Chris Piech
CS109

Handout #31
May 9th, 2016

Midterm Solution

This handout goes over the solutions to the midterm. Overall the class performed above expectation – but still (almost) everyone walked out with room for improvement. By the final you will want to make sure you are comfortable with all the problems on the midterm.

1 Counting Cards

a

This is a permutation with indistinguishable elements question. There are 208 cards. Each value (of which there are 52) is repeated four times, because it shows up in each of four decks:

$$\text{Answer} = \frac{208!}{(4!)^{52}}$$

b

Think of this as a two step experiment. The probability of getting two good cards is the probability that the first card you get is a good card (G_1) times the probability that the second card you get is a good card (G_2):

$$\begin{aligned}\text{Answer} &= P(G_1) \cdot P(G_2) \\ &= \frac{5 \cdot 4 \cdot 4}{208} \cdot \frac{5 \cdot 4 \cdot 4 - 1}{207} \\ &= 0.147\end{aligned}$$

c

This is just like the previous question, except that the number of total cards in the deck is 100 less and the number of "good" cards in the deck is 15 less.

$$\begin{aligned}\text{Answer} &= P(G_1) \cdot P(G_2) \\ &= \frac{5 \cdot 4 \cdot 4 - 15}{108} \cdot \frac{5 \cdot 4 \cdot 4 - 16}{107} \\ &= 0.360\end{aligned}$$

d

Let X_i be a random variable which is the number of steps until the ith success. A success is the event that in a "step", the top card gets placed below the ace of spades.

The probability of a particular success depends on i. At first, when $i = 1$, the probability of a success is $1/52$ because only one position is below the ace of spades. When $i = 2$ the probability of a success is $2/52$ since there are now two positions below the ace of spades. The probability of a success in general is $i/52$.

The number of steps until a success with known probability can be modeled as a Geometric Random Variable. Thus:

$$X_i \sim Geo(i/52).$$

Let X be the total number of steps until 51 successes, plus the single extra step to shuffle the ace of spades:

$$X = \left(\sum_{i=1}^{51} X_i\right) + 1$$
$$E[X] = E[\sum_{i=1}^{51} X_i] + 1 = \left(\sum_{i=1}^{51} E[X_i]\right) + 1$$
$$= \left(\sum_{i=1}^{51} \frac{52}{i}\right) + 1$$

The last term can be reworked into a Harmonic series. Evaluating it would be nice, but unnecessary.

2 Sampling

a

This is the combination of three steps. You must have the median in your sample. Then you need five numbers from below the median (of which there are 50 numbers) and five numbers from above the median (of which there are again 50 numbers).

$$\text{Answer } = \binom{50}{5} \cdot \binom{50}{5}$$

b

$$\text{Answer } = \frac{\binom{50}{5} \cdot \binom{50}{5}}{\binom{101}{11}}$$

c

Let X be the number of samples with the same median. $X \sim Bin(100, p)$

$$P(X < 10) = \sum_{i=0}^{9} P(X = i)$$
$$= \sum_{i=0}^{9} \binom{100}{i} p^i (1-p)^{100-i}$$

Given that the value of n is large and p is very small, it would be reasonable to approximate the count using a Poisson random variable $A \sim Poi(100p)$.

$$P(X < 10) \approx P(A < 10) = \sum_{i=0}^{9} P(X = i) = \sum_{i=0}^{9} \frac{(100p)^i}{i!} e^{-(100p)}$$

3 Netflix

a

$$\text{Answer } = p_1 \cdot p_2 \cdot p_3$$

b

Let $L_i|G$ be the event that they like movie T_i given that they like the genre.

$$\begin{aligned}\text{Answer} &= P(L_1|G \cup L_2|G \cup L_3|G) \\ &= 1 - P(L_1|G \cup L_2|G \cup L_3|G)^C \\ &= 1 - P(L_1^C|G \cdot L_2^C|G \cdot L_3^C|G) \\ &= 1 - (1-p_1)(1-p_2)(1-p_3)\end{aligned}$$

