

Chapter 4

Limit Theorems

4.1 Laws of Large Numbers

If X_1, X_2, \dots are independent and have the same distribution then we say the X_i are **independent and identically distributed**, or i.i.d. for short. Such sequences arise if we repeat some experiment such as flipping a coin or rolling a die, or if we stop people at random and measure their height or ask them how they will vote in an upcoming election. As we mentioned at the beginning of Chapter 3, if X_1, X_2, \dots are i.i.d. with $EX_i = \mu$ then when n is large, the average of the first n observations, $(X_1 + \dots + X_n)/n$, will be close to EX with high probability.

Our first goal in this section will be to prove this result, which is called the **law of large numbers**. Let

$$\bar{X}_n = (X_1 + \dots + X_n)/n$$

\bar{X}_n is called the **sample mean** because if we assigned probability $1/n$ to each of the first n observations then \bar{X}_n would be the mean of that distribution. If we suppose that the X_i are i.i.d. with $EX_i = \mu$ then using the facts that $E(cY) = cEY$ and the expected value of the sum is the sum of the expected values, we have

$$\begin{aligned} E\bar{X}_n &= \frac{1}{n}E(X_1 + \dots + X_n) \\ &= \frac{1}{n}\{EX_1 + \dots + EX_n\} = \mu \end{aligned} \tag{4.1}$$

If we suppose that $\text{var}(X_i) = \sigma^2$ then using the facts that $\text{var}(cY) = c^2 \text{var}(Y)$ and that for independent X_1, \dots, X_n the variance of the sum is the sum of the variances, we have

$$\text{var}(\bar{X}_n) = \frac{1}{n^2} \text{var}(X_1 + \dots + X_n)$$

$$= \frac{1}{n^2} \{ \text{var}(X_1) + \cdots + \text{var}(X_n) \} = \sigma^2/n \quad (4.2)$$

Taking square roots we see that the standard deviation of \bar{X}_n is σ/\sqrt{n} . We have earlier called this the size of a typical deviation from the mean. The key to proving the law of large numbers is to show that the probability of a deviation that is a large number of standard deviations is small. To motivate the inequality we will use to prove this, we consider a

Puzzle. Suppose $EX = 0$ and $EX^2 = 1$. How large can $P(|X| \geq 3)$ be?

Solution. On the set $\{|X| \geq 3\}$, $X^2 \geq 9$. Since $X^2 \geq 0$, EX^2 must be larger than what we get from considering only values with $|X| \geq 3$. That is,

$$1 = EX^2 \geq 9P(|X| \geq 3)$$

or $P(|X| \geq 3) \leq 1/9$. To see that this can be achieved we let $P(X = 3) = 1/18$, $P(X = -3) = 1/18$, $P(X = 0) = 8/9$ and note that

$$\begin{aligned} EX &= 3 \cdot \frac{1}{18} + (-3) \cdot \frac{1}{18} = 0 \\ EX^2 &= 9 \cdot \frac{1}{18} + 9 \cdot \frac{1}{18} = 1 \end{aligned}$$

Generalizing leads to

Chebyshev's inequality. If $y > 0$ then

$$P(|Y - EY| \geq y) \leq \text{var}(Y)/y^2 \quad (4.3)$$

Proof. Again since $|Y - EY|^2 \geq 0$, $E|Y - EY|^2$ must be larger than what we get from considering only values with $|Y - EY| \geq y$ so

$$\text{var}(Y) = E|Y - EY|^2 \geq y^2 P(|Y - EY| \geq y)$$

and rearranging gives the inequality. \square

If we let $\sigma^2 = \text{var}(Y)$ and take $y = k\sigma$ with $k \geq 1$ then (4.3) implies that

$$P(|Y - EY| \geq k\sigma) \leq 1/k^2 \quad (4.4)$$

This reinforces our notion that σ is the size of the typical deviation from the mean by showing that a deviation of k standard deviations has probability smaller than $1/k^2$.

Proof of the law of large numbers. Let $Y = \bar{X}_n$. (4.1) implies $EY = \mu$ and (4.2) implies $\text{var}(Y) = \sigma^2/n$ so using Chebyshev's inequality, we see that if $\epsilon > 0$ then

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\text{var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2 n} \rightarrow 0 \quad (4.5)$$

as $n \rightarrow \infty$.

(4.5) could be called the fundamental theorem of statistics because it says that the sample mean is close to the mean μ of the underlying population when the sample is large. The last conclusion does not rule out the possibility that the sequence of sample means $\bar{X}_1, \bar{X}_2, \dots$ stays close to EX most of the time but occasionally wanders off because of a streak of bad luck. Our next result says that this does not happen.

Strong law of large numbers. Suppose X_1, X_2, \dots are i.i.d. with $E|X_i| < \infty$. Then with probability one the sequence of numbers \bar{X}_n converges to EX_i as $n \rightarrow \infty$.

The first thing we have to explain is the phrase “with probability one.” To do this we first consider flipping a coin and letting X_i be 1 if the i th toss results in Heads, and 0 otherwise. The strong law of large numbers says that with probability one

$$(X_1 + \dots + X_n)/n \rightarrow 1/2 \quad \text{as } n \rightarrow \infty$$

It is easy to write down sequences of tosses for which this is false:

$$H, H, T, H, H, TH, H, TH, H, T, \dots$$

However, the strong law of large numbers implies that the collection of “bad sequences” (i.e., those for which the asymptotic frequency of Heads is not $1/2$) has probability zero.

To see what the law of large numbers says we turn to simulation. Figure 5.1 shows the fraction of heads versus time in three simulations of flipping a fair coin. As predicted the frequency of heads is approaching $1/2$ in each case. Figure 5.2 shows three simulations of a person playing roulette 1000 times and betting \$1 on black each time. The straightline gives the mean loss which is $-1/19$ of a dollar per play. After 1000 plays this is \$52.63 but the actual amounts lost vary from \$10 to \$100 in the three simulations.

4.2 Central Limit Theorem

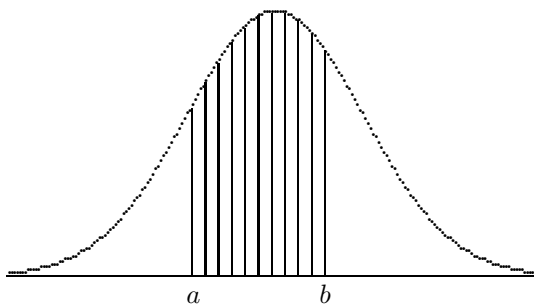
The limit theorem in this section gets its name not only from the fact that it is of central importance but also because it shows that if you add up a large number of random variables with a fixed distribution with finite variance then, if we subtract the mean and divide by the standard deviation the result has approximately a normal distribution.

To show that this is true and to motivate the formal statement, Figure 4.3 gives the distribution of the number of heads in 100 coin flips, Figure 4.4 gives the distribution of the sum of 100 dice, and Figure 4.5 gives the distribution of the net winnings in 100 plays of roulette for a player betting on black. In each case the distribution has the same shape.

Central limit theorem. *Suppose X_1, X_2, \dots are i.i.d. and have $EX_i = \mu$ and $\text{var}(X_i) = \sigma^2$ with $0 < \sigma^2 < \infty$. Let $S_n = X_1 + \dots + X_n$. As $n \rightarrow \infty$*

$$P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \rightarrow \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \quad (4.6)$$

The function $f(x) = (2\pi)^{-1/2} e^{-x^2/2}$ is the normal density function.



A random variable χ is said to have this distribution if for all $a \leq b$ we have

$$P(a \leq \chi \leq b) = \int_a^b f(x) dx \quad (4.7)$$

Geometrically, $P(a \leq \chi \leq b)$ is the area under the curve f between a and b . Letting $a \rightarrow -\infty$ in the last formula we have

$$P(\chi \leq b) = \int_{-\infty}^b (2\pi)^{-1/2} e^{-x^2/2} dx$$

This is the normal distribution function. Calculus cannot be used to evaluate the integral so must instead use a table like the one at the back of the book. The next example illustrates the use of (4.6) and the table.

Example 4.1. Suppose we flip a coin 900 times. What is the probability we get at least 465 heads?

The mean number of heads $n/2 = 450$ and the standard deviation $\sqrt{n}/2 = 15$ so (4.6) implies

$$P(S_n \geq 465) = P\left(\frac{S_n - 450}{15} \geq \frac{15}{15}\right) \approx P(\chi \geq 1) = 1 - 0.8413 = 0.1587$$

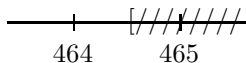
If the question in the problem had been formulated as “What is the probability of at most 464 heads?” we would have computed

$$P(S_n \leq 464) = P\left(\frac{S_n - 450}{15} \leq \frac{14}{15}\right) \approx P(\chi \leq 0.933) = 0.8238$$

which does not quite agree with our first answer since

$$0.8238 + (1 - 0.8413) = 0.9825 < 1$$

whereas $P(S_n \leq 464) + P(S_n \geq 465) = 1$. The solution to this problem is to regard $\{S_n \geq 465\}$ as $\{S_n \geq 464.5\}$, that is, the integers 464 and 465 split up the territory that lies between them.



When we do this, the answer to our original question becomes

$$\begin{aligned} P(S_n \geq 464.5) &= P\left(\frac{S_n - 450}{15} \geq \frac{14.5}{15}\right) \\ &\approx P(\chi \geq 0.966) = 1 - 0.8340 = 0.1660 \end{aligned}$$

which is a much better approximation of the exact probability 0.1669 than was our first answer, 0.1587.

The last correction, which is called the **histogram correction**, should be used whenever we apply (4.6) to integer-valued random variables. As we did in the last example, if k is an integer we regard $P(S_n \geq k)$ as $P(S_n \geq k - 0.5)$ and $P(S_n \leq k)$ as $P(S_n \leq k + 0.5)$. More generally, we replace each integer k in the set of interest by the interval $[k - 0.5, k + 0.5]$. The next example shows that the histogram correction is not only a device to get more accurate estimates, it also allows us to get answers in cases where a naive application of the central limit theorem would give a senseless answer.

Example 4.2. Suppose we flip 16 coins. Use (4.6) to estimate the probability that we get exactly 8 heads.

The mean number of heads is $n/2 = 16$ while the standard deviation is $\sqrt{n}/2 = 2$. To use the normal approximation we write $\{S_{16} = 8\}$ as $\{7.5 \leq S_{16} \leq 8.5\}$. By (2.1)

$$P\left(\frac{7.5 - 8}{2} \leq \frac{S_{16} - 8}{2} \leq \frac{8.5 - 8}{2}\right) \approx P(-0.25 \leq \chi \leq 0.25)$$

Since $P(\chi = -0.25) = 0$, the probability of interest is $P(\chi \leq 0.25) - P(\chi \leq -0.25)$. The table tells us that $P(\chi \leq 0.25) = 0.5987$. There are no negative numbers in the table but the normal distribution is symmetric so

$$P(\chi \leq -0.25) = P(\chi \geq 0.25) = 1 - P(\chi \leq 0.25) = 1 - 0.5987 = 0.4013$$

and we have

$$P(S_{16} = 8) \approx 0.5987 - 0.4013 = 0.1974$$

The exact answer is

$$2^{-16} \frac{16!}{8!8!} = 0.1964$$

Similar reasoning shows that

$$P(S_{16} = 8 + k) \approx P(\chi \leq (k + 1/2)/2) - P(\chi \leq (k - 1/2)/2)$$

As the next table shows these probabilities are fairly close to the exact answers obtained from the binomial distribution.

k	9	10	11	12
normal approx.	.1747	.1209	.0616	.0319
exact ans.	.1746	.1221	.0666	.0277

Example 4.3. Suppose we roll a die 420 times. What is the probability that the sum of the numbers $S_{420} \geq 1500$?

