

Exam I

STAT 105, Section A SPRING 2016

Instructions

- The exam is scheduled for 80 minutes, from 9:30 to 10:50 AM. At 10:50 AM the exam will end.
- A formula sheet is attached to the end of the exam. Feel free to tear it off.
- You may use a calculator during this exam.
- Answer the questions in the space provided. If you run out of room, continue on the back of the page.
- If you have any questions about, or need clarification on the meaning of an item on this exam, please ask your instructor. No other form of external help is permitted attempting to receive help or provide help to others will be considered cheating.
- **Do not cheat on this exam.** Academic integrity demands an honest and fair testing environment. Cheating will not be tolerated and will result in an immediate score of 0 on the exam and an incident report will be submitted to the dean's office.

Name: _____

Student ID: _____

1. Netflix recommendations

A Recommender System is a statistical tool many websites use to recommend new items, products, or services to a user based on the user's history. For example, the popular movie streaming service Netflix recommends movies a user might like by assigning a predicting a rating between 0 and 5 to movies the user has not previously rated on the site and recommending the movies with the highest predicted rating. Netflix is testing a few changes in their recommendation system in an attempt to increase its accuracy. A specific user's true ratings are recorded below as well as the predicted ratings from the current recommender system and the predicted ratings from the new recommender system:

	Ratings									
User's True Ratings	4.30	4.10	4.10	4.20	4.20	4.20	4.30	4.30	4.20	4.20
Current System Predictions	4.50	4.20	3.70	4.80	4.30	4.00	4.40	3.80	4.40	3.70
New System Predictions	0.40	0.20	0.20	0.30	0.30	0.30	0.40	0.40	0.30	0.30

- (a) (2 points) Is the new recommender system more or less accurate than the current system?
- (b) (2 points) Is there a possible calibration-like adjustment that would improve the accuracy of either system? Explain.

2. Chilling fluid

The production of a certain type of metal beam involves rapidly cooling the beam to 10 degrees Celsius by submerging it in a bath of chilling fluid. The time required to reduce the beams temperature should be short or the cooling process can cause problems on the production line. During a single working week, six beams were randomly chosen and the time it took to cool them to 10 degrees Celsius is recorded below (in minutes):

12.1, 12.8, 13.3, 13.4, 14, 14.2

Using these observed values, report the following:

- (a) (2 points) the mean
- (b) (2 points) the variance
- (c) (2 points) the standard deviation
- (d) (2 points) the value of $Q(.30)$
- (e) (2 points) the interquartile range

3. Coloring books

Power4People (P4P), an electrical utility company, has consulted with workplace psychologists to find ways to reduce the stress levels of their engineering employees. The psychologists recommended two techniques: (1) practicing transcendental meditation and (2) spending time on simple arts and crafts projects. P4P has asked one of their engineers, KW, to determine which (if either) of two techniques is most effective. P4P's engineers belong to one of three departments: computing (22 engineers); electrical (26 engineers); and construction (30 engineers). It is believed that stress levels will be different for each department (for instance, the computer engineers travel less frequently). The psychologists have created an exam that measures stress levels from 0 (lowest) to 100 (highest), with all scores between 0 and 100 possible.

After a little thought, KW comes up with the following plan: 18 engineers are randomly selected from each of the three departments. All the selected engineers take the exam to measure their stress level. The 18 engineers selected from each department are then randomly split into three groups of six engineers each.

- The first group will meet each morning from 10:00 to 10:30 to practice transcendental meditation,
- The second group will meet each morning from 10:00 to 10:30 to quietly color in the coloring books,
- The third group will have no adjustment to their work schedule.

After two weeks, the engineers are again given the exam to measure their stress level and the change in their stress level is recorded.

- (a) (3 points) Is this an experiment or an observational study? Explain.
- (b) For each of the following variables types, identify whether it is numeric or qualitative. If it is numeric, state whether it is continuous or discrete. If it qualitative, provide the possible values.
- i. (3 points) Response variable(s) (if there is no response variable, put "Not Used"):
 - ii. (3 points) Experimental variable(s) (if there is no experimental variable, put "Not Used"):
 - iii. (3 points) Controlled variable(s) (if there is no control variable, put "Not Used"):
 - iv. (3 points) Blocking variable(s) (if there is no blocking, put "Not Used"):

4. Tidal torrents

Within 24 hours of its release, The Life of Pablo, a new album by Kanye West, had been illegally downloaded over 500,000 times by people using BitTorrent, a system in which users download a file to their own computers and upload the file to the computers of others.

