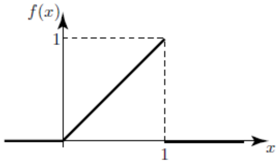
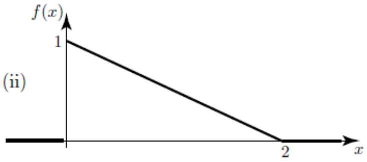
Which of the following are not probability density functions?

(i)



(ii)



(iii) $f(x) = \begin{cases} x^2 - 4x + \frac{10}{3}, & 0 \le x \le 3 \\ 0, & \text{elsewhere} \end{cases}$

Binomial Probability-Mass Function. . .

Let X be a binomial random variable. Then, its probability-mass function is:

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \qquad (1)$$

for $x = 0, 1, 2, \ldots, n$.

The values of n and p are called the *parameters* of the distribution.

To understand (1), note that:

- The probability for observing *any* sequence of n independent trials that contains x successes and $n - x$ failures is $p^n(1-p)^{n-x}$.
- The total number of such sequences is equal to

$$\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}$$

(i.e., the total number of possible combinations when we randomly select x objects out of n objects).

Binomial Mean and Variance. . .

It can be shown that

$$\mu = E(X) = np$$

and

$$\sigma^2 = V(X) = np(1-p) .$$

For the previous example, we have

- $E(X) = 10 \cdot 0.25 = 2.5$.
- $V(X) = 10 \cdot (0.25) \cdot (1 - 0.25) = 1.875$.

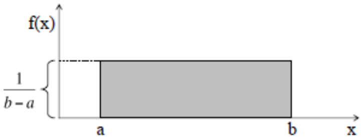
Uniform Distribution. . .

The uniform distribution is the simplest example of a continuous probability distribution. A random variable X is said to be uniformly distributed if its density function is given by:

$$f(x) = \frac{1}{b-a} \qquad (5)$$

for $-\infty < a \leq x \leq b < \infty$.

Visually, we have



where the shaded region has area $(b-a)[1/(b-a)] = 1$ (width times height).

The values a and b are the parameters of the uniform distribution. It can be shown that

$$E(X) = \frac{a+b}{2} \quad \text{and} \quad V(X) = \frac{(b-a)^2}{12}.$$

The *standard* uniform density has parameters $a = 0$ and $b = 1$; and hence $f(x) = 1$ for $0 \leq x \leq 1$ and 0 otherwise. The Excel function RAND() "pretends" to generate independent samples from this density function.

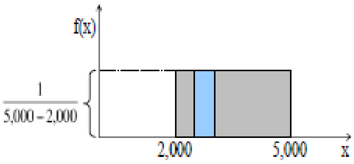
Example: Gasoline Sales

Suppose the amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2,000 gallons and a maximum of 5,000 gallons.

What is the probability that daily sales will fall between 2,500 gallons and 3,000 gallons? Answer:

$$\begin{aligned} P(2500 < X \leq 3000) &= \frac{1}{5000-2000}(3000-2500) \\ &= 0.1667\,. \end{aligned}$$

Visually, we have



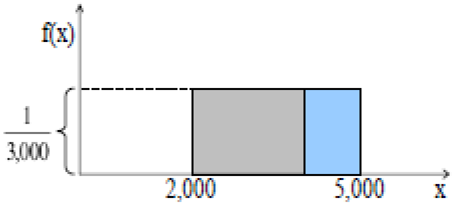
and the answer corresponds to the area in blue.

What is the probability that the service station will sell *at least* 4,000 gallons? Answer:

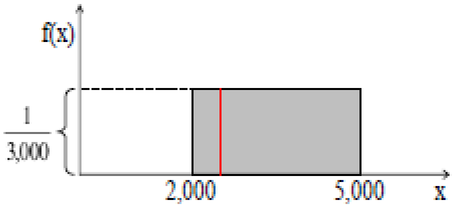
$$P(X > 4000) = \frac{1}{5000 - 2000}(5000 - 4000)$$

