STAT 830

Solutions: Assignment 1

1. The concentration of cadmium in a lake is measured 17 times. The measurements average 211 parts per billion with an SD of 15 parts per billion. Could the real concentration of cadmium be below the standard of 200 ppb? I want an answer in the form of a paragraph with NO formulas, no Greek letters. An answer is not 1 word long. I also want this turned in in the form of a document produced in LaTeX.

I would prefer to have answers to this question which acknowledged that **this is a statistics course and you are graduate level professional statisticians**. *You are being asked to answer a yes or no question on the basis of data. First, you need to ask how the data is connected to the question. How were the 17 measurements obtained? One bucket dipped in off a dock in Thunder Bay (on Lake Superior)? Two buckets in two places in mid November of 2010 with 9 subsamples from each, one of which was contaminated before processing? Seventeen samples at random locations in the lake? Should we worry about time effects? Should we worry about bias in the measuring device? You will need to ask questions like these and being a graduate student in statistics you need to be asking these questions all the time!*

Discussion: Before a statistician tells people the results of a t-test or produces a confidence interval s/he needs to find out if the proposed procedure is appropriate. When we think about that in the context of the question we need to ask a number of questions about the real world. I want to try to make a list of some of the points you might raise with the person bringing you the data – your client:

- You need to know how the data were collected.
- You need to know whether the 'concentration of cadmium' is a single number for the whole lake or whether it might vary from place to place. Is the concentration

higher near some places where water flows in to the lake? Is the concentration the same at every depth?

- If the concentration were known to be the same everywhere (or to vary by quite a bit less than the reported SD of the measurements) then the sampling design is less important. It would seem that the variability in the measurements is due to variability in the process of measuring the cadmium content of a sample. If the concentration is different in different places then you need to get clear what 'the concentration of cadmium' means. Is it the average concentration in the lake? Is it the highest concentration found anywhere in the lake? If the average concentration is the quantity of interest then you have to have a conversation about where, when, and how the 17 samples were gathered.
- Many of you said they would assume that the 17 measurements are a 'random sample' but this assumption needs to be faced up to in real world terms. Assuming that the concentration in the lake varies from place to place (or that you are worried it might vary) you want to make sure the random sample was in fact gathered by somehow dividing up the lake into many possible sampling locations and selecting a simple random sample of these locations.

For your answer I was hoping you would raise the issues in a couple of sentences. Announcing that you will make a technical, mathematical assumption is not the same as discussing why that assumption might or might not be a good match for what actually happened.

Beyond that I consider that testing the null hypothesis that the true concentration is less than or equal to 200 ppb against the one-sided alternative that it is larger and providing a tiny P-value is likely the right way forward (with the caveat that the method is likely flawed if there are outliers in the sample – say one big measurement and 16 others much smaller). The P-value you get from a one sided t-test is quite small so there is strong evidence that the standard is not met. Of

course you are not saying it could not possibly be met, only that the evidence against that is strong. I would not use a confidence interval of the form 211 plus or minus 2 standard errors because the question doesn't ask you to rule out high values – just low values; the plus or minus form is not one-tailed. Moreover it doesn't assess the strength of the evidence in the way a P-value does. I also would not do a formal 5% level test. I don't think 'the null hypothesis is rejected' is a real world conclusion. Of course, for a regulatory body, rather than for a person summarizing some evidence, there would need to be a clear rule for making a decision and a level α test for some sensible value of α might be a reasonable strategy.

2. Suppose X and Y are independent Geometric(p) random variables. In other words for non-negative integers j and k

$$\mathrm{P}(X = j \text{ and } Y = k) = \mathrm{P}(X = j)\mathrm{P}(Y = k) = p^2(1-p)^{j+k}\,.$$

WARNING: there are two standard definitions of Geometric distributions. The formula above specifies which I am talking about.

(a) Let $U = \min(X, Y)$, $V = \max(X, Y)$ and $W = V - U$. Express the event $U = j$ and $W = k$ in terms of X and Y.

For $k = 0$ the event $U = j, W = 0$ is the event $X = j, Y = j$. For $k > 0$ the event $U = j, W = k$ is the union of the two events $X = j, Y = j + k$ and $X = j + k, Y = j$. Several people did the two events separately, though the English is quite clear, I think.

(b) Compute $\mathrm{P}(U = j)$ and $\mathrm{P}(W = k)$ and prove that the event $U = j$ and the event $W = k$ are independent.

