Chapter 4: Pairs of Random Variables

M. George Akritas
Introduction

The Joint Probability Mass Function

Marginal and Conditional Probability Mass Functions

Independence

Mean Value of Functions of Random Variables

Expected Value of Sums

Variance of Sums

Quantifying Dependence: Pearson’s Linear Correlation Coefficient

The Bivariate Normal and Multinomial Distributions

Hierarchical Models

Regression Models

The Simple Linear Regression Model

Multinomial Distribution
When we record two characteristics from each population unit, the outcome variable is a pair of random variables.

- \( Y \) = age of a tree, \( X \) = the tree’s diameter at breast height.
- \( Y \) takes the value 0, or 1 if a child survives a car accident, or not. \( X \) takes the value 0, 1, or 2, if the child uses no seat belt, adult seat belt, or child seat.
- \( Y \) = propagation of an ultrasonic wave, \( X \) = tensile strength of substance.

Of interest here is not only each variable separately, but also to a) quantify the relationship between them, and b) be able to predict one from another. The *joint distribution* forms the basis for doing so.
The joint probability mass function of two discrete random variables, \( X \) and \( Y \), is denoted as \( p(x, y) \) and is defined as

\[
p(x, y) = P(X = x, Y = y)
\]

If \( X \) and \( Y \) take on only a few values, the joint pmf is typically given in a table like:

<table>
<thead>
<tr>
<th>( p(x, y) )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.034</td>
</tr>
<tr>
<td>2</td>
<td>0.066</td>
</tr>
<tr>
<td>3</td>
<td>0.100</td>
</tr>
</tbody>
</table>

The Axioms of probability imply that \( \sum_{i} p(x_i, y_i) = 1. \)
Chapter 4: Pairs of Random Variables

- The Joint Probability Mass Function
- Mean Value of Functions of Random Variables
- Quantifying Dependence: Pearson’s Linear Correlation Coefficient
- The Bivariate Normal and Multinomial Distributions
- Marginal and Conditional Probability Mass Functions
- Independence
Example

The joint pmf of $X =$ amount of drug administered to a randomly selected rat, and $Y =$ the number of tumors the rat develops, is:

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>$y$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0.0 mg/kg</td>
<td>.388</td>
<td>.009</td>
<td>.003</td>
</tr>
<tr>
<td>1.0 mg/kg</td>
<td>.485</td>
<td>.010</td>
<td>.005</td>
</tr>
<tr>
<td>2.0 mg/kg</td>
<td>.090</td>
<td>.008</td>
<td>.002</td>
</tr>
</tbody>
</table>

Thus 48.5% of the rats will receive the 1.0 mg dose and will develop 0 tumors, while 40% of the rats will receive the 0.0 mg dose.
The individual pmfs of $X$ and $Y$ are called the **marginal** pmfs.

We saw that $p_X(0) = P(X = 0)$ is obtained by summing the probabilities in the first row. Similarly, for $p_X(1)$ and $p_X(2)$.

Finally, the marginal pmf of $Y$ is obtained by summing the columns.

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>0 mg/kg</td>
<td>.388</td>
</tr>
</tbody>
</table>

$X$
The individual pmfs of $X$ and $Y$ are called the **marginal** pmfs.

We saw that $p_X(0) = P(X = 0)$ is obtained by summing the probabilities in the first row. Similarly, for $p_X(1)$ and $p_X(2)$.

Finally, the marginal pmf of $Y$ is obtained by summing the columns.

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>$y$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 \text{ mg/kg}$</td>
<td>$0$</td>
<td>.388</td>
<td>.009</td>
<td>.003</td>
<td>.400</td>
</tr>
<tr>
<td>$x$</td>
<td>$1 \text{ mg/kg}$</td>
<td>.485</td>
<td>.010</td>
<td>.005</td>
<td>.500</td>
</tr>
<tr>
<td>$2 \text{ mg/kg}$</td>
<td>.090</td>
<td>.008</td>
<td>.002</td>
<td>.100</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>.963</td>
<td>.027</td>
<td>.010</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Formulae: $p_X(x) = \sum_y p(x, y)$, $p_Y(y) = \sum_x p(x, y)$.
The concept of a *conditional distribution* provides answers to questions regarding the value of one variable, say \( Y \), given that the value of the other, say \( X \), has been observed.

From the definition of conditional probability, we have

\[
P(Y = y | X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p(x, y)}{p_X(x)},
\]

When we think of \( P(Y = y | X = x) \) as a function of \( y \) with \( x \) being kept fixed, we call it the *conditional* pmf of \( Y \) given that \( X = x \), and is denoted by \( p_{Y|X}(y|x) \) or \( p_{Y|X=x}(y) \).
When the joint pmf of \((X, Y)\) is given in a table form, \(p_{Y|X=x}(y)\) is found by dividing the joint probabilities in the row that corresponds to the \(X\)-value \(x\) by the marginal probability that \(X = x\).

**Example**

Find the conditional pmf of the number of tumors when the dosage is 0 mg and when the dosage is 2 mg.

**Solution:**

\[
\begin{array}{c|ccc}
  y & 0 & 1 & 2 \\
\hline
  p_{Y|X}(y|X = 0) & 0.388/0.4 = 0.97 & 0.009/0.4 = 0.0225 & 0.003/0.4 = 0.0075 \\
  p_{Y|X}(y|X = 2) & 0.090/0.1 = 0.9 & 0.008/0.1 = 0.08 & 0.002/0.1 = 0.02 \\
\end{array}
\]
Marginal and Conditional PMFs through R

\[ p_{XY} = (0.388, 0.485, 0.090, 0.009, 0.010, 0.008, 0.003, 0.005, 0.002) \] #
the vector of joint probabilities from the rat-drug example

\[ p_{XYm} = \text{matrix}(p_{XY}, 3, 3) \] #
the matrix of joint probabilities

\[ px = \text{rowSums}(p_{XYm}); \quad py = \text{colSums}(p_{XYm}) \] #
the marginal PMFs

\[ \text{diag}(1/px) \% \% p_{XYm} \] #
the matrix \( p_{Y|X=x}(y) \)

\[ p_{XYm}[1,]/px[1] \] #
the first row of \( p_{Y|X=x}(y) \)

\[ p_{XYm} \% \% \text{diag}(1/py) \] #
the matrix \( p_{X|Y=y}(x) \)

\[ p_{XYm}[1,]/py[1] \] #
the first column of \( p_{X|Y=y}(x) \)
Proposition

1. The joint pmf of \((X, Y)\) can be obtained as

\[
p(x, y) = p_{Y|X}(y|x)p_X(x) \quad \text{Multiplication rule for joint probabilities}
\]

2. The marginal pmf of \(Y\) can be obtained as

\[
p_Y(y) = \sum_{x \in S_X} p_{Y|X}(y|x)p_X(x) \quad \text{Law of Total Probability for marginal PMFs}
\]
The conditional pmf is a proper pmf. Thus,

\[ p_{Y|X}(y_j|x) \geq 0, \quad \text{for all } j = 1, 2, \ldots, \quad \text{and } \sum_j p_{Y|X}(y_j|x) = 1. \]

The **conditional expected value** of \( Y \) given \( X = x \) is the mean of the conditional pmf of \( Y \) given \( X = x \). It is denoted by \( E(Y|X = x) \) or \( \mu_{Y|X=x} \) or \( \mu_{Y|X}(x) \).

