Name: _____

I.D. #: _____

University of Waterloo

STAT 341
Exam

Instructor: R.P. Browne

Time: 2:30 minutes

Instructions

1. Pay attention to how much each question is worth (about 1 mark per point to make). Do not provide overly long answers.

Question	Marks Available	Marks Earned
1	7	
2	6	
3	10	
4	10	
5	6	
6	7	
7	5	
8	6	
9	7	
10	6	
11	6	
TOTAL	76	

1. Explicitly defined Population Attributes

   (a) *(2 marks)* Let $a_1(\mathcal{P})$ and $a_2(\mathcal{P})$ be location invariant and scale equivariant then let $a_R(\mathcal{P}) = a_1(\mathcal{P})/a_2(\mathcal{P})$. Cross out any the following properties that does not apply to $a_R(\mathcal{P})$.

location equivariant	location invariant	scale equivariant	scale invariant

   (b) *(2 mark)* Contrast and compare the properties of the standard deviation and the coefficient of variation as an attribute. **Note:** the coefficient of variation is the standard deviation divided by the population average.

   (c) *(2 mark)* Contrast and compare using a histogram or boxplot to summarize a population.

   (d) *(1 mark)* Given a population what does skewness measure?

2. Implicitly defined Population Attributes

(a) *(5 marks)* Given a population $\mathcal{P} = \{(x_1, y_1), ..., (x_N, y_N)\}$, consider the problem of calculating β for the regression through origin, i.e. $y_u = \beta x_u + r_u$. We would like to calculate the parameter β of this model using

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}} \sum_{u=1}^{N} \rho(r_u) \qquad \text{where} \qquad \rho_k(r) = \begin{cases} \frac{1}{2} r^2 & \text{for } |r| \leq k, \\ k\,|r| - \frac{1}{2}k^2, & \text{otherwise.} \end{cases}$$

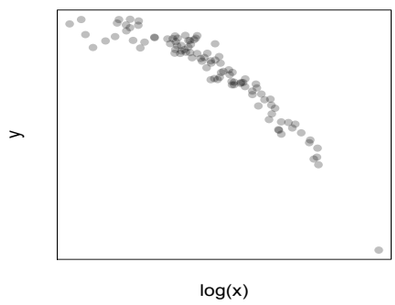
Using bullet points, describe the gradient descent algorithm to find $\widehat{\beta}$ in this particular problem. Any function/derivative should be calculated explicitly as the ρ function has been provided in the question.

(b) *(1 mark)* How do you adjust the above objective function to make it scale invariant?

3. Power transformations and sensitivity curves.

(a) *(4 marks)* Power transformations.

i. Consider the scatterplot below of y versus $\log x$:

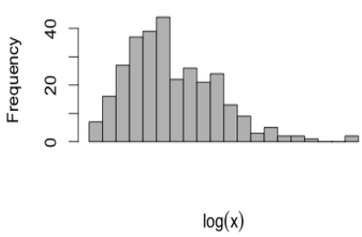


Which of the following transformations might straighten the scatterplot at left?
Circle those that **do apply**.
Cross out those that **do not**.

(i) $log(y)$ vs $log(x)$ (ii) y^2 vs $log(x)$

(iii) $log(y)$ vs x^2 (iv) y vs $\sqrt{x}$

ii. Consider the histogram below of $\log x$:



Which of the following transformations might make the histogram at left more symmetric?
Circle those that **do apply**.
Cross out those that **do not**.

(i) $\sqrt{x}$ (ii) $-\frac{1}{x}$

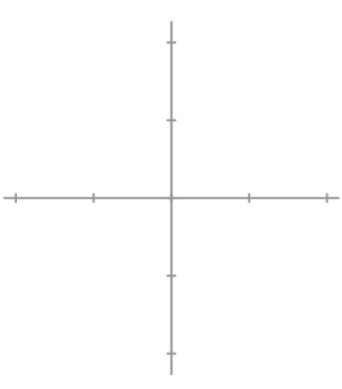
(iii) x^2 (iv) x

(b) *(6 marks)* Consider the range statistic:

$$T_N(y_1, \ldots, y_N) = \max_i \ (y_i) \quad - \quad \min_i \ (y_i).$$

i. Derive this statistic's sensitivity curve:

ii. Give a sketch of the sensitivity curve:



iii. From this curve, what do you conclude about this statistic's resistance to outliers in y?

4. *(10 marks)* Classify each statement by circling TRUE or FALSE.

i)	A scatterplot is not an attribute.	TRUE	FALSE
ii)	An unbiased estimator is always preferable to a biased estimator.	TRUE	FALSE
iii)	The observed significance level measures evidence against the null hypothesis.	TRUE	FALSE
iv)	The Horvitz-Thompson estimator is unbiased given a probability sampling design.	TRUE	FALSE
v)	When sampling with replacement the sample average is a Horvitz-Thompson estimate.	TRUE	FALSE
vi)	We can quantify the sampling error by calculating all possible samples.	TRUE	FALSE
vii)	The exact value of an estimator's bias is calculable based on all possible samples.	TRUE	FALSE
viii)	The standard bootstrap confidence interval is invariant to any one-to-one transformation.	TRUE	FALSE
ix)	One purpose of bootstrap method is to approximate the sampling distribution of an attribute.	TRUE	FALSE
x)	Increasing the complexity of a statistical model results in an increase in its prediction power.	TRUE	FALSE

5. Horvitz-Thompson (HT) estimator

(a) *(2 marks)* Show that the HT estimator $\sum_{u\in\mathcal{S}} y_u/\pi_u$ is unbiased for the population total. Clearly define any notation used.