Another approach is to use the inclusion/exclusion principle to expand the first line of the previous answer:

$$\begin{aligned}\text{Answer} &= P(L_1|G \cap L_2|G \cap L_3|G) \\ &= p_1 + p_2 + p_3 - (p_1p_2) - (p_1p_3) - (p_2p_3) + (p_1p_2p_3)\end{aligned}$$

c

Solve this question using Bayes theorem:

$$\begin{aligned}P(G|L_1L_2L_3) &= \frac{P(L_1L_2L_3|G)P(G)}{P(L_1L_2L_3|G)P(G) + P(L_1L_2L_3|G^C)P(G^C)} \\ &= \frac{(p_1p_2p_3)(0.6)}{(p_1p_2p_3)(0.6) + (q_1q_2q_3)(0.4)}\end{aligned}$$

4 Wind Power

a

$$\begin{aligned}P(X > 4) &= 1 - P(X < 4) \\ &= 1 - F_X(4) \\ &= 1 - \phi(\frac{4-2}{\sqrt{64}}) \\ &= 1 - \phi(\frac{1}{4}) \\ &= 1 - 0.5987 \\ &= 0.4013\end{aligned}$$

b

Approximate the binomial with a Normal that matches the mean and variance of the binomial.

$$\begin{aligned}\mu &= np = 40,000 \cdot \frac{1}{2} = 20,000 \\ \sigma^2 &= np(1-p) = 40,000 \cdot \frac{1}{2} \cdot \frac{1}{2} = 10,000 \\ \sigma &= \sqrt{(10,000)} = 100\end{aligned}$$

When approximating normally we use the continuity correction. In this case it doesn't matter because the numbers are so big. Let Y be the number of people using electricity. $Y \sim N(\mu = 20,000, \sigma^2 = 10,000)$

$$\begin{aligned}
P(Y > 20,300) &= 1 - P(Y < 20,300) \\
&= 1 - F_Y(20,300) \\
&= 1 - \phi(\frac{20,300 - 20,000}{\sqrt{10,000}}) \\
&= 1 - \phi(\frac{300}{100}) \\
&= 1 - 0.9987 \\
&= 0.0013
\end{aligned}$$

c

Let Z be the amount of wind produced by the two wind farms. It is easier to incorporate the fact that wind may or may not blow at each wind-farm if you break the probability of wind being produced into mutually exclusive cases of wind in both farms, wind in farm-1, wind in farm-2 and no wind. Let E_1 be the event that wind is blowing at farm-1 and let E_2 at farm-2.

$$\begin{aligned}
P(Z < \theta) =& P(Z < \theta, E_1E_2) + P(Z < \theta, E_1E_2^C) + P(Z < \theta, E_1^C E_2) + P(Z < \theta, E_1^C, E_2^C) \\
=& P(Z < \theta | E_1E_2)P(E_1E_2) \\
&+ P(Z < \theta | E_1E_2^C)P(E_1E_2^C) \\
&+ P(Z < \theta | E_1^C E_2)P(E_1^C E_2) \\
&+ P(Z < \theta | E_1^C E_2^C)P(E_1^C E_2^C) \\
=& p_1p_2[F_{W_1+W_2}(\theta)] \\
&+ p_1(1-p_2)[F_{W_1}(\theta)] \\
&+ (1-p_1)(p_2)[F_{W_2}(\theta)] \\
&+ (1-p_1)(1-p_2) \\
=& p_1p_2\phi(\frac{\theta - \mu_1 - \mu_2}{\sigma_1^2 + \sigma_2^2}) \\
&+ p_1(1-p_2)\phi(\frac{\theta - \mu_1}{\sigma_1^2}) \\
&+ (1-p_1)(p_2)\phi(\frac{\theta - \mu_2}{\sigma_2^2})) \\
&+ (1-p_1)(1-p_2)
\end{aligned}$$

Germany tried to turn their grid 100% into wind. As they learned, it is more important to chose locations where wind is generated independently then it is to chose a wind farm location where wind generates a lot of electricity.