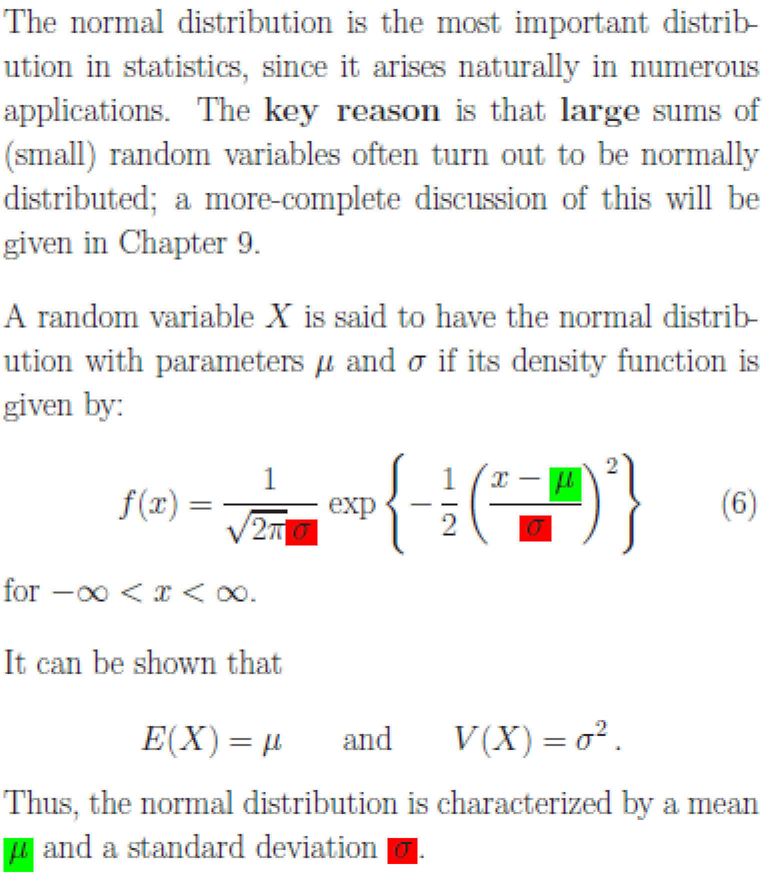
$$= 0.3333\,.$$

Visually, we have



What is the probability that the service station will sell *exactly* 2,500 gallons? Answer: $P(X = 2500) = 0$, since the area of a “vertical line” at 2,500 is 0.





The normal distribution is the most important distribution in statistics, since it arises naturally in numerous applications. The **key reason** is that **large** sums of (small) random variables often turn out to be normally distributed; a more-complete discussion of this will be given in Chapter 9.

A random variable X is said to have the normal distribution with parameters μ and σ if its density function is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{6}$$

for $-\infty < x < \infty$.

It can be shown that

$$E(X) = \mu \qquad \text{and} \qquad V(X) = \sigma^2 .$$

Thus, the normal distribution is characterized by a mean μ and a standard deviation σ.

Calculating Normal Probabilities...

A normal distribution whose mean is 0 and standard deviation is 1 is called the **standard** normal distribution. In this case, the density function assumes the simpler form:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{7}$$

for $-\infty < x < \infty$.

Table 3 in Appendix B of the text can be used to calculate probabilities associated with the standard normal distribution. The Excel function NORMSDIST() (where "S" is for "standard") can also be used.

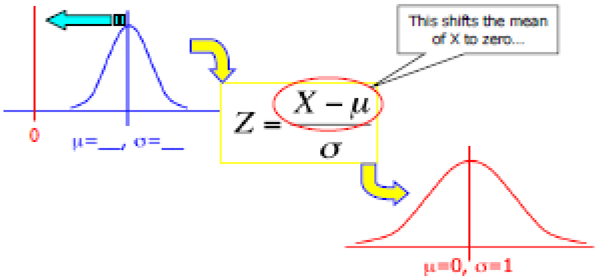
Denote by Z a random variable that follows the standard normal distribution. Then, Table 3 gives the probability $P(0 < Z \leq z)$ for any nonnegative value z; whereas NORMSDIST() returns $P(Z \leq z)$ for any z from $-\infty$ to ∞, i.e., values of the *cumulative* distribution function.

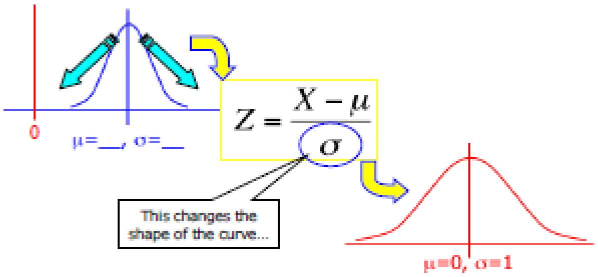
For general parameter values, the Excel function NORMDIST() (without "S" in the middle) can be used directly. However, ...

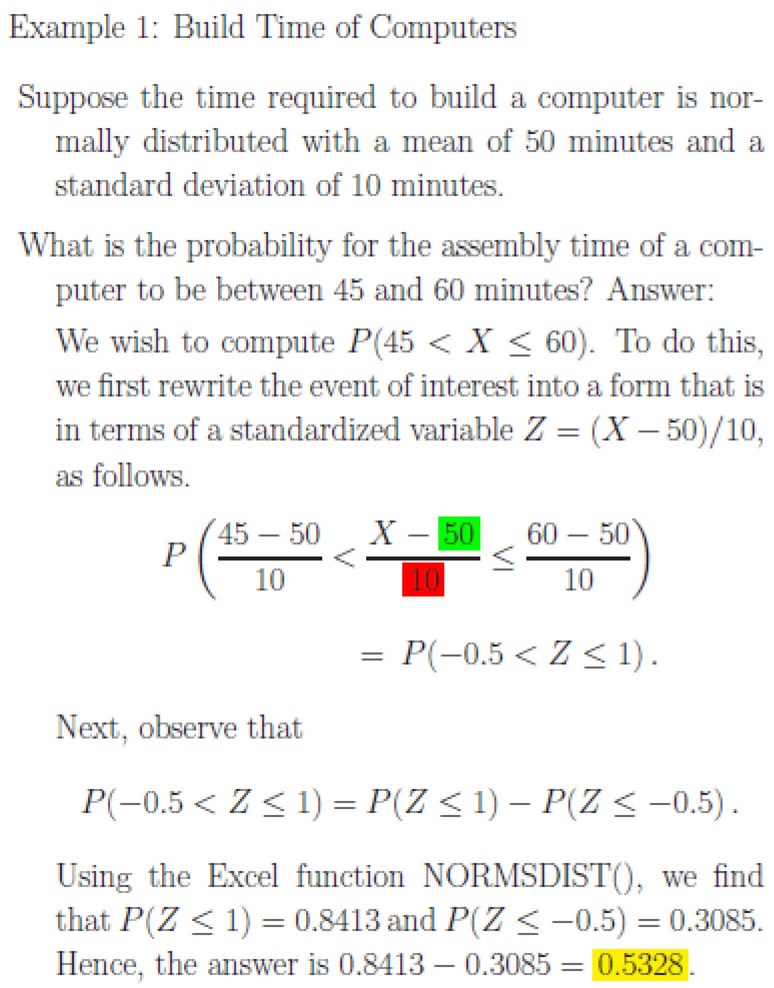
A standard practice is to convert a normal random variable X with arbitrary parameters μ and σ into a **standardized** normal random variable Z with parameters 0 and 1 via the transformation:

$$Z = \frac{X - \mu}{\sigma}; \tag{8}$$

this is illustrated in:


This shifts the mean of X to zero...
0
μ=__, σ=__
$Z = \frac{X - \mu}{\sigma}$
μ=0, σ=1


0
μ=__, σ=__
$Z = \frac{X - \mu}{\sigma}$
This changes the shape of the curve...
μ=0, σ=1



Example 1: Build Time of Computers

Suppose the time required to build a computer is normally distributed with a mean of 50 minutes and a standard deviation of 10 minutes.

What is the probability for the assembly time of a computer to be between 45 and 60 minutes? Answer:

We wish to compute $P(45 < X \leq 60)$. To do this, we first rewrite the event of interest into a form that is in terms of a standardized variable $Z = (X - 50)/10$, as follows.

$$P\left(\frac{45-50}{10} < \frac{X-50}{10} \leq \frac{60-50}{10}\right)$$

$$= P(-0.5 < Z \leq 1)\,.$$

Next, observe that

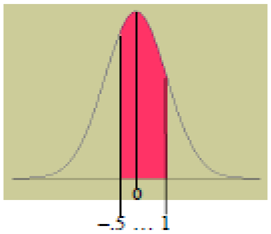
$$P(-0.5 < Z \leq 1) = P(Z \leq 1) - P(Z \leq -0.5)\,.$$

Using the Excel function NORMSDIST(), we find that $P(Z \leq 1) = 0.8413$ and $P(Z \leq -0.5) = 0.3085$. Hence, the answer is $0.8413 - 0.3085 = 0.5328$.

Table 3 can also be used for this calculation:

$$
\begin{aligned}
&P(-0.5 < Z \leq 1) \\
&\quad = P(-0.5 < Z \leq 0) + P(0 < Z \leq 1) \\
&\quad = P(0 < Z \leq 0.5) + P(0 < Z \leq 1) \\
&\quad = 0.1915 + 0.3414 \\
&\quad = 0.5328\ ,
\end{aligned}
$$

where the first equality follows from

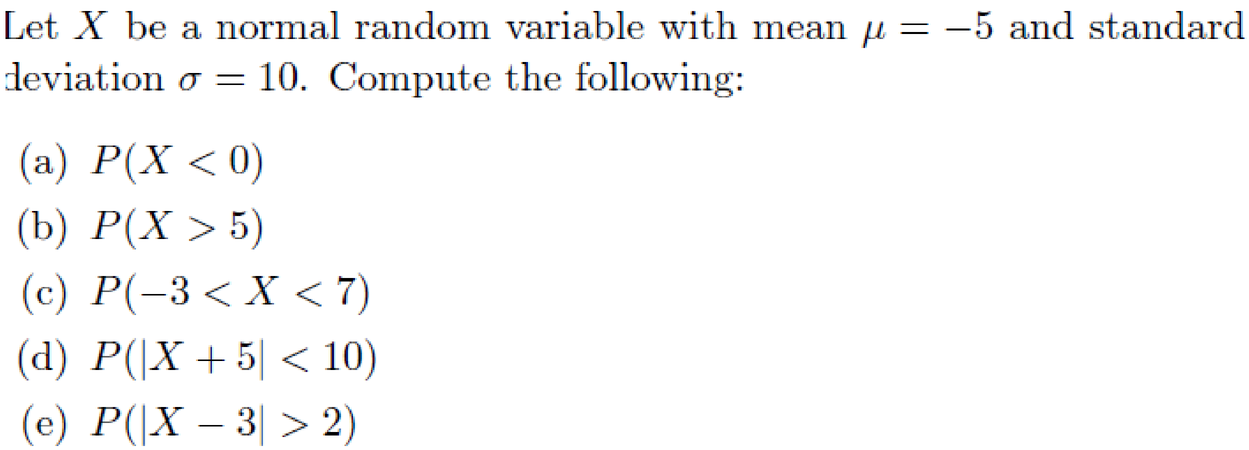


the second equality is due to the fact that the normal density curve is *symmetric*, and the third equality is from Table 3.