(b) *(1 mark)* The variance of the HT estimator is $Var(\widetilde{a}_{HT}(\mathcal{S})) = \sum_{u\in\mathcal{P}}\sum_{v\in\mathcal{P}}(\pi_{uv} - \pi_u\pi_v)\frac{y_u}{\pi_u}\frac{y_v}{\pi_v}$. Give the HT estimate of the variance based on the sample, $\mathcal{S}$.

(c) *(1 mark)* Suppose we sample with replacement where probability of selecting unit u is p_u. Derive the inclusion probability for unit u.

(d) A sample ($n = 6$) was randomly selected from a population ($N = 124$). The inclusion probabilities and the observed y values from the sample are

y_i	-2	-1	0	1	2	3
π_i	4/50	4/50	4/50	4/50	1/73	1

i. *(1 mark)* Give the HT estimate for the proportion of units less than or equal to -2.

ii. *(1 mark)* Give the HT estimate of the population median.

6. Inductive Inference

(a) *(3 marks)* Define the three possible sources of error in a scientific study performed within the inductive inference framework.

-
-
-

(b) *(1 marks)* What is coverage probability?

(c) *(3 marks)* Show how to construct a confidence interval using the pivotal quantity, $\left[\widetilde{a}(\mathcal{S}) - a(\mathcal{P})\right] / \widetilde{SD}\left((\widetilde{a}(\mathcal{S})\right)$ which has distribution F.

7. Comparing sub-populations

   (a) *(5 marks)* Describe in point form a test of significance using a discrepancy measure $D(\mathcal{P}_1, \mathcal{P}_2)$

8. Discrepancy measures and significance levels.

City of Baltimore is interested in determining if the Neighbourhood property crime rate per 1,000 residents has decreased from 2011 to 2014.

(a) *(1 mark)* State an appropriate discrepancy measure.

(b) *(1 marks)* Why should we randomly swap property crime rates within a neighbourhood instead of randomly shuffling all the property crime rates.

(c) *(1 marks)* Describe in words what a significance level is.

(d) *(3 marks)* If we had multiple discrepancy measures, describe a way to combine them into a single discrepancy measure?

9. Bootstrap

(a) *(1 mark)* Give an advantage of the bootstrap estimate of the standard error.

(b) *(3 marks)* Describe the bootstrap percentile method for constructing a confidence interval.

(c) *(3 marks)* For regression describe how to generate bootstrap samples by resampling from the errors.

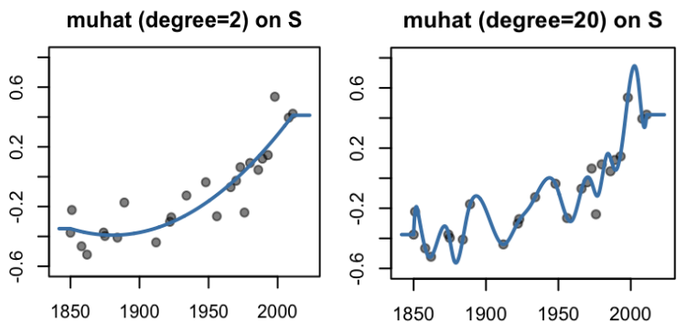
10. *(6 marks)* Show that the average prediction squared error, $APSE(\mathcal{P}, \widetilde{\mu})$, can be written as the sum of three components:

$$\frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \frac{1}{N} \sum_{i \in \mathcal{P}} (y_i - \widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_i))^2 = \frac{1}{N} \sum_{i \in \mathcal{P}} [y_i - \mu(\mathbf{x}_i)]^2 + \frac{1}{m} \sum_{j=1}^{m} \frac{1}{N} \sum_{i \in \mathcal{P}} \left[\widehat{\mu}_{\mathcal{S}_j}(\mathbf{x}_i) - \overline{\widehat{\mu}}(\mathbf{x}_i)\right]^2 + \frac{1}{N} \sum_{i \in \mathcal{P}} \left[\overline{\widehat{\mu}}(\mathbf{x}_i) - \mu(\mathbf{x}_i)\right]^2$$

Define any notation introduced.

11. Model selection and Prediction Error.

(a) *(2 marks)* A sample and two fitted functions are displayed in the graphs below. Based on this which function do you prefer? Why?



(b) *(1 marks)* Why are many samples preferred than a single sample when calculating the prediction error?

(c) *(3 marks)* Describe k-fold cross-validation.