Expected Values and Variances of Sums and Averages of Random Variables

The expected value of a sum of random variables, say $S_n = \sum_{i=1}^{n} X_i$, is the sum of the expected values:

$$E(S_n) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

This is always true (as long as the *X*s have expected values). The *X*s **do not** have to be independent,

... or to have the same distribution,
... or to have the same expected values,
... or anything.

For example, if X_1 has mean μ_1 and X_2 has mean μ_2 then

$$E(X_1 + X_2)= \mu_1+ \mu_2 \quad \text{and} \quad E(X_1 - X_2)= \mu_1 - \mu_2$$

On the other hand,

The variance of a sum of **<u>independent</u>** random variables, $S_n = \sum_1^n X_i$ is the sum of the variances:

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

For example, if X_1 has variance σ_1^2 and

X_2 has variance σ_2^2 and

X_1 and X_2 are independent,

then Var($X_1 + X_2$) = $\sigma_1^2 + \sigma_2^2$ and

Var($X_1 - X_2$) = $\sigma_1^2 + \sigma_2^2$

- Why?

⇨ If $X_1, X_2, \ldots, X_n$ are independent random variables with common mean μ and variance, σ^2, then for their sum, S_n ,

$$E(S_n) = n\mu$$

$$\mathrm{Var}(S_n) = n\sigma^2$$

⇨ and their average, $\overline{X}_n = \frac{\sum_1^n X_i}{n} = \frac{S_n}{n}$,

$$E(\overline{X}_n) = \mu$$

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

- Why?

The variance of the mean from a sample of size n (i.e., observations on n independent random variables, all with the same probability distribution) is 1/n times the variance of a single observation.

The standard deviation of the mean, $\overline{X}_n$, is $\sigma/\sqrt{n}$

$$\text{SD}\left(\overline{X}_n\right) = \frac{\sigma}{\sqrt{n}}$$

⇨ NOTICE:

Many writers use the term "standard deviation" to refer to the square root of the variance for a single observation **only** (say σ).

For a mean, $\overline{X}_n$, they call the square root of the variance the "standard error." (say $\sigma/\sqrt{n}$)

That is, they call the standard deviation of $\overline{X}_n$ the "standard error." It equals $\sigma/\sqrt{n}$, regardless of which name is used.

$\text{SD}\left(X\right) = \sigma$ is the ‘standard deviation’

$\text{SD}\left(\overline{X}_n\right) = \dfrac{\sigma}{\sqrt{n}}$ is the ‘standard Error’

Question:

If $X_1, X_2, \ldots, X_m$ all have mean μ_{x} and variance σ_{x}^2,

and $Y_1, Y_2, \ldots, Y_n$ all have mean μ_{y} and variance σ_{y}^2,

and if all of the *X*s and *Y*s are independent, then what is the mean, variance, and standard deviation of $\overline{X}_m - \overline{Y}_n$?

Answer:

$$E(\overline{X}_m - \overline{Y}_n) = \mu_x - \mu_y$$

$$Var(\overline{X}_m - \overline{Y}_n) = Var(\overline{X}_m) + Var(\overline{Y}_n)$$

$$= \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$$

$$SD(\overline{X}_m - \overline{Y}_n) = \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}.$$

- Why?

Sums and Averages of Normal Random Variables

The above facts about expected values and variances of random variables are true for all probability distributions.

Additionally, there are two very important facts that are specific to normally distributed random variables. These are

⇨ **(1) If X ~ normal (μ,σ^2), then**

$$\mathbf{a}X + \mathbf{b} = \mathbf{Q} \sim \text{normal}\ (\,a\mu+b,\ a^2\sigma^2\,)$$

The mean and variance of the transformed random variable Q are just what we know they must be.

What is new in **(1)** is that the distribution of Q is also normal with that mean and variance.

⇨ **(2) If *X* and *Y* are normal R.V.s, then so is their sum, *X* + *Y* .**

So:

If Q=X+Y then Q ~ N(E(X)+E(Y), Var(X+Y))

Note that so far, we can only calculate Var(X+Y) when X and Y are independent. Nevertheless **(2)** holds even if X and Y are not independent.

Although we will not prove number **(2)** it is especially important because it tells us that the distribution of the sample average of normal RVs is itself normal.

So when we have normal random variables, $X_1, X_2, \ldots, X_n$, their sum and average must also have a normal distribution.

And when the *X*s and *Y*s are normal, the difference between the averages, $\overline{X} - \overline{Y}$, also has a normal distribution.

Example

If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ and $\overline{X}_n = \frac{\sum_1^n X_i}{n}$ then

(1) $E\left(\overline{X}_n\right) = E\left(\frac{\sum_1^n X_i}{n}\right) = \frac{1}{n} \sum_1^n E\left(X_i\right) = \mu$

Justification: property of expected values

(2) $Var\left(\overline{X}_n\right) = Var\left(\frac{\sum_1^n X_i}{n}\right) = \frac{1}{n^2} \sum_1^n Var\left(X_i\right) = \frac{\sigma^2}{n}$

Justification: property of variances for independent RVs

(3) $\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(E(X_i), \frac{Var(X_i)}{n}\right)$

Justification: numbers (1) and (2) above, sum of normal RVs is also normal.

