

**Stat 321 - Day 27**  
**Joint Probability Distributions, Covariance, Linear Combinations**

The joint probability distribution of a pair of *discrete* random variables  $X$  and  $Y$  is represented by a *joint probability mass function*  $p(x,y)$  that reports  $P(X=x, Y=y)$ . This function can also be represented in a joint probability table.

**Example: Random Letters**

Suppose that three letters are to be chosen at random from the word “statistics.” Let the random variable  $X = \#$  of  $t$ 's chosen and the random variable  $Y = \#$  of  $i$ 's chosen.

- (a) Determine the probability of getting two  $t$ 's and two  $i$ 's.
- (b) Determine the probability of getting two  $t$ 's and one  $i$ . [*Hint*: The denominator is the total number of ways to choose three letters from the ten. The numerator is the number of ways to get two  $t$ 's and one  $i$ . Remember how to work with combinations.]
- (c) Determine the probability of getting exactly one  $t$  and one  $i$ .

The following table reveals the joint pmf of  $(X,Y)$ :

$p(x,y)$	$y=0$	$y=1$	$y=2$
$x=0$	10/120	20/120	5/120
$x=1$	30/120		3/120
$x=2$	15/120		
$x=3$	1/120	0	0

- (d) Fill in the missing probabilities in the table.
- (e) What is the probability that the same number of each letter is chosen? (Also express this probability in terms of the random variables  $X$  and  $Y$ .)
- (f) What is the probability that there are more  $t$ 's than  $i$ 's chosen?
- (g) What is the probability that there zero  $t$ 's are chosen? Repeat for one  $t$  and then for two  $t$ 's and then for three  $t$ 's.

You have found that the *marginal pmf* is found by summing the joint pmf over the other variable.

- (h) What is the conditional probability that zero  $t$ 's are chosen given that zero  $i$ 's are chosen?

- (i) Would you say that  $X$  and  $Y$  are *independent*? Explain. [Hint: Base your answer on the probabilities in (g) and (h).]

Two discrete random variables are said to be *independent* if their joint pmf is equal to the product of their marginal pmf's for *all* pairs of possible values.

The expected value of a function of a pair of discrete random variables is found as you would expect:  $E[h(X,Y)] = \sum_x \sum_y [h(x,y)p(x,y)]$ . The *covariance* between two random variables is defined by  $\text{Cov}(X,Y) = E\{[X-E(X)][Y-E(Y)]\}$ , which can be shown to equal  $E(XY)-E(X)E(Y)$ . The *correlation coefficient* is the covariance divided by the product of the standard deviations, which forces the correlation to be between -1 and +1 (inclusive).

- (j) One can show that the expected number of  $t$ 's to be chosen is  $E(X) = 0.9$ , and the expected number of  $i$ 's to be chosen is  $E(Y) = 0.6$ . Use this information to calculate the covariance between  $X$  and  $Y$ , using the short-cut formula.
- (k) One can further show that the variance of the number of  $t$ 's chosen is  $E(X^2) = 1.30$ , and the variance of the number of  $i$ 's chosen is  $E(Y^2) = 11/15$ . Use this information to calculate the correlation coefficient between  $X$  and  $Y$ .
- (l) Judging from the context and from the joint probability table, explain why it makes sense that the correlation turns out to be negative.

In the continuous case, the joint distribution of two random variables is described by a *joint probability density function*  $f(x,y)$ , which has the property that  $P(a \leq X \leq b, c \leq Y \leq d) =$

$\int_a^b \int_c^d f(x,y) dy dx$ . In other words, the probability of a region corresponds to the *volume* under the joint pdf over that region. To be a legitimate joint pdf,  $f(x,y)$  must be nonnegative, and its double integral over the real plane must equal one. The *marginal* pdf of  $X$  is found from the joint pdf as:  $f_X(x) = \int_{-\infty}^{\infty} f(x,y) dy$ , and the marginal pdf of  $Y$  as:  $f_Y(y) = \int_{-\infty}^{\infty} f(x,y) dx$ . If  $X$  and  $Y$  are independent, then the joint pdf is the product of the marginal pdf's.

**Linear Combinations:**

Consider random variables  $X_1, X_2, \dots, X_n$ , and consider the *linear combination*  
 $Y = a_1X_1 + a_2X_2 + \dots + a_nX_n$ . Some results about expectations of this linear combination are:

1.  $E(Y) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n)$
2. If  $X_1, X_2, \dots, X_n$  are independent, then  $V(Y) = a_1^2V(X_1) + a_2^2V(X_2) + \dots + a_n^2V(X_n)$

Now suppose that  $X_1, X_2, \dots, X_n$  are independent with common mean  $E(X_i)=\mu$  and common variance  $V(X_i)=\sigma^2$ . Consider the random variable  $T = X_1 + X_2 + \dots + X_n$ .

(m) What do these results say about  $E(T)$ ?

(n) What do these results say about  $V(T)$ ?

(o) What do these results say about  $SD(T)$ ?

Now suppose that  $X$  and  $Y$  are independent with  $E(X) = \mu_x$ ,  $E(Y) = \mu_y$ ,  $V(X) = \sigma_x$ , and  $V(Y) = \sigma_y$ . Consider the random variables sum  $S=X+Y$  and difference  $D=X-Y$ .

(p) What do these results say about  $E(S)$ ? What about  $E(D)$ ?

(q) What do these results say about  $V(S)$ ? What about  $V(D)$ ?

(r) What do these results say about  $SD(S)$ ? What about  $SD(D)$ ?

**Example: Exam Scores**

Suppose that students' scores on a midterm exam follow a normal distribution with mean 75 and standard deviation 8, while scores on the final exam independently follow a normal distribution with mean 70 and standard deviation 12.

(s) What can you say about the mean and standard deviation of the combined exam scores?

(t) Use simulation to confirm that your answers to (s) are reasonable:

```
MTB> random 1000 c1;  
SUBC> normal 75 8.  
MTB> random 1000 c2;  
SUBC> normal 70 12.  
MTB> let c3=c1+c2  
MTB> name c1 'midterm' c2 'final' c3 'combined'  
MTB> describe c3
```

Do the mean and standard deviation of these simulated combined exam scores come close to the theoretical values that you calculated in (s)?

(u) Is the probability distribution of the combined exam scores normal? Investigate this question by looking at visual displays of the simulated combined exam scores. Does this distribution appear to be normal? Explain.

It turns out that any linear combination of normally distributed random variables follows a normal distribution.

(v) Use this result to determine the probability that the combined score exceeds 160.

(w) Use this result to determine the probability that a student scores higher on the final exam than on the midterm exam. [*Hint: Work with the difference between the two scores.*]