Homework2

Jeffrey Liang

10/4/2020

Problem 1

A new test is being developed for the detection of carcinoma of the prostate (Foti et al. N Engl. JMed. 1977). When it is tested in a group of 113 patients with prostatic cancer, 79 have a positive diagnosis; in a group of 217 individuals without prostatic cancer, 10 have a positive diagnosis.

1) Calculate the specificity and sensitivity of the test. (4p)
2) In another hypothetical scenario, it is planned to use the test to screen a large sample of subjects for prostatic cancer where the test results will be the only data available. Is this information enough to assess the test characteristics, i.e., sensitivity & specificity? Yes, no, what is missing (if the case)? (2p)
3) In patients with palpable prostatic nodules, the pretest likelihood of prostatic cancer is 50%. The test under these conditions has a sensitivity of 80% and a specificity of 95%.

a) Calculate the probability of a patient with a palpable prostatic nodule and a positive test result having prostatic cancer. What is the epidemiological term of this probability? (3.5p)
b) Re-calculate the probability in 2-a), using a pretest likelihood of prostatic cancer of 10%. Compare the two values and comment on their differences. (3.5p)

1.

$$\begin{aligned} sensitivity &= \frac{true\ positive}{true\ positive + \ false\ negative} \\ &= \frac{79}{113} \\ &= 0.6991 \end{aligned}$$

$$\begin{aligned} specificity &= \frac{true\ negative}{true\ nagative + \ false\ positive} \\ &= \frac{79}{113} \\ &= 0.0461 \end{aligned}$$

2. NO, the sensitivity and specificity are computed base on comparision to the test result of gold standard. And thus, given the data of this test only are not sufficient to calculate sensitivity and specificity of this test.

3. a.

$$\begin{aligned} &P(prostatic\ cancer \mid Positive \cap Palpable\ Prostatic\ Nodule) \\ &= \frac{P(Positive \cap Palpable\ Prostatic\ Nodule\ \cap prostatic\ cancer)}{P(Positive \cap Palpable\ Prostatic\ Nodule)} \\ &= P(Positive \cap Palpable\ Prostatic\ Nodule\ \cap prostatic\ cancer) \div \\ &\{P(Positive \mid Palpable\ Prostatic\ Nodule \cap prostatic\ cancer) \times \\ &P(Palpable\ Prostatic\ Nodule \cap prostatic\ cancer) + \\ &P(Positive \mid Palpable\ Prostatic\ Nodule \cap prostatic\ cancer^c) \times \\ &P(Palpable\ Prostatic\ Nodule \cap prostatic\ cancer^c)\} \end{aligned}$$

$$= \frac{80\% \times 50\%}{80\% \times 50\% + 5\% \times 50\%}$$
$$= 0.9412$$

This is the positive predictive value of a test.

b.

$$P(prostatic\ cancer' \mid Positive \cap Palpable\ Prostatic\ Nodule)$$
$$= \frac{P(Positive \cap Palpable\ Prostatic\ Nodule\ \cap prostatic\ cancer')}{P(Positive \cap Palpable\ Prostatic\ Nodule)}$$
$$= P(Positive \cap Palpable\ Prostatic\ Nodule\ \cap prostatic\ cancer') \div$$
$$\{P(Positive \mid Palpable\ Prostatic\ Nodule \cap prostatic\ cancer') \times$$
$$P(Palpable\ Prostatic\ Nodule \cap prostatic\ cancer) +$$
$$P(Positive \mid Palpable\ Prostatic\ Nodule \cap prostatic\ cance'r^c) \times$$
$$P(Palpable\ Prostatic\ Nodule \cap prostatic\ cancer'^c)\}$$
$$= \frac{80\% \times 10\%}{80\% \times 10\% + 5\% \times 90\%}$$
$$= 0.64$$

Noticing that lower disease prevalence/ prior produce a lower positive predictive value. This is similar to the example given on class regarding the low positive predictive value base on the low prevalence of COVID-19 in some communities. ■

Problem 2

According to the Center for Disease Control (CDC), about 34.5% of the adult US population are prediabetic (National Diabetes Statistics Report, 2020). Suppose we randomly select a group of 50 patients seen at Columbia University Medical Center. Calculate the following exact probabilities based on the national statistics:

1) Probability that none of these patients are prediabetic. (2.5p)
2) Probability that less than 10 patients are prediabetic. (2.5p)
3) Probability that 34.5% of these patients are prediabetic (round to the nearest integer). (2.5p)
4) Could you use an approximation method to calculate the probabilities above? If yes, calculate the probabilities using approximations and compare to the exact values; otherwise, explain why approximations methods are not appropriate. (2.5p)