As a function in \( x \), \( \mu_{Y|X}(x) \) is called the **regression function** of \( Y \) on \( X \).

The **conditional variance** of \( Y \) given \( X = x \) is the variance of the conditional pmf of \( Y \) given \( X = x \). It is denoted by \( \sigma^2_{Y|X=x} \) or \( \sigma^2_{Y|X}(x) \).
Example

Find the conditional expected value and variance of the number of tumors when $X = 2$.

**Solution.** Using the conditional pmf that we found before, we have,

$$E(Y|X = 2) = 0 \times (.9) + 1 \times (.08) + 2 \times (.02) = .12.$$ 

Compare this with $E(Y) = .047$. Next,

$$E(Y^2|X = 2) = 0 \times (.9) + 1 \times (.08) + 2^2 \times (.02) = .16.$$ 

so that $\sigma^2_{Y|X}(2) = .16 - .12^2 = .1456$
Example

In the example where $X =$ amount of drug administered and $Y =$ number of tumors developed we have

<table>
<thead>
<tr>
<th>$y$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{Y</td>
<td>X=0}(y)$</td>
<td>.97</td>
<td>.0225</td>
</tr>
<tr>
<td>$p_{Y</td>
<td>X=1}(y)$</td>
<td>.97</td>
<td>.02</td>
</tr>
<tr>
<td>$p_{Y</td>
<td>X=2}(y)$</td>
<td>.9</td>
<td>.08</td>
</tr>
</tbody>
</table>

Find the regression function of $Y$ on $X$.

**Solution:** Here, $E(Y|X = 0) = 0.0375$, $E(Y|X = 1) = 0.04$, $E(Y|X = 2) = 0.12$. Thus the regression function of $Y$ on $X$ is

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{Y</td>
<td>X}(x)$</td>
<td>0.0375</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Proposition (Law of Total Probability for Expectations)

\[ E(Y) = \sum_{x \in S_X} E(Y|X = x)p_X(x). \]

Example

Use the regression function obtained in the previous example, and the pmf of \( X \), which is \( x \) \begin{tabular}{c|ccc}
\( p_X(x) \) & 0 & 1 & 2 \\
\hline
0.4 & 0.5 & 0.1
\end{tabular}, to find \( \mu_Y \).

Solution: \( 0.0375 \times 0.4 + 0.04 \times 0.5 + 0.12 \times 0.1 = 0.047 \)
The discrete r.v.s $X, Y$ are called **independent** if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \text{ for all } x, y.$$

The r.v.s $X_1, X_2, \ldots, X_n$ are called **independent** if

$$p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \cdots p_{X_n}(x_n).$$

If $X_1, X_2, \ldots, X_n$ are independent and also have the same distribution they are called **independent and identically distributed**, or **iid** for short.
Example

Let $X = 1$ or 0, according to whether component $A$ works or not, and $Y = 1$ or 0, according to whether component $B$ works or not. From their repair history it is known that the joint pmf of $(X, Y)$ is

<table>
<thead>
<tr>
<th></th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$X$</td>
<td>0.0098</td>
</tr>
<tr>
<td></td>
<td>0.98</td>
</tr>
<tr>
<td>1</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
</tr>
</tbody>
</table>

Are $X$, $Y$ independent?
The joint PMF of independent RVs in R

For \( X \sim Bin(2, 0.5) \), \( Y \sim Bin(3, 0.5) \) independent, do:

\[
p_{XY} = \text{dbinom}(0:2, 2, 0.5) \times\text{t(dbinom}(0:3, 3, 0.5))
\]

Alternatively:

\[
P = \text{expand.grid}(px=\text{dbinom}(0:2, 2, 0.5), py=\text{dbinom}(0:3, 3, 0.5));
\]

\[
P_{pxy} = P_{px} \times P_{py}; \quad P_{XY} = \text{matrix}(P_{pxy}, 3, 4)
\]

Check: \( P_{pxy} \) same as \( P_{XY} = \text{as.vector}(P_{XY}) \)

Also check:

\[
px = \text{rowSums}(p_{XY}) \quad \text{and} \quad py = \text{colSums}(p_{XY}) \quad \text{same as} \quad \text{dbinom}(0:2, 2, 0.5) \text{ and } \text{dbinom}(0:3, 3, 0.5), \text{ respectively.}
\]
Proposition

1. If $X$ and $Y$ are independent, so are $g(X)$ and $h(Y)$ for any functions $g$, $h$.

2. If $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent, so are $g(X_1, \ldots, X_{n_1})$ and $h(Y_1, \ldots, Y_{n_2})$ for any functions $g$, $h$.

Example

Consider the two-component system of the previous example and suppose that the failure of component $A$ incurs a cost of $500.00, while the failure of component $B$ incurs a cost of $750.00. Let $C_A$ and $C_B$ be the costs incurred by the failures of components $A$ and $B$ respectively. Are $C_A$ and $C_B$ independent?
Example

You and a friend each flip a fair coin. Let $X$ be the number of heads you get in the $n_1$ flips, and $Y$ be the number of heads your friend gets in $n_2$ flips.

1. Are $X$ and $Y$ independent?
2. What is the distribution of the total number of heads $X + Y$?

- The pmf of the sum of two independent RVs is called the convolution of their PMFs.
- The convolution is not always easy to obtain analytically. For example take $X \sim Bin(n_1, p_1)$ and $Y \sim Bin(n_2, p_2)$.
- The R commands given in the next slide compute the pmf of $X + Y$ from their joint distribution.
Let $X \sim Bin(3, 0.3)$ and $Y \sim Bin(4, 0.6)$, $X$, $Y$ independent.

\[
P = \text{expand.grid}(px = \text{dbinom}(0:3, 3, 0.3), py = \text{dbinom}(0:4, 4, 0.6)));
P$pxy = P.px \times P.py
\]

\[
G = \text{expand.grid}(X = 0:3, Y = 0:4); P$X = G$X; P$Y = G$Y; \text{attach}(P)
\]

\[
\text{sum}(pxy[\text{which}(X+Y == 4)]) \# \text{pmf of } X + Y \text{ at } x = 4
\]

\[
\text{sum}(pxy[\text{which}(X+Y < 4)]) \# \text{cdf of } X + Y \text{ at } x = 4
\]

\[
\text{sum}(pxy[\text{which}(3 < X+Y \& X+Y < 5)]) \# P(3 < X + Y \leq 5)
\]

Try: \[v = \text{seq}(0, 20, 2); v[5]; \text{which}(v == 8); \text{which}(2 < v \& v < 8); \text{sum}(v[2:5]); \text{sum}(v[\text{which}(2 < v \& v < 8)])]
Example

Let $X$ and $Y$ be independent random variables, both having the $\text{Bin}(n, p)$ distribution. Find the conditional pmf of $X$ given that $X + Y = m$.

