Noah Arthurs
CS109

Section #3
July 19, 2019

Section #3 Solutions

1. **Website Visits**: On average, visitors leave your website after 5 minutes. Assume that the length of stay is exponentially distributed. What is the probability that a user stays more than 10 minutes?

Let X be the number of minutes that a user stays. $X \sim \text{Exp}(\lambda = \frac{1}{5})$.

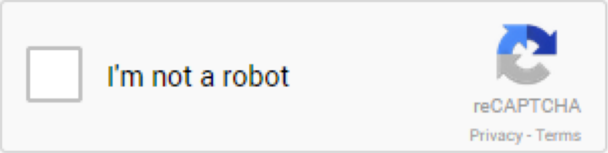
$$\begin{aligned} P(X > 10) &= 1 - F_X(10) \\ &= 1 - (1 - e^{\lambda 10}) = e^{-2} \approx 0.1353 \end{aligned}$$

2. **Approximating Normal**: Your website has 100 users and each day each user independently has a 20% chance of logging into your website. Use a normal approximation to estimate the probability that more than 21 users log in.

The number of users that log in B is binomial: $B \sim \text{Bin}(n = 100, p = 0.2)$. It can be approximated with a normal that matches the mean and variance. Let C be the normal that approximates B. We have $E[B] = np = 20$ and $Var(B) = np(1 - p) = 16$, so $C \sim N(\mu = 20, \sigma^2 = 16)$. Note that because we are approximating a discrete value with a continuous random variable, we need to use the continuity correction:

$$\begin{aligned} P(B > 21) &\approx P(C > 21.5) \\ &= P\Big(\frac{C - 20}{\sqrt{16}} > \frac{21.5 - 20}{\sqrt{16}}\Big) \\ &= P(Z > 0.375) \\ &= 1 - P(Z < 0.375) \\ &= 1 - \phi(0.375) = 1 - 0.6462 = 0.3538 \end{aligned}$$

3. **ReCaptcha**: Based on browser history, Google believes that there is a 0.2 probability that a particular visitor to a website is a robot. They decide to give the visitor a recaptcha:



Google presents the visitor with a box, 10mm by 10mm. The visitor must click inside the box to show that they are not a robot. You have observed that robots click uniformly in the box. However, the distance location of a human click has X location (mm from the left) and the Y location (mm from the right) distributed as independent normals both with mean $\mu = 5$ and $\sigma^2 = 4$:

a. What the the probability density function of a robot clicking $X = x$ mm from the left of the box and $Y = y$ mm from the top of the box?

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{100} & \text{if } 0 < x, y < 10 \\ 0 & \text{else} \end{cases}$$

b. What the the probability density function of a human clicking $X = x$ mm from the left of the box and $Y = y$ mm from the top of the box?

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x) f_Y(y) && \text{independence} \\ &= \frac{1}{(2\sqrt{2\pi})^2} e^{-\frac{(x-5)^2}{8}} e^{-\frac{(y-5)^2}{8}} && \text{normal PDF} \\ &= \frac{1}{8\pi} e^{-\frac{(x-5)^2}{8}} e^{-\frac{(y-5)^2}{8}} && \text{normal PDF} \end{aligned}$$

c. The visitor clicks in the box at ($x = 6$ mm, $y = 6$ mm). What is Google's new belief that the visitor is a robot?

Let Click to be the event that the user clicked at location $X = 6, Y = 6$. We can then use Bayes Rule (with law of total probability in the denominator):

$$\begin{aligned} P(\text{Robot}|\text{Click}) &= \frac{f(\text{Click}|\text{Robot})P(\text{Robot})}{f(\text{Click})} \\ &= \frac{f(\text{Click}|\text{Robot})P(\text{Robot})}{f(\text{Click}|\text{Robot})P(\text{Robot}) + f(\text{Click}|\text{Human})P(\text{Human})} \\ &= \frac{\frac{1}{100} \cdot 0.2}{\frac{1}{100} \cdot 0.2 + \frac{1}{8\pi} e^{-\frac{(1)^2}{8}} e^{-\frac{(1)^2}{8}} \cdot 0.8} \approx 0.075 \end{aligned}$$

4. **It's Complicated**

This probability table shows the joint distribution between two random variables: the year of the student at Stanford (Y) and their relationship status (R). The data was volunteered last year by over 200 anonymous students:

	Single	**In a Relationship**	**It's Complicated**
Freshman	0.12	0.07	0.02
Sophomore	0.17	0.12	0.02
Junior	0.10	0.11	0.02
Senior	0.01	0.07	0.00
5+	0.04	0.10	0.03

a. What is the marginal probability distribution for relationship status at Stanford (R)? Provide your result as a mapping between the values that R can take on and the corresponding probabilities.