Review

For any constants, a and b,

$$E(\,aX + b\,) \;=\; aE(\,X\,) + b$$
$$\mathrm{Var}(\,aX + b\,) \;=\; a^2\,\mathrm{Var}(\,X\,)$$

If $X_1, X_2, \ldots, X_n$ are independent random variables that have the same distribution with E(X) and Var(X), then for the average, $\overline{X}_n = \sum_1^n X_i \;/n$,

$$\mathrm{E}(\;\overline{X}_n\;) = E(X)$$

$$\mathrm{Var}(\;\overline{X}_n\;) = \frac{\mathrm{Var(X)}}{\mathrm{n}}$$

$$\mathrm{SE}(\;\overline{X}_n\;) = \sqrt{\frac{\mathrm{Var(X)}}{\mathrm{n}}}$$

This is true regardless of what probability distribution *X* has (Binomial, Normal, etc.).

For example, if *X* has the Bernoulli(θ) distribution (whose mean and variance are θ and θ(1-θ)), then

$$E(\overline{X}_n) = \theta$$

$$Var(\overline{X}_n) = \theta(1-\theta)/n$$

$$SE(\overline{X}_n) = \sqrt{\theta(1-\theta)/n}$$

In words, these last three results are:

(1) The mean (or average), $\overline{X}_n$, of a sample of size n has the same expected value as a single observation.

(2) The mean of a sample of size n, $\overline{X}_n$, has a variance, $\mathrm{Var}(\overline{X}_n)$, that is 1/n times the variance of a single observation.

(3) The mean of a sample of size n, $\overline{X}_n$, has a standard deviation (or standard error), $\mathrm{SE}(\overline{X}_n)$, that is $1/\sqrt{n}$ times the standard deviation of a single observation.

Covariance and Correlation

When two Random Variables, say X and Y, are not independent we can measure their dependence by assessing their covariance.

$$Cov(X,Y) = E\left[\left(X - E\left(X\right)\right)\left(Y - E\left(Y\right)\right)\right]$$

⇨ $Cov(X,X) = Var(X)$

The covariance measures the strength of the linear relationship between X and Y. It is directly related to a more familiar measure of dependence called the correlation between X and Y:

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

and $-1 \leq Corr(X,Y) \leq 1$

Simply put, correlation is scaled covariance. They are the same measure but on different scales.

Covariance is important because it allows us to account for the dependence between random variables. For any two Random Variables, X and Y,

$$\Rightarrow \quad \begin{cases} Var\left(X+Y\right)=Var\left(X\right)+Var\left(Y\right)+2Cov\left(X,Y\right) \\ Var\left(X-Y\right)=Var\left(X\right)+Var\left(Y\right)-2Cov\left(X,Y\right) \end{cases}$$

This is always true (as long as the *X*s have expected values). The *X*s **do not** have to be independent,

... or to have the same distribution,
... or to have the same expected values,
... or anything.

Important Note:

It is always true that if two RVs are independent, then their covariance is ZERO and so is their correlation.
(If X independent Y then Cov(X,Y)=0).

However, the reverse is **not** true!!
That is, zero covariance or zero correlation does **not** imply that two variables are independent.

Just like expected values and variances there are some general rules for calculating the covariance.

For any constants a,b,c,d, and RV's X and Y:

$$Cov(aX + c, bY + d) = abCov(X, Y)$$

and for RV's X,Y,Z,W

$$\begin{aligned} Cov(X + W, Y + Z) &= Cov(X, Y) + Cov(X, Z) \\ &\quad + Cov(W, Y) + Cov(W, Z) \end{aligned}$$

Example:

Suppose we have two samples X_1,X_2,…,X_n i.d. (identically distributed) and Y_1, Y_2,…,Y_m i.d. with Cov(X,Y)=$\rho\sigma_X\sigma_Y$ (why?). Then

$$Cov\left(\overline{X}_n, \overline{Y}_m\right) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{m}\sum_{j=1}^{m} Y_j\right) = \frac{1}{nm} Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right)$$

$$= \frac{1}{nm} Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} Cov\left(X_i, Y_j\right) = \rho\sigma_X\sigma_Y$$

Example:

Suppose we have two samples X_1,X_2,…,X_n i.d. (identically distributed) and Y_1, Y_2,…,Y_m i.d. with Cov(X,Y)=$\rho\sigma_x\sigma_y$ (where σ_x=sd(x) and σ_y=sd(Y)). Then

$$Var\left(\overline{X}_n - \overline{Y}_m\right) = Var\left(\overline{X}_n\right) + Var\left(\overline{Y}_m\right) - 2Cov\left(\overline{X}_n, \overline{Y}_m\right)$$

And we know:

$$Var\left(\overline{X}_n\right) = \frac{Var\,(X\,)}{n}$$
$$Var\left(\overline{Y}_m\right) = \frac{Var\,(Y\,)}{m}$$
$$Cov\left(\overline{X}_n, \overline{Y}_m\right) = \rho\sqrt{Var\,(X\,)}\sqrt{Var\,(Y\,)}$$

Now we have for the two means:

$$Var\left(\overline{X}_n - \overline{Y}_m\right) = \frac{Var\,(X\,)}{n} + \frac{Var\,(Y\,)}{m} - 2\rho\sqrt{Var\,(X\,)}\sqrt{Var\,(Y\,)}$$

This formula has particular use when the X's are baseline measurements and the Y's represent a measurement on the same subject after treatment is applied.