DSE 210 (Probability and Statistics Using Python)
Homework 2, due Friday 2/5 9:00am

Instructions

Please follow these instructions when completing your assignment:

- Please upload your written answers to Gradescope by the due date. **Late submissions will not be graded**.
- You can write up your answers using pencil and paper or using document editing software (LaTeX, Word, etc...). If you write your answers using pencil and paper, you can scan your answers and upload the resulting file to Gradescope or take pictures of each page and upload those.
- Please associate each problem with a page on your gradescope submission
- For written answers you are not required to show work. However, showing work will enable better feedback.
- Collaboration is encouraged, but all submissions should be in your own writing and completed with your own understanding.

Problems

- **Worksheet 4**: $1, 2, 3, 8, 9, 11, 12, 18, 20$
- **Worksheet 5**: $1, 2, 4, 5, 8$

Worksheet 4 Solutions

1. Let $R_1, R_2 \in \{1, ..., 6\}$ be the outcomes of the first and second roll respectively and let $X = \min(R_1, R_2)$. Let's compute $\Pr(X = 1) = \Pr(R_1 = 1, R_2 \geq 1) + \Pr(R_1 \geq 1, R_2 = 1) - \Pr(R_1 = 1, R_2 = 1) = \frac{11}{36}$. Then we can compute the rest analogously:

x	1	2	3	4	5	6
$\Pr(x)$	11/36	9/36	7/36	5/36	3/36	1/36

   Let's also check $\sum_{x=1}^{6} \mathbb{P}(x) = 1$ which is true.

2. Let $X = 1$ if the roll was a six and zero otherwise. Then $\Pr(X = 1) = 1/6$ and so the expected number of rolls is $1/\Pr(X = 1) = 6$.

3. a) $\mathbb{E}[Z] = \sum_{x=1}^{6} x \cdot \Pr(X = x) = 4$
   b) $\texttt{Var}(Z) = \sum_{x=1}^{6} \Pr(X = x) \cdot (x - \mathbb{E}[Z])^2 = 3$
   c) $\binom{10}{5} \left(\frac{1}{4}\right)^5 \left(\frac{3}{4}\right)^5 = 0.058$
   d) Using the same logic as in problem (2), $\Pr(X = 6) = 1/4$ and so the expected number of rolls is $1/\Pr(X = 6) = 4$.
   e) The rolls are independent so we would expect to wait for more tosses before the second six - or a total of eight tosses for two sixes.

8. a) The standard deviation is the square root of the variance so: $\sqrt{\texttt{Var}(X)} = 2 \Rightarrow \texttt{Var}(X) = 4$.
   b) By linearity of expectation: $\mathbb{E}[10X] = 10\mathbb{E}[X] = 50$

c) For any real number a, $\texttt{Var}(aX) = a^2\texttt{Var}(X)$, therefore $\texttt{Std}(aZ) = \sqrt{\texttt{Var}(aZ)} = a\sqrt{\texttt{Var}(X)} = a\texttt{Std}(X)$. So $\texttt{Std}(10X) = 10\texttt{Std}(X) = 20$.

d) By the rule used above: $\texttt{Var}(10X) = 100\texttt{Var}(X) = 400$.

9. This is another biased coin flip problem in disguise:

a) Let $X_i^j = 1$ if the j-th person gets off on the i-th floor. Then $\Pr(X_i^j = 1) = \frac{1}{10}$ since each person picks one of the ten floors at random. So this is exactly the same as asking for the probability that exactly one out of n tosses of a coin with bias $1/10$ came up heads. We know we can compute this as:

$$\Pr\left(\sum_{j=1}^{n} X_i^j = 1\right) = \binom{n}{1}\left(\frac{1}{10}\right)^1\left(\frac{9}{10}\right)^{n-1} = \frac{n \cdot 9^{n-1}}{10^n}$$

b) Let $X_i = 1$ if exactly one person gets out on the i-th floor and zero otherwise and let $S = \sum_{i=1}^{10} X_i$. Then by linearity of expectation:

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=1}^{10} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Then by part (a):

$$\mathbb{E}[S] = 10 \cdot \frac{n \cdot 9^{n-1}}{10^n} = \frac{n \cdot 9^{n-1}}{10^{n-1}}.$$

11. a) Before we look at the number in the first position, any of the n numbers are equally likely to be in the second position, so $\Pr(Y = y) = \frac{1}{n}$. Now suppose we look at the first number, since each number appears in the sequence exactly once, there are $n-1$ numbers which could be in the second position and they are all equally likely, so $\Pr(Y = y|X = x) = \frac{1}{n-1}$. Therefore, the events are **dependent**.

b) Language is structured and words do not appear at random. We should expect these events to be **dependent**.

c) We can check all four cases: for example $\Pr(X = 1, Y = 1) = \frac{1}{52}$ since it is the singleton event of getting the nine of hearts. Similarly $\Pr(X = 1) = \frac{1}{13}$ and $\Pr(Y = 1) = \frac{1}{4}$ and so $\Pr(X = 1, Y = 1) = 1/52 = \Pr(X = 1)\Pr(Y = 1)$. We can verify the same property holds for $\Pr(X = 1, Y = 0)$, $\Pr(X = 0, Y = 1)$, and $\Pr(X = 0, Y = 0)$. Therefore, the events are **independent**.

d) Consider:

$$\Pr(X = 1|Y = 1) = 1 - \Pr(X = 0|Y = 1) = 1 - \left(\frac{12}{13} \cdot \frac{11}{12} \cdot \ldots \cdot \frac{3}{4}\right) = 0.769$$

But:

$$\Pr(X = 1) = 1 - \left(\frac{48}{52} \cdot \frac{47}{51} \cdot \ldots \cdot \frac{39}{43}\right) = 0.587$$

So the events are **dependent**.

12. a) Let $X_i = 1$ if the i-th accident occurred on a Sunday and zero otherwise. Let $p = 0.05$ be the probability an accident occurred on Sunday. So $\mathbb{E}[X_i] = p$ and $\texttt{Var}(X_i) = p(1-p) = 0.0475$. Then the total number of accidents occurring on Sunday can be modeled as:

$$S = \sum_{i=1}^{200} X_i$$

By linearity of expectation:

$$\mathbb{E}[S] = \sum_{i=1}^{200} \mathbb{E}[X_i] = 200 \cdot p = 10$$

where we have used the fact that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = ... = \mathbb{E}[X_{200}] = p$. We can then use the fact that the accidents are independent to write the variance as:

$$\texttt{Var}(S) = \sum_{i=1}^{200} \texttt{Var}(X_i) = 200p(1-p) = 9.5.$$

b) Recognizing that this problem can be modeled as flips of a biased coin, we are looking for the probability that exactly 10 flips of a coin with bias 0.05 came up heads. This can be computed as:

$$\Pr(S = 10) = \binom{200}{10} p^{10}(1-p)^{190} = 0.1284$$

18. Let's rewrite the table with variables instead of ??? to make it more intuitive:

		Y = 1	2	3
X	1	1/12	1/24	1/8
	2	p_1	p_2	p_3
	3	1/12	1/24	1/8

So we need to solve for p_1, p_2, p_3. To start, let's compute the marginal distributions $\Pr(X = x) = \sum_{y=1}^{3} \Pr(X = x, Y = y)$ and $\Pr(Y = y) = \sum_{x=1}^{3} \Pr(X = x, Y = y)$. So we fix one variable and sum over the others:

x	$\Pr(X = x)$
1	1/4
2	1/2
3	1/4

and

y	$\Pr(Y = y)$
1	$1/6 + p_1$
2	$1/12 + p_2$
3	$1/4 + p_3$

Since X and Y are independent, we know $\Pr(X, Y) = \Pr(X)\Pr(Y)$. So then:

$$\Pr(X = 1, Y = 1) = \Pr(X = 1)\Pr(Y = 1) = \frac{1}{4}(\frac{1}{6} + p_1) \Rightarrow p_1 = \frac{1}{6}$$
$$\Pr(X = 1, Y = 2) = \Pr(X = 1)\Pr(Y = 2) = \frac{1}{4}(\frac{1}{12} + p_2) \Rightarrow p_2 = \frac{1}{12}$$
$$\Pr(X = 1, Y = 3) = \Pr(X = 1)\Pr(Y = 3) = \frac{1}{4}(\frac{1}{4} + p_3) \Rightarrow p_3 = \frac{1}{4}$$

Finally let's check $\frac{1}{6} + \frac{1}{12} + \frac{1}{4} = \frac{1}{2}$ as required.

20. Let $X_i = +1$ if the i-th step was to the right and -1 if it was to the left. We can model their final position as:

$$X = \sum_{i=1}^{n} X_i$$

(a) We first note $\mathbb{E}[X_i] = -\frac{1}{3} + \frac{2}{3} = \frac{1}{3}$. Then, by linearity of expectation: $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n}{3}$.

(b) We first note: $\texttt{Var}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = 1 - \frac{1}{9} = \frac{8}{9}$. We assume the person is so incapacitated that their steps are independent. We can then use the decomposition $\texttt{Var}(X) = \sum_{i=1}^{n} \texttt{Var}(X_i) = \frac{8n}{9}$.

c) For n reasonably large, the distribution of X will be approximately Gaussian with a mean of $\mathbb{E}[X]$ and a variance $\texttt{Var}(X)$. So using the fact that roughly 95% of the Gaussian distribution's mass lies within 2 standard deviations of its mean, we would expect that, with probability ~ 0.95, the person's position is $\frac{n}{3} \pm \frac{2\sqrt{8n}}{3}$).

Worksheet 5 Solutions

1. a) By symmetry of the Gaussian distribution about the mean, $\Pr(X \geq 10) = 1/2$.

   b) The probability of exactly attaining any specific value in a continuous distribution is 0.

   c) First note: $\sigma^2 = 16 \Rightarrow \sigma = 4$. Then: $\frac{|X-\mu|}{\sigma} = \frac{14-10}{4} = 1$. Using the 68-95-99 rule for the Gaussian distribution: $\Pr(X \geq 14) = (1 - 0.683)/2 \approx 0.1585$, where we divide by two to account for the lower tail.

   d) As before: $\frac{|X-\mu|}{\sigma} = 2$. Therefore $\Pr(X \leq 2) \approx (1 - 0.955)/2 = 0.0225$

2. a) There are an average of 3.154 calls per hour, so we should choose $\hat{\lambda} = 3.154$.

   b) The expected number of intervals (out of 500) in which k calls were received is: $\Pr(X = k) \cdot 500$, where:

$$\Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

   Plugging in our estimate of λ from part (a), and applying the process above, we get the following:

k	0	1	2	3	4	5	6	7	8	≥ 9
N_k	22	66	106	115	85	55	28	13	10	0
$\mathbb{E}[N_k]$	21.34	67.31	106.14	111.59	87.99	55.51	29.17	13.14	5.18	2.61

   Note that we can compute $\Pr(X \geq 9) = 1 - \Pr(X < 9)$.

4. a) The MLE for the bias is just the empirical frequency of heads, so we would estimate $\hat{p} = 1$.

   b) The Laplace smoothing rule uses:

$$\hat{p} = \frac{k+1}{n+2} = \frac{21}{22}$$

   c) The sequence has 4 heads, so it's $\Pr(\texttt{HHTTHH}) = \left(\frac{21}{22}\right)^4 \left(\frac{1}{22}\right)^2 \approx 0.0017$

5. a) The bag-of-words representation just counts the number of occurrences of each word:

w	`a`	`rose`	`is`	`flower`
N_w	3	3	2	0

   The vector itself is just the bottom row of the table.

