03. Introduction to Probability Distribution

1. Random Variable and Probability Distribution
2. Mathematical Expectation

Part 1

Random Variable and Probability Distribution

Adapted From :

Probability & Statistics for Engineers & Scientists, 9th Ed.

Walpole/Myers/Myers/Ye (c)2012

Probability & Statistics for Engineering & The Science, 9th Ed.

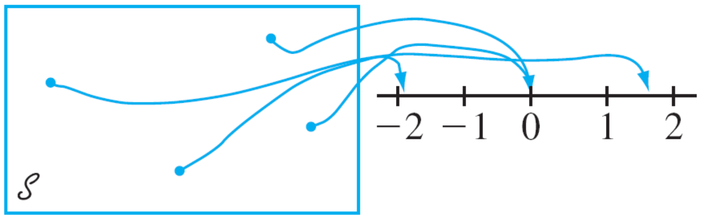
J.L.Devore © 2014

Random Variables

- In general, each outcome of an experiment can be associated with a number by specifying a rule of association (e.g., the number among the sample of ten components that fail to last 1000 hours or the total weight of baggage for a sample of 25 airline passengers).

Random Variables

- Such a rule of association is called a **random variable**—a variable because different numerical values are possible and random because the observed value depends on which of the possible experimental outcomes results (Figure 3.1).



A random variable

Figure 3.1

A Random Variable

- **A <u>random variable</u> is a function that associates a real number with each element in a sample space.**
  - Example: 3 components can either be defective (D) or not (N).
  - What is the sample space for this situation?
  - S = {NNN, NND, NDN, NDD, DNN, DND, DDN, DDD}
  - Let random variable X denote the number of defective components in each sample point.
  - Describe $P(X \le 2)$ in words.
  - What are the values of $P(X = x)$, for $x = 0, 1, 2, 3$?
  - .125, .375, .375, .125.

For a given sample space $\mathcal{S}$ of some experiment, a **random variable (rv)** is any rule that associates a number with each outcome in $\mathcal{S}$. In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.

Random Variables

- Random variables are customarily denoted by uppercase letters, such as X and Y, near the end of our alphabet.
- In contrast to our previous use of a lowercase letter, such as x, to denote a variable, we will now use lowercase letters to represent some particular value of the corresponding random variable.
- The notation $X(s) = x$ *means that* x *is the value associated with the outcome* s *by the* rv X.

Example 3.1

- When a student calls a university help desk for technical support, he/she will either immediately be able to speak to someone (*S*, for success) or will be placed on hold (*F*, for failure).

  With $\mathcal{S} = \{S, F\}$, define an rv X by

  $$X(S) = 1 \qquad X(F) = 0$$

  The rv X indicates whether (1) or not (0) the student can immediately speak to someone.

Random Variables

- **Definition**

> Any random variable whose only possible values are 0 and 1 is called a Bernoulli random variable.

Example 3.3

- Example 2.3 described an experiment in which the number of pumps in use at each of two six-pump gas stations was determined. Define rv's *X, Y,* and *U* by

  X = the total number of pumps in use at the two stations

  Y = the difference between the number of pumps in use at station 1 and the number in use at station 2

  U = the maximum of the numbers of pumps in use at the two stations

Example 3.3

cont'd

- If this experiment is performed and $s = (2, 3)$ results, then $X((2, 3)) = 2 + 3 = 5$, so we say that the observed value of X was $x = 5$.

- Similarly, the observed value of Y would be $y = 2 - 3 = -1$, and the observed value of U would be $u = \max(2, 3) = 3$.

Two Types of Random Variables

- **Definition**

A **discrete** random variable is an rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on ("countably" infinite).

A random variable is **continuous** if *both* of the following apply:

1. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from $-\infty$ to ∞) or all numbers in a disjoint union of such intervals (e.g., $[0, 10] \cup [20, 30]$).
2. No possible value of the variable has positive probability, that is, $P(X = c) = 0$ for any possible value c.

Discrete and Continuous Sample Spaces

- **A <u>discrete sample space</u> has a finite or countably infinite number of points (outcomes).**
  - Example of countably infinite: experiment consists of flipping a coin until a heads occurs.
  - S = ?
  - S = {H, TH, TTH, TTTH, TTTTH, TTTTTH, …}
  - S has a countably infinite number of sample points.
- **A <u>continuous sample space</u> has an infinite number of points, equivalent to the number of points on a line segment.**
  - For discrete sample spaces, we know that the sum of all of the probabilities of the points in the sample space equals 1.
  - Addition won't work for continuous sample spaces. Here, $P(X = x) = 0$ for any given value of x.

Example 3.6

- All random variables in Examples 3.1 –3.3 are discrete.
- As another example, suppose we select married couples at random and do a blood test on each person until we find a husband and wife who both have the same Rh factor.
- With X = the number of blood tests to be performed, possible values of X are $D = \{2, 4, 6, 8, \ldots\}$.
- Since the possible values have been listed in sequence, X is a discrete rv.

Discrete Distributions

- For a discrete random variable X, we generally look at the probability $P(X = x)$ of X taking on each value x.
- Often, the probability can be expressed in a formula, $f(x) = P(X = x)$.
- The set of ordered pairs $(x, f(x))$, is called the <u>probability distribution</u> or probability function of X.
- Note that $f(x) \geq 0$, and $\Sigma_x f(x) = 1$.

Probability Distributions for Discrete Random Variables

- Probabilities assigned to various outcomes in $\mathcal{S}$ in turn determine probabilities associated with the values of any particular rv *X*.

- The *probability distribution of X* says how the total probability of 1 is distributed among (allocated to) the various possible *X* values.

- Suppose, for example, that a business has just purchased four laser printers, and let *X* be the number among these that require service during the warranty period.

Probability Distributions for Discrete Random Variables

- Possible *X* values are then 0, 1, 2, 3, and 4. The probability distribution will tell us how the probability of 1 is subdivided among these five possible values— how much probability is associated with the *X* value 0, how much is apportioned to the *X* value 1, and so on.

  We will use the following notation for the probabilities in the distribution:

  $p\ (0)$ = the probability of the *X* value 0 = $P(X = 0)$

  $p\ (1)$ = the probability of the *X* value 1 = $P(X = 1)$

and so on. In general, $p\ (x)$ will denote the probability assigned to the value x.

Example 3.7

- The Cal Poly Department of Statistics has a lab with six computers reserved for statistics majors.
- Let *X* denote the number of these computers that are in use at a particular time of day.
- Suppose that the probability distribution of *X* is as given in the following table; the first row of the table lists the possible *X* values and the second row gives the probability of each such value.

x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

Example 3.7

cont'd

- We can now use elementary probability properties to calculate other probabilities of interest. For example, the probability that at most 2 computers are in use is

$$P(X \leq 2) = P(X = 0 \text{ or } 1 \text{ or } 2)$$

$$= p(0) + p(1) + p(2)$$

$$= .05 + .10 + .15$$

$$= .30$$

Example 3.7

cont'd

- Since the event *at least 3 computers are in use* is complementary to *at most 2 computers are in use,*

$$P(X \geq 3) = 1 - P(X \leq 2)$$

$$= 1 - .30$$

$$= .70$$

- which can, of course, also be obtained by adding together probabilities for the values, 3, 4, 5, and 6.

Example 3.7

cont'd

- The probability that between 2 and 5 computers *inclusive* are in use is

$$P(2 \leq X \leq 5) = P(X = 2, 3, 4, \text{ or } 5)$$

$$= .15 + .25 + .20 + .15$$

$$= .75$$

- whereas the probability that the number of computers in use is *strictly between* 2 and 5 is

$$P(2 < X < 5) = P(X = 3 \text{ or } 4)$$

$$= .25 + .20$$

$$= .45$$

Probability Distributions for Discrete Random Variables

- **Definition**

The probability distribution or probability mass function (pmf) of a discrete rv is defined for every number x by $p(x) = P(X = x) = P(\text{all } \omega \in \mathcal{S}: X(\omega) = x)$.

Probability Distributions for Discrete Random Variables

- In words, for every possible value *x* of the random variable, the pmf specifies the probability of observing that value when the experiment is performed.
- The conditions $p(x) \geq 0$ and $\Sigma_{\text{all possible } x}\, p(x) = 1$ are required of any pmf.
- The pmf of *X* in the previous example was simply given in the problem description.
- We now consider several examples in which various probability properties are exploited to obtain the desired distribution.

A Parameter of a Probability Distribution

- The pmf of the Bernoulli rv *X in Example 3.9* was $p(0) = .8$ and $p(1) = .2$ because 20% of all purchasers selected a desktop computer.

  At another store, it may be the case that $p(0) = .9$ and $p(1) = .1$.

- More generally, the pmf of any Bernoulli rv can be expressed in the form $p(1) = \alpha$ and $p(0) = 1 - \alpha$, where $0 < \alpha < 1$. Because the pmf depends on the particular value of α we often write $p(x; \alpha)$ rather than just $p(x)$:

$$p(x;\alpha) = \begin{cases} 1-\alpha & \text{if } x = 0 \\ \alpha & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \textbf{(3.1)}$$

A Parameter of a Probability Distribution

- Then each choice of a in Expression (3.1) yields a different pmf.

Suppose $p(x)$ depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a **parameter** of the distribution. The collection of all probability distributions for different values of the parameter is called a **family** of probability distributions.

A Parameter of a Probability Distribution

- The quantity α in Expression (3.1) is a parameter. Each different number α between 0 and 1 determines a different member of the Bernoulli family of distributions.

Example 3.12

Starting at a fixed time, **we observe the gender of each newborn child at a certain hospital until a boy (*B*) is born.**

Let $p = P(B)$, assume that successive births are independent, and define the rv X by x = number of births observed.

Then

$$p(1) = P(X = 1)$$

$$= P(B)$$

$$= p$$

Example 3.12

cont'd

$$p(2) = P(X = 2)$$
$$= P(GB)$$
$$= P(G) \cdot P(B)$$
$$= (1 - p)p$$

and

$$p(3) = P(X = 3)$$
$$= P(GGB)$$
$$= P(G) \cdot P(G) \cdot P(B)$$
$$= (1 - p)^2 p$$

Example 3.12

cont'd

- Continuing in this way, a general formula emerges:

$$p(x) = \begin{cases} (1-p)^{x-1}p & x = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(3.2)}$$

- The parameter *p* can assume any value between 0 and 1. Expression (3.2) describes the family of *geometric* distributions.
- In the gender example, *p* = .51 might be appropriate, but if we were looking for the first child with Rh-positive blood, then we might have *p* = .85.