For one hour, starting at 6:00 pm and continuing until 7:00 pm Eastern, a systems engineer working for the company that stands to lose money from the piracy used a combination of tools to gather the following information on every 500th individual who began downloading a torrent of the album.

- Whether the individual was in the United States or not,
- The individual's average upload speed,
- The individual's average download speed,
- The number of users the individual shared the album with,

Tidal plans to use the data gathered to illustrate the extent of their loss in future legal action.

(a) (2 points) Is this an experiment or an observational study?

(b) (2 points) What is the population being considered?

(c) It turns out that average upload speed and average download speed were very strongly, positively, correlated. Answer each of the following TRUE or FALSE.

- i. (1 point) We can conclude that higher values of average **download** speed *cause* the value of average **upload** speed to be **higher**.
- ii. (1 point) We can conclude that higher values of average **download** speed *cause* the value of average **upload** speed to be **lower**.
- iii. (1 point) We can conclude that higher values of average **upload** speed *cause* the value of average **download** speed to be **higher**.
- iv. (1 point) We can conclude that higher values of average **upload** speed *cause* the value of average **download** speed to be **lower**.

(d) (2 points) After noticing that most of the users are from the United States, one executive comments that most Kanye West fans are Americans. Another executive points out that most of Europe is asleep at 6:00 pm eastern and wouldn't be awake to download the album. In this case, whether or not a person is awake could be considered what kind of variable?

5. Highlights

The lifetime of a highlighter can be measured in the area of paper it can cover with highlighting fluid until either the marker falls apart or it runs out of highlighter fluid. A sample of 26 highlighters were used until the end of their lifetime. The areas covered by the highlighters (in square meters) are recorded below:

The decimal point is at the |

```

0 | 1
1 | 28
2 | 7
3 | 113588
4 | 055688
5 | 011789
6 | 9
7 | 0
8 |
9 | 24

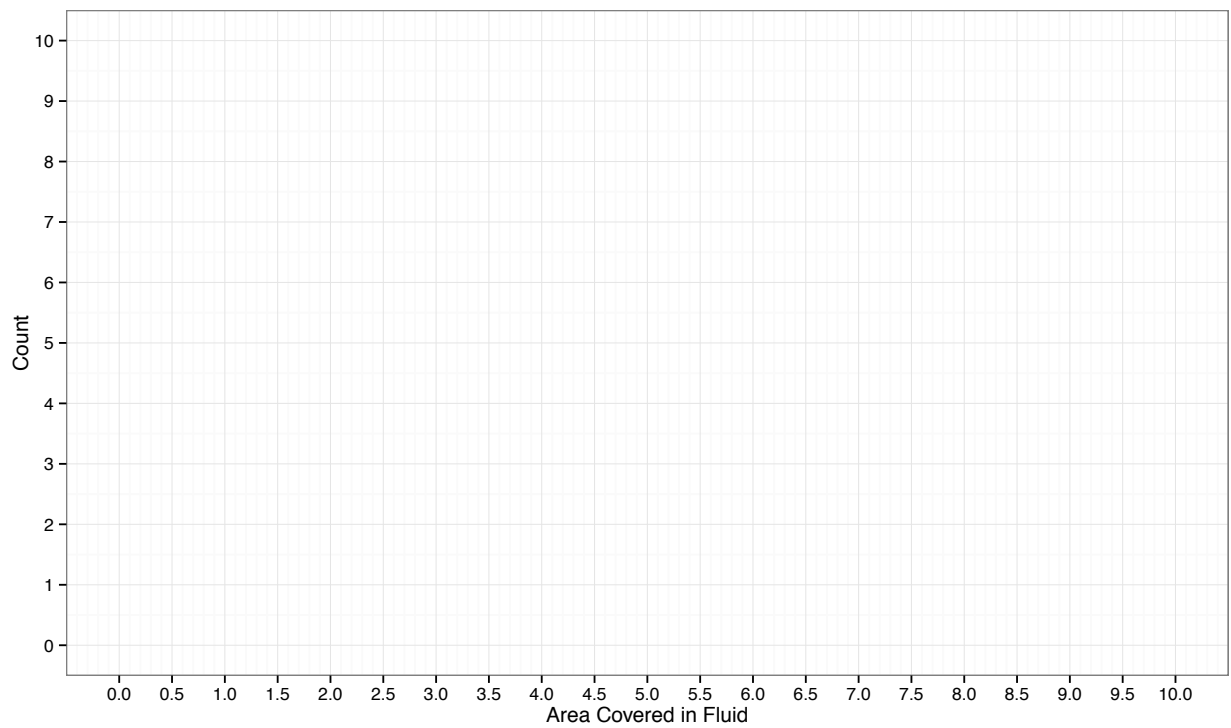
```

