PHP 2500 Introduction to Biostatistics

Problem Set Three Solutions

---

1. (b) The number of ways of selecting four from a group of seven is

$$\binom{7}{4} = \frac{7\times6\times5\times4\times3\times2\times1}{(4\times3\times2\times1)(3\times2\times1)} = 35$$

(c) The number in the sample who have diabetes is a random variable, X, which has a binomial probability distribution with n = 7 and θ = 0.125. From the Binomial formula we have

$$P(X = 2) = \binom{7}{2}(0.125)^2(1-0.125)^5$$

$$= 21(0.0156)(0.5129)$$
$$= 0.1683$$

(d) $$P(X = 4) = \binom{7}{4}(0.125)^4(1-0.125)^3$$

$$= 35(0.000244)(0.66992)$$
$$= 0.0057$$

2. (b) The number of ways of selecting four from a group of 10 is

$$\binom{10}{4} = \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{(4\times3\times2\times1)(6\times5\times4\times3\times2\times1)} = 210$$

(c) The number of left-handers in a random sample of size 10 (with replacement) is a binomial random variable with n = 10 and θ = 0.098. We are asked to find the probability that this variable, *X*, will equal three:

$$P(X = 3) = \binom{10}{3}(0.098)^3(1-0.098)^7$$
$$= 120(0.00094)(0.4858) = 0.0549$$

2. (d) P($X \geq 6$) = P(X=6 or X=7 or X=8 or X=9 or X=10)

= P(X=6) + P(X=7) + ... + P(X=10)

$$= \binom{10}{6}(0.098)^6(1-0.098)^4 + \binom{10}{7}(0.098)^7(1-0.098)^3 + \ldots$$
$$\ldots + \binom{10}{10}(0.098)^{10}(1-0.098)^0$$

= 0.000123 + 0.0000076 + 0.0000003...
= 0.0001311

(e) "At most two" means two or fewer:

P($X \leq 2$) = P(X=0) + P(X=1) + P(X=2)

$$= (1-0.098)^{10} + 10(0.098)(1-0.098)^9 + 45(0.098)^2(1-0.098)^8$$

= 0.3565 + 0.3873 + 0.1894

= 0.9332

3. The number who "adhere to a sedentary lifestyle" (do not exercise regularly) in a random sample of size twelve is a binomial(12, 0.58) random variable.

(a) The probability that you will get a sample in which two or fewer do not exercise regularly is

P($X \leq 2$) = P(X = 2) + P(X = 1) + P(X = 0)

$$= 66(0.58)^2(1-0.58)^{10} + 12(0.58)^1(1-0.58)^{11} + (1-0.58)^{12}$$

= 0.0038 + 0.0005 + 0.00003

= 0.0043

3. (b) The probability that you will get a sample in which two or fewer exercise regularly (ten or more do not exercise) is

$$P(X \geq 10) = P(X = 10) + P(X = 11) + P(X = 12)$$

$$= 66 (0.58)^{10} (1 - 0.58)^{2} + 12 (0.58)^{11} (1 - 0.58)^{1} + (0.58)^{12}$$

$$= 0.050 + 0.013 + 0.001$$

$$= 0.064$$

(c) The probability of observing the sequence SFFSSSFFFSFF is the same as for FFFFFFFSSSSS because the events are independent and both sequences have the same number of S's and F's.

$$P(\text{SFFSSSFFFSFF}) = (P(S))^{5} (P(F))^{7}$$

$$= (0.58)^{7} (1 - 0.58)^{5}$$

$$= 0.0003$$

(d) Let X be the number of observed sedentary individuals. Then the probability of observing 10 individuals with a sedentary lifestyle given that three sedentary individuals were observed earlier is

$$P(X = 10 \mid X \geq 3) = P(X = 10 \text{ and } X \geq 3) / P(X \geq 3)$$

$$= P(X = 10) / (1 - P(X \leq 2))$$

$$= 0.050 / (1 - 0.0043) \quad \text{(from (a) and (b))}$$

$$= 0.0502$$

4. The number of cases is a random variable, X, and we are told that X has a Poisson probability distribution with λ= 4.5. This means that the probability that X takes the value k is