The Cumulative Distribution Function

- For some fixed value *x*, we often wish to compute the probability that the observed value of *X* will be at most *x*. For example, let *X* be the number of number of beds occupied in a hospital's emergency room at a certain time of day; suppose the pmf of *X* is given by

x	0	1	2	3	4
$p(x)$	.20	.25	.30	.15	.10

- Then the probability that at most two beds are occupied is

$$P(X \le 2) = p(0) + p(1) + p(2) = .75$$

The Cumulative Distribution Function

- Furthermore, since $X \leq 2.7$ if and only if $X \leq 2$, we also have $P(X \leq 2.7) = .75$, and similarly $P(X \leq 2.999) = .75$.
- Since 0 is the smallest possible X value, $P(X \leq -1.5) = 0$,
- $P(X \leq -10) = 0$, and in fact for any negative number x, $P(X \leq x) = 0$.
- And because 4 is the largest possible value of X, $P(X \leq 4) = 1$, $P(X \leq 9.8) = 1$, and so on.

The Cumulative Distribution Function

- Very importantly,

$$P(X < 2) = p(0) + p(1) = .45 < .75 = P(X \le 2)$$

- because the latter probability includes the probability mass at the *x* value 2 whereas the former probability does not.

- More generally, $P(X < x) < P(X \le x)$ whenever x is a possible value of X. Furthermore, $P(X \le x)$ is a well-defined and computable probability for *any*

The Cumulative Distribution Function

- **Definition**

The **cumulative distribution function** (cdf) $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \le x) = \sum_{y:\, y \le x} p(y) \qquad (3.3)$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.

Example 3.13

- A store carries flash drives with either 1 GB, 2 GB, 4 GB, 8 GB, or 16 GB of memory.

- The accompanying table gives the distribution of *Y* = the amount of memory in a purchased drive:

y	1	2	4	8	16
$p(y)$	.05	.10	.35	.40	.10

Example 3.13

cont'd

- Let's first determine $F(y)$ for each of the five possible values of Y:

$$F(1) = P(Y \leq 1)$$
$$= P(Y = 1)$$
$$= p(1)$$
$$= .05$$

$$F(2) = P(Y \leq 2)$$
$$= P(Y = 1 \text{ or } 2)$$
$$= p(1) + p(2)$$
$$= .15$$

cont'd

Example 3.13

$$
\begin{aligned}
F(4) &= P(Y \le 4) \\
&= P(Y = 1 \text{ or } 2 \text{ or } 4) \\
&= p(1) + p(2) + p(4) \\
&= .50
\end{aligned}
$$

$$
\begin{aligned}
F(8) &= P(Y \le 8) \\
&= p(1) + p(2) + p(4) + p(8) \\
&= .90
\end{aligned}
$$

$$
\begin{aligned}
F(16) &= P(Y \le 16) \\
&= 1
\end{aligned}
$$

Example 3.13

cont'd

- Now for any other number y, $F(y)$ will equal the value of F at the closest possible value of Y to the left of y. For example,

$$\begin{aligned} F(2.7) &= P(Y \leq 2.7) \\ &= P(Y \leq 2) \\ &= F(2) \\ &= .15 \end{aligned}$$

$$\begin{aligned} F(7.999) &= P(Y \leq 7.999) \\ &= P(Y \leq 4) \\ &= F(4) \\ &= .50 \end{aligned}$$

Example 3.13

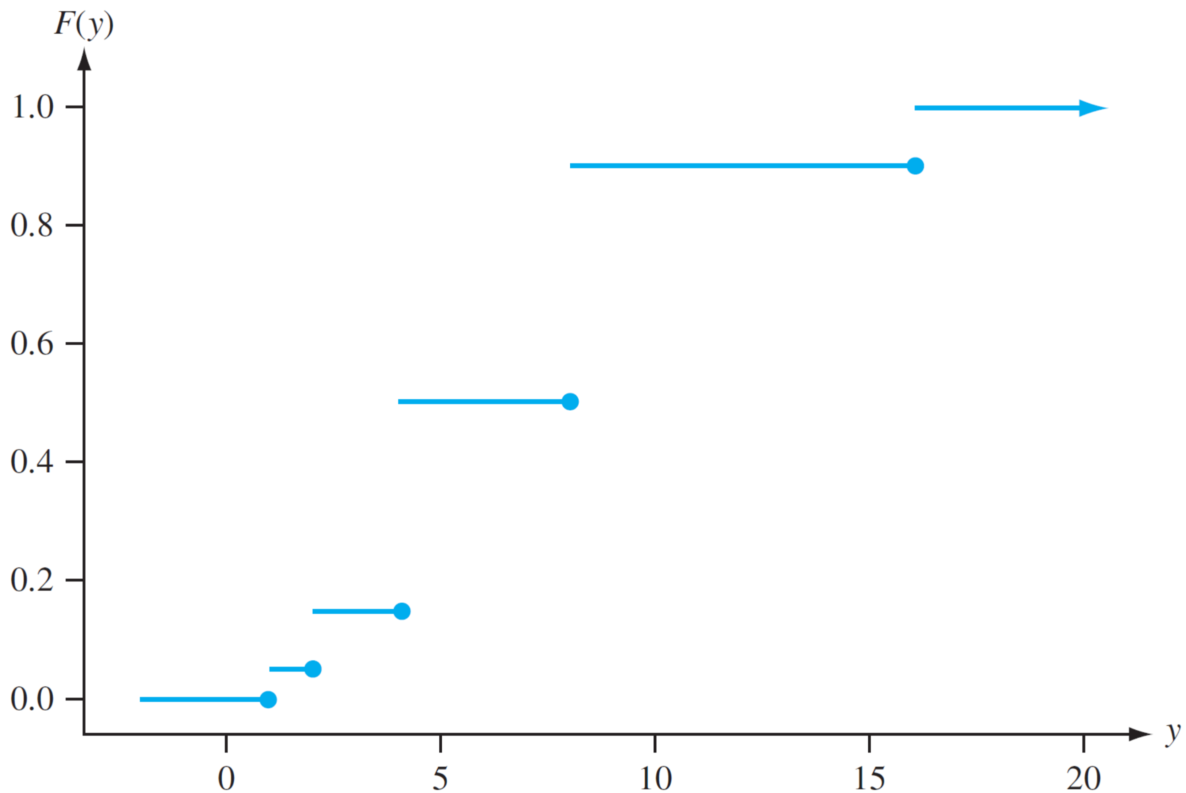
cont'd

- If y is less than 1, $F(y) = 0$ [e.g. $F(.58) = 0$], and if y is at least 16, $F(y) = 1$[e.g. $F(25) = 1$]. The cdf is thus

$$F(y) = \begin{cases} 0 & y < 1 \\ .05 & 1 \le y < 2 \\ .15 & 2 \le y < 4 \\ .50 & 4 \le y < 8 \\ .90 & 8 \le y < 16 \\ 1 & 16 \le y \end{cases}$$

Example 3.13

- A graph of this cdf is shown in Figure 3.5.



A graph of the cdf of Example 3.13

Figure 3.13

The Cumulative Distribution Function

- For X a discrete rv, the graph of $F(x)$ will have a jump at every possible value of X and will be flat between possible values. Such a graph is called a **step function.**
- **Proposition**

For any two numbers a and b with $a \le b$,

$$P(a \le X \le b) = F(b) - F(a-)$$

where "$a-$" represents the largest possible X value that is strictly less than a. In particular, if the only possible values are integers and if a and b are integers, then

$$\begin{aligned} P(a \le X \le b) &= P(X = a \text{ or } a + 1 \text{ or} \dots \text{ or } b) \\ &= F(b) - F(a - 1) \end{aligned}$$

Taking $a = b$ yields $P(X = a) = F(a) - F(a - 1)$ in this case.

The Cumulative Distribution Function

- The reason for subtracting $F(a-)$rather than $F(a)$ is that we want to include $P(X = a)$ $F(b) - F(a)$; gives P $(a < X \leq b)$.

- This proposition will be used extensively when computing binomial and Poisson probabilities in Sections 3.4 and 3.6.

Example 3.15

- Let X = the number of days of sick leave taken by a randomly selected employee of a large company during a particular year.

- If the maximum number of allowable sick days per year is 14, possible values of X are 0, 1, . . . , 14.

Example 15

cont'd

- With $F(0) = .58$, $F(1) = .72$, $F(2) = .76$, $F(3) = .81$, $F(4) = .88$, $F(5) = .94$,

$$
\begin{aligned}
P(2 \leq X \leq 5) &= P(X = 2, 3, 4, \text{ or } 5) \\
&= F(5) - F(1) \\
&= .22
\end{aligned}
$$

and

$$
\begin{aligned}
P(X = 3) &= F(3) - F(2) \\
&= .05
\end{aligned}
$$

Continuous Distributions

- Continuous distributions have an infinite number of points in the sample space, so for a given value of x, what is $P(X = x)$?
  - $P(X = x) = 0$.
  - Otherwise the probabilities couldn't sum to 1.
- What we can calculate is the probability that X lies in a given interval, such as $P(a < X < b)$, or $P(X < C)$.
  - Since the probability of any individual point is 0,
    $P(a < X < b) = P(a \leq X \leq b)$
    on, the endpoints can be included or not.
- For continuous distributions, f(x) is called a <u>probability density function</u>.

Probability Density Functions

- If $f(x)$ is a continuous probability density function,
  - $f(x) \geq 0$, as before.
  - What corresponds to $\Sigma_x f(x) = 1$ for discrete distributions?
  - $\int_{-\infty \ldots \infty} f(x)\, dx = 1$.
  - and, $P(a \leq X \leq b) = ?$
  - $P(a \leq X \leq b) = \int_{a \ldots b} f(x)\, dx$
- The cumulative distribution of a continuous random variable X is?
  - $F(x) = P(X \leq x) = ?$
  - $F(x) = \int_{-\infty \ldots x} f(x)\, dx$
- What is $P(a < X < b)$ in terms of $F(x)$?
  - $P(a < X \leq b) = F(b) - F(a)$
  - If discrete, must use "$a < X$", and not "$a \leq X$", above.

Probability Density Functions

- A discrete random variable (rv) is one whose possible values either constitute a finite set or else can be listed in an infinite sequence (a list in which there is a first element, a second element, etc.).

- A random variable whose set of possible values is an entire interval of numbers is not discrete.

Probability Density Functions

- Recall from Chapter 3 that a random variable X is continuous if
- (1) possible values comprise either a single interval on the number line (for some $A < B$, any number x between A and B is a possible value) or a union of disjoint intervals,
  and
- (2) $P(X = c) = 0$ for any number c that is a possible value of X.

Example 4.1

- If in the study of the ecology of a lake, we make depth measurements at randomly chosen locations, then
  X = the depth at such a location is a continuous rv.

- Here A is the minimum depth in the region being sampled, and B is the maximum depth.

Probability Density Functions

- One might argue that although in principle variables such as height, weight, and temperature are continuous, in practice the limitations of our measuring instruments restrict us to a discrete (though sometimes very finely subdivided) world.

