Classwork 2

Two More Questions on Covariance

1) Prove that if two random variables are independent, they have 0 covariance. (2 points)

Hint: start with this: if two random variables X and Y are independent, we know that $E[XY] = E[X]E[Y]$. You want to show that $Cov(X, Y) = 0$. Remember that $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$. You can also use the fact that μ_X and μ_Y are constants, so they can be pulled outside of an expectation.

2) (Continuing from the previous question) Show that the reverse is not true: two random variables can have zero covariance without being independent. This makes "independence" a stronger assumption than "0 covariance". (2 points)

Hint: One example is all that is required here: you just need to show that for some X and Y, X and Y are dependent and they have 0 covariance. Try this example: let X be a random variable that takes on {-1, 0, 1} each with equal probability. Let $Y = X^2$. You should be able to show that X and Y have 0 covariance but are not independent, because covariance only measures linear associations between variables.

Moving from Probability to Statistics

In workbook chapter 1, you learned how to calculate a random variable's expected value and variance using its probability distribution.

But what if you don't know the probabilitiy distribution? It turns out that you can *estimate* a random variable's expected value and variance as long as you have a sample of outcomes for that random variable. That's what the rest of this classwork is all about.

The procedure: suppose you have a random variable X and you take a sample of n observations with the intention of learning about some property of it, say for example its expected value. An *estimator* is a formula for estimating that value. For example, the best estimator for the expected value of X is the sample mean.

Consider the random variable "a dice roll". Recall from the workbook chapter 1 that E[dice roll] is 3.5. Let's see how close our estimates get to that value. We'll take $n = 3$: roll the dice 3 times to generate a sample.

```r
sample(1:6, size = 3, replace = T)
```

```
## [1] 3 6 3
```

The sample mean is defined as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where x_i is an observation in the sample (in classwork 1 it had a different meaning, it was a potential outcome for X):

```r
(3 + 6 + 3) / 3
```

```
## [1] 4
```

Actually a pretty good estimate! But is this just a fluke? Let's take 10 samples.

3) Calculate the sample mean $\bar{x}$ for each of these 10 samples (your answer should be 10 numbers). (1 point)

```r
for (i in 1:10) {
  sample(1:6, size = 3, replace = T) %>% print()
}
```

```
## [1] 2 2 6
## [1] 3 5 4
## [1] 6 6 1
## [1] 2 3 5
## [1] 3 3 1
## [1] 4 1 1
## [1] 5 3 2
## [1] 2 1 6
## [1] 3 4 6
## [1] 1 3 5
```

The sample mean is not always on target with 3.5: sometimes it's over, sometimes it's under, but in the long-run, its average seems to be around 3.5. Also notice that the sample mean $\bar{x}$ cannot be predicted exactly, so it is itself a random variable! Since it's a random variable, it has its own expected value and variance. We'll explore that in the next two problems.

4) Fill out the proof below. (1 point)

Definition: Unbiasedness: If the expected value of an estimator is on target (it's equal to the value of the true parameter), then we say that the estimator is unbiased. So for example, if $E[\bar{x}] = \mu_X$, the sample mean is said to be an **unbiased estimator** of the expected value of the random variable X. Show that the expected value of the sample mean is indeed an unbiased estimator for the random variable's expectation by filling in the empty steps of the proof.

Proof: We want to show that $E[\bar{x}] = \mu_X$. We'll start with $E[\bar{x}]$ and show that it ends up being equal to μ_X:

$$
\begin{aligned}
E[\bar{x}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] && (1)\\
&= \frac{1}{n}E\left[\qquad\right] && (2)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\qquad\right] && (3)\\
&= \frac{1}{n}\sum_{i=1}^{n}\mu_X && (4)\\
&= \left[\qquad\right] n\mu_X && (5)\\
&= \mu_X && (6)
\end{aligned}
$$

Hints:

- Constants can be pulled outside of expectations: $E[bX] = bE[X]$
- Expectations can be distributed across sums: $E[X + Y + Z] = E[X] + E[Y] + E[Z]$
- $E[x_i] = \mu_X$ by definition
- Since $5 + 5 + 5 = 3 \times 5$, it is true that $\sum_{i=1}^{n} 5 = n \times 5$

5) Consider a second estimator for μ_X: the first observation of the sample, which I'll call x_1.

In this question you'll show that while x_1 is also an unbiased for μ_X (that is, $E[x_1] = \mu_X$), x_1 is not the best estimator for μ_X because it has a larger variance than the sample mean $\bar{x}$, that is, it's a **less efficient** estimator for μ_X.

Definition: Efficiency: If one estimator has a lower variance than another estimator, it is said to be more **efficient** because theoretically, when you have a lower variance estimator, less data is required to get a precise estimate.

5a) Argue why $E[x_1] = \mu_X$. (1 point) This will be very short: a single sentence will suffice.

5b) Show that $Var(\bar{x}) = \frac{1}{n}\sigma_X^2$ (1 point) Recall that σ_X^2 is the variance of the random variable X.

5c) Argue why $Var(x_1) = \sigma_X^2$. (1 point) This will be very short: a single sentence will suffice.

5d) Reflecting on the previous two answers, argue why the sample mean $\bar{x}$ is a more efficient (lower variance) estimator of μ_X compared to the first observation of the sample x_1. (1 point)