Statistics for PP&A
Spring 2004 Problem Set #1 Suggested Answers

Stock and Watson Exercises 2.2, 2.3, 2.5.d, 2.6.b, 2.8

2.2 Using the random variables $X$ and $Y$ from Table 2.2, consider two new random variables $W = 3 + 6X$ and $V = 20 - 7Y$. Compute (a) $E(W)$ and $E(V)$; (b) $\sigma_W^2$ and $\sigma_V^2$; and (c) $\sigma_{WV}$ and $\text{corr}(W, V)$.

<table>
<thead>
<tr>
<th></th>
<th>Rain ($X = 0$)</th>
<th>No Rain ($X = 1$)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long Commute ($Y = 0$)</td>
<td>0.15</td>
<td>0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Short Commute ($Y = 1$)</td>
<td>0.15</td>
<td>0.63</td>
<td>0.78</td>
</tr>
<tr>
<td>Total</td>
<td>0.30</td>
<td>0.70</td>
<td>1.00</td>
</tr>
</tbody>
</table>

a It might be helpful to give $W$ and $V$ interpretations. For example, you could think of $V$ as the cost of a cab ride contingent on the length of the commute: $V$ is $13 (= 20 - 7 \cdot 1)$ if the commute is short and $20 (= 20 - 7 \cdot 0)$ if the commute is long.

The answer uses the mean of a linear function of a random variable (end of section 2.2).

\[
E(W) = E(3 + 6X) = 3 + 6E(X) \text{ by equation 2.12} = 3 + 6(0.7) \text{ because } E(X) = 0.70 = 7.20
\]
\[
E(V) = E(20 - 7Y) = 20 - 7E(Y) \text{ by equation 2.12} = 20 - 7(0.78) \text{ because } E(Y) = 0.78 = 14.54
\]

b The answer uses the variance of a linear function of a random variable (end of section 2.2) and the variance of a Bernoulli random variable (section 2.2).

\[
\sigma_W^2 = 6^2 \sigma_X^2 \text{ by equation 2.13} = 6^2(0.7)(0.3) \text{ by equation 2.7, variance for a Bernoulli variable} = 7.56
\]
\[
\sigma_V^2 = 7^2 \sigma_Y^2 \text{ by equation 2.13} = 7^2(0.78)(0.2) \text{ by equation 2.7, variance for a Bernoulli variable} = 8.4084
\]

c The answer uses the definition of covariance (equation 2.22) and the probability distribution given in the example.
\[ \sigma_{WV} = E[(W - \mu_W)(V - \mu_V)] \]
\[ = \sum_{i=1}^{k} \sum_{j=1}^{l} (w_j - \mu_W)(v_i - \mu_V) \Pr(W = w_j, V = v_j) \]

<table>
<thead>
<tr>
<th></th>
<th>Rain (W = 3)</th>
<th>No Rain (W = 9)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long Commute (V = 20)</td>
<td>0.15</td>
<td>0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Short Commute (V = 13)</td>
<td>0.15</td>
<td>0.63</td>
<td>0.78</td>
</tr>
<tr>
<td>Total</td>
<td>0.30</td>
<td>0.70</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Recall that \( \mu_W = 7.20 \) and \( \mu_V = 14.54 \).

\[
\sigma_{WV} = \sum_{i=1}^{k} \sum_{j=1}^{l} (w_j - \mu_W)(v_i - \mu_V) \Pr(W = w_j, V = v_j) \\
= (3 - 7.20)(20 - 14.54)(0.15) \text{ because } 0.15 \text{ of the time } W = 3 \text{ and } V = 20 \\
+ (3 - 7.20)(13 - 14.54)(0.15) \text{ because } 0.15 \text{ of the time } W = 3 \text{ and } V = 13 \\
+ (9 - 7.20)(20 - 14.54)(0.07) \text{ because } 0.07 \text{ of the time } W = 9 \text{ and } V = 20 \\
+ (9 - 7.20)(13 - 14.54)(0.63) \text{ because } 0.63 \text{ of the time } W = 9 \text{ and } V = 20 \\
= -3.53
\]

With covariance in hand, correlation is easy to compute.

\[
\text{corr}(W, V) = \frac{\sigma_{WV}}{\sigma_W \sigma_V} = -3.53 / (\sqrt{7.56} \sqrt{8.4084}) = -0.44
\]

Is there any intuition for why \( W \) and \( V \) are negatively correlated? \( X \) and \( Y \) are positively correlated: short commutes are associated with no rain. Because \( W = 3 + 6X \) increases with \( X \), and \( V = 20 - 7Y \) decreases with \( Y \), it’s not surprising that \( W \) and \( V \) are negatively correlated. When \( X \) and \( Y \) are both up, which happens frequently, \( W \) is up and \( V \) is down. Is this consistent with your interpretation of \( W \) and \( V \)?

2.3 I covered this in some detail in class.

2.5.d If \( Y \) is distributed \( N(5, 2) \), find \( \Pr(6 \leq Y \leq 8) \).