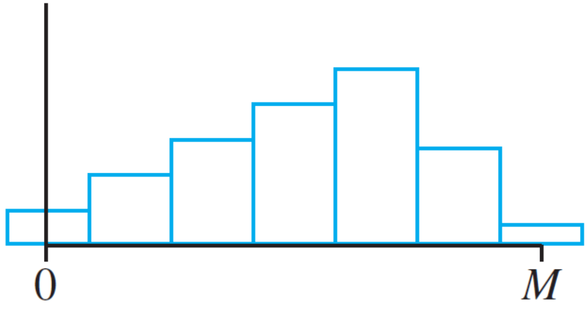
- However, continuous models often approximate real-world situations very well, and continuous mathematics (the calculus) is frequently easier to work with than mathematics of discrete variables and distributions.

Probability Distributions for Continuous Variables

- Suppose the variable *X* of interest is the depth of a lake at a randomly chosen point on the surface.
- Let *M* = the maximum depth (in meters), so that any number in the interval [0, *M*] is a possible value of *X*.
- If we "discretize" *X* by measuring depth to the nearest meter, then possible values are nonnegative integers less than or equal to *M*.
- The resulting discrete distribution of depth can be pictured using a probability histogram.

Probability Distributions for Continuous Variables

- If we draw the histogram so that the area of the rectangle above any possible integer *k* is the proportion of the lake whose depth is (to the nearest meter) *k*, then the total area of all rectangles is 1. A possible histogram appears in Figure 4.1(a).

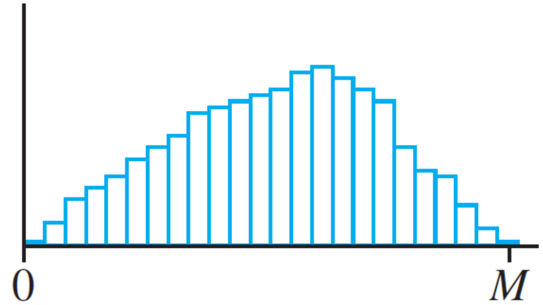


Probability histogram of depth measured to the nearest meter

Figure 4.1(a)

Probability Distributions for Continuous Variables

- If depth is measured much more accurately and the same measurement axis as in Figure 4.1(a) is used, each rectangle in the resulting probability histogram is much narrower, though the total area of all rectangles is still 1. A possible histogram is pictured in Figure 4.1(b).

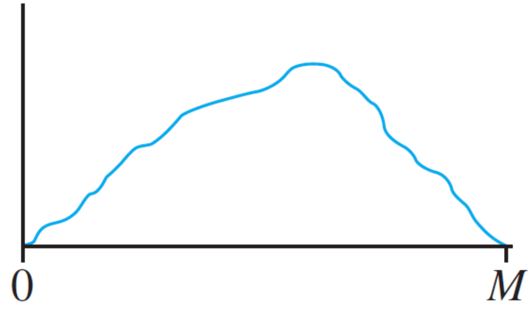


Probability histogram of depth measured to the nearest centimeter

Figure 4.1(b)

Probability Distributions for Continuous Variables

- It has a much smoother appearance than the histogram in Figure 4.1(a). If we continue in this way to measure depth more and more finely, the resulting sequence of histograms approaches a smooth curve, such as is pictured in Figure 4.1(c).



A limit of a sequence of discrete histograms

Figure 4.1(c)

Probability Distributions for Continuous Variables

- Because for each histogram the total area of all rectangles equals 1, the total area under the smooth curve is also 1.

- The probability that the depth at a randomly chosen point is between *a* and *b* is just the area under the smooth curve between *a* and *b*. It is exactly a smooth curve of the type pictured in Figure 4.1(c) that specifies a continuous probability distribution.

Probability Distributions for Continuous Variables

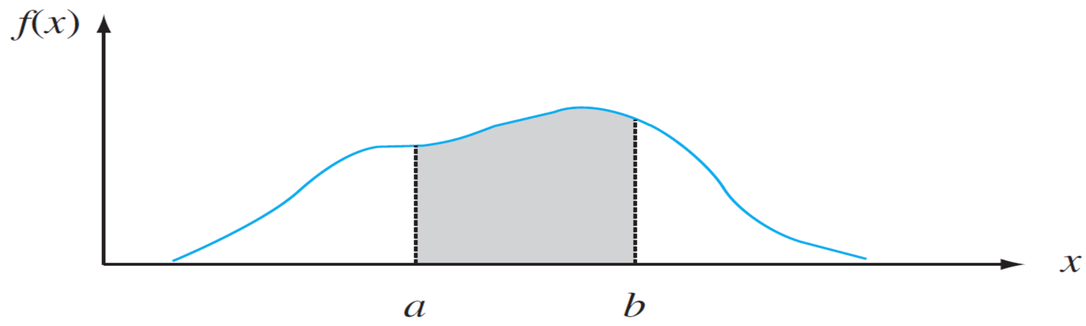
- **Definition**

Let X be a continuous rv. Then a **probability distribution** or **probability density function** (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

That is, the probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function, as illustrated in Figure 4.2. The graph of $f(x)$ is often referred to as the *density curve*.

Probability Distributions for Continuous Variables



$P(a \le X \le b)$ = the area under the density curve between a and b

Figure 4.2

- For $f(x)$ to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \geq 0$ for all x

2. $\int_{-\infty}^{\infty} f(x)\,dx$ = area under the entire graph of $f(x)$

= 1

Example 4.4

- The direction of an imperfection with respect to a reference line on a circular object such as a tire, brake rotor, or flywheel is, in general, subject to uncertainty.

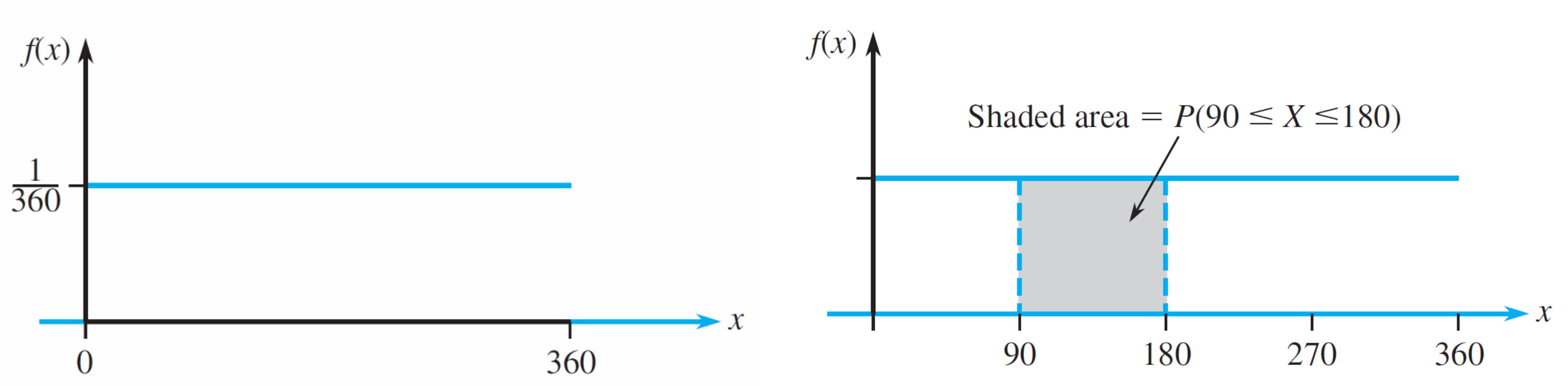
- Consider the reference line connecting the valve stem on a tire to the center point, and let *X* be the angle measured clockwise to the location of an imperfection. One possible pdf for *X* is

$$f(x) = \begin{cases} \dfrac{1}{360} & 0 \le x < 360 \\ 0 & \text{otherwise} \end{cases}$$

Example 4.4

cont'd

- The pdf is graphed in Figure 4.3.



The pdf and probability from Example 4

Figure 4.3

Example 4.4

- Clearly $f(x) \geq 0$. The area under the density curve is just the area of a rectangle:

$$(\text{height})(\text{base}) = \left(\frac{1}{360}\right)(360) = 1.$$

- The probability that the angle is between 90° and 180° is

$$P(90 \leq X \leq 180) = \int_{90}^{180} \frac{1}{360}\, dx = \frac{x}{360}\Bigg|_{x=90}^{x=180}$$

$$= \frac{1}{4} = .25$$

cont'd

Example 4.4

- The probability that the angle of occurrence is within 90° of the reference line is
- $P(0 \leq X \leq 90) + P(270 \leq X < 360) = .25 + .25 = .50$

Probability Distributions for Continuous Variables

- Because whenever $0 \leq a \leq b \leq 360$ in Example 4.4 and
  $P(a \leq X \leq b)$ depends only on the width $b - a$ of the interval, X is said to have a uniform distribution.

A continuous rv X is said to have a **uniform distribution** on the interval $[A, B]$ if the pdf of X is

$$f(x; A, B) = \begin{cases} \dfrac{1}{B - A} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

Probability Distributions for Continuous Variables

- The graph of any uniform pdf looks like the graph in Figure 4.3 except that the interval of positive density is [*A*, *B*] rather than [0, 360].

- In the discrete case, a probability mass function (pmf) tells us how little "blobs" of probability mass of various magnitudes are distributed along the measurement axis.

- In the continuous case, probability density is "smeared" in a continuous fashion along the interval of possible values. When density is smeared uniformly over the interval, a uniform pdf, as in Figure 4.3, results.

Probability Distributions for Continuous Variables

- When *X* is a discrete random variable, each possible value is assigned positive probability.

- This is not true of a continuous random variable (that is, the second condition of the definition is satisfied) because the area under a density curve that lies above any single value is zero:

$$P(X = c) = \int_c^c f(x)\,dx = \lim_{\varepsilon \to 0} \int_{c-\varepsilon}^{c+\varepsilon} f(x)\,dx = 0$$

Probability Distributions for Continuous Variables

- The fact that $P(X = c) = 0$ when X is continuous has an important practical consequence: The probability that X lies in some interval between a and b does not depend on whether the lower limit a or the upper limit b is included in the probability calculation:

$$P(a \leq X \leq b) = P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) \quad (4.1)$$

- If X is discrete and both a and b are possible values (e.g., X is binomial with $n = 20$ and $a = 5$, $b = 10$), then all four of the probabilities in (4.1) are different.

Probability Distributions for Continuous Variables

- The zero probability condition has a physical analog. Consider a solid circular rod with cross-sectional area = 1 in^2.
- Place the rod alongside a measurement axis and suppose that the density of the rod at any point *x* is given by the value *f* (*x*) of a density function. Then if the rod is sliced at points *a* and *b* and this segment is removed, the amount of mass removed is $\int_a^b f(x)\,dx$; if the rod is sliced just at the point *c*, no mass is removed.
- Mass is assigned to interval segments of the rod but not to individual points.

Example 5.5

- "Time headway" in traffic flow is the elapsed time between the time that one car finishes passing a fixed point and the instant that the next car begins to pass that point.