Note that 1 | 9 represents 1.9. In this case, the first quartile is $Q(.25) = 3.3$, the median is 4.55, and the third quartile is $Q(.75) = 5.7$.

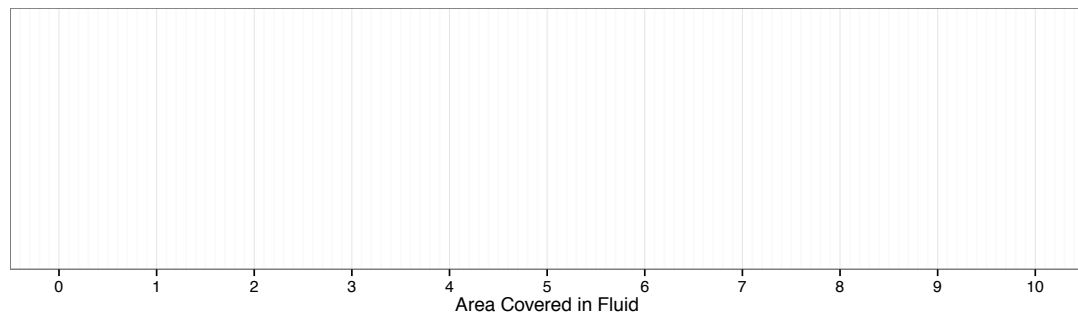
(a) (10 points) Complete the following frequency table:

Value Range	Frequency	Relative Frequency	Cumulative Relative Frequency
0.00 - 2.00			
2.01 - 4.00			
4.01 - 6.00			
6.01 - 8.00			
8.01 - 10.00			

- (b) (5 points) Using the area below, create a **dot plot** to summarize the data.



- (c) (10 points) Using the area below, create a **box plot** to summarize the data. Carefully label the important parts.



- (d) (4 points) Are there any unusually low or high observations? If so, what was the area covered by those highlighters?

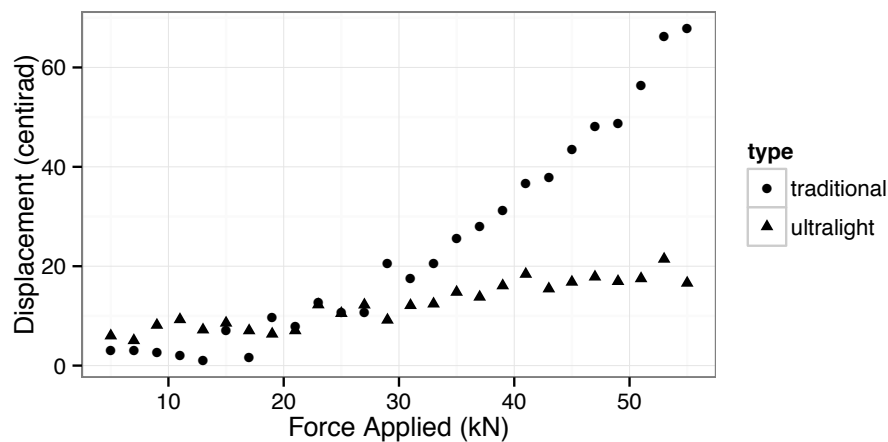
6. Ultralight beams

An ultralight beam is a type of metal beam used as secondary support in the facade of tall structures. Its low weight and high flexibility make it ideal for environments exposed to high amounts of pressure.