From a) $P(U = j, W = k) = 2(1-p)^2p^{2j+k}$ for $k > 0$. For $k = 0$ you get $P(U = j, W = k) = (1-p)^2p^{2j}$. This can be written as the product

$$(1-p^2)(p^2)^j \times \frac{2(1-p)}{1-p^2}p^k$$

for $k > 0$ and as

$$(1-p^2)(p^2)^j \times \frac{1-p}{1-p^2}$$

for $k = 0$ which is a product the Geometric(p^2) pmf and the pmf $P(W = k) = \frac{2}{1+p}p^k$ for $k > 0$ and $P(W = 0) = \frac{1}{1+p}$. This shows that U and W are independent and gives the margins. It was a common mistake not to separate out the case $k = 0$.
Here is a theorem. Suppose X and Y have a joint distribution on the set of all pairs of integers i, j. If there are functions g and h such that for every pair j, k we have

$$P(X = j, Y = k) = g(j)h(k)$$

then X and Y are independent and

$$P(X = j) = \frac{g(j)}{\sum_{j'} g(j')}$$

and

$$P(Y = k) = \frac{h(k)}{\sum_{k'} h(k')}.$$

Moreover,

$$\sum_{j'} g(j') \sum_{k'} h(k') = 1.$$

3. Each month Statistics Canada publishes data on employment and unemployment in Canada. My home page has a link to the Daily. In early September there will be a release of August data. Navigate from the Daily page to the LFS page and download the pdf – about 360 Kilobytes. I want you each to get one estimate from the tables in that document and compare it to the standard error for that estimate as follows:

   - Use the last digit of the day of the month which is your birthday to pick a province: 0 for Newfoundland, 1 for PEI, 2 for Nova Scotia, 3 for New Brunswick, 4 for Quebec, 5 for Ontario, 6 for

Manitoba, 7 for Saskatchewan, 8 for Alberta, and 9 for BC. The data for that province is in one of tables 4 or 5. Go to the section for men, 25 years and over, or women, 25 years and over according to whether or not you are male or female OR toss a coin to pick between those two possibilities. Please tell me what table and row you end up at.

- Get the estimated change, August minus July, and the associated Standard Error, the column labelled 'S.E.'.

  There is nothing to write about the first two parts of this except to say that 2020 is wildly different than other years with much bigger change from month to month than are usual.

- Does a 1 standard error confidence interval include the value 'no change'?

- Suppose that X has a normal distribution with mean μ and standard deviation σ. (In this question I have in mind that σ is known and you are making a confidence interval for μ which is unknown.) Give a formula, as a function of μ and σ, for the probability that a one standard deviation confidence interval includes the value 0. Your answer should be expressed as an integral involving the standard normal density

$$\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

  The confidence interval has ends $X \pm \sigma$ so we want

  $$P(X-\sigma < 0 < X+\sigma) = P(\sigma Z+\mu-\sigma < 0 < \sigma Z+\mu+\sigma)$$

  where $Z \sim N(), 1)$. Rewrite this in the form

  $$P(-\mu/\sigma - 1 < Z < 1 - \mu/\sigma) = \int_a^b \phi(u)du$$

  where $a = -1 - \mu/\sigma$ and $b = 1 - \mu/\sigma$.

- Show that this probability is maximized when $\mu = 0$ and tell me what the maximum probability is.

To maximize this probability we differentiate with respect to μ to get

$$\phi(b)\frac{db}{d\mu} - \phi(a)\frac{da}{d\mu} = -\frac{\phi(b) - \phi(a)}{\sigma}$$

Set this equal to 0 to discover $\phi(b) = \phi(a)$ which evidently implies $b = \pm a$. The equation $b = a$ is impossible while

$$b = -a \text{ iff } 1 - \mu/\sigma = -(-1 - \mu/\sigma) \text{ iff } -\mu = \mu$$

which means $\mu = 0$. The second derivative is easily found to be negative at $\mu = 0$ showing that this is a local maximum. Since the probability in question is a continuously differentiable function of μ which converges to 0 as $\mu \to \pm\infty$ and is positive at $\mu = 0$ the global maximum must occur at a critical point. There is only one critical point, at $\mu = 0$ so this must be the global maximum. For $\mu = 0$ we have

$$P(-1 < Z < 1) \approx 0.68.$$

4. The summary to lecture 1 contains a link to the `R` code I used for the Pearson-Lee height data. I used vertical strips centred at round numbers of inches for the fathers' heights in approximating the graph of $\mathrm{E}(S|F)$. I want you to try bars with widths 0.2, 0.4 ,0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0. Present one graph, properly labelled and captioned for the width 0.2 case which is otherwise like the one I showed in class. Add to that graph the corresponding approximation to $\mathrm{E}(F|S)$. (This is a different line than the other one.) Present another graph which permits comparison of the effects of bin width on the plots; you might use colours or line types in `R`, for instance. You can use my R code to get the data I used.

   I mark this looking for weaknesses like no units on axes, graphs which were unreadable, graphs which did not meaningfully summarize the differences between the plots for different bin widths. I wanted you to see that small bin widths

give more variable estimates – they jump around as you move across the picture. I wanted you to see that $E(F|S)$ and $E(S|F)$ are different. Some people seemed to say that the straight lines fitted by least squares were the 'true' regressions but this is unknowable. It is true that the binning method – a non-parametric estimation method – and the straight line ordinary least squares method produce similar plots. For this data the truth appears to be pretty close to a straight line.