The first step in the solution is to compute the mean and variance of S_{420} . $ES_{420} = 420 \cdot 7/2 = 1470$. Since $420 = 35 \cdot 12$ $\text{var}(S_{420}) = 420 \cdot 35/12 = 35^2$ and the standard deviation is 35. Using (4.6) now we have

$$\begin{aligned} P(S_{420} \geq 1499.5) &= P\left(\frac{S_{420} - 1470}{35} \geq 0.84\right) \\ &\approx P(\chi \geq 0.84) = 1 - P(\chi \leq 0.84) = 0.2005 \end{aligned}$$

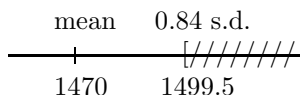
z-score. The key to finding the solution of this and the previous two problems is to compute the number of standard deviations separating the observed value from the expected value. The z-score is defined by

$$z = \frac{\text{observed value} - \text{expected value}}{\text{standard deviation}}$$

In the preceding example this is

$$z = \frac{1499.5 - 1470}{35} = 0.84$$

so the normal approximation is $P(\chi \geq 0.84)$. Here a picture is worth a hundred words.

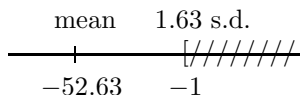


Example 4.4. Consider now the roulette simulations in Figure 4.2. If we bet on black on 1000 times what is the probability our net winnings are ≥ 0 ?

In this case the outcome of the i th play $P(X_i = 1) = 18/38$ and $P(X_i = -1) = 20/38$ so $EX_i = -1/19 = -0.05263$, $EX_i^2 = 1$ and $\text{var}(X_i) = 1 - (1/19)^2 \approx 1$. The mean of 1000 plays is -52.63 , while the standard deviation is $\sqrt{1000} = 31.62$. Writing ≥ 0 as ≥ -1 since only even numbers are possible values for S_{1000} , the z -score is

$$z = \frac{-1 - (-52.63)}{31.62} = 1.63$$

so the normal approximation is $P(\chi \geq 1.63) = 1 - P(\chi \leq 1.63) = 0.0516$



In this book we have restricted our attention to discrete random variables, but the central limit theorem applies equally well to continuous random variables.

Example 4.5. A manufacturing plant produces boxes of biscuit mix that are one pound (454 grams). However due to the poor flow properties of the powder the standard deviation of the weight of a box is 10 grams. A sample of 25 boxes had an average weight of 449.4 grams. Does this indicate a problem with the manufacturing process?

The average weight \bar{X}_{25} has mean 454 grams and standard deviation $\sigma/\sqrt{n} = 10/\sqrt{25} = 2$. The observed weight is 4.6 grams below the mean, which is 2.3 standard deviations. From the normal table $P(\chi \geq 2.3) = 1 - 0.9893 = 0.0107$, so a deviation this large by chance is rather unlikely.

Note: Continuous random variables can take on any value so there is no “histogram correction.”

Example 4.6. Suppose that the average weight of a person is 182 pounds with a standard deviation of 40 pounds. A large plane can hold 400 people. What is the probability the total weight of the people, S_{400} will be more than 75,000 pounds?

The expected value of S_{400} is $400 \cdot 182 = 72,800$. The standard deviation is $\sigma\sqrt{n} = 40 \cdot 20 = 800$.

$$\begin{aligned} P(S_{400} \geq 75,000) &= P\left(\frac{S_{420} - 72,800}{800} \geq 2.75\right) \\ &\approx P(\chi \geq 2.75) = 1 - P(\chi \leq 2.75) = 0.003 \end{aligned}$$

In designing airplanes one cannot afford to make a mistake 3 times out of a 1000 if the error will have disastrous consequences like a crash. Our table stops at 3.09. For larger values one can use the following approximation

$$\int_x^\infty e^{-y^2/2} dy \leq \frac{1}{x} e^{-x^2/2} \quad (4.8)$$

Proof. Multiplying by y/x which is ≥ 1 when $y \geq x$ we have

$$\begin{aligned} \int_x^\infty e^{-y^2/2} dy &\leq \frac{1}{x} \int_x^\infty ye^{-y^2/2} dy \\ &= \frac{1}{x} (-e^{-y^2/2}) \Big|_x^\infty = \frac{1}{x} e^{-x^2/2} \end{aligned}$$

which proves the desired estimate. \square

From this we see that the probability that the total weight is more than 77,600 pounds, which is six standard deviations above the mean is at most

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{6} e^{-18} \approx 1 \times 10^{-9}$$

For our last example we return to the discrete setting.

Example 4.7. *Each year in Mythica, an average of 64 letter carriers are bitten by dogs. In the past year, 88 incidents were reported. Is this number exceptionally high?*

Assuming that dog bites are a rare event, we will use the Poisson distribution for the number of dog bites. As we observed in Example 3.37, a Poisson with mean 64 is the sum of 64 independent Poisson mean 1 random variables, so we can use the Normal to approximate the Poisson. The mean is 64 while the standard deviation is $\sqrt{64}$. Writing the observed event as ≥ 87.5 we see that this is $(87.5 - 64)/8 = 2.94$ standard deviations above the mean, so the normal approximation is $P(\chi \geq 2.94) = 1 - 0.9982 = 0.0018$ so this is an unusual event.

4.3 Hypothesis Testing

The common feature of the examples in this section is that we have a belief about the state of nature and we ask if the data we have observed is consistent with this belief.

Example 4.8 (Spinning a penny). *Part of the probability folklore (see e.g., page 75 of Paulos' 1955 book, *A Mathematician Reads the Newspaper*) that if you spin a penny the result is heads about 30% of the time. To test this out Sally Sievers 1999 probability class at Wells College spun 650 pennies and got 321 heads. What is the probability we would get this many heads if the true probability was $p = 0.3$?*

If S_{650} is the number of heads then the mean is $np = 650 \cdot (0.3) = 195$ and the standard deviation is $\sqrt{np(1-p)} = \sqrt{650(0.3)(0.7)} = 11.7$. The event of interest which with the histogram correction translates into ≥ 320.5 represents a deviation of $(320.5 - 195)/11.7 = 10.7$ standard deviations, so it is extremely unlikely that this claim is correct.

As reported in *Chance News*, Peter Doyle from Dartmouth tried this experiment as well. After getting 953 heads in 2000 trials, he wrote to Princeton mathematician John Conway who responded "My guess is that you didn't pay much attention to the surface you spun 'em on — it makes a LOT of difference. In my own investigation — the celebrated Burger King study — 50 pennies were spun 20 times over and I got almost exactly 2:1 tails to heads. I didn't keep the records of the individual runs of 50, but remember that they got slowly better, showing that you can learn to spin them better."

Example 4.9 (1970 Draft Lottery). *For the most part drawing balls out of urns is a mathematical idealization. However on December 1, 1969, the government put 366 balls with birthdays on them in a giant urn and the order in which the numbers were drawn determined the order in which men were called for service in Vietnam. However, looking at the average rank in the draw for different months shows that something was wrong.*

January	201.1	July	181.5
February	202.9	August	173.4
March	225.8	September	157.3
April	203.6	October	182.4
May	207.9	November	147.8
June	195.7	December	121.5

The source of the problem was that the birthdays were put in systematically: first the January birthdays, then February, and so on until the December birthdays were put in last. It should be clear from the outcome that the balls were not mixed well enough, so those from latter months were more likely to be near the top and drawn earlier.

To see what we would expect from a random drawing, we if X is that the rank of a given birthday, then X is uniform on $1, \dots, 366$ so $EX = 183.5$. To compute the variance we need the formula

$$\sum_{j=1}^k j^2 = \frac{k(k+1)(2k+1)}{6} \quad (4.9)$$

which can be checked by noting that it is correct when $k = 1$ and

$$\frac{k+1(k+2)(2k+3)}{6} - \frac{k(k+1)(2k+1)}{6} = (k+1)^2$$

Using (4.9) with $k = 366$

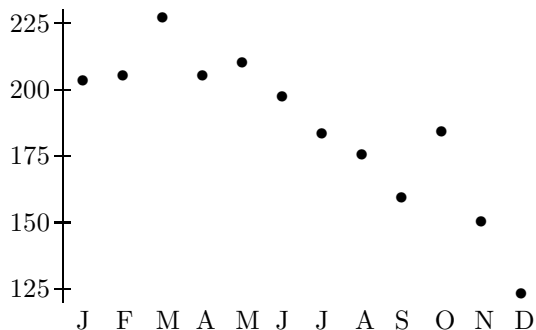
$$EX^2 = \frac{366(367)(733)}{6(366)} = 44,835$$

so $\text{var}(X) = 44,835 - (183.5)^2 = 11,162$ and the standard deviation $\sigma(X) = 105$.

Let $Y = (X_1 + \dots + X_{30})/30$ be the average rank of dates in a 30 day month. The sum rule for expected value implies $EY = 183.5$. The X_i are not independent but since a large value for X_i reduces the mean for X_j , $\text{cov}(X_i, X_j) \leq 0$ and it follows that $\text{var}(Y) \leq 11,162/(30) = 372$, and the standard deviation is 19.2. Using the normal approximation we see that 95% of the observations should lie in $[145.1, 221.9]$ and that the December observation which represents a deviation of 3.22 standard deviations is quite unusual.

The last number in our table is 3.09. Consulting another larger table the probability of a deviation of more than 3.22 standard deviations is approximately .00064. To put this in the proper perspective we should multiply this probability by 12 to account for the fact that any of the 12 months could have shown this pattern and by 2 since we would have been equally surprised by an observation that is too low. The result $0.00064 \cdot 24 = 0.01536$ is still small.

In our analysis we have focussed on the month of December. Another stronger indication that the lottery was not fair can be seen in the trend of monthly averages:



A formal analysis of the probability of observing such a pattern is beyond the scope of this book, so we will content ourselves to observe that it is consistent with the fact that the birthdays were put in one month at a time and the urn was not shaken up enough so December birthdays were closer to the top and drawn first.

Example 4.10 (Home field advantage). *In 2004 the Yankees record was 101–61 for a winning percentage of 62.3%. However, their record in home games was 57–24 or a winning percentage of 70.3%. Is this difference due to chance or are they more successful in home games?*

If each home game was independent and a win with probability 0.623 then they would expect to win $81(0.623) = 50.463$ games and the standard deviation would be $\sqrt{81(0.623)(0.377)} = 4.36$. Winning ≥ 57 games translates into a cutoff of 56.5 which is

$$\frac{56.5 - 50.463}{4.36} = 1.384$$

standard deviations. The probability of being ≥ 1.38 standard deviations is 0.0838, so this observation by itself is not very unusual.

A clear picture of home field advantage appears if we look at the records of all 30 teams. Only three teams (Anaheim, Philadelphia, and San Diego) won more away games than home games. If the two winning probabilities were equal then the home and away winning percentages would each have probability 1/2 of being larger. To be able to analyze the data, we assume that the outcome of the home vs. away comparison is independent for different teams. This is not true since when one team wins a home game the other loses an away game. However it should provide a reasonable approximation.

Having assumed we have 30 independent coin flips, we can analyze this data two ways. Using the binomial distribution, the probability of only getting 3 or fewer heads in 30 coin flips is 4.21×10^{-6} . On the other hand if we use the normal approximation, the mean is 15 and the standard deviation is $\sqrt{30(1/2)(1/2)} = 2.7386$. Writing the event of interest as ≤ 3.5 we see that the deviation is

$$\frac{3.5 - 15}{2.7386} = -3.46$$

standard deviations which has probability 2.7×10^{-4} .

Example 4.11 (Salk Vaccine). *To test the effectiveness of the Salk Vaccine, 400,000 children were divided into two groups of equal size. One group got the vaccine, the other a placebo. 57 children in the treatment group went on to contract polio, compared with 142 in the control group. Based on these results how reliable is the claim that the vaccine worked?*

If there was no difference between the two groups then the 199 cases would have been distributed at random between the two groups. Since the number of cases is much smaller than the size of the population, the number of cases in the treatment group is approximately Binomial(199, 1/2). The mean is $199/2 = 99.5$

and the standard deviation is $\sqrt{199(1/2)(1/2)} = 7.053$ so the observed deviation represents

$$\frac{57 - 99.5}{7.053} = -6.025 \text{ standard deviations}$$

As we noted in Example 4.6 the probability of a deviation this large is very small ($\approx 10^{-9}$).

4.4 Confidence Intervals

The common feature of examples in this section is that we are taking a random sample to estimate an unknown quantity and we would like to quantify the uncertainty in our estimate.

Example 4.12. *Suppose that we have been hired to take a poll to forecast the outcome of a presidential election. How accurate will our results be if we ask 900 people how they are going to vote?*

Suppose that we ask n voters how they are going to vote. Letting $X_i = 1$ if the i th person is for the Republican candidate and 0 otherwise, $X_1 + \cdots + X_n$ counts the number of people who say they are going to vote Republican, while the sample mean

$$\bar{X}_n = (X_1 + \cdots + X_n)/n$$

estimates the unknown fraction of Republican voters p . The standard deviation of \bar{X}_n is $1/n$ times that of $S_n = X_1 + \cdots + X_n$ so it is $\sqrt{p(1-p)/n}$. Dividing the numerator and denominator of (4.6) by n we have

$$P\left(\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \leq x\right) \rightarrow P(\chi \leq x)$$

Consulting the normal table we see that

$$P(-2 \leq \chi \leq 2) = P(\chi \leq 2) - P(\chi \leq -2) = 0.9772 - (1 - 0.9772) = 0.9544$$

In words our estimate \bar{X}_n will be within $2\sqrt{p(1-p)/n}$ of the true value 95% of the time. We do not know what p is, but $p(1-p) = p - p^2$ is a quadratic that is symmetric about $1/2$ and hence achieves its maximum there. When $p = 1/2$ we have $2\sqrt{p(1-p)/n} = 1/\sqrt{n}$. Thus if $n = 900$ our estimate will 95% of the time be accurate to within $1/30 = 0.033$. In opinion polls you sometimes see this reported as “there is a margin of error of ± 3.3 percent. If we want 1% accuracy, we would need to take a sample of size $100^2 = 10,000$.”

Example 4.13 (Weldon’s dice data). *An English biologist named Weldon was interested in the “pip effect” in dice – the idea that the spots, or “pips,” which on some dice are produced by cutting small holes in the surface, make the sides with more spots lighter and more likely to turn up. Weldon threw 12 dice 26,306 times for a total of 315,672 throws and observed that a 5 or 6 came up on 106,602 throws.*

His estimate of the probability that a 5 or 6 appears is thus $\hat{p} = 106,602/315,672 = 0.33770$. Now if the true probability is p then the variance of each observation is $\sigma^2 = p(1-p)$ so we estimate σ by $\sqrt{\hat{p}(1-\hat{p})}$. Using the fact that $P(-2 \leq \chi \leq 2) \approx 0.95$ we are 95% confident that the true value of p lies in the interval

$$(4.1) \quad \left[\hat{p} - \frac{2\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + \frac{2\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right]$$

For obvious reasons this is called a **95% confidence interval**. Plugging in our estimate \hat{p} we have

$$0.33770 \pm 2\sqrt{\frac{0.3377 \cdot 0.6623}{315,672}} = 0.33770 \pm 0.00168 = [0.33602, 0.33938]$$

so, for Weldon's dice at least, the true probability of a 5 or 6 is somewhat larger than $1/3$. This difference is not enough to be noticeable by people who play dice games for amusement but is perhaps large enough to be of concern for a casino that entertains tens of thousands of gamblers a year. For this reason most casinos use dice with no pips.

Example 4.14. *It is commonly presumed that an unborn child has a 50% chance of being female. But is this really the case? According to the Central Bureau of Statistics in the Netherlands, during the years 1989, 1990, and 1991, there were 585,609 children born of which 286,114 were girls. What is a 95% confidence interval for the percentage of female births?*

The sample mean is $\hat{p} = 286,114/585,609 = .4886$. The standard deviation of \hat{p} is

$$\frac{\sqrt{(.4886)(.5114)}}{\sqrt{585,609}} = 6.53 \times 10^{-4}$$

so the 95% confidence interval is $.4886 \pm .0013 = [.4873, .4899]$. A probability $p = 0.5$ represents $.0114/(6.53 \times 10^{-4}) = 17.45$ standard deviations, so we are very confident that $p < 1/2$.

Example 4.15 (The Literary Digest poll). *In order to forecast the outcome of the 1936 election, Literary Digest polled 2.4 million people and found that 57% of them were going to vote for Alf Landon and 43% were going to vote for F. D. Roosevelt.*

A 95% confidence interval for the true fraction of people voting for Landon based on this sample would be 0.57 ± 0.00064 but Roosevelt won, getting 62% of the vote to Landon's 38%. To explain how this happened we have to look at the methods *Literary Digest* used. They sent 10 million questionnaires to people whose names came from telephone books and club membership lists. Since many of the 9 million unemployed did not belong to clubs or have telephones the sample was not representative of the population as a whole. A second bias came from the fact that only 24% of the people filled out the form. This problem was mentioned in our discussion of exit polls in Chapter 2. If, for example, 36% of Landon voters and 16.6% of Roosevelt voters responded then the fraction of people who responded would be $0.62(0.166) + 0.38(0.36) = 0.24$ and the fraction in the sample for Landon would be

$$\frac{0.38(0.36)}{0.62(0.166) + 0.38(0.36)} = \frac{0.1368}{0.24} = 0.57$$

in agreement with the data.

Finally, we would like to observe that *Literary Digest*, which soon after went bankrupt, could have saved a lot of money by taking a smaller sample. George Gallup, who was just then getting started in the polling business, predicted based on a sample of size 50,000 that Roosevelt would get 56% of the vote. His 95% confidence interval for the election result would be 0.56 ± 0.0045 , compared with the election result of 62%. Again there could be some bias in his sample, or perhaps Landon voters, discouraged by the predicted outcome, were less likely to vote. The moral of our story is: It is much better to take a good sample than a large one.

The case of estimating a proportion is special since there is only one parameter that determines both the mean and the variance. However, if we were to estimate the average height of Cornell freshmen we do not know the mean or the variance. The solution is simple – we use the sample to estimate σ^2 . To motivate the estimate that we will use, we note that

$$\bar{X}_n = (X_1 + \cdots + X_n)/n$$

is the mean of the so-called **empirical distribution**, which assigns probability $1/n$ to each of the X_i . We use the mean of the empirical distribution \bar{X}_n to estimate the mean μ , so it is natural to estimate the variance σ^2 by the variance of the empirical distribution:

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) - \bar{X}_n^2 \quad (4.10)$$

Here the second equality follows from formula $E(Y - EY)^2 = EY^2 - (EY)^2$ applied to the random variable Y that is equal to X_i with probability $1/n$ and hence has $EY = \bar{X}_n$.

The first thing we want to show is

With probability one,

$$s_n^2 \rightarrow \sigma^2 \quad \text{as } n \rightarrow \infty \quad (4.11)$$

Proof. The strong law of large numbers tells us that with probability one $\bar{X}_n \rightarrow \mu$, so $\bar{X}_n^2 \rightarrow \mu^2$ with probability one. Applying the strong law to the i.i.d. variables $Y_i = X_i^2$ we have

$$\frac{1}{n} \sum_{i=1}^n X_i^2 = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}_n \rightarrow EY = EX^2$$

with probability one. Combining this with the conclusion for \bar{X}_n^2 we have

$$s_n^2 = \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) - \bar{X}_n^2 \rightarrow EX^2 - (EX)^2 = \text{var}(X)$$

with probability one. □

Being the variance of the empirical distribution, our estimator of σ^2 is a natural one, but from another point of view it is not. s_n^2 is a **biased estimator**, that is, $Es_n^2 \neq \sigma^2$. To compute Es_n^2 we note that

$$\begin{aligned} ns_n^2 &= \sum_{m=1}^n (X_m - \bar{X}_n)^2 = \sum_{m=1}^n (X_m - \mu - (\bar{X}_n - \mu))^2 \\ &= \sum_{m=1}^n \{(X_m - \mu)^2 - 2(X_m - \mu)(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2\} \\ &= \sum_{m=1}^n (X_m - \mu)^2 - 2(\bar{X}_n - \mu) \sum_{m=1}^n (X_m - \mu) + n(\bar{X}_n - \mu)^2 \\ &= \sum_{m=1}^n (X_m - \mu)^2 - n(\bar{X}_n - \mu)^2 \end{aligned}$$

since $\sum_{m=1}^n (X_m - \mu) = n\bar{X}_n - n\mu$. Taking expected value, now we have

$$\begin{aligned} nEs_n^2 &= \sum_{m=1}^n E(X_m - \mu)^2 - nE(\bar{X}_n - \mu)^2 \\ &= n\sigma^2 - n \operatorname{var}(\bar{X}_n) = (n-1)\sigma^2 \end{aligned}$$

since $E\bar{X}_n = \mu$ and $\operatorname{var}(\bar{X}_n) = \sigma^2/n$.