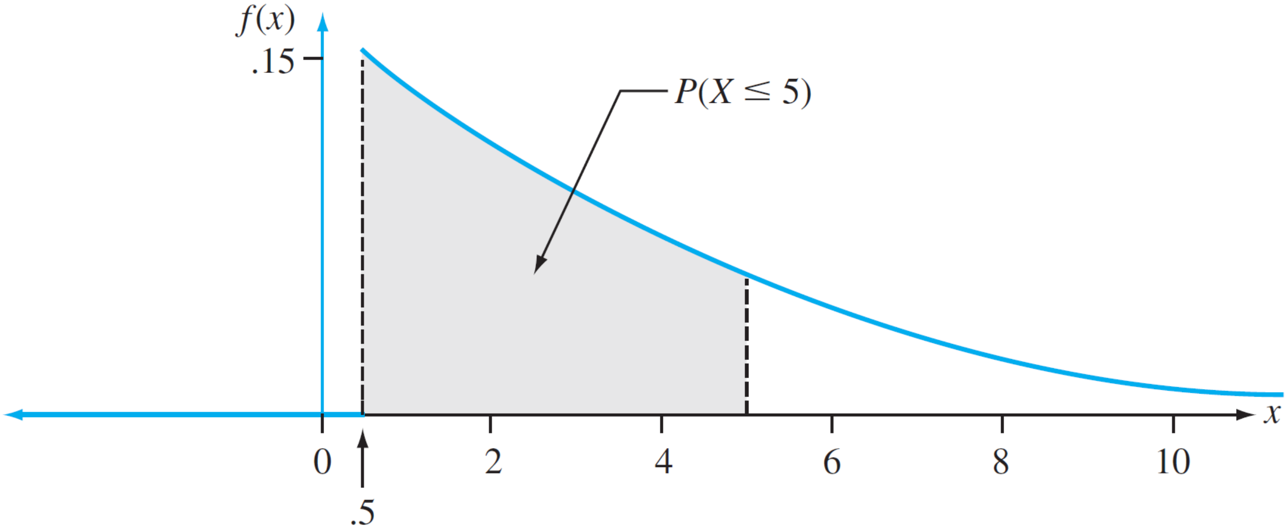
- Let *X* = the time headway for two randomly chosen consecutive cars on a freeway during a period of heavy flow. The following pdf of *X* is essentially the one suggested in "The Statistical Properties of Freeway Traffic" (*Transp. Res.,* vol. 11: 221–228):

$$f(x) = \begin{cases} .15e^{-.15(x-.5)} & x \geq .5 \\ 0 & \text{otherwise} \end{cases}$$

Example 5.5

cont'd

- The graph of $f(x)$ is given in Figure 4.4; there is no density associated with headway times less than .5, and headway density decreases rapidly



The density curve for time headway in Example 5

Figure 4.4

cont'd

Example 5.5

- Clearly, $f(x) \geq 0$; to show that $\int_{-\infty}^{\infty} f(x)dx = 1$, we use the calculus result

$$\int_{a}^{\infty} e^{-kx}\,dx = (1/k)e^{-k \cdot a}.$$

- Then

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{.5}^{\infty} .15e^{-.15(x-.5)}\,dx$$

$$= .15e^{.075} \int_{.5}^{\infty} e^{-.15x}\,dx$$

$$= .15e^{.075} \cdot \frac{1}{.15} e^{-(.15)(.5)}$$

$$= 1$$

Example 5.5

cont'd

- The probability that headway time is at most 5 sec is

$$P(X \le 5) = \int_{-\infty}^{5} f(x)\,dx$$

$$= \int_{.5}^{5} .15e^{-.15}(x - .5)\,dx$$

$$= .15e^{.075} \int_{.5}^{5} e^{-15x}\,dx$$

$$= .15e^{.075} \cdot \left(-\frac{1}{.15} e^{-.15x} \Big|_{x=.5}^{x=5} \right)$$

Example 5.5

cont'd

$$= e^{.075}(-e^{-.75} + e^{-.075})$$

$$= 1.078(-.472 + .928)$$

$$= .491$$

$$= P(\text{less than 5 sec})$$

$$= P(X < 5)$$

Probability Distributions for Continuous Variables

- Unlike discrete distributions such as the binomial, hypergeometric, and negative binomial, the distribution of any given continuous rv cannot usually be derived using simple probabilistic arguments.

- Just as in the discrete case, it is often helpful to think of the population of interest as consisting of X values rather than individuals or objects.

- The pdf is then a model for the distribution of values in this numerical population, and from this model various population characteristics (such as the mean) can be calculated.

Joint Probability Distributions

- Given a pair of discrete random variables on the same sample space, X and Y, the joint probability distribution of X and Y is

  $$f(x,y) = P(X = x, Y = y)$$

  $f(x,y)$ equals the probability that both x and y occur.

- The usual rules hold for joint probability distributions:
  - $f(x,y) \geq 0$
  - $\Sigma_x \Sigma_y f(x, y) = 1$
  - For any region A in the xy plane,
    $P[(X,Y) \in A] = \Sigma\Sigma_A f(x, y)$
- For continuous joint probability distributions, the sums above are replaced with integrals.

Marginal Distributions

- The marginal distribution of X alone or Y alone can be calculated from the joint distribution function as follows:
  - $g(x) = \Sigma_y f(x,y)$ and $h(y) = \Sigma_x f(x,y)$ if discrete
  - $g(x) = \int_y f(x,y)$ and $h(y) = \int_x f(x,y)$ if continuous
- In other words, for example, $g(x) = P(X = x)$ is the sum (or integral) of $f(x,y)$ over all values of y.

Two Discrete Random Variables

- The probability mass function (pmf) of a single discrete rv X specifies how much probability mass is placed on each possible X value.

- The joint pmf of two discrete rv's X and Y describes how much probability mass is placed on each possible pair of values (x, y).

Two Discrete Random Variables

- **Definition**

Let X and Y be two discrete rv's defined on the sample space $\mathcal{S}$ of an experiment. The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

It must be the case that $p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$.

Now let A be any particular set consisting of pairs of (x, y) values (e.g., $A = \{(x, y): x + y = 5\}$ or $\{(x, y): \max(x, y) \leq 3\}$). Then the probability $P[(X, Y) \in A]$ that the random pair (X, Y) lies in the set A is obtained by summing the joint pmf over pairs in A:

$$P[(X, Y) \in A] = \sum_{(x, y) \in A} \sum p(x, y)$$

Example 5.1

- Anyone who purchases an insurance policy for a home or automobile must specify a deductible amount, the amount of loss to be absorbed by the policyholder before the insurance company begins paying out.

- Suppose that a particular company offers auto deductible amounts of \$100, \$500, and \$1000, and homeowner deductible amounts of \$500, \$1000, and \$2000. Consider randomly selecting someone who has both auto and homeowner insurance with this company, and let X =the amount of the auto policy deductible and Y = the amount of the homeowner policy deductible.

Example 5.1

cont'd

- The joint pmf of these two variables appears in the accompanying *joint probability table*:

	$p(x, y)$	y: 500	1000	5000
	100	.30	.05	0
x	500	.15	.20	.05
	1000	.10	.10	.05

- According to this joint pmf, there are nine possible (X, Y) pairs: (100, 500), (100, 1000), … , and finally (1000, 5000). The probability of (100, 500) is $p(100, 500) = P(X = 100, Y = 500) = .30$. Clearly $p(x, y) \geq 0$, and it is easily confirmed that the sum of the nine displayed probabilities is 1

Example 5.1

cont'd

- The probability $P(X = Y)$ is computed by summing $p(x, y)$ over the two (x, y) pairs for which the two deductible amounts are identical:

- $P(X = Y) = p(500, 500) + p(1000, 1000) = .15 + .10 = .25$

- Similarly, the probability that the auto deductible amount is at least \$500 is the sum of all probabilities corresponding to (x, y) pairs for which $x \geq 500$; this is the sum of the probabilities in the bottom two rows of the joint probability table:

Two Discrete Random Variables

- Once the joint pmf of the two variables X and Y is available, it is in principle straightforward to obtain the distribution of just one of these variables.

  As an example, let X and Y be the number of statistics and mathematics courses, respectively, currently being taken by a randomly selected statistics major.
- Suppose that we wish the distribution of X, and that when $X = 2$, the only possible values of Y are 0, 1, and 2.

Two Discrete Random Variables

- Then

$$p_X(2) = P(X = 2) = P[(X, Y) = (2, 0) \text{ or } (2, 1) \text{ or } (2, 2)]$$

$$= p(2, 0) + p(2, 1) + p(2, 2)$$

- That is, the joint pmf is summed over all pairs of the form $(2, y)$. More generally, for any possible value x of X, the probability $p_X(x)$ results from holding x fixed and summing the joint pmf $p(x, y)$ over all y for which the pair (x, y) has positive probability mass.

Two Discrete Random Variables

- **Definition**

The **marginal probability mass function of** X, denoted by $p_X(x)$, is given by

$$p_X(x) = \sum_{y:\, p(x,y)>0} p(x, y) \quad \text{for each possible value } x$$

Similarly, the **marginal probability mass function of** Y is

$$p_Y(y) = \sum_{x:\, p(x,y)>0} p(x, y) \quad \text{for each possible value } y.$$

Example 5.2

- **Example 5.1 continued…**
  The possible X values are x = 100, 500 and x = 1000, so computing row totals in the joint probability table yields

$$p_X(100) = p(100, 500) + p(100, 1000) + p(100, 5000) = .30 + .05 + 0 = .35$$

$$p_X(500) = .15 + .20 + .05 = .40, \ p_X(1000) = 1 - (.35 + .40) = .25$$

The marginal pmf of X is then

$$p_X(x) = \begin{cases} .35 & x = 100 \\ .40 & x = 500 \\ .25 & x = 1000 \\ 0 & \text{otherwise} \end{cases}$$

Example 5.2

cont'd

- Similarly, the marginal pmf of *X* is then

$$p_X(x) = \begin{cases} .35 & x = 100 \\ .40 & x = 500 \\ .25 & x = 1000 \\ 0 & \text{otherwise} \end{cases}$$

- From this pmf, $P(X \geq 500) = .40 + .25 = .65$, which we already calculated in Example 5.1. Similarly, the marginal pmf of *Y* is obtained from the column totals as

$$p_Y(y) = \begin{cases} .55 & y = 500 \\ .35 & y = 1000 \\ .10 & y = 5000 \\ 0 & \text{otherwise} \end{cases}$$

Two Continuous Random Variables

- The probability that the observed value of a continuous rv X lies in a one-dimensional set A (such as an interval) is obtained by integrating the pdf $f(x)$ over the set A.

  Similarly, the probability that the pair (X, Y) of continuous rv's falls in a two-dimensional set A (such as a rectangle) is obtained by integrating a function called the *joint density function*.