$$P(X = k) = (4.5)^k e^{-4.5} / k! \quad \text{for } k = 0,1,2,3, \ldots$$

(a) $P(X = 1) = (4.5)^1 e^{-4.5} / 1! = (4.5)(0.0111)/1 = 0.050$

(b) $P(X=0 \text{ or } X=1 \text{ or } X=2) = P(X=0) + P(X=1) + P(X=2)$

$$= (4.5)^0 e^{-4.5}/0! + (4.5)^1 e^{-4.5}/1! + (4.5)^2 e^{-4.5}/2!$$

$$= [(4.5)^0/0! + (4.5)^1/1! + (4.5)^2/2]\ e^{-4.5}$$

$$= [1 + 4.5 + (4.5)^2/2](0.0111)$$

$$= 0.0111 + 0.0500 + 0.1125$$

$$= 0.1736$$

(c) $P(X \geq 5) = ?$ This is an infinite sum. It will take a long time to calculate if we don't stop and think: The <u>complement</u> of "$X \geq 4$" is "$X \leq 3$." That is, $P(X \geq 4) + P(X \leq 3) = 1$, so

$$P(X \geq 4) = 1 - P(X \leq 3)$$

$$= 1 - [P(X=0) + P(X=1) + \ldots + P(X=3)]$$

$$= 1 - [0.0111 + 0.0500 + 0.1125 + 0.1687]$$

$$= 1 - 0.3423$$

$$= 0.6577$$

5. The number of suicides in a month is a random variable, which we will denote by X. We are told that X has a Poisson probability distribution with $\lambda = 2.75$. (This is because the quantity, λ, in the Poisson distribution is the average number of occurrences, and we are told that this is 2.75.)

$$P(X = k) = (2.75)^k e^{-2.75} / k! \quad \text{for } k = 0, 1, 2, 3, \ldots$$

(a) $P(X = 0) = (2.75)^0 e^{-2.75} / 0! = 0.0639$

(b) $P(X \le 4) = P(X = 0) + \ldots + P(X = 4)$

$$= (2.75)^0 e^{-2.75} / 0! + \ldots + (2.75)^4 e^{-2.75} / 4!$$

$$= 0.06393 + 0.17580 + 0.24173 + 0.22158 + 0.15234$$

$$= 0.855$$

(c) $P(X \ge 6) = ?$
This is just like (4c): $P(X \ge 6) + P(X \le 5) = 1$, so

$$P(X \ge 6) = 1 - P(X \le 5)$$

$$= 1 - [P(X=0) + P(X=1) + \ldots + P(X=5)]$$

$$= 1 - [0.06393 + 0.17580 + 0.24173 + 0.22158 + 0.15234 + 0.08378]$$

$$= 1 - 0.939 = 0.061$$

The following calculations are the same as above. The only difference is that our Poisson process has four months to observe events. Hence $\lambda = 2.75(4) = 11$ and the process is now over four months instead of one.

(d) $P(X = 0) = (11)^0 e^{-11} / 0! = 0.000017$

(e) $P(X \le 4) = P(X = 0) + \ldots + P(X = 4)$

$$= 0.0151$$

(f) $P(X \le 16) = P(X = 0) + \ldots + P(X = 16)$

$$= 0.9441$$

Notice that Parts (b) and (d) have different probabilities. This occurs because the variance of the Poisson implicitly changes with λ.

