Response to Assignment #6

## Why worry about the midterm answers?

The midterm test revealed a lot of shortcomings in students' understanding of the course material. Often students would know parts of what was needed in a question, but not the whole idea. A review of the midterm should be very instructive since the missing pieces can be filled in, in time for the rest of the course and the final exam.

## Why are these "concept" questions so common on my tests?

My objective is setting a test or exam is two fold: 1. To assess the degree to which students have gained a useful understanding of the material, and 2. By announcing the nature of the questions I will be asking, to motivate the mastery of a useful understanding of the material.

What motivates me to take this approach is what I view as the common but pointless emphasis on students learning calculation rituals: the approach to problems is then – try to recognize the type of problem, then fit the current numbers into that procedure, do the arithmetic, and hope that the answer is correct. This sort of "pluginski" learning is worse than useless – not only does it fail to guide practical application of the strategies (which requires understanding), but it makes students resent and scorn the anti-intellectual nature of the subject as they perceive it. Calculation exercises have a role in this course: it is to demonstrate what the calculations accomplish, and to reinforce the relationships among the technical components in a calculation. But, of course, there is much more to learn, and the additional stuff is much more interesting and more useful as well.

## Isn't the midterm material history now?

I will go over the midterm questions again and identify the concepts being examined as well as the common misunderstandings. A careful review of these comments should be instructive. Obviously, I will be re-examining many of these ideas on the final exam, partly because of the cumulative nature of the course material, and partly because I want to ensure that the pre-midterm concepts are mastered by all students.

## Related Feedback

These comments supplement those posted after the midterm titled "Midterm Feedback and Assignment #6". Please review those comments if you have not already done so.

## About the Midterm – Question by Question

**Q1.** My objective in this question was to see if conditioning was understood – I could have asked a question that required A and B and P(A|B), and the formal manipulations of symbols is useful since it is a check on heuristic understanding, but the heuristic understanding itself is also useful and in simple situations like the one in this question, it is safe. We did do in class (and the notes) an example just like this, and so I expected students to work directly from the data table provided.

Some students had trouble identifying the condition as a subset of the table. The wording needs to be carefully read. In part c, the "show method" instruction was just so that "Over 64" would not be a complete answer – I wanted to know how you got that outcome. There was a "?" missing here but the words did constitute a question, and so the method without the answer was not enough.

**Q2.** The sampling distribution of the mean is the key to getting sample-based information about the relationship of the sample mean to the population mean. As I said many times, it is the most useful and important bit of theory in the course. It is frustrating to me that many students have so far been unable to master this idea to the extent that they can use it in an application. So here is another try at conveying the idea:

**How to understand and make use of the Sampling Distribution of the Sample Mean**

An unknown population has an unknown mean μ and an unknown SD σ. We attempt to get some information about the unknown mean μ by observing the values in a random sample of size n from this population. We summarize the sample by calculating a sample mean $\bar{X}$ and a sample SD $s$. We want to use the statistic $\bar{X}$ to estimate the population mean μ. To judge the precision we can expect in estimating μ using $\bar{X}$, we need to think about this sampling process we have used, and imagine what would have happened in other random samples of size n. These imagined random samples would have resulted in other imagined values of $\bar{X}$. This collection of imagined values of $\bar{X}$ is referred to as the "sampling distribution of $\bar{X}$" for this situation. This sampling distribution of $\bar{X}$ is the key to judging how far $\bar{X}$ is likely to be from μ. If only we knew the mean and SD of the sampling distribution of $\bar{X}$, we would be able to say something about how far $\bar{X}$ is likely to be from μ. (Why? because the SD of $\bar{X}$ is the typical deviation of $\bar{X}$ from its mean, and its mean is μ.) But we **do** know the mean of $\bar{X}$, and we have a pretty good estimate of the SD of $\bar{X}$! The sample mean $\bar{X}$ is an unbiased estimate of μ and so the expected value (mean) of $\bar{X}$ is μ. And the SD of $\bar{X}$ is just the population SD divided by the square root of n (i.e. ${}^{\sigma}\!/\!_{\sqrt{n}}$), and this is estimated by ${}^{s}\!/\!_{\sqrt{n}}$. So we can say that $\bar{X}$ varies around μ and has a SD around μ in this sampling distribution of about ${}^{s}\!/\!_{\sqrt{n}}$. We might write that $\bar{X}$ is μ ± ${}^{s}\!/\!_{\sqrt{n}}$, which does tell us something about how close $\bar{X}$ is to μ.

But we can actually say more when n is large: Since the CLT tells us that this "sampling distribution of $\bar{X}$" is approximately a normal distribution, and since we have info about the mean and SD of this distribution, we can actually compute probabilities about the possible distances from $\bar{X}$ to μ. For example, P($\mu - {}^{s}\!/\!_{\sqrt{n}} < \bar{X} < \mu + {}^{s}\!/\!_{\sqrt{n}}$) = approx

P(-1<z<+1) = .68. The calculation on the midterm was P( $\bar{X}$ -μ < $s/\sqrt{n}$) where μ, $s$ and n were known.

