statistics is an inseparable aspect of scientific research.	47	20	3	0	0	4.63
I am excited at the prospect of using statistics in my work.	51	16	3	0	0	4.69
One becomes a more effective "consumer" of research findings if one has some training in statistics.	31	35	4	0	0	4.39
Statistical training is relevant to my performance in my field of study.	41	26	2	1	0	4.53
Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.	10	20	31	9	0	3.44

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	1	2	21	44	4.49 **R**
Dealing with numbers makes me uneasy.	1	12	15	33	9	3.53 **R**
Statistical analysis is best left to the "experts" and should not be part of a typical scientist's job.	1	2	10	32	25	4.11 **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.

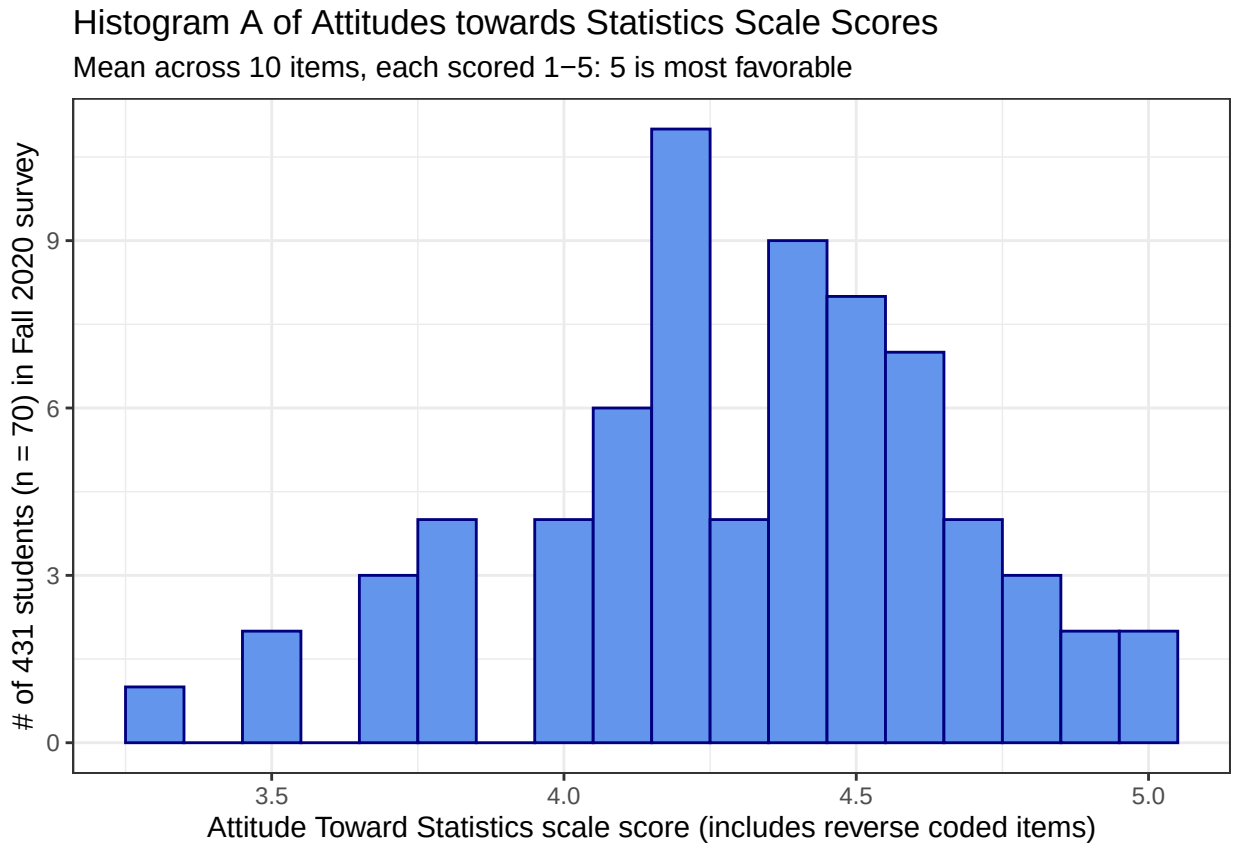
```
scores <- c(5, 5, 4.9, 4.9, 4.8, 4.8, 4.8, 4.7, 4.7, 4.7, 4.7,
            4.6, 4.6, 4.6, 4.6, 4.6, 4.6, 4.6, 4.5, 4.5, 4.5,
            4.5, 4.5, 4.5, 4.5, 4.5, 4.4, 4.4, 4.4, 4.4, 4.4,
            4.4, 4.4, 4.4, 4.4, 4.3, 4.3, 4.3, 4.3, 4.2, 4.2,
            4.2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.1,
            4.1, 4.1, 4.1, 4.1, 4.1, 4, 4, 4, 4, 3.8, 3.8, 3.8,
            3.8, 3.7, 3.7, 3.7, 3.5, 3.5, 3.3)

ats <- tibble(student = 1:70, 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 = 70) in Fall 2020 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)
```

