Homework 2 Output

Emil Hafeez (eh2928)

10/3/2020

Problem 1

I use the shorthand $P(T) = P(TestPositive)$, and $P(T-) = P(TestNegative)$, and similar for $P(D)$ where D is the disease at hand.

Part 1

Given that "in a group of 113 patients with prostatic cancer, 79 have a positive diagnosis" and that "in a group of 217 individuals without prostatic cancer, 10 have a positive diagnosis,", we have the following table.

	With Prostatic Cancer	Without Prostatic Cancer
Test Positive (T)	79	10
Test Negative (T-)	34	207
Total	113	217

To calculate the sensitivity and the specificity of the test, we take the sensitivity to be equal to $P(Test|Disease)$ and the specificity to be $P(T-|D-)$. For the sensitivity, equates to $\frac{79}{113}$ and for the specificity to be $\frac{207}{217}$, which is to say the sensitivity equals 0.6991 or 69.91% and the specificity equals 0.9539 or 95.39%.

Part 2

In this other hypothetical scenario, it will not be enough to use only the data provided by this new test being developed to assess its test characteristics like sensitivity and specificity. This is because determining sensitivity and specificity requires reference to a gold-standard test, so that we can compare the accuracy of the new test data to a standard for which the "truth" is known. This reference is missing. With just the new test data available, this investigation cannot be made.

Part 3

a) In this example, the $P(T+|D+)$ is given as 0.8 and the $P(T-|D-)$ as 0.95. The $P(D)$ is 0.5, also given. Therefore, we can use Bayes' Theorem and then apply the LTP to the denominator to determine the following $P(D|T)$.

$$P(D|T) = \frac{P(T|D)*P(D)}{P(T)} = \frac{P(T|D)*P(D)}{P(T)} = \frac{P(T|D)*P(D)}{P(T|D)P(D)+(P(T|D-)P(D-)} = \frac{(0.8)(0.5)}{(0.8)(0.5)+(0.05)(0.5)} = 0.94117.$$

This is known as the positive predictive value (PPV).The value of the PPV is higher when the pretest likelihood of prostatic cancer is higher, and conversely, the value of the PPV is lower when the pretest likelihood of prostatic cancer is lower.

b) We now calculate the same test characteristic as above in Part 3a, but now we decrease the pre-test probability of the disease to 0.1 instead of 0.5.

$$P(D|T) = \frac{P(T|D)*P(D)}{P(T)} = \frac{P(T|D)*P(D)}{P(T)} = \frac{P(T|D)*P(D)}{P(T|D)P(D)+(P(T|D-)P(D-)} = \frac{(0.8)(0.1)}{(0.8)(0.1)+(0.05)(0.9)} = 0.64.$$

Problem 2

Part 1 The probability that zero of these 50 patients are prediabetic is given by the following equation. We observe that the scenario follows a binomial distribution where the number of patients is $n = 50$ and the probability of success $ p = 0.345)$, X~$Bin(50, 0.345)$ where X is a random variable denoting the number of patients selected that are prediabetic among the 50 randomly selected group. We observe the binomial distribution given that there are a fixed number of trials, trials are independent, there are only two possible outcomes (success or failure on each draw), and the probability of success if fixed for each trial.

We observe that the scenario follows a binomial distribution X~$Bin(50, 0.0345)$ where X is a random variable denoting the number of patients that are prediabetic among the 50 randomly selected group, where n is the number of patients total, p is the probability of success, and $1 - p$ is the probability of failure, such that $n = 50, p = 0.345, (1 - p) = 0.655$,

$\mathrm{P}(X = x) = \sum_x^n \frac{n!}{x!(n-x)!}(p)^x(1-p)^{n-x}$

$P(X = 0) = \frac{50!}{0!(50-0)!}(0.345)^x(1-0.345)^{50-0}$

Using R `dbinom(0,50,0.345)` we can computer that $P(X = 0)= 6.487315e-10.

Part 2 $P(X < 9) = \sum_{x=0}^{9} \frac{9!}{x!(9-x)!}(0.345)^x(1-0.345)^{9-x} = 0.008153186$

We can use r to compute the code "pbinom(9, 50, 0.345)". We use 9 rather than 10 because we are computing less than 10, not equal to 10.

Part 3 The probability that 34.5% or $\frac{17}{50}$ of these patients are prediabetic is given by the expected value of the binomial distribution, similar to Question 2 Part 1, where $P(X = 17) = \frac{50!}{17!(50-17)!}(0.345)^{17}(1-0.345)^{50-17}$. This value is given by `dbinom(17,50,.345)` $= 0.1180887$.

Part 4 In some conditions, the Binomial distribution can be approximated by the Poisson distribution: when n is large (> 100) and when the probability of success is small $p < 0.01$. These conditions are not met, and so it would not be appropriate to use this approximation method.