For $X \sim \text{Bin}(3, 0.3)$ and $Y \sim \text{Bin}(4, 0.6)$, given $X + Y = 6$, the possible values of $X$, which are 3 and 2, are obtained from $X[\text{which}(X+Y==6)]$, and the conditional pmf, $P(X = 3|X + Y = 6) = 0.57$ and $P(X = 2|X + Y = 6) = 0.43$, is obtained from

$$p_x[\text{which}(X+Y==6)]*p_y[\text{which}(X+Y==6)]/\text{sum}(p_{xy}[\text{which}(X+Y==6)])$$

Example

If $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ are independent random variables, find the distribution of $Z = X + Y$. 
Proposition

Each of the following statements implies, and is implied by, the independence of $X$ and $Y$.

1. $p_{Y|X}(y|x) = p_Y(y)$.

2. $p_{Y|X}(y|x)$ does not depend on $x$, i.e. is the same for all possible values of $X$.

3. $p_{X|Y}(x|y) = p_X(x)$.

4. $p_{X|Y}(x|y)$ does not depend on $y$, i.e. is the same for all possible values of $Y$. 
Verification of the proposition with R

Use the joint PMF of $X \sim Bin(2, 0.5)$ and $Y \sim Bin(3, 0.5)$, $X$, $Y$ independent.

$p_{XY}[1,]/\text{rowSums}(p_{XY})[1]$ # gives $P(Y = y|X = 0)$, $y = 0:3$

Check that this is the same as

dbinom(0:3,3,.5) # why?

Next verify that the following are the same:

$p_{XY}[2,]/\text{rowSums}(p_{XY})[2]$ # gives $P(Y = y|X = 1)$, $y = 0:3$

$p_{XY}[3,]/\text{rowSums}(p_{XY})[3]$ # gives $P(Y = y|X = 2)$, $y = 0:3$
Example

$X$ takes the value 0, 1, or 2 if a child under 5 uses no seat belt, uses adult seat belt, or uses child seat. $Y$ takes the value 0, or 1 if a child survives car accident, or not. Suppose that

<table>
<thead>
<tr>
<th>$y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{Y</td>
<td>X=0}(y)$</td>
<td>0.69</td>
</tr>
<tr>
<td>$p_{Y</td>
<td>X=1}(y)$</td>
<td>0.85</td>
</tr>
<tr>
<td>$p_{Y</td>
<td>X=2}(y)$</td>
<td>0.84</td>
</tr>
</tbody>
</table>

and

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{X}(x)$</td>
<td>0.54</td>
<td>0.17</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Are $X$ and $Y$ independent?
Example (Continued)

Here \( p_{Y|X=x}(y) \) depends on \( x \) and so \( X, Y \) are not independent.

Alternatively, the joint pmf of \( X, Y \) is

<table>
<thead>
<tr>
<th>( p(x, y) )</th>
<th>( y )</th>
<th>( 0 )</th>
<th>( 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>0.3726</td>
<td>0.1674</td>
<td></td>
</tr>
<tr>
<td>( x )</td>
<td>1</td>
<td>0.1445</td>
<td>0.0255</td>
</tr>
<tr>
<td>2</td>
<td>0.2436</td>
<td>0.0464</td>
<td></td>
</tr>
<tr>
<td>( p_Y(y) )</td>
<td>0.7607</td>
<td>0.2393</td>
<td></td>
</tr>
</tbody>
</table>

Here \( p_X(0)p_Y(0) = 0.54 \times 0.7607 = 0.4108 \neq p(0, 0) = 0.3726 \).
Thus \( X, Y \) are not independent.
Example

Show that if $X$ and $Y$ are independent then

$$E(Y|X = x) = E(Y).$$

Thus, if $X$, $Y$ are independent then the regression function of $Y$ on $X$ is a constant function.
As in the univariate case the expected value and, consequently, the variance of a function of random variables (statistic) can be obtained without having to first obtain its distribution.

**Proposition**

\[ E(h(X, Y)) = \sum_x \sum_y h(x, y)p_{X,Y}(x, y) \]

\[ \sigma^2_{h(X,Y)} = E[h^2(X, Y)] - [E(h(X, Y))]^2 \]
Example

Find $E(X + Y)$ and $\sigma^2_{X+Y}$ if the joint pmf of $(X, Y)$ is

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0  1  2</td>
</tr>
</tbody>
</table>
| 0         | 0.10 0.04 0.02 | (4.1)  
| 1         | 0.08 0.20 0.06 |
| 2         | 0.06 0.14 0.30 |

**Solution:** By the previous formula,

$$E(X + Y) = (0)0.1 + (1)0.04 + (2)0.02 + (1)0.08 + (2)0.2$$
$$+ (3)0.06 + (2)0.06 + (3)0.14 + (4)0.3 = 2.48.$$
Example (Continued)

\[ E[(X + Y)^2] = (0)0.1 + (1)0.04 + (2^2)0.02 + (1)0.08 + (2^2)0.2 \]
\[ + (3^2)0.06 + (2^2)0.06 + (3^2)0.14 + (4^2)0.3 = 7.84. \]

Thus \( \sigma_{X+Y}^2 = 7.84 - 2.48^2 = 1.69. \)

Example

Find \( E(\min\{X, Y\}) \) and \( \text{Var}(\min\{X, Y\}) \) if the joint pmf of \( (X, Y) \) is as in the previous example.
Mean value of a function in R

```r
pxy=c(0.10, 0.08, 0.06, 0.04, 0.20, 0.14, 0.02, 0.06, 0.30)
pxym=matrix(pxy,3,3)
px=rowSums(pxym) ; py=colSums(pxym)
P=expand.grid(px=px,py=py) ; P$pxy=pxy
G=expand.grid(x=0:2,y=0:2) ; P$x=G$x; P$y=G$y

attach(P) ; sum((x+y)*pxy)
sum(x*pxy) ; sum(y*pxy)
sum((x+y)**2*pxy)
sum(pmin(x,y)*pxy) ; sum(pmin(x,y)**2*pxy)

Also, $E(e^{X+Y}) = \text{sum}(\exp(x+y)*pxy)$ etc.
```
Proposition

If \( X \) and \( Y \) are independent, then

\[
E(g(X)h(Y)) = E(g(X))E(h(Y))
\]

holds for any functions \( g(x) \) and \( h(y) \).
h(X_1, \ldots, X_n) is a linear combination of X_1, \ldots, X_n if

\[ h(X_1, \ldots, X_n) = a_1 X_1 + \cdots + a_n X_n. \]

- \overline{X} is a linear combination with all \( a_i = 1/n \).
- \[ T = \sum_{i=1}^{n} X_i \] is a linear combination with all \( a_i = 1 \).