Is it reasonable to assume that the build time is normally distributed? Reasoning: The build time can be thought of as the sum of times needed to build many individual components.

GAUSSIAN PROCESSES

- Gaussian processes are non-parametric.
- They provide a structured method of model and parameter selection.
- A Gaussian process is defined by a mean and covariance function.
- Learning takes the form of setting the hyper-parameters. Occam's Razor is implicit.
- GP's can be used for regression or classification.



Let X be a normal random variable with mean $\mu = -5$ and standard deviation $\sigma = 10$. Compute the following:

(a) $P(X < 0)$

(b) $P(X > 5)$

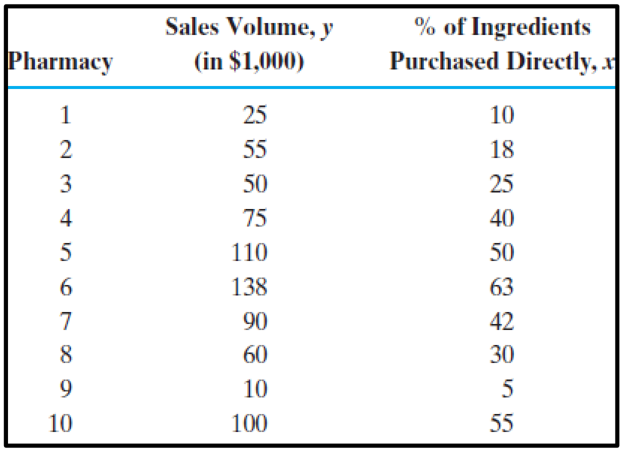
(c) $P(-3 < X < 7)$

(d) $P(|X + 5| < 10)$

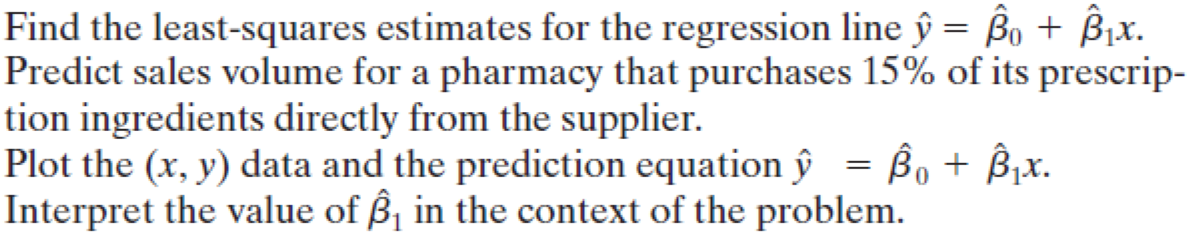
(e) $P(|X - 3| > 2)$

[statistical_inference.PDF]

Data from a sample of 10 pharmacies are used to examine the relation between prescription sales volume and the percentage of prescription ingredients purchased directly from the supplier. The sample data are shown in Table below:



Pharmacy	Sales Volume, y (in \$1,000)	% of Ingredients Purchased Directly, x
1	25	10
2	55	18
3	50	25
4	75	40
5	110	50
6	138	63
7	90	42
8	60	30
9	10	5
10	100	55



Find the least-squares estimates for the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
Predict sales volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.
Plot the (x, y) data and the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
Interpret the value of $\hat{\beta}_1$ in the context of the problem.

(a) Find the least-squares estimates for the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

(b) Predict sales volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.

(c) Plot the (x, y) data and the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

(d) Interpret the value of $\hat{\beta}_1$ in the context of the problem.

a. The equation can be calculated by virtually any statistical computer package; for example, here is abbreviated Minitab output:

```
MTB > Regress 'Sales' on 1 variable 'Directly'

The regression equation is
Sales = 4.70 + 1.97 Directly

Predictor     Coef     Stdev   t-ratio       p
Constant     4.698     5.952      0.79   0.453
Directly    1.9705    0.1545     12.75   0.000
```

To see how the computer does the calculations, you can obtain the least-squares estimates from Table 11.2.

	y	x	$y - \bar{y}$	$x - \bar{x}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	25	10	−46.3	−23.8	1,101.94	566.44
	55	18	−16.3	−15.8	257.54	249.64
	50	25	−21.3	−8.8	187.44	77.44
	75	40	3.7	6.2	22.94	38.44
	110	50	38.7	16.2	626.94	262.44
	138	63	66.7	29.2	1,947.64	852.64
	90	42	18.7	8.2	153.34	67.24
	60	30	−11.3	−3.8	42.94	14.44
	10	5	−61.3	−28.8	1,765.44	829.44
	100	55	28.7	21.2	608.44	449.44
Totals	713	338	0	0	6,714.60	3,407.60
Means	71.3	33.8				

$$S_{xx} = \sum (x - \overline{x})^2 = 3{,}407.6$$

$$S_{xy} = \sum (x - \overline{x})(y - \overline{y}) = 6{,}714.6$$

Substituting into the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{6{,}714.6}{3{,}407.6} = 1.9704778 \qquad \text{rounded to 1.97}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 71.3 - 1.9704778(33.8) = 4.6978519 \qquad \text{rounded to 4.70}$$

b. When $x = 15\%$, the predicted sales volume is $\hat{y} = 4.70 + 1.97(15) = 34.25$ (that is, \$34,250).

c. The (x, y) data and prediction equation are shown in Figure 11.11.

d. From $\hat{\beta}_1 = 1.97$, we conclude that if a pharmacy would increase by 1% the percentage of ingredients purchased directly, then the estimated increase in average sales volume would be \$1,970.

