

## Chapter 5

# From Probability to Inference

The probability distribution of a statistic obtained from an experiment is called its sampling distribution. An important class of statistics arises when the observations are counts of some variable. This leads to the binomial model for sample counts and sample proportions. These sampling distributions can be approximated by normal curves, and they directly demonstrate several important results about sample means  $\bar{x}$  in general:

1.  $\bar{x}$  is an unbiased estimate of the population mean  $\mu$ .
2. The standard deviation of  $\bar{x}$  is equal to  $\frac{\sigma}{\sqrt{n}}$  where  $n$  is the sample size and  $\sigma$  the population standard deviation.
3. The sampling distribution of  $\bar{x}$  is approximately  $N(\mu, \frac{\sigma}{\sqrt{n}})$ .

### 5.1 Sampling Distributions for Counts and Proportions

#### The Binomial Distribution

A binomial distribution is associated with an experiment comprising  $n$  independent trials each of which has the same success probability  $p$ . The random variable  $X$  counts the number of successes.

It is known that

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, 2, \dots, n$$

$$\text{mean} = \mu_X = np$$

and

$$\text{standard deviation} = \sigma_X = \sqrt{np(1-p)}$$

The corresponding Excel function is `BINOMDIST( $k, n, p, \text{cumulative}$ )`. If the parameter `cumulative` is set to “false,” Excel returns the probabilities  $P(X = k)$ , while if it is set to “true,” Excel returns the cumulative probabilities  $P(X \leq k)$ .

**Example 5.1.** Construct a binomial table for  $n = 15$  and  $p = 0.03$ , including both individual and cumulative probabilities.

### Solution

1. Enter the label  $k$  in cell A1 and the label  $P(X = k)$  in cell B1 of a new workbook. In A2:A17 enter the values  $\{0, 1, 2, \dots, 15\}$ .
2. Activate cell B2. Using either the **Formula Palette** or the **Function Wizard**, construct the binomial function by selecting **Statistical** for Function Category and **BINOMDIST** for Function Name.
3. Input the following into the dialog box

number\_s Enter the cell address A2  
 trials Enter the value 15  
 probability\_s Enter the value 0.3  
 cumulative Enter the value 0

to create the formula `BINOMDIST(A2,15,0.3,0)`. Click Finish or OK.

4. Activate cell B2, click the fill handle in the lower right corner, and drag to cell B17 to fill the column with individual binomial probabilities (Figure 5.1).

	A	B	C
1	k	$P(X=k)$	$P(X \leq k)$
2	0	0.00475	0.00475
3	1	0.03052	0.03527
4	2	0.09156	0.12683
5	3	0.17004	0.29687
6	4	0.21862	0.51549
7	5	0.20613	0.72162
8	6	0.14724	0.86886
9	7	0.08113	0.94999
10	8	0.03477	0.98476
11	9	0.01159	0.99635
12	10	0.00298	0.99933
13	11	0.00058	0.99991
14	12	0.00008	0.99999
15	13	0.00001	1.00000
16	14	0.00000	1.00000
17	15	0.00000	1.00000

Figure 5.1: Binomial Probabilities

5. Next label cell C1 as  $P(X \leq k)$  and repeat Steps 2, 3, and 4. Activate C2 instead of B2 in Steps 2 and 4 and enter the value 1 for the cumulative distribution in Step 3.

The resulting table of individual and cumulative binomial probabilities appears in Figure 5.1.

### Binomial Distribution Chart

We can quickly construct a histogram using the **ChartWizard** displaying the binomial probabilities just calculated. As the procedure is identical to earlier constructions of charts, we omit the details. This histogram appears in Figure 5.2.

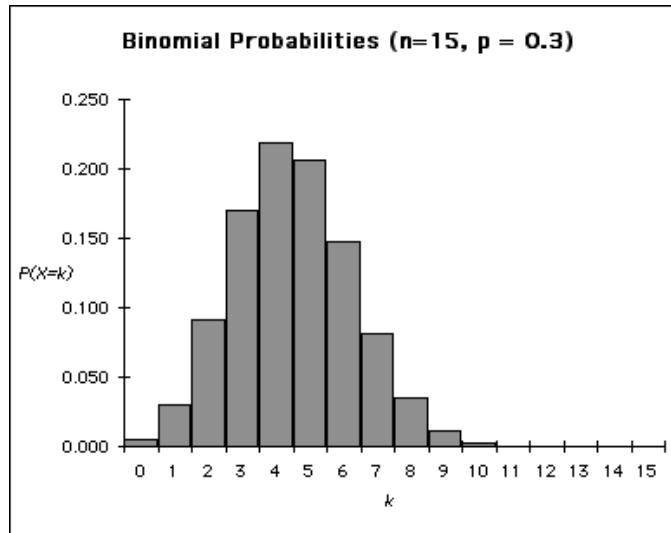


Figure 5.2: Binomial Histogram

### Inverse Cumulative Binomial

`CRITBINOM(trials, probability_s, alpha)` returns the smallest  $x$  for which the binomial cumulative distribution function (c.d.f.) is greater than or equal to alpha; that is, if  $B(x)$  represents the binomial c.d.f. on  $n$  trials and success probability  $p$ , then  $= \text{CRITBINOM}(n, p, \alpha)$  returns

$$B^{-1}(\alpha) = \inf\{x : B(x) \geq \alpha\}, \quad 0 < \alpha \leq 1$$

For  $\alpha = 0$ , this definition gives  $-\infty$  and Excel gives the error message **#NUM!**.