**Proposition**

Let \( X_1, \ldots, X_n \) be any r.v.s (i.e, discrete or continuous, independent or dependent), with \( E(X_i) = \mu_i \). Then

\[ E (a_1 X_1 + \cdots + a_n X_n) = a_1 \mu_1 + \cdots + a_n \mu_n. \]
Corollary

1. Let $X_1, X_2$ be any two r.v.s. Then

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

and

$$E(X_1 + X_2) = \mu_1 + \mu_2.$$

2. Let $X_1, \ldots, X_n$ be such that $E(X_i) = \mu$ for all $i$. Then

$$E(T) = n\mu, \quad E(\bar{X}) = \mu.$$

Example

Let $X, Y$ have the joint pmf of the first example of this lecture. Then, $E(X) = 1.34, E(Y) = 1.14$, and, according to the corollary,

$$E(X + Y) = 1.34 + 1.14 = 2.48,$$

as before.
Example

Tower 1 is constructed by stacking 30 segments of concrete vertically. Tower 2 is constructed similarly. The height, in inches, of a randomly selected segment is uniformly distributed in (35.5,36.5). Find a) the expected value of the height, $T_1$, of tower 1, b) the expected value of the height, $T_2$ of tower 2, and c) the expected value of the difference $T_1 - T_2$.

Solution: a) Let $X_1, \ldots, X_{30}$ denote the heights of the segments used in tower 1. Since the $X_i$ are iid with $E(X_i) = 36$,

$$E(T_1) = 30 \times 36 = 1080 \text{ inches.}$$

b) $E(T_2) = 1080 \text{ inches.}$

c) $E(T_1 - T_2) = E(T_1) - E(T_2) = 0.$
Example

Let $N$ be the number of accidents per month in an industrial complex, and $X_i$ the number of injuries caused by the $i$th accident. Suppose the $X_i$ are independent from $N$ and $E(X_i) = 1.5$ for all $i$. If $E(N) = 7$, find the expected number of injuries, $Y = \sum_{i=1}^{N} X_i$, in a month.

Solution: Using the Law of Total Probability for Expectations and the independence of the $X_i$ from $N$, we have,

$$E(Y) = \sum_{n \in S_N} E(Y|N = n)P(N = n) = \sum_{n \in S_N} 1.5nP(N = n)$$

$$= 1.5 \sum_{n \in S_N} nP(N = n) = 1.5E(N) = 10.5$$
The above example dealt with the expected value of the sum of a random number, $N$, of random variables when these variables are independent of $N$. In general we have:

**Proposition**

Suppose that $N$ is an integer valued random variable, and the random variables $X_i$ are independent from $N$ and have common mean value $\mu$. Then,

$$E \left( \sum_{i=1}^{N} X_i \right) = E(N)\mu$$

The expected value of a sum of a random number of random variables.
Example (Expected Value of the Binomial R.V.)

Let $X \sim \text{Bin}(n, p)$. Find $E(X)$.

Solution: Recall that $X = \sum_{i=1}^{n} X_i$, where each $X_i$ is a Bernoulli r.v. with probability of 1 equal to $p$. Since $E(X_i) = p$, for all $i$,

$$E(X) = np.$$ 

Example (Expected Value of the Negative Binomial R.V.)

Let $X \sim \text{NBin}(r, p)$. Thus, $X$ is the number of Bernoulli trials up to an including the $r$th 1. Find $E(X)$.

Solution: Recall that $X = \sum_{i=1}^{r} X_i$, where each $X_i$ is a geometric, or $\text{NBin}(1, p)$ r.v. Since $E(X_i) = 1/p$, for all $i$,

$$E(X) = r/p.$$
The variance of a sum is the sum of the variances only in the case of independent variables. Here is why:

\[
\text{Var}(X + Y) = E \left\{ [X + Y - E(X + Y)]^2 \right\} \\
= E \left\{ [(X - E(X)) + (Y - E(Y))]^2 \right\} \\
= \text{Var}(X) + \text{Var}(Y) + 2E[(X - E(X))(Y - E(Y))].
\]

If \(X\) and \(Y\) are independent then,

\[
E[(X - E(X))(Y - E(Y))] = E[(X - E(X))] E[(Y - E(Y))] = 0
\]
Thus, if $X$ and $Y$ are independent

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

But if $X$ and $Y$ are not independent (or only correlated)

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2E[(X - E(X))(Y - E(Y))].$$

- The quantity $E[(X - E(X))(Y - E(Y))]$ is called the covariance of $X$ and $Y$. A computational formula is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X\mu_Y$$
Example

Let $X =$ deductible in car insurance, and $Y =$ deductible in home insurance, of a randomly chosen home and car owner. Suppose that the joint pmf of $X$, $Y$ is

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>.20</td>
<td>.10</td>
<td>.20</td>
</tr>
<tr>
<td>250</td>
<td>.05</td>
<td>.15</td>
<td>.30</td>
</tr>
</tbody>
</table>

where the deductible amounts are in dollars. Find $\sigma_{XY}$. 
Example (Continued)

Solution: First,

\[ E(XY) = \sum_x \sum_y xy p(x, y) = 23,750, \]

and

\[ E(X) = \sum_x xp_X(x) = 175, \quad E(Y) = \sum_y yp_Y(y) = 125. \]

Thus,

\[ \sigma_{XY} = 23,750 - 175 \times 125 = 1875. \]
Example ($\sigma_{XY} = 0$ for dependent $X$, $Y$)

Find Cov$(X, Y)$, where $X$, $Y$ have joint pmf given by

<table>
<thead>
<tr>
<th></th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1/3</th>
<th>0</th>
<th>1/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>1/3</td>
<td>0</td>
<td>1/3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1/3</td>
<td>0</td>
<td>1/3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2/3</td>
<td>1/3</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

Solution: Since $E(X) = 0$, the computational formula gives Cov$(X, Y) = E(XY)$. However, the product $XY$ takes the value zero with probability 1. Thus, Cov$(X, Y) = E(XY) = 0$.\[\Box\]
Proposition

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
2. $\text{Cov}(X, X) = \text{Var}(X)$.
3. If $X, Y$ are independent, then $\text{Cov}(X, Y) = 0$.
4. $\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$, for any real numbers $a, b, c$ and $d$.
5. $\text{Cov}\left(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} \text{Cov}(X_i, Y_j)$. 
Example

Consider the example with the car and home insurance deductibles, but suppose that the deductible amounts are given in cents. Find the covariance of the two deductibles.