6. The number of infants that die in the first year is a binomial random variable with n = 2000 and θ = 0.00075. We are asked to find the probability that this variable, *X*, will be at most two:

(a) $P(X \leq 2) = P(X=2 \text{ or } X=1 \text{ or } X=0)$

$= P(X=2) + P(X=1) + P(X=0)$

$$= \binom{2000}{2}(0.00075)^2(1-0.00075)^{2000-2} + \binom{2000}{1}(0.00075)^1(1-0.00075)^{2000-1} + \binom{2000}{0}(0.00075)^0(1-0.00075)^{2000}$$

$= 0.2511 + 0.3348 + 0.223$

$= 0.8089$

(b) We have from (a): The probability that in a year no more than two infant deaths will occur is 0.8089. Because the years are independent, we have that

P(10 consecutive years with $\leq$ 2 deaths) = $(0.8089)^{10} = 0.1199$

(c) Let Y be a random variable that represents the number of years that no more than five infant deaths occur during their first year of life. We have from (a): The probability that in a year no more than two infant deaths will occur is 0.8089. Because the years are independent, we have that Y~bin(15,0.8089). So the probability of exactly 7 years is

$$P(X = 7) = \binom{15}{7}(0.8089)^7(1-0.8089)^8$$

$= 0.0026$

(d) Because $\lambda=n\theta$, we have that $\lambda=1.5$, and using X~Poiss(1.5):

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$=(1.5)^0 e^{-1.5} / 0! + (1.5)^1 e^{-1.5}/ 1! + (1.5)^2 e^{-1.5}/ 2!$

$= 0.2231 + 0.3347 + 0.251$

$= 0.8088$

The accuracy of the Poisson approximation seems good.

7. (a) The number of doctoral students among the four selected at random to receive A's has a Hypergeometric probability distribution.

(b)

$$P(\ X = 4\ ) = \frac{\binom{7}{4}\binom{73}{0}}{\binom{80}{4}} = \frac{35}{1581580} = 0.00002.$$

(c)

$$P(\ X \geq 1\ ) = 1 - P(\ X = 0\ ) = 1 - \frac{\binom{7}{0}\binom{73}{4}}{\binom{80}{4}} = 1 - \frac{(1)(1088430)}{1581580} = 1 - 0.688 = 0.312\ .$$

8. (a) There are $\binom{N}{n}$ equally probable samples, and you are in $\binom{N-1}{n-1}$ of them.

Thus the probability that the sample selected will be one of those that you are in is

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!\,(N-1-(n-1))!}}{\frac{(N)!}{(n)!\,(N-n)!}} = \frac{(N-1)!\,n!}{N!\,(n-1)!} = \frac{n}{N}$$

(The probability that you will be selected equals the sampling fraction, n/N.) Another approach is to imagine that we put N marbles into a jar -- N-1 white ones to represent the other students, and one black one to represent you. The probability that a simple random sample of size n will contain the black one is given by the Hypergeometric probability,

$$\frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}}$$

8. (b) One way to solve is

$$P(\text{Both selected}) = \frac{no.\,samples\,containing\,both}{total\,no.of\ samples}$$

$$= \frac{\binom{N-2}{n-2}}{\binom{N}{n}}$$

Another way to solve it uses the hypergeometric distribution: There are two black marbles, representing you and your friend. The probability of getting them both in a sample of size n is

$$= \frac{\binom{2}{2}\binom{N-2}{n-2}}{\binom{N}{n}}$$

Yet another way is (using part (a)):

$$P(\text{Both selected}) = P(\text{You selected})P(\text{Friend selected} \mid \text{You selected})$$

$$= \left(\frac{n}{N}\right)\left(\frac{n-1}{N-1}\right)$$

(You can verify that all of these answers are actually the same.)

8. (c) One way: $P(\ \text{Neither selected}) = \dfrac{\binom{N-2}{n}}{\binom{N}{n}}$.

