Mean and Variance of a Linear Combination of Random Variables (Section 9.5)

Results of Expectation:

1. $E(aX + bY) = aE(X) + bE(Y)$.

2. For each $i = 1,2, \dots, n$, let a_i be a constant and $E(X_i) = \mu_i$.

Then $E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i \mu_i$.

3. Let $X_1, X_2, \dots, X_n$ be random variables which each have mean μ (i.e. $E(X_i) = \mu \; \forall \, i$).

Then, the sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ satisfies

$$E(\bar{X}) = \mu.$$

Proof:

Results for Covariance:

1. $Cov(X,X) = Var(X)$

2. Consider $Cov(aX + bY, cU + dV)$

3. More generally, we have

$$Cov\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j Cov(X_i, Y_j).$$

Results for Variance:

1. Variance of a linear combination of 2 random variables:

$$Var(aX + bY)$$
$$= a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$$

Note: X and Y are not independent, so the Cov terms are needed.

2. Variance of a sum of 2 independent random variables:

Let X and Y be **independent** random variables.

Since $Cov(X, Y) = 0$, then

$$Var(X + Y) = Var(X) + Var(Y).$$

In a similar fashion, we also have

$$Var(X - Y) = Var(X) + Var(Y).$$

This second result seems a little counter-intuitive at first, but look more closely:

$$Var(X - Y) = Var(X) + (-1)^2 Var(Y)$$
$$= Var(X) + Var(Y)$$

Also, remember that Var cannot be negative!

3. Variance of a general linear combination:

For each $i = 1,2,\dots,n$, let a_i be a constant and $Var(X_i) = \sigma_i^2$. Then,

$$Var\left(\sum_{i=1}^{n} a_i X_i\right)$$

$$= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j\, Cov(X_i, X_j)$$

Again, the r.v.'s are not independent, so the Cov terms need to be included. The next slide will show the case for independent r.v.'s.

4. Variance of a linear combination of independent random variables:

a) If $X_1, X_2, \dots, X_n$ are independent random variables, then $Cov(X_i, X_j) = 0$, so that:

$$\boldsymbol{Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \sigma_i^2}$$

b) If $X_1, X_2, \dots, X_n$ are independent random variables and all have the **same** variance σ^2, then:

$$\boldsymbol{Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sigma^2 \sum_{i=1}^{n} a_i^2}$$

Remark: Using the previous results, we now know that if $X_1, X_2, \ldots, X_n$ are independent random variables with the **same** mean μ and **same** variance σ^2, then the sample mean $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ has

$$\boldsymbol{E(\bar{X}) = \mu} \text{ and } \boldsymbol{Var(\bar{X}) = \frac{\sigma^2}{n}}.$$

What does this tell us about $\bar{X}$?

1. If we consider $\mathrm{Var}(\bar{\mathrm{X}})$, we notice that it is a **scaled version** of $\mathrm{Var}(\mathrm{X_i}) = \sigma^2$.

In other words, $\bar{\mathrm{X}}$ is less variable as compared to $\mathrm{X_i}$.

Does this make sense?

2. We also note that $Var(\bar{X})$ will become smaller as the sample size, n gets larger. In other words, $\bar{X}$ becomes less variable as n gets larger.

This happens because as n increases (i.e. as we collect more data points in our sample), we are obtaining more information, so our sample average, $\bar{X}$, is becoming **more precise** in the sense that we have

$$\boldsymbol{Var(\bar{X}) \to 0 \; as \; n \to \infty.}$$

This implies that as $\boldsymbol{n \to \infty, \bar{X} \to \mu.}$

This is sometimes called the **"law of averages"** or the **"law of large numbers"**.

Note: There is no mention of what the specific distribution of $X_1, X_2, \dots, X_n$ is.

Linear Combinations of Independent Normal Random Variables (Section 9.6)

1. Let $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, where a and b are constant real numbers. Then,

$$\boldsymbol{Y \sim N(a\mu + b, a^2\sigma^2).}$$

2. Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent random variables. If a and b are constants, then

$$\boldsymbol{aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).}$$

In general, if $X_i \sim N(\mu_i, \sigma_i^2), i = 1,2, \ldots, n$, are independent random variables and $a_1, a_2, \ldots, a_n$ are constants, then

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

3. Let $X_1, X_2, \dots, X_n$ be independent $N(\mu, \sigma^2)$ random variables.

Then:

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

and

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

Example: Let $X \sim N(3,\ 5)$ and $Y \sim N(6,\ 14)$ be independent random variables.

Find $P(X > Y)$.

You Try:

Let X ~ N(5,4). Y ~ N(7,9). X and Y are independent random variables.

a) Find P(2*X* – Y > 4).

b) Find the minimum number, n, of independent observations needed on X so that $P(|\bar{X} - 5| < 0.1) \geq 0.98$.

a) We want P(2X – Y > 4).

Let W = 2X – Y. E(W) = 2(5) - 7 = 3. Var(Y) = 4(4) + 9 = 25. So, W ~ N(3, 25).

So, P(2X - Y > 4) = P(W > 4) = P(Z > $\frac{4-3}{5}$) = P(Z > 0.2) = 1 – 0.57926 = 0.42074

b) X ~ N(5,4). It follows that $\bar{X} \sim N(5, \frac{4}{n})$.

$$P(|\bar{X} - 5| < 0.1) \geq 0.98 \Rightarrow P\left(|Z| < \frac{0.1}{\sqrt{\frac{4}{n}}}\right) \geq 0.98$$

We need to determine the 0.99 quantile of the distribution of Z. From the quantile table, we see that z = 2.3263. Remember that 0.98 probability between these values implies that there is 0.01 in the left and right tails.

So, we have 2.3263 = $\frac{0.1\sqrt{n}}{2}$. We can solve this for n.

This gives us n = $\left(\frac{2 * 2.3263}{0.1}\right)^2 = 2{,}164.67$.

Therefore, a minimum of 2,165 independent observations are needed.

Indicator Random Variables (Section 9.7)

An indicator random variable is a binary variable (0 or 1) that indicates whether or not an event has occurred.

It can allow us to take more complicated scenarios and break them into simpler ones.

Example: Suppose that $X \sim Bin(n, p)$.

For $i = 1,2, \ldots, n$, define a new random variable X_i as follows:

$$X_i = \begin{cases} 0 & if\ i^{th}\ trial\ was\ a\ failure, \\ 1 & if\ i^{th}\ trial\ was\ a\ success. \end{cases}$$

Then:

Example: We have *N* letters to *N* different people, and *N* envelopes addressed to those *N* people. One letter is randomly put in each envelope.

Find the mean and variance of the number of letters placed in the correct envelope.

The results are fascinating! If we let X represent the number of correct matches, we showed that:

E(X) = 1 and Var(X) = 1 regardless of the value of N.

Exercise: Work this out numerically with N = 3. Consider the sequence A, B, C to be a perfect match (i.e. letter A is in envelope A, letter B is in envelope B, and letter C is in envelope C. So, the sequence A, C, B would only have 1 match (letter A is in envelope A).

There are 6 possible outcomes. List them, with their probabilities, then set up the pf of X, noting that it's not possible to get exactly 2 matches, as 2 matches guarantees that the third is a match!

You will see that E(X) and Var(X) = 1!