Stat 135, Fall 2006 A. Adhikari

HOMEWORK 1 SOLUTIONS

1.

$$\begin{aligned}\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 &= \frac{1}{n}(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\mu^2)\\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\mu\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}\mu^2\\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\mu^2 + \mu^2\end{aligned}$$

2. a) (iii) greater than 10

Because the section averages are different, the whole class will have somewhat more variability than each section. After all, the data now spread from the lowest scores in the weaker section to the highest scores in the strongest section.

b) Use the result of Problem 1. For Section 1,

$$\sum_{i=1}^{30} y_i^2 = 171750$$

For Section 2,

$$\sum_{i=1}^{20} w_i^2 = 74000$$

Now the class average, μ_{class}, is the weighted average

$$\frac{30\times 75 \ + \ 20\times 60}{50} = 69$$

Now,

$$\sigma^2 = \frac{1}{50}\sum_{i=1}^{50}x_i^2 - \mu_{class}^2 = \frac{1}{50}\times(171750+74000) - 69^2 = 154 \quad \Rightarrow \quad \sigma = 12.41$$

3. Write $x_i - c = (x_i-\mu)+(\mu-c)$ and play the same game as in Problem 1:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[(x_i-\mu)]+(\mu-c)]^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \ + \ \frac{2}{n}\sum_{i=1}^{n}(x_i-\mu)(\mu-c) \ + \ \frac{1}{n}\sum_{i=1}^{n}(\mu-c)^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \ + \ 0 \ + \ (\mu-c)^2\end{aligned}$$

(Why is the middle term equal to 0, you ask? Pull out the constant and see!) So you're left with the variance plus the square of something. That's always going to be at least as large as the variance, because the square is non-negative.

An alternative proof uses the heavier machinery of calculus:

$$\frac{\partial}{\partial c} mse_c = \frac{1}{n}(-2)\sum_{i=1}^{n}(x_i - c) = 0 \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} c = 0 \quad \Rightarrow \quad c = \frac{1}{n}\sum_{i=1}^{n} x_i = \mu$$

Convince yourself that this gives a minimum, not a maximum!

Now if we plug in μ for c, by definition we get that $mse_\mu = \sigma^2$.

4. a) $\hat{p}$=0.525 and the bootstrap standard error is

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.02497$$

Compare this with the conservative estimate for the standard error which is 0.025. So, the bootstrap confidence interval is $0.525 \pm (1.96)(0.02497)$ while the conservative confidence interval is $0.525 \pm (1.96)(0.025)$.

b) $\hat{p}$=0.275 and so the bootstrap standard error is computed as 0.0223. The conservative standard error is still 0.025. So, the bootstrap confidence interval is $0.275 \pm (1.96)(0.0223)$ while the conservative confidence interval is $0.275 \pm (1.96)(0.025)$. Note that in both cases the conservative confidence interval was centered around the same number as the bootstrap, but, it was a little wider. And, the further $\hat{p}$ is from 0.5, the bigger the difference in the interval width will be. However, they will be pretty close to one another until $\hat{p}$ gets near 0 or 1.

5. a) The standard deviation for the sample mean is estimated by $25/\sqrt{10} = 2.5$. So the approx. confidence interval is $75348 \pm 1.96 \times 2.5$ or (75343, 75353), roughly.

b) (ii) is true, the rest are false. First, (i) is false because the average of the measurements is simply known to be 75348 and there's nothing to estimate. Next, (iii) and (iv) are about the variability in the measurements themselves, not the variability in the average of the measurements. The variability in the measurements is of the order of 25, not 2.5.

c) Can't do it. There's nothing in the model about the shape of the distribution of the errors, so you don't know what the shape of the distribution of the measurements will be.

d) $n = 2,400$, roughly. We want the half-width of the confidence interval equal to 1. The half-width of the interval is just $1.96\frac{\sigma}{\sqrt{n}}$ so setting this equal to 1 we get (approximating from the data at hand)

$$1.96 \times \frac{25}{\sqrt{n}} = 1 \Rightarrow \sqrt{n} = 49$$

.

