Hi18

Seat

Please print in pen: STC 0010

Waterloo Student ID Number:

WatIAM/Quest Login Userid:

UNIVERSITY OF WATERLOO

Examination
Test
Winter 2020
STAT 341

Times: Friday 2020-02-07 at 09:30 to 10:20

Duration: 50 minutes

Exam ID: 4449268

Sections: STAT 341 LEC 001

Instructors: Nathaniel Stevens

Special Materials

Candidates may bring only the listed aids.

- Calculator - Pink Tie

Instructions:

- You have 50 minutes to complete this test.
- This test consists of 6 questions and 8 pages (including this cover page).
- Page 8 contains additional space for rough work. DO NOT use this page for anything that you would like to have marked.
- Numeric answers should be rounded to four decimal places (unless the answer is exact to fewer than four decimal places).
- Incorrect answers may receive partial credit if your work is shown. An incorrect answer with no work shown will receive 0 points.

Question	Points
Q1	4
Q2	6
Q3	4
Q4	3
Q5	5
Q6	8
Total	**30**

- Please identify yourself by signing here: ______________________

SOLUTIONS

1. **[4 points]** Consider the population $\mathcal{P} = \{y_1, \ldots, y_N\}$ and the attribute $a(\mathcal{P})$.

(a) [1 point] What does it mean for $a(\mathcal{P})$ to be **location-scale equivariant**?

$$a(m\mathcal{P}+b) = a(my_1+b, \ldots, my_N+b)$$

1 point $$= m\,a(y_1, \ldots, y_N)+b$$

$$= m\,a(\mathcal{P})+b$$

* note that just the second equivalence is sufficient

* note that just the first and last equivalence are sufficient

(b) [3 points] Suppose that the attribute $a(\mathcal{P})$ is location equivariant **and** scale equivariant. Show that $a(\mathcal{P})$ is therefore location-scale equivariant.

- We know $a(m\mathcal{P}) = m\,a(\mathcal{P})$ and $a(\mathcal{P}+b) = a(\mathcal{P})+b$ for any population $\mathcal{P}$.

- Define $\mathcal{P}^* = \{x_1, \ldots, x_N\} = \{my_1, \ldots, my_N\} = m\mathcal{P}$. Then

$$a(m\mathcal{P}+b) = a(\mathcal{P}^*+b)$$

$$= a(\mathcal{P}^*)+b \quad \text{by location equivariance}$$

$$= a(m\mathcal{P})+b$$

$$= m\,a(\mathcal{P})+b \quad \text{by scale equivariance}$$

$\therefore$ $a(\mathcal{P})$ is location-scale equivariant

* note this is just one version of a correct response.

Points are allocated as follows:

- 3 points for an entirely correct solution
- 2 points if there is something wrong
- 1 point if there is a lot wrong
- 0 points if there is no attempt.

2. **[6 points]** This question concerns sensitivity analysis.

(a) [1 points] Given a population $\mathcal{P} = \{y_1, \ldots, y_{N-1}\}$ and an attribute $a(\mathcal{P})$, define the sensitivity curve for the attribute.

1 point

$$SC(y; a(\mathcal{P})) = N\left[a(y_1, \ldots, y_{N-1}, y) - a(y_1, \ldots, y_{N-1})\right]$$

For the full point it must be clear this is a function of y. Give a half-point otherwise.

(b) [5 points] Consider the population above and assume N is odd (i.e., $N = 2m + 1$). The order statistics in this case are $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N-1)}$ and the median can be written as

$$a(\mathcal{P}) = a(y_1, \ldots, y_{N-1}) = \frac{y_{(m)} + y_{(m+1)}}{2}$$

i. [3 points] Derive the median's sensitivity curve.

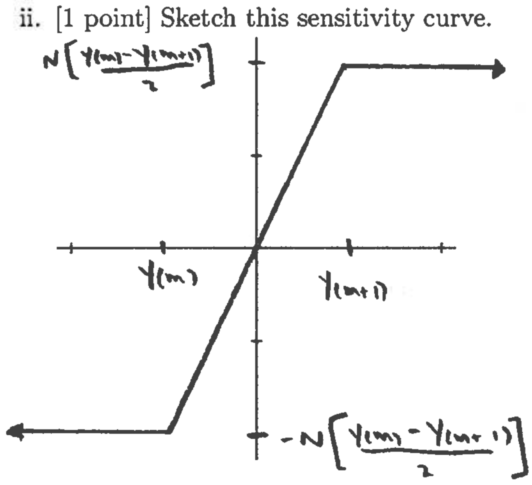
- If $y < y_{(m)}$ then $a(y_1, \ldots, y_{N-1}, y) = y_{(m)}$
- If $y_{(m)} \leq y \leq y_{(m+1)}$ then $a(y_1, \ldots, y_{N-1}, y) = y$
- If $y > y_{(m+1)}$ then $a(y_1, \ldots, y_{N-1}, y) = y_{(m+1)}$