5. Monte Hall was a television game show host from North Vancouver. He hosted, long ago, a popular show called Let's Make a Deal. In the show he would give contestants small prizes and offer to trade them the prize for some other thing that might or might not be more valuable. At the end of the show the two contestants who had one the most valuable prizes were allowed to trade in those prizes for a chance to play for the big prize of the day which was hidden behind one of three doors.

   This setting prompted someone, I don't know who, to dream up the following simplified version — a version which does not match the TV show as I remember it. In this version a single contestant picks a door, Monte then opens another door to show the contestant that this other door does not have a prize and then offers the contestant the opportunity to switch to the third door – the one which is not the one the contestant first picked and not the one Monte Hall opened.

   Please describe a sample space for this experiment and make a choice of probabilities for all the elementary outcomes. Explain clearly exactly why you choose these probabilities. Are they based on long run relative frequency ideas or are they subjectively motivated?

   Finally: should the contestant take the offer? Explain carefully why. I want to see the formal rules of probability used here — no hand waving.

   In doing this problem you have to begin by deciding what to include in the sample space. One comprehensive version has 4 elements: the door where Monte hides the prize, the door you pick, the door Monte opens and your decision 'switch' or 'do not switch'. It is okay to eliminate the last of these if you think it cannot be random and similarly you can eliminate

the door you pick by writing, say, 'Consider the case where the contestant picks door 1. The other two cases are analogous.' But I think there is merit in making sure that there is no impact of randomization for the contestant.

So a typical outcome is a sequence of 4 symbols: (h, c, o, s). I have h standing for the door where Monte Hall Hides the prize ($h \in \{1, 2, 3\}$) and $c \in \{1, 2, 3\}$ is the door the contestant chooses. Then o is the door to be opened. Again $o \in \{1, 2, 3\}$. Finally s is either S for switch or something like N for Not Switching.

Altogether Ω would have $3^3 \times 2 = 54$ outcomes. It is okay to omit the outcomes where $h = o$ since that is not permitted. It is also okay to omit the outcomes where $o = c$ since that is not permitted by the description. Or you keep all these outcomes and give them probability 0. If you do the elimination then altogether you rule out 30 of the 54 outcomes.

Now for some probabilities. Let H be the door Monte Hall picks. Students almost always assume $P(H = k) = 1/3$ for $k = 1, 2, 3$. This is sensible but deserves some explanation. It could be you are saying that this strategy is optimal for Monte. If Monte uses any other probabilities and you know those probabilities then you can improve your chances of winning beyond 2/3. **Please think about how!**

It also could be that you are saying that you don't know how Monte picks the doors and so are expressing your ignorance by saying no door is more likely than any other. That is, this might be a subjective probability.

Next let C be the door the contestant chooses. It doesn't matter what probability rule you use for $P(C = c)$ since the answer to the question doesn't depend on this rule. BUT you should recognize that you will assume that C and H are independent. That will be true unless there is some sort of collusion. Most students compute probabilities using this assumption but it is rare for them to make it explicit.

Next consider O, the door Monte opens. On the event that $H \neq C$ Monte has no choice about which door to open so for instance $P(O = 3|H = 1, C = 2) = 1$. Some students

did the problem by assuming that when Monte has a choice he makes the choice by tossing a fair coin. So for instance $P(O = 2|H = 1, C = 1) = 1/2$. This is another modelling assumption I wanted explained. Again you can either imagine Monte doesn't want to give you information or treat it as a subjective probability which you assess at $1/2$.

Now let us analyze the strategy: "whatever door Monte opens I switch to the other door" (That is I have made $P(D = S) = 1$ where D is the choice I make to switch or not.) Consider the event W that I win. Then the event I win is the event $H \neq C$. The probability of this event is

$$1 - P(H = C) = 1 - \sum_{j=1}^{3} P(H = j, C = j).$$

If H and C are independent then this simplifies to

$$P(W) = 1 - \sum_{j=1}^{3} P(H = j)P(C = j).$$

If we assume that $P(H = j) = 1/3$ as suggested above this becomes

$$P(W) = 1/(1/3) \sum_{j=1}^{3} P(C = j) = 2/3$$

which is more than the chance I don't win which is $1/3$.

Conversely if I use the strategy: "whatever door Monte opens I won't switch". Now the event W is just $\{H = C\}$ whose probability is $1/3$. So the chance of winning is larger for the always switch strategy.

Other questions might be asked: if I switch from door 1 if Monte opens door 2 and not if he opens door 3 how do I do? That chance depends on $P(O = 2|C = 1, H = 1)$ which *might* be $1/2$ if Monte tosses a fair coin to decide which door to open if he has a choice.

Other questions might be: given that Monte opens door 2 what is the chance that the prize is behind door 3? What

should you do if Monte hides the door behind door i with probability p_i which is not $1/3$?

Once upon a time a student considered the possibility of proving that no strategy of any kind had a higher probability of winning that always switching (this is true). Conceivably you might use a strategy where your probability of switching depended on the door you chose first and the door Monte opens. So there are 6 probabilities to specify to determine a complete strategy. But it will be seen that our analysis above was: unconditional, used independence of H and C, and relied on the assumption $P(H = j) = 1/3$ for each j.