Final exam    PS 200A    December 2019

You can use class notes for this exam and also computers and calculators, although heavy-duty calculation should not be necessary. If you use external sources like books, webpages, youtube videos, and so forth, please cite your sources in the exam itself. Getting help from another living being other than Michael or Ashley is not allowed. If you notice any typos or errors, please let Michael know and he will send out corrections via email.
Problem 1. Recall that the covariance of two random variables $X$ and $Y$ is defined as

$$Cov(X, Y) = E(XY) - E(X)E(Y).$$

a. Show that $Cov(X, X) = Var(X)$.

b. Show that $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$. 
c. Show that $Var(X + Y + Z) = Var(X) + Var(Y) + Var(Z) + 2Cov(X, Y) + 2Cov(X, Z) + 2Cov(Y, Z)$. Hint: We know that $Var(X + Y + Z) = Var(X + (Y + Z))$, and we can use the result from b. above.

d. Show that $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$.  

Recall that the correlation of two random variables $X$ and $Y$ is defined as

$$
\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.
$$

e. Assume that $Var(X) \neq 0$. Show that $\rho(X, 6X + 2) = 1$.

f. Assume that $Var(X) \neq 0$. Show that $\rho(X, -2X + 4) = -1$. 

Say that $X$ is the number of dogs a person has and $Y$ is the number of cats a person has. The total number of pets a person has is $X + Y$. Say that in a population, $Var(X) = 1$.

g. The correlation between the number of dogs a person has and the number of pets a person has is $\rho(X, X + Y)$. Say that $Var(Y) = 3$ and $Cov(X,Y) = 0$. Compute $\rho(X, X + Y)$.

h. The correlation between the number of dogs a person has and the number of pets a person has is $\rho(X, X + Y)$. Say that $Var(Y) = 99$ and $Cov(X,Y) = 0$. Compute $\rho(X, X + Y)$. 
i. Show that if $\rho(X, Y) = 1$, then $\rho(X, X + Y) = 1$. 
Problem 2. Recall that the number of (say) industrial accidents in a month \( X \) is usually modeled with the Poisson distribution, which is given by

\[
Prob(X = m) = e^{-\lambda} \frac{\lambda^m}{m!}
\]

where \( \lambda \) (lambda) is a parameter which corresponds to the average number of accidents per month. Remember that \( m! \) is the factorial of \( m \) (for example, \( 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120 \)), and \( e = 2.71828 \).

a. Say that \( \lambda = 1 \). We can compute the probability that \( X \) takes on various values in the table below. Note that some of the table entries are left blank. Please fill in the blank entries in the table, explaining how you got your numbers. Three decimal places is fine.

<table>
<thead>
<tr>
<th>( X )</th>
<th>if lambda is 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.368</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.061</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.003</td>
</tr>
<tr>
<td>6</td>
<td>0.001</td>
</tr>
</tbody>
</table>

b. Using your completed table, calculate the expected value of \( X \) (assume for convenience that \( X \) takes only the values 0, 1, 2, 3, 4, 5, 6).

c. Using your completed table, calculate \( E(X^2) \).

d. Calculate \( Var(X) \).
e. Now let $\lambda$ take on the values 1, 2, 3, 4, 5, 6, 7, 8, 9 as shown in the table below.

<table>
<thead>
<tr>
<th>$X = x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.368</td>
<td>0.135</td>
<td>0.018</td>
<td>0.007</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>0.368</td>
<td>0.271</td>
<td>0.149</td>
<td>0.073</td>
<td>0.034</td>
<td>0.006</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.184</td>
<td>0.271</td>
<td>0.224</td>
<td>0.147</td>
<td>0.084</td>
<td>0.045</td>
<td>0.011</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>3</td>
<td>0.061</td>
<td>0.180</td>
<td>0.224</td>
<td>0.195</td>
<td>0.140</td>
<td>0.089</td>
<td>0.052</td>
<td>0.029</td>
<td>0.015</td>
</tr>
<tr>
<td>4</td>
<td>0.015</td>
<td>0.090</td>
<td>0.168</td>
<td>0.195</td>
<td>0.175</td>
<td>0.134</td>
<td>0.091</td>
<td>0.057</td>
<td>0.034</td>
</tr>
<tr>
<td>5</td>
<td>0.003</td>
<td>0.101</td>
<td>0.156</td>
<td>0.175</td>
<td>0.161</td>
<td>0.128</td>
<td>0.092</td>
<td>0.061</td>
<td>0.061</td>
</tr>
<tr>
<td>6</td>
<td>0.001</td>
<td>0.012</td>
<td>0.050</td>
<td>0.104</td>
<td>0.161</td>
<td>0.149</td>
<td>0.122</td>
<td>0.091</td>
<td>0.091</td>
</tr>
</tbody>
</table>

f. Please fill in the blank entries in the table, explaining how you got your numbers.

