

PubH 7401: Elements of Biostatistical Inference I
Homework 3

1. Let X be a random variable denoting out-of-pocket health care expenditures in a year for a randomly selected Minnesotan. Define X as follows

$$X = \begin{cases} 1 & \text{if \$0 spent} \\ 2 & \text{if } \$0 < \text{spent} \leq \$100 \\ 3 & \text{if } \$100 < \text{spent} \leq \$500 \\ 4 & \text{if over \$500 spent} \end{cases}$$

X is of course discrete. In fact, X is said to be a *categorical* variable with *ordinal* outcomes. The pmf of X is

$$P(X = 1) = 0.1, P(X = 2) = 0.2, P(X = 3) = 0.45, P(X = 4) = 0.25.$$

- (a) What is the probability of not spending anything?

Answer: $P(X = 1) = 0.1$

- (b) What is the probability of spending more than \$100?

Answer: $P(X \geq 3) = P(X = 3) + P(X = 4) = 0.45 + 0.25 = 0.7.$

- (c) What is median of X ? What is the mode of X ?

The mode is 3 because $p(3) > p(j)$ for $j = 1, 2, 4$. The median is also 3 because $P(X \geq 3) = 0.7 \geq 0.5$ and $P(X \leq 3) = 0.75 \geq 0.5$.

2. Diabetes is a growing concern among Inuit, who as a people have gradually shifted to a diet rich in carbohydrates from their traditional diet of seafood. A small town in Labrador (part of a Canadian province) has $n = 243$ Inuit living in it. Let X be the number of Inuit in this town with diabetes. In Labrador, the probability that an individual has diabetes is $p = 0.04$. Assume that individuals have diabetes independently of each other in this town.

- (a) What is the range of X ?

Answer: $R_X = \{0, 1, 2, \dots, 242, 243\}$

- (b) What distribution best describes X ?

Answer: $X \sim \text{binomial}(243, 0.04).$

- (c) What is $P(X \leq 2)$?

Answer:

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \binom{243}{0} 0.04^0 (0.96)^{243} + \binom{243}{1} 0.04^1 (0.96)^{242} + \binom{243}{2} 0.04^2 (0.96)^{241} \\ &= 0.96^{243} + (243)(0.04)(0.96^{242}) + (29403)(0.04^2)(0.96^{241}) \\ &= 0.00306. \end{aligned}$$

- (d) Use the Poisson approximation to estimate $P(X \leq 2)$. How does this compare to (c)?

Answer: $X \overset{\bullet}{\sim} \text{Poisson}(np) = \text{Poisson}(243 \times 0.04) = \text{Poisson}(9.72).$

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{e^{-9.72} 9.72^0}{0!} + \frac{e^{-9.72} 9.72^1}{1!} + \frac{e^{-9.72} 9.72^2}{2!} \\ &= 0.00348. \end{aligned}$$

This is a large relative difference $(0.00348 - 0.00306)/(0.00306) \approx 14\%$.

3. Consider screening a patient for HIV. Let D denote the event that a randomly selected individual from the population has the HIV virus and let S be the event that the HIV screening test comes up positive. The ELISA test has very high sensitivity, about 99.7%, or probabilistically $P(S|D) = 0.997$. The specificity is $P(S^C|D^C) = 0.985$. The prevalence of HIV in the general population is about 0.51%, yielding $P(D) = 0.0051$.

- (a) A randomly selected individual tests positive for HIV; what is the probability that the individual really has HIV? That is, find $P(D|S)$.

$$\begin{aligned} P(D|S) &= \frac{P(S|D)P(D)}{P(S|D)P(D) + P(S|D^C)P(D^C)} \\ &= \frac{(0.997)(0.0051)}{(0.997)(0.0051) + (1 - 0.985)(1 - 0.0051)} = 0.254. \end{aligned}$$

- (b) A randomly selected individual tests negative for HIV; what is the probability the individual has HIV? That is, find $P(D|S^C)$.

$$\begin{aligned} P(D|S^C) &= \frac{P(S^C|D)P(D)}{P(S^C|D)P(D) + P(S^C|D^C)P(D^C)} \\ &= \frac{(1 - 0.997)(0.0051)}{(1 - 0.997)(0.0051) + (0.985)(1 - 0.0051)} = 0.000016. \end{aligned}$$

4. Let X be the survival time in days (after diagnosis) of a randomly selected 60 year-old male diagnosed with small cell lung cancer and treated with a regimen of cisplatin followed by etoposide. The pdf of X is estimated to be $f(x) = 0.00125 \exp(-0.00125x)$ for $x \geq 0$.

- (a) What distribution is this?

Answer: $X \sim \exp(0.00125)$?

- (b) What is the probability that a randomly selected individual will live less than 30 days?