Two Continuous Random Variables

- **Definition**

Let X and Y be continuous rv's. A **joint probability density function** $f(x, y)$ for these two variables is a function satisfying $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$. Then for any two-dimensional set A

$$P[(X, Y) \in A] = \iint_A f(x, y)\, dx\, dy$$

In particular, if A is the two-dimensional rectangle $\{(x, y): a \leq x \leq b, c \leq y \leq d\}$, then

$$P[(X,Y) \in A] = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y)\, dy\, dx$$

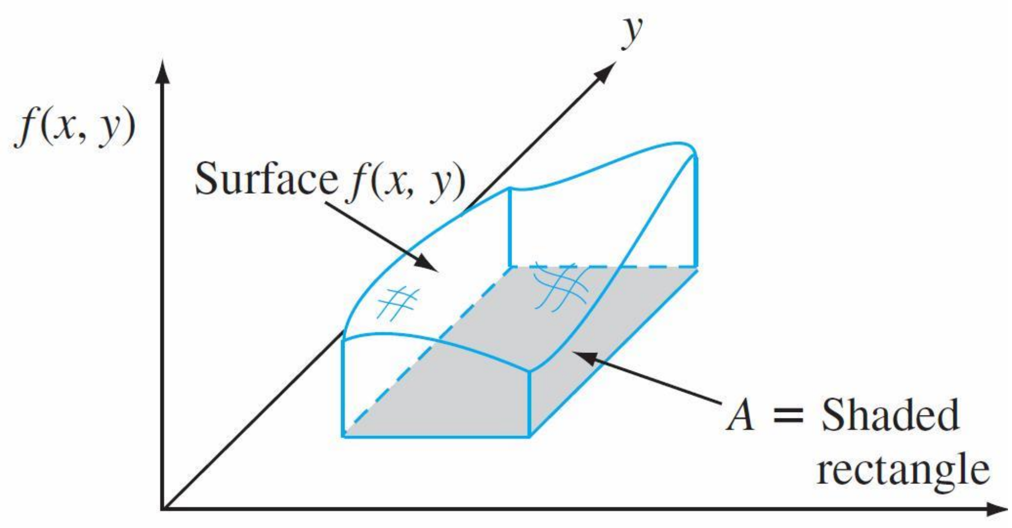
Two Continuous Random Variables

- We can think of $f(x, y)$ as specifying a surface at height
  $f(x, y)$ above the point (x, y) in a three-dimensional coordinate system.

  Then $P[(X, Y) \in A]$ is the volume underneath this surface and above the region A, analogous to the area under a curve in the case of a single rv.

Two Continuous Random Variables

- This is illustrated in Figure 5.1.



$P[(X, Y) \in A]$ = volume under density surface above A

Figure 5.1

Example 5.3

- A bank operates both a drive-up facility and a walk-up window. On a randomly selected day, let X = the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y = the proportion of time that the walk-up window is in use.

- Then the set of possible values for (X, Y) is the rectangle
- $D = \{(x, y): 0 \leq x \leq 1, 0 \leq y \leq 1\}$.

Example 5.3

cont'd

- Suppose the joint pdf of (X, Y) is given by

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- To verify that this is a legitimate pdf, note that $f(x, y) \ge 0$ and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = \int_0^1\int_0^1 \frac{6}{5}(x + y^2)\,dx\,dy$$

$$= \int_0^1\int_0^1 \frac{6}{5}x\,dx\,dy + \int_0^1\int_0^1 \frac{6}{5}y^2\,dx\,dy$$

Example 5.3

cont'd

$$= \int_0^1 \frac{6}{5} x \, dx + \int_0^1 \frac{6}{5} y^2 \, dy = \frac{6}{10} + \frac{6}{15} = 1$$

$$= \frac{6}{10} + \frac{6}{15}$$

$$= 1$$

The probability that neither facility is busy more than one-quarter of the time is

$$P\left(0 \le X \le \frac{1}{4}, 0 \le Y \le \frac{1}{4}\right) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5} (x + y^2) \, dx \, dy$$

$$= \frac{6}{5} \int_0^{1/4} \int_0^{1/4} x \, dx \, dy + \frac{6}{5} \int_0^{1/4} \int_0^{1/4} y^2 \, dx \, dy$$

Example 5.3

cont'd

$$= \frac{6}{20} \cdot \left.\frac{x^2}{2}\right|_{x=0}^{x=1/4} + \frac{6}{20} \cdot \left.\frac{y^3}{3}\right|_{y=0}^{y=1/4}$$

$$= \frac{7}{640}$$

$$= .0109$$

Two Continuous Random Variables

- The marginal pdf of each variable can be obtained in a manner analogous to what we did in the case of two discrete variables.

  The marginal pdf of X at the value x results from holding x fixed in the pair (x, y) and *integrating* the joint pdf over y. Integrating the joint pdf with respect to x gives the marginal pdf of Y.

Two Continuous Random Variables

- **Definition**

The **marginal probability density functions** of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy \qquad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx \qquad \text{for } -\infty < y < \infty$$

Example 5.4

- The marginal pdf of *X*, which gives the probability distribution of busy time for the drive-up facility without reference to the walk-up window, is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy = \int_0^1 \frac{6}{5}(x + y^2)\,dy = \frac{6}{5}x + \frac{2}{5}$$

- for $0 \leq x \leq 1$ and 0 of
  Y is

$$f_Y(y) = \begin{cases} \frac{6}{5}y^2 + \frac{3}{5} & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Example 5.4

- Then

$$P(.25 \le Y \le .75) = \int_{.25}^{.75} f_Y(y)\,dy = \frac{37}{80} = .4625$$

Independent Random Variables

- In many situations, information about the observed value of one of the two variables *X* and *Y* gives information about the value of the other variable.

  In Example 5.1, the marginal probability of *X* at *x* = 100 was .35, and at *X* = 1000 is .25. However, we learn that Y = 5000 the last column of the joint probability table tells us that X can't possible be 100 and the other two possibilities, 500 and 1000, are now equally likely. Thus knowing the value is a dependence between two variables.

- In Chapter 2, we pointed out that one way of defining independence of two events is via the condition
- $P(A \cap B) = P(A) \cdot P(B)$.

Independent Random Variables

- Here is an analogous definition for the independence of two rv's.

Two random variables X and Y are said to be **independent** if for every pair of x and y values

$$p(x, y) = p_X(x) \cdot p_Y(y) \qquad \text{when } X \text{ and } Y \text{ are discrete}$$

or $\hfill (5.1)$

$$f(x, y) = f_X(x) \cdot f_Y(y) \qquad \text{when } X \text{ and } Y \text{ are continuous}$$

If (5.1) is not satisfied for all (x, y), then X and Y are said to be **dependent.**

Independent Random Variables

- The definition says that two variables are independent if their joint pmf or pdf is the product of the two marginal pmf's or pdf's.

  Intuitively, independence says that knowing the value of one of the variables does not provide additional information about what the value of the other variable might be.

Example 5.6

- In the insurance situation of Examples 5.1 and 5.2,

$$p(1000, 5000) = .05 \neq (.10)(.25) = p_X(1000) \cdot p_Y(5000)$$

so X and Y are not independent.

- In fact, the joint probability table has an entry which is 0, yet the corresponding row and column totals are both positive.

- Independence of X and Y requires that *every* entry in the joint probability table be the product of the corresponding row and column marginal probabilities.

Independent Random Variables

- Independence of two random variables is most useful when the description of the experiment under study suggests that X and Y have no effect on one another.

- Then once the marginal pmf's or pdf's have been specified, the joint pmf or pdf is simply the product of the two marginal functions. It follows that

  - $P(a \leq X \leq b, c \leq Y \leq d) = P(a \leq X \leq b) \cdot P(c \leq Y \leq d)$

More Than Two Random Variables

- To model the joint behavior of more than two random variables, we extend the concept of a joint distribution of two variables.

If $X_1, X_2, \ldots, X_n$ are all discrete random variables, the joint pmf of the variables is the function

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

If the variables are continuous, the joint pdf of $X_1, \ldots, X_n$ is the function $f(x_1, x_2, \ldots, x_n)$ such that for any n intervals $[a_1, b_1], \ldots, [a_n, b_n]$,

$$P(a_1 \le X_1 \le b_1, \ldots, a_n \le X_n \le b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)\, dx_n \ldots dx_1$$

Example 5.9

- A binomial experiment consists of *n* dichotomous (success–failure), homogenous (constant success probability) independent trials.

- Now consider a *trinomial* experiment in which each of the *n* trials can result in one of *three* possible outcomes. For example, each successive customer at a store might pay with cash, a credit card, or a debit card. The trials are assumed independent.

- Let p_1 = *P*(trial results in a type 1 outcome) and define p_2 and p_3 analogously for type 2 and type 3 outcomes. The random variables of interest here are X_i= the number of trials that result in a type *i* outcome for *i* = 1, 2, 3.

Example 5.9

- In n = 10 trials, the probability that the first five are type 1 outcomes, the next three are type 2, and the last two are type 3—that is, the probability of the experimental outcome 1111122233—is $p_1^5 \cdot p_2^3 \cdot p_3^2$.

- This is also the probability of the outcome 1122311123, and in fact the probability of any outcome that has exactly five 1's, three 2's, and two 3's.

- Now to determine the probability $P(X_1 = 5, X_2 = 3, \text{and } X_3 = 2)$, we have to count the number of outcomes that have exactly five 1's, three 2's, and two 3's.

Example 5.9

- First, there are $\binom{10}{5}$ ways to choose five of the trials to be the type 1 outcomes. Now from the remaining five trials, we choose three to be the type 2 outcomes, which can be done in $\binom{5}{3}$ ways.

- This determines the remaining two trials, which consist of type 3 outcomes. So the total number of ways of choosing five 1's, three 2's, and two 3's is

$$\binom{10}{5} \cdot \binom{5}{3} = \frac{10!}{5!5!} \cdot \frac{5!}{3!2!} = \frac{10!}{5!3!2!} = 2520$$

Example 5.9

- Thus we see that $P(X_1 = 5, X_2 = 3, X_3 = 2) = 2520\, p_1^5 \cdot p_2^3 \cdot p_3^2.$ Generalizing this to *n* trials gives

$$p(x_1, x_2, x_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

- for $x_1 = 0, 1,2, \ldots;\ x_2 = 0, 1, 2, \ldots;\ x_3 = 0, 1, 2, \ldots$ such that $x_1 + x_2 + x_3 = n.$
- Notice that whereas there are three random variables here, the third variable x_3 is actually redundant. For example, in the case *n* = 10, having x_1 = 5 and $x_2 = 3$ implies that x_3 = 2 (just as in a binomial experiment there are actually two rv's—the number of successes and number of failures—but

Example 5.9

- As a specific example, the genetic allele of a pea section can be either AA, Aa, or aa.
- A simple genetic model specifies $P(\text{AA}) = .25$, $P(\text{Aa}) = .50$, and $P(\text{aa}) = .25$.
- If the alleles of 10 independently obtained sections are determined, the probability that exactly five of these are Aa and two are AA is

$$p(2, 5, 3) = \frac{10!}{2!5!3!}(.25)^2(.50)^5(.25)^3 = 0.769$$

Example 5.9

- A natural extension of the trinomial scenario is an experiment consisting of *n* independent and identical trials,
- in which each trial can result in any one of *r* possible outcomes.
- Let p_i = *P*(outcome *i* on any particular trial), and define random variables by X_i = the number of trials resulting in outcome *i* (*i* = 1, … , *r*).

