17D Sample proportions

So far, we have recorded the result of a binomial experiment as the number of successes. Often, however, it is more interesting to record instead the *proportion of the Bernoulli trials that were successes* — this is called the *sample proportion*.

Sample proportions have many purposes, but one important purpose is to estimate the probability p of each Bernoulli trial in the binomial experiment. For example, before every election, surveys estimate the proportion of Australians voting for the Working Together Party. If a survey of 2000 finds 700 WTP voters (the number of successes), then the the sample proportion is 0.35 (the proportion of successes), so the survey has estimated the WTP vote to be 35% of the population.

Population proportions

The probability p that a voter chosen at random will vote for the Working Together Party is a *population proportion*,

$$p = \frac{\text{number of Australians voting WTP}}{\text{number of Australian voters}}.$$

Many probabilities arise in this way, as we have often seen. Sample proportions apply to all binomial situations, but may be intuitively easier to understand if we think of the probability p arising as a population proportion, as in this voting example.

Sample proportions

Let $X \sim B(n, p)$ be a binomial variable consisting of n independent Bernoulli trials, each with probability p of success. The *sample proportion* $\hat{p}$ is the proportion of successes when the experiment is run. Thus *the sample proportion is a new random variable denoted by* $\hat{p}$ *and defined by*

$$\hat{p} = \frac{X}{n}.$$

For example, I survey 20 voters, and 7 say that they will vote WTP. This is a binomial experiment with 20 independent Bernoulli stages, and random variable X with 21 possible values $0,\ 1,\ 2,\ 3, \ldots, 20$. I can record my result as

$$X = 7 \qquad \text{or as} \qquad \hat{p} = \frac{7}{20}.$$

The binomial random variable X has 21 possible values $0,\ 1,\ 2,\ 3, \ldots, 20$, and the new random variable $\hat{p}$ has 21 possible values $0 = \frac{0}{20}, \frac{1}{20}, \frac{2}{20}, \ldots, 1 = \frac{20}{20}$.

The sample proportion distribution — mean and variance

A sample proportion random variable $\hat{p}$ is not a binomial random variable X because its values are fractions between 0 and 1 rather than whole numbers from 0 to n. But it is closely related to the corresponding binomial variable — the probability of each value of $\hat{p}$ is the same as the probability of the corresponding value of X,

$$P\left(\hat{p} = \frac{x}{n}\right) = P(X = x) = {}^{n}\mathrm{C}_{x}\, p^{x} q^{n-x},$$

so the various graphs of a sample proportion distribution can be obtained from the graphs of the corresponding binomial distributions by relabelling the axes and changing the numbers on the horizontal axis.

ISBN 978-1-108-76630-2

The mean and variance of $\hat{p}$ come directly from the mean and variance of X,

$$\begin{aligned} \mathrm{E}(\hat{p}) &= \mathrm{E}\left(\frac{X}{n}\right) \\ &= \frac{E(X)}{n} \\ &= \frac{np}{n} \\ &= p, \end{aligned} \qquad \begin{aligned} \mathrm{Var}(\hat{p}) &= \mathrm{Var}\left(\frac{X}{n}\right) \\ &= \frac{\mathrm{Var}(X)}{n^2} \\ &= \frac{npq}{n^2} \\ &= \frac{pq}{n}, \end{aligned} \qquad \begin{aligned} \mathrm{SD} &= \sqrt{\frac{npq}{n^2}} \\ &= \frac{\sqrt{npq}}{n}. \end{aligned}$$

The first formula $\mathrm{E}(\hat{p}) = p$ proves what we said in the introduction to this section — running a binomial experiment and recording the sample proportion $\hat{p}$ gives an estimate of the Bernoulli probability p.

7 THE SAMPLE PROPORTION $\hat{p}$

Suppose that a binomial experiment with random variable X consists of n independent Bernoulli trials, each with probability p.

- The *sample proportion* is the random variable $\hat{p} = \dfrac{X}{n}$.
- The values of $\hat{p}$ are the $n + 1$ values $0, 1, 2, \ldots, n$ of X divided by n. Thus the values of $\hat{p}$ are $0 = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, 1 = \frac{n}{n}$.
- The probability of each value of $\hat{p}$ equals the probability of the corresponding value of X,
$$P\left(\hat{p} = \frac{x}{n}\right) = P(X = x) = {}^{n}\mathrm{C}_x p^x q^{n-x}.$$
- The mean, variance and standard deviation of $\hat{p}$ are
$$\mathrm{E}(\hat{p}) = p, \qquad \mathrm{Var}(\hat{p}) = \frac{pq}{n}, \qquad \text{standard deviation} = \frac{\sqrt{npq}}{n}.$$
- The sample proportion $\hat{p}$ gives an estimate of the probability p.