*Solution:* Let $X'$, $Y'$ denote the deductibles of a randomly chosen home and car owner in cents. If $X$, $Y$ denote the deductibles in dollars, then $X' = 100X$, $Y' = 100Y$. According to part 4 of the proposition,

$$\sigma_{X'Y'} = 100 \times 100 \times \sigma_{XY} = 18,750,000.$$
Proposition

1. For any rv’s $X_1, \ldots, X_n$,

$$\text{Var}(a_1X_1 + \cdots + a_nX_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_ia_j \text{Cov}(X_i, X_j).$$

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

$$\text{Var}(a_1X_1 + \cdots + a_nX_n) = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2,$$

where $\sigma_i^2 = \text{Var}(X_i)$. 
Corollary

1. If $X_1, X_2$ are independent,

$$Var(X_1 + X_2) = \sigma_1^2 + \sigma_2^2, \text{ and } Var(X_1 - X_2) = \sigma_1^2 + \sigma_2^2.$$  

2. Without independence,

$$Var(X_1 + X_2) = \sigma_1^2 + \sigma_2^2 + 2\text{Cov}(X_1, X_2)$$

$$Var(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 - 2\text{Cov}(X_1, X_2)$$

3. If $X_1, \ldots, X_n$ are independent and $\sigma_1 = \cdots = \sigma_n = \sigma$, then

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$
Example

Let $X$ denote the number of short fiction books sold at the State College, PA, location of the Barnes and Noble bookstore in a given week, and let $Y$ denote the corresponding number sold on line from State College residents. We are given that $E(X) = 14.80$, $E(Y) = 14.51$, $E(X^2) = 248.0$, $E(Y^2) = 240.7$, and $E(XY) = 209$. Find $\text{Var}(X + Y)$.

Solution: Since $\text{Var}(X) = 248.0 - 14.8^2 = 28.96$, $\text{Var}(Y) = 240.7 - 14.51^2 = 30.16$, and $\text{Cov}(X, Y) = 209 - (14.8)(14.51) = -5.75$, we have

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$= 28.96 + 30.16 - 2 \times 5.75 = 47.62.$$
Example

On the first day of a wine tasting event three randomly selected judges are to taste and rate a particular wine before tasting any other wine. On the second day the same three judges are to taste and rate the wine after tasting other wines. Let $X_1, X_2, X_3$ be the ratings, on a 100 point scale, in the first day, and $Y_1, Y_2, Y_3$ be the ratings on the second day. We are given that the variance of each $X_i$ is $\sigma^2_{X_i} = 9$, the variance of each $Y_i$ is $\sigma^2_{Y_i} = 4$, the covariance $\text{Cov}(X_i, Y_i) = 5$, for all $i = 1, 2, 3$, and $\text{Cov}(X_i, Y_j) = 0$ for all $i \neq j$. Find the variance of combined rating $\overline{X} + \overline{Y}$. 
Example (Continued)

**Solution.** First note that

\[ \text{Var}(\bar{X} + \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) + 2\text{Cov}(\bar{X}, \bar{Y}). \]

Next, with the information given, we have \( \text{Var}(\bar{X}) = 9/3 \) and \( \text{Var}(\bar{Y}) = 4/3 \). Finally, using the proposition we have

\[
\text{Cov}(\bar{X}, \bar{Y}) = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{1}{3} \cdot \frac{1}{3} \cdot \text{Cov}(X_i, Y_j) = \sum_{i=1}^{3} \frac{1}{3} \cdot \frac{1}{3} \cdot \text{Cov}(X_i, Y_i) = 3 \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot 5.
\]

Combining these partial results we obtain \( \text{Var}(\bar{X} + \bar{Y}) = 23/3 \).
Example (Variance of the Binomial RV and of $\hat{p}$)

Let $X \sim \text{Bin}(n, p)$. Find $\text{Var}(X)$, and $\text{Var}(\hat{p})$.

Solution: Use the expression

$$X = \sum_{i=1}^{n} X_i,$$

where the $X_i$ are iid Bernoulli($p$) RVs. Since $\text{Var}(X_i) = p(1 - p)$,

$$\text{Var}(X) = np(1 - p), \quad \text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{p(1 - p)}{n}.$$
Example (Variance of the Negative Binomial RV)

Let $X \sim \text{NBin}(r, p)$. Thus, $X$ is the number of Bernoulli trials up to an including the $r$th 1. Find $\text{Var}(X)$.

*Solution:* Use the expression

$$X = \sum_{i=1}^{r} X_i,$$

where the $X_i$ are iid Geometric($p$), or $\text{NBin}(1, p)$ RVs. Since $\text{Var}(X_i) = (1 - p)/p^2$,

$$\text{Var}(X) = \sum_{i=1}^{n} \sigma_{X_i}^2 = r(1 - p)/p^2$$
Proposition

Let $X_1, \ldots, X_n$ be a simple random sample from a population having variance $\sigma^2$. Then

$$E(S^2) = \sigma^2, \text{ where } \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$ 

The next slide gives a numerical verification of this proposition.
Proof by Simulation

- Random number generation can be used to provide numerical evidence in support of, or in contradiction to, certain probabilistic/statistical claims.

- Evidence for the above proposition can be obtained by generating a large number of samples (of any given size and from any distribution), compute the sample variance for each sample, and average the sample variances.

- We will use 10,000 samples of size 5 from the standard normal distribution. The R commands are:

```r
m = matrix(rnorm(50000), ncol=10000)
mean(apply(m, 2, var))
```
When two variables are not independent, it is of interest to qualify and quantify dependence.

- \( X, Y \) are **positively dependent** or **positively correlated** if "large" values of \( X \) are associated with "large" values of \( Y \), and "small" values of \( X \) are associated with "small" values of \( Y \).

- In the opposite case, \( X, Y \) are **negatively dependent** or **negatively correlated**.

- If the dependence is either positive or negative it is called **monotone**.
Proposition (Monotone dependence and regression function)

The dependence is monotone (positive or negative), if and only if the regression function, $\mu_{Y|X}(x) = E(Y|X = x)$, is monotone (increasing or decreasing) in $x$.

Example

1. $X =$ height and $Y =$ weight of a randomly selected adult male, are positively dependent.
   1.1 If $X =$ height and $Y =$ weight, then $\mu_{Y|X}(1.82) < \mu_{Y|X}(1.90)$.