3 points to be allocated as in 1(b).

$$\therefore SC(y, a(\mathcal{P})) = \begin{cases} N\left[\frac{y_{(m)} - y_{(m+1)}}{2}\right] & \text{if } y < y_{(m)} \\ N\left[\frac{2y - y_{(m)} - y_{(m+1)}}{2}\right] & \text{if } y_{(m)} \leq y \leq y_{(m+1)} \\ -N\left[\frac{y_{(m)} - y_{(m+1)}}{2}\right] & \text{if } y > y_{(m+1)} \end{cases}$$

* note that equivalent but less simplified solutions will also be accepted.

ii. [1 point] Sketch this sensitivity curve.



- $\frac{1}{2}$ point if the plot has the correct shape
- $\frac{1}{2}$ point the location of the elbows is somehow labelled.

iii. [1 point] From this curve, what do you conclude about the median's resistance to outliers in y?

Since this curve is bounded (i.e., constant) for $y \to \pm\infty$, we can conclude that the median is not sensitive to extreme values and hence resistant to outliers.

1 point for anything that says the median is robust <u>because</u> the SC is bounded. Award only a half-point if they miss this justification

3. [4 points] Consider the population $\mathcal{P} = \{y_1, \ldots, y_N\}$, and the following two attributes that measure its spread. The first attribute is the standard deviation defined as

$$a_1(\mathcal{P}) = \sqrt{\frac{\sum_{u \in \mathcal{P}} (y_u - \overline{y})^2}{N}}$$

and the second attribute is the interdecile range (the difference between the first and ninth deciles)

$$a_2(\mathcal{P}) = Q_y(0.9) - Q_y(0.1)$$

where $Q_y(p)$ is the $100 \times p\%$ quantile of the population.

(a) [1 point] Define the breakdown point for a population attribute.

The breakdown point for a population attribute is the smallest proportion of observations that would need to be changed to $\pm\infty$ to make the difference between the attribute value with and without this change infinitely large. 1 point for something along these lines.

(b) [2 points] Determine the breakdown point for both attributes $a_1(\mathcal{P})$ and $a_2(\mathcal{P})$. Provide rationale for your answer.

- $a_1(P)$ would evaluate to ∞ if even 1 observation was changed to ∞. Thus its breakdown point is:

  $\frac{1}{N}$ $\frac{1}{2}$ point

  Justification = $\frac{1}{2}$ point

- $a_2(P)$ would evaluate to ∞ if 10% of the observations were changed to ∞. Thus its breakdown point is:

  0.1 $\frac{1}{2}$ point

  Justification = $\frac{1}{2}$ point.

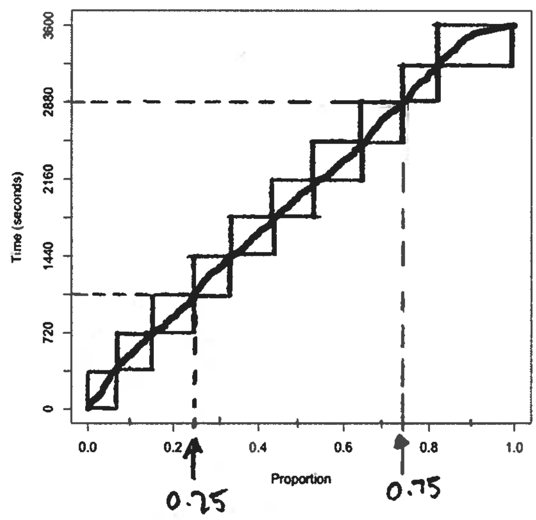
(c) [1 point] Given your answers in part (b), which attribute is more robust (i.e., resistent to outliers) and why?

For $N > 10$ the interdecile range is more resistant to outliers because its breakdown point is larger ($0.1 > \frac{1}{N}$). If $N \leq 10$, they're equally robust.

$\frac{1}{2}$ point for each of the two cases.

Hi18

4. **[3 points]** Wayne Gretzky "The Great One" is a Canadian former professional ice hockey player. He played 20 seasons in the National Hockey League (NHL) and he is considered to be the greatest hockey player ever. Below is a quantile plot of the $N = 894$ goals he scored during his time in the NHL. In particular, we examine the times (in seconds) at which the goals occurred in a sixty-minute game.



(a) [1 point] Using this plot provide an estimate of the interquartile range.

$IQR = Q_y(0.75) - Q_y(0.25)$

$\approx 2880 - 1080$

$= 1800$ 1 point.