Here is a simple example comparing a binomial distribution with the corresponding sample proportion distribution.

Example 15 17D

A coin is tossed six times. Let X be the binomial variable and $\hat{p}$ the corresponding sample proportion random variable.

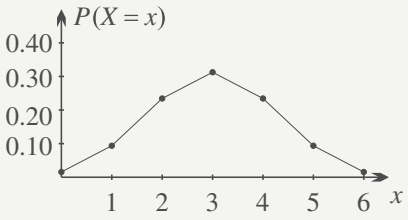
a Tabulate the probabilities of X, find the mean and variance, and draw the frequency polygon.

b Tabulate the probabilities of $\hat{p}$, find the mean and variance, and draw the frequency polygon.

SOLUTION

a Using the formula $P(x \text{ heads}) = {}^{6}\mathrm{C}_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{6-x}$,

x	0	1	2	3	4	5	6
prob	$\frac{1}{64}$	$\frac{6}{64}$	$\frac{15}{64}$	$\frac{20}{64}$	$\frac{15}{64}$	$\frac{6}{64}$	$\frac{1}{64}$



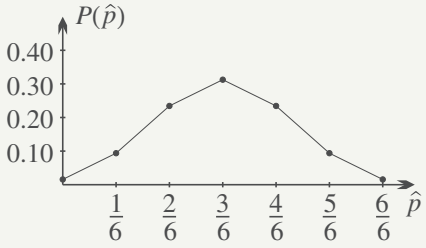
For the binomial variable X,

$\mathrm{E}(X) = np = 3$ and $\mathrm{Var}(X) = npq = 1\frac{1}{2}$.

ISBN 978-1-108-76630-2

b The sample proportion is $\hat{p} = \frac{X}{n}$, and the corresponding probabilities are the same, so

$\hat{p}$	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1
prob	$\frac{1}{64}$	$\frac{6}{64}$	$\frac{15}{64}$	$\frac{20}{64}$	$\frac{15}{64}$	$\frac{6}{64}$	$\frac{1}{64}$



For the sample proportion $\hat{p}$,

$$\mathrm{E}(\hat{p}) = p = \tfrac{1}{2} \text{ and } \mathrm{Var}(\hat{p}) = \frac{pq}{n} = \tfrac{1}{24}.$$

Why all this machinery of sample proportions?

Marketers and public opinion pollsters are very keen to find sample proportions for the Australian population. But surveys are expensive, and pollsters need to know how effective they can be with the small samples that they can afford or have time to carry out.

These things are also crucial for health research and public policy. Estimating a quantity such as the monthly unemployment statistics, or the number of Australians who require disability assistance services, is dependent on sample survey methods.

Working with the sample proportion not only provides an immediate estimate of p, but it also provides a straightforward approach to assessing how accurate the estimate is likely to be.

For example, the next worked example calculates the probability that a pollster will get within 5% and 10% of the correct result when using a sample of 20 voters to estimate the proportion voting for the WTP.

Example 16 17D

A binomial experiment with 20 Bernoulli trials is being used to estimate the probability p of success in the Bernoulli experiment. If the actual value of p is $\frac{1}{3}$, find the probability that the experiment will yield an estimate within:

a 0.05 of the actual value, **b** 0.1 of the actual value.

SOLUTION

Notice first that the estimate cannot be exactly $p = \frac{1}{3} \fallingdotseq 0.333$ because the size of the sample is 20, which is not a multiple of 3.

a To obtain an estimate within 0.05 of the actual value $p \fallingdotseq 0.333$, the sample proportion would need to be $\hat{p} = \frac{6}{20} = 0.30$ or $\hat{p} = \frac{7}{20} = 0.35$.

$$P\left(\hat{p} = \tfrac{6}{20} \text{ or } \hat{p} = \tfrac{7}{20}\right) = {}^{20}\mathrm{C}_6 \left(\tfrac{1}{3}\right)^6 \left(\tfrac{2}{3}\right)^{14} + {}^{20}\mathrm{C}_7 \left(\tfrac{1}{3}\right)^7 \left(\tfrac{2}{3}\right)^{13}$$
$$= 0.36425 \ldots \text{ (save this value in the memory)}.$$

Hence the probability that the estimate is within 0.05 of p is about 0.364.

ISBN 978-1-108-76630-2

b To obtain an estimate within 0.1 of the actual value $p \doteqdot 0.333$, the sample proportion would need to be $\hat{p} = \frac{5}{20} = 0.25$ or $\frac{6}{20}$ or $\frac{7}{20}