5 Joint Probability Distributions and Random Samples

5.5 The Distribution of a Linear Combination

The Distribution of a Linear Combination

The sample mean $\overline{X}$ and sample total T_o are special cases of a type of random variable that arises very frequently in statistical applications.

Definition

Given a collection of n random variables $X_1, \ldots, X_n$ and n numerical constants $a_1, \ldots, a_n$, the rv

$$Y = a_1X_1 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i \qquad \textbf{(5.7)}$$

is called a **linear combination** of the X_i's.

3

The Distribution of a Linear Combination

For example, $4X_1 - 5X_2 + 8X_3$ is a linear combination of X_1, X_2, and X_3 with $a_1 = 4$, $a_2 = -5$, and $a_3 = 8$.

Taking $a_1 = a_2 = \ldots = a_n = 1$ gives $Y = X_1 + \ldots + X_n = T_o$, and $a_1 = a_2 = \ldots = a_n = \frac{1}{n}$ yields

$$Y = \frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}T_o = \overline{X}$$

Notice that we are not requiring the X_i's to be independent or identically distributed. All the X_i's could have different distributions and therefore different mean values and variances. We first consider the expected value and variance of a linear combination.

4

The Distribution of a Linear Combination

Proposition

Let $X_1, X_2, \ldots, X_n$ have mean values $\mu_1, \ldots, \mu_n$, respectively, and variances $\sigma_1^2, \ldots, \sigma_n^2$, respectively.

1. Whether or not the X_i's are independent,

$$E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$$
$$= a_1\mu_1 + \ldots + a_n\mu_n \qquad \textbf{(5.8)}$$

2. If $X_1, \ldots, X_n$ are independent,

$$V(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n)$$
$$= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2 \qquad \textbf{(5.9)}$$

The Distribution of a Linear Combination

And

$$\sigma_{a_1X_1+\cdots+a_nX_n} = \sqrt{a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2} \qquad \textbf{(5.10)}$$

3. For any $X_1, \ldots, X_n$,

$$V(a_1X_1 + \cdots + a_nX_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\text{Cov}(X_i, X_j) \qquad \textbf{(5.11)}$$

The Distribution of a Linear Combination

Proofs are sketched out at the end of the section. A paraphrase of (5.8) is that the expected value of a linear combination is the same as the linear combination of the expected values—for example, $E(2X_1 + 5X_2) = 2\mu_1 + 5\mu_2$.

The result (5.9) in Statement 2 is a special case of (5.11) in Statement 3; when the X_i's are independent, $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$ and $= V(X_i)$ for $i = j$ (this simplification actually occurs when the X_i's are uncorrelated, a weaker condition than independence).

Specializing to the case of a random sample (X_i's iid) with $a_i = 1/n$ for every i gives $E(\overline{X}) = \mu$ and $V(\overline{X}) = \sigma^2/n$. A similar comment applies to the rules for T_o.

Example 29

A gas station sells three grades of gasoline: regular, extra, and super.

These are priced at \$3.00, \$3.20, and \$3.40 per gallon, respectively.

Let X_1, X_2, and X_3 denote the amounts of these grades purchased (gallons) on a particular day.

Suppose the X_i's are independent with $\mu_1 = 1000$, $\mu_2 = 500$, $\mu_3 = 300$, $\sigma_1 = 100$, $\sigma_2 = 80$, and $\sigma_3 = 50$.

Example 29

cont'd

The revenue from sales is $Y = 3.0X_1 + 3.2X_2 + 3.4X_3$, and

$$E(Y) = 3.0\mu_1 + 3.2\mu_2 + 3.4\mu_3$$

$$= \$5620$$

$$V(Y) = (3.0)^2\sigma_1^2 + (3.2)^2\sigma_2^2 + (3.4)^2\sigma_3^2$$

$$= 184{,}436$$

$$\sigma_Y = \sqrt{184{,}436}$$

$$= \$429.46$$

9

The Difference Between Two Random Variables

10

The Difference Between Two Random Variables

An important special case of a linear combination results from taking $n = 2$, $a_1 = 1$, and $a_2 = -1$:

$$Y = a_1X_1 + a_2X_2 = X_1 - X_2$$

We then have the following corollary to the proposition.

Corollary

$E(X_1 - X_2) = E(X_1) - E(X_2)$ for any two rv's X_1 and X_2.
$V(X_1 - X_2) = V(X_1) + V(X_2)$ if X_1 and X_2 are independent rv's.

11

The Difference Between Two Random Variables

The expected value of a difference is the difference of the two expected values, but the variance of a difference between two independent variables is the *sum*, *not* the difference, of the two variances.

There is just as much variability in $X_1 - X_2$ as in $X_1 + X_2$ [writing $X_1 - X_2 = X_1 + (-1)X_2$, $(-1)X_2$ has the same amount of variability as X_2 itself].

12

Example 30

A certain automobile manufacturer equips a particular model with either a six-cylinder engine or a four-cylinder engine.

Let X_1 and X_2 be fuel efficiencies for independently and randomly selected six-cylinder and four-cylinder cars, respectively. With $\mu_1 = 22$, $\mu_2 = 26$, $\sigma_1 = 1.2$, and $\sigma_2 = 1.5$,

$$E(X_1 - X_2) = \mu_1 - \mu_2$$

$$= 22 - 26$$

$$= -4$$

Example 30

cont'd

$$V(X_1 - X_2) = \sigma_1^2 + \sigma_2^2$$
$$= (1.2)^2 + (1.5)^2$$
$$= 3.69$$
$$\sigma_{X_1 - X_2} = \sqrt{3.69}$$
$$= 1.92$$

If we relabel so that X_1 refers to the four-cylinder car, then $E(X_1 - X_2) = 4$, but the variance of the difference is still 3.69.

The Case of Normal Random Variables

The Case of Normal Random Variables

When the X_i's form a random sample from a normal distribution, $\overline{X}$ and T_o are both normally distributed. Here is a more general result concerning linear combinations.

Proposition

If $X_1, X_2, \ldots, X_n$ are independent, normally distributed rv's (with possibly different means and/or variances), then any linear combination of the X_i's also has a normal distribution. In particular, the difference $X_1 - X_2$ between two independent, normally distributed variables is itself normally distributed.

Example 31

The total revenue from the sale of the three grades of gasoline on a particular day was $Y = 3.0X_1 + 3.2X_2 + 3.4X_3$, and we calculated $\mu_Y = 5620$ and (assuming independence) $\sigma_Y = 429.46$. If the X_is are normally distributed, the probability that revenue exceeds 4500 is

$$\begin{aligned} P(Y > 4500) &= P\left(Z > \frac{4500 - 5620}{429.46}\right) \\ &= P(Z > -2.61) \\ &= 1 - \Phi(-2.61) \\ &= .9955 \end{aligned}$$

17

The Case of Normal Random Variables

The CLT can also be generalized so it applies to certain linear combinations. Roughly speaking, if n is large and no individual term is likely to contribute too much to the overall value, then Y has approximately a normal distribution.

18