Proof

1. With the information given, any sample from the population has probability of $p = 34.5$ suffer from prediabetic. Define $\mathcal{X}$ as the random variable(R.V.) , then the probability of number of samples from population suffer from prediabetic follows the distribution of Binomial of $Bin\ (50, .345)$. Therefore we have the distribution of:
$$P(\mathcal{X} = k)\ = \binom{n}{k} p^k (1-p)^{n-k}$$
And the probability of none of these patients have prediabetics is:
$$P(\mathcal{X} = 0)\ = \binom{n}{0} p^0 (1-p)^{n-0}\ = 6.4873 \times 10^{-10}$$

2. Use the CDF of binomial distribution:
$$P(\mathcal{X} < 10)\ = \sum_{k=0}^{9} \binom{n}{k} p^k (1-p)^{n-k}\ = 0.0082$$

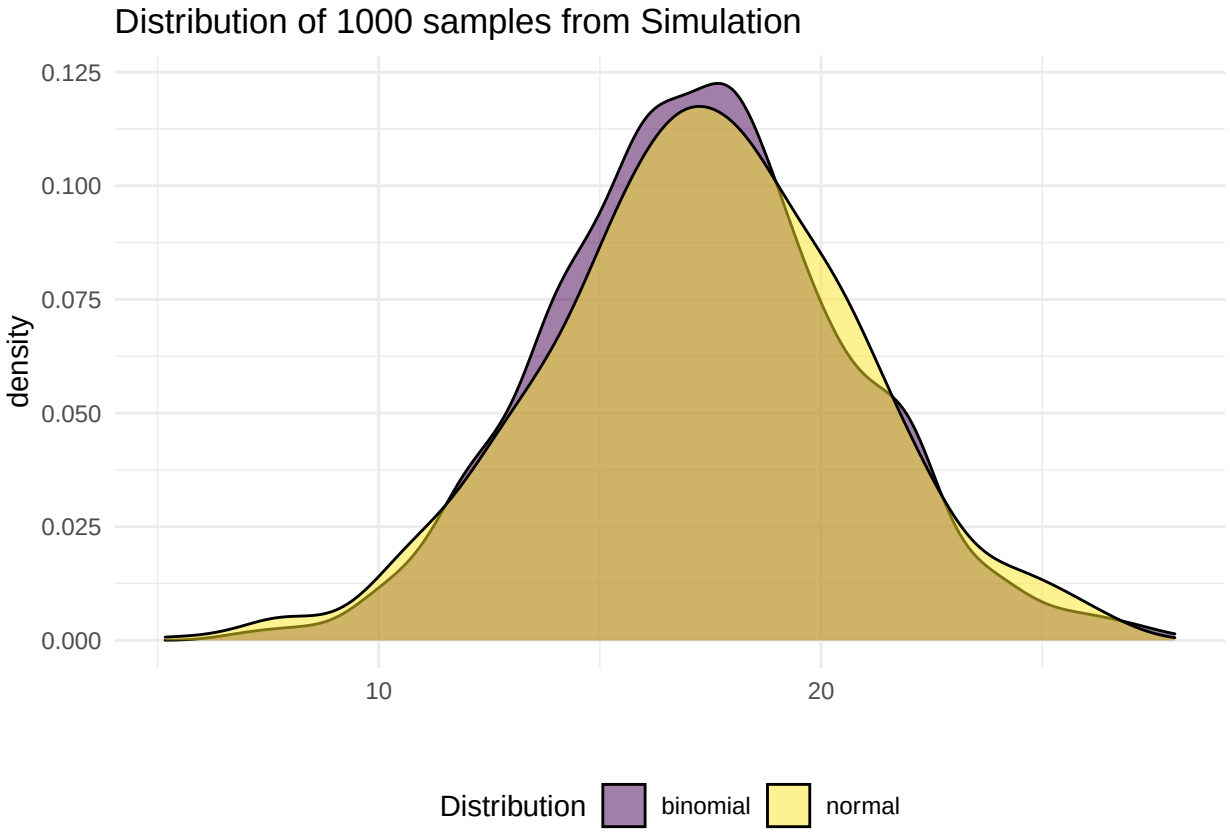
3. The probability 34.5% proportion of the patients are prediabetic is: $P(\mathcal{X} = 50 \times 34.5\% \approx 18) = \binom{n}{18} p^{18}(1-p)^{n-18} = 0.114$

4. the given $np = 17.25 > 10$, and $n(1-p) = 32.75 > 10$, it can be approximate by Normal distribution of $X \sim N(np, np(1-p))$.