2. $X =$ stress applied and $Y =$ time to failure, are negatively dependent.
   2.1 If $X =$ stress applied and $Y =$ time to failure, then $\mu_{Y|X}(10) > \mu_{Y|X}(20)$. 
The regression function, however, is not designed to measure the degree of dependence of $X$ and $Y$.

**Proposition (Monotone dependence and covariance)**

The monotone dependence is positive or negative, if and only if the covariance is positive or negative.

- This proposition DOES NOT say that if the covariance is positive then the dependence is positive.
- Positive covariance implies positive dependence only if we know that the dependence is monotone.
Intuition for the Covariance

Consider a population of \( N \) units and let \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\) denote the values of the bivariate characteristic of the \( N \) units.

Let \((X, Y)\) denote the bivariate characteristic of a randomly selected unit.

In this case, the covariance is computed as

\[
\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y),
\]

where \(\mu_X = \frac{1}{N} \sum_{i=1}^{N} x_i\) and \(\mu_Y = \frac{1}{N} \sum_{i=1}^{N} y_i\) are the marginal expected values of \(X\) and \(Y\).
Intuition for Covariance - Continued

- If $X$, $Y$ are positively dependent (think of $X = \text{height}$ and $Y = \text{weight}$), then
  \[(x_i - \mu_X)(y_i - \mu_Y)\] will be mostly positive.

- If $X$, $Y$ are negatively dependent (think of $X = \text{stress}$ and $Y = \text{time to failure}$), then
  \[(x_i - \mu_X)(y_i - \mu_Y)\] will be mostly negative.

- Therefore, $\sigma_{XY}$ will be positive or negative according to whether the dependence of $X$ and $Y$ is positive or negative.
Quantification of dependence should be scale-free!

- Proper quantification of dependence should be not depend on the units of measurement.
  - For example, a quantification of the dependence between height and weight should not depend on whether the units are meters and kilograms or feet and pounds.

- The scale-dependence of the covariance, implied by the property

\[
\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)
\]

makes it unsuitable as a measure of dependence.
Definition (Pearson’s (or linear) correlation coefficient)

\[ \text{Corr}(X, Y) \text{ or } \rho_{XY}, \text{ of } X \text{ and } Y \text{ is defined as} \]

\[ \rho_{XY} = \text{Corr}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \]

Proposition (Properties of Correlation)

1. \text{If } ac > 0, \text{ then } Corr(aX + b, cY + d) = Corr(X, Y).
2. \text{ } -1 \leq \rho(X, Y) \leq 1.
3. \text{If } X, Y \text{ are independent, then } \rho_{XY} = 0.
4. \rho_{XY} = 1 \text{ or } -1 \text{ if and only if } Y = aX + b, \text{ for some constants } a, b.
Example

Find the correlation coefficient of the deductibles in car and home insurance of a randomly chosen car and home owner, when the deductibles are expressed a) in dollars, and b) in cents.

Solution: If $X$, $Y$ denote the deductibles in dollars, we saw that $\sigma_{XY} = 1875$. Omitting the details, it can be found that $\sigma_X = 75$ and $\sigma_Y = 82.92$. Thus,

$$\rho_{XY} = \frac{1875}{75 \times 82.92} = 0.301.$$ 

Next, the deductibles expressed in cents are $(X', Y') = (100X, 100Y)$. According to the proposition, $\rho_{X'Y'} = \rho_{XY}$
Correlation as a Measure of Linear Dependence

Commentaries on Pearson’s correlation coefficient:

- It is independent of scale - - highly desirable.
- It quantifies dependence well:
  - If $X$ and $Y$ are independent, then $\rho_{XY} = 0$, and
  - if one is a linear function of the other (so knowing one amounts to knowing the other), then $\rho_{XY} = \pm 1$.
- Try the applet CorrelationPicture in http://www.bolderstats.com/jmsl/doc/ to see how the correlation changes with the scatterplot
- HOWEVER, it measures only linear dependence:
  - It is possible to have perfect dependence (i.e. knowing one amounts to knowing the other) but $\rho_{XY} \neq \pm 1$. 
Hierarchical models use the Multiplication Rules, i.e.

\[ p(x, y) = p_{Y|x=x}(y)p_X(x), \quad \text{for the joint pmf} \]
\[ f(x, y) = f_{Y|x=x}(y)f_X(x), \quad \text{for the joint pdf} \]

in order to specify the joint distribution of \( X, Y \) by first specifying the conditional distribution of \( Y \) given \( X = x \), and then specifying the marginal distribution of \( X \).

This hierarchical method of modeling yields a very rich and flexible class of joint distributions.
Example

Let $X$ be the number of eggs an insect lays and $Y$ the number of eggs that survive. Model the joint pmf of $X$ and $Y$.

**Solution:** With the principle of hierarchical modeling, we need to specify the conditional pmf of $Y$ given $X = x$, and the marginal pmf of $X$. One possible specification is

$$Y|X = x \sim \text{Bin}(x, p), \quad X \sim \text{Poisson}(\lambda),$$

which leads to the following joint pmf of $X$ and $Y$.

$$p(x, y) = p_{Y|X=x}(y)p_X(x) = \binom{x}{y} p^y (1 - p)^{x-y} \frac{e^{-\lambda} \lambda^x}{x!}$$
Calculation of the marginal expected value of \( Y \), i.e. \( E(Y) \), and the marginal pmf of \( Y \) can be done with the laws of total probability.

**Example**

For the previous example, find \( E(Y) \).

**Solution:** Since \( Y|X \sim \text{Bin}(X, p) \), \( E(Y|X) = Xp \). Thus, according to the Law of Total Probability for Expectations,

\[
E(Y) = E[E(Y|X)] = E[Xp] = pE[X] = p\lambda
\]
With hierarchical modeling it is possible to describe the joint distribution of a discrete and a continuous random variable.

Example

Let the joint distribution of $X$ and $P$ be described hierarchically by specifying the conditional distribution of $X$ given $P = p$ to be binomial with parameters $n, p$ and the marginal distribution of $P$ to be uniform in $(0, 1)$. That is,

$$X|P = p \sim \text{Bin}(n, p), \quad P \sim \text{U}(0, 1).$$

Find the marginal pmf of $X$.

