# STAT 450

**Solutions: Assignment 1**

**My comments below on answers received do not apply to this class; I have not yet marked these papers.**

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? Imagine that you are answering this question for someone who is not a statistician and who brought you these numbers. Your answer **must** be in the form of a paragraph explaining the statistical points you are making and addressing any issues which might need to be clarified before giving advice. You may use 2 or 3 very simple formulas at most but a good answer can be given without any formulas at all.

   **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 $\alpha$ test for some sensible value of $\alpha$ might be a reasonable strategy.

2. Consider a population of 200 million people of whom 200 thousand have a certain condition. A test is available with the following properties. Assuming that a person has the condition the probability that the test detects the condition is 0.9. Assuming that a person does not have the condition the test detects (incorrectly) the condition with probability 0.001. A person is picked at random from the 200 million people and the test is administered.

   (a) What is the chance that the test detects the condition for this randomly selected person?

   I mean what is the chance the test is positive. Answer is

   $$0.9 \times \frac{200,000}{200,000,000} + 0.001 \times \frac{199,800,000}{200,000,000} = 0.001899.$$

   (b) Assuming that the condition is detected by the test for this randomly selected person what is the chance that the person has the condition?

   This is a Bayes theorem problem.

   $$\frac{P(\text{positive test}|\text{ has condition})P(\text{has condition})}{P(\text{positive test})}$$

   The denominator is in a) and the two pieces of the numerator are given in the problem. The answer is

   $$\frac{(0.9)(0.001)}{0.001899} = 0.474.$$

   (c) A mandatory testing program is contemplated. If all 200 million are tested about how many positive results should be expected? Of these about how many will not have the condition?

You multiply 200 million by answer in a) for the first. Of these the fraction in b) will have the condition and the rest won't. The whole point of the problem is that even a fairly accurate test applied to a group in which people don't have the condition will turn up a lot of false positives compared to the true positives. This is relevant both to mandatory testing programs and to early detection programs for conditions like cancer. If you start these programs at an early age a high fraction of the cases detected are false detections which can generate un-needed and unhealthy treatment.

3. You are presented with 2 boxes. One is known to contain two real diamonds and 1 fake. The other has two fakes and 1 real diamond. You are allowed to pick a box and test one stone picked at random from the box and then decide whether or not to take that box or switch for the other. Suppose you decide to switch if the tested stone is a fake. What is the chance that you will end up with two real diamonds?

   Let $A$ be the the event that the first box you pick has two diamonds. If you don't have any idea which box is which then you should assume $P(A) = 1/2$. Let $B$ be the event that the stone you pick is a diamond. Then

   $$P(B|A) = \frac{2}{3}$$

   and

   $$P(B|A^c) = \frac{1}{3}$$

   because in each case the fraction is the number of diamonds in the box you have picked out of the number of stones in the box. If you follow the stragegy of switching when the tested stone is a fake then you switch when $B^c$ happens. So you win if $A^cB^c$ happens (your first pick was the wrong box and the stone you tested was a fake) or if $AB$ happens (you picked the right box initially and the tested stone was a diamond).

So the chance I asked for was

$$\begin{aligned} P(AB) + P(A^cB^c) &= P(B|A)P(A) + P(B^c|A^c)P(A^c) \\ &= \frac{2}{3} \cdot \frac{1}{2} + \left(1 - \frac{1}{3}\right)\left(1 - \frac{1}{2}\right) \\ &= \frac{2}{3}. \end{aligned}$$

4. BONUS: From the text. Chapter 1, number 25, page 40. HINT: this problem is hard. The number of children in a family depends on cultural and personal attitudes concerning family size and gender distribution. You might think about what happens if the parents are determined to go on having children until at least one is a boy, for instance.

   The point here is that it is NOT POSSIBLE to answer this one without knowing something about culture and something about how the Smith family was selected for consideration. Imagine that every family in the Smith's culture wants to have at least 1 boy and 1 girl. Every family which gets one boy and one girl in the first two births stops having children. Those who get two the same try one more time. In this case either the Smith's have one of each or they are planning to try again but aren't yet old enough to have done so. If you knew the Smiths were my age the probability the other kid is a girl would be essentially 1 — most people my age are finished having kids (but then there is the possibility of infertility). If you knew they were 25 then it would be much harder to guess whether or not they have had time to try a third time.

   To make it simpler to see the problem consider a society in which every family has children until they get their first boy and then they stop. Since you know the Smiths have stopped (because you know they have a boy) and you know they have two kids, then their first kid was a girl and the second is a boy. You can make up lots of other behavioural patterns and

get lots of probabilities for the chance the other child is a boy – anything from 0 to 1 (which would apply if the family had kids until they had two boys and you knew they had quit).

If all families have exactly two children OR all families use a stopping rule which does not depend on the sexes of the kids they have so far then there is a straightforward answer. There are 4 possible outcomes: $\{GG, GB, BG, BB\}$ in obvious notation. Each has chance $1/4$ (assuming births are independent events with chance $1/2$ of having a boy. The chance of having a boy is, in fact, more than $1/2$ at birth. Then let A be the event "at least one boy" so that $A = \{GB, BG, BB\}$. The event $B$, that the other one is a girl" occurs if you get $GB$ or $BG$; it doesn't matter to the answer if you think "the other one is a girl" also includes $GG$. In either case

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{2}{4}}{\frac{3}{4}} = \frac{2}{3}.$$

5. Suppose $X$ is Poisson($\theta$). After observing $X$ a coin landing Heads with probability $p$ is tossed $X$ times. Let $Y$ be the number of Heads and $Z$ be the number of Tails. Find the joint and marginal distributions of $Y$ and $Z$.