5 Autonomous Car

This problem required a few steps. First use Bayes theorem to rewrite the probability density function. Then recognize that two of the three terms are constant. Finally plug in the density functions for the independent

instrument readings given the true direction.

$$\begin{aligned}
f(T=t|D-1=57,D_2=59) &= \frac{f(D_1=57,D_2=59|T)f(T)}{f(D_1=57,D_2=59)} \\
&= Kf(D_1=57,D_2=59|T=t) \\
&= Kf(D_1=57|T=t)\cdot f(D_2=59|T=t) \\
&= K\cdot\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(57-t)^2}{2}}\right)\left(\frac{1}{2\sqrt{2\pi}}e^{-\frac{(57-t)^2}{2\cdot 4}}\right) \\
&= K\cdot e^{-\frac{(57-t)^2}{2}}\cdot e^{-\frac{(57-t)^2}{8}}
\end{aligned}$$

There are a few ways to figure out $f(D_1=57|T=t)\cdot f(D_2=59|T=t)$. The one used above requires recognizing that adding a constant to a normal random variable produces a new normal with a shifted mean. The direction instruments are conditionally independent of one another given the true direction since all of their randomness comes from the independent noise terms. The other approach is to plug in the equation for the D variables and solve for the noise term X. The noise terms are defined as independent.

6 Uber Pool

a

Part (a) is mysteriously similar to a question on Problem Set 3, so its solution has been redacted. In case you are curious and want to check your math, the evaluated answer is approximately 0.865.

b

Part (b) is also mysteriously similar to a question on Problem Set 3, so its solution has also been redacted. In case you are curious and want to check your math, the evaluated answer is \$7.75.

c

Let Y be the number of minutes until a user requests. You could either model Y as a geometric random variable or as an exponential.

We have redacted the exponential approach, because it is mysteriously similar to a question on Problem Set 3.

If you model Y as a geometric distribution, you must first calculate the probability that you get a request in one minute. Let this probability be p. Thus $Y \sim \text{Geo}(p)$.

$$\begin{aligned}
P(Y>8) &= 1-P(Y\leq 8) \\
&= 1-\sum_{i=1}^{8} P(Y=i) \\
&= 1-\sum_{i=1}^{8} [1-p]^{i-1}\, p
\end{aligned}$$

We've redacted our calculation for p because it has a similar approach to part (a). In case you are curious and want to check your math, the evaluated answer for both approaches is $e^{-16/5} \approx 0.041$.

c

Let Y be the number of minutes until a user requests. You could either model Y as a geometric random variable or as an exponential.

If you model Y as an exponential, first calculate the rate per minute, which is $\frac{2}{5}$. Thus $Y \sim Exp(\lambda = \frac{2}{5})$.

$$\begin{aligned} P(Y > 8) &= 1 - P(Y < 8) \\ &= 1 - F_Y(8) \\ &= 1 - (1 - e^{-\frac{2 \cdot 8}{5}}) \\ &= e^{-\frac{16}{5}} \end{aligned}$$

If you model Y as a geometric distribution, you first have to calculate the probability that you get a request in one minute. The probability of a "success" in one minute is the probability that we worked out in part a: $1 - e^{-2/5}$. Thus $Y \sim Geo(1 - e^{-2/5})$:

$$\begin{aligned} P(Y > 8) &= 1 - P(Y < 8) \\ &= 1 - \sum_{i=1}^{8} P(Y = i) \\ &= 1 - \sum_{i=1}^{8} \left[1 - (1 - e^{-2/5})\right]^{i-1} (1 - e^{-2/5}) \\ &= 1 - \sum_{i=1}^{8} \left[e^{-2/5}\right]^{i-1} (1 - e^{-2/5}) \end{aligned}$$

This also works out to be $e^{-\frac{16}{5}}$. Just in case you are curious the evaluated answer is: 0.041