Stat 305
Midterm 2

1. The CDC estimates that about 10% of Americans have diabetes. A fasting plasma glucose (FPG) screen can be used to test for the presence of diabetes. If a person has diabetes, the probability that the FPG test is positive is about 0.6. However, when a person does not have diabetes, there is still a probability that the FPG test returns a (false) positive of about 0.2.
   (a) **(4 points.)** Suppose a randomly selected American takes the FPG test. Compute the probability that the test is positive.
   Solution: Law of total probability (or two-way table): $\mathrm{P}(T) = \mathrm{P}(T|D)\mathrm{P}(D)+\mathrm{P}(T|D^c) = (0.6)(0.1) + (0.2)(0.9) = 0.24$.
   (b) **(4 points.)** Suppose a randomly selected American has a positive FPG test result. Compute the probability that the person has diabetes.
   Solution: Bayes rule (or two-way table): $\mathrm{P}(D|T) = \mathrm{P}(T|D)\mathrm{P}(D)/\mathrm{P}(T) = (0.6)(0.1)/(0.24) = 0.25$

2. Each of three basketball players attempts a three pointer. The probability that Amir makes the attempt is 0.3, that Barack makes the attempt is 0.4, and that Casey makes the attempt is 0.5. Assume the attempts are independent. Let X be the number of players that successfully make their attempt.
   (a) **(4 points.)** Compute $\mathrm{P}(X = 3|X \geq 1)$.
   Solution: $\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X = 0) = 1 - (1 - 0.3)(1 - 0.4)(1 - 0.5) = 0.79$. Note $\{X = 3, X \geq 1\} = \{X = 3\}$.

$$\mathrm{P}(X = 3|X \geq 1) = \frac{\mathrm{P}(X = 3, X \geq 1)}{\mathrm{P}(X \geq 1)} = \frac{\mathrm{P}(X = 3)}{\mathrm{P}(X \geq 1)} = \frac{(0.3)(0.4)(0.5)}{0.79} = 0.076$$

   (b) **(6 points.)** Describe **in detail** how you could (in principle) **perform by hand a simulation involving physical objects** (coins, dice, spinners, cards, boxes, etc) to estimate $\mathrm{E}(X|X \geq 1)$. Be sure you detail how you would set up and perform the simulation, what one repetition of the simulation entails, and how you would use the simulation results to estimate the quantity of interest. Note: you do NOT need to compute any numerical values.
   Solution:
   - Have a separate spinner for each player: Amir has 30% of its area labeled success, 40% for B, 50% for C
   - Spin each spin once and count the number of spins that land in the success area; this count is one realization of X.
   - If $X = 0$ discard this repetition; otherwise keep it and record X.
   - Repeat mamy times until obtaining say 10000 reps in which $X \geq 1$. The average of the X values for these reps approximates $\mathrm{E}(X|X \geq 1)$.

3. **(5 points each.)** Suppose that X and Y are random variables such that
   - The marginal distribution of X is Uniform(2, 5).
   - For any x, the conditional distribution of Y given $X = x$ is Uniform$(-x, x)$.

   (a) Explain how you could simulate a single (X, Y) pair using only a Uniform(0, 1) spinner.
   Solution: Spin the Uniform(0, 1) spinner once to get U_1. Let $X = 2 + (5 - 2)U_1$. Spin the Uniform(0, 1) spinner again to get U_2. Let $Y = -X + (2X)U_2$.

(b) For each of the following, draw an appropriate picture representing the distribution. The picture does not have to be exact, but it should explicitly illustrate the most important features. Clearly label relevant axes with appropriate values. You might also want to explain the most important features in words if they are not clear from your plot.

i. Joint distribution of (X, Y).
Solution: See Symbulate in Github for plots. The possible x values are (2, 5). The possible y values are $(-5, 5)$ but given x must have $-x < y < x$. The X values are Uniformly distributed over (2, 5). For any particular x, the Y values are Uniformly distributed over $(-x, x)$. The density of (x, y) pairs will be higher for x near 2 than for x near 5.

ii. Conditional distribution of X given $Y = 3$.
Solution: Slice the joint plot horizontally at $Y = 3$. The possible values of X given $Y = 3$ lie in (3, 5). Since the density of (x, y) pairs is higher for smaller x, the conditional density will be highest for x near 3 and decrease with x.

iii. Marginal distribution of Y.
Solution: Marginally, the possible values of Y lie in $(-5, 5)$. Collapse/stack the x values in the joint plot. All the stacks between $(-2, 2)$ will be the same, so the height of the marginal density will be constant for y in $(-2, 2)$ and then decrease from there.

(c) Compute $\mathrm{Cov}(X, Y)$ **without calculus**. Your final answer should be a number. Explain your reasoning fully in terms of probability facts we have learned in this class. (If you don't know how to compute it, for some partial credit explain if the covariance is positive, negative, or zero.)
Solution: Since the distribution of $(Y|X = x)$ is Uniform$(-x, x)$, $\mathrm{E}(Y|X = x) = 0$ so by LTE $\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y|X)) = \mathrm{E}(0) = 0$. So $\mathrm{Cov}(X, Y) = \mathrm{E}(XY)$. Using LTE and TOWIK

$$\mathrm{E}(XY) = \mathrm{E}(\mathrm{E}(XY|X)) = \mathrm{E}(X\mathrm{E}(Y|X)) = \mathrm{E}(X(0)) = 0$$

4. **(5 points each.)** Each of the following situations contains Symbulate commands. For each situation, provide an expression representing your single best estimate of the numerical value that the last line of code will produce (e.g. e^{-1}). Your answer should be a single number/expression (you don't have to worry about simulation margin of error), or "error". **Explain your reasoning.** If the code will produce an error, explain why.