Problem 3

Part a)

Here, the incidence of uveal melanoma in the US can be assumed to follow a Poisson distribution characterizing the number of occurrences of an event (uveal melanoma) given that the events occur one at a time (one case a time per person, no one person can get it twice simultaneously), and the number of expected events is given as a constant. In this case, the probability that exactly 30 cases occur in a given year in NYC is given by the following equation.

$P(X = x) = f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, \ldots, n$

In this circumstance, we use λ as the overall rate for the country adjusted to the NYC population, such that 5 per million at a population of 8.3 million expects a rate of $\lambda = 41.5$ cases in a given year. Therefore, $P(X = 30) = \frac{41.5^{30}e^{-41.5}}{30!}$

We can calculate this in R using the code "dpois(30, 41.5)", so $P(X = 30) = 0.01243$

Part b)

To conduct this part, we use λ as the expected number of events in a given year, the race-specific rate of disease per million computed a population of 8.3 million (where Asian people are .14 of the population, non-Hispanic whites are 0.428, and Black people are 0.243).

For example, for the 6.02 per million in non-Hispanic whites, which constitute 0.428 of the NYC population, $\frac{6.02}{1000000} = \frac{\lambda}{8300000*0.428}$ such that $\lambda = 3.5529$. We can do this for each racial group, so:

Racial group	Racial Demographic proportion in the NYC population	Race-specific λ
Asian	0.14	0.45318
non-Hispanic white	0.428	3.5524
Black	0.243	0.d

Now, given the parameter λ, we can compute the corresponding probabilities in the Poisson distribution using the R code cited previously, "dpois(x,λ"). For Asian people, $P(X = 30) = \frac{0.45318^{30}e^{-0.45318}}{30!} = 1.168553e - 43$, for non-Hispanic white people it's $P(X = 30) = \frac{3.5524^{30}e^{-3.5524}}{30!} = 3.54167e - 188$, and for Black people it's $P(X = 30) = \frac{0.625239^{30}e^{-0.625239}}{30!} = 1.535267e - 39$.

While each of these probabilities are very small, we can see that the probability of 30 cases in a given year among non-Hispanic white people is much larger than Asian or Black people. While we hold X=30 constant, the higher rate per million and the higher representation of non-Hispanic white people in the NYC population combine to form a higher λ than either other group. This is reflected in the corresponding probabilities computed for each racial group.

Problem 4

Part 1

Given that an independent random sample of size > 30 (sufficiently large) is drawn from the population of sick individuals at Columbia University Medical Center, the shape of the sampling distribution is approximately normal ($\overline{X}$~μ, $\frac{\sigma^2}{n}$) such that $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$.

The probability that the sample mean is less than 98 degrees Fahrenheit is given by determining the corresponding critical value and its one-sided probability on the t-distribution, associated with the given sample statistics. The population mean is $\mu = 99.9$, the standard deviation is $\sigma = 0.73$ which we can square to get the variance, and in this case the sample size $n = 40$ randomly chosen individuals.

The $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = Z = \frac{98-99.9}{0.73/\sqrt{40}}$ so that the $P(X < 98) = P(Z < -16.46117138)$ = "pnorm(98,99.9,sd=.1154231,lower.tail = TRUE)" = 3.486602e-61.

Part 2

The probability that the sample mean is greater than 100.5 degrees Fahrenheit is given by determining the corresponding critical value and its one-sided probability on the t-distribution, associated with the given sample statistics. The population mean is still $\mu = 99.9$, the standard deviation is $\sigma = 0.73$ which we can square to get the variance, and in this case the sample size $n = 40$ randomly chosen individuals.

The $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = Z = \frac{100.5-99.9}{0.73/\sqrt{40}}$ so that the $P(X > 100.5) = P(Z < -5.19826)$ = "pnorm(100.5, 99.9,sd=.1154231,lower.tail = F)" = 1.00578e-07.

Part 3

The 90th percentile of the sampling distribution of the sample mean is given by the value corresponding to the critical value where 90% of the AUC is to the left; this is given by qnorm(0.9, mean = 99.9, sd=.1154231, lower.tail = T) = 100.0479 degrees Fahrenheit.

Part 4

The cutoff values for the middle 50% of the sampling distribution of the sample mean.

To get the cutoff values for the middle 50% of the sampling distribution of the sample mean $\overline{X}$, we need to find the critical values which correspond to containing the 25% AUC probability to the left of the sample mean and 25% to the right of the sample mean. These are the 25th and 75th percentiles. Therefore, from `"qnorm(0.25, mean = 99.9, sd=.1154231, lower.tail = T)" and "qnorm(0.75, mean = 99.9, sd=.1154231, lower.tail = T)"` or by taking the first one and applying symmetry.

These give us cutoff values for the middle 50% of the sampling distribution of the sample mean $\overline{X}$ of 99.82215 and 99.97785.