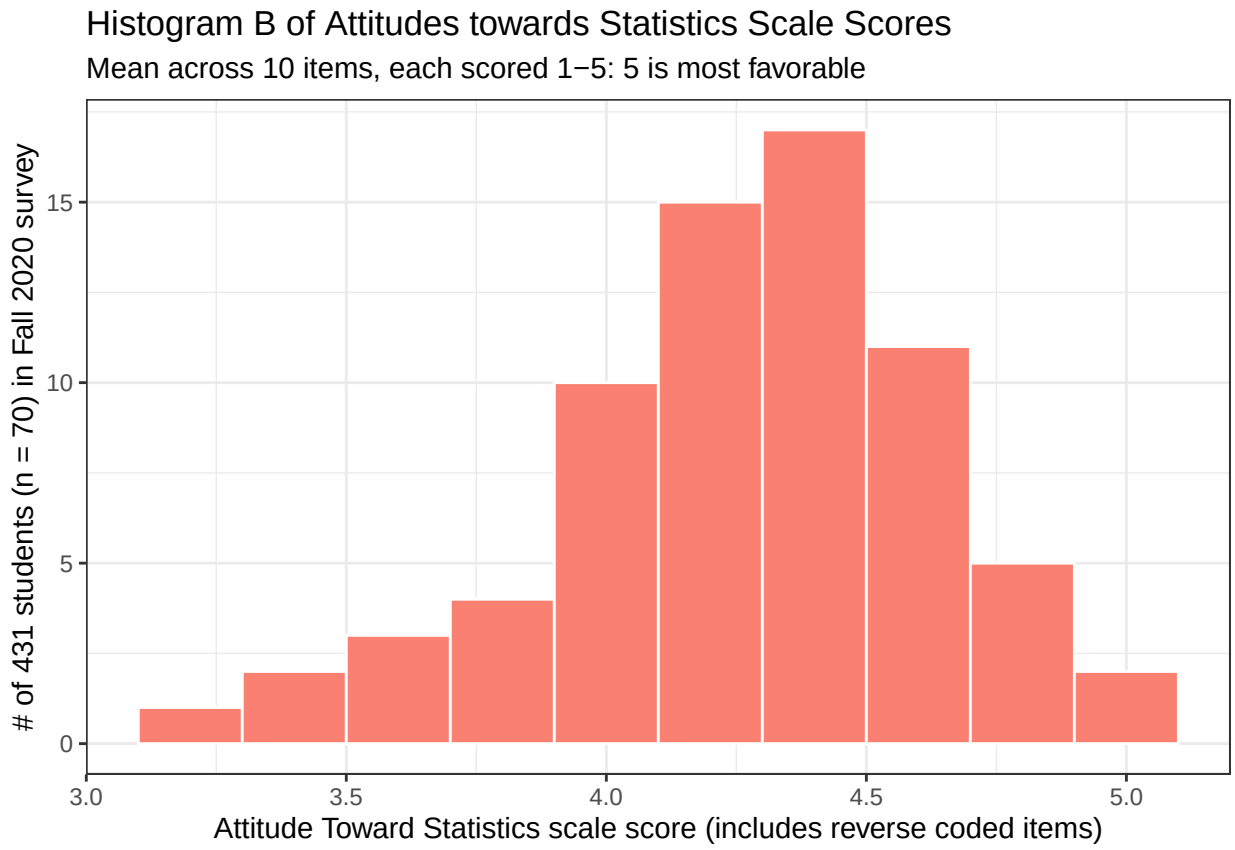
```
  The decimal point is 1 digit(s) to the left of the |

  33 | 0
  34 |
  35 | 00
  36 |
  37 | 000
  38 | 0000
  39 |
  40 | 0000
  41 | 000000
  42 | 00000000000
  43 | 0000
  44 | 000000000
  45 | 00000000
  46 | 0000000
  47 | 0000
  48 | 000
  49 | 00
  50 | 00
```

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 = "white") +
    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 = 70) in Fall 2020 survey")
```



Numerical Summaries

Using `summary()`

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

```
   ats_score
 Min.   :3.300
 1st Qu.:4.100
 Median :4.350
 Mean   :4.306
 3rd Qu.:4.575
 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
 3.3 4.1   4.35 4.575   5 4.305714 0.365103 70       0
```

Using `Hmisc::describe()`

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

```
.

 1  Variables      70  Observations
---------------------------------------------------------------------------
ats_score
       n  missing distinct     Info     Mean      Gmd      .05      .10
      70        0       15     0.99    4.306   0.4104    3.700    3.800
     .25      .50      .75      .90      .95
   4.100    4.350    4.575    4.710    4.855

lowest : 3.3 3.5 3.7 3.8 4.0, highest: 4.6 4.7 4.8 4.9 5.0

Value        3.3   3.5   3.7   3.8   4.0   4.1   4.2   4.3   4.4   4.5   4.6
Frequency      1     2     3     4     4     6    11     4     9     8     7
Proportion 0.014 0.029 0.043 0.057 0.057 0.086 0.157 0.057 0.129 0.114 0.100

Value        4.7   4.8   4.9   5.0
Frequency      4     3     2     2
Proportion 0.057 0.043 0.029 0.029
```

Using `psych::describe()`

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

```
   vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 70 4.31 0.37   4.35    4.32 0.37 3.3   5   1.7 -0.44    -0.08 0.04
```