**Solution.** Note that $(X, P)$ are neither jointly discrete nor jointly continuous. However, the Multiplication Rule can still be used to describe their joint distribution as
\[ P(X = k | P = p) f_P(p) = \binom{n}{k} p^k (1 - p)^{n-k}. \]

This expression plays the role of the joint pdf or joint pmf of \((X, P)\) in the sense that the marginal pmf of \(X\) is obtained by integrating over all possible values of \(P\). Thus we obtain

\[
P(X = k) = \int_0^1 \binom{n}{k} p^k (1 - p)^{n-k} \, dp
\]

\[
= \binom{n}{k} \frac{k!(n - k)!}{(n + 1)!} = \frac{1}{n + 1}, \quad k = 0, \ldots, n,
\]

where the value of the integral is obtained by recognizing the integrant as proportional to a beta probability density function (see p. 503 of the book). Hence, the marginal distribution of \(X\) is uniform on the set of values \(\{0, 1, \ldots, n\}\). \(\square\)
Definition

If the joint distribution of \((X, Y)\) is specified by the assumptions that the conditional distribution of \(Y\) given \(X = x\) follows the normal linear regression model, and \(X\) has a normal distribution:

\[
Y|X = x \sim N(\beta_0 + \beta_1(x - \mu_X), \sigma_\varepsilon^2), \quad \text{and} \quad X \sim N(\mu_X, \sigma_X^2),
\]

then \((X, Y)\) is said to have a \textbf{bivariate normal distribution}.

\textit{Commentaries:}

1. The regression function of \(Y\) on \(X\) is linear in \(x\):

\[
\mu_{Y|X}(x) = \beta_0 + \beta_1(x - \mu_X).
\]

2. The marginal mean \(E(Y)\) of \(Y\) is

\[
E(Y) = E[E(Y|X)] = E[\beta_0 + \beta_1(X - \mu_X)] = \beta_0
\]
It follows that the joint pdf of \((X, Y)\) is

\[
f_{X,Y}(x, y) = \frac{1}{\sqrt{2\pi \sigma_\varepsilon^2}} \exp \left\{ - \frac{(y - \beta_0 - \beta_1(x - \mu_X))^2}{2\sigma_\varepsilon^2} \right\} \\
\cdot \frac{1}{\sqrt{2\pi \sigma_X^2}} \exp \left\{ - \frac{(x - \mu_X)^2}{2\sigma_X^2} \right\}
\]

It can be shown that the following is true

**Proposition**

*If \((X, Y)\) have a bivariate normal distribution, then the marginal distribution of \(Y\) is also normal with mean \(\mu_Y\) and variance \(\sigma_Y^2 = \sigma_\varepsilon^2 + \beta_1^2 \sigma_X^2\).*
A bivariate normal distribution is completely specified by $\mu_X$, $\mu_Y$, $\sigma_X^2$, $\sigma_Y^2$ and $\sigma_{XY}$.

The two variances and the covariance are typically arranged in a symmetric matrix, called the covariance matrix.

$$
\Sigma = \begin{pmatrix}
\sigma_X^2 & \sigma_{XY} \\
\sigma_{XY} & \sigma_Y^2
\end{pmatrix}
$$

With some algebra, the pdf of $(X, Y)$ can be rewritten as

$$
f(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} \exp \left\{ \frac{-1}{2(1 - \rho^2)} \left[ \frac{\tilde{x}^2}{\sigma_X^2} - \frac{2\rho \tilde{x} \tilde{y}}{\sigma_X \sigma_Y} + \frac{\tilde{y}^2}{\sigma_Y^2} \right] \right\}
$$

$$
= \frac{1}{2\pi \sqrt{|\Sigma|}} \exp \left\{ -\frac{1}{2} (x - \mu_X, y - \mu_Y) \Sigma^{-1} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix} \right\},
$$

where $|\Sigma| = \det(\Sigma)$ and $\tilde{x} = x - \mu_X$, $\tilde{y} = y - \mu_Y$. 
From the first of the above expressions of the joint pdf of $X, Y$ it is readily seen that if $\rho = 0$, then $f(x, y)$ becomes the product of the two marginal distributions. Thus we have

**Proposition**

*If $(X, Y)$ have a bivariate normal distribution with $\rho = 0$, $X$ and $Y$ are independent.*
Joint PMF of two independent N(0,1) RVs
Joint PMF of two N(0,1) RVs with $\rho = 0.5$
Regression models focus primarily on the regression function of a variable $Y$ on another variable $X$.

- For example, $X =$ speed of an automobile and $Y =$ the stopping distance.

- Because of this the marginal distribution of $X$, which is of little interest in such studies, is left unspecified.

- Moreover, the regression function is highlighted by writing the conditional distribution of $Y$ given $X = x$ as

$$Y = \mu_{Y|X}(x) + \varepsilon \quad (6.1)$$

and $\varepsilon$ is called the (intrinsinc) error variable.

- In regression models, $Y$ is called the response variable and $X$ is interchangeably referred to as covariate, or independent variable, or predictor, or explanatory variable.
The simple linear regression model specifies that the regression function in (6.1) is linear in $x$, i.e.

$$\mu_{Y|X}(x) = \alpha_1 + \beta_1 x, \quad (6.2)$$

and that $\text{Var}(Y|X = x)$ is the same for all values $x$.

An alternative expression of the simple linear regression model is

$$\mu_{Y|X}(x) = \beta_0 + \beta_1 (x - \mu_X). \quad (6.3)$$

The straight line in (6.2) (or (6.3)) is called the regression line.
The following picture illustrates the meaning of the slope of the regression line.

![Illustration of Regression Parameters](image)

**Figure:** Illustration of Regression Parameters
Proposition

In the linear regression model

1. The intrinsic error variable, $\epsilon$, has zero mean and is uncorrelated from the explanatory variable, $X$.

2. The slope is related to the correlation by

$$\beta_1 = \frac{\sigma_{X,Y}}{\sigma_X^2} = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}.$$  

3. Two additional relationships are

$$\mu_Y = \beta_0, \quad \sigma^2_Y = \sigma^2_\epsilon + \beta_1^2 \sigma^2_X.$$
Sketch of proof

1. \[ E(\epsilon|X) = E(Y|X) - \mu_{Y|X}(X) = 0. \] Thus, also \( E(\epsilon) = 0. \)

2. \[ E(X\epsilon) = E[E(X\epsilon|X)] = E[XE(\epsilon|X)] = 0. \] Thus, \( \text{Cov}(X, \epsilon) = 0. \)

Thus,

\[ \beta_1 = \frac{\sigma_{X,Y}}{\sigma_X^2} = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}. \]
Example

Suppose $Y = 5 - 2x + \varepsilon$, and let $\sigma_\varepsilon = 4$, $\mu_X = 7$ and $\sigma_X = 3$.

a) Find $\sigma_Y^2$, and $\rho_{X,Y}$.

b) Find $\mu_Y$.