For each assignment to R, sum over all the values that S can take on that are consistent with that assignment.

$$\begin{aligned} P(\text{Single}) &= 0.12 + 0.17 + 0.10 + 0.01 + 0.04 = 0.44 \\ P(\text{Relationship}) &= 0.07 + 0.12 + 0.11 + 0.07 + 0.10 = 0.47 \\ P(\text{Complicated}) &= 0.02 + 0.02 + 0.02 + 0.00 + 0.03 = 0.09 \end{aligned}$$

b. What is the conditional probability of relationship status (R) given that a student is a Senior (Y = Senior)? Provide your result as a mapping between the values that R can take on and the corresponding probabilities.

$$P(\text{Single}|\text{Senior}) = \frac{P(\text{Single, Senior})}{P(\text{Senior})} = \frac{0.01}{0.08} = 0.125$$

$P(\text{Relationship}|\text{Senior}) = 0.88$ and $P(\text{Complicated}|\text{Senior}) = 0$ (same approach).

c. What is the conditional probability that someone is "In a Relationship" given their year in school, $P(R = \text{In a Relationship}|Y)$? Give your answer as a mapping between the values that Y can take on and the corresponding probabilities.

$$P(\text{Relationship}|\text{Freshman}) = \frac{P(\text{Relationship, Freshman})}{P(\text{Freshman})} = \frac{0.07}{0.21} = 0.33$$

Same approach yields $P(\text{Relationship}|\text{Sophomore}) = 0.39$,
$P(\text{Relationship}|\text{Junior}) = 0.48$, $P(\text{Relationship}|\text{Senior}) = 0.875$, and
$P(\text{Relationship}|5+) = 0.59$

5. Student Heights: Adult heights can be considered to be normally distributed, but the distributions are different between men and women.

a. Adult women have a mean height of 65 inches and a standard deviation of 3.5 inches. What is the probability that a randomly selected adult woman is over 72 inches? What is the probability that a randomly selected woman is between 63 and 65 inches?

$$P(h > 72) = 1 - P(h < 72) = 1 - \Phi(\tfrac{72-65}{3.5}) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228$$

$$\begin{aligned} P(63 < h < 65) = P(h < 65) - P(h < 63) &= \Phi(\frac{65-65}{3.5}) - \Phi(\frac{65-63}{3.5}) \\ &= \Phi(0) - (1 - \Phi(0.57)) = 0.5 - (1 - 0.7157) = 0.216 \end{aligned}$$

b. Adult men are slightly taller, and the distribution of their heights has a slightly different spread. We know that the average adult man is 70 inches tall, and that 10 percent of adult men are under 65 inches tall. What is the standard deviation of adult men's heights, if they are normally distributed?

$$\Phi(\tfrac{x-\mu}{\sigma}) = \Phi(\tfrac{65-70}{\sigma}) = 0.1 \implies \tfrac{-5}{\sigma} = -1.29 \implies \sigma = 3.88.$$

c. Chris Piech is 6'5" (77 inches) tall. What is the probability that, of the six men in a typical section, at least one of them is taller than Chris? What is the probability that there another man is exactly the same height as Chris? What is the probability that another man is approximately his same height (i.e. within +/- half an inch)?

P(at least one man taller than Chris) = 1 – P(all men shorter)

P(any one man in the section is shorter = $\Phi(\frac{77-70}{3.88}) = \Phi(1.80) = 0.9641$

1 – P(all men shorter) = $1 - (0.9641)^6 = 0.803$.

P(height = 77") = 0 because we cannot talk about precise equalities when dealing with continuous distributions.

