Statistics 403 Problem Set 5

Due in lab on Friday, October 15th

1. The following questions are about a population of people, some of whom are home owners, and some of whom are renters.

(a) Suppose 65% of the population are home-owners, the home-owners have a mean annual income of $52,000, and the mean annual income of the population is $45,000. What is the mean annual income of the renters?

**Solution:** Apply the double expectation theorem:

\[
45000 = E(\text{Income}) = E_{\text{Own}}E(\text{Income}|\text{Own}) = 0.65 \cdot E(\text{Income}|\text{Own} = 1) + 0.35 \cdot E(\text{Income}|\text{Own} = 0) = 0.65 \cdot 52000 + 0.35 \cdot E(\text{Income}|\text{Own} = 0),
\]

where \( \text{Own} = 1 \) means the person owns their home, and \( \text{Own} = 0 \) means that the person rents their home.

Solving we get

\[
E(\text{Income}|\text{Own} = 0) = 32000.
\]

(b) Suppose the mean annual incomes of the home-owners and renters are $43,000 and $29,000, respectively, and the mean annual income in the population is $33,000. What proportion of the population owns their home?

**Solution:**

\[
33000 = E_{\text{Own}}E(\text{Income}|\text{Own}) = p \cdot E(\text{Income}|\text{Own} = 1) + (1 - p) \cdot E(\text{Income}|\text{Own} = 0) = p \cdot 43000 + (1 - p) \cdot 29000 = 14000p + 29000.
\]

Thus \( p = 4000/14000 \approx 0.29 \).

(c) Suppose the home owners have a mean annual income of $47,000 and the standard
deviation of home owners’ incomes is $7,000. Then suppose that the renters have a mean annual income of $31,000 and the standard deviation of renters’ incomes is $4,500. If 72% of the population are renters, what is the standard deviation of incomes in the population?

**Solution:** Apply the law of total variation to get the variance of incomes in the population:

\[
\text{var}(\text{Income}) = E_{\text{Own}} \text{var}(\text{Income}|\text{Own}) + \text{var}_{\text{Own}} E(\text{Income}|\text{Own})
\]

The first term is

\[
E_{\text{Own}} \text{var}(\text{Income}|\text{Own}) = 0.28 \cdot 7000^2 + 0.72 \cdot 4500^2 = 28300000.
\]

The overall mean is

\[
0.28 \cdot 47000 + 0.72 \cdot 31000 = 35480.
\]

Thus the second term is

\[
0.28(47000 - 35480)^2 + 0.72(31000 - 35480)^2 = 51609600.
\]

Thus the standard deviation of incomes is

\[
\sqrt{28300000 + 51609600} \approx 8939.
\]

(d) Suppose the home owners have a mean annual income of $52,000 and the standard deviation of home owners’ incomes is $5,500. Then suppose that the renters have a mean annual income of $33,000, and 55% of the population are renters. Not knowing the variance of renters’ incomes, what is the smallest possible value for the overall population variance in incomes?

**Solution:** We can apply the law of total variation, using zero as the variance of the renters’ incomes. The mean is

\[
\sqrt{28300000 + 51609600} \approx 8939.
\]
\[ 0.45 \cdot 52000 + 0.55 \cdot 33000 = 41550. \]

The \( \text{var}_{\text{Own}} E(\text{Income}|\text{Own}) \) term is
\[ 0.45(33000 - 41550)^2 + 0.55(52000 - 41550)^2 = 92957500. \]

Thus the smallest possible variance for the incomes is
\[ 0.45 \cdot 5500^2 + 92957500 = 106570000, \]
which corresponds to a standard deviation of 10323.

(e) Suppose the home owners have a mean annual income of $55,000 and the standard deviation of home owners’ incomes is $8,500. Then suppose that the renters have a mean annual income of $36,000, and the standard deviation of renters’ incomes is $5,500. If 79\% of the population are renters, what fraction of the overall variability in incomes is explained by home ownership status?

**Solution:** The \( E_{\text{Own}} \text{var}(\text{Income}|\text{Own}) \) term is
\[ 0.21 \cdot 8500^2 + 0.79 \cdot 5500^2 = 39070000. \]

The overall mean is
\[ 0.21 \cdot 55000 + 0.79 \cdot 36000 = 39990. \]

The \( \text{var} E_{\text{Own}}(\text{Income}|\text{Own}) \) term is
\[ 0.21(55000 - 39990)^2 + 0.79(36000 - 39990)^2 = 59889900. \]

Thus the fraction of variance explained by home ownership status is
\[ 59889900/(59889900 + 39070000) \approx 0.61. \]

2. Suppose that \( A \) and \( B \) are independent, standardized random variables. What are the values of (i) \( \text{cor}(A + B, A - B) \), and (ii) \( \text{cor}(A + B, A) \)?
Solution:

(i)

\[
\text{cov}(A + B, A - B) = \text{cov}(A, A) - \text{cov}(A, B) + \text{cov}(A, B) - \text{cov}(B, B) \\
= \text{var}(A) - \text{var}(B) \\
= 0.
\]

Note that we do not even need \( A \) and \( B \) to be independent for this to be true. Since the covariance is zero, the correlation must also be zero.

(ii)

\[
\text{cov}(A + B, A) = \text{cov}(A, A) + \text{cov}(B, A) \\
= \text{var}(A) \\
= 1.
\]

Also,

\[
\text{var}(A + B) = \text{var}(A) + \text{var}(B) = 2.
\]

(here we are using the independence of \( A \) and \( B \)).