Example 5.9

- This is called a **multinomial experiment**, and the joint pmf of $X_1, \ldots, X_r$ is called the **multinomial distribution**. An argument analogous to the one used to derive the trinomial pmf gives the multinomial pmf as

$$p(x_1, \ldots, x_r) = \begin{cases} \dfrac{n!}{(x_1!)(x_2!) \cdot \cdots \cdot (x_r!)} \, p_1^{x_1} \cdot \ldots \cdot p_r^{x_r} & x_i = 0, 1, 2, \ldots, \text{ with } x_1 + \cdots + x_r = n \\ 0 & \text{otherwise} \end{cases}$$

More Than Two Random Variables

- The notion of independence of more than two random variables is similar to the notion of independence of more than two events.

The random variables $X_1, X_2, \ldots, X_n$ are said to be **independent** if for *every* subset $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ of the variables (each pair, each triple, and so on), the joint pmf or pdf of the subset is equal to the product of the marginal pmf's or pdf's.

More Than Two Random Variables

- Thus if the variables are independent with $n = 4$, then the joint pmf or pdf of any two variables is the product of the two marginals, and similarly for any three variables and all four variables together.

- Intuitively, independence means that learning the values of some variables doesn't change the distribution of the remaining variables.

- Most importantly, once we are told that n variables are independent, then the joint pmf or pdf is the product of the n marginals.

Conditional Distributions

- Suppose X = the number of major defects in a randomly selected new automobile and Y = the number of minor defects in that same auto.

  If we learn that the selected car has one major defect, what now is the probability that the car has at most three minor defects—that is, what is $P(Y \leq 3 \mid X = 1)$?

Conditional Distributions

- Similarly, if X and Y denote the lifetimes of the front and rear tires on a motorcycle, and it happens that $X = 10{,}000$ miles, what now is the probability that Y is at most 15,000 miles, and what is the expected lifetime of the rear tire “conditional on” this value of X?

- Questions of this sort can be answered by studying conditional probability distributions.

Conditional Distributions

- **Definition**

Let X and Y be two continuous rv's with joint pdf $f(x, y)$ and marginal X pdf $f_X(x)$. Then for any X value x for which $f_X(x) > 0$, the **conditional probability density function of Y given that $X = x$** is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} \qquad -\infty < y < \infty$$

If X and Y are discrete, replacing pdf's by pmf's in this definition gives the **conditional probability mass function of Y when $X = x$.**

Conditional Distributions

- Notice that the definition of $f_{Y \mid X}(y \mid x)$ parallels that of $P(B \mid A)$, the conditional probability that B will occur, given that A has occurred.

  Once the conditional pdf or pmf has been determined, questions of the type posed at the outset of this subsection can be answered by integrating or summing over an appropriate set of Y values.

Example 5.12

- Reconsider the situation of example 5.3 and 5.4 involving X = the proportion of time that a bank's drive-up facility is busy and Y = the analogous proportion for the walk-up window.

- The conditional pdf of Y given that X = .8 is

$$f_{Y|X}(y\,|\,.8) = \frac{f(.8, y)}{f_X(.8)} = \frac{1.2(.8 + y^2)}{1.2(.8) + .4} = \frac{1}{34}(24 + 30y^2) \quad 0 < y < 1$$

Example 5.12

- The probability that the walk-up facility is busy at most half the time given that $X = .8$ is then

$$P(Y \leq .5 \mid X = .8) = \int_{-\infty}^{.5} f_{Y|X}(y \mid .8)\, dy$$

$$= \int_{0}^{.5} \frac{1}{34}(24 + 30y^2)\, dy$$

$$= .390$$

Example 5.12

cont'd

- Using the marginal pdf of Y gives $P(Y \leq .5) = .350$. Also $E(Y) = .6$, whereas the expected proportion of time that the walk-up facility is busy given that $X = .8$ (a *conditional* expectation) is

$$E(Y \mid X = .8) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid .8)\, dy$$

$$= \frac{1}{34} \int_0^1 y(24 + 30y^2)\, dy$$

$$= .574$$

Conditional Distributions

- For either discrete or continuous random variables, X and Y, the conditional distribution of Y, given that X = x, is

$$f(y \mid x) = f(x,y) / g(x) \qquad \text{if } g(x) > 0$$

and

$$f(x \mid y) = f(x,y) / h(y) \qquad \text{if } h(y) > 0$$

- X and Y are statistically independent if

  $$f(x,y) = g(x)\, h(y)$$

  for all x and y within their range.

- A similar equation holds for n mutually statistically independent jointly distributed random variables.

Statistical Independence

- The definition of independence is as before:
  - Previously, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.
  - How about terms of the conditional distribution?
  - $f(x \mid y) = g(x)$ and $f(y \mid x) = h(y)$.
  - The other way to demonstrate independence?
  - $f(x, y) = g(x)\,h(y) \qquad \forall\, x, y$ in range.
- Similar formulas also apply to more than two mutually independent random variables.

Part 2

Mathematical Expectation

Adapted From :

Probability & Statistics for Engineers & Scientists, 9th Ed.

Walpole/Myers/Myers/Ye (c)2012

Probability & Statistics for Engineering & The Science, 9th Ed.

J.L.Devore © 2014

Mean of a Set of Observations

- Suppose an experiment involves tossing 2 coins. The result is either 0, 1, or 2 heads. Suppose the experiment is repeated 15 times, and suppose that 0 heads is observed 3 times, 1 head 8 times, and 2 heads 4 times.
  - What is the average number of heads flipped?
  - x_{bar} = (0+0+0+1+1+1+1+1+1+1+1+2+2+2+2) / 15
    = ((0)(3) + (1)*(8) + (2)*(4)) / 15 = 1.07
  - This could also be written as a weighted average,
  - x_{bar} = (0)(3/15) + (1)(8/15) + (2)(4/15) = 1.07
    where 3/15, 8/15, etc. are the fraction of times the given number of heads came up.
- The average is also called the <u>mean</u>.

Mean of a Random Variable

- A similar technique, taking the probability of an outcome times the value of the random variable for that outcome, is used to calculate the mean of a random variable.
- The <u>mean</u> or <u>expected value</u> μ of a random variable X with probability distribution f (x), is

$$\mu = E(X) = \Sigma_x x\, f(x) \qquad \text{if discrete, or}$$

$$\mu = E(X) = \int_x x\, f(x)\, dx \qquad \text{if continuous}$$

Mean of a Random Variable Depending on X

- If X is a random variable with distribution f(x). The mean $\mu_{g(X)}$ of the random variable g(X) is

$$\mu_{g(X)} = E[g(X)] = \Sigma_x g(x) f(x) \qquad \text{if discrete, or}$$

$$\mu_{g(X)} = E[g(X)] = \int_x g(x) f(x)\, dx \qquad \text{if continuous}$$

Example 3.16

- Consider a university having 15,000 students and let X = of courses for which a randomly selected student is registered. The pmf of X follows.

x	1	2	3	4	5	6	7
$p(x)$	.01	.03	.13	.25	.39	.17	.02
Number registered	150	450	1950	3750	5850	2550	300

$$\mu = 1 \cdot p(1) + 2 \cdot p(2) + \ldots + 7 \cdot p(7)$$

$$= (1)(.01) + 2(.03) + \ldots + (7)(.02)$$

$$= .01 + .06 + .39 + 1.00 + 1.95 + 1.02 + .14$$

$$= 4.57$$

Example 3.16

cont'd

- If we think of the population as consisting of the X values 1, 2, . . . , 7, then $\mu = 4.57$ is the population mean.
- In the sequel, we will often refer to μ as the *population mean* rather than the mean of X in the population.
- Notice that μ here is not 4, the ordinary average of 1, . . . , 7, because the distribution puts more weight on 4, 5, and 6 than on other X values.

The Expected Value of a Function

- Sometimes interest will focus on the expected value of some function $h(X)$ rather than on just $E(X)$.

If the rv X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(X)$, denoted by $E[h(X)]$ or $\mu_{h(X)}$, is computed by

$$E[h(X)] = \sum_D h(x) \cdot p(x)$$

That is, $E[h(X)]$ is computed in the same way that $E(X)$ itself is, except that $h(x)$ is substituted in place of x.

Example 3.23

- A computer store has purchased three computers of a certain type at \$500 apiece. It will sell them for \$1000 apiece.
- The manufacturer has agreed to repurchase any computers still unsold after a specified period at \$200 apiece.
- Let X denote the number of computers sold, and suppose that $p(0) = .1$, $p(1) = .2$, $p(2) = .3$ and $p(3) = .4$.

Example 3.23

- With $h(X)$ denoting the profit associated with selling X units, the given information implies that

$$h(X) = \text{revenue} - \text{cost}$$

$$= 1000X + 200(3 - X) - 1500$$

$$= 800X - 900$$

The expected profit is then

$$E[h(X)] = h(0) \cdot p(0) + h(1) \cdot p(1) + h(2) \cdot p(2) + h(3) \cdot p(3)$$

$$= (-900)(.1) + (-100)(.2) + (700)(.3) + (1500)(.4)$$

$$= \$700$$

Expected Value of a Linear Function

- The *h* (*X*) function of interest is quite frequently a linear function *aX* + *b*. In this case, *E* [*h* (*X*)] is easily computed from *E* (*X*).

Proposition

$$E(aX + b) = a \cdot E(X) + b$$

(Or, using alternative notation, $\mu_{aX+b} = a \cdot \mu_X + b$)

To paraphrase, the expected value of a linear function equals the linear function evaluated at the expected value *E*(*X*). Since *h*(*X*) in Example 3.23 is linear and *E*(*X*) = 2, *E*[*h*(*x*)] = 800(2) – 900 = $700, as before.

The Variance of *X*

- **Definition**

Let X have pmf $p(x)$ and expected value μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2, or just σ^2, is

$$V(X) = \sum_{D} (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

The **standard deviation** (SD) of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

The Variance of X

- The quantity $h(X) = (X - \mu)^2$ is the squared deviation of X from its mean, and σ^2 is the expected squared deviation—i.e., the weighted average of squared deviations, where the weights are probabilities from the distribution.

  If most of the probability distribution is close to μ, then σ^2 will be relatively small.

  However, if there are x values far from μ that have large $p(x)$, then σ^2 will be quite large.

  Very roughly σ can be interpreted as the size of a representative deviation from the mean value μ.

The Variance of X

- So if $\sigma = 10$, then in a long sequence of observed X values, some will deviate from μ by more than 10 while others will be closer to the mean than that—a typical deviation from the mean will be something on the order of 10.

Example 3.24

- A library has an upper limit of 6 on the number of videos that can be checked out to an individual at one time. Consider only those who check out videos, and let *X* denote the number of videos checked out to a randomly selected individual. The pmf of *X* is as follows:

x	1	2	3	4	5	6
$p(x)$	.30	.25	.15	.05	.10	.15

The expected value of *X* is easily seen to be μ = 2.85.