$P(76.5 < h < 77.5) = \Phi(\frac{77.5-70}{3.88}) - \Phi(\frac{76.5-70}{3.88}) = 0.020$

6. Elections: We would like to see how we could predict an election between two candidates in France (A and B), given data from 10 polls. For each of the 10 polls, we report below their sample size, how many people said they would vote for candidate A, and how many people said they would vote for candidate B. Not all polls are created equal, so for each poll we also report a value "weight" which represents how accurate we believe the poll was. The data for this problem can be found on the class website in polls.csv:

Poll	N samples	A votes	B votes	Weight
1	862	548	314	0.93
2	813	542	271	0.85
3	984	682	302	0.82
4	443	236	207	0.87
5	863	497	366	0.89
6	648	331	317	0.81
7	891	552	339	0.98
8	661	479	182	0.79
9	765	609	156	0.63
10	523	405	118	0.68
Totals:	**7453**	**4881**	**2572**	

a. First, assume that each sample in each poll is an independent experiment of whether or not a random person in France would vote for candidate A (disregard weights).

- Calculate the probability that a random person in France votes for candidate A.
- Assume each person votes for candidate A with the probability you've calculated and otherwise votes for candidate B. If the population of France is 64,888,792, what is the probability that candidate A gets more than half of the votes?

P(random person votes for A) = $\frac{votesforA}{totalvotes} = \frac{4881}{7453} = 0.655$

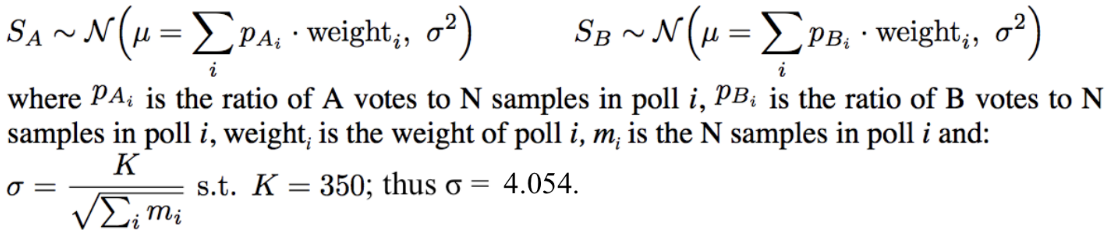
Now, let X be the number of votes for candidate A. We assume that X $\sim Bin(64888792, 0.655)$.

- Since n is so large, we can approximate X using a normal Y $\sim N(np, np(1-p))$.

- $\mu = np = 42502158.76$, Variance $= np(1 - p) = 14663244.77$ Std Dev $= 3829.26$
- Votes to win $= \frac{64888792}{2} = 32444396$
- $P(\text{A gets enough votes}) = P(X > 32444396) \approx P(Y > 32444396.5) = 1.00$

b. Nate Silvers at fivethirtyeight pioneered an approach called the "Poll of Polls" to predict elections. For each candidate A or B, we have a random variable S_A or S_B which represents their strength on election night (like ELO scores). The probability that A wins is $P(S_A > S_B)$.

- Identify the parameters for the random variables S_A and S_B. Both S_A and S_B are defined to be normal with the following parameters:



$$S_A \sim \mathcal{N}\Big(\mu = \sum_i p_{A_i} \cdot \text{weight}_i,\ \sigma^2\Big) \qquad S_B \sim \mathcal{N}\Big(\mu = \sum_i p_{B_i} \cdot \text{weight}_i,\ \sigma^2\Big)$$

where p_{A_i} is the ratio of A votes to N samples in poll i, p_{B_i} is the ratio of B votes to N samples in poll i, weight$_i$ is the weight of poll i, m_i is the N samples in poll i and:

$\sigma = \frac{K}{\sqrt{\sum_i m_i}}$ s.t. $K = 350$; thus σ = 4.054.

S_A ~$N(5.324, 16.436)$
S_B ~$N(2.926, 16.436)$

$P(S_A > S_B) \approx 0.66$

We can figure this out through simulation by drawing from S_A and S_B 100,000 times and seeing how often the S_A value is greater than the S_B value. Later in the quarter, when we learn the convolution of independent normals, you will be able to figure this out mathematically.

- We will calculate $P(S_A > S_B)$ by simulating 100,000 fake elections. In each fake election, we draw a random sample for the strength of A from S_A and a random sample for the strength of B from S_B. If S_A is greater than S_B, candidate A wins. What do we expect to see if we simulate so many times? What do we actually see?

c. Which model, the one from (a) or the model from (b) seems more appropriate? Why might that be the case? On election night candidate A wins. Was your prediction from part (b) "correct"?

Algorithm (a) makes very few assumptions, and simplicity can be useful, but it does assume that each voter is independent - which we definitely know isn't the case in real elections. Algorithm (b) allows us to model bias (using the weights we incorporated), and doesn't think of each voter as necessarily independent.