Or using the Hypergeometric probability distribution, the probability of getting no black marbles in a sample of n is:

$$\frac{\binom{2}{0}\binom{N-2}{n-0}}{\binom{N}{n}} = \frac{\binom{N-2}{n}}{\binom{N}{n}}$$

Still another way:

$$P(\ \text{Neither}) = P(\ \text{Not you}\ )P(\ \text{Not friend} \mid \text{Not you}\ )$$

$$= \left(1 - \frac{n}{N}\right)\left(1 - \frac{n}{(N-1)}\right)$$

9. P(defective) = 0.04, so P(good) = 0.96.

(a) P(All good) = $(0.96)^{10}$ = 0.66

(b) P(10 or more good) = P(10 good) + P(11 good)

= $11(0.96)^{10}(0.04) + (0.96)^{11}$ = 0.93

(c) We showed in part (a) that P(10 or more good) is only 0.66, so that ten connectors are not enough. Similarly (b) showed that if we have eleven P(10 or more good) = 0.93 < 0.95, so eleven are not enough. Trying n=12 give us

P(10 or more good) = P(10 good) + P(11 good) + P(12 good)

$$= \binom{12}{10}(0.96)^{10}(0.04)^2 + 12(0.96)^{11}(0.04) + (0.96)^{12} = 0.99,$$

You need twelve connectors in order to be at least 95% certain that you have at least ten good ones.

10. From the table in Pagano we get:

    (a) $P(Z \geq 1.96) = 0.025$
    (b) $P(Z \geq -1.645) = P(Z < 1.645) = 1 - P(Z \geq 1.645) = 0.95$
    (c) $P(|Z| \geq 1.96) = 2P(Z \geq 1.96) = 0.05$ (by symmetry)
    (d) $P(Z \geq q)=0.80$ $q= -0.84$
    (by using the interior of the table to find the lower tail
    $P(Z < q)=0.20$ or $P(Z > -q)=0.20$. It is easiest to sketch a picture.)

11. From Appendix A:

    (a) $P(Z > 2.6) = 0.005$

    (b) $P(Z < 1.35) = 1 - P(Z > 1.35) = 1 - 0.0885 = 0.9115.$

    (c) $P(-1.7 < Z < 3.1) = 1 - P(Z < -1.7) - P(Z > 3.1)$
    $= 1 - P(Z > 1.7) - P(Z > 3.1)$
    $= 1 - 0.045 - 0.001$
    $= 0.954$

    (d) We must find the value, z, that satisfies this equation:

    $P(Z > z) = 0.15.$

    Looking in the body of table A.1 we find that the probability is 0.15 when z is (slightly less than) 1.04. (If we need greater accuracy, we can find a better table or use a computer to find z = 1.03643.)

    (e) We must find the value that cuts off probability 0.20 in the lower tail of the standard normal distribution. By symmetry, this is the negative of the value that cuts off the same probability in the upper tail. Again looking in the body of table A.1 we find that that value is 0.84. Therefore our answer is -0.84.

12. We're told that the random variable of interest, X, has a normal probability distribution with mean 77 and standard deviation 11.6

(a) $P(X < 60) = P(X - 77 < 60 - 77)$
$= P((X - 77)/ 11.6 < (60 - 77)/ 11.6)$
$= P(Z < -1.47)$
$= P(Z > 1.47)$
$= 0.071.$

(b) $P(X > 90) = P(X - 77 > 90 - 77)$
$= P((X - 77)/ 11.6 > (90 - 77)/ 11.6)$
$= P(Z > 1.12)$
$= 0.131.$

(c) $P(60 < Z < 90) = 1 - P(X < 60) - P(X > 90)$
$= 1 - 0.071 - 0.131$
$= 0.798$

13. Parts (a) and (b) are just like (12). The answers are

(a) 0.078
(b) 0.102

(c) From (a) and (b) we have that for each of the five males, the probability of falling inside the interval (130, 210) is 1 - 0.078 - 0.102 = 0.820.

Now the probability that least one will fall outside the interval is one minus the probability that they are all <u>inside</u> the interval. Since by saying that they are "selected at random from the population" Pagano is saying that they are independent, the probability that at least one will fall outside is
$1- (0.820)^5 = 1- (0.371) = 0.629.$

14. Our data are binomial. Let Y=number of participants who survived one year, then we have that Y~Bin(20, θ), where θ=P(surviving 1 year). We are interested in learning about θ.