6. (9.05%, 13.19%). If you can find the percent of at-risk people in the sample, then the confidence interval is easily calculated as in Problem 4. So let's look at the sample of blood pressures. Its distribution is normal, with center equal to the center of the 99% confidence interval, which is 127.5 mm. You can figure out the SD of the sample by noticing that the half-width of the 99% confidence interval is $1.05 = 2.57\frac{SD}{\sqrt{n}}$ so the SD of the sample is 10.21. Since the sample closely follows the normal distribution we can figure out the percent of at-risk patients in the sample. In standard units, at-risk corresponds to $\frac{140-127.5}{10.21} = 1.22$ and higher. $P[Z \geq 1.22] = 0.1112$ so we estimate that 11.12% of the population is "at risk." The bootstrap standard error for this percentage is $\sqrt{\frac{0.1112(1-0.1112)}{n}} \times 100\% = 1.26\%$ and then the confidence interval is $11.12\% \pm (1.645)(1.26\%)$

It's fine to convert use 139.5 instead of 140 in the normal curve calculation above. Your answer will not be very different.

7. a) $n!$

b) $(n-1)!$ Fix card m_1 in place k_1 and permute the rest.

c) $\frac{(n-1)!}{n!} = \frac{1}{n}$

d) $\frac{1}{n(n-1)}$

There are (n-2)! permutations in which card number m_1 falls in place number k_1 and card number m_2 falls in place number k_2. So the probability is $\frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$

e) $E(M) = Var(M) = 1$

Write M as the sum of n (dependent) indicator variables $I_k, 1 \leq k \leq n$. I_k is one if there is a "match" at place number k and zero otherwise. Then each I_k has probability of success equal to $\frac{1}{n}$ so

$$E(M) = \sum_{k=1}^{n} \frac{1}{n} = 1$$

To find the variance we need to find $E(M^2)$. Recall that

$$M^2 \;=\; \sum_{k=1}^{n} I_k^2 \;+\; \sum_{j \neq k} I_j I_k$$

So

$$E(M^2) \;=\; \sum_{k=1}^{n} \frac{1}{n} \;+\; \sum\sum_{j \neq k} \frac{1}{n(n-1)} \;=\; n\frac{1}{n} \;+\; n(n-1)\frac{1}{n(n-1)} \;=\; 1+1 \;=\; 2$$

So $Var(M) = 2 - 1^2 = 1$.

f) The distribution converges to Poisson(1).

If the I_k were independent then M would be Binomial$(n, \frac{1}{n})$, which has expectation 1 and variance $1 - 1/n$ which goes to 1 as n gets large. The dependence goes away as n increases: if n is very large and we observe that the top card is card number 1 (a match), then that knowledge tells us very little about the chance that there is a match at any other spot in the deck. Now all you have to do is review the result that says that if $p_n \to 0$ as $n \to \infty$ while np_n remains constant at λ, then the limiting distribution of this Binomial (n, p_n) variable is Poisson(λ).

8. a) Random variable.

b) Real number. $E(X) = 3 - 4\theta$.

c) Random variable.

9. a) False. $\bar{X}_{(n)}$ is a random variable and $E(X)$ is a constant.

b) True. This is our old familiar result that says the expectation of the mean of an i.i.d. sample is just the population mean.

c) False. If n is large, there's only a small probability that $\bar{X}_{(n)}$ is exactly equal to $E(X)$, or

exactly equal to any other number for that matter.

d) True, by the Law of Large numbers, or by the CLT.

10. Set $\bar{X}_{(n)}$ approximately equal to $E(X) = 3 - 4\theta$ and solve for θ to get your estimate

$$\hat{\theta} = \frac{3 - \bar{X}_{(n)}}{4}$$

This construction has the comforting property that the expectation of the estimate is θ, which is the number you're trying to estimate:

$$E(\hat{\theta}) = \frac{3 - E(\bar{X}_{(n)})}{4} = \frac{3 - E(X))}{4} == \theta$$

In other words $\hat{\theta}$ is an unbiased estimate of θ.