(a)
```
P = BoxModel({1: 20, 0: 30}, size = 10, replace = False)
X = RV(P, sum)
X.sim(1000000).count_eq(6) / 1000000
```
Solution: An outcome in the probability space is a sequence of 10 success/failure (1/0) trials with probability of success 0.6 on each trial. The trials are not independent because the sampling is without replacement. The RV X sums the 1/0 values, so it counts the number of successes. Therefore, X has a Hypergeometric distribution with $n = 10, N_1 = 20, N_0 = 30$. The last line simulates many values of X and finds the relative frequency of 6, so it is estimating

$$\mathrm{P}(X = 6) = \frac{\binom{20}{6}\binom{30}{4}}{\binom{50}{10}}$$

(b)
```
X, Y = RV(Binomial(10, 0.6) * Exponential(1))
```

```
(X | (abs(Y - 0.5) < 0.01) ).sim(10000).var()
```

Solution: Here X and Y are independent because in the first line their joint distribution is defined as the product of marginal distributions. The last line is simulating the conditional distribution of X given $Y = 0.5$, which by independence will be the same as the marginal distribution of X which is Binomial(10, 0.6) and so its variance is $10(0.6)(1 - 0.6) = 2.4$

(c)
```
X, Y = RV(BivariateNormal(mean1 = 1, mean2 = 2, sd1 = 10, sd2 = 5, corr = -0.4))
Z = X + Y + 3
Z.sim(1000000).sd()
```

Solution: $\mathrm{Var}(X + Y + 3) = \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y) = 10^2 + 5^2 + 2(-0.4)(10)(5) = 85$; $\mathrm{SD}(Z) = \sqrt{85}$

(d)
```
X, Y = RV(Binomial(10, 0.6) * Binomial(5, 0.6))
(X + Y).sim(1000000).count_eq(7) / 1000000
```

Solution: Here X and Y are independent because in the first line their joint distribution is defined as the product of marginal distributions. Since the probability of success is the same, $X + Y$ has a Binomial(15, 0.6) distribution.

$$\mathrm{P}(X + Y = 7) = \binom{15}{7}(0.6)^7(0.4)^8$$

5. **(1 point each.)** In each of the following situations, which quantity — 1 or 2 — is **larger**, or are they exactly the same? Or is there not enough information? No explanation necessary.

(a) Suppose A and B are independent events.

1) $\mathrm{P}(A|B)$
2) $\mathrm{P}(A|B^c)$
3) (1) and (2) of this part are exactly the same
4) Not enough information to determine which of (1) or (2) is larger

Solution: (3). This is one of the equivalent conditions for independence.

(b) Randomly select an American and let A be the event that the person is a California resident and let B be the event that that the person is a Cal Poly graduate.

1) $\mathrm{P}(A|B)$
2) $\mathrm{P}(B|A)$
3) (1) and (2) of this part are exactly the same
4) Not enough information to determine which of (1) or (2) is larger

Solution: (2).

(c) A deck of 20 cards, of which some are red and the others are blue, is shuffled and 10 cards are dealt. Let X be the number of red cards among the 10 dealt.

1) $\mathrm{E}(X)$ if the deal is performed with replacement
2) $\mathrm{E}(X)$ if the deal is performed without replacement
3) (1) and (2) of this part are exactly the same
4) Not enough information to determine which of (1) or (2) is larger

Solution: (3).

(d) A deck of 20 cards, of which some are red and the others are blue, is shuffled and 10 cards are dealt. Let X be the number of red cards among the 10 dealt.

1) $\mathrm{Var}(X)$ if the deal is performed with replacement
2) $\mathrm{Var}(X)$ if the deal is performed without replacement
3) (1) and (2) of this part are exactly the same

4) Not enough information to determine which of (1) or (2) is larger

   Solution: (1). This is the idea of the finite population correction when sampling without replacement.

(e) Roll a fair die 48 times and let X be the number of rolls which land on 1

1) P($X > 18$) if the die is four-sided
2) P($X > 18$) if the die is six-sided
3) (1) and (2) of this part are exactly the same
4) Not enough information to determine which of (1) or (2) is larger

   Solution: (1) is larger. The mean in (2) is 8 while in (1) it is 12 and the variability is less in 2 since p is closer to 0.

(f) Random variables X, Y, and Z each have a Binomial(3, 0.5) distribution.

1) E($Y|X = 2$)
2) E($Z|X = 2$)
3) (1) and (2) of this part are exactly the same
4) Not enough information to determine which of (1) or (2) is larger

   Solution: (4). It's not necessarily true that the joint distribution of (X, Y) is the same as the joint distribution of (X, Z). See the related problem from HW 6.

(g) Shuffle a standard deck of 52 cards (13 hearts, 39 other cards) and deal 2.

1) The probability that the first card dealt is a heart
2) The probability that the second card dealt is a heart
3) (1) and (2) of this part are exactly the same
4) Not enough information to determine which of (1) or (2) is larger

Solution: (3) — (1) and (2) are both 13/52. This same question was on midterm 1, and we talked about the issue again with Hypergeometric distributions. We talked about this in class with the Harry problem. The *unconditional* probability that the second card is a heart, not knowing what the first card is, is 13/52 regardless if the cards are dealt with or without replacement. (If we knew the result of the first card then the conditional probability that the second card is a heart is 13/51 if the first card is not a heart or 11/51 if the first card is a heart. But this is NOT what (2) represents.)