    Now, we are told that if the chemo-regimen works then θ=P(surviving 1 year)=0.4, but if it does not work then θ=P(surviving 1 year)=0.28. There are our two scientific hypotheses, H_1: θ=0.28 and H_2: θ=0.4, and we want to find out which hypothesis is better supported by the data at hand: 8 survivors out of 20 participants.

    (a) We are also told that P(regimen works)=P(θ=0.4)=P(H_2)=.90 for the inventors. This also implies that P(regimen does not work)= P(θ=0.28)= P(H_1)=0.1.

    We are asked to compute P(regimen works | data)= P(H_2 | 8 success)=P(θ=0.4 | Y=8):

$$P(\theta = 0.4 \mid Y = 8) = \frac{P(Y = 8 \mid \theta = 0.4)P(\theta = 0.4)}{P(Y = 8 \mid \theta = 0.4)P(\theta = 0.4) + P(Y = 8 \mid \theta = 0.28)P(\theta = 0.28)}$$

$$= \frac{P(\theta = 0.4)}{P(\theta = 0.4) + \frac{P(Y = 8 \mid \theta = 0.28)}{P(Y = 8 \mid \theta = 0.4)} P(\theta = 0.28)}$$

$$= \frac{0.9}{0.9 + 0.514 \times 0.1} = 0.946$$

    So after observing these data, the inventors' guess that the regimen works increases from 90% to 95%.

(b) This is the same calculation as in Part (a), but now were are told that P(regimen works)=P(θ=0.4)=P(H_2)=.35 (these are the experts). This also implies that P(regimen does not work)= P(θ=0.28)= P(H_1)=0.65.

    We are asked to compute P(regimen works | data)= P(H_2 | 8 success)=P(θ=0.4 | Y=8):

$$P(\theta = 0.4 \mid Y = 8) = \frac{P(Y = 8 \mid \theta = 0.4)P(\theta = 0.4)}{P(Y = 8 \mid \theta = 0.4)P(\theta = 0.4) + P(Y = 8 \mid \theta = 0.28)P(\theta = 0.28)}$$

$$= \frac{P(\theta = 0.4)}{P(\theta = 0.4) + \frac{P(Y = 8 \mid \theta = 0.28)}{P(Y = 8 \mid \theta = 0.4)} P(\theta = 0.28)}$$

$$= \frac{0.35}{0.35 + 0.514 \times 0.65} = 0.512$$

So after observing these data, the experts' guess that the regimen works increases from 35% to 51%.

(c) The prior odds are 0.9/0.1=9 and 0.35/0.65=0.538 for inventors and experts respectively. The posterior odds are 0.946/(1-0.946)=17.52 and 0.512/(1-0.512)=1.05 for inventors and experts respectively. The posterior to prior odds ratios are identical for the inventors and experts: 17.52/9= 1.95 and 1.05/0.538=1.95 (this should be identical, but may vary slightly due to rounding).

(d) The likelihood ratio is LR= P(data| regimen works)/ P(data| regimen does not work)= P(8 success| H_2)/ P(8 success|H_1)= P(Y=8|θ=0.4)/P(Y=8| θ=0.28)=1.95=1/0.514. Notice where the LR shows up in the Bayes theorem calculation above (actually it is 1/LR=0.514).

(e) It is the likelihood ratio that describes how the prior odds of the regimen working changes into the posteriors odds of the regimen working. That is, it is the likelihood ratio that describes how the data modify the probability that the regimen works. So, in order to accurately describe what the data are saying, we need to report the likelihood ratio. Reporting the probability that the regimen works does not describe what the data themselves say, since it depends on the prior probability (what we thought to begin with).

(f) Yes the experts and inventors will agree now. The LR is now 776.98 in favor of the hypothesis that the regimen works (over 398 times greater than the small study – much stronger evidence). This has a dramatic effect on the posterior probability that the regimen works. Now, the investors' posterior probability is 0.999 (up from 0.90) and the experts' posterior probability that the regimen works is 0.9976 (up from 0.35). Notice how more data make a greater impact in this case, in an evidential sense, even though the estimated probability of survival stays the same.