g. One interesting thing in the table is that for each value of $\lambda$, there are two entries in the column which are exactly equal. For example, when $\lambda = 5$, we can see that $\text{Prob}(X = 4)$ and $\text{Prob}(X = 5)$ are both 0.175. Is this a coincidence? Please explain why we see this pattern.
We are going to introduce the idea of estimation.

Say that we observe $X = 2$, in other words, we had two industrial accidents last month. We do not know what $\lambda$ is—we do not know the average number of industrial accidents per month. All we know is that last month we had two.

It is possible that $\lambda = 9$, in which case $\Prob(X = 2) = 0.005$. In other words, it is possible that we have 9 accidents per month on average and last month we were extremely lucky to have far fewer than normal.

It is possible that $\lambda = 1$, in which case $\Prob(X = 2) = 0.184$. In other words, it is possible that we have 1 accident per month on average and last month we were unlucky.

Given that all we know is that $X = 2$ last month, what is our best guess about what $\lambda$ is? If we say that $\lambda = 9$, then $X = 2$ would be very unlikely, so $\lambda = 9$ is not a great guess.

The idea behind maximum likelihood estimation is that given that we observe $X = 2$, our best guess about $\lambda$ is whatever $\lambda$ maximizes $\Prob(X = 2)$ given $\lambda$. For example, when $X = 2$, we look at the row of probabilities 0.184, 0.271, 0.224, 0.147, 0.084 and so forth, and find that the largest value in this row is 0.271, which corresponds to $\lambda = 2$. So if $X = 2$, then the maximum likelihood estimate is $\lambda = 2$.

h. When $X = 0$, find the maximum likelihood estimate of $\lambda$ using your table above (you only have to consider values of $\lambda$ in the table).

i. Find the maximum likelihood estimate of $\lambda$ when $X = 1$, when $X = 2$, when $X = 3$, and so forth up to $X = 6$, again using the table.
Here is the table again (you can fill in the blank entries from your work on the previous page).

<table>
<thead>
<tr>
<th></th>
<th>if lambda is</th>
<th>if lambda is</th>
<th>if lambda is</th>
<th>if lambda is</th>
<th>if lambda is</th>
<th>if lambda is</th>
<th>if lambda is</th>
<th>if lambda is</th>
<th>if lambda is</th>
</tr>
</thead>
<tbody>
<tr>
<td>X = 0</td>
<td>0.368</td>
<td>0.135</td>
<td></td>
<td>0.018</td>
<td>0.007</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>X = 1</td>
<td>0.368</td>
<td>0.271</td>
<td>0.149</td>
<td>0.073</td>
<td>0.034</td>
<td>0.006</td>
<td>0.003</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>X = 2</td>
<td>0.184</td>
<td>0.271</td>
<td>0.224</td>
<td>0.147</td>
<td>0.084</td>
<td>0.045</td>
<td>0.011</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>X = 3</td>
<td>0.061</td>
<td>0.180</td>
<td>0.224</td>
<td>0.195</td>
<td>0.140</td>
<td>0.089</td>
<td>0.052</td>
<td>0.029</td>
<td>0.015</td>
</tr>
<tr>
<td>X = 4</td>
<td>0.015</td>
<td>0.090</td>
<td>0.168</td>
<td>0.195</td>
<td>0.175</td>
<td>0.134</td>
<td>0.091</td>
<td>0.057</td>
<td>0.034</td>
</tr>
<tr>
<td>X = 5</td>
<td>0.003</td>
<td>0.101</td>
<td>0.156</td>
<td>0.175</td>
<td>0.161</td>
<td>0.128</td>
<td>0.092</td>
<td>0.061</td>
<td></td>
</tr>
<tr>
<td>X = 6</td>
<td>0.001</td>
<td>0.012</td>
<td>0.050</td>
<td>0.104</td>
<td></td>
<td>0.161</td>
<td>0.149</td>
<td>0.122</td>
<td>0.091</td>
</tr>
</tbody>
</table>

Now we will introduce the idea of confidence intervals.