Thus the correlation is

\[
\frac{\text{cov}(A + B, A)}{\text{SD}(A + B) \cdot \text{SD}(A)} = 1/\sqrt{2}.
\]

3. (a) Suppose that \( A \) and \( B \) are independent, standardized random variables. Describe all constants \( c \) and \( d \) for which \( cA + dB \) is standardized.

Solution: The expected value is

\[
E(cA + dB) = cE(A) + dE(B) = 0.
\]

Thus \( cA + dB \) is always centered, regardless of the values of \( c \) and \( d \).
The variance is
\[
\text{var}(cA + dB) = \text{var}(cA) + \text{var}(dB)
\]
\[
= c^2 \text{var}(A) + d^2 \text{var}(B)
\]
\[
= c^2 + d^2.
\]
Thus we need \(c^2 + d^2 = 1\) for the \(cA + dB\) to be standardized. These points form a circle centered at the origin with radius 1 if plotted in the \(c,d\) plane.

(b) Suppose that \(A\) and \(B\) are independent, standardized random variables. Show that \(cA + \sqrt{1 - c^2}B\) is standardized for any value \(-1 \leq c \leq 1\).

**Solution:** As in part (a), \(cA + \sqrt{1 - c^2}B\) has expected value zero since \(A\) and \(B\) have expected value zero and expectations are linear.

The variance of \(cA + \sqrt{1 - c^2}B\) is
\[
c^2 \text{var}(A) + (1 - c^2) \text{var}(B) = 1.
\]

(c) Suppose that \(A\) and \(B\) are standardized, random variables. For a given value \(-1 \leq r \leq 1\), find a constant \(c\) such that the correlation between \(cA + \sqrt{1 - c^2}B\) and \(A\) is equal to \(r\).

**Solution:** From part (b), we know that \(cA + \sqrt{1 - c^2}B\) is standardized, and you are given that \(A\) is standardized. Thus, the correlation and the covariance are the same. The covariance is
\[
\text{cov}(A, cA + \sqrt{1 - c^2}B) = c \cdot \text{var}(A) + c\sqrt{1 - c^2} \text{cov}(A, B) = c.
\]
Thus if we set \(c = r\), we will get \(\text{cor}(cA + \sqrt{1 - c^2}B, A) = r\), as desired.

(d) Suppose that \(A\), \(B\), and \(U\) are independent. For a given value \(0 \leq r \leq 1\), find a constant \(c\) such that the correlation between \(\sqrt{1 - c^2}A + cU\) and \(\sqrt{1 - c^2}B + cU\) is equal to \(r\).
Solution: From part (b) above, we know that $\sqrt{1-c^2}A + cU$ and $\sqrt{1-c^2}B + cU$ are both standardized, regardless of the value of $c$. Thus the correlation is the covariance, which is

$$\text{cov}(\sqrt{1-c^2}A + cU, \sqrt{1-c^2}B + cU) = c^2 \text{var}(U) = c^2.$$ 

Thus we set $c^2 = r$ and solve for $c$ to get $c = \sqrt{r}$.

4. Suppose we plan to collect data on peoples’ diets by conducting a survey in which siblings will be interviewed. We will collect data $X_1, \ldots, X_{2n}$ on $n$ sibling pairs, ($X_1$ and $X_2$ are data from a sibling pair, $X_3$ and $X_4$ are data from another sibling pair, and so on). Unrelated people respond independently, but the responses of siblings are correlated at level $r$.

(a) Determine the standard deviation of $\bar{X}$.

Solution: The covariance matrix has $2n$ rows and $2n$ columns (corresponding to the $2n$ observed data points). It is entirely zero, except for $n$ $2 \times 2$ blocks that look like this

$$
\begin{pmatrix}
\sigma^2 & r\sigma^2 \\
r\sigma^2 & \sigma^2
\end{pmatrix}
$$

The sum of values in each of these blocks is $2\sigma^2(1 + r)$. Thus the sum of all values in the covariance matrix is $2n\sigma^2(1 + r)$. Thus, based on a result given in the notes,

$$\text{var}(\bar{X}) = \frac{2n\sigma^2(1 + r)}{(2n)^2} = \frac{\sigma^2(1 + r)}{2n}.$$ 

So the standard deviation of $\bar{X}$ is $\sigma \sqrt{(1 + r)/(2n)}$.

Note some special cases:

- When $r = 0$ we have independent data, and the standard error is $\sigma/\sqrt{2n}$ which is “standard deviation of the data”/“square root of the sample size”, as we have known for a long time.

- When $r = 1$ the two people in a household are perfectly correlated, so the second person interviewed in each household provides no new information. Thus we really only have a sample size of $n$ (the number
of households), so the standard error is $\sigma/\sqrt{n}$, which is also “standard deviation of the data”/“square root of the sample size”.

(b) Suppose another researcher conducts a survey by interviewing $2n$ unrelated people. What is the ratio of the length of the confidence interval obtained in this way to the length of the confidence interval obtained using sibling data?

**Solution:** The length of the CI from part (a) is $4\sigma\sqrt{(1 + r)/2n}$. If we had independent data, the CI would be $\bar{X} \pm \sigma/\sqrt{2n}$, so the length of the interval is $4\sigma/\sqrt{2n}$. The ratio of these two lengths is $1/\sqrt{1 + r}$. Thus when $r = 0$, the lengths are the same, so the ratio is 1. When $r > 0$, the ratio is less than 1, reflecting the fact that the interval based on independent data is shorter. When $r = 1$, the interval based on independent data is exactly $1/\sqrt{2}$ times as long as the interval based on dependent data, which is the same ratio you get when reducing the sample size by half.