    1. $P(\mathcal{X} = 0) \approx P_N(\mathcal{X} < 0 + 1/2)$
    $= 3.1287 \times 10^{-7}$

    2. $P(\mathcal{X} < 10) \approx P_N(\mathcal{X} < 9 + 1/2)$
    $= 0.0106$

    3. $P(\mathcal{X} = 18) \approx P_N(18 - 1/2 < \mathcal{X} < 18 + 1/2)$
    $= 0.1154$



We observed a heavier tail on both side of normal distribution compared with binomial.The normal tail is heavier than the binomial, at least far enough from the mean, where the binomial tail is just 0. And thus the probability approximation from normal distribution of the first two results are higher than the actual probabilities from binomial. ■

Problem 3

The incidence of uveal melanoma in the US is approximately 5 per million individuals per year, with a significantly higher incidence in non-Hispanic Whites (6.02 per million), when compared to Blacks and Asians: 0.31 and 0.39 per million, respectively.

a) What is the probability that in NYC (population of 8.3 million reported in 2020), exactly 30 cases occur in a given year? (4p)

b) Compute the same probability in a) by the mentioned racial/ethnic groups and comment on the findings. Demographic data of NYC in 2020: 14.0% Asians, 42.8% non-Hispanic Whites, 24.3% Black. (6p)

Proof

1. There isn't evidence that the total incidence of uveal melanoma in NYC is different from the US. And given the incidence is rare(one digit per million a small p) and calculating the probability in large population(a large n) in period of year 2020, a Poisson distribution can be used to approximate the distribution of population. with the information in the text, $\lambda = 5 \times 8.3 = 41.5$, so the number of cases (defined as the R.V. of $\mathcal{X}$)is follow $\mathcal{X} \sim Poi(41.5)$.

$$P_X(x = 30) = e^{-\lambda}\lambda^k/k! = 0.0124$$

2. There're 1.162 million asians, 3.5524 million non-Hispanic white, 2.0169 million black and 1.5687 million other unspecified ethnic/racial groups. The λ for these subpopulation within their group are 0.4532, 21.3854, 0.6252 and 7.8435 respectively.

$$P_{X\ asian}(x = 30) = \frac{e^{-0.4532} \times 0.4532^{30}}{30!} = 1.1686 \times 10^{-43}$$

$$P_{X\ non\ hispanic\ white}(x = 30) = \frac{e^{-21.3854} \times 21.3854^{30}}{30!} = 0.0156$$

$$P_{X\ black}(x = 30) = \frac{e^{-0.6252} \times 0.6252^{30}}{30!} = 1.5353 \times 10^{-39}$$

The sum of λ' is 30.3074 for the updated NYC population by joint independent distribution of poisson random variables. The pdf of the Poisson distribution is

$$p(k) = \frac{\lambda^k}{k!}\exp(-\lambda)$$

The joint pdf is the product of the pdfs of all independent variables $z = \sum x_i$. Which is given by

$$P_Z(z) = \sum P_Z(x_1 + x_2) = \sum P_{X_1}(x_1) \times P_{X_2}(z - x_1)$$

which can be siplified to

$$\begin{aligned}\sum P_{X_1}(x_1) \times P_{X_2}(z - x_1) &= \sum \frac{\lambda_1^{x_1}\exp(-\lambda_1)}{x_1!} \times \frac{\lambda_2^{z-x_1}\exp(-\lambda_2)}{z - x_1!} \\ &= \sum \frac{z!}{(z - x_1)!x_1!} \times \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \times \lambda_1^{x_1} \times \lambda_2^{z-x_1}\end{aligned}$$

by binomial theorem, this is equal to:

$$\begin{aligned}P_Z(z) = \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1 + \lambda_2)^z}{z!} \\ let\ \lambda = \lambda_1 + \lambda_2 \\ Z \sim Poi(\lambda)\end{aligned}$$

So, we have

$$P_X(x = 30) = e^{-\lambda'}\lambda'^k/k! = 0.0725$$

Looking at the result, we noticing that, given the high incidence in Non-hispanic white population, its probablitly of having 30 cases is higher than the original probability of nyc and other subpopulations. Comparing the asian and black population, they have similar incidence, but the proportion of black population is some 1.5 times of asian, the probability of 30 cases in black is 10^4 higher than asian's.

Considering the whole updated population, because updating the λ' is lower than the US's incidence and closer to the event we seek, therefore the pobability we have 30 cases is higher in updated nyc than in the non-updated nyc. ■

Problem 4

People with COVID-19 have had a wide range of symptoms, but high temperature is one of the most important indications. Based on current data, the average temperature of patients diagnosed with COVID-19 follow a normal distribution with a mean of 99.9 degrees Fahrenheit and a standard deviation of 0.73 degrees Fahrenheit. Of course, this differs from person to person, based on factors like body weight, height, the weather, age or gender. Let $X_1 ... X_{40}$ be the body temperatures of 40 randomly chosen individuals returning to the Columbia University Medical Center. Calculate the following probabilities:

