Feedback About 2022's Exam in
DATA70121: Machine Learning and Statistics I

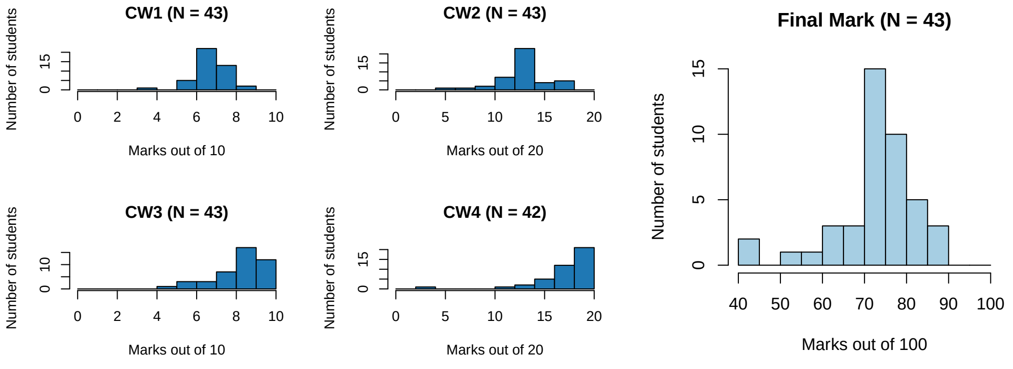

Figure 1: *Histograms of scores from the four coursework assignments and of the final mark for the course: note that the vertical scale and the total number of marks available varies from panel to panel. These are "raw" marks in the sense that they have neither been moderated by the Exam Board, nor adjusted to account for any Mitigating Circumstances, DASS allowances or late penalties.*

General remarks

- The assessment this year had five components: four courswork assignments that contributed a total of 60% of the final mark and an exam worth 40%.
- Many people did very welle: here is a summary of the final marks

Result:	Distinction	Merit	Pass	Fail
Range:	70–100	60–69	50–59	0–49
Number of students:	34	5	2	2
Fraction of students:	79%	11.6%	4.7%	4.7%

Remarks about individual exam questions

Q1 People did reasonably well here: the average was around $9.7/15$, with 2 perfect scores. Termeh Shafie, who marked the question, writes:

- In parts (a) and (b), many students missed that the x_i are constants (as stated 'fixed observables') and not random variables such that $E(c) = c$ and $\text{Var}(c) = 0$.
- Overall, everyone did rather well in part (c), mentioning that an intercept-free model leads to a bias, but that there are circumstances where this might be justified
- I was very impressed with how well students did in parts (d) and (e), with very good derivations offered by the majority.
- In the last part, I wished more students would have mentioned that MSE is equal to the variance of the estimators when they are unbiased and that is why we look for the one with smallest variance. Some even tried to compute the variances so as to compare them, but this is very difficult and only a handful of students got it right

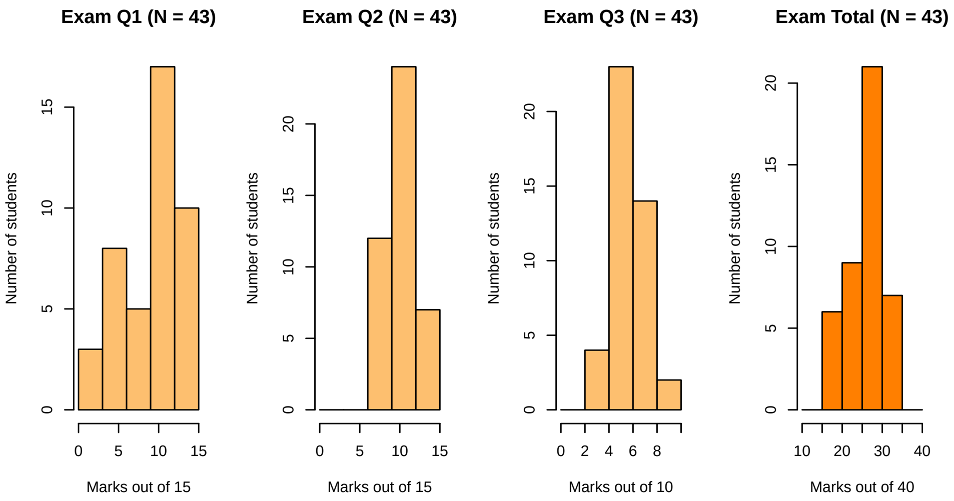

Figure 2: *Histograms of marks for the individual questions and the total for the exam: note that the vertical scale and the total number of marks available varies from question to question.*

