

Part I. (200 points) The exam has two problems. You are to choose **ONE problem** to solve and turn in your analysis. Try to be thorough yet succinct with your writeup. This is a final exam, so you are on the honor system to work on this alone.

Due: Thurs Dec 8 at 12:30 PM (beginning of our last day of class).

Your project should be broken down into the following sections:

1. Summary
2. Description of data
3. Scientific questions of interest and statistical techniques that are needed to answer the questions
4. Statistical analysis
5. Conclusions
6. Appendix

You should include in the body of the report only those aspects of the analysis that you feel are necessary to the discussion. All other supporting documents should be placed and labeled in the Appendix.

Sections 1 through 5 should be no longer than 10 word processed pages. The report should be similar in structure to any scientific paper, but with more emphasis placed on the statistical issues.

(200^{pts})

1. Word length: Mark Twain is often credited with being the author of ten letters published in the New Orleans Daily Crescent under the name "Quintius Curtis Snodgrass". In 1963, C. Brinegar used statistical methods to compare the Snodgrass letters to works known to be written by Twain in an attempt to decide whether Twain did write the Snodgrass letters.

Brinegar used a very simple test of authorship based on the distributions of word lengths of the two authors. If the distributions of word lengths are very different, we have some evidence that the authors are probably different people. Other tests of authorship have been subsequently developed.

Brinegar counted the number of two-letter words, the number of three-letter words, and so on, for the ten Snodgrass letters. A similar summarization was done for a collection of works known to be written by Twain, including seven letters written to friends between 1858 and 1867, and samples of approximately 2500 words each from his 1872 work "Roughing It" and his 1897 work "Following the Equator." One-letter words, such as "I" and "a" were omitted because they tend to characterize content more so than style. Brinegar then compared the proportions of words of a given length across authors, that is, he tested homogeneity of proportions.

Initial analyses of Twain's and Snodgrass's works showed no significant differences in the word length proportions across the works of either author. All of the works for Twain and Snodgrass were combined for the final analysis.

In the table below, the first column gives the word lengths, with words of 13 or more letters combined together. The second and third columns give the word length counts for Snodgrass's and Twain's works, respectively.

The goal for this problem is to devise an appropriate statistical analysis to assess the agreement between the word length distributions for Twain and Snodgrass.

WordLength	Snodgrass	Twain
2	2685	2989
3	2752	3917
4	2302	3224
5	1431	1954
6	992	1283
7	896	1026
8	638	693
9	465	495
10	276	287
11	152	134
12	101	62
13	61	47

200 pts

200 pts

- (a) Use MINITAB or another package, for example EXCEL, to compute and plot, for each author, the proportion or percentage of words of lengths 2 through 13+.

For example, in MINITAB, if you enter the 3 columns of data into the spreadsheet, you can generate a graphical summary via:

GRAPH > BAR CHART; select BARS REPRESENT: VALUES FROM A TABLE; then choose TWO-WAY TABLE CLUSTER; from next dialog box specify SNODGRASS and TWAIN as the GRAPH VARIABLES and WORDLENGTH as ROW LABELS. On TABLE ARRANGEMENT click COLUMNS ARE...; for BAR CHART OPTIONS choose Show Y as PCT and TAKE PERCENTAGES Within Categories....

Describe what you see, keeping in mind that the goal is to compare the word length distributions for works by the two authors.

- (b) Carry out a formal test that the word length distributions for Twain and Snodgrass are identical.
(c) Carry out any additional analyses that you might deem appropriate.
(d) Summarize your conclusions.
-

Part II. (200 points) The exam has two problems. You are to choose **ONE problem** to solve and turn in your analysis. Try to be thorough yet succinct with your writeup. This is a final exam, so you are on the honor system to work on this alone.

Due: Thurs Dec 8 at 12:30 PM (start of class).

Your project should be broken down into the following sections:

1. Summary
2. Description of data
3. Scientific questions of interest and statistical techniques that are needed to answer the questions
4. Statistical analysis
5. Conclusions
6. Appendix

You should include in the body of the report only those aspects of the analysis that you feel are necessary to the discussion. All other supporting documents should be placed and labeled in the Appendix.

Sections 1 through 5 should be no longer than 10 word processed pages. The report should be similar in structure to any scientific paper, but with more emphasis placed on the statistical issues.

- (200^{pts}) **1. Birth at 40+ years:** The following data were collected from 48 women who were at least 40 years old when they gave birth to their first child. The data concern the gestation period of that pregnancy, and related variables on the child and mother.