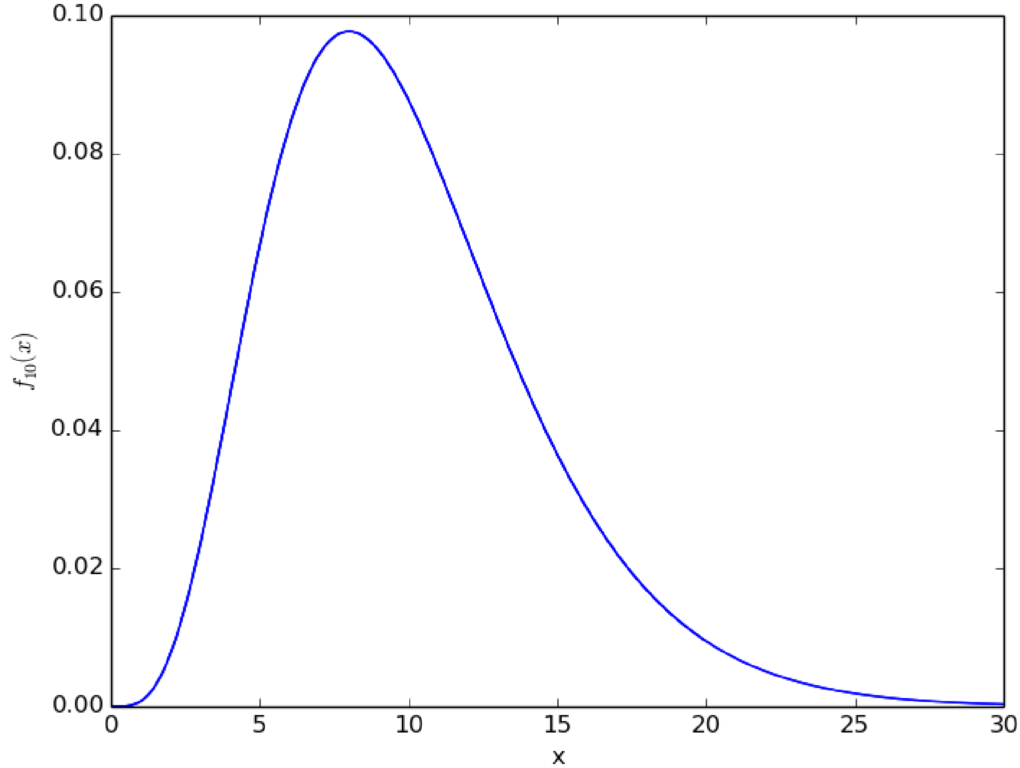
   b) Let $N = \sum_{w \in V} N_w$. The MLE of the multinomial is just the empirical frequency so $\Pr(X = w) = \frac{N_w}{N}$ so:

w	`a`	`rose`	`is`	`flower`
$\Pr(X = w)$	3/8	0.3/8	1/4	0

   c) Then, with Laplace smoothing $\Pr(X = w) = \frac{N_w+1}{N+|V|}$. So:

w	`a`	`rose`	`is`	`flower`
$\Pr(X = w)$	1/3	1/3	1/4	1/12

8. a) Here is an example plot - it's important to choose the X-axis range wide enough to see the entire shape of the distribution:



Here is the Python code I used to generate this (and for part (b) below):

```python
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

np.random.seed(13298)

chi2 = stats.chi2(10)

# Plot the density at a range of query points

xi   = np.linspace(0, 30, 1000)
f    = chi2.pdf(xi)

plt.plot(xi, f, 'b-')
plt.xlabel("x")
plt.ylabel("$f_{10}(x)$")
plt.savefig("chi2.png")

# estimate the median from samples

samples = chi2.rvs(size=10000)
print("Estimated Median: {}".format(np.median(samples)))
```

b) I got 9.23 as the estimated median - your value may be a bit different, but should be pretty close. If your answer was different by > 0.5, try running your code again with a larger number of samples.