Q2 Here too, people did reasonably well: the average was around $10.5/15$, with one perfect score. Mark Muldoon, who marked the question, writes:

- In part (a), students generally recognized that

  $$g(k) = \frac{N!}{k!\,(N-k)!}\, p^k (1-p)^{N-k}$$

  is the probability mass function for the Binomial distribution and said that it gives the probability of seeing k successes over N trials, where the probability of success on a single trial is p. Many also knew that

  $$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1}$$

  is the probability density function for the Beta distribution, which is supported on the interval $0 \leq p \leq 1$ and mentioned that the shape parameters α and β determine the form of the density. Fewer said that the mean of the distribution is $E(p) = \alpha/(\alpha+\beta)$ and the mode $(\alpha-1)/(\alpha+\beta-2)$, but I didn't deduct any marks for this.

  More problematic was widespread muddle about the use of the Beta distribution as a conjugate prior in Bayesian inference of a proportion or probability. Although the Beta distribution *can* be used this way—and we did so in lectures—it exists independently of this application.

- Part (b) depends on knowing that if one does a Bayesian estimate of the proportion p of students who prefer R to Python and if one uses the Beta distribution with $\alpha = \beta = 1$ as an (uninformative) prior, then if one polls N students and finds that k of them prefer R, the posterior for p is Beta-distributed with shape parameters

  $$\alpha' = k+1 \qquad \text{and} \qquad \beta' = N-k+1.$$

  The variance of the posterior is then

  $$\mathrm{Var}(p) = \frac{\alpha'\beta'}{(\alpha'+\beta')^2(\alpha'+\beta'+1)} = \frac{(k+1)(N-k+1)}{(N+2)^2(N+3)}$$

  which decreases like $1/N$. Thus we expect the *narrowest* posterior—in this case (c)—to be the one that corresponds to the largest sample. Many students chose (c), but argued that the key point was that it had the *tallest* peak, or the highest maximum. I deducted a mark for this unless it was accompanied by an argument explaining that the tallest peak must—as the total area under the posterior is always 1.0—necessarily be the narrowest too.

The Laplace approximation to a posterior density is essentially an appropriately shifted and scaled normal distribution and so the Laplace approximation works best when the posterior being approximated is symmetric about its mode. Of the examples in the exam, the most symmetric posterior is (b).

- The data behind these plots is a subset of that available at

  `https://archive.ics.uci.edu/ml/datasets/Leaf`

  I wanted to see the following points, made in a clear, but concise way: unduly terse or long-winded answers were penalised.

  – Most of the univariate distributions are bimodal, as are most of the joint distributions, especially those involving the Isoperimetric Factor, which suggests that there are data from two populations here.

  – Entropy does not seem an especially informative measurement.

  – The relationship between Elongation and Isoperimetric Factor seems most straightforward in that these variables are negatively correlated and lie, roughly, along a line.

Q3 The average on this question was around $6.3/10$ and here too, one student got a prefect score. Ke Chen, who marked the question, writes:

> Part (a) was set to test book knowledge while parts (b) and (c) require a certain level of understanding about model selection and assessment. Overall, the performance on Q3 was adequate, but there were some common issues as follows:
>
> - For part (a), most students properly described the under-fitting/over-fitting phenomena, but a number of students did not explain their cause and implications in an accurate manner.
> - For part (b), quite a few of students did not give an operable, yet practical procedure in their answers. In other words, one cannot see how to use the method given in their answer to detect under-fitting and over-fitting in statistical learning.
> - Similarly, a common issue in students' answers to part (c) is that they did not describe a generic, yet operable way to prevent both under-fitting and over-fitting in statistical learning. Several students suggested using "regularisation" techniques. While a regularisation technique is often effective to prevent over-fitting, it cannot prevent under-fitting in general.