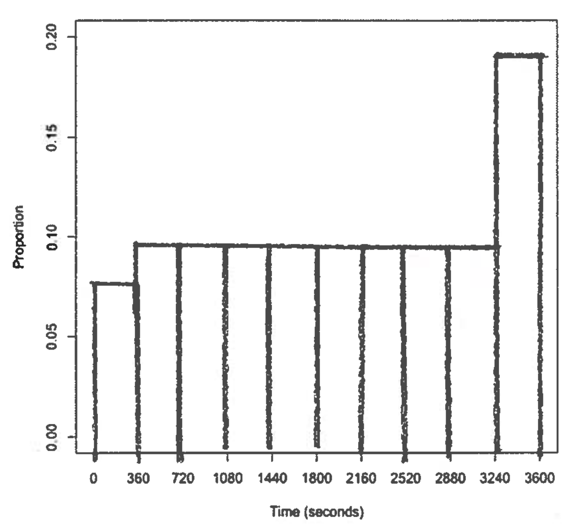
* any answers reasonably close to this will be accepted.

(b) [1 points] Using your answer from (a) together with the Freedman-Diaconis Rule, explain why when creating a histogram of these data, 10 bins would be a sensible choice.

FD Bin Width $= \dfrac{2\, IQR}{N^{1/3}} = \dfrac{2 \times 1800}{894^{1/3}} = 373.7$ ← $\frac{1}{2}$ point

bins should then be $\dfrac{3600-0}{373.7} = 9.6 \approx 10$ $\frac{1}{2}$ point (i.e., 10 bins is sensible)

(c) [1 point] Using the axes below, construct a histogram for these data using 10 bins. (Hint: Consider concentration boxes and the connection between quantile plots and histograms.)



1 point

* note I will accept anything that looks reasonably uniform over [0, 3240] with bar heights close to 0.1 as long as the final bar is noticeably taller and almost 0.2 high.

5. [**5 points**] This question concerns the implicitly defined attribute $\boldsymbol{\theta} \in \mathbb{R}^k$ in population $\mathcal{P}$.

(a) [1 point] Provide the objective function-based definition of $\widehat{\boldsymbol{\theta}}$. Define any notation that you use.

1 point

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^k}{\arg\min}\ \rho(\theta; \mathcal{P})$$ where $\rho(\theta; \mathcal{P})$ is the objective function.

word-based definitions also okay.

(b) [1 point] Provide the system of equations-based definition of $\widehat{\boldsymbol{\theta}}$. Define any notation that you use.

$\hat{\theta}$ is the solution to $\vec{\Psi}(\theta; \mathcal{P}) = \vec{0}$

1 point

where $\vec{\Psi}$ is a $k \times 1$ vector of equations and $\vec{0}$ is the $k \times 1$ zero vector.

(c) [3 points] In point form, describe the batch-stochastic gradient descent algorithm. Define any notation that you use.

- Given a starting point $\hat{\theta}_0$ and fixed step size λ^*
- Initialize $i = 0$
- LOOP over i
    - Draw a random sample S from population $\mathcal{P}$
    - calculate the gradient $g_i = \nabla \rho(\theta; \mathcal{P})|_{\theta = \hat{\theta}_i}$
    - Calculate the direction $d_i = \dfrac{g_i}{\|g_i\|}$
    - Update the iterate $\hat{\theta}_{i+1} = \hat{\theta}_i - \lambda^* d_i$
    - Check convergence
        - IF converged RETURN
        - ELSE $i = i + 1$
- RETURN $\hat{\theta} = \hat{\theta}_i$

3 points to ~~be~~ allocated as in 1(b).

6. **[8 points]** Determine whether the following statements are True or False. In each case circle the correct answer.

(a) [1 point] The ratio of two scale equivariant attributes is a scale invariant attribute.

i. True

ii. False

1

(b) [1 point] A relative-frequency histogram (where the height of each bar reflects the *proportion* of data lying in the bin) is replication invariant.

i. True

ii. False

1

(c) [1 point] When evaluating a population attribute, the notions of *influence* and *sensitivity* are the same.

i. True

ii. False

1

(d) [1 point] Consider the population $\mathcal{P} = \{y_1, \ldots, y_N\}$. Suppose that a histogram of this data is left-skewed (i.e., negatively skewed). If we wanted to use a power transformation to make the histogram more symmetric, we should use a power that is less than 1.

1

i. True

ii. False

(e) [1 point] Compared to batch-sequential gradient descent, batch-stochastic gradient descent is less senstive to outliers in the data.

i. True

ii. False

1

(f) [1 point] Consider the objective function $\rho(\theta) = 2\theta^2 - 5\theta + 3$. Stochastic gradient descent would be a useful technique for determining

$$\underset{\theta \in \mathbb{R}}{\operatorname{argmin}}\, \rho(\theta)$$

1

i. True

ii. False

(g) [1 point] In objective function-minimization problems, the Newton-Raphson algorithm can be viewed as a form of gradient descent.

i. True

ii. False

1

(h) [1 point] Iteratively reweighted least squares is a root-finding algorithm.

i. True

ii. False I will accept either answer as correct here.

1

Hi18

This space is left for rough work