Problem set 2: More probability

Due October 14, 2024, at 10am

(Your name here)

NOTE: Start with the file `ps2_2024_more_probability.qmd` (available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F24/tree/main/assignments). Modify that file to include your answers. Make sure you can "render" the file (e.g. in RStudio by clicking on the `Render` button). Submit both the qmd file and the PDF via Canvas.

Problem 1: Bayes' Rule

Bayes Rule expresses the relationship between a conditional probability, e.g. $P(A \mid B)$, and the "reverse" conditional probability $P(B \mid A)$.

One formulation of Bayes' Rule states that, if $\{A_1, A_2, \dots A_n\}$ is a partition of Ω with $P(A_i) > 0$ for $i = 1, 2, \dots, n$ and $B \in S$ with $P(B) > 0$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

$\forall i$.

Here is a proof of Bayes Rule that is missing explanations for the steps:

Step 1:

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

Step 2:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}$$

Step 3:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

Explain each step in the proof: what definition(s)/rule(s)/law(s)/axiom(s)/condition(s)/mathematical operation(s) is the proof relying on?

Answer:

Problem 2: Error rates in hypothesis testing

You have a fancy device that tests null hypotheses. Null hypotheses are statements about the world that can be either true or false. The device is designed to turn red when a null hypothesis is false and green when it is true, but it doesn't work perfectly: when a null hypothesis is false it turns red with probability $3/4$ (i.e. it mistakenly turns green with probability $1/4$), and when a null hypothesis is true it turns green with probability $19/20$ (i.e. it mistakenly turns red with probability $1/20$). Tests of different null hypotheses are independent, and $4/5$ of the null hypotheses you test are true.

(1a) If you test 12 true null hypotheses in a row, what is the probability that the alarm turns red at least once? Explain your solution with reference to any axioms/definitions/rules/laws of probability you use.

Answer:

(1b) Write a simulation to check your answer to (1a). That is, use R to generate many draws according to the random process described (each time testing twelve true null hypotheses in a row), and confirm that the proportion of draws with at least one red light is approximately the same as in your answer above.

```
# your code here
```

(1c) Now suppose you come upon a null hypothesis at random; you don't know if it is true or false in this case. What is the probability of getting a red light when you run your test? Explain your solution with reference to any axioms/definitions/rules/laws of probability you use.

Answer:

(1d) If the light turns red in a given test, what is the probability that the null hypothesis is true? Explain your solution with reference to any axioms/definitions/rules/laws of probability you use.

Answer:

(1e) Write a simulation to check your answer to (1c) and (1d). That is, use R to generate many draws according to the random process described (testing null hypotheses), and confirm your answer about the proportion of red-light-producing draws (1c) and the proportion of red-light-producing draws in which the null hypothesis is actually true (1d).

Answer:

(1f) Using a similar calculation, John Ioannides reported in a famous 2005 paper that "most published research findings are false". By this he meant that the probability of the null hypothesis being true, given that the researcher rejected the null hypothesis (i.e. the light turns red), is above $1/2$. Show one way to change the assumptions in the problem to produce that result.

Answer:

Problem 3: discrete random variables

Suppose I am sampling a voter at random from a population. I care about two characteristics of the voter, which I characterize using numbers: whether she supports the populist candidate ($X = 1$ if so, 0 otherwise) and her level of education ($Y = 0$ if less than college degree, $Y = 1$ if college degree, $Y = 2$ if more than a college degree).

You happen to know the joint distribution of these characteristics in the population from which I am sampling. The resulting joint PMF for my random variables is:

$$f_{X,Y}(x,y) = \begin{cases} 1/12 & x = 0, y = 0 \\ 1/6 & x = 1, y = 0 \\ 1/4 & x = 0, y = 1 \\ 1/4 & x = 1, y = 1 \\ 3/16 & x = 0, y = 2 \\ 1/16 & x = 1, y = 2 \\ 0 & \text{otherwise} \end{cases}$$

(3a) What is the marginal PMF of X, $f_X(x)$? Replace the question marks with your answers.

$$f_X(x) = \begin{cases} ? & x = 0 \\ ? & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

3

(3b) What is the marginal PMF of Y, $f_Y(y)$?

$$f_Y(y) = \begin{cases} ? & y = 0 \\ ? & y = 1 \\ ? & y = 2 \\ 0 & \text{otherwise} \end{cases}$$

(3c) What is the conditional PMF $f_{X|Y}(x|y)$?

$$f_{X|Y}(x|y) = \begin{cases} ? & x = 0, y = 0 \\ ? & x = 1, y = 0 \\ ? & x = 0, y = 1 \\ ? & x = 1, y = 1 \\ ? & x = 0, y = 2 \\ ? & x = 1, y = 2 \\ 0 & \text{otherwise} \end{cases}$$

(3d) What is the conditional PMF $f_{Y|X}(y|x)$?

$$f_{X|Y}(x|y) = \begin{cases} ? & x = 0, y = 0 \\ ? & x = 1, y = 0 \\ ? & x = 0, y = 1 \\ ? & x = 1, y = 1 \\ ? & x = 0, y = 2 \\ ? & x = 1, y = 2 \\ 0 & \text{otherwise} \end{cases}$$

Problem 4: continuous random variables

(4a) Let X be uniformly distributed between -5 and 3. Compute $\Pr[X < 2]$ and $\Pr[-4 < X < -1/2]$ analytically (e.g. by computing the length and height of the area to be integrated) and confirm your results using a simulation in R.

Answer:

Confirmation in R:

(4b) Let X be normally distributed with mean 2 and standard deviation 1.25. Using R, compute (i) $\Pr[X < 0]$, (ii) $\Pr[1 < X < 3]$, and (iii) $\Pr[X > 3.5]$

Answer:

(4c) As we discussed in class, if X is a continuous random variable, $f(x)$ can be used to approximate the probability of getting a value near x. Suppose X is normally distributed with mean 3 and standard deviation 1.5. Use $f(x)$ (`dnorm()`) to approximate the probability of obtaining a value within .01 of 2. Then use $F(x)$ (`pnorm()`) to obtain the exact value. Finally, draw a large number of random samples using `rnorm()` to obtain a numerical estimate of the same value.

Answer: