Chapter 8

Law of Large Numbers

8.1 Law of Large Numbers for Discrete Random Variables

We are now in a position to prove our first fundamental theorem of probability. We have seen that an intuitive way to view the probability of a certain outcome is as the frequency with which that outcome occurs in the long run, when the experiment is repeated a large number of times. We have also defined probability mathematically as a value of a distribution function for the random variable representing the experiment. The Law of Large Numbers, which is a theorem proved about the mathematical model of probability, shows that this model is consistent with the frequency interpretation of probability. This theorem is sometimes called the *law of averages.* To find out what would happen if this law were not true, see the article by Robert M. Coates.[1]

Chebyshev Inequality

To discuss the Law of Large Numbers, we first need an important inequality called the *Chebyshev Inequality.*

Theorem 8.1 (Chebyshev Inequality) Let X be a discrete random variable with expected value $\mu = E(X)$, and let $\epsilon > 0$ be any positive real number. Then

$$P(|X - \mu| \geq \epsilon) \leq \frac{V(X)}{\epsilon^2} \ .$$

Proof. Let $m(x)$ denote the distribution function of X. Then the probability that X differs from μ by at least ϵ is given by

$$P(|X - \mu| \geq \epsilon) = \sum_{|x-\mu| \geq \epsilon} m(x) \ .$$

[1] R. M. Coates, "The Law," *The World of Mathematics,* ed. James R. Newman (New York: Simon and Schuster, 1956.

We know that

$$V(X) = \sum_x (x-\mu)^2 m(x) \ ,$$

and this is clearly at least as large as

$$\sum_{|x-\mu|\geq\epsilon} (x-\mu)^2 m(x) \ ,$$

since all the summands are positive and we have restricted the range of summation in the second sum. But this last sum is at least

$$\begin{aligned} \sum_{|x-\mu|\geq\epsilon} \epsilon^2 m(x) &= \epsilon^2 \sum_{|x-\mu|\geq\epsilon} m(x) \\ &= \epsilon^2 P(|X-\mu|\geq\epsilon) \ . \end{aligned}$$

So,

$$P(|X-\mu|\geq\epsilon) \leq \frac{V(X)}{\epsilon^2} \ .$$

□

Note that X in the above theorem can be any discrete random variable, and ϵ any positive number.

Example 8.1 Let X by any random variable with $E(X) = \mu$ and $V(X) = \sigma^2$. Then, if $\epsilon = k\sigma$, Chebyshev's Inequality states that

$$P(|X-\mu|\geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \ .$$

Thus, for any random variable, the probability of a deviation from the mean of more than k standard deviations is $\leq 1/k^2$. If, for example, $k = 5$, $1/k^2 = .04$. □

Chebyshev's Inequality is the best possible inequality in the sense that, for any $\epsilon > 0$, it is possible to give an example of a random variable for which Chebyshev's Inequality is in fact an equality. To see this, given $\epsilon > 0$, choose X with distribution

$$p_X = \begin{pmatrix} -\epsilon & +\epsilon \\ 1/2 & 1/2 \end{pmatrix} \ .$$

Then $E(X) = 0$, $V(X) = \epsilon^2$, and

$$P(|X-\mu|\geq\epsilon) = \frac{V(X)}{\epsilon^2} = 1 \ .$$

We are now prepared to state and prove the Law of Large Numbers.

Law of Large Numbers

Theorem 8.2 (Law of Large Numbers) Let X_1, X_2, ..., X_n be an independent trials process, with finite expected value $\mu = E(X_j)$ and finite variance $\sigma^2 = V(X_j)$. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$. Equivalently,

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \to 1$$

as $n \to \infty$.

Proof. Since X_1, X_2, ..., X_n are independent and have the same distributions, we can apply Theorem 6.9. We obtain

$$V(S_n) = n\sigma^2 \ ,$$

and

$$V(\frac{S_n}{n}) = \frac{\sigma^2}{n} \ .$$

Also we know that

$$E(\frac{S_n}{n}) = \mu \ .$$

By Chebyshev's Inequality, for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \ .$$

Thus, for fixed ϵ,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$, or equivalently,

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \to 1$$

as $n \to \infty$. □

Law of Averages

Note that S_n/n is an average of the individual outcomes, and one often calls the Law of Large Numbers the "law of averages." It is a striking fact that we can start with a random experiment about which little can be predicted and, by taking averages, obtain an experiment in which the outcome can be predicted with a high degree of certainty. The Law of Large Numbers, as we have stated it, is often called the "Weak Law of Large Numbers" to distinguish it from the "Strong Law of Large Numbers" described in Exercise 15.

Consider the important special case of Bernoulli trials with probability p for success. Let $X_j = 1$ if the jth outcome is a success and 0 if it is a failure. Then $S_n = X_1 + X_2 + \cdots + X_n$ is the number of successes in n trials and $\mu = E(X_1) = p$. The Law of Large Numbers states that for any $\epsilon > 0$

$$P\left(\left|\frac{S_n}{n} - p\right| < \epsilon\right) \to 1$$

as $n \to \infty$. The above statement says that, in a large number of repetitions of a Bernoulli experiment, we can expect the proportion of times the event will occur to be near p. This shows that our mathematical model of probability agrees with our frequency interpretation of probability.

Coin Tossing

Let us consider the special case of tossing a coin n times with S_n the number of heads that turn up. Then the random variable S_n/n represents the fraction of times heads turns up and will have values between 0 and 1. The Law of Large Numbers predicts that the outcomes for this random variable will, for large n, be near $1/2$.

In Figure 8.1, we have plotted the distribution for this example for increasing values of n. We have marked the outcomes between .45 and .55 by dots at the top of the spikes. We see that as n increases the distribution gets more and more concentrated around .5 and a larger and larger percentage of the total area is contained within the interval $(.45, .55)$, as predicted by the Law of Large Numbers.

Die Rolling

Example 8.2 Consider n rolls of a die. Let X_j be the outcome of the jth roll. Then $S_n = X_1 + X_2 + \cdots + X_n$ is the sum of the first n rolls. This is an independent trials process with $E(X_j) = 7/2$. Thus, by the Law of Large Numbers, for any $\epsilon > 0$

$$P\left(\left|\frac{S_n}{n} - \frac{7}{2}\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$. An equivalent way to state this is that, for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \frac{7}{2}\right| < \epsilon\right) \to 1$$

as $n \to \infty$. □

Numerical Comparisons

It should be emphasized that, although Chebyshev's Inequality proves the Law of Large Numbers, it is actually a very crude inequality for the probabilities involved. However, its strength lies in the fact that it is true for any random variable at all, and it allows us to prove a very powerful theorem.

In the following example, we compare the estimates given by Chebyshev's Inequality with the actual values.

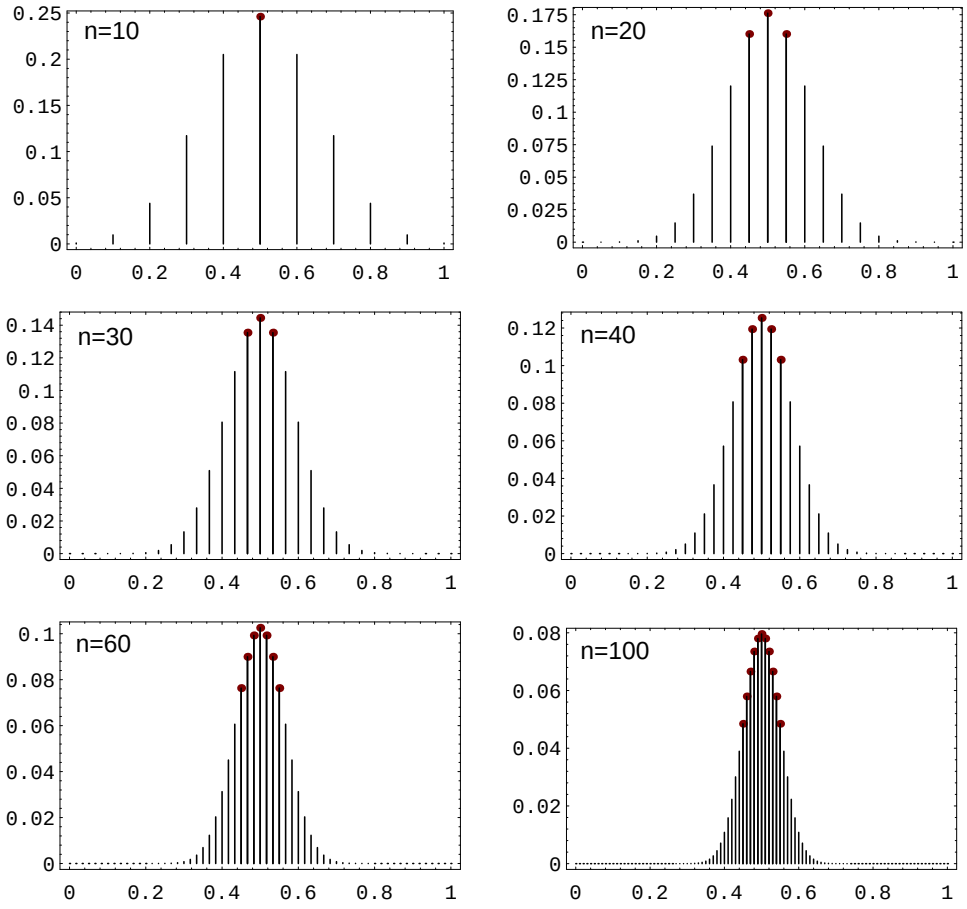


Figure 8.1: Bernoulli trials distributions.

Example 8.3 Let $X_1, X_2, \ldots, X_n$ be a Bernoulli trials process with probability .3 for success and .7 for failure. Let $X_j = 1$ if the jth outcome is a success and 0 otherwise. Then, $E(X_j) = .3$ and $V(X_j) = (.3)(.7) = .21$. If

$$A_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is the *average* of the X_i, then $E(A_n) = .3$ and $V(A_n) = V(S_n)/n^2 = .21/n$. Chebyshev's Inequality states that if, for example, $\epsilon = .1$,

$$P(|A_n - .3| \geq .1) \leq \frac{.21}{n(.1)^2} = \frac{21}{n} \ .$$

Thus, if $n = 100$,

$$P(|A_{100} - .3| \geq .1) \leq .21 \ ,$$

or if $n = 1000$,

$$P(|A_{1000} - .3| \geq .1) \leq .021 \ .$$

These can be rewritten as

$$\begin{aligned} P(.2 < A_{100} < .4) &\geq .79 \ , \\ P(.2 < A_{1000} < .4) &\geq .979 \ . \end{aligned}$$

These values should be compared with the actual values, which are (to six decimal places)

$$\begin{aligned} P(.2 < A_{100} < .4) &\approx .962549 \\ P(.2 < A_{1000} < .4) &\approx 1 \ . \end{aligned}$$

The program **Law** can be used to carry out the above calculations in a systematic way. □

Historical Remarks

The Law of Large Numbers was first proved by the Swiss mathematician James Bernoulli in the fourth part of his work *Ars Conjectandi* published posthumously in 1713.[2] As often happens with a first proof, Bernoulli's proof was much more difficult than the proof we have presented using Chebyshev's inequality. Chebyshev developed his inequality to prove a general form of the Law of Large Numbers (see Exercise 12). The inequality itself appeared much earlier in a work by Bienaymé, and in discussing its history Maistrov remarks that it was referred to as the Bienaymé-Chebyshev Inequality for a long time.[3]

In *Ars Conjectandi* Bernoulli provides his reader with a long discussion of the meaning of his theorem with lots of examples. In modern notation he has an event

[2] J. Bernoulli, *The Art of Conjecturing IV,* trans. Bing Sung, Technical Report No. 2, Dept. of Statistics, Harvard Univ., 1966

[3] L. E. Maistrov, *Probability Theory: A Historical Approach,* trans. and ed. Samual Kotz, (New York: Academic Press, 1974), p. 202

that occurs with probability p but he does not know p. He wants to estimate p by the fraction $\bar{p}$ of the times the event occurs when the experiment is repeated a number of times. He discusses in detail the problem of estimating, by this method, the proportion of white balls in an urn that contains an unknown number of white and black balls. He would do this by drawing a sequence of balls from the urn, replacing the ball drawn after each draw, and estimating the unknown proportion of white balls in the urn by the proportion of the balls drawn that are white. He shows that, by choosing n large enough he can obtain any desired accuracy and reliability for the estimate. He also provides a lively discussion of the applicability of his theorem to estimating the probability of dying of a particular disease, of different kinds of weather occurring, and so forth.

In speaking of the number of trials necessary for making a judgement, Bernoulli observes that the "man on the street" believes the "law of averages."

> Further, it cannot escape anyone that for judging in this way about any event at all, it is not enough to use one or two trials, but rather a great number of trials is required. And sometimes the stupidest man—by some instinct of nature *per se* and by no previous instruction (this is truly amazing)— knows for sure that the more observations of this sort that are taken, the less the danger will be of straying from the mark.[4]

But he goes on to say that he must contemplate another possibility.

> Something futher must be contemplated here which perhaps no one has thought about till now. It certainly remains to be inquired whether after the number of observations has been increased, the probability is increased of attaining the true ratio between the number of cases in which some event can happen and in which it cannot happen, so that this probability finally exceeds any given degree of certainty; or whether the problem has, so to speak, its own asymptote—that is, whether some degree of certainty is given which one can never exceed.[5]

Bernoulli recognized the importance of this theorem, writing:

> Therefore, this is the problem which I now set forth and make known after I have already pondered over it for twenty years. Both its novelty and its very great usefullness, coupled with its just as great difficulty, can exceed in weight and value all the remaining chapters of this thesis.[6]

Bernoulli concludes his long proof with the remark:

> Whence, finally, this one thing seems to follow: that if observations of all events were to be continued throughout all eternity, (and hence the ultimate probability would tend toward perfect certainty), everything in

[4]Bernoulli, op. cit., p. 38.
[5]ibid., p. 39.
[6]ibid., p. 42.

> the world would be perceived to happen in fixed ratios and according to a constant law of alternation, so that even in the most accidental and fortuitous occurrences we would be bound to recognize, as it were, a certain necessity and, so to speak, a certain fate.
>
> I do now know whether Plato wished to aim at this in his doctrine of the universal return of things, according to which he predicted that all things will return to their original state after countless ages have past.[7]

Exercises

1 A fair coin is tossed 100 times. The expected number of heads is 50, and the standard deviation for the number of heads is $(100 \cdot 1/2 \cdot 1/2)^{1/2} = 5$. What does Chebyshev's Inequality tell you about the probability that the number of heads that turn up deviates from the expected number 50 by three or more standard deviations (i.e., by at least 15)?

2 Write a program that uses the function binomial(n, p, x) to compute the exact probability that you estimated in Exercise 1. Compare the two results.

3 Write a program to toss a coin 10,000 times. Let S_n be the number of heads in the first n tosses. Have your program print out, after every 1000 tosses, $S_n - n/2$. On the basis of this simulation, is it correct to say that you can expect heads about half of the time when you toss a coin a large number of times?

4 A 1-dollar bet on craps has an expected winning of $-.0141$. What does the Law of Large Numbers say about your winnings if you make a large number of 1-dollar bets at the craps table? Does it assure you that your losses will be small? Does it assure you that if n is very large you will lose?

5 Let X be a random variable with $E(X) = 0$ and $V(X) = 1$. What integer value k will assure us that $P(|X| \geq k) \leq .01$?

6 Let S_n be the number of successes in n Bernoulli trials with probability p for success on each trial. Show, using Chebyshev's Inequality, that for any $\epsilon > 0$

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2} \ .$$

7 Find the maximum possible value for $p(1-p)$ if $0 < p < 1$. Using this result and Exercise 6, show that the estimate

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2}$$

is valid for any p.

[7] ibid., pp. 65–66.

8 A fair coin is tossed a large number of times. Does the Law of Large Numbers assure us that, if n is large enough, with probability $> .99$ the number of heads that turn up will not deviate from $n/2$ by more than 100?

9 In Exercise 6.2.15, you showed that, for the hat check problem, the number S_n of people who get their own hats back has $E(S_n) = V(S_n) = 1$. Using Chebyshev's Inequality, show that $P(S_n \geq 11) \leq .01$ for any $n \geq 11$.

10 Let X by any random variable which takes on values 0, 1, 2, ..., n and has $E(X) = V(X) = 1$. Show that, for any positive integer k,

$$P(X \geq k+1) \leq \frac{1}{k^2} \ .$$

11 We have two coins: one is a fair coin and the other is a coin that produces heads with probability 3/4. One of the two coins is picked at random, and this coin is tossed n times. Let S_n be the number of heads that turns up in these n tosses. Does the Law of Large Numbers allow us to predict the proportion of heads that will turn up in the long run? After we have observed a large number of tosses, can we tell which coin was chosen? How many tosses suffice to make us 95 percent sure?

12 (Chebyshev[8]) Assume that $X_1, X_2, \ldots, X_n$ are independent random variables with possibly different distributions and let S_n be their sum. Let $m_k = E(X_k)$, $\sigma_k^2 = V(X_k)$, and $M_n = m_1 + m_2 + \cdots + m_n$. Assume that $\sigma_k^2 < R$ for all k. Prove that, for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \frac{M_n}{n}\right| < \epsilon\right) \to 1$$

as $n \to \infty$.

13 A fair coin is tossed repeatedly. Before each toss, you are allowed to decide whether to bet on the outcome. Can you describe a betting system with infinitely many bets which will enable you, in the long run, to win more than half of your bets? (Note that we are disallowing a betting system that says to bet until you are ahead, then quit.) Write a computer program that implements this betting system. As stated above, your program must decide whether to bet on a particular outcome before that outcome is determined. For example, you might select only outcomes that come after there have been three tails in a row. See if you can get more than 50% heads by your "system."

***14** Prove the following analogue of Chebyshev's Inequality:

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

[8] P. L. Chebyshev, "On Mean Values," *J. Math. Pure. Appl.,* vol. 12 (1867), pp. 177–184.

***15** We have proved a theorem often called the "Weak Law of Large Numbers." Most people's intuition and our computer simulations suggest that, if we toss a coin a sequence of times, the proportion of heads will really approach 1/2; that is, if S_n is the number of heads in n times, then we will have

$$A_n = \frac{S_n}{n} \to \frac{1}{2}$$

as $n \to \infty$. Of course, we cannot be sure of this since we are not able to toss the coin an infinite number of times, and, if we could, the coin could come up heads every time. However, the "Strong Law of Large Numbers," proved in more advanced courses, states that

$$P\left(\frac{S_n}{n} \to \frac{1}{2}\right) = 1 \ .$$

Describe a sample space Ω that would make it possible for us to talk about the event

$$E = \left\{ \omega : \frac{S_n}{n} \to \frac{1}{2} \right\} \ .$$

Could we assign the equiprobable measure to this space? (See Example 2.18.)

***16** In this exercise, we shall construct an example of a sequence of random variables that satisfies the weak law of large numbers, but not the strong law. The distribution of X_i will have to depend on i, because otherwise both laws would be satisfied. (This problem was communicated to us by David Maslen.)

Suppose we have an infinite sequence of mutually independent events $A_1, A_2, \ldots$. Let $a_i = P(A_i)$, and let r be a positive integer.

(a) Find an expression of the probability that none of the A_i with $i > r$ occur.

(b) Use the fact that $x - 1 \leq e^{-x}$ to show that

$$P(\text{No } A_i \text{ with } i > r \text{ occurs}) \leq e^{-\sum_{i=r}^{\infty} a_i}$$

(c) (The first Borel-Cantelli lemma) Prove that if $\sum_{i=1}^{\infty} a_i$ diverges, then

$$P(\text{infinitely many } A_i \text{ occur}) = 1.$$

Now, let X_i be a sequence of mutually independent random variables such that for each positive integer $i \geq 2$,

$$P(X_i = i) = \frac{1}{2i \log i}, \quad P(X_i = -i) = \frac{1}{2i \log i}, \quad P(X_i = 0) = 1 - \frac{1}{i \log i}.$$

When $i = 1$ we let $X_i = 0$ with probability 1. As usual we let $S_n = X_1 + \cdots + X_n$. Note that the mean of each X_i is 0.

(d) Find the variance of S_n.

(e) Show that the sequence $\langle X_i \rangle$ satisfies the Weak Law of Large Numbers, i.e. prove that for any $\epsilon > 0$

$$P\left(\left|\frac{S_n}{n}\right| \geq \epsilon\right) \to 0 \ ,$$

as n tends to infinity.

We now show that $\{X_i\}$ does not satisfy the Strong Law of Large Numbers. Suppose that $S_n/n \to 0$. Then because

$$\frac{X_n}{n} = \frac{S_n}{n} - \frac{n-1}{n}\frac{S_{n-1}}{n-1} \ ,$$

we know that $X_n/n \to 0$. From the definition of limits, we conclude that the inequality $|X_i| \geq \frac{1}{2}i$ can only be true for finitely many i.

(f) Let A_i be the event $|X_i| \geq \frac{1}{2}i$. Find $P(A_i)$. Show that $\sum_{i=1}^{\infty} P(A_i)$ diverges (use the Integral Test).

(g) Prove that A_i occurs for infinitely many i.

(h) Prove that

$$P\left(\frac{S_n}{n} \to 0\right) = 0,$$

and hence that the Strong Law of Large Numbers fails for the sequence $\{X_i\}$.

***17** Let us toss a biased coin that comes up heads with probability p and assume the validity of the Strong Law of Large Numbers as described in Exercise 15. Then, with probability 1,

$$\frac{S_n}{n} \to p$$

as $n \to \infty$. If $f(x)$ is a continuous function on the unit interval, then we also have

$$f\left(\frac{S_n}{n}\right) \to f(p) \ .$$

Finally, we could hope that

$$E\left(f\left(\frac{S_n}{n}\right)\right) \to E(f(p)) = f(p) \ .$$

Show that, if all this is correct, as in fact it is, we would have proven that any continuous function on the unit interval is a limit of polynomial functions. This is a sketch of a probabilistic proof of an important theorem in mathematics called the *Weierstrass approximation theorem.*

8.2 Law of Large Numbers for Continuous Random Variables

In the previous section we discussed in some detail the Law of Large Numbers for discrete probability distributions. This law has a natural analogue for continuous probability distributions, which we consider somewhat more briefly here.

Chebyshev Inequality

Just as in the discrete case, we begin our discussion with the Chebyshev Inequality.

Theorem 8.3 (Chebyshev Inequality) Let X be a continuous random variable with density function $f(x)$. Suppose X has a finite expected value $\mu = E(X)$ and finite variance $\sigma^2 = V(X)$. Then for any positive number $\epsilon > 0$ we have

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \ .$$

□

The proof is completely analogous to the proof in the discrete case, and we omit it.

Note that this theorem says nothing if $\sigma^2 = V(X)$ is infinite.

Example 8.4 Let X be any continuous random variable with $E(X) = \mu$ and $V(X) = \sigma^2$. Then, if $\epsilon = k\sigma = k$ standard deviations for some integer k, then

$$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \ ,$$

just as in the discrete case. □

Law of Large Numbers

With the Chebyshev Inequality we can now state and prove the Law of Large Numbers for the continuous case.

Theorem 8.4 (Law of Large Numbers) Let X_1, X_2, ..., X_n be an independent trials process with a continuous density function f, finite expected value μ, and finite variance σ^2. Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of the X_i. Then for any real number $\epsilon > 0$ we have

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0 \ ,$$

or equivalently,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1 \ .$$

□

Note that this theorem is not necessarily true if σ^2 is infinite (see Example 8.8).

As in the discrete case, the Law of Large Numbers says that the average value of n independent trials tends to the expected value as $n \to \infty$, in the precise sense that, given $\epsilon > 0$, the probability that the average value and the expected value differ by more than ϵ tends to 0 as $n \to \infty$.

Once again, we suppress the proof, as it is identical to the proof in the discrete case.

Uniform Case

Example 8.5 Suppose we choose at random n numbers from the interval $[0,1]$ with uniform distribution. Then if X_i describes the ith choice, we have

$$\begin{aligned} \mu &= E(X_i) = \int_0^1 x\,dx = \frac{1}{2}\ , \\ \sigma^2 &= V(X_i) = \int_0^1 x^2\,dx - \mu^2 \\ &= \frac{1}{3} - \frac{1}{4} = \frac{1}{12}\ . \end{aligned}$$

Hence,

$$\begin{aligned} E\left(\frac{S_n}{n}\right) &= \frac{1}{2}\ , \\ V\left(\frac{S_n}{n}\right) &= \frac{1}{12n}\ , \end{aligned}$$

and for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \frac{1}{2}\right| \geq \epsilon\right) \leq \frac{1}{12n\epsilon^2}\ .$$

This says that if we choose n numbers at random from $[0,1]$, then the chances are better than $1 - 1/(12n\epsilon^2)$ that the difference $|S_n/n - 1/2|$ is less than ϵ. Note that ϵ plays the role of the amount of error we are willing to tolerate: If we choose $\epsilon = 0.1$, say, then the chances that $|S_n/n - 1/2|$ is less than 0.1 are better than $1 - 100/(12n)$. For $n = 100$, this is about .92, but if $n = 1000$, this is better than .99 and if $n = 10{,}000$, this is better than .999.

We can illustrate what the Law of Large Numbers says for this example graphically. The density for $A_n = S_n/n$ is determined by

$$f_{A_n}(x) = n f_{S_n}(nx)\ .$$

We have seen in Section 7.2, that we can compute the density $f_{S_n}(x)$ for the sum of n uniform random variables. In Figure 8.2 we have used this to plot the density for A_n for various values of n. We have shaded in the area for which A_n would lie between .45 and .55. We see that as we increase n, we obtain more and more of the total area inside the shaded region. The Law of Large Numbers tells us that we can obtain as much of the total area as we please inside the shaded region by choosing n large enough (see also Figure 8.1). □

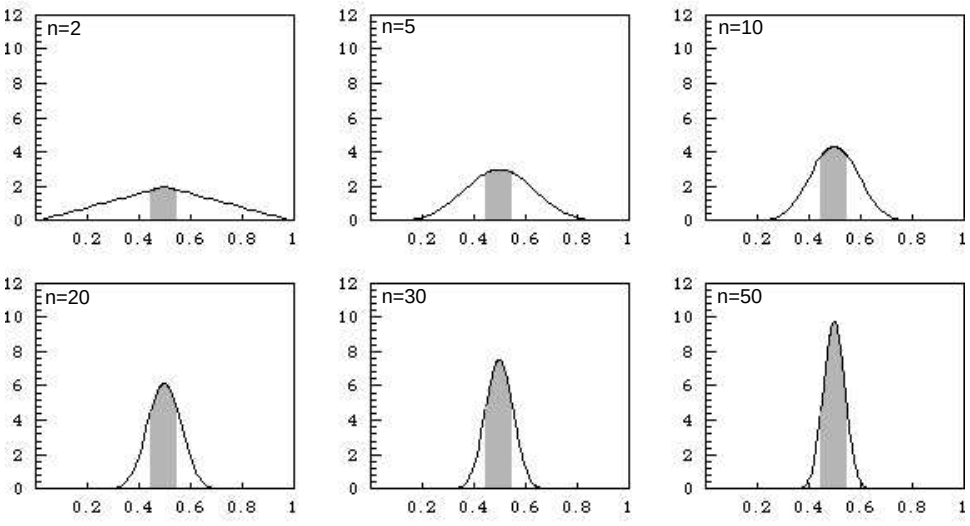


Figure 8.2: Illustration of Law of Large Numbers — uniform case.

Normal Case

Example 8.6 Suppose we choose n real numbers at random, using a normal distribution with mean 0 and variance 1. Then

$$\begin{aligned} \mu &= E(X_i) = 0 \ , \\ \sigma^2 &= V(X_i) = 1 \ . \end{aligned}$$

Hence,

$$\begin{aligned} E\left(\frac{S_n}{n}\right) &= 0 \ , \\ V\left(\frac{S_n}{n}\right) &= \frac{1}{n} \ , \end{aligned}$$

and, for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - 0\right| \geq \epsilon\right) \leq \frac{1}{n\epsilon^2} \ .$$

In this case it is possible to compare the Chebyshev estimate for $P(|S_n/n - \mu| \geq \epsilon)$ in the Law of Large Numbers with exact values, since we know the density function for S_n/n exactly (see Example 7.9). The comparison is shown in Table 8.1, for $\epsilon = .1$. The data in this table was produced by the program **LawContinuous**. We see here that the Chebyshev estimates are in general *not* very accurate. □

n	$P(\lvert S_n/n\rvert \geq .1)$	Chebyshev
100	.31731	1.00000
200	.15730	.50000
300	.08326	.33333
400	.04550	.25000
500	.02535	.20000
600	.01431	.16667
700	.00815	.14286
800	.00468	.12500
900	.00270	.11111
1000	.00157	.10000

Table 8.1: Chebyshev estimates.

Monte Carlo Method

Here is a somewhat more interesting example.

Example 8.7 Let $g(x)$ be a continuous function defined for $x \in [0,1]$ with values in $[0,1]$. In Section 2.1, we showed how to estimate the area of the region under the graph of $g(x)$ by the Monte Carlo method, that is, by choosing a large number of random values for x and y with uniform distribution and seeing what fraction of the points $P(x,y)$ fell inside the region under the graph (see Example 2.2).

Here is a better way to estimate the same area (see Figure 8.3). Let us choose a large number of independent values X_n at random from $[0,1]$ with uniform density, set $Y_n = g(X_n)$, and find the average value of the Y_n. Then this average is our estimate for the area. To see this, note that if the density function for X_n is uniform,

$$\begin{aligned} \mu &= E(Y_n) = \int_0^1 g(x)f(x)\,dx \\ &= \int_0^1 g(x)\,dx \\ &= \text{average value of } g(x)\ , \end{aligned}$$

while the variance is

$$\sigma^2 = E((Y_n - \mu)^2) = \int_0^1 (g(x) - \mu)^2\,dx < 1\ ,$$

since for all x in $[0,1]$, $g(x)$ is in $[0,1]$, hence μ is in $[0,1]$, and so $|g(x) - \mu| \leq 1$. Now let $A_n = (1/n)(Y_1 + Y_2 + \cdots + Y_n)$. Then by Chebyshev's Inequality, we have

$$P(|A_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} < \frac{1}{n\epsilon^2}\ .$$

This says that to get within ϵ of the true value for $\mu = \int_0^1 g(x)\,dx$ with probability at least p, we should choose n so that $1/n\epsilon^2 \leq 1 - p$ (i.e., so that $n \geq 1/\epsilon^2(1-p)$). Note that this method tells us how large to take n to get a desired accuracy. □

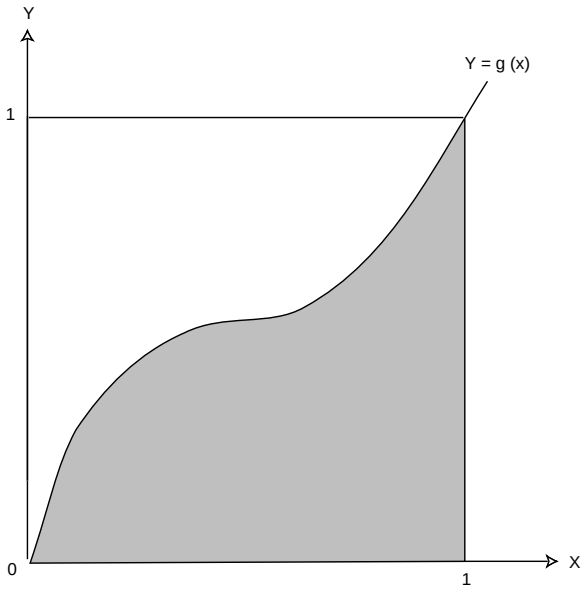


Figure 8.3: Area problem.

The Law of Large Numbers requires that the variance σ^2 of the original underlying density be finite: $\sigma^2 < \infty$. In cases where this fails to hold, the Law of Large Numbers may fail, too. An example follows.

Cauchy Case

Example 8.8 Suppose we choose n numbers from $(-\infty, +\infty)$ with a Cauchy density with parameter $a = 1$. We know that for the Cauchy density the expected value and variance are undefined (see Example 6.28). In this case, the density function for

$$A_n = \frac{S_n}{n}$$

is given by (see Example 7.6)

$$f_{A_n}(x) = \frac{1}{\pi(1 + x^2)} \ ,$$

that is, *the density function for A_n is the same for all n.* In this case, as n increases, the density function does not change at all, and the Law of Large Numbers does not hold. □

Exercises

1 Let X be a continuous random variable with mean $\mu = 10$ and variance $\sigma^2 = 100/3$. Using Chebyshev's Inequality, find an upper bound for the following probabilities.

(a) $P(|X - 10| \geq 2)$.

(b) $P(|X - 10| \geq 5)$.

(c) $P(|X - 10| \geq 9)$.

(d) $P(|X - 10| \geq 20)$.

2 Let X be a continuous random variable with values unformly distributed over the interval $[0, 20]$.

(a) Find the mean and variance of X.

(b) Calculate $P(|X - 10| \geq 2)$, $P(|X - 10| \geq 5)$, $P(|X - 10| \geq 9)$, and $P(|X - 10| \geq 20)$ exactly. How do your answers compare with those of Exercise 1? How good is Chebyshev's Inequality in this case?

3 Let X be the random variable of Exercise 2.

(a) Calculate the function $f(x) = P(|X - 10| \geq x)$.

(b) Now graph the function $f(x)$, and on the same axes, graph the Chebyshev function $g(x) = 100/(3x^2)$. Show that $f(x) \leq g(x)$ for all $x > 0$, but that $g(x)$ is not a very good approximation for $f(x)$.

4 Let X be a continuous random variable with values exponentially distributed over $[0, \infty)$ with parameter $\lambda = 0.1$.

(a) Find the mean and variance of X.

(b) Using Chebyshev's Inequality, find an upper bound for the following probabilities: $P(|X - 10| \geq 2)$, $P(|X - 10| \geq 5)$, $P(|X - 10| \geq 9)$, and $P(|X - 10| \geq 20)$.

(c) Calculate these probabilities exactly, and compare with the bounds in (b).

5 Let X be a continuous random variable with values normally distributed over $(-\infty, +\infty)$ with mean $\mu = 0$ and variance $\sigma^2 = 1$.

(a) Using Chebyshev's Inequality, find upper bounds for the following probabilities: $P(|X| \geq 1)$, $P(|X| \geq 2)$, and $P(|X| \geq 3)$.

(b) The area under the normal curve between -1 and 1 is .6827, between -2 and 2 is .9545, and between -3 and 3 it is .9973 (see the table in Appendix A). Compare your bounds in (a) with these exact values. How good is Chebyshev's Inequality in this case?

6 If X is normally distributed, with mean μ and variance σ^2, find an upper bound for the following probabilities, using Chebyshev's Inequality.

(a) $P(|X - \mu| \geq \sigma)$.

(b) $P(|X - \mu| \geq 2\sigma)$.

(c) $P(|X - \mu| \geq 3\sigma)$.

(d) $P(|X - \mu| \geq 4\sigma)$.

Now find the exact value using the program **NormalArea** or the normal table in Appendix A, and compare.

7 If X is a random variable with mean $\mu \neq 0$ and variance σ^2, define the *relative deviation* D of X from its mean by

$$D = \left| \frac{X - \mu}{\mu} \right| .$$

(a) Show that $P(D \geq a) \leq \sigma^2/(\mu^2 a^2)$.

(b) If X is the random variable of Exercise 1, find an upper bound for $P(D \geq .2)$, $P(D \geq .5)$, $P(D \geq .9)$, and $P(D \geq 2)$.

8 Let X be a continuous random variable and define the *standardized version* X^* of X by:

$$X^* = \frac{X - \mu}{\sigma} .$$

(a) Show that $P(|X^*| \geq a) \leq 1/a^2$.

(b) If X is the random variable of Exercise 1, find bounds for $P(|X^*| \geq 2)$, $P(|X^*| \geq 5)$, and $P(|X^*| \geq 9)$.

9 (a) Suppose a number X is chosen at random from $[0, 20]$ with uniform probability. Find a lower bound for the probability that X lies between 8 and 12, using Chebyshev's Inequality.

(b) Now suppose 20 real numbers are chosen independently from $[0, 20]$ with uniform probability. Find a lower bound for the probability that their average lies between 8 and 12.

(c) Now suppose 100 real numbers are chosen independently from $[0, 20]$. Find a lower bound for the probability that their average lies between 8 and 12.

10 A student's score on a particular calculus final is a random variable with values of $[0, 100]$, mean 70, and variance 25.

(a) Find a lower bound for the probability that the student's score will fall between 65 and 75.

(b) If 100 students take the final, find a lower bound for the probability that the class average will fall between 65 and 75.

11 The Pilsdorff beer company runs a fleet of trucks along the 100 mile road from Hangtown to Dry Gulch, and maintains a garage halfway in between. Each of the trucks is apt to break down at a point X miles from Hangtown, where X is a random variable uniformly distributed over $[0, 100]$.

(a) Find a lower bound for the probability $P(|X - 50| \leq 10)$.

(b) Suppose that in one bad week, 20 trucks break down. Find a lower bound for the probability $P(|A_{20} - 50| \leq 10)$, where A_{20} is the average of the distances from Hangtown at the time of breakdown.

12 A share of common stock in the Pilsdorff beer company has a price Y_n on the nth business day of the year. Finn observes that the price change $X_n = Y_{n+1} - Y_n$ appears to be a random variable with mean $\mu = 0$ and variance $\sigma^2 = 1/4$. If $Y_1 = 30$, find a lower bound for the following probabilities, under the assumption that the X_n's are mutually independent.

(a) $P(25 \leq Y_2 \leq 35)$.

(b) $P(25 \leq Y_{11} \leq 35)$.

(c) $P(25 \leq Y_{101} \leq 35)$.

13 Suppose one hundred numbers $X_1, X_2, \ldots, X_{100}$ are chosen independently at random from $[0, 20]$. Let $S = X_1 + X_2 + \cdots + X_{100}$ be the sum, $A = S/100$ the average, and $S^* = (S - 1000)/(10/\sqrt{3})$ the standardized sum. Find lower bounds for the probabilities

(a) $P(|S - 1000| \leq 100)$.

(b) $P(|A - 10| \leq 1)$.

(c) $P(|S^*| \leq \sqrt{3})$.

14 Let X be a continuous random variable normally distributed on $(-\infty, +\infty)$ with mean 0 and variance 1. Using the normal table provided in Appendix A, or the program **NormalArea**, find values for the function $f(x) = P(|X| \geq x)$ as x increases from 0 to 4.0 in steps of .25. Note that for $x \geq 0$ the table gives $NA(0, x) = P(0 \leq X \leq x)$ and thus $P(|X| \geq x) = 2(.5 - NA(0, x)$. Plot by hand the graph of $f(x)$ using these values, and the graph of the Chebyshev function $g(x) = 1/x^2$, and compare (see Exercise 3).

15 Repeat Exercise 14, but this time with mean 10 and variance 3. Note that the table in Appendix A presents values for a standard normal variable. Find the standardized version X^* for X, find values for $f^*(x) = P(|X^*| \geq x)$ as in Exercise 14, and then rescale these values for $f(x) = P(|X - 10| \geq x)$. Graph and compare this function with the Chebyshev function $g(x) = 3/x^2$.

16 Let $Z = X/Y$ where X and Y have normal densities with mean 0 and standard deviation 1. Then it can be shown that Z has a Cauchy density.

(a) Write a program to illustrate this result by plotting a bar graph of 1000 samples obtained by forming the ratio of two standard normal outcomes. Compare your bar graph with the graph of the Cauchy density. Depending upon which computer language you use, you may or may not need to tell the computer how to simulate a normal random variable. A method for doing this was described in Section 5.2.

(b) We have seen that the Law of Large Numbers does not apply to the Cauchy density (see Example 8.8). Simulate a large number of experiments with Cauchy density and compute the average of your results. Do these averages seem to be approaching a limit? If so can you explain why this might be?

17 Show that, if $X \geq 0$, then $P(X \geq a) \leq E(X)/a$.

18 (Lamperti[9]) Let X be a non-negative random variable. What is the best upper bound you can give for $P(X \geq a)$ if you know

(a) $E(X) = 20$.

(b) $E(X) = 20$ and $V(X) = 25$.

(c) $E(X) = 20$, $V(X) = 25$, and X is symmetric about its mean.

[9]Private communication.

Chapter 9

Central Limit Theorem

9.1 Central Limit Theorem for Bernoulli Trials

The second fundamental theorem of probability is the *Central Limit Theorem.* This theorem says that if S_n is the sum of n mutually independent random variables, then the distribution function of S_n is well-approximated by a certain type of continuous function known as a normal density function, which is given by the formula

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \ ,$$

as we have seen in Chapter 4.3. In this section, we will deal only with the case that $\mu = 0$ and $\sigma = 1$. We will call this particular normal density function the *s*tandard normal density, and we will denote it by $\phi(x)$:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \ .$$

A graph of this function is given in Figure 9.1. It can be shown that the area under any normal density equals 1.

The Central Limit Theorem tells us, quite generally, what happens when we have the sum of a large number of independent random variables each of which contributes a small amount to the total. In this section we shall discuss this theorem as it applies to the Bernoulli trials and in Section 9.2 we shall consider more general processes. We will discuss the theorem in the case that the individual random variables are identically distributed, but the theorem is true, under certain conditions, even if the individual random variables have different distributions.

Bernoulli Trials

Consider a Bernoulli trials process with probability p for success on each trial. Let $X_i = 1$ or 0 according as the ith outcome is a success or failure, and let $S_n = X_1 + X_2 + \cdots + X_n$. Then S_n is the number of successes in n trials. We know that S_n has as its distribution the binomial probabilities $b(n, p, j)$. In Section 3.2,

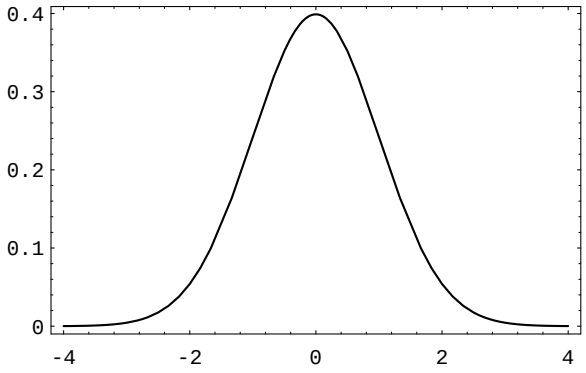


Figure 9.1: Standard normal density.

we plotted these distributions for $p = .3$ and $p = .5$ for various values of n (see Figure 3.5).

We note that the maximum values of the distributions appeared near the expected value np, which causes their spike graphs to drift off to the right as n increased. Moreover, these maximum values approach 0 as n increased, which causes the spike graphs to flatten out.

Standardized Sums

We can prevent the drifting of these spike graphs by subtracting the expected number of successes np from S_n, obtaining the new random variable $S_n - np$. Now the maximum values of the distributions will always be near 0.

To prevent the spreading of these spike graphs, we can normalize $S_n - np$ to have variance 1 by dividing by its standard deviation $\sqrt{npq}$ (see Exercise 6.2.12 and Exercise 6.2.16).

Definition 9.1 The *standardized sum* of S_n is given by

$$S_n^* = \frac{S_n - np}{\sqrt{npq}} \ .$$

S_n^* always has expected value 0 and variance 1. □

Suppose we plot a spike graph with the spikes placed at the possible values of S_n^*: x_0, x_1, ..., x_n, where

$$x_j = \frac{j - np}{\sqrt{npq}} \ . \tag{9.1}$$

We make the height of the spike at x_j equal to the distribution value $b(n, p, j)$. An example of this standardized spike graph, with $n = 270$ and $p = .3$, is shown in Figure 9.2. This graph is beautifully bell-shaped. We would like to fit a normal density to this spike graph. The obvious choice to try is the standard normal density, since it is centered at 0, just as the standardized spike graph is. In this figure, we

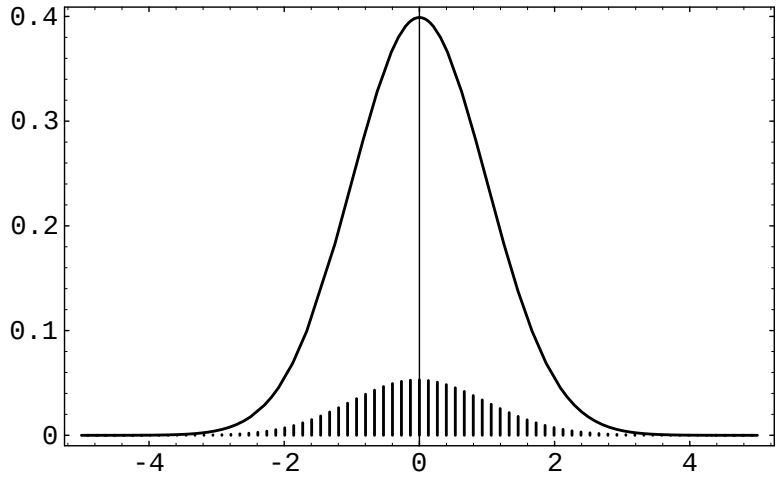


Figure 9.2: Normalized binomial distribution and standard normal density.

have drawn this standard normal density. The reader will note that a horrible thing has occurred: Even though the shapes of the two graphs are the same, the heights are quite different.

If we want the two graphs to fit each other, we must modify one of them; we choose to modify the spike graph. Since the shapes of the two graphs look fairly close, we will attempt to modify the spike graph without changing its shape. The reason for the differing heights is that the sum of the heights of the spikes equals 1, while the area under the standard normal density equals 1. If we were to draw a continuous curve through the top of the spikes, and find the area under this curve, we see that we would obtain, approximately, the sum of the heights of the spikes multiplied by the distance between consecutive spikes, which we will call ϵ. Since the sum of the heights of the spikes equals one, the area under this curve would be approximately ϵ. Thus, to change the spike graph so that the area under this curve has value 1, we need only multiply the heights of the spikes by $1/\epsilon$. It is easy to see from Equation 9.1 that

$$\epsilon = \frac{1}{\sqrt{npq}} \ .$$

In Figure 9.3 we show the standardized sum S_n^* for $n = 270$ and $p = .3$, after correcting the heights, together with the standard normal density. (This figure was produced with the program **CLTBernoulliPlot**.) The reader will note that the standard normal fits the height-corrected spike graph extremely well. In fact, one version of the Central Limit Theorem (see Theorem 9.1) says that as n increases, the standard normal density will do an increasingly better job of approximating the height-corrected spike graphs corresponding to a Bernoulli trials process with n summands.

Let us fix a value x on the x-axis and let n be a fixed positive integer. Then, using Equation 9.1, the point x_j that is closest to x has a subscript j given by the

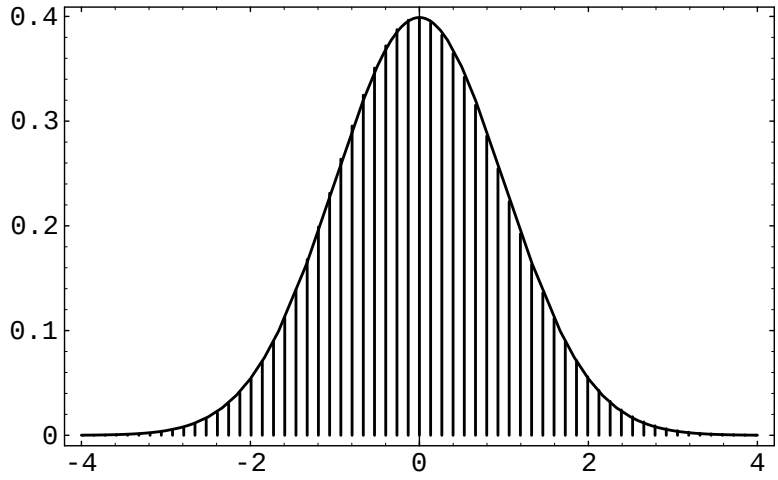


Figure 9.3: Corrected spike graph with standard normal density.

formula

$$j = \langle np + x\sqrt{npq} \rangle \ ,$$

where $\langle a \rangle$ means the integer nearest to a. Thus the height of the spike above x_j will be

$$\sqrt{npq}\, b(n,p,j) = \sqrt{npq}\, b(n,p,\langle np + x_j\sqrt{npq} \rangle) \ .$$

For large n, we have seen that the height of the spike is very close to the height of the normal density at x. This suggests the following theorem.

Theorem 9.1 (Central Limit Theorem for Binomial Distributions) For the binomial distribution $b(n,p,j)$ we have

$$\lim_{n\to\infty} \sqrt{npq}\, b(n,p,\langle np + x\sqrt{npq} \rangle) = \phi(x) \ ,$$

where $\phi(x)$ is the standard normal density.

The proof of this theorem can be carried out using Stirling's approximation from Section 3.1. We indicate this method of proof by considering the case $x = 0$. In this case, the theorem states that

$$\lim_{n\to\infty} \sqrt{npq}\, b(n,p,\langle np \rangle) = \frac{1}{\sqrt{2\pi}} = .3989\ldots \ .$$

In order to simplify the calculation, we assume that np is an integer, so that $\langle np \rangle = np$. Then

$$\sqrt{npq}\, b(n,p,np) = \sqrt{npq}\, p^{np} q^{nq} \frac{n!}{(np)!\,(nq)!} \ .$$

Recall that Stirling's formula (see Theorem 3.3) states that

$$n! \sim \sqrt{2\pi n}\, n^n e^{-n} \qquad \text{as} \quad n \to \infty \ .$$

Using this, we have

$$\sqrt{npq}\,b(n,p,np) \sim \frac{\sqrt{npq}\,p^{np}q^{nq}\sqrt{2\pi n}\,n^n e^{-n}}{\sqrt{2\pi np}\sqrt{2\pi nq}\,(np)^{np}(nq)^{nq}e^{-np}e^{-nq}} ,$$

which simplifies to $1/\sqrt{2\pi}$. □

Approximating Binomial Distributions

We can use Theorem 9.1 to find approximations for the values of binomial distribution functions. If we wish to find an approximation for $b(n,p,j)$, we set

$$j = np + x\sqrt{npq}$$

and solve for x, obtaining

$$x = \frac{j-np}{\sqrt{npq}} .$$

Theorem 9.1 then says that

$$\sqrt{npq}\,b(n,p,j)$$

is approximately equal to $\phi(x)$, so

$$\begin{aligned} b(n,p,j) &\approx \frac{\phi(x)}{\sqrt{npq}} \\ &= \frac{1}{\sqrt{npq}}\phi\left(\frac{j-np}{\sqrt{npq}}\right) . \end{aligned}$$

Example 9.1 Let us estimate the probability of exactly 55 heads in 100 tosses of a coin. For this case $np = 100 \cdot 1/2 = 50$ and $\sqrt{npq} = \sqrt{100 \cdot 1/2 \cdot 1/2} = 5$. Thus $x_{55} = (55-50)/5 = 1$ and

$$\begin{aligned} P(S_{100} = 55) \sim \frac{\phi(1)}{5} &= \frac{1}{5}\left(\frac{1}{\sqrt{2\pi}}e^{-1/2}\right) \\ &= .0484 . \end{aligned}$$

To four decimal places, the actual value is .0485, and so the approximation is very good. □

The program **CLTBernoulliLocal** illustrates this approximation for any choice of n, p, and j. We have run this program for two examples. The first is the probability of exactly 50 heads in 100 tosses of a coin; the estimate is .0798, while the actual value, to four decimal places, is .0796. The second example is the probability of exactly eight sixes in 36 rolls of a die; here the estimate is .1093, while the actual value, to four decimal places, is .1196.

The individual binomial probabilities tend to 0 as n tends to infinity. In most applications we are not interested in the probability that a specific outcome occurs, but rather in the probability that the outcome lies in a given interval, say the interval $[a, b]$. In order to find this probability, we add the heights of the spike graphs for values of j between a and b. This is the same as asking for the probability that the standardized sum S_n^* lies between a^* and b^*, where a^* and b^* are the standardized values of a and b. But as n tends to infinity the sum of these areas could be expected to approach the area under the standard normal density between a^* and b^*. The *Central Limit Theorem* states that this does indeed happen.

Theorem 9.2 (Central Limit Theorem for Bernoulli Trials) Let S_n be the number of successes in n Bernoulli trials with probability p for success, and let a and b be two fixed real numbers. Then

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) = \int_a^b \phi(x)\,dx \ .$$

□

This theorem can be proved by adding together the approximations to $b(n, p, k)$ given in Theorem 9.1.It is also a special case of the more general Central Limit Theorem (see Section 10.3).

We know from calculus that the integral on the right side of this equation is equal to the area under the graph of the standard normal density $\phi(x)$ between a and b. We denote this area by $\mathrm{NA}(a^*, b^*)$. Unfortunately, there is no simple way to integrate the function $e^{-x^2/2}$, and so we must either use a table of values or else a numerical integration program. (See Figure 9.4 for values of $\mathrm{NA}(0, z)$. A more extensive table is given in Appendix A.)

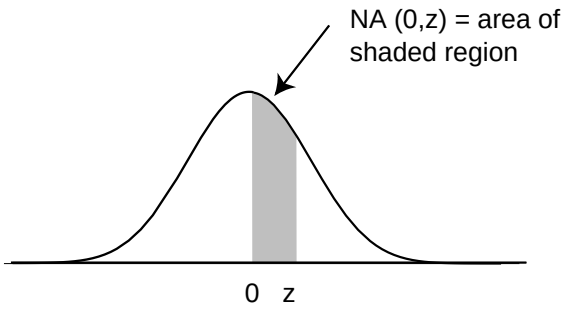
It is clear from the symmetry of the standard normal density that areas such as that between -2 and 3 can be found from this table by adding the area from 0 to 2 (same as that from -2 to 0) to the area from 0 to 3.

Approximation of Binomial Probabilities

Suppose that S_n is binomially distributed with parameters n and p. We have seen that the above theorem shows how to estimate a probability of the form

$$P(i \le S_n \le j) \ , \tag{9.2}$$

where i and j are integers between 0 and n. As we have seen, the binomial distribution can be represented as a spike graph, with spikes at the integers between 0 and n, and with the height of the kth spike given by $b(n, p, k)$. For moderate-sized values of n, if we standardize this spike graph, and change the heights of its spikes, in the manner described above, the sum of the heights of the spikes is approximated by the area under the standard normal density between i^* and j^*. It turns out that a slightly more accurate approximation is afforded by the area under the standard



z	NA(z)	z	NA(z)	z	NA(z)	z	NA(z)
.0	.0000	1.0	.3413	2.0	.4772	3.0	.4987
.1	.0398	1.1	.3643	2.1	.4821	3.1	.4990
.2	.0793	1.2	.3849	2.2	.4861	3.2	.4993
.3	.1179	1.3	.4032	2.3	.4893	3.3	.4995
.4	.1554	1.4	.4192	2.4	.4918	3.4	.4997
.5	.1915	1.5	.4332	2.5	.4938	3.5	.4998
.6	.2257	1.6	.4452	2.6	.4953	3.6	.4998
.7	.2580	1.7	.4554	2.7	.4965	3.7	.4999
.8	.2881	1.8	.4641	2.8	.4974	3.8	.4999
.9	.3159	1.9	.4713	2.9	.4981	3.9	.5000

Figure 9.4: Table of values of $\mathrm{NA}(0, z)$, the normal area from 0 to z.

normal density between the standardized values corresponding to $(i - 1/2)$ and $(j + 1/2)$; these values are

$$i^* = \frac{i - 1/2 - np}{\sqrt{npq}}$$

and

$$j^* = \frac{j + 1/2 - np}{\sqrt{npq}} \ .$$

Thus,

$$P(i \leq S_n \leq j) \approx \mathrm{NA}\left(\frac{i - \frac{1}{2} - np}{\sqrt{npq}}, \frac{j + \frac{1}{2} - np}{\sqrt{npq}}\right) .$$

It should be stressed that the approximations obtained by using the Central Limit Theorem are only approximations, and sometimes they are not very close to the actual values (see Exercise 12).

We now illustrate this idea with some examples.

Example 9.2 A coin is tossed 100 times. Estimate the probability that the number of heads lies between 40 and 60 (the word "between" in mathematics means inclusive of the endpoints). The expected number of heads is $100 \cdot 1/2 = 50$, and the standard deviation for the number of heads is $\sqrt{100 \cdot 1/2 \cdot 1/2} = 5$. Thus, since $n = 100$ is reasonably large, we have

$$\begin{aligned} P(40 \leq S_n \leq 60) &\approx P\left(\frac{39.5 - 50}{5} \leq S_n^* \leq \frac{60.5 - 50}{5}\right) \\ &= P(-2.1 \leq S_n^* \leq 2.1) \\ &\approx \mathrm{NA}(-2.1, 2.1) \\ &= 2\mathrm{NA}(0, 2.1) \\ &\approx .9642 \ . \end{aligned}$$

The actual value is .96480, to five decimal places.

Note that in this case we are asking for the probability that the outcome will not deviate by more than two standard deviations from the expected value. Had we asked for the probability that the number of successes is between 35 and 65, this would have represented three standard deviations from the mean, and, using our 1/2 correction, our estimate would be the area under the standard normal curve between -3.1 and 3.1, or $2\mathrm{NA}(0, 3.1) = .9980$. The actual answer in this case, to five places, is .99821. □

It is important to work a few problems by hand to understand the conversion from a given inequality to an inequality relating to the standardized variable. After this, one can then use a computer program that carries out this conversion, including the 1/2 correction. The program **CLTBernoulliGlobal** is such a program for estimating probabilities of the form $P(a \leq S_n \leq b)$.

Example 9.3 Dartmouth College would like to have 1050 freshmen. This college cannot accommodate more than 1060. Assume that each applicant accepts with

probability .6 and that the acceptances can be modeled by Bernoulli trials. If the college accepts 1700, what is the probability that it will have too many acceptances?

If it accepts 1700 students, the expected number of students who matriculate is $.6 \cdot 1700 = 1020$. The standard deviation for the number that accept is $\sqrt{1700 \cdot .6 \cdot .4} \approx 20$. Thus we want to estimate the probability

$$\begin{aligned} P(S_{1700} > 1060) &= P(S_{1700} \geq 1061) \\ &= P\left(S^*_{1700} \geq \frac{1060.5 - 1020}{20}\right) \\ &= P(S^*_{1700} \geq 2.025) \ . \end{aligned}$$

From Table 9.4, if we interpolate, we would estimate this probability to be $.5 - .4784 = .0216$. Thus, the college is fairly safe using this admission policy. □

Applications to Statistics

There are many important questions in the field of statistics that can be answered using the Central Limit Theorem for independent trials processes. The following example is one that is encountered quite frequently in the news. Another example of an application of the Central Limit Theorem to statistics is given in Section 9.2.

Example 9.4 One frequently reads that a poll has been taken to estimate the proportion of people in a certain population who favor one candidate over another in a race with two candidates. (This model also applies to races with more than two candidates A and B, and two ballot propositions.) Clearly, it is not possible for pollsters to ask everyone for their preference. What is done instead is to pick a subset of the population, called a sample, and ask everyone in the sample for their preference. Let p be the actual proportion of people in the population who are in favor of candidate A and let $q = 1 - p$. If we choose a sample of size n from the population, the preferences of the people in the sample can be represented by random variables $X_1,\ X_2,\ \ldots,\ X_n$, where $X_i = 1$ if person i is in favor of candidate A, and $X_i = 0$ if person i is in favor of candidate B. Let $S_n = X_1 + X_2 + \cdots + X_n$. If each subset of size n is chosen with the same probability, then S_n is hypergeometrically distributed. If n is small relative to the size of the population (which is typically true in practice), then S_n is approximately binomially distributed, with parameters n and p.

The pollster wants to estimate the value p. An estimate for p is provided by the value $\bar{p} = S_n/n$, which is the proportion of people in the sample who favor candidate B. The Central Limit Theorem says that the random variable $\bar{p}$ is approximately normally distributed. (In fact, our version of the Central Limit Theorem says that the distribution function of the random variable

$$S^*_n = \frac{S_n - np}{\sqrt{npq}}$$

is approximated by the standard normal density.) But we have

$$\bar{p} = \frac{S_n - np}{\sqrt{npq}}\sqrt{\frac{pq}{n}} + p \ ,$$

i.e., $\bar{p}$ is just a linear function of S_n^*. Since the distribution of S_n^* is approximated by the standard normal density, the distribution of the random variable $\bar{p}$ must also be bell-shaped. We also know how to write the mean and standard deviation of $\bar{p}$ in terms of p and n. The mean of $\bar{p}$ is just p, and the standard deviation is

$$\sqrt{\frac{pq}{n}} \ .$$

Thus, it is easy to write down the standardized version of $\bar{p}$; it is

$$\bar{p}^* = \frac{\bar{p} - p}{\sqrt{pq/n}} \ .$$

Since the distribution of the standardized version of $\bar{p}$ is approximated by the standard normal density, we know, for example, that 95% of its values will lie within two standard deviations of its mean, and the same is true of $\bar{p}$. So we have

$$P\left(p - 2\sqrt{\frac{pq}{n}} < \bar{p} < p + 2\sqrt{\frac{pq}{n}}\right) \approx .954 \ .$$

Now the pollster does not know p or q, but he can use $\bar{p}$ and $\bar{q} = 1 - \bar{p}$ in their place without too much danger. With this idea in mind, the above statement is equivalent to the statement

$$P\left(\bar{p} - 2\sqrt{\frac{\bar{p}\bar{q}}{n}} < p < \bar{p} + 2\sqrt{\frac{\bar{p}\bar{q}}{n}}\right) \approx .954 \ .$$

The resulting interval

$$\left(\bar{p} - \frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}},\ \bar{p} + \frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}}\right)$$

is called the *95 percent confidence interval* for the unknown value of p. The name is suggested by the fact that if we use this method to estimate p in a large number of samples we should expect that in about 95 percent of the samples the true value of p is contained in the confidence interval obtained from the sample. In Exercise 11 you are asked to write a program to illustrate that this does indeed happen.

The pollster has control over the value of n. Thus, if he wants to create a 95% confidence interval with length 6%, then he should choose a value of n so that

$$\frac{2\sqrt{\bar{p}\bar{q}}}{\sqrt{n}} \leq .03 \ .$$

Using the fact that $\bar{p}\bar{q} \leq 1/4$, no matter what the value of $\bar{p}$ is, it is easy to show that if he chooses a value of n so that

$$\frac{1}{\sqrt{n}} \leq .03 \ ,$$

he will be safe. This is equivalent to choosing

$$n \geq 1111 \ .$$

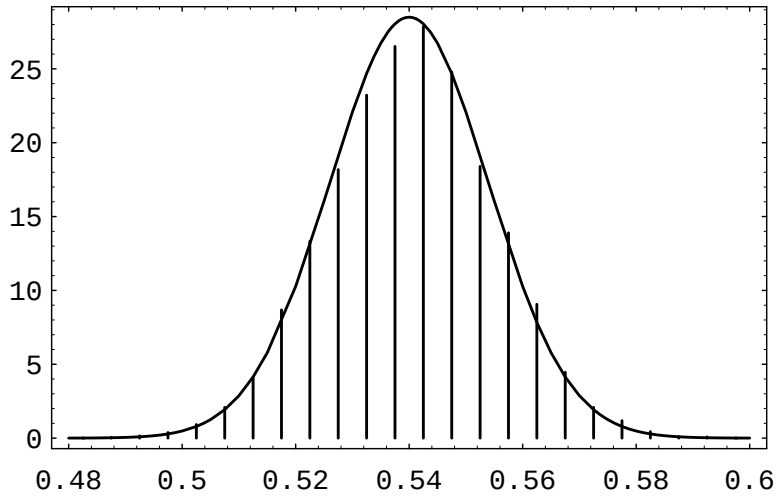


Figure 9.5: Polling simulation.

So if the pollster chooses n to be 1200, say, and calculates $\bar{p}$ using his sample of size 1200, then 19 times out of 20 (i.e., 95% of the time), his confidence interval, which is of length 6%, will contain the true value of p. This type of confidence interval is typically reported in the news as follows: this survey has a 3% margin of error. In fact, most of the surveys that one sees reported in the paper will have sample sizes around 1000. A somewhat surprising fact is that the size of the population has apparently no effect on the sample size needed to obtain a 95% confidence interval for p with a given margin of error. To see this, note that the value of n that was needed depended only on the number .03, which is the margin of error. In other words, whether the population is of size 100,000 or 100,000,000, the pollster needs only to choose a sample of size 1200 or so to get the same accuracy of estimate of p. (We did use the fact that the sample size was small relative to the population size in the statement that S_n is approximately binomially distributed.)

In Figure 9.5, we show the results of simulating the polling process. The population is of size 100,000, and for the population, $p = .54$. The sample size was chosen to be 1200. The spike graph shows the distribution of $\bar{p}$ for 10,000 randomly chosen samples. For this simulation, the program kept track of the number of samples for which $\bar{p}$ was within 3% of .54. This number was 9648, which is close to 95% of the number of samples used.

Another way to see what the idea of confidence intervals means is shown in Figure 9.6. In this figure, we show 100 confidence intervals, obtained by computing $\bar{p}$ for 100 different samples of size 1200 from the same population as before. The reader can see that most of these confidence intervals (96, to be exact) contain the true value of p.

The Gallup Poll has used these polling techniques in every Presidential election since 1936 (and in innumerable other elections as well). Table 9.1[1] shows the results

[1]The Gallup Poll Monthly, November 1992, No. 326, p. 33. Supplemented with the help of

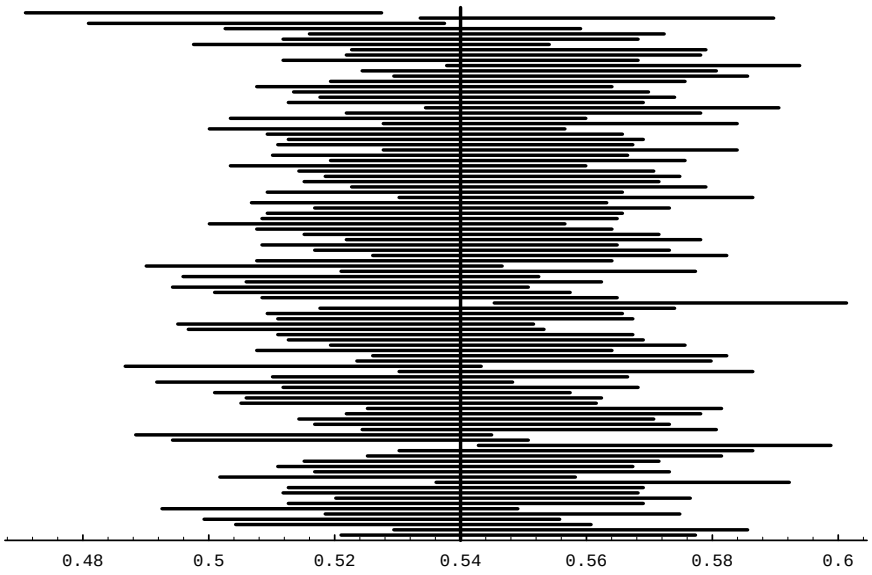


Figure 9.6: Confidence interval simulation.

of their efforts. The reader will note that most of the approximations to p are within 3% of the actual value of p. The sample sizes for these polls were typically around 1500. (In the table, both the predicted and actual percentages for the winning candidate refer to the percentage of the vote among the "major" political parties. In most elections, there were two major parties, but in several elections, there were three.)

This technique also plays an important role in the evaluation of the effectiveness of drugs in the medical profession. For example, it is sometimes desired to know what proportion of patients will be helped by a new drug. This proportion can be estimated by giving the drug to a subset of the patients, and determining the proportion of this sample who are helped by the drug. □

Historical Remarks

The Central Limit Theorem for Bernoulli trials was first proved by Abraham de Moivre and appeared in his book, *The Doctrine of Chances,* first published in 1718.[2]

De Moivre spent his years from age 18 to 21 in prison in France because of his Protestant background. When he was released he left France for England, where he worked as a tutor to the sons of noblemen. Newton had presented a copy of his *Principia Mathematica* to the Earl of Devonshire. The story goes that, while de Moivre was tutoring at the Earl's house, he came upon Newton's work and found that it was beyond him. It is said that he then bought a copy of his own and tore

Lydia K. Saab, The Gallup Organization.

[2]A. de Moivre, *The Doctrine of Chances,* 3d ed. (London: Millar, 1756).

Year	Winning Candidate	Gallup Final Survey	Election Result	Deviation
1936	Roosevelt	55.7%	62.5%	6.8%
1940	Roosevelt	52.0%	55.0%	3.0%
1944	Roosevelt	51.5%	53.3%	1.8%
1948	Truman	44.5%	49.9%	5.4%
1952	Eisenhower	51.0%	55.4%	4.4%
1956	Eisenhower	59.5%	57.8%	1.7%
1960	Kennedy	51.0%	50.1%	0.9%
1964	Johnson	64.0%	61.3%	2.7%
1968	Nixon	43.0%	43.5%	0.5%
1972	Nixon	62.0%	61.8%	0.2%
1976	Carter	48.0%	50.0%	2.0%
1980	Reagan	47.0%	50.8%	3.8%
1984	Reagan	59.0%	59.1%	0.1%
1988	Bush	56.0%	53.9%	2.1%
1992	Clinton	49.0%	43.2%	5.8%
1996	Clinton	52.0%	50.1%	1.9%

Table 9.1: Gallup Poll accuracy record.

it into separate pages, learning it page by page as he walked around London to his tutoring jobs. De Moivre frequented the coffeehouses in London, where he started his probability work by calculating odds for gamblers. He also met Newton at such a coffeehouse and they became fast friends. De Moivre dedicated his book to Newton.

The Doctrine of Chances provides the techniques for solving a wide variety of gambling problems. In the midst of these gambling problems de Moivre rather modestly introduces his proof of the Central Limit Theorem, writing

> A Method of approximating the Sum of the Terms of the Binomial $(a+b)^n$ expanded into a Series, from whence are deduced some practical Rules to estimate the Degree of Assent which is to be given to Experiments.[3]

De Moivre's proof used the approximation to factorials that we now call Stirling's formula. De Moivre states that he had obtained this formula before Stirling but without determining the exact value of the constant $\sqrt{2\pi}$. While he says it is not really necessary to know this exact value, he concedes that knowing it "has spread a singular Elegancy on the Solution."

The complete proof and an interesting discussion of the life of de Moivre can be found in the book *Games, Gods and Gambling* by F. N. David.[4]

[3] ibid., p. 243.
[4] F. N. David, *Games, Gods and Gambling* (London: Griffin, 1962).

Exercises

1 Let S_{100} be the number of heads that turn up in 100 tosses of a fair coin. Use the Central Limit Theorem to estimate

(a) $P(S_{100} \leq 45)$.

(b) $P(45 < S_{100} < 55)$.

(c) $P(S_{100} > 63)$.

(d) $P(S_{100} < 57)$.

2 Let S_{200} be the number of heads that turn up in 200 tosses of a fair coin. Estimate

(a) $P(S_{200} = 100)$.

(b) $P(S_{200} = 90)$.

(c) $P(S_{200} = 80)$.

3 A true-false examination has 48 questions. June has probability 3/4 of answering a question correctly. April just guesses on each question. A passing score is 30 or more correct answers. Compare the probability that June passes the exam with the probability that April passes it.

4 Let S be the number of heads in 1,000,000 tosses of a fair coin. Use (a) Chebyshev's inequality, and (b) the Central Limit Theorem, to estimate the probability that S lies between 499,500 and 500,500. Use the same two methods to estimate the probability that S lies between 499,000 and 501,000, and the probability that S lies between 498,500 and 501,500.

5 A rookie is brought to a baseball club on the assumption that he will have a .300 batting average. (Batting average is the ratio of the number of hits to the number of times at bat.) In the first year, he comes to bat 300 times and his batting average is .267. Assume that his at bats can be considered Bernoulli trials with probability .3 for success. Could such a low average be considered just bad luck or should he be sent back to the minor leagues? Comment on the assumption of Bernoulli trials in this situation.

6 Once upon a time, there were two railway trains competing for the passenger traffic of 1000 people leaving from Chicago at the same hour and going to Los Angeles. Assume that passengers are equally likely to choose each train. How many seats must a train have to assure a probability of .99 or better of having a seat for each passenger?

7 Assume that, as in Example 9.3, Dartmouth admits 1750 students. What is the probability of too many acceptances?

8 A club serves dinner to members only. They are seated at 12-seat tables. The manager observes over a long period of time that 95 percent of the time there are between six and nine full tables of members, and the remainder of the

time the numbers are equally likely to fall above or below this range. Assume that each member decides to come with a given probability p, and that the decisions are independent. How many members are there? What is p?

9 Let S_n be the number of successes in n Bernoulli trials with probability .8 for success on each trial. Let $A_n = S_n/n$ be the average number of successes. In each case give the value for the limit, and give a reason for your answer.

(a) $\lim_{n\to\infty} P(A_n = .8)$.

(b) $\lim_{n\to\infty} P(.7n < S_n < .9n)$.

(c) $\lim_{n\to\infty} P(S_n < .8n + .8\sqrt{n})$.

(d) $\lim_{n\to\infty} P(.79 < A_n < .81)$.

10 Find the probability that among 10,000 random digits the digit 3 appears not more than 931 times.

11 Write a computer program to simulate 10,000 Bernoulli trials with probability .3 for success on each trial. Have the program compute the 95 percent confidence interval for the probability of success based on the proportion of successes. Repeat the experiment 100 times and see how many times the true value of .3 is included within the confidence limits.

12 A balanced coin is flipped 400 times. Determine the number x such that the probability that the number of heads is between $200 - x$ and $200 + x$ is approximately .80.

13 A noodle machine in Spumoni's spaghetti factory makes about 5 percent defective noodles even when properly adjusted. The noodles are then packed in crates containing 1900 noodles each. A crate is examined and found to contain 115 defective noodles. What is the approximate probability of finding at least this many defective noodles if the machine is properly adjusted?

14 A restaurant feeds 400 customers per day. On the average 20 percent of the customers order apple pie.

(a) Give a range (called a 95 percent confidence interval) for the number of pieces of apple pie ordered on a given day such that you can be 95 percent sure that the actual number will fall in this range.

(b) How many customers must the restaurant have, on the average, to be at least 95 percent sure that the number of customers ordering pie on that day falls in the 19 to 21 percent range?

15 Recall that if X is a random variable, the *cumulative distribution function* of X is the function $F(x)$ defined by

$$F(x) = P(X \leq x) \ .$$

(a) Let S_n be the number of successes in n Bernoulli trials with probability p for success. Write a program to plot the cumulative distribution for S_n.

(b) Modify your program in (a) to plot the cumulative distribution $F_n^*(x)$ of the standardized random variable

$$S_n^* = \frac{S_n - np}{\sqrt{npq}} \ .$$

(c) Define the *normal distribution* $N(x)$ to be the area under the normal curve up to the value x. Modify your program in (b) to plot the normal distribution as well, and compare it with the cumulative distribution of S_n^*. Do this for $n = 10, 50$, and 100.

16 In Example 3.11, we were interested in testing the hypothesis that a new form of aspirin is effective 80 percent of the time rather than the 60 percent of the time as reported for standard aspirin. The new aspirin is given to n people. If it is effective in m or more cases, we accept the claim that the new drug is effective 80 percent of the time and if not we reject the claim. Using the Central Limit Theorem, show that you can choose the number of trials n and the critical value m so that the probability that we reject the hypothesis when it is true is less than .01 and the probability that we accept it when it is false is also less than .01. Find the smallest value of n that will suffice for this.

17 In an opinion poll it is assumed that an unknown proportion p of the people are in favor of a proposed new law and a proportion $1 - p$ are against it. A sample of n people is taken to obtain their opinion. The proportion $\bar{p}$ in favor in the sample is taken as an estimate of p. Using the Central Limit Theorem, determine how large a sample will ensure that the estimate will, with probability .95, be correct to within .01.

18 A description of a poll in a certain newspaper says that one can be 95% confident that error due to sampling will be no more than plus or minus 3 percentage points. A poll in the New York Times taken in Iowa says that "according to statistical theory, in 19 out of 20 cases the results based on such samples will differ by no more than 3 percentage points in either direction from what would have been obtained by interviewing all adult Iowans." These are both attempts to explain the concept of confidence intervals. Do both statements say the same thing? If not, which do you think is the more accurate description?

9.2 Central Limit Theorem for Discrete Independent Trials

We have illustrated the Central Limit Theorem in the case of Bernoulli trials, but this theorem applies to a much more general class of chance processes. In particular, it applies to any independent trials process such that the individual trials have finite variance. For such a process, both the normal approximation for individual terms and the Central Limit Theorem are valid.

Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of n independent discrete random variables of an independent trials process with common distribution function $m(x)$ defined on the integers, with mean μ and variance σ^2. We have seen in Section 7.2 that the distributions for such independent sums have shapes resembling the normal curve, but the largest values drift to the right and the curves flatten out (see Figure 7.6). We can prevent this just as we did for Bernoulli trials.

Standardized Sums

Consider the standardized random variable

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \ .$$

This standardizes S_n to have expected value 0 and variance 1. If $S_n = j$, then S_n^* has the value x_j with

$$x_j = \frac{j - n\mu}{\sqrt{n\sigma^2}} \ .$$

We can construct a spike graph just as we did for Bernoulli trials. Each spike is centered at some x_j. The distance between successive spikes is

$$b = \frac{1}{\sqrt{n\sigma^2}} \ ,$$

and the height of the spike is

$$h = \sqrt{n\sigma^2} P(S_n = j) \ .$$

The case of Bernoulli trials is the special case for which $X_j = 1$ if the jth outcome is a success and 0 otherwise; then $\mu = p$ and $\sigma^2 = \sqrt{pq}$.

We now illustrate this process for two different discrete distributions. The first is the distribution m, given by

$$m = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ .2 & .2 & .2 & .2 & .2 \end{pmatrix} \ .$$

In Figure 9.7 we show the standardized sums for this distribution for the cases $n = 2$ and $n = 10$. Even for $n = 2$ the approximation is surprisingly good.

For our second discrete distribution, we choose

$$m = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ .4 & .3 & .1 & .1 & .1 \end{pmatrix} \ .$$

This distribution is quite asymmetric and the approximation is not very good for $n = 3$, but by $n = 10$ we again have an excellent approximation (see Figure 9.8). Figures 9.7 and 9.8 were produced by the program **CLTIndTrialsPlot**.

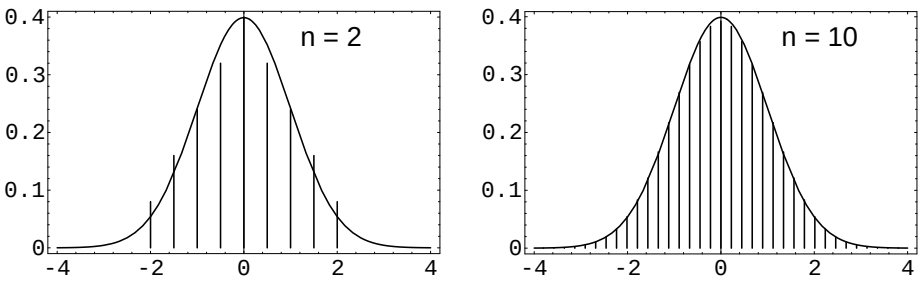


Figure 9.7: Distribution of standardized sums.

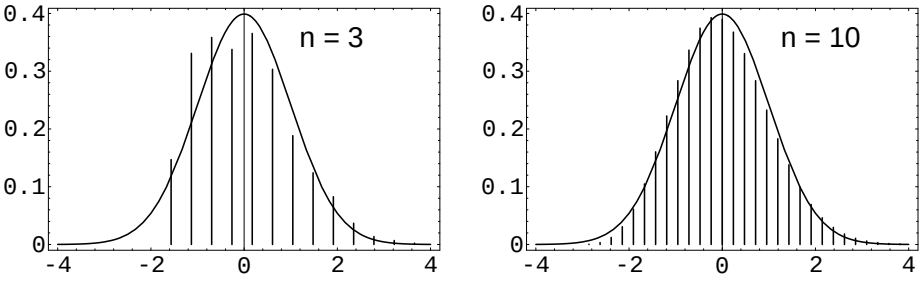


Figure 9.8: Distribution of standardized sums.

Approximation Theorem

As in the case of Bernoulli trials, these graphs suggest the following approximation theorem for the individual probabilities.

Theorem 9.3 Let X_1, X_2, ..., X_n be an independent trials process and let $S_n = X_1 + X_2 + \cdots + X_n$. Assume that the greatest common divisor of the differences of all the values that the X_j can take on is 1. Let $E(X_j) = \mu$ and $V(X_j) = \sigma^2$. Then for n large,

$$P(S_n = j) \sim \frac{\phi(x_j)}{\sqrt{n\sigma^2}} \ ,$$

where $x_j = (j - n\mu)/\sqrt{n\sigma^2}$, and $\phi(x)$ is the standard normal density. □

The program **CLTIndTrialsLocal** implements this approximation. When we run this program for 6 rolls of a die, and ask for the probability that the sum of the rolls equals 21, we obtain an actual value of .09285, and a normal approximation value of .09537. If we run this program for 24 rolls of a die, and ask for the probability that the sum of the rolls is 72, we obtain an actual value of .01724 and a normal approximation value of .01705. These results show that the normal approximations are quite good.

Central Limit Theorem for a Discrete Independent Trials Process

The Central Limit Theorem for a discrete independent trials process is as follows.

Theorem 9.4 (Central Limit Theorem) Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of n discrete independent random variables with common distribution having expected value μ and variance σ^2. Then, for $a < b$,

$$\lim_{n\to\infty} P\left(a < \frac{S_n - n\mu}{\sqrt{n\sigma^2}} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \ .$$

□

We will give the proofs of Theorems 9.3 and Theorem 9.4 in Section 10.3. Here we consider several examples.

Examples

Example 9.5 A die is rolled 420 times. What is the probability that the sum of the rolls lies between 1400 and 1550?

The sum is a random variable

$$S_{420} = X_1 + X_2 + \cdots + X_{420} \ ,$$

where each X_j has distribution

$$m_X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}$$

We have seen that $\mu = E(X) = 7/2$ and $\sigma^2 = V(X) = 35/12$. Thus, $E(S_{420}) = 420 \cdot 7/2 = 1470$, $\sigma^2(S_{420}) = 420 \cdot 35/12 = 1225$, and $\sigma(S_{420}) = 35$. Therefore,

$$\begin{aligned} P(1400 \le S_{420} \le 1550) &\approx P\left(\frac{1399.5 - 1470}{35} \le S^*_{420} \le \frac{1550.5 - 1470}{35}\right) \\ &= P(-2.01 \le S^*_{420} \le 2.30) \\ &\approx \mathrm{NA}(-2.01, 2.30) = .9670 \ . \end{aligned}$$

We note that the program **CLTIndTrialsGlobal** could be used to calculate these probabilities. □

Example 9.6 A student's grade point average is the average of his grades in 30 courses. The grades are based on 100 possible points and are recorded as integers. Assume that, in each course, the instructor makes an error in grading of k with probability $|p/k|$, where $k = \pm 1, \pm 2, \pm 3, \pm 4, \pm 5$. The probability of no error is then $1 - (137/30)p$. (The parameter p represents the inaccuracy of the instructor's grading.) Thus, in each course, there are two grades for the student, namely the

"correct" grade and the recorded grade. So there are two average grades for the student, namely the average of the correct grades and the average of the recorded grades.

We wish to estimate the probability that these two average grades differ by less than .05 for a given student. We now assume that $p = 1/20$. We also assume that the total error is the sum S_{30} of 30 independent random variables each with distribution

$$m_X : \begin{Bmatrix} -5 & -4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 & 5 \\ \frac{1}{100} & \frac{1}{80} & \frac{1}{60} & \frac{1}{40} & \frac{1}{20} & \frac{463}{600} & \frac{1}{20} & \frac{1}{40} & \frac{1}{60} & \frac{1}{80} & \frac{1}{100} \end{Bmatrix} .$$

One can easily calculate that $E(X) = 0$ and $\sigma^2(X) = 1.5$. Then we have

$$\begin{aligned} P\left(-.05 \le \tfrac{S_{30}}{30} \le .05\right) &= P(-1.5 \le S_{30} \le 1.5) \\ &= P\left(\tfrac{-1.5}{\sqrt{30 \cdot 1.5}} \le S^*_{30} \le \tfrac{1.5}{\sqrt{30 \cdot 1.5}}\right) \\ &= P(-.224 \le S^*_{30} \le .224) \\ &\approx \mathrm{NA}(-.224, .224) = .1772 . \end{aligned}$$

This means that there is only a 17.7% chance that a given student's grade point average is accurate to within .05. (Thus, for example, if two candidates for valedictorian have recorded averages of 97.1 and 97.2, there is an appreciable probability that their correct averages are in the reverse order.) For a further discussion of this example, see the article by R. M. Kozelka.[5] □

A More General Central Limit Theorem

In Theorem 9.4, the discrete random variables that were being summed were assumed to be independent and identically distributed. It turns out that the assumption of identical distributions can be substantially weakened. Much work has been done in this area, with an important contribution being made by J. W. Lindeberg. Lindeberg found a condition on the sequence $\{X_n\}$ which guarantees that the distribution of the sum S_n is asymptotically normally distributed. Feller showed that Lindeberg's condition is necessary as well, in the sense that if the condition does not hold, then the sum S_n is not asymptotically normally distributed. For a precise statement of Lindeberg's Theorem, we refer the reader to Feller.[6] A sufficient condition that is stronger (but easier to state) than Lindeberg's condition, and is weaker than the condition in Theorem 9.4, is given in the following theorem.

[5] R. M. Kozelka, "Grade-Point Averages and the Central Limit Theorem," *American Math. Monthly,* vol. 86 (Nov 1979), pp. 773-777.

[6] W. Feller, *Introduction to Probability Theory and its Applications,* vol. 1, 3rd ed. (New York: John Wiley & Sons, 1968), p. 254.

Theorem 9.5 (Central Limit Theorem) Let $X_1,\ X_2,\ \ldots,\ X_n\ ,\ \ldots$ be a sequence of independent discrete random variables, and let $S_n = X_1 + X_2 + \cdots + X_n$. For each n, denote the mean and variance of X_n by μ_n and σ_n^2, respectively. Define the mean and variance of S_n to be m_n and s_n^2, respectively, and assume that $s_n \to \infty$. If there exists a constant A, such that $|X_n| \leq A$ for all n, then for $a < b$,

$$\lim_{n\to\infty} P\left(a < \frac{S_n - m_n}{s_n} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx\ .$$

□

The condition that $|X_n| \leq A$ for all n is sometimes described by saying that the sequence $\{X_n\}$ is uniformly bounded. The condition that $s_n \to \infty$ is necessary (see Exercise 15).

We illustrate this theorem by generating a sequence of n random distributions on the interval $[a, b]$. We then convolute these distributions to find the distribution of the sum of n independent experiments governed by these distributions. Finally, we standardize the distribution for the sum to have mean 0 and standard deviation 1 and compare it with the normal density. The program **CLTGeneral** carries out this procedure.

In Figure 9.9 we show the result of running this program for $[a, b] = [-2, 4]$, and $n = 1,\ 4$, and 10. We see that our first random distribution is quite asymmetric. By the time we choose the sum of ten such experiments we have a very good fit to the normal curve.

The above theorem essentially says that anything that can be thought of as being made up as the sum of many small independent pieces is approximately normally distributed. This brings us to one of the most important questions that was asked about genetics in the 1800's.

The Normal Distribution and Genetics

When one looks at the distribution of heights of adults of one sex in a given population, one cannot help but notice that this distribution looks like the normal distribution. An example of this is shown in Figure 9.10. This figure shows the distribution of heights of 9593 women between the ages of 21 and 74. These data come from the Health and Nutrition Examination Survey I (HANES I). For this survey, a sample of the U.S. civilian population was chosen. The survey was carried out between 1971 and 1974.

A natural question to ask is "How does this come about?". Francis Galton, an English scientist in the 19th century, studied this question, and other related questions, and constructed probability models that were of great importance in explaining the genetic effects on such attributes as height. In fact, one of the most important ideas in statistics, the idea of regression to the mean, was invented by Galton in his attempts to understand these genetic effects.

Galton was faced with an apparent contradiction. On the one hand, he knew that the normal distribution arises in situations in which many small independent effects are being summed. On the other hand, he also knew that many quantitative

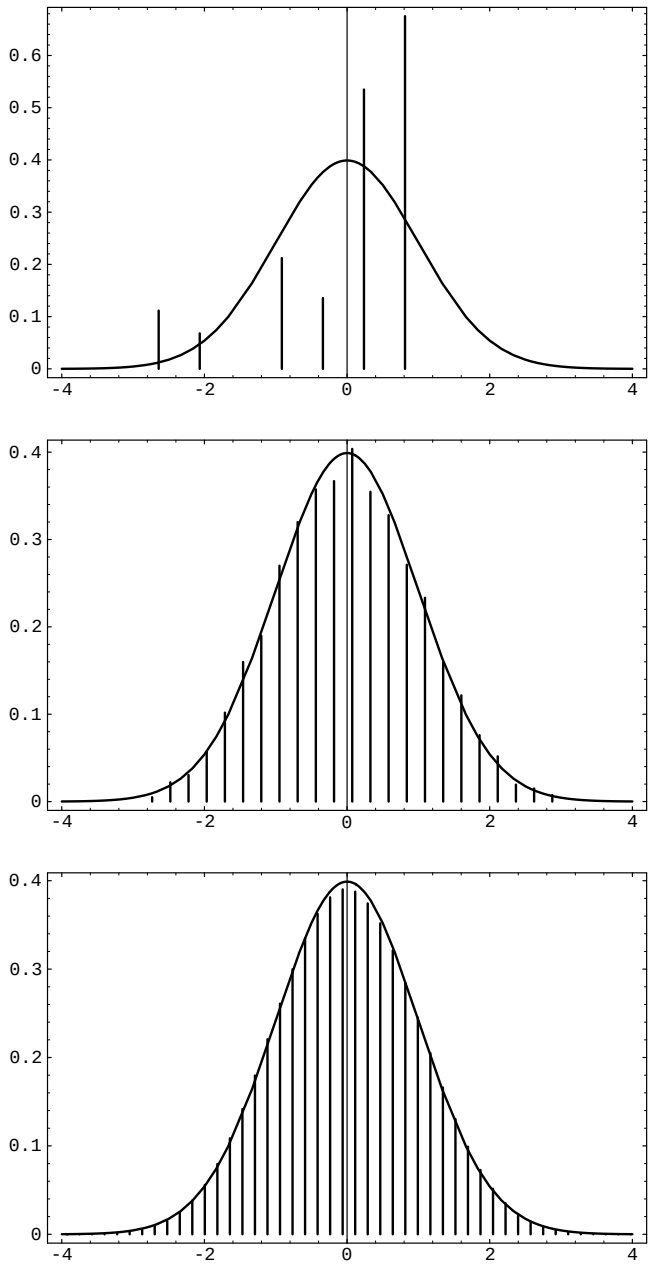


Figure 9.9: Sums of randomly chosen random variables.

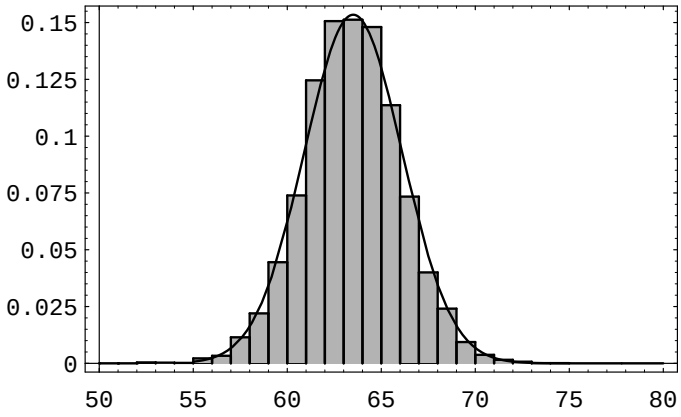


Figure 9.10: Distribution of heights of adult women.

attributes, such as height, are strongly influenced by genetic factors: tall parents tend to have tall offspring. Thus in this case, there seem to be two large effects, namely the parents. Galton was certainly aware of the fact that non-genetic factors played a role in determining the height of an individual. Nevertheless, unless these non-genetic factors overwhelm the genetic ones, thereby refuting the hypothesis that heredity is important in determining height, it did not seem possible for sets of parents of given heights to have offspring whose heights were normally distributed.

One can express the above problem symbolically as follows. Suppose that we choose two specific positive real numbers x and y, and then find all pairs of parents one of whom is x units tall and the other of whom is y units tall. We then look at all of the offspring of these pairs of parents. One can postulate the existence of a function $f(x, y)$ which denotes the genetic effect of the parents' heights on the heights of the offspring. One can then let W denote the effects of the non-genetic factors on the heights of the offspring. Then, for a given set of heights $\{x, y\}$, the random variable which represents the heights of the offspring is given by

$$H = f(x, y) + W \ ,$$

where f is a deterministic function, i.e., it gives one output for a pair of inputs $\{x, y\}$. If we assume that the effect of f is large in comparison with the effect of W, then the variance of W is small. But since f is deterministic, the variance of H equals the variance of W, so the variance of H is small. However, Galton observed from his data that the variance of the heights of the offspring of a given pair of parent heights is not small. This would seem to imply that inheritance plays a small role in the determination of the height of an individual. Later in this section, we will describe the way in which Galton got around this problem.

We will now consider the modern explanation of why certain traits, such as heights, are approximately normally distributed. In order to do so, we need to introduce some terminology from the field of genetics. The cells in a living organism that are not directly involved in the transmission of genetic material to offspring are called somatic cells, and the remaining cells are called germ cells. Organisms of

a given species have their genetic information encoded in sets of physical entities, called chromosomes. The chromosomes are paired in each somatic cell. For example, human beings have 23 pairs of chromosomes in each somatic cell. The sex cells contain one chromosome from each pair. In sexual reproduction, two sex cells, one from each parent, contribute their chromosomes to create the set of chromosomes for the offspring.

Chromosomes contain many subunits, called genes. Genes consist of molecules of DNA, and one gene has, encoded in its DNA, information that leads to the regulation of proteins. In the present context, we will consider those genes containing information that has an effect on some physical trait, such as height, of the organism. The pairing of the chromosomes gives rise to a pairing of the genes on the chromosomes.

In a given species, each gene can be any one of several forms. These various forms are called alleles. One should think of the different alleles as potentially producing different effects on the physical trait in question. Of the two alleles that are found in a given gene pair in an organism, one of the alleles came from one parent and the other allele came from the other parent. The possible types of pairs of alleles (without regard to order) are called genotypes.

If we assume that the height of a human being is largely controlled by a specific gene, then we are faced with the same difficulty that Galton was. We are assuming that each parent has a pair of alleles which largely controls their heights. Since each parent contributes one allele of this gene pair to each of its offspring, there are four possible allele pairs for the offspring at this gene location. The assumption is that these pairs of alleles largely control the height of the offspring, and we are also assuming that genetic factors outweigh non-genetic factors. It follows that among the offspring we should see several modes in the height distribution of the offspring, one mode corresponding to each possible pair of alleles. This distribution does not correspond to the observed distribution of heights.

An alternative hypothesis, which does explain the observation of normally distributed heights in offspring of a given sex, is the multiple-gene hypothesis. Under this hypothesis, we assume that there are many genes that affect the height of an individual. These genes may differ in the amount of their effects. Thus, we can represent each gene pair by a random variable X_i, where the value of the random variable is the allele pair's effect on the height of the individual. Thus, for example, if each parent has two different alleles in the gene pair under consideration, then the offspring has one of four possible pairs of alleles at this gene location. Now the height of the offspring is a random variable, which can be expressed as

$$H = X_1 + X_2 + \cdots + X_n + W \ ,$$

if there are n genes that affect height. (Here, as before, the random variable W denotes non-genetic effects.) Although n is fixed, if it is fairly large, then Theorem 9.5 implies that the sum $X_1 + X_2 + \cdots + X_n$ is approximately normally distributed. Now, if we assume that the X_i's have a significantly larger cumulative effect than W does, then H is approximately normally distributed.

Another observed feature of the distribution of heights of adults of one sex in

a population is that the variance does not seem to increase or decrease from one generation to the next. This was known at the time of Galton, and his attempts to explain this led him to the idea of regression to the mean. This idea will be discussed further in the historical remarks at the end of the section. (The reason that we only consider one sex is that human heights are clearly sex-linked, and in general, if we have two populations that are each normally distributed, then their union need not be normally distributed.)

Using the multiple-gene hypothesis, it is easy to explain why the variance should be constant from generation to generation. We begin by assuming that for a specific gene location, there are k alleles, which we will denote by $A_1,\ A_2,\ \ldots,\ A_k$. We assume that the offspring are produced by random mating. By this we mean that given any offspring, it is equally likely that it came from any pair of parents in the preceding generation. There is another way to look at random mating that makes the calculations easier. We consider the set S of all of the alleles (at the given gene location) in all of the germ cells of all of the individuals in the parent generation. In terms of the set S, by random mating we mean that each pair of alleles in S is equally likely to reside in any particular offspring. (The reader might object to this way of thinking about random mating, as it allows two alleles from the same parent to end up in an offspring; but if the number of individuals in the parent population is large, then whether or not we allow this event does not affect the probabilities very much.)

For $1 \leq i \leq k$, we let p_i denote the proportion of alleles in the parent population that are of type A_i. It is clear that this is the same as the proportion of alleles in the germ cells of the parent population, assuming that each parent produces roughly the same number of germs cells. Consider the distribution of alleles in the offspring. Since each germ cell is equally likely to be chosen for any particular offspring, the distribution of alleles in the offspring is the same as in the parents.

We next consider the distribution of genotypes in the two generations. We will prove the following fact: the distribution of genotypes in the offspring generation depends only upon the distribution of alleles in the parent generation (in particular, it does not depend upon the distribution of genotypes in the parent generation). Consider the possible genotypes; there are $k(k+1)/2$ of them. Under our assumptions, the genotype A_iA_i will occur with frequency p_i^2, and the genotype A_iA_j, with $i \neq j$, will occur with frequency $2p_ip_j$. Thus, the frequencies of the genotypes depend only upon the allele frequencies in the parent generation, as claimed.

This means that if we start with a certain generation, and a certain distribution of alleles, then in all generations after the one we started with, both the allele distribution and the genotype distribution will be fixed. This last statement is known as the Hardy-Weinberg Law.

We can describe the consequences of this law for the distribution of heights among adults of one sex in a population. We recall that the height of an offspring was given by a random variable H, where

$$H = X_1 + X_2 + \cdots + X_n + W\ ,$$

with the X_i's corresponding to the genes that affect height, and the random variable

W denoting non-genetic effects. The Hardy-Weinberg Law states that for each X_i, the distribution in the offspring generation is the same as the distribution in the parent generation. Thus, if we assume that the distribution of W is roughly the same from generation to generation (or if we assume that its effects are small), then the distribution of H is the same from generation to generation. (In fact, dietary effects are part of W, and it is clear that in many human populations, diets have changed quite a bit from one generation to the next in recent times. This change is thought to be one of the reasons that humans, on the average, are getting taller. It is also the case that the effects of W are thought to be small relative to the genetic effects of the parents.)

Discussion

Generally speaking, the Central Limit Theorem contains more information than the Law of Large Numbers, because it gives us detailed information about the *shape* of the distribution of S_n^*; for large n the shape is approximately the same as the shape of the standard normal density. More specifically, the Central Limit Theorem says that if we standardize and height-correct the distribution of S_n, then the normal density function is a very good approximation to this distribution when n is large. Thus, we have a computable approximation for the distribution for S_n, which provides us with a powerful technique for generating answers for all sorts of questions about sums of independent random variables, even if the individual random variables have different distributions.

Historical Remarks

In the mid-1800's, the Belgian mathematician Quetelet[7] had shown empirically that the normal distribution occurred in real data, and had also given a method for fitting the normal curve to a given data set. Laplace[8] had shown much earlier that the sum of many independent identically distributed random variables is approximately normal. Galton knew that certain physical traits in a population appeared to be approximately normally distributed, but he did not consider Laplace's result to be a good explanation of how this distribution comes about. We give a quote from Galton that appears in the fascinating book by S. Stigler[9] on the history of statistics:

> First, let me point out a fact which Quetelet and all writers who have followed in his paths have unaccountably overlooked, and which has an intimate bearing on our work to-night. It is that, although characteristics of plants and animals conform to the law, the reason of their doing so is as yet totally unexplained. The essence of the law is that differences should be wholly due to the collective actions of a host of independent *petty* influences in various combinations...Now the processes of heredity...are not petty influences, but very important ones...The conclusion

[7] S. Stigler, *The History of Statistics,* (Cambridge: Harvard University Press, 1986), p. 203.
[8] ibid., p. 136
[9] ibid., p. 281.

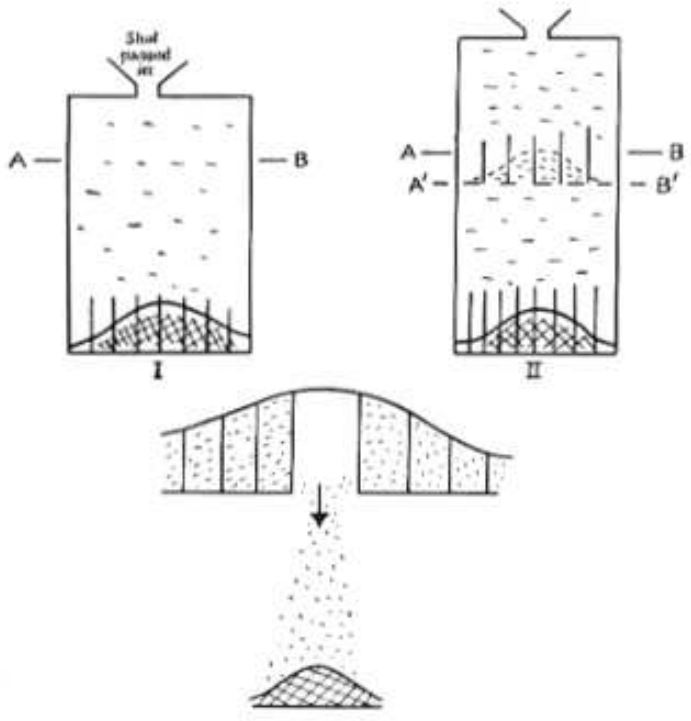


Figure 9.11: Two-stage version of the quincunx.

> is...that the processes of heredity must work harmoniously with the law of deviation, and be themselves in some sense conformable to it.

Galton invented a device known as a quincunx (now commonly called a Galton board), which we used in Example 3.10 to show how to physically obtain a binomial distribution. Of course, the Central Limit Theorem says that for large values of the parameter n, the binomial distribution is approximately normal. Galton used the quincunx to explain how inheritance affects the distribution of a trait among offspring.

We consider, as Galton did, what happens if we interrupt, at some intermediate height, the progress of the shot that is falling in the quincunx. The reader is referred to Figure 9.11. This figure is a drawing of Karl Pearson,[10] based upon Galton's notes. In this figure, the shot is being temporarily segregated into compartments at the line AB. (The line A′B′ forms a platform on which the shot can rest.) If the line AB is not too close to the top of the quincunx, then the shot will be approximately normally distributed at this line. Now suppose that one compartment is opened, as shown in the figure. The shot from that compartment will fall, forming a normal distribution at the bottom of the quincunx. If now all of the compartments are

[10]Karl Pearson, *The Life, Letters and Labours of Francis Galton,* vol. IIIB, (Cambridge at the University Press 1930.) p. 466. Reprinted with permission.

opened, all of the shot will fall, producing the same distribution as would occur if the shot were not temporarily stopped at the line AB. But the action of stopping the shot at the line AB, and then releasing the compartments one at a time, is just the same as convoluting two normal distributions. The normal distributions at the bottom, corresponding to each compartment at the line AB, are being mixed, with their weights being the number of shot in each compartment. On the other hand, it is already known that if the shot are unimpeded, the final distribution is approximately normal. Thus, this device shows that the convolution of two normal distributions is again normal.

Galton also considered the quincunx from another perspective. He segregated into seven groups, by weight, a set of 490 sweet pea seeds. He gave 10 seeds from each of the seven group to each of seven friends, who grew the plants from the seeds. Galton found that each group produced seeds whose weights were normally distributed. (The sweet pea reproduces by self-pollination, so he did not need to consider the possibility of interaction between different groups.) In addition, he found that the variances of the weights of the offspring were the same for each group. This segregation into groups corresponds to the compartments at the line AB in the quincunx. Thus, the sweet peas were acting as though they were being governed by a convolution of normal distributions.

He now was faced with a problem. We have shown in Chapter 7, and Galton knew, that the convolution of two normal distributions produces a normal distribution with a larger variance than either of the original distributions. But his data on the sweet pea seeds showed that the variance of the offspring population was the same as the variance of the parent population. His answer to this problem was to postulate a mechanism that he called *reversion*, and is now called *regression to the mean*. As Stigler puts it:[11]

> The seven groups of progeny were normally distributed, but not about their parents' weight. Rather they were in every case distributed about a value that was closer to the average population weight than was that of the parent. Furthermore, this reversion followed "the simplest possible law," that is, it was linear. The average deviation of the progeny from the population average was in the same direction as that of the parent, but only a third as great. The mean progeny reverted to type, and the increased variation was just sufficient to maintain the population variability.

Galton illustrated reversion with the illustration shown in Figure 9.12.[12] The parent population is shown at the top of the figure, and the slanted lines are meant to correspond to the reversion effect. The offspring population is shown at the bottom of the figure.

[11] ibid., p. 282.

[12] Karl Pearson, *The Life, Letters and Labours of Francis Galton,* vol. IIIA, (Cambridge at the University Press 1930.) p. 9. Reprinted with permission.

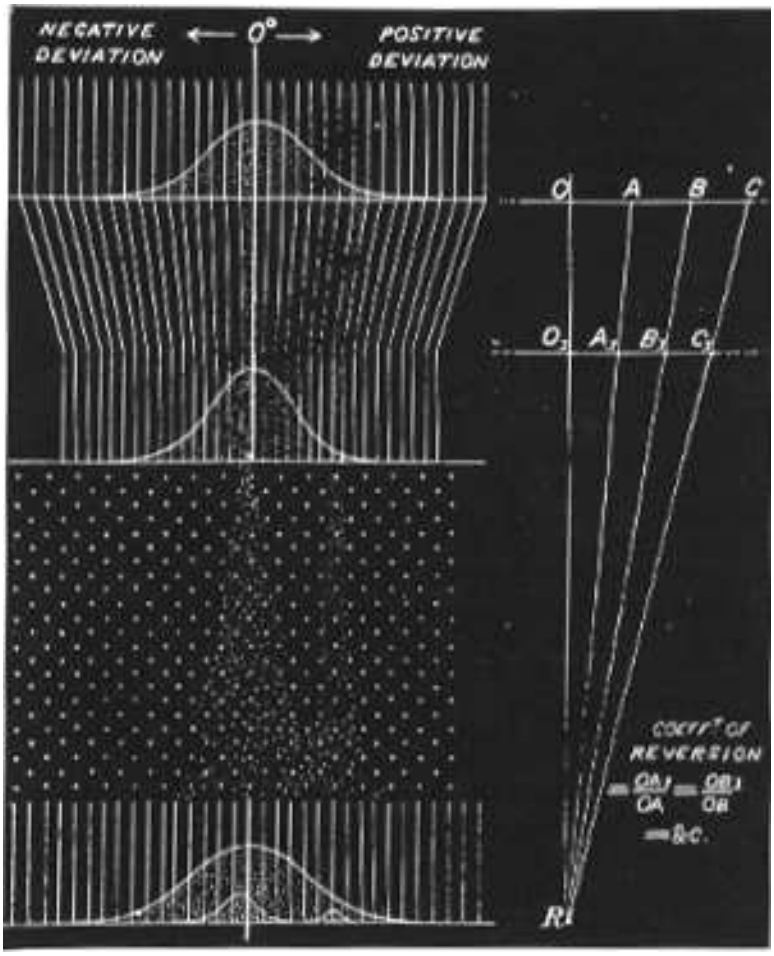


Figure 9.12: Galton's explanation of reversion.

Exercises

1 A die is rolled 24 times. Use the Central Limit Theorem to estimate the probability that

(a) the sum is greater than 84.

(b) the sum is equal to 84.

2 A random walker starts at 0 on the x-axis and at each time unit moves 1 step to the right or 1 step to the left with probability 1/2. Estimate the probability that, after 100 steps, the walker is more than 10 steps from the starting position.

3 A piece of rope is made up of 100 strands. Assume that the breaking strength of the rope is the sum of the breaking strengths of the individual strands. Assume further that this sum may be considered to be the sum of an independent trials process with 100 experiments each having expected value of 10 pounds and standard deviation 1. Find the approximate probability that the rope will support a weight

(a) of 1000 pounds.

(b) of 970 pounds.

4 Write a program to find the average of 1000 random digits 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. Have the program test to see if the average lies within three standard deviations of the expected value of 4.5. Modify the program so that it repeats this simulation 1000 times and keeps track of the number of times the test is passed. Does your outcome agree with the Central Limit Theorem?

5 A die is thrown until the first time the total sum of the face values of the die is 700 or greater. Estimate the probability that, for this to happen,

(a) more than 210 tosses are required.

(b) less than 190 tosses are required.

(c) between 180 and 210 tosses, inclusive, are required.

6 A bank accepts rolls of pennies and gives 50 cents credit to a customer without counting the contents. Assume that a roll contains 49 pennies 30 percent of the time, 50 pennies 60 percent of the time, and 51 pennies 10 percent of the time.

(a) Find the expected value and the variance for the amount that the bank loses on a typical roll.

(b) Estimate the probability that the bank will lose more than 25 cents in 100 rolls.

(c) Estimate the probability that the bank will lose exactly 25 cents in 100 rolls.

(d) Estimate the probability that the bank will lose any money in 100 rolls.

(e) How many rolls does the bank need to collect to have a 99 percent chance of a net loss?

7 A surveying instrument makes an error of -2, -1, 0, 1, or 2 feet with equal probabilities when measuring the height of a 200-foot tower.

(a) Find the expected value and the variance for the height obtained using this instrument once.

(b) Estimate the probability that in 18 independent measurements of this tower, the average of the measurements is between 199 and 201, inclusive.

8 For Example 9.6 estimate $P(S_{30} = 0)$. That is, estimate the probability that the errors cancel out and the student's grade point average is correct.

9 Prove the Law of Large Numbers using the Central Limit Theorem.

10 Peter and Paul match pennies 10,000 times. Describe briefly what each of the following theorems tells you about Peter's fortune.

(a) The Law of Large Numbers.

(b) The Central Limit Theorem.

11 A tourist in Las Vegas was attracted by a certain gambling game in which the customer stakes 1 dollar on each play; a win then pays the customer 2 dollars plus the return of her stake, although a loss costs her only her stake. Las Vegas insiders, and alert students of probability theory, know that the probability of winning at this game is 1/4. When driven from the tables by hunger, the tourist had played this game 240 times. Assuming that no near miracles happened, about how much poorer was the tourist upon leaving the casino? What is the probability that she lost no money?

12 We have seen that, in playing roulette at Monte Carlo (Example 6.13), betting 1 dollar on red or 1 dollar on 17 amounts to choosing between the distributions

$$m_X = \begin{pmatrix} -1 & -1/2 & 1 \\ 18/37 & 1/37 & 18/37 \end{pmatrix}$$

or

$$m_X = \begin{pmatrix} -1 & 35 \\ 36/37 & 1/37 \end{pmatrix}$$

You plan to choose one of these methods and use it to make 100 1-dollar bets using the method chosen. Using the Central Limit Theorem, estimate the probability of winning any money for each of the two games. Compare your estimates with the actual probabilities, which can be shown, from exact calculations, to equal .437 and .509 to three decimal places.

13 In Example 9.6 find the largest value of p that gives probability .954 that the first decimal place is correct.

14 It has been suggested that Example 9.6 is unrealistic, in the sense that the probabilities of errors are too low. Make up your own (reasonable) estimate for the distribution $m(x)$, and determine the probability that a student's grade point average is accurate to within .05. Also determine the probability that it is accurate to within .5.

15 Find a sequence of uniformly bounded discrete independent random variables $\{X_n\}$ such that the variance of their sum does not tend to ∞ as $n \to \infty$, and such that their sum is not asymptotically normally distributed.

9.3 Central Limit Theorem for Continuous Independent Trials

We have seen in Section 9.2 that the distribution function for the sum of a large number n of independent discrete random variables with mean μ and variance σ^2 tends to look like a normal density with mean $n\mu$ and variance $n\sigma^2$. What is remarkable about this result is that it holds for *any* distribution with finite mean and variance. We shall see in this section that the same result also holds true for continuous random variables having a common density function.

Let us begin by looking at some examples to see whether such a result is even plausible.

Standardized Sums

Example 9.7 Suppose we choose n random numbers from the interval $[0, 1]$ with uniform density. Let X_1, X_2, ..., X_n denote these choices, and $S_n = X_1 + X_2 + \cdots + X_n$ their sum.

We saw in Example 7.9 that the density function for S_n tends to have a normal shape, but is centered at $n/2$ and is flattened out. In order to compare the shapes of these density functions for different values of n, we proceed as in the previous section: we *standardize* S_n by defining

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n}\sigma} \ .$$

Then we see that for all n we have

$$\begin{aligned} E(S_n^*) &= 0 \ , \\ V(S_n^*) &= 1 \ . \end{aligned}$$

The density function for S_n^* is just a standardized version of the density function for S_n (see Figure 9.13). □

Example 9.8 Let us do the same thing, but now choose numbers from the interval $[0, +\infty)$ with an exponential density with parameter λ. Then (see Example 6.26)

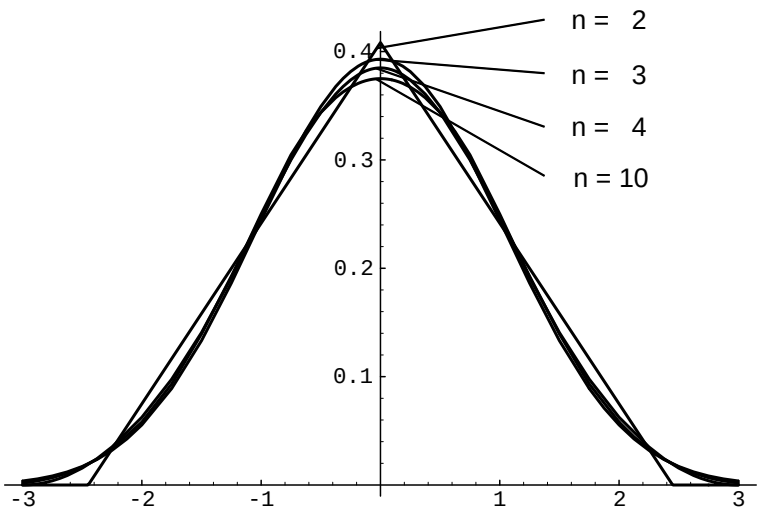


Figure 9.13: Density function for S_n^* (uniform case, $n = 2, 3, 4, 10$).

$$\begin{aligned} \mu &= E(X_i) = \frac{1}{\lambda} , \\ \sigma^2 &= V(X_j) = \frac{1}{\lambda^2} . \end{aligned}$$

Here we know the density function for S_n explicitly (see Section 7.2). We can use Corollary 5.1 to calculate the density function for S_n^*. We obtain

$$\begin{aligned} f_{S_n}(x) &= \frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!} , \\ f_{S_n^*}(x) &= \frac{\sqrt{n}}{\lambda} f_{S_n}\left(\frac{\sqrt{n}x + n}{\lambda}\right) . \end{aligned}$$

The graph of the density function for S_n^* is shown in Figure 9.14. □

These examples make it seem plausible that the density function for the normalized random variable S_n^* for large n will look very much like the normal density with mean 0 and variance 1 in the continuous case as well as in the discrete case. The Central Limit Theorem makes this statement precise.

Central Limit Theorem

Theorem 9.6 (Central Limit Theorem) Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of n independent continuous random variables with common density function p having expected value μ and variance σ^2. Let $S_n^* = (S_n - n\mu)/\sqrt{n}\sigma$. Then we have,

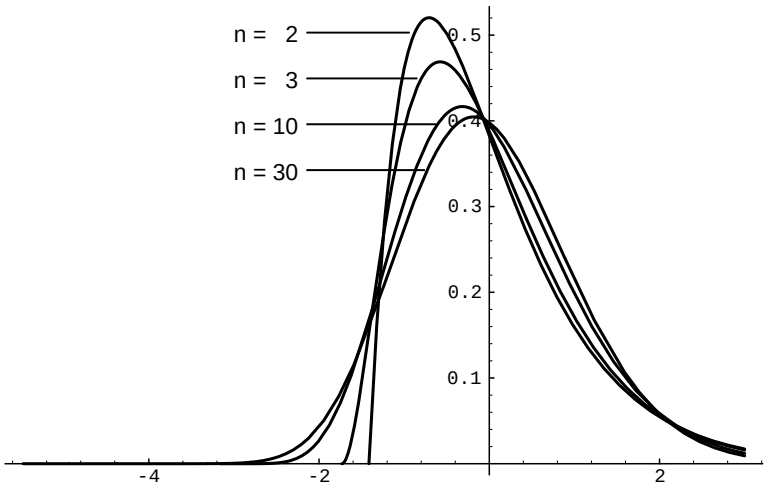


Figure 9.14: Density function for S_n^* (exponential case, $n = 2, 3, 10, 30$, $\lambda = 1$).

for all $a < b$,

$$\lim_{n\to\infty} P(a < S_n^* < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \ .$$

□

We shall give a proof of this theorem in Section 10.3. We will now look at some examples.

Example 9.9 Suppose a surveyor wants to measure a known distance, say of 1 mile, using a transit and some method of triangulation. He knows that because of possible motion of the transit, atmospheric distortions, and human error, any one measurement is apt to be slightly in error. He plans to make several measurements and take an average. He assumes that his measurements are independent random variables with a common distribution of mean $\mu = 1$ and standard deviation $\sigma = .0002$ (so, if the errors are approximately normally distributed, then his measurements are within 1 foot of the correct distance about 65% of the time). What can he say about the average?

He can say that if n is large, the average S_n/n has a density function that is approximately normal, with mean $\mu = 1$ mile, and standard deviation $\sigma = .0002/\sqrt{n}$ miles.

How many measurements should he make to be reasonably sure that his average lies within .0001 of the true value? The Chebyshev inequality says

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq .0001\right) \leq \frac{(.0002)^2}{n(10^{-8})} = \frac{4}{n} \ ,$$

so that we must have $n \geq 80$ before the probability that his error is less than .0001 exceeds .95.

We have already noticed that the estimate in the Chebyshev inequality is not always a good one, and here is a case in point. If we assume that n is large enough so that the density for S_n is approximately normal, then we have

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu\right| < .0001\right) &= P\big(-.5\sqrt{n} < S_n^* < +.5\sqrt{n}\big) \\ &\approx \frac{1}{\sqrt{2\pi}} \int_{-.5\sqrt{n}}^{+.5\sqrt{n}} e^{-x^2/2}\, dx \ , \end{aligned}$$

and this last expression is greater than .95 if $.5\sqrt{n} \geq 2$. This says that it suffices to take $n = 16$ measurements for the same results. This second calculation is stronger, but depends on the assumption that $n = 16$ is large enough to establish the normal density as a good approximation to S_n^*, and hence to S_n. The Central Limit Theorem here says nothing about how large n has to be. In most cases involving sums of independent random variables, a good rule of thumb is that for $n \geq 30$, the approximation is a good one. In the present case, if we assume that the errors are approximately normally distributed, then the approximation is probably fairly good even for $n = 16$. □

Estimating the Mean

Example 9.10 (Continuation of Example 9.9) Now suppose our surveyor is measuring an unknown distance with the same instruments under the same conditions. He takes 36 measurements and averages them. How sure can he be that his measurement lies within .0002 of the true value?

Again using the normal approximation, we get

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu\right| < .0002\right) &= P\big(|S_n^*| < .5\sqrt{n}\big) \\ &\approx \frac{2}{\sqrt{2\pi}} \int_{-3}^{3} e^{-x^2/2}\, dx \\ &\approx .997 \ . \end{aligned}$$

This means that the surveyor can be 99.7 percent sure that his average is within .0002 of the true value. To improve his confidence, he can take more measurements, or require less accuracy, or improve the quality of his measurements (i.e., reduce the variance σ^2). In each case, the Central Limit Theorem gives quantitative information about the confidence of a measurement process, assuming always that the normal approximation is valid.

Now suppose the surveyor does not know the mean or standard deviation of his measurements, but assumes that they are independent. How should he proceed?

Again, he makes several measurements of a known distance and averages them. As before, the average error is approximately normally distributed, but now with unknown mean and variance. □

Sample Mean

If he knows the variance σ^2 of the error distribution is .0002, then he can estimate the mean μ by taking the *average,* or *sample mean* of, say, 36 measurements:

$$\bar{\mu} = \frac{x_1 + x_2 + \cdots + x_n}{n} \ ,$$

where $n = 36$. Then, as before, $E(\bar{\mu}) = \mu$. Moreover, the preceding argument shows that

$$P(|\bar{\mu} - \mu| < .0002) \approx .997 \ .$$

The interval $(\bar{\mu} - .0002, \bar{\mu} + .0002)$ is called *the 99.7% confidence interval* for μ (see Example 9.4).

Sample Variance

If he does not know the variance σ^2 of the error distribution, then he can estimate σ^2 by the *sample variance*:

$$\bar{\sigma}^2 = \frac{(x_1 - \bar{\mu})^2 + (x_2 - \bar{\mu})^2 + \cdots + (x_n - \bar{\mu})^2}{n} \ ,$$

where $n = 36$. The Law of Large Numbers, applied to the random variables $(X_i - \bar{\mu})^2$, says that for large n, the sample variance $\bar{\sigma}^2$ lies close to the variance σ^2, so that the surveyor can use $\bar{\sigma}^2$ in place of σ^2 in the argument above.

Experience has shown that, in most practical problems of this type, the sample variance is a good estimate for the variance, and can be used in place of the variance to determine confidence levels for the sample mean. This means that we can rely on the Law of Large Numbers for estimating the variance, and the Central Limit Theorem for estimating the mean.

We can check this in some special cases. Suppose we know that the error distribution is *normal,* with unknown mean and variance. Then we can take a sample of n measurements, find the sample mean $\bar{\mu}$ and sample variance $\bar{\sigma}^2$, and form

$$T_n^* = \frac{S_n - n\bar{\mu}}{\sqrt{n}\bar{\sigma}} \ ,$$

where $n = 36$. We expect T_n^* to be a good approximation for S_n^* for large n.

t-Density

The statistician W. S. Gosset[13] has shown that in this case T_n^* has a density function that is not normal but rather a *t-density* with n degrees of freedom. (The number n of degrees of freedom is simply a parameter which tells us which t-density to use.) In this case we can use the t-density in place of the normal density to determine confidence levels for μ. As n increases, the t-density approaches the normal density. Indeed, even for $n = 8$ the t-density and normal density are practically the same (see Figure 9.15).

[13]W. S. Gosset discovered the distribution we now call the t-distribution while working for the Guinness Brewery in Dublin. He wrote under the pseudonym "Student." The results discussed here first appeared in Student, "The Probable Error of a Mean," *Biometrika,* vol. 6 (1908), pp. 1-24.

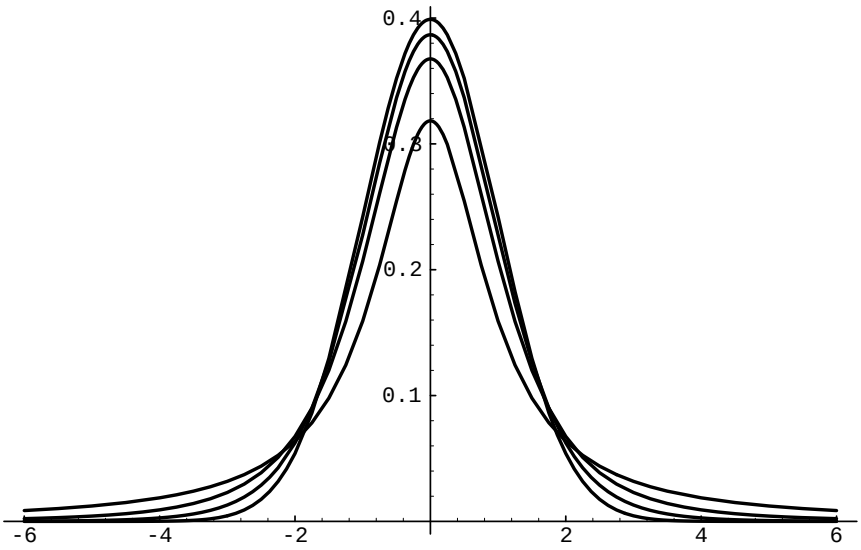


Figure 9.15: Graph of $t-$density for $n = 1, 3, 8$ and the normal density with $\mu = 0, \sigma = 1$.

Exercises

Notes on computer problems:

(a) Simulation: Recall (see Corollary 5.2) that

$$X = F^{-1}(rnd)$$

will simulate a random variable with density $f(x)$ and distribution

$$F(X) = \int_{-\infty}^{x} f(t)\, dt \ .$$

In the case that $f(x)$ is a normal density function with mean μ and standard deviation σ, where neither F nor F^{-1} can be expressed in closed form, use instead

$$X = \sigma\sqrt{-2\log(rnd)}\cos 2\pi(rnd) + \mu \ .$$

(b) Bar graphs: you should aim for about 20 to 30 bars (of equal width) in your graph. You can achieve this by a good choice of the range $[x\text{min}, x\text{min}]$ and the number of bars (for instance, $[\mu - 3\sigma, \mu + 3\sigma]$ with 30 bars will work in many cases). Experiment!

1 Let X be a continuous random variable with mean $\mu(X)$ and variance $\sigma^2(X)$, and let $X^* = (X - \mu)/\sigma$ be its standardized version. Verify directly that $\mu(X^*) = 0$ and $\sigma^2(X^*) = 1$.

2 Let $\{X_k\}$, $1 \le k \le n$, be a sequence of independent random variables, all with mean 0 and variance 1, and let S_n, S_n^*, and A_n be their sum, standardized sum, and average, respectively. Verify directly that $S_n^* = S_n/\sqrt{n} = \sqrt{n}A_n$.

3 Let $\{X_k\}$, $1 \le k \le n$, be a sequence of random variables, all with mean μ and variance σ^2, and $Y_k = X_k^*$ be their standardized versions. Let S_n and T_n be the sum of the X_k and Y_k, and S_n^* and T_n^* their standardized version. Show that $S_n^* = T_n^* = T_n/\sqrt{n}$.

4 Suppose we choose independently 25 numbers at random (uniform density) from the interval $[0, 20]$. Write the normal densities that approximate the densities of their sum S_{25}, their standardized sum S_{25}^*, and their average A_{25}.

5 Write a program to choose independently 25 numbers at random from $[0, 20]$, compute their sum S_{25}, and repeat this experiment 1000 times. Make a bar graph for the density of S_{25} and compare it with the normal approximation of Exercise 4. How good is the fit? Now do the same for the standardized sum S_{25}^* and the average A_{25}.

6 In general, the Central Limit Theorem gives a better estimate than Chebyshev's inequality for the average of a sum. To see this, let A_{25} be the average calculated in Exercise 5, and let N be the normal approximation for A_{25}. Modify your program in Exercise 5 to provide a table of the function $F(x) = P(|A_{25} - 10| \ge x) =$ fraction of the total of 1000 trials for which $|A_{25} - 10| \ge x$. Do the same for the function $f(x) = P(|N - 10| \ge x)$. (You can use the normal table, Table 9.4, or the procedure **NormalArea** for this.) Now plot on the same axes the graphs of $F(x)$, $f(x)$, and the Chebyshev function $g(x) = 4/(3x^2)$. How do $f(x)$ and $g(x)$ compare as estimates for $F(x)$?

7 The Central Limit Theorem says the sums of independent random variables tend to look normal, no matter what crazy distribution the individual variables have. Let us test this by a computer simulation. Choose independently 25 numbers from the interval $[0, 1]$ with the probability density $f(x)$ given below, and compute their sum S_{25}. Repeat this experiment 1000 times, and make up a bar graph of the results. Now plot on the same graph the density $\phi(x) =$ normal $(x, \mu(S_{25}), \sigma(S_{25}))$. How well does the normal density fit your bar graph in each case?

(a) $f(x) = 1$.

(b) $f(x) = 2x$.

(c) $f(x) = 3x^2$.

(d) $f(x) = 4|x - 1/2|$.

(e) $f(x) = 2 - 4|x - 1/2|$.

8 Repeat the experiment described in Exercise 7 but now choose the 25 numbers from $[0, \infty)$, using $f(x) = e^{-x}$.

9 How large must n be before $S_n = X_1 + X_2 + \cdots + X_n$ is approximately normal? This number is often surprisingly small. Let us explore this question with a computer simulation. Choose n numbers from $[0, 1]$ with probability density $f(x)$, where $n = 3$, 6, 12, 20, and $f(x)$ is each of the densities in Exercise 7. Compute their sum S_n, repeat this experiment 1000 times, and make up a bar graph of 20 bars of the results. How large must n be before you get a good fit?

10 A surveyor is measuring the height of a cliff known to be about 1000 feet. He assumes his instrument is properly calibrated and that his measurement errors are independent, with mean $\mu = 0$ and variance $\sigma^2 = 10$. He plans to take n measurements and form the average. Estimate, using (a) Chebyshev's inequality and (b) the normal approximation, how large n should be if he wants to be 95 percent sure that his average falls within 1 foot of the true value. Now estimate, using (a) and (b), what value should σ^2 have if he wants to make only 10 measurements with the same confidence?

11 The price of one share of stock in the Pilsdorff Beer Company (see Exercise 8.2.12) is given by Y_n on the nth day of the year. Finn observes that the differences $X_n = Y_{n+1} - Y_n$ appear to be independent random variables with a common distribution having mean $\mu = 0$ and variance $\sigma^2 = 1/4$. If $Y_1 = 100$, estimate the probability that Y_{365} is

(a) ≥ 100.

(b) ≥ 110.

(c) ≥ 120.

12 Test your conclusions in Exercise 11 by computer simulation. First choose 364 numbers X_i with density $f(x) = \text{normal}(x, 0, 1/4)$. Now form the sum $Y_{365} = 100 + X_1 + X_2 + \cdots + X_{364}$, and repeat this experiment 200 times. Make up a bar graph on $[50, 150]$ of the results, superimposing the graph of the approximating normal density. What does this graph say about your answers in Exercise 11?

13 Physicists say that particles in a long tube are constantly moving back and forth along the tube, each with a velocity V_k (in cm/sec) at any given moment that is normally distributed, with mean $\mu = 0$ and variance $\sigma^2 = 1$. Suppose there are 10^{20} particles in the tube.

(a) Find the mean and variance of the average velocity of the particles.

(b) What is the probability that the average velocity is $\geq 10^{-9}$ cm/sec?

14 An astronomer makes n measurements of the distance between Jupiter and a particular one of its moons. Experience with the instruments used leads her to believe that for the proper units the measurements will be normally

distributed with mean d, the true distance, and variance 16. She performs a series of n measurements. Let

$$A_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

be the average of these measurements.

(a) Show that

$$P\left(A_n - \frac{8}{\sqrt{n}} \leq d \leq A_n + \frac{8}{\sqrt{n}}\right) \approx .95.$$

(b) When nine measurements were taken, the average of the distances turned out to be 23.2 units. Putting the observed values in (a) gives the *95 percent confidence interval* for the unknown distance d. Compute this interval.

(c) Why not say in (b) more simply that the probability is .95 that the value of d lies in the computed confidence interval?

(d) What changes would you make in the above procedure if you wanted to compute a 99 percent confidence interval?

15 Plot a bar graph similar to that in Figure 9.10 for the heights of the mid-parents in Galton's data as given in Appendix B and compare this bar graph to the appropriate normal curve.