Example 3.24

cont'd

- The variance of *X* is then

$$V(X) = \sigma^2 = \sum_{x=1}^{6} (x - 2.85)^2 \cdot p(x)$$

$$= (1 - 2.85)^2(.30) + (2 - 2.85)^2(.25) + \ldots + (6 - 2.85)^2(.15) = 3.2275$$

- The standard deviation of *X* is $\sigma = \sqrt{3.2275} = 1.800$.

The Variance of X

- When the pmf $p(x)$ specifies a mathematical model for the distribution of population values, both σ^2 and σ measure the spread of values in the population; σ^2 is the population variance, and σ is the population standard deviation.

A Shortcut Formula for σ^2

- The number of arithmetic operations necessary to compute σ^2 can be reduced by using an alternative formula.

$$V(X) = \sigma^2 = \left[\sum_D x^2 \cdot p(x)\right] - \mu^2 = E(X^2) - [E(X)]^2$$

- In using this formula, $E(X^2)$ is computed first without any subtraction; then $E(X)$ is computed, squared, and subtracted (once) from $E(X^2)$.

Variance of a Linear Function

- The variance of $h(X)$ is the expected value of the squared difference between $h(X)$ and its expected value:

$$V[h(X)] = \sigma^2_{h(X)} = \sum_D \{h(x) - E[h(X)]\}^2 \cdot p(x) \qquad \textbf{(3.13)}$$

When $h(X) = aX + b$, a linear function,

$$h(x) - E[h(X)] = ax + b - (a\mu + b) = a(x - \mu)$$

Substituting this into (3.13) gives a simple relationship between $V[h(X)]$ and $V(X)$:

Variance of a Linear Function

Proposition

$$V(aX + b) = \sigma^2_{aX+b} = a^2 \cdot \sigma^2_X \quad \text{and} \quad \sigma_{aX+b} = |a| \cdot \sigma_X$$

In particular,

$$\sigma_{aX} = |a| \cdot \sigma_X, \quad \sigma_{X+b} = \sigma_X \tag{3.14}$$

The absolute value is necessary because *a* might be negative, yet a standard deviation cannot be.

Usually multiplication by *a* corresponds to a change in the unit of measurement (e.g., kg to lb or dollars to euros).

Rules of Variance

- According to the first relation in (3.14), the sd in the new unit is the original sd multiplied by the conversion factor.

- The second relation says that adding or subtracting a constant does not impact variability; it just rigidly shifts the distribution to the right or left.

Example 3.26

- In the computer sales scenario of Example 3.23, $E(X) = 2$ and

$E(X^2) = (0)^2(.1) + (1)^2(.2) + (2)^2(.3) + (3)^2(.4) = 5$

so, $V(X) = 5 - (2)^2 = 1$. The profit function $h(X) = 800X - 900$ then has variance $(800)^2 \cdot V(X) = (640{,}000)(1) = 640{,}000$ and standard deviation 800.

Variance

- What was the variance of a set of observations?
- The variance σ^2 of a random variable X with distribution f(x) is

$$\sigma^2 = E[(X - \mu)^2] = \Sigma_x (x - \mu)^2 f(x) \quad \text{if discrete, or}$$

$$\sigma^2 = E[(X - \mu)^2] = \int_x (x - \mu)^2 f(x)\,dx \quad \text{if continuous}$$

- An equivalent and easier computational formula, also easy to remember, is

$$\sigma^2 = E[X^2] - E[X]^2 = E[X^2] - \mu^2$$

  - "The expected value of X^2 - the expected value of X...squared."
  - Derivation from the previous formula is simple.

Expected Values, Covariance, and Correlation of Joint Distribution

- Any function $h(X)$ of a single rv X is itself a random variable.

- However, to compute $E[h(X)]$, it is not necessary to obtain the probability distribution of $h(X)$; instead, $E[h(X)]$ is computed as a weighted average of $h(x)$ values, where the weight function is the pmf $p(x)$ or pdf $f(x)$ of X.

- A similar result holds for a function $h(X, Y)$ of two jointly distributed random variables.

Expected Values, Covariance, and Correlation

- **Proposition**

Let X and Y be jointly distributed rv's with pmf $p(x, y)$ or pdf $f(x, y)$ according to whether the variables are discrete or continuous. Then the expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X, Y)}$, is given by

$$E[h(X, Y)] = \begin{cases} \displaystyle\sum_x \sum_y h(x, y) \cdot p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y)\, dx\, dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

Expected Value for a Joint Distribution

- If X and Y are random variables with joint probability distribution f (x,y). The mean or expected value $\mu_{g(X,Y)}$ of the random variable g (X,Y) is

  $\mu_{g(X,Y)} = E\,[g(X,Y)] = \Sigma_x \Sigma_y\, g(x,y)\, f(x,y)$ if discrete, or

  $\mu_{g(X,Y)} = E\,[g(X,Y)] = \int_x \int_y g(x,y)\, f(x,y)\, dy\, dx$ if continuous

- Note that the mean of a distribution is a single value, so it doesn't make sense to talk of the mean the distribution f (x,y).

Example 5.13

- Five friends have purchased tickets to a certain concert. If the tickets are for seats 1–5 in a particular row and the tickets are randomly distributed among the five, what is the expected number of seats separating any particular two of the five?
- Let *X* and *Y* denote the seat numbers of the first and second individuals, respectively. Possible (*X*, *Y*) pairs are {(1, 2), (1, 3), . . . , (5, 4)}, and the joint pmf of (*X*, *Y*) is

$$p(x, y) = \begin{cases} \frac{1}{20} & x = 1, \ldots, 5;\ y = 1, \ldots, 5;\ x \neq y \\ 0 & \text{otherwise} \end{cases}$$

cont'd

Example 5.13

- The number of seats separating the two individuals is $h(X, Y) = |X - Y| - 1$.
- The accompanying table gives $h(x, y)$ for each possible (x, y) pair.

$h(x, y)$		x = 1	2	3	4	5
	1	—	0	1	2	3
	2	0	—	0	1	2
y	3	1	0	—	0	1
	4	2	1	0	—	0
	5	3	2	1	0	—

Example 5.13

cont'd

- Thus

$$E[h(X, Y)] = \sum_{(x,\, y)} \sum h(x, y) \cdot p(x, y)$$

$$= \sum_{\substack{x=1 \\ x \neq y}}^{5} \sum_{y=1}^{5} (|x - y| - 1) \cdot \frac{1}{20}$$

$$= 1$$

Covariance

- If X and Y are random variables with joint probability distribution f (x,y), the covariance, σ_{XY} , of X and Y is defined as

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

- The better computational formula for covariance is

$$\sigma_{XY} = E\,(XY) - \mu_X \mu_Y$$

- Note that although the standard deviation σ can't be negative, the covariance σ_{XY} can be negative.
- Covariance will be useful later when looking at the linear relationship between two random variables.

Covariance

- When two random variables *X* and *Y* are not independent, it is frequently of interest to assess how strongly they are related to one another.

The **covariance** between two rv's X and Y is

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \begin{cases} \displaystyle\sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) & X, Y \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x, y)\,dx\,dy & X, Y \text{ continuous} \end{cases}$$

Covariance

- That is, since $X - \mu_X$ and $Y - \mu_Y$ are the deviations of the two variables from their respective mean values, the covariance is the expected product of deviations. Note that $\text{Cov}(X, X) = E[(X - \mu_X)^2] = V(X)$.

- The rationale for the definition is as follows.

- Suppose X and Y have a strong positive relationship to one another, by which we mean that large values of X tend to occur with large values of Y and small values of X with small values of Y.

Covariance

- Then most of the probability mass or density will be associated with $(x - \mu_X)$ and $(y - \mu_Y)$, either both positive (both X and Y above their respective means) or both negative, so the product $(x - \mu_X)(y - \mu_Y)$ will tend to be positive.

- Thus for a strong positive relationship, Cov(X, Y) should be quite positive.

- For a strong negative relationship, the signs of $(x - \mu_X)$ and $(y - \mu_Y)$ will tend to be opposite, yielding a negative product.

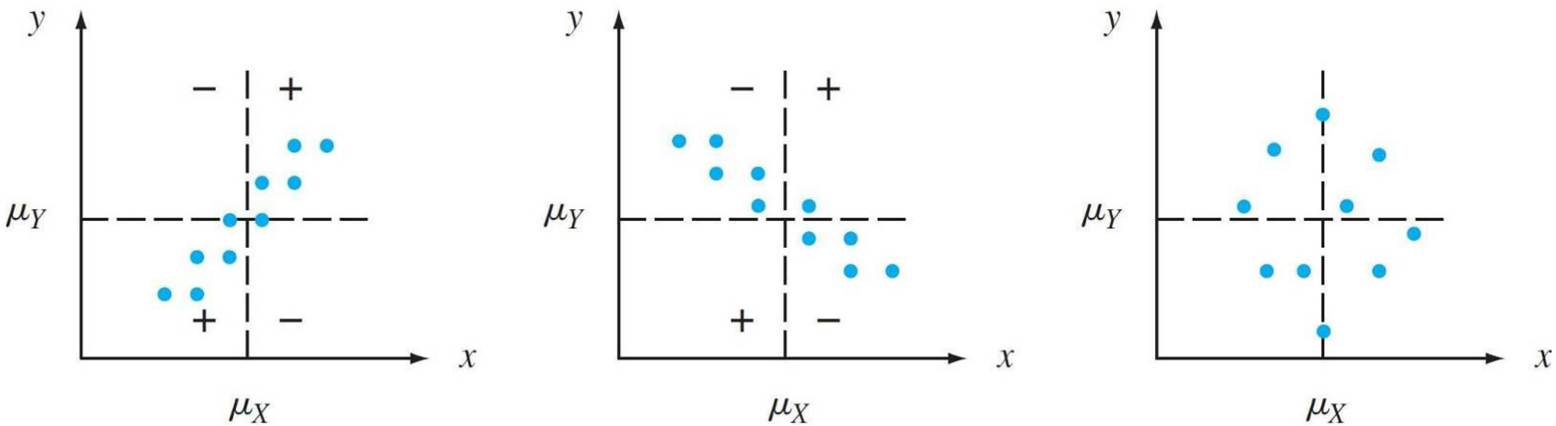
Covariance

- Thus for a strong negative relationship, Cov(X, Y) should be quite negative.

- If X and Y are not strongly related, positive and negative products will tend to cancel one another, yielding a covariance near 0.

Covariance

- Figure 5.4 illustrates the different possibilities. The covariance depends on *both* the set of possible pairs and the probabilities. In Figure 5.4, the probabilities could be changed without altering the set of possible pairs, and this could drastically change the value of Cov(*X*, *Y*).



