Sheldon_M_Ross-Introduction_to_Probability_and_Statistics_FOR_ENGINEERS_AND_SCIENTISTS, Third Edition and_ALL_SOLVED_EXAMPLES_OF_CHAPTERS_in_order_8_7_9_6_5_4_3_2_1_TO_BE_PRACTICED/SOLVED_AGAIN_AS_GIVEN_IN_BOOK_TEXT.

Chapter_8: HYPOTHESIS TESTING: 8.1 A random sample from a population distribution, to test some particular hypothesis concerning set of parameters of population distribution eg. Parameter population mean. A statistical hypothesis is usually a statement about a set of parameters of a population distribution. It is called a hypothesis because it is not known whether or not it is true. A primary problem is to develop a procedure for determining whether or not the values of a random sample from this population are consistent with the hypothesis. For instance, consider a particular normally distributed population having an unknown mean value θ and known variance 1. The statement "θ is less than 1" is a statistical hypothesis that we could try to test by observing a random sample from this population. 8.2 SIGNIFICANCE LEVELS: Consider a population having distribution Fθ , where θ is unknown, and suppose we want to test a specific hypothesis about θ. We shall denote this hypothesis by H0 and call it the null hypothesis. (a) H0 : θ = 1 (b) H0 : θ ≤ 1 Note that the **null** hypothesis in (a), when true, completely specifies the population distribution, simple hypothesis; whereas the **null** hypothesis in (b) does not, composite hypothesis. The statistical test determined by the critical region C is the one that accepts H0 if (X1, X2, . . . , Xn) $\notin$ C and rejects H0 if (X1, . . . , Xn) ∈ C. $C = \left\{(X_1,\ldots,X_n) : \left|\frac{\sum_{i=1}^{n} X_i}{n} - 1\right| > \frac{1.96}{\sqrt{n}}\right\}$ This test calls for rejection of the null hypothesis that θ = 1 when the **sample average** differs from 1 by more than 1.96 divided by the square root of the sample size. A type I error, is said to result if the test **incorrectly** calls for rejecting H0 when it is indeed **correct**. The second, called a type II error, results if the test calls for **accepting** H0 when it is **false**. Level of significance of the test(**α**): when H0 is true, probability of a type **I error occurring** i.e. **being rejected** is **never** greater than α. For a given set of parameter values w, suppose we are interested in testing H0 : θ ∈ w. Approach to developing a test of H0, say at level of significance α, start by determining a **point** estimator of θ — say d(X). The hypothesis is then rejected if d (X) is "far away" from the region w. However, to determine how "far away" it need be to justify rejection of H0, we need to determine the probability **distribution** of d(X) when H0 is true. rejection **when** the point estimate of θ —that is, the sample average—is farther than $1.96/\sqrt{n}$ **away from 1** to meet a level of significance of α =.05 8.3 TESTS CONCERNING THE MEAN OF A NORMAL POPULATION 8.3.1 Case of Known Variance: Suppose that X1, . . . , Xn is a sample of size n from a normal distribution having an unknown mean μ and a known variance σ^2 and suppose we are interested in testing the null hypothesis H0 : μ = μ0 against the alternative hypothesis H1 : μ$\neq$ μ0 where μ0 is some specified constant. Since $\overline{X} = \sum_{i=1}^{n} X_i/n$ is a natural point estimator of μ, it seems reasonable to accept H0 if X is **not too far** from μ0. That is, the critical region of the test would be of the form $C = \{X_1,\ldots,X_n : |\overline{X} - \mu_0| > c\}$ for some suitably chosen value c. Thus, the significance level α test is to reject H0 and accept

$$\text{reject} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| > z_{\alpha/2}$$

$$\text{accept} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| \le z_{\alpha/2}$$

otherwise; or, equivalently, to, TEST when $z_{\alpha/2}$ the standard normal **density** function which is the density of the test statistic $\sqrt{n}(\overline{X} - \mu_0)/\sigma$ when H0 is true. at say α = .05, α/2=.025 less than $z_{.025} = 1.96$ say α = .1, α/2=.05 less stringent $z_{.05} = 1.645$The "CORRECT" level of significance to use in a given situation depends on the individual circumstances involved in that situation. For instance, if rejecting a null hypothesis H0 would result in large costs that would thus be lost if H0 were indeed true, then we might elect to be quite conservative and so choose a significance level of .05 or .01. Also, if we initially feel strongly that H0 was correct, then we would require very stringent data evidence to the contrary for us to reject H0. (That is, we would set a very low significance level in this situation.) The TEST can be described as follows: For any observed value of the test statistic $\sqrt{n}|\overline{X} - \mu_0|/\sigma$, call it v, the test calls for rejection of the null hypothesis if the **probability** that the **test statistic** would be **as large as** v when H0 is true is **less** than **or equal** to the significance level **α**. From this, it follows that we can determine whether or not to accept the null hypothesis by computing, first, the value of the test statistic and, second, the probability that a unit normal would (in absolute value) **exceed** that quantity. This probability — called the **p-value** of the test — gives the **CRITICAL** SIGNIFICANCE LEVEL(p-value) in the sense that H0 will be **accepted** if the significance level **α is less than the p**-value and rejected if it is greater than **or** equal. In practice, the **significance** level is often **not** set in advance but rather the data are **looked at** to determine the **resultant** p-value. Sometimes, this critical significance level is clearly much larger than any we would want to use, and so the null hypothesis can be readily accepted. At other times the p-value is so small that it is clear that the hypothesis should be rejected. The probability of a type II error—that is, the probability of accepting the null hypothesis when the **true** mean μ is **unequal** to μ0. This probability will depend on the value of μ, and so let us define β(μ) by $\beta(\mu) = P_\mu\{\text{acceptance of } H_0\} = P_\mu\left\{-z_{\alpha/2} \le \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right\}$ The function **β**(μ) is called the operating characteristic (or **OC**) curve and represents the probability that H0 will be accepted when the true mean is μ. [$\overline{X}$ is normal with mean μ and variance σ^2/n, μ is which $\overline{X}$ is estimating]. The function 1 – β(μ) is called the **power-function** of the test. Thus, for a **given** value μ, the **power** of the test is **equal** to the **probability** of rejection when **μ** is the **true** value. The operating characteristic function is useful in **determining how large** the random sample need be to meet certain specifications concerning type II errors. For instance, suppose that we desire to determine the **sample size n** necessary to ensure that the probability of accepting H0 : μ = μ0 when the **true mean** is actually **μ1** is **approximately β**. That is, we want n to be such that $\beta(\mu_1) \approx \beta$. To start, **suppose** that μ1 > μ0, In fact, the same approximation would result when μ1 < μ0, so, in all cases a reasonable approximation to the sample size necessary to ensure that the type **II** error at the value μ = μ1 is approximately equal to β. ONE-SIDED TESTS: In testing the null hypothesis that μ = μ0, we have chosen a test that calls for rejection when $\overline{X}$ is far from μ0. That is, a **very small** value of $\overline{X}$ **or** a **very large** value appears to make it **unlikely** that μ (**which** $\overline{X}$ is estimating) **could** equal μ0. However, what happens when the only alternative to μ being equal to μ0 is for μ to be greater than μ0? We would not want to reject H0 when $\overline{X}$ is **small** (since a small $\overline{X}$ is more likely when H0 is true than when **H1** is true). This is called a one-sided critical region (since it calls for **rejection only when** $\overline{X}$ is **large**). Hence, the TEST is to reject H0 if

$$\text{accept} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) \le z_\alpha$$

$$\text{reject} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) > z_\alpha$$

$\overline{X} - \mu_0 > z_\alpha\sigma/\sqrt{n}$ and accept otherwise; or, equivalently, to Correspondingly, the hypothesis testing problem $H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$ is called a one-sided testing problem (in contrast to the two-sided problem that results **when the alternative hypothesis is H1** : μ= μ0).
To compute the p-value in the one-sided test, we first use the data to determine the value of the statistic $\sqrt{n}(\overline{X} - \mu_0)/\sigma$. The p-value is then equal to the probability that a **standard normal** would be **at least** as large as this value **, i.e.,** p-value is the probability that a **standard normal** would **exceed** the test statistic, i.e. **p-value** = 1 – (test statistic). Since the test would call for **rejection** at all significance levels greater than or equal to p-value, it would, for instance, **reject** the null hypothesis at the α = α-value level of significance. The operating characteristic function of the one-sided test $\beta(\mu) = P_\mu\{\text{accepting } H_0\}$, $\beta(\mu) = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right)$. To verify that it **remains** a level α test, we need show that the probability of rejection is **never** greater than α when H0 is true. β(μ) ≥ β(μ0) = 1 – α for all μ ≤ μ0, which shows that the TEST remains a level α test for H0 : μ ≤ μ0 against the alternative hypothesis H1 : μ ≤ μ0. We can also test the one-sided hypothesis

$$\text{accepting} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) \ge -z_\alpha$$

$$\text{rejecting} \quad H_0 \quad \text{otherwise}$$

H0 : μ = μ0 (or μ ≥ μ0) versus H1 : μ < μ0 at significance level α by This test can alternatively be performed by first computing the value of the test statistic $\sqrt{n}(\overline{X} - \mu_0)/\sigma$. The **p**-value would then equal the probability that a **standard normal** would be **less** than this value, and the hypothesis would be **rejected** at any significance level greater than **or** equal to this p-value. REMARKS: (a) There is a **direct analogy** between confidence interval estimation and hypothesis testing. For instance, for a normal population having mean μ and known variance σ^2, we have shown in Section 7.3 that a 100(1 – α) percent confidence interval for μ is given by $\mu \in \left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ More formally, the preceding confidence **interval** statement is equivalent to $P\left\{\mu \in \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right\} = 1 - \alpha$ Hence, if μ = μ0, then the probability that **μ0** will fall in the interval $\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is **1 – α**, implying that a significance level α test of H0 : μ = μ0 versus H1 : $\mu \neq \mu_0$ is to **reject** H0 when $\mu_0 \notin \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ **Similarly**, since a 100(1 – α) percent **one**-sided confidence **interval** for μ is given by $\mu \in \left(\overline{X} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty\right)$ it follows that an α-level significance test of H0 : μ ≤ μ0 versus H1 : μ > μ0 is to reject H0 when, $\mu_0 \notin (\overline{X} - z_\alpha\sigma/\sqrt{n}, \infty)$ — that is, when $\mu_0 < \overline{X} - z_\alpha\sigma/\sqrt{n}$. (b) A Remark on Robustness A test that performs well even when the underlying assumptions on which it is based are violated is said to be robust.

TABLE 8.1 $X_1,\ldots,X_n$ *Is a Sample from a* $\mathcal{N}(\mu,\sigma^2)$ *Population* σ^2 *Is Known* $\overline{X} = \sum_{i=1}^{n} X_i/n$

H_0	H_1	Test Statistic TS	Significance Level α Test	p-Value if $TS = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/\sigma$	Reject if $\lvert TS\rvert > z_{\alpha/2}$	$2P\{Z \ge \lvert t\rvert\}$
$\mu \le \mu_0$	$\mu > \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/\sigma$	Reject if $TS > z_\alpha$	$P\{Z \ge t\}$
$\mu \ge \mu_0$	$\mu < \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/\sigma$	Reject if $TS < -z_\alpha$	$P\{Z \le t\}$

8.3.2 Case of Unknown Variance: The t-Test: Up to now we have supposed that the only unknown parameter of the normal population distribution is its mean. However, the more common situation is one where the mean μ and variance σ^2 are both unknown. Let The t-Test: again consider a test of the hypothesis that the mean is equal to some specified value μ0. As before, it seems reasonable to reject H0 when the sample mean $\overline{X}$ is far from μ0. However, how far away it need be to justify rejection will depend on the variance σ^2. Recall that when the value of σ^2 was known, the test called for rejectingH0 when $|\overline{X} - \mu 0|$ exceeded $z_{\alpha/2}\sigma/\sqrt{n}$ or, equivalently, when $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}$, to reject H0 when, $\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right|$ is large. To determine how large a value of the statistic $\left|\frac{\sqrt{n}(\overline{X}-\mu_0)}{S}\right|$ to require for rejection, in order that the resulting test have significance level α, we must determine the probability distribution of this statistic when H0 is true. The appropriate significance level α test of H0 : μ = μ0 versus H1 : μ= μ0 is, when σ^2 is unknown $\text{accept} \quad H_0 \quad \text{if} \quad \left|\frac{\sqrt{n}(\overline{X}-\mu_0)}{S}\right| \le t_{\alpha/2,n-1}$ $\text{reject} \quad H_0 \quad \text{if} \quad \left|\frac{\sqrt{n}(\overline{X}-\mu_0)}{S}\right| > t_{\alpha/2,n-1}$. If we let t denote the observed value of the test statistic T = $\sqrt{n}(\overline{X}-\mu_0)/S$ then the p-value of the test is the probability that |T | would exceed |t | when H0 is true. That is, the p-value is the probability that the absolute value of a t-random variable with n – 1 degrees of freedom would exceed |t |. The test then calls for rejection at all significance levels higher than the p-value and acceptance at all lower significance levels. Program 8.3.2 computes the value of the test statistic and the corresponding p-value. We can use a one-sided t –test to test the hypothesis $H_0 : \mu = \mu_0 \quad (\text{or } H_0 : \mu \le \mu_0)$ against the one-sided alternative $H_1 : \mu > \mu_0$ The significance level α test is to $\text{accept} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}(\overline{X}-\mu_0)}{S} \le t_{\alpha,n-1}$ $\text{reject} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}(\overline{X}-\mu_0)}{S} > t_{\alpha,n-1}$ If $\sqrt{n}(\overline{X}-\mu_0)/S = v$, then the p-value of the test is the probability that a t-random variable with n – 1 degrees of freedom would be at least as large as v. The significance level α test of $H_0 : \mu = \mu_0 \quad (\text{or } H_0 : \mu \ge \mu_0)$ versus the alternative $H_1 : \mu < \mu_0$ is to $\text{accept} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}(\overline{X}-\mu_0)}{S} \ge -t_{\alpha,n-1}$ $\text{reject} \quad H_0 \quad \text{if} \quad \frac{\sqrt{n}(\overline{X}-\mu_0)}{S} < -t_{\alpha,n-1}$. The p-value of this test is the probability that a t-random variable with n – 1 degrees of freedom would be less than or equal to the observed value of $\sqrt{n}(\overline{X}-\mu_0)/S$

TABLE 8.2 $X_1, \ldots, X_n$ *Is a Sample from a* $\mathcal{N}(\mu, \sigma^2)$ *Population* σ^2 *Is Unknown* $\overline{X} = \sum_{i=1}^{n} X_i/n$ $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$

H_0	H_1	Test Statistic TS	Significance Level α Test	p-Value if $TS = t$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\sqrt{n}(\overline{X}-\mu_0)/S$	Reject if $\lvert TS\rvert > t_{\alpha/2,n-1}$	$2P\{T_{n-1} \ge \lvert t\rvert\}$
$\mu \le \mu_0$	$\mu > \mu_0$	$\sqrt{n}(\overline{X}-\mu_0)/S$	Reject if $TS > t_{\alpha,n-1}$	$P\{T_{n-1} \ge t\}$
$\mu \ge \mu_0$	$\mu < \mu_0$	$\sqrt{n}(\overline{X}-\mu_0)/S$	Reject if $TS < -t_{\alpha,n-1}$	$P\{T_{n-1} \le t\}$

8.4 TESTING THE EQUALITY OF MEANS OF TWO NORMAL POPULATIONS: A common situation faced by a practicing engineer is one in which she must decide whether two different approaches lead to the same solution. Often such a situation can be modelled as a test of the hypothesis that two normal populations have the same mean value. 8.4.1 Case of Known Variances: 8.4.2 Case of Unknown Variances 8.4.3 Case of Unknown and Unequal VariancesBehrens-Fisher problem. There is no completely satisfactory solution known.

TABLE 8.4 $X_1, \ldots, X_n$ *Is a Sample from a* $\mathcal{N}(\mu_1, \sigma_1^2)$ *Population:* $Y_1, \ldots, Y_m$ *Is a Sample from a* $\mathcal{N}(\mu_2, \sigma_2^2)$ *Population*

The Two Population Samples Are Independent

To Test $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 \ne \mu_2$

Assumption	Test Statistic TS	Significance Level α Test	p-Value if $TS = t$
σ_1, σ_2 known	$\frac{\overline{X}-\overline{Y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}}$	Reject if $\lvert TS\rvert > z_{\alpha/2}$	$2P\{Z \ge \lvert t\rvert\}$
$\sigma_1 = \sigma_2$	$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}\sqrt{1/n+1/m}}$	Reject if $\lvert TS\rvert > t_{\alpha/2,n+m-2}$	$2P\{T_{n+m-2} \ge \lvert t\rvert\}$
n, m large	$\frac{\overline{X}-\overline{Y}}{\sqrt{S_1^2/n+S_2^2/m}}$	Reject if $\lvert TS\rvert > z_{\alpha/2}$	$2P\{Z \ge \lvert t\rvert\}$

8.4.4 The Paired t-Test: Suppose we are interested in determining whether the installation of a certain antipollution device will affect a car's mileage. To test this, a collection of n cars that do not have this device are gathered. Each car's mileage per gallon is then determined both before and after the device is installed. How can we test the hypothesis that the antipollution control has no effect on gas consumption?

The data can be described by the n pairs (Xi , Yi), I = 1, . . . , n, where Xi is the gas consumption of the ith car before installation of the pollution control device, and Yi of the same car after installation. It is important to note that, since each of the n cars will be inherently different, we cannot treat X1, . . . , Xn and Y1, . . . , Yn as being independent samples. For example, if we know that X1 is large (say, 40 miles per gallon), we would certainly expect that Y1 would also probably be large. Thus, we cannot employ the earlier methods presented in this section. One way in which we can test the hypothesis that the antipollution device does not affect gas mileage is to let the data consist of each car's difference in gas mileage. That is, letWi = Xi –Yi , I = 1, . . . , n. Now, if there is no effect from the device, it should follow that the Wi would have mean 0. Hence, we can test the hypothesis of no effect by testing $H_0 : \mu_w = 0 \quad \text{versus} \quad H_1 : \mu_w \ne 0$ whereW1, . . . ,Wn are assumed to be a sample from a normal population having unknown mean μw and unknown variance σ^2w. But the t –test described in Section 8.3.2 shows that this can be tested by $\text{accepting} \quad H_0 \quad \text{if} \quad -t_{\alpha/2,n-1} < \sqrt{n}\frac{\overline{W}}{S_w} < t_{\alpha/2,n-1}$ $\text{rejecting} \quad H_0 \quad \text{otherwise}$ 8.5 HYPOTHESIS TESTS CONCERNING THE VARIANCE(σ^2)OF A NORMAL POPULATION: Let X1, . . . , Xn denote a sample from a normal population having unknown mean μ and unknown variance σ^2, and suppose we desire to test the hypothesis $H_0 : \sigma^2 = \sigma_0^2$ versus the alternative $H_1 : \sigma^2 \ne \sigma_0^2$ for some specified value σ_0^2. The preceding test can be implemented by first computing the value of the test statistic (n−1)S2/σ2 0 —call it c. Then compute the probability that a chi-square random variable with n – 1 degrees of freedom would be (a) less than and (b) greater than c. If either of these probabilities is less than α/2, then the hypothesis is rejected. In other words, the p-value of the test data is $p\text{-value} = 2\min(P\{\chi_{n-1}^2 < c\}, 1 - P\{\chi_{n-1}^2 < c\})$ 8.5.1 Testing for the Equality of Variances of Two Normal Populations: Let X1, . . . , Xn and Y1, . . . , Ym denote independent samples from two normal populations having respective (unknown) parameters μx , σ_0^2x and μy , σ_0^2y and consider a test of . $H_0 : \sigma_x^2 = \sigma_y^2 \quad \text{versus} \quad H_1 : \sigma_x^2 \ne \sigma_y^2$ Thus, a significance level α test of H0 against H1 is to $\text{accept} \quad H_0 \quad \text{if} \quad F_{1-\alpha/2,n-1,m-1} < S_x^2/S_y^2 < F_{\alpha/2,n-1,m-1}$ $\text{reject} \quad H_0 \quad \text{otherwise}$. $p\text{-value} = 2\min(P\{F_{n-1,m-1} < v\}, 1 - P\{F_{n-1,m-1} < v\})$. The test now calls for rejection whenever the significance level α is at least as large as the p-value. 8.6 HYPOTHESIS TESTS IN BERNOULLI POPULATIONS: An assumption often made is that each item produced will, independently, be defective with probability p, binomial distribution. Consider a test of $H_0 : p \le p_0 \quad \text{versus} \quad H_1 : p > p_0$ where p0 is some specified value. If we let X denote the number of defects in the sample of size n, then it is clear that we wish to reject H0 when X is large. To see how large it need be to justify rejection at the α level of significance. $P\{X \ge k\} = \sum_{i=k}^{n} P\{X = i\} = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$ Now it is certainly intuitive (and can be proven) that P{X ≥ k} is an increasing function of p—that is, the probability that the sample will contain at least k errors increases in the defect probability p. k^* is the smallest value of k for which $k^* = \min\left\{k : \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \le \alpha\right\}$ This test can best be performed by first determining the value of the test statistic —say, X = x — and then computing the p-value given by $p\text{-value} = P\{B(n, p_0) \ge x\} = \sum_{i=x}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$ .When the sample size n is large, we can derive an approximate significance level α test of H0 : p ≤ p0 versus H1 : p > p0 by using the normal approximation to the binomial. It works as follows: Because when n is large X will have approximately a normal distribution with mean and variance $E[X] = np, \quad \text{Var}(X) = np(1-p)$ it follows that $\frac{X - np}{\sqrt{np(1-p)}}$ will have approximately a standard normal distribution. Therefore, an approximate significance level α test would be to reject H0 if $\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \ge z_\alpha$ .Equivalently, one can use the normal approximation to approximate the p-value. Suppose now that we want to test the null hypothesis that p is equal to some specified value; that is, we want to test $H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \ne p_0$. If X , a binomial random variable with parameters n and p, is observed to equal x, then a significance level α test would reject H0 if the value x was either significantly larger or significantly smaller than what

would be expected when p is equal to p0. More precisely, the test would reject H0 if either . $P\{\text{Bin}(n,p_0) \geq x\} \leq \alpha/2$ or $P\{\text{Bin}(n,p_0) \leq x\} \leq \alpha/2$ In other words, the p-value when X = x is $p\text{-value} = 2\min(P\{\text{Bin}(n,p_0) \geq x\}, P\{\text{Bin}(n,p_0) \leq x\})$ 8.6.1 Testing the Equality of Parameters in Two Bernoulli Populations: Suppose there are two distinct methods for producing a certain type of transistor; and suppose that transistors produced by the first method will, independently, be defective with probability p1, with the corresponding probability being p2 for those produced by the second method. To test the hypothesis that p1 = p2, a sample of n1 transistors is produced using method 1 and n2 using method 2. Suppose that X1 +X2 = k and so there have been a total of k defectives. Now, if H0 is true, then each of the n1 +n2 transistors produced will have the same probability of being defective, and so the determination of the k defectives will have the same distribution as a random selection of a sample of size k from a population of n1 + n2 items of which n1 are white and n2 are black. In other words, given a total of k defectives, the conditional distribution of the number of defective transistors obtained from method 1 will, when H0 is true, have the following hypergeometric distribution.

$$P_{H_0}\{X_1 = i | X_1 + X_2 = k\} = \frac{\binom{n_1}{i}\binom{n_2}{k-i}}{\binom{n_1+n_2}{k}}, \quad i = 0, 1, \ldots, k$$

Now, in testing, $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$ it seems reasonable to reject the null hypothesis when the proportion of defective transistors produced by method 1 is much different than the proportion of defectives obtained under method 2. Therefore, if there is a total of k defectives, then we would expect, when H0 is true, that X1/n1 (the proportion of defective transistors produced by method 1) would be close to (k – X1)/n2 (the proportion of defective transistors produced by method 2). Because X1/n1 and (k – X1)/n2 will be farthest apart when X1 is either very small or very large, it thus seems that a reasonable significance level α test of Equation 8.6.1 is as follows. If X1 + X2 = k, then one should

$$\begin{matrix}\text{reject } H_0 \text{ if either } P\{X \leq x_1\} \leq \alpha/2 \text{ or } P\{X \geq x_1\} \leq \alpha/2 \\ \text{accept } H_0 \text{ otherwise}\end{matrix}$$

where X is a hypergeometric random variable with probability mass function

$$P\{X = i\} = \frac{\binom{n_1}{i}\binom{n_2}{k-i}}{\binom{n_1+n_2}{k}} \quad i = 0, 1, \ldots, k$$

In other words, this test will call for rejection if the significance level is at least as large as the p-value given by $p\text{-value} = 2\min(P\{X \leq x_1\}, P\{X \geq x_1\})$. This is called the Fisher-Irwin test. To compute the hypergeometric distribution function:

$$\begin{aligned}\frac{P\{X = i+1\}}{P\{X = i\}} &= \frac{\binom{n_1}{i+1}\binom{n_2}{k-i-1}}{\binom{n_1}{i}\binom{n_2}{k-i}} \\ &= \frac{(n_1 - i)(k - i)}{(i+1)(n_2 - k + i + 1)}\end{aligned}$$

Program 8.6.1 uses the preceding identity to compute the p-value of the data for the Fisher-Irwin test of the equality of two Bernoulli probabilities. The program will work best if the Bernoulli outcome that is called unsuccessful (or defective) is the one whose probability is less than .5. For instance, if over half the items produced are defective, then rather than testing that the defect probability is the same in both samples, one should test that the probability of producing an acceptable item is the same in both samples. Observational Study: The ideal way to test the hypothesis that the results of two different treatments are identical is to randomly divide a group of people into a set that will receive the first treatment and one that will receive the second. However, such randomization is not always possible. For instance, if we want to study whether drinking alcohol increases the risk of prostate cancer, we cannot instruct a randomly chosen sample to drink alcohol. An alternative way to study the hypothesis is to use an observational study that begins by randomly choosing a set of drinkers and one of nondrinkers. These sets are followed for a period of time and the resulting data is then used to test the hypothesis that members of the two groups have the same risk for prostate cancer. Our next sample illustrates another way of performing an observational study. In 1970, the researchers Herbst, Ulfelder, and Poskanzer (H-U-P) suspected that vaginal cancer in young women, a rather rare disease, might be caused by one's mother having taken the drug diethylstilbestrol (usually referred to as DES) while pregnant. To study this possibility, the researchers could have performed an observational study by searching for a (treatment) group of women whose mothers took DES when pregnant and a (control) group of women whose mothers did not. They could then observe these groups for a period of time and use the resulting data to test the hypothesis that the probabilities of contracting vaginal cancer are the same for both groups. However, because vaginal cancer is so rare (in both groups) such a study would require a large number of individuals in both groups and would probably have to continue for many years to obtain significant results. Consequently, H-U-P decided on a different type of observational study. They uncovered 8 women between the ages of 15 and 22 who had vaginal cancer. Each of these women (called cases) was then matched with 4 others, called referents or controls. Each of the referents of a case was free of the cancer and was born within 5 days in the same hospital and in the same type of room (either private or public) as the case. Arguing that if DES had no effect on vaginal cancer then the probability, call it pc , that the mother of a case took DES would be the same as the probability, call it pr , that the mother of a referent took DES, the researchers H-U-P decided to test H0 : pc = pr against H1 : pc = pr Discovering that 7 of the 8 cases had mothers who took DES while pregnant, while none of the 32 referents had mothers who took the drug, the researchers (see Herbst, A., Ulfelder, H., and Poskanzer, D., "Adenocarcinoma of the Vagina: Association of Maternal Stilbestrol Therapy with Tumor Appearance in Young Women," New England Journal of Medicine, 284, 878–881, 1971) concluded that there was a strong association between DES and vaginal cancer. (The p-value for these data is approximately 0.) When n1 and n2 are large, an approximate level α test of H0 : p1 = p2, based on the normal approximation to the binomial, is outlined in Problem 63. 8.7 TESTS CONCERNING THE MEAN OF A POISSON DISTRIBUTION: Let X1 and X2 be independent Poisson random variables with respective means λ1 and λ2, and consider a test of $H_0 : \lambda_2 = c\lambda_1$ versus $H_1 : \lambda_2 \neq c\lambda_1$ for a given constant c. Our test of this is a conditional test (similar in spirit to the Fisher-Irwin test of Section 8.6.1), which is based on the fact that the conditional distribution of X1 given the sum of X1 and X2 is binomial. More specifically, we have the following proposition. $P\{X_1 = k | X_1 + X_2 = n\} = \binom{n}{k}[\lambda_1/(\lambda_1+\lambda_2)]^k[\lambda_2/(\lambda_1+\lambda_2)]^{n-k}$ It follows from Proposition that, if H0 is true, then the conditional distribution of X1 given that X1+X2 = n is the binomial distribution with parameters n and p = 1/(1+c). From this we can conclude that if X1+X2 = n, then H0 should be rejected if the observed value of X1, call it x1, is such that either $P\{\text{Bin}(n, 1/(1+c)) \geq x_1\} \leq \alpha/2$ or $P\{\text{Bin}(n, 1/(1+c)) \leq x_1\} \leq \alpha/2$

Chapter_7: PARAMETER ESTIMATION: 7.1 INTRODUCTION: Let X1, . . . , Xn be a random sample from a distribution Fθ that is specified up to a vector of unknown parameters θ. Whereas in probability theory it is usual to suppose that all of the parameters of a distribution are known, the opposite is true in statistics, where a central problem is to use the observed data to make inferences about the unknown parameters. The maximum likelihood method for determining estimators of unknown parameters. The estimates so obtained are called point estimates, because they specify a single quantity as an estimate of θ. In Section 7.3, we consider the problem of obtaining interval estimates, Additionally, we consider the question of how much confidence we can attach to such an interval estimate. The general problem of obtaining point estimates of unknown parameters and show how to evaluate an estimator by considering its mean square error. The bias of an estimator is discussed, and its relationship to the mean square error is explored. We consider the problem of determining an estimate of an unknown parameter when there is some prior information available. This is the Bayesian approach, which supposes that prior to observing the data, information about θ is always available to the decision maker, and that this information can be expressed in terms of a probability distribution on θ. In such a situation, we show how to compute the Bayes estimator, which is the estimator whose expected squared distance from θ is minimal. 7.2 MAXIMUM LIKELIHOOD ESTIMATORS: Any statistic used to estimate the value of an unknown parameter θ is called an estimator of θ. The usual estimator of the mean of a normal population, based on a sample X1, . . . , Xn from that population, is the sample $\overline{X} = \sum_i X_i/n$ Suppose that the random variables X1, . . . , Xn, whose joint distribution is assumed(observed) given except for an unknown parameter θ, are to be observed. The problem of interest is to use the observed values to estimate θ. For example, the Xi 's might be Independent, exponential random variables each having the same unknown mean θ. In this case, the joint density function of the random variables would be given by

$$\begin{aligned}f(x_1, x_2, \ldots, x_n) &= f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) \\ &= \frac{1}{\theta}e^{-x_1/\theta}\frac{1}{\theta}e^{-x_2/\theta}\cdots\frac{1}{\theta}e^{-x_n/\theta}, \quad 0 < x_i < \infty, i = 1, \ldots, n \\ &= \frac{1}{\theta^n}\exp\left\{-\sum_1^n x_i/\theta\right\}, \quad 0 < x_i < \infty, i = 1, \ldots, n\end{aligned}$$

and the objective would be to estimate θ from the observed data X1, X2, . . . , Xn. Maximum Likelihood Estimator: Let f (x1, . . . , xn|θ) denote the joint probability mass function of the random variables X1, X2, . . . , Xn when they are discrete, and let it be their joint probability density function when they are jointly continuous random variables. Because θ is assumed unknown, we also write f as a function of θ. Now since f (x1, . . . , xn|θ) represents the likelihood that the values x1, x2, . . . , xn will be observed when θ is the true value of the parameter, it would seem that a reasonable estimate of θ would be that value yielding the largest likelihood of the observed values. In other words, the maximum likelihood estimate $\hat{\theta}$ is defined to be that value of θ maximizing f (x1, . . . , xn|θ) where x1, . . . , xn are the observed values. The function f (x1, . . . , xn|θ) is often referred to as the likelihood function of θ. In determining the maximizing value of θ, it is often useful to use the fact that f (x1, . . . , xn|θ) and log[f (x1, . . . , xn|θ)] have their maximum at the same value of θ. Hence, we may also obtain ˆ θ by maximizing log[f (x1, . . . , xn|θ)]. EXAMPLE 7.2a Maximum Likelihood Estimator of a Bernoulli Parameter: Suppose that n independent trials, each of which is a success with probability p, are performed. What is the maximum likelihood estimator of p? The data consist of the values of X1, . . . , Xn where

$$X_i = \begin{cases}1 & \text{if trial } i \text{ is a success} \\ 0 & \text{otherwise}\end{cases} \qquad P\{X_i = 1\} = p = 1 - P\{X_i = 0\}$$

which can be succinctly expressed as $P\{X_i = x\} = p^x(1-p)^{1-x}, \quad x = 0, 1$ Hence, by the assumed independence of the trials, the likelihood (that is, the joint probability mass function) of the data is

$$f(x_1,\ldots,x_n|p) = P\{X_1 = x_1,\ldots,X_n = x_n|p\}$$
$$= p^{x_1}(1-p)^{1-x_1}\cdots p^{x_n}(1-p)^{1-x_n}$$
$$= p^{\sum_1^n x_i}(1-p)^{n-\sum_1^n x_i}, \quad x_i = 0,1, \quad i = 1,\ldots,n$$

given by . To determine the value of p that maximizes the likelihood, first take logs to obtain

$$\log f(x_1,\ldots,x_n|p) = \sum_1^n x_i \log p + \left(n - \sum_1^n x_i\right)\log(1-p)$$

Differentiation yields $\frac{d}{dp}\log f(x_1,\ldots,x_n|p) = \frac{\sum_1^n x_i}{p} - \frac{\left(n-\sum_1^n x_i\right)}{1-p}$ Upon equating to zero and solving, we obtain that the maximum likelihood estimate $\hat{p}$ satisfies $\frac{\sum_1^n x_i}{\hat{p}} = \frac{n-\sum_1^n x_i}{1-\hat{p}}$ or $\hat{p} = \frac{\sum_{i=1}^n x_i}{n}$ Hence, the maximum likelihood estimator of the unknown mean of a Bernoulli distribution is given by $d(X_1,\ldots,X_n) = \frac{\sum_{i=1}^n X_i}{n}$. Since $\sum_{i=1}^n X_i$ is the number of successful trials, we see that the maximum likelihood estimator of p is equal to the proportion of the observed trials that result in successes. Maximum Likelihood Estimator of a Poisson Parameter: Suppose X1, . . . , Xn

$$f(x_1,\ldots,x_n|\lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\cdots\frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$
$$= \frac{e^{-n\lambda}\lambda^{\sum_1^n x_i}}{x_1!\ldots x_n!}$$

are independent Poisson random variables each having mean λ. The likelihood function is given by Thus,

$$\log f(x_1,\ldots,x_n|\lambda) = -n\lambda + \sum_1^n x_i \log\lambda - \log c \quad \text{where } c = \prod_{i=1}^n x_i! \text{ does not depend on } \lambda, \text{ and} \quad \frac{d}{d\lambda}\log f(x_1,\ldots,x_n|\lambda) = -n + \frac{\sum_1^n x_i}{\lambda}$$

By equating to zero, we obtain that the maximum likelihood estimate $\hat{\lambda}$ equals $\hat{\lambda} = \frac{\sum_1^n x_i}{n}$ and so the maximum likelihood estimator is given by $d(X_1,\ldots,X_n) = \frac{\sum_{i=1}^n X_i}{n}$ Maximum Likelihood Estimator in a Normal Population: Suppose X1, . . . , Xn are independent, normal random variables each with unknown mean μ and unknown standard deviation σ. The joint density is given by

$$f(x_1,\ldots,x_n|\mu,\sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(x_i-\mu)^2}{2\sigma^2}\right] = \left(\frac{1}{2\pi}\right)^{n/2}\frac{1}{\sigma^n}\exp\left[\frac{-\sum_1^n(x_i-\mu)^2}{2\sigma^2}\right]$$

The logarithm of the likelihood is thus given by $\log f(x_1,\ldots,x_n|\mu,\sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{\sum_1^n(x_i-\mu)^2}{2\sigma^2}$ In order to find the value of μ and σ maximizing the foregoing, we compute $\frac{\partial}{\partial\mu}\log f(x_1,\ldots,x_n|\mu,\sigma) = \frac{\sum_{i=1}^n(x_i-\mu)}{\sigma^2}$ $\frac{\partial}{\partial\sigma}\log f(x_1,\ldots,x_n|\mu,\sigma) = -\frac{n}{\sigma} + \frac{\sum_1^n(x_i-\mu)^2}{\sigma^3}$ Equating these equations to zero yields that $\hat{\mu} = \sum_{i=1}^n x_i/n$ and $\hat{\sigma} = \left[\sum_{i=1}^n (x_i-\hat{\mu})^2/n\right]^{1/2}$ Hence, the maximum likelihood estimators of μ and σ are given, respectively, by $\overline{X}$ and $\left[\sum_{i=1}^n (X_i-\overline{X})^2/n\right]^{1/2}$ It should be noted that the maximum likelihood estimator of the standard deviation σ differs from the sample standard deviation $S = \left[\sum_{i=1}^n (X_i-\overline{X})^2/(n-1)\right]^{1/2}$ in that the denominator is $\sqrt{n}$ rather than $\sqrt{n-1}$. However, for n of reasonable size, these two estimators of σ will be approximately equal. Estimating the Mean of a Uniform Distribution: Suppose X1, . . . , Xn constitute a sample from a uniform distribution on (0, θ), where θ is unknown. Their joint density is thus

$$f(x_1,x_2,\ldots,x_n|\theta) = \begin{cases}\frac{1}{\theta^n} & 0 < x_i < \theta, \quad i = 1,\ldots,n\\ 0 & \text{otherwise}\end{cases}$$

This density is maximized by choosing θ as small as possible. Since θ must be at least as large as all of the observed values xi , it follows that the smallest possible choice of θ is equal to max(x1, x2, . . . , xn). Hence, the maximum likelihood estimator of θ is $\hat{\theta} = \max(X_1,X_2,\ldots,X_n)$. It easily follows from the foregoing that the maximum likelihood estimator of θ/2, the mean of the distribution, is max(X1, X2, . . . , Xn)/2. 7.2.1 Estimating Life Distributions: Let X denote the age at death of a randomly chosen child born today. That is, X = I if the newborn dies in its ith year, I ≥ 1. To estimate the probability mass function of X, let λi denote the probability that a newborn who has survived his or her first I – 1 years dies in year i. That is, $\lambda_i = P\{X = i|X > i-1\} = \frac{P\{X=i\}}{P\{X>i-1\}}$ Also, let $s_i = 1-\lambda_i = \frac{P\{X>i\}}{P\{X>i-1\}}$ be the probability that a newborn who survives her first I – 1 years also survives year i.The quantity λi is called the failure rate, and si is called the survival rate, of an individual who is entering his or her ith year. Now,

$$s_1 s_2\cdots s_i = P\{X>1\}\frac{P\{X>2\}P\{X>3\}}{P\{X>1\}P\{X>2\}}\cdots\frac{P\{X>i\}}{P\{X>i-1\}} = P\{X>i\}$$

Therefore, $P\{X=n\} = P\{X>n-1\}\lambda_n = s_1\cdots s_{n-1}(1-s_n)$. Consequently, we can estimate the probability mass function of X by estimating the quantities si , I = 1, . . . , n. The value si can be estimated by looking at all individuals in the population who reached age I one year ago, and then letting the estimate $\hat{s}_i$ be the fraction of them who are alive today. We would then use $\hat{s}_1\hat{s}_2\cdots\hat{s}_{n-1}(1-\hat{s}_n)$ as the estimate of P{X = n}. (Note that although we are using the most recent possible data to estimate the quantities si , our estimate of the probability mass function of the lifetime of a newborn assumes that the survival rate of the newborn when it reaches age I will be the same as last year's survival rate of someone of age i.) The use of the survival rate to estimate a life distribution is also of importance in health studies with partial information. For instance, consider a study in which a new drug is given to a random sample of 12 lung cancer patients. Suppose that after some time we have the following data on the number of months of survival after starting the new drug: $4, 7^*, 9, 11^*, 12, 3, 14^*, 1, 8, 7, 5, 3^*$ where x means that the patient died in month x after starting the drug treatment, and x^* means that the patient has taken the drug for x months and is still alive. Let X equal the number of months of survival after beginning the drug treatment, and let $s_i = P\{X>i|X>i-1\} = \frac{P\{X>i\}}{P\{X>i-1\}}$ To estimate si , the probability that a patient who has survived the first I – 1 months will also survive month I, we should take the fraction of those patients who began their ith month of drug taking and survived the month. For instance, because 11 of the 12 patients survived month 1, $\hat{s}_1$ = 11/12. Because all 11 patients who began month 2 survived, $\hat{s}_2$ = 11/11. Because 10 of the 11 patients who began month 3 survived, $\hat{s}_3$ = 10/11. Because 8 of the 9 patients who began their fourth month of taking the drug (all but the ones labelled 1, 3, and 3∗) survived month 4, $\hat{s}_4$ = 8/9. Similar reasoning holds for the others, giving the following survival rate estimates: $\hat{s}_5 = 7/8$, $\hat{s}_6 = 7/7$, $\hat{s}_7 = 6/7$, $\hat{s}_8 = 4/5$, $\hat{s}_9 = 3/4$, $\hat{s}_{10} = 3/3$, $\hat{s}_{11} = 3/3$, $\hat{s}_{12} = 1/2$, $\hat{s}_{13} = 1/1$, $\hat{s}_{14} = 1/2$ We can now use $\prod_{i=1}^j \hat{s}_i$ to estimate the probability that a drug taker survives at least j time periods, j = 1, . . . , 14. For instance, our estimate of P{X > 6} is 35/54. 7.3 INTERVAL ESTIMATES: Suppose that X1, . . . , Xn is a sample from a normal population having unknown mean μ and known variance σ2. It has been shown that $\overline{X} = \sum_{i=1}^n X_i/n$ X is the maximum likelihood estimator for μ. However, we don't expect that the sample mean X will exactly equal μ, but rather that it will "be close." Hence, rather than a point estimate, it is sometimes more valuable to be able to specify an interval for which we have a certain degree of confidence that μ lies within. To obtain such an interval estimator, we make use of the probability distribution of the point estimator. Let us see how it works for the preceding situation. In the foregoing, since the point estimator X is normal with mean μ and variance σ^2/n, it follows that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{(\overline{X}-\mu)}{\sigma}$ has a standard normal distribution. Therefore,

$$P\left\{-1.96 < \sqrt{n}\frac{(\overline{X}-\mu)}{\sigma} < 1.96\right\} = .95$$

or, equivalently,

$$P\left\{-1.96\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

Multiplying through by −1 yields the equivalent statement

$$P\left\{-1.96\frac{\sigma}{\sqrt{n}} < \mu-\overline{X} < 1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

or, equivalently,

$$P\left\{\overline{X}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

That is, 95 percent of the time μ will lie within $1.96\sigma/\sqrt{n}$ units of the sample average. If we now observe the sample and it turns out that X = x, then we say that "with 95 percent confidence" $\overline{x}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x}+1.96\frac{\sigma}{\sqrt{n}}$ That is, "with 95 percent confidence" we assert that the true mean lies within $1.96\sigma/\sqrt{n}$ of the observed sample mean. The interval $\left(\overline{x}-1.96\frac{\sigma}{\sqrt{n}}, \overline{x}+1.96\frac{\sigma}{\sqrt{n}}\right)$ is called a 95 percent confidence interval estimate of μ. **Sometimes, however, we are interested in determining a value so that we can assert with, say, 95 percent confidence, that μ is at least as large as that value.** To determine such a value, note that if Z is a standard normal random variable then $P\{Z < 1.645\} = .95$ As a result,

$$P\left\{\sqrt{n}\frac{(\overline{X}-\mu)}{\sigma} < 1.645\right\} = .95$$

or

$$P\left\{\overline{X}-1.645\frac{\sigma}{\sqrt{n}} < \mu\right\} = .95$$

Thus, a 95 percent one-sided upper confidence interval for μ is $\left(\overline{x}-1.645\frac{\sigma}{\sqrt{n}}, \infty\right)$ where $\overline{x}$ is the observed value of the sample mean. A one-sided

lower confidence interval is obtained similarly; when the observed value of the sample mean is x, then the 95 percent one-sided lower confidence interval for μ is $\left(-\infty, \bar{x}+1.645\frac{\sigma}{\sqrt{n}}\right)$. Sometimes we are interested in a two-sided confidence interval of a certain level, say 1 – α, and the problem is to choose the sample size n so that the interval is of a certain size. For instance, suppose that we want to compute an interval of length .1 that we can assert, with 99 percent confidence, contains μ. How large need n be? To solve this, note that as z.005 = 2.58 it follows that the 99 percent confidence interval for μ from a sample of size n is $\left(\bar{x}-2.58\frac{\sigma}{\sqrt{n}}, \quad \bar{x}+2.58\frac{\sigma}{\sqrt{n}}\right)$ Hence, its length is $5.16\frac{\sigma}{\sqrt{n}}$ Thus, to make the length of the interval equal to .1, we must choose $5.16\frac{\sigma}{\sqrt{n}}=.1$ or $n=(51.6\sigma)^2$ REMARK: The interpretation of "a 100(1−α) percent confidence interval" can be confusing. It should be noted that we are not asserting that the probability that $\mu \in (\bar{x}-1.96\sigma/\sqrt{n}, \bar{x}+1.96\sigma/\sqrt{n})$ is .95, for there are no random variables involved in this assertion. What we are asserting is that the technique utilized to obtain this interval is such that 95 percent of the time that it is employed it will result in an interval in which μ lies. In other words, before the data are observed we can assert that with probability .95 the interval that will be obtained will contain μ, whereas after the data are obtained we can only assert that the resultant interval indeed contains μ "with confidence .95." 7.3.1 Confidence Interval for a Normal Mean When the Variance Is Unknown: Suppose now that X1, . . . , Xn is a sample from a normal distribution with unknown mean μ and unknown variance σ^2, and that we wish to construct a 100(1−α) percent confidence interval for μ. Since σ is unknown, we can no longer base our interval on the fact that $\sqrt{n}(\bar{X}-\mu)/\sigma$ is a standard normal random variable. However, by letting $S^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ denote the sample variance, then it follows that $\sqrt{n}\frac{(\bar{X}-\mu)}{S}$ is a t-random variable with n – 1 degrees of freedom. Hence, from the symmetry of the t –density function, we have that for any α ∈ (0, ½), $P\left\{-t_{\alpha/2,n-1}<\sqrt{n}\frac{(\bar{X}-\mu)}{S}<t_{\alpha/2,n-1}\right\}=1-\alpha$ or, equivalently, $P\left\{\bar{X}-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}<\mu<\bar{X}+t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right\}=1-\alpha$ Thus, if it is observed that X = x and S = s, then we can say that "with 100(1 – α) percent confidence" $\mu \in \left(\bar{x}-t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x}+t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$ REMARKS: (a) The confidence interval for μ when σ is known is based on the fact that $\sqrt{n}(\bar{X}-\mu)/\sigma$ has a standard normal distribution. When σ is unknown, the foregoing approach is to estimate it by S and then use the fact that $\sqrt{n}(\bar{X}-\mu)/S$ has a t –distribution with n – 1 degrees of freedom.
(b) The length of a 100(1 – α) percent confidence interval for μ is not always larger when the variance is unknown. For the length of such an interval is $2z_{\alpha}\sigma/\sqrt{n}$ when σ is known, whereas it is $2t_{\alpha,n-1}S/\sqrt{n}$ when σ is unknown; and it is certainly possible that the sample standard deviation S can turn out to be much smaller than σ. However, it can be shown that the mean length of the interval is longer when σ is unknown. That is, it can be shown that $t_{\alpha,n-1}E[S]\geq z_{\alpha}\sigma$. Indeed, E[S] is evaluated in Chapter 14 and it is shown, for instance, that $E[S]=\begin{cases}.94\sigma & \text{when } n=5\\ .97\sigma & \text{when } n=9\end{cases}$ Since $z_{.025}=1.96, \quad t_{.025,4}=2.78, \quad t_{.025,8}=2.31$ the length of a 95 percent confidence interval from a sample of size 5 is 2×1.96σ/√5 = 1.75σ when σ is known, whereas its expected length is 2×2.78×.94σ/√5 = 2.34σ when σ is unknown—an increase of 33.7 percent. If the sample is of size 9, then the two values to compare are 1.31σ and 1.49σ —a gain of 13.7 percent. A one-sided upper confidence interval can be obtained by noting that $P\left\{\sqrt{n}\frac{(\bar{X}-\mu)}{S}<t_{\alpha,n-1}\right\}=1-\alpha$ or $P\left\{\bar{X}-\mu<\frac{S}{\sqrt{n}}t_{\alpha,n-1}\right\}=1-\alpha$ or $P\left\{\mu>\bar{X}-\frac{S}{\sqrt{n}}t_{\alpha,n-1}\right\}=1-\alpha$. Hence, if it is observed that X = x, S = s, then we can assert "with 100(1 – α) percent confidence" that $\mu \in \left(\bar{x}-\frac{s}{\sqrt{n}}t_{\alpha,n-1}, \infty\right)$ Similarly, a 100(1 – α) lower confidence interval would be $\mu \in \left(-\infty, \bar{x}+\frac{s}{\sqrt{n}}t_{\alpha,n-1}\right)$ Program 7.3.1 will compute both one- and two-sided confidence intervals for the mean of a normal distribution when the variance is unknown. 7.3.2 Confidence Intervals for the Variance of a Normal Distribution: If X1, . . . , Xn is a sample from a normal distribution having unknown parameters μ and σ2, then we can construct a confidence interval for σ2 by using the fact that $(n-1)\frac{S^2}{\sigma^2}\sim\chi^2_{n-1}$ Hence, $P\left\{\chi^2_{1-\alpha/2,n-1}\leq(n-1)\frac{S^2}{\sigma^2}\leq\chi^2_{\alpha/2,n-1}\right\}=1-\alpha$ or, equivalently, $P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\leq\sigma^2\leq\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\}=1-\alpha$ Hence when $S^2=s^2$, a $100(1-\alpha)$ percent confidence interval for σ^2 is $\left\{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right\}$.

TABLE 7.1 $100(1-\alpha)$ *Percent Confidence Inservals*

$$X_1,\ldots,X_n\sim\mathcal{N}(\mu,\sigma^2)$$

$$\bar{X}=\sum_{i=1}^{n}X_i/n, \qquad S=\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)}$$

Assumption	Parameter	Confidence Interval	Lower Interval	Upper Interval
σ^2 known	μ	$\bar{X}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\left(-\infty,\bar{X}+z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$	$\left(\bar{X}+z_{\alpha}\frac{\sigma}{\sqrt{n}},\infty\right)$
σ^2 unknown	μ	$\bar{X}\pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$	$\left(-\infty,\bar{X}+t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right)$	$\left(\bar{X}-t_{\alpha,n-1}\frac{S}{\sqrt{n}},\infty\right)$
μ unknown	σ^2	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$	$\left(0,\frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}},\infty\right)$

7.4 ESTIMATING THE DIFFERENCE IN MEANS OF TWO NORMAL POPULATIONS: Let X1, X2, . . . , Xn be a sample of size n from a normal population having mean μ1 and variance σ_1^2 and let Y1, . . . , Ym be a sample of size m from a different normal population having mean μ2 and variance σ_2^2 and suppose that the two samples are independent of each other. We are interested in estimating μ1 – μ2. Since $\bar{X}=\sum_{i=1}^{n}X_i/n$ and $\bar{Y}=\sum_{i=1}^{m}Y_i/m$ are the maximum likelihood estimators of μ1 and μ2 it seems intuitive (and can be proven) that X – Y is the maximum likelihood estimator of μ1 – μ2. To obtain a confidence interval estimator, we need the distribution of X –Y . Because $\bar{X}\sim\mathcal{N}(\mu_1,\sigma_1^2/n)$ $\bar{Y}\sim\mathcal{N}(\mu_2,\sigma_2^2/m)$ it follows from the fact that the sum of independent normal random variables is also normal, that $\bar{X}-\bar{Y}\sim\mathcal{N}\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}\right)$. Hence, assuming σ_1^2 and σ_2^2 are known, we have that $\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\sim\mathcal{N}(0,1)$ and so $P\left\{-z_{\alpha/2}<\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}<z_{\alpha/2}\right\}=1-\alpha$ or, equivalently, $P\left\{\bar{X}-\bar{Y}-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}<\mu_1-\mu_2<\bar{X}-\bar{Y}+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right\}=1-\alpha$ Hence, if X and Y are observed to equal x and y, respectively, then a 100(1 – α) two-sided confidence interval estimate for μ1 – μ2 is $\mu_1-\mu_2\in\left(\bar{x}-\bar{y}-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}, \bar{x}-\bar{y}+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right)$. One-sided confidence intervals for μ1 – μ2 are obtained in a similar fashion, and we leave it for the reader to verify that a 100(1 – α) percent one-sided interval is given by $\mu_1-\mu_2\in\left(-\infty,\bar{x}-\bar{y}+z_{\alpha}\sqrt{\sigma_1^2/n+\sigma_2^2/m}\right)$ Program 7.4.1 will compute both one- and two-sided confidence intervals for μ1−μ2. Let us suppose now that we again desire an interval estimator of μ1 – μ2 but that the population variances σ_1^2 and σ_2^2 are unknown. In this case, it is natural to try to replace σ_1^2 and σ_2^2 in Equation 7.4.1 by the sample variances $S_1^2=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1}$ $S_2^2=\sum_{i=1}^{m}\frac{(Y_i-\bar{Y})^2}{m-1}$. That is, it is natural to base our interval estimate on something like $\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{S_1^2/n+S_2^2/m}}$ However, to utilize the foregoing to obtain a confidence interval, we need its distribution and it must not depend on any of the unknown parameters σ_1^2 and σ_2^2 .Unfortunately, this distribution is both complicated and does indeed depend on the unknown parameters σ_1^2 and σ_2^2. In fact, it is only in the special case when $\sigma_1^2=\sigma_2^2$ that we will be able to obtain an interval estimator. So let us suppose that the population variances, though unknown, are equal and let σ2 denote their common value. Now, from Theorem 6.5.1 it follows that $(n-1)\frac{S_1^2}{\sigma^2}\sim\chi^2_{n-1}$ and $(m-1)\frac{S_2^2}{\sigma^2}\sim\chi^2_{m-1}$ Also, because the samples are independent, it follows that these two chi-square random variables are independent. Hence, from the additive property of chi-square random variables, which states that the sum of independent chi-square random variables is also chi-square with a degree of freedom equal to the sum of their degrees of freedom, it follows that $(n-1)\frac{S_1^2}{\sigma^2}\sim\chi^2_{n-1}$ and $(m-1)\frac{S_2^2}{\sigma^2}\sim\chi^2_{m-1}$Also, because the samples are independent, it follows that these two chi-square random variables are independent. Hence, from the additive property of chi-square random variables, which states that the sum of independent chi-square random variables is also chi-square with a degree of freedom equal to the sum of their degrees of freedom, it follows that $(n-1)\frac{S_1^2}{\sigma^2}+(m-1)\frac{S_2^2}{\sigma^2}\sim\chi^2_{n+m-2}$ Also, since $\bar{X}-\bar{Y}\sim\mathcal{N}\left(\mu_1-\mu_2,\frac{\sigma^2}{n}+\frac{\sigma^2}{m}\right)$ we see

$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}} \sim \mathcal{N}(0,1)$ that Now it follows from the fundamental result that in normal sampling $\overline{X}$ and S^2 are independent (Theorem 6.5.1), that $\overline{X}_1, S_1^2, \overline{X}_2, S_2^2$ are independent random variables. Hence, using the definition of a t-random variable (as the ratio of two independent random variables, the numerator being a standard normal and the denominator being the square root of a chi-square random variable divided by its degree of freedom parameter), it follows from Equations 7.4.2 and 7.4.3 that if we let $S_p^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$ then $\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma^2(1/n+1/m)}} \div \sqrt{S_p^2/\sigma^2} = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{S_p^2(1/n+1/m)}}$ has a t –distribution with n + m – 2 degrees of freedom. Consequently,

$$P\left\{-t_{\alpha/2,n+m-2} \le \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n+1/m}} \le t_{\alpha/2,n+m-2}\right\} = 1-\alpha$$

Therefore, when the data result in the values X = x, Y = y, Sp = sp, we obtain the following 100(1 – α) percent confidence interval for μ1 – μ2: $\left(\overline{x}-\overline{y}-t_{\alpha/2,n+m-2}s_p\sqrt{1/n+1/m},\quad \overline{x}-\overline{y}+t_{\alpha/2,n+m-2}s_p\sqrt{1/n+1/m}\right)$ One-sided confidence intervals are similarly obtained. Program 7.4.2 can be used to obtain both one- and two-sided confidence intervals for the difference in means in two normal populations having unknown but equal variances. REMARK: The confidence interval given by Equation 7.4.4 was obtained under the assumption that the population variances are equal; with σ2 as their common value, it follows that $\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma^2/n+\sigma^2/m}} = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n+1/m}}$ has a standard normal distribution. However, since σ^2 is unknown this result cannot be immediately applied to obtain a confidence interval; σ^2 must first be estimated. To do so, note that both sample variances are estimators of σ^2 ; moreover, since S_1^2 has n−1 degrees of freedom and S_2^2 has m−1, the appropriate estimator is to take a weighted average of the two sample variances, with the weights proportional to these degrees of freedom. That is, the estimator of σ^2 is the pooled estimator $S_p^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$ and the confidence interval is then based on the statistic $\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{S_p^2}\sqrt{1/n+1/m}}$ which, by our previous analysis, has a t –distribution with n + m – 2 degrees of freedom. The results of this section are summarized up in Table 7.2.

TABLE 7.2 $100(1-\sigma)$ *Percens Confidence Inuervals for* $\mu_1-\mu_2$

$$X_1,\ldots,X_n \sim N(\mu_1,\sigma_1^2)$$
$$Y_1,\ldots,Y_m \sim N(\mu_2,\sigma_2^2)$$
$$\overline{X}=\sum_{i=1}^{n}X_i/n, \quad S_1^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)$$
$$\overline{Y}=\sum_{i=1}^{m}Y_i/n, \quad S_2^2=\sum_{i=1}^{m}(Y_i-\overline{Y})^2/(m-1)$$

Assumption	Confidence Interval
σ_1,σ_2 known	$\overline{X}-\overline{Y} \pm z_{\alpha/2}\sqrt{\sigma_1^2/n+\sigma_2^2/m}$
σ_1,σ_2 unknown but equal	$\overline{X}-\overline{Y} \pm t_{\alpha/2,n+m-2}\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}$

Assumption	Lower Confidence Interval
σ_1,σ_2 known	$(-\infty, \overline{X}-\overline{Y}+z_\alpha\sqrt{\sigma_1^2/n+\sigma_2^2/m})$
σ_1,σ_2 unknown but equal	$\left(-\infty, \overline{X}-\overline{Y}+t_{\alpha,n+m-2}\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}\right)$

7.5 APPROXIMATE CONFIDENCE INTERVAL FOR THE MEAN OF A BERNOULLI RANDOM VARIABLE: Consider a population of items, each of which independently meets certain standards with some unknown probability p. If n of these items are tested to determine whether they meet the standards, how can we use the resulting data to obtain a confidence interval for p? If we let X denote the number of the n items that meet the standards, then X is a binomial random variable with parameters n and p. Thus, when n is large, it follows by the normal approximation to the binomial that X is approximately normally distributed with mean np and variance np(1 – p). Hence, $\frac{X-np}{\sqrt{np(1-p)}} \dot{\sim} \mathcal{N}(0,1)$ where $\dot{\sim}$ means "is approximately distributed as." Therefore, for any α ∈ (0, 1), $P\left\{-z_{\alpha/2} < \frac{X-np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right\} \approx 1-\alpha$ and so if X is observed to equal x, then an approximate 100(1 – α) percent confidence region for p is $\left\{p : -z_{\alpha/2} < \frac{x-np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right\}$ The foregoing region, however, is not an interval. To obtain a confidence interval for p, let $\hat{p}$ = X /n be the fraction of the items that meet the standards. From Example 7.2a, $\hat{p}$ is the maximum likelihood estimator of p, and so should be approximately equal to p. As a result, + $\sqrt{n\hat{p}(1-\hat{p})}$ will be approximately equal to $\sqrt{np(1-p)}$ and so from so, $\frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}} \dot{\sim} \mathcal{N}(0,1)$ .Hence, for any α ∈ (0, 1) we have that $P\left\{-z_{\alpha/2} < \frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{\alpha/2}\right\} \approx 1-\alpha$ or, equivalently, $P\{-z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < np-X < z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})}\} \approx 1-\alpha$ .Since $\hat{p}$ = X /n, the preceding can be written as $P\{\hat{p}-z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p}+z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\} \approx 1-\alpha$ which yields an approximate 100(1 – α) percent confidence interval for p. We often want to specify an approximate 100(1 – α) percent confidence interval for p that is no greater than some given length, say b. The problem is to determine the appropriate sample size n to obtain such an interval. To do so, note that the length of the approximate 100(1 – α) percent confidence interval for p from a sample of size n is $2z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$, which is approximately equal to $2z_{\alpha/2}\sqrt{p(1-p)/n}$. Unfortunately, p is not known in advance, and so we cannot just set $2z_{\alpha/2}\sqrt{p(1-p)/n}$ equal to b to determine the necessary sample size n. What we can do, however, is to first take a preliminary sample to obtain a rough estimate of p, and then use this estimate to determine n. That is, we use p∗, the proportion of the preliminary sample that meets the standards, as a preliminary estimate of p; we then determine the total sample size n by solving the equation $2z_{\alpha/2}\sqrt{p^*(1-p^*)/n} = b$ Squaring both sides of the preceding yields that $(2z_{\alpha/2})^2p^*(1-p^*)/n = b^2$ or $n = \frac{(2z_{\alpha/2})^2p^*(1-p^*)}{b^2}$ .That is, if k items were initially sampled to obtain the preliminary estimate of p, then an additional n – k (or 0 if n ≤ k) items should be sampled. REMARK: As shown, a 100(1 – α) percent confidence interval for p will be of approximate length b when the sample size is $n = \frac{(2z_{\alpha/2})^2}{b^2}p(1-p)$. Now it is easily shown that the function g (p) = p(1 – p) attains its maximum value of ¼, in the interval 0 ≤ p ≤ 1, when p = ½. Thus an upper bound on n is $n \le \frac{(z_{\alpha/2})^2}{b^2}$ and so by choosing a sample whose size is at least as large as $(z_{\alpha/2})^2/b^2$, one can be assured of obtaining a confidence interval of length no greater than b without need of any additional sampling. One-sided approximate confidence intervals for p are also easily obtained; Table 7.3 gives the results.

TABLE 7.3 *Approximate* $100(1-\alpha)$ *Percens Confidence Inuervals for* p

X Is a Binomial (n,p) Random Variable
$\hat{p}=X/n$

Type of Interval	Confidence Interval
Two-sided	$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
One-sided lower	$\left(-\infty, \hat{p}+z_\alpha\sqrt{\hat{p}(1-\hat{p})/n}\right)$
One-sided upper	$\left(\hat{p}-z_\alpha\sqrt{\hat{p}(1-\hat{p})/n}, \infty\right)$

7.6 CONFIDENCE INTERVAL OF THE MEAN OF THE EXPONENTIAL DISTRIBUTION: If X1, X2, . . . , Xn are independent exponential random variables each having mean θ, then it can be shown that the maximum likelihood estimator of θ is the sample mean $\sum_{i=1}^{n}X_i/n$. To obtain a confidence interval estimator of θ, recall from Section 5.7 that $\sum_{i=1}^{n}X_i$ has a gamma distribution with parameters n, 1/θ. This in turn implies (from the relationship between the gamma and chi-square distribution shown in Section 5.8.1.1) that $\frac{2}{\theta}\sum_{i=1}^{n}X_i \sim \chi^2_{2n}$. Hence, for any α ∈ (0, 1) $P\left\{\chi^2_{1-\alpha/2,2n} < \frac{2}{\theta}\sum_{i=1}^{n}X_i < \chi^2_{\alpha/2,2n}\right\} = 1-\alpha$ or, equivalently, $P\left\{\frac{2\sum_{i=1}^{n}X_i}{\chi^2_{\alpha/2,2n}} < \theta < \frac{2\sum_{i=1}^{n}X_i}{\chi^2_{1-\alpha/2,2n}}\right\} = 1-\alpha$ Hence, a 100(1 – α) percent

confidence interval for θ is $\theta \in \left(\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\alpha/2,2n}}, \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\alpha/2,2n}} \right)$ 7.7 EVALUATING A POINT ESTIMATOR: Let X = (X1, . . . , Xn) be a sample from a population whose distribution is specified up to an unknown parameter θ, and let d = d (X) be an estimator of θ. How are we to determine its worth as an estimator of θ? One way is to consider the square of the difference between d (X) and θ. However, since $(d(\mathbf{X}) - \theta)^2$ is a random variable, let us agree to consider r(d , θ), **the mean square error** of the estimator d , which is defined by $r(d, \theta) = E[(d(\mathbf{X}) - \theta)^2]$ as an indication of the worth of d as an estimator of θ. It would be nice if there were a single estimator d that minimized r(d , θ) for all possible values of θ. However, except in trivial situations, this will never be the case. For example, consider the estimator d∗ defined by $d^*(X_1, \ldots, X_n) = 4$. That is, no matter what the outcome of the sample data, the estimator d∗ chooses 4 as its estimate of θ. While this seems like a silly estimator (since it makes no use of the data), it is, however, true that when θ actually equals 4, the mean square error of this estimator is 0. Thus, the mean square error of any estimator different than d ∗ must, in most situations, be larger than the **mean square error** of d∗ when θ = 4. Although minimum mean square estimators rarely exist, it is sometimes possible to find an estimator having the smallest mean square error among all estimators that satisfy a certain property. One such property is that of unbiasedness. Definition: Let d = d (X) be an estimator of the parameter θ. Then $b_\theta(d) = E[d(\mathbf{X})] - \theta$ is called the bias of d as an estimator of θ. If bθ (d) = 0 for all θ, then we say that d is an unbiased estimator of θ. In other words, an estimator is unbiased if its expected value always equals the value of the parameter it is attempting to estimate. If d (X1, . . . , Xn) is an unbiased estimator, then its mean square error is given by

$$\begin{aligned} r(d, \theta) &= E[(d(\mathbf{X}) - \theta)^2] \\ &= E[(d(\mathbf{X}) - E[d(\mathbf{X})])^2] \quad \text{since } d \text{ is unbiased} \\ &= \text{Var}(d(\mathbf{X})) \end{aligned}$$

Thus the mean square error of an unbiased estimator is equal to its variance. EXAMPLE 7.7b: Combining Independent Unbiased Estimators. Let d1 and d2 denote independent unbiased estimators of θ, having known variances σ_1^2 and σ_2^2. That is, for I = 1, 2, $E[d_i] = \theta, \quad \text{Var}(d_i) = \sigma_i^2$ Any estimator of the form $d = \lambda d_1 + (1 - \lambda) d_2$ will also be unbiased. To determine the value of λ that results in d having the smallest possible mean square error, note that $r(d, \theta) = \text{Var}(d) = \lambda^2 \text{Var}(d_1) + (1 - \lambda)^2 \text{Var}(d_2)$ by the independence of d_1 and d_2 $= \lambda^2 \sigma_1^2 + (1 - \lambda)^2 \sigma_2^2$ Differentiation yields that $\frac{d}{d\lambda} r(d, \theta) = 2\lambda\sigma_1^2 - 2(1 - \lambda)\sigma_2^2$ To determine the value of λ that minimizes r(d , θ) — call it $\hat{\lambda}$ — set this equal to 0 and solve for λ to obtain $2\hat{\lambda}\sigma_1^2 = 2(1 - \hat{\lambda})\sigma_2^2$ or $\hat{\lambda} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}$ In words, the optimal weight to give an estimator is inversely proportional to its variance (when all the estimators are unbiased and independent). For an application of the foregoing, suppose that a conservation organization wants to determine the acidity content of a certain lake. To determine this quantity, they draw some water from the lake and then send samples of this water to n different laboratories. These laboratories will then, independently, test for acidity content by using their respective titration equipment, which is of differing precision. Specifically, suppose that di , the result of a titration test at laboratory I, is a random variable having mean θ, the true acidity of the sample water, and variance σ_i^2, I = 1, . . . , n. If the quantities σ_i^2, I = 1, . . . , n are known to the conservation organization, then they should estimate the acidity of the sampled water from the lake by $d = \frac{\sum_{i=1}^{n} d_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}$ **The mean square error of d** is as follows:

$$\begin{aligned} r(d, \theta) &= \text{Var}(d) \quad \text{since } d \text{ is unbiased} \\ &= \left(\sum_{i=1}^{n} 1/\sigma_i^2 \right)^{-2} \sum_{i=1}^{n} \left(\frac{1}{\sigma_i^2} \right)^2 \sigma_i^2 = \frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2} \end{aligned}$$

A generalization of the result that the mean square error of an unbiased estimator is equal to its variance is that the mean square error of any estimator is equal to its variance plus the square of its bias. This follows since $r(d, \theta) = E[(d(\mathbf{X}) - \theta)^2] = E[(d - E[d] + E[d] - \theta)^2] = E[(d - E[d])^2 + (E[d] - \theta)^2 + 2(E[d] - \theta)(d - E[d])] = E[(d - E[d])^2] + E[(E[d] - \theta)^2] + 2E[(E[d] - \theta)(d - E[d])] = E[(d - E[d])^2] + (E[d] - \theta)^2 + 2(E[d] - \theta)E[d - E[d]]$ since $E[d] - \theta$ is constant $= E[(d - E[d])^2] + (E[d] - \theta)^2$ The last equality follows since E[d – E[d]] = 0 Hence $r(d, \theta) = \text{Var}(d) + b_\theta^2(d)$ 7.8 THE BAYES ESTIMATOR: In certain situations it seems reasonable to regard an unknown parameter θ as being the value of a random variable from a given probability distribution. This usually arises when, prior to the observance of the outcomes of the data X1, . . . , Xn, we have some information about the value of θ and this information is expressible in terms of a probability distribution (called appropriately the **prior** distribution of θ). For instance, suppose that from past experience we know that θ is equally likely to be near any value in the interval (0, 1). Hence, we could reasonably assume that θ is chosen from a uniform distribution on (0, 1). Suppose now that our **prior feelings** about θ are that it can be regarded as being the value of a continuous random variable having probability density function p(θ); and suppose that we are about to observe the value of a sample whose distribution depends on θ. Specifically, suppose that f (x|θ) represents the likelihood — that is, it is the probability mass function in the discrete case or the probability density function in the continuous case — that a data value is equal to x when θ is the value of the parameter. If the observed data values are Xi = xi , I = 1, . . . , n, then the **updated, or conditional**, probability density function of θ is as follows:

$$\begin{aligned} f(\theta|x_1, \ldots, x_n) &= \frac{f(\theta, x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)} \\ &= \frac{p(\theta) f(x_1, \ldots, x_n|\theta)}{\int f(x_1, \ldots, x_n|\theta) p(\theta)\, d\theta} \end{aligned}$$

The conditional density function f (θ|x1, . . . , xn) is called the posterior density function. (Thus, before observing the data, one's feelings about θ are expressed in terms of the prior distribution, whereas once the data are observed, this prior distribution is updated to yield the posterior distribution.) The conditional density function f (θ|x1, . . . , xn) is called **the posterior density function**. (Thus, before observing the data, one's feelings about θ are expressed in terms of the prior distribution, whereas once the data are observed, this prior distribution is updated to yield the posterior distribution.) Now we have shown that whenever we are given the probability distribution of a random variable, the best estimate of the value of that random variable, in the sense of **minimizing the expected squared error**, is its mean. Therefore, it follows that **the best estimate of θ**, given the data values Xi = xi , I = 1, . . . , n, **is the mean of the posterior distribution** f (θ|x1, . . . , xn). This estimator, called **the Bayes estimator**, is written as E[θ|X1, . . . , Xn]. That is, if Xi = xi , I = 1, . . . , n, then the value of the Bayes estimator is $E[\theta|X_1 = x_1, \ldots, X_n = x_n] = \int \theta f(\theta|x_1, \ldots, x_n)\, d\theta$ EXAMPLE 7.8a Suppose that X1, . . . , Xn are independent Bernoulli random variables, each having probability mass function given by $f(x|\theta) = \theta^x (1 - \theta)^{1-x}, \quad x = 0, 1$ where θ is unknown. Further, suppose that θ is chosen from a uniform distribution on (0, 1). Compute the Bayes estimator of θ. We must compute E[θ|X1, . . . , Xn]. Since the prior density of θ is the uniform Density $p(\theta) = 1, \quad 0 < \theta < 1$ we have that the conditional density of θ given X1, . . . , Xn is given by

$$\begin{aligned} f(\theta|x_1, \ldots, x_n) &= \frac{f(x_1, \ldots, x_n, \theta)}{f(x_1, \ldots, x_n)} \\ &= \frac{f(x_1, \ldots, x_n|\theta) p(\theta)}{\int_0^1 f(x_1, \ldots, x_n|\theta) p(\theta)\, d\theta} \\ &= \frac{\theta^{\Sigma_1^n x_i} (1 - \theta)^{n - \Sigma_1^n x_i}}{\int_0^1 \theta^{\Sigma_1^n x_i} (1 - \theta)^{n - \Sigma_1^n x_i}\, d\theta} \end{aligned}$$

Now it can be shown that for integral values m and r $\int_0^1 \theta^m (1 - \theta)^r\, d\theta = \frac{m!r!}{(m + r + 1)!}$ Hence, upon letting $x = \sum_{i=1}^{n} x_i$ $f(\theta|x_1, \ldots, x_n) = \frac{(n + 1)!}{x!(n - x)!} \theta^x (1 - \theta)^{n-x}$ Therefore, $E[\theta|x_1, \ldots, x_n] = \frac{(n + 1)!}{x!(n - x)!} \int_0^1 \theta^{1+x} (1 - \theta)^{n-x}\, d\theta = \frac{(n + 1)!}{x!(n - x)!} \frac{(1 + x)!(n - x)!}{(n + 2)!}$ from Equation 7.8.1 $= \frac{x + 1}{n + 2}$ Thus, the Bayes estimator is given by $E[\theta|X_1, \ldots, X_n] = \frac{\sum_{i=1}^{n} X_i + 1}{n + 2}$ As an illustration, if 10 independent trials, each of which results in a success with probability θ, result in 6 successes, then assuming a uniform (0, 1) prior distribution on θ, the **Bayes estimator of θ is** 7/12 (**as opposed**, for instance, to the **maximum likelihood estimator** of 6/10). REMARK: The conditional distribution of θ given that Xi = xi , I = 1, . . . , n, whose density function is given by Equation 7.8.2, is called the **beta distribution** with parameters $\sum_{i=1}^{n} x_i + 1, n - \sum_{i=1}^{n} x_i + 1$. EXAMPLE 7.8b Suppose X1, . . . , Xn are independent normal random variables, each having unknown mean θ and known variance σ_0^2. If θ is itself selected from a normal population having known mean μ and known variance σ^2, what is the Bayes estimator of θ? In order to determine E[θ|X1, . . . , Xn], the Bayes estimator, we need first determine the conditional density of θ given the values of X1, . . . , Xn. Now $f(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta) p(\theta)}{f(x_1, \ldots, x_n)}$ where $f(x_1, \ldots, x_n|\theta) = \frac{1}{(2\pi)^{n/2} \sigma_0^n} \exp\left\{ -\sum_{i=1}^{n} (x_i - \theta)^2 / 2\sigma_0^2 \right\}$ $p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(\theta - \mu)^2 / 2\sigma^2\}$ and $f(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} f(x_1, \ldots, x_n|\theta) p(\theta)\, d\theta$ With the help of a little algebra, it can now be shown that this

conditional density is a normal density with mean $E[\theta|X_1,\ldots,X_n]=\frac{n\sigma^2}{n\sigma^2+\sigma_0^2}\overline{X}+\frac{\sigma_0^2}{n\sigma^2+\sigma_0^2}\mu=\frac{\frac{n}{\sigma_0^2}}{\frac{n}{\sigma_0^2}+\frac{1}{\sigma^2}}\overline{X}+\frac{\frac{1}{\sigma^2}}{\frac{n}{\sigma_0^2}+\frac{1}{\sigma^2}}\mu$ and variance $\mathrm{Var}(\theta|X_1,\ldots,X_n)=\frac{\sigma_0^2\sigma^2}{n\sigma^2+\sigma_0^2}$ Writing the Bayes estimator as we did in Equation 7.8.3 is informative, for it shows that it is a weighted average of X , the sample mean, and μ, the a priori mean. In fact, the weights given to these two quantities are in proportion to the inverses of σ_0^2/n (the conditional variance of the sample mean X given θ) and σ^2 (the variance of the prior distribution). REMARK: **ON CHOOSING A NORMAL PRIOR:** As illustrated by Example 7.8b, it is computationally very convenient to choose a normal prior for the unknown mean θ of a normal distribution — for then the Bayes estimator is simply given by Equation 7.8.3. This raises the question of how one should go about determining whether there is a normal prior that reasonably represents one's prior feelings about the unknown mean. To begin, it seems reasonable to determine the value — call it μ — that you a priori feel is most likely to be near θ. That is, we start with the mode (which equals the mean when the distribution is normal) of the prior distribution. We should then try to ascertain whether or not we believe that the prior distribution is symmetric about μ. That is, for each a > 0 do we believe that it is just as likely that θ will lie between μ–a and μ as it is that it will be between μ and μ+a? If the answer is positive, then we accept, as a working hypothesis, that our **prior feelings** about θ can be expressed in terms of a prior distribution that is normal with mean μ. To determine σ, the standard deviation of the normal prior, think of an interval centered about μ that you a priori feel is 90 percent certain to contain θ. For instance, suppose you feel 90 percent (no more and no less) certain that θ will lie between μ – a and μ + a. Then, since a normal random variable θ with mean μ and variance σ^2 is such that $P\left\{-1.645<\frac{\theta-\mu}{\sigma}<1.645\right\}=.90$ or $P\{\mu-1.645\sigma<\theta<\mu+1.645\sigma\}=.90$ it seems reasonable to take $1.645\sigma=a \quad \text{or} \quad \sigma=\frac{a}{1.645}$ Thus, if your prior feelings can indeed be reasonably described by a normal distribution, then that distribution would have mean μ and standard deviation σ = a/1.645. As a test of whether this distribution indeed fits your prior feelings you might ask yourself such questions as whether you are 95 percent certain that θ will fall between μ – 1.96σ and μ+1.96σ, or whether you are 99 percent certain that θ will fall between μ–2.58σ and μ + 2.58σ, where these intervals are determined by the equalities $P\left\{-1.96<\frac{\theta-\mu}{\sigma}<1.96\right\}=.95$, $P\left\{-2.58<\frac{\theta-\mu}{\sigma}<2.58\right\}=.99$ which hold when θ is normal with mean μ and variance σ^2. EXAMPLE 7.8c Consider the likelihood function f (x1, . . . , xn|θ) and suppose that θ is

$$f(\theta|x_1,\ldots,x_n)=\frac{f(x_1,\ldots,x_n|\theta)p(\theta)}{\int_a^b f(x_1,\ldots,x_n|\theta)p(\theta)\,d\theta}=\frac{f(x_1,\ldots,x_n|\theta)}{\int_a^b f(x_1,\ldots,x_n|\theta)\,d\theta}\qquad a<\theta<b$$

uniformly distributed over some interval (a, b). The posterior density of θ given X1, . . . , Xn equals Now the **mode** of a density f (θ) was defined to be that value of θ that maximizes f (θ). By the foregoing, it follows that the mode of the density f (θ|x1, . . . , xn) is that value of θ maximizing f (x1, . . . , xn|θ); that is, it is just the maximum likelihood estimate of θ [when it is constrained to be in (a, b)]. In other words, the maximum likelihood estimate equals the mode of the posterior distribution when a uniform prior distribution is assumed. If, rather than a point estimate, we desire an interval in which θ lies with a specified probability —say 1 – α —we can accomplish this by choosing values a and b such that $\int_a^b f(\theta|x_1,\ldots,x_n)\,d\theta=1-\alpha$

Chapter_9: REGRESSION: 9.1 INTRODUCTION: Many engineering and scientific problems are concerned with determining a relationship between a set of variables. In many situations, there is a single response variable Y , also called the dependent variable, which depends on the value of a set of input, also called independent, variables x1, . . . , xr . The simplest type of relationship between the dependent variable Y and the input variables x1, . . . , xr is a linear relationship. That is, for some constants β0, β1, . . . , βr the equation $Y=\beta_0+\beta_1x_1+\cdots+\beta_rx_r$ It would be possible (once the βi were learned) to exactly predict the response for any set of input values. However, in practice, such precision is almost never attainable, subject to random error. $Y=\beta_0+\beta_1x_1+\cdots+\beta_rx_r+e$, linear regression equation. We say that it describes the regression of Y on the set of independent variables x1, . . . , xr, The quantities β0, β1, . . . , βr are called the regression coefficients, and must usually be estimated from a set of data. A regression equation containing a single independent variable — that is, one in which r = 1 — is called a simple regression equation, It can be expressed as $Y=\alpha+\beta x+e$ where x is the value of the independent variable, also called the input level, Y is the response, and e, representing the random error, is a random variable having mean 0; whereas one containing many independent variables is called a multiple regression equation. Another way of expressing $E[Y|\mathbf{x}]=\beta_0+\beta_1x_1+\cdots+\beta_rx_r$.where x = (x1, . . . , xr) is the set of independent variables, and E[Y |x] is the expected response given the inputs x. Consider the following 10 data pairs (xi , yi), I = 1, . . . , 10, relating y to x. A plot of yi versus xi — called a scatter diagram Figure 9.1—As this scatter diagram appears to reflect, subject to random error, a linear relation between y and x, it seems that a simple linear regression model would be appropriate. 9.2 LEAST SQUARES ESTIMATORS OF THE REGRESSION PARAMETERS: Suppose that the responses Yi corresponding to the input values xi , I = 1, . . . , n are to be observed and used to estimate α and β in a simple linear regression model.If A is the estimator of α and B of β, then the **estimator** of the response corresponding to the input variable xi would be A + Bxi . The **actual** response is Yi , the squared difference is $(Y_i-A-Bx_i)^2$, and so if A and B are the estimators of α and β, then the sum of the squared differences between the estimated responses and the actual response values—call it SS —is given by $SS=\sum_{i=1}^{n}(Y_i-A-Bx_i)^2$, This is method of least squares, now minimize SS, we differentiate SS first with respect to A and then to B as follows: $\frac{\partial SS}{\partial A}=-2\sum_{i=1}^{n}(Y_i-A-Bx_i)\quad\frac{\partial SS}{\partial B}=-2\sum_{i=1}^{n}x_i(Y_i-A-Bx_i)$ Setting these partial derivatives equal to zero yields the following equations for the minimizing values A and B: $\sum_{i=1}^{n}Y_i=nA+B\sum_{i=1}^{n}x_i\quad\sum_{i=1}^{n}x_iY_i=A\sum_{i=1}^{n}x_i+B\sum_{i=1}^{n}x_i^2$, These Equations are known as the **normal equations**. Now, let $\overline{Y}=\sum_i Y_i/n,\qquad\overline{x}=\sum_i x_i/n$ The first normal equation as: $A=\overline{Y}-B\overline{x}$,Substituting this value of A into the second normal equation yields $\sum_i x_iY_i=(\overline{Y}-B\overline{x})n\overline{x}+B\sum_i x_i^2$, i.e., $B=\frac{\sum_i x_iY_i-n\overline{x}\,\overline{Y}}{\sum_i x_i^2-n\overline{x}^2}$ using $\text{the fact that } n\overline{Y}=\sum_{i=1}^{n}Y_i$, i.e. The least squares estimators of β and α corresponding to the data set xi , Yi , I = 1, . . . , n are, B(for β) & A(for α), respectively, The straight line A + Bx is called the **estimated regression line**. 9.3 DISTRIBUTION OF THE ESTIMATORS: To specify the distribution of the estimators A and B, it is necessary to make additional assumptions about the random errors aside from just assuming that their mean is 0. The usual approach is to assume that the random errors are independent normal random variables having mean 0 and variance σ^2, it is supposed that the variance of the random error does not depend on the input value but rather is a constant. This value σ^2 is not assumed to be known but rather **must be estimated** from the data. Now, B is a linear combination of the independent normal random variables Yi , I = 1, . . . , n and so is itself normally distributed. The mean and variance of B are computed as follows: $E[B]=\frac{\sum_i(x_i-\overline{x})E[Y_i]}{\sum_i x_i^2-n\overline{x}^2}$ = β. Thus E[B] = β and so B is an unbiased estimator of β. We will now compute the variance of B $\mathrm{Var}(B)=\frac{\mathrm{Var}\left(\sum_{i=1}^{n}(x_i-\overline{x})Y_i\right)}{\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)^2}$ using identity $\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-n\overline{x}^2$ .A can also be expressed as a linear combination of the independent normal random variables Yi , I = 1, . . . , n and is thus also normally distributed. $A=\sum_{i=1}^{n}\frac{Y_i}{n}-B\overline{x}$. Its mean is obtained from $E[A]=\sum_{i=1}^{n}\frac{E[Y_i]}{n}-\overline{x}E[B]$ = α. Thus A is also an unbiased estimator. The variance of A is computed by first

$$\mathrm{Var}(A) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

expressing A as a linear combination of the Yi. The quantities Yi –A–Bxi , I = 1, . . . , n, which represent the differences between the actual responses (that is, the Yi) and their least squares estimators (that is, A + Bxi) are called the **residuals**. The sum of squares of the residuals $SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$ can be utilized to estimate the unknown error variance σ^2. Indeed, it can be shown that $\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$. That is, SS_R/σ^2 has a chi-square distribution with n−2 degrees of freedom, which implies that $E\left[\frac{SS_R}{\sigma^2}\right] = n-2$ or $E\left[\frac{SS_R}{n-2}\right] = \sigma^2$. Thus $SS_R/(n-2)$ is an unbiased estimator of σ2. In addition, it can be shown that SSR is independent of the pair A and B. Why SSR/σ^2 might have a chi-square distribution with n−2 degrees of freedom and be independent of A and B runs as follows. Because the Yi are independent normal random variables, it follows that $(Y_i - E[Y_i])/\sqrt{\mathrm{Var}(Y_i)}, i = 1, \ldots, n$ are independent standard 9ormal and so $\sum_{i=1}^{n} \frac{(Y_i - E[Y_i])^2}{\mathrm{Var}(Y_i)} = \sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2} \sim \chi^2_n$ .Now if we substitute the estimators A and B for α and β, then 2 degrees of freedom are **lost**, and so it is not an altogether surprising result that SSR/σ^2 has a chi-square distribution with n – 2 degrees of freedom. [1.The fact that SSR is independent of A and B is quite similar to the fundamental result that in normal sampling X and S^2 are independent.2. if Y1, . . . , Yn is a normal sample with population mean μ and variance σ2, then if in the sum of squares $\sum_{i=1}^{n}(Y_i - \mu)^2/\sigma^2$, which has a chi-square distribution with n degrees of freedom, one substitutes the estimator $\bar{Y}$ for μ to obtain the new sum of squares $\sum_i (Y_i - \bar{Y})^2/\sigma^2$, then this quantity [equal to $(n-1)S^2/\sigma^2$] will be independent of Y and will have a chi-square distribution with n – 1 degrees of freedom.3. Since SS_R/σ^2 is obtained by substituting the estimators A and B for α and β in the sum of squares $\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2/\sigma^2$, it is not unreasonable to expect that this quantity might be independent of A and B.] When the Yi are normal random variables, the **least square estimators** are **also** the **maximum likelihood estimators.** To verify this remark, note that the joint density of Y1, . . . , Yn is given by

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \alpha - \beta x_i)^2/2\sigma^2}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2/2\sigma^2}$$

Consequently, the maximum likelihood estimators of α and β are precisely the values of α and β that minimize $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$That is, they are the least squares estimators. **Notation**: If we let $S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$ $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$ $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$, then the least squares estimators can be expressed as $B = \frac{S_{xY}}{S_{xx}}$ $A = \bar{Y} - B\bar{x}$. The following computational identity for SSR, the sum of squares of the residuals, can be established. $SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$. 9.4 STATISTICAL INFERENCES ABOUT THE REGRESSION PARAMETERS: Using above **notations**, it is a simple matter to devise hypothesis tests and confidence intervals for the regression parameters. 9.4.1 Inferences Concerning β: An important hypothesis to consider regarding the simple linear regression model Y = α + βx + e, i.e., is the hypothesis that β = 0. Its importance derives from the fact that it is equivalent to stating that the mean response does not depend on the input, or, equivalently, that there is no regression on the input variable. To test H0:β=0 versus H1:β $\neq$ 0 Note $\frac{B - \beta}{\sqrt{\sigma^2/S_{xx}}} = \sqrt{S_{xx}}\frac{(B-\beta)}{\sigma} \sim \mathcal{N}(0,1)$ and is independent of $\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$. Hence, from the definition of a t-random variable it follows that $\frac{\sqrt{S_{xx}}(B-\beta)/\sigma}{\sqrt{\frac{SS_R}{\sigma^2(n-2)}}} = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) \sim t_{n-2}$ That is, $\sqrt{(n-2)S_{xx}/SS_R}(B-\beta)$ has a t-distribution with n – 2 degrees of freedom. Therefore, if H0 is true (and so β = 0), then $\sqrt{\frac{(n-2)S_{xx}}{SS_R}}B \sim t_{n-2}$ which gives rise to the following test of H0. Hypothesis Test of H0: β = 0, A significance level γ test of H0 is to

$$\begin{array}{ll} \text{reject} & H_0 \quad \text{if } \sqrt{\frac{(n-2)S_{xx}}{SS_R}}|B| > t_{\gamma/2,n-2} \\ \text{accept} & H_0 \quad \text{otherwise} \end{array}$$

. This test can be performed by first computing the value of the test statistic $\sqrt{(n-2)S_{xx}/SS_R}|B|$ —call its value v— $p\text{-value} = P\{|T_{n-2}| > v\}$ $= 2P\{T_{n-2} > v\}$. Where T_{n-2} is a t-random variable with n – 2 degrees of freedom. This latter probability can be obtained by using Program 5.8.2a**. A confidence interval estimator for β** is easily obtained from definition of a t-random variable & that for any a, 0 < a < 1, $P\left\{-t_{a/2,n-2} < \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) < t_{a/2,n-2}\right\} = 1-a$ or, equivalently, $P\left\{B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2} < \beta < B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right\} = 1-a$ which yields the following: Confidence Interval for β: A 100(1 – a) percent confidence interval estimator of β is $\left(B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}, B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right)$ **REMARK**:The result that $\frac{B-\beta}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0,1)$ cannot be immediately applied to make inferences about β since it involves the unknown parameter σ^2. Instead, what we do is use the preceding statistic with σ^2 replaced by its estimator $SS_R/(n-2)$, which has the **effect of changing the distribution of the statistic** from the standard normal to the t-distribution with n – 2 degrees of freedom. 9.4.1.1 REGRESSION TO THE MEAN: The term regression was originally employed by Francis Galton while describing the laws of inheritance. Galton believed that these laws caused population extremes to “regress toward the mean.” By this he meant that children of individuals having extreme values of a certain characteristic would tend to have less extreme values of this characteristic than their parent. If we assume a linear regression relationship between the characteristic of the offspring (Y), and that of the parent (x), then a regression to the mean will occur when the regression parameter β is between 0 and 1. That is, if $E[Y] = \alpha + \beta x$ and 0 < β < 1, then E[Y] will be smaller than x when x is large and greater than x when x is small. That this statement is true can be easily checked either algebraically or by plotting the two straight lines y = α + βx and y = x. A plot indicates that, when 0 < β < 1, the line y = α + βx is above the line y = x for small values of x and is below it for large values of x. EXAMPLE 9.4c To illustrate Galton’s thesis of regression to the mean, the British statistician Karl Pearson plotted the heights of 10 randomly chosen sons versus that of their fathers. The resulting data (in inches) were as follows.

Fathers' height	60	62	64	65	66	67	68	70	72	74
Sons' height	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

A scatter diagram representing these data is presented in Figure 9.6.

Note that whereas the data appear to indicate that taller fathers tend to have taller sons, it also appears to indicate that the sons of fathers that are either extremely short or extremely tall tend to be more "average" than their fathers—that is, there is a "regression toward the mean."
We will determine whether the preceding data are strong enough to prove that there is a regression toward the mean by taking this statement as the alternative hypothesis. That is, we will use the above data to test $H_0 : \beta \geq 1$ versus $H_1 : \beta < 1$ which is equivalent to a test of $H_0 : \beta = 1$ versus $H_1 : \beta < 1$ Hence, from the definition of a t-random variable it follows that when β = 1, the test statistic $TS = \sqrt{8S_{xx}/SS_R}(B-1)$ has a t-distribution with 8 degrees of freedom. The significance level α test will reject H0 when the value of TS is sufficiently small (since this will occur when B, the estimator of β, is sufficiently smaller than 1). Specifically, the test is to reject H_0 if $\sqrt{8S_{xx}/SS_R}(B-1) < -t_{\alpha,8}$ Program 9.2 gives that $\sqrt{8S_{xx}/SS_R}(B-1) = 30.2794(.4646-1) = -16.21$ Since $t_{.01,8} = 2.896$, we see that $TS < -t_{.01,8}$,and so the null hypothesis that β ≥ 1 is rejected at the 1 percent level of significance. In fact, the p-value is p-value = P{T8 ≤ −16.213} ≈ 0 and so the null hypothesis that β ≥ 1 is rejected at almost any significance level, thus establishing a regression toward the mean (see Figure 9.7). A modern biological explanation for the regression to the mean phenomenon would roughly go along the lines of noting that as an offspring obtains a random selection of one-half of its parents' genes, it follows that the offspring of, say, a very tall parent would, by chance, tend to have fewer "tall" genes than its parent. While the most important applications of the regression to the mean phenomenon concern the relationship between the biological characteristics of an offspring and that of its parents, this phenomenon also arises in situations where we have two sets of data referring to the same variables. EXAMPLE 9.4d The data of Table 9.1 relate the number of motor vehicle deaths occurring
in 12 counties in the northwestern United States in the years 1988 and 1989. A glance at Figure 9.8 indicates that in 1989 there was, for the most part, a reduction in the number of deaths in those counties that had a large number of motor deaths in 1988. Similarly, there appears to have been an increase in those counties that had a low value in 1988. Thus, we would expect that a regression to the mean is in effect. In fact, running Program 9.2 yields that the estimated regression equation is y = 74.589 + .276x showing that the estimated value of β indeed appears to be less than 1. One must be careful when considering the reason behind the regression to the mean phenomenon in the preceding data. For instance, it might be natural to suppose that those counties that had a large number of deaths caused by motor vehicles in 1988 would have made a large effort — perhaps by improving the safety of their roads or by making people more aware of the potential dangers of unsafe driving — to reduce this number. In addition, we might suppose that those counties that had the fewest number of deaths in 1988 might have "rested on their laurels" and not made much of an effort to further improve their numbers — and as a result had an increase in the number of casualties the following year. While the above supposition might be correct, it is important to realize that a regression to the mean would probably have occurred even if none of the counties had done anything out of the ordinary. Indeed, it could very well be the case that those counties having large numbers of casualties in 1988 were just very unlucky in that year and thus a decrease in the next year was just a return to a more normal result for them. (For an analogy, if 9 heads results when 10 fair coins are flipped then it is quite likely that another flip of these
10 coins will result in fewer than 9 heads.) Similarly, those counties having few deaths in 1988 might have been "lucky" that year and a more normal result in 1989 would thus lead to an increase. The mistaken belief that regression to the mean is due to some **outside influence** when
it is in reality just due to "**chance**" occurs frequently enough that it is often referred to as the **regression fallacy**. 9.4.2 Inferences Concerning α: The determination of confidence intervals and hypothesis tests for α is accomplished in exactly the same manner as was done for β. Specifically, using
expression for **B** can be used to show that Equation_X $\sqrt{\frac{n(n-2)S_{xx}}{\sum_i x_i^2 SS_R}}(A-\alpha) \sim t_{n-2}$ which leads to the following confidence interval estimator of α.
Confidence Interval Estimator of α: The 100(1 – a) percent confidence interval for α is the interval: $A \pm \sqrt{\frac{\sum_i x_i^2 SS_R}{n(n-2)S_{xx}}} t_{a/2,n-2}$ Hypothesis tests concerning α are easily obtained from Equation_X. 9.4.3 Inferences Concerning the Mean Response α+βx0: It is often of interest to use the data pairs (xi , Yi), I = 1, . . . , n, to estimate α + βx0, the mean response for a given input level x0. If it is a **point estimator** that is desired, then the
natural estimator is A + Bx0, which is an unbiased estimator since E[A + Bx0] = E[A] + x0E[B] = α + βx0 However, if **we desire a confidence interval, or are interested in testing some hypothesis about this mean response**, then it is necessary to first determine the probability distribution of the
estimator A + Bx0 . We now do so. Using the expression for **B**, yields that $B = c\sum_{i=1}^{n}(x_i - \bar{x})Y_i$ where $c = \frac{1}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{1}{S_{xx}}$ Since $A = \bar{Y} - B\bar{x}$ we see that
$A + Bx_0 = \frac{\sum_{i=1}^{n} Y_i}{n} - B(\bar{x} - x_0) = \sum_{i=1}^{n} Y_i \left[\frac{1}{n} - c(x_i - \bar{x})(\bar{x} - x_0)\right]$ Since the Yi are independent normal random variables, the foregoing equation shows that A + Bx0 can be expressed as a linear combination of independent normal random variables and is thus itself normally distributed. Because we already
know its mean, we need only compute its variance, which is accomplished as follows:
$\mathrm{Var}(A + Bx_0) = \sum_{i=1}^{n}\left[\frac{1}{n} - c(x_i - \bar{x})(\bar{x} - x_0)\right]^2 \mathrm{Var}(Y_i)$
$= \sigma^2\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]$ where the last equality followed from $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 1/c = S_{xx}, \quad \sum_{i=1}^{n}(x_i - \bar{x}) = 0$
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 1/c = S_{xx}, \quad \sum_{i=1}^{n}(x_i - \bar{x}) = 0$ Hence, we have shown that $A + Bx_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$,In addition, because A + Bx0
is independent of $SS_R/\sigma^2 \sim \chi^2_{n-2}$ it follows that $\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}$, this can now be used to obtain the following confidence interval estimator
of α + βx0. Confidence Interval Estimator of α + βx0: With 100(1 – a) percent confidence, α + βx0 will lie within: $A + Bx_0 \pm \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}} t_{a/2,n-2}$
9.4.4 Prediction Interval of a Future Response: It is often the case that it is more important to estimate the actual value of a future response rather than its mean value. For instance, if an experiment is to be performed at temperature level x0, then we would probably be more interested in predicting Y (x0), the yield from this experiment, than we would be in estimating the expected yield — E[Y (x0)] = α + βx0. (On the other hand, if a series of experiments were to be performed at input level x0, then we would probably want to estimate α + βx0, the mean yield.) Suppose first that we are interested in a single value (as opposed to an interval) to use as a predictor of Y (x0), the response at level x0. Now, it is clear that the best predictor of Y (x0) is its mean value α + βx0. [Actually, this is not so immediately obvious since one could argue that the best predictor of a random variable is (1) its **mean**—which minimizes the expected square of the difference between the predictor and the actual value; or (2) its **median** — which minimizes the expected absolute difference between the predictor and the actual value; or (3) its **mode**—which is the most likely value to occur. However, as **the mean, median, and mode of a normal random variable are all equal**—and the response is, by assumption, normally distributed—there is no doubt in this situation.] Since α and β are not known, it seems reasonable to use their estimators A and B and thus use A + Bx0 as the predictor of a new response at input level x0. Let us now suppose that rather than being concerned with determining a single value to predict a response, we are interested in finding a prediction interval that, with a given degree of confidence, will contain the response. To obtain such

an interval, let Y denote the future response whose input level is x0 and consider the probability distribution of the response minus its predicted value— that is, the distribution of Y – A – Bx0. Now, $Y \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2)$, and, as was shown before above, $A + Bx_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$, Hence, because Y is independent of the earlier data values Y1, Y2, . . . , Yn that were used to determine A and B, it follows that Y is independent of A + Bx0 and so $Y - A - Bx_0 \sim \mathcal{N}\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$, or, equivalently, $\frac{Y - A - Bx_0}{\sigma\sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim \mathcal{N}(0, 1)$ equation_x1, Now, using once again the result that SSR is independent of A and B (and also of Y) and $\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$, we obtain, by the usual argument, upon replacing σ^2 in equation_x1 by its estimator $SS_R/(n-2)$ that $\frac{Y - A - Bx_0}{\sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}$ and so, for any value a, 0 < a < 1, $p\left\{-t_{a/2,n-2} < \frac{Y - A - Bx_0}{\sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} < t_{a/2,n-2}\right\} = 1 - a$

That is, we have just established the following. Prediction Interval for a Response at the Input Level x0: Based on the response values Yi corresponding to the input values xi , I = 1, 2, . . . , n:

With 100(1−a) percent confidence, the response Y at the input level x0 will be contained in the interval $A + Bx_0 \pm t_{a/2,n-2}\sqrt{\left[\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\frac{SS_R}{n-2}}$

REMARKS:

(a) There is often some confusion about the difference between a confidence and a prediction interval. A confidence interval is an interval that does contain, with a given degree of confidence, a **fixed parameter** of interest. A prediction interval, on the other hand, is an interval that will contain, again with a given degree of confidence, a **random variable** of interest.

(b**) One should not make predictions about responses at input levels that are far from those used to obtain the estimated regression line**. For instance, the data of Example 9.4c should not be used to predict the height of a male whose father is 42 inches tall.

9.4.5 Summary of Distributional Results:

Model: $Y = \alpha + \beta x + e, \quad e \sim \mathcal{N}(0, \sigma^2)$

Data: $(x_i, Y_i), \quad i = 1, 2, \ldots, n$

Inferences About	Use the Distributional Result
β	$\sqrt{\frac{(n-2)S_{xx}}{SS_r}}(B - \beta) \sim t_{n-2}$
α	$\sqrt{\frac{n(n-2)S_{xx}}{\sum_i x_i^2 SS_R}}(A - \alpha) \sim t_{n-2}$

Inferences About	Use the Distributional Result
$\alpha + \beta x_0$	$\frac{A + Bx_0 - \alpha - \beta x_0}{\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\left(\frac{SS_R}{n-2}\right)}} \sim t_{n-2}$
$Y(x_0)$	$\frac{Y(x_0) - A - Bx_0}{\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\left(\frac{SS_R}{n-2}\right)}} \sim t_{n-2}$

9.5 THE COEFFICIENT OF DETERMINATION AND THE SAMPLE CORRELATION COEFFICIENT: Suppose we wanted to measure the amount of variation in the set of response values Y1, . . . , Yn corresponding to the set of input values x1, . . . , xn. A standard measure in statistics of the amount of variation in a set of values Y1, . . . , Yn is given by the quantity $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ For instance, if all the Yi are equal — and thus are all equal to $\bar{Y}$ — then S_{YY} would equal 0. The variation in the values of the Yi arises from two factors. **First**, because the input values xi are different, the response variables Yi all have different mean values, which will result in some variation in their values. **Second**, the variation also arises from the fact that even when the differences in the input values are taken into account, each of the response variables Yi has variance σ2 and thus will not exactly equal the predicted value at its input xi. Let us consider now the question as to how much of the variation in the values of the response variables is due to the different input values, and how much is due to the inherent variance of the responses even when the input values are taken into account. To answer this question, note that the quantity $SS_R = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2$ measures the remaining amount of variation in the response values after the different input values have been taken into account. Thus, $S_{YY} - SS_R$ represents the amount of variation in the response variables that is explained by the different input values; and so the quantity R^2 defined by $R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}}$, represents the proportion of the variation in the response variables that is explained by the different input values. R^2 is called the coefficient of determination. The coefficient of determination R^2 will have a value between 0 and 1. A value of R^2 near 1 indicates that most of the variation of the response data is explained by the different input values, whereas a value of R^2 near 0 indicates that little of the variation is explained by the different input values. The value of R^2 is often used as an indicator of how well the regression model fits the data, with a value near 1 indicating a good fit, and one near 0 indicating a poor fit. In other words, if the regression model is able to explain most of the variation in the response data, then it is considered to fit the data well. Recall that in **Section 2.6** we defined the sample correlation coefficient r of the set of data pairs (xi , Yi), I = 1, . . . , n, by $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$ It was noted that r provided a measure of the degree to which high values of x are paired with high values of Y and low values of x with low values of Y . A value of r near +1 indicated that large x values were strongly associated with large Y values and small x values were strongly associated with small Y values, whereas a value near −1 indicated that large x values were strongly associated with small Y values and small x values with large Y values. In the notation of this chapter, $r = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}}$ Upon using identity for SS_R , $SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$, we see that, $r^2 = \frac{S_{xY}^2}{S_{xx}S_{YY}} = \frac{S_{xx}S_{YY} - SS_R S_{xx}}{S_{xx}S_{YY}} = 1 - \frac{SS_R}{S_{YY}} = R^2$ That is, $|r| = \sqrt{R^2}$ and so, except for its sign indicating whether it is positive or negative, the sample correlation coefficient is equal to the square root of the coefficient of determination. The sign of r is the same as that of B. The above gives additional meaning to the sample correlation coefficient. For instance, if a data set has its sample correlation coefficient r equal to .9, then this implies that a simple linear regression model for these data explains 81 percent (since R^2 = .92 = .81) of the variation in the response values. That is, 81 percent of the variation in the response values is explained by the different input values.

9.6 ANALYSIS OF RESIDUALS: ASSESSING THE MODEL: The initial step for ascertaining whether or not the simple linear regression model $Y = \alpha + \beta x + e, \quad e \sim \mathcal{N}(0, \sigma^2)$ is appropriate in a given situation is to investigate the scatter diagram. Indeed, this is often sufficient to convince one that the regression model is or is not correct. When the scatter diagram does not by itself rule out the preceding model, then the least square

estimators A and B should be computed and the residual Yi – (A + Bxi), I = 1, . . . , n 12ormal1212. The analysis begins by normalizing, or standardizing, the residuals by dividing them by $\sqrt{SS_R/(n-2)}$, the estimate of the standard deviation of the Yi . The resulting quantities $\frac{Y_i-(A+Bx_i)}{\sqrt{SS_R/(n-2)}}, \quad i=1,\ldots,n$ are called the **standardized residuals.** When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables, and thus should be randomly distributed about 0 with about 95 percent of their values being between −2 and +2 (since P{−1.96 < Z < 1.96} = .95). In addition, a plot of the standardized residuals should not indicate any distinct pattern. Indeed, any indication of a distinct pattern should make one suspicious about the validity of the assumed simple linear regression model. 9.7 TRANSFORMING TO LINEARITY: In many situations, it is clear that the mean response is not a linear function of the input level. In such cases, if the form of the relationship can be determined it is sometimes possible, by a change of variables, to transform it into a linear form. For instance, in certain applications it is known that W(t), the amplitude of a signal a time t after its origination, is approximately related to t by the functional form $W(t) \approx ce^{-dt}$ On taking logarithms, this can be expressed as $\log W(t) \approx \log c - dt$ If we now let $Y = \log W(t)$, $\alpha = \log c$, $\beta = -d$ then the foregoing can be 12ormal1212 as a regression of the form $Y = \alpha + \beta t + e$ The regression parameters α and β would then be estimated by the usual least squares approach and the original functional relationships can be predicted from $W(t) \approx e^{A+Bt}$ 9.8 WEIGHTED LEAST SQUARES: In the regression model Y = α + βx + e it often turns out that the variance of a response is not constant but rather depends on its input level. If these variances are known — at least up to a proportionality constant — then the regression parameters α and β should be stimated by minimizing a weighted sum of squares. Specifically, if $\text{Var}(Y_i) = \frac{\sigma^2}{w_i}$ then the estimators A and B should be chosen to minimize $\sum_i \frac{[Y_i-(A+Bx_i)]^2}{\text{Var}(Y_i)} = \frac{1}{\sigma^2}\sum_i w_i(Y_i-A-Bx_i)^2$ On taking partial derivatives with respect to A and B and setting them equal to 0, we obtain the following equations for the minimizing A and B. $\sum_i w_iY_i = A\sum_i w_i + B\sum_i w_ix_i \quad \sum_i w_ix_iY_i = A\sum_i w_ix_i + B\sum_i w_ix_i^2$ These equations are easily solved to yield the least squares estimators. EXAMPLE 9.8a To develop a feel as to why the estimators should be obtained by minimizing the weighted sum of squares rather than the ordinary sum of squares, consider the following situation. Suppose that X1, . . . , Xn are independent normal random variables each having mean μ and variance σ2. Suppose further that the Xi are not directly observable but rather only Y1 and Y2, defined by Y1 = X1 +· · ·+Xk , Y2 = Xk+1 +· · ·+Xn, k < n are directly observable. Based on Y1 and Y2, how should we estimate μ? Whereas the best estimator of μ is clearly $\overline{X} = \sum_{i=1}^n X_i/n = (Y_1+Y_2)/n$, let us see what the ordinary least squares estimator would be. Since E[Y1] = kμ, E[Y2] = (n – k)μ the least squares estimator of μ would be that value of μ that minimizes $(Y_1-k\mu)^2 + (Y_2-[n-k]\mu)^2$ On differentiating and setting equal to zero, we see that the least squares estimator of μ—call it $\hat{\mu}$ —is such that $-2k(Y_1-k\hat{\mu}) - 2(n-k)[Y_2-(n-k)\hat{\mu}] = 0$ or $[k^2+(n-k)^2]\hat{\mu} = kY_1+(n-k)Y_2$ or $\hat{\mu} = \frac{kY_1+(n-k)Y_2}{k^2+(n-k)^2}$ Thus we see that while the ordinary least squares estimator is an unbiased estimator of μ—since $E[\hat{\mu}] = \frac{kE[Y_1]+(n-k)E[Y_2]}{k^2+(n-k)^2} = \frac{k^2\mu+(n-k)^2\mu}{k^2+(n-k)^2} = \mu$ **it is not the best estimator** $\overline{X}$ Now let us determine the estimator produced by minimizing the weighted sum of squares. That is, let us determine the value of μ—call it μ_w — that minimizes $\frac{(Y_1-k\mu)^2}{\text{Var}(Y_1)} + \frac{[Y_2-(n-k)\mu]^2}{\text{Var}(Y_2)}$,since, $\text{Var}(Y_1) = k\sigma^2, \quad \text{Var}(Y_2) = (n-k)\sigma^2$ this is equivalent to choosing μ to minimize $\frac{(Y_1-k\mu)^2}{k} + \frac{[Y_2-(n-k)\mu]^2}{n-k}$ Upon differentiating and then equating to 0, we see that μ_w , the minimizing value, satisfies $\frac{-2k(Y_1-k\mu_w)}{k} - \frac{2(n-k)[Y_2-(n-k)\mu_w]}{n-k} = 0$ or or $\mu_w = \frac{Y_1+Y_2}{n}$ $Y_1+Y_2 = n\mu_w$ That is, the weighted least squares estimator is indeed the **preferred** estimator $(Y_1+Y_2)/n = \overline{X}$. REMARKS: (a) Assuming normally distributed data, the **weighted least squares estimators are precisely the maximum likelihood estimators**. This follows because the joint density of the data Y1, . . . , Yn is

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}(\sigma/\sqrt{w_i})} e^{-(y_i-\alpha-\beta x_i)^2/(2\sigma^2/w_i)}$$
$$= \frac{\sqrt{w_1\ldots w_n}}{(2\pi)^{n/2}\sigma^n} e^{-\sum_{i=1}^n w_i(y_i-\alpha-\beta x_i)^2/2\sigma^2}$$

Consequently, the maximum likelihood estimators of α and β are precisely the values of α and β that minimize the weighted sum of squares $\sum_{i=1}^n w_i(y_i-\alpha-\beta x_i)^2$.(b) The weighted sum of squares can also be seen as the relevant quantity to be minimized by multiplying the regression equation Y = α + βx + e by $\sqrt{w}$ This results in the equation $Y\sqrt{w} = \alpha\sqrt{w} + \beta x\sqrt{w} + e\sqrt{w}$ Now, in this latter equation the error term $e\sqrt{w}$ has mean 0 and constant variance. Hence, the natural least squares estimators of α and β would be the values of A and B that minimize $\sum_i (Y_i\sqrt{w_i} - A\sqrt{w_i} - Bx_i\sqrt{w_i})^2 = \sum_i w_i(Y_i-A-Bx_i)^2$ I The weighted least squares approach puts the greatest emphasis on those data pairs having the greatest weights (and thus the smallest variance in their error term). At this point it might appear that the weighted least squares approach is not particularly useful since it requires a knowledge, up to a constant, of the variance of a response at an arbitrary input level. However, by analyzing the model that generates the data, it is often possible to determine these values. EXAMPLE 9.8c Consider the relationship between Y , the number of accidents on a heavily traveled highway, and x, the number of cars traveling on the highway. After a little thought it would probably seem to most that the linear model Y = α + βx + e would be appropriate. However, as there does not appear to be any a priori reason why Var(Y) should not depend on the input level x, it is not clear that we would be justified in using the ordinary least squares approach to estimate α and β. Indeed, we will now argue that a weighted least squares approach with weights 1/x should be employed—that is, we should choose A and B to minimize $\sum_i \frac{(Y_i-A-Bx_i)^2}{x_i}$ The rationale behind this claim is that it seems reasonable to suppose that Y has approximately a Poisson distribution. This is so since we can imagine that each of the x cars will have a small probability of causing an accident and so, for large x, the number of accidents should be approximately a Poisson random variable. **Since the variance of a Poisson random variable is equal to its mean**, we see that $\text{Var}(Y) \simeq E[Y] \quad \text{since } Y \text{ is approximately Poisson} = \alpha+\beta x \simeq \beta x \quad \text{for large } x$ REMARKS: (a) Another technique that is often employed when the variance of the response depends on the input level is to attempt to stabilize the variance by an appropriate transformation. For example, if Y is a Poisson random variable with mean λ, then it can be shown [seeRemark (b)] that $\sqrt{Y}$ has approximate variance .25 no matter what the value of λ. Based on this fact, one might try to model E[$\sqrt{Y}$] as a linear function of the input. That is, one might consider the model $\sqrt{Y}$ = α + βx + e (b) Proof that Var $\sqrt{Y}$ ≈ .25 when Y is Poisson with mean λ. Consider the Taylor series expansion of g (y) = $\sqrt{Y}$ about the value λ. By ignoring all terms beyond the second derivative term, we obtain that $g(y) \approx g(\lambda) + g'(\lambda)(y-\lambda) + \frac{g''(\lambda)(y-\lambda)^2}{2}$, since, $g'(\lambda) = \frac{1}{2}\lambda^{-1/2}, \quad g''(\lambda) = -\frac{1}{4}\lambda^{-3/2}$ we obtain, on evaluating at y = Y , that $\sqrt{Y} \approx \sqrt{\lambda} + \frac{1}{2}\lambda^{-1/2}(Y-\lambda) - \frac{1}{8}\lambda^{-3/2}(Y-\lambda)^2$ Taking expectations, and using the results that, $E[Y-\lambda] = 0, \quad E[(Y-\lambda)^2] = \text{Var}(Y) = \lambda$ yields that, $E[\sqrt{Y}] \approx \sqrt{\lambda} - \frac{1}{8\sqrt{\lambda}}$ Hence $(E[\sqrt{Y}])^2 \approx \lambda + \frac{1}{64\lambda} - \frac{1}{4} \approx \lambda - \frac{1}{4}$ and so $\text{Var}(\sqrt{Y}) = E[Y] - (E[\sqrt{Y}])^2$ $\approx \lambda - \left(\lambda - \frac{1}{4}\right) = \frac{1}{4}$ 9.9 POLYNOMIAL REGRESSION: In situations where the functional relationship between the response Y and the independent variable x **cannot be adequately approximated** by a linear relationship, it is sometimes possible to obtain a reasonable fit by considering a **polynomial** relationship. That is, we might try to fit to the data set a functional relationship of the form $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + e$ where β0, β1, . . . , βr are regression coefficients that would have to be

estimated. If the data set consists of the n pairs (xi , Yi), I = 1, . . . , n, then the least square estimators of β0, . . . , βr — call them B0, . . . , Br — are those values that minimize $\sum_{i=1}^{n}(Y_i - B_0 - B_1x_i - B_2x_i^2 - \cdots - B_rx_i^r)^2$ To determine these estimators, we take partial derivatives with respect to B0 . . . Br of the foregoing sum of squares, and then set these equal to 0 so as to determine the minimizing values. On doing so, and then rearranging the resulting equations, we obtain that the least square estimators B0, B1, . . . , Br satisfy the following set of r + 1 linear equations called the **normal equations**. $\sum_{i=1}^{n} Y_i = B_0 n + B_1\sum_{i=1}^{n} x_i + B_2\sum_{i=1}^{n} x_i^2 + \cdots + B_r\sum_{i=1}^{n} x_i^r$

$\sum_{i=1}^{n} x_iY_i = B_0\sum_{i=1}^{n} x_i + B_1\sum_{i=1}^{n} x_i^2 + B_2\sum_{i=1}^{n} x_i^3 + \cdots + B_r\sum_{i=1}^{n} x_i^{r+1}$ $\sum_{i=1}^{n} x_i^2Y_i = B_0\sum_{i=1}^{n} x_i^2 + B_1\sum_{i=1}^{n} x_i^3 + \cdots + B_r\sum_{i=1}^{n} x_i^{r+2}$ ..... $\sum_{i=1}^{n} x_i^rY_i = B_0\sum_{i=1}^{n} x_i^r + B_1\sum_{i=1}^{n} x_i^{r+1} + \cdots + B_r\sum_{i=1}^{n} x_i^{2r}$ In fitting a polynomial to a set of data pairs, it is often possible to determine the necessary degree of the polynomial by a study of the scatter diagram. We emphasize that one should always use the lowest possible degree that appears to adequately describe the data. [Thus, for instance, whereas it is usually possible to find a polynomial of degree n that passes through all the n pairs (xi , Yi), I = 1, . . . , n, it would be hard to ascribe much confidence to such a fit.] Even more so than in linear regression, it is extremely risky to use a polynomial fit to predict the value of a response at an input level x0 that is far away from the input levels xi , I = 1, . . . , n used in finding the polynomial fit. (For one thing, the polynomial fit may be valid only in a region around the xi , I = 1, . . . , n and not including x0.) 9.10 MULTIPLE LINEAR REGRESSION: In the majority of applications, the response of an experiment can be predicted more adequately not on the basis of a single independent input variable but on a collection of such variables. Indeed, a typical situation is one in which there are a set of, say, k input variables and the response Y is related to them by the relation $Y = \beta_0 + \beta_1x_1 + \cdots + \beta_kx_k + e$ where xj , j = 1, . . . , k is the level of the jth input variable and e is a random error that we shall assume is normally distributed with mean 0 and (constant) variance σ^2. The parameters β0, β1, . . . , βk and σ^2 are assumed to be unknown and must be estimated from the data, which we shall suppose will consist of the values of Y1, . . . , Yn where Yi is the response level corresponding to the k input levels xi1, . . . , xi2, . . . , xik . That is, the Yi are related to these input levels through $E[Y_i] = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \cdots + \beta_kx_{ik}$ If we let B0, B1, . . . , Bk denote estimators of β0, . . . , βk , then the sum of the squared differences between the Yi and their estimated expected values is $\sum_{i=1}^{n}(Y_i - B_0 - B_1x_{i1} - B_2x_{i2} - \cdots - B_kx_{ik})^2$ The least squares estimators are those values of B0, B1, . . . , Bk that minimize the foregoing. To determine the least squares estimators, we repeatedly take partial derivatives of the preceding sum of squares first with respect to B0, then to B1, . . . , then to Bk . On equating these k + 1 equations to 0, we obtain the following set of equations: $\sum_{i=1}^{n}(Y_i - B_0 - B_1x_{i1} - B_2x_{i2} - \cdots - B_kx_{ik}) = 0$ $\sum_{i=1}^{n} x_{i1}(Y_i - B_0 - B_1x_{i1} - \cdots - B_kx_{ik}) = 0$

$\sum_{i=1}^{n} x_{i2}(Y_i - B_0 - B_1x_{i1} - \cdots - B_kx_{ik}) = 0$ ..... $\sum_{i=1}^{n} x_{ik}(Y_i - B_0 - B_1x_{i1} - \cdots - B_ix_{ik}) = 0$ **Rewriting** these equations yields that the least squares estimators B0, B1, . . . , Bk satisfy the following set of linear equations, called the **normal** equations: $\sum_{i=1}^{n} Y_i = nB_0 + B_1\sum_{i=1}^{n} x_{i1} + B_2\sum_{i=1}^{n} x_{i2} + \cdots + B_k\sum_{i=1}^{n} x_{ik}$

$\sum_{i=1}^{n} x_{i1}Y_i = B_0\sum_{i=1}^{n} x_{i1} + B_1\sum_{i=1}^{n} x_{i1}^2 + B_2\sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + B_k\sum_{i=1}^{n} x_{i1}x_{ik}$ ..... $\sum_{i=1}^{k} x_{ik}Y_i = B_0\sum_{i=1}^{n} x_{ik} + B_1\sum_{i=1}^{n} x_{ik}x_{i1} + B_2\sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + B_k\sum_{i=1}^{n} x_{ik}^2$ Before solving the normal equations, it is convenient to introduce matrix notation. If we let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

then Y is an n × 1, X an n × p, β a p × 1, and e an n × 1 matrix where p ≡ k + 1. The multiple regression model can now be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ In addition, if we let $\mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix}$ be the matrix of least squares estimators, then the normal Equations can be written as $\mathbf{X'XB} = \mathbf{X'Y}$ where $\mathbf{X'}$ is the transpose of $\mathbf{X}$. To see this Equation is equivalent to the normal Equations, note that

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_i x_{ik} & \sum_i x_{ik}x_{i1} & \sum_i x_{ik}x_{i2} & \cdots & \sum_i x_{ik}^2 \end{bmatrix}$$

and $\mathbf{X'Y} = \begin{bmatrix} \sum_i Y_i \\ \sum_i x_{i1}Y_i \\ \vdots \\ \sum_i x_{ik}Y_i \end{bmatrix}$ It is now easy to see that the matrix equation $\mathbf{X'XB} = \mathbf{X'Y}$ is equivalent to the set of normal Equations 9.10.1. Assuming that $(\mathbf{X'X})^{-1}$ exists, which is usually the case, we obtain, upon multiplying it by both sides of the foregoing, that the least squares estimators are given by $\mathbf{B} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ Program 9.10 computes the least squares estimates, the inverse matrix $(\mathbf{X'X})^{-1}$, and SS_R. It follows from Equation $\mathbf{B} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ that the least squares estimators B0, B1, . . . , Bk —the elements of the matrix B — are all linear combinations of the independent normal random variables Y1, . . . , Yn and so will also be normally distributed. Indeed in such a situation — namely, when each member of a set of random variables can be expressed as a linear combination of independent normal random variables—we say that the set of random variables has **a joint multivariate normal distribution**. The least squares estimators turn out to be unbiased. This can be shown as follows: $E[\mathbf{B}] = E[(\mathbf{X'X})^{-1}\mathbf{X'Y}] = E[(\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e})]$ since $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

$= E[(\mathbf{X'X})^{-1}\mathbf{X'X}\boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'e}] = E[\boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'e}] = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'}E[\mathbf{e}] = \boldsymbol{\beta}$ .The variances of the least squares estimators can be obtained from the matrix $(\mathbf{X'X})^{-1}$. Indeed, the values of this matrix are related to the covariances of the B_i's. Specifically, the element in the (I + 1)st row, (j + 1)st column of $(\mathbf{X'X})^{-1}$ is equal to $\mathrm{Cov}(B_i, B_j)/\sigma^2$. To verify the preceding statement concerning Cov(Bi , Bj), let $\mathbf{C} = (\mathbf{X'X})^{-1}\mathbf{X'}$ Since X is an n×p matrix and $\mathbf{X'}$ a p×n matrix, it follows that $\mathbf{X'X}$ is p×p, as is $(\mathbf{X'X})^{-1}$, and so C will be a p × n matrix. Let Cij denote the element in row I, column j of this matrix. Now

$$\begin{bmatrix} B_0 \\ \vdots \\ B_{i-1} \\ \vdots \\ B_k \end{bmatrix} = \mathbf{B} = \mathbf{CY} = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & & \vdots \\ C_{i1} & \cdots & C_{in} \\ \vdots & & \vdots \\ C_{p1} & \cdots & C_{pn} \end{bmatrix} \begin{bmatrix} Y_1 \\ \\ \vdots \\ Y_n \end{bmatrix}$$

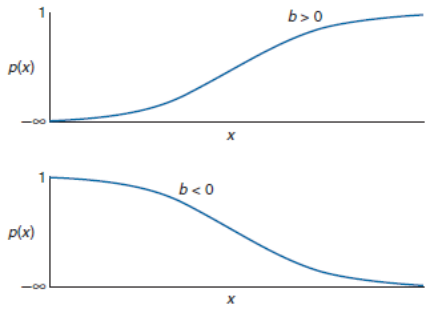
and so,

$B_{i-1} = \sum_{l=1}^{n} C_{il}Y_l \quad B_{j-1} = \sum_{r=1}^{n} C_{jr}Y_r$ Hence $\mathrm{Cov}(B_{i-1}, B_{j-1}) = \mathrm{Cov}\left(\sum_{l=1}^{n} C_{il}Y_l, \sum_{r=1}^{n} C_{jr}Y_r\right) = \sum_{r=1}^{n}\sum_{l=1}^{n} C_{il}C_{jr}\,\mathrm{Cov}(Y_l, Y_r)$ Now Yl and Yr are independent when $l \neq r$, and so

$$\mathrm{Cov}(Y_l, Y_r) = \begin{cases} 0 & \text{if } l \neq r \\ \mathrm{Var}(Y_r) & \text{if } l = r \end{cases}$$ Since $\mathrm{Var}(Y_r) = \sigma^2$, we see that $\mathrm{Cov}(B_{i-1}, B_{j-1}) = \sigma^2\sum_{r=1}^{n} C_{ir}C_{jr} = \sigma^2(\mathbf{CC'})_{ij}$ where $(\mathbf{CC'})_{ij}$ is the element in row i, column j of $\mathbf{CC'}$. If

we now let Cov(B) denote the matrix of covariances —that is, $\mathrm{Cov}(\mathbf{B}) = \begin{bmatrix} \mathrm{Cov}(B_0, B_0) & \cdots & \mathrm{Cov}(B_0, B_k) \\ \vdots & & \vdots \\ \mathrm{Cov}(B_k, B_0) & \cdots & \mathrm{Cov}(B_k, B_k) \end{bmatrix}$ then it follows from Equation $\mathrm{Cov}(B_{i-1}, B_{j-1}) = \sigma^2 \sum_{r=1}^{n} C_{ir} C_{jr} = \sigma^2 (\mathbf{CC}')_{ij}$ that $\mathrm{Cov}(\mathbf{B}) = \sigma^2 \mathbf{CC}'$ Now, $\mathbf{C}' = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' = \mathbf{X}\left((\mathbf{X}'\mathbf{X})^{-1}\right)' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ where the last equality follows since $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric (since $\mathbf{X}'\mathbf{X}$ is) and so is equal to its transpose. Hence $\mathbf{CC}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$ and so we can conclude from Equation $\mathrm{Cov}(\mathbf{B}) = \sigma^2 \mathbf{CC}'$ that $\mathrm{Cov}(\mathbf{B}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ Since Cov(Bi , Bi) = Var(Bi), it follows that the variances of the least squares estimators are given by σ^2 multiplied by the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$. The quantity σ^2 can be estimated by using the sum of squares of the **residuals**. That is, if we let $SS_R = \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2$ then it can be shown that $\frac{SS_r}{\sigma^2} \sim \chi^2_{n-(k+1)}$ and so $E\left[\frac{SS_R}{\sigma^2}\right] = n - k - 1$ or $E[SS_R/(n-k-1)] = \sigma^2$ That is, SS_R /(n – k – 1) is an unbiased estimator of σ2. In addition, as in the case of simple linear regression, SS_R will be independent of the least squares estimators B0, B1, . . . , Bk . REMARK: If we let ri denote the ith $r_i = Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}$, I = 1, . . . , n then r = Y – XB where $\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}$ Hence, we may write $SS_R = \sum_{i=1}^{n} r_i^2 = \mathbf{r}'\mathbf{r} = (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})$ $= [\mathbf{Y}' - (\mathbf{XB})'](\mathbf{Y} - \mathbf{XB}) = (\mathbf{Y}' - \mathbf{B}'\mathbf{X}')(\mathbf{Y} - \mathbf{XB}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{XB} - \mathbf{B}'\mathbf{X}'\mathbf{Y} + \mathbf{B}'\mathbf{X}'\mathbf{XB} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{XB}$ where the last equality follows from the **normal** equations $\mathbf{X}'\mathbf{XB} = \mathbf{X}'\mathbf{Y}$ Because $\mathbf{Y}'$ is $1 \times n$, $\mathbf{X}$ is $n \times p$, and $\mathbf{B}$ is $p \times 1$, it follows that $\mathbf{Y}'\mathbf{XB}$ is a 1×1 matrix. That is, $\mathbf{Y}'\mathbf{XB}$ is a scalar and thus is equal to its transpose, which shows that $\mathbf{Y}'\mathbf{XB} = (\mathbf{Y}'\mathbf{XB})' = \mathbf{B}'\mathbf{X}'\mathbf{Y}$ Hence, using Equation $SS_R = \sum_{i=1}^{n} r_i^2$ we have proven the following identity: $SS_R = \mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y}$ The foregoing is a **useful** computational formula for SSR (though one must be careful of possible roundoff error when using it). REMARK: The quantity $R^2 = 1 - \frac{SS_R}{\sum_i (Y_i - \overline{Y})^2}$ which measures the amount of reduction in the sum of squares of the residuals when using the model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + e$ **as opposed to the model** $Y = \beta_0 + e$ **is called the coefficient of multiple determination**. 9.10.1 Predicting Future Responses: Let us now suppose that a series of experiments is to be performed using the input levels x1, . . . , xk . Based on our data, consisting of the prior responses Y1, . . . , Yn, suppose we would like to estimate the mean response. Since the mean response is $E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ a point estimate of it is simply $\sum_{i=0}^{k} B_i x_i$ where $x_0 \equiv 1$. To determine a confidence interval estimator, we need the distribution of $\sum_{i=0}^{k} B_i x_i$. Because it can be expressed as a linear combination of the independent normal random variables Yi , I = 1, . . . , n, it follows that it is also normally distributed. Its mean and variance are obtained as follows: $E\left[\sum_{i=0}^{k} x_i B_i\right] = \sum_{i=0}^{k} x_i E[B_i] = \sum_{i=0}^{k} x_i \beta_i$ since $E[B_i] = \beta_i$ That is, it is an unbiased estimator. Also, using the fact that the variance of a random variable is equal to the covariance between that random variable and itself, we see that equation_x1 $\mathrm{Var}\left(\sum_{i=0}^{k} x_i B_i\right) = \mathrm{Cov}\left(\sum_{i=0}^{k} x_i B_i, \sum_{j=0}^{k} x_j B_j\right)$ $= \sum_{i=0}^{k} \sum_{j=0}^{k} x_i x_j \mathrm{Cov}(B_i, B_j)$ If we let x denote the matrix $\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_k \end{bmatrix}$ then, recalling that $\mathrm{Cov}(B_i, B_j)/\sigma^2$ is the element in the (i+1)st row and (j+1)st column of $(\mathbf{X}'\mathbf{X})^{-1}$ we can express equation_x1 as $\mathrm{Var}\left(\sum_{i=0}^{k} x_i B_i\right) = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\sigma^2$, $\frac{\sum_{i=0}^{k} x_i B_i - \sum_{i=0}^{k} x_i \beta_i}{\sigma\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim N(0,1)$ If we now replace σ by its estimator $\sqrt{SS_R/(n-k-1)}$ we obtain, by the usual argument, that $\frac{\sum_{i=0}^{k} x_i B_i - \sum_{i=0}^{k} x_i \beta_i}{\sqrt{\frac{SS_R}{(n-k-1)}}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim t_{n-k-1}$ which gives rise to the following confidence interval estimator of $\sum_{i=0}^{k} x_i \beta_i$. Confidence Interval Estimate of E [Y|x] = $\sum_{i=0}^{k} x_i \beta_i$, $(x_0 \equiv 1)$ A 100(1 – a) percent confidence interval estimate of $\sum_{i=0}^{k} x_i \beta_i$ is given by $\sum_{i=0}^{k} x_i b_i \pm \sqrt{\frac{ss_r}{(n-k-1)}}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}\ t_{a/2,n-k-1}$ where b0, . . . , bk are the values of the least squares estimators B0, B1, . . . , Bk , and ssr is the value of SS_R. **When it is only a single experiment** that is going to be performed at the input levels x1, . . . , xk , we are usually more concerned with predicting the actual response than its mean value. That is, we are interested in utilizing our data set Y1, . . . , Yn to predict $Y(\mathbf{x}) = \sum_{i=0}^{k} \beta_i x_i + e, \quad \text{where } x_0 = 1$ A point prediction is given by $\sum_{i=0}^{k} B_i x_i$ where Bi is the least squares estimator of βi based on the set of prior responses Y1, . . . , Yn, I = 1, . . . , k. To determine a prediction interval for Y (x), note first that since B0, . . . , Bk are based on prior responses, it follows that they are independent of Y (x). Hence, it follows that $Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i$ is normal with mean 0 and variance given by $\mathrm{Var}\left[Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i\right] = \mathrm{Var}[Y(\mathbf{x})] + \mathrm{Var}\left(\sum_{i=0}^{k} B_i x_i\right)$ by independence $= \sigma^2 + \sigma^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$ and so, $\frac{Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i}{\sigma\sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim N(0,1)$ which yields, upon replacing σ by its estimator, that $\frac{Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i}{\sqrt{\frac{SS_R}{(n-k-1)}}\sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim t_{n-k-1}$ We thus have: Prediction Interval for Y(x): With 100(1 – a) percent confidence Y (x) will lie between $\sum_{i=0}^{k} x_i b_i \pm \sqrt{\frac{ss_r}{(n-k-1)}}\sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}\ t_{a/2,n-k-1}$ where b0, . . . , bk are the values of the least squares estimators B0, B1, . . . , Bk , and ssr is

the value of SS_R. 9.11 LOGISTIC REGRESSION MODELS FOR BINARY OUTPUT DATA: In this section we consider experiments that result in either a success or a failure. We will suppose that these experiments can be performed at various levels, and that an experiment performed at level x will result in a success with probability p(x), $-\infty < x < \infty$. If p(x) is of the **form** $p(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ then the experiments are said to come from a logistic regression model and p(x) is called the logistics regression function. If b > 0, then p(x) = $1/[e^{-(a+bx)}+1]$ is an increasing function that converges to 1 as x →∞; if b < 0, then p(x) is a decreasing function that converges to 0 as x → ∞. (When b = 0, p(x) is constant.) Plots of logistics regression functions are given in Figure Notice the s-shape of these curves.



Logistic regression functions. Writing p(x) = 1 – $[1/(1+e^{a+bx})]$ and differentiating gives that

$$\frac{\partial}{\partial x}p(x) = \frac{be^{a+bx}}{(1+e^{a+bx})^2} = bp(x)[1-p(x)]$$

Thus the rate of change of p(x) depends on x and is largest at those values of x for which p(x) is near .5. For instance, at the value x such that p(x) = .5, the rate of change is $\frac{\partial}{\partial x}p(x) = .25b$, whereas at that value x for which p(x) = .8 the rate of change is .16b. If we let o(x) be the odds for success

$$o(x) = \frac{p(x)}{1-p(x)} = e^{a+bx}$$

when the experiment is run at level x, then Thus, when b > 0, the odds increase exponentially in the input level x; when b < 0, the odds decrease exponentially in the input level x. Taking logs of the preceding shows the the log odds, called the **logit**, is a linear function: $\log[o(x)] = a + bx$ The parameters a and b of the logistic regression function are assumed to be unknown and need to be estimated. This can be accomplished by using the maximum likelihood approach. That is, suppose that the experiment is to be performed at levels x1, . . . , xk . Let Yi be the result (either 1 if a success, or 0 if a failure) of the experiment when performed at level xi . Then, using the Bernoulli density function (that is, the binomial density for a single trial), gives

$$P\{Y_i = y_i\} = [p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i} = \left(\frac{e^{a+bx}}{1+e^{a+bx}}\right)^{y_i}\left(\frac{1}{1+e^{a+bx}}\right)^{1-y_i}, \quad y_i = 0, 1$$

Thus, the probability that the experiment at level xi results in outcome yi , for all I = 1, . . . , k, is

$$P\{Y_i = y_i, i = 1, \ldots, k\} = \prod_i \left(\frac{e^{a+bx_i}}{1+e^{a+bx_i}}\right)^{y_i}\left(\frac{1}{1+e^{a+bx_i}}\right)^{1-y_i}$$
$$= \prod_i \frac{\left(e^{a+bx_i}\right)^{y_i}}{1+e^{a+bx_i}}$$

Taking logarithms gives that

$$\log\left(P\{Y_i = y_i, i = 1, \ldots, k\}\right) = \sum_{i=1}^{k} y_i(a+bx_i) - \sum_{i=1}^{k}\log\left(1+e^{a+bx_i}\right)$$

The maximum likelihood estimates can now be obtained by numerically finding the values of a and b that maximize the preceding likelihood. However, because the likelihood is **nonlinear** this requires an iterative approach; consequently, one typically resorts to specialized software to obtain the estimates. Whereas the logistic regression model is the most frequently used model when the response data are **binary**, other models are often employed. For instance in situations where it is reasonable to suppose that p(x), the probability of a **positive response** when the input level is x, is an **increasing** function of x, it is often supposed that p(x) has the form of a **specified** probability distribution function. Indeed, when b > 0, the **logistic regression model** is of this form because p(x) is equal to the distribution function of a **logistic random variable** (Section 5.9) with parameters μ = −a/b, ν = 1/b. Another model of this type is the **probit** model, which supposes that for some constants, α,β > 0

$$p(x) = \Phi(\alpha+\beta x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\alpha+\beta x} e^{-y^2/2}\,dy$$

In other words p(x) is equal to the probability that a standard normal random variable is less than α + βx. EXAMPLE 9.11a A common assumption for whether an animal becomes sick when exposed to a chemical at dosage level x is to assume a threshold model, which supposes that each animal has a random threshold and will become ill if the dosage level exceeds that threshold. The exponential distribution has sometimes been used as the threshold distribution. For instance, a model considered in Freedman and Zeisel ("From Mouse to Man: The Quantitative Assessment of Cancer Risks," Statistical Science, 1988, 3, 1, 3–56) supposes that a mouse exposed to x units of DDT (measured in ppm) will contract cancer of the liver with probability $p(x) = 1 - e^{-ax}, \quad x > 0$ Because of the **lack of memory of the exponential distribution**, this is **equivalent to assuming** that if the mouse who is still healthy **after** receiving a (partial) dosage of level x is as good as it was **before** receiving any dosage. It was reported in Freedman and Zeisel that 84 of 111 mice exposed to DDT at a level of 250 ppm developed cancer. Therefore, α can be estimated from $1 - e^{-250\hat{\alpha}} = \frac{84}{111}$ or $\hat{\alpha} = -\frac{\log(27/111)}{250} = .005655$

Chapter_6: DISTRIBUTIONS OF SAMPLING STATISTICS: 6.1 INTRODUCTION: The science of statistics deals with drawing conclusions from observed data. For instance, a typical situation in a technological study arises when one is confronted with a large collection, or population, of items that have measurable values associated with them. By suitably sampling from this collection, and then 15ormal1515g the sampled items, one hopes to be able to draw some conclusions about the collection as a whole. To use sample data to make inferences about an entire population, it is necessary to make some assumptions about the relationship between the two. One such assumption,which is often quite reasonable, is that there is an underlying (population) probability distribution such that the measurable values of the items in the population can be thought of as being independent random variables having this distribution. If the sample data are then chosen in a random fashion, then it is reasonable to suppose that they too are independent values from the distribution. Definition: If X1, . . . , Xn are independent random variables having a common distribution F, then we say that they constitute a sample (sometimes called a random sample) from the distribution F. In most applications, the population distribution F will not be completely specified and one will attempt to use the data to make inferences about F. Sometimes it will be supposed that F is specified up to some unknown parameters (for instance, one might suppose that F was a normal distribution function having an unknown mean and variance, or that it is a Poisson distribution function whose mean is not given), and at other times it might be assumed that almost nothing is known about F (except maybe for assuming that it is a continuous, or a discrete, distribution). Problems in which the form of the underlying distribution is specified up to a set of unknown parameters are called parametric inference problems, whereas those in which nothing is assumed about the form of F are called nonparametric inference problems. Physical reasons sometimes suggest the parametric form of the distribution F; for instance, it may lead us to believe that F is a normal distribution, or that F is an exponential distribution. In such cases, we are confronted with a parametrical statistical problem in which we would want to use the observed data to estimate the parameters of F. For instance, if F were assumed to be a normal distribution, then we would want to estimate its mean and variance; if F were assumed to be exponential, we would want to estimate its mean. In other situations, there might not be any physical justification for supposing that F has any particular form; in this case the problem of making inferences about F would constitute a nonparametric inference problem. In this chapter, we will be concerned with the probability distributions of certain statistics that arise from a sample, where a statistic is a random variable whose value is determined by the sample data. Two important statistics that we will discuss are the sample mean and the sample variance. 6.2 THE SAMPLE MEAN: Consider a population of elements, each of which has a numerical value attached to it. For instance, the population might consist of the adults of a specified community and the value attached to each adult might be his or her annual income, or height, or age, and so on. We often suppose that the value associated with any member of the population can be regarded as being the value of a random variable having expectation μ and variance . The quantities μ and σ^2 are called the population mean and the population variance, respectively. Let X1, X2, . . . , Xn be a sample of values from this population. The sample mean is defined

$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ by Since the value of the sample mean X is determined by the values of the random variables in the sample, it follows that X is also a random variable. Its expected value and variance

$$E[\overline{X}] = E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}(E[X_1] + \cdots + E[X_n]) = \mu \quad \text{and} \quad \text{Var}(\overline{X}) = \text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right)$$

are obtained as follows:

$= \frac{1}{n^2}[\text{Var}(X_1) + \cdots + \text{Var}(X_n)]$ by independence $= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$ where μ and σ^2 are the population mean and variance, respectively. Hence, the expected value of the sample mean is the population mean μ whereas its variance is 1/n times the population variance. As a result, we can conclude that X is also centered about the population mean μ, but its spread becomes more and more reduced as the sample size increases. Figure 6.1 plots the probability density function of the sample mean from a standard normal population for a variety of sample sizes.

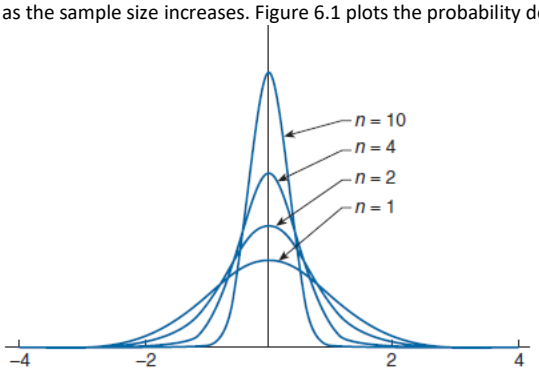


FIGURE 6.1 Densities of sample means from a standard normal population. 6.3 THE CENTRAL LIMIT THEOREM: In this section, we will consider one of the most remarkable results in probability — namely, the central limit theorem. Loosely speaking, this theorem asserts that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but it also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit a bell-shaped (that is, a normal) curve. In its simplest form, the central limit theorem is as follows: Theorem 6.3.1 The Central Limit Theorem: Let X1, X2, . . . , Xn be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of X1 +· · ·+Xn, is approximately normal with mean nμ and variance nσ^2. It follows from the central limit theorem that $\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$ is approximately a standard normal random variable; thus, for n large,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} \approx P\{Z < x\}$$

where Z is a standard normal random variable. The central limit theorem is illustrated by Program 6.1 on the text disk. This program plots the probability mass function of the sum of n independent and identically distributed random variables that each take on one of the values 0, 1, 2, 3, 4. The central limit theorem is illustrated by Program 6.1 on the text disk. This program plots the probability mass function of the sum of n independent and identically distributed random variables that each take on one of the values 0, 1, 2, 3, 4. One of the **most important applications of the central limit theorem** is in regard **to binomial random variables**. Since such a random variable X having parameters (n, p) represents the number of successes in n independent trials when each trial is a success with probability p, **we can express** it as X = X1 +· · ·+Xn where Xi = $\begin{cases} 1 & \text{if the } i\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$ Because E[Xi] = p, Var(Xi) = p(1 – p), it follows from the central limit theorem that for n large $\frac{X - np}{\sqrt{np(1-p)}}$ will approximately be a standard normal random variable [the probability mass function of a binomial (n, p) random variable becomes more and more "normal" as n becomes larger and larger]. It should be noted that we now **have two possible approximations to binomial probabilities**: The Poisson approximation, which yields a good approximation **when n is large and p small**, and the **normal approximation**, which can be shown to be quite good **when np(1 – p) is large**. [The normal approximation will, in general, be quite good for values **of n satisfying np(1 – p) ≥ 10.**] 6.3.1 Approximate Distribution of the Sample Mean: Let X1, . . . , Xn be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ Since a **constant multiple** of a normal random variable is **also** normal, it follows from the central limit theorem that X will be approximately normal when the sample size n is large. Since the sample mean has expected value μ and standard deviation $\sigma/\sqrt{n}$, it then follows that $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ has approximately a standard normal distribution. 6.3.2 **How Large** a Sample Is Needed? The central limit theorem leaves open the question of how large the sample size n needs to be for the normal approximation to be valid, and indeed the answer depends on the population distribution of the sample data. For instance, if the underlying population distribution is normal, then the sample mean X will also be normal regardless of the sample size. **A general rule of thumb** is that one can be confident of the normal approximation whenever the sample size n is **at least 30**. That is, practically speaking, no matter how nonnormal the underlying population distribution is, the sample mean of a sample of size at least 30 will be approximately normal. In most cases, the normal approximation is valid for much smaller sample sizes. Indeed, a sample of size 5 will often suffice for the approximation to be valid. Figure 6.4 presents the distribution of the sample means from an exponential population distribution for samples of sizes n = 1, 5, 10. 6.4 THE SAMPLE VARIANCE: Let X1, . . . , Xn be a random sample from a distribution with mean μ and variance σ^2. Let X be the sample mean, and recall the following definition from Section 2.3.2. Definition: The statistic S^2, defined by $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ is called the sample variance. $S = \sqrt{S^2}$ is called the sample standard deviation. To compute $E[S^2]$, we use an identity that was proven in Section 2.3.2: For any numbers x1, . . . , xn

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

where $\overline{x} = \sum_{i=1}^{n} x_i/n$ It follows from this identity that $(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$ Taking expectations of both sides of the preceding yields, upon using the fact that for any random variable $W, E[W^2] = \text{Var}(W) + (E[W])^2$

$$(n-1)E[S^2] = E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2]$$

$= nE[X_1^2] - nE[\overline{X}^2] = n\text{Var}(X_1) + n(E[X_1])^2 - n\text{Var}(\overline{X}) - n(E[\overline{X}])^2 = n\sigma^2 + n\mu^2 - n(\sigma^2/n) - n\mu^2 = (n-1)\sigma^2$ or $E[S^2] = \sigma^2$ That is, the **expected** value of the **sample variance** S^2 is **equal** to the **population variance** σ^2. 6.5 SAMPLING DISTRIBUTIONS FROM A NORMAL POPULATION : Let X1, X2, . . . , Xn be a

sample from a normal population having mean μ and variance σ^2. That is, they are independent **and** Xi **~ N**(μ, σ^2), I = 1, . . . , n. Also let $\overline{X} = \sum_{i=1}^{n} X_i/n$ and $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ denote the sample mean and sample variance, respectively. We would like to compute their distributions. 6.5.1 Distribution of the Sample Mean: Since the sum of independent normal random variables is normally distributed, it follows that X is normal with mean $E[\overline{X}] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \mu$ and variance $\text{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) = \sigma^2/n$ That is, X , the average of the sample, is normal with a mean equal to the population mean but with a variance **reduced by a factor of 1/n**. It follows from this that $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is a standard normal random variable. 6.5.2 **Joint Distribution** of $\overline{X}$ and S^2 : In this section, we not only obtain the distribution of the sample variance S2, but we also discover a **fundamental fact** about normal samples—namely, that X and S^2 **are independent** with (n – 1) S^2/σ^2 having a chi-square distribution with n – 1 degrees of freedom. To start, for numbers x1, . . . , xn, let yi = xi – μ, I = 1, . . . , n. Then as y = x – μ, it follows from the identity $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2$ Now, if X1, . . . , Xn is a sample from a normal population having mean μ variance σ^2, then we obtain from the preceding identity that $\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}$ or, equivalently, $\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} + \left[\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right]^2$ Because (Xi – μ)/σ, I = 1, . . . , n are independent standard 17ormal, it follows that the left side of Equation is a chi-square random variable with n degrees of freedom. Also, as shown earlier $\sqrt{n}(\overline{X} - \mu)/\sigma$ is a standard normal random variable and so its square is a chi-square random variable with 1 degree of freedom. Thus Equation, equates a chi-square random variable having **n degrees** of freedom to the **[sum of two random variables], one** of which is chi-square with **1 degree** of freedom. But it has been established that the sum of two independent chi-square random variables **is also** chi-square with a degree of freedom equal to the sum of the two degrees of freedom. Thus, it would seem that there is a **reasonable possibility** that the two terms on the right side of Equation are **independent**, with $\sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2$ having a chi-square distribution with **n – 1 degrees** of freedom. Since this result can indeed be established, we have the following fundamental result. [i.e. n=[n-1] sum [1]] Theorem 6.5.1: If X1, . . . , Xn is a sample from a normal population having mean μ and variance σ^2, then $\overline{X}$ and S^2 are independent random variables, with X being normal with mean μ and variance σ2/n and (n – 1) S^2/σ^2 being chi-square with n – 1 degrees of freedom. Theorem 6.5.1 not only provides the distributions of $\overline{X}$ and S^2 for a normal population but also establishes the important fact that they are independent. In fact, it turns out that this independence of $\overline{X}$ and is a unique property of the normal distribution. Its importance will become evident in the following chapters. Corollary 6.5.2 Let Xi , . . . , Xn be a sample from a normal population with mean μ. If $\overline{X}$ denotes the sample mean and S the sample standard deviation, then $\sqrt{n}\frac{(\overline{X} - \mu)}{S} \sim t_{n-1}$ That is, $\sqrt{n}(\overline{X} - \mu)/S$ has a t-distribution with n – 1 degrees of freedom. Proof: Recall that a t-random variable with n degrees of freedom is defined as the distribution of $\frac{Z}{\sqrt{\chi_n^2/n}}$ where Z is a standard normal random variable that is independent of χ_n^2 , a chi-square random variable with n degrees of freedom. It then follows from Theorem 6.5.1 $\frac{\sqrt{n}(\overline{X} - \mu)/\sigma}{\sqrt{S^2/\sigma^2}} = \sqrt{n}\frac{(\overline{X} - \mu)}{S}$ -is a t-random variable with n-1 degrees of freedom. 6.6 SAMPLING FROM A **FINITE POPULATION**: Consider a population of N elements, and suppose that p is the proportion of the population that has a certain characteristic of interest; that is, **Np elements have this characteristic, and N(1– p) do not**. A sample of size n from this population is said to be a random sample if it is chosen in such a manner that each of the $\binom{N}{n}$ population **subsets** of size n is equally likely to be the sample. For instance, if the population consists of the three elements a, b, c, then a random sample of size 2 is one that is chosen so that each of the subsets {a, b}, {a, c}, and {b, c} is equally likely to be the sample. A random subset can be chosen sequentially by letting its first element be equally likely to be any of the N elements of the population, then letting its second element be equally likely to be any of the remaining N – 1 elements of the population, and so on. Suppose now that a random sample of size n has been chosen from a population of size N. For I = 1, . . . , n, let $X_i = \begin{cases} 1 & \text{if the } i\text{th member of the sample has the characteristic} \\ 0 & \text{otherwise} \end{cases}$ Consider now the sum of the Xi ; that is, consider X = X1 + X2 +· · ·+Xn **Because** the term Xi **contributes 1** to the sum if the ith member of the sample has the **characteristic** and 0 **otherwise**, it follows that X is equal to the number of members of the sample that possess the characteristic. In addition, the sample mean $\overline{X} = X/n = \sum_{i=1}^{n} X_i/n$ is equal to the **proportion** of the members of the sample that **possess** the **characteristic**. Let us now consider the **probabilities** associated with the statistics X **and** $\overline{X}$. To begin, note that since each of the N members of the population is equally likely to be the ith member of the sample, it follows that $P\{X_i = 1\} = \frac{Np}{N} = p$ Also, P{Xi = 0} = 1 – P{Xi = 1} = 1 – p That is, each Xi is equal to either 1 or 0 with respective probabilities p and 1 – p. It should be noted that the random variables X1, X2, . . . , Xn are **not independent**. For instance, since the second selection is equally likely to be any of the N members of the population, of which Np have the characteristic, it follows that the probability that the second selection has the characteristic is Np/N = p. That is, without any knowledge of the outcome of the first selection, P{X2 = 1} = p However, the **conditional probability** that X2 = 1, given that the **first selection has the characteristic**, is $P\{X_2 = 1|X_1 = 1\} = \frac{Np - 1}{N - 1}$ which is seen by noting that if the first selection has the characteristic, then the second selection is equally likely to be any of the remaining N −1 elements, of which Np−1 have the characteristic. Similarly, the probability that the **second selection** has the characteristic given that the **first one does not** is $P\{X_2 = 1|X_1 = 0\} = \frac{Np}{N - 1}$ Thus, knowing whether or not the first element of the random sample has the characteristic changes the probability for the next element. However, when the population size N Thus, knowing whether or not the first element of the random sample has the characteristic changes the probability for the next element. However, when the population size N $P\{X_2 = 1|X_1 = 1\} = \frac{399}{999} = .3994$ which is **very close to** the unconditional probability that X2 = 1; namely, P{X2 = 1} = .4 Similarly, the probability that the second element of the sample has the characteristic given that the first does not is $P\{X_2 = 1|X_1 = 0\} = \frac{400}{999} = .4004$ **which is again very close to .4.**

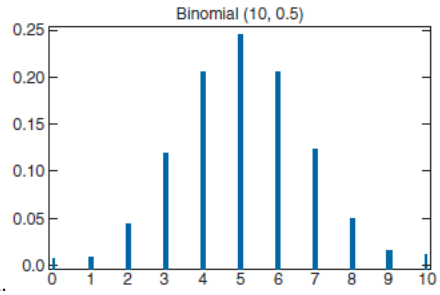
Indeed, it can be shown that when the population size N is large with respect to the sample size n, then X1, X2, . . . , Xn are [[**approximately independent]]**. Now if we think of each Xi as representing the result of a trial that is a success if Xi equals 1 and a failure otherwise, it follows that $X = \sum_{i=1}^{n} X_i$ can be **thought of as** representing the total number of successes in n trials. Hence, **if** the Xi were independent**, then X would be a binomial random variable** with parameters n and p. In other words, when the population size N is large in relation to the sample size n, then the distribution of the number of members of the sample that possess the characteristic is **approximately** that of a binomial random variable with parameters n and p. **REMARK**: Of course, X is a hypergeometric random variable (Section 5.4); and so the preceding shows that a **hypergeometric can be approximated by a binomial random varia**ble when
the number chosen is small in relation to the total number of elements. For the remainder of this text, we will suppose that the underlying population is large in relation to the sample size and we will take the distribution of X to be binomial. By using the formulas given in Section 5.1 for the mean and standard deviation of
a binomial random variable, we see that $E[X] = np \quad \text{and} \quad SD(X) = \sqrt{np(1-p)}$ Since $\overline{X}$, the proportion of the sample that has the characteristic, is equal to X
/n, we see from the preceding that $E[\overline{X}] = E[X]/n = p \; SD(\overline{X}) = SD(X)/n = \sqrt{p(1-p)/n}$ **Even when** each element of the population has more than **two possible values**, **it still remains true** that if the population size is large in relation to the sample size, then the sample data can be regarded as being independent random variables from the population distribution.

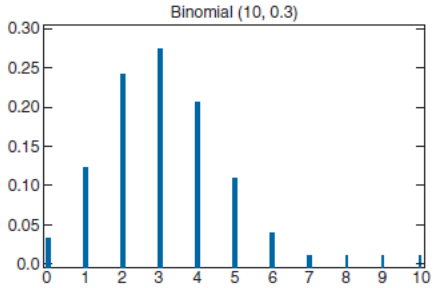
Chapter_5: SPECIAL RANDOM VARIABLES: 5.1 Certain **types of random variables** occur over and over again in applications. In this chapter, we will study a **variety** of them. 5.1 THE BERNOULLI AND BINOMIAL RANDOM VARIABLES: Suppose that a **trial**, or an **experiment**, whose outcome can be classified as either a "**success**" or as a "**failure**" is performed. If we let X = 1 when the outcome is a success and X = 0 when it is a failure, then the **probability mass function** of X is given by P{X = 0} = 1 – p, P{X = 1} = p where p, 0 ≤ p ≤ 1 (5.1.1), is the probability that the trial is a "success." A random variable X is said to be a Bernoulli random variable (after the Swiss mathematician James Bernoulli) if its probability mass function is given by Equations 5.1.1 for some p ∈ (0, 1). Its expected value is E[X] = 1 · P{X = 1} + 0 · P{X = 0} = p. That is, the **expectation** of a Bernoulli random variable is the **probability** that the **random variable equals 1**. Suppose now that n independent trials, each of which results in a "success" with probability p and in a "failure" with probability 1 – p, are to be performed. **If X** represents the **number of successes** that occur in the **n trials**, then X is said to be a binomial random variable with parameters (n, p). The probability mass function of a binomial random variable with parameters n and p is given by

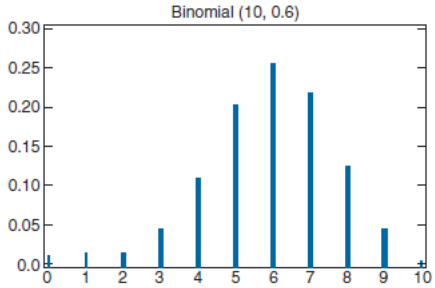
$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n$$ Equation 5.1.2 , where $\binom{n}{i} = n!/[i!(n-i)!]$ is the **number of different groups of I objects** that can be **chosen** from a set of n objects. The validity of Equation 5.1.2 may be verified by first noting that the probability **of any particular sequence of the n outcomes containing I** successes and **n – I** failures is, by the **assumed independence of trials**, noting that there are $\binom{n}{i}$ different selections of the I trials that result in successes. Note that, **by** the

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = [p + (1-p)]^n = 1$$

binomial theorem, the **probabilities sum to 1**, that is, The probability mass function of three binomial random variables with respective parameters (10, .5), (10, .3), and (10, .6) are presented in Figure 5.1.







EXAMPLE 5.1a It is known that disks produced by a certain company will be defective with probability .01 independently of each other. The company sells the disks in packages of 10 and offers a money-back guarantee that at most 1 of the 10 disks is defective. What proportion of packages is returned? If someone buys three packages, what is the probability that exactly one of them will be returned? SOLUTION If X is the number of defective disks in a package, then assuming that customers always take advantage of the guarantee, it follows that X is a binomial random variable with parameters (10, .01). Hence the probability that a package will have to be replaced is

$$P\{X > 1\} = 1 - P\{X = 0\} - P\{X = 1\} = 1 - \binom{10}{0}(.01)^0(.99)^{10} - \binom{10}{1}(.01)^1(.99)^9 \approx .005$$ Because each package will, **independently**, **have to be replaced with probability .005**, it follows from the **law of large numbers** that **in the long run** .5 **percent** of the packages will
have to be **replaced**. It follows from the foregoing that the number of packages that the person will have to return is a binomial random variable with parameters n = 3
and p = .005. Therefore, the probability that exactly one of the three packages will be returned is & $\binom{3}{1}(.005)(.995)^2 =$ 0.15. EXAMPLE 5.1c A communications system consists of n components, each of which will, **independently**, function with **probability p**. The total system will be able to operate
effectively if **at least one-half** of its components function. (a) For what values of p is a **5-component** system more likely to operate effectively than a **3-component** system? (b) In general, when is a **2k +** 1 component system better than a **2k – 1** component system? SOLUTION: (a) Because the number of functioning components is a
binomial random variable with parameters (n, p), it follows that the **probability** that a **5**-component system will be effective is $\binom{5}{3}p^3(1-p)^2 + \binom{5}{4}p^4(1-p) + p^5$
whereas the corresponding **probability** for a **3**-component system is $\binom{3}{2}p^2(1-p) + p^3$ Hence, the **5**-component system is better if

$10p^3(1-p)^2 + 5p^4(1-p) + p^5 \geq 3p^2(1-p) + p^3$ which reduces to $3(p-1)^2(2p-1) \geq 0$ or $p \geq \frac{1}{2}$ (b) In general, a system with 2k + 1 components will be better than one with 2k – 1 components if (and only if) p ≥ ½ . To prove this, consider a system of 2k + 1 components and let X denote the **number**(quantity) of the **first** 2k – 1**[quantity 3]** that function. Then $P_{2k+1}(\text{effective}) = P\{X \geq k+1\} + P\{X = k\}(1-(1-p)^2) + P\{X = k-1\}p^2$ which follows since the 2k + 1 component system will be effective if either (1) X ≥ k + 1; (2) X = k and **at least one** of the **remaining** 2 components function; or (3) X = k – 1 and **both of the next 2** function. Because

$P_{2k-1}(\text{effective}) = P\{X \geq k\} = P\{X = k\} + P\{X \geq k+1\}$ we obtain that $P_{2k+1}(\text{effective}) - P_{2k-1}(\text{effective}) = P\{X = k-1\}p^2 - (1-p)^2 P\{X = k\}$

$$= \binom{2k-1}{k-1}p^{k-1}(1-p)^k p^2 - (1-p)^2\binom{2k-1}{k}p^k(1-p)^{k-1} = \binom{2k-1}{k}p^k(1-p)^k[p-(1-p)] \quad \text{since} \binom{2k-1}{k-1} = \binom{2k-1}{k} \geq 0 \Leftrightarrow p \geq \frac{1}{2}$$ Since a binomial random variable X, with parameters n and p, represents the number of successes in n independent trials, each having success probability p, we can **represent X** as
follows: $X = \sum_{i=1}^{n} X_i$ (5.1.3) where $X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$ Because the Xi , I = 1, . . . , n are **independent Bernoulli** random variables, we have that

$E[X_i] = P\{X_i = 1\} = p \quad \text{Var}(X_i) = E[X_i^2] - p^2 \quad = p(1-p)$ where the last equality follows since $X_i^2 = X_i$, and so $E[X_i^2] = E[X_i] = p$. Using the representation

$$E[X] = \sum_{i=1}^{n} E[X_i] \qquad = np \qquad \text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) \qquad = np(1-p)$$

Equation 5.1.3, it is now an easy matter to compute the mean and variance of X: **since the Xi are independent**. If X1 and X2 are independent binomial random variables having respective parameters (ni , p), I = 1, 2, then their sum is binomial with parameters (n1 + n2, p). This can most

easily be seen by noting that because Xi , I = 1, 2, represents the number of successes in ni independent trials each of which is a success with probability p, then X1 + X2 represents the number of successes in n1 + n2 independent trials each of which is a success with probability p. Therefore, X1 + X2 is binomial with parameters (n1 + n2, p). 5.1.1 Computing the Binomial Distribution Function: Suppose that X is binomial with parameters (n, p). The **key to computing** its distribution function

$$P\{X \le i\} = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}, \qquad i = 0, 1, \ldots, n$$

is to **utilize the following relationship** between P{X = k + 1} and P{X = k}:

$$P\{X = k+1\} = \frac{p}{1-p} \frac{n-k}{k+1} P\{X = k\}$$

5.2 THE **POISSON** RANDOM VARIABLE: A random variable X, taking on one of the values 0, 1, 2, . . . , is said to be a Poisson

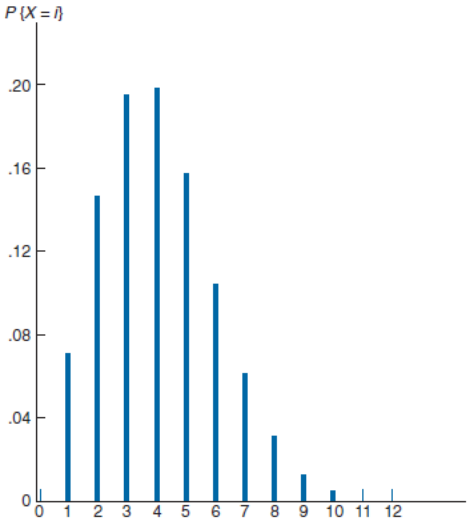
random variable with parameter λ,λ > 0, if its probability mass function is given by equation (5.2.1)

$$P\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = 0, 1, \ldots$$

The symbol e stands for a constant approximately equal to **2.7183.** It is a famous constant in mathematics, named after the Swiss mathematician L. Euler, and it is **also the base of the**

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda} \sum_{i=0}^{\infty} \lambda^i / i! = e^{-\lambda} e^{\lambda} = 1$$

so-called natural logarithm. Equation 5.2.1 defines a probability mass function, since A graph of this mass function when λ = 4 is



given in Figure 5.3. As a prelude to determining the mean and variance of a Poisson random variable, let us first determine its moment generating function. $\phi(t) = E[e^{tX}] = \sum_{i=0}^{\infty} e^{ti} e^{-\lambda} \lambda^i / i! = e^{-\lambda} \sum_{i=0}^{\infty} (\lambda e^t)^i / i! \quad = e^{-\lambda} e^{\lambda e^t} \quad = \exp\{\lambda(e^t - 1)\}$ Differentiation yields

$\phi'(t) = \lambda e^t \exp\{\lambda(e^t - 1)\}$ $\phi''(t) = (\lambda e^t)^2 \exp\{\lambda(e^t - 1)\} + \lambda e^t \exp\{\lambda(e^t - 1)\}$ Evaluating at t = 0 gives that $E[X] = \phi'(0) = \lambda \quad \text{Var}(X) = \phi''(0) - (E[X])^2$

$= \lambda^2 + \lambda - \lambda^2 = \lambda$ **Thus both the mean and the variance** of a Poisson random variable are **equal** to the **parameter λ**. The Poisson random variable has a wide range of applications in a variety of areas because it may be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small. To see

this, suppose that X is a binomial random variable with parameters (n, p) and let λ = np. Then $P\{X = i\} = \frac{n!}{(n-1)!i!} p^i (1-p)^{n-i} = \frac{n!}{(n-1)!i!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i}$

$= \frac{n(n-1)\ldots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}$ Now, for n large and p small, $\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda} \quad \frac{n(n-1)\ldots(n-i+1)}{n^i} \approx 1 \quad \left(1 - \frac{\lambda}{n}\right)^i \approx 1$ Hence, for n large and p small,

$P\{X = i\} \approx e^{-\lambda} \frac{\lambda^i}{i!}$ In other words, if n independent trials, each of which results in a "success" with probability p, are performed, then when n is large and p small, the number of successes occurring is approximately a Poisson random variable with mean λ = np. Some examples of random variables that usually obey, to a good approximation, the Poisson probability law (that is, they usually obey Equation 5.2.1 for some value of λ) are: 1. The number of misprints on a page (or a group of pages) of a book. 2. The number of people in a community living to 100 years of age. 3. The number of wrong telephone numbers that are 19ormal19 in a day. 4. The number of transistors that fail on their first day of use. 5. The number of customers entering a post office on a given day. 6. The number of α-particles discharged in a fixed period of time from some radioactive

particle. Each of the foregoing, and numerous other random variables, is approximately Poisson for the same reason—**namely, because of the Poisson approximation to the binomial**. For instance, we can suppose that there is a small probability p that each letter typed on a page will be misprinted, and so the number of misprints on a given page will be approximately Poisson with mean λ = np where n is the (presumably) large number of letters on that page. Similarly, we can suppose that each person in a given community, independently, has a small probability p of reaching the age 100, and so the number of people that do will have approximately a Poisson distribution with mean np where n is the large number of people in the community. We leave it for the reader to reason out why the remaining random variables in examples 3 through 6 should have approximately a Poisson distribution. The Poisson approximation result can be shown to be valid under even more general conditions than those so far mentioned. For instance, suppose that n independent trials are to be performed, with the ith trial resulting in a success with probability pi , I = 1, . . . , n. Then it can be shown that if n is large and each pi is small, then the number of successful

trials is approximately Poisson distributed with mean equal to $\sum_{i=1}^{n} p_i$. In fact**, this result will sometimes remain true even when the trials are not independent**, **provided that their dependence is "weak."** For **instance**, consider the following **example**. EXAMPLE 5.2e At a party n people put their hats in the center of a room, where the hats are mixed together. Each person then randomly chooses a hat. If X denotes the number of people who select their own hat then, for large n, it can be

$$X_i = \begin{cases} 1 & \text{if the } i\text{th person selects his or her own hat} \\ 0 & \text{otherwise} \end{cases}$$

shown that X has approximately a Poisson distribution with mean 1. To see why this might be true, let Then we can express X as X = X1 +· · ·+Xn and so X can be regarded as representing the number of "successes" in n "trials" where trial I is said to be a success if the ith person chooses

his own hat. Now, since the ith person is equally likely to end up with any of the n hats, one of which is his own, it follows that (5.2.2) $P\{X_i = 1\} = \frac{1}{n}$ Suppose now that $i \neq j$ and consider the conditional probability that the ith person chooses his own hat given that the jth person does—that is, consider P{Xi = 1|Xj = 1}. Now given that the jth person indeed selects his own hat, it follows that the ith individual is equally likely to end up with any of the remaining n−1, one of which is his own. Hence, it follows that (5.2.3) Thus, we see from Equations 5.2.2 and 5.2.3 that whereas the trials are not independent, their dependence is rather weak [since, if the above conditional probability were equal to 1/n rather than 1/(n – 1), then trials I and j would be independent]; and thus it is not at all surprising that X has approximately a

Poisson distribution. The fact that E[X] = 1 follows since E[X] = E[X1 +· · ·+Xn] = E[X1]+· · ·+E[Xn] $= n\left(\frac{1}{n}\right) = 1$ The last equality follows since, from Equation 5.2.2, E[Xi] =

P{Xi = 1} = 1/n. The **Poisson** distribution **possesses the reproductive property** that the sum of independent Poisson random variables is also a Poisson random variable. To see this, suppose that
X1 and X2 are independent Poisson random variables having respective means λ1 and λ2**. Then the moment generating function** of X1 + X2 is as follows:

$$E[e^{t(X_1+X_2)}] = E[e^{tX_1}e^{tX_2}] = E[e^{tX_1}]E[e^{t}X_2] \quad \text{by independence} = \exp\{\lambda_1(e^t-1)\}\ \exp\{\lambda_2(e^t-1)\} = \exp\{(\lambda_1+\lambda_2)(e^t-1)\}$$

Because $\exp\{(\lambda_1+\lambda_2)(e^t-1)\}$ is the moment generating function of a Poisson random variable having mean λ1+λ2, we may conclude, from the fact that the **moment generating function uniquely specifies the distribution**, that X1 + X2 is Poisson with mean λ1 + λ2. Consider now a situation in which a random number, call it N, of events will occur, and suppose that each of these events will independently be a type 1 event with probability p or a type 2 event with probability 1−p. Let N1 and N2 denote, respectively, the numbers of type 1 and type 2 events that occur. (So N = N1 + N2.) **If N** is **Poisson** distributed with mean λ, then **the joint probability mass function** of N1 and N2 is

$$= P\{N_1 = n, N_2 = m | N = n+m\}e^{-\lambda}\frac{\lambda^{n+m}}{(n+m)!}$$

obtained as follows. P{N1 = n,N2 = m} = P{N1 = n,N2 = m,N = n + m} = P{N1 = n,N2 = m | N = n + m}P{N = n + m} Now, given a total of n + m events, because each one of these events is independently type 1 with probability p, it follows that the conditional probability that there are exactly n type 1 events (and m type 2 events) is the probability that a binomial (n+m, p) random variable is equal to n. Consequently, (5.2.4)

$$P\{N_1 = n, N_2 = m\} = \frac{(n+m)!}{n!m!}p^n(1-p)^m e^{-\lambda}\frac{\lambda^{n+m}}{(n+m)!} = e^{-\lambda p}\frac{(\lambda p)^n}{n!}e^{-\lambda(1-p)}\frac{(\lambda(1-p))^m}{m!}$$

The probability mass function of N1 is thus (5.2.5)

$$P\{N_1 = n\} = \sum_{m=0}^{\infty} P\{N_1 = n, N_2 = m\} = e^{-\lambda p}\frac{(\lambda p)^n}{n!}\sum_{m=0}^{\infty} e^{-\lambda(1-p)}\frac{(\lambda(1-p))^m}{m!} = e^{-\lambda p}\frac{(\lambda p)^n}{n!}$$

Similarly, (5.2.6)

$$P\{N_2 = m\} = \sum_{n=0}^{\infty} P\{N_1 = n, N_2 = m\} = e^{-\lambda(1-p)}\frac{(\lambda(1-p))^m}{m!}$$

It now follows from Equations 5.2.4, 5.2.5, and 5.2.6, that N1 and N2 are independent Poisson random variables with respective means λp and λ(1 – p). The preceding result generalizes **when each of the Poisson number of events can be classified into any of r categories**, to yield the following important property of the Poisson distribution: If each of a Poisson number of events having mean λ is **independently classified** as being of **one of the types** 1, . . . , r, with respective probabilities p1, . . . , pr, $\sum_{i=1}^{r} p_i = 1$, **then the numbers of** type 1, . . . , r events are **independent** Poisson random variables with **respective** means λp1, . . . , λpr . 5.2.1 Computing the Poisson Distribution Function: If X is Poisson with mean λ, then (5.2.7)

$$\frac{P\{X = i+1\}}{P\{X = i\}} = \frac{e^{-\lambda}\lambda^{i+1}/(i+1)!}{e^{-\lambda}\lambda^i/i!} = \frac{\lambda}{i+1}$$

Starting with P{X = 0} =$e^{-\lambda}$, we can use Equation 5.2.7 to successively compute P{X = 1} = λP{X = 0} , P{X = 2} = λ/2 P{X = 1} .....P{X = I + 1} = λ/I + 1 P{X = i} The text disk includes a program that uses Equation 5.2.7 to compute Poisson probabilities. 5.3 THE **HYPERGEOMETRIC** RANDOM VARIABLE: A bin contains N +M batteries, of which N are of acceptable quality and the other M are defective**. A sample of size n** is to be randomly chosen (without replacements) in the sense that the **set of sampled batteries** is **equally likely to be any of the** $\binom{N+M}{n}$**subsets of size n**. If we let X denote the number of acceptable batteries in the sample, then

$$P\{X = i\} = \frac{\binom{N}{i}\binom{M}{n-i}}{\binom{N+M}{n}}, \quad i = 0, 1, \ldots, \min(N, n)^*$$

(5.3.1) Any random variable X whose probability mass function is given by Equation 5.3.1 is said to be a hypergeometric random variable with **parameters N,M, n.** To compute the mean and variance of a hypergeometric random variable whose probability mass function is given by Equation 5.3.1, imagine that the batteries are drawn sequentially and let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th selection is acceptable} \\ 0 & \text{otherwise} \end{cases}$$

Now, since the ith selection is equally likely to be any of the N + M batteries, of which N are acceptable, it follows that

$$P\{X_i = 1\} = \frac{N}{N+M} \quad (5.3.2)$$

Also, for $i \neq j$, $P\{X_i = 1, X_j = 1\} = P\{X_i = 1\}P\{X_j = 1 | X_i = 1\} = \frac{N}{N+M}\frac{N-1}{N+M-1}$ (5.3.3) which follows since, given that the ith selection is acceptable, the jth selection is equally likely to be any of the other N + M – 1 batteries of which N – 1 are acceptable. To compute the mean and variance of X, the number of acceptable batteries in thesample of size n, use the representation

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

This gives

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P\{X_i = 1\} = \frac{nN}{N+M}$$

(5.3.4) Also,

$$\text{Var}(X) = \sum_{i=1}^{n}\text{Var}(X_i) + 2\sum\sum_{1 \le i < j \le n}\text{Cov}(X_i, X_j)$$

Corollary 4.7.3 for the variance of a sum of random variables gives (5.3.5) Now, Xi is a Bernoulli random variable and so Var(Xi) = P{Xi = 1}(1 – P{Xi = 1}) = $\frac{N}{N+M}\frac{M}{N+M}$ (5.3.6) Also, for I < j, Cov(Xi , Xj) = E[XiXj] – E[Xi]E[Xj] , Now, because both Xi and Xj are Bernoulli (that is, 0 – 1) random variables, it follows that XiXj is a Bernoulli random variable, and so

$$E[X_iX_j] = P\{X_iX_j = 1\} = P\{X_i = 1, X_j = 1\} = \frac{N(N-1)}{(N+M)(N+M-1)}$$

from (5.3.3) So from Equation 5.3.2 and the foregoing we see that for

$$i \neq j, \ \text{Cov}(X_i, X_j) = \frac{N(N-1)}{(N+M)(N+M-1)} - \left(\frac{N}{N+M}\right)^2 = \frac{-NM}{(N+M)^2(N+M-1)}$$

Hence, since there are $\binom{n}{2}$ terms in the second sum on the right side of Equation 5.3.5, we obtain from Equation 5.3.6

$$\text{Var}(X) = \frac{nNM}{(N+M)^2} - \frac{n(n-1)NM}{(N+M)^2(N+M-1)}$$

$$= \frac{nNM}{(N+M)^2}\left(1 - \frac{n-1}{N+M-1}\right)$$

(5.3.8) If we let p = N/(N +M) denote the proportion of batteries in the bin that are acceptable, we can rewrite Equations 5.3.4 and 5.3.8 as follows. E(X) = np

$$\text{Var}(X) = np(1-p)\left[1 - \frac{n-1}{N+M-1}\right]$$

It should be noted that**, for fixed p, as N + M increases to ∞, Var(X) converges to np(1 – p)**, which is the **variance** of a **binomial random variable with parameters (n, p).** There is a **relationship** between **binomial** random variables **and** the **hypergeometric** distribution that will be useful to us in developing a **statistical test concerning two binomial populations**. EXAMPLE 5.3c Let X and Y be independent binomial random variables having respective parameters (n, p) and (m, p). The **conditional probability mass function of X given that X + Y = k** is as follows.

$$P\{X = i | X + Y = k\} = \frac{P\{X = i, X + Y = k\}}{P\{X + Y = k\}} = \frac{P\{X = i, Y = k - i\}}{P\{X + Y = k\}} = \frac{P\{X = i\}P\{Y = k - i\}}{P\{X + Y = k\}}$$

$$= \frac{\binom{n}{i}p^i(1-p)^{n-i}\binom{m}{k-i}p^{k-i}(1-p)^{m-(k-i)}}{\binom{n+m}{k}p^k(1-p)^{n+m-k}} = \frac{\binom{n}{i}\binom{m}{k-i}}{\binom{n+m}{k}}$$

where the next-to-last equality **used the fact that X + Y is binomial with parameters (n + m, p)**. Hence, we see that the **conditional distribution of X given the value of X + Y is hypergeometric**. It is worth noting that the preceding is quite intuitive. For suppose that n + m independent trials, each of which has the same probability of being a success, are performed; let X be the number of successes in the first n trials, and let Y be the number of successes in the final m trials. Given a total of k successes in the n + m trials, it is quite intuitive that each subgroup of k trials is equally likely to consist of those trials that resulted in successes. That is, the k success trials are distributed as a random selection of k of the n + m trials, and so the number that are from the first n trials is hypergeometric. 5.4 THE **UNIFORM** RANDOM VARIABLE: A random variable X is said to be uniformly distributed over the interval [α, β] if its **probability density function** is given by

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$$

A graph of this function is given in Figure 5.4. Note that the foregoing meets the requirements of being a probability density function since

$$\frac{1}{\beta - \alpha}\int_{\alpha}^{\beta} dx = 1$$

The uniform distribution arises in practice when we suppose a certain random variable is equally likely to be near any value in the interval [α, β]. The probability that X lies in any subinterval of [α, β] is equal to the length of that subinterval divided by the length of the interval [α, β]. This follows since when [a, b] is a subinterval of [α, β] (see Figure 5.5),

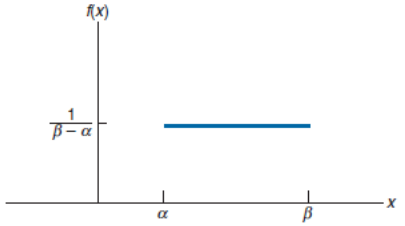


FIGURE 5.4 Graph of f (x) for a uniform [α, β].

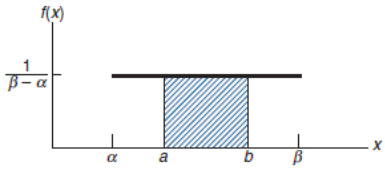


FIGURE 5.5 Probabilities of a uniform random variable.

$$P\{a < X < b\} = \frac{1}{\beta - \alpha}\int_a^b dx = \frac{b-a}{\beta-\alpha}$$ The mean of a uniform [α, β] random variable is $$E[X] = \int_\alpha^\beta \frac{x}{\beta-\alpha}\,dx = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)}$$ $$= \frac{(\beta-\alpha)(\beta+\alpha)}{2(\beta-\alpha)}$$ or $$E[X] = \frac{\alpha+\beta}{2}$$ Or, in other words, the expected value of a uniform [α, β] random variable is equal to the **midpoint** of the interval [α, β], which is clearly what one would expect. The variance is computed as follows.

$$E[X^2] = \frac{1}{\beta-\alpha}\int_\alpha^\beta x^2\,dx = \frac{\beta^3-\alpha^3}{3(\beta-\alpha)} = \frac{\beta^2+\alpha\beta+\alpha^2}{3}$$ and so,

$$\mathrm{Var}(X) = \frac{\beta^2+\alpha\beta+\alpha^2}{3} - \left(\frac{\alpha+\beta}{2}\right)^2 = \frac{\alpha^2+\beta^2-2\alpha\beta}{12} = \frac{(\beta-\alpha)^2}{12}$$ **The value of a uniform (0, 1) random variable is called a random number.** Most computer systems have a built-in subroutine for generating (to a high level of approximation) sequences of independent random numbers — for instance, Table 5.1 presents a set of independent random numbers generated by an IBM personal computer. Random numbers are quite useful in probability and statistics because their use enables one to empirically estimate various probabilities and expectations.

TABLE 5.1 *A Random Number Table*

.68587	.25848	.85227	.78724	.05302	.70712	.76552	.70326	.80402	.49479
.73253	.41629	.37913	.00236	.60196	.59048	.59946	.75657	.61849	.90181
.84448	.42477	.94829	.86678	.14030	.04072	.45580	.36833	.10783	.33199
.49564	.98590	.92880	.69970	.83898	.21077	.71374	.85967	.20857	.51433
.68304	.46922	.14218	.63014	.50116	.33569	.97793	.84637	.27681	.04354
.76992	.70179	.75568	.21792	.50646	.07744	.38064	.06107	.41481	.93919
.37604	.27772	.75615	.51157	.73821	.29928	.62603	.06259	.21552	.72977
.43898	.06592	.44474	.07517	.44831	.01337	.04538	.15198	.50345	.65288
.86039	.28645	.44931	.59203	.98254	.56697	.55897	.25109	.47585	.59524
.28877	.84966	.97319	.66633	.71350	.28403	.28265	.61379	.13886	.78325
.44973	.12332	.16649	.88908	.31019	.33358	.68401	.10177	.92873	.13065
.42529	.37593	.90208	.50331	.37531	.72208	.42884	.07435	.58647	.84972
.82004	.74696	.10136	.35971	.72014	.08345	.49366	.68501	.14135	.15718
.67090	.08493	.47151	.06464	.14425	.28381	.40455	.87302	.07135	.04507
.62825	.83809	.37425	.17693	.69327	.04144	.00924	.68246	.48573	.24647
.10720	.89919	.90448	.80838	.70997	.98438	.51651	.71379	.10830	.69984
.69854	.89270	.54348	.22658	.94233	.08889	.52655	.83351	.73627	.39018
.71460	.25022	.06988	.64146	.69407	.39125	.10090	.08415	.07094	.14244
.69040	.33461	.79399	.22664	.68810	.56303	.65947	.88951	.40180	.87943
.13452	.36642	.98785	.62929	.88509	.64690	.38981	.99092	.91137	.02411
.94232	.91117	.98610	.71605	.89560	.92921	.51481	.20016	.56769	.60462
.99269	.98876	.47254	.93637	.83954	.60990	.10353	.13206	.33480	.29440
.75323	.86974	.91355	.12780	.01906	.96412	.61320	.47629	.33890	.22099
.75003	.98538	.63622	.94890	.96744	.73870	.72527	.17745	.01151	.47200

For an illustration of the use of random numbers, suppose that a medical center is planning to test a new drug designed to reduce its users' blood cholesterol levels. To test its effectiveness, the medical center has recruited 1,000 volunteers to be subjects in the test. To take into account the possibility that the subjects' blood cholesterol levels may be affected by factors external to the test (such as changing weather conditions), it has been decided to split the volunteers into 2 groups of size 500 — a **treatment** group that will be given the drug and a **control** group that will be given a placebo. Both the volunteers and the administrators of the drug will not be told who is in each group (such a test is called a **double-blind** test). It remains to determine which of the volunteers should be chosen to constitute the treatment group. Clearly, one would want the treatment group and the control group to be as similar as possible in all respects with the exception that members in the first group are to receive the drug while those in the other group receive a placebo; then it will be possible to conclude that any difference in response between the groups is indeed due to the drug. There is general agreement that the best way to accomplish this is to choose the 500 volunteers to be in the treatment group in **a completely random fashion**. That is, the choice should be made so that each of the $\binom{1000}{500}$ **subsets** of 500 volunteers is **equally likely** to constitute the **control** group. **How** can this be **accomplished**? EXAMPLE 5.4d **Choosing** a **Random Subset**: From a set of n elements — numbered 1, 2, . . . , n — suppose we want to generate a random subset of size k that is to be chosen in such a manner that each of the subsets $\binom{n}{k}$ is equally likely to be the subset chosen. How can we do this? **To answer this question**, let us work backwards and suppose that we have indeed randomly generated such a subset of size k. Now for each j = 1, . . . , n, we set

$$I_j = \begin{cases} 1 & \text{if element } j \text{ is in the subset} \\ 0 & \text{otherwise} \end{cases}$$

and compute the conditional distribution of Ij given $I_1, \ldots, I_{j-1}$ To start, note that the probability that element 1 is in the subset of size k is clearly k/n (which can be seen either by noting that there is probability 1/n that element 1 would have been the jth element chosen, j = 1, . . . , k; or by noting that the proportion of outcomes of the random selection that results in element 1 being chosen is $\binom{1}{1}\binom{n-1}{k-1}/\binom{n}{k} = k/n$) Therefore, we have that $P\{I_1 = 1\} = k/n$ (5.4.1) To compute the conditional probability that element 2 is in the subset given I1, note that if I1 = 1, then aside from element 1 the remaining k – 1 members of the subset would have been chosen "at random" from the remaining n−1 elements (in the sense that each of the subsets of size k – 1 of the numbers 2, . . . , n is equally likely to be the other elements of the subset). Hence, we have that $P\{I_2 = 1 | I_1 = 1\} = \frac{k-1}{n-1}$ (5.4.2), Similarly, if element 1 is not in the subgroup, then the k members of the subgroup would have been chosen "at random" from the other n – 1 elements, and thus $P\{I_2 = 1 | I_1 = 0\} = \frac{k}{n-1}$ (5.4.3), From Equations 5.4.2 and 5.4.3, we see that $P\{I_2 = 1 | I_1\} = \frac{k - I_1}{n-1}$ In general, we have that

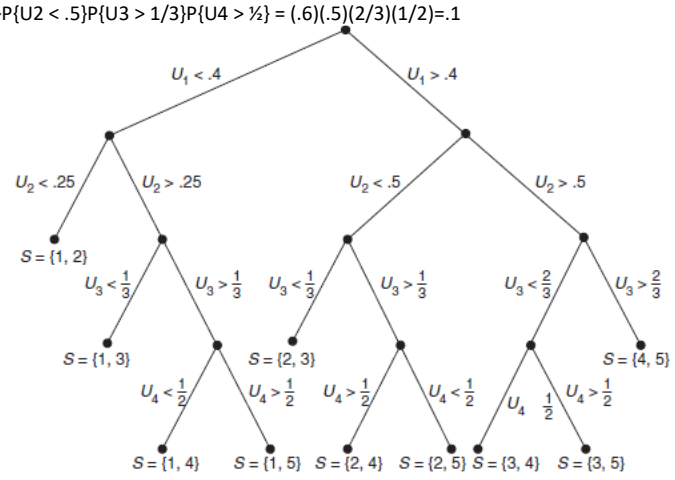
$$P\{I_j = 1 | I_1, \ldots, I_{j-1}\} = \frac{k - \sum_{i=1}^{j-1} I_i}{n-j+1}, \quad j = 2, \ldots, n$$ (5.4.4), The preceding formula follows since $\sum_{i=1}^{j-1} I_i$ represents the number of the first j – 1 elements that are included in the subset, and so given I1, . . . , Ij−1 there remain $k - \sum_{i=1}^{j-1} I_i$ elements to be selected from the remaining n – (j – 1). Since P{U < a} = a, 0 ≤ a ≤ 1, when U is a uniform (0, 1) random variable, Equations 5.4.1 and 5.4.4 lead to the following method

$$I_1 = \begin{cases} 1 & \text{if } U_1 < \frac{k}{n} \\ 0 & \text{otherwise} \end{cases}$$

for generating a random subset of size k from a set of n elements: Namely, generate a sequence of (at most n) random numbers U1,U2, . . . and set

$$I_2 = \begin{cases} 1 & \text{if } U_2 < \frac{k - I_1}{n-1} \\ 0 & \text{otherwise} \end{cases} \quad I_j = \begin{cases} 1 & \text{if } U_j < \frac{k - I_1 - \cdots - I_{j-1}}{n-j+1} \\ 0 & \text{otherwise} \end{cases}$$

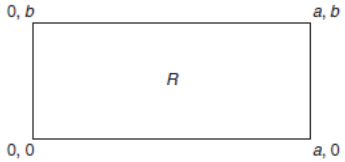
, ... This process stops when I1+· · ·+Ij = k and the random subset consists of the k elements whose I-value equals 1. That is, $S = \{i : I_i = 1\}$ is the subset. For instance, if k = 2, n = 5, then the tree diagram of Figure 5.6 illustrates the foregoing technique. The random subset S is given by the final position on the tree. Note that the probability of ending up in any given final position is equal to 1/10, which can be seen by multiplying the probabilities of moving through the tree to the desired endpoint. For instance, the probability of ending at the point labeled S = {2, 4} is P{U1 >.4}P{U2 < .5}P{U3 > 1/3}P{U4 > ½} = (.6)(.5)(2/3)(1/2)=.1



= .1. FIGURE 5.6 Tree diagram. As indicated in the tree diagram (see the rightmost branches that result in S = {4, 5}), **we can stop generating random numbers when the number of remaining places in the subset to be chosen is equal to the remaining number of elements.** That is, the general procedure would stop whenever either $\sum_{i=1}^{j} I_i = k$ or $\sum_{i=1}^{j} I_i = k - (n - j)$. In the latter case, $S = \{i \le j : I_i = 1, j+1, \ldots, n\}$. EXAMPLE 5.4e The random vector X, Y is said to have a uniform distribution over the two-dimensional region R if its joint density function is constant for points in R, and is 0 for points outside of R. That is, if

$$f(x,y) = \begin{cases} c & \text{if } (x,y) \in R \\ 0 & \text{if otherwise} \end{cases}$$

Because, $1 = \int_R f(x,y)\,dx\,dy = \int_R c\,dx\,dy$ $= c \times \text{Area of } R$ it follows that $c = \frac{1}{\text{Area of } R}$ For any region A ⊂ R, $P\{(X,Y) \in A\} = \int\int_{(x,y)\in A} f(x,y)\,dx\,dy = \int\int_{(x,y)\in A} c\,dx\,dy = \frac{\text{Area of } A}{\text{Area of } R}$ Suppose now that X, Y is uniformly distributed over the following rectangular region R:



Its joint density function is

$$f(x,y) = \begin{cases} c & \text{if } 0 \le x \le a,\ 0 \le y \le b \\ 0 & \text{otherwise} \end{cases}$$

where $c = \frac{1}{\text{Area of rectangle}} = \frac{1}{ab}$. In this case, X and Y are independent uniform random variables. To show this, note that for 0 ≤ x ≤ a, 0 ≤ y ≤ b $P\{X \le x, Y \le y\} = c\int_0^x \int_0^y dy\,dx = \frac{xy}{ab}$ (5.4.5) First letting y = b, and then letting x = a, in the preceding shows that $P\{X \le x\} = \frac{x}{a}, \quad P\{Y \le y\} = \frac{y}{b}$ (5.4.6). Thus, from Equations 5.4.5 and 5.4.6 we can conclude that X and Y are independent, with X being uniform on (0, a) and Y being uniform on (0, b).

5.5 **NORMAL** RANDOM VARIABLES: A random variable is said to be normally distributed with parameters μ and σ2, and we write X ~ N(μ, σ^2), if its density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty^*$$

The normal density f (x) is a bell-shaped curve that is symmetric about μ and that attains its **maximum value of** $1/\sigma\sqrt{2\pi}$ ≈ 0. 399/σ at x = μ (see Figure 5.7).

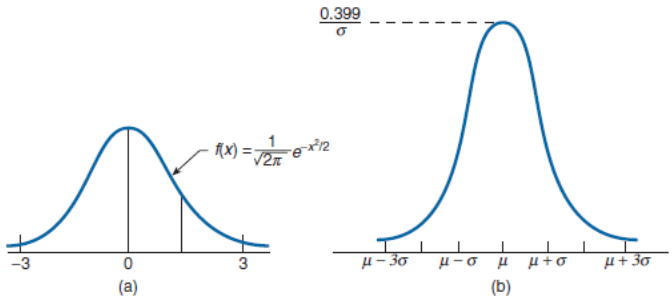


FIGURE 5.7 *The normal density function (a) with $\mu = 0, \sigma = 1$ and (b) with arbitrary μ and σ^2.* The normal distribution can be used to **approximate** probabilities associated with binomial random variables **when** the **binomial** parameter n is **large**. The **moment generating function** of a normal random variable with parameters μ and σ^2 σ2 is derived as follows:

$$\phi(t) = E[e^{tX}] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2\sigma^2}\,dx = \frac{1}{\sqrt{2\pi}} e^{\mu t}\int_{-\infty}^{\infty} e^{t\sigma y} e^{-y^2/2}\,dy \qquad \text{by letting } y = \frac{x-\mu}{\sigma}$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left\{-\left[\frac{y^2 - 2t\sigma y}{2}\right]\right\} dy = \frac{e^{\mu t}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left\{-\frac{(y - t\sigma)^2}{2} + \frac{t^2\sigma^2}{2}\right\} dy = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(y-t\sigma)^2/2}\,dy = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \quad (5.5.1)$$

where the last equality follows since $\frac{1}{\sqrt{2\pi}}e^{-(y-t\sigma)^2/2}$ is the density of a normal random variable (having parameters tσ and 1) and its integral must thus equal 1. Upon differentiating Equation 5.5.1, we obtain $\phi'(t) = (\mu + t\sigma^2)\exp\left\{\mu t + \sigma^2\frac{t^2}{2}\right\}$, $\phi''(t) = \sigma^2\exp\left\{\mu t + \sigma^2\frac{t^2}{2}\right\} + \exp\left\{\mu t + \sigma^2\frac{t^2}{2}\right\}(\mu + t\sigma^2)^2$ Hence, $E[X] = \phi'(0) = \mu$ $E[X^2] = \phi''(0) = \sigma^2 + \mu^2$ and so E[X] = μ $\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \sigma^2$ Thus μ and σ^2 represent respectively the mean and variance of the distribution. An important fact about normal random variables is that **if X** is normal with mean μ and variance σ^2 , **then Y** = αX + β is normal with mean αμ + β and variance $\alpha^2\sigma^2$. That this is so can easily be seen by using moment generating functions as follows. $E[e^{t(\alpha X+\beta)}] = e^{t\beta}E[e^{\alpha tX}] = e^{t\beta}\exp\{\mu\alpha t + \sigma^2(\alpha t)^2/2\}$ from Equation 5.5.1 $= \exp\{(\beta + \mu\alpha)t + \alpha^2\sigma^2t^2/2\}$ Because the final equation is the **moment generating function** of the **normal random variable** with **mean** β + μα and **variance** $\alpha^2\sigma^2$, the result follows. It follows from the foregoing that if X ~ N(μ, σ^2), then $Z = \frac{X-\mu}{\sigma}$ is a normal random variable with **mean 0** and **variance 1**. Such a random variable Z is said to have a **standard**, **or unit**, normal distribution. **Let (·)** denote its distribution function. That is, $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}dy, \quad -\infty < x < \infty$ This result that Z = (X – μ)/σ has a standard normal distribution when X is normal with parameters μ and σ^2 is quite **important**, for it enables us to write all probability statements about X in terms of probabilities for Z. For instance, to obtain P{X < b}, we note that X will be less than b if and only if (X – μ)/σ is less than (b – μ)/σ, and so $P\{X < b\} = P\left\{\frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right\} = \Phi\left(\frac{b-\mu}{\sigma}\right)$ Similarly, for any a < b, $P\{a < X < b\} = P\left\{\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right\} = P\left\{\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right\} = P\left\{Z < \frac{b-\mu}{\sigma}\right\} - P\left\{Z < \frac{a-\mu}{\sigma}\right\} = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

It remains for us to compute $\Phi(x)$**.** This has been accomplished by an **approximation** and the results are presented in Table A1 of the Appendix, which tabulates $\Phi(x)$ (to a 4-digit level of accuracy) for a wide range of nonnegative values of x. In addition, Program 5.5a of the text disk can be used to obtain $\Phi(x)$. While Table A1 tabulates $\Phi(x)$ only for nonnegative values of x, we can also obtain $\Phi(-x)$ from the table by making use of the **symmetry** (about 0) of the standard normal probability **density** function. That is, for x > 0, if Z represents a standard normal random variable, then (see Figure 5.8)

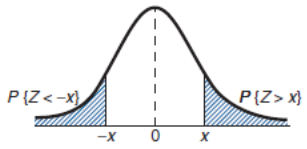


FIGURE 5.8 *Standard normal probabilities.*

$\Phi(-x) = P\{Z < -x\} = P\{Z > x\}$ **by symmetry** $= 1 - \Phi(x)$ **,** Thus, for instance, $P\{Z < -1\} = \Phi(-1) = 1 - \Phi(1) = 1 - .8413 = .1587$ **Another important** result is that the **sum** of independent normal random variables is **also** a normal random variable. To see this, suppose that Xi , i = 1, . . . , n, are independent, with Xi being normal with mean μ_i and variance σ_i^2 . The **moment generating function** of $\sum_{i=1}^n X_i$ is as follows. $E\left[\exp\left\{t\sum_{i=1}^n X_i\right\}\right] = E[e^{tX_1}e^{tX_2}\cdots e^{tX_n}] = \prod_{i=1}^n E[e^{tX_i}]$ by independence $= \prod_{i=1}^n e^{\mu_i t + \sigma_i^2 t^2/2}$ where

$\mu = \sum_{i=1}^n \mu_i, \qquad \sigma^2 = \sum_{i=1}^n \sigma_i^2$ Therefore, $\sum_{i=1}^n X_i$ has the **same** moment generating function **as** a **normal random variable** having mean μ and variance σ2. Hence, from the **one-to-one correspondence** between moment generating functions and distributions, we can conclude that $\sum_{i=1}^n X_i$ is normal with **mean** $\sum_{i=1}^n \mu_i$ and **variance** $\sum_{i=1}^n \sigma_i^2$. That is, the **probability** that a **standard** normal random variable is **greater than** z_α is **equal** to **α** (see Figure 5.9.)

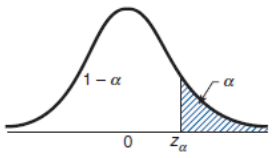


FIGURE 5.9 $P\{Z > z_\alpha\} = \alpha.$

The value of z_α can, for any α, be obtained from **Table A1(appendix)**. For instance, since $1 - \Phi(1.645) = .05$, $1 - \Phi(1.96) = .025$, $1 - \Phi(2.33) = .01$ it follows that $z_{.05} = 1.645, \quad z_{.025} = 1.96, \quad z_{.01} = 2.33$ Program 5.5b on the text disk can also be used to obtain the value of z_α Since $P\{Z < z_\alpha\} = 1 - \alpha$ it follows that 100(1 – –) percent of the time a **standard normal** random variable **will be less than** z_α. As a result, we call z_α the **100(1 – –) percentile of the standard normal distribution.** 5.6 **EXPONENTIAL** RANDOM VARIABLES: A continuous random variable whose probability **density** function is given, for some λ > 0, by $f(x) = \begin{cases}\lambda e^{-\lambda x} & \text{if } x \geq 0\\ 0 & \text{if } x < 0\end{cases}$ is said to be an exponential random variable (or, more simply, is said to be exponentially distributed) with parameter λ. The **cumulative distribution function F (x)** of an exponential random variable is given by F (x) = P{X ≤ x} $= \int_0^x \lambda e^{-\lambda y}\,dy$ $= 1 - e^{-\lambda x}, \qquad x \geq 0$ The exponential distribution often arises, in practice, as **being the distribution of the amount of time until some specific event occurs**. For instance, the amount of time (starting from now) **until** an **earthquake** occurs, or **until** a new **war** breaks out, or **until** a **telephone** call you receive turns out to be a wrong number are all random variables that tend in practice to have exponential distributions (see Section 5.6.1 for an explanation). The **moment generating function** of the exponential is given by $\phi(t) = E[e^{tX}] = \int_0^\infty e^{tx}\lambda e^{-\lambda x}\,dx = \lambda\int_0^\infty e^{-(\lambda-t)x}\,dx = \frac{\lambda}{\lambda - t}, \qquad t < \lambda$ Differentiation yields $\phi'(t) = \frac{\lambda}{(\lambda-t)^2}$ $\phi''(t) = \frac{2\lambda}{(\lambda-t)^3}$ and so, $E[X] = \phi'(0) = 1/\lambda$ $\mathrm{Var}(X) = \phi''(0) - (E[X])^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$ Thus **λ** is the **reciprocal** of the **mean**, and the **variance** is **equal** to the **square** of the **mean**. The **key property** of an exponential random variable is that it is **memoryless**, where we say that a **nonnegative** random variable X is memoryless **if** P{X > s + t |X > t} = P{X > s} for all s, t ≥ 0 (5.6.1) To understand why Equation 5.6.1 is called the **memoryless property**, imagine that X represents the length of time that a certain item functions before failing. Now let us consider the probability that an item that is **still** functioning at age t will continue to function for at least an additional time s. Since this will be the case if the total functional lifetime of the item **exceeds t + s** given that the item is **still** functioning at **t** , we see that, **P{additional** functional life of **t**-unit-old item **exceeds s**} = P{X > t + s|X > t } Thus, we see that Equation 5.6.1 states that the **distribution** of additional functional life of an item of **age t** is the **same as** that of a **new** item — in other words, when Equation 5.6.1 is satisfied, there is **no need to**

remember the age of a functional **item** since as long as it is still functional it **is "as good as new."** The condition in Equation 5.6.1 is equivalent to $\frac{P\{X > s+t, X > t\}}{P\{X > t\}} = P\{X > s\}$ or $P\{X > s+t\} = P\{X > s\}P\{X > t\}$ (5.6.2) When X is an exponential random variable, then $P\{X > x\} = e^{-\lambda x}, \quad x > 0$ and so Equation 5.6.2 is satisfied $(\text{since } e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t})$. Hence, exponentially **distributed** random variables are memoryless (and in fact it can be shown that they are **the only random variables** that are memoryless). EXAMPLE 5.6a Suppose that a number of miles that a car can run before its battery wears out is exponentially distributed with an average value of 10,000 miles. If a person desires to take a 5,000-mile trip, what is the probability that she will be able to complete her trip without having to replace her car battery? What can be said when the distribution is not exponential? SOLUTION It follows, **by** the **memoryless** property of the exponential distribution, that the **remaining** lifetime (in thousands of miles) of the battery is exponential with parameter λ = 1/10. Hence the desired probability is P{remaining lifetime > 5} = 1 – – (5) $= e^{-5\lambda} = e^{-1/2} \approx .604$ However, **if** the **lifetime distribution F is not exponential**, then the relevant probability is $P\{\text{lifetime} > t+5 | \text{lifetime} > t\} = \frac{1-F(t+5)}{1-F(t)}$ where t is the number of miles that the battery had been in use prior to the start of the trip. Therefore, **if** the distribution is **not exponential**, **additional** information is needed (namely, t) **before** the desired probability can be calculated. EXAMPLE 5.6b A crew of workers has 3 interchangeable machines, of which 2 must be working for the crew to do its job. When in use, each machine will function for an exponentially distributed time having parameter λ before breaking down. The workers decide to initially use machines A and B and keep machine C in reserve to replace whichever of A or B breaks down first. They will then be able to continue working until one of the remaining machines breaks down. When the crew is forced to stop working because only one of the machines has not yet broken down, what is the probability that the still operable machine is machine C? SOLUTION This can be easily answered, without any need for computations, **by invoking** the memoryless property of the exponential distribution. The argument is as follows: Consider the moment at which machine C is first put in use. At that time either A or B would have just broken down and the other one — call it machine 0 — will still be functioning. Now even though 0 would have already been functioning for some time, by the memoryless property of the exponential distribution, it follows that its remaining lifetime has the same distribution as that of a machine that is just being put into use. Thus, the remaining lifetimes of machine 0 and machine C have the same distribution and so, by symmetry, the probability that 0 will fail before C is ½. The following proposition presents another property of the exponential distribution. PROPOSITION 5.6.1 If X1, X2, . . . , Xn are independent exponential random variables having **respective** parameters λ1, λ2, . . . , λn, then **min** (X1, X2, . . . , Xn) is exponential with parameter $\sum_{i=1}^{n} \lambda_i$. Proof: **Since** the **smallest** value of a set of numbers is **greater** than x **if and only if** all values are **greater** than x, we have $P\{\min(X_1, X_2, \ldots, X_n) > x\} = P\{X_1 > x, X_2 > x, \ldots, X_n > x\} = \prod_{i=1}^{n} P\{X_i > x\} \quad \text{by independence}$ $= \prod_{i=1}^{n} e^{-\lambda_i x} = e^{-\sum_{i=1}^{n} \lambda_i x}$ EXAMPLE 5.6c A series system is one that needs all of its components to function in order for the system itself to be functional. For an n-component series system in which the component lifetimes are independent exponential random variables with respective parameters λ1, λ2, . . . , λn, what is the probability that the system survives for a time t ? SOLUTION Since the system life is equal to the minimal component life, it follows from Proposition 5.6.1 that $P\{\text{system life exceeds } t\} = e^{-\sum_i \lambda_i t}$ **Another** useful property of exponential random variables is that **cX** is exponential **with** parameter **λ/c when X** is exponential **with** parameter **λ**, and c > 0. This follows since $P\{cX \le x\} = P\{X \le x/c\} = 1 - e^{-\lambda x/c}$ The parameter **λ** is called the **rate** of the exponential distribution. 5.6.1 The **Poisson Process**: Suppose that "events" are occurring at random time points, and letN(t) denote the number of events that occurs in the time interval [0, t]. These events are said to constitute a Poisson process having rate λ, λ > 0, if

(a) N(0) = 0

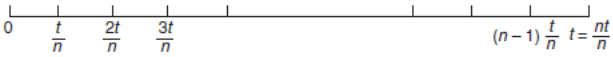
(b) The numbers of events that occur in **disjoint** time intervals are independent.

(clhe distribution of the number of events that occur in a given **interval depends** only on the **length** of the interval and **not** on its **location**.

(d) $\lim_{h \to 0} \frac{P\{N(h) = 1\}}{h} = \lambda$

(e) $\lim_{h \to 0} \frac{P\{N(h) \ge 2\}}{h} = 0$

Thus, Condition (a) states that the process begins at time 0. Condition (b), the **independent increment assumption**, states for instance that the number of events by time t [that is, N(t)] is independent of the number of events that occurs between t and t + s [that is, N(t + s) – –(t)]. Condition (c), the **stationary increment assumption**, states that probability distribution of N(t + s) – –(t) is the same for all values of t . Conditions (d) and (eItate that in a small interval of length h, the probability of one event occurring is approximately λh, whereas the probability of 2 or more is approximately 0. We will now show that **these assumptions imply** that the number of events occurring in any interval of length t is a Poisson random variable with parameter λt . To be precise, let us call the interval [0, t] and denote by N(t) the number of events occurring in that interval. To obtain an expression for P{N(t) = k}, we start by **breaking** the interval [0, t] into n **nonoverlapping subintervals** each of length t /n



(Figure 5.10). FIGURE 5.10

Now there will be k events in [0, t] if either (i) N(t) equals k and there is **at most one** event in each subinterval; (ii) N(t) equals k and **at least one** of the subintervals **contains 2 or more** events. Since these two possibilities are clearly **mutually exclusive**, and since Condition (i) is equivalent to the statement that k of the n subintervals contain exactly **1** event and the other n – – contain **0** events, we have that P{N(t) = k} = P{k of the n subintervals contain exactly 1 event (5.6.3) and the other n – – contain 0 events} + P{N(t) = k and at least 1 subinterval contains 2 or more events} Now it can be shown, using Condition (e), that P{N(t) = k and at least 1 subinterval contains 2 or more events} –→ 0 as n→∞ (5.6.4). Also, it follows from Conditions (d) and (eIhat P{exactly 1 event in a subinterval} ≈ λt/n P{0 events in a subinterval} ≈ 1 – λt/n Hence, since the **numbers** of events that occur in **different** subintervals are independent [from Condition (b)], it follows that P{k of the subintervals contain exactly 1 event and the other n – – contain 0 events} $\approx \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k}$ with the **approximation** becoming exact as the number of subintervals, n, goes to ∞.

However, the probability in Equation 5.6.5 is just the probability that a **binomial** random variable with parameters n and p = λt /n equals k. Hence, as **n** becomes **larger and larger**, this **approaches** the probability that a **Poisson** random variable with mean nλt /n = λt equals k. Hence, from Equations 5.6.3, 5.6.4, and 5.6.5, we see upon letting n approach ∞ that $P\{N(t) = k\} = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$ PROPOSITION 5.6.2 For a Poisson process having rate λ $P\{N(t) = k\} = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad k = 0, 1, \ldots$ That is, the number of events in any interval of length t has a Poisson distribution with mean λt .

For a Poisson process, let X1 denote the time of the first event. Further, for n > 1, let Xn denote the elapsed time between (n – –)st and the nth events. The sequence {Xn, n = 1, 2, . . .} is called the **sequence of interarrival times**. For instance, if X1 = 5 and X2 = 10, then the **first** event of the **Poisson** process would have occurred at time 5 and the **second** at time 15. We now determine the distribution of the Xn. To do so, we first note that the event {X1 > t } takes **place if and only if no** events of the Poisson process occur in the interval [0, t] and thus, $P\{X_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}$ Hence, **X1** has an **exponential** distribution with mean 1/λ. To obtain the distribution of X2, note that $P\{X_2 > t | X_1 = s\} = P\{0 \text{ events in } (s, s+t] | X_1 = s\} = P\{0 \text{ events in } (s, s+t]\} = e^{-\lambda t}$ where **the last two equations followed** from **independent** and **stationary** increments. Therefore, from the foregoing we conclude that X2 is also an exponential random variable with mean 1/λ, and furthermore, that X2 is independent of X1. **Repeating** the same argument yields: PROPOSITION 5.6.3 X1, X2, . . . are independent **exponential** random variables each with mean 1/λ.

5.7 THE GAMMA DISTRIBUTION : A random variable is said to have a gamma distribution with parameters (α, λ),λ > 0, α > 0, if its density function is given by

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$ where $$\Gamma(\alpha) = \int_0^\infty \lambda e^{-\lambda x}(\lambda x)^{\alpha-1}\,dx = \int_0^\infty e^{-y} y^{\alpha-1}\,dy \quad \text{(by letting } y = \lambda x\text{)}$$ The integration by parts formula

$\int u\,dv = uv - \int v\,du$ yields, with $u = y^{\alpha-1}$, $dv = e^{-y}dy$, $v = -e^{-y}$, that for $\alpha > 1$ $\int_0^\infty e^{-y}y^{\alpha-1}\,dy = -e^{-y}y^{\alpha-1}\Big|_{y=0}^{y=\infty} + \int_0^\infty e^{-y}(\alpha-1)y^{\alpha-2}\,dy$

$= (\alpha-1)\int_0^\infty e^{-y}y^{\alpha-2}\,dy$ or $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ (5.7.1) When α is an integer— say, α = n — we can iterate the foregoing to obtain that

$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2)$ by letting α = n – – in Eq. 5.7.1 $= (n-1)(n-2)(n-3)\Gamma(n-3)$ by letting α = n – – in Eq. 5.7.1 ..... $= (n-1)!\Gamma(1)$

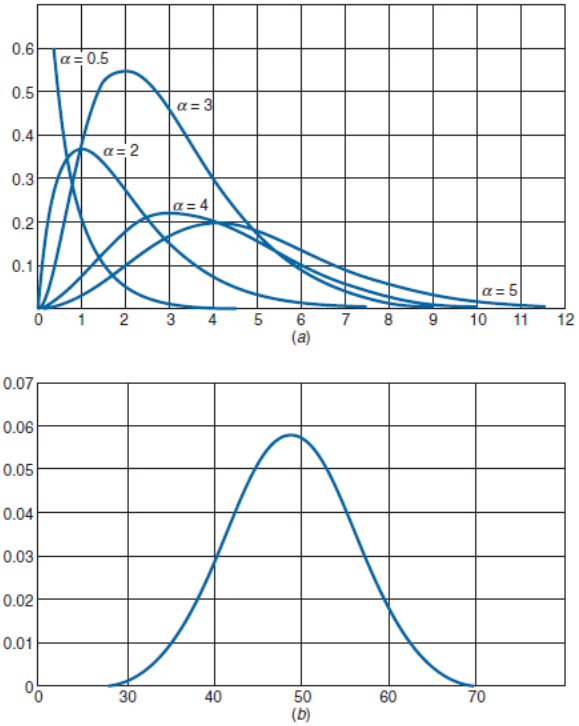
Because $\Gamma(1) = \int_0^\infty e^{-y}\,dy = 1$ we see that $\Gamma(n) = (n-1)!$ The function $\Gamma(\alpha)$ is called the gamma function. It should be noted that when $\alpha = 1$, the gamma distribution reduces to the exponential with mean $1/\lambda$. The moment generating function of a gamma random variable X with parameters (α, λ) is obtained as follows: $\phi(t) = E[e^{tX}] = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty e^{tx}e^{-\lambda x}x^{\alpha-1}\,dx = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty e^{-(\lambda-t)x}x^{\alpha-1}\,dx = \left(\frac{\lambda}{\lambda-t}\right)^\alpha \frac{1}{\Gamma(\alpha)}\int_0^\infty e^{-y}y^{\alpha-1}\,dy \quad [\text{by } y = (\lambda-t)x] = \left(\frac{\lambda}{\lambda-t}\right)^\alpha$

(5.7.2) Differentiation of Equation 5.7.2 yields $\phi'(t) = \frac{\alpha\lambda^\alpha}{(\lambda-t)^{\alpha+1}}$ $\phi''(t) = \frac{\alpha(\alpha+1)\lambda^\alpha}{(\lambda-t)^{\alpha+2}}$ Hence, $E[X] = \phi'(0) = \frac{\alpha}{\lambda}$ (5.7.3) $\text{Var}(X) = E[X^2] - (E[X])^2 = \phi''(0) - \left(\frac{\alpha}{\lambda}\right)^2$

$= \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}$ (5.7.4) An important property of the gamma is that if X1 and X2 are independent gamma random variables having respective parameters (α1, λ) and (α2, λ), then X1 + X2 is a gamma random variable with parameters (α1 + α2, λ). This result easily follows since

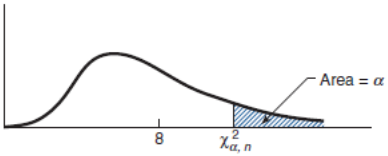
$\phi_{X_1+X_2}(t) = E[e^{t(X_1+X_2)}]$ (5.7.5) $= \phi_{X_1}(t)\phi_{X_2}(t) = \left(\frac{\lambda}{\lambda-t}\right)^{\alpha_1}\left(\frac{\lambda}{\lambda-t}\right)^{\alpha_2}$ from Equation 5.7.2 $= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha_1+\alpha_2}$ which is seen to be the moment generating function of a gamma (α1 + α2, λ) random variable. Since a moment generating function **uniquely** characterizes a distribution, the result entails.

The foregoing result easily generalizes to yield the following proposition. PROPOSITION 5.7.1 If Xi , i = 1, . . . , n are independent gamma random variables with respective parameters (αi , λ), then $\sum_{i=1}^n X_i$ is gamma with parameters $\sum_{i=1}^n \alpha_i, \lambda$. Since the gamma distribution with parameters (1, λ) **reduces** to the exponential with the rate λ, we have thus shown the following useful result. Corollary 5.7.2: **If** X1, . . . , Xn are independent exponential random variables, each having rate λ, **then** $\sum_{i=1}^n X_i$ is a gamma random variable with parameters (n, λ). EXAMPLE 5.7a The **lifetime** of a battery is exponentially distributed with rate λ. **If** a stereo cassette requires one battery to operate, **then** the total playing time one can obtain from a total of n batteries is a gamma random variable with parameters (n, λ). Figure 5.11 presents a



graph of the gamma (α, 1) density for a variety of values of α. FIGURE 5.11 *Graphs of the gamma (α, 1) density for (a) α = .5, 2, 3, 4, 5 and (b) α = 50.* It should be noted that as α becomes **large**, the density starts to **resemble** the normal density. This is theoretically explained by the **central limit theorem**. 5.8 DISTRIBUTIONS **ARISING** FROM THE NORMAL: 5.8.1 The **Chi-Square** Distribution: Definition: If Z1, Z2, . . . , Zn are independent standard normal random variables, then X , defined by

$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ (5.8.1) is said to have a chi-square distribution with **n** degrees of freedom. We will use the notation $X \sim \chi_n^2$ to signify that X has a chi-square distribution with **n** degrees of freedom. The chi-square distribution has the **additive** property that if X1 and X2 are independent chi-square random variables with n1 and n2 degrees of freedom, respectively, then X1+X2 is chi-square with n1 + n2 degrees of freedom. This can be formally shown either by the use of moment generating functions or, most easily, by noting that X1 + X2 is the sum of squares of n1 + n2 independent standard no25ormalnd thus has a chi-square distribution with **n1 + n2** degrees of freedom. If **X** is a chi-square random variable with **n** degrees of freedom, then for any α ∈ (0, 1), the quantity $\chi^2_{\alpha,n}$ is defined to be such that $P\{X \geq \chi^2_{\alpha,n}\} = \alpha$



This is illustrated in Figure 5.12. FIGURE 5.12 *The chi-square density function with 8 degrees of freedom.* In Table A2 of the Appendix, we list $\chi^2_{\alpha,n}$ for a variety of values of α and n (including all those needed to solve problems and examples in this text). In addition, Programs 5.8.1a and 5.8.1b on the text disk can be used to obtain chi-square probabilities and the values of $\chi^2_{\alpha,n}$. 5.8.1.1 THE **RELATION** BETWEEN CHI-SQUARE AND GAMMA RANDOM VARIABLES: Let us compute the moment generating function of a chi-square random variable with n degrees of freedom. To begin, we have, when n = 1, that $E[e^{tX}] = E[e^{tZ^2}]$ where $Z \sim \mathcal{N}(0,1)$ (5.8.2) $= \int_{-\infty}^{\infty} e^{tx^2} f_Z(x)\,dx$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2(1-2t)/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2\bar{\sigma}^2}\, dx \quad \text{where } \bar{\sigma}^2 = (1-2t)^{-1} = (1-2t)^{-1/2} \frac{1}{\sqrt{2\pi}\bar{\sigma}} \int_{-\infty}^{\infty} e^{-x^2/2\bar{\sigma}^2}\, dx = (1-2t)^{-1/2}$ where the last equality follows since the integral of the normal $(0, \bar{\sigma}^2)$ density equals 1. Hence, in the general case of n degrees of freedom $E[e^{tX}] = E\left[e^{t\sum_{i=1}^{n} Z_i^2}\right]$ $= E\left[\prod_{i=1}^{n} e^{tZ_i^2}\right] = \prod_{i=1}^{n} E[e^{tZ_i^2}]$ by independence of the Z_i $= (1-2t)^{-n/2}$ from Equation 5.8.2 However, we recognize [1/(1 − −t)]n/2 as being the moment generating function of a gamma random variable with parameters (n/2, 1/½ Hence, by the **uniqueness** of moment generating functions, it follows that **these two distributions** — **chi-square** with **n** degrees of freedom and **gamma** with **parameters n/2 and 1/½** are identical, and thus we can **conclude** that the **density** of X is given by

$$f(x) = \frac{\frac{1}{2} e^{-x/2} \left(\frac{x}{2}\right)^{(n/2)-1}}{\Gamma\left(\frac{n}{2}\right)}, \quad x > 0$$

The chi-square density functions having 1, 3, and 10 degrees of freedom, respectively, are plotted in Figure 5.13.

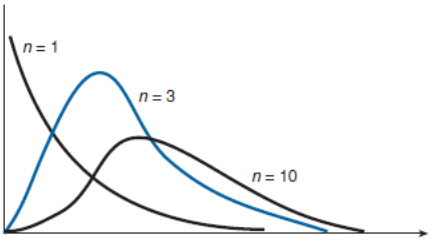


FIGURE 5.13 *The chi-square density function with n degrees of freedom.*

Let us reconsider Example 5.8c, this time supposing that the target is located in the two-dimensional plane. SOLUTION If D is the distance and Xi , i = 1, 2 are the coordinate errors, then $D^2 = X_1^2 + X_2^2$. Since, $Z_i = X_i/2$, i = 1, 2, are standard normal random variables, we obtain $P\{D^2 > 9\} = P\{Z_1^2 + Z_2^2 > 9/4\} = P\{\chi_2^2 > 9/4\} = e^{-9/8} \approx .3247$ where the preceding calculation used the fact that the chi-square distribution with 2 degrees of freedom is the same as the exponential distribution with parameter 1/½Since the chi-square distribution with n degrees of freedom is identical to the gamma distribution with parameters α = n/2 and λ = 1/½it follows from Equations 5.7.3 and 5.7.4 that the mean and variance of a random variable X having this distribution is E[X] = n, Var(X) = 2n **5.8.2 The t-Distribution** : **If** Z and χ_n^2 are independent random variables, with Z having a standard normal distribution and χ_n^2 having a chi-square distribution with **n** degrees of freedom, then the random variable **Tn** defined by $T_n = \frac{Z}{\sqrt{\chi_n^2/n}}$ is said to have a t-distribution with **n** degrees of freedom. A graph of the density function of Tn is given in Figure 5.14 for n = 1, 5, and 10.

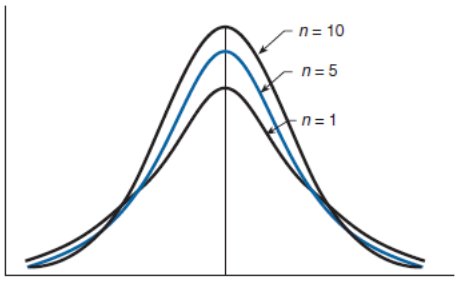


FIGURE 5.14 *Density function of T_n.*

Like the standard normal density, the t-density is symmetric about zero. In addition, as n becomes larger, it becomes more and more like a standard normal density. To understand why, recall that χ_n^2 can be expressed as the sum of the squares of n standard no26ormaland so $\frac{\chi_n^2}{n} = \frac{Z_1^2 + \cdots + Z_n^2}{n}$ where Z1, . . . , Zn are independent standard normal random variables. It now follows from the weak law of large numbers that, for large n, χ_n^2/n will, with probability close to 1, be approximately equal to $E[Z_i^2]$ = 1. Hence, for n large, $T_n = Z/\sqrt{\chi_n^2/n}$ will have approximately the same distribution as Z. Figure 5.15 shows a graph of the t -d–nsity function with 5 degrees of freedom compared with the standard normal density. Notice that the t -d–nsity has thicker “tails,” indicating greater variability, than does the normal density.

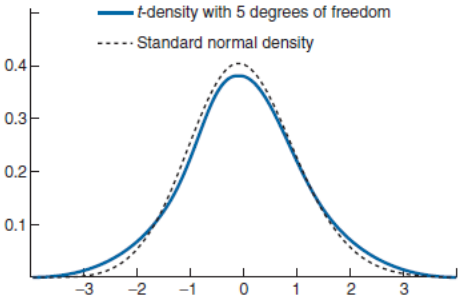
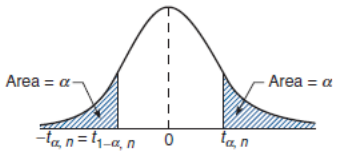


FIGURE 5.15 *Comparing standard normal density with the density of T_5.*

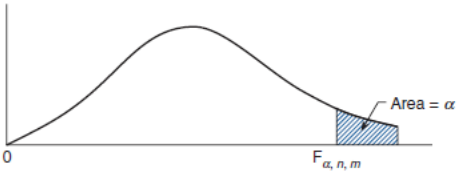
The mean and variance of Tn can be shown to equal $E[T_n] = 0, \qquad n > 1$ $\text{Var}(T_n) = \frac{n}{n-2}, \qquad n > 2$ Thus the variance of Tn decreases to 1 — the variance of a standard normal random variable— as n increases to∞. For α, 0 < α < 1, let $t_{\alpha,n}$ be such that $P\{T_n \geq t_{\alpha,n}\} = \alpha$ It follows from the symmetry about zero of the t -d–nsity function that −T– has the same distribution as Tn, and so $\alpha = P\{-T_n \geq t_{\alpha,n}\}$ $= P\{T_n \leq -t_{\alpha,n}\} \quad = 1 - P\{T_n > -t_{\alpha,n}\}$ Therefore, $P\{T_n \geq -t_{\alpha,n}\} = 1 - \alpha$ leading to the conclusion that $-t_{\alpha,n} = t_{1-\alpha,n}$ which is illustrated in Figure 5.16. The

values of $t_{\alpha,n}$ for a variety of values of n and α have been tabulated in **Table A3** in the **Appendix**. In addition, Programs 5.8.2a and 5.8.2b on the text disk compute the t -



d–stribution function and the values $t_{\alpha,n}$ respectively. FIGURE 5.16 $t_{1-\alpha,n} = -t_{\alpha,n}$.

5.8.3 The F-Distribution: If χ^2_n and χ^2_m are independent chi-square random variables with n and m degrees of freedom, respectively, **then** the random variable **Fn,m** defined by $F_{n,m} = \dfrac{\chi^2_n/n}{\chi^2_m/m}$ is said to have an F-distribution with **n and m degrees** of freedom. For any α ∈ (0, 1), let Fα,n,m be such that $P\{F_{n,m} > F_{\alpha,n,m}\} = \alpha$ This is



illustrated in Figure 5.17. FIGURE 5.17 *Density function of* $F_{n,m}$. The quantities Fα,n,m are tabulated in **Table A4 of the Appendix** for

$$\alpha = P\left\{\frac{\chi^2_n/n}{\chi^2_m/m} > F_{\alpha,n,m}\right\}$$

different values of n,m, and α ≤ ½. If Fα,n,m is desired when α > 1/½ it can be obtained by using the following equalities:

$$= P\left\{\frac{\chi^2_m/m}{\chi^2_n/n} < \frac{1}{F_{\alpha,n,m}}\right\} = 1 - P\left\{\frac{\chi^2_m/m}{\chi^2_n/n} \geq \frac{1}{F_{\alpha,n,m}}\right\}$$ or, equivalently, $$P\left\{\frac{\chi^2_m/m}{\chi^2_n/n} \geq \frac{1}{F_{\alpha,n,m}}\right\} = 1 - \alpha$$ (5.8.3) But because $(\chi^2_m/m)/(\chi^2_n/n)$ has an F -d–stribution with

$$1 - \alpha = P\left\{\frac{\chi^2_m/m}{\chi^2_n/n} \geq F_{1-\alpha,m,n}\right\}$$

degrees of freedom m and n, it follows that implying, from Equation 5.8.3, that $\dfrac{1}{F_{\alpha,n,m}} = F_{1-\alpha,m,n}$ Thus, for instance, F.9,5,7 = 1/F.1,7,5 = 1/3.37 = .2967 where the value of F.1,7,5 was obtained from **Table A4 of the Appendix**. Program 5.8.3 computes the distribution function of Fn,m**. 5.9 THE LOGISTICS DISTRIBUTION:** A random variable X is said to have a **logistics** distribution with parameters μ and v > 0 **if** its **distribution** function is

$$F(x) = \frac{e^{(x-\mu)/v}}{1 + e^{(x-\mu)/v}}, \quad -\infty < x < \infty$$ Differentiating $F(x) = 1 - 1/(1 + e^{(x-\mu)/v})$ yields the density function $$f(x) = \frac{e^{(x-\mu)/v}}{v(1 + e^{(x-\mu)/v})^2}, \quad -\infty < x < \infty$$ To obtain

the **mean** of a logistics random variable, $$E[X] = \int_{-\infty}^{\infty} x\frac{e^{(x-\mu)/v}}{v(1 + e^{(x-\mu)/v})^2}\,dx$$ make the substitution y = (x – –)/v. This yields

$$E[X] = v\int_{-\infty}^{\infty} \frac{ye^y}{(1 + e^y)^2}\,dy + \mu\int_{-\infty}^{\infty} \frac{e^y}{(1 + e^y)^2}\,dy = v\int_{-\infty}^{\infty} \frac{ye^y}{(1 + e^y)^2}\,dy + \mu$$ (5.9.1) where the preceding equality used that $e^y/((1 + e^y)^2)$ is the density function of a logistic random variable with parameters μ = 0, v = 1 (such a random variable is called a **standard logistic**) and thus integrates to 1. Now,

$$\int_{-\infty}^{\infty} \frac{ye^y}{(1 + e^y)^2}\,dy = \int_{-\infty}^{0} \frac{ye^y}{(1 + e^y)^2}\,dy + \int_{0}^{\infty} \frac{ye^y}{(1 + e^y)^2}\,dy = -\int_{0}^{\infty} \frac{xe^{-x}}{(1 + e^{-x})^2}\,dx + \int_{0}^{\infty} \frac{ye^y}{(1 + e^y)^2}\,dy = -\int_{0}^{\infty} \frac{xe^x}{(e^x + 1)^2}\,dx + \int_{0}^{\infty} \frac{ye^y}{(1 + e^y)^2}\,dy$$ = 0 (5.9.2)

where the second equality is obtained by making the substitution x = −y, and the third by multiplying the numerator and denominator by e^{2x}. From Equations 5.9.1 and 5.9.2 we obtain E[X] = μ Thus μ is the **mean** of the logistic**; v is called the dispersion parameter**.

Chapter_4: RANDOM VARIABLES AND EXPECTATION: 4.1 RANDOM VARIABLES: When a random experiment is performed, we are often not interested in all of the details of the experimental result but only in the value of some numerical quantity determined by the result. For instance, in tossing dice we are often interested in the sum of the two dice and are not really concerned about the values of the individual dice. That is, we may be interested in knowing that the sum is 7 and not be concerned over whether the actual outcome was (1, 6) or (2, 5) or (3, 4) or (4, 3) or (5, 2) or (6, 1). Also, a civil engineer may not be directly concerned with the daily risings and declines of the water level of a reservoir (which we can take as the experimental result) but may only care about the level at the end of a rainy season. These quantities of interest that are determined by the result of the experiment are known as random variables. Since the value of a random variable is determined by the outcome of the experiment, we may assign probabilities of its possible values. EXAMPLE 4.1b Suppose that an individual purchases two electronic components each of which may be either defective or acceptable. In addition, suppose that the four possible results — (d, d), (d, a), (a, d), (a, a) — have respective probabilities .09, .21, .21, .49 [where (d, d) means that both components are defective, (d, a) that the first component is defective and the second acceptable, and so on]. If we let X denote the number of acceptable components obtained in the purchase, then X is a random variable taking on one of the values 0, 1, 2 with respective probabilities. P{X = 0} = .09 P{X = 1} = .42 P{X = 2} = .49 If we were mainly concerned with whether there was at least one acceptable component, we could define the random variable I by

$$I = \begin{cases} 1 & \text{if } X = 1 \text{ or } 2 \\ 0 & \text{if } X = 0 \end{cases}$$ If A denotes the event that at least one acceptable component is obtained, then the random variable I is called the indicator random variable for the event A, since I will equal 1. or 0 depending upon whether A occurs. The probabilities attached to the possible values of I are P{I =1} = .91 P{I =0} = .09 In the two foregoing examples, the random variables of interest took on a finite number of possible values. Random variables whose set of possible values can be written either as a finite sequence x1, . . . , xn, or as an infinite sequence x1, . . . are said to be discrete. For instance, a random variable whose set of possible values is the set of nonnegative integers is a discrete random variable. However, there also exist random variables that take on a continuum of possible values. These are known as continuous random variables. One example is the random variable denoting the lifetime of a car, when the car's lifetime is assumed to take on any value in some interval (a, b). The cumulative distribution function, or more simply the distribution function, F of the random variable X is defined for any real number x by F (x) = P{X ≤ x} That is, F (x) is the probability that the random variable X takes on a value that is less than or equal to x. Notation: We will use the notation X ∼ F to signify that F is the distribution function of X. All probability questions about X can be answered in terms of its distribution function F . For example, suppose we wanted to compute P{a < X ≤ b}. This can be accomplished

by first noting that the event {X ≤ b} can be expressed as the union of the two mutually exclusive events {X ≤ a} and {a < X ≤ b}. Therefore, applying Axiom 3, we obtain that P{X ≤ b} = P{X ≤ a} + P{a < X ≤ b} or P{a < X ≤ b} = F (b) − F (a) 4.2 TYPES OF RANDOM VARIABLES: As was previously mentioned, a random variable whose set of possible values is a sequence is said to be discrete. For a discrete random variable X , we define the probability mass function p(a) of X by p(a) = P{X = a} The probability mass function p(a) is positive for at most a countable number of values of a. That is, if X must assume one of the values x1, x2, . . . , then

p(xi) > 0, i = 1, 2, . . . p(x) = 0, all other values of x Since X must take on one of the values xi , we have $\sum_{i=1}^{\infty} p(x_i) = 1$ The cumulative distribution function F can be expressed in terms of p(x) by $F(a) = \sum_{\text{all } x \le a} p(x)$ If X is a discrete random variable whose set of possible values are x1, x2, x3, . . ., where x1 < x2 < x3 < · · · , then its distribution function F is a step function. That is, the value of F is constant in the intervals [xi−1, xi) and then takes a step (or jump) of size p(xi) at xi.

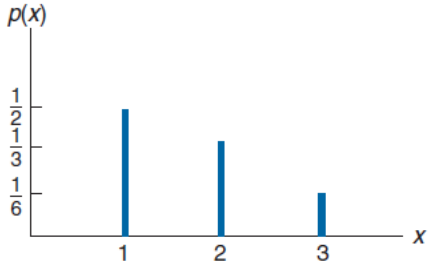


FIGURE 4.1 *Graph of (p)x, Example 4.2a.*

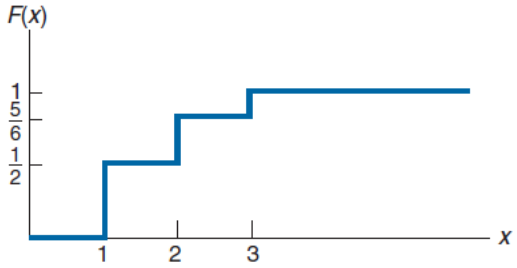


FIGURE 4.2 *Graph of F(x).*

For instance, suppose X has a probability mass function given (as in Example 4.2a) by p(1) = ½ , p(2) = 1/3 , p(3) = 1/6 then the cumulative distribution function F of X is given by

$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{2} & 1 \le a < 2 \\ \frac{5}{6} & 2 \le a < 3 \\ 1 & 3 \le a \end{cases}$$

This is graphically presented in Figure 4.2. Whereas the set of possible values of a discrete random variable is a sequence, we often must consider random variables whose set of possible values is an interval. Let X be such a random variable. We say that X is a continuous random variable if there exists a nonnegative function f (x), defined for all real x ∈ (−∞,∞), having the property that for any set B of real numbers $P\{X \in B\} = \int_B f(x)\,dx$ (4.2.1) The function f (x) is called the probability density function of the random variable X. In words, Equation 4.2.1 states that the probability that X will be in B may be obtained by integrating the probability density function over the set B. Since X must assume some value, f (x) must satisfy $1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\,dx$ All probability statements about X can be answered in terms of f (x).

For instance, letting B = [a, b], we obtain from Equation 4.2.1 that $P\{a \le X \le b\} = \int_a^b f(x)\,dx$ (4.2.2) If we let a = b in the above, then

$$P\{X = a\} = \int_a^a f(x)\,dx = 0$$

In words, this equation states that the probability that a continuous random variable will assume any particular value is zero. (See Figure 4.3.) The relationship between the cumulative distribution F (·) and the probability density f (·) is expressed by $F(a) = P\{X \in (-\infty, a]\} = \int_{-\infty}^{a} f(x)\,dx$ Differentiating both sides yields $\frac{d}{da}F(a) = f(a)$

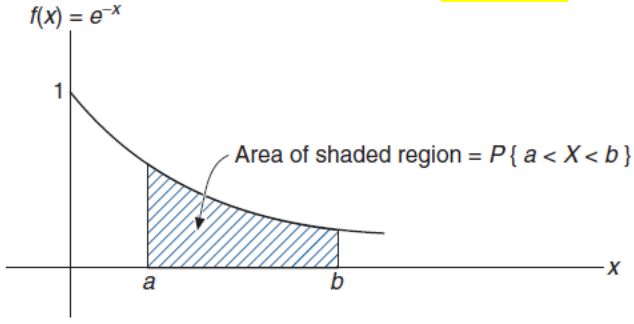


FIGURE 4.3 *The probability density function* $f(x) = \begin{cases} e^{-x} & x \ge 0 \\ 0 & x < 0 \end{cases}$.

That is, the density is the derivative of the cumulative distribution function. A somewhat more intuitive interpretation of the density function may be obtained from Equation 4.2.2 as follows:

$$P\left\{a - \frac{\varepsilon}{2} \le X \le a + \frac{\varepsilon}{2}\right\} = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)\,dx \approx \varepsilon f(a)$$

when ε is small. In other words, the probability that X will be contained in an interval

of length ε around the point a is approximately εf (a). From this, we see that f (a) is a measure of how likely it is that the random variable will be near a. 4.3 JOINTLY DISTRIBUTED RANDOM VARIABLES: For a given experiment, we are often interested not only in probability distribution functions of individual random variables but also in the relationships between two or more random variables. For instance, in an experiment into the possible causes of cancer, we might be interested in the relationship between the average number of cigarettes smoked daily and the age at which an individual contracts cancer. Similarly, an engineer might be interested in the relationship between the shear strength and the diameter of a spot weld in a fabricated sheet steel specimen. To specify the relationship between two random variables, we define the joint cumulative probability distribution function of X and Y by $F(x, y) = P\{X \leq x, Y \leq y\}$ A knowledge of the joint probability distribution function enables one, at least in theory, to compute the probability of any statement concerning the values of X and Y . For instance, the distribution function of X —call it F_X —can be obtained from the joint distribution function F of X and Y as follows:

$$\begin{aligned} F_X(x) &= P\{X \leq x\} \\ &= P\{X \leq x, Y < \infty\} \\ &= F(x, \infty) \end{aligned}$$

Similarly, the cumulative distribution function of Y is given by $F_Y(y) = F(\infty, y)$ In the case where X and Y are both discrete random variables whose possible values are, respectively, x1, x2, . . ., and y1, y2, . . ., we define the joint probability mass function of X and Y , p(xi , yj), by $p(x_i, y_j) = P\{X = x_i, Y = y_j\}$ The individual probability mass functions of X and Y are easily obtained from the joint probability mass function by the following reasoning. Since Y must take on some value yj, it follows that the event {X = xi } can be written as the union, over all j, of the mutually exclusive events {X = xi , Y = yj }. That is,

$$\{X = x_i\} = \bigcup_j \{X = x_i, Y = y_j\}$$

$$P\{X = x_i\} = P\left(\bigcup_j \{X = x_i, Y = y_j\}\right)$$

and so, using Axiom 3 of the probability function, we see that

$$= \sum_j P\{X = x_i, Y = y_j\} = \sum_j p(x_i, y_j)$$

Similarly, we can obtain P{Y = yj } by summing p(xi , yj) over all possible values of xi, that is,

$$P\{Y = y_j\} = \sum_i P\{X = x_i, Y = y_j\} = \sum_i p(x_i, y_j) \quad (4.3.2)$$

Hence, specifying the joint probability mass function always determines the individual mass functions. However, it should be noted that the reverse is not true. Namely, knowledge of P{X = xi } and P{Y = yj } does not determine the value of P{X = xi , Y = yj }. EXAMPLE 4.3a Suppose that 3 batteries are randomly chosen from a group of 3 new, 4 used but still working, and 5 defective batteries. If we let X and Y denote, respectively, the number of new and used but still working batteries that are chosen, then the joint probability mass function of X and Y , p(i, j) = P{X = i, Y = j}, is given by

$$p(0,0) = \binom{5}{3} \Big/ \binom{12}{3} = 10/220$$

$$p(0,1) = \binom{4}{1}\binom{5}{2} \Big/ \binom{12}{3} = 40/220$$

$$p(0,2) = \binom{4}{2}\binom{5}{1} \Big/ \binom{12}{3} = 30/220$$

$$p(0,3) = \binom{4}{3} \Big/ \binom{12}{3} = 4/220$$

$$p(1,0) = \binom{3}{1}\binom{5}{2} \Big/ \binom{12}{3} = 30/220$$

$$p(1,1) = \binom{3}{1}\binom{4}{1}\binom{5}{1} \Big/ \binom{12}{3} = 60/220$$

$$p(1,2) = \binom{3}{1}\binom{4}{2} \Big/ \binom{12}{3} = 18/220$$

$$p(2,0) = \binom{3}{2}\binom{5}{1} \Big/ \binom{12}{3} = 15/220$$

$$p(2,1) = \binom{3}{2}\binom{4}{1} \Big/ \binom{12}{3} = 12/220$$

$$p(3,0) = \binom{3}{3} \Big/ \binom{12}{3} = 1/220$$

These probabilities can most easily be expressed in tabular form as shown in Table 4.1.

TABLE 4.1 $P\{X = i, Y = j\}$

i \ j	0	1	2	3	Row Sum $= P\{X = i\}$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
Column Sums $= P\{Y = j\}$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

The reader should note that the probability mass function of X is obtained by computing the row sums, in accordance with the Equation 4.3.1, whereas the probability mass function of Y is obtained by computing the column sums, in accordance with Equation 4.3.2. Because the individual probability mass functions of X and Y thus appear in the margin of such a table, they are often referred to as being the marginal probability mass functions of X and Y , respectively. It should be noted that to check the correctness of such a table we could sum the marginal row (or the marginal column) and verify that its sum is 1. (Why must the sum of the entries in the marginal row (or column) equal 1?) These probabilities are obtained as follows:

$$P\{B = 0, G = 1\} = P\{1 \text{ girl and total of 1 child}\}$$

$$P\{B = 0, G = 0\} = P\{\text{no children}\} = .15$$

$$= P\{1 \text{ child}\}P\{1 \text{ girl}|1 \text{ child}\} = (.20)\left(\tfrac{1}{2}\right) = .1$$

$$P\{B = 0, G = 2\} = P\{2 \text{ girls and total of 2 children}\} = P\{2 \text{ children}\}P\{2 \text{ girls}|2 \text{ children}\} = (.35)\left(\tfrac{1}{2}\right)^2 = .0875$$

$$P\{B = 0, G = 3\} = P\{3 \text{ girls and total of 3 children}\} = P\{3 \text{ children}\}P\{3 \text{ girls}|3 \text{ children}\} = (.30)\left(\tfrac{1}{2}\right)^3 = .0375$$

We leave it to the reader to verify the remainder of Table 4.2, which tells us, among other things, that the family chosen will have at least 1 girl with probability .625. We say that X and Y are jointly continuous if there exists a function f (x, y) defined for all real x and y, having the property that for every set C of pairs of real numbers (that is, C is a set in the two-dimensional plane)

$$P\{(X, Y) \in C\} = \iint_{(x,y)\in C} f(x, y)\, dx\, dy \quad (4.3.3)$$

The function f (x, y) is called the joint probability density function of X and Y . If A and B are any sets of real numbers, then by defining C = {(x, y) : x ∈ A, y ∈ B}, we see from Equation 4.3.3 that

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x, y)\, dx\, dy \quad (4.3.4)$$

Because $F(a, b) = P\{X \in (-\infty, a], Y \in (-\infty, b]\} = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y)\, dx\, dy$ it follows, upon differentiation, that

$$f(a, b) = \frac{\partial^2}{\partial a\, \partial b} F(a, b)$$

wherever the partial derivatives are defined. Another interpretation of the joint density function is obtained from Equation 4.3.4 as follows:

$$P\{a < X < a + da, b < Y < b + db\} = \int_b^{d+db} \int_a^{a+da} f(x, y)\, dx\, dy \approx f(a, b)da\, db$$

when da and db are small and f (x, y) is continuous at a, b. Hence f (a, b) is a measure of how likely it is that the random vector (X , Y) will be near (a, b). If X and Y are jointly continuous, they are individually continuous, and their probability density functions can be obtained as follows:

$$P\{X \in A\} = P\{X \in A, Y \in (-\infty, \infty)\} = \int_A \int_{-\infty}^{\infty} f(x, y)\, dy\, dx \quad (4.3.5)$$

$$= \int_A f_X(x)\, dx \qquad f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

where is thus the probability density function of X . Similarly, the probability density function of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \quad (4.3.6)$$

EXAMPLE 4.3c The joint density function of X and Y is given by

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Compute (a) P{X > 1, Y < 1}; (b) P{X < Y }; and (c) P{X < a}. SOLUTION (a)

$$P\{X > 1, Y < 1\} = \int_0^1 \int_1^{\infty} 2e^{-x}e^{-2y}\, dx\, dy = e^{-1}\int_0^1 2e^{-2y}\, dy = e^{-1}(1 - e^{-2})$$

(b)

$$P\{X < Y\} = \iint_{(x,y):x<y} 2e^{-x}e^{-2y}\, dx\, dy = \int_0^{\infty} \int_0^{y} 2e^{-x}e^{-2y}\, dx\, dy = \int_0^{\infty} 2e^{-2y}(1 - e^{-y})\, dy = \int_0^{\infty} 2e^{-2y}\, dy - \int_0^{\infty} 2e^{-3y}\, dy = 1 - \tfrac{2}{3} = \tfrac{1}{3}$$

(c)

$$P\{X < a\} = \int_0^a \int_0^{\infty} 2e^{-2y}e^{-x}\, dy\, dx = \int_0^a e^{-x}\, dx = 1 - e^{-a}$$

4.3.1 Independent Random Variables: The random variables X and Y are said to be independent if for any two sets of real numbers A and B

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \quad (4.3.7)$$

In other words, X and Y are independent if, for all A and B, the events E_A = {X ∈ A} and F_B = {Y ∈ B} are independent. It can be shown by using the three axioms of probability that Equation 4.3.7 will follow if and only if for all a, b P{X ≤ a, Y ≤ b} = P{X ≤ a}P{Y ≤ b} Hence, in terms of the joint distribution function F of X and Y , we have that X and Y are independent if

$$F(a, b) = F_X(a)F_Y(b) \text{ for all } a, b$$

When X

and Y are discrete random variables, the condition of independence Equation 4.3.7 is equivalent to $p(x, y) = p_X(x)p_Y(y) \quad \text{for all } x, y$ (4.3.8) where p_X and p_Y are the probability mass functions of X and Y . The equivalence follows because, if Equation 4.3.7 is satisfied, then we obtain Equation 4.3.8 by letting A and B be, respectively, the one-point sets A = {x}, B = {y}. Furthermore, if Equation 4.3.8 is valid, then for any sets A, B

$$P\{X \in A, Y \in B\} = \sum_{y \in B} \sum_{x \in A} p(x, y) = \sum_{y \in B} \sum_{x \in A} p_X(x) p_Y(y) = \sum_{y \in B} p_Y(y) \sum_{x \in A} p_X(x)$$

$$= P\{Y \in B\} P\{X \in A\}$$

and thus Equation 4.3.7 is established. In the jointly continuous case, the condition of independence is equivalent to

$$f(x, y) = f_X(x) f_Y(y) \quad \text{for all } x, y$$

Loosely speaking, X and Y are independent if knowing the value of one does not change the distribution of the other. Random variables that are not independent are said to be dependent. We can also define joint probability distributions for n random variables in exactly the same manner as we did for n = 2. For instance, the joint cumulative probability distribution function F (a1, a2, . . . , an) of the n random variables X1, X2, . . . , Xn is defined by $F(a_1, a_2, \ldots, a_n) = P\{X_1 \le a_1, X_2 \le a_2, \ldots, X_n \le a_n\}$ If these random variables are discrete, we define their joint probability mass function p(x1, x2, . . . , xn) by $p(x_1, x_2, \ldots, x_n) = P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$ Further, the n random variables are said to be jointly continuous if there exists a function f (x1, x2, . . . , xn), called the joint probability density function, such that for any set C in n-space

$$P\{(X_1, X_2, \ldots, X_n) \in C\} = \int \int_{(x_1, \ldots, x_n) \in C} \cdots \int f(x_1, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

In particular, for any n sets of real numbers A1, A2, . . . , An.

$$= \int_{A_n} \int_{A_{n-1}} \cdots \int_{A_1} f(x_1, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

P{X1 ∈ A1, X2 ∈ A2, . . . , Xn ∈ An} The concept of independence may, of course, also be defined for more than two random variables. In general, the n random variables X1, X2, . . . , Xn are said to be independent if, for all sets of real numbers A1, A2, . . . , An,

$$P\{X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\} = \prod_{i=1}^{n} P\{X_i \in A_i\}$$

As before, it can be shown that this condition is equivalent to

$$P\{X_1 \le a_1, X_2 \le a_2, \ldots, X_n \le a_n\} = \prod_{i=1}^{n} P\{X_1 \le a_i\} \quad \text{for all } a_1, a_2, \ldots, a_n$$

Finally, we say that an infinite collection of random variables is independent if every finite subcollection of them is independent. 4.3.2 Conditional Distributions: The relationship between two random variables can often be clarified by consideration ofthe conditional distribution of one given the value of the other. Recall that for any two events E and F , the conditional probability of E given F is defined, provided that P(F) > 0, by $P(E|F) = \frac{P(EF)}{P(F)}$ Hence, if X and Y are discrete random variables, it is natural to define the conditional probability mass function of X given that Y = y, by $p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{P\{X = x, Y = y\}}{P\{Y = y\}} = \frac{p(x, y)}{p_Y(y)}$ for all values of y such that pY (y) > 0. If X and Y have a joint probability density function f (x, y), then the conditional probability density function of X , given that Y = y, is defined for all values of y such that $f_Y(y)$ > 0, by $f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$ To motivate this definition, multiply the left-hand side by dx and the right-hand side by (dx dy)/dy to obtain

$$f_{X|Y}(x|y)\, dx = \frac{f(x, y)\, dx\, dy}{f_Y(y)\, dy} \approx \frac{P\{x \le X \le x + dx, y \le Y \le y + dy\}}{P\{y \le Y \le y + dy\}}$$

$$= P\{x \le X \le x + dy | y \le Y \le y + dy\}$$

In other words, for small values of dx and dy, $f_{X|Y}(x|y)\, dx$ represents the conditional probability that X is between x and x + dx, given that Y is between y and y + dy. The use of conditional densities allows us to define conditional probabilities of events associated with one random variable when we are given the value of a second random variable. That is, if X and Y are jointly continuous, then, for any set A,

$$P\{X \in A|Y = y\} = \int_A f_{X|Y}(x|y)\, dx$$

4.4 EXPECTATION: One of the most important concepts in probability theory is that of the expectation of a random variable. If X is a discrete random variable taking on the possible values x1, x2, . . ., then the expectation or expected value of X , denoted by E[X], is defined by $E[X] = \sum_i x_i P\{X = x_i\}$ In words, the expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it. For instance, if the probability mass function of X is given by $p(0) = \frac{1}{2} = p(1)$ then $E[X] = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}$ Another motivation of the definition of expectation is provided by the frequency interpretation of probabilities. This interpretation assumes that if an infinite sequence of independent replications of an experiment is performed, then for any event E, the proportion of time that E occurs will be P(E). Now, consider a random variable X that must take on one of the values x1, x2, . . . , xn with respective probabilities p(x1), p(x2), . . . , p(xn); and think of X as representing our winnings in a single game of chance. That is, with probability p(xi) we shall win xi units i = 1, 2, . . . , n. Now by the frequency interpretation, it follows that if we continually play this game, then the proportion of time that we win xi will be p(xi). Since this is true for all i, i = 1, 2, . . . , n, it follows that our average winnings per game will be $\sum_{i=1}^{n} x_i p(x_i) = E[X]$ To see this argument more clearly, suppose that we play N games where N is very large. Then in approximately $Np(x_i)$ of these games, we shall win xi , and thus our total winnings in the N games will be $\sum_{i=1}^{n} x_i N_p(x_i)$ implying that our average winnings per game are

$$\sum_{i=1}^{n} \frac{x_i N_p(x_i)}{N} = \sum_{i=1}^{n} x_i p(x_i) = E[X]$$

EXAMPLE 4.4a Find E[X] where X is the outcome when we roll a fair die. SOLUTION Since p(1) = p(2) = p(3) =

p(4) = p(5) = p(6) = 1/6 , we obtain that $E[X] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{7}{2}$ The reader should note that, for this example, the expected value of X is not a value that X could possibly assume. (That is, rolling a die cannot possibly lead to an outcome of 7/2.) Thus, even though we call E[X] the expectation of X , it should not be interpreted as the value that we expect X to have but rather as the average value of X in a large number of repetitions of the experiment. That is, if we continually roll a fair die, then after a large number of rolls the average of all the outcomes will be approximately 7/2. (The interested reader should try this as an experiment.) EXAMPLE 4.4b If I is an indicator random variable for the event A, that is, if

$$I = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases}$$

then $E[I] = 1P(A) + 0P(A^c) = P(A)$ Hence, the expectation of the indicator random variable for the event A is just the probability that A occurs. EXAMPLE 4.4c Entropy For a given random variable X , how much information is conveyed in the message that X = x? Let us begin our attempts at quantifying this statement by agreeing that the amount of information in the message that X = x should depend on how likely it was that X would equal x. In addition, it seems reasonable that the more unlikely it was that X would equal x, the more informative would be the message. For
instance, if X represents the sum of two fair dice, then there seems to be more information in the message that X equals 12 than there would be in the message that X equals 7, since the former event has probability 1/36 and the latter 1/6. Let us denote by I (p) the amount of information contained in the message that an event, whose probability is p, has occurred. Clearly I (p) should be a nonnegative, decreasing function of p. To determine its form, let X and Y be independent random variables, and suppose that P{X = x} = p and P{Y = y} = q. How much information is contained in the message that X equals x and Y equals y? To answer this, note first that the amount of information in the statement that X equals x is I (p). Also, since knowledge of the fact that X is equal to x does not affect the probability that Y will equal y (since X and Y are independent), it seems reasonable that the additional amount of information contained in the statement that Y = y should equal I (q). Thus, it seems that the amount of information in the message that X equals x and Y equals y is I (p)+I (q). On the other hand, however, we have that

$$P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\} = pq$$

which implies that the amount of information in the message that X equals x and Y equals y is I (pq). Therefore, it seems that the function I should satisfy the identity I(pq) = I(p) + I(q) However, if we define the function G by $G(p) = I(2^{-p})$ then we see from the above that $G(p + q) = I(2^{-(p+q)}) = I(2^{-p}2^{-q}) = I(2^{-p}) + I(2^{-q}) = G(p) + G(q)$ However, it can be shown that the only (monotone) functions G that satisfy the foregoing functional relationship are those of the form $G(p) = cp$ for some constant c. Therefore, we must have that $I(2^{-p}) = cp$ or, letting $q = 2^{-p}$ $I(q) = -c \log_2(q)$ for some positive constant c. It is traditional to let c = 1 and to say that the information is measured in units of bits (short for binary digits). Consider now a random variable X , which must take on one of the values x1, . . . , xn with respective probabilities p1, . . . , pn. As log2(pi) represents the information conveyed by the message that X is equal to xi , it follows that the expected amount of information that will be conveyed when the value of X is transmitted is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

The quantity H(X) is known in information theory as the entropy of the random variable X . We can also define the expectation of a continuous random variable. Suppose that X is a continuous random variable with probability density function f . Since, for dx small

$$f(x)\,dx \approx P\{x < X < x + dx\}$$

it follows that a weighted average of all possible values of X , with the weight given to x equal to the probability that X is near x, is just the integral over all x of xf (x) dx. Hence,

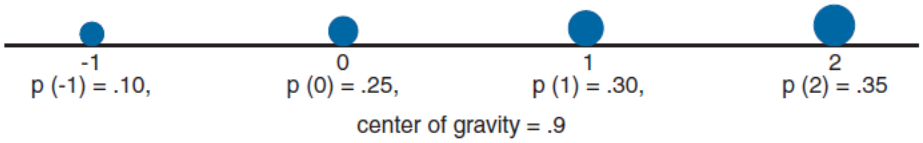


FIGURE 4.4

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx$$

it is natural to define the expected value of X by REMARKS : (a) The concept of expectation is analogous to the physical concept of the center of gravity of a distribution of mass. Consider a discrete random variable X having probability mass function $P(x_i), i \geq 1$. If we now imagine a weightless rod in which weights with mass $P(x_i), i \geq 1$ are located at the points xi , i ≥ 1 (see Figure 4.4), then the point at which the rod would be in balance is known as the center of gravity. For those readers acquainted with elementary statics, it is now a simple matter to show that this point is at E[X]. To prove this, we must show that the sum of the torques tending to turn the point around E[X] is equal to 0. That is, we must show that 0 = i (xi − E[X])p(xi), which is immediate.
(b) E[X] has the same units of measurement as does X .
4.5 PROPERTIES OF THE EXPECTED VALUE: Suppose now that we are given a random variable X and its probability distribution (that is, its probability mass function in the discrete case or its probability density function in the continuous case). Suppose also that we are interested in calculating, not the expected value of X , but the expected value of some function of X , say g (X). How do we go about doing this? One way is as follows. Since g (X) is itself a random variable, it must have a probability distribution, which should be computable from a knowledge of the distribution of X . Once we have obtained the distribution of g (X), we can then compute E[g (X)] by the definition of the expectation. While the foregoing procedure will, in theory, always enable us to compute the expectation
of any function of X from a knowledge of the distribution of X , there is an **easier way** of doing this. Suppose, for instance, that we wanted to compute the expected value of g (X). Since g (X) takes on the value g (X) when X = x, it seems intuitive that E[g (X)] should be a weighted average of the possible values g (X) with, for a given x, the weight given to g (x) being equal to the probability (or probability density in the continuous case) that X will equal x. Indeed, the foregoing can be shown to be true and we thus have the following proposition. PROPOSITION 4.5.1 EXPECTATION OF A FUNCTION OF A RANDOM VARIABLE: (a) If X is a discrete random variable with probability mass function p(x), then for any real-valued function g ,

$$E[g(X)] = \sum_x g(x)p(x)$$

(b) If X is a continuous random variable with probability density function f (x), then for any real-valued function g,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx$$

An immediate corollary of Proposition 4.5.1 is the following. Corollary 4.5.2 If a and b are constants, then E[aX + b] = aE[X] + b Proof: In the discrete case,

$$E[aX + b] = \sum_x (ax + b)p(x) = a\sum_x xp(x) + b\sum_x p(x)$$

$= aE[X] + b$ In the continuous case,

$$E[aX + b] = \int_{-\infty}^{\infty} (ax + b)f(x)\,dx = a\int_{-\infty}^{\infty} xf(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx = aE[X] + b$$

If we take a = 0 in Corollary 4.5.2, we see that E[b] = b That is, the expected value of a constant is just its value. (Is this intuitive?) Also, if we take b = 0, then we obtain E[aX] = aE[X] or, in words, the expected value of a constant multiplied by a random variable is just the constant times the expected value of the random variable. The expected

value of a random variable X , E[X], is also referred to as the mean or the first moment of X . The quantity $E[X^n]$, n ≥ 1, is called the nth moment of X . By Proposition

$$E[X^n] = \begin{cases} \sum_x x^n p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

4.5.1, we note that 4.5.1 Expected Value of Sums of Random Variables: The two-dimensional version

$$E[g(X, Y)] = \sum_y \sum_x g(x, y)p(x, y)$$

of Proposition 4.5.1 states that if X and Y are random variables and g is a function of two variables, then in the discrete

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)\,dx\,dy$$

case in the continuous case For example, if g (X , Y) = X + Y , then, in the continuous case,

$$E[X + Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)f(x, y)\,dx\,dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x, y)\,dx\,dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x, y)\,dx\,dy = E[X] + E[Y]$$

A similar result can be shown in the discrete case and indeed, for any random variables X and Y, E[X + Y] = E[X] + E[Y] (4.5.1) By repeatedly applying Equation 4.5.1 we can show that the expected value of the sum of any number of random variables equals the sum of their individual expectations. For instance, E[X + Y + Z] = E[(X + Y) + Z] = E[X + Y] + E[Z] by Equation 4.5.1 = E[X] + E[Y] + E[Z] again by Equation 4.5.1 And in general, for any n, E[X1 + X2 · · ·+Xn] = E[X1] + E[X2]+· · ·+E[Xn] (4.5.2) Equation 4.5.2 is an extremely useful formula whose utility will now be illustrated by examples. EXAMPLE 4.5f A secretary has typed N letters along with their respective envelopes. The envelopes get mixed up when they fall on the floor. If the letters are placed in the mixed-up envelopes in a completely random manner (that is, each letter is equally likely to end up in any of the envelopes), what is the expected number of letters that are placed in the correct envelopes? SOLUTION Letting X denote the number of letters that are placed in the correct envelope, we can most easily compute E[X] by noting that

$$X_i = \begin{cases} 1 & \text{if the } i\text{th letter is placed in its proper envelope} \\ 0 & \text{otherwise} \end{cases}$$

X = X1 + X2 +· · ·+XN where Now, since the ith letter is equally likely to be put in any of the N envelopes, it follows that P{Xi = 1} = P{ith letter is in its proper envelope} = 1/N and so E[Xi] = 1P{Xi = 1} + 0P{Xi = 0} = 1/N Hence, from Equation 4.5.2 we obtain that

$$E[X] = E[X_1] + \cdots + E[X_N] = \left(\frac{1}{N}\right) N = 1$$

Hence, no matter how many letters there are, on the average, exactly one of the letters will be in its own envelope. EXAMPLE 4.5g Suppose there are 20 different types of coupons and suppose that each time one obtains a coupon it is equally likely to be any one of the types. Compute the expected number of different types that are contained in a set for 10 coupons. SOLUTION Let X denote the number of different types in the set of 10 coupons. We compute E[X] by using the representation X = X1 +· · ·+X20 where

$$X_i = \begin{cases} 1 & \text{if at least one type } i \text{ coupon is contained in the set of 10} \\ 0 & \text{otherwise} \end{cases}$$

Now $E[X_i] = P\{X_i = 1\}$ = P{at least one type i coupon is in the set of 10} = 1 − P{no type i coupons are contained in the set of 10} $= 1 - \left(\frac{19}{20}\right)^{10}$ when the last equality follows since each of the 10 coupons will (independently) not be a type i with probability 19/20 . Hence, E[X] = E[X1]+· · ·+E[X20] = $20\left[1 - \left(\frac{19}{20}\right)^{10}\right] = 8.025$ An important property of the mean arises when one must predict the value of a random variable. That is, suppose that the value of a random variable X is to be predicted. If we predict that X will equal c, then the square of the "error" involved will be (X − c)2. We will now show that the average squared error is minimized when we predict that X will equal its mean μ. To see this, note that for any constant c.

$$E[(X - c)^2] = E[(X - \mu + \mu - c)^2] = E[(X - \mu)^2 + 2(\mu - c)(X - \mu) + (\mu - c)^2]$$

$$= E[(X - \mu)^2] + 2(\mu - c)E[X - \mu] + (\mu - c)^2 = E[(X - \mu)^2] + (\mu - c^2) \quad \text{since} \quad E[X - \mu] = E[X] - \mu = 0$$

$$\geq E[(X - \mu)^2]$$

Hence, the best predictor of a random variable, in terms of minimizing its mean square error, is just its mean. 4.6 VARIANCE: Given a random variable X along with its probability distribution function, it would be extremely useful if we were able to summarize the essential properties of the mass function by certain suitably defined measures. One such measure would be E[X], the expected value of X . However, while E[X] yields the weighted average of the possible values of X , it does not tell us anything about the variation, or spread, of these values. For instance, while the following random variables W, Y, and Z having probability mass functions determined by

$$W = 0 \quad \text{with probability 1} \qquad Y = \begin{cases} -1 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases} \qquad Z = \begin{cases} -100 & \text{with probability } \frac{1}{2} \\ 100 & \text{with probability } \frac{1}{2} \end{cases}$$

all have the same expectation—namely, 0—there is much greater spread in the possible values of Y than in those of W (which is a constant) and in the possible values of Z than in those of Y. Because we expect X to take on values around its mean E[X], it would appear that a reasonable way of measuring the possible variation of X would be to look at how far apart X would be from its mean on the average. One possible way to measure this would be to consider the quantity E[|X − μ|], where μ = E[X], and [X − μ] represents the absolute value of X −μ. However, it turns out to be mathematically inconvenient to deal with this quantity and so a more tractable quantity is usually considered — namely, the expectation of the square of the difference between X and its mean. We thus have the following definition. Definition If X is a random variable with mean μ, then the variance of X , denoted by Var(X), is defined by

$$\text{Var}(X) = E[(X - \mu)^2]$$

An alternative formula for Var(X) can be derived as follows:

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

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

That is, $\text{Var}(X) = E[X^2] - (E[X])^2$ (4.6.1) or, in words, the variance of X is equal to the expected value of the square of X minus the square of the expected value of X . This is, in practice, often the easiest way to compute Var(X). A useful identity concerning variances is that for any constants a and b, $\text{Var}(aX + b) = a^2\text{Var}(X)$ (4.6.2) To prove Equation 4.6.2, let μ = E[X] and recall that E[aX + b] = aμ + b. Thus, by the definition of variance, we have

$$\text{Var}(aX + b) = E[(aX + b - E[aX + b])^2] = E[(aX + b - a\mu - b)^2] = E[(aX - a\mu)^2] = E[a^2(X - \mu)^2]$$

$$= a^2E[(X - \mu)^2]$$

Specifying particular values for a and b in Equation 4.6.2 leads to some interesting corollaries. For instance, by setting a = 0 in Equation 4.6.2 we obtain that Var(b) = 0 That is, the variance of a constant is 0. (Is this intuitive?) Similarly, by setting a = 1 we obtain Var(X + b) = Var(X) That is, the variance of a constant plus a random variable is equal to the variance of the random variable. (Is this intuitive? Think about it.)

Finally, setting b = 0 yields $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$ The quantity $\sqrt{\mathrm{Var}(X)}$ is called the standard deviation of X . The standard deviation has the same units as does the mean. REMARK: Analogous to the mean's being the center of gravity of a distribution of mass, the variance represents, in the terminology of mechanics, the moment of inertia. 4.7 COVARIANCE AND VARIANCE OF SUMS OF RANDOM VARIABLES: We showed in Section 4.5 that the expectation of a sum of random variables is equal to the sum of their expectations. The corresponding result for variances is, however, not generally valid. Consider Var(X + X) = Var(2X) = 2^2 Var(X) = 4 Var(X) = Var(X) + Var(X) There is, however, an important case in which the variance of a sum of random variables is equal to the sum of the variances; and this is when the random variables are independent. Before proving this, however, let us define the concept of the covariance of two random variables. Definition: The covariance of two random variables X and Y , written Cov(X , Y) is defined by $\mathrm{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$ where μx and μy are the means of X and Y , respectively. A useful expression for Cov(X , Y) can be obtained by expanding the right side of the definition. This yields

$$\mathrm{Cov}(X, Y) = E[XY - \mu_x Y - \mu_y X + \mu_x\mu_y] = E[XY] - \mu_x E[Y] - \mu_y E[X] + \mu_x\mu_y$$

$$= E[XY] - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y = E[XY] - E[X]E[Y]$$ (4.7.1) From its definition we see that covariance satisfies the following properties: Cov(X , Y) = Cov(Y , X) (4.7.2) and Cov(X , X) = Var(X) (4.7.3) Another property of covariance, which immediately follows from its definition, is that, for any constant a, Cov(aX , Y) = a Cov(X , Y) (4.7.4) Covariance, like expectation, possesses an additive property. Lemma 4.7.1 Cov(X + Z, Y) = Cov(X , Y) + Cov(Z, Y) Proof: Cov(X + Z, Y) = E[(X + Z)Y] − E[X + Z]E[Y] from Equation 4.7.1 = E[XY] + E[ZY] − (E[X] + E[Z])E[Y] = E[XY] − E[X]E[Y] + E[ZY] − E[Z]E[Y] = Cov(X , Y) + Cov(Z, Y) Lemma 4.7.1 can be easily generalized (see Problem 48) to show that

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n} \mathrm{Cov}(X_i, Y)$$ (4.7.5) which gives rise to the following. PROPOSITION 4.7.2:

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$$

Proof:

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \mathrm{Cov}\left(X_i, \sum_{j=1}^{m} Y_j\right) \quad \text{from Equation 4.7.5}$$

$$= \sum_{i=1}^{n} \mathrm{Cov}\left(\sum_{j=1}^{m} Y_j, X_i\right) \quad \text{by the symmetry property Equation 4.7.2} \quad = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(Y_j, X_i) \quad \text{again from Equation 4.7.5}$$

and the result now follows by again applying the property Equation 4.7.2. Using Equation 4.7.3 gives rise to the following formula for the variance of a sum of random variables. Corollary 4.7.3:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{Cov}(X_i, X_j)$$

Proof: The proof follows directly from Proposition 4.7.2 upon setting m = n, and Yj = Xj for j = 1, . . . , n. In the case of n = 2, Corollary 4.7.3: yields that Var(X + Y) = Var(X) + Var(Y) + Cov(X , Y) + Cov(Y , X) or, using Equation 4.7.2, Var(X + Y) = Var(X) + Var(Y) + 2Cov(X , Y) (4.7.6) Theorem 4.7.4: If X and Y are independent random variables, then Cov(X , Y) = 0 and so for independent X1, . . . , Xn,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Proof: We need to prove that E[XY] = E[X]E[Y]. Now, in the discrete case,

$$E[XY] = \sum_j\sum_i x_i y_j P\{X = x_i, Y = y_j\} = \sum_j\sum_i x_i y_j P\{X = x_i\}P\{Y = y_j\} \quad \text{by independence}$$

$$= \sum_y y_j P\{Y = y_j\}\sum_i x_i P\{X = x_i\} = E[Y]E[X]$$ Because a similar argument holds in all other cases, the result is proven. EXAMPLE 4.7a Compute the variance of the sum obtained when 10 independent rolls of a fair die are made. SOLUTION Letting X_i denote the outcome of the ith roll, we have that

$$\mathrm{Var}\left(\sum_{1}^{10} X_i\right) = \sum_{1}^{10} \mathrm{Var}(X_i) = 10\tfrac{35}{12} \quad \text{from Example 4.6a} \quad = \tfrac{175}{6}$$ EXAMPLE 4.7b Compute the variance of the number of heads resulting from 10 independent tosses of a fair coin. SOLUTION Letting

$$I_j = \begin{cases} 1 & \text{if the } j\text{th toss lands heads} \\ 0 & \text{if the } j\text{th toss lands tails} \end{cases}$$ then the total number of heads is equal to $\sum_{j=1}^{10} I_j$ Hence, from Theorem 4.7.4,

$$\mathrm{Var}\left(\sum_{j=1}^{10} I_j\right) = \sum_{j=1}^{10} \mathrm{Var}(I_j)$$ Now, since I_j is an indicator random variable for an event having probability 1/2

$$\mathrm{Var}(I_j) = \tfrac{1}{2}\left(1 - \tfrac{1}{2}\right) = \tfrac{1}{4}$$, it follows from Example 4.6b that and thus

$$\mathrm{Var}\left(\sum_{j=1}^{10} I_j\right) = \frac{10}{4}$$ The covariance of two random variables is important as an indicator of the relationship between them. For instance, consider the situation where X and Y are indicator variables for whether or not the events A and B occur. That is, for events A and B, define

$$X = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}, \qquad Y = \begin{cases} 1 & \text{if } B \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

and note that

$$XY = \begin{cases} 1 & \text{if } X = 1, Y = 1 \\ 0 & \text{otherwise} \end{cases}$$ Thus, Cov(X , Y) = E[XY] − E[X]E[Y] = P{X = 1, Y = 1} − P{X = 1}P{Y = 1} From this we see that Cov(X , Y) > 0 ⇔ P{X = 1, Y = 1} > P{X = 1}P{Y = 1}

$$\Leftrightarrow \frac{P\{X = 1, Y = 1\}}{P\{X = 1\}} > P\{Y = 1\} \Leftrightarrow P\{Y = 1|X = 1\} > P\{Y = 1\}$$ That is, the covariance of X and Y is positive if the outcome X = 1 makes it more likely that Y = 1 (which, as is easily seen by symmetry, also implies the reverse). In general, it can be shown that a positive value of Cov(X , Y) is an

indication that Y tends to increase as X does, whereas a negative value indicates that Y tends to decrease as X increases. The strength of the relationship between X and Y is indicated by the correlation between X and Y , a dimensionless quantity obtained by dividing the covariance

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

by the product of the standard deviations of X and Y . That is, It can be shown (see Problem 49) that this quantity always has a value between−1 and+1. 4.8 MOMENT GENERATING FUNCTIONS: The moment generating function ϕ(t) of the random variable X is defined for all values t by

$$\phi(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx} p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

We call ϕ(t) the moment generating function because all of the moments of X can be obtained by successively differentiating ϕ(t). For example, $\phi'(t) = \frac{d}{dt}E[e^{tX}] = E\left[\frac{d}{dt}(e^{tX})\right] = E[Xe^{tX}]$ Hence, $\phi'(0) = E[X]$ Similarly,

$$\phi''(t) = \frac{d}{dt}\phi'(t) = \frac{d}{dt}E[Xe^{tX}] = E\left[\frac{d}{dt}(Xe^{tX})\right] = E[X^2e^{tX}]$$ and so, $\phi''(0) = E[X^2]$ In general, the nth derivative of ϕ(t) evaluated at t = 0 equals $E[X^n]$; that is, $\phi^n(0) = E[X^n], \quad n \geq 1$ An important property of moment generating functions is that the moment generating function of the sum of independent random variables is just the product of the individual moment generating functions. To see this, suppose that X and Y are independent and have moment generating functions $\phi_X(t)$ and $\phi_Y(t)$, respectively. Then $\phi_{X+Y}(t)$, the moment generating function of X + Y , is given by $\phi_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = \phi_X(t)\phi_Y(t)$ where the next to the last equality follows from Theorem 4.7.4 since X and Y , and thus etX and etY , are independent. Another important result is that the moment generating function uniquely determines the distribution. That is, there exists a one-to-one correspondence between the moment generating function and the distribution function of a random variable. 4.9 CHEBYSHEV'S INEQUALITY AND THE WEAK LAW OF LARGE NUMBERS: We start this section by proving a result known as Markov's inequality. PROPOSITION 4.9.1 MARKOV'S INEQUALITY If X is a random variable that takes only nonnegative values, then for any value a > 0 Proof:

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

$$E[X] = \int_0^{\infty} xf(x)\, dx = \int_0^a xf(x)\, dx + \int_a^{\infty} xf(x)\, dx \geq \int_a^{\infty} xf(x)\, dx$$

$$\geq \int_a^{\infty} af(x)\, dx = a\int_a^{\infty} f(x)\, dx = aP\{X \geq a\}$$

and the result is proved. As a corollary, we obtain Proposition 4.9.2. PROPOSITION 4.9.2 CHEBYSHEV'S INEQUALITY: If X is a random variable with mean μ and variance σ^2, then for any value k > 0 Proof: Since

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

$(X - \mu)^2$ is a nonnegative random variable, we can apply Markov's inequality $(\text{with } a = k^2)$ to obtain

$$P\{(X - \mu)^2 \geq k^2\} \leq \frac{E[(X - \mu)^2]}{k^2}$$

(4.9.1) But since $(X - \mu) \geq k^2$ if and only if $|X - \mu| \geq k$, Equation 4.9.1 is equivalent to

$$P\{|X - \mu| \geq k\} \leq \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

and the proof is complete. The importance of Markov's and Cheybyshev's inequalities is that they enable us to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known. Of course, if the actual distribution were known, then the desired probabilities could be exactly computed and we would not need to resort to bounds. By replacing k by kσ in Equation 4.9.1, we can write Chebyshev's inequality as

$$P\{|X - \mu| > k\sigma\} \leq 1/k^2$$

Thus it states that the probability a random variable differs from its mean by more than k standard deviations is bounded by $1/k^2$. We will end this section by using Chebyshev's inequality to prove the weak law of large numbers, which states that the probability that the average of the first n terms in a sequence of independent and identically distributed random variables differs by its mean by more than ε goes to 0 as n goes to infinity. Theorem 4.9.3 The Weak Law of Large Numbers: Let X1, X2, . . . , be a sequence of independent and identically distributed random variables, each having mean E[Xi] = μ. Then, for any ε > 0,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right\} \to 0 \quad \text{as } n \to \infty$$

Proof: We shall prove the result only under the additional assumption that the random variables have a finite variance σ^2. Now, as

$$E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu \quad \text{and} \quad \text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

it follows from Chebyshev's inequality that

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}$$

and the result is proved. For an application of the above, suppose that a sequence of independent trials is performed. Let E be a fixed event and denote by P(E) the probability that E occurs on a given trial. Letting

$$X_i = \begin{cases} 1 & \text{if } E \text{ occurs on trial } i \\ 0 & \text{if } E \text{ does not occur on trial } i \end{cases}$$

it follows that X1+X2+· · ·+Xn represents the number of times that E occurs in the first n trials. Because E[Xi] = P(E), it thus follows from the weak law of large numbers that for any positive number ε, no matter how small, the probability that the proportion of the first n trials in which E occurs differs from P(E) by more than ε goes to 0 as n increases.

Chapter_3: ELEMENTS OF PROBABILITY: 3.1 INTRODUCTION The concept of the probability of a particular event of an experiment is subject to various meanings or interpretations. For instance, if a geologist is quoted as saying that "there is a 60 percent chance of oil in a certain region," we all probably have some intuitive idea as to what is being said. Indeed, most of us would probably interpret this statement in one of two possible ways: either by imagining that 1. the geologist feels that, over the long run, in 60 percent of the regions whose outward environmental conditions are very similar to the conditions that prevail in the region under consideration, there will be oil; or, by imagining that 2. the geologist believes that it is more likely that the region will contain oil than it is that it will not; and in fact .6 is a measure of the

geologist's belief in the hypothesis that the region will contain oil. The two foregoing interpretations of the probability of an event are referred to as being the frequency interpretation and the subjective (or personal) interpretation of probability. It is imagined that this property can be operationally determined by continual repetition of the experiment — the probability of the outcome will then be observable as being the proportion of the experiments that result in the outcome. This is the interpretation of probability that is most prevalent. In this chapter, we present the accepted rules, or axioms, used in probability theory. 3.2 SAMPLE SPACE AND EVENTS: Consider an experiment whose outcome is not predictable with certainty in advance. Although the outcome of the experiment will not be known in advance, let us suppose that the set of all possible outcomes is known. This set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by S. Some examples are the following. Any subset E of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in E, then we say that E has occurred. Some examples of events are the following. For any two events E and F of a sample space S, we define the new event E UF , called the union of the events E and F, to consist of all outcomes that are either in E or in F or in both E and F. That is, the event E UF will occur if either E or F occurs. For instance, in Example 1 if E = {g } and F = {b}, then E UF = {g , b}. That is, E UF would be the whole sample space S. In Example 2 if E = {all outcomes starting with 6} is the event that the number 6 horse wins and F = {all outcomes having 6 in the second position} is the event that the number 6 horse comes in second, then E UF is the event that the number 6 horse comes in either first or second. Similarly, for any two events E and F, we may also define the new event EF, called the intersection of E and F, to consist of all outcomes that are in both E and F. That is, the event EF will occur only if both E and F occur. For instance, in Example 3 if E = (0, 5) is the event that the required dosage is less than 5 and F = (2, 10) is the event that it is between 2 and 10, then EF = (2, 5) is the event that the required dosage is between 2 and 5. In Example 2 if E = {all outcomes ending in 5} is the event that horse number 5 comes in last and F = {all outcomes starting with 5} is the event that horse number 5 comes in first, then the event EF does not contain any outcomes and hence cannot occur. To give such an event a name, we shall refer to it as the null event and denote it by Ø. Thus Ø refers to the event consisting of no outcomes. If EF = Ø, implying that E and F cannot both occur, then E and F are said to be mutually exclusive. For any event E, we define the event Ec , referred to as the complement of E, to consist of all outcomes in the sample space S that are not in E. That is, Ec will occur if and only if E does not occur. In Example 1 if E = {b} is the event that the child is a boy, then Ec = {g } is the event that it is a girl. Also note that since the experiment must result in some outcome, it follows that Sc = Ø. For any two events E and F, if all of the outcomes in E are also in F, then we say that E is contained in F and write E ⊂ F (or equivalently, F ⊃E). Thus if E ⊂ F , then the occurrence of E necessarily implies the occurrence of F. If E ⊂ F and F ⊂ E, then we say that E and F are equal (or identical) and we write E = F . We can also define unions and intersections of more than two events. In particular, the union of the events E1, E2, . . . , En, denoted either by E1 UE2 U · · · UEn or by Un1 Ei , is defined to be the event consisting of all outcomes that are in Ei for at least one i = 1, 2, . . . , n. Similarly, the intersection of the events Ei , i = 1, 2, . . . , n, denoted by E1E2 · · · En, is defined to be the event consisting of those outcomes that are in all of the events Ei , i = 1, 2, . . . , n. In other words, the union of the Ei occurs when at least one of the events Ei occurs; the intersection occurs when all of the events Ei occur. 3.3 VENN DIAGRAMS AND THE ALGEBRA OF EVENTS: The operations of forming unions, intersections, and complements of events obey certain rules not dissimilar to the rules of algebra. We list a few of these.

Commutative law E ∪ F = F UE EF = FE

Associative law (E ∪ F) ∪ G = E ∪ (F ∪ G) (EF)G = E(FG)

Distributive law (E ∪ F)G = EG ∪ FG EF ∪ G = (E ∪ G)(F ∪ G)

3.4 AXIOMS OF PROBABILITY: It appears to be an empirical fact that if an experiment is continually repeated under the exact same conditions, then for any event E, the proportion of time that the outcome is contained in E approaches some constant value as the number of repetitions increases. For instance, if a coin is continually flipped, then the proportion of flips resulting in heads will approach some value as the number of flips increases. It is this constant limiting frequency that we often have in mind when we speak of the probability of an event. From a purely mathematical viewpoint, we shall suppose that for each event E of an experiment having a sample space S there is a number, denoted by P(E), that is in accord with the following three axioms. AXIOM 1 0 ≤ P(E) ≤ 1

AXIOM 2 P(S) = 1

AXIOM 3 For any sequence of mutually exclusive events E1, E2, . . . (that is, events for which $E_iE_j = \emptyset$ when $i \neq j$), $P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i), \qquad n = 1, 2, \ldots, \infty$ We call P(E) the probability of the event E. Thus, Axiom 1 states that the probability that the outcome of the experiment is contained in E is some number between 0 and 1. Axiom 2 states that, with probability 1, the outcome will be a member of the sample space S. Axiom 3 states that for any set of mutually exclusive events the probability that **at least** one of these events occurs is equal to the sum of their respective probabilities. It should be noted that if we interpret P(E) as the relative frequency of the event E when a large number of repetitions of the experiment are performed, then P(E) would indeed satisfy the above axioms. For instance, the proportion (or frequency) of time that the outcome is in E is clearly between 0 and 1, and the proportion of time that it is in S is 1 (since all outcomes are in S). Also, if E and F have no outcomes in common, then the proportion of time that the outcome is in either E or F is the sum of their respective frequencies. As an illustration of this last statement, suppose the experiment consists of the rolling of a pair of dice and suppose that E is the event that the sum is 2, 3, or 12 and F is the event that the sum is 7 or 11. Then if outcome E occurs 11 percent of the time and outcome F 22 percent of the time, then 33 percent of the time the outcome will be either 2, 3, 12, 7, or 11. These axioms will now be used to prove two simple propositions concerning probabilities. We first note that E and Ec are always mutually exclusive, and since E UEc = S,we have by Axioms 2 and 3 that $1 = P(S) = P(E \cup E^c) = P(E) + P(E^c)$ Or equivalently, we have the following: PROPOSITION 3.4.1: $P(E^c) = 1 - P(E)$ In other words, Proposition 3.4.1 states that the probability that an event does not occur is 1 minus the probability that it does occur. For instance, if the probability of obtaining a head on the toss of a coin is 38 , the probability of obtaining a tail must be 58. Our second proposition gives the relationship between the probability of the union of two events in terms of the individual probabilities and the probability of the intersection. PROPOSITION 3.4.2: P(E ∪ F) = P(E)+P(F)−P(EF) The odds of an event A is defined by $\frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}$ Thus the odds of an event A tells how much more likely it is that A occurs than that it does not occur. For instance if P(A) = 3/4, then P(A)/(1 − P(A)) = 3, so the odds is 3. Consequently, it is 3 times as likely that A occurs as it is that it does not. 3.5 SAMPLE SPACES HAVING EQUALLY LIKELY OUTCOMES: For a large number of experiments, it is natural to assume that each point in the sample space is equally likely to occur. That is, for many experiments whose sample space S is a finite set, say S = {1, 2, . . . ,N}, it is often natural to assume that $P(\{1\}) = P(\{2\}) = \cdots = P(\{N\}) = p \quad \text{(say)}$ Now it follows from Axioms 2 and 3 that 1 = P(S) = P({1})+ · · · +P({N}) = Np which shows that P({i}) = p = 1/N From this it follows from Axiom 3 that for any event E, $P(E) = \frac{\text{Number of points in } E}{N}$ In words, if we assume that each outcome of an experiment is equally likely to occur, then the probability of any event E equals the proportion of points in the sample space that are contained in E. Thus, to compute probabilities it is often necessary to be able to effectively count the number of different ways that a given event can occur. To do this, we will make use of the following rule. BASIC PRINCIPLE OF COUNTING: Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of m possible outcomes and if, for each outcome of experiment 1, there are n possible outcomes of experiment 2, then together there are mn possible outcomes of the two experiments. Generalized Basic Principle of Counting: If r experiments that are to be performed are such that the first one may result in any of n1 possible outcomes, and if for each of these n1 possible outcomes there are n2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments there are n3 possible outcomes of the third experiment, and if, . . . , then there are a total of n1 · n2 · · · nr possible outcomes of the r experiments. Suppose now that we are interested in determining the number of different groups of r objects that could be formed from a total of n objects. For instance, how many different groups of three could be selected from the five items A, B, C, D, E? To answer this, reason as follows. Since there are 5 ways to select the initial item, 4 ways to then select the next item, and 3 ways to then select the final item, there are thus 5 · 4 · 3 ways of selecting the group of 3 when the order in which the items are selected is relevant. However, since every group of 3, say the group consisting of items A, B, and C, will be counted 6 times (that is, all of the permutations ABC, ACB, BAC, BCA, CAB, CBA will be counted when the order of selection is relevant), it follows that the total number of different groups that can be formed is (5 · 4 · 3)/(3 · 2 · 1) = 10. In general, as n(n−1) · · · (n−r +1) represents the number of different ways that a group of r items could be selected from n items when the order of selection is considered relevant (since the first one selected can be any one of the n, and the second selected any one of the remaining n−1, etc.), and since each group of r items will be counted r! times in this count, it follows that the number of different groups of r items that could be formed from a set of n items is $\frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!}$ NOTATION AND TERMINOLOGY: We define $\binom{n}{r}$, for $r \leq n$, by $\binom{n}{r} = \frac{r!}{(n-r)!r!}$ and call $\binom{n}{r}$ the number of combinations of n objects taken r at a time. Thus $\binom{n}{r}$ represents the number of different groups of size r that can be selected from a set of size n when the order of

selection is not considered relevant. 3.6 CONDITIONAL PROBABILITY: In this section, we introduce one of the most important concepts in all of probability theory — that of conditional probability. Its importance is twofold. In the first place, we are often interested in calculating probabilities when some partial information concerning the result of the experiment is available, or in recalculating them in light of additional information. In such situations, the desired probabilities are conditional ones. Second, as a kind of a bonus, it often turns out that the easiest way to compute the probability of an event is to first "condition" on the occurrence or nonoccurrence of a secondary event. A general formula for P(E|F) that is valid for all events E and F is derived in the same manner as just described. Namely, if the event F occurs, then in order for E to occur it is necessary that the actual occurrence be a point in both E and F; that is, it must be in EF. Now, since we know that F has occurred, it follows that F becomes our new (reduced) sample space and hence the probability that the event EF occurs will equal the probability of EF relative to the probability of F. That is,

$$P(E|F) = \frac{P(EF)}{P(F)}$$

(3.6.1) Note that Equation 3.6.1 is well defined only when P(F)>0 and hence P(E|F) is defined only when P(F)>0. (See Figure 3.5.)

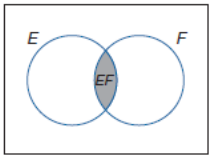


FIGURE 3.5 $P(E|F) = \frac{P(EF)}{P(F)}$.

The definition of conditional probability given by Equation 3.6.1 is consistent with the interpretation of probability as being a long-run relative frequency. To see this, suppose that a large number n of repetitions of the experiment are performed. Then, since P(F) is the long-run proportion of experiments in which F occurs, it follows that F will occur approximately nP(F) times. Similarly, in approximately nP(EF) of these experiments, both E and F will occur. Hence, of the approximately nP(F) experiments whose outcome is in F, approximately nP(EF) of them will also have their outcome in E. That is,

$$\frac{nP(EF)}{nP(F)} = \frac{P(EF)}{P(F)}$$

for those experiments whose outcome is in F, the proportion whose outcome is also in E is approximately Since this approximation becomes exact as n becomes larger and larger, it follows that (3.6.1) gives the appropriate definition of the conditional probability of E given that F has occurred. 3.7 BAYES' FORMULA: Let E and F be events. We may express E as: $E = EF \cup EF^c$ for, in order for a point to be in E, it must either be in both E and F or be in E but not in F. (See Figure 3.6.)

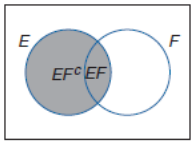


FIGURE 3.6 $E = EF \cup EF^c$.

As EF and EF c are clearly mutually exclusive, we have by Axiom 3 that $P(E) = P(EF) + P(EF^c)$

$$= P(E|F)P(F) + P(E|F^c)P(F^c) \quad = P(E|F)P(F) + P(E|F^c)[1 - P(F)]$$

(3.7.1) Equation 3.7.1 states that the probability of the event E is a weighted average of the conditional probability of E given that F has occurred and the conditional probability of E given that F has not occurred: Each conditional probability is given as much weight as the event it is conditioned on has of occurring. It is an extremely useful formula, for its use often enables us to determine the probability of an event by first "conditioning" on whether or not some second event has occurred. That is, there are many instances where
it is difficult to compute the probability of an event directly, but it is straightforward to compute it once we know whether or not some second event has occurred.

Equation 3.7.1 may be generalized in the following manner. Suppose that F1, F2, . . . , Fn are mutually exclusive events such that $\bigcup_{i=1}^{n} F_i = S$ In other words, exactly

$$E = \bigcup_{i=1}^{n} EF_i$$

one of the events F1, F2, . . . , Fn must occur. By writing and using the fact that the events EFi , i = 1, . . . , n are mutually exclusive, we obtain that

$$P(E) = \sum_{i=1}^{n} P(EF_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

(3.7.2) Thus, Equation 3.7.2 shows how, for given events F1, F2, . . . , Fn of which one and only one must occur, we can compute P(E) by first "conditioning" on which one of the Fi occurs. That is, it states that P(E) is equal to a weighted average of P(E|Fi), each term
being weighted by the probability of the event on which it is conditioned. Suppose now that E has occurred and we are interested in determining which one of

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$$

Fj also occurred. By Equation 3.7.2, we have that (3.7.3) Equation 3.7.3 is known as Bayes' formula, after the English philosopher Thomas Bayes. If we think of the events Fj as being possible "hypotheses" about some subject matter, then Bayes' formula may be interpreted as showing us how opinions about these hypotheses held before the experiment [that is, the P(Fj)] should be modified by the evidence of the experiment. EXAMPLE 3.7f A plane is missing and it is presumed that it was equally likely to have gone down in any of three possible regions. Let 1−αi denote the probability the plane
will be found upon a search of the ith region when the plane is, in fact, in that region, i =1, 2, 3. (The constants αi are called overlook probabilities because they represent the probability of overlooking the plane; they are generally attributable to the geographical and environmental conditions of the regions.) What is the conditional probability that the plane is in the ith region, given that a search of region 1 is unsuccessful, i = 1, 2, 3? SOLUTION Let Ri , i = 1, 2, 3, be the event that the

$$P(R_1|E) = \frac{P(ER_1)}{P(E)} = \frac{P(E|R_1)P(R_1)}{\sum_{i=1}^{3} P(E|R_i)P(R_i)}$$

plane is in region i; and let E be the event that a search of region 1 is unsuccessful. From Bayes' formula, we obtain

$$= \frac{(\alpha_1)(1/3)}{(\alpha_1)(1/3) + (1)(1/3) + (1)(1/3)} = \frac{\alpha_1}{\alpha_1 + 2}$$ For j = 2, 3, $$P(R_j|E) = \frac{P(E|R_j)P(R_j)}{P(E)} = \frac{(1)(1/3)}{(\alpha_1)1/3 + 1/3 + 1/3} = \frac{1}{\alpha_1 + 2}, \qquad j = 2, 3$$

Thus, for instance, if α1 = .4, then the conditional probability that the plane is in region 1 given that a search of that region did not uncover it is 1/6 3.8 INDEPENDENT EVENTS: The previous examples in this chapter show that P(E|F), the conditional probability of E given F, is not generally equal to P(E), the unconditional probability of E. In other words, knowing that F has occurred generally changes the chances of E 's occurrence. In the special cases where P(E|F) does in fact equal P(E), we say that E is independent of F. That is, E is independent of F if knowledge that F has occurred does not change the probability that E occurs. Since P(E|F) = P(EF)/P(F), we see that E is independent of F if P(EF) = P(E)P(F) (3.8.1) Since this equation is symmetric in E and F, it shows that whenever E is independent of F so is F of E. We thus have the following. Definition: Two events E and F are said to be independent if Equation 3.8.1 holds. Two events E and F that are not independent are said to be dependent. PROPOSITION 3.8.1 If E and F are independent, then so are E and F^c Proof: Assume that E and F are independent. Since $E = EF \cup EF^c$, and EF and EF^c are obviously mutually exclusive, we have that

$$P(E) = P(EF) + P(EF^c) \quad = P(E)P(F) + P(EF^c) \qquad \text{by the independence of } E \text{ and } F$$

or equivalently, $$P(EF^c) = P(E)(1 - P(F)) \quad = P(E)P(F^c)$$ and the result is proven. Thus if E is independent of F, then the probability of E 's occurrence is unchanged by information as to whether or not F has occurred. Suppose now that E is independent of F and is also independent of G. Is E then necessarily independent of FG ? The answer, somewhat surprisingly, is no. Consider the following example. Definition

The three events E, F, and G are said to be independent if
P(EFG) = P(E)P(F)P(G)
P(EF) = P(E)P(F)
P(EG) = P(E)P(G)
P(FG) = P(F)P(G)
It should be noted that if the events E, F, G are independent, then E will be independent of any event formed from F and G. For instance, E is independent of F ∪ G since
P(E(F ∪ G)) = P(EF ∪ EG)
= P(EF)+P(EG)−P(EFG)
= P(E)P(F)+P(E)P(G)−P(E)P(FG)
= P(E)[P(F)+P(G)−P(FG)]
= P(E)P(F ∪ G) Of course we may also extend the definition of independence to more than three events. The events E1, E2, . . . , En are said to be independent if for every subset $E_{1'}, E_{2'}, \ldots, E_{r'}, r \leq n$, of these events $P(E_{1'}E_{2'} \cdots E_{r'}) = P(E_{1'})P(E_{2'}) \cdots P(E_{r'})$.

Chapter_2: DESCRIPTIVE STATISTICS: 2.1 INTRODUCTION: In this chapter we introduce the subject matter of descriptive statistics, and in doing so learn ways to describe and summarize a set of data. Section 2.2 deals with ways of describing a data set. Subsections 2.2.1 and 2.2.2 indicate how data that take on only a relatively few distinct values can be described by using frequency tables or graphs, whereas Subsection 2.2.3 deals with data whose set of values is grouped into different intervals. Section 2.3 discusses ways of summarizing data sets by use of statistics, which are numerical quantities whose values are determined by the data. Subsection 2.3.1 considers three statistics that are used to indicate the "center" of the data set: the sample mean, the sample median, and the sample mode. Subsection 2.3.2 introduces the sample variance and its square root, called the sample standard deviation. These statistics are used to indicate the
spread of the values in the data set. Subsection 2.3.3 deals with sample percentiles, which are statistics that tell us, for instance, which data value is greater than 95 percent of all the data. In Section 2.4 we present Chebyshev's inequality for sample data. This famous inequality gives a lower bound to the proportion of the data that can differ from the sample mean by more than k times the sample standard deviation. Whereas Chebyshev's inequality holds for all data sets, we can in certain situations, which are discussed in Section 2.5, obtain more precise estimates of the proportion of the data that is within k sample standard deviations of the sample mean. In Section 2.5 we note that when a graph of the data follows a bell-shaped form the data set is said to be approximately normal, and
more precise estimates are given by the so-called empirical rule. Section 2.6 is concerned with situations in which the data consist of paired values. A graphical technique, called the scatter diagram, for presenting such data is introduced, as is the sample correlation coefficient, a statistic that indicates the degree to which a large value of the first member of the pair tends to go along with a large value of the second. 2.2 DESCRIBING DATA SETS: 2.2.1 Frequency Tables and Graphs: A data set having a relatively small number of distinct values can be conveniently presented in a frequency table. Data from a frequency table can be graphically represented by a line graph that plots the distinct data values on the horizontal axis and indicates their frequencies by the heights of vertical lines. A line graph of the data presented in Table 2.1 is shown in Figure 2.1.

TABLE 2.1 *Starting Yearly Salaries*

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

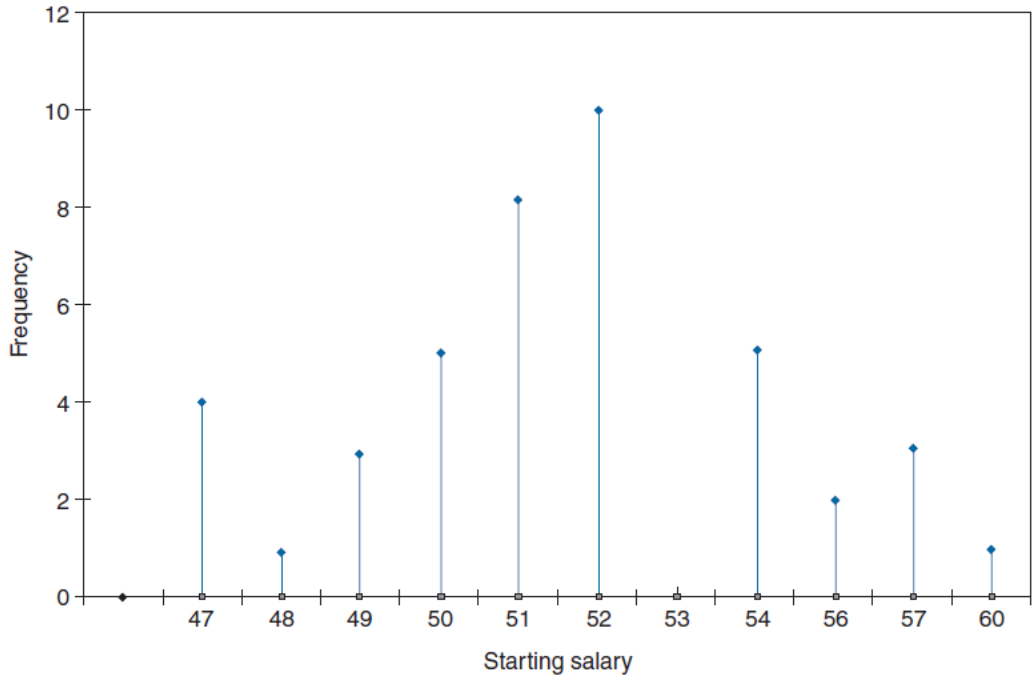


FIGURE 2.1 *Starting salary data.*

When the lines in a line graph are given added thickness, the graph is called a bar graph. Figure 2.2 presents a bar graph.

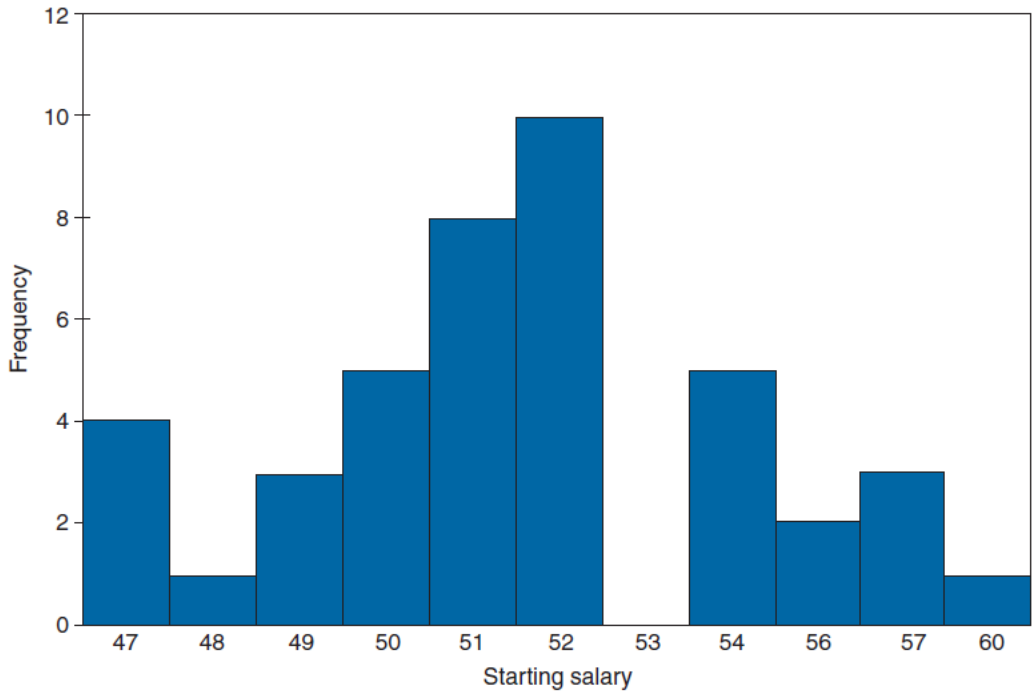


FIGURE 2.2 *Bar graph for starting salary data.*

Another type of graph used to represent a frequency table is the frequency polygon, which plots the frequencies of the different data values on the vertical axis, and then connects the plotted points with straight lines. Figure 2.3 presents a frequency polygon for the data of Table 2.1.

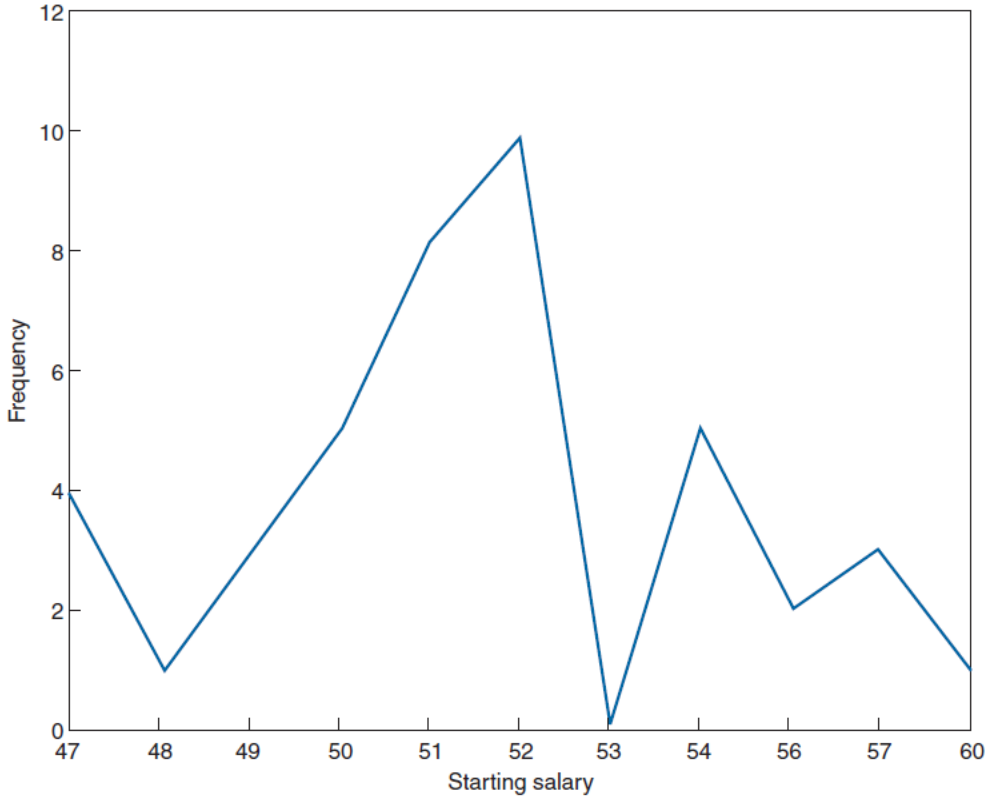


FIGURE 2.3 *Frequency polygon for starting salary data.*

2.2.2 Relative Frequency Tables and Graphs: Consider a data set consisting of n values. If f is the frequency of a particular value, then the ratio f /n is called its relative frequency. That is, the relative frequency of a data value is the proportion of the data that have that value. The relative frequencies can be represented graphically by a relative frequency line or bar graph or by a relative frequency polygon. Indeed, these relative frequency graphs will look like the corresponding graphs of the absolute frequencies except that the labels on the vertical axis are now the old labels (that gave the frequencies) divided by the total number of data points. EXAMPLE 2.2a Table 2.2 is a relative frequency table for the

data of Table 2.1. The relative frequencies are obtained by dividing the corresponding frequencies of Table 2.1 by 42, the size of the data set.

TABLE 2.2

Starting Salary	Frequency
47	4/42 = .0952
48	1/42 = .0238
49	3/42
50	5/42
51	8/42
52	10/42
53	0
54	5/42
56	2/42
57	3/42
60	1/42

A pie chart is often used to indicate relative frequencies when the data are not numerical in nature. A circle is constructed and then sliced into different sectors; one for each distinct type of data value. The relative frequency of a data value is indicated by the area of its sector, this area being equal to the total area of the circle multiplied by the relative frequency of the data value. EXAMPLE 2.2b The following data relate to the different types of cancers affecting the 200 most recent patients to enroll at a clinic specializing in cancer. These data are represented

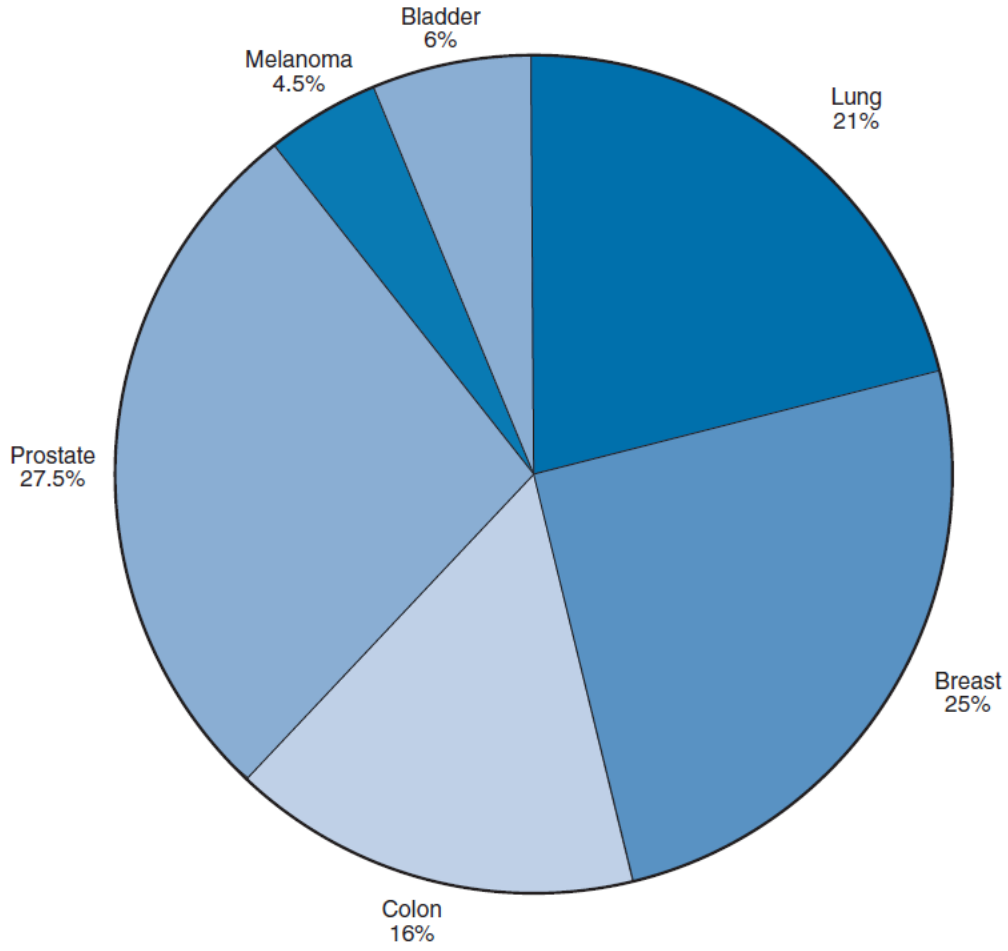


FIGURE 2.4

in the pie chart presented in Figure 2.4.

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

2.2.3 Grouped Data, Histograms, Ogives, and Stem and Leaf Plots: However, for some data sets the number of distinct values is too large to utilize this approach. Instead, in such cases, it is useful to divide the values into groupings, or class intervals, and then plot the number of data values falling in each class interval. The number of class intervals chosen should be a trade-off between (1) choosing too few classes at a cost of losing too much information about the actual data values in a class and (2) choosing too many classes, which will result in the frequencies of each class being too small for a pattern to be discernible. Although 5 to 10 class intervals are typical, the appropriate number is a subjective choice, and of course, you can try different numbers of class intervals to see which of the resulting charts appears to be most revealing about the data. It is common, although not essential, to choose class intervals of equal length. The endpoints of a class interval are called the class boundaries. We will adopt the left-end inclusion convention,

which stipulates that a class interval contains its left-end but not its right-end boundary point. Thus, for instance, the class interval 20–30 contains all values that are both greater than or equal to 20 and less than 30. Table 2.3 presents the lifetimes of 200 incandescent lamps.

TABLE 2.3 *Life in Hours of 200 Incandescent Lamps*

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

A class frequency table for the data of Table 2.3 is presented in Table 2.4.

TABLE 2.4 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

The class intervals are of length 100, with the first one starting at 500. A bar graph plot of class data, with the bars placed adjacent to each other, is called a histogram. The vertical axis of a histogram can represent either the class frequency or the relative class frequency; in the former case the graph is called a frequency histogram and in the latter a relative frequency histogram. Figure 2.5 presents a frequency histogram of the data in Table 2.4.

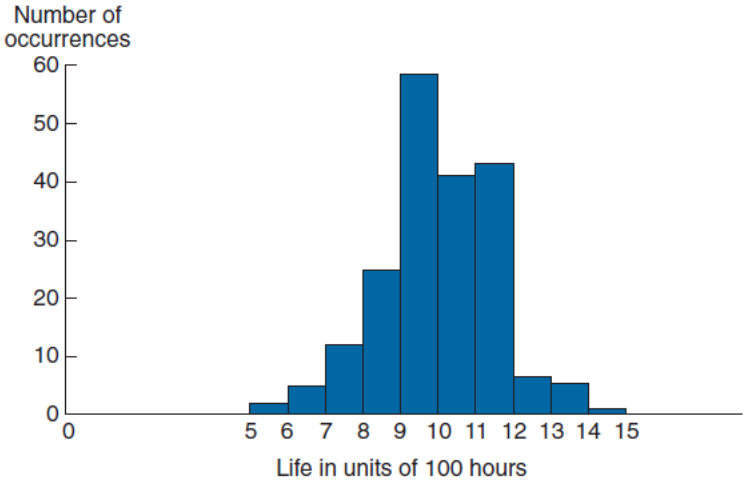


FIGURE 2.5 *A frequency histogram.*

We are sometimes interested in plotting a cumulative frequency (or cumulative relative frequency) graph. A point on the horizontal axis of such a graph represents a possible data value; its corresponding vertical plot gives the number (or proportion) of the data whose values are less than or equal to it. A cumulative relative frequency plot of the data of Table 2.3 is given in Figure 2.6. We can conclude from this figure that 100 percent of the data values are less than 1,500, approximately 40 percent are less than or equal to 900, approximately 80 percent are less than or equal to 1,100, and so on. A cumulative frequency plot is called an ogive.

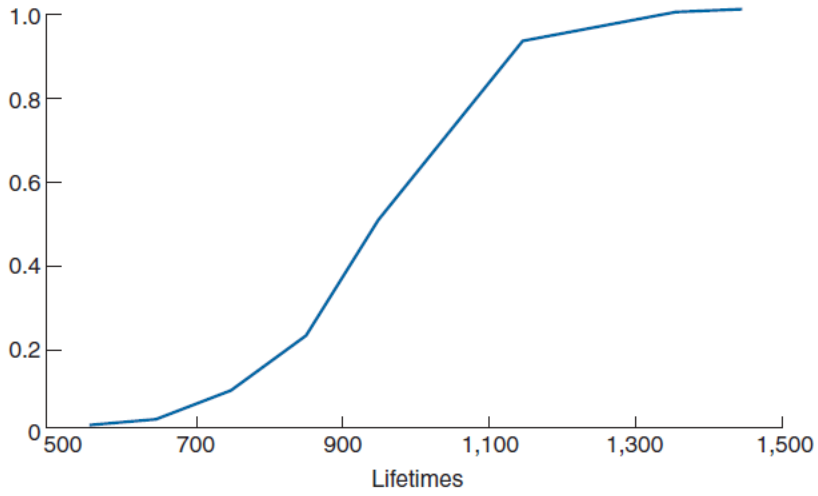


FIGURE 2.6 *A cumulative frequency plot.*

An efficient way of organizing a small- to moderate-sized data set is to utilize a stem and leaf plot. Such a plot is obtained by first dividing each data value into two parts — its stem and its leaf. For instance, if the data are all two-digit numbers, then we could let the stem part of a data value be its tens digit and let the leaf be its ones digit. Thus, for instance, the value 62 is expressed as Stem Leaf 6 2 and the two data values 62 and 67 can be represented as Stem Leaf 6 2,7 EXAMPLE 2.2c Table 2.5 gives the monthly and yearly average daily minimum temperatures in 35 U.S. cities. The annual average daily minimum temperatures from Table 2.5 are represented in the following stem and leaf plot.

TABLE 2.5 *Normal Daily Minimum Temperature — Selected Cities*

[In Fahrenheit degrees. Airport data except as noted. Based on standard 30-year period, 1961 through 1990]

State	Station	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual avg.
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5
DE	Wilmington	22.4	24.8	33.1	41.8	52.2	61.6	67.1	65.9	58.2	45.7	37.0	27.6	44.8
DC	Washington	26.8	29.1	37.7	46.4	56.6	66.5	71.4	70.0	62.5	50.3	41.1	31.7	49.2
FL	Jacksonville	40.5	43.3	49.2	54.9	62.1	69.1	71.9	71.8	69.0	59.3	50.2	43.4	57.1
	Miami	59.2	60.4	64.2	67.8	72.1	75.1	76.2	76.7	75.9	72.1	66.7	61.5	69.0
GA	Atlanta	31.5	34.5	42.5	50.2	58.7	66.2	69.5	69.0	63.5	51.9	42.8	35.0	51.3
HI	Honolulu	65.6	65.4	67.2	68.7	70.3	72.2	73.5	74.2	73.5	72.3	70.3	67.0	70.0
ID	Boise	21.6	27.5	31.9	36.7	43.9	52.1	57.7	56.8	48.2	39.0	31.1	22.5	39.1
IL	Chicago	12.9	17.2	28.5	38.6	47.7	57.5	62.6	61.6	53.9	42.2	31.6	19.1	39.5
	Peoria	13.2	17.7	29.8	40.8	50.9	60.7	65.4	63.1	55.2	43.1	32.5	19.3	41.0
IN	Indianapolis	17.2	20.9	31.9	41.5	51.7	61.0	65.2	62.8	55.6	43.5	34.1	23.2	42.4
IA	Des Moines	10.7	15.6	27.6	40.0	51.5	61.2	66.5	63.6	54.5	42.7	29.9	16.1	40.0
KS	Wichita	19.2	23.7	33.6	44.5	54.3	64.6	69.9	67.9	59.2	46.6	33.9	23.0	45.0
KY	Louisville	23.2	26.5	36.2	45.4	54.7	62.9	67.3	65.8	58.7	45.8	37.3	28.6	46.0
LA	New Orleans	41.8	44.4	51.6	58.4	65.2	70.8	73.1	72.8	69.5	58.7	51.0	44.8	58.5
ME	Portland	11.4	13.5	24.5	34.1	43.4	52.1	58.3	57.1	48.9	38.3	30.4	17.8	35.8
MD	Baltimore	23.4	25.9	34.1	42.5	52.6	61.8	66.8	65.7	58.4	45.9	37.1	28.2	45.2
MA	Boston	21.6	23.0	31.3	40.2	49.8	59.1	65.1	64.0	56.8	46.9	38.3	26.7	43.6
MI	Detroit	15.6	17.6	27.0	36.8	47.1	56.3	61.3	59.6	52.5	40.9	32.2	21.4	39.0
	Sault Ste. Marie	4.6	4.8	15.3	28.4	38.4	45.5	51.3	51.3	44.3	36.2	25.9	11.8	29.8
MN	Duluth	−2.2	2.8	15.7	28.9	39.6	48.5	55.1	53.3	44.5	35.1	21.5	4.9	29.0
	Minneapolis-St. Paul	2.8	9.2	22.7	36.2	47.6	57.6	63.1	60.3	50.3	38.8	25.2	10.2	35.3
MS	Jackson	32.7	35.7	44.1	51.9	60.0	67.1	70.5	69.7	63.7	50.3	42.3	36.1	52.0
MO	Kansas City	16.7	21.8	32.6	43.8	53.9	63.1	68.2	65.7	56.9	45.7	33.6	21.9	43.7
	St. Louis	20.8	25.1	35.5	46.4	56.0	65.7	70.4	67.9	60.5	48.3	37.7	26.0	46.7
MT	Great Falls	11.6	17.2	22.8	31.9	40.9	48.6	53.2	52.2	43.5	35.8	24.3	14.6	33.1

Source: U.S. National Oceanic and Atmospheric Administration, Climatography of the United States, No. 81.

7 0.0
6 9.0
5 1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4 0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3 3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5

2 9.0, 9.8
2.3 SUMMARIZING DATA SETS: In this section we present some summarizing statistics, where a statistic is a numerical quantity whose value is determined by the data.
2.3.1 Sample Mean, Sample Median, and Sample Mode: In this section we introduce some statistics that are used for describing the center of a set of data values. To begin, suppose that we have a data set consisting of the n numerical values x1, x2, . . . , xn. The sample mean is the arithmetic average of these values.

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

Definition: The sample mean, designated by $\bar{x}$, is defined by The computation of the sample mean can often be simplified by noting that if for

$$\bar{y} = \sum_{i=1}^{n} (ax_i + b)/n = \sum_{i=1}^{n} ax_i/n + \sum_{i=1}^{n} b/n = a\bar{x} + b$$

constants a and b yi =axi +b, i =1, . . . , n then the sample mean of the data set y1, . . . , yn is
Sometimes we want to determine the sample mean of a data set that is presented in a frequency table listing the k distinct values $v_1, \ldots, v_k$ having corresponding frequencies f1, . . . , fk . Since such a data set consists of $n = \sum_{i=1}^{k} f_i$ observations,

$$\bar{x} = \sum_{i=1}^{k} v_i f_i/n$$

with the value vi appearing f_i times, for each i =1, . . . , k, it follows that the sample mean of these n data values is By writing the preceding as

$$\bar{x} = \frac{f_1}{n} v_1 + \frac{f_2}{n} v_2 + \cdots + \frac{f_k}{n} v_k$$

we see that the sample mean is a weighted average of the distinct values, where the weight given to the value vi is equal to the proportion of the n data values that are equal to vi , i =1, . . . , k. Another statistic used to indicate the center of a data set is the sample median; loosely speaking, it is the middle value when the data set is arranged in increasing order. Definition Order the values of a data set of size n from smallest to largest. If n is odd, the sample median is the value in position (n+1)/2; if n is even, it is the average of the values in positions n/2 and n/2+1.
Thus the sample median of a set of three values is the second smallest; of a set of four values, it is the average of the second and third smallest. The sample mean and sample median are both useful statistics for describing the central tendency of a data set. The sample mean makes use of all the data values and is affected by extreme values that are much larger or smaller than the others; the sample median makes use of only one or two of the middle values and is thus not affected by extreme values. Which of them is more useful depends on what one is trying to learn from the data. For instance, if a city government has a flat rate income tax and is trying to estimate its total revenue from the tax, then the sample mean of its residents' income would be a more useful statistic. On the other hand, if the city was thinking about constructing middle-income housing, and wanted to determine the proportion of its population able to afford it, then the sample median would probably be more useful. Another statistic that has been used to indicate the central tendency of a data set is the sample mode, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called modal values. 2.3.2 Sample Variance and Sample Standard Deviation: Whereas we have presented statistics that describe the central tendencies of a data set, we are also interested in ones that describe the spread or variability of the data values. A statistic that could be used for this purpose would be one that measures the average value of the squares of the distances between the data values and the sample mean. This is accomplished by the sample variance, which for technical reasons divides the sum of the squares of the differences by n−1 rather than n, where n is the size of the data set. Definition: The sample variance, call it s^2 , of the data set x1, . . . , xn is defined by

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$$

The following algebraic identity is often useful for computing the sample variance:
An Algebraic Identity:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

The identity is proven as follows:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \quad = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \quad = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

The computation of the sample variance can also be eased by noting that if

$$y_i = a + bx_i, \qquad i = 1, \ldots, n \quad \text{then } \bar{y} = a + b\bar{x}, \text{ and so } \sum_{i=1}^{n} (y_i - \bar{y})^2 = b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

That is, if s_y^2 and s_x^2 are the respective sample variances, then $s_y^2 = b^2 s_x^2$ In other words, adding a constant to each data value does not change the sample variance; whereas multiplying each data value by a constant results in a new sample variance that is equal to the old one multiplied by the square of the constant. The positive square root of the sample variance is called the sample standard deviation. Definition: The quantity s, defined by

$$s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)}$$

is called the sample standard deviation.
The sample standard deviation is measured in the same units as the data. 2.3.3 Sample Percentiles and Box Plots: Loosely speaking, the sample 100p percentile of a data set is that value such that 100p percent of the data values are less than or equal to it, 0 ≤ p ≤ 1. More formally, we have the following definition. Definition: The sample 100p percentile is that data value such that 100p percent of the data are less than or equal to it and 100(1−p) percent are greater than or equal to it. If two data values satisfy this condition, then the sample 100p percentile is the arithmetic average of these two values. To determine the sample 100p percentile of a data set of size n, we need to determine the data values such that
1. At least np of the values are less than or equal to it.
2. At least n(1−p) of the values are greater than or equal to it.
To accomplish this, first arrange the data in increasing order. Then, note that if np is not an integer, then the only data value that satisfies the preceding conditions is the one whose position when the data are ordered from smallest to largest is the smallest integer exceeding np. For instance, if n=22, p = .8, then we require a data value such that at least 17.6 of the values are less than or equal to it, and at least 4.4 of them are greater than or equal to it. Clearly, only the 18th smallest value satisfies both conditions and this is the sample 80 percentile. On the other hand, if np is an integer, then it is easy to check that both the values in positions np and np+1 satisfy the preceding conditions, and so the sample 100p percentile is the average of these values.
The sample 50 percentile is, of course, just the sample median. Along with the sample 25 and 75 percentiles, it makes up the sample quartiles. Definition
The sample 25 percentile is called the first quartile; the sample 50 percentile is called the sample median or the second quartile; the sample 75 percentile is called the third quartile. The quartiles break up a data set into four parts, with roughly 25 percent of the data being less than the first quartile, 25 percent being between the first and second quartile, 25 percent being between the second and third quartile, and 25 percent being greater than the third quartile. EXAMPLE 2.3i Noise is measured in decibels, denoted as dB. One decibel is about the level of the weakest sound that can be heard in a quiet surrounding by someone with good

hearing; a whisper measures about 30 dB; a human voice in normal conversation is about 70 dB; a loud radio is about 100 dB. Ear discomfort usually occurs at a noise level of about 120 dB. The following data give noise levels measured at 36 different times directly outside of Grand Central Station in Manhattan.

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85
69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

Determine the quartiles.

SOLUTION A stem and leaf plot of the data is as follows:

6 0, 5, 5, 8, 9
7 2, 4, 4, 5, 7, 8
8 2, 3, 3, 5, 7, 8, 9
9 0, 0, 1, 4, 4, 5, 7
10 0, 2, 7, 8
11 0, 2, 4, 5
12 2, 4, 5

The first quartile is 74.5, the average of the 9th and 10th smallest data values; the second quartile is 89.5, the average of the 18th and 19th smallest values; the third quartile is 104.5, the average of the 27th and 28th smallest values.

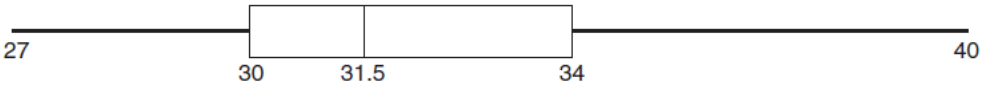


FIGURE 2.7 *A box plot.*

A box plot is often used to plot some of the summarizing statistics of a data set. A straight line segment stretching from the smallest to the largest data value is drawn on a horizontal axis; imposed on the line is a "box," which starts at the first and continues to the third quartile, with the value of the second quartile indicated by a vertical line. For instance, the 42 data values presented in Table 2.1 go from a low value of 27 to a high value of 40. The value of the first quartile (equal to the value of the 11th smallest on the list) is 30; the value of the second quartile (equal to the average of the 21st and 22nd smallest values) is 31.5; and the value of the third quartile (equal to the value of the 32nd smallest on the list) is 34. The box plot for this data set is shown in Figure 2.7. The length of the line segment on the box plot, equal to the largest minus the smallest data value, is called the range of the data. Also, the length of the box itself, equal to the third quartile minus the first quartile, is called the interquartile range.

2.4 CHEBYSHEV'S INEQUALITY: Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set. Assuming that s >0, Chebyshev's inequality states that for any value of k ≥ 1, greater than 100(1−1/k2) percent of the data lie within the interval from $\bar{x}$ −ks to $\bar{x}$ +ks. Thus, by letting k =3/2, we obtain from Chebyshev's inequality that greater than 100(5/9)=55.56 percent of the data from any data set lies within a distance 1.5s of the sample mean $\bar{x}$; letting k =2 shows that greater than 75 percent of the data lies within 2s of the sample mean; and letting k =3 shows that greater than 800/9≈88.9 percent of the data lies within 3 sample standard deviations of $\bar{x}$. When the size of the data set is specified, Chebyshev's inequality can be sharpened, as indicated in the following formal statement and proof.

Chebyshev's Inequality: Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data x1, . . . , xn, where s >0. Let

$$S_k = \{i, 1 \le i \le n : |x_i - \bar{x}| < ks\}$$

and let $N(S_k)$ be the number of elements in the set S_k. Then, for any k ≥ 1,

$$\frac{N(S_k)}{n} \ge 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k_2}$$

Proof:

$$(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i \in S_k}(x_i - \bar{x})^2 + \sum_{i \notin S_k}(x_i - \bar{x})^2 \ge \sum_{i \notin S_k}(x_i - \bar{x})^2$$

$$\ge \sum_{i \notin S_k} k^2 s^2 = k^2 s^2 (n - N(S_k))$$

where the first inequality follows because all terms being summed are nonnegative, and the second follows since $(x_1 - \bar{x})^2 \ge k^2 s^2$ when $i \notin S_k$. Dividing both sides of the preceding inequality by $nk^2 s^2$ yields that $\frac{n-1}{nk^2} \ge 1 - \frac{N(S_k)}{n}$ and the result is proven.

Because Chebyshev's inequality holds universally, it might be expected for given data that the actual percentage of the data values that lie within the interval from $\bar{x}$–ks to $\bar{x}$+ks might be quite a bit larger than the bound given by the inequality. Suppose now that we are interested in the fraction of data values that exceed the sample mean by at least k sample standard deviations, where k is positive. That is, suppose that $\bar{x}$ and s are the sample mean and the sample standard deviation of the data set x1, x2, . . . , xn. Then, with

$$N(k) = \text{number of } i : x_i - \bar{x} \ge ks$$

what can we say about N(k)/n? Clearly,

$$\frac{N(k)}{n} \le \frac{\text{number of } i : |x_i - \bar{x}| \ge ks}{n}$$

$$\le \frac{1}{k^2} \quad \text{by Chebyshev's inequality}$$

However, we can make a stronger statement, as is shown in the following one-sided version of Chebyshev's inequality.

The One-Sided Chebyshev Inequality: For k >0,

$$\frac{N(k)}{n} \le \frac{1}{1+k^2}$$

Proof: Let yi = xi – $\bar{x}$, i = 1, . . . , n. For any b >0, we have that

$$\sum_{i=1}^{n}(y_i + b)^2 \ge \sum_{i:y_i \ge ks}(y_i + b)^2 \ge \sum_{i:y_i \ge ks}(ks + b)^2 = N(k)(ks + b)^2 \quad (2.4.1)$$

where the first inequality follows because $(y_i + b)^2 \ge 0$, and the second because both ks and b are positive. However,

$$\sum_{i=1}^{n}(y_i + b)^2 = \sum_{i=1}^{n}(y_i^2 + 2by_i + b^2) = \sum_{i=1}^{n} y_i^2 + 2b\sum_{i=1}^{n} y_i + nb^2$$

$$= (n-1)s^2 + nb^2$$

where the final equation used that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0.$ Therefore, we obtain from Equation (2.4.1) that

$$N(k) \le \frac{(n-1)s^2 + nb^2}{(ks+b)^2}$$

implying that

$$\frac{N(k)}{n} \le \frac{s^2 + b^2}{(ks+b)^2}$$

Because the preceding is valid for all b >0, we can set b = s/k (which is

the value of b that minimizes the right-hand side of the preceding) to obtain that $\frac{N(k)}{n} \leq \frac{s^2 + s^2/k^2}{(ks + s/k)^2}$ Multiplying the numerator and the denominator of the right side of the preceding by k2/s2

Gives $\frac{N(k)}{n} \leq \frac{k^2 + 1}{(k^2 + 1)^2} = \frac{1}{k^2 + 1}$ and the result is proven. Thus, for instance, where the usual Chebyshev inequality shows that at most 25 percent of data values are at least 2 standard deviations greater than the sample mean, the one-sided Chebyshev inequality lowers the bound to "at most 20 percent."

2.5 NORMAL DATA SETS: Many of the large data sets observed in practice have histograms that are similar in shape. These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion. Such data sets are said to be normal and their histograms are called normal histograms. Figure 2.8 is the histogram of a normal data set. If the histogram of a data set is close to being a normal histogram, then we say that the data set is approximately normal. For instance, we would say that the histogram given in Figure 2.9 is from an approximately normal data set, whereas the ones presented in Figures 2.10 and 2.11 are not (because each is too nonsymmetric). Any data set that is not approximately symmetric about its sample median is said to be skewed. It is "skewed to the right" if it has a long tail to the right and "skewed to the left" if it has a long tail to the left. Thus the data set presented in Figure 2.10 is skewed to the left and the one of Figure 2.11 is skewed to the right. It follows from the symmetry of the normal histogram that a data set that is approximately normal will have its sample mean and sample median approximately equal.

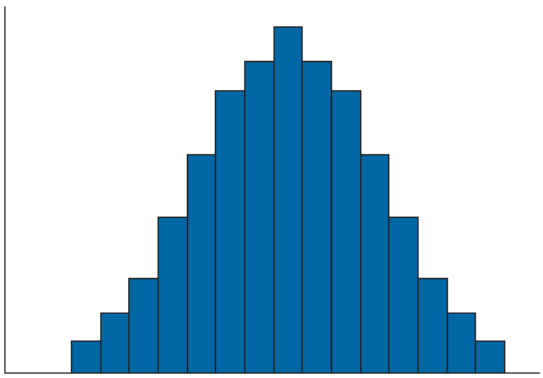


FIGURE 2.8 *Histogram of a normal data set.*

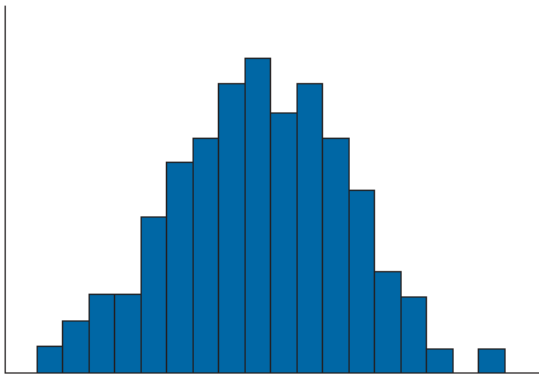


FIGURE 2.9 *Histogram of an approximately normal data set.*

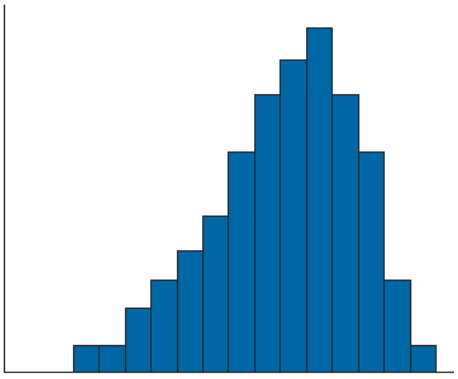


FIGURE 2.10 *Histogram of a data set skewed to the left.*

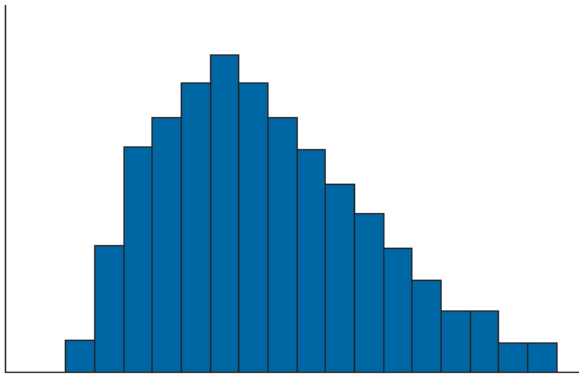


FIGURE 2.11 *Histogram of a data set skewed to the right.*

Suppose that $\bar{x}$ and s are the sample mean and sample standard deviation of an approximately normal data set. The following rule, known as the empirical rule, specifies the approximate proportions of the data observations that are within s, 2s, and 3s of the sample mean $\bar{x}$. The Empirical Rule :

If a data set is approximately normal with sample mean $\bar{x}$ and sample standard deviation , then the following statements are true.

1. Approximately 68 percent of the observations lie within $\bar{x} \pm s$
2. Approximately 95 percent of the observations lie within $\bar{x} \pm 2s$
3. Approximately 99.7 percent of the observations lie within $\bar{x} \pm 3s$ A data set that is obtained by sampling from a population that is itself made up of subpopulations of different types is usually not normal. Rather, the histogram from such a data set often appears to resemble a combining, or superposition, of normal histograms and thus will often have more than one local peak or hump. Because the histogram will be higher at these local peaks than at their neighboring values, these peaks are similar to modes. A data set whose histogram has two local peaks is said to be bimodal. The data set represented in Figure 2.12 is bimodal.

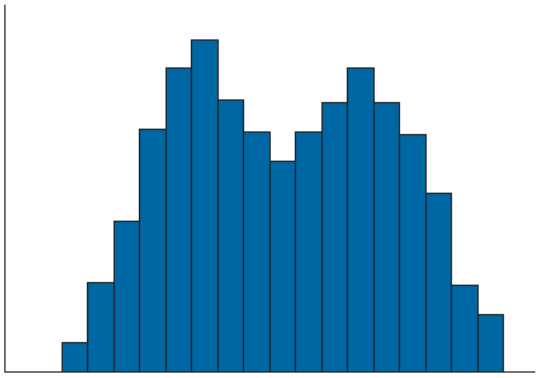


FIGURE 2.12 *Histogram of a bimodal data set.*

2.6 PAIRED DATA SETS AND THE SAMPLE CORRELATION COEFFICIENT: We are often concerned with data sets that consist of pairs of values that have some relationship to each other. If each element in such a data set has an x value and a y value, then we represent the ith data point by the pair (xi , yi). For instance, in an attempt to determine the relationship between the daily midday temperature (measured in degrees Celsius) and the number of defective parts produced during that day, a company recorded the data presented in Table 2.8. For this data set, xi represents the temperature in degrees Celsius and yi the number of defective parts

TABLE 2.8 *Temperature and Defect Data*

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

produced on day i. A useful way of portraying a data set of paired values is to plot the data on a two-dimensional graph, with the x-axis representing the x value of the data and the y-axis representing the y value. Such a plot is called a scatter diagram. Figure 2.13 presents a scatter diagram for the data of Table 2.8.

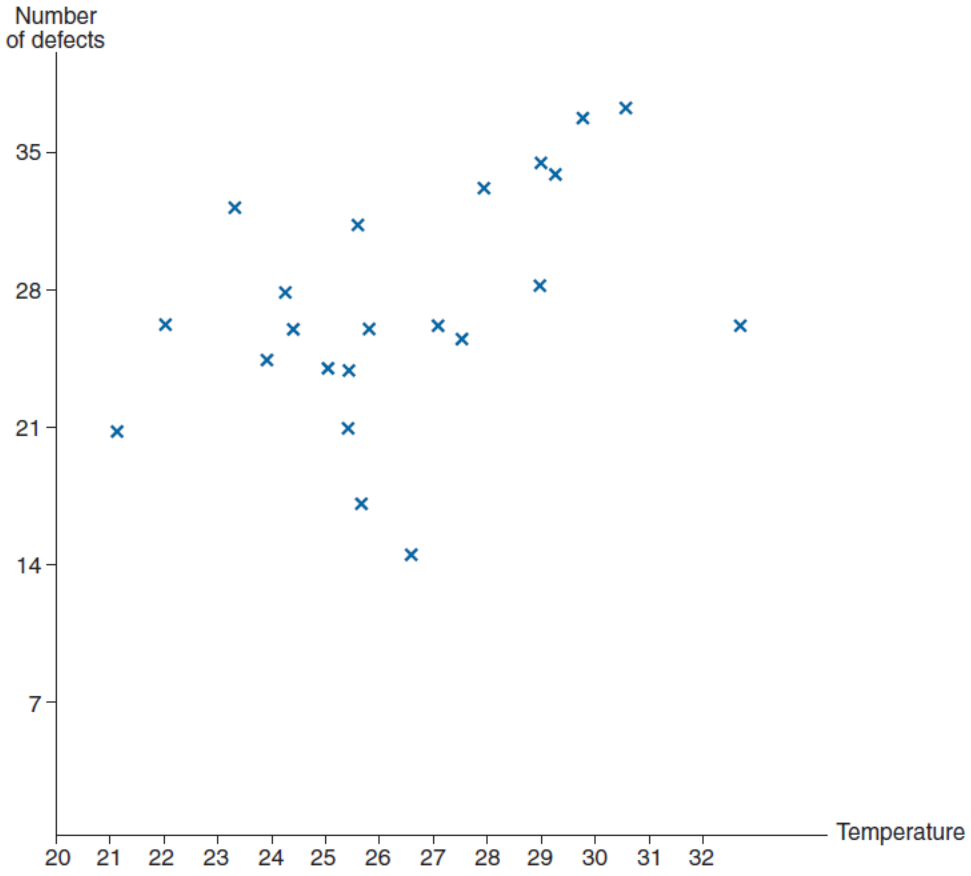


FIGURE 2.13 *A scatter diagram.*

A question of interest concerning paired data sets is whether large x values tend to be paired with large y values, and small x values with small y values; if this is not the case, then we might question whether large values of one of the variables tend to be paired with small values of the other. A rough answer to these questions can often be provided by the scatter diagram. For instance, Figure 2.13 indicates that there appears to be some connection between high temperatures and large numbers of defective items. To obtain a quantitative measure of this relationship, we now develop a statistic that attempts to measure the degree to which larger x values go with larger y values and smaller x values with smaller y values. Suppose that the data set consists of the paired values (xi , yi), i =1, . . . , n. To obtain a statistic that can be used to measure the association between the individual values of a set of paired data, let $\bar{x}$ and $\bar{y}$ denote the sample means of the x values and the y values, respectively. For data pair i, consider xi – $\bar{x}$ the deviation of its x value from the sample mean, and yi – $\bar{y}$ the deviation of its y value from the sample mean. Now if xi is a large x value, then it will be larger than the

average value of all the x's, so the deviation xi – $\bar{y}$ will be a positive value. Similarly, when xi is a small x value, then the deviation xi – $\bar{x}$ will be a negative value. Because the same statements are true about the y deviations, we can conclude the following:
When large values of the x variable tend to be associated with large values of the y variable and small values of the x variable tend to be associated with small values of the y variable, then the signs, either positive or negative, of xi – $\bar{x}$ and yi – $\bar{y}$ will tend to be the same.
Now, if xi – $\bar{x}$ and yi – $\bar{y}$ both have the same sign (either positive or negative), then their product (xi – $\bar{x}$)(yi – $\bar{y}$) will be positive. Thus, it follows that when large x values tend to be associated with large y values and small x values are associated with small y values, then $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ will tend to be a large positive number. [In fact, not only will all the products have a positive sign when large (small) x values are paired with large (small) y values, but it also follows from a mathematical result known as Hardy's lemma that the largest possible value of the sum of paired products will be obtained when the largest xi – $\bar{x}$ is paired with the largest yi – $\bar{y}$, the second largest xi – $\bar{x}$ is paired with the second largest yi – $\bar{y}$, and so on.] In addition, it similarly follows that when large values of xi tend to be paired with small values of yi then the signs of xi – $\bar{x}$ and yi – $\bar{y}$ will be opposite and so $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ will be a large negative number. To determine what it means for $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ to be "large," we standardize this sum first by dividing by n–1 and then by dividing by the product of the two sample standard deviations. The resulting statistic is called the sample correlation coefficient. Definition: Let sx and sy denote, respectively, the sample standard deviations of the x values and the y values. The sample correlation coefficient, call it r, of the data pairs (xi , yi), i =1, . . . , n is defined by

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

When r >0 we say that the sample data pairs are positively correlated, and when r <0 we say that they are negatively correlated. The following are properties of the sample correlation coefficient. Properties of r:
1. −1 ≤ r ≤ 1
2. If for constants a and b, with b >0, $y_i=a+bx_i$, i =1, . . . , n then r =1.
3. If for constants a and b, with b <0, $y_i=a+bx_i$, i =1, . . . , n then r =−1.
4. If r is the sample correlation coefficient for the data pairs xi , yi , i =1, . . . , n then it is also the sample correlation coefficient for the data pairs $a+bx_i,\quad c+dy_i$, i =1, . . . , n provided that b and d are both positive or both negative. Property 1 says that the sample correlation coefficient r is always between −1 and +1. Property 2 says that r will equal+1 when there is a straight line (also called a linear) relation between the paired data such that large y values are attached to large x values. Property 3 says that r will equal −1 when the relation is linear and large y values are attached to small x values. Property 4 states that the value of r is unchanged when a constant is added to each of the x variables (or to each of the y variables) or when each x variable (or each y variable) is multiplied by a positive constant. This property implies that r does not depend on the dimensions chosen to measure the data. For instance, the sample correlation coefficient between a person's height and weight does not depend on whether the height is measured in feet or in inches nor whether the weight is measured in pounds or in kilograms. Also, if one of the values in the pair is temperature, then the sample correlation coefficient is the same whether it is measured in Fahrenheit or in Celsius. The absolute value of the sample correlation coefficient r (that is, |r|, its value without regard to its sign) is a measure of the strength of the linear relationship between the x and the y values of a data pair. A value of |r| equal to 1 means that there is a perfect linear relation—that is, a straight line can pass through all the data points (xi , yi), i =1, . . . , n. A value of |r| of around .8 means that the linear relation is relatively strong; although there is no straight line that passes through all of the data points, there is one that is "close" to them all. A value for |r| of around .3 means that the linear relation is relatively weak. The sign of r gives the direction of the relation. It is positive when the linear relation is such that smaller y values tend to go with smaller x values and larger y values with larger x values (and so a straight line approximation points upward), and it is negative when larger y values tend to go with smaller x values and smaller y values with larger x values (and so a straight line approximation points downward). Figure 2.14 displays scatter diagrams for data sets with various values of r.

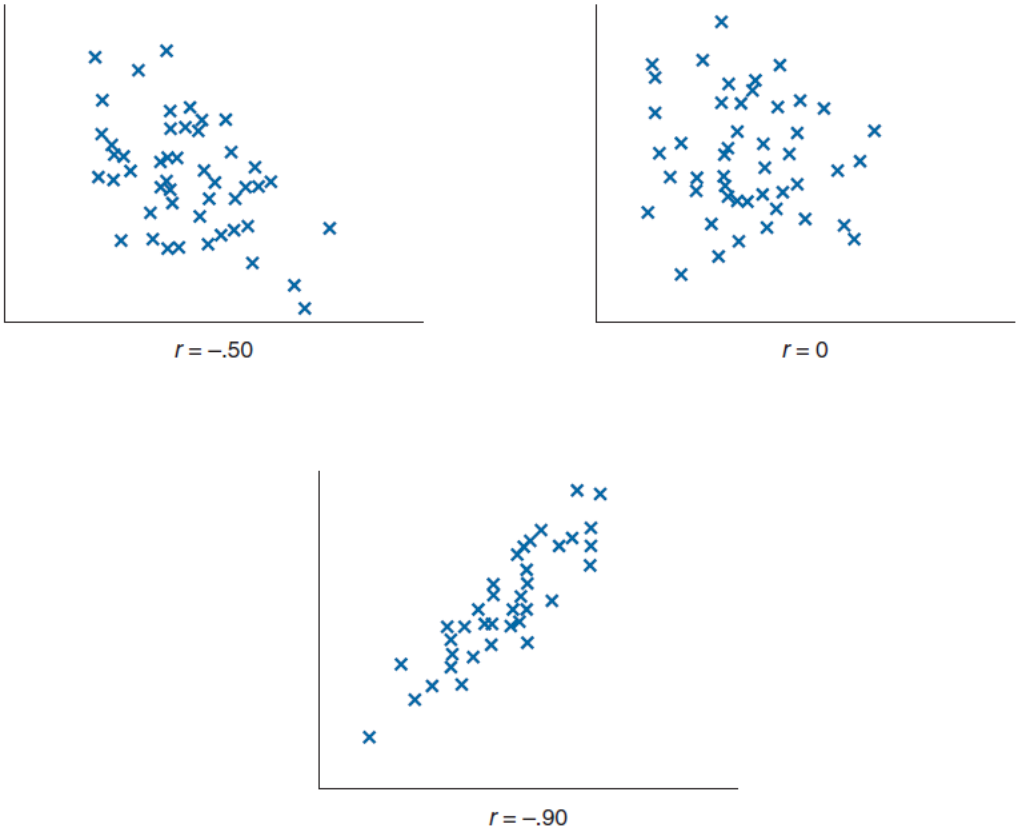


FIGURE 2.14 *Sample correlation coefficients.*

EXAMPLE 2.6b The following data give the resting pulse rates (in beats per minute) and the years of schooling of 10 individuals. A scatter diagram of these data is presented in Figure 2.15. The sample correlation coefficient for these data is r =–.7638. This negative correlation indicates that for this data set a high pulse rate is strongly associated with a small number of years in school, and a low pulse rate with a large number of years in school.

Person 1 2 3 4 5 6 7 8 9 10
Years of School 12 16 13 18 19 12 18 19 12 14
Pulse Rate 73 67 74 63 73 84 60 62 76 71

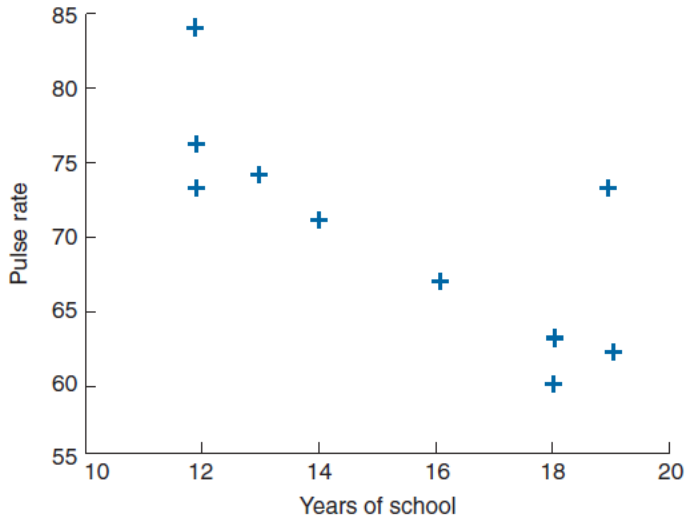


FIGURE 2.15 *Scatter diagram of years in school and pulse rate.*

Correlation Measures Association, Not Causation: The results of Example 2.6b indicate a strong negative correlation between an individual's years of education and that individual's resting pulse rate. However, this does not imply that additional years of school will directly reduce one's pulse rate. That is, whereas additional years of school tend to be associated with a lower resting pulse rate, this does not mean that it is a direct cause of it. Often, the explanation for such an association lies with an unexpressed factor that is related to both variables under consideration. In this instance, it may be that a person who has spent additional time in school is more aware of the latest findings in the area of health, and thus may be more aware of the importance of exercise and good nutrition; or it may be that it is not knowledge that is making the difference but rather it is that people who have had more education tend to end up in jobs that allow them more time for exercise and money for good nutrition. The strong negative correlation between years in school and resting pulse rate probably results from a combination of these as well as other underlying factors.

Chapter_1: INTRODUCTION TO STATISTICS: 1.1 INTRODUCTION: It has become accepted in today's world that in order to learn about something, you must first collect data. Statistics is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions. 1.2 DATA COLLECTION AND DESCRIPTIVE STATISTICS: Sometimes a statistical analysis begins with a given set of data: For instance, the government regularly collects and publicizes data concerning yearly precipitation totals, earthquake occurrences, the unemployment rate, the gross domestic product, and the rate of inflation. Statistics can be used to describe, summarize, and analyze these data. In other situations, data are not yet available; in such cases statistical theory can be used to design an appropriate experiment to generate data. The experiment chosen should depend on the use that one wants to make of the data. For instance, suppose that an instructor is interested in determining which of two different methods for teaching computer programming to beginners is most effective. To study this question, the instructor might divide the students into two groups, and use a different teaching method for each group.

At the end of the class the students can be tested and the scores of the members of the different groups compared. If the data, consisting of the test scores of members of each group, are significantly higher in one of the groups, then it might seem reasonable to suppose that the teaching method used for that group is superior. It is important to note, however, that in order to be able to draw a valid conclusion from the data, it is essential that the students were divided into groups in such a manner that neither group was more likely to have the students with greater natural aptitude for programming. For instance, the instructor should not have let the male class members be one group and the females the other. For if so, then even if the women scored significantly higher than the men, it would not be clear whether this was due to the method used to teach them, or to the fact that women may be inherently better than men at learning programming skills. The accepted way of avoiding this pitfall is to divide the class members into the two groups "at random." This term means that the division is done in such a manner that all possible choices of the members of a group are equally likely. At the end of the experiment, the data should be described. For instance, the scores of the two groups should be presented. In addition, summary measures such as the average score of members of each of the groups should be presented. This part of statistics, concerned with the description and summarization of data, is called descriptive statistics. 1.3 INFERENTIAL STATISTICS AND PROBABILITY MODELS: After the preceding experiment is completed and the data are described and summarized, we hope to be able to draw a conclusion about which teaching method is superior. This part of statistics, concerned with the drawing of conclusions, is called inferential statistics. To be able to draw a onclusion from the data, we must take into account the possibility of chance. For instance, suppose that the average score of members of the first group is quite a bit higher than that of the second. Can we conclude that this increase is due to the teaching method used? Or is it possible that the teaching method was not responsible for the increased scores but rather that the higher scores of the first group were just a chance occurrence? For instance, the fact that a coin comes up heads 7 times in 10 flips does not necessarily mean that the coin is more likely to come up heads than tails in future flips. Indeed, it could be a perfectly ordinary coin that, by chance, just happened to land heads 7 times out of the total of 10 flips. (On the other hand, if the coin had landed heads 47 times out of 50 flips, then we would be quite certain that it was not an ordinary coin.) To be able to draw logical conclusions from data, we usually make some assumptions about the chances (or probabilities) of obtaining the different data values. The totality of these assumptions is referred to as a probability model for the data. Sometimes the nature of the data suggests the form of the probability model that is assumed. For instance, suppose that an engineer wants to find out what proportion of computer chips, produced by a new method, will be defective. The engineer might select group of these chips, with the resulting data being the number of defective chips in this group. Provided that the chips selected were "randomly" chosen, it is reasonable to suppose that each one of them is defective with probability p, where p is the unknown proportion of all the chips produced by the new method that will be defective. The resulting data can then be used to make inferences about p. In other situations, the appropriate probability model for a given data set will not be readily apparent. However, careful description and presentation of the data sometimes enable us to infer a reasonable model, which we can then try to verify with the use of additional data. Because the basis of statistical inference is the formulation of a probability model to describe the data, an understanding of statistical inference requires some knowledge of the theory of probability. In other words, statistical inference starts with the assumption that important aspects of the phenomenon under study can be described in terms of probabilities; it then draws conclusions by using data to make inferences about these probabilities. 1.4 POPULATIONS AND SAMPLES: In statistics, we are interested in obtaining information about a total collection of elements, which we will refer to as the population. The population is often too large for us to examine each of its members. For instance, we might have all the residents of a given state, or all the television sets produced in the last year by a particular manufacturer, or all the households in a given community. In such cases, we try to learn about the population by choosing and then examining a subgroup of its elements. This subgroup of a population is called a sample. If the sample is to be informative about the total population, it must be, in some sense, representative of that population. For instance, suppose that we are interested in learning about the age distribution of people residing in a given city, and we obtain the ages of the first 100 people to enter the town library. If the average age of these 100 people is 46.2 years, are we justified in concluding that this is approximately the average age of the entire population? Probably not, for we could certainly argue that the sample chosen in this case is probably not representative of the total population because usually more young students and senior citizens use the library than do working-age citizens. In certain situations, such as the library illustration, we are presented with a sample and must then decide whether this sample is reasonably representative of the entire population. In practice, a given sample generally cannot be assumed to be representative of a populationunless that sample has been chosen in a random manner. This is because any specific non random rule for selecting a sample often results

in one that is inherently biased toward some data values as opposed to others. Thus, although it may seem paradoxical, we are most likely to obtain a representative sample by choosing its members in a totally random fashion without any prior considerations of the elements that will be chosen. In other words, we need not attempt to deliberately choose the sample so that it contains, for instance, the same gender percentage and the same percentage of people in each profession as found in the general population. Rather, we should just leave it up to "chance" to obtain roughly the correct percentages. Once a random sample is chosen, we can use statistical inference to draw conclusions about the entire population by studying the elements of the sample. 1.5 A BRIEF HISTORY OF STATISTICS: A systematic collection of data on the population and the economy was begun in the Italian city states of Venice and Florence during the Renaissance. The term statistics, derived from the word state, was used to refer to a collection of facts of interest to the state. The idea of collecting data spread from Italy to the other countries of Western Europe. Indeed, by the first half of the 16th century it was common for European governments to require parishes
to register births, marriages, and deaths. Because of poor public health conditions this last statistic was of particular interest. The high mortality rate in Europe before the 19th century was due mainly to epidemic diseases, wars, and famines. Among epidemics, the worst were the plagues. Starting with the Black Plague in 1348, plagues recurred frequently for nearly 400 years. In 1562, as a way to alert the King's court to consider moving to the countryside, the City of London
began to publish weekly bills of mortality. Initially these mortality bills listed the places of death and whether a death had resulted from plague. Beginning in 1625 the bills were expanded to include all causes of death. In 1662 the English tradesman John Graunt published a book entitled Natural and Political Observations Made upon the Bills of Mortality. Table 1.1, which notes the total number of deaths in England and the number due to the plague for five different plague years, is taken from this book.

TABLE 1.1 *Total Deaths in England*

Year	Burials	Plague Deaths
1592	25,886	11,503
1593	17,844	10,662
1603	37,294	30,561
1625	51,758	35,417
1636	23,359	10,400

Source: John Graunt, Observations Made upon the Bills of Mortality. 3rd ed. London: John Martyn and James Allestry (1st ed. 1662).

Graunt used London bills of mortality to estimate the city's population. For instance, to estimate the population of London in 1660, Graunt surveyed households in certain London parishes (or neighborhoods) and discovered that, on average, there were approximately 3 deaths for every 88 people. Dividing by 3 shows that, on average, there was roughly 1 death for every 88/3 people. Because the London bills cited 13,200 deaths in London for that year, Graunt estimated the London population to be about 13,200 × 88/3 = 387,200 Graunt used this estimate to project a figure for all England. In his book he noted that these figures would be of interest to the rulers of the country, as indicators of both the number of men who could be drafted into an army and the number who could be taxed. Graunt also used the London bills of mortality—and some intelligent guesswork as to what diseases killed whom and at what age — to infer ages at death. (Recall that the bills of mortality listed only causes and places at death, not the ages of those dying.) Graunt then used this information to compute tables giving the proportion of the population that dies at various ages. Table 1.2 is one of Graunt's mortality tables. It states, for instance, that of 100 births, 36 people will die before reaching age 6, 24 will die between the age of 6 and 15, and so on.

TABLE 1.2 *John Graunt's Mortality Table*

Age at Death	Number of Deaths per 100 Births
0–6	36
6–16	24
16–26	15
26–36	9
36–46	6
46–56	4
56–66	3
66–76	2
76 and greater	1

Note: The categories go up to but do not include the right-hand value. For instance, 0–6 means all ages from 0 up through 5.

Graunt's estimates of the ages at which people were dying were of great interest to those in the business of selling annuities. Annuities are the opposite of life insurance in that one pays in a lump sum as an investment and then receives regular payments for as long as one lives. Graunt's work on mortality tables inspired further work by Edmund Halley in 1693. Halley, the discoverer of the comet bearing his name (and also the man who was most responsible, by both his encouragement and his financial support, for the publication of Isaac Newton's famous Principia Mathematica), used tables of mortality to compute the odds that a person of any age would live to any other particular age. Halley was influential in convincing the insurers of the time that an annual life insurance premium should depend on the age of the person being insured. Following Graunt and Halley, the collection of data steadily increased throughout the remainder of the 17th and on into the 18th century. For instance, the city of Paris began collecting bills of mortality in 1667; and by 1730 it had become common practice throughout Europe to record ages at death. The term statistics, which was used until the 18th century as a shorthand for the descriptive science of states, became in the 19th century increasingly identified with numbers. By the 1830s the term was almost universally regarded in Britain and France as being synonymous with the "numerical science" of society. This change in meaning was caused by the large availability of census records and other tabulations that began to be systematically collected and published by the governments of Western Europe and the
United States beginning around 1800. Throughout the 19th century, although probability theory had been developed by such mathematicians as Jacob Bernoulli, Karl Friedrich Gauss, and Pierre-Simon Laplace, its use in studying statistical findings was almost nonexistent, because most social statisticians at the time were content to let the data speak for themselves. In particular, statisticians of that time were not interested in drawing inferences about individuals, but rather were
concerned with the society as a whole. Thus, they were not concerned with sampling but rather tried to obtain censuses of the entire population. As a result, probabilistic inference from samples to a population was almost unknown in 19th century social statistics. It was not until the late 1800s that statistics became concerned with inferring conclusions from numerical data. The movement began with Francis Galton's work on analysing hereditary genius through the uses of what we would now call regression and correlation analysis (see Chapter 9), and obtained much of its impetus from the work of Karl Pearson. Pearson, who developed the chi-square goodness of fit tests (see Chapter 11), was the first director of the Galton Laboratory, endowed by Francis Galton in 1904. There Pearson originated a research program aimed at developing new methods of using statistics in inference. His laboratory invited advanced students from science and industry to learn statistical methods that could then be applied in their fields. One of his earliest visiting researchers was W. S. Gosset, a chemist by training, who showed his devotion to Pearson by publishing his own works under the name "Student." (A famous story has it that Gosset was afraid to publish under his own name for fear that his employers, the Guinness brewery, would be unhappy to discover that one of its chemists was doing research in statistics.) Gosset is famous for his development of the t-test (see Chapter 8). Two of the most important areas of applied statistics in the early 20th century were population biology and agriculture. This was due to the interest of Pearson and others at his laboratory and also to the remarkable accomplishments of the English scientist Ronald
A. Fisher. The theory of inference developed by these pioneers, including among others Karl Pearson's son Egon and the Polish born mathematical statistician Jerzy Neyman, was general enough to deal with a wide range of quantitative and practical problems. As a result, after the early years of the 20th century a rapidly increasing number of people in science, business, and government began to regard statistics as a tool that was able to provide quantitative solutions to scientific and practical problems (see Table 1.3). Nowadays the ideas of statistics are everywhere. Descriptive statistics are featured in every newspaper and magazine. Statistical inference has

become indispensable to public health and medical research, to engineering and scientific studies, to marketing and quality control, to education, to accounting, to economics, to meteorological forecasting, to polling and surveys, to sports, to insurance, to gambling, and to all research that makes any claim to being scientific. Statistics has indeed become ingrained in our intellectual heritage.

TABLE 1.3 *The Changing Definition of Statistics*

Statistics has then for its object that of presenting a faithful representation of a state at a determined epoch. (Quetelet, 1849)

Statistics are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man. (Galton, 1889)

Statistics may be regarded (i) as the study of populations, (ii) as the study of variation, and (iii) as the study of methods of the reduction of data. (Fisher, 1925)

Statistics is a scientific discipline concerned with collection, analysis, and interpretation of data obtained from observation or experiment. The subject has a coherent structure based on the theory of Probability and includes many different procedures which contribute to research and development throughout the whole of Science and Technology. (E. Pearson, 1936)

Statistics is the name for that science and art which deals with uncertain inferences — which uses numbers to find out something about nature and experience. (Weaver, 1952)

Statistics has become known in the 20th century as the mathematical tool for analyzing experimental and observational data. (Porter, 1986)

Statistics is the art of learning from data. (this book, 2004)

Sheldon_M_Ross-Introduction_to_Probability_Models_Tenth_Edition, and_ALL_SOLVED_EXAMPLES_OF_CHAPTERS_in_order_4_6_5_1_2_3_TO_BE_PRACTICED/SOLVED_AGAIN_AS_GIVEN_IN_BOOK_TEXT.

Ch-4 Markov Chains:

4.1 Introduction: EXPLANATION_1: Consider a process that has a value in each time period. Let Xn denote its value in time period n, and suppose we want to make a probability model for the sequence of successive values X0,X1,X2 . . .. The simplest model would probably be to assume that the Xn are independent random variables, but often such an assumption is clearly unjustified. For instance, starting at some time suppose that Xn represents the price of one share of some security, such as Google, at the end of n additional trading days. Then it certainly seems unreasonable to suppose that the price at the end of day n+1 is independent of the prices on days n, n − 1, n − 2 and so on down to day 0. However, it might be reasonable to suppose that the price at the end of trading day n + 1 depends on the previous end-of-day prices only through the price at the end of day n. That is, it might be reasonable to assume that the conditional distribution of Xn+1 given all the past end-of-day prices Xn,Xn − 1, . . . ,X0 depends on these past prices only through the price at the end of day n. Such an assumption defines a Markov chain, a type of stochastic process that will be studied in this chapter, and which we now formally define. Let {Xn, n = 0, 1, 2, . . . , } be a stochastic process(SP) that takes on a finite or countable number of possible values. Unless otherwise mentioned, this set of possible values of the process will be denoted by the set of nonnegative integers {0, 1, 2, . . .}. If Xn = i, then the process is said to be in state i(or state T={1,2,…,t}), at time n(or stage n).

$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} = P_{ij}$ (4.1) for all states i0, i1, . . . , in−1, i, j and all $n \geqslant 0$. Such a stochastic process is known as a Markov chain(MC). Equation (4.1) may be interpreted as stating that, for a Markov chain, the conditional distribution of any future state Xn+1, given the past states X0,X1, . . . ,Xn−1 and the present state Xn, is independent of the past states and depends only on the present state. Equation (4.1) may be interpreted as stating that, for a Markov chain, the conditional distribution of any future state Xn+1, given the past states X0,X1, . . . ,Xn−1 and the present state Xn, is independent of the past states and depends only on the present state.

$$P_{ij} \geqslant 0, \quad i,j \geqslant 0; \quad \sum_{j=0}^{\infty} P_{ij} = 1, \quad i = 0, 1, \ldots$$

Let **P** denote the matrix of one-step transition probabilities $P_{i,j}$, so that

$$\mathbf{P} = \begin{Vmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \vdots & \vdots & \vdots & \\ P_{i0} & P_{i1} & P_{i2} & \cdots \\ \vdots & \vdots & \vdots & \end{Vmatrix}$$

EXPLANATION_2: Markov chain is a mathematical model for movement between states. A process starts in one of these states and moves from state to state. The moves between states are called steps or transitions. The terms "chain" and "process" are used interchangeably, so the chain can be said to move between states and to be "at a state" or "in a state" after a certain number of steps. The state of the chain at any given step is not known; what is known is the probability that the chain moves from state j to state i in one step. This probability is called a transition probability for the Markov chain. The transition probabilities are placed in a matrix called the transition matrix P for the chain by entering the probability of a transition from state j to state i at the (i, j)-entry of P. So if there were m states named 1, 2, . . .m, the transition matrix would be the

m × m matrix.

$$P = \begin{array}{c} \text{From:} \\ \begin{matrix} 1 & j & m \end{matrix} \\ \begin{bmatrix} & \vdots & \\ & \downarrow & \\ p_{ij} & \cdots & \rightarrow \end{bmatrix} \end{array} \begin{array}{l} \text{To:} \\ 1 \\ \\ i \\ m \end{array}$$

The probabilities that the chain is in each of the possible states after n steps are listed in a state vector $\mathbf{x}_n$. If there are m possible states the state vector would be

$$\mathbf{x}_n = \begin{bmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_m \end{bmatrix} \longleftarrow \text{Probability that the chain is at state } j \text{ after } n \text{ steps}$$

State vectors are probability vectors since their entries must sum to 1. The state vector $\mathbf{x}_0$ is called the initial probability vector. Notice that the jth column of P is a probability vector – its entries list the probabilities of a move from state j to the states of the Markov chain. The transition matrix is thus a stochastic matrix since all of its columns are probability vectors. The state vectors for the chain are related by the equation $\mathbf{x}_{n+1} = P\mathbf{x}_n$ Equation (1), for n = 1, 2, ... Notice that Equation (1) may be used to show that $\mathbf{x}_n = P^n\mathbf{x}_0$ Equation (2) Thus any state vector $\mathbf{x}_n$ may be computed from the initial probability vector $\mathbf{x}_0$ and an appropriate power of the transition matrix P. This chapter concerns itself with Markov chains with a finite number of states; that is, those chains for which the transition matrix P is of finite size. To use a finite-state Markov chain to model a process, the process must have the following properties, which are implied by Equations (1) and (2).

1. Since the values in the vector depend only on the transition matrix P and on, the state of the chain before time n must have no effect on its state at time n + 1 and beyond.
2. Since the transition matrix P does not change with time, the probability of a transition from one state to another must not depend upon how many steps the chain has taken. Even with these restrictions, Markov chains may be used to model an amazing variety of processes.

EXPLANATION_3: A Markov chain (MC) is a SP such that whenever the process is in state i, there is a fixed transition probability $P_{i,j}$ that its next state will be j. Denote the "current" state (at time n) by $X_n = i$. . Let the event $A = \{X_0 = i_0, X_1 = i_1, \ldots X_{n-1} = i_{n-1}\}$ be the previous history of the MC (before time n). Denote the "current" state (at time n) by $X_n = i$. . A Markov chain is a SP such that Pr(Xn+1 = j|A ∩ Xn = i) = $P_{i,j}$, i.e., the next state depends only on the current state (and is independent of the time). {Xn} has the Markov property if it forgets about its past, i.e., Pr(Xn+1 = j|A ∩ Xn = i) = Pr(Xn+1 = j|Xn = i). {Xn} is time homogeneous if Pr(Xn+1 = j|Xn = i) = Pr(X1 = j|X0 = i) = $P_{i,j}$, i.e., if the transition probabilities are independent of n.

Example 4.1 (Forecasting the Weather): Suppose that the chance of rain tomorrow depends on previous weather conditions only through whether or not it is raining today and not on past weather conditions. Suppose also that if it rains today, then it will rain tomorrow with probability α; and if it does not rain today, then it will rain tomorrow with probability β. If we say that the process is in state 0 when it rains and state 1 when it does not rain, then the preceding is a two-state Markov chain whose transition probabilities are given by $\mathbf{P} = \begin{Vmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{Vmatrix}$ Alpha=probability-of-being-in-State0-Rain. Beta=probability-of-being-in-State0-Rain. State1-is-notRain. [[[[Note:This-**Example-4.1**-is-1-step-matrix-**CONDITIONAL**____Link-this-with-n-step-matrix-**CONDITIONAL-Example4.8**____See-the-difference-**UNCONDITIONAL-supposing-Initial-Probabilities-of-being-in-States-at-time-n**-is-**Known**-Page-200-**InstanceExample**]]]]

Example 4.2 (A Communications System): Consider a communications system that transmits the digits 0 and 1. Each digit transmitted must pass through several stages, at each of which there is a probability p that the digit entered will be unchanged when it leaves. Letting Xn denote the digit entering the nth stage, then {Xn, n = 0, 1, . . .} is a two-state Markov chain having a transition probability matrix

$$\mathbf{P} = \begin{Vmatrix} p & 1-p \\ 1-p & p \end{Vmatrix}$$

In other words (see Figure1) below, Suppose that each bit of data is either a 0 or a 1, and at each stage there is a probability **p** that the bit will pass through the stage **unchanged**. Thus the probability is **1 – p** that the bit will be **transposed**. The transmission process is modeled by a Markov chain, with states 0 and 1 and transition matrix. **Explanation_for_obtaining**_probability_transition_matrix: [[[[Xn denote the digit entering the nth stage, where there is a probability p that the digit entered will be unchanged at stage_0_in_Figure1_below or stage_1_in_Figure1_below when it leaves.]]]]

From: 0 1 To:

$$P = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix} \begin{matrix} 0 \\ 1 \end{matrix}$$

Transition diagram for signal transmission:

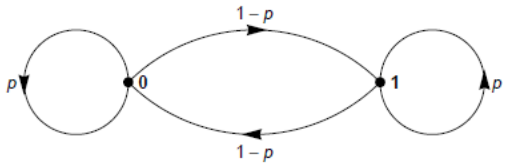


Figure 1: Transition diagram for signal transmission.

Example 4.3 On any given day Gary is either cheerful (C), so-so (S), or glum (G). If he is cheerful today, then he will be C, S, or G tomorrow with respective probabilities 0.5, 0.4, 0.1. If he is feeling so-so today, then he will be C, S, or G tomorrow with probabilities 0.3, 0.4, 0.3. If he is glum today, then he will be C, S, or G tomorrow with probabilities 0.2, 0.3, 0.5.Letting Xn denote Gary's mood on the nth day, then {Xn, $n \geqslant 0$ } is a three-state Markov chain (state 0 = C, state 1 = S, state 2 = G) with transition probability matrix

$$\mathbf{P} = \begin{Vmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{Vmatrix}$$

Example 4.4 (**Transforming** a Process into a Markov Chain): Suppose that whether or not it rains today depends on previous weather conditions through the last two days. Specifically, suppose that if it has rained for the past two days, then it will rain tomorrow with probability 0.7; if it rained today but not yesterday, then it will rain tomorrow with probability 0.5; if it rained yesterday but not today, then it will rain tomorrow with probability 0.4; if it has not rained in the past two days, then it will rain tomorrow with probability 0.2. If we let the state at time n depend only on whether or not it is raining at time n, then the preceding model is not a Markov chain (why not?). However, we can transform this model into a Markov chain by saying that the state at any time is determined by the weather conditions during both that day and the previous day. In other words, we can say that the process is in

state 0 if it rained both today and yesterday,
state 1 if it rained today but not yesterday,
state 2 if it rained yesterday but not today,
state 3 if it did not rain either yesterday or today.

The preceding would then represent a four-state Markov chain having a transition probability matrix: X0,X1, . . . isn't quite a MC, since the probability that it'll rain tomorrow depends on Xi and Xi−1. We'll transform the process into a Markov Chain by defining the following states in terms of today and yesterday.

0 : Xi−1 = R, Xi = R
1 : Xi−1 = S, Xi = R
2 : Xi−1 = R, Xi = S
3 : Xi−1 = S, Xi = S

Thus, we have, e.g.,

Pr(Xi+1 = R | Xi−1 = R,Xi = R) = P00 = 0.7
Pr(Xi+1 = S | Xi−1 = R,Xi = R) = P02 = 0.3

Using similar reasoning, we get

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0 & 0.3 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0.2 & 0 & 0.8 \end{pmatrix}$$

See Explanatory_**Picture-A**_below_for_**obtatining**_probability_transition_matrix:

Yesterday,today Combined into a single state to transform into a Markov chain. **Picture-A**

State	yesterday(y)	today(t)
1	Rain(R)	Rain(R)
2	Sun(S)	Rain(R)
3	Rain(R)	Sun(S)
4	Sun(S)	Sun(S)
	Xi-1	Xi

Transition to next state,i.e., matching_yesterday_with_today_to_make_it_as_a_transition_from_a_single_combined_state_to_another.

State1 transition to State 1: R[R R] Pr=0.7 State1 transition to State 2: R[R S] R Pr=0 State1 transition to State 3: R[R R] S Pr=0.3 State1 transition to State 4: R[R S] S Pr=0

State2 transition to state 1: S[R R] R Pr=0.5 State2 transition to State 2:S[R S] R Pr=0 State2 transition to State 3:S[R R] S Pr=0.5 State2 transition to State 4: S[R S] S Pr=0

State3 transition to state 1: R[S R] R Pr=0 State3 transition to State 2:R[S S] R Pr=0.4 State3 transition to State 3:R[S R] S Pr=0 State3 transition to State 4:R[S S] S Pr=0.6

State4 transition to state 1: S[S R] R Pr=0 State4 transition to State 2:S[S S] R Pr=0.2 State4 transition to State 3:S[S R] S Pr=0 State4 transition to State 4:S[S S] S Pr=0.8

Example 4.5 (A Random Walk Model) A Markov chain whose state space is given by the **integers** i = 0,±1,±2, . . . is said to be a random walk if, for some number 0 < p < 1, Pi,i+1 = p = 1 − Pi,i−1, i = 0,±1, . . .The preceding Markov chain is called a random walk for we may think of it as being a model for an individual walking on a straight line who at each point of time either takes one step to the right with probability p or one step to the left with probability 1 − p. A MC whose state space is given by the integers(coordinates) is called a random walk if Pi,i+1 = p and Pi,i−1 = 1 − p.

$$\mathbf{P} = \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1-p & 0 & p & 0 & 0 & \cdots \\ \cdots & 0 & 1-p & 0 & p & 0 & \cdots \\ \cdots & 0 & 0 & 1-p & 0 & p & \cdots \\ \cdots & 0 & 0 & 0 & 1-p & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

Example 4.6 (A Gambling Model) Consider a gambler who, at each play of the game, either wins $1 with probability p or loses $1 with probability 1 − p. If we suppose that our gambler quits playing either when he goes broke or he attains a fortune of $N, then the gambler's fortune is a Markov chain having transition probabilities Pi,i+1 = p = 1 − Pi,i−1, i = 1, 2, . . . ,N − 1, $P_{00} = P_{NN} = 1$, States 0 and N are called absorbing states since once entered they are never left.
Note that the this is a **finite** state random walk with absorbing barriers (states 0 and N).

Diffusion: Consider two compartments filled with different gases which are separated only by a membrane which allows molecules of each gas to pass from one container to the other. The two gases will then diffuse into each other over time, so that each container will contain some mixture of the gases. The major question of interest is what mixture of gases is in each container at a time after the containers are joined. A famous mathematical model for this process was described originally by the physicists Paul and Tatyana Ehrenfest. Since their preferred term for "container" was urn, the model is called the Ehrenfest urn model for diffusion. Label the two urns A and B, and place k molecules of gas in each urn. At each time step, select one of the 2k molecules at random and move it from its urn to the other urn, and keep track of the number of molecules in urn A. This process can be modeled by a finite-state Markov chain: the number of molecules in urn A after n + 1 time steps depends only on the number in urn A after n time steps, the transition probabilities do not change with time, and the number of states is finite.
For this example, let k = 3. Then the two urns contain a total of 6 molecules, and the possible states for the Markov chain are 0, 1, 2, 3, 4, 5, and 6. Notice first that if there are 0 molecules in urn A at time n, then there must be 1 molecule in urn A at time n+1, and if there are 6 molecules in urn A at time n, then there must be 5 molecules in urn A at time n + 1. In terms of the transition matrix P, this means that the columns in P corresponding to states 0 and 6 are

$$\mathbf{p}_0 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \mathbf{p}_6 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

If there are i molecules in urn A at time n, with 0 < i < 6, then there must be either i − 1 or i + 1 molecules in urn A at time n + 1. In order for a transition from i to i−1 molecules to occur, one of the i molecules in urn A must be selected to move; this event happens with probability i/6.
Likewise a transition from i to i + 1 molecules occurs when one of the 6 − i molecules in urn B is selected, and this occurs with probability (6 − i)/6. Allowing i to range from 1 to 5 creates the columns of P corresponding to these states, and the transition matrix for the Ehrenfest urn model with k = 3 is thus

$$P = \begin{array}{c} \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{bmatrix} 0 & 1/6 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 5/6 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 2/3 & 0 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 5/6 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1/6 & 0 \end{bmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}$$

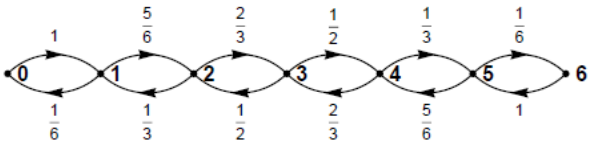


Figure 2: Transition diagram of the Ehrenfest urn model.

Explanation_for_**obtaining**_probability_transition_matrix: [[[[A selected_molecule_i_having_state_i_transitions_to_state_i-1_with_probability_(i/6)_or_transitions_to_state_i+1_with_probability_((6-i)/6)]]]].

Random Walks on {1, . . . , n}: Molecular motion has long been an important issue in physics. Einstein and others investigated Brownian motion, which is a mathematical model for the motion of a molecule exposed to collisions with other molecules. The analysis of Brownian motion turns out to be quite **complicated**, but a discrete version of Brownian motion called a random walk provides an introduction to this important model. Think of the states {1, 2, . . . , n} as lying on a line. Place a molecule at a point that is not on the end of the line. At each step the molecule moves left one unit with probability p and right one unit with probability 1 − p. See Figure 3, below. The molecule thus "walks randomly" along the line. If p = 1/2, the walk is called simple, or **unbiased**. If p $\neq$ 1/2, the walk is said to be **biased**.

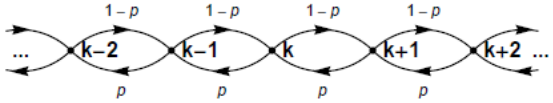


Figure 3: A graphical representation of a random walk.

The molecule must move either to the left or right at the states 2, . . . , n − 1, but it cannot do this at the endpoints 1 and n. The molecule's possible movements at the endpoints 1 and n must be specified. One possibility is to have the molecule stay at an endpoint forever once it reaches either end of the line. This is called a random walk with **absorbing boundaries**, and the endpoints 1 and n are called absorbing states. Another possibility is to have the molecule bounce back one unit when an endpoint is reached. This is called a random walk with **reflecting boundaries**.

For this example, A random walk on {1, 2, 3, 4, 5} with **absorbing boundaries** has a transition matrix of $P = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{matrix} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{bmatrix} 1 & p & 0 & 0 & 0 \\ 0 & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & 0 \\ 0 & 0 & 0 & 1-p & 1 \end{bmatrix} \end{matrix}$ since the molecule at state 1 has probability 1 of staying at state 1, and a molecule at state 5 has probability 1 of staying at state 5. [[[[**Note:** This is similar to Example 4.6 i.e., (A Gambling Model)]]]]
Another possibility is to have the molecule bounce back one unit when an endpoint is reached. This is called a random walk with reflecting boundaries. A random walk on {1, 2, 3, 4, 5} with **reflecting boundaries** has a transition matrix of $P = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{matrix} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{bmatrix} 0 & p & 0 & 0 & 0 \\ 1 & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & 1 \\ 0 & 0 & 0 & 1-p & 0 \end{bmatrix} \end{matrix}$ since the molecule at state 1 has probability 1 of moving to state 2, and a molecule at state 5 has probability 1 of moving to state 4.
In addition to their use in physics, random walks also occur in problems related to gambling and its more socially acceptable variants: the stock market and the insurance industry.
Example 4.7 In most of Europe and Asia annual automobile insurance premiums are determined by use of a Bonus Malus (Latin for Good-Bad) system. Each policyholder is given a positive integer valued state and the annual premium is a function of this state (along, of course, with the type of car being insured and the level of insurance). A policyholder's state changes from year to year in response to the number of claims made by that policyholder. Because lower numbered states correspond to lower annual premiums, a policyholder's state will usually decrease if he or she had no claims in the preceding year, and will generally increase if he or she had at least one claim. (Thus, no claims is good and typically results in a decreased premium, while claims are bad and typically result in a higher premium.) For a given Bonus Malus system, let $s_i(k)$ denote the next state of a policyholder who was in state i in the previous year and who made a total of k claims in that year. If we suppose that the number of yearly claims made by a particular policyholder is a Poisson random variable with parameter λ, then the successive states of this policyholder will constitute a Markov chain with transition probabilities $P_{i,j} = \sum_{k:s_i(k)=j} e^{-\lambda}\frac{\lambda^k}{k!}, \quad j \geqslant 0$ Whereas there are usually many states (20 or so is not atypical), the following table specifies a hypothetical Bonus Malus system having four states.

		Next state if			
State	Annual Premium	0 claims	1 claim	2 claims	⩾ 3 claims
1	200	1	2	3	4
2	250	1	3	4	4
3	400	2	4	4	4
4	600	3	4	4	4

Thus, for instance, the table indicates that s2(0)=1; s2(1)=3; s2(k)=4, $k \geqslant 2$. Consider a policyholder whose annual number of claims is a Poisson random variable with parameter λ. If a_k is the probability that such a policyholder makes k claims in a year, then $a_k = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \geqslant 0$ For the Bonus Malus system specified in the preceding table, the transition probability matrix of the successive states of this policyholder is $\mathbf{P} = \begin{Vmatrix} a_0 & a_1 & a_2 & 1-a_0-a_1-a_2 \\ a_0 & 0 & a_1 & 1-a_0-a_1 \\ 0 & a_0 & 0 & 1-a_0 \\ 0 & 0 & a_0 & 1-a_0 \end{Vmatrix}$ **Explanation_for_obtaining_probability_transition_matrix:** [[[[In row1, one can make 0(k)claim and transition to itself i.e., state 1 with probability a0, or one can make 1(k)-claims and transition to state 2 with probability a1 or make 2(k)claims and transition to state 3 with probability a2, or make 3(k)claimsorMore and transition to state 4 with probability 1-a0-a1-a2. In row 2, one can make 0(k)-claims and transition to state 1 with probability a0 or make 1(k)claim and transition to state 3 with probability a1 or make 2(k)claimsorMore and transition to state 4 with probability 1-a0-a1. In row 3, one can make 0(k)-claims and transition to state 2 with probability a0 or make 1(k)orMore-claims and transition to state 4 with probability 1-a0. In row 4, one can make 0(k)-claims and transition to state 3 with probability a0 or one can make 1(k)orMore claims and transition to state 4 with probability 1-a0.]]]]
4.2 Chapman–Kolmogorov Equations: We have already defined the one-step transition probabilities Pij. We now define the n-step transition probabilities P_{ij}^n to be the probability that a process in state I will be in state j after n additional transitions. That is, $P_{ij}^n = P\{X_{n+k} = j | X_k = i\}, \quad n \geqslant 0, \; i, j \geqslant 0$ Of course $P_{ij}^1 = P_{ij}$ The Chapman–Kolmogorov equations provide a method for computing these n-step transition probabilities. These equations are $P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m \quad \text{for all } n, m \geqslant 0, \text{ all } i, j$ Equation (4.2) and are most easily understood by noting that $P_{ik}^n P_{kj}^m$ represents the probability that starting in i the process will go to state j in n + m transitions through a path which takes it into state k at the nth transition. Hence, summing over all intermediate states k yields the probability that the process will be in state j after n + m transitions. Formally, we have $P_{ij}^{n+m} = P\{X_{n+m} = j | X_0 = i\} = \sum_{k=0}^{\infty} P\{X_{n+m} = j, X_n = k | X_0 = i\} = \sum_{k=0}^{\infty} P\{X_{n+m} = j | X_n = k, X_0 = i\} P\{X_n = k | X_0 = i\} = \sum_{k=0}^{\infty} P_{kj}^m P_{ik}^n$ If we let $\mathbf{P}^{(n)}$ denote the matrix of n-step transition probabilities P_{ij}^n, then Equation (4.2) asserts that $\mathbf{P}^{(n+m)} = \mathbf{P}^{(n)} \cdot \mathbf{P}^{(m)}$ where the dot represents matrix multiplication.* Hence, in particular, $\mathbf{P}^{(2)} = \mathbf{P}^{(1+1)} = \mathbf{P} \cdot \mathbf{P} = \mathbf{P}^2$ and by induction $\mathbf{P}^{(n)} = \mathbf{P}^{(n-1+1)} = \mathbf{P}^{n-1} \cdot \mathbf{P} = \mathbf{P}^n$ That is, the *n*-step transition matrix may be obtained by multiplying the matrix **P** by itself *n* times.
Example 4.8: Consider Example 4.1 in which the weather is considered as a two-state Markov chain. If α = 0.7 and β = 0.4, then calculate the probability that it will rain four days from today given that it is raining today. Solution: The one-step transition probability matrix is given by $\mathbf{P} = \begin{Vmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{Vmatrix} \quad \mathbf{P}^{(2)} = \mathbf{P}^2 = \begin{Vmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{Vmatrix} \cdot \begin{Vmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{Vmatrix} = \begin{Vmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{Vmatrix}, \quad \mathbf{P}^{(4)} = (\mathbf{P}^2)^2 = \begin{Vmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{Vmatrix} \cdot \begin{Vmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{Vmatrix} = \begin{Vmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{Vmatrix}$ and the desired probability[CONDITIONAL_PROBABILITY] P_{00}^4 equals 0.5749.
For instance, if α0 = 0.4, α1 = 0.6, in Example 4.8, **(As given on Page 200)** then the (unconditional) probability that it will rain four days after we begin keeping weather records is $\begin{aligned} P\{X_4 = 0\} &= 0.4P_{00}^4 + 0.6P_{10}^4 \\ &= (0.4)(0.5749) + (0.6)(0.5668) \\ &= 0.5700 \end{aligned}$ [UNCONDITIONAL_PROBABILITY]
Example 4.9: Consider Example 4.4, Given that it rained on Monday and Tuesday, what is the probability that it will rain on Thursday? Solution: The two-step transition matrix is given by $\mathbf{P}^{(2)} = \mathbf{P}^2 = \begin{Vmatrix} 0.7 & 0 & 0.3 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0.2 & 0 & 0.8 \end{Vmatrix} \cdot \begin{Vmatrix} 0.7 & 0 & 0.3 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0.2 & 0 & 0.8 \end{Vmatrix} = \begin{Vmatrix} 0.49 & 0.12 & 0.21 & 0.18 \\ 0.35 & 0.20 & 0.15 & 0.30 \\ 0.20 & 0.12 & 0.20 & 0.48 \\ 0.10 & 0.16 & 0.10 & 0.64 \end{Vmatrix}$ Since rain on Thursday is equivalent to the process being in either state 0 or state 1 on Thursday, the desired probability is given by $P_{00}^2 + P_{01}^2$ = 0.49 + 0.12 = 0.61.
Example 4.10: An urn always contains 2 balls. Ball colors are red and blue. At each stage a ball is randomly chosen and then replaced by a new ball, which with probability 0.8 is the same color, and with probability 0.2 is the opposite color, as the ball it replaces. If initially both balls are red, find the probability that the fifth ball selected is red. Solution: To find the desired probability we first define an appropriate Markov chain. This can be accomplished by noting that the probability that a selection is red is determined by the composition of the urn at the time of the selection. So, let us define X_n to be the number of red balls in the urn after the nth selection and subsequent replacement. Then $X_n, n \geqslant 0$, is a Markov chain with states 0, 1, 2 and with transition probability matrix P given by $\begin{pmatrix} 0.8 & 0.2 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 0.2 & 0.8 \end{pmatrix}$

To understand the preceding, consider for instance P1,0. Now, to go from 1 red ball in the urn to 0 red balls, the ball chosen must be red (which occurs with probability 0.5) and it must then be replaced by a ball of opposite color (which occurs with probability 0.2), showing that $P_{1,0} = (0.5)(0.2) = 0.1$
Explanation_for_obtaining_probability_transition_matrix: Consider another instance P0,1: going from 0_red_balls to 1_red_ball = Probability of choosing 1_blue_ball_out_of_2_blue_balls_is_1 x 0.2(which is replacement probability for opposite) = 1 x 0.2 = 0.2. Consider another instance P0,0: 0.8 (transition to itself with replacement probability same = 0.8)
To determine the probability that the fifth selection is red, **condition** on the number of red balls in the urn after the fourth selection. This yields

$$P(\text{fifth selection is red}) = \sum_{i=0}^{2} P(\text{fifth selection is red}|X_4 = i)P(X_4 = i|X_0 = 2) = (0)P^4_{2,0} + (0.5)P^4_{2,1} + (1)P^4_{2,2} = 0.5P^4_{2,1} + P^4_{2,2}$$

(**P(fifth selection is red)**, is conditional probability sum over **all possible transitions** to states 0,1,2 from state 2 and **As urn always contains 2 balls, so,** $X_0 = 2$). To calculate the preceding we compute P^4. Doing so yields $P^4_{2,1} = 0.4352, \quad P^4_{2,2} = 0.4872$ giving the answer P(fifth selection is red) = 0.7048.
Example 4.11 Suppose that balls are successively distributed among 8 urns, with each ball being equally likely to be put in any of these urns. What is the probability that there will be exactly 3 nonempty urns after 9 balls have been distributed? Solution: If we let X_n be the number of nonempty urns after n balls have been distributed, then $X_n, n \geqslant 0$ is a Markov chain with states 0, 1, . . . , 8 and transition probabilities $P_{i,j} = i/8 = 1 - P_{i,i+1}, \quad i = 0, 1, \ldots, 8$ The desired probability is $P^9_{0,3} = P^8_{1,3}$, where the equality follows because $P_{0,1} = 1.$ Now, starting with 1 occupied urn, if we had wanted to determine the entire probability distribution of the number of occupied urns after 8 additional balls had been distributed we would need to consider the transition probability matrix with states 1, 2, . . . , 8. However, because we only require the probability, starting with a single occupied urn, that there are 3 occupied urns after an additional 8 balls have been distributed we can make use of the fact that the state of the Markov chain cannot decrease to collapse all states 4, 5, . . . , 8 into a single state 4 with the interpretation that the state is 4 whenever four or more of the urns are occupied. Consequently, we need only determine the eight-step transition probability $P^8_{1,3}$ of the Markov chain with states 1, 2, 3, 4 having transition probability matrix P given by $\begin{pmatrix} 1/8 & 7/8 & 0 & 0 \\ 0 & 2/8 & 6/8 & 0 \\ 0 & 0 & 3/8 & 5/8 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ **Explanation_for_obtaining_**probability_transition_matrix: [[[[As, $P^9_{0,3} = P^8_{1,3}$, i.e., going from state 0 to state 3(3UrnsAreOccupied) in 9 transitions is equal_to going from state 1 to state 3(3UrnsAreOccupied) in 8 transitions. As $P_{0,1} = 1$ (i.e., with Probability 1, going from state0 to state1(1UrnIsOccupied). **So, we start in state1 with 1UrnOccupied with Probability 1, and then the transitions start**.
For_state1_transition_to_itself_i=1_1Urn_is_occupied_with_probability_(i/8)_=_(1/8)_or_For_state1_transition_to_state2_another_1Urn_is_occupied_with_probability_(8-i/8)_=_(7/8).
Similarly_For_state2_transition_to_itself_i=2_2Urns_are_occupied_with_probability_(i/8)_=_(2/8)_or_For_state2_transition_to_state3_another_1Urn_is_occupied_with_probability_(8-i/8)_=_(6/8).
Similarly_For_state3_transition_to_itself_i=3_3Urns_are_occupied_with_probability_(i/8)_=_(3/8)_or_For_state3_transition_to_state4_another_1Urn_is_occupied_with_probability_(8-i/8)_=_(5/8)_and_FourOrMoreUrns_are_occupied_with_probability_1]]]]
Raising the preceding matrix to the power 4 yields the matrix $\mathbf{P}^4$ given by $\begin{pmatrix} 0.0002 & 0.0256 & 0.2563 & 0.7178 \\ 0 & 0.0039 & 0.0952 & 0.9009 \\ 0 & 0 & 0.0198 & 0.9802 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ Hence,

$$P^8_{1,3} = 0.0002 \times 0.2563 + 0.0256 \times 0.0952 + 0.2563 \times 0.0198 + 0.7178 \times 0 = 0.00756$$

So far, all of the probabilities we have considered are conditional probabilities. For instance, P^n_{ij} is the probability that the state at time n is j given that the initial state at time 0 is i. If the unconditional distribution of the state at time n is desired, it is necessary to specify the probability distribution of the initial state.
Let us denote this by $\alpha_i \equiv P\{X_0 = i\}, \quad i \geqslant 0 \quad \left(\sum_{i=0}^{\infty} \alpha_i = 1\right)$ All **un**conditional probabilities may be computed by conditioning on the initial state. That is,

$$P\{X_n = j\} = \sum_{i=0}^{\infty} P\{X_n = j|X_0 = i\}P\{X_0 = i\} = \sum_{i=0}^{\infty} P^n_{ij}\alpha_i$$

For instance, if α0 = 0.4, α1 = 0.6, in Example 4.8, then the (unconditional) probability that it will rain four days after we begin keeping weather records is

$$\begin{aligned} P\{X_4 = 0\} &= 0.4P^4_{00} + 0.6P^4_{10} \\ &= (0.4)(0.5749) + (0.6)(0.5668) \\ &= 0.5700 \end{aligned}$$

Consider a Markov chain with transition probabilities P_{ij}. Let $\mathscr{A}$ be a set of states, and suppose we are interested in the probability that the Markov chain ever enters any of the states in $\mathscr{A}$ by time m. That is, for a given state i / ∈ $\mathscr{A}$, we are interested in determining **[Beta]** $\beta = P(X_k \in \mathscr{A} \text{ for some } k = 1, \ldots, m|X_0 = i)$ To determine the preceding probability we will define a Markov chain {Wn, n $\geqslant$ 0} whose states are the states that are not in $\mathscr{A}$ plus an additional state, which we will call $\mathscr{A}$ in our general discussion (though in specific examples we will usually give it a different name). Once the {Wn} Markov chain enters state A it remains there forever. The new Markov chain is defined as follows. Letting Xn denote the state at time n of the Markov chain with transition probabilities P_{ij}, define $N = \min\{n : X_n \in \mathscr{A}\}$ and let N = ∞ if Xn $\notin$ $\mathscr{A}$ for all n. In words, N is the first time the Markov chain enters the set of states $\mathscr{A}$. Now, define

$$W_n = \begin{cases} X_n, & \text{if } n < N \\ A, & \text{if } n \geqslant N \end{cases}$$

So the state of the {Wn} process is equal to the state of the original Markov chain up to the point when the original Markov chain enters a state in A . At that time the new process goes to state A and remains there forever. From this description it follows that Wn, n $\geqslant$ 0 is a Markov chain with states i, $\notin$ $\mathscr{A}$, A and with transition probabilities Qi,j given by

$$Q_{i,j} = P_{i,j}, \quad \text{if} \quad i \notin \mathscr{A}, j \notin \mathscr{A}$$
$$Q_{i,A} = \sum_{j \in \mathscr{A}} P_{i,j}, \quad \text{if} \quad i \notin \mathscr{A}$$
$$Q_{A,A} = 1$$

Because the original Markov chain will have entered a state in $\mathscr{A}$ by time m if and only if the state at time m of the new Markov chain is A, we see that

$$P(X_k \in \mathscr{A} \text{ for some } k = 1, \ldots, m|X_0 = i) = P(W_m = A|X_0 = i) = P(W_m = A|W_0 = i) = Q^m_{i,A}$$

That is, the desired probability is equal to an m-step transition probability of the new chain.
Example 4.12 A pensioner receives 2 (thousand dollars) at the beginning of each month. The amount of money he needs to spend during a month is independent of the amount he has and is equal to i with probability P_i, i = 1, 2, 3, 4 $\sum_{i=1}^{4} P_i = 1$. If the pensioner has more than 3 at the end of a month, he gives the amount greater than 3 to his son. If, after receiving his payment at the beginning of a month, the pensioner has a capital of 5, what is the probability that his capital is ever 1 or less at any time within the following four months? Solution: To find the desired probability, we consider a Markov chain with the state equal to the amount the pensioner has at the end of a month. Because we are interested in whether this amount ever falls as low as 1, we will let 1 mean that the pensioner's end-of-month fortune has ever been less than or equal to 1. Because the pensioner will give any end-of-month amount greater than 3 to his son, we need only consider the Markov chain with states 1, 2, 3 and transition probability matrix $\mathbf{Q} = [Q_{i,j}]$ given by $\begin{Vmatrix} 1 & 0 & 0 \\ P_3 + P_4 & P_2 & P_1 \\ P_4 & P_3 & P_1 + P_2 \end{Vmatrix}$ To understand the preceding, consider $Q_{2,1}$, the probability that a month that ends with the pensioner having the amount 2 will be followed by a month that ends with the pensioner having less than or equal to 1. Because the pensioner will begin the new month with the amount 2 + 2 = 4, his ending capital will be less than or equal to 1 if his expenses are either 3 or 4. Thus, $Q_{2,1} = P_3 + P_4$. The other transition probabilities are similarly explained. **Explanation_for_obtaining_**probability_transition_matrix:[[[[Q2,2= begin the new month with the amount 2 + 2 = 4, For end of a

month amount 2 or less, Expenses = probability P2; Q2,3= begin the new month with the amount 2 + 2 = 4, For end of a month amount 3 or less, Expenses = probability P1; Q3,1= begin the new month with the amount 3 + 2 = 5, For end of a month amount 1 or less, Expenses = probability P4; Q3,2= begin the new month with the amount 3 + 2 = 5, For end of a month amount 2 or less, Expenses = probability P3; Q3,3= begin the new month with the amount 3 + 2 = 5, For end of a month amount 3 or less, Expenses = remaining probability P1+P2 as row sum is 1; Q1,1= begin the new month with the amount 1 + 2 = 3, For end of a month amount 1 or less, Expenses = with probability 1; Q1,2= begin the new month with the amount 1 + 2 = 3, For end of a month amount 2 or less, Expenses = with remaining probability 0 as row sum is 1; ; Q1,3= begin the new month with the amount 1 + 2 = 3, For end of a month amount 3 or less, Expenses = with remaining probability 0 as row sum is 1]]]]

Suppose now that P_i = 1/4, i = 1, 2, 3, 4. The transition probability matrix is $\begin{Vmatrix} 1 & 0 & 0 \\ 1/2 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{Vmatrix}$ Squaring this matrix and then squaring the result gives the matrix $\begin{Vmatrix} 1 & 0 & 0 \\ \frac{222}{256} & \frac{13}{256} & \frac{21}{256} \\ \frac{201}{256} & \frac{21}{256} & \frac{34}{256} \end{Vmatrix}$ Because the pensioner's initial end-of-month capital was 3, the desired answer is $Q^4_{3,1}$ = 201/256.

Suppose now that we want to compute the probability that the $\{X_n, n \geq 0\}$ chain, starting in state i, enters state j at time m without ever entering any of the states in $\mathscr{A}$, where neither i nor j is in $\mathscr{A}$. That is, for $i, j \notin \mathscr{A}$, we are interested in **[alpha]** $\alpha = P(X_m = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-1 | X_0 = i)$

Noting that the event that $X_m = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-1$ is equivalent to the event that $W_m = j$, it follows that for $i, j \notin \mathscr{A}$,

$$P(X_m = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-1 | X_0 = i) = P(W_m = j | X_0 = i) = P(W_m = j | W_0 = i) = Q^m_{i,j}.$$

For instance, in Example 4.12, starting with 5 at the beginning of January, the probability that the pensioner's capital is 4 at the beginning of May without ever having been less than or equal to 1 in that time is $Q^4_{3,2} = 21/256.$

Example 4.13 Consider a Markov chain with states 1, 2, 3, 4, 5, and suppose that we want to compute $P(X_4 = 2, X_3 \leq 2, X_2 \leq 2, X_1 \leq 2 | X_0 = 1)$ That is, we want the probability that, starting in state 1, the chain is in state 2 at time 4 and has never entered any of the states in the set A = {3, 4, 5}. To compute this probability all we need to know are the transition probabilities $P_{11}, P_{12}, P_{21}, P_{22}$. So, suppose that $P_{11} = 0.3 \quad P_{12} = 0.3$ $P_{21} = 0.1 \quad P_{22} = 0.2$ Then we consider the Markov chain having states 1, 2, 3 (we are giving state A the name 3), and having the transition probability matrix Q as follows: $\begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.1 & 0.2 & 0.7 \\ 0 & 0 & 1 \end{pmatrix}$ The desired probability is Q^4_{12} Raising Q to the power 4 yields the matrix $\begin{pmatrix} 0.0219 & 0.0285 & 0.9496 \\ 0.0095 & 0.0124 & 0.9781 \\ 0 & 0 & 1 \end{pmatrix}$ Hence, the desired probability is α = 0.0285.

A_When $i \notin \mathscr{A}$ but $j \in \mathscr{A}$ we can determine the probability $\alpha = P(X_m = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-1 | X_0 = i)$ as follows.

$$\alpha = \sum_{r \notin \mathscr{A}} P(X_m = j, X_{m-1} = r, X_k \notin \mathscr{A}, k = 1, \ldots, m-2 | X_0 = i)$$
$$= \sum_{r \notin \mathscr{A}} P(X_m = j | X_{m-1} = r, X_k \notin \mathscr{A}, k = 1, \ldots, m-2, X_0 = i) \times P(X_{m-1} = r, X_k \notin \mathscr{A}, k = 1, \ldots, m-2 | X_0 = i)$$
$$= \sum_{r \notin \mathscr{A}} P_{r,j} P(X_{m-1} = r, X_k \notin \mathscr{A}, k = 1, \ldots, m-2 | X_0 = i)$$

Also, when $i \in \mathscr{A}$ we could determine

$$\alpha = P(X_m = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-1 | X_0 = i) = \sum_{r \notin \mathscr{A}} P_{r,j} Q^{m-1}_{i,r}$$

B_Also, when $i \in \mathscr{A}$ we could determine $\alpha = P(X_m = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-1 | X_0 = i)$ by conditioning on the first transition to obtain

$$\alpha = \sum_{r \notin \mathscr{A}} P(X_m = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-1 | X_0 = i, X_1 = r) P(X_1 = r | X_0 = i) = \sum_{r \notin \mathscr{A}} P(X_{m-1} = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-2 | X_0 = r) P_{i,r}$$

For instance, if $i \in \mathscr{A}, j \notin \mathscr{A}$ then the preceding equation yields

$$P(X_m = j, X_k \notin \mathscr{A}, k = 1, \ldots, m-1 | X_0 = i) = \sum_{r \notin \mathscr{A}} Q^{m-1}_{r,j} P_{i,r}$$

C_We can also compute the conditional probability of X_n given that the chain starts in state i and has not entered any state in $\mathscr{A}$ by time n, as follows.

For $i, j \notin \mathscr{A}$,

$$P\{X_n = j | X_0 = i, X_k \notin \mathscr{A}, k = 1, \ldots, n\} = \frac{P\{X_n = j, X_k \notin \mathscr{A}, k = 1, \ldots, n | X_0 = i\}}{P\{X_k \notin \mathscr{A}, k = 1, \ldots, n | X_0 = i\}} = \frac{Q^n_{i,j}}{\sum_{r \notin \mathscr{A}} Q^n_{i,r}}$$

4.3 Classification of States: State j is said to be accessible from state i if $P^n_{ij} > 0$ for some $n \geq 0$. Note that this implies that state j is accessible from state i if and only if, starting in i, it is possible that the process will ever enter state j. This is true since if j is not accessible from i, then

$$P\{\text{ever enter } j | \text{start in } i\} = P\left\{\bigcup_{n=0}^{\infty} \{X_n = j\} \middle| X_0 = i\right\} \leq \sum_{n=0}^{\infty} P\{X_n = j | X_0 = i\} = \sum_{n=0}^{\infty} P^n_{ij} = 0$$

Two states i and j that are accessible to each other are said to communicate, and we write i ↔ j. Note that any state communicates with itself since, by definition, $P^0_{ii} = P\{X_0 = i | X_0 = i\} = 1$ (i) State i communicates with state i, all $i \geq 0$.

(ii) If state i communicates with state j, then state j communicates with state i.

(iii) If state i communicates with state j, and state j communicates with state k, then state i communicates with state k.

Properties (i) and (ii) follow immediately from the definition of communication. To prove (iii) suppose that i communicates with j, and j communicates with k.

Thus, there exist integers n and m such that $P_{ij}^n > 0,\ P_{jk}^m > 0$. Now by the Chapman–Kolmogorov equations, we have $P_{ik}^{n+m} = \sum_{r=0}^{\infty} P_{ir}^n P_{rk}^m \geqslant P_{ij}^n P_{jk}^m > 0$ Hence, state k is accessible from state i. Similarly, we can show that state i is accessible from state k. Hence, states i and k communicate. Two states that communicate are said to be in the same class. It is an easy consequence of (i), (ii), and (iii) that any two classes of states are either identical or disjoint. In other words, the concept of communication divides the state space up into a number of separate classes. The Markov chain is said to be irreducible if there is only one class, that is, if all states communicate with each other.

Example 4.14 Consider the Markov chain consisting of the three states 0, 1, 2 and having transition probability matrix $\mathbf{P} = \begin{Vmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{Vmatrix}$ It is easy to verify that this Markov chain is irreducible. For example, it is possible to go from state 0 to state 2 since $0 \to 1 \to 2$ That is, one way of getting from state 0 to state 2 is to go from state 0 to state 1 (with probability 1/2) and then go from state 1 to state 2 (with probability 1/4).

Example 4.15 Consider a Markov chain consisting of the four states 0, 1, 2, 3 and having transition probability matrix

$\mathbf{P} = \begin{Vmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{Vmatrix}$ The classes of this Markov chain are {0, 1}, {2}, and {3}. Note that while state 0 (or 1) is accessible from state 2, the reverse is not true. Since state 3 is an absorbing state, that is, P33 = 1, no other state is accessible from it.

For any state i we let fi denote the probability that, starting in state i, the process will ever reenter state i. State i is said to be recurrent if fi = 1 and transient if fi < 1.

A_Suppose that the process starts in state i and i is recurrent. Hence, with probability 1, the process will eventually reenter state i. However, by the definition of a Markov chain, it follows that the process will be starting over again when it reenters state i and, therefore, state i will eventually be visited again. Continual repetition of this argument leads to the conclusion that if state i is recurrent then, starting in state i, the process will reenter state i again and again and again—in fact, infinitely often.

B_On the other hand, suppose that state i is transient. Hence, each time the process enters state i there will be a positive probability, namely, 1 − fi, that it will never again enter that state. Therefore, starting in state i, the probability that the process will be in state i for exactly n time periods equals $f_i^{n-1}(1 - f_i),\ n \geqslant 1$. In other words, if state i is transient then, starting in state i, the number of time periods that the process will be in state i has a geometric distribution with finite mean 1/(1 − fi).

From the preceding two paragraphs, it follows that state i is recurrent if and only if, starting in state i, the expected number of time periods that the process is in state i is infinite. But, letting $I_n = \begin{cases} 1, & \text{if } X_n = i \\ 0, & \text{if } X_n \neq i \end{cases}$ we have that $\sum_{n=0}^{\infty} I_n$ represents the number of periods that the process is in state i. Also,

$$E\left[\sum_{n=0}^{\infty} I_n | X_0 = i\right] = \sum_{n=0}^{\infty} E[I_n | X_0 = i]$$
$$= \sum_{n=0}^{\infty} P\{X_n = i | X_0 = i\}$$
$$= \sum_{n=0}^{\infty} P_{ii}^n$$

We have thus proven the following.

Proposition 4.1 State i is

recurrent if $\sum_{n=1}^{\infty} P_{ii}^n = \infty$,

transient if $\sum_{n=1}^{\infty} P_{ii}^n < \infty$

The argument leading to the preceding proposition is doubly important because it also shows that a transient state will only be visited a finite number of times (hence the name transient). This leads to the conclusion that in a finite-state Markov chain not all states can be transient. To see this, suppose the states are 0, 1, . . . ,M and suppose that they are all transient. Then after a finite amount of time (say, after time T0) state 0 will never be visited, and after a time (say, T1) state 1 will never be visited, and after a time (say, T2) state 2 will never be visited, and so on. Thus, after a finite time T = max{T0, T1, . . . , TM} no states will be visited. But as the process must be in some state after time T we arrive at a contradiction, which shows that at least one of the states must be recurrent. Another use of Proposition 4.1 is that it enables us to show that recurrence is a class property. Corollary 4.2 If state i is recurrent, and state i communicates with state j, then state j is recurrent. Proof. To prove this we first note that, since state i communicates with state j, there exist integers k and m such that $P_{ij}^k > 0, P_{ji}^m > 0$. Now, for any integer n $P_{jj}^{m+n+k} \geqslant P_{ji}^m P_{ii}^n P_{ij}^k$ This follows since the left side of the preceding is the probability of going from j to j in m + n + k steps, while the right side is the probability of going from j to j in m + n + k steps via a path that goes from j to i in m steps, then from i to i in an additional n steps, then from i to j in an additional k steps. From the preceding we obtain, by summing over n, that

$\sum_{n=1}^{\infty} P_{jj}^{m+n+k} \geqslant P_{ji}^m P_{ij}^k \sum_{n=1}^{\infty} P_{ii}^n = \infty$ since $P_{ji}^m P_{ij}^k > 0$ and $\sum_{n=1}^{\infty} P_{ii}^n$ is infinite since state i is recurrent. Thus, by Proposition 4.1 it follows that state j is also recurrent. Remarks:

(i) Corollary 4.2 also implies that transience is a class property. For if state i is transient and communicates with state j, then state j must also be transient. For if j were recurrent then, by Corollary 4.2, i would also be recurrent and hence could not be transient.

(ii) Corollary 4.2 along with our previous result that not all states in a finite Markov chain can be transient leads to the conclusion that all states of a finite irreducible Markov chain are recurrent.

Example 4.16 Let the Markov chain consisting of the states 0, 1, 2, 3 have the transition probability matrix $\mathbf{P} = \begin{Vmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{Vmatrix}$ Determine which states are transient and which are recurrent. Solution: It is a simple matter to check that all states communicate and, hence, since this is a finite chain, all states must be recurrent.

Example 4.17 Consider the Markov chain having states 0, 1, 2, 3, 4 and $\mathbf{P} = \begin{Vmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{2} \end{Vmatrix}$ Determine the recurrent state. Solution: This chain consists of the three classes {0, 1}, {2, 3}, and {4}. The first two classes are recurrent and the third transient.

Example 4.18 (A Random Walk): Consider a Markov chain whose state space consists of the integers i = 0,±1,±2, . . . , and has transition probabilities given by $P_{i,i+1} = p = 1 - P_{i,i-1}, \quad i = 0, \pm 1, \pm 2, \ldots$ where 0 < p < 1. In other words, on each transition the process either moves one step to the right (with probability p) or one step to the left (with probability 1−p). One colorful interpretation of this process is that it represents the wanderings of a drunken man as he walks along a straight line. Another is that it represents the winnings of a gambler who on each play of the game either wins or loses one dollar. Since all states clearly communicate, it follows from Corollary 4.2 that they are either all transient or all recurrent. So let us consider state 0 and attempt to determine if $\sum_{n=1}^{\infty} P_{00}^n$ is finite or infinite. Since it is impossible to be even (using the gambling model interpretation) after an odd number of plays we must, of course, have that $P_{00}^{2n-1} = 0, \quad n = 1, 2, \ldots$ On the other hand, we would be even after 2n trials if and only if we won n of these and lost n of these. Because each play of the game results in a win with probability p and a loss with probability 1−p, the desired probability is thus the binomial probability $P_{00}^{2n} = \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{n!n!}(p(1-p))^n, \quad n = 1, 2, 3, \ldots$ By using an approximation, due to Stirling,

which asserts that $n! \sim n^{n+1/2}e^{-n}\sqrt{2\pi}$ where we say that an $a_n \sim b_n$ when $\lim_{n\to\infty} a_n/b_n = 1$, we obtain $P_{00}^{2n} \sim \frac{(4p(1-p))^n}{\sqrt{\pi n}}$ Now it is easy to verify, for positive $a_n, b_n,$ that if $a_n \sim b_n,$ then $\sum_n a_n < \infty$ if and only if $\sum_n b_n < \infty.$ Hence, $\sum_{n=1}^{\infty} P_{00}^n$ will converge if and only if $\sum_{n=1}^{\infty} \frac{(4p(1-p))^n}{\sqrt{\pi n}}$ does. However, $4p(1-p) \leqslant 1$ with equality holding if and only if p = 1/2 . Hence, $\sum_{n=1}^{\infty} P_{00}^n = \infty$ if and only if p = 1/2 . Thus, the chain is recurrent when p = 1/2 and transient if $p \neq \frac{1}{2}.$ When p = 1/2 , the preceding process is called a symmetric random walk. We could also look at symmetric random walks in more than one dimension. For instance, in the two-dimensional symmetric random walk the process would, at each transition, either take one step to the left, right, up, or down, each having probability 1/4 . That is, the state is the pair of integers (i, j) and the transition probabilities are given by $P_{(i,j),(i+1,j)} = P_{(i,j),(i-1,j)} = P_{(i,j),(i,j+1)} = P_{(i,j),(i,j-1)} = \frac{1}{4}$ By using the same method as in the one-dimensional case, we now show that this Markov chain is also recurrent. Since the preceding chain is irreducible, it follows that all states will be recurrent if state 0 = (0, 0) is recurrent. So consider $P_{00}^{2n}.$. Now after 2n steps, the chain will be back in its original location if for some i, $0 \leqslant i \leqslant n,$ the 2n steps consist of i steps to the left, i to the right, n − i up, and n − i down. Since each step will be either of these four types with probability 1/4 , it follows that the desired probability is a multinomial probability. That is,

$$P_{00}^{2n} = \sum_{i=0}^{n} \frac{(2n)!}{i!i!(n-i)!(n-i)!}\left(\frac{1}{4}\right)^{2n} = \sum_{i=0}^{n} \frac{(2n)!}{n!n!}\frac{n!}{(n-i)!i!}\frac{n!}{(n-i)!i!}\left(\frac{1}{4}\right)^{2n} = \left(\frac{1}{4}\right)^{2n}\binom{2n}{n}\sum_{i=0}^{n}\binom{n}{i}\binom{n}{n-i} = \left(\frac{1}{4}\right)^{2n}\binom{2n}{n}\binom{2n}{n} \quad (4.4)$$

where the last equality uses the combinatorial identity $\binom{2n}{n} = \sum_{i=0}^{n}\binom{n}{i}\binom{n}{n-i}$ which follows upon noting that both sides represent the number of subgroups of size n one can select from a set of n white and n black objects. Now, $\binom{2n}{n} = \frac{(2n)!}{n!n!} \sim \frac{(2n)^{2n+1/2}e^{-2n}\sqrt{2\pi}}{n^{2n+1}e^{-2n}(2\pi)}$ by Stirling's approximation $= \frac{4^n}{\sqrt{\pi n}}$ Hence, from Equation (4.4) we see that $P_{00}^{2n} \sim \frac{1}{\pi n}$ which shows that $\sum_n P_{00}^{2n} = \infty,$ and thus all states are recurrent. Interestingly enough, whereas the symmetric random walks in one and two dimensions are both recurrent, all higher-dimensional symmetric random walks turn out to be transient. (For instance, the three-dimensional symmetric random walk is at each transition equally likely to move in any of six ways—either to the left, right, up, down, in, or out.)

Remark: For the one-dimensional random walk of Example 4.18 here is a direct argument for establishing recurrence in the symmetric case, and for determining the probability that it ever returns to state 0 in the nonsymmetric case. Let β = P{ever return to 0}

To determine β, start by conditioning on the initial transition to obtain $\beta = P\{\text{ever return to } 0 | X_1 = 1\}p + P\{\text{ever return to } 0 | X_1 = -1\}(1-p)$ (4.5)

Now, let α denote the probability that the Markov chain will ever return to state 0 given that it is currently in state 1. Because the Markov chain will always increase by 1 with probability p or decrease by 1 with probability 1−p no matter what its current state, note that α is also the probability that the Markov chain currently in state i will ever enter state i−1, for any i. To obtain an equation for α, condition on the next transition to obtain

$$\alpha = P\{\text{ever return} | X_1 = 1, X_2 = 0\}(1-p) + P\{\text{ever return} | X_1 = 1, X_2 = 2\}p = 1 - p + P\{\text{ever return} | X_1 = 1, X_2 = 2\}p = 1 - p + p\alpha^2$$

where the final equation follows by noting that in order for the chain to ever go from state 2 to state 0 it must first go to state 1—and the probability of that ever happening is α—and if it does eventually go to state 1 then it must still go to state 0—and the conditional probability of that ever happening is also α. Therefore, $\alpha = 1 - p + p\alpha^2$ The two roots of this equation are α = 1 and α = (1−p)/p. Consequently, in the case of the symmetric random walk where p = 1/2 we can conclude that α = 1. By symmetry, the probability that the symmetric random walk will ever enter state 0 given that it is currently in state −1 is also 1, proving that the symmetric random walk is recurrent. Suppose now that p > 1/2. In this case, it can be shown (see Exercise 17 at the end of this chapter) that P{ever return to 0 | X1 = −1} = 1. Consequently, Equation (4.5) reduces to $\beta = \alpha p + 1 - p$ Because the random walk is transient in this case we know that β < 1, showing that $\alpha \neq 1.$ Therefore, α = (1 − p)/p, yielding that

$$\beta = 2(1-p), \quad p > 1/2$$

Similarly, when p < 1/2 we can show that β = 2p. Thus, in general $P\{\text{ever return to } 0\} = 2\min(p, 1-p)$

Example 4.19 (On the Ultimate Instability of the Aloha Protocol) Consider a communications facility in which the numbers of messages arriving during each of the time periods n = 1, 2, . . . are independent and identically distributed random variables. Let a_i = P{i arrivals}, and suppose that a0 + a1 < 1. Each arriving message will transmit at the end of the period in which it arrives. If exactly one message is transmitted, then the transmission is successful and the message leaves the system. However, if at any time two or more messages simultaneously transmit, then a collision is deemed to occur and these messages remain in the system. Once a message is involved in a collision it will, independently of all else, transmit at the end of each additional period with probability p—the socalled Aloha protocol (because it was first instituted at the University of Hawaii). We will show that such a system is asymptotically unstable in the sense that the number of successful transmissions will, with probability 1, be finite. To begin let X_n denote the number of messages in the facility at the beginning of the nth period, and note that $\{X_n, n \geqslant 0\}$ is a Markov chain. Now for $k \geqslant 0$ define the indicator variables I_k by

$$I_k = \begin{cases} 1, & \text{if the first time that the chain departs state } k \text{ it} \\ & \text{directly goes to state } k-1 \\ 0, & \text{otherwise} \end{cases}$$

and let it be 0 if the system is never in state k, $k \geqslant 0$. (For instance, if the successive states are 0, 1, 3, 4, . . . , then I_3 = 0 since when the chain first departs state 3 it goes to state 4; whereas, if they are 0, 3, 3, 2, . . . , then I3 = 1 since this time it goes to state 2.)

$$E\left[\sum_{k=0}^{\infty} I_k\right] = \sum_{k=0}^{\infty} E[I_k] = \sum_{k=0}^{\infty} P\{I_k = 1\} \leqslant \sum_{k=0}^{\infty} P\{I_k = 1 | k \text{ is ever visited}\} \quad (4.6)$$

Now, $P\{I_k = 1 | k \text{ is ever visited}\}$ is the probability that when state k is departed the next state is k − 1. That is, it is the conditional probability that a transition from k is to k − 1 given that it is not back into k, and so

$$P\{I_k = 1 | k \text{ is ever visited}\} = \frac{P_{k,k-1}}{1 - P_{k,k}}$$

Because

$$P_{k,k-1} = a_0 kp(1-p)^{k-1},$$
$$P_{k,k} = a_0[1 - kp(1-p)^{k-1}] + a_1(1-p)^k$$

which is seen by noting that if there are k messages present on the beginning of a day, then (a) there will be k − 1 at the beginning of the next day if there are no new messages that day and exactly one of the k messages transmits; and (b) there will be k at the beginning of the next day if either (i) there are no new messages and it is not the case that exactly one of the existing k messages transmits, or (ii) there is exactly one new message (which automatically transmits) and none of the other k messages transmits.

Substitution of the preceding into Equation (4.6) yields

$$E\left[\sum_{k=0}^{\infty} I_k\right] \leqslant \sum_{k=0}^{\infty} \frac{a_0 kp(1-p)^{k-1}}{1 - a_0[1 - kp(1-p)^{k-1}] - a_1(1-p)^k} < \infty$$

where the convergence follows by noting that when k is large the denominator of the expression in the preceding sum converges to 1 − a0 and so the convergence or divergence of the sum is determined by whether or not the sum of the terms in the numerator converge and $\sum_{k=0}^{\infty} k(1-p)^{k-1} < \infty.$ Hence, $E[\sum_{k=0}^{\infty} I_k] < \infty,$ which implies that $\sum_{k=0}^{\infty} I_k < \infty$ with probability 1 (for if there was a positive probability that $\sum_{k=0}^{\infty} I_k$ could be ∞, then its mean would be ∞). Hence, with probability 1, there will be only a finite number of states that are initially departed via a successful transmission; or equivalently, there will be some finite integer N such that whenever there are N or more messages in the system, there will never again be a successful transmission. From this (and the fact that such higher states will eventually be reached—why?) it follows that, with probability 1, there will only be a finite number of successful transmissions.

Remark: For a (slightly less than rigorous) probabilistic proof of Stirling's approximation, let X1X2, . . . be independent Poisson random variables each having mean 1. Let $S_n = \sum_{i=1}^{n} X_i$, and note that both the mean and variance of Sn are equal to n. Now,

$$\begin{aligned} P\{S_n = n\} &= P\{n-1 < S_n \leqslant n\} \\ &= P\{-1/\sqrt{n} < (S_n - n)/\sqrt{n} \leqslant 0\} \\ &\approx \int_{-1/\sqrt{n}}^{0} (2\pi)^{-1/2} e^{-x^2/2}\, dx \quad \text{when } n \text{ is large, by the central limit theorem} \\ &\approx (2\pi)^{-1/2}(1/\sqrt{n}) \\ &= (2\pi n)^{-1/2} \end{aligned}$$

But Sn is Poisson with mean n, and so $P\{S_n = n\} = \frac{e^{-n}n^n}{n!}$ Hence, for n large

$\frac{e^{-n}n^n}{n!} \approx (2\pi n)^{-1/2}$ or, equivalently $n! \approx n^{n+1/2}e^{-n}\sqrt{2\pi}$ which is Stirling's approximation.

4.4 Limiting Probabilities: In Example 4.8, we calculated $\mathbf{P}^{(4)}$ for a two-state Markov chain; it turned out to be $\mathbf{P}^{(4)} = \begin{Vmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{Vmatrix}$ From this it follows that $\mathbf{P}^{(8)} = \mathbf{P}^{(4)} \cdot \mathbf{P}^{(4)}$ is given (to three significant places) by $\mathbf{P}^{(8)} = \begin{Vmatrix} 0.572 & 0.428 \\ 0.570 & 0.430 \end{Vmatrix}$ Note that the matrix $\mathbf{P}^{(8)}$ is almost identical to the matrix $\mathbf{P}^{(4)}$, and secondly, that each of the rows of $\mathbf{P}^{(8)}$ has almost identical entries. In fact it seems that P_{ij}^n is converging to some value (as n→∞) that is the same for all i. In other words, there seems to exist a limiting probability that the process will be in state j after a large number of transitions, and this value is independent of the initial state. To make the preceding heuristics more precise, two additional properties of the states of a Markov chain need to be considered. State i is said to have period d if $P_{ii}^n = 0$ whenever n is not divisible by d, and d is the largest integer with this property. For instance, starting in i, it may be possible for the process to enter state i only at the times 2, 4, 6, 8, . . . , in which case state i has period 2. A state with period 1 is said to be aperiodic. It can be shown that periodicity is a class property. That is, if state i has period d, and states i and j communicate, then state j also has period d. If state i is recurrent, then it is said to be positive recurrent if, starting in i, the expected time until the process returns to state i is finite. It can be shown that positive recurrence is a class property. While there exist recurrent states that are not positive recurrent,∗ it can be shown that in a finite-state Markov chain all recurrent states are positive recurrent. Positive recurrent, aperiodic states are called ergodic**[ERGODIC]**. ∗ Such states are called null recurrent. We are now ready for the following important theorem, which we state without proof. Theorem 4.1 For an irreducible ergodic Markov chain $\lim_{n\to\infty} P_{ij}^n$ exists and is independent of i. Furthermore, letting $\pi_j = \lim_{n\to\infty} P_{ij}^n, \quad j \geqslant 0$ then πj is the unique nonnegative solution of $\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, \quad j \geqslant 0, \quad \sum_{j=0}^{\infty} \pi_j = 1$ Remarks:

(i) Given that $\pi_j = \lim_{n\to\infty} P_{ij}^n$ exists and is independent of the initial state i, it is not difficult to (heuristically) see that the π's must satisfy Equation (4.7). Let us derive an expression for $P\{X_{n+1} = j\}$ by conditioning on the state at time n. That is,

$$\begin{aligned} P\{X_{n+1} = j\} &= \sum_{i=0}^{\infty} P\{X_{n+1} = j | X_n = i\} P\{X_n = i\} \\ &= \sum_{i=0}^{\infty} P_{ij} P\{X_n = i\} \end{aligned}$$

Letting n → ∞, and assuming that we can bring the limit inside the summation, leads to

$$\pi_j = \sum_{i=0}^{\infty} P_{ij}\pi_i$$

(ii) It can be shown that πj , the limiting probability that the process will be in state j at time n, also equals the long-run proportion of time that the process will be in state j.

(iii) If the Markov chain is irreducible, then there will be a solution to $\pi_j = \sum_i \pi_i P_{ij}, \quad j \geqslant 0, \quad \sum_j \pi_j = 1$ if and only if the Markov chain is positive recurrent. If a solution exists then it will be unique, and πj will equal the long-run proportion of time that the Markov chain is in state j. If the chain is aperiodic, then πj is also the limiting probability that the chain is in state j.

Example 4.20 Consider Example 4.1, in which we assume that if it rains today, then it will rain tomorrow with probability α; and if it does not rain today, then it will rain tomorrow with probability β. If we say that the state is 0 when it rains and 1 when it does not rain, then by Equation (4.7) the limiting probabilities π0 and π1 are given by

$$\begin{aligned} \pi_0 &= \alpha\pi_0 + \beta\pi_1, \\ \pi_1 &= (1-\alpha)\pi_0 + (1-\beta)\pi_1, \\ \pi_0 + \pi_1 &= 1 \end{aligned}$$

which yields that $\pi_0 = \frac{\beta}{1+\beta-\alpha}, \quad \pi_1 = \frac{1-\alpha}{1+\beta-\alpha}$ For example if α = 0.7 and β = 0.4, then the limiting probability of rain is $\pi_0 = \frac{4}{7} = 0.571.$

Example 4.21 Consider Example 4.3 in which the mood of an individual is considered as a three-state Markov chain having a transition probability matrix

$$\mathbf{P} = \begin{Vmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{Vmatrix}$$

In the long run, what proportion of time is the process in each of the three states? Solution: The limiting probabilities π_i, i = 0, 1, 2, are obtained by solving the set of equations in Equation (4.1). In this case these equations are

$$\begin{aligned} \pi_0 &= 0.5\pi_0 + 0.3\pi_1 + 0.2\pi_2, \\ \pi_1 &= 0.4\pi_0 + 0.4\pi_1 + 0.3\pi_2, \\ \pi_2 &= 0.1\pi_0 + 0.3\pi_1 + 0.5\pi_2, \quad \pi_0 + \pi_1 + \pi_2 = 1 \end{aligned}$$

Solving yields

$$\pi_0 = \frac{21}{62}, \quad \pi_1 = \frac{23}{62}, \quad \pi_2 = \frac{18}{62}$$

Example 4.22 (A Model of Class Mobility) A problem of interest to sociologists is to determine the proportion of society that has an upper- or lower-class occupation. One possible mathematical model would be to assume that transitions between social classes of the successive generations in a family can be regarded as transitions of a Markov chain. That is, we assume that the occupation of a child depends only on his or her parent's occupation. Let us suppose that such a model is appropriate and that the transition probability matrix is given by

$$\mathbf{P} = \begin{Vmatrix} 0.45 & 0.48 & 0.07 \\ 0.05 & 0.70 & 0.25 \\ 0.01 & 0.50 & 0.49 \end{Vmatrix}$$

That is, for instance, we suppose that the child of a middle-class worker will

attain an upper-, middle-, or lower-class occupation with respective probabilities 0.05, 0.70, 0.25. The limiting probabilities π_i thus satisfy

$$\pi_0 = 0.45\pi_0 + 0.05\pi_1 + 0.01\pi_2,$$
$$\pi_1 = 0.48\pi_0 + 0.70\pi_1 + 0.50\pi_2,$$
$$\pi_2 = 0.07\pi_0 + 0.25\pi_1 + 0.49\pi_2, \qquad \pi_0 + \pi_1 + \pi_2 = 1$$

Hence, $\pi_0 = 0.07, \quad \pi_1 = 0.62, \quad \pi_2 = 0.31$ In other words, a society in which social mobility between classes can be described by a Markov chain with transition probability matrix given by Equation (4.8) has, in the long run, 7 percent of its people in upper-class jobs, 62 percent of its people in middle-class jobs, and 31 percent in lower-class jobs.

Example 4.23 (The Hardy–Weinberg Law and a Markov Chain in Genetics) Consider a large population of individuals, each of whom possesses a particular pair of genes, of which each individual gene is classified as being of type A or type a. Assume that the proportions of individuals whose gene pairs are AA, aa, or Aa are, respectively, p_0, q_0, and r_0 $(p_0 + q_0 + r_0 = 1)$. When two individuals mate, each contributes one of his or her genes, chosen at random, to the resultant offspring. Assuming that the mating occurs at random, in that each individual is equally likely to mate with any other individual, we are interested in determining the proportions of individuals in the next generation whose genes are AA, aa, or Aa. Calling these proportions p, q, and r, they are easily obtained by focusing attention on an individual of the next generation and then determining the probabilities for the gene pair of that individual. To begin, note that randomly choosing a parent and then randomly choosing one of its genes is equivalent to just randomly choosing a gene from the total gene population. By conditioning on the gene pair of the parent, we see that a randomly chosen gene will be type A with probability $P\{A\} = P\{A|AA\}p_0 + P\{A|aa\}q_0 + P\{A|Aa\}r_0 \; = p_0 + r_0/2$ Similarly, it will be type a with probability $P\{a\} = q_0 + r_0/2$ Thus, under random mating a randomly chosen member of the next generation will be type AA with probability p, where $p = P\{A\}P\{A\} = (p_0 + r_0/2)^2$ Similarly, the randomly chosen member will be type aa with probability $q = P\{a\}P\{a\} = (q_0 + r_0/2)^2$ and will be type Aa with probability $r = 2P\{A\}P\{a\} = 2(p_0 + r_0/2)(q_0 + r_0/2)$ Since each member of the next generation will independently be of each of the three gene types with probabilities p, q, r, it follows that the percentages of the members of the next generation that are of type AA, aa, or Aa are respectively p, q, and r. If we now consider the total gene pool of this next generation, then $p + r/2$, the fraction of its genes that are A, will be unchanged from the previous generation. This follows either by arguing that the total gene pool has not changed from generation to generation or by the following simple algebra:

$$p + r/2 = (p_0 + r_0/2)^2 + (p_0 + r_0/2)(q_0 + r_0/2) \; = (p_0 + r_0/2)[p_0 + r_0/2 + q_0 + r_0/2] \; = p_0 + r_0/2 \quad \text{since } p_0 + r_0 + q_0 = 1 = P\{A\}$$

Thus, the fractions of the gene pool that are A and a are the same as in the initial generation. From this it follows that, under random mating, in all successive generations after the initial one the percentages of the population having gene pairs AA, aa, and Aa will remain fixed at the values p, q, and r. This is known as the Hardy–Weinberg law. Suppose now that the gene pair population has stabilized in the percentages p, q, r, and let us follow the genetic history of a single individual and her descendants. (For simplicity, assume that each individual has exactly one offspring.) So, for a given individual, let Xn denote the genetic state of her descendant in the nth generation.

$$\begin{array}{c|ccc} & AA & aa & Aa \\ AA & p + \frac{r}{2} & 0 & q + \frac{r}{2} \\ aa & 0 & q + \frac{r}{2} & p + \frac{r}{2} \\ Aa & \frac{p}{2} + \frac{r}{4} & \frac{q}{2} + \frac{r}{4} & \frac{p}{2} + \frac{q}{2} + \frac{r}{2} \end{array}$$

The transition probability matrix of this Markov chain, namely, is easily verified by conditioning on the state of the randomly chosen mate. It is quite intuitive (why?) that the limiting probabilities for this Markov chain (which also equal the fractions of the individual's descendants that are in each of the three genetic states) should just be p, q, and r. To verify this we must show that they satisfy Equation (4.7). Because one of the equations in Equation (4.7) is redundant, it suffices to show that

$$p = p\left(p + \frac{r}{2}\right) + r\left(\frac{p}{2} + \frac{r}{4}\right) = \left(p + \frac{r}{2}\right)^2,$$
$$q = q\left(q + \frac{r}{2}\right) + r\left(\frac{q}{2} + \frac{r}{4}\right) = \left(q + \frac{r}{2}\right)^2, \qquad p + q + r = 1$$

But this follows from Equation (4.9), and thus the result is established.

Example 4.24 Suppose that a production process changes states in accordance with an irreducible, positive recurrent Markov chain having transition probabilities $P_{ij}, i,j = 1,\ldots,n,$ and suppose that certain of the states are considered acceptable and the remaining unacceptable. Let A denote the acceptable states and A^c the unacceptable ones. If the production process is said to be "up" when in an acceptable state and "down" when in an unacceptable state, determine

1. the rate at which the production process goes from up to down (that is, the rate of breakdowns);
2. the average length of time the process remains down when it goes down; and
3. the average length of time the process remains up when it goes up.

Solution: Let π_k, k = 1, . . . , n, denote the long-run proportions. Now for i ∈ A and $j \in A^c$ the rate at which the process enters state j from state i is rate enter j from i = $\pi_i P_{ij}$ and so the rate at which the production process enters state j from an acceptable state is $\text{rate enter } j \text{ from } A = \sum_{i \in A} \pi_i P_{ij}$ Hence, the rate at which it enters an unacceptable state from an acceptable one (which is the rate at which breakdowns occur) is $\text{rate breakdowns occur} = \sum_{j \in A^c} \sum_{i \in A} \pi_i P_{ij}$ (4.10) Now let $\bar{U} + \bar{D}$ denote the average time the process remains up when it goes up and down when it goes down. Because there is a single breakdown every $\bar{U}$ and $\bar{D}$ time units on the average, it follows heuristically that $\text{rate at which breakdowns occur} = \frac{1}{\bar{U} + \bar{D}}$ and so from Equation (4.10), $\frac{1}{\bar{U} + \bar{D}} = \sum_{j \in A^c} \sum_{i \in A} \pi_i P_{ij}$ (4.11) To obtain a second equation relating $\bar{U} + \bar{D}$, consider the percentage of time the process is up, which, of course, is equal to $\sum_{i \in A} \pi_i$. However, since the process is up on the average $\bar{U}$ out of every $\bar{U} + \bar{D}$ time units, it follows (again somewhat heuristically) that the $\text{proportion of up time} = \frac{\bar{U}}{\bar{U} + \bar{D}}$ and so $\frac{\bar{U}}{\bar{U} + \bar{D}} = \sum_{i \in A} \pi_i$ (4.12)

$$\bar{U} = \frac{\sum_{i \in A} \pi_i}{\sum_{j \in A^c} \sum_{i \in A} \pi_i P_{ij}},$$
$$\bar{D} = \frac{1 - \sum_{i \in A} \pi_i}{\sum_{j \in A^c} \sum_{i \in A} \pi_i P_{ij}} \; = \frac{\sum_{i \in A^c} \pi_i}{\sum_{j \in A^c} \sum_{i \in A} \pi_i P_{ij}}$$

Hence, from Equations (4.11) and (4.12) we obtain For example, suppose the transition probability matrix is

$$\mathbf{P} = \begin{Vmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \end{Vmatrix}$$ where the acceptable (up) states are 1, 2 and the unacceptable (down) ones are 3, 4. The limiting probabilities satisfy

$\pi_1 = \pi_1\frac{1}{4} + \pi_3\frac{1}{4} + \pi_4\frac{1}{4}$,

$\pi_2 = \pi_1\frac{1}{4} + \pi_2\frac{1}{4} + \pi_3\frac{1}{4} + \pi_4\frac{1}{4}$,

$\pi_3 = \pi_1\frac{1}{2} + \pi_2\frac{1}{2} + \pi_3\frac{1}{4}$, $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$ These solve to yield $\pi_1 = \frac{3}{16}, \quad \pi_2 = \frac{1}{4}, \quad \pi_3 = \frac{14}{48}, \quad \pi_4 = \frac{13}{48}$ and thus

rate of breakdowns $= \pi_1(P_{13} + P_{14}) + \pi_2(P_{23} + P_{24}) \quad = \frac{9}{32}, \quad \bar{U} = \frac{14}{9}$ and $\bar{D} = 2$ Hence, on the average, breakdowns occur about 9/32 (or 28 percent) of the time. They last, on the average, 2 time units, and then there follows a stretch of (on the average) 14/9 time units when the system is up.

Remark The long run proportions $\pi_j, j \geqslant 0$, are often called stationary probabilities. The reason being that if the initial state is chosen according to the probabilities

$$P\{X_0 = j\} = \pi_j, \quad j \geqslant 0$$

then

$\pi_j, j \geqslant 0$, then the probability of being in state j at any time n is also equal to πj. That is, if $P\{X_n = j\} = \pi_j \quad$ for all $n, j \geqslant 0$ The preceding is easily proven by induction, for if we suppose it true for n − 1, then writing

$$\begin{aligned} P\{X_n = j\} &= \sum_i P\{X_n = j | X_{n-1} = i\} P\{X_{n-1} = i\} \\ &= \sum_i P_{ij}\pi_i \quad \text{by the induction hypothesis} \\ &= \pi_j \quad \text{by Equation (4.7)} \end{aligned}$$

Example 4.25 Suppose the numbers of families that check into a hotel on successive days are independent Poisson random variables with mean λ. Also suppose that the number of days that a family stays in the hotel is a geometric random variable with parameter p, 0 < p < 1. (Thus, a family who spent the previous night in the hotel will, independently of how long they have already spent in the hotel, check out the next day with probability p.) Also suppose that all families act independently of each other. Under these conditions it is easy to see that if X_n denotes the number of families that are checked in the hotel at the beginning of day n then $\{X_n, n \geqslant 0\}$ is a Markov chain. Find

(a) the transition probabilities of this Markov chain;

(b) $E[X_n | X_0 = i]$;

(c) the stationary probabilities of this Markov chain.

Solution: (a) To find $P_{i,j}$, suppose there are i families checked into the hotel at the beginning of a day. Because each of these i families will stay for another day with probability q = 1−p it follows that R_i, the number of these families that remain another day, is a binomial (i, q) random variable. So, letting N be the number of new families that check in that day, we see that $P_{i,j} = P(R_i + N = j)$ Conditioning on R_i and using that N is Poisson with mean λ, we obtain

$$P_{i,j} = \sum_{k=0}^{i} P(R_i + N = j | R_i = k)\binom{i}{k} q^k p^{i-k} = \sum_{k=0}^{i} P(N = j - k | R_i = k)\binom{i}{k} q^k p^{i-k} = \sum_{k=0}^{\min(i,j)} P(N = j - k)\binom{i}{k} q^k p^{i-k}$$

$$= \sum_{k=0}^{\min(i,j)} e^{-\lambda}\frac{\lambda^{j-k}}{(j-k)!}\binom{i}{k} q^k p^{i-k}$$

(b) Using the preceding representation $R_i + N$ for the next state from state i, we see that $E[X_n | X_{n-1} = i] = E[R_i + N] = iq + \lambda$ Consequently, $E[X_n | X_{n-1}] = X_{n-1}q + \lambda$ Taking expectations of both sides yields $E[X_n] = \lambda + qE[X_{n-1}]$ Iterating the preceding gives $E[X_n] = \lambda + qE[X_{n-1}]$

$= \lambda + q(\lambda + qE[X_{n-2}]) \quad = \lambda + q\lambda + q^2E[X_{n-2}] = \lambda + q\lambda + q^2(\lambda + qE[X_{n-3}]) = \lambda + q\lambda + q^2\lambda + q^3E[X_{n-3}]$ showing that

$E[X_n] = \lambda\left(1 + q + q^2 + \ldots + q^{n-1}\right) + q^nE[X_0]$ and yielding the result $E[X_n | X_0 = i] = \frac{\lambda(1 - q^n)}{p} + q^n i$

(c) To find the stationary probabilities we will not directly use the complicated transition probabilities derived in part (a). Rather we will make use of the fact that the stationary probability distribution is the only distribution on the initial state that results in the next state having the same distribution. Now, suppose that the initial state X_0 has a Poisson distribution with mean α. That is, assume that the number of families initially in the hotel is Poisson with mean α. Let R denote the number of these families that remain in the hotel at the beginning of the next day. Then, using the result of Example 3.23 that if each of a Poisson distributed (with mean α) number of events occurs with probability q, then the total number of these events that occur is Poisson distributed with mean αq, it follows that R is a Poisson random variable with mean αq. In addition, the number of new families that check in during the day, call it N, is Poisson with mean λ, and is independent of R. Hence, since the sum of independent Poisson random variables is also Poisson distributed, it follows that R + N, the number of guests at the beginning of the next day, is Poisson with mean λ + αq. Consequently, if we choose α so that $\alpha = \lambda + \alpha q$ then the distribution of X_1 would be the same as that of X_0. But this means that when the initial distribution of X_0 is Poisson with mean $\alpha = \frac{\lambda}{p}$, then so is the distribution of X_1, implying that this is the stationary distribution. That is, the stationary probabilities are $\pi_i = e^{-\lambda/p}(\lambda/p)^i/i!, \quad i \geqslant 0$ The preceding model has an important generalization. Namely, consider an organization whose workers are of r distinct types. For instance, the organization could be a law firm and its lawyers could either be juniors, associates, or partners. Suppose that a worker who is currently type i will in the next period become type j with probability $q_{i,j}$ for j = 1, . . . , r or will leave the organization with probability $1 - \sum_{j=1}^{r} q_{i,j}$. In addition, suppose that new workers are hired each period, and that the numbers of types 1, . . . , r workers hired are independent Poisson random variables with means $\lambda_1, \ldots, \lambda_r$. If we let $\mathbf{X}_n = (X_n(1), \ldots, X_n(r))$, where $X_n(i)$ is the number of type i workers in the organization at the beginning of period n, then $\{X_n, n \geqslant 0\}$ is a Markov chain. To compute its stationary probability distribution, suppose that the initial state is chosen so that the number of workers of different types are independent Poisson random variables, with α_i being the mean number of type i workers. That is, suppose that $X_0(1), \ldots, X_0(r)$ are independent Poisson random variables with respective means α1, . . . , αr. Also, let $N_j, j = 1, \ldots, r$, be the number of new type j workers hired during the initial period. Now, fix i, and for j = 1, . . . , r, let $M_i(j)$ be the number of the $X_0(i)$ type i workers who become type j in the next period. Then because each of the Poisson number $X_0(i)$ of type i workers will independently become type j with probability $q_{i,j}, j = 1, \ldots, r$, it follows from the remarks following Example 3.23 that $M_i(1), \ldots, M_i(r)$ are independent Poisson random variables with $M_i(j)$ having mean $\alpha_i q_{i,j}$. Because $X_0(1), \ldots, X_0(r)$ are, by assumption, independent, we can also conclude that the random variables $M_i(j), i, j = 1, \ldots, r$ are all

independent. Because the sum of independent Poisson random variables is also Poisson distributed, the preceding yields that the random variables

$$X_1(j) = N_j + \sum_{i=1}^{r} M_i(j), \quad j = 1, \ldots, r$$

are independent Poisson random variables with means

$$E[X_1(j)] = \lambda_j + \sum_{i=1}^{r} \alpha_i q_{i,j}$$

Hence, if $\alpha_1, \ldots, \alpha_r$ satisfied

$$\alpha_j = \lambda_j + \sum_{i=1}^{r} \alpha_i q_{i,j}, \quad j = 1, \ldots, r$$

then X_1 would have the same distribution as X_0. Consequently, if we let $\alpha_1^o, \ldots, \alpha_r^o$ be such that

$$\alpha_j^o = \lambda_j + \sum_{i=1}^{r} \alpha_i^o q_{i,j}, \quad j = 1, \ldots, r$$

then the stationary distribution of the Markov chain is the distribution that takes the number of workers in each type to be independent Poisson random variables with means $\alpha_1^o, \ldots, \alpha_r^o$. That is,

$$\lim_{n\to\infty} P\{\mathbf{X}_n = (k_1, \ldots, k_r)\} = \prod_{i=1}^{r} e^{-\alpha_i^o} (\alpha_i^o)^{k_i}/k_i!$$

It can be shown that there will be such values $\alpha_j^o, j = 1, \ldots, r,$ provided that, with probability 1, each worker eventually leaves the organization. Also, because there is a unique stationary distribution, there can only be one such set of values.

For state j, define m_{jj} to be the expected number of transitions until a Markov chain, starting in state j, returns to that state. Since, on the average, the chain will spend 1 unit of time in state j for every m_{jj} units of time, it follows that

$$\pi_j = \frac{1}{m_{jj}}$$

In words, the proportion of time in state j equals the inverse of the mean time between visits to j. (The preceding is a special case of a general result, sometimes called the strong law for renewal processes, which will be presented in Chapter 7.)

Example 4.26: (Mean Pattern Times in Markov Chain Generated Data) Consider an irreducible Markov chain $\{X_n, n \geqslant 0\}$ with transition probabilities $P_{i,j}$ and stationary probabilities $\pi_j, j \geqslant 0$. Starting in state r, we are interested in determining the expected number of transitions until the pattern $i_1, i_2, \ldots, i_k$ appears. That is, with $N(i_1, i_2, \ldots, i_k) = \min\{n \geqslant k: X_{n-k+1} = i_1, \ldots, X_n = i_k\}$ we are interested in $E[N(i_1, i_2, \ldots, i_k)|X_0 = r]$ Note that even if $i_1 = r$, the initial state X0 is not considered part of the pattern sequence.

Let $\mu(i, i_1)$ be the mean number of transitions for the chain to enter state i_1, given that the initial state is $i, i \geqslant 0$. The quantities $\mu(i, i_1)$ can be determined as the solution of the following set of equations, obtained by conditioning on the first transition out of state i:

$$\mu(i, i_1) = 1 + \sum_{j \neq i_1} P_{i,j}\mu(j, i_1), \quad i \geqslant 0$$

For the Markov chain $\{X_n, n \geqslant 0\}$ associate a corresponding Markov chain, which we will refer to as the k-chain, whose state at any time is the sequence of the most recent k states of the original chain. (For instance, if k = 3 and $X_2 = 4, X_3 = 1, X_4 = 1,$, then the state of the k-chain at time 4 is (4, 1, 1).) Let $\pi(j_1, \ldots, j_k)$ be the stationary probabilities for the k-chain. Because $\pi(j_1, \ldots, j_k)$ is the proportion of time that the state of the original Markov chain k units ago was j_1 and the following k−1 states, in sequence, were $j_2, \ldots, j_k$, we can conclude that

$$\pi(j_1, \ldots, j_k) = \pi_{j_1} P_{j_1,j_2} \cdots P_{j_{k-1},j_k}$$

Moreover, because the mean number of transitions between successive visits of the k-chain to the state $i_1, i_2, \ldots, i_k$ is equal to the inverse of the stationary probability of that state, we have that

$$E[\text{number of transitions between visits to } i_1, i_2, \ldots, i_k] = \frac{1}{\pi(i_1, \ldots, i_k)} \quad (4.13)$$

Let $A(i_1, \ldots, i_m)$ be the additional number of transitions needed until the pattern appears, given that the first m transitions have taken the chain into states $X_1 = i_1, \ldots, X_m = i_m$. We will now consider whether the pattern has overlaps, where we say that the pattern $i_1, i_2, \ldots, i_k$ has an overlap of size j, j < k, if the sequence of its final j elements is the same as that of its first j elements. That is, it has an overlap of size j if

$$(i_{k-j+1}, \ldots, i_k) = (i_1, \ldots, i_j), \quad j < k$$

Case 1 The pattern $i_1, i_2, \ldots, i_k$ has no overlaps. Because there is no overlap, Equation (4.13) yields

$$E[N(i_1, i_2, \ldots, i_k)|X_0 = i_k] = \frac{1}{\pi(i_1, \ldots, i_k)}$$

Because the time until the pattern occurs is equal to the time until the chain enters state i_1 plus the additional time, we may write

$$E[N(i_1, i_2, \ldots, i_k)|X_0 = i_k] = \mu(i_k, i_1) + E[A(i_1)]$$

The preceding two equations imply

$$E[A(i_1)] = \frac{1}{\pi(i_1, \ldots, i_k)} - \mu(i_k, i_1)$$

Using that

$$E[N(i_1, i_2, \ldots, i_k)|X_0 = r] = \mu(r, i_1) + E[A(i_1)]$$

gives the result

$$E[N(i_1, i_2, \ldots, i_k)|X_0 = r] = \mu(r, i_1) + \frac{1}{\pi(i_1, \ldots, i_k)} - \mu(i_k, i_1)$$

where

$$\pi(i_1, \ldots, i_k) = \pi_{i_1} P_{i_1,i_2} \cdots P_{i_{k-1},i_k}$$

Case 2 Now suppose that the pattern has overlaps and let its largest overlap be of size s. In this case the number of transitions between successive visits of the k-chain to the state $i_1, i_2, \ldots, i_k$ is equal to the additional number of transitions of the original chain until the pattern appears given that it has already made s transitions with the results $X_1 = i_1, \ldots, X_s = i_s$. Therefore, from Equation (4.13)

$$E[A(i_1, \ldots, i_s)] = \frac{1}{\pi(i_1, \ldots, i_k)}$$

But because

$$N(i_1, i_2, \ldots, i_k) = N(i_1, \ldots, i_s) + A(i_1, \ldots, i_s)$$

we have

$$E[N(i_1, i_2, \ldots, i_k)|X_0 = r] = E[N(i_1, i_2, \ldots, i_s)|X_0 = r] + \frac{1}{\pi(i_1, \ldots, i_k)}$$

We can now repeat the same procedure on the pattern $i_1, \ldots, i_s$, continuing to do so until we reach one that has no overlap, and then apply the result from Case 1. For instance, suppose the desired pattern is 1, 2, 3, 1, 2, 3, 1, 2. Then

$$E[N(1,2,3,1,2,3,1,2)|X_0 = r] = E[N(1,2,3,1,2)|X_0 = r] + \frac{1}{\pi(1,2,3,1,2,3,1,2)}$$

Because the largest overlap of the pattern (1, 2, 3, 1, 2) is of size 2, the same argument as in the preceding gives

$$E[N(1,2,3,1,2)|X_0 = r] = E[N(1,2)|X_0 = r] + \frac{1}{\pi(1,2,3,1,2)}$$

Because the pattern (1, 2) has no overlap, we obtain from Case 1 that

$$E[N(1,2)|X_0 = r] = \mu(r,1) + \frac{1}{\pi(1,2)} - \mu(2,1)$$

Putting it together yields

$$E[N(1,2,3,1,2,3,1,2)|X_0 = r] = \mu(r,1) + \frac{1}{\pi_1 P_{1,2}} - \mu(2,1) + \frac{1}{\pi_1 P_{1,2}^2 P_{2,3} P_{3,1}} + \frac{1}{\pi_1 P_{1,2}^3 P_{2,3}^2 P_{3,1}^2}$$

If the generated data is a sequence of independent and identically distributed random variables, with each value equal to j with probability P_j, then the Markov chain has $P_{i,j} = P_j$. In this case, $\pi_j = P_j$. Also, because the

time to go from state i to state j is a geometric random variable with parameter P_j, we have $\mu(i,j) = 1/P_j$. Thus, the expected number of data values that need be generated before the pattern 1, 2, 3, 1, 2, 3, 1, 2 appears would be $\frac{1}{P_1} + \frac{1}{P_1P_2} + \frac{1}{P_1^2P_2^2P_3} + \frac{1}{P_1^3P_2^3P_3^2} = \frac{1}{P_1P_2} + \frac{1}{P_1^2P_2^2P_3} + \frac{1}{P_1^3P_2^3P_3^2}$

The following result is quite useful.

Proposition 4.3 Let $\{X_n, n \geq 1\}$ be an irreducible Markov chain with stationary probabilities $\pi_j, j \geq 0$, and let r be a bounded function on the state space. Then, with probability 1, $\lim_{N\to\infty} \frac{\sum_{n=1}^{N} r(X_n)}{N} = \sum_{j=0}^{\infty} r(j)\pi_j$

Proof. If we let $a_j(N)$ be the amount of time the Markov chain spends in state j during time periods 1, . . . ,N, then $\sum_{n=1}^{N} r(X_n) = \sum_{j=0}^{\infty} a_j(N) r(j)$

Since, $a_j(N)/N \to \pi_j$ the result follows from the preceding upon dividing by N and then letting $N \to \infty$.

If we suppose that we earn a reward r(j) whenever the chain is in state j, then Proposition 4.3 states that our average reward per unit time is $\sum_j r(j)\pi_j$.

Example 4.27 For the four state Bonus Malus automobile insurance system specified in Example 4.7, find the average annual premium paid by a policyholder whose yearly number of claims is a Poisson random variable with mean 1/2. Solution: With $a_k = e^{-1/2}\frac{(1/2)^k}{k!}$, we have $a_0 = 0.6065, \quad a_1 = 0.3033, \quad a_2 = 0.0758$

Therefore, the Markov chain of successive states has the following transition probability matrix: $\begin{Vmatrix} 0.6065 & 0.3033 & 0.0758 & 0.0144 \\ 0.6065 & 0.0000 & 0.3033 & 0.0902 \\ 0.0000 & 0.6065 & 0.0000 & 0.3935 \\ 0.0000 & 0.0000 & 0.6065 & 0.3935 \end{Vmatrix}$ The stationary probabilities are given as the solution of

$$\pi_1 = 0.6065\pi_1 + 0.6065\pi_2,$$
$$\pi_2 = 0.3033\pi_1 + 0.6065\pi_3,$$
$$\pi_3 = 0.0758\pi_1 + 0.3033\pi_2 + 0.6065\pi_4, \quad \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

Rewriting the first three of these equations gives $\pi_2 = \frac{1 - 0.6065}{0.6065}\pi_1, \quad \pi_3 = \frac{\pi_2 - 0.3033\pi_1}{0.6065}, \quad \pi_4 = \frac{\pi_3 - 0.0758\pi_1 - 0.3033\pi_2}{0.6065}$

or

$$\pi_2 = 0.6488\pi_1,$$
$$\pi_3 = 0.5697\pi_1,$$
$$\pi_4 = 0.4900\pi_1$$

Using that $\sum_{i=1}^{4} \pi_i = 1$ gives the solution (rounded to four decimal places) $\pi_1 = 0.3692, \quad \pi_2 = 0.2395, \quad \pi_3 = 0.2103, \quad \pi_4 = 0.1809$

Therefore, the average annual premium paid is

$$200\pi_1 + 250\pi_2 + 400\pi_3 + 600\pi_4 = 326.375$$

4.5 Some Applications:

4.5.1 The Gambler's Ruin Problem: Consider a gambler who at each play of the game has probability p of winning one unit and probability q = 1 − p of losing one unit. Assuming that successive plays of the game are independent, what is the probability that, starting with I units, the gambler's fortune will reach N before reaching 0? If we let X_n denote the player's fortune at time n, then the process $\{X_n$, n = 0, 1, 2, . . .} is a Markov chain with transition probabilities

$$P_{00} = P_{NN} = 1,$$
$$P_{i,i+1} = p = 1 - P_{i,i-1}, \quad i = 1, 2, \ldots, N-1$$

This Markov chain has three classes, namely, {0}, {1, 2, . . . ,N −1}, and {N}; the first and third class being recurrent and the second transient. Since each transient state is visited only finitely often, it follows that, after some finite amount of time, the gambler will either attain his goal of N or go broke. Let P_i, i = 0, 1, . . . ,N, denote the probability that, starting with i, the gambler's fortune will eventually reach N. By conditioning on the outcome of the initial play of the game we obtain

$$P_i = pP_{i+1} + qP_{i-1}, \quad i = 1, 2, \ldots, N-1$$

or equivalently, since p + q = 1,

$$pP_i + qP_i = pP_{i+1} + qP_{i-1}$$

or

$$P_{i+1} - P_i = \frac{q}{p}(P_i - P_{i-1}), \quad i = 1, 2, \ldots, N-1$$

Hence,

$$P_2 - P_1 = \frac{q}{p}(P_1 - P_0) = \frac{q}{p}P_1,$$
$$P_3 - P_2 = \frac{q}{p}(P_2 - P_1) = \left(\frac{q}{p}\right)^2 P_1,$$
$$\vdots$$
$$P_i - P_{i-1} = \frac{q}{p}(P_{i-1} - P_{i-2}) = \left(\frac{q}{p}\right)^{i-1} P_1,$$
$$\vdots$$
$$P_N - P_{N-1} = \left(\frac{q}{p}\right)(P_{N-1} - P_{N-2}) = \left(\frac{q}{p}\right)^{N-1} P_1$$

since P_0 = 0, we obtain from the preceding line that

$$P_i - P_1 = P_1\left[\left(\frac{q}{p}\right) + \left(\frac{q}{p}\right)^2 + \cdots + \left(\frac{q}{p}\right)^{i-1}\right]$$

Adding the first i − 1 of these equations yields

$$P_i = \begin{cases} \frac{1 - (q/p)^i}{1 - (q/p)} P_1, & \text{if } \frac{q}{p} \neq 1 \\ iP_1, & \text{if } \frac{q}{p} = 1 \end{cases}$$

or

Now, using the fact that P_N = 1, we obtain

$$P_1 = \begin{cases} \frac{1 - (q/p)}{1 - (q/p)^N}, & \text{if } p \neq \frac{1}{2} \\ \frac{1}{N}, & \text{if } p = \frac{1}{2} \end{cases}$$

(4.14) and hence Note that, as N→ ∞,

$$P_i \to \begin{cases} 1 - \left(\frac{q}{p}\right)^i, & \text{if } p > \frac{1}{2} \\ 0, & \text{if } p \leqslant \frac{1}{2} \end{cases}$$

Thus, if p > 1/2 , there is a positive probability that the gambler's fortune will increase indefinitely; while if p $\leqslant$ 1/2 , the gambler will, with probability 1, go broke against an infinitely rich adversary.

Example 4.28 Suppose Max and Patty decide to flip pennies; the one coming closest to the wall wins. Patty, being the better player, has a probability 0.6 of winning on each flip. (a) If Patty starts with five pennies and Max with ten, what is the probability that Patty will wipe Max out? (b) What if Patty starts with 10 and Max with 20?

Solution: (a) The desired probability is obtained from Equation (4.14) by letting i = 5, N = 15, and p = 0.6. Hence, the desired probability is $\frac{1 - (\frac{2}{3})^5}{1 - (\frac{2}{3})^{15}} \approx 0.87$ (b) The

$$\frac{1-\left(\frac{2}{3}\right)^{10}}{1-\left(\frac{2}{3}\right)^{30}} \approx 0.98$$

desired probability is For an application of the gambler's ruin problem to drug testing, suppose that two new drugs have been developed for treating a certain disease. Drug i has a cure rate P_i, i = 1, 2, in the sense that each patient treated with drug i will be cured with probability P_i. These cure rates, however, are not known, and suppose we are interested in a method for deciding whether P1 > P2 or P2 > P1. To decide upon one of these alternatives, consider the following test: Pairs of patients are treated sequentially with one member of the pair receiving drug 1 and the other drug 2. The results for each pair are determined, and the testing stops when the cumulative number of cures using one of the drugs exceeds the cumulative number of cures when using the other by some fixed predetermined

$$X_j = \begin{cases} 1, & \text{if the patient in the } j\text{th pair to receive drug number 1 is cured} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_j = \begin{cases} 1, & \text{if the patient in the } j\text{th pair to receive drug number 2 is cured} \\ 0, & \text{otherwise} \end{cases}$$

number. More formally, let For a predetermined positive integer M the test stops after pair N where N is the first value of n such that either $X_1 + \cdots + X_n - (Y_1 + \cdots + Y_n) = M$ or $X_1 + \cdots + X_n - (Y_1 + \cdots + Y_n) = -M$ In the former case we then assert that P1 > P2, and in the latter that P2 > P1. In order to help ascertain whether the preceding is a good test, one thing we would like to know is the probability of it leading to an incorrect decision. That is, for given P1 and P2 where P1 > P2, what is the probability that the test will incorrectly assert that P2 > P1? To determine this probability, note that after each pair is checked the cumulative difference of cures using drug 1 versus drug 2 will either go up by 1 with probability P1(1−P2)—since this is the probability that drug 1 leads to a cure and drug 2 does not—or go down by 1 with probability (1−P1)P2, or remain the same with probability P1P2 + (1−P1)(1−P2). Hence, if we only consider those pairs in which the cumulative difference changes, then the difference will go up 1 with probability $p = P\{\text{up 1} | \text{up 1 or down 1}\}$ $= \frac{P_1(1-P_2)}{P_1(1-P_2) + (1-P_1)P_2}$ and down 1 with probability $q = 1 - p = \frac{P_2(1-P_1)}{P_1(1-P_2) + (1-P_1)P_2}$ Hence, the probability that the test will assert that P2 > P1 is equal to the probability that a gambler who wins each (one unit) bet with probability p will go down M before going up M. But Equation (4.14) with i = M, N = 2M, shows

$$P\{\text{test asserts that } P_2 > P_1\} = 1 - \frac{1-(q/p)^M}{1-(q/p)^{2M}} = \frac{1}{1+(p/q)^M}$$

that this probability is given by Thus, for instance, if P1 = 0.6 and P2 = 0.4 then the probability of an incorrect decision is 0.017 when M = 5 and reduces to 0.0003 when M = 10.

4.5.2 A Model for Algorithmic Efficiency: The following optimization problem is called a linear program

$$\begin{aligned} &\text{minimize } \mathbf{cx}, \\ &\text{subject to } \mathbf{Ax} = \mathbf{b}, \\ &\mathbf{x} \geqslant 0 \end{aligned}$$

where A is an m × n matrix of fixed constants; c = (c1, . . . , cn) and b = (b1, . . . , bm) are vectors of fixed constants; and x = (x1, . . . , xn) is the n-vector of nonnegative values that is to be chosen to minimize $\mathbf{cx} \equiv \sum_{i=1}^{n} c_i x_i$. Supposing that n > m, it can be shown that the optimal x can always be chosen to have at least n − m components equal to 0—that is, it can always be taken to be one of the so-called extreme points of the feasibility region. The simplex algorithm solves this linear program by moving from an extreme point of the feasibility region to a better (in terms of the objective function cx) extreme point (via the pivot operation) until the optimal is reached. Because there can be as many as $N \equiv \binom{n}{m}$ such extreme points, it would seem that this method might take many iterations, but, surprisingly to some, this does not appear to be the case in practice. To obtain a feel for whether or not the preceding statement is surprising, let us consider a simple probabilistic (Markov chain) model as to how the algorithm moves along the extreme points. Specifically, we will suppose that if at any time the algorithm is at the jth best extreme point then after the next pivot the resulting extreme point is equally likely to be any of the j−1 best. Under this assumption, we show that the time to get from the Nth best to the best extreme point has approximately, for large N, a normal distribution with mean and variance equal to the logarithm (base e) of N. Consider a Markov chain for which P11 = 1 and

$$P_{ij} = \frac{1}{i-1}, \quad j = 1, \ldots, i-1, \; i > 1$$

and let Ti denote the number of transitions needed to go from state i to state 1. A recursive formula for E[Ti] can be obtained by conditioning on the initial transition:

$$E[T_i] = 1 + \frac{1}{i-1}\sum_{j=1}^{i-1} E[T_j]$$

Starting with E[T1] = 0, we successively see that

$$\begin{aligned} E[T_2] &= 1, \\ E[T_3] &= 1 + \tfrac{1}{2}, \\ E[T_4] &= 1 + \tfrac{1}{3}(1 + 1 + \tfrac{1}{2}) = 1 + \tfrac{1}{2} + \tfrac{1}{3} \end{aligned}$$

and it is not difficult to guess and then prove inductively that

$$E[T_i] = \sum_{j=1}^{i-1} 1/j$$

However, to obtain a more complete description of T_N, we will use the representation

$$T_N = \sum_{j=1}^{N-1} I_j$$

where

$$I_j = \begin{cases} 1, & \text{if the process ever enters } j \\ 0, & \text{otherwise} \end{cases}$$

The importance of the preceding representation stems from the following:

Proposition 4.4: $I_1, \ldots, I_{N-1}$ are independent and $P\{I_j = 1\} = 1/j, \quad 1 \leqslant j \leqslant N-1$ Proof. Given $I_{j+1}, \ldots, I_N$, let n = $\min\{i: i > j, I_i = 1\}$ denote the lowest numbered state, greater than j, that is entered. Thus we know that the process enters state n and the next state entered is one of the states 1, 2, . . . , j. Hence, as the next state from state n is equally likely to be any of the lower number states 1, 2, . . . , n − 1 we see that

$$P\{I_j = 1 | I_{j+1}, \ldots, I_N\} = \frac{1/(n-1)}{j/(n-1)} = 1/j$$

. Hence, $P\{I_j = 1\} = 1/j$, and independence follows since the preceding conditional probability does not depend on $I_{j+1}, \ldots, I_N$.

Corollary 4.5: (i) $E[T_N] = \sum_{j=1}^{N-1} 1/j$. (ii) $\text{Var}(T_N) = \sum_{j=1}^{N-1} (1/j)(1-1/j)$. (iii) For N large, T_N has approximately a normal distribution with mean logN and variance logN. Proof. Parts (i) and (ii) follow from Proposition 4.4 and the representation $T_N = \sum_{j=1}^{N-1} I_j$. Part (iii) follows from the central limit theorem since

$$\int_1^N \frac{dx}{x} < \sum_1^{N-1} 1/j < 1 + \int_1^{N-1} \frac{dx}{x}$$

or

$$\log N < \sum_1^{N-1} 1/j < 1 + \log(N-1)$$

and so

$$\log N \approx \sum_{j=1}^{N-1} 1/j$$

Returning to the simplex algorithm, if we assume that

$$N = \binom{n}{m} \sim \frac{n^{n+1/2}}{(n-m)^{n-m+1/2} m^{m+1/2} \sqrt{2\pi}}$$

n, m, and n − m are all large, we have by Stirling's approximation that and so, letting c = n/m,

$$\begin{aligned} \log N \sim &\left(mc + \tfrac{1}{2}\right)\log(mc) - \left(m(c-1) + \tfrac{1}{2}\right)\log(m(c-1)) \\ &- \left(m + \tfrac{1}{2}\right)\log m - \tfrac{1}{2}\log(2\pi) \end{aligned}$$

or

$$\log N \sim m\left[c \log \frac{c}{c-1} + \log(c-1)\right]$$

Now, as $\lim_{x\to\infty} x \log[x/(x-1)] = 1$, 1)], it follows that, when c is large,

$$\log N \sim m[1 + \log(c-1)]$$

Thus, for instance, if n = 8000, m = 1000, then the number of necessary transitions is approximately normally distributed with mean and variance equal to 1000(1 + log 7) ≈ 3000. Hence, the number of necessary transitions would be roughly between $3000 \pm 2\sqrt{3000}$ or roughly 3000 ± 110, 95 percent of the time.

4.5.3 Using a Random Walk to Analyze a Probabilistic Algorithm for the Satisfiability Problem: Consider a Markov chain with states 0, 1, . . . , n having $P_{0,1} = 1, \quad P_{i,i+1} = p, \quad P_{i,i-1} = q = 1 - p, \quad 1 \leqslant i < n$ and suppose that we are interested in studying the time that it takes for the chain to go from state 0 to state n. One approach to obtaining the mean time to reach state n would be to let m_i denote the mean time to go from state i to state n, i = 0, . . . , n − 1. If we then condition on the initial transition, we obtain the following set of equations: $m_0 = 1 + m_1,$

$$\begin{aligned} m_i &= E[\text{time to reach } n | \text{next state is } i+1]p \\ &\quad + E[\text{time to reach } n | \text{next state is } i-1]q \\ &= (1 + m_{i+1})p + (1 + m_{i-1})q \\ &= 1 + pm_{i+1} + qm_{i-1}, \quad i = 1, \ldots, n-1 \end{aligned}$$

Whereas the preceding equations can be solved for $m_i, i = 0, \ldots, n-1,$ we do not pursue their solution; we instead make use of the special structure of the Markov chain to obtain a simpler set of equations. To start, let N_i denote the number of additional transitions that it takes the chain when it first enters state i until it enters state i + 1. By the Markovian property, it follows that these random variables $N_i, i = 0, \ldots, n-1$ are independent. Also, we can express $N_{0,n}$, the number of transitions that it takes the chain to go from state 0 to state n, as $N_{0,n} = \sum_{i=0}^{n-1} N_i$ (4.15), Letting μi = E[Ni] we obtain, upon conditioning on the next transition after the chain enters state i, that for i = 1, . . . , n − 1 $\mu_i = E[N_i]$, $\mu_i = 1 + E[\text{number of additional transitions to reach } i+1 | \text{chain to } i-1]q$, Now, if the chain next enters state i−1, then in order for it to reach i + 1 it must first return to state i and must then go from state i to state i + 1. Hence, we have from the preceding that $\mu_i = 1 + E[N_{i-1}^* + N_i^*]q$ where N_{i-1}^* and N_i^* are, respectively, the additional number of transitions to return to state i from i − 1 and the number to then go from i to i + 1. Now, it follows from the Markovian property that these random variables have, respectively, the same distributions as N_{i-1} and N_i. In addition, they are independent (although we will only use this when we compute the variance of $N_{0,n}$). Hence, we see that $\mu_i = 1 + q(\mu_{i-1} + \mu_i)$ or

$$\mu_i = \frac{1}{p} + \frac{q}{p}\mu_{i-1}, \quad i = 1, \ldots, n-1$$

. Starting with μ0 = 1, and letting α = q/p, we obtain from the preceding recursion that

$$\begin{aligned} \mu_1 &= 1/p + \alpha, \\ \mu_2 &= 1/p + \alpha(1/p + \alpha) = 1/p + \alpha/p + \alpha^2, \\ \mu_3 &= 1/p + \alpha(1/p + \alpha/p + \alpha^2) \\ &= 1/p + \alpha/p + \alpha^2/p + \alpha^3 \end{aligned}$$

In general, we see that $\mu_i = \frac{1}{p}\sum_{j=0}^{i-1}\alpha^j + \alpha^i, \quad i = 1, \ldots, n-1$ (4.16), Using Equation (4.15), we now get

$$E[N_{0,n}] = 1 + \frac{1}{p}\sum_{i=1}^{n-1}\sum_{j=0}^{i-1}\alpha^j + \sum_{i=1}^{n-1}\alpha^i$$

When p = 1/2, and so α = 1, we see from the preceding that $E[N_{0,n}] = 1 + (n-1)n + n - 1 = n^2$ When p ≠ 1/2

$$E[N_{0,n}] = 1 + \frac{1}{p(1-\alpha)}\sum_{i=1}^{n-1}(1-\alpha^i) + \frac{\alpha - \alpha^n}{1-\alpha} = 1 + \frac{1+\alpha}{1-\alpha}\left[n - 1 - \frac{(\alpha - \alpha^n)}{1-\alpha}\right] + \frac{\alpha - \alpha^n}{1-\alpha} = 1 + \frac{2\alpha^{n+1} - (n+1)\alpha^2 + n - 1}{(1-\alpha)^2}$$

, we obtain where the second equality used the fact that p = 1/(1 + α). Therefore, we see that when α > 1, or equivalently when p < 1/2, the expected number of transitions to reach n is an exponentially increasing function of n. On the other hand, when p = 1/2 , $E[N_{0,n}] = n^2$, and when p > 1/2, $E[N_{0,n}]$ is, for large n, essentially linear in n. Let us now compute $\text{Var}(N_{0,n})$. To do so, we will again make use of the representation given by Equation (4.15). Letting $v_i = \text{Var}(N_i)$, we start by determining the v_i recursively by using the conditional variance formula. Let S_i = 1 if the first transition out of state i is into state i + 1, and let S_i = −1 if the transition is into state i − 1, i = 1, . . . , n − 1. Then, given that S_i = 1: $N_i = 1$, given that S_i = −1: $N_i = 1 + N_{i-1}^* + N_i^*$. Hence, $E[N_i|S_i = 1] = 1, \quad E[N_i|S_i = -1] = 1 + \mu_{i-1} + \mu_i$ implying that $\text{Var}(E[N_i|S_i]) = \text{Var}(E[N_i|S_i] - 1) = (\mu_{i-1} + \mu_i)^2 q - (\mu_{i-1} + \mu_i)^2 q^2 = qp(\mu_{i-1} + \mu_i)^2$. Also, since N_{i-1}^* and N_i^*, the numbers of transitions to return from state i − 1 to i and to then go from state i to state i + 1 are, by the Markovian property, independent random variables having the same distributions as N_{i-1} and N_i, respectively, we see that $\text{Var}(N_i|S_i = 1) = 0, \quad \text{Var}(N_i|S_i = -1) = v_{i-1} + v_i$. Hence, $E[\text{Var}(N_i|S_i)] = q(v_{i-1} + v_i)$ From the conditional variance formula, we thus obtain $v_i = pq(\mu_{i-1} + \mu_i)^2 + q(v_{i-1} + v_i)$ or, equivalently, $v_i = q(\mu_{i-1} + \mu_i)^2 + \alpha v_{i-1}, \quad i = 1, \ldots, n-1$. Starting with v_0 = 0, we obtain from the preceding recursion that

$$\begin{aligned} v_1 &= q(\mu_0 + \mu_1)^2, \\ v_2 &= q(\mu_1 + \mu_2)^2 + \alpha q(\mu_0 + \mu_1)^2, \\ v_3 &= q(\mu_2 + \mu_3)^2 + \alpha q(\mu_1 + \mu_2)^2 + \alpha^2 q(\mu_0 + \mu_1)^2 \end{aligned}$$

In general, we have for i > 0,

$$v_i = q\sum_{j=1}^{i}\alpha^{i-j}(\mu_{j-1} + \mu_j)^2$$

(4.17) Therefore, we see that

$$\text{Var}(N_{0,n}) = \sum_{i=0}^{n-1} v_i = q\sum_{i=1}^{n-1}\sum_{j=1}^{i}\alpha^{i-j}(\mu_{j-1} + \mu_j)^2$$

where μ_j is given by Equation (4.16).

We see from Equations (4.16) and (4.17) that when $p \geqslant \frac{1}{2}$, , and so $\alpha \leqslant 1,$ that μ_i and v_i, the mean and variance of the number of transitions to go from state i to i + 1, do not increase too rapidly in i. For instance, when $p = \frac{1}{2}$ it follows from Equations (4.16) and (4.17) that $\mu_i = 2i + 1$ and $v_i = \frac{1}{2}\sum_{j=1}^{i}(4j)^2 = 8\sum_{j=1}^{i} j^2$

Hence, since $N_{0,n}$ is the sum of independent random variables, which are of roughly similar magnitudes when $p \geqslant \frac{1}{2}$, , it follows in this case from the central limit theorem that N0,n is, for large n, approximately normally distributed. In particular, when $p = \frac{1}{2}, N_{0,n}$ is approximately normal with mean n^2 and variance

$$\text{Var}(N_{0,n}) = 8\sum_{i=1}^{n-1}\sum_{j=1}^{i} j^2 = 8\sum_{j=1}^{n-1}\sum_{i=j}^{n-1} j^2 = 8\sum_{j=1}^{n-1}(n-j)j^2 \approx 8\int_1^{n-1}(n-x)x^2\,dx \approx \tfrac{2}{3}n^4$$

Example 4.29 (The Satisfiability Problem): A Boolean variable x is one that takes on either of two values: TRUE or FALSE. If $x_i, i \geqslant 1$ are Boolean variables, then a Boolean clause of the form $x_1 + \bar{x}_2 + x_3$ is TRUE if x1 is TRUE, or if x2 is FALSE, or if x3 is TRUE. That is, the symbol "+" means "or" and $\bar{x}$ is TRUE if x is FALSE and vice versa. A Boolean formula is a combination of clauses such as $(x_1 + \bar{x}_2) * (x_1 + x_3) * (x_2 + \bar{x}_3) * (\bar{x}_1 + \bar{x}_2) * (x_1 + x_2)$ In the preceding, the terms between the parentheses represent clauses, and the formula is TRUE if all the clauses are TRUE, and is FALSE otherwise. For a given Boolean formula, the satisfiability problem is either to determine values for the variables that result in the formula being TRUE, or to determine that the formula is never true. For instance, one set of values that makes the preceding formula TRUE is to set x1 =TRUE, x2 =FALSE, and x3 =FALSE. Consider a formula of the n Boolean variables $x_1, \ldots, x_n$ and suppose that each

clause in this formula refers to exactly two variables. We will now present a probabilistic algorithm that will either find values that satisfy the formula or determine to a high probability that it is not possible to satisfy it. To begin, start with an arbitrary setting of values. Then, at each stage choose a clause whose value is FALSE, and randomly choose one of the Boolean variables in that clause and change its value. That is, if the variable has value TRUE then change its value to FALSE, and vice versa. If this new setting makes the formula TRUE then stop, otherwise continue in the same fashion. If you have not stopped after $n^2(1 + 4\sqrt{\frac{2}{3}})$ repetitions, then declare that the formula cannot be satisfied. We will now argue that if there is a satisfiable assignment then this algorithm will find such an assignment with a probability very close to 1. Let us start by assuming that there is a satisfiable assignment of truth values and let $\mathcal{A}$ be such an assignment. At each stage of the algorithm there is a certain assignment of values. Let Y_j denote the number of the n variables whose values at the jth stage of the algorithm agree with their values in $\mathcal{A}$. For instance, suppose that n = 3 and $\mathcal{A}$ consists of the settings x1 = x2 = x3 = TRUE. If the assignment of values at the jth step of the algorithm is x1 =TRUE, x2 =x3 =FALSE, then Y_j =1. Now, at each stage, the algorithm considers a clause that is not satisfied, thus implying that at least one of the values of the two variables in this clause does not agree with its value in $\mathcal{A}$. As a result, when we randomly choose one of the variables in this clause then there is a probability of at least 1/2 that $Y_{j+1} = Y_j + 1$ and at most 1/2 that $Y_{j+1} = Y_j - 1$. That is, independent of what has previously transpired in the algorithm, at each stage the number of settings in agreement with those in A will either increase or decrease by 1 and the probability of an increase is at least 1/2 (it is 1 if both variables have values different from their values in $\mathcal{A}$). Thus, even though the process $Y_j, j \geq 0$ is not itself a Markov chain (why not?) it is intuitively clear that both the expectation and the variance of the number of stages of the algorithm needed to obtain the values of $\mathcal{A}$ will be less than or equal to the expectation and variance of the number of transitions to go from state 0 to state n in the Markov chain of Section 4.5.2.

Hence, if the algorithm has not yet terminated because it found a set of satisfiable values different from that of $\mathcal{A}$, it will do so within an expected time of at most n^2 and with a standard deviation of at most $n^2\sqrt{\frac{2}{3}}$. In addition, since the time for the Markov chain to go from 0 to n is approximately normal when n is large we can be quite certain that a satisfiable assignment will be reached by $n^2 + 4(n^2\sqrt{\frac{2}{3}})$ stages, and thus if one has not been found by this number of stages of the algorithm we can be quite certain that there is no satisfiable assignment. Our analysis also makes it clear why we assumed that there are only two variables in each clause. For if there were k, k > 2, variables in a clause then as any clause that is not presently satisfied may have only one incorrect setting, a randomly chosen variable whose value is changed might only increase the number of values in agreement with $\mathcal{A}$ with probability 1/k and so we could only conclude from our prior Markov chain results that the mean time to obtain the values in $\mathcal{A}$ is an exponential function of n, which is not an efficient algorithm when n is large.

4.6 Mean Time Spent in Transient States: Consider now a finite state Markov chain and suppose that the states are numbered so that T = {1, 2, . . . , t} denotes the set of transient states. Let $\mathbf{P}_T = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1t} \\ \vdots & \vdots & \vdots & \vdots \\ P_{t1} & P_{t2} & \cdots & P_{tt} \end{bmatrix}$ and note that since $\mathbf{P}_T$ specifies only the transition probabilities from transient states into transient states, some of its row sums are less than 1 (otherwise, T would be a closed class of states). For transient states i and j, let s_{ij} denote the expected number of time periods that the Markov chain is in state j, given that it starts in state i. Let $\delta_{i,j}$ = 1 when i = j and let it be 0 otherwise. Condition on the initial transition to obtain

$$\begin{aligned} s_{ij} &= \delta_{i,j} + \sum_k P_{ik} s_{kj} \\ &= \delta_{i,j} + \sum_{k=1}^{t} P_{ik} s_{kj} \end{aligned} \qquad (4.18)$$

where the final equality follows since it is impossible to go from a recurrent to a transient state, implying that s_{kj} = 0 when k is a recurrent state. Let S denote the matrix of values $s_{ij}, i, j = 1, \ldots, t.$ That is, $\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1t} \\ \vdots & \vdots & \vdots & \vdots \\ s_{t1} & s_{t2} & \cdots & s_{tt} \end{bmatrix}$ In matrix notation, Equation (4.18) can be written as $\mathbf{S} = \mathbf{I} + \mathbf{P}_T\mathbf{S}$ where I is the identity matrix of size t. Because the preceding equation is equivalent to $(\mathbf{I} - \mathbf{P}_T)\mathbf{S} = \mathbf{I}$ we obtain, upon multiplying both sides by $(\mathbf{I} - \mathbf{P}_T)^{-1}$, $\mathbf{S} = (\mathbf{I} - \mathbf{P}_T)^{-1}$ That is, the quantities s_{ij}, i ∈ T, j ∈ T, can be obtained by inverting the matrix $\mathbf{I} - \mathbf{P}_T$. (The existence of the inverse is easily established.)

Example 4.30: Consider the gambler's ruin problem with p = 0.4 and N = 7. Starting with 3 units, determine
(a) the expected amount of time the gambler has 5 units,
(b) the expected amount of time the gambler has 2 units.

Solution: The matrix $\mathbf{P}_T$, which specifies P_{ij}, i, j ∈ {1, 2, 3, 4, 5, 6}, is as follows:

$$\mathbf{P}_T = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 0.4 & 0 & 0 & 0 & 0 \\ 2 & 0.6 & 0 & 0.4 & 0 & 0 & 0 \\ 3 & 0 & 0.6 & 0 & 0.4 & 0 & 0 \\ 4 & 0 & 0 & 0.6 & 0 & 0.4 & 0 \\ 5 & 0 & 0 & 0 & 0.6 & 0 & 0.4 \\ 6 & 0 & 0 & 0 & 0 & 0.6 & 0 \end{array}$$

Inverting $\mathbf{I} - \mathbf{P}_T$ gives

$$\mathbf{S} = (\mathbf{I} - \mathbf{P}_T)^{-1} = \begin{bmatrix} 1.6149 & 1.0248 & 0.6314 & 0.3691 & 0.1943 & 0.0777 \\ 1.5372 & 2.5619 & 1.5784 & 0.9228 & 0.4857 & 0.1943 \\ 1.4206 & 2.3677 & 2.9990 & 1.7533 & 0.9228 & 0.3691 \\ 1.2458 & 2.0763 & 2.6299 & 2.9990 & 1.5784 & 0.6314 \\ 0.9835 & 1.6391 & 2.0763 & 2.3677 & 2.5619 & 1.0248 \\ 0.5901 & 0.9835 & 1.2458 & 1.4206 & 1.5372 & 1.6149 \end{bmatrix}$$

Hence, $s_{3,5} = 0.9228, \quad s_{3,2} = 2.3677$

For i ∈ T, j ∈ T, the quantity f_{ij}, equal to the probability that the Markov chain ever makes a transition into state j given that it starts in state i, is easily determined from $\mathbf{P}_T$. To determine the relationship, let us start by deriving an expression for s_{ij} by conditioning on whether state j is ever entered. This yields

$$\begin{aligned} s_{ij} &= E[\text{time in } j | \text{start in } i, \text{ ever transit to } j] f_{ij} \\ &\quad + E[\text{time in } j | \text{start in } i, \text{ never transit to } j](1 - f_{ij}) \\ &= (\delta_{i,j} + s_{ij}) f_{ij} + \delta_{i,j}(1 - f_{i,j}) \\ &= \delta_{i,j} + f_{ij} s_{jj} \end{aligned}$$

since s_{ij} is the expected number of additional time periods spent in state j given that it is eventually entered from state i. Solving the preceding equation yields $f_{ij} = \frac{s_{ij} - \delta_{i,j}}{s_{jj}}$.

Example 4.31: In Example 4.30, what is the probability that the gambler ever has a fortune of 1? Solution: Since $s_{3,1}$ = 1.4206 and $s_{1,1}$ = 1.6149, then

$$f_{3,1} = \frac{s_{3,1}}{s_{1,1}} = 0.8797$$

As a check, note that $f_{3,1}$ is just the probability that a gambler starting with 3 reaches 1 before 7. That is, it is the probability that the gambler's fortune will go down 2 before going up 4; which is the probability that a gambler starting with 2 will go broke before reaching 6. Therefore,

$$f_{3,1} = \frac{1 - (0.6/0.4)^2}{1 - (0.6/0.4)^6} = 0.8797$$

which checks with our earlier answer.

Suppose we are interested in the expected time until the Markov chain enters some sets of states A, which need not be the set of recurrent states. We can reduce this back to the previous situation by making all states in A absorbing states. That is, reset the transition probabilities of states in A to satisfy $P_{i,i} = 1, \quad i \in A$ This transforms the states of A into recurrent states, and transforms any state outside of A from which an eventual transition into A is possible into a transient state. Thus, our previous approach can be used.

4.7 Branching Processes: In this section we consider a class of Markov chains, known as branching processes, which have a wide variety of applications in the biological, sociological, and engineering sciences. Consider a population consisting of individuals able to produce offspring of the same kind. Suppose that each individual will, by the end of its lifetime, have produced j new offspring with probability $P_j, j \geq 0$, independently of the numbers produced by other individuals. We suppose that $P_j < 1$ for all $j \geq 0$ The number of individuals initially present, denoted by X0, is called the size of the zeroth generation. All offspring of the zeroth generation constitute the first generation and their number is denoted by X_1. In general, let X_n denote the size of the nth generation. It follows that $\{X_n, n = 0, 1, \ldots\}$ is a Markov chain having as its state space the set of nonnegative integers. Note that state 0 is a recurrent state, since clearly $P_{00} = 1$. Also, if $P_0 > 0$, all other states are transient. This follows since $P_{i0} = P_0^i$, which implies that starting with i individuals there is a positive probability of at least P_0^i that no later generation will ever consist of i individuals. Moreover, since any finite set of transient states {1, 2, . . . , n} will be visited only finitely often, this leads to the important conclusion that, if $P_0 > 0$, then the population will either die out or its size will converge to infinity. Let $\mu = \sum_{j=0}^{\infty} jP_j$ denote the mean number of offspring of a single individual, and let $\sigma^2 = \sum_{j=0}^{\infty} (j-\mu)^2 P_j$ be the variance of the number of offspring produced by a single individual. Let us suppose that X0 = 1, that is, initially there is a single individual present. We calculate E[Xn] and Var(Xn) by first noting that we may write $X_n = \sum_{i=1}^{X_{n-1}} Z_i$ where Z_i represents the number of offspring of the ith individual of the (n−1)st generation. By conditioning on Xn−1, we obtain

$$\begin{aligned} E[X_n] &= E[E[X_n|X_{n-1}]] \\ &= E\left[E\left[\sum_{i=1}^{X_{n-1}} Z_i | X_{n-1}\right]\right] \\ &= E[X_{n-1}\mu] \\ &= \mu E[X_{n-1}] \end{aligned}$$

where we have used the fact that $E[Z_i] = \mu$. Since E[X0] = 1, the preceding yields

$$\begin{aligned} E[X_1] &= \mu, \\ E[X_2] &= \mu E[X_1] = \mu^2, \\ &\vdots \\ E[X_n] &= \mu E[X_{n-1}] = \mu^n \end{aligned}$$

Similarly, Var(Xn) may be obtained by using the conditional variance formula $\mathrm{Var}(X_n) = E[\mathrm{Var}(X_n|X_{n-1})] + \mathrm{Var}(E[X_n|X_{n-1}])$ Now, given Xn−1,Xn is just the sum of Xn−1 independent random variables each having the distribution $\{P_j, j \geq 0\}$. Hence, $E[X_n|X_{n-1}] = X_{n-1}\mu, \quad \mathrm{Var}(X_n|X_{n-1}) = X_{n-1}\sigma^2$ The conditional variance formula now yields

$$\begin{aligned} \mathrm{Var}(X_n) &= E[X_{n-1}\sigma^2] + \mathrm{Var}(X_{n-1}\mu) \\ &= \sigma^2\mu^{n-1} + \mu^2\mathrm{Var}(X_{n-1}) \\ &= \sigma^2\mu^{n-1} + \mu^2(\sigma^2\mu^{n-2} + \mu^2\mathrm{Var}(X_{n-2})) \\ &= \sigma^2(\mu^{n-1} + \mu^n) + \mu^4\mathrm{Var}(X_{n-2}) \\ &= \sigma^2(\mu^{n-1} + \mu^n) + \mu^4(\sigma^2\mu^{n-3} + \mu^2\mathrm{Var}(X_{n-3})) \\ &= \sigma^2(\mu^{n-1} + \mu^n + \mu^{n+1}) + \mu^6\mathrm{Var}(X_{n-3}) \\ &= \cdots \\ &= \sigma^2(\mu^{n-1} + \mu^n + \cdots + \mu^{2n-2}) + \mu^{2n}\mathrm{Var}(X_0) \\ &= \sigma^2(\mu^{n-1} + \mu^n + \cdots + \mu^{2n-2}) \end{aligned}$$

Therefore,

$$\mathrm{Var}(X_n) = \begin{cases} \sigma^2\mu^{n-1}\left(\frac{1-\mu^n}{1-\mu}\right), & \text{if } \mu \neq 1 \\ n\sigma^2, & \text{if } \mu = 1 \end{cases} \quad (4.19)$$

Let π_0 denote the probability that the population will eventually die out (under the assumption that $X_0 = 1$). More formally, $\pi_0 = \lim_{n\to\infty} P\{X_n = 0 | X_0 = 1\}$ The problem of determining the value of π_0 was first raised in connection with the extinction of family surnames by Galton in 1889. We first note that $\pi_0 = 1$ if $\mu < 1$. This follows since

$$\begin{aligned} \mu^n = E[X_n] &= \sum_{j=1}^{\infty} jP\{X_n = j\} \\ &\geq \sum_{j=1}^{\infty} 1 \cdot P\{X_n = j\} \\ &= P\{X_n \geq 1\} \end{aligned}$$

Since $\mu^n \to 0$ when $\mu < 1$, it follows that $P\{X_n \geq 1\} \to 0$, and hence $P\{X_n = 0\} \to 1$.

In fact, it can be shown that $\pi_0 = 1$ even when $\mu = 1$. When $\mu > 1$, it turns out that $\pi_0 < 1$, and an equation determining π_0 may be derived by conditioning on the number of offspring of the initial individual, as follows:

$$\begin{aligned} \pi_0 &= P\{\text{population dies out}\} \\ &= \sum_{j=0}^{\infty} P\{\text{population dies out} | X_1 = j\}P_j \end{aligned}$$

Now, given that $X_1 = j$, the population will eventually die out if and only if each of the j families started by the members of the first generation eventually dies out. Since each family is assumed to act independently, and since the probability that any particular family dies out is just π_0, this yields

$$P\{\text{population dies out} | X_1 = j\} = \pi_0^j$$

and thus π_0 satisfies

$$\pi_0 = \sum_{j=0}^{\infty} \pi_0^j P_j \quad (4.20)$$

In fact when $\mu > 1$, it can be shown that π_0 is the smallest positive number satisfying Equation (4.20).

Example 4.32: If $P_0 = \frac{1}{2}, P_1 = \frac{1}{4}, P_2 = \frac{1}{4}$, then determine π_0. **Solution:** Since $\mu = \frac{3}{4} \leq 1$, it follows that $\pi_0 = 1$.

Example 4.33: If $P_0 = \frac{1}{4}, P_1 = \frac{1}{4}, P_2 = \frac{1}{2}$, then determine π_0. **Solution:** π_0 satisfies $\pi_0 = \frac{1}{4} + \frac{1}{4}\pi_0 + \frac{1}{2}\pi_0^2$ or $2\pi_0^2 - 3\pi_0 + 1 = 0$. The smallest positive solution of this quadratic equation is $\pi_0 = \frac{1}{2}$.

Example 4.34: In Examples 4.32 and 4.33, what is the probability that the population will die out if it initially consists of n individuals?

Solution: Since the population will die out if and only if the families of each of the members of the initial generation die out, the desired probability is π_0^n. For Example 4.32 this yields $\pi_0^n = 1$, and for Example 4.33, $\pi_0^n = (\frac{1}{2})^n$.

4.8 Time Reversible Markov Chains: Consider a stationary ergodic Markov chain (that is, an ergodic Markov chain that has been in operation for a long time) having transition probabilities P_{ij} and stationary probabilities π_i, and suppose that starting at some time we trace the sequence of states going backward in time. That is,

starting at time n, consider the sequence of states Xn,Xn−1,Xn−2, ... It turns out that this sequence of states is itself a Markov chain with transition probabilities Q_{ij}

$$\begin{aligned} Q_{ij} &= P\{X_m = j | X_{m+1} = i\} \\ &= \frac{P\{X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\}} \\ &= \frac{P\{X_m = j\}P\{X_{m+1} = i | X_m = j\}}{P\{X_{m+1} = i\}} \\ &= \frac{\pi_j P_{ji}}{\pi_i} \end{aligned}$$

defined by To prove that the reversed process is indeed a Markov chain, we must verify that

$$P\{X_m = j | X_{m+1} = i, X_{m+2}, X_{m+3}, \ldots\} = P\{X_m = j | X_{m+1} = i\}$$

To see that this is so, suppose that the present time is m+1. Now, since X0,X1,X2, . . . is a Markov chain, it follows that the conditional distribution of the future Xm+2,Xm+3, ….. given the present state Xm+1 is independent of the past state Xm. However, independence is a symmetric relationship (that is, if A is independent of B, then B is independent of A), and so this means that given Xm+1, Xm is independent of Xm+2,Xm+3, ….. But this is exactly what we had to verify. Thus, the reversed process is also a Markov chain with transition probabilities given by $Q_{ij} = \frac{\pi_j P_{ji}}{\pi_i}$ If Qij = Pij for all i, j, then the Markov chain is said to be time reversible. The condition for time reversibility, namely, Qij = Pij, can also be expressed as

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \text{for all } i, j$$

(4.21) The condition in Equation (4.21) can be stated that, for all states i and j, the rate at which the process goes from i to j (namely, $\pi_i P_{ij}$) is equal to the rate at which it goes from j to i (namely, $\pi_j P_{ji}$). It is worth noting that this is an obvious necessary condition for time reversibility since a transition from i to j going backward in time is equivalent to a transition from j to i going forward in time; that is, if Xm = i and Xm−1 = j, then a transition from i to j is observed if we are looking backward, and one from j to i if we are looking forward in time. Thus, the rate at which the forward process makes a transition from j to i is always equal to the rate at which the reverse process makes a transition from i to j; if time reversible, this must equal the rate at which the forward process makes a transition from i to j. If we can find nonnegative numbers, summing to one, that satisfy Equation (4.21), then it follows that the Markov chain is time reversible and the numbers represent the limiting probabilities. This is so since if

$$x_i P_{ij} = x_j P_{ji} \quad \text{for all } i, j, \sum_i x_i = 1$$

(4.22) then summing over i yields

$$\sum_i x_i P_{ij} = x_j \sum_i P_{ji} = x_j, \quad \sum_i x_i = 1$$

and, because the limiting probabilities π_i are the unique solution of the preceding, it follows that $x_i = \pi_i$ for all i.

Example 4.35: Consider a random walk with states 0, 1, . . . ,M and transition probabilities

$$\begin{aligned} P_{i,i+1} &= \alpha_i = 1 - P_{i,i-1}, \quad i = 1, \ldots, M-1, \\ P_{0,1} &= \alpha_0 = 1 - P_{0,0}, \\ P_{M,M} &= \alpha_M = 1 - P_{M,M-1} \end{aligned}$$

Without the need for any computations, it is possible to argue that this Markov chain, which can only make transitions from a state to one of its two nearest neighbors, is time reversible. This follows by noting that the number of transitions from i to i + 1 must at all times be within 1 of the number from i + 1 to i. This is so because between any two transitions from i to i + 1 there must be one from i + 1 to i (and conversely) since the only way to re-enter i from a higher state is via state i + 1. Hence, it follows that the rate of transitions from i to i + 1 equals the rate from i + 1 to i, and so the process is time reversible. We can easily obtain the limiting probabilities by equating for each state i = 0, 1, . . . ,M−1 the rate at which the process goes from i to i + 1 with the rate at which it goes from i + 1 to i. This yields

$$\begin{aligned} \pi_0\alpha_0 &= \pi_1(1-\alpha_1), \\ \pi_1\alpha_1 &= \pi_2(1-\alpha_2), \\ &\vdots \\ \pi_i\alpha_i &= \pi_{i+1}(1-\alpha_{i+1}), \quad i = 0, 1, \ldots, M-1 \end{aligned}$$

Solving in terms of π_0 yields

$$\begin{aligned} \pi_1 &= \frac{\alpha_0}{1-\alpha_1}\pi_0, \\ \pi_2 &= \frac{\alpha_1}{1-\alpha_2}\pi_1 = \frac{\alpha_1\alpha_0}{(1-\alpha_2)(1-\alpha_1)}\pi_0 \end{aligned}$$

and, in general,

$$\pi_i = \frac{\alpha_{i-1}\cdots\alpha_0}{(1-\alpha_i)\cdots(1-\alpha_1)}\pi_0, \quad i = 1, 2, \ldots, M$$

. Since $\sum_0^M \pi_i = 1$, we obtain

$$\pi_0\left[1 + \sum_{j=1}^{M} \frac{\alpha_{j-1}\cdots\alpha_0}{(1-\alpha_j)\cdots(1-\alpha_1)}\right] = 1$$

or

$$\pi_0 = \left[1 + \sum_{j=1}^{M} \frac{\alpha_{j-1}\cdots\alpha_0}{(1-\alpha_j)\cdots(1-\alpha_1)}\right]^{-1}$$

(4.23) and

$$\pi_i = \frac{\alpha_{i-1}\cdots\alpha_0}{(1-\alpha_i)\cdots(1-\alpha_1)}\pi_0, \quad i = 1, \ldots, M$$

(4.24) For instance, if $\alpha_i \equiv \alpha$, then

$$\pi_0 = \left[1 + \sum_{j=1}^{M}\left(\frac{\alpha}{1-\alpha}\right)^j\right]^{-1} = \frac{1-\beta}{1-\beta^{M+1}}$$

and, in general,

$$\pi_i = \frac{\beta^i(1-\beta)}{1-\beta^{M+1}}, \quad i = 0, 1, \ldots, M \qquad \beta = \frac{\alpha}{1-\alpha}$$

where

Another special case of Example 4.35: is the following urn model, proposed by the physicists P. and T. Ehrenfest to describe the movements of molecules. Suppose thatMmolecules are distributed among two urns; and at each time point one of the molecules is chosen at random, removed from its urn, and placed in the other one. The number of molecules in urn I is a special case of the Markov chain of Example 4.35 having $\alpha_i = \frac{M-i}{M}, \quad i = 0, 1, \ldots, M$ Hence, using Equations (4.23) and (4.24) the limiting probabilities in this case are

$$\pi_0 = \left[1 + \sum_{j=1}^{M} \frac{(M-j+1)\cdots(M-1)M}{j(j-1)\cdots 1}\right]^{-1} = \left[\sum_{j=0}^{M}\binom{M}{j}\right]^{-1} = \left(\frac{1}{2}\right)^M$$

where we have used the identity

$$1 = \left(\frac{1}{2} + \frac{1}{2}\right)^M = \sum_{j=0}^{M}\binom{M}{j}\left(\frac{1}{2}\right)^M$$

Hence, from Equation (4.24)

$$\pi_i = \binom{M}{i}\left(\frac{1}{2}\right)^M, \quad i = 0, 1, \ldots, M$$

Because the preceding are just the binomial probabilities, it follows that in the long run, the positions of each of the M balls are independent and each one is equally likely to be in either urn. This, however, is quite intuitive, for if we focus on any one ball, it becomes quite clear that its position will be independent of the positions of the other balls (since no matter where the other M− 1 balls are, the ball under consideration at each stage will be moved with probability 1/M) and by symmetry, it is equally likely to be in either urn.

Example 4.36: Consider an arbitrary connected graph (see Section 3.6 for definitions) having a number w_{ij} associated with arc (i, j) for each arc. One instance of such a graph is given by Figure 4.1. Now consider a particle moving from node to node in this manner: If at any time the particle resides at node i, then it will next move to node j with probability P_{ij} where $P_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$ and where w_{ij} is 0 if (i, j) is not an arc. For instance, for the graph of Figure 4.1, P_{12} = 3/(3 + 1 + 2) = 1/2.

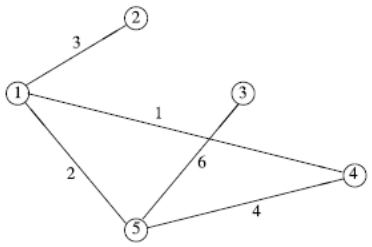


Figure 4.1 A connected graph with arc weights.

The time reversibility equations $\pi_i P_{ij} = \pi_j P_{ji}$ reduce to $\pi_i \frac{w_{ij}}{\sum_j w_{ij}} = \pi_j \frac{w_{ji}}{\sum_i w_{ji}}$ or, equivalently, since $w_{ij} = w_{ji}$, $\frac{\pi_i}{\sum_j w_{ij}} = \frac{\pi_j}{\sum_i w_{ji}}$ which is equivalent to $\frac{\pi_i}{\sum_j w_{ij}} = c$ or $\pi_i = c \sum_j w_{ij}$ or, since $1 = \sum_i \pi_i$, $\pi_i = \frac{\sum_j w_{ij}}{\sum_i \sum_j w_{ij}}$. Because the π_is given by this equation satisfy the time reversibility equations, it follows that the process is time reversible with these limiting probabilities.

For the graph of Figure 4.1 we have that $\pi_1 = \frac{6}{32}, \quad \pi_2 = \frac{3}{32}, \quad \pi_3 = \frac{6}{32}, \quad \pi_4 = \frac{5}{32}, \quad \pi_5 = \frac{12}{32}$

If we try to solve Equation (4.22) for an arbitrary Markov chain with states 0, 1, . . . ,M, it will usually turn out that no solution exists. For example, from Equation (4.22), $x_i P_{ij} = x_j P_{ji}$, $x_k P_{kj} = x_j P_{jk}$ implying $(\text{if } P_{ij} P_{jk} > 0)$ that $\frac{x_i}{x_k} = \frac{P_{ji} P_{kj}}{P_{ij} P_{jk}}$ which in general need not equal P_{ki}/P_{ik}. Thus, we see that a necessary condition for time reversibility is that

$$P_{ik} P_{kj} P_{ji} = P_{ij} P_{jk} P_{ki} \quad \text{for all } i, j, k \tag{4.25}$$

which is equivalent to the statement that, starting in state i, the path i → k →j → i has the same probability as the reversed path i → j → k → i. To understand the necessity of this, note that time reversibility implies that the rate at which a sequence of transitions from i to k to j to i occurs must equal the rate of ones from i to j to k to i (why?), and so we must have $\pi_i P_{ik} P_{kj} P_{ji} = \pi_i P_{ij} P_{jk} P_{ki}$ implying Equation (4.25) when $\pi_i > 0$. In fact, we can show the following.

Theorem 4.2 An ergodic Markov chain for which $P_{ij} = 0$ whenever $P_{ji} = 0$ time reversible if and only if starting in state i, any path back to i has the same probability as the reversed path. That is, if

$$P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,i} = P_{i,i_k} P_{i_k,i_{k-1}} \cdots P_{i_1,i} \tag{4.26}$$

for all states $i, i_1, \ldots, i_k$.

Proof. We have already proven necessity. To prove sufficiency, fix states i and j and rewrite (4.26) as

$$P_{i,i_1} P_{i_1,i_2} \cdots P_{i_k,j} P_{ji} = P_{ij} P_{j,i_k} \cdots P_{i_1,i}$$

Summing the preceding over all states $i_1, \ldots, i_k$ yields $P_{ij}^{k+1} P_{ji} = P_{ij} P_{ji}^{k+1}$ Letting $k \to \infty$ yields $\pi_j P_{ji} = P_{ij} \pi_i$ which proves the theorem.

Example 4.37: Suppose we are given a set of n elements, numbered 1 through n, which are to be arranged in some ordered list. At each unit of time a request is made to retrieve one of these elements, element i being requested (independently of the past) with probability P_i. After being requested, the element then is put back but not necessarily in the same position. In fact, let us suppose that the element requested is moved one closer to the front of the list; for instance, if the present list ordering is 1, 3, 4, 2, 5 and element 2 is requested, then the new ordering becomes 1, 3, 2, 4, 5. We are interested in the long-run average position of the element requested.

For any given probability vector $P = (P_1, \ldots, P_n)$, the preceding can be modelled as a Markov chain with $n!$ states, with the state at any time being the list order at that time. We shall show that this Markov chain is time reversible and then use this to show that the average position of the element requested when this one-closer rule is in effect is less than when the rule of always moving the requested element to the front of the line is used. The time reversibility of the resulting Markov chain when the one-closer reordering rule is in effect easily follows from Theorem 4.2. For instance, suppose n = 3 and consider the following path from state (1, 2, 3) to itself:

$(1,2,3) \to (2,1,3) \to (2,3,1) \to (3,2,1) \to (3,1,2) \to (1,3,2) \to (1,2,3)$. The product of the transition probabilities in the forward direction is $P_2 P_3 P_3 P_1 P_1 P_2 = P_1^2 P_2^2 P_3^2$ whereas in the reverse direction, it is $P_3 P_3 P_2 P_2 P_1 P_1 = P_1^2 P_2^2 P_3^2$ Because the general result follows in much the same manner, the Markov chain is indeed time reversible. (For a formal argument note that if f_i denotes the number of times element i moves forward in the path, then as the path goes from a fixed state back to itself, it follows that element i will also move backward f_i times. Therefore, since the backward moves of element i are precisely the times that it moves forward in the reverse path, it follows that the product of the transition probabilities for both the path and its reversal will equal $\prod_i P_i^{f_i + r_i}$, where r_i is equal to the number of times that element i is in the first position and the path (or the reverse path) does not change states.) For any permutation $i_1, i_2, \ldots, i_n$ of 1, 2, . . . , n, let $\pi(i_1, i_2, \ldots, i_n)$ denote the limiting probability under the one-closer rule. By time reversibility we have

$$P_{i_{j+1}} \pi(i_1, \ldots, i_j, i_{j+1}, \ldots, i_n) = P_{i_j} \pi(i_1, \ldots, i_{j+1}, i_j, \ldots, i_n) \tag{4.27}$$

for all permutations. Now the average position of the element requested can be expressed (as in Section 3.6.1) as

$$\text{Average position} = \sum_i P_i E[\text{Position of element } i] = \sum_i P_i \left[1 + \sum_{j \neq i} P\{\text{element } j \text{ precedes element } i\} \right] = 1 + \sum_i \sum_{j \neq i} P_i P\{e_j \text{ precedes } e_i\}$$

$$= 1 + \sum_{i<j} [P_i P\{e_j \text{ precedes } e_i\} + P_j P\{e_i \text{ precedes } e_j\}] = 1 + \sum_{i<j} [P_i P\{e_j \text{ precedes } e_i\} + P_j (1 - P\{e_j \text{ precedes } e_i\})]$$

$$= 1 + \sum_{i<j} \sum (P_i - P_j) P\{e_j \text{ precedes } e_i\} + \sum_{i<j} \sum P_j$$

Hence, to minimize the average position of the element requested, we would want to make $P\{e_j \text{ precedes } e_i\}$ as large as possible when $P_j > P_i$ and as small as possible when $P_i > P_j$. Under the front-of-the-line rule we showed in Section 3.6.1,

$$P\{e_j \text{ precedes } e_i\} = \frac{P_j}{P_j + P_i}$$

(since under the front-of-the-line rule element j will precede element i if and only if the last request for either i or j was for j). Therefore, to show that the one-closer rule is better than the front-of-the-line rule, it suffices to show that under the one-closer rule

$$P\{e_j \text{ precedes } e_i\} > \frac{P_j}{P_j + P_i} \quad \text{when } P_j > P_i$$

, Now consider any state where element i precedes element j, say, $(\ldots, i, i_1, \ \ldots, i_k, j, \ldots)$. By successive transpositions using Equation (4.27), we have

$$\pi(\ldots, i, i_1, \ldots, i_k, j, \ldots) = \left(\frac{P_i}{P_j} \right)^{k+1} \pi(\ldots, j, i_1, \ldots, i_k, i, \ldots) \tag{4.28}$$

, For instance,

$\pi(1,2,3) = \frac{P_2}{P_3}\pi(1,3,2) = \frac{P_2}{P_3}\frac{P_1}{P_3}\pi(3,1,2) = \frac{P_2}{P_3}\frac{P_1}{P_3}\frac{P_1}{P_2}\pi(3,2,1) = \left(\frac{P_1}{P_3}\right)^2 \pi(3,2,1)$, Now when $P_j > P_i$, Equation (4.28) implies that

$\pi(\ldots,i,i_1,\ldots,i_k,j,\ldots) < \frac{P_i}{P_j}\pi(\ldots,j,i_1,\ldots,i_k,i,\ldots)$ Letting $\alpha(i,j) = P\{e_i \text{ precedes } e_j\}$, we see by summing over all states for which I precedes j and by using the preceding that $\alpha(i,j) < \frac{P_i}{P_j}\alpha(j,i)$ which, since $\alpha(i,j) = 1 - \alpha(j,i)$, yields $\alpha(j,i) > \frac{P_j}{P_j + P_i}$, Hence, the average position of the element requested is indeed smaller under the one-closer rule than under the front-of-the-line rule. The concept of the reversed chain is useful even when the process is not time reversible. To illustrate this, we start with the following proposition whose proof is left as an exercise.

Proposition 4.6: Consider an irreducible Markov chain with transition probabilities P_{ij}. If we can find positive numbers $\pi_i, i \geqslant 0$, summing to one, and a transition probability matrix $\mathbf{Q} = [Q_{ij}]$ such that $\pi_i P_{ij} = \pi_j Q_{ji}$ (4.29) then the Q_{ij} are the transition probabilities of the reversed chain and the π_i are the stationary probabilities both for the original and reversed chain. The importance of the preceding proposition is that, by thinking backward, we can sometimes guess at the nature of the reversed chain and then use the set of Equations (4.29) to obtain both the stationary probabilities and the Q_{ij}.

Example 4.38: A single bulb is necessary to light a given room. When the bulb in use fails, it is replaced by a new one at the beginning of the next day. Let X_n equal i if the bulb in use at the beginning of day n is in its ith day of use (that is, if its present age is i). For instance, if a bulb fails on day n−1, then a new bulb will be put in use at the beginning of day n and so $X_n = 1$. If we suppose that each bulb, independently, fails on its ith day of use with probability $p_i, i \geqslant 1$, then it is easy to see that $\{X_n, n \geqslant 1\}$ is a Markov chain whose transition probabilities are as follows: $P_{i,1} = P\{\text{bulb, on its } i\text{th day of use, fails}\} = P\{\text{life of bulb} = i | \text{life of bulb} \geqslant i\} = \frac{P\{L = i\}}{P\{L \geqslant i\}}$, where L, a random variable representing the lifetime of a bulb, is such that P{L = i} = p_i. Also, $P_{i,i+1} = 1 - P_{i,1}$. Suppose now that this chain has been in operation for a long (in theory, an infinite) time and consider the sequence of states going backward in time. Since, in the forward direction, the state is always increasing by 1 until it reaches the age at which the item fails, it is easy to see that the reverse chain will always decrease by 1 until it reaches 1 and then it will jump to a random value representing the lifetime of the (in real time) previous bulb. Thus, it seems that the reverse chain should have transition probabilities given by

$$Q_{i,i-1} = 1, \quad i > 1$$
$$Q_{1,i} = p_i, \quad i \geqslant 1$$

To check this, and at the same time determine the stationary probabilities, we must see if we can find, with the $Q_{i,j}$ as previously given, positive numbers $\{\pi_i\}$ such that $\pi_i P_{i,j} = \pi_j Q_{j,i}$ To begin, let j = 1 and consider the resulting equations: $\pi_i P_{i,1} = \pi_1 Q_{1,i}$ This is equivalent to $\pi_i \frac{P\{L = i\}}{P\{L \geqslant i\}} = \pi_1 P\{L = i\}$ or $\pi_i = \pi_1 P\{L \geqslant i\}$ Summing over all i yields $1 = \sum_{i=1}^{\infty} \pi_i = \pi_1 \sum_{i=1}^{\infty} P\{L \geqslant i\} = \pi_1 E[L]$ and so, for the preceding Q_{ij} to represent the reverse transition probabilities, it is necessary for the stationary probabilities to be $\pi_i = \frac{P\{L \geqslant i\}}{E[L]}, \quad i \geqslant 1$ To finish the proof that the reverse transition probabilities and stationary probabilities are as given, all that remains is to show that they satisfy $\pi_i P_{i,i+1} = \pi_{i+1} Q_{i+1,i}$ which is equivalent to $\frac{P\{L \geqslant i\}}{E[L]}\left(1 - \frac{P\{L = i\}}{P\{L \geqslant i\}}\right) = \frac{P\{L \geqslant i+1\}}{E[L]}$ and which is true since $P\{L \geqslant i\} - P\{L = i\} = P\{L \geqslant i+1\}$.

4.9 Markov Chain Monte Carlo Methods: Let X be a discrete random vector whose set of possible values is $\mathbf{x}_j, j \geqslant 1$. Let the probability mass function of X be given by $P\{\mathbf{X} = \mathbf{x}_j\}, j \geqslant 1$, and suppose that we are interested in calculating $\theta = E[h(\mathbf{X})] = \sum_{j=1}^{\infty} h(\mathbf{x}_j) P\{\mathbf{X} = \mathbf{x}_j\}$ for some specified function h. In situations where it is computationally difficult to evaluate the function $h(\mathbf{x}_j), j \geqslant 1$, we often turn to simulation to approximate θ. The usual approach, called Monte Carlo simulation, is to use random numbers to generate a partial sequence of independent and identically distributed random vectors X1,X2, . . . ,Xn having the mass function $P\{\mathbf{X} = \mathbf{x}_j\}, j \geqslant 1$ (see Chapter 11 for a discussion as to how this can be accomplished). Since the strong law of large numbers yields $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{h(\mathbf{X}_i)}{n} = \theta$ (4.30) it follows that we can estimate θ by letting n be large and using the average of the values of $h(\mathbf{X}_i)$, i = 1, . . . , n as the estimator. It often, however, turns out that it is difficult to generate a random vector having the specified probability mass function, particularly if X is a vector of dependent random variables. In addition, its probability mass function is sometimes given in the form $P\{\mathbf{X} = \mathbf{x}_j\} = Cb_j, j \geqslant 1$, where the bj are specified, but C must be computed, and in many applications it is not computationally feasible to sum the bj so as to determine C. Fortunately, however, there is another way of using simulation to estimate θ in these situations. It works by generating a sequence, not of independent random vectors, but of the successive states of a vector-valued Markov chain X1,X2, . . . whose stationary probabilities are $P\{\mathbf{X} = \mathbf{x}_j\}, j \geqslant 1$. If this can be accomplished, then it would follow from Proposition 4.4 that Equation (4.30) remains valid, implying that we can then use $\sum_{i=1}^{n} h(\mathbf{X}_i)/n$ as an estimator of θ. We now show how to generate a Markov chain with arbitrary stationary probabilities that may only be specified up to a multiplicative constant. Let b(j), j = 1, 2, . . . be positive numbers whose sum $B = \sum_{j=1}^{\infty} b(j)$ is finite. The following, known as the Hastings–Metropolis algorithm, can be used to generate a time reversible Markov chain whose stationary probabilities are $\pi(j) = b(j)/B, \quad j = 1, 2, \ldots$ To begin, let Q be any specified irreducible Markov transition probability matrix on the integers, with q(i, j) representing the row i column j element of Q. Now define a Markov chain $\{X_n, n \geqslant 0\}$ as follows. When $X_n = i$, generate a random variable Y such that P{Y = j} = q(i, j), j = 1, 2, ... If Y = j, then set X_{n+1} equal to j with probability α(i, j), and set it equal to i with probability $1 - \alpha(i, j)$. Under these conditions, it is easy to see that the sequence of states constitutes a Markov chain with transition probabilities $P_{i,j}$ given by

$$P_{i,j} = q(i,j)\alpha(i,j), \quad \text{if } j \neq i$$
$$P_{i,i} = q(i,i) + \sum_{k \neq i} q(i,k)(1 - \alpha(i,k))$$

This Markov chain will be time reversible and have stationary probabilities π(j) if $\pi(i)P_{i,j} = \pi(j)P_{j,i} \quad \text{for } j \neq i$ which is

But if we take $\pi_j = b(j)/B$ and set

$$\alpha(i,j) = \min\left(\frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}, 1\right)$$

equivalent to $\pi(i)q(i,j)\alpha(i,j) = \pi(j)q(j,i)\alpha(j,i)$ (4.31) (4.32) then Equation (4.31) is easily seen to be satisfied. For if

$\alpha(i,j) = \dfrac{\pi(j)q(j,i)}{\pi(i)q(i,j)}$ then $\alpha(j,i) = 1$ and Equation (4.31) follows, and if $\alpha(i,j) = 1$ then $\alpha(j,i) = \dfrac{\pi(i)q(i,j)}{\pi(j)q(j,i)}$ and again Equation (4.31) holds, thus showing that the Markov chain is time reversible with stationary probabilities π(j). Also, since $\pi(j) = b(j)/B$, we see from (4.32) that

$$\alpha(i,j) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right)$$

which shows that the value of B is not needed to define the Markov chain, because the values b(j) suffice. Also, it is almost always the case that $\pi(j), j \geq 1$ will not only be stationary probabilities but will also be limiting probabilities. (Indeed, a sufficient condition is that $P_{i,i} > 0$ for some i.)

Example 4.39: Suppose that we want to generate a uniformly distributed element in $\mathcal{S}$, the set of all permutations $(x_1, \ldots, x_n)$ of the numbers (1, . . . , n) for which $\sum_{j=1}^{n} jx_j > a$ for a given constant a. To utilize the Hastings–Metropolis algorithm we need to define an irreducible Markov transition probability matrix on the state space $\mathcal{S}$. To accomplish this, we first define a concept of "neighboring" elements of $\mathcal{S}$, and then construct a graph whose vertex set is $\mathcal{S}$. We start by putting an arc between each pair of neighboring elements in $\mathcal{S}$, where any two permutations in $\mathcal{S}$ are said to be neighbors if one results from an interchange of two of the positions of the other. That is, (1, 2, 3, 4) and (1, 2, 4, 3) are neighbors whereas (1, 2, 3, 4) and (1, 3, 4, 2) are not. Now, define the q transition probability function as follows. With N(s) defined as the set of neighbors of s, and $|N(s)|$ equal to the number of elements in the set N(s), let $q(s,t) = \dfrac{1}{|N(s)|} \quad \text{if } t \in N(s)$ That is, the candidate next state from s is equally likely to be any of its neighbors. Since the desired limiting probabilities of the Markov chain are π(s) = C, it follows that π(s) = π(t), and so $\alpha(s,t) = \min(|N(s)|/|N(t)|, 1)$ That is, if the present state of the Markov chain is s then one of its neighbors is randomly chosen, say, t. If t is a state with fewer neighbors than s (in graph theory language, if the degree of vertex t is less than that of vertex s), then the next state is t. If not, a uniform (0,1) random number U is generated and the next state is t if $U < |(N(s)|/|N(t)|$ and is s otherwise. The limiting probabilities of this Markov chain are $\pi(s) = 1/|\mathcal{S}|$, where $|\mathcal{S}|$ is the (unknown) number of permutations in $\mathcal{S}$.

The most widely used version of the Hastings–Metropolis algorithm is the Gibbs sampler. Let X = (X1, . . . ,Xn) be a discrete random vector with probability mass function p(x) that is only specified up to a multiplicative constant, and suppose that we want to generate a random vector whose distribution is that of X. That is, we want to generate a random vector having mass function $p(x) = Cg(x)$ where g(x) is known, but C is not. Utilization of the Gibbs sampler assumes that for any i and values x_j, $j \neq i$, we can generate a random variable X having the probability mass function $P\{X = x\} = P\{X_i = x | X_j = x_j, j \neq i\}$ It operates by using the Hasting–Metropolis algorithm on a Markov chain with states x=(x1, . . . , xn), and with transition probabilities defined as follows. Whenever the present state is x, a coordinate that is equally likely to be any of 1, . . . , n is chosen. If coordinate i is chosen, then a random variable X with probability mass function $P\{X = x\} = P\{X_i = x | X_j = x_j, j \neq i\}$ is generated. If X =x, then the state $y = (x_1, \ldots x_{i-1}, x, x_{i+1}, \ldots, x_n)$ is considered as the candidate next state. In other words, with x and y as given, the Gibbs sampler uses the Hastings–Metropolis algorithm with $q(x,y) = \dfrac{1}{n} P\{X_i = x | X_j = x_j, j \neq i\} = \dfrac{p(y)}{nP\{X_j = x_j, j \neq i\}}$ Because we want the limiting mass function to be p, we see from Equation (4.32) that the vector y is then accepted as the new state with probability

$$\begin{aligned}\alpha(x,y) &= \min\left(\frac{p(y)q(y,x)}{p(x)q(x,y)}, 1\right)\\ &= \min\left(\frac{p(y)p(x)}{p(x)p(y)}, 1\right)\\ &= 1\end{aligned}$$

Hence, when utilizing the Gibbs sampler, the candidate state is always accepted as the next state of the chain.

Example 4.40 Suppose that we want to generate n uniformly distributed points in the circle of radius 1 centered at the origin, conditional on the event that no two points are within a distance d of each other, when the probability of this conditioning event is small. This can be accomplished by using the Gibbs sampler as follows. Start with any n points $x_1, \ldots, x_n$ in the circle that have the property that no two of them are within d of the other; then generate the value of I, equally likely to be any of the values 1, . . . , n. Then continually generate a random point in the circle until you obtain one that is not within d of any of the other n – 1 points excluding x_I. At this point, replace x_I by the generated point and then repeat the operation. After a large number of iterations of this algorithm, the set of n points will approximately have the desired distribution.

Example 4.41: Let $X_i, i = 1, \ldots, n,$ be independent exponential random variables with respective rates $\lambda_i, i = 1, \ldots, n.$ Let $S = \sum_{i=1}^{n} X_i,$ and suppose that we want to generate the random vector $X = (X_1, \ldots, X_n),$ conditional on the event that S > c for some large positive constant c. That is, we want to generate the value of a random vector whose density function is

$$f(x_1, \ldots, x_n) = \frac{1}{P\{S > c\}} \prod_{i=1}^{n} \lambda_i e^{-\lambda_i x_i}, \quad x_i \geq 0, \sum_{i=1}^{n} x_i > c$$

This is easily accomplished by starting with an initial vector $x = (x_1, \ldots, x_n)$ satisfying $x_i > 0, i = 1, \ldots, n, \sum_{i=1}^{n} x_i > c.$ Then generate a random variable I that is equally likely to be any of 1, . . . , n. Next, generate an exponential random variable X with rate λ_I conditional on the event that $X + \sum_{j \neq I} x_j > c.$ This latter step, which calls for generating the value of an exponential random variable given that it exceeds $c - \sum_{j \neq I} x_j,$ is easily accomplished by using the fact that an exponential conditioned to be greater than a positive constant is distributed as the constant plus the exponential. Consequently, to obtain X, first generate an exponential random variable Y with rate λ_I, and then set $X = Y + \left(c - \sum_{j \neq I} x_j\right)^+$, where $a^+ = \max(a, 0).$ The value of x_I should then be reset as X and a new iteration of the algorithm begun.

Remark: As can be seen by Examples 4.40 and 4.41, although the theory for the Gibbs sampler was represented under the assumption that the distribution to be generated was discrete, it also holds when this distribution is continuous.

4.10 Markov Decision Processes: Consider a process that is observed at discrete time points to be in any one of M possible states, which we number by 1, 2, . . . ,M. After observing the state of the process, an action must be chosen, and we let A, assumed finite, denote the set of all possible actions. If the process is in state i at time n and action a is chosen, then the next state of the system is determined according to the transition probabilities $P_{ij}(a)$. If we let Xn denote the state of the process at time n and an the action chosen at time n, then the preceding is equivalent to stating that $P\{X_{n+1} = j | X_0, a_0, X_1, a_1, \ldots, X_n = i, a_n = a\} = P_{ij}(a)$ Thus, the transition probabilities are functions only of the present state and the subsequent action. By a policy, we mean a rule for choosing actions. We shall restrict ourselves to policies that are of the form that the action they prescribe at any time depends only on the state of the process at that time (and not on any information concerning prior states and actions). However, we shall allow the policy to be "randomized" in that its instructions may be to choose actions according to a probability distribution. In other words, a policy β is a set of numbers β = {βi(a), a ∈ A, i = 1, . . . ,M} with the interpretation that if the process is in state i, then action a is to be chosen with probability

$$0 \leqslant \beta_i(a) \leqslant 1, \quad \text{for all } i, a$$
$$\sum_a \beta_i(a) = 1, \quad \text{for all } i$$

βi(a). Of course, we need have Under any given policy β, the sequence of states {Xn, n = 0, 1, . . .} constitutes a Markov chain with

$$P_{ij}(\boldsymbol{\beta}) = P_{\boldsymbol{\beta}}\{X_{n+1} = j | X_n = i\}^* = \sum_a P_{ij}(a)\beta_i(a)$$

transition probabilities $P_{ij}(\boldsymbol{\beta})$ given by ∗ We use the notation $P_{\boldsymbol{\beta}}$ to signify that the probability is conditional on the fact that policy β is used. where the last equality follows by conditioning on the action chosen when in state i. Let us suppose that for every choice of a policy β, the resultant Markov chain {Xn, n = 0, 1, . . .} is ergodic. For any policy β, let π_{ia} denote the limiting (or steady-state) probability that the process will be in state i and action a will be chosen if

$$\pi_{ia} = \lim_{n\to\infty} P_{\boldsymbol{\beta}}\{X_n = i, a_n = a\}$$

policy β is employed. That is,

The vector $\boldsymbol{\pi} = (\pi_{ia})$ must satisfy
(i) $\pi_{ia} \geqslant 0$ for all i, a,
(ii) $\sum_i \sum_a \pi_{ia} = 1$,
(iii) $\sum_a \pi_{ja} = \sum_i \sum_a \pi_{ia} P_{ij}(a)$ for all j

(4.33) Equations (i) and (ii) are obvious, and Equation (iii), which is an analogue of Equation (4.7), follows as the left-hand side equals the steady-state probability of being in state j and the right-hand side is the same probability computed by conditioning on the state and action chosen one stage earlier. Thus for any policy β, there is a vector $\boldsymbol{\pi} = (\pi_{ia})$ that satisfies (i)–(iii) and with the interpretation that π_{ia} is equal to the steady-state probability of being in state I and choosing action a when policy β is employed. Moreover, it turns out that the reverse is also true. Namely, for any vector $\boldsymbol{\pi} = (\pi_{ia})$ that satisfies (i)–(iii), there exists a policy β such that if β is used, then the steady-state probability of being in i and choosing action a equals π_{ia}. To verify this last statement, suppose that $\boldsymbol{\pi} = (\pi_{ia})$ is a vector that satisfies (i)–(iii). Then, let the policy β = (βi(a)) be

$$\beta_i(a) = P\{\boldsymbol{\beta} \text{ chooses } a | \text{state is } i\} = \frac{\pi_{ia}}{\sum_a \pi_{ia}}$$

Now let P_{ia} denote the limiting probability of being in i and choosing a when policy β is employed. We need to show that $P_{ia} = \pi_{ia}$. To do so, first note that {Pia, i = 1, . . . ,M, a ∈ A} are the limiting probabilities of the two-dimensional Markov chain $\{(X_n, a_n), n \geqslant 0\}$. Hence, by the fundamental Theorem 4.1, they are the unique solution of

(i′) $P_{ia} \geqslant 0$,
(ii′) $\sum_i \sum_a P_{ia} = 1$,
(iii′) $P_{ja} = \sum_i \sum_{a'} P_{ia'} P_{ij}(a')\beta_j(a)$

where (iii′) follows since

$$P\{X_{n+1} = j, a_{n+1} = a | X_n = i, a_n = a'\} = P_{ij}(a')\beta_j(a)$$

Because

$$\beta_j(a) = \frac{\pi_{ja}}{\sum_a \pi_{ja}}$$

we see that (Pia) is the unique solution of

$$P_{ia} \geqslant 0,$$
$$\sum_i \sum_a P_{ia} = 1,$$
$$P_{ja} = \sum_i \sum_{a'} P_{ia'} P_{ij}(a') \frac{\pi_{ja}}{\sum_a \pi_{ja}}$$

Hence, to show that $P_{ia} = \pi_{ia}$, we need show that

$$\pi_{ia} \geqslant 0,$$
$$\sum_i \sum_a \pi_{ia} = 1,$$
$$\pi_{ja} = \sum_i \sum_{a'} \pi_{ia'} P_{ij}(a') \frac{\pi_{ja}}{\sum_a \pi_{ja}}$$

The top two equations follow from (i) and (ii) of Equation (4.33), and the third, which is equivalent to

$$\sum_a \pi_{ja} = \sum_i \sum_{a'} \pi_{ia'} P_{ij}(a')$$

follows from condition (iii) of Equation (4.33). Thus we have shown that a vector $\boldsymbol{\beta} = (\pi_{ia})$ will satisfy (i), (ii), and (iii) of Equation (4.33) if and only if there exists a policy β such that π_{ia} is equal to the steady-state probability of being in state i and choosing action a when β is used. In fact, the policy β is defined by $\beta_i(a) = \pi_{ia} / \sum_a \pi_{ia}$. The preceding is quite important in the determination of "optimal" policies. For instance, suppose that a reward R(i, a) is earned whenever action a is chosen in state i. Since $R(X_i, a_i)$ would then represent the reward earned at time i, the expected average reward per unit time under policy β can be expressed as

$$\text{expected average reward under } \boldsymbol{\beta} = \lim_{n\to\infty} E_{\boldsymbol{\beta}}\left[\frac{\sum_{i=1}^n R(X_i, a_i)}{n}\right]$$

Now, if π_{ia} denotes the steady-state probability of being in state i and choosing action a, it follows that the limiting expected reward at time n equals

$$\lim_{n\to\infty} E[R(X_n, a_n)] = \sum_i \sum_a \pi_{ia} R(i, a)$$

which implies that

$$\text{expected average reward under } \boldsymbol{\beta} = \sum_i \sum_a \pi_{ia} R(i, a)$$

Hence, the problem of determining the policy that maximizes the expected average reward is

$$\begin{aligned} &\underset{\boldsymbol{\pi}=(\pi_{ia})}{\text{maximize}} \sum_i \sum_a \pi_{ia} R(i, a) \\ &\text{subject to } \pi_{ia} \geqslant 0, \quad \text{for all } i, a, \\ &\sum_i \sum_a \pi_{ia} = 1, \\ &\sum_a \pi_{ja} = \sum_i \sum_a \pi_{ia} P_{ij}(a), \quad \text{for all } j \end{aligned}$$

(4.34) However, the preceding maximization problem is a special case of what is known as a linear program and can be solved by a standard linear programming algorithm known as the simplex algorithm.∗ If $\boldsymbol{\beta}^* = (\pi_{ia}^*)$ maximizes the preceding, then the optimal policy will be given by $\boldsymbol{\beta}^*$ where

$$\beta_i^*(a) = \frac{\pi_{ia}^*}{\sum_a \pi_{ia}^*}$$

∗ It is called a linear program since the objective function $\sum_i \sum_a R(i, a)\pi_{ia}$ and the constraints are all linear functions of the πia. For a heuristic analysis of the simplex algorithm, see 4.5.2.

Remarks:

(i) It can be shown that there is a π^* maximizing Equation (4.34) that has the property that for each i, π_{ia}^* is zero for all but one value of a, which implies that the optimal policy is nonrandomized. That is, the action it prescribes when in state i is a deterministic function of i.

(ii) The linear programming formulation also often works when there are restrictions placed on the class of allowable policies. For instance, suppose there is a restriction on the fraction of time the process spends in some state, say, state 1. Specifically, suppose that we are allowed to consider only policies having the property that their use results in the process being in state 1 less than 100α percent of time. To determine the optimal policy subject to this requirement, we add to the linear programming problem the additional constraint

$$\sum_a \pi_{1a} \leqslant \alpha$$

since $\sum_a \pi_{1a}$ represents the proportion of time that the process is in state 1.

4.11 Hidden Markov Chains: Let {Xn, n = 1, 2, . . .} be a Markov chain with transition probabilities $P_{i,j}$ and initial state probabilities $p_i = P\{X_1 = i\},\ i \geqslant 0$. Suppose that there is a finite set $\mathscr{S}$ of signals, and that a signal from $\mathscr{S}$ is emitted each time the Markov chain enters a state. Further, suppose that when the Markov chain enters state j then, independently of previous Markov chain states and signals, the signal emitted is s with probability $p(s|j),\ \sum_{s\in\mathscr{S}} p(s|j) = 1.$ That is, if S_n

$$P\{S_1 = s|X_1 = j\} = p(s|j),$$

represents the nth signal emitted, then $P\{S_n = s|X_1, S_1, \ldots, X_{n-1}, S_{n-1}, X_n = j\} = p(s|j)$ A model of the preceding type in which the sequence of signals $S_1, S_2, \ldots$ is observed, while the sequence of underlying Markov chain states $X_1, X_2, \ldots$ is unobserved, is called a hidden Markov chain model.

Example 4.42: Consider a production process that in each period is either in a good state (state 1) or in a poor state (state 2). If the process is in state 1 during a period then, independent of the past, with probability 0.9 it will be in state 1 during the next period and with probability 0.1 it will be in state 2. Once in state 2, it remains in that state forever. Suppose that a single item is produced each period and that each item produced when the process is in state 1 is of acceptable quality with probability 0.99, while each item produced when the process is in state 2 is of acceptable quality with probability 0.96. If the status, either acceptable or unacceptable, of each successive item is observed, while the process states are unobservable, then the preceding is a hidden Markov chain model. The signal is the status of the item produced, and has value either a or u, depending on whether the item is acceptable or unacceptable. The signal probabilities are

$$p(u|1) = 0.01, \quad p(a|1) = 0.99,$$
$$p(u|2) = 0.04, \quad p(a|2) = 0.96$$

while the transition probabilities of the underlying Markov chain are $P_{1,1} = 0.9 = 1 - P_{1,2}, \quad P_{2,2} = 1$

Although $\{S_n, n \geqslant 1\}$ is not a Markov chain, it should be noted that, conditional on the current state X_n, the sequence $S_n, X_{n+1}, S_{n+1}, \cdots$ of future signals and states is independent of the sequence $X_1, S_1, \ldots, X_{n-1}, S_{n-1}$ of past states and signals. Let $S^n = (S_1, \ldots, S_n)$ be the random vector of the first n signals. For a fixed sequence of signals $s_1, \ldots, s_n$, let $s_k = (s_1, \ldots, s_k),\ k \leqslant n.$ To begin, let us determine the conditional probability of the Markov chain state at time n given that $S^n = s_n$. To obtain this probability, let $F_n(j) = P\{S^n = s_n, X_n = j\}$ and note that $P\{X_n = j|S^n = s_n\} = \dfrac{P\{S^n = s_n, X_n = j\}}{P\{S^n = s_n\}} = \dfrac{F_n(j)}{\sum_i F_n(i)}$. Now,

$$F_n(j) = P\{S^{n-1} = s_{n-1}, S_n = s_n, X_n = j\} = \sum_i P\{S^{n-1} = s_{n-1}, X_{n-1} = i, X_n = j, S_n = s_n\} = \sum_i F_{n-1}(i) P\{X_n = j, S_n = s_n|S^{n-1} = s_{n-1}, X_{n-1} = i\}$$
$$= \sum_i F_{n-1}(i) P\{X_n = j, S_n = s_n|X_{n-1} = i\} = \sum_i F_{n-1}(i) P_{i,j} p(s_n|j) = p(s_n|j) \sum_i F_{n-1}(i) P_{i,j}$$

(4.35), where the preceding used that,

$$P\{X_n = j, S_n = s_n|X_{n-1} = i\} = P\{X_n = j|X_{n-1} = i\} \times P\{S_n = s_n|X_n = j, X_{n-1} = i\} = P_{i,j} P\{S_n = s_n|X_n = j\} = P_{i,j} p(s_n|j)$$

. Starting with

$$F_1(i) = P\{X_1 = i, S_1 = s_1\} = p_i p(s_1|i)$$

we can use Equation (4.35) to recursively determine the functions $F_2(i),\ F_3(i), \ldots,$ up to $F_n(i).$

Example 4.43: Suppose in Example 4.42 that $P\{X_1 = 1\} = 0.8.$ It is given that the successive conditions of the first three items produced are a, u, a.
(a) What is the probability that the process was in its good state when the third item was produced?
(b) What is the probability that X4 is 1?
(c) What is the probability that the next item produced is acceptable?

Solution: With $s_3 = (a, u, a)$, we have

$$F_1(1) = (0.8)(0.99) = 0.792,$$
$$F_1(2) = (0.2)(0.96) = 0.192$$

$$F_2(1) = 0.01[0.792(0.9) + 0.192(0)] = 0.007128,$$
$$F_2(2) = 0.04[0.792(0.1) + 0.192(1)] = 0.010848$$

$$F_3(1) = 0.99[(0.007128)(0.9)] \approx 0.006351,$$
$$F_3(2) = 0.96[(0.007128)(0.1) + 0.010848] \approx 0.011098$$

Therefore, the answer to part (a) is $P\{X_3 = 1|s_3\} \approx \dfrac{0.006351}{0.006351 + 0.011098} \approx 0.364$ . . To compute $P\{X_4 = 1|s_3\},$ condition on X_3 to obtain

$$P\{X_4 = 1|s_3\} = P\{X_4 = 1|X_3 = 1, s_3\} P\{X_3 = 1|s_3\} + P\{X_4 = 1|X_3 = 2, s_3\} P\{X_3 = 2|s_3\}$$
$$= P\{X_4 = 1|X_3 = 1, s_3\}(0.364) + P\{X_4 = 1|X_3 = 2, s_3\}(0.636) = 0.364 P_{1,1} + 0.636 P_{2,1} = 0.3276$$

. To compute $P\{S_4 = a|s_3\},$ condition on X4 to obtain

$$P\{S_4 = a|s_3\} = P\{S_4 = a|X_4 = 1, s_3\} P\{X_4 = 1|s_3\} + P\{S_4 = a|X_4 = 2, s_3\} P\{X_4 = 2|s_3\}$$
$$= P\{S_4 = a|X_4 = 1\}(0.3276) + P\{S_4 = a|X_4 = 2\}(1 - 0.3276) = (0.99)(0.3276) + (0.96)(0.6724) = 0.9698$$

To compute $P\{S^n = s_n\}$, use the identity $P\{S^n = s_n\} = \sum_i F_n(i)$ along with Equation (4.35). If there are N states of the Markov chain, this requires computing nN quantities $F_n(i)$, with each computation requiring a summation over N terms. This can be compared with a computation of $P\{S^n = s_n\}$ based on conditioning on the first n states of the Markov chain to obtain

$$P\{S^n = s_n\} = \sum_{i_1,\ldots,i_n} P\{S^n = s_n|X_1 = i_1, \ldots, X_n = i_n\} P\{X_1 = i_1, \ldots, X_n = i_n\}$$
$$= \sum_{i_1,\ldots,i_n} p(s_1|i_1) \cdots p(s_n|i_n) p_{i_1} P_{i_1,i_2} P_{i_2,i_3} \cdots P_{i_{n-1},i_n}$$

The use of the preceding identity to compute $P\{S^n = s_n\}$ would thus require a summation over N^n terms, with each term being a product of 2n values, indicating that it is not competitive with the previous approach. The computation of P{Sn = sn} by recursively determining the functions $F_k(i)$ is known as the forward approach. There also is a backward approach, which is based on the quantities $B_k(i)$, defined by $B_k(i) = P\{S_{k+1} = s_{k+1}, \ldots, S_n = s_n|X_k = i\}$ A recursive formula for Bk(i) can be obtained by conditioning on X_{k+1}.

$$B_k(i) = \sum_j P\{S_{k+1} = s_{k+1}, \ldots, S_n = s_n|X_k = i, X_{k+1} = j\} P\{X_{k+1} = j|X_k = i\} = \sum_j P\{S_{k+1} = s_{k+1}, \ldots, S_n = s_n|X_{k+1} = j\} P_{i,j}$$
$$= \sum_j P\{S_{k+1} = s_{k+1}|X_{k+1} = j\}$$
$$\times P\{S_{k+2} = s_{k+2}, \ldots, S_n = s_n|S_{k+1} = s_{k+1}, X_{k+1} = j\} P_{i,j} = \sum_j p(s_{k+1}|j) P\{S_{k+2} = s_{k+2}, \ldots, S_n = s_n|X_{k+1} = j\} P_{i,j} = \sum_j p(s_{k+1}|j) B_{k+1}(j) P_{i,j}$$

(4.36)

Starting with

$$B_{n-1}(i) = P\{S_n = s_n | X_{n-1} = i\} = \sum_j P_{i,j} p(s_n|j)$$

we would then use Equation (4.36) to determine the function $B_{n-2}(i)$, then $B_{n-3}(i)$, and so on, down to $B_1(i)$. This would then yield $P\{S^n = s_n\}$ via

$$P\{S^n = s_n\} = \sum_i P\{S_1 = s_1, \ldots, S_n = s_n | X_1 = i\} p_i = \sum_i P\{S_1 = s_1 | X_1 = i\} P\{S_2 = s_2, \ldots, S_n = s_n | S_1 = s_1, X_1 = i\} p_i$$
$$= \sum_i p(s_1|i) P\{S_2 = s_2, \ldots, S_n = s_n | X_1 = i\} p_i = \sum_i p(s_1|i) B_1(i) p_i$$

Another approach to obtaining $P\{S^n = s_n\}$ is to combine both the forward and backward approaches. Suppose that for some k we have computed both functions $F_k(j)$ and $B_k(j)$. Because

$$P\{S^n = s_n, X_k = j\} = P\{S^k = s_k, X_k = j\} \times P\{S_{k+1} = s_{k+1}, \ldots, S_n = s_n | S^k = s_k, X_k = j\} = P\{S^k = s_k, X_k = j\} P\{S_{k+1} = s_{k+1}, \ldots, S_n = s_n | X_k = j\} = F_k(j) B_k(j)$$

we see that

$$P\{S^n = s_n\} = \sum_j F_k(j) B_k(j)$$

The beauty of using the preceding identity to determine $P\{S^n = s_n\}$ is that we may simultaneously compute the sequence of forward functions, starting with F1, as well as the sequence of backward functions, starting at B_{n-1}. The parallel computations can then be stopped once we have computed both F_k and B_k for some k.

4.11.1 Predicting the States: Suppose the first n observed signals are $s_n = (s_1, \ldots, s_n)$, and that given this data we want to predict the first n states of the Markov chain. The best predictor depends on what we are trying to accomplish. If our objective is to maximize the expected number of states that are correctly predicted, then for each k = 1, . . . , n we need to compute $P\{X_k = j | S^n = s_n\}$ and then let the value of j that maximizes this quantity be the predictor of X_k. (That is, we take the mode of the conditional probability mass function of X_k, given the sequence of signals, as the predictor of X_k.) To do so, we must first compute this conditional probability mass function, which is accomplished as follows. $k \leqslant n$,

$$P\{X_k = j | S^n = s_n\} = \frac{P\{S^n = s_n, X_k = j\}}{P\{S^n = s_n\}} = \frac{F_k(j) B_k(j)}{\sum_j F_k(j) B_k(j)}$$

Thus, given that $S^n = s_n$, the optimal predictor of Xk is the value of j that maximizes $F_k(j) B_k(j)$. A different variant of the prediction problem arises when we regard the sequence of states as a single entity. In this situation, our objective is to choose that sequence of states whose conditional probability, given the sequence of signals, is maximal. For instance, in signal processing, while $X_1, \ldots, X_n$ might be the actual message sent, $S_1, \ldots, S_n$ would be what is received, and so the objective would be to predict the actual message in its entirety. Letting $X_k = (X_1, \ldots, X_k)$ be the vector of the first k states, the problem of interest is to find the sequence of states $i_1, \ldots, i_n$ that maximizes $P\{X_n = (i_1, \ldots, i_n) | S^n = s_n\}$. Because

$$P\{X_n = (i_1, \ldots, i_n) | S^n = s_n\} = \frac{P\{X_n = (i_1, \ldots, i_n), S^n = s_n\}}{P\{S^n = s_s\}}$$

this is equivalent to finding the sequence of states i1, . . . , in that maximizes $P\{X_n = (i_1, \ldots, i_n), S^n = s_n\}$. To solve the preceding problem let, for $k \leqslant n$,

$$V_k(j) = \max_{i_1, \ldots, i_{k-1}} P\{X_{k-1} = (i_1, \ldots, i_{k-1}), X_k = j, S^k = s_k\}$$

To recursively solve for $V_k(j)$, use that

$$V_k(j) = \max_i \max_{i_1, \ldots, i_{k-2}} P\{X_{k-2} = (i_1, \ldots, i_{k-2}), X_{k-1} = i, X_k = j, S^k = s_k\}$$
$$= \max_i \max_{i_1, \ldots, i_{k-2}} P\{X_{k-2} = (i_1, \ldots, i_{k-2}), X_{k-1} = i, S^{k-1} = s_{k-1}, X_k = j, S_k = s_k\}$$
$$= \max_i \max_{i_1, \ldots, i_{k-2}} P\{X_{k-2} = (i_1, \ldots, i_{k-2}), X_{k-1} = i, S^{k-1} = s_{k-1}\} \times P\{X_k = j, S_k = s_k | X_{k-1} = i\}$$
$$= \max_i P_{i,j} p(s_k|j) V_{k-1}(i) = p(s_k|j) \max_i P_{i,j} V_{k-1}(i)$$
$$= \max_i \max_{i_1, \ldots, i_{k-2}} P\{X_{k-2} = (i_1, \ldots, i_{k-2}), X_{k-1} = i, S^{k-1} = s_{k-1}\} \times P\{X_k = j, S_k = s_k | X_{k-2} = (i_1, \ldots, i_{k-2}), X_{k-1} = i, S^{k-1} = s_{k-1}\}$$
$$= \max_i P\{X_k = j, S_k = s_k | X_{k-1} = i\} \times \max_{i_1, \ldots, i_{k-2}} P\{X_{k-2} = (i_1, \ldots, i_{k-2}), X_{k-1} = i, S^{k-1} = s_{k-1}\}$$

(4.37) Starting with $V_1(j) = P\{X_1 = j, S_1 = s_1\} = p_j p(s_1|j)$

we now use the recursive identity (4.37) to determine $V_2(j)$ for each j; then $V_3(j)$ for each j; and so on, up to $V_n(j)$ for each j. To obtain the maximizing sequence of states, we work in the reverse direction. Let j_n be the value (or any of the values if there are more than one) of j that maximizes $V_n(j)$. Thus j_n is the final state of a maximizing state sequence. Also, for k < n, let $i_k(j)$ be a value of i that maximizes $P_{i,j} V_k(i)$. Then

$$\max_{i_1, \ldots, i_n} P\{X_n = (i_1, \ldots, i_n), S^n = s_n\} = \max_j V_n(j) = V_n(j_n) = \max_{i_1, \ldots, i_{n-1}} P\{X_n = (i_1, \ldots, i_{n-1}, j_n), S^n = s_n\} = p(s_n|j_n) \max_i P_{i,j_n} V_{n-1}(i) = p(s_n|j_n) P_{i_{n-1}(j_n), j_n} V_{n-1}(i_{n-1}(j_n))$$

Thus, $i_{n-1}(j_n)$ is the next to last state of the maximizing sequence. Continuing in this manner, the second from the last state of the maximizing sequence is $i_{n-2}(i_{n-1}(j_n))$, and so on. The preceding approach to finding the most likely sequence of states given a prescribed sequence of signals is known as the Viterbi Algorithm.

Ch-5 The Exponential Distribution and the Poisson Process: (Remaining_As_Highlighted_in_TextBook)

5.1 Introduction: In making a mathematical model for a real-world phenomenon it is always necessary to make certain simplifying assumptions so as to render the mathematics tractable. On the other hand, however, we cannot make too many simplifying assumptions, for then our conclusions, obtained from the mathematical

model, would not be applicable to the real-world situation. Thus, in short, we must make enough simplifying assumptions to enable us to handle the mathematics but not so many that the mathematical model no longer resembles the real-world phenomenon. One simplifying assumption that is often made is to assume that certain random variables are exponentially distributed. The reason for this is that the exponential distribution is both relatively easy to work with and is often a good approximation to the actual distribution. The property of the exponential distribution that makes it easy to analyze is that it does not deteriorate with time. By this we mean that if the lifetime of an item is exponentially distributed, then an item that has been in use for ten (or any number of) hours is as good as a new item in regards to the amount of time remaining until the item fails. This will be formally defined in Section 5.2 where it will be shown that the exponential is the only distribution that possesses this property. In Section 5.3 we shall study counting processes with an emphasis on a kind of counting process known as the Poisson process. Among other things we shall discover about this process is its intimate connection with the exponential distribution.

Ch-6 Continuous-Time Markov Chains:

6.1 Introduction
6.2 Continuous-Time Markov Chains
6.3 Birth and Death Processes
6.4 The Transition Probability Function $P_{ij}(t)$
6.5 Limiting Probabilities
6.6 Time Reversibility
6.7 Uniformization
6.8 Computing the Transition Probabilities

6.1 Introduction: In this chapter we consider a class of probability models that has a wide variety of applications in the real world. The members of this class are the continuous-time analogs of the Markov chains of Chapter 4 and as such are characterized by the Markovian property that, given the present state, the future is independent of the past.

One example of a continuous-time Markov chain is the Poisson process of Chapter 5. For if we let the total number of arrivals by time t (that is, N(t)) be the state of the process at time t, then the Poisson process is a continuous-time Markov chain having states 0, 1, 2, . . . that always proceeds from state n to state n + 1, where $n \geqslant 0$. Such a process is known as a pure birth process since when a transition occurs the state of the system is always increased by one. More generally, an exponential model that can go (in one transition) only from state n to either state n − 1 or state n + 1 is called a birth and death model. For such a model, transitions from state n to state n + 1 are designated as births, and those from n to n − 1 as deaths. Birth and death models have wide applicability in the study of biological systems and in the study of waiting line systems in which the state represents the number of customers in the system. These models will be studied extensively in this chapter. In Section 6.2 we define continuous-time Markov chains and then relate them to the discrete-time Markov chains of Chapter 4. In Section 6.3 we consider birth and death processes and in Section 6.4 we derive two sets of differential equations—the forward and backward equations—that describe the probability laws for the system. The material in Section 6.5 is concerned with determining the limiting (or long-run) probabilities connected with a continuous-time Markov chain. In Section 6.6 we consider the topic of time reversibility. We show that all birth and death processes are time reversible, and then illustrate the importance of this observation to queueing systems. In the final section we show how to "uniformize" Markov chains, a technique useful for numerical computations.

6.2 Continuous-Time Markov Chains: Suppose we have a continuous-time stochastic process {X(t), $t \geqslant 0$} taking on values in the set of nonnegative integers. In analogy with the definition of a discrete-time Markov chain, given in Chapter 4, we say that the process {X(t), $t \geqslant 0$} is a continuous-time Markov chain if for all s, $t \geqslant 0$ and

$$P\{X(t+s) = j | X(s) = i,\ X(u) = x(u),\ 0 \leqslant u < s\}$$

nonnegative integers i, j, x(u), $0 \leqslant u < s$

$$= P\{X(t+s) = j | X(s) = i\}$$

In other words, a continuous-time Markov chain is a stochastic process having the Markovian property that the conditional distribution of the future X(t + s) given the present X(s) and the past X(u), $0 \leqslant u < s$, depends only on the present and is independent of the past. If, in addition, P{X(t + s) = j|X(s) = i} is independent of s, then the continuous-time Markov chain is said to have stationary or homogeneous transition probabilities. All Markov chains considered in this text will be assumed to have stationary transition probabilities.

Point1: That is, if we let Ti denote the amount of time that the process stays in state i before making a transition into a different state ….. Hence, the random variable Ti is memoryless and must thus (see Section 5.2.2) be exponentially distributed.

Point2: In fact, the preceding gives us another way of defining a continuous-time Markov chain. Namely, it is a stochastic process having the properties that each time it enters state i.

(i) the amount of time it spends in that state before making a transition into a different state is exponentially distributed with mean, say, 1/vi , and

(ii) when the process leaves state i, it next enters state j with some probability, say, Pij . Of course, the Pij must satisfy

$$P_{ii} = 0, \quad \text{all } i$$
$$\sum_j P_{ij} = 1, \quad \text{all } i$$

Point3: In other words, a continuous-time Markov chain is a stochastic process that moves from state to state in accordance with a (discrete-time) Markov chain, but is such that the amount of time it spends in each state, before proceeding to the next state, is exponentially distributed. In addition, the amount of time the process spends in state i, and the next state visited, must be independent random variables. For if the next state visited were dependent on Ti, then information as to how long the process has already been in state i would be relevant to the prediction of the next state—and this contradicts the Markovian assumption.

6.3 Birth and Death Processes: Consider a system whose state at any time is represented by the number of people in the system at that time. Suppose that whenever there are n people in the system, then (i) new arrivals enter the system at an exponential rate λn, and (ii) people leave the system at an exponential rate μn. That is, whenever there are n persons in the system, then the time until the next arrival is exponentially distributed with mean 1/λn and is independent of the time until the next departure, which is itself exponentially distributed with mean 1/μn. Such a system is called a birth and death process. The parameters $\{\lambda_n\}_{n=0}^{\infty}$ and $\{\mu_n\}_{n=1}^{\infty}$ are called, respectively, the arrival (or birth) and departure (or death) rates. Thus, a birth and death process is a continuous-time Markov chain with states {0, 1, . . .} for which transitions from state n may go only to either state n − 1 or state n + 1. The relationships between the birth and death rates and the state transition rates and probabilities are

$$v_0 = \lambda_0,$$
$$v_i = \lambda_i + \mu_i, \quad i > 0$$
$$P_{01} = 1,$$
$$P_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i}, \quad i > 0$$
$$P_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i}, \quad i > 0$$

The preceding follows, because if there are i in the system, then the next state will be i + 1 if a birth occurs before a death, and the probability that an exponential random variable with rate λi will occur earlier than an (independent) exponential with rate μi is λi/(λi + μi). Moreover, the time until either a birth or a death occurs is exponentially distributed with rate λi + μi (and so, vi = λi + μi).

6.4 The Transition Probability Function $P_{ij}(t)$: Let Pij(t) = P{X(t + s) = j|X(s) = i} denote the probability that a process presently in state i will be in state j a time t later. These quantities are often called the transition probabilities of the continuous-time Markov chain. We can explicitly determine Pij(t) in the case of a pure birth process having distinct birth rates. **Proposition 6.1 Lemma 6.2 Lemma 6.3 (6.8)** The set of Equations (6.8) is known as the **Chapman–Kolmogorov equations.** Now, assuming that we can interchange the limit and the summation in the preceding and applying Lemma 6.2, we obtain $P'_{ij}(t) = \sum_{k \neq i} q_{ik} P_{kj}(t) - v_i P_{ij}(t)$ It turns out that this **interchange** can indeed be **justified** and, hence, we have the following theorem**. Theorem 6.1 (Kolmogorov's Backward Equations):** Another set of differential equations, different from the backward equations, may also be derived. This set of equations, known **as Kolmogorov's forward equations** is derived as follows. From the Chapman–Kolmogorov equations (Lemma 6.3), and, assuming that we can interchange limit with summation. Unfortunately, we cannot always justify the interchange of limit and summation and thus the preceding is not always valid. However, they do hold in most models, including all birth and death processes and all finite state models. We thus have the following**. Theorem 6.2 (Kolmogorov's Forward Equations)** Under suitable regularity conditions **Proposition 6.4** For a pure birth process

6.5 Limiting Probabilities: In analogy with a basic result in discrete-time Markov chains, the probability that a continuous-time Markov chain will be in state j at time t often converges to a limiting value that is independent of the initial state. That is, if we call this value Pj, then $P_j \equiv \lim_{t \to \infty} P_{ij}(t)$ where we are assuming that the limit exists and is independent of the initial state i.

6.6 Time Reversibility: Consider a continuous-time Markov chain that is ergodic and let us consider the limiting probabilities Pi from a different point of view than previously. If we consider the sequence of states visited, ignoring the amount of time spent in each state during a visit, then this sequence constitutes a discrete-time Markov chain withtransition probabilities Pij. Let us assume that this discrete-time Markov chain, called the embedded chain, is ergodic and denote by πi its limiting probabilities.

Since Pi is the proportion of time in state i and qij is the rate when in state i that the process goes to j, the condition of time reversibility is that the rate at which the process goes directly from state i to state j is equal to the rate at which it goes directly from j to i. It should be noted that this is exactly the same condition needed for an ergodic discrete-time Markov chain to be time reversible (see Section 4.8). An application of the preceding condition for time reversibility yields the following proposition concerning birth and death processes.**Proposition 6.5** An ergodic birth and death process is time reversible. Proposition 6.5 can be used to prove the important result that the output process of an M/M/s queue is a Poisson process. We state this as a corollary. **Corollary 6.6** Consider an **M/M/s** queue in which customers arrive in accordance with a Poisson process having rate λ and are served by any one of s servers— each having an exponentially distributed service time with rate μ. If λ < sμ, then the output process of customers departing is, after the process has been in operation for a long time, a Poisson process with rate λ.

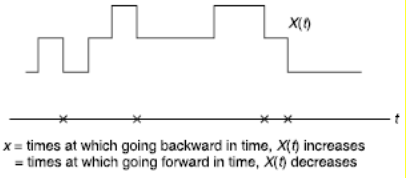


Figure 6.1 The number in the system.

Proof. Let X(t) denote the number of customers in the system at time t. Since the **M/M/s** process is a birth and death process, it follows from Proposition 6.5 that {X(t), $t \geqslant 0$} is time reversible. Going forward in time, the time points at which X(t) increases by 1 constitute a Poisson process since these are just the arrival times of customers. Hence, by time reversibility the time points at which X(t) increases by 1 when we go backward in time also constitute a Poisson process. But these latter points are exactly the points of time when customers depart (see Figure 6.1). Hence, the departure times constitute a Poisson process with rate λ.

Analogous to the result for discrete-time Markov chains, if we can find a probability vector P that satisfies the preceding then the Markov chain is time reversible and the Pis are the long-run probabilities. That is, we have the following proposition. **Proposition 6.7** If for some set {Pi}

$$\sum_i P_i = 1, \qquad P_i \geqslant 0$$

and

$$P_i q_{ij} = P_j q_{ji} \qquad \text{for all } i \neq j$$

then the continuous-time Markov chain is time reversible and Pi represents the limiting probability of being in state i. Proof. For fixed i we obtain upon summing Equation (6.27) over all $j : j \neq i$

$$\sum_{j \neq i} P_i q_{ij} = \sum_{j \neq i} P_j q_{ji}$$

$$\text{or, since } \sum_{j \neq i} q_{ij} = v_i,$$

$$v_i P_i = \sum_{j \neq i} P_j q_{ji}$$

Hence, the Pis satisfy the balance equations and thus represent the limiting probabilities. Because Equation (6.27) holds, the chain is time reversible. **Proposition 6.8** A time reversible chain with limiting probabilities Pj, j ∈ S that is truncated to the set A ⊂ S and remains irreducible is also time reversible and has limiting probabilities PA j given by

$$P_j^A = \frac{P_j}{\sum_{i \in A} P_i}, \qquad j \in A$$

Proof. By Proposition 6.7 we need to show that, with P_j^A as given,

$$P_i^A q_{ij} = P_j^A q_{ji} \qquad \text{for } i \in A,\ j \in A$$

or, equivalently,

$$P_i q_{ij} = P_j q_{ji} \qquad \text{for } i \in A,\ j \in A$$

But this follows since the original chain is, by assumption, time reversible.

Proposition 6.9 If {Xi(t), t ≥ 0} are, for i = 1, . . . , n, independent time reversible continuous-time Markov chains, then the vector process {(Xi(t), . . . ,Xn(t)), t ≥ 0} is also a time reversible continuous-time Markov chain.

6.7 Uniformization: Consider a continuous-time Markov chain in which the mean time spent in a state is the same for all states. That is, suppose that vi = v, for all states i. In this case since the amount of time spent in each state during a visit is exponentially distributed with rate v, it follows that if we let N(t) denote the number of state transitions by time t, then {N(t), t ≥ 0} will be a Poisson process with rate v. To compute the transition probabilities Pij(t), we can condition on N(t):

$$\begin{aligned} P_{ij}(t) &= P\{X(t) = j | X(0) = i\} \\ &= \sum_{n=0}^{\infty} P\{X(t) = j | X(0) = i,\ N(t) = n\} P\{N(t) = n | X(0) = i\} \\ &= \sum_{n=0}^{\infty} P\{X(t) = j | X(0) = i,\ N(t) = n\} e^{-vt} \frac{(vt)^n}{n!} \end{aligned}$$

Now, the fact that there have been n transitions by time t tells us something about the amount of time spent in each of the first n states visited, but since the distribution of time spent in each state is the same for all states, it follows that knowing that N(t) = n gives us no information about which states were visited. Hence,

$$P\{X(t) = j | X(0) = i,\ N(t) = n\} = P_{ij}^n$$

where P_{ij}^n is just the n-stage transition probability associated with the discrete-time Markov chain with transition probabilities Pij; and so when vi ≡ v

$$P_{ij}(t) = \sum_{n=0}^{\infty} P_{ij}^n e^{-vt} \frac{(vt)^n}{n!} \qquad (6.29)$$

Equation (6.29) is often useful from a computational point of view since it enables us to approximate Pij(t) by taking a partial sum and then computing (by matrix multiplication of the transition probability matrix) the relevant n stage probabilities P_{ij}^n .

6.8 Computing the Transition Probabilities: For any pair of states i and j, let

$$r_{ij} = \begin{cases} q_{ij}, & \text{if } i \neq j \\ -v_i, & \text{if } i = j \end{cases}$$

Using this notation, we can rewrite the Kolmogorov backward equations

$$P'_{ij}(t) = \sum_{k \neq i} q_{ik} P_{kj}(t) - v_i P_{ij}(t)$$

and the forward equations

$$P'_{ij}(t) = \sum_{k \neq j} q_{kj} P_{ik}(t) - v_j P_{ij}(t)$$

as follows:

$$P'_{ij}(t) = \sum_k r_{ik} P_{kj}(t) \qquad \text{(backward)}$$

$$P'_{ij}(t) = \sum_k r_{kj} P_{ik}(t) \qquad \text{(forward)}$$

This representation is especially revealing when we introduce matrix notation. Define the matrices R and P(t), $\mathbf{P}'(t)$ by letting the element in row i, column j of these matrices be, respectively, rij, Pij(t), and $P'_{ij}(t)$. Since the backward equations say that the element in row i, column j of the matrix $\mathbf{P}'(t)$ can be obtained by multiplying the ith row of the matrix R by the jth column of the matrix P(t), it is equivalent to the matrix equation

$$\mathbf{P}'(t) = \mathbf{RP}(t) \qquad (6.32)$$

Similarly, the forward equations can be written as

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{R} \qquad (6.33)$$

Now,

$$f'(t) = cf(t)$$

$$\text{(or, equivalent, } f'(t) = f(t)c) \text{ is}$$

$$f(t) = f(0)e^{ct}$$

just as the solution of the scalar differential equation it can be shown that the solution of the matrix differential Equations (6.32) and (6.33) is given by

$$\mathbf{P}(t) = \mathbf{P}(0)e^{\mathbf{R}t}$$

Since P(0) = I (the identity matrix), this yields that

$$\mathbf{P}(t) = e^{\mathbf{R}t} \qquad (6.34)$$

where the matrix $e^{\mathbf{R}t}$ is defined by

$$e^{\mathbf{R}t} = \sum_{n=0}^{\infty} \mathbf{R}^n \frac{t^n}{n!} \qquad (6.35)$$

with $\mathbf{R}^n$ being the (matrix) multiplication of R by itself n times.

The direct use of Equation (6.35) to compute P(t) turns out to be very inefficient for two reasons. First, since the matrix R contains both positive and negative

elements (remember the off-diagonal elements are the qij while the ith diagonal element is −vi), there is the problem of computer round-off error when we compute the powers of R. Second, we often have to compute many of the terms in the infinite sum (6.35) to arrive at a good approximation. However, there are certain indirect ways that we can utilize the relation in (6.34) to efficiently approximate the matrix P(t). We now present two of these methods.

Approximation Method 1: Rather than using Equation (6.35) to compute $e^{\mathbf{R}t}$, we can use the matrix equivalent of the identity $e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$ which states that $e^{\mathbf{R}t} = \lim_{n\to\infty}\left(\mathbf{I}+\mathbf{R}\frac{t}{n}\right)^n$ Thus, if we let n be a power of 2, say, 2^k, then we can approximate **P**(t) by raising the matrix M = **I** + **R**t/n to the nth power, which can be accomplished by k matrix multiplications (by first multiplying **M** by itself to obtain $\mathbf{M}^2$ and then multiplying that by itself to obtain $\mathbf{M}^4$ and so on). In addition, since only the diagonal elements of **R** are negative (and the diagonal elements of the identity matrix **I** are equal to 1), by choosing n large enough we can guarantee that the matrix **I** + **R**t/n has all nonnegative elements.

Approximation Method 2: A second approach to approximating $e^{\mathbf{R}t}$ uses the identity

$$e^{-\mathbf{R}t} = \lim_{n\to\infty}\left(\mathbf{I}-\mathbf{R}\frac{t}{n}\right)^n \approx \left(\mathbf{I}-\mathbf{R}\frac{t}{n}\right)^n \quad \text{for } n \text{ large}$$

and thus

$$\mathbf{P}(t) = e^{\mathbf{R}t} \approx \left(\mathbf{I}-\mathbf{R}\frac{t}{n}\right)^{-n} = \left[\left(\mathbf{I}-\mathbf{R}\frac{t}{n}\right)^{-1}\right]^n$$

Hence, if we again choose n to be a large power of 2, say, n = 2^k, we can approximate **P**(t) by first computing the inverse of the matrix **I** − **R**t/n and then raising that matrix to the nth power (by utilizing k matrix multiplications). It can be shown that the matrix $(\mathbf{I}-\mathbf{R}t/n)^{-1}$ will have only nonnegative elements.

Remark: Both of the preceding computational approaches for approximating **P**(t) have probabilistic interpretations (see Exercises 41 and 42).

Ch-1 Introduction to Probability Theory: (Remaining_As_Highlighted_in_TextBook)

Ch-2 Random Variables: (Remaining_As_Highlighted_in_TextBook)

Ch-3 Conditional Probability and Conditional Expectation: (Remaining_As_Highlighted_in_TextBook)