Say we observe $X = 2$, in other words, we had two accidents last month. Again, this is possible for many values of $\lambda$. But if $\lambda = 9$, then $X = 2$ is very unlikely, occurring with probability $0.005$.

The idea of confidence intervals is simply that we will say something is possible only if it occurs with at least some reasonable level of probability. If $X = 2$, it is possible that $\lambda = 9$ but that means we are OK with saying that an event occurred with probability 0.005, which is kind of a miracle. We don’t usually believe in miracles.

Everyone has a different idea of what a miracle is. Say that for me, a miracle is an event that happens with probability of less than 0.1. In other words, I am willing to accept an event can happen if its probability is at least 0.1. So if $X = 2$, the possible values of $\lambda$ in the table above which make the probability of $X = 2$ greater than 0.10 are $\lambda = 1, 2, 3, 4$. We would then say that the confidence interval for $\lambda$ is $[1, 4]$, the interval from 1 to 4. If $\lambda$ is outside this interval, then $X = 2$ would be a miracle (a miracle is an event that happens with probability less than 0.1). We call this confidence interval a 90% confidence interval because $0.90 = 1 - 0.1$.

j. When $X = 4$, find the 90% confidence interval for $\lambda$.

k. Find the 90% confidence intervals for $\lambda$ when $X = 0$, when $X = 1$, and so forth up to $X = 6$, using the table.
Here is the table again.

<table>
<thead>
<tr>
<th></th>
<th>if lambda is 1</th>
<th>if lambda is 2</th>
<th>if lambda is 3</th>
<th>if lambda is 4</th>
<th>if lambda is 5</th>
<th>if lambda is 6</th>
<th>if lambda is 7</th>
<th>if lambda is 8</th>
<th>if lambda is 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>X = 0</td>
<td>0.368</td>
<td>0.135</td>
<td>0.018</td>
<td>0.007</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>X = 1</td>
<td>0.368</td>
<td>0.271</td>
<td>0.149</td>
<td>0.073</td>
<td>0.034</td>
<td>0.006</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>X = 2</td>
<td>0.184</td>
<td>0.271</td>
<td>0.224</td>
<td>0.147</td>
<td>0.084</td>
<td>0.045</td>
<td>0.011</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>X = 3</td>
<td>0.061</td>
<td>0.180</td>
<td>0.224</td>
<td>0.195</td>
<td>0.140</td>
<td>0.089</td>
<td>0.052</td>
<td>0.029</td>
<td>0.015</td>
</tr>
<tr>
<td>X = 4</td>
<td>0.015</td>
<td>0.090</td>
<td>0.168</td>
<td>0.195</td>
<td>0.175</td>
<td>0.134</td>
<td>0.091</td>
<td>0.057</td>
<td>0.034</td>
</tr>
<tr>
<td>X = 5</td>
<td>0.003</td>
<td>0.101</td>
<td>0.156</td>
<td>0.175</td>
<td>0.161</td>
<td>0.128</td>
<td>0.092</td>
<td>0.061</td>
<td></td>
</tr>
<tr>
<td>X = 6</td>
<td>0.001</td>
<td>0.012</td>
<td>0.050</td>
<td>0.104</td>
<td>0.161</td>
<td>0.149</td>
<td>0.122</td>
<td>0.091</td>
<td></td>
</tr>
</tbody>
</table>

1. Now let’s find 95% confidence intervals, in other words, we believe that a miracle is an event that happens with probability of less than 0.05. We have a greater tolerance for unlikely events than before. Find the 95% confidence intervals for \( \lambda \) when \( X = 0 \), when \( X = 1 \), and so forth up to \( X = 6 \), using the table.

m. Now let’s find 99% confidence intervals, in other words, we believe that a miracle is an event that happens with probability of less than 0.01. We have an even greater tolerance for unlikely events, and will reject something only if its probability is lower than 0.01. For example, we now are willing to entertain the possibility of Bigfoot, love at first sight, and Santa Claus. Find the 99% confidence intervals for \( \lambda \) when \( X = 0 \), when \( X = 1 \), and so forth up to \( X = 6 \), using the table.