$p(x, y) = 1/10$ for each of ten pairs corresponding to indicated points:

(a) positive covariance; (b) negative covariance; (c) covariance near zero

Figure 5.4

Example 5.15

- The joint and marginal pmf's for
  - *X* = automobile policy deductible amount and
  - *Y* = homeowner policy deductible amount in Example 5.1 were

	$p(x, y)$	y = 500	1000	5000
	100	.30	.05	0
x	500	.15	.20	.05
	1000	.10	.10	.05

x	100	500	1000
$p_X(x)$	.35	.40	.25

y	500	1000	5000
$p_Y(y)$	.55	.35	.10

from which $\mu_X = \Sigma x p_X(x) = 485$ and $\mu_Y = 1125$.

cont'd

Example 5.15

- Therefore,

$$\text{Cov}(X, Y) = \sum_{(x,\,y)}\sum (x - 485)(y - 1125)p(x, y)$$

$$= (100 - 485)(500 - 1125)(.30) + \ldots$$

$$+ (1000 - 485)(5000 - 1125)(.05)$$

$$= 136{,}875$$

Covariance

- The following shortcut formula for Cov(X, Y) simplifies the computations.

$$\text{Cov}(X, Y) = E(XY) - \mu_X \cdot \mu_Y$$

- According to this formula, no intermediate subtractions are necessary; only at the end of the computation is $\mu_X \cdot \mu_Y$ subtracted from $E(XY)$. The proof involves expanding $(X - \mu_X)(Y - \mu_Y)$ and then carrying the summation or integration through to

Correlation Coefficient

- If X and Y are random variables with covariance σ_{XY} and standard deviations σ_X and σ_Y respectively, the correlation coefficient ρ_{XY} is defined as

$$\rho_{XY} = \sigma_{XY} / (\sigma_X \sigma_Y)$$

Correlation coefficient notes:

- What are the units of ρ_{XY} ?
- What is the possible range of ρ_{XY} ?
- What is the meaning of the correlation coefficient?
- If $\rho_{XY} = 1$ or -1, then there is an exact linear relationship between Y and X (i.e., $Y = a + bX$). If $\rho_{XY} = 1$, then $b > 0$, and if $\rho_{XY} = -1$, then $b < 0$.
- Can show this by calculating the covariance of X and a + bX, which simplifies to $b / \sqrt{b^2} = 1$.

Correlation

- **Definition**

The **correlation coefficient** of X and Y, denoted by Corr(X, Y), $\rho_{X,Y}$, or just ρ, is defined by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Example 5.17

- It is easily verified that in the insurance scenario of Example 5.15, $E(X^2) = 2{,}987{,}500$

$\sigma_X^2 = 353{,}500 - (485)^2 = 118{,}275$,

$\sigma_X = 343.911$, $E(Y^2) = 2{,}987{,}500$,

σ_Y^2 =1,721,875, and $\sigma_Y = 1312.202$.

- This gives $$\rho = \frac{136.875}{(343{,}911)(1312.202)} = .303$$

Correlation

- The following proposition shows that ρ remedies the defect of Cov(X, Y) and also suggests how to recognize the existence of a strong (linear) relationship.

1. If a and c are either both positive or both negative,

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

2. For any two rv's X and Y, $-1 \leq \rho \leq 1$. The two variables are said to be **uncorrelated** when $\rho = 0$.

Correlation

- If we think of $p(x, y)$ or $f(x, y)$ as prescribing a mathematical model for how the two numerical variables X and Y are distributed in some population (height and weight, verbal SAT score and quantitative SAT score, etc.), then ρ is a population characteristic or parameter that measures how strongly X and Y are related in the population.

- In Chapter 12, we will consider taking a sample of pairs $(x_1, y_1), \ldots, (x_n, y_n)$ from the population.

- The sample correlation coefficient r will then be defined and used to make inferences about ρ.

Correlation

- The correlation coefficient ρ is actually not a completely general measure of the strength of a relationship.

1. If X and Y are independent, then $\rho = 0$, but $\rho = 0$ does not imply independence.
2. $\rho = 1$ or -1 iff $Y = aX + b$ for some numbers a and b with $a \neq 0$.

Correlation

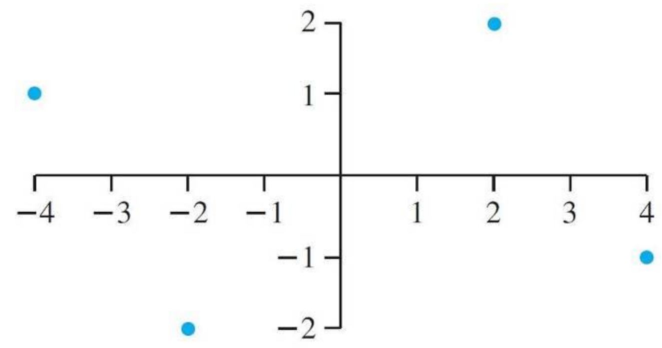
- This proposition says that ρ is a measure of the degree of **linear** relationship between *X* and *Y*, and only when the two variables are perfectly related in a linear manner will ρ be as positive or negative as it can be.
- However, if / *p* / << 1, there may still be a strong relationship between the two variables, just one that is not linear.
- And even if | *p* | is close to 1, it may be that the relationship is really nonlinear but can be well approximated by a straight line.

Example 5.18

- Let X and Y be discrete rv's with joint pmf

$$p(x, y) = \begin{cases} \frac{1}{4} & (x, y) = (-4, 1), (4, -1), (2, 2), (-2, -2) \\ 0 & \text{otherwise} \end{cases}$$

- The points that receive positive probability mass are identified on the (x, y) coordinate system in Figure 5.5.



The population of pairs for Example 18

Figure 5.5

Example 5.18

cont'd

- It is evident from the figure that the value of X is completely determined by the value of Y and vice versa, so the two variables are completely dependent. However, by symmetry $\mu_X = \mu_Y = 0$ and

$$E(XY) = (-4)\frac{1}{4} + (-4)\frac{1}{4} + (4)\frac{1}{4} + (4)\frac{1}{4} = 0$$

- The covariance is then $\text{Cov}(X,Y) = E(XY) - \mu_X \cdot \mu_Y = 0$ and thus $\rho_{X,Y} = 0$. Although there is perfect dependence, there is also complete absence of any linear relationship!

Correlation

- A value of ρ near 1 does not necessarily imply that increasing the value of X *causes* Y to increase. It implies only that large X values are *associated* with large Y values.
- For example, in the population of children, vocabulary size and number of cavities are quite positively correlated, but it is certainly not true that cavities cause vocabulary to grow. Instead, the values of both these variables tend to increase as the value of age, a third variable, increases. For children of a fixed age, there is probably a low correlation between number of cavities and vocabulary size.
- In summary, association (a high correlation) is not the same as causation.

The Bivariate Normal Distribution

- Just as the most useful univariate distribution in statistical practice is the normal distribution, the most useful joint distribution for two rv's *X* and *Y* is the bivariate normal distribution. The pdf is somewhat complicated:

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp$$

$$\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right)$$

The Bivariate Normal Distribution

- A graph of this pdf, the density surface, appears in Figure 5.6. It follows (after some tricky integration) that the marginal distribution of *X* is normal with mean value μ_1 and standard deviation σ_1, and similarly the marginal distribution of *Y* is normal with mean μ_2 and standard deviation σ_2. The fifth parameter of the distribution
- is ρ, which can be shown to be the correlation coefficient between *X* and *Y*.

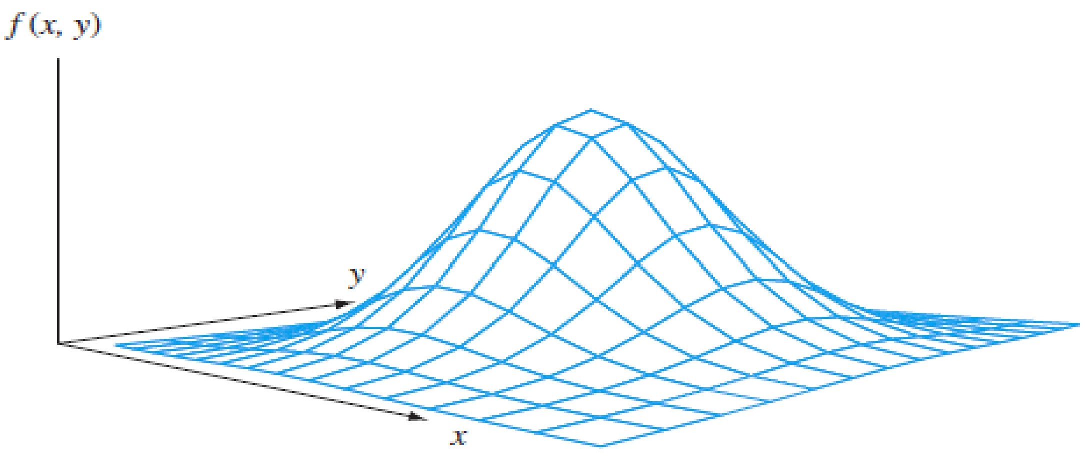


Figure 5.6 A graph of the bivariate normal pdf

The Bivariate Normal Distribution

- It is not at all straightforward to integrate the bivariate normal pdf in order to calculate probabilities. Instead, selected software packages employ numerical integration techniques for this purpose.
- Many students applying for college take the SAT, which for a few years consisted of three components: Critical Reading, Mathematics, and Writing. While some colleges used all three components to determine admission, many only looked at the first two (reading and math).

The Bivariate Normal Distribution

- Let *X* and *Y* denote the Critical Reading and Mathematics scores, respectively, for a randomly selected student. According to the College Board website, the $\mu_1 = 496, \sigma_1 = 114, \mu_2 = 514, \sigma_2 = 117$.king the exam in Fall 2012 had the following characteristics:
  - Suppose that *X* and *Y* have (approximately, since both variables are discrete) a bivariate normal distribution with correlation coefficient $\rho = .25$. The Matlab software package gives $P(X \leq 650, Y \leq 650) = P$(both scores are at most 650) = .8097.

The Bivariate Normal Distribution

- It can also be shown that the conditional distribution of *Y* given that *X* = *x* is normal. This can be seen geometrically by slicing the density surface with a plane perpendicular to the (*x*, *y*) passing through the value *x* on that axis; the result is a normal curve sketched out on the slicing plane. The conditional mean value is $\mu_{Y\cdot x} = (\mu_2 - \rho\mu_1\sigma_2/\sigma_1) + \rho\sigma_2 x/\sigma_1$
- a linear function of *x*, and the conditional variance is $\sigma^2_{Y\cdot x} = (1 - \rho^2)\sigma^2_2.$
- The closer the correlation coefficient is to 1 or 21, the less variability there is in the conditional distribution. Analogous results hold for the conditional distribution of *X* given that *Y* = *y*.

The Bivariate Normal Distribution

- The bivariate normal distribution can be generalized to the *multivariate normal distribution*. Its density function is quite complicated, and the only way to write it compactly is to employ matrix notation.
- If a collection of variables has this distribution, then the marginal distribution of any single variable is normal, the conditional distribution of any single variable given values of the other variables is normal, the joint marginal distribution of any pair of variables is bivariate normal, and the joint marginal distribution of any subset of three or more of the variables is again multivariate normal.

Thank You

"We trust in GOD, all others must bring data"