A manufacturer of ultralight beams performed the following analysis to illustrate the flexible nature of the beams. 26 ultralight beams were fixed in place vertically. Each beam had a different amount of force applied to it, from 5 kN 55 kN, for 20 minutes in an attempt to bend the bar. After the twenty minutes, the beam had a chance to “bounce back” the final displacement from its original position was measured (in centi-radians). The smaller the displacement, the better the bar did in returning to its original position.

The experiment was repeated, this time using the steel beams traditionally used as secondary support in the facade of tall buildings (referred to from here on as “traditional beams”).

The results are depicted below (with the force applied on the x-axis, and with ultralight beam observations represented with triangles and traditional beam results represented by dots):



Here are some summaries of the ultralight data (again using the force applied as the x-value and the displacement of the beam as the y-value):

$$\sum_{i=1}^{50} x_i = 780$$

$$\sum_{i=1}^{50} x_i^2 = 29250$$

$$\sum_{i=1}^{50} y_i = 319$$

$$\sum_{i=1}^{50} y_i^2 = 4455$$

$$\sum_{i=1}^{50} x_i y_i = 11227$$

- (a) Using the summaries above, we plan to fit a linear relationship between **force applied** (x) and **displacement** (y) for the ultralight beam.
- i. (5 points) Write the equation of the fitted linear relationship.
 - ii. (5 points) Find and interpret the value of R^2 for the fitted linear relationship.
 - iii. (5 points) Using the fitted line, what do we suppose the displacement of an ultralight beam will be if the force applied is 41.
 - iv. (5 points) The ultralight beam that actually underwent 41 kN had a displacement of 18.41 centiradians. What is the residual associated with this observation?

- (b) The JMP output below comes from fitting a quadratic model using the force applied (`force_app`) and the square of the force applied (`force_app^2`) for the displacement of the traditional beams.

▼ Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	10785.781	5392.89	1081.437
Error	23	114.696	4.99	Prob > F
C. Total	25	10900.477		<.0001*
▼ Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	3.9781742	1.768155	2.25	0.0343*
force_app	-0.401014	0.13408	-2.99	0.0065*
force_app^2	0.0284085	0.002181	13.03	<.0001*

- i. (5 points) Write the equation of the fitted quadratic relationship.

- ii. (5 points) Find and interpret the value of R^2 for the fitted quadratic relationship.

- iii. (5 points) Using the fitted line, what do we suppose the displacement of an traditional beam will be if the force applied is 41.

- iv. (5 points) The traditional beam that actually underwent 41 kN had a displacement of 36.68 centiradians. What is the residual associated with this observation?

STAT 105 Exam I

Reference Sheet

Numeric Summaries

mean	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
population variance	$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
population standard deviation	$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$
sample variance	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
sample standard deviation	$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

Linear Relationships

Form	$y \approx \beta_0 + \beta_1 x$
Fitted linear relationship	$\hat{y} = b_0 + b_1 x$
Least squares estimates	$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ $b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$ $b_0 = \bar{y} - b_1 \bar{x}$
Residuals	$e_i = y_i - \hat{y}_i$
sample correlation coefficient	$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$ $r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n \bar{x}^2)(\sum_{i=1}^n y_i^2 - n \bar{y}^2)}}$
coefficient of determination	$R^2 = (r)^2$
	$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

Multivariate Relationships

Form	$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$
Fitted relationship	$\hat{y} \approx b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$
Residuals	$e_i = y_i - \hat{y}_i$
Sums of Squares	$SSTO = \sum_{i=1}^n (y_i - \bar{y})^2$
	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
	$SSR = SSTO - SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
coefficient of determination	$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$
	$R^2 = \frac{SSTO - SSE}{SSTO}$
	$R^2 = \frac{SSR}{SSTO}$
	$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

Functions

Quantile Function $Q(p)$ For a dataset consisting of n values that are ordered so that $x_1 \leq x_2 \leq \dots \leq x_n$ and value p where $0 \leq p \leq 1$, let $i = \lfloor n \cdot p + 0.5 \rfloor$. Then the quantile function at p is:

$$Q(p) = x_i + (n \cdot p + 0.5 - i)(x_{i+1} - x_i)$$