As a check on our calculation of Es_n^2 we note that when $n = 1$, $\bar{X}_1 = X_1$ so $s_1^2 = (X_1 - X_1)^2 = 0$, which agrees with our computation that $Es_1^2 = 0$. Since $Es_n^2 = \sigma^2(n-1)/n$, it follows that for $n > 1$,

an **unbiased estimator** of σ^2 is

$$\frac{n}{n-1}s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \equiv \sigma_n^2$$

Using (4.10) we can write the estimator in a form more convenient for calculations:

$$\sigma_n^2 = \frac{1}{n-1} \sum_{i=1}^n X_i^2 - n(\bar{X}_n)^2$$

Having settled on a way of estimating σ^2 there is very little more to say about 95% confidence intervals when σ^2 is unknown. We let $s_n = \sqrt{s_n^2}$ and replace σ by our estimate σ_n to get our 95% confidence interval

$$(4.3) \quad \left[\bar{X}_n - \frac{2\sigma_n}{\sqrt{n}}, \bar{X}_n + \frac{2\sigma_n}{\sqrt{n}} \right]$$

Example 4.16. In 1998 as an advertising campaign, the Nabisco Company announced a “1000 Chips Challenge” claiming that every 18-ounce bag of Chips Ahoy cookies contained at least 1000 chocolate chips. Dedicated Statistics students at the Air Force Academy purchased 16 bags of cookies and obtained the following counts for the number of chocolate chips.

1087	1121	1132	1135	1191	1200	1214	1219
1244	1258	1270	1295	1325	1345	1356	1419

For this data set $\bar{X}_n = 1238.2$ and $\sigma_n = 94.3$, so $2\sigma/\sqrt{16} = 47.15$ and the 95% confidence interval is [1191, 1285]. Having fewer than 1000 chips represents

$$\frac{1238.2 - 99.5}{94.3} = 2.53 \quad \text{standard deviations}$$

so about 5.73% of the the bags will have too few chocolate chips.

Example 4.17 (Speed of light). *In 1882 Michelson performed experiments to measure the speed of light. 23 trials gave an average of 299,756.2 km/sec with a stadard deviation of 107.12. Find a 95% confidence interval.*

$2\sigma/\sqrt{23} = 44.7$ so the interval is [299, 711.5, 299, 800.9].

In his 1897 experiment 100 trials resulted in a mean of 299,852.4 and a standard deviation of 79.0. In this case $2\sigma/\sqrt{100} = 15.8$ so the confidence interval was [299, 836.6, 299, 868.2]. Ntoice that this interval is disjoint from the first one. The correct answer is 299,710.5 so there must have be some bias in his experiments.

4.5 Exercises

Chebyshev's inequality

1. Suppose that it is known that the number of items produced at a factory per week is a random variable X with mean 50. (a) What can we say about the probability $X \geq 75$? (b) Suppose that the variance of X is 25. What can we say about $P(40 < X < 60)$?
2. Let $X = \text{binomial}(4, 1/2)$. Use Chebyshev's inequality to estimate $P(|X - 2| \geq 2)$ and compare with the exact probability.
3. Let \bar{X}_{10000} be the fraction of heads in 10,000 tosses. Use Chebyshev's inequality to bound $P(|\bar{X}_n - 1/2| \geq 0.01)$ and the normal approximation to estimate this probability.
4. Let X have a Poisson distribution with mean 16. Estimate $P(X \geq 28)$ using (a) Chebyshev's inequality, (b) the normal approximation.

Central limit theorem

5. Suppose that each of 300 patients has a probability of $1/3$ of being helped by a treatment. Find approximately the probability that more than 120 patients are helped by the treatment.
6. A person bets you that in 100 tosses of a fair coin the number of Heads will differ from 50 by 4 or more. What is the probability you will win this bet?
7. Suppose we toss a coin 100 times. Which is bigger, the probability of exactly 50 Heads or at least 60 Heads?
8. Suppose that 10% of a certain brand of jelly beans are red. Use the normal approximation to estimate the probability that in a bag of 400 jelly beans there are at least 45 red ones.
9. To estimate the percent of voters who oppose a certain ballot measure, a survey organization takes a random sample of 200 voters. If 45% of the voters oppose the measure, estimate the chance that (a) exactly 90 voters in the sample oppose the measure, (b) more than half the voters in the sample oppose the measure.
10. A basketball player makes 80% of his free throws on the average. Use the normal approximation to compute the probability that in 25 attempts he will make at least 23.
11. In a 162 game season find the approximate probability that a team with a .5 chance of winning will win at least 87 games.
12. Suppose we roll a die 600 times. What is the approximate probability that the number of 1's obtained lies between 90 and 110?
13. British Airways and United offer identical service on two flights from New York to London that leave at the same time. Suppose that they are competing

for the same pool of 400 customers who choose an airline at random. What is the probability United will have more customers than its 230 seats?

14. An insurance company has 10,000 automobile policyholders. The expected yearly claim per policyholder is \$240 with a standard deviation of \$800. Approximate the probability that the yearly claim exceeds \$2.7 million.

15. On each bet a gambler loses \$1 with probability 0.7, loses \$2 with probability 0.2, and wins \$10 with probability 0.1. Estimate the probability that the gambler will be losing after 100 bets.

16. Suppose we roll a die 10 times. What is the approximate probability that the sum of the numbers obtained lies between 30 and 40?

17. An airline knows that in the long run only 90% of passengers who book a seat show up for their flight. On a particular flight with 300 seats there are 324 reservations. (a) Assuming passengers make independent decisions what is the chance that the flight will be over booked? (b) Redo (a) assuming passengers travel in pairs and each pair flips a coin with probability 0.9 of heads to see if they will both show up or both stay home.

18. A student is taking a true/false test with 48 questions. (a) Suppose she has a probability $p = 3/4$ of getting each question right. What is the probability she will get at least 38 right? (b) Answer the last question if she knows the answers to half the questions and flips a coin to answer the other half. Notice that in each case the expected number of questions she gets right is 36.

19. The number of students who enroll in a psychology class is Poisson with mean 100. If the enrollment is > 120 then the class will be split into two sections. Estimate the probability that this will occur.