Solution: For a) use the formulas in the above proposition:

$$\sigma_Y^2 = 4^2 + (-2)^2 \cdot 3^2 = 52$$

$$\rho_{X,Y} = \beta_1 \sigma_X / \sigma_Y = -2 \times 3 / \sqrt{52} = -0.832.$$ 

For b) use the Law of Total Probability for Expectations:

$$E(Y) = E[E(Y|X)] = E[5 - 2X] = 5 - 2E(X) = -9$$
The **normal simple linear regression model** specifies that the intrinsic error variable in (6.1) is normally distributed (in addition to the regression function being linear in \(x\)), i.e.

\[
Y = \beta_0 + \beta_1(X - \mu_X) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2_\varepsilon).
\]  

(6.4)

An alternative expression of the normal simple linear regression model is

\[
Y | X = x \sim N(\beta_0 + \beta_1(x - \mu_X), \sigma^2_\varepsilon).
\]  

(6.5)

In the above, \(\beta_0 + \beta_1(x - \mu_X)\) can also be replaced by \(\alpha_1 + \beta_1x\).

An illustration of the normal simple linear regression model is given in the following figure.
Figure: Illustration of Intrinsic Scatter in Regression
Proposition

If $(X_1, X_2)$ have a bivariate normal distribution with means $\mu_1, \mu_2$, variances $\sigma_1^2, \sigma_2^2$ and covariance $\sigma_{X,Y}$, then

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\sigma_{X,Y}).$$
Example

Suppose \( Y = 5 - 2x + \varepsilon \), and let \( \varepsilon \sim N(0, \sigma_\varepsilon^2) \) where \( \sigma_\varepsilon = 4 \). Let \( Y_1, Y_2 \) be observations taken at \( X = 1 \) and \( X = 2 \), respectively.

a) Find the 95th percentile of \( Y_1 \) and \( Y_2 \).

b) Find \( P(Y_1 > Y_2) \).

Solution:

a) \( Y_1 \sim N(3, 4^2) \), so its 95th percentile is \( 3 + 4 \times 1.645 = 9.58 \).

\( Y_2 \sim N(1, 4^2) \), so its 95th percentile is \( 1 + 4 \times 1.645 = 7.58 \).

b) \( Y_1 - Y_2 \sim N(2, 32) \) (why?). Thus,

\[
P(Y_1 > Y_2) = P(Y_1 - Y_2 > 0) = 1 - \Phi \left( \frac{-2}{\sqrt{32}} \right) = 0.6382.
\]
The multinomial distribution arises in cases where a basic experiment, which has \( r \) possible outcomes, is repeated independently \( n \) times.

For example, the basic experiment can be life testing of an electric component, with \( r = 3 \) possible outcomes: 1, if the life time is short (less than 50 time units), 2, if the life time is medium (between 50 and 90 time units), or 3, if the life time is long (exceeds 90 time units).

When this basic experiment is repeated \( n \) times, one can record the outcome of each basic experiment, resulting in \( n \) iid random variables \( X_1, \ldots, X_n \).

Alternatively, we may record

\[ N_1, \ldots, N_r, \]

where \( N_i \) = the number of times outcome \( i \) occurred.
Definition

Let $p_1, \ldots, p_r$ denote the probability of the first, $\ldots, r$th outcome. Then, the random variables $N_1, \ldots, N_r$, defined above, are said to have the \textbf{multinomial distribution} with parameters, $n$, $r$, and $p_1, \ldots, p_r$.

The joint probability mass function of the multinomial random variables $N_1, \ldots, N_r$ is

\[ P(N_1 = x_1, \ldots, N_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r}, \]

if $x_1 + \cdots x_r = n$, and zero otherwise, where $p_i$ is the probability of outcome $i$. 
The *multinomial experiment* generalizes the binomial one.

- The binomial experiment consists of \( n \) independent repetitions of an experiment whose sample space has two possible outcomes.
- Instead of recording the individual Bernoulli outcomes, \( X_1, \ldots, X_n \), the binomial experiment records \( T = \) the number of times one of the outcomes occurred.
- To see the analogy with the multinomial experiment, note that recording \( T \) is equivalent to recording

  \[
  N_1 = T, \quad \text{and} \quad N_2 = n - T.
  \]

- Similarly, in the multinomial experiment one need not record \( N_r \) since

  \[
  N_r = n - N_1 - \cdots - N_{r-1}.
  \]
Proposition

If $N_1, \ldots, N_r$ have the multinomial distribution with parameters, $n$, $r$, and $p_1, \ldots, p_r$, the marginal distribution of each $N_i$ is binomial with probability of 1 equal to $p_i$, i.e. $N_i \sim \text{Bin}(n, p_i)$. Thus,

$$E(N_i) = np_i \quad \text{and} \quad \text{Var}(N_i) = np_i(1 - p_i).$$

Moreover, it can be shown that, for $i \neq j$,

$$\text{Cov}(N_i, N_j) = -np_ip_j.$$
Example

Suppose that 60% of the supply of raw material kits used in a chemical reaction can be classified as recent, 30% as moderately aged, 8% as aged, and 2% unusable. 16 kits are randomly chosen to be used for 16 chemical reactions. Let $N_1, N_2, N_3, N_4$ denote the number of chemical reactions performed with recent, moderately aged, aged, and unusable materials.

1. Find the probability that:
   1.1 Exactly one of the 16 planned chemical reactions will not be performed due to unusable raw materials.
   1.2 10 chemical reactions will be performed with recent materials, 4 with moderately aged, and 2 with aged materials.

2. Do you expect $N_1$ and $N_2$ to be positively or negatively related? Explain at an intuitive level.

3. Find $\text{Cov}(N_1, N_2)$. 
Example (Continued)

Solution:

1. According to Proposition 19, \( N_4 \sim \text{Bin}(16, 0.02) \). Thus,

\[
P(N_4 = 1) = 16(0.02)(0.98)^{15} = 0.2363.
\]

\[
P(N_1 = 10, N_2 = 4, N_3 = 2, N_4 = 0)
= \frac{16!}{10!4!2!}0.6^{10}0.3^40.08^2 = 0.0377.
\]

2. Expect them to be negatively related: The larger \( N_1 \) is, the smaller \( N_2 \) is expected to be.

3. According to Proposition 19,

\[
\text{Cov}(N_1, N_2) = -16(0.6)(0.3) = -2.88.
\]
Example

Faxes sent by a company can be up to 3 pages long. The pmf of the fax length is $p(1) = 0.25$, $p(2) = 0.625$, $p(3) = 0.125$. Find the probability that the total page length of the next three faxes is 6 pages.

Solution: The three faxes sent must either all be 2 pages, or have 1, 2, and 3 pages. The number of faxes of 1, 2 and 3 pages, $N_1, N_2, N_3$, is multinomial. The desired probability is

$$P(N_1 = 1, N_2 = 1, N_3 = 1) + P(N_1 = 0, N_2 = 3, N_3 = 0)$$

$$= \frac{3!}{1!1!1!} 0.25 \times 0.625 \times 0.125 + 0.625^3 = .36$$

[See also Example 5.2 p. 213 of the book.]