Elementary Statistical Techniques of Analysis

Most commonly used statistical techniques of analysis data are:

1. Calculating frequency of distribution in percentages of items under study.
2. Testing data for normality of distribution Skewness Kurtosis and mode.
3. Calculating percentiles and percentile ranks.
4. Calculating measures of central tendency-Mean, Median and Mode and establishing Norms.
5. Calculating measures of dispersion-Standard deviation, Mean deviation, quartile deviation and range.
6. Calculating measures of relationship-Coefficients of Correlation, Reliability by the Rank difference and Product moment method.
7. Graphical presentation of data-Frequency polygon curve, Histogram, Cumulative frequency polygon and Ogive, etc.

While analysis their data investigator usually makes use of as many of the above simple statistical devices as necessary for the purpose for their study. There are some other complicated devices of statistical analysis listed below which researcher use in particular experimental or complex casual-comparative studies and investigations.

Special Statistical Techniques of Analysis

The following are the special statistical techniques of analysis:

1. Test of students '*t*' and analysis of variance for testing significance of differences between statistics especially between Means.
2. Chi-square test for testing null hypothesis.
3. Calculation of Biserial '*r*' and Tetrachoric '*r*' for finding out relationship between different phenomena in complex situations.
4. Calculation of partial and multiple correlation and of Bivariate and Multivariate Regression Equations for findings out casual relationship between various phenomena involved in a situation.
5. Factorial Analysis for the purpose of analysing the composition of certain complex phenomena.
6. Analysis of co-variance for estimating the true effect of the treatment after adjusting the initial effect.

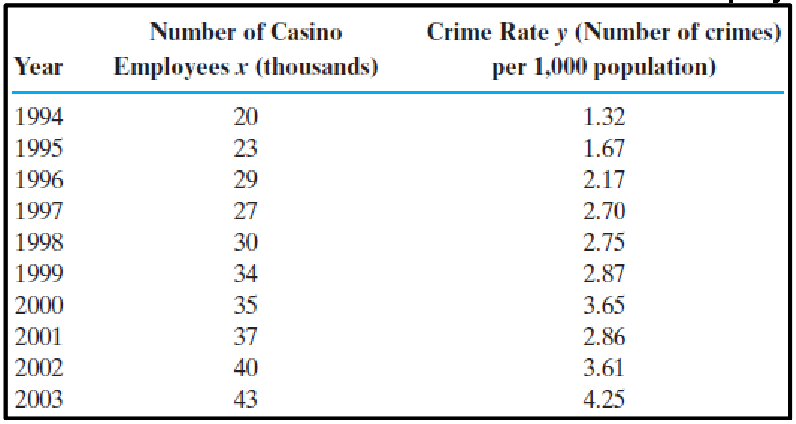
The **correlation coefficient** measures the strength of the linear relationship between two quantitative variables. The correlation coefficient is usually denoted as *r*.

Suppose we have data on two variables x and y collected from n individuals or objects with means and standard deviations of the variables given as $\overline{x}$ and s_x for the x-variable and $\overline{y}$ and s_y for the y-variable. The correlation r between x and y is computed as

$$r = \frac{1}{n - 1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

In computing the correlation coefficient, the two variables x and y are standardized to be unit-free variables. The standardized x-variable for the ith individual, $\left(\frac{x_i - \overline{x}}{s_x}\right)$, measures how many standard deviations x_i is above or below the x-mean. Thus, the correlation coefficient, r, is a unit-free measure of the strength of linear relationship between the quantitative variables, x and y.

Table: Crime rate as a function of number of casino employees



Year	Number of Casino Employees x (thousands)	Crime Rate y (Number of crimes) per 1,000 population)
1994	20	1.32
1995	23	1.67
1996	29	2.17
1997	27	2.70
1998	30	2.75
1999	34	2.87
2000	35	3.65
2001	37	2.86
2002	40	3.61
2003	43	4.25

Generally, the correlation coefficient, r, is a positive number if y tends to increase as x increases; r is negative if y tends to decrease as x increases; and r is nearly zero if there is either no relation between changes in x and changes in y or there is a nonlinear relation between x and y such that the patterns of increase and decrease in y (as x increases) cancel each other.

Some properties of r that assist us in the interpretation of relationship between two variables include the following:

1. A positive value for r indicates a positive association between the two variables, and a negative value for r indicates a negative association between the two variables.
2. The value of r is a number between -1 and $+1$. When the value of r is very close to ± 1, the points in the scatterplot will lie close to a straight line.
3. Because the two variables are standardized in the calculation of r, the value of r does not change if we alter the units of x or y. The same value of r will be obtained no matter what units are used for x and y. Correlation is a unit-free measure of association.
4. Correlation measures the degree of straight line relationship between two variables. The correlation coefficient does *not* describe the closeness of the points (x, y) to a curved relationship, no matter how strong the relationship.

