Sample exam questions for Inf2 - FDS

1. Consider a histogram of the height of people in a particular cohort from university.

- Would you expect the histogram to be right or left skewed? Why?
- Which statistic is expected to be larger: the mean or the median?
- What transform would you suggest to reduce the effect of outliers?

2. PCA

- Show that the eigenvalues of (a) X^TX and (b) XX^T are all non-negative and real.
- Write down two assumptions that PCA makes about the data.
- If you are given data in 10 dimensions with features X1,..,X10 and you run PCA with the top three components "explaining 100% of the variance," what can you conclude about the data?

3. For each of the following, discuss in 2-4 sentences whether there is a possible source for sample bias and what the implications of any conclusions are:

- A well known Tory politician reported that letters into her office were running 3 to 1 in opposition to a second referendum. She concluded that the constituents were strongly opposed to a second referendum.
- An academic researcher is evaluating the effect of a drug on the circadian rhythm of the general population. He enlists undergraduates from the Informatics Forum and uses a table of random numbers to select a sub-population of students. His results suggest that the drug improves sleep duration by 1 hour for the general population.

4. If your sample size increases by a factor of ten, would you expect each of the following to (a) increase (b) decrease (c) neither (d) either (i.e., the statistic might increase or decrease)?
- Population standard deviation
- Sample standard deviation
- Standard error of the mean
- Estimate of the mean

5. For a particular sample of data, you estimate the standard error of the mean as x. You want it to be x/2. Roughly how much more data would you need to collect?

6. [Give example of a particular prediction problem].
(a) Would it be more appropriate to use linear or logistic regression for this problem? Explain why.
(b) [maybe some question about feature independence, or which features you might use?]

7. Suppose you have a dataset divided into training, development, and test partitions and you are exploring different machine learning methods . As we've seen in class, some methods are deterministic: they always produce the same results when given the same inputs and hyperparameters, while others are non-deterministic: the results depend on initialization.
(a) Name one deterministic ML method and one non-deterministic method.
(b) [some question about what to use the different partitions for]
(c) [some question about how to evaluate whether one system has higher accuracy than the other, given that results in one case are non-deterministic]

[The above q requires discussing some ML method that's non-deterministic, such as random forests. There may be other ways to ask something related without this, for example randomness from the data itself as in x-validation. It needs some thought but the point is to have some connection between stats and the results of ML experiments.]

8. The mean of 10 numbers is 3.8. Another number is added to this sample and the mean increases to 4.0. What number was added?

[This question, or some variant, would make sense if there is some discussion of algorithmic strategies for dealing with large data sets; in particular streaming data]

The following are a few questions from past Inf2b papers that would also be relevant here:

9. Consider a document classication system which was trained with a large data set D. An evaluation experiment was carried out using the same data set D. State the main problem with this evaluation, and describe how it can be resolved.

10. Considering the two vectors, a = (6 2 4 0) and b = (-3 1 1 3), find the Pearson correlation coe cient, r(a, b).

11. Explain what linearly separable means in pattern recognition. [or: give a diagram of data points and ask if it is linearly separable and what the implications are.]

[Also consider some of the harder questions, but many of these not appropriate because they're about mathematical details of specific methods we likely won't cover.]

The following are suggestions but are more likely to be exercises, because we probably don’t want to require calculators for the exam:

Example 12.9 The following table shows the marks of 10 candidates in Physics and Mathematics. Find the product-moment correlation coefficient and comment on your value.

Mark in Physics (x)	18	20	30	40	46	54	60	80	88	92
Mark in Mathematics (y)	42	54	60	54	62	68	80	66	80	100

Example 2.6 A random sample of 51 people were asked to record the number of miles they travelled by car in a given week. The distances, to the nearest mile, are shown below.

67	76	85	42	93	48	93	46	52
72	77	53	41	48	86	78	56	80
70	70	66	62	54	85	60	58	43
58	74	44	52	74	52	82	78	47
66	50	67	87	78	86	94	63	72
63	44	47	57	68	81			

(a) Construct a stem and leaf diagram to represent these data.

(b) Find the median and the quartiles of this distribution.

(c) Draw a box plot to represent these data.

(d) Give one advantage of using
(*i*) a stem and leaf diagram,
(*ii*) a box plot,
to illustrate data such as that given above. (L)

The accountant of a company monitors the number of items produced per month by the company, together with the total cost of production. The following table shows the data collected for a random sample of 12 months.

f items (x) (1000 s)	21	39	48	24	72	75	15	35	62	81	12	56
n cost (y) (£ 1000)	40	58	67	45	89	96	37	53	83	102	35	75

(a) Plot these data on a scatter diagram. Explain why this diagram would support the fitting of a regression equation of y on x.

(b) Find an equation for the regression line of y on x in the form $y = a + bx$.
(Use $\sum x^2 = 30\,786$; $\sum xy = 41\,444$)
The selling price of each item produced is £2.20.

(c) Find the level of output at which total income and total costs are equal. Interpret this value.