We know how to compute \( \Pr(d_1 \leq Z \leq d_2) \) when \( Z \) is a standard normal variable. Can we convert \( \Pr(6 \leq Y \leq 8) \), an example of a \( \Pr(c_1 \leq Y \leq c_2) \) type problem into this problem? The standard deviation of \( Y \), \( \sigma_Y \) is \( \sqrt{2} \), or about 1.41.
\[ d_1 = \frac{(c_1 - \mu_Y)}{\sigma_Y} \]
\[ = \frac{(6 - 5)}{\sqrt{2}} \]
\[ \approx 0.71 \]
\[ d_2 = \frac{(c_2 - \mu_Y)}{\sigma_Y} \]
\[ = \frac{(8 - 5)}{\sqrt{2}} \]
\[ \approx 2.12 \]

So we have converted the problem into \( \Pr(0.71 \leq Z \leq 2.12) \) and we can turn to the standard normal tables.

\[ \Pr(0.71 \leq Z \leq 2.12) = \Pr(Z \leq 2.12) - \Pr(Z \leq 0.71) = 0.2218 \]

Around 22 percent of the time, \( Y \sim N(5, 2) \) will fall between 6 and 8.

2.6.b If \( Y \) is distributed \( \chi^2_4 \), find \( \Pr(Y \leq 7.78) \).
This problem is a direct look-up from the \( \chi^2 \) table and takes advantage of the fact that 7.78 happens to be a critical value. Ten percent of the time, \( Y \sim \chi^2_4 \) will exceed 7.78; so 90 percent \( Y \leq 7.78 \).

2.8 In any year, the weather can inflict storm damage to a home. From year to year, the damage is random. Let \( Y \) denote the dollar value of damage in any given year. Suppose that in 95 percent of the years, \( Y = \$0 \), but in 5 percent of the years \( Y = \$20,000 \).

a What is the mean and standard deviation of the damage in any year?
The mean of the damage is \( E(Y) \).
\[ E(Y) = p_1y_1 + p_2y_2 \]
\[ = 0.95(\$0) + 0.05(\$20,000) \]
\[ = \$1,000 \]

The variance of the damage is \( E[(Y - E(Y))^2] \)
\[ E[(Y - E(Y))^2] = p_1(y_1 - E(Y))^2 + p_2(y_2 - E(Y))^2 \]
\[ = 0.95(\$0 - \$1,000)^2 + 0.05(\$20,000 - \$1,000)^2 \]
\[ = 19,000,000 \]

The standard deviation of the damage is the square root of the variance.
\[ \sigma_Y = \sqrt{\text{var}(Y)} = \sqrt{19,000,000} \approx 4,358.90 \]
A key feature of the insurance pool is that the damage to each house is uncorrelated with damage to any other house. Consider: homeowners on the Florida coast do not want to form an insurance pool together.

In this case, the average damage can be thought of as adding up the total damage \( Y_1 + Y_2 + \cdots + Y_{100} \) and dividing the cost over the 100 participating households, or

\[
\overline{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i
\]

i. First we can compute the expected value of the average damage:

\[
E(\overline{Y}) = E\left(\frac{1}{n} \sum_{i=1}^{n} Y_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(Y_i) = \frac{1}{n} \times n \times E(Y_i) \text{ because all of the } E(Y_i) \text{ are equal.} = E(Y_i) = $1,000 \text{ (we computed } E(Y_i) = $1,000 \text{ in part a)}
\]

ii. We would like to compute \( \Pr(\overline{Y} > 2000) \). We need to know the probability distribution for the average damage.

Is it possible that this is a normal distribution problem? \( \overline{Y} \) is a mean of random variables, and the Central Limit Theorem states that means of random variables are normally distributed. So \( \overline{Y} \) is normally distributed with mean \( E(\overline{Y}) \) and variance \( \sigma^2_{\overline{Y}} \). Before we worry about these values, let’s convert the normal problem into a standard normal problem.

\[
\Pr \left[ \overline{Y} > 2000 \right] \text{ with } \overline{Y} \sim N(E(\overline{Y}), \sigma^2_{\overline{Y}})
\]

is the same as

\[
\Pr \left[ Z > \frac{2000 - E(\overline{Y})}{\sigma_{\overline{Y}}} \right] \text{ with } Z \sim N(0, 1)
\]

Again, don’t worry about \( E(\overline{Y}) \) and variance \( \sigma^2_{\overline{Y}} \) yet. We know that if \( \overline{Y} \sim N(E(\overline{Y}), \sigma^2_{\overline{Y}}) \), then a standardization of \( \overline{Y} \) (by subtracting the mean and dividing by the standard deviation) is distributed \( N(0, 1) \).

Now, let’s fill in the pieces that we need. We have already computed \( E(\overline{Y}) = $1,000 \) in part b.i.

The tricky part is \( \sigma^2_{\overline{Y}} \), but the Central Limit Theorem gives us guidance. Because \( \overline{Y} \) is an average of i.i.d. random variables, its variance is \( \sigma^2_{\overline{Y}} = \sigma^2_Y/n \). The standard deviation of \( \overline{Y} \) is \( \sqrt{\sigma^2_Y/n} \).

We have the elements that we need: \( \sigma_{\overline{Y}} = 4358.90/10 = $435.89 \)

Before we proceed with the analytic answer, this is a good opportunity for some estimation. The standard deviation of average damage is slightly under $500. How often will average damage
exceeds its average ($1,000) by $1,000, or about two standard deviations ($2 \times \$500$)? This should not happen often, definitely less than 2.5 percent of the time.

$$\Pr\left[Z > \frac{2000 - 1000}{435.89}\right] = \Pr(Z > 2.29) = 1 - \Pr(Z < 2.29) = 0.0109.$$  

Indeed, this is a very rare event: average damage exceeds $2,000 around 1 percent of the time. (Consider how important independence of each $Y_i$ was for this result.)