n. Why, for each realization of \( X \), does the 99% confidence interval contain (is wider than) the 95% confidence interval, which in turn contains the 90% confidence interval?

o. Why, for each realization of \( X \), is the maximum likelihood estimate of \( \lambda \) you found in part i. earlier contained in the confidence intervals for \( \lambda \)?
Problem 3. Say that there are 100 people in a town. Each person is either a Democrat (D) or Republican (R), a Couch-surfer (C) or a Walker (W), a Boomer (B) or a Millennial (M), and a Smoker (S) or a Non-smoker (N). The distribution of people is shown by the table below.

<table>
<thead>
<tr>
<th></th>
<th>Smoker</th>
<th>Non-smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat, Couch, Boomer</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Democrat, Couch, Millennial</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Democrat, Walker, Boomer</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Democrat, Walker, Millennial</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Republican, Couch, Boomer</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Republican, Couch, Millennial</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Republican, Walker, Boomer</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Republican, Walker, Millennial</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

a. What is \( \text{Prob(Smoker)} \)? \( \text{Prob(Non-smoker)} \)? \( \text{Prob(Democrat)} \)? \( \text{Prob(Republican)} \)? \( \text{Prob(Couch)} \)? \( \text{Prob(Walker)} \)? \( \text{Prob(Boomer)} \)? \( \text{Prob(Millennial)} \)?

b. What is \( \text{Prob(Smoker, Democrat)} \) (in other words, the probability that a person is a Smoker and a Democrat)? What is \( \text{Prob(Smoker, Couch)} \)? \( \text{Prob(Smoker, Boomer)} \)?
c. Say we are trying to predict whether a person in this town is a smoker. Which factor (party affiliation, exercise habits, or age) best predicts whether a person in this town is a smoker or not?

If party affiliation is a big predictor of whether a person smokes, we would expect that Prob(\text{Smoker}|\text{Democrat}) is a lot different from Prob(\text{Smoker}|\text{Republican}). In other words, a change in party affiliation is related to a large change in smoking habits.

In other words, we can calculate the following three numbers.

\[
\begin{align*}
\text{Prob(\text{Smoker}|\text{Democrat})} & - \text{Prob(\text{Smoker}|\text{Republican})} \\
\text{Prob(\text{Smoker}|\text{Couch})} & - \text{Prob(\text{Smoker}|\text{Walker})} \\
\text{Prob(\text{Smoker}|\text{Boomer})} & - \text{Prob(\text{Smoker}|\text{Millennial})}
\end{align*}
\]

Calculate these three numbers. Which characteristic (party affiliation, exercise habits, or age) is the most important predictor of whether a person in the general population is a smoker? Which characteristic (party affiliation, exercise habits, or age) is the second most important predictor of whether a person in the general population is a smoker? Which characteristic (party affiliation, exercise habits, or age) is the least important predictor whether a person in the general population is a smoker?
d. Now say that we consider not the general population but only the population of couch-surfers. Among these people, which characteristic (party affiliation or age) is more important in predicting whether a person is a smoker? Answer this question by calculating the following two numbers.

\[
\text{Prob(Smoker|Couch, Democrat)} - \text{Prob(Smoker|Couch, Republican)}
\]

\[
\text{Prob(Smoker|Couch, Boomer)} - \text{Prob(Smoker|Couch, Millennial)}
\]

e. Now say that we consider only the population of walkers. Among these people, which characteristic (party affiliation or age) is more important in predicting whether a person is a smoker? Answer this question by calculating the following two numbers.

\[
\text{Prob(Smoker|Walker, Democrat)} - \text{Prob(Smoker|Walker, Republican)}
\]

\[
\text{Prob(Smoker|Walker, Boomer)} - \text{Prob(Smoker|Walker, Millennial)}
\]

f. Among the couch-surfers, which factor best predicts smoking, party affiliation or age? Among the walkers, which factor best predicts smoking, party affiliation or age? In other words, once you have predicted smoking using a person’s exercise habits, what factor would you say best explains the remaining variation: party affiliation or age?