Start with the joint:

$$\begin{aligned} P(Y = y, Z = z) &= P(X = y + z, Y = y) \\ &= P(Y = y | X = y + z) P(X = y + z) \\ &= \left( \begin{array}{c} y + z \\ y \end{array} \right) p^y (1 - p)^{y+z-y} \exp(-\lambda) \lambda^{y+z} / (y + z)! \\ &= (p\lambda)^y \exp(-\lambda p) [(1 - p)\lambda]^z \exp(-(1 - p)\lambda) / (y! z!) \end{aligned}$$

which clearly factors into the product of two Poisson probability mass functions. Thus $Y$ and $Z$ are independent Poissons with means $p\lambda$ and $(1 - p)\lambda$.

**Chapters 2 and 4**

You should be able to do all the questions numbered 2.1 to 2.9 although 2.7b and 2.8 are related to things I didn't cover explicitly.

6. Suppose $X$ has cdf $F_X$. Let $Y = X1(X > 0) = \max\{0, X\}$. What is the cdf of $Y$?

$$P(\max\{0, X\} \le y) = \begin{cases} 0 & y < 0 \\ F_X(y) & y \ge 0 \end{cases}$$

7. Suppose the density of $X$ is

$$f(x) = \begin{cases} Kx^2/(1+x)^5 & x > 0 \\ 0 & x \le 0 \end{cases}$$

Find the constant $K$ and the density of $Y = 1/(1+X)$.

Solve $Y = 1/(1+X)$ to get $X = 1/Y - 1$ and $dx/dy = -1/y^2$. The Jacobian is then $1/y^2$ and we get

$$f_Y(y) = K\frac{(1/y-1)^2}{(1/y)^5}\left|\frac{dx}{dy}\right| 1(1/y - 1 > 0)$$

Now do some algebraic simplification to get

$$f_Y(y) = Ky(1-y)^2 1(0 < y < 1)$$

Notice the simplification of the indicator. If $y < 0$ then $1/y - 1$ is negative so the indicator above must be 0. If $y > 0$ then $1/y - 1 > 0$ means $1/y > 1$ or $1 > y$. Compare this to the Beta$(a, b)$ density

$$f_Y(y) = \frac{1}{B(a,b)} y^{b-1}(1-y)^{a-1}1(0 < y < 1)$$
$$+ = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{b-1}(1-y)^{a-1}1(0 < y < 1)$$

to see that $K = 1/B(2,3) = \frac{4!}{1!2!} = 12$. Alternatively, you can multiply out $y(1-y)^2 = y - 2y^2 + y^3$ and do the integrals easily.

8. Suppose $X$ has density

$$f(x) = \begin{cases} 2xe^{-x^2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Find the distribution of $Y = X^2$.

Notice that $Y = X^2$ will be positive and that then $x = \sqrt{y}$ is the inverse. Then $dx/dy = \frac{1}{2}y^{-1/2}$ and the change of variables formula gives

$$f_Y(y) = 2\sqrt{y}e^{-(\sqrt{y})^2}1(0 < y < \infty)\frac{1}{2}y^{-1/2}$$

which simplifies to

$$f_Y(y) = e^{-y}1(y > 0)$$

which is the standard exponential density.

9. Suppose $X, Y$ have joint density

$$f_{XY}(x,y) = \begin{cases} ke^{-y}(1-e^{-x}) & 0 < x < y \\ ke^{-x}(1-e^{-y}) & 0 < y < x \\ 0 & \text{otherwise} \end{cases}$$

(a) Find $k$ so that $f_{XY}$ is a density.

You must show that $f$ is non-negative which is nearly obvious and then compute

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx &= k\int_0^{\infty}\int_0^{y} e^{-y}(1-e^{-x})dxdy \\ &\quad + k\int_0^{\infty}\int_0^{x} e^{-x}(1-e^{-y})dydx \\ &= k\int_0^{\infty} e^{-y}(y-(1-e^{-y}))dy \\ &\quad + k\int_0^{\infty} e^{-x}(x-(1-e^{-x}))dx \\ &= 2k\int_0^{\infty}\left(e^{-2x} - e^{-x} + xe^{-x}\right)dx \\ &= 2k(1/2 - 1 + 1) \\ &= k \end{aligned}$$

This shows that the function integrates to 1 if $k = 1$. Hence $f_{X,Y}$ is a density if $k = 1$.

(b) Find the marginal densities of $X$ and $Y$.

When I did the problem above I split the integral into two pieces and did the two pieces in opposite orders. To compute marginals I can't do that. For $x > 0$ I get:

$$\begin{aligned} f_X(x) &= \int_0^\infty f_{XY}(x,y)\,dy \\ &= \int_0^x ke^{-x}(1-e^{-y})\,dy + \int_x^\infty ke^{-y}(1-e^{-x})\,dy \\ &= e^{-x}\int_0^x (1-e^{-y})\,dy + (1-e^{-x})\int_x^\infty e^{-y}\,dy \\ &= e^{-x}(x-(1-e^{-x})) + (1-e^{-x})e^{-x} \\ &= xe^{-x} \end{aligned}$$

and by symmetry $f_Y = f_X$. Of course the densities are 0 at negative $x$ or $y$.