20. A gymnast has a difficult trick with a 10% chance of success. She tries the trick 25 times and wants to know the probability she will get exactly two successes. Compute the (a) exact answer, (b) Poisson approximation, (c) normal approximation.

21. Suppose that we roll two dice 180 times and we are interested in the probability that we get exactly 5 double sixes. Find (a) the normal approximation, (b) the exact answer, (c) the Poisson approximation.

22. A seed manufacturer sells seeds in packets of 50. Assume that each seed germinates with probability 0.99 independently of all the others. The manufacturer promises to replace, at no cost to the buyer, any packet with 3 or more seeds that do not germinate. (a) Use the Poisson to estimate the probability a packet must be replaced. (b) Use the normal to estimate the probability that the manufacturer has to replace more than 70 of the last 4000 packets sold.

23. A probability class has 30 students. As part of an assignment, each student tosses a coin 200 times and records the number of heads. What is the probability no student gets exactly 100 heads?

24. A die is rolled repeatedly until the sum of the numbers obtained is larger than 200. What is the probability that you can do this in 66 rolls or fewer?
25. Suppose that the checkout time at a grocery store has a mean of 5 minutes and a standard deviation of 2 minutes. Estimate the probability that a checker will serve at least 49 customers during her 4-hour shift.
26. A fair coin is tossed 2500 times. Find a number m so that the chance that the number of heads is between $1250-m$ and $1250+m$ is approximately $2/3$.
27. Members of the Beta Upsilon Tau fraternity each drink a random number of beers with mean 6 and standard deviation 3. If there are 81 fraternity members, how much should they buy so that using the normal approximation they are 93.32% sure they will not run out?
28. For a class project, you are supposed to take a poll to forecast the outcome of an election. How many people do you have to ask so that with probability .95 your estimate will not differ from the true outcome by more than 5%?
29. Suppose we take a poll of 2,500 people. What percentage should the leader have for us to be 99% confident that the leader will be the winner?
30. An electronics company produces devices that work properly 95% of the time. The new devices are shipped in boxes of 400. The company wants to guarantee that k or more devices per box work. What is the largest k so that at least 95% of the boxes meet the warranty?

Hypothesis Testing

31. A softball player brags that he is a .300 hitter, yet at the end of the season he has gotten 21 hits in 84 at bats. Is this just bad luck? To decide, compute the probability that he would get 21 hits or fewer if his probability of getting a hit were $p = 0.3$.
32. A tobacco company claims that the amount of nicotine in one of its cigarettes has mean 2.2 mg and a standard deviation of 0.3 mg. However the average nicotine content of 100 randomly chose cigarettes was 3.1 mg. Calculate the probability the average would be this high or higher if the company's claims were true.
33. In 1960, census results indicated that the age at which American men first married had a mean of 23.3 years. A sample of 40 men taken in 2000 married at an average age of 24.2 years and a standard deviation of 5.3 years. Does this support the belief that people are waiting longer to get married?
34. An article in *Fortune* magazine claimed that 50% of engineering graduates continue their studies to get an advanced degree. However a study of 484 graduates revealed only 198 who were planning graduate study. Is the observation consistent with the claim?
35. A researcher claims that at least 10% of all football helmets have manufacturing flaws that could potentially cause injury. A sample of 200 helmets revealed 13 that were defective. Does this finding support the researcher's claim.

36. A baseball player whose batting average is .150 gets glasses. In his next 50 at bats, he gets 15 hits. Did the glasses help?

Confidence Intervals

37. Of the first 10,000 votes cast in an election, 5,180 were for candidate A. Find a 95% confidence interval for the fraction of votes that candidate A will receive.

38. A bank examines the records of 150 patrons and finds that 63 have savings accounts. Find a 95% confidence interval for the fraction of people with savings accounts.

39. Among 625 randomly chosen Swedish citizens, it was found that 25 had previously been citizens of another country. Find a 95% confidence interval for the true proportion.

40. On 384 out of 600 randomly selected farms, the operator was also the owner. Find a 95% confidence interval for the true proportion of owner operated farms.

41. A sample of 2,809 hand-held video games revealed that 212 broke within the first three months of operation. Find a 95% confidence interval for the true proportion that break in the first three months.

42. A test of a medicine for reducing blood pressure revealed that 16 patients had a mean reduction of 11 points in their diastolic blood pressure (the smaller number in the reading) with a standard deviation of 2 points. Find a 95% confidence interval for the mean reduction.

43. Inspection of 25 trout revealed a sample mean of 2.10 and a standard deviation of 0.25 pounds. Find a 95% confidence interval for the mean.

44. A machine measures the bounce of 36 tennis balls and finds a mean of 1.7 ft and a standard deviation of 0.3 ft. What is a 95% confidence interval for the bounce of tennis balls?

45. The mean length of time 64 butterflies spent in the pupa stage was 49 hours with a standard deviation of 10 hours. Find a 95% confidence interval for the mean duration of the pupa stage.

46. During a two month period (44 weekdays) a parking garage collected an average of \$126 with a standard deviation of \$15. Find a 95% confidence interval for the mean revenue.

47. A nutrition laboratory tested 40 reduced sodium hotdogs and found the at the mean sodium content was 310 mg with a standard deviation of 36 mg. Find a 95% confidence interval for the mean sodium content.

48. A survey of 60 travel management professionals revealed an average salary of \$74,000 and a standard deviation of \$30,000. Find a 95% confidence interval for the mean salary.

49. A survey showed that tutoring increased the SAT math score of a group of 100 students by an average of 19 points with a standard deviation of 65. Find a 95% confidence interval for the mean improvement.
50. The tread life of 25 tires was an average of 31,485 miles with a standard deviation of 5,120 miles. Find a 95% confidence interval for the mean tread life.