[By the way, the new step in Ch 7 is to reorganize P(μ – $s/\sqrt{n}$ < $\bar{X}$ <μ + $s/\sqrt{n}$) into P($\bar{X}$ – $s/\sqrt{n}$ < μ < $\bar{X}$ + $s/\sqrt{n}$) to get a 68% CI]

The key to this calculation is to understand what $s/\sqrt{n}$ is the SD of $\bar{X}$ - hopefully you know this now?

The main mistake that was made in this question was to use $s$ instead of $s/\sqrt{n}$ in assessing the distance of $\bar{X}$ from μ. This was the most important indicator of understanding the theory, and was marked as a major error.

**Q3.** Nobody got a single mark on this question. Everyone apparently saw the word 'Poisson' and wrote down the formula for the Poisson probability, a discrete probability law that does not involve a density at all. The point that was missed was that the Poisson process provides a way to understand the connections between the Poisson, Exponential and Gamma distribution (and is the continuous analogue of our detailed discussion Binomial, Geometric and Negative Binomial models). As this theory is a personal favorite of mine, and which I highlighted in our discussions of Ch 3 and 4, I would have thought that it was an obvious candidate for the midterm. It is probably fair to count on it for the final exam too!

The key to the question was to realize that the waiting time from the occurrence of the third event to the occurrence of the fifth event was the same as the waiting time to the second event. If you realize that a Poisson process is made up of a sequence of event times that are separated by exponentially distributed inter-arrival times, then this is a consequence. Once you have identified the gamma with alpha = 2 as a model for this waiting time (3 marks), then the only remaining step is to figure out what beta should be (1 mark) and specializing the gamma density for this alpha and beta (1 mark).

I thought that the assignment in which you had to invent an example for each distribution would have reinforced understanding of these models. Apparently not. Please review the notes about the Poisson process, and also about the other model connections that I discussed in class and the notes.

**Q4.** This question was fairly well done as I would have expected. When I look carefully at some of the answers, there are hints that the answer was partially copied from the notes without real understanding, so if that is your situation you may wish to review the idea. Note that the bootstrap provides a way of assessing the variability of any statistic, not only the sample mean. Also, from the discussion of question 2, note that the variability of a statistic is often the key to finding a confidence interval for a parameter it estimates.

**Q5.** This question about the SD of the discrete uniform distribution might be though of as a "pluginski" question – but in order to use the formula, you had to have a couple of ideas in your head: that the discrete uniform spreads the total probability of 1 throughout all possible values in the sample space, and that the SD would only depend on the extent of the equally spaced values of the sample space. So this was a "conceptual understanding" question even for people who were aware of the formula provided during the review session.
There is another issue here. The several models we have studied: Bernoulli, Binomial, Negative Binomial, Geometric, Hypergeometric, Poisson, Uniform (Discrete and Continuous), Normal, Gamma, Chi-square and Lognormal all have a probability law (pmf or pdf and sometimes a cdf, explicitly given), a mean and an SD. The mean and the SD express the result of a summation or integral. These results should be assumed to be known – you can use them. If someone wants you to compute them from scratch, you will be asked to do that explicitly. Sometimes these sum/integral results are useful in other settings. For example, you can write down the integral over (0,inf) of $x^k$exp(-ax) as a function of k without actually doing any integration. (Use Gamma mean).

**Q6.** Ch 4 sections 1 and 2 are all about the relationship between pdf and cdf and how they are used to describe probability distributions of continuous RVs. It is not enough to know where to find the formulas relating these two – you need to understand (and be able to verbalize what it is) that the cdf at x is a sort of sum of probabilities up to x ($P(X \leq x)$) and that the pdf is the rate of increase of cdf with increase in x. In the quiz I asked you to draw the cdf of the continuous uniform distribution on (0,1) – did you understand why it was a straight line starting at (0,0) and ending at (1,1)? Calculus is a prerequisite-corequisite to this course – hopefully you have learned this idea: that integrals are like sums and derivatives are like differences.

One problem in this question is that many students did not see it was the "density" that was to be approximated at 5.5, not the cdf at 5.5. Perhaps this last question was being done in a hurry. Needless to say, everyone knows how important it is to allow enough time to at least try each question without rushing.

## General Advice

When you look over the questions, you will see that once the method is determined, the execution of the answer takes only a minute or two. There were no time-consuming calculations required. The midterm was really focused on assessing the understanding of the ideas, as promised. This is why I have stressed making the posted notes (and the lectures) a priority rather than the drill of exercises or ploughing through all the details in the text. These latter can be helpful, but they should not be the main or only strategy.