g. Are your answers to part c. above and part f. above the same or different? Meditate on this and explain why your answers are the same or are different.
h. Using our methodology from part c. above, which factor best explains the party affiliation of a person in the general population: exercise habits, age, or smoking habits?
Problem 4. Say that the random variable $Z$ is distributed $N(0, 1)$, in other words $Z$ has a normal distribution with mean 0 and standard deviation 1. Sometimes we call this a “standard normal” distribution.

a. Use a standard normal distribution table (for example at https://bit.ly/359cTOz) to calculate the probability that $Z$ is less than or equal to 1.5.

b. What is $\text{Prob}(Z \leq -0.4)$?

c. What is $\text{Prob}(Z > 1)$?

d. What is $\text{Prob}(Z > 1.96 \text{ or } Z < -1.96)$?

e. What is $\text{Prob}(|Z| < 1.96)$?
f. Say that $Y = Z + 12$. The random variable $Y$ also has a normal distribution. Remember that $Z$ is distributed $N(0, 1)$. What is the mean of $Y$? What is the standard deviation of $Y$?

g. Use a standard normal distribution table (for example at https://bit.ly/359cTOz) to calculate the probability that $Y$ is less than or equal to 13.5.

h. What is $\text{Prob}(Y \leq 11.6)$?

i. What is $\text{Prob}(Y > 13)$?

j. What is $\text{Prob}(Y > 13.96$ or $Y < 10.04$)?

k. What is $\text{Prob}(|Y - 12| < 1.96)$?
1. Say that $X = 2Z + 12$. The random variable $X$ also has a normal distribution. Remember that $Z$ is distributed $N(0, 1)$. What is the mean of $X$? What is the standard deviation of $X$?

m. Use a standard normal distribution table (for example at https://bit.ly/359cTOz) to calculate the probability that $X$ is less than or equal to 15.

n. What is $\text{Prob}(X \leq 11.2)$?

o. What is $\text{Prob}(X > 14)$?

p. What is $\text{Prob}(X > 15.92 \text{ or } X < 8.08)$?

q. What is $\text{Prob}(|(X - 12)/2| < 1.96)$?
Say that there are four kids and each kid has Pokémon cards. Kid 1 has $X_1$ cards, Kid 2 has $X_2$ cards, and so forth. The random variables $X_1$, $X_2$, $X_3$, and $X_4$ are independent and are all distributed $N(100,24)$. In other words, $X_i$ is distributed according to a normal distribution with mean 100 and standard deviation 24.

r. Say that $S = X_1 + X_2 + X_3 + X_4$ is the total number of Pokémon cards the four kids own. What is $E(S)$? What is $Var(S)$? (Hint: you can use the reasoning from part c. of problem 1 on this exam).

s. Say that $M = (X_1 + X_2 + X_3 + X_4)/4$ is the average number of Pokémon cards the four kids own. What is $E(M)$? What is $Var(M)$? (Hint: note that $M = S/4$.)

t. It turns out that $M$ is distributed according to a normal distribution. What is the mean and standard deviation of this normal distribution?

u. What is $\text{Prob}(M > 123.52 \text{ or } M < 76.48)$?

v. What is $\text{Prob}(|(M - 100)/12| < 1.96)$?
Now say that there are nine kids and each kid has Pokémon cards. Kid 1 has $X_1$ cards, Kid 2 has $X_2$ cards, and so forth. The random variables $X_1, X_2, X_3, \ldots, X_9$ are independent and are all distributed N(100,24), just as before.

w. Say that $S = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9$ is the total number of Pokémon cards the nine kids own. What is $E(S)$? What is $Var(S)$? (Hint: you can use the reasoning from part c. of problem 1 on this exam).

x. Say that $M = (X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9)/9$ is the average number of Pokémon cards the four kids own. What is $E(M)$? What is $Var(M)$? (Hint: note that $M = S/9$.)

y. It turns out that $M$ is distributed according to a normal distribution. What is the mean and standard deviation of this normal distribution?

z. What is Prob($M > 115.68$ or $M < 84.32$)?

zz. What is Prob($|(M - 100)/8| < 1.96$)?

zzz. When the number of kids in the group goes from 4 to 9, what happens to the distribution of $M$, the average number of Pokémon cards in the group?