The columns are, from left to right:

- 1) ID
- 2) The child's gestation period, in weeks
- 3) Sex of the child (0=Male, 1=Female)
- 4) Birth Weight of child, in grams
- 5) Number of cigarettes smoked per day (on average) by the mother
- 6) Height of mother in cm
- 7) Weight of mother in kilograms at first prenatal visit
- 8) Weight of mother in kilograms at final prenatal visit

```
id GEST sex BW numcig htmom(in) wtmom_first(kg) wtmom_last(kg)
1 36 0 3300 0 160.0 67.3 82.7
2 38 0 3300 60 167.6 52.7 76.0
3 38 0 4100 20 167.6 64.2 79.6
4 38 1 2900 10 163.9 72.7 95.8
5 39 0 2820 0 161.3 50.0 63.3
6 39 0 3040 0 158.8 49.1 61.5
7 39 0 4120 0 160.0 57.7 73.5
8 39 0 4200 0 174.0 68.0 86.8
9 39 1 3100 0 171.5 67.3 85.6
10 39 1 3330 0 160.0 74.0 90.5
11 39 1 3410 0 165.1 55.9 70.7
12 39 1 3420 0 162.6 52.3 66.0
13 40 0 2450 20 167.6 61.4 72.5
14 40 0 2885 0 167.7 60.0 78.6
15 40 0 3235 0 170.2 50.0 65.5
16 40 0 3320 0 165.1 63.6 80.2
17 40 0 3600 0 165.1 53.2 69.7
18 40 0 3720 0 165.0 57.7 74.4
19 40 0 3720 0 172.7 61.4 80.0
20 40 0 3820 0 175.3 60.8 78.1
21 40 0 3840 0 167.0 60.5 83.9
22 40 0 3880 0 156.2 57.3 73.7
23 40 0 3960 0 157.5 52.7 68.2
24 40 0 4465 0 157.5 51.4 66.4
25 40 1 2980 0 160.0 47.7 55.2
26 40 1 3040 0 162.0 49.0 60.3
27 40 1 3060 20 157.5 61.0 75.0
28 40 1 3100 0 170.2 55.5 64.6
29 40 1 3120 0 160.3 56.8 75.4
30 40 1 3205 0 172.7 58.2 75.5
31 40 1 3220 0 170.0 64.6 86.0
32 40 1 4100 40 167.0 67.0 85.0
33 41 0 3100 0 168.9 61.4 69.2
34 41 0 3720 0 170.2 57.7 67.7
35 41 0 3720 20 170.2 57.7 80.5
36 41 0 3900 0 167.0 68.0 85.4
37 41 0 3990 0 165.1 52.3 71.2
38 41 0 4050 0 167.6 61.0 78.5
```

200 pts

200 pts

```
39 41 0 4080 0 162.6 59.1 83.1
40 41 0 4100 0 165.1 60.5 86.5
41 41 0 4460 20 165.1 56.8 88.0
42 41 0 5220 0 157.5 56.8 68.2
43 41 1 3300 40 162.6 74.1 89.7
44 41 1 3400 0 172.7 71.4 87.8
45 41 1 4000 0 165.1 90.0 100.8
46 41 1 4030 0 166.0 63.0 95.3
47 43 1 3220 0 166.4 60.9 72.0
48 43 1 4270 0 162.6 54.5 70.3
```

- (a) Plot the birth weight (BW) against the length of gestation (GEST). Describe the relationship. Looking at the plot, should the sample correlation between BW and GEST be positive, negative, or nearly zero?
- (b) Compute the Pearson and Spearman correlations between BW and GEST. Comment. Test the hypothesis that the population correlation between BW and GEST is zero. Comment on the tests.
- (c) Provide an equation for the least squares line for predicting BW from GEST. Test the hypothesis that the slope of the population regression line is zero. We can think of this as a test that GEST is important for explaining the observed variation in BW. Superimpose the LS line on the data plot and comment on whether the simple linear regression model appears to adequately summarize the relationship between BW and GEST.
- (d) What percentage (or proportion) of the variability in BW is explained by the linear relationship between BW and GEST?
- (e) Compute the Cook's distance and the studentized residuals for each case. Make appropriate residual plots and an index plot of Cook's D to check for inadequacies with the model, and for potentially influential cases. Comment on what you find.
- (f) Calculate a 95% CI for the mean of all birth weights for a gestation period of 38 weeks. Also calculate a 95% prediction interval for a birth weight at 38 weeks. Which is wider, and why?
- (g) Carry out any further analyses that you feel are needed.
- (h) Provide a short summary of your analysis.