Inverse probabilities are useful for finding  $P$ -values and in **simulation** because from the definition, if  $U$  is uniform  $(0, 1)$  and  $F(x)$  is an arbitrary c.d.f. with inverse defined by

$$F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}$$

then

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

has the specified distribution  $F(x)$ . Thus, for instance

$$= \text{CRITBINOM}(u, p, U)$$

is a binomial random variable on  $n$  trials and success probability  $p$ , and

$$= \text{NORMINV}(U, \mu, \sigma)$$

is a  $N(\mu, \sigma)$  random variable.

## 5.2 Sampling Distribution of a Sample Mean

Simulation followed by a histogram of the results provides an insightful view of the *central limit theorem*.

### Simulating the Central Limit Theorem

**Example 5.2.** (Exercise 5.36, page 404 in the text.) A roulette wheel has 38 slots – 18 are black, 18 are red, and 2 are green. When the wheel is spun, a ball is equally likely to come to rest in any of the slots. Gamblers can place a number of different bets in roulette. One of the simplest wagers chooses red or black. A bet of one dollar on red will pay off an additional dollar if the ball lands in a red slot. Otherwise the player loses his dollar. When a gambler bets on red or black, the two green slots belong to the house. A gambler's winnings on a \$1 bet are either \$1 or -\$1.

- (a) Simulate a gambler's winnings after 50 bets and compare the gambler's mean winnings per bet with the theoretical results.
- (b) Compare the results with the normal approximation.

**Solution.** The number of wins after 50 bets  $X$  is a binomial  $B(50, 10/38)$  random variable with

$$\text{mean} \quad \mu_X = 50 \left( \frac{18}{38} \right) = 23.684$$

$$\text{standard deviation} \quad \sigma_X = \sqrt{50 \left( \frac{18}{38} \right) \left( \frac{20}{38} \right)} = 3.5306$$

The proportion of wins after 50 bets is  $\hat{p} = X/50$  with

$$\text{mean} \quad \mu_{\hat{p}} = \frac{18}{38} = 0.4737$$

$$\text{standard deviation} \quad \sigma_{\hat{p}} = \sqrt{\left( \frac{18}{38} \right) \left( \frac{20}{38} \right) / 50} = 0.0706$$

The gambler either wins \$1 or loses \$1. His average winnings per game, denoted by  $\bar{w}$ , are therefore  $1 \times$  the proportion of times that  $X = 1$  minus  $1 \times$  the proportion of times that  $X = -1$ , that is  $\bar{w} = \hat{p}(1) + (1 - \hat{p})(-1) = 2\hat{p} - 1$ . By the rules for means and standard deviations

$$\begin{array}{ll} \text{mean} & \mu_{\bar{w}} = 2\mu_{\hat{p}} - 1 = -0.0527 \\ \text{standard deviation} & \sigma_{\bar{w}} = 2\sigma_{\hat{p}} = 0.14123 \end{array}$$

By the **Central Limit Theorem**  $\hat{p}$  is approximately normal, and therefore

$$\bar{w} \text{ is approximately } N(-0.0527, 0.14123)$$

We can simulate a binomial random variable  $X$ , convert it first to  $\hat{p} = \frac{X}{n}$  and then to  $\bar{w} = 2\hat{p} - 1$ , after which we construct a histogram of the simulation results.

The following steps, referring to Figure 5.3, show how to develop a workbook to simulate 500 replications of 50 games. We will use the **RAND** function, which *links* the output to a histogram.

	A	B	C	D	E
1	<b>Demonstrating the Central Limit Theorem by Simulation</b>				
2					
3	true mean =	-0.0527	simulated mean =	-0.0534	
4	true st_dev =	0.14123	simulated st_dev =	0.14122	
5					
6	Formula entered in column B		=2*CRITBINOM(50,18/38,RAND())/50 - 1		
7	Simulation	Average per Game			
8	1	0.040	Bin Formulas	Bin	Freq.
9	2	-0.080	=-0.0527-3.5*0.14123	-0.55	0
10	3	-0.160	=-0.0527-3*0.14123	-0.48	1
11	4	0.000	=-0.0527-2.5*0.14123	-0.41	0
12	5	-0.040	=-0.0527-2*0.14123	-0.34	10
13	6	-0.080	=-0.0527-1.5*0.14123	-0.26	27
14	7	0.040	=-0.0527-0.14123	-0.19	44
15	8	-0.080	=-0.0527-0.5*0.14123	-0.12	51
16	9	0.000	=-0.0527	-0.05	107
17	10	0.120	=-0.0527+0.5*0.14123	0.02	118
18	11	0.080	=-0.0527+0.14123	0.09	74
19	12	-0.120	=-0.0527+1.5*0.14123	0.16	26
20	13	-0.160	=-0.0527+2*0.14123	0.23	24
21	14	-0.200	=-0.0527+2.5*0.14123	0.30	16
22	15	-0.160	=-0.0527+3*0.14123	0.37	2
23	16	0.040	=-0.0527+3.5*0.14123	0.44	0
24	17	0.000			500

