Lisa Yan
CS109

Section #4
May 6-8, 2020

Section #4

Based on the work of many prior CS109 staffs

1. **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.

2. **Fairness in AI.** In their 2018 paper "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Joy Buolamwini and Timnit Gebru showed that commercial gender classifiers performed significantly worse on face images of darker-skinned females (error rates up to 34.7%) than lighter-skinned males (maximum error rate of 0.8%.) This disparity may result from training the classifiers on unbalanced datasets. To evaluate the classifiers, the authors designed their own dataset by collecting photos of national parliamentarians from three African countries and three European ones.

   The probability table below shows the joint distribution of the dataset between two random variables: the demographic (D) of the photo subject and their country (C).

Demographic	**South Africa**	**Senegal**	**Rwanda**	**Sweden**	**Finland**	**Iceland**
Darker Female	0.12	0.05	0.04	0.01	0	0
Darker Male	0.15	0.07	0.02	0.01	0	0
Lighter Female	0.02	0	0	0.12	0.06	0.02
Lighter Male	0.05	0	0	0.14	0.09	0.03

   a. What is the marginal probability distribution for demographic D? Provide your result as a mapping from values that D can take to probabilities.

   b. What is the conditional probability of country given that the subject is a lighter female, $P(C|D = \text{Lighter Female})$? Provide your result as a mapping from values that C can take to probabilities. Is this mapping a probability distribution?

   c. What is the conditional probability that the subject is from Senegal given their demographic, $P(C = \text{Senegal}|D)$? Provide your answer as a mapping from values that D can take to probabilities. Is this mapping a probability distribution?

   d. What are the pitfalls in using this dataset for a purpose beyond what the authors intended?

3. **Hat-Check Again??** Recall the hat-check problem from section 2: n people go to a party and drop off their hats to a hat-check person. When the party is over, a different hat-check person is on duty, and returns the n hats randomly back to each person. Let X be the random variable representing the number of people who get their own hat back. We showed last time that $E[X] = 1$ for any n. What is $Var(X)$? Hint: Be careful when taking the variance of a sum of random variables.