1) $(\bar{X} < 98)$, the probability that the sample mean is less than 98 (average normal temperature). (2.5p)
2) $(\bar{X} > 100.5)$, the probability that the sample mean is greater than 100.5 (alarming zone for COVID-19). (2.5p)
3) The 90th quantile of the sampling distribution of the sample mean $\bar{X}$. (2.5p)
4) The cutoff values for the middle 50% of the sampling distribution of the sample mean $\bar{X}$(2.5p)

Proof

1. We know the sample mean and its standard deviation of the population. Thus the sample distribution of these 40 patients given the null hypothesis following a $\mathcal{N}(99.9, \frac{0.73^2}{40})$ distribution.

$$\begin{aligned} P_{\bar{X}}(\bar{x} < 98) &= \phi(k < \frac{98 - 99.9}{.73/\sqrt{40}}) \\ &= 3.4869 \times 10^{-61} \end{aligned}$$

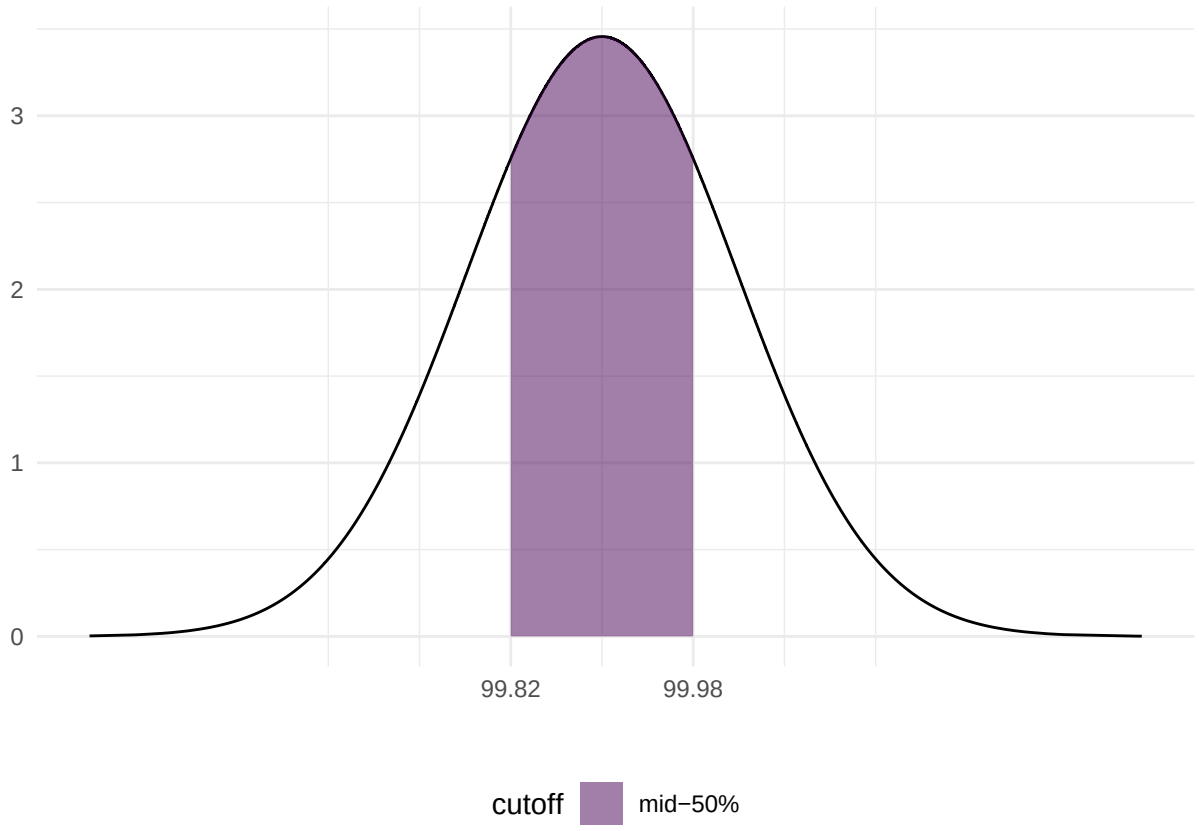
2.

$$\begin{aligned} P_{\bar{X}}(\bar{x} > 100.5) &= 1 - \phi(k < \frac{100.5 - 99.9}{.73/\sqrt{40}}) \\ &= 1.0058 \times 10^{-7} \end{aligned}$$

3.

$$\begin{aligned} p_{0.75} = \Phi(Z \leq \frac{k - 99.9}{.73/\sqrt{40}}) &= 0.75 \\ \Rightarrow \ Z &\leq 1.2816 \\ \Rightarrow \ k = \frac{1.2816 \times .73}{\sqrt{40}} &+ 99.9 \\ &= 100.0479 \end{aligned}$$

4. left cutoff is 25^{th} percentile of the sampling distribution, which is 99.8221. And the right is the 75^{th} 99.9779.


3
2
1
0
99.82
99.98
cutoff
mid−50%

■