Figure 5.3: Simulating the Central Limit Theorem

Bin intervals for the histogram will be located at multiples of the standard deviation from the mean

1. Prepare a new workbook by entering “Simulating the Central Limit Theorem” in cell A1 and centering the heading across A1:E1. Enter “Simulation” in A7 and “Average per Game” in B7.
2. Enter the values 1, 2, ..., 500 in cells A8:A507 as follows: Enter “1” in A8. Select A8 and choose **Edit – Fill – Series...** from the Menu Bar. In

the **Series** dialog box, check Series in **Columns** and Type **Linear**. Clear the **Trend** box and type “1” and “500” for the **Step** and **Stop** values, respectively.

3. Now simulate 50 games. In cell B8 enter `=2*CRITBINOM(50,18/38,RAND())/50-1` to generate the random variable  $\bar{w}$ . We have shown the formula on line 6 beginning in column C. Select B8, click the fill handle at the lower right corner of B8 and drag down to cell B507. Cells B8:B507 are now filled with 500 replications of the gambler’s average net gain per game after each simulated 50 games.
4. Next we prepare the simulations for output into a histogram. Enter the labels “Bin” in C8 and “Freq.” in D8. The bin endpoints for the histogram appear in cells in cells D9:D22 and the corresponding formulas behind the values are shown in cells C9:C23. These are based on theoretical true mean and standard deviation and the bin endpoints are expressed in simple multiples of the standard deviation from the mean, Enter `= -0.0527 - 3.5 * 0.14123` in D9, `= -0.0527 - 3.0 * 0.14123` in D10, and so on. Refer to Figure 5.3 where we have shown in cells C9:C23 the formulas to be entered in D9:D23. Note that you do not require a column C in your own worksheet.
5. Select E9:E23. Then type `= FREQUENCY(B8:B507,D9:D23)` in the entry area of the **Formula Bar**. Hold down the **Shift and Control** keys (either **Macintosh** or **Windows**) and press enter/return to **array-enter** the formula. The formula will appear **surrounded by braces { }** in the **Formula Bar**, and the bin frequencies will appear in cells E6:E19. the **Chart Wizard** as discussed previously. The resulting histogram appears in Figure 5.4.
6. Note that have also located the sample mean  $\bar{w}$  and sample standard deviation  $s_{\bar{w}}$  on the worksheet for comparison with the true values. These will change whenever the worksheet is re-evaluated. The formula behind cell D3 is `=AVERAGE(B8:B507)` and behind cell D4 it is `= SQRT((1-D3*D3)/50)`.

## Excel Output

The sample mean and sample standard deviation appear in D3:D4, and the population mean and standard deviation appear in B3:B4 for comparison purposes. For the simulation shown

$$\begin{array}{ll} \bar{w} & = -0.0534 & \mu_{\bar{w}} & = -0.0527 \\ s_{\bar{w}} & = 0.14122 & \sigma_{\bar{w}} & = 0.14123 \end{array}$$

The table of frequency counts appears in E9:E23 with the corresponding histogram in Figure 5.4. The histogram appears normal shaped with no unusual features.

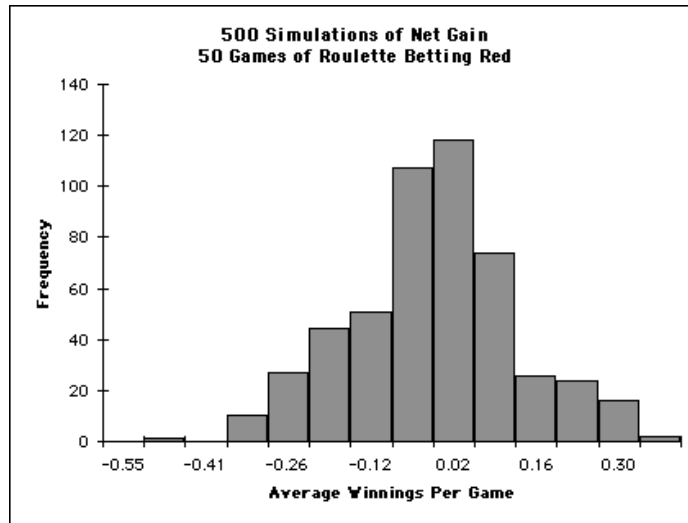


Figure 5.4: Histogram of 500 Simulations of 50 Roulette Games

Recalling that the bin entries are the right endpoints of the bin interval, we can determine the proportion of counts within 1, 2, and 3 standard deviation units of the mean. The simulation results are alarmingly good.

	Actual	Theoretical
Within 1 s	.700	.68
Within 2 s	.942	.95
Within 3 s	.998	.997