What values of r indicate a "strong" relationship between y and x? Figure 3.31 displays 15 scatterplots obtained by randomly selecting 1,000 pairs (x_i, y_i) from 15 populations having bivariate normal distributions with correlations ranging from $-.99$ to $.99$. We can observe that unless $|r|$ is greater than .6, there is very little trend in the scatterplot.

Consider two events A and B with nonzero probabilities, $P(A)$ and $P(B)$. The **conditional probability** of event A given event B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The conditional probability of event B given event A is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

This definition for conditional probabilities gives rise to what is referred to as the *multiplication law*.

The **probability of the intersection** of two events A and B is

$$\begin{aligned} P(A \cap B) &= P(A)P(B|A) \\ &= P(B)P(A|B) \end{aligned}$$

calculated. When the intersection probability $P(A \cap B)$ and the individual probability $P(A)$ are known, we can compute $P(B|A)$. When we know $P(A)$ and $P(B|A)$, we can compute $P(A \cap B)$.

Bayes' Formula

In this section, we will show how Bayes' Formula can be used to update conditional probabilities by using sample data when available. These "updated" conditional probabilities are useful in decision making. A particular application of these techniques involves the evaluation of diagnostic tests. Suppose a meat inspector must decide whether a randomly selected meat sample contains *E. coli* bacteria. The inspector conducts a diagnostic test. Ideally, a positive result (Pos) would mean that the meat sample actually has *E. coli,* and a negative result (Neg) would imply that the meat sample is free of *E. coli.* However, the diagnostic test is occasionally in error. The results of the test may be a **false positive,** for which the test's indication of *E. coli* presence is incorrect, or a **false negative,** for which the test's conclusion of *E. coli* absence is incorrect. Large-scale screening tests are conducted to evaluate the accuracy of a given diagnostic test. For example, *E. coli* (E) is placed in 10,000 meat samples, and the diagnostic test yields a positive result for 9,500 samples and a negative result for 500 samples; that is, there are 500 false negatives out of the 10,000 tests. Another 10,000 samples have all traces of *E. coli* (NE) removed, and the diagnostic test yields a positive result for 100 samples and a negative result for 9,900 samples. There are 100 false positives out of the 10,000 tests. We can summarize the results in Table 4.3.

TABLE 4.3
E. coli test data

Diagnostic Test Result	Meat Sample Status	
	E	NE
Positive	9,500	100
Negative	500	9,900
Total	10,000	10,000

Evaluation of test results is as follows:

$$\text{True positive rate} = P(\text{Pos}|\text{E}) = \frac{9{,}500}{10{,}000} = .95$$

$$\text{False positive rate} = P(\text{Pos}|\text{NE}) = \frac{100}{10{,}000} = .01$$

$$\text{True negative rate} = P(\text{Neg}|\text{NE}) = \frac{9{,}900}{10{,}000} = .99$$

$$\text{False negative rate} = P(\text{Neg}|\text{NE}) = \frac{500}{10{,}000} = .05$$

The **sensitivity** of the diagnostic test is the true positive rate—that is, *P*(test is positive|disease is present). The **specificity** of the diagnostic test is the true negative rate—that is, *P*(test is negative|disease is not present).

The primary question facing the inspector is to evaluate the probability of *E. coli* being present in the meat sample when the test yields a positive result—that is, the inspector needs to know *P*(E|Pos). Bayes' Formula provides us with a method to obtain this probability.

If *A* and *B* are any events whose probabilities are not 0 or 1, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}$$

The above formula was developed by Thomas Bayes in a book published in 1763. We will illustrate the application of Bayes' Formula by returning to the meat inspection example. We can use Bayes' Formula to compute *P*(E|Pos) for the meat inspection example. To make this calculation, we need to know the *rate* of *E. coli* in the type of meat being inspected. For this example, suppose that *E. coli* is present in 4.5% of all meat samples; that is, *E. coli* has prevalence *P*(E) = .045. We can then compute *P*(E|Pos) as follows:

$$\begin{aligned} P(\text{E}|\text{Pos}) &= \frac{P(\text{Pos}|\text{E})P(\text{E})}{P(\text{Pos}|\text{E})P(\text{E}) + P(\text{Pos}|\text{NE})P(\text{NE})} \\ &= \frac{(.95)(.045)}{(.95)(.045) + (.01)(1 - .045)} = .817 \end{aligned}$$

Thus, *E. coli* is truly present in 81.7% of the tested samples in which a positive test result occurs. Also, we can conclude that 18.3% of the tested samples indicated *E. coli* was present when in fact there was no *E. coli* in the meat sample.

A book club classifies members as heavy, medium, or light purchasers, and separate mailings are prepared for each of these groups. Overall, 20% of the members are heavy purchasers, 30% medium, and 50% light. A member is not classified into a group until 18 months after joining the club, but a test is made of the feasibility of using the first 3 months' purchases to classify members. The following percentages are obtained from existing records of individuals classified as heavy, medium, or light purchasers (Table 4.4):

First 3 Months'	Group (%)		
Purchases	Heavy	Medium	Light
0	5	15	60
1	10	30	20
2	30	40	15
3+	55	15	5

If a member purchases no books in the first 3 months, what is the probability that the member is a light purchaser? (*Note:* This table contains "conditional" percentages for each column.)

Solution Using the conditional probabilities in the table, the underlying purchase probabilities, and Bayes' Formula, we can compute this conditional probability.

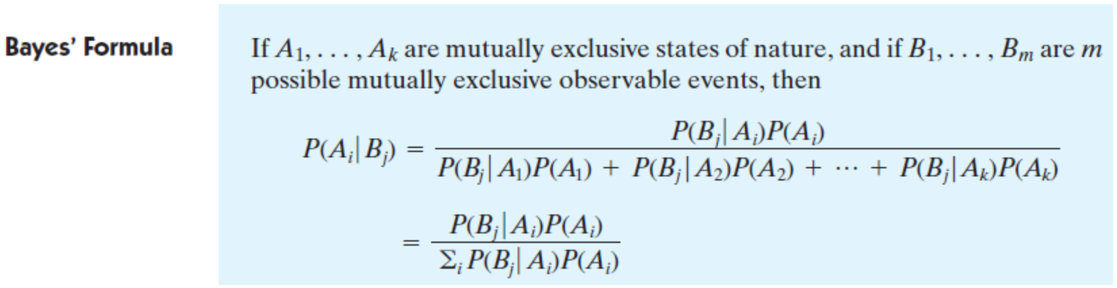
$$P(\text{light}|0)$$

$$= \frac{P(0|\text{light})P(\text{light})}{P(0|\text{light})P(\text{light}) + P(0|\text{medium})P(\text{medium}) + P(0|\text{heavy})P(\text{heavy})}$$

$$= \frac{(.60)(.50)}{(.60)(.50) + (.15)(.30) + (.05)(.20)}$$

$$= .845$$

These examples indicate the basic idea of Bayes' Formula. There is some number k of possible, mutually exclusive, underlying events $A_1, \ldots, A_k$, which are sometimes called the **states of nature.** Unconditional probabilities $P(A_1), \ldots,$ $P(A_k)$, often called **prior probabilities,** are specified. There are m possible, mutually exclusive, **observable events** $B_1, \ldots, B_m$. The conditional probabilities of each observable event given each state of nature, $P(B_i|A_i)$, are also specified, and these probabilities are called **likelihoods.** The problem is to find the **posterior probabilities** $P(A_i|B_i)$. *Prior* and *posterior* refer to probabilities before and after observing an event B_i.

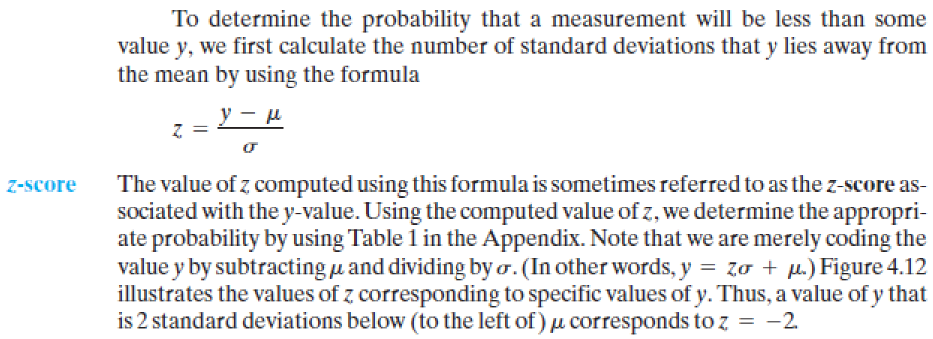


Bayes' Formula

If $A_1, \ldots, A_k$ are mutually exclusive states of nature, and if $B_1, \ldots, B_m$ are m possible mutually exclusive observable events, then

$$P(A_i|B_j) = \frac{P(B_j|A_i)P(A_i)}{P(B_j|A_1)P(A_1) + P(B_j|A_2)P(A_2) + \cdots + P(B_j|A_k)P(A_k)}$$

$$= \frac{P(B_j|A_i)P(A_i)}{\Sigma_i P(B_j|A_i)P(A_i)}$$

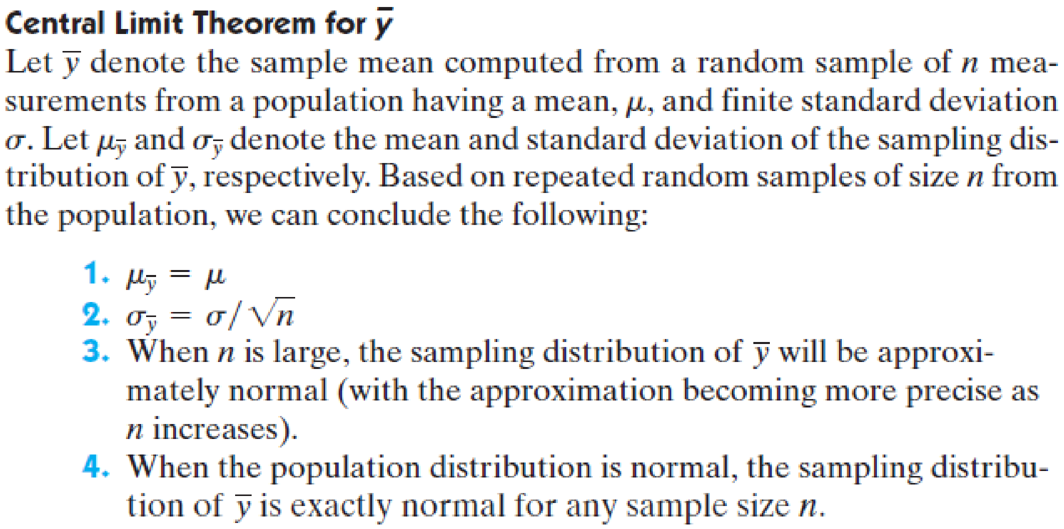


To determine the probability that a measurement will be less than some value y, we first calculate the number of standard deviations that y lies away from the mean by using the formula

$$z = \frac{y - \mu}{\sigma}$$

z-score

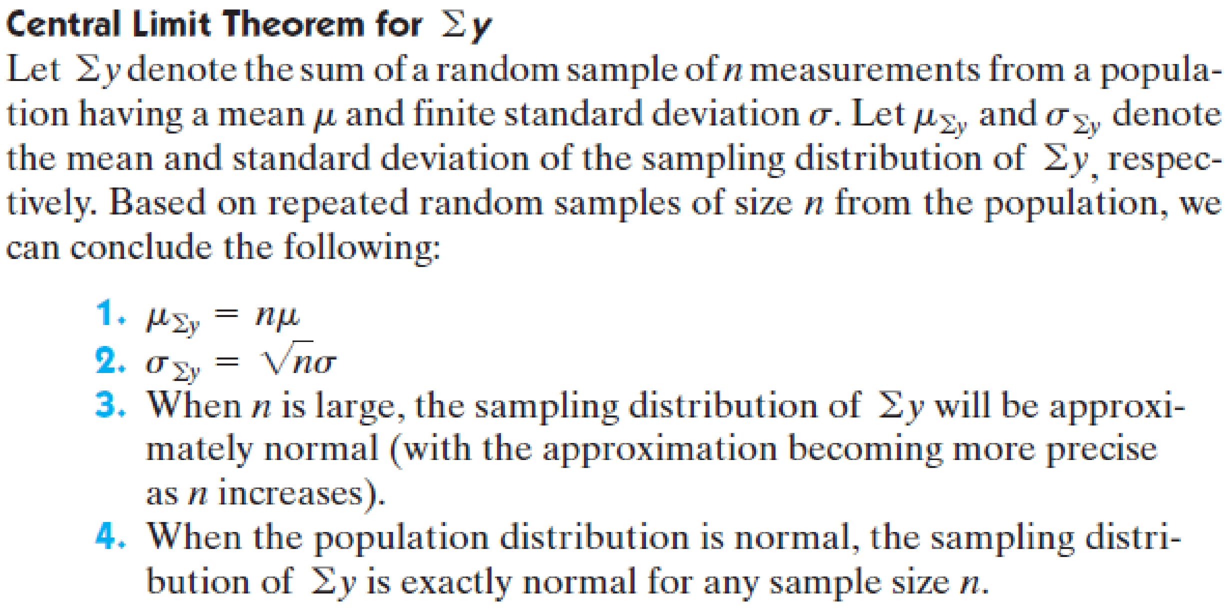
The value of z computed using this formula is sometimes referred to as the ***z*-score** associated with the y-value. Using the computed value of z, we determine the appropriate probability by using Table 1 in the Appendix. Note that we are merely coding the value y by subtracting μ and dividing by σ. (In other words, $y = z\sigma + \mu$.) Figure 4.12 illustrates the values of z corresponding to specific values of y. Thus, a value of y that is 2 standard deviations below (to the left of) μ corresponds to $z = -2$.



Central Limit Theorem for $\bar{y}$

Let $\bar{y}$ denote the sample mean computed from a random sample of n measurements from a population having a mean, μ, and finite standard deviation σ. Let $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ denote the mean and standard deviation of the sampling distribution of $\bar{y}$, respectively. Based on repeated random samples of size n from the population, we can conclude the following:

1. $\mu_{\bar{y}} = \mu$
2. $\sigma_{\bar{y}} = \sigma/\sqrt{n}$
3. When n is large, the sampling distribution of $\bar{y}$ will be approximately normal (with the approximation becoming more precise as n increases).
4. When the population distribution is normal, the sampling distribution of $\bar{y}$ is exactly normal for any sample size n.



Central Limit Theorem for Σy

Let Σy denote the sum of a random sample of n measurements from a population having a mean μ and finite standard deviation σ. Let $\mu_{\Sigma y}$ and $\sigma_{\Sigma y}$ denote the mean and standard deviation of the sampling distribution of Σy, respectively. Based on repeated random samples of size n from the population, we can conclude the following:

1. $\mu_{\Sigma y} = n\mu$
2. $\sigma_{\Sigma y} = \sqrt{n}\sigma$
3. When n is large, the sampling distribution of Σy will be approximately normal (with the approximation becoming more precise as n increases).
4. When the population distribution is normal, the sampling distribution of Σy is exactly normal for any sample size n.