Answer:

$$P(X < 30) = F(30) = 1 - e^{-(30)(0.00125)} = 0.037.$$

- (c) What is the probability that a randomly selected individual will live past two years?

Answer:

$$P(X > 730) = 1 - P(X < 730) = 1 - [1 - e^{-(730)(0.00125)}] = 0.402.$$

- (d) What is the median survival time after diagnosis?

Answer: Recall that the p^{th} quantile is given by $x_p = F^{-1}(p)$. In the notes we derived $F^{-1}(p) = -\log(1 - p)/\lambda$. So $x_{0.5} = F^{-1}(0.5) = -\log(0.5)/0.00125 = 554.5$ days, or about 1.5 years.

- (e) What survival time do 90% of the population being studied live less than?

Answer: This number is $x_{0.9}$, also called the '90th percentile,' and solves $P(X \leq x_{0.9}) = 0.9$. This is given by $x_{0.9} = F^{-1}(0.9) = -\log(0.1)/0.00125 = 1842.1$ days, or about 5 years.

5. Chachugi, a particularly good hunter from a Paraguayan Ache' tribe, is going on a three-day armadillo hunting trek. Hunting success contributes to a tribe member's status within the tribe; the number of armadillos killed over time is well-modeled as a Poisson random variable. Say the Chachugi's rate of killing armadillos is $\lambda = 3.5$ armadillos per three days. Let X be the number of armadillos killed by Chachugi over the three day hunting trek.

(a) What is the probability that Chachugi kills no armadillos over the three days?

Answer: $P(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-3.5} = 0.030$.

(b) What is the median number of armadillos killed over all treks?

Answer: Verify that

$$\sum_{j=0}^3 \frac{e^{-\lambda}\lambda^j}{j!} = p(0) + p(1) + p(2) + p(3) = 0.537,$$

and

$$\sum_{j=3}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} = 1 - [p(0) + p(1) + p(2)] = 0.679,$$

so $P(X \leq 3) \geq 0.5$ and $P(X \geq 3) \geq 0.5$. Therefore the median number of armadillos killed is 3.

6. The probability of a randomly selected pig being infected with toxoplasmosis is modeled as $X \sim \text{beta}(3, 2)$, for a particular large herd.

(a) Write down the pdf $f(x)$ of this random variable X . Use the fact that for an integer n , $\Gamma(n) = (n - 1)! = (n - 1)(n - 2) \cdots (3)(2)(1)$.

Answer:

$$\begin{aligned} f(x) &= \frac{\Gamma(3 + 2)}{\Gamma(3)\Gamma(2)} x^{3-1}(1 - x)^{2-1} \\ &= \frac{4!}{2!1!} x^2(1 - x) \\ &= 12x^2(1 - x). \end{aligned}$$

(b) Show that the pdf in (a) integrates to one over $R = [0, 1]$.

Answer:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_0^1 12x^2(1 - x)dx \\ &= 12 \int_0^1 [x^2 - x^3]dx \\ &= 12 \int_0^1 \left[\frac{1}{3}x^3 - \frac{1}{4}x^4 \right]' dx \\ &= 12 \left[\frac{1}{3}x^3 - \frac{1}{4}x^4 \right]_0^1 \\ &= 12 \left\{ \left[\frac{1}{3}1^3 - \frac{1}{4}1^4 \right] - \left[\frac{1}{3}0^3 - \frac{1}{4}0^4 \right] \right\} \\ &= 12 \left\{ \left[\frac{1}{3} - \frac{1}{4} \right] - [0 - 0] \right\} \\ &= \frac{12}{12} = 1. \end{aligned}$$

(c) Find the cdf $F(x) = P(X \leq x)$.

Answer: Since the pdf $f(x)$ is zero when $x < 0$, we know $F(x) = 0$ for all $x < 0$.

When $0 \leq x \leq 1$,

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= P(X \leq 0) + P(0 \leq x) \\ &= 0 + \int_0^x 12s^2(1-s)ds \\ &= \dots\text{similar to above...} \\ &= 12 \left\{ \left[\frac{1}{3}x^3 - \frac{1}{4}x^4 \right] - \left[\frac{1}{3}0^3 - \frac{1}{4}0^4 \right] \right\} \\ &= 12 \left[\frac{1}{3}x^3 - \frac{1}{4}x^4 \right]. \end{aligned}$$

And finally, of course when $x > 1$ we have $F(x) = 1$. So

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 12 \left[\frac{1}{3}x^3 - \frac{1}{4}x^4 \right] & 0 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}.$$

(d) Find the mode of X .

Answer: $f'(x) = 24x - 36x^2$. Setting this equal to zero and solving for x gives a mode of $\frac{2}{3}$.