

Chapter 4: Expected Values

October 11, 2009

1 The expected value of a random variable

Definition 1.1 *If X is a discrete random variable with frequency function $p(x)$, the expected value of X , denoted by $E(X)$, is*

$$E(X) = \sum_i x_i p(x_i)$$

provided that $\sum_i |x_i| p(x_i) < \infty$. If the sum diverges, the expectation is undefined.

$E(X)$ is also referred to as the mean of X and is often denoted by μ or μ_X .

EX A (Roulette) A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet \$1 that an odd number comes up, you win or lose \$1 according to whether that event occurs. Let X denote your net gain.

$$E(X) = 1 \times \frac{18}{38} + (-1) \times \frac{20}{38} = -\frac{1}{19}.$$

EX B (Expectation of a geometric random variable)

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} k p q^{k-1} = p \sum_{k=1}^{\infty} k q^{k-1} \\ &= p \frac{d}{dq} \sum_{k=1}^{\infty} q^k \quad (\text{because } k q^{k-1} = \frac{d}{dq} q^k) \\ &= p \frac{d}{dq} \frac{q}{1-q} = \frac{1}{p} \end{aligned}$$

Thus, for example, if 10% of the items are defective, an average of 10 items must be examined to find one that is defective.

EX C (Poisson distribution)

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \\ &= \lambda \end{aligned}$$

EX D (St. Petersburg Paradox) A gambler has the following strategy for playing a sequence of games: He starts off being \$1; if he loses, he doubles his bet; and he continues to double his bet until he wins. Suppose that the game is fair and that he wins or loses the amount he bets.

Let X denote the amount of money bet on the very last game (the game he wins).

$$P(X = 1) = \frac{1}{2}, \quad P(X = 2) = \frac{1}{2^2}, \quad \dots, \quad P(X = 2^k) = \frac{1}{2^{k+1}}.$$

Then

$$E(X) = \sum_{k=0}^{\infty} 2^k \frac{1}{2^{k+1}} = \infty.$$

Formally, $E(X)$ is not defined. This strategy requires an enormous amount of capital.

Definition 1.2 If X is a continuous random variable with density $f(x)$, then

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

provided that $\int |x| f(x) dx < \infty$. If the integral diverges, the expectation is undefined.

EX E (Gamma density) If X follows a gamma density with parameters α and λ ,

$$\begin{aligned} E(X) &= \int_0^{\infty} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha} e^{-\lambda x} dx \\ &= \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 1)}{\lambda^{\alpha+1}} \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

EX F (Normal density)

$$\begin{aligned} E(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-z^2/2\sigma^2} dz + \frac{\mu}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} dz, \\ &\quad \text{(setting } z = x - \mu) \\ &= 0 + \mu = \mu \end{aligned}$$

EX G (Cauchy density) Recall that the Cauchy density is

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad -\infty < x < \infty$$

Since $\int_{-\infty}^{\infty} \frac{|x|}{1+x^2} dx = \infty$, the expectation does not exist.

The expected value can be interpreted as a long-run average. In chapter 5, it will be shown that if $E(X)$ exists and if X_1, X_2, \dots is a sequence of independent random variables with the same distribution as X , and if $S_n = \sum_{i=1}^n X_i$, then, as $n \rightarrow \infty$,

$$\frac{S_n}{n} \rightarrow E(X).$$

Theorem 1.1 (*Markov's inequality*) IF X is a random variable with $P(X \geq 0) = 1$ and for which $E(X)$ exists, then $P(X \geq t) \leq E(X)/t$.

PROOF:

$$E(X) \geq \sum_{x \geq t} xp(x) \geq t \sum_{x \geq t} p(x) = tP(X \geq t).$$

□

1.1 Expectations of Functions of Random Variables

We often need to find $Eg(X)$, where X is a random variable and g is a fixed function.

Theorem 1.2 *Suppose that $Y = g(X)$.*

(a) *If X is discrete with frequency function $p(x)$, then*

$$E(Y) = \sum_x g(x)p(x).$$

provided that $\sum |g(x)|p(x) < \infty$.

(b) *If X is continuous with density function $f(x)$, then*

$$E(Y) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

provided that $\int |g(x)|f(x)dx < \infty$.

PROOF: We will prove this result for the discrete case. The basic argument is the same for the continuous case. By definition,

$$E(Y) = \sum_i y_i p_Y(y_i).$$

Let A_i denote the set of x 's mapped to y_i by g ; that is, $x \in A_i$, if $g(x) = y_i$. Then

$$\begin{aligned} E(Y) &= \sum_i y_i \sum_{x \in A_i} p(x) \\ &= \sum_i \sum_{x \in A_i} y_i p(x) \\ &= \sum_i \sum_{x \in A_i} g(x) p(x) \\ &= \sum_x g(x) p(x) \end{aligned}$$

This last step follows because the A_i are disjoint and every x belongs to some A_i . \square

EX A Suppose that the pdf of a random variable X is

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E(X) = \int_0^1 x(2x)dx = \int_0^1 2x^2 dx = \frac{2}{3}.$$

Let $Y = \sqrt{X}$. Then

$$E(Y) = \int_0^1 x^{1/2}(2x)dx = 2 \int_0^1 x^{3/2} dx = \frac{4}{5}.$$

Theorem 1.3 Suppose that X_1, \dots, X_n are jointly distributed random variables and $Y = g(X_1, \dots, X_n)$.

(a) If the X_i are discrete with frequency function $p(x_1, \dots, x_n)$, then

$$E(Y) = \sum_{x_1, \dots, x_n} g(x_1, \dots, x_n) p(x_1, \dots, x_n),$$

provided that $\sum_{x_1, \dots, x_n} |g(x_1, \dots, x_n)| p(x_1, \dots, x_n) < \infty$.

(b) If the X_i are continuous with joint density function $f(x_1, \dots, x_n)$, then

$$E(Y) = \int \int \cdots \int g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \cdots dx_n,$$

provided that $\int \int \cdots \int |g(x_1, \dots, x_n)| p(x_1, \dots, x_n) dx_1 \cdots dx_n < \infty$.

EX B A stick of unit length is broken randomly in two places. What is the average length of the middle piece?

We interpret this question to mean that the locations of the two break points are independent uniform random variables U_1 and U_2 .

$$\begin{aligned} E|U_1 - U_2| &= \int_0^1 \int_0^1 |u_1 - u_2| du_1 du_2 \\ &= \int_0^1 \int_0^{u_1} (u_1 - u_2) du_2 du_1 + \int_0^1 \int_{u_1}^1 (u_2 - u_1) du_2 du_1 \\ &= \frac{1}{3} \end{aligned}$$

Corollary 1.1 *If X and Y are independent random variables and g and h are fixed functions, then $E[g(X)h(Y)] = \{E[g(X)]\}\{E[h(Y)]\}$, provided that the expectations on the right-hand side exist.*

In particular, if X and Y are independent, then $E(XY) = E(X)E(Y)$.

1.2 Expectations of Linear Combinations of Random Variables

One of the most useful properties of the expectation is that it is a linear operation.

Theorem 1.4 *If $Y = aX + b$, where a and b are constants, then*

$$E(Y) = aE(X) + b.$$

Theorem 1.5 *If there exists a constant such that $P(X \geq a) = 1$, then $E(X) \geq a$.
If there exists a constant b such that $P(X \leq b) = 1$, then $E(X) \leq b$.*

PROOF: We shall assume, for convenience, that X has a continuous distribution for which the pdf is f , and we shall suppose first that $P(X \geq a) = 1$. Then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_a^{\infty} x f(x) dx \\ &\geq \int_a^{\infty} a f(x) dx = aP(X \geq a) = a. \end{aligned}$$

The proof of the other part of the theorem and the proof for a discrete distribution are similar. \square

Theorem 1.6 *If X_1, \dots, X_n are n random variables such that each expectation $E(X_i)$ exists ($i = 1, \dots, n$), then*

$$E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n).$$

PROOF: We shall assume that $n = 2$ and also, for convenience, that X_1 and X_2 have a continuous joint distribution for which the joint pdf is f . Then

$$\begin{aligned} E(X_1 + X_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 + x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1, x_2) dx_1 dx_2 + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 f(x_1, x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{\infty} x_1 f_1(x_1) dx_1 + \int_{-\infty}^{\infty} x_2 f_2(x_2) dx_2 \\ &= E(X_1) + E(X_2). \end{aligned}$$

The proof for a discrete distribution is similar. \square

Theorem 1.7 If X_1, \dots, X_n are jointly distributed random variables with expectations $E(X_i)$ and Y is a linear function of the X_i , $Y = a + \sum_{i=1}^n b_i X_i$, then

$$E(Y) = a + \sum_{i=1}^n b_i E(X_i).$$

EX A Suppose that a box contains red and blue balls and that the proportion of red balls in the box is p ($0 \leq p \leq 1$). Suppose that n balls are selected from the box at random without replacement, and let X denote the number of red balls that are selected. We shall determine the value of $E(X)$.

Let $X_i = 1$ if the i th ball that is selected is red, and let $X_i = 0$ if the i th ball selected is blue. We can imagine that all the balls are arranged in the box in some random order, and that the first n balls in this arrangement are selected. Because of randomness, the probability that the i th ball in the arrangement will be red is simply p . Hence, for $i = 1, \dots, n$,

$$P(X_i = 1) = p, \quad \text{and} \quad P(X_i = 0) = 1 - p.$$

Therefore $E(X_i) = p$, and

$$E(X) = E(X_1) + \dots + E(X_n) = np.$$

EX B (Coupon collection) Suppose that you collect coupons, that there are n distinct types of coupons, and that on each trial you are equally likely to get a coupon of any of the types. How many trials would you expect to go through until you had a complete set of coupons?

Let X_1 be the number of trials up to and including the trial on which the first coupon is collected: $X_1 = 1$. Let X_2 be the number of trials from that point up to and including the trial on which the next coupon different from the first is obtained; and so on, up to X_n .

We now find the distribution of X_r . At this point, $r - 1$ of n coupons have been collected, so on each trial the probability of success is $(n - r + 1)/n$. Therefore, X_r is a geometric random variable, with $E(X_r) = n/(n - r + 1)$. Thus,

$$E(X) = \sum_{r=1}^n E(X_r) = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = n \sum_{r=1}^n \frac{1}{r}.$$

EX C Suppose that a large number, n , of blood samples are to be screened for a relatively rare disease. If each sample is assayed individually, n tests will be required. On other hand, if each sample is divided in half and one of the halves is put into a pool with all the other halves, the pooled lot can be tested. If this test is negative, no further assays are necessary and only one test has to be performed. If the test on the pooled blood is positive, each reserved half-sample can be tested individually. In this case, a total of $n + 1$ tests will be required. It is therefore plausible, assuming that the disease is rare, that some savings can be achieved through this pooling procedure.

Let us first generalize the scheme and suppose that the n samples are first grouped into m groups of k samples each, or $n = mk$. Each group is tested; if a group tests positively, each individual in the group is tested. If X_i is the number of tests run on the i th group, then

$$E(X_i) = p^k + (k + 1)(1 - p^k) = k + 1 - kp^k,$$

where p is the probability of a negative on any individual sample. We now have

$$E(N) = m(k + 1) - mkp^k = n\left(1 + \frac{1}{k} - p^k\right).$$

Figure 1 shows the factor $1 + \frac{1}{k} - p^k$, the number of samples used in a group testing as a proportion of n , for $p = 0.99$.

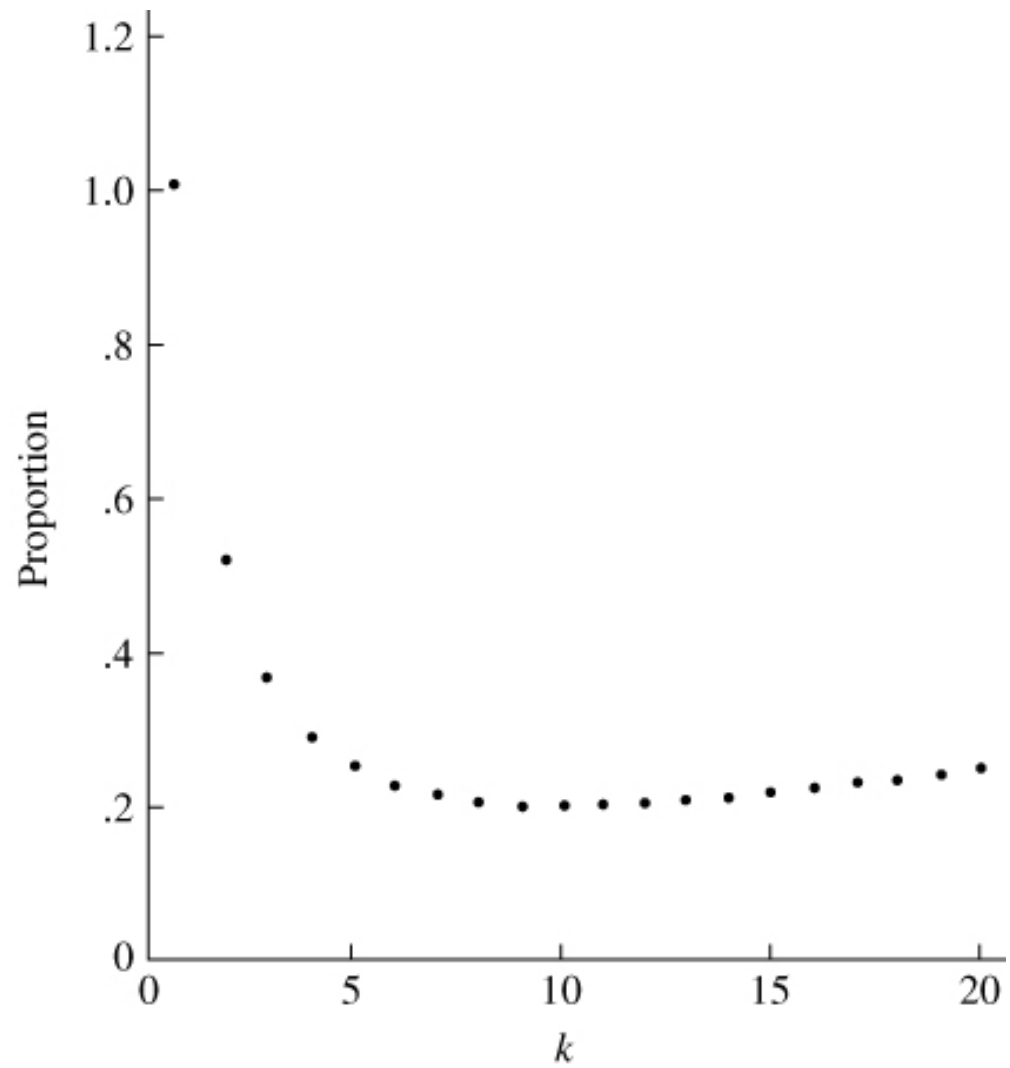


Figure 1: The proportion of n in the average number of samples tested using group testing as a function of k .

2 Variance and Standard Deviation

Definition 2.1 *If X is random variable with expected value $E(X)$, the variance of X is*

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

provided that the expectation exists. The standard deviation of X is the square root of the variance.

If the expectation $E[(X - \mu)^2]$ does not exist, it is said that $\text{Var}(X)$ does not exist. The variance of a random variable X depends only on the distribution of X . The variance of a distribution provides a measure of the spread or dispersion of the distribution around its mean μ . A small value of the variance indicates that the distribution is tightly concentrated around μ ; and a large value of the variance typically indicates that the distribution has a wide spread around μ .

The standard deviation of a given random variable is commonly denoted by the symbol σ , and the variance is denoted by σ^2 .

EX A Suppose that a random variable X can take each of the five values -2, 0, 1, 3, and 4 with equal probability.

The mean of X is

$$E(X) = \frac{1}{5}(-2 + 0 + 1 + 3 + 4) = 1.2.$$

The variance of X is

$$\text{Var}(X) = \frac{1}{5}[(-2-1.2)^2 + (0-1.2)^2 + (1-1.2)^2 + (3-1.2)^2 + (4-1.2)^2] = 4.56.$$

The standard deviation of X is $\sqrt{\text{Var}(X)} = 2.135$.

Theorem 2.1 $\text{Var}(X) = 0$ if and only if there exists a constant c such that $P(X = c) = 1$.

PROOF: If $P(X = c) = 1$, then $P((X - c)^2 = 0) = 1$. Therefore,

$$\text{Var}(X) = E(X - c)^2 = 0.$$

Conversely, if $\text{Var}(X) = 0$, then for any a given $\epsilon > 0$,

$$0 = E(X - \mu)^2 \geq \epsilon P((X - \mu)^2 \geq \epsilon) + \int_{\{x:(x-\mu)^2 < \epsilon\}} (x - \mu)^2 f(x) dx.$$

Hence, $P((X - \mu)^2 \geq \epsilon) = 0$ and thus, $P((X - \mu)^2 < \epsilon) = 1$. Due to the arbitrariness of ϵ , we have

$$P((X - \mu)^2 = 0) = 1,$$

that is, $P(X = \mu) = 1$. \square

Theorem 2.2 For constants a and b ,

$$\text{Var}(aX + b) = a^2 \text{Var}(X).$$

Theorem 2.3 If X_1, \dots, X_n are independent random variables, then

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n).$$

It should be emphasized that the random variables in Theorem 2.3 must be independent, otherwise, the equality will not hold.

EX B (Binomial distribution) Suppose that a box contains red and blue balls and that the proportion of red balls in the box is p ($0 \leq p \leq 1$). Suppose now, however, that a random sample of n balls is selected from the box with replacement. As before, let $X_i = 1$ if the i th ball that is selected is red, and let $X_i = 0$ otherwise. Then $X = X_1 + \cdots + X_n$ is the number of red balls in the sample, and X has a binomial distribution with parameters n and p .

Since X_1, \dots, X_n are independent,

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) = n[E(X_1^2) - (EX_1)^2] = np(1 - p).$$

EX C (Normal distribution) We have seen that $E(X) = \mu$. Then

$$\text{Var}(X) = E(X - \mu)^2 = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} dx$$

Making the change of variables $x = (x - \mu)/\sigma$, we get

$$\text{Var}(X) = \sigma^2.$$

Theorem 2.4 For every random variable X ,

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

provided that the variance of X exists.

EX D (Uniform distribution) Let X be a uniform random variable on $[0,1]$. Then $E(X) = 1/2$ and

$$E(X^2) = \int_0^1 x^2 dx = 1/3.$$

Thus, $\text{Var}(X) = 1/3 - (1/2)^2 = 1/12$.

Theorem 2.5 (Chebyshev Inequality). Let X be a random variable for which $\text{Var}(X)$ exists. Then for every number $t > 0$,

$$P(|X - E(X)| \geq t) \leq \frac{\text{Var}(X)}{t^2}.$$

PROOF: Let $Y = [X - E(X)]^2$. Then $P(Y \geq 0) = 1$ and $E(Y) = \text{Var}(X)$. By applying the Markov inequality of Y , we obtain the following result:

$$P(|X - E(X)| \geq t) = P(Y \geq t^2) \leq \frac{\text{Var}(X)}{t^2}.$$

□

Markov and Chebyshev inequalities are very useful. For example, if $\text{Var}(X) = \sigma^2$ and we let $t = 3\sigma$, then the Chebyshev inequality yields the result that

$$P(|X - E(X)| \geq 3\sigma) \leq \frac{1}{9}.$$

This states that the probability that *any* given random variable will differ from its mean by more than 3 standard deviations cannot exceed 1/9. The inequality is tight in the sense that it cannot be made any smaller and still hold for all distributions.

3 Covariance and Correlation

When we are interested in the joint distribution of two random variables, it is useful to have a summary of how much the two random variables depend on each other. The covariance and correlation are attempts to measure that dependence, but they only capture a particular type of dependence, namely linear dependence.

Definition 3.1 *Let X and Y be random variables having a specified joint distribution; and let $E(X) = \mu_X$, $E(Y) = \mu_Y$, $\text{Var}(X) = \sigma_X^2$, and $\text{Var}(Y) = \sigma_Y^2$. The covariance of X and Y , which is denoted by $\text{Cov}(X, Y)$, is defined as follows:*

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

As shown below, if $\sigma_X^2 < \infty$ and $\sigma_Y^2 < \infty$, then the covariance will be finite.

EX A Let X and Y has a continuous distribution with joint pdf

$$f(x, y) = \begin{cases} x + y & \text{for } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

It is easy to compute that $\mu_X = \mu_Y = 7/12$. The covariance of X and Y is

$$\int_0^1 \int_0^1 (x - 7/12)(y - 7/12)(x + y) dx dy = -1/144.$$

Theorem 3.1 For all random variables X and Y such that $\sigma_X^2 < \infty$ and $\sigma_Y^2 < \infty$,

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

Theorem 3.2 Suppose that $U = a + \sum_{i=1}^n b_i X_i$ and $V = c + \sum_{j=1}^m d_j Y_j$. Then

$$\text{Cov}(U, V) = \sum_{i=1}^n \sum_{j=1}^m b_i d_j \text{Cov}(X_i, Y_j).$$

Theorem 3.3 *If X and Y are random variables such that $\text{Var}(X) < \infty$ and $\text{Var}(Y) < \infty$, then*

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

Theorem 3.4 *If X_1, \dots, X_n are random variables such that $\text{Var}(X_i) < \infty$ for $i = 1, \dots, n$, then*

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \sum_{j=1}^n \text{Cov}(X_i, X_j).$$

EX B A drunken walker starts out at a point x_0 on the real line. He takes a step on length X_1 , which is a random variable with expected value μ and variance σ^2 , and his position at that time is $S(1) = x_0 + X_1$. He takes another step of length X_2 , which is independent of X_1 with the same mean and standard deviation. His position after n such steps is $S(n) = x_0 + \sum_{i=1}^n X_i$. Then

$$E(S(n)) = x_0 + E\left(\sum_{i=1}^n X_i\right) = x_0 + n\mu,$$

$$\text{Var}(S(n)) = \text{Var}\left(\sum_{i=1}^n X_i\right) = n\sigma^2.$$

Random walks have found applications in many areas of science. Brownian motion is a continuous time version of a random walk with the steps being normally distributed random variables. The theory of Brownian motion was developed by Louis Bachelier in 1900 in his PhD thesis “The theory of speculation”, which related random walks to the evolution of stock prices. If the value of a stock evolves through time as a random walk, its short-term behavior is unpredictable. The efficient market hypothesis states that stock prices already reflect all known information so that the future price is random and unknowable.

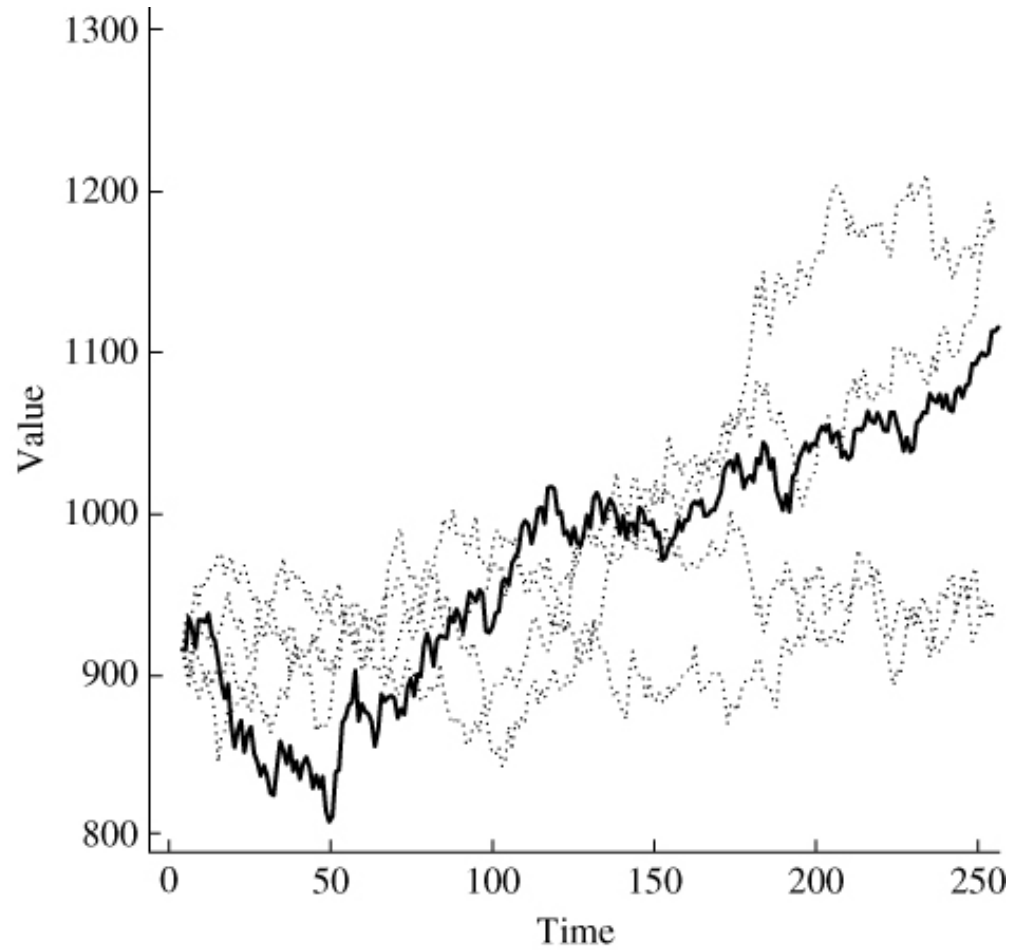


Figure 2: The solid line is the value of the S&P500 during 20003. The dashed lines are simulations of random walks.

Definition 3.2 If $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$, then the correlation of X and Y , which is denoted by $\rho(X, Y)$, is defined as follows:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Theorem 3.5 For any random variables X and Y ,

- a. $-1 \leq \rho(X, Y) \leq 1$.
- b. $|\rho(X, Y)| = 1$ if and only if there exist numbers $a \neq 0$ and b such that $P(Y = aX + b) = 1$. If $\rho(X, Y) = 1$, then $a > 0$, and if $\rho(X, Y) = -1$, then $a < 0$.

PROOF: Consider the function $h(t)$ defined by

$$\begin{aligned} h(t) &= E((X - \mu_X)t + (Y - \mu_Y))^2 \\ &= t^2 \sigma_X^2 + 2t \text{Cov}(X, Y) + \sigma_Y^2. \end{aligned}$$

Since $h(t) \geq 0$ and it is quadratic function,

$$(2\text{Cov}(X, Y))^2 - 4\sigma_X^2 \sigma_Y^2 \leq 0.$$

This is equivalent to

$$-\sigma_X\sigma_Y \leq \text{Cov}(X, Y) \leq \sigma_X\sigma_Y.$$

That is,

$$-1 \leq \rho(X, Y) \leq 1.$$

Also, $|\rho(X, Y)| = 1$ if and only if the discriminant is equal to 0, that is, if and only if $h(t)$ has a single root. But since $((X - \mu_X)t + (Y - \mu_Y))^2 \geq 0$, $h(t) = 0$ if and only if

$$P((X - \mu_X)t + (Y - \mu_Y) = 0) = 1.$$

This $P(Y = aX + b) = 1$ with $a = -t$ and $b = \mu_X t + \mu_Y$, where t is the root of $h(t)$. Using the quadratic formula, we see that this root is $t = -\text{Cov}(X, Y)/\sigma_X^2$. Thus $a = -t$ has the same sign as $\rho(X, Y)$, proving the final assertion. \square

EX C We will show that the covariance of X and y when they follow a bivariate normal distribution is $\rho\sigma_X\sigma_Y$, which means that ρ is the correlation coefficient.

The covariance is

$$\text{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy.$$

Making the changes $u = (x - \mu_X)/\sigma_X$ and $v = (y - \mu_Y)/\sigma_Y$, we have

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{\sigma_X \sigma_Y}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} uv \exp \left[-\frac{1}{2(1 - \rho^2)} (u^2 + v^2 - 2\rho uv) \right] dudv \\ &= \frac{\sigma_X \sigma_Y}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} v \exp(-v^2/2) dv \int_{-\infty}^{\infty} u \exp \left[-\frac{1}{2(1 - \rho^2)} (u - \rho v)^2 \right] du \\ &= \frac{\rho \sigma_X \sigma_Y}{\sqrt{2\pi}} \int_{-\infty}^{\infty} v \exp(-v^2/2) dv \\ &= \rho \sigma_X \sigma_Y. \end{aligned}$$

The correlation coefficient ρ measures the strength of the linear relationship between X and Y . Correlation often affects the appearance of a scatterplot.

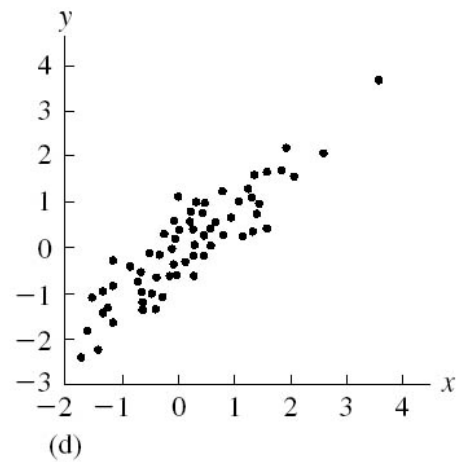
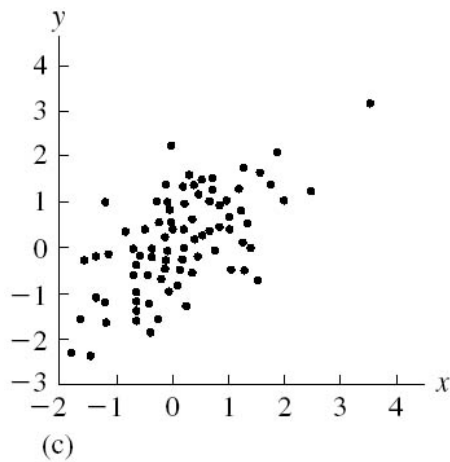
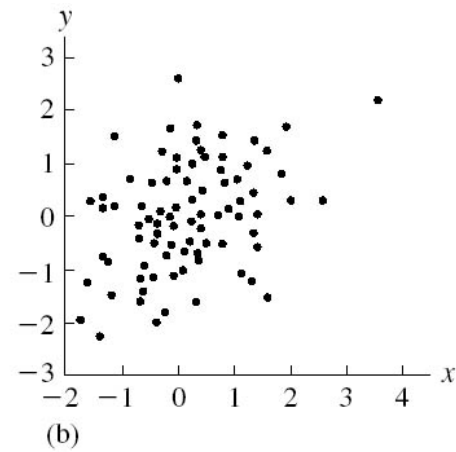
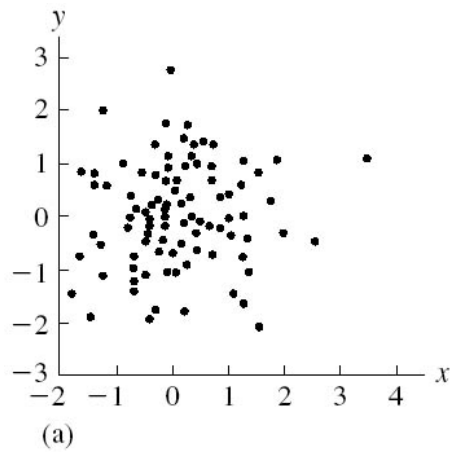


Figure 3: Scatterplots of 100 independent pairs of bivariate normal random variables. (a) $\rho = 0$, (b) $\rho = 0.3$, (c) $\rho = 0.6$, (d) $\rho = 0.9$.

4 Conditional Expectation and Prediction

Definition and Basic Properties

Suppose that X and Y are random variables with a continuous joint distribution. The conditional expectation of Y given $X = x$ is denoted by $E(Y|x)$ and is defined to be the expectation of the distribution whose pdf is $f_{Y|X}(y|x)$. That is,

$$E(Y|x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy.$$

Similarly, if X and Y have a discrete joint distribution, then

$$E(Y|x) = \sum_y y p_{Y|X}(y|x).$$

$E(Y|x)$ is a function of x . Hence, $E(Y|X)$ is a random variable (a function of X) whose value is $E(Y|x)$ when $X = x$.

More generally, the conditional expectation of a function $h(Y)$ is

$$E(h(Y)|X = x) = \int h(y) f_{Y|X}(y|x) dy.$$

EX A An insect lays a large number of eggs, each surviving with probability p . On the average, how many eggs will survive?

The large number of eggs laid is a random variable, often taken to be $\text{Poisson}(\lambda)$. Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials. Therefore, if we let X =number of survivors and Y =number of eggs laid, we have

$$X|Y \text{ binomial}(Y, p), \quad Y \sim \text{Poisson}(\lambda).$$

EX B (Continuation of Example A) The random variable X has the distribution given by

$$\begin{aligned}
 P(X = x) &= \sum_{y=0}^{\infty} P(X = x, Y = y) = \sum_{y=0}^{\infty} P(X = x|Y = y)P(Y = y) \\
 &= \sum_{y=x}^{\infty} \left[\binom{y}{x} p^x (1-p)^{y-x} \right] \left[\frac{e^{-y} \lambda^y}{y!} \right] \\
 &\quad \text{(conditional probability is 0 if } y < x \text{)} \\
 &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\
 &= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} \\
 &= \frac{(\lambda p)^x}{x!} e^{-\lambda p},
 \end{aligned}$$

so $X \sim \text{Poisson}(\lambda p)$. Thus, any marginal inference on X is with respect to a $\text{Poisson}(\lambda p)$ distribution, with Y playing no part at all. Introducing Y in the

hierarchy was mainly to aid our understanding of the model. On the average,

$$EX = \lambda p$$

eggs will survive.

EX C If X and Y follow a bivariate normal distribution, the conditional density of Y given X is

$$f_{Y|X}(y|x) = \frac{1}{\sigma_Y \sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{1}{2} \frac{[y - \mu_Y - \rho\sigma_Y/\sigma_X(x - \mu_X)]^2}{\sigma_Y^2(1-\rho^2)}\right).$$

This is normal density with mean $\mu_Y + \rho(x - \mu_X)\sigma_Y/\sigma_X$ and variance $\sigma_Y^2(1 - \rho^2)$.

Theorem 4.1 *Let X and Y be random variables such that Y has finite mean. Then*

$$E[E(Y|X)] = E(Y).$$

PROOF: We shall assume, for convenience, that X and Y have a continuous joint

distribution. Then

$$\begin{aligned} E[E(Y|X)] &= \int_{-\infty}^{\infty} E(Y|x) f_X(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dy dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy \\ &= E(Y). \end{aligned}$$

The proof for a discrete distribution is similar. \square

Using Theorem 4.1, we have

$$EX = E(E(X|Y)) = E(pY) = p\lambda$$

for Example B.

EX D Suppose that a point X is chosen in accordance with a uniform distribution on the interval $[0, 1]$. Also, suppose that after the value $X = x$ has been observed ($0 < x < 1$), a point Y is chosen in accordance with a uniform distribution on the interval $[x, 1]$. We shall determine the value of Y .

$$E(Y) = E[E(Y|X)] = E\left[\frac{1}{2}(X + 1)\right] = \frac{1}{2}\left(\frac{1}{2} + 1\right) = \frac{3}{4}.$$

Theorem 4.1 implies that for two arbitrary random variables X and Y ,

$$E\{E[r(X, Y)|X]\} = E[r(X, Y)],$$

by letting $Z = r(X, Y)$ and noting that $E\{E(Z|X)\} = E(Z)$.

EX E Suppose that $E(Y|X) = aX + b$ for some constants a and b . We shall determine the value of $E(XY)$ in terms of $E(X)$ and $E(X^2)$.

$$E(XY|X) = XE(Y|X) = X(aX + b) = aX^2 + bX.$$

Therefore,

$$E(XY) = E[E(XY|X)] = E(aX^2 + bX) = aE(X^2) + bE(X).$$

Theorem 4.2 (*Conditional variance identity*) For any two random variables X and Y ,

$$\text{Var}X = E(\text{Var}(X|Y)) + \text{Var}(E(X|Y)),$$

provided that the expectations exist.

PROOF: By definition, we have

$$\begin{aligned}\text{Var}X &= E([X - EX]^2) = E([X - E(X|Y) + E(X|Y) - EX]^2) \\ &= E([X - E(X|Y)]^2) + E([E(X|Y) - EX]^2) \\ &\quad + 2E([X - E(X|Y)][E(X|Y) - EX]).\end{aligned}$$

The last term in this expression is equal to 0, however, which can easily be seen by iterating the expectation:

$$E([X - E(X|Y)][E(X|Y) - EX]) = E(E\{[X - E(X|Y)][E(X|Y) - EX]|Y\})$$

In the conditional distribution $X|Y$, X is the random variable. Conditional on Y , $E(X|Y)$ and EX are constants. Thus,

$$E\{[X - E(X|Y)][E(X|Y) - EX]|Y\} = (E(X|Y) - E(X|Y))(E(X|Y) - EX) = 0$$

Since

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

and

$$E([E(X|Y) - EX]^2) = \text{Var}(E(X|Y)),$$

Theorem 4.2 is proved. \square

EX F (Beta-binomial hierarchy) One generalization of the binomial distribution is to allow the success probability to vary according to a distribution. A standard model for this situation is

$$\begin{aligned}X|P &\sim \text{binomial}(P), \quad i = 1, \dots, n, \\P &\sim \text{beta}(\alpha, \beta).\end{aligned}$$

The mean of X is then

$$EX = E[E(X|p)] = E[nP] = \frac{n\alpha}{\alpha + \beta}.$$

Since $P \sim \text{beta}(\alpha, \beta)$,

$$\text{Var}(E(X|P)) = \text{Var}(nP) = n^2 \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Also, since $X|P$ is binomial(n, P), $\text{Var}(X|P) = nP(1 - P)$. We then have

$$\begin{aligned} E[\text{Var}(X|P)] &= nE[P(1 - P)] \\ &= n \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p(1 - p)p^{\alpha-1}(1 - p)^{\beta-1} dp \\ &= n \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + 1)\Gamma(\beta + 1)}{\Gamma(\alpha + \beta + 2)} \\ &= \frac{n\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)}. \end{aligned}$$

Adding together the two pieces, we get

$$\text{Var}X = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

4.1 Prediction

Consider predicting Y by some function $h(X)$ in order to minimize $MSE = E[Y - h(X)]^2$. Theorem 4.1 implies that

$$E[Y - h(X)]^2 = E(E\{[Y - h(X)]^2 | X\}).$$

The outer expectation is with respect to X . For every x , the inner expectation is minimized by setting $h(x)$ equal to the constant $E(Y|X = x)$. We thus have that the minimizing function $h(X)$ is

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

EX A For the bivariate normal distribution, we found that

$$E(Y|X) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (X - \mu_X).$$

This linear function of X is thus the minimum mean squared error predictor of Y from X .

A practical limitation of the optimal prediction scheme is that its implementation depends on knowing the joint distribution of Y and X in order to find $E(Y|X)$, and this information is not available, not even approximately. For this reason, we can try to attain

the more modest goal of finding the optimal linear predictor of Y . Let $h(x) = \alpha + \beta x$.
Now,

$$\begin{aligned} E[(Y - \alpha - \beta X)^2] &= \text{Var}(Y - \alpha - \beta X) + [E(Y - \alpha - \beta X)]^2 \\ &= \text{Var}(Y - \beta X) + [E(Y - \alpha - \beta X)]^2 \end{aligned}$$

The first term does not depend on α , so α can be chosen so as to minimize the second term. Hence, $\alpha = E(Y - \beta X) = \mu_Y - \beta\mu_X$.

As for the first term,

$$\text{Var}(Y - \beta X) = \sigma_Y^2 + \beta^2 \sigma_X^2 - 2\beta \sigma_{XY},$$

where $\sigma_{XY} = \text{Cov}(X, Y)$. This is a quadratic function of β , and it is minimized at

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X},$$

where ρ is the correlation coefficient. In summary, we have

$$\hat{Y} = \alpha + \beta X = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2} (X - \mu_X).$$

The mean squared prediction error is then

$$\begin{aligned}\text{Var}(Y - \beta X) &= \sigma_Y^2 + \frac{\sigma_{XY}^2}{\sigma_X^4} \sigma_X^2 - 2 \frac{\sigma_{XY}}{\sigma_X^2} \sigma_{XY} \\ &= \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} \\ &= \sigma_Y^2 - \rho^2 \sigma_Y^2 = \sigma_Y^2 (1 - \rho^2).\end{aligned}$$

5 The Moment-Generating Function

We shall now consider a given random variable X ; and for each real number t , we shall let

$$\psi(t) = E(e^{tX}).$$

The function ψ is called the moment generating function (mgf) of X .

If x is bounded, then the mgf of X will exist for all values of t . On the other hand, if X is not bounded, then the mgf might exist for some values of t and might not exist for others. However, for every random variable X , the mgf $\psi(t)$ must exist at point $t = 0$.

Proposition 5.1 *If the m.g.f. exists for t in an open interval containing zero, it uniquely determines the probability distribution.*

Suppose that the mgf of a random variable X exists for all values of t in some open interval around the point $t = 0$. It can be shown that the derivative $\psi'(t)$ then exists at the point $t = 0$; and that at $t = 0$, the derivative of the expectation must be equal to the expectation of the derivative. (The proof is beyond the scope of this course.) Thus,

$$\begin{aligned}\psi'(0) &= \left[\frac{d}{dt} E(e^{tX}) \right]_{t=0} = E \left[\left(\frac{d}{dt} e^{tX} \right)_{t=0} \right] \\ &= E \left[(X e^{tX})_{t=0} \right] = E(X).\end{aligned}$$

In other words, the derivative of the mgf $\psi(t)$ at $t = 0$ is the mean of X .

More generally, if the mgf $\psi(t)$ of X exists for all values of t in an open interval around the point $t = 0$, then it can be shown that all moments $E(X^k)$ of X must exist ($k = 1, 2, \dots$). Furthermore, it can be shown that it is possible to differentiate $\psi(t)$ an arbitrary number of times at the point $t = 0$. For $n = 1, 2, \dots$, the n th derivative $\psi^{(n)}(0)$ at $t = 0$ will satisfy the following relation:

$$\psi^{(n)}(0) = \left[\frac{d^n}{dt^n} E(e^{tX}) \right]_{t=0} = E \left[\left(\frac{d^n}{dt^n} e^{tX} \right)_{t=0} \right] = E(X^n).$$

Thus, $\psi'(0) = E(X)$, $\psi''(0) = E(X^2)$, $\psi'''(0) = E(X^3)$, and so on.

Proposition 5.2 *If the m.g.f. exists in an open interval containing zero, then $\psi^{(r)}(0) = E(X^r)$.*

EX A (Poisson distribution)

$$\begin{aligned}\psi(t) &= \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} e^{-\lambda} \\ &= e^{\lambda(e^t - 1)}\end{aligned}$$

The sum converges for all t . Differentiating, we have

$$\begin{aligned}\psi'(t) &= \lambda e^t e^{\lambda(e^t - 1)} \\ \psi''(t) &= \lambda e^t e^{\lambda(e^t - 1)} + \lambda^2 e^{2t} e^{\lambda(e^t - 1)}\end{aligned}$$

Evaluating these derivatives at $t = 0$, we find

$$E(X) = \lambda, \quad E(X^2) = \lambda^2 + \lambda.$$

Thus, $\text{Var}(X) = \lambda$.

EX B (Gamma Distribution) The mgf of a gamma distribution is

$$\psi(t) = \int_0^{\infty} e^{tx} \frac{\lambda^{\alpha}}{\Gamma(x)} x^{\alpha-1} e^{-\lambda x} dx = \left(\frac{\lambda}{\lambda - t} \right)^{\alpha}$$

for $t < \lambda$. Differentiating, we find

$$\psi'(0) = E(X) = \frac{\alpha}{\lambda}$$

$$\psi''(0) = E(X^2) = \frac{\alpha(\alpha + 1)}{\lambda^2}$$

From these equations, we find that

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{\alpha}{\lambda^2}.$$

EX C (Binomial mgf) The binomial mgf is

$$\begin{aligned} \psi(t) &= \sum_{x=0}^n e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^n (pe^t)^x (1-p)^{n-x} \end{aligned}$$

The binomial formula gives

$$\sum_{x=0}^n \binom{n}{x} u^x v^{n-x} = (u + v)^n.$$

Hence, letting $u = pe^t$ and $v = 1 - p$, we have

$$\psi(t) = [pe^t + (1 - p)]^n.$$

EX D (standard normal) For this distribution, we have

$$\psi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx - x^2/2} dx = e^{t^2/2},$$

for all t . From this result, we have $E(X) = 0$ and $\text{Var}(X) = 1$.

Theorem 5.1 *Let X be a random variable for which the mgf is ψ_X ; let $Y = aX + b$, where a and b are given constants; and let ψ_Y denote the mgf of Y . Then for every value of t such that $\psi_X(at)$ exists,*

$$\psi_Y(t) = e^{bt} \psi_X(at).$$

EX E If $Y \sim N(\mu, \sigma^2)$, then

$$\psi_Y(t) = e^{\mu t} \psi_X(\sigma t) = e^{\mu t + \sigma^2 t^2 / 2}.$$

Theorem 5.2 Suppose that X_1, \dots, X_n are n independent random variables; and for $i = 1, \dots, n$, let ψ_i denote the mgf of X_i . Let $Y = X_1 + \dots + X_n$, and let the mgf of Y be denoted by ψ . Then for every value of t such that $\psi_i(t)$ exists for $i = 1, \dots, n$,

$$\psi(t) = \prod_{i=1}^n \psi_i(t).$$

EX F The sum of independent Poisson random variables is a Poisson random variable: If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, then $X + Y \sim \text{Poisson}(\lambda + \mu)$, since

$$e^{\lambda(e^t-1)} e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}.$$

EX G If $X \sim N(\mu, \sigma^2)$ and, independent of X , $Y \sim N(\nu, \tau^2)$, then the mgf of $X + Y$ is

$$e^{\mu t + t^2 \sigma^2 / 2} e^{\nu t + t^2 \tau^2 / 2} = e^{(\mu + \nu)t + (\sigma^2 + \tau^2)t^2 / 2},$$

which is a mgf of a normal distribution with mean $\mu + \nu$ and variance $\sigma^2 + \tau^2$.

6 Approximate Methods

Suppose that we know the expectation and the variance of a random variable X but not the entire distribution, and that we are interested in the mean and variance of $Y = g(X)$ for some fixed function g . From the results given in this chapter, we can not in general find $E(Y)$ and $\text{Var}(Y)$ from $E(X)$ and $\text{Var}(X)$, unless the function g is linear. However, if g is nearly linear in a range in which X has high probability, it can be approximated by a linear function and approximate moments of Y can be found.

When confronted with a nonlinear problem that we can not solve, we linearize. In Statistics, this method is called **propagation of error**, or the **δ -method**. Applying Taylor expansion to g , we have

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X).$$

Thus,

$$\mu_Y \approx g(\mu_X), \quad \sigma_Y^2 \approx \sigma_X^2 [g'(\mu_X)]^2.$$

We can carry out the Taylor expansion to the second order to get an improved approximation of μ_Y :

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 g''(\mu_X),$$

which implies that

$$E(Y) \approx g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X).$$

EX A This example examines the accuracy of the approximations using a simple test case. We choose the function $g(x) = \sqrt{x}$ and consider two cases: X uniform on $[0,1]$, and X uniform on $[1, 2]$. The graph (Figure 4) shows that g is more nearly linear in the latter case, so we would expect the approximations to work better there.

Let $Y = \sqrt{X}$ and $X \sim \text{Unif}[0, 1]$, then

$$E(Y) = \int_0^1 \sqrt{x} dx = \frac{2}{3}, \quad E(Y^2) = \int_0^1 x dx = \frac{1}{2}.$$

So $\text{Var}(Y) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$ and $\sigma_Y = 0.236$.

Using the approximation method, we first calculate

$$g'(x) = \frac{1}{2}x^{-1/2}, \quad g''(x) = -\frac{1}{4}x^{-3/2}, \quad \mu_X = 1/2,$$

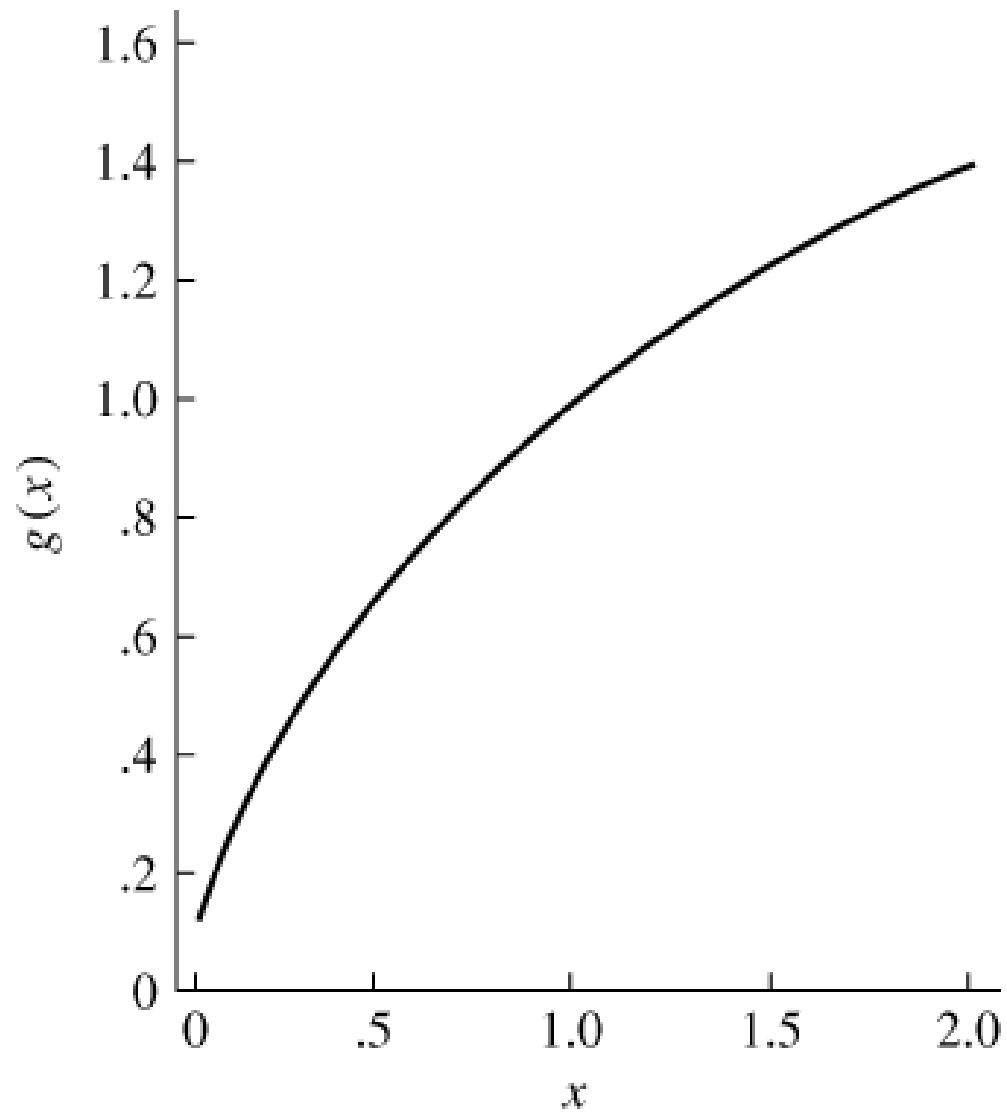


Figure 4: The function $g(x) = \sqrt{x}$.

which leads to that

$$E(Y) \approx \sqrt{\frac{1}{2}} - \frac{1}{2} \frac{\sqrt{2}}{2} \frac{1}{12} = 0.678,$$
$$\text{Var}(Y) \approx \frac{1}{2} \frac{1}{12} = 0.042,$$
$$\sigma_Y \approx 0.204.$$

Consider the case in which X is uniform on $[1, 2]$. We find that $y = \sqrt{x}$ has mean 1.219, the variance 0.0142, the standard deviation 0.119, $\mu_X = 1.5$, $\sigma_X^2 = 1/12$, $g'(\mu_X) = 0.408$, and $g''(\mu_X) = -0.136$. So the approximations are

$$E(Y) \approx \sqrt{\frac{3}{2}} - \frac{1}{2} \frac{0.136}{12} = 1.219,$$
$$\text{Var}(Y) \approx \frac{0.408^2}{12} = 0.0138,$$
$$\sigma_Y \approx 0.118.$$

These values are much closer to the actual values than are the approximations for the first case.

Suppose that we have $Z = g(X, Y)$, a function of two random variables. Let $\mu = (\mu_X, \mu_Y)$, we have

$$Z = g(X, Y) \approx g(\mu) + (X - \mu_X) \frac{\partial g(\mu)}{\partial x} + (Y - \mu_Y) \frac{\partial g(\mu)}{\partial y},$$

which implies that

$$E(Z) \approx g(\mu)$$

and

$$\text{Var}(Z) \approx \sigma_X^2 \left(\frac{\partial g(\mu)}{\partial x} \right)^2 + \sigma_Y^2 \left(\frac{\partial g(\mu)}{\partial y} \right)^2 + 2\sigma_{XY} \left(\frac{\partial g(\mu)}{\partial x} \right) \left(\frac{\partial g(\mu)}{\partial y} \right).$$

For the second order approximation, we have

$$E(Z) \approx g(\mu) + \frac{1}{2} \sigma_X^2 \frac{\partial^2 g(\mu)}{\partial x^2} + \frac{1}{2} \sigma_Y^2 \frac{\partial^2 g(\mu)}{\partial y^2} + \sigma_{XY} \frac{\partial^2 g(\mu)}{\partial x \partial y}.$$

EX B Consider the case $Z = Y/X$.

For $g(x, y) = y/x$, we have

$$\begin{aligned}\frac{\partial g}{\partial x} &= \frac{-y}{x}, & \frac{\partial g}{\partial y} &= \frac{1}{x} \\ \frac{\partial^2 g}{\partial x^2} &= \frac{2y}{x^3}, & \frac{\partial^2 g}{\partial y^2} &= 0 \\ \frac{\partial^2 g}{\partial x \partial y} &= \frac{-1}{x^2}.\end{aligned}$$

Using the preceding results, we find, if $\mu_X \neq 0$,

$$E(Z) \approx \frac{\mu_Y}{\mu_X} + \sigma_X^2 \frac{\mu_Y}{\mu_X^3} - \frac{\sigma_{XY}}{\mu_X^2}.$$

and

$$\text{Var}(Z) \approx \sigma_X^2 \frac{\mu_Y^2}{\mu_X^4} + \frac{\sigma_Y^2}{\mu_X^2} - 2\sigma_{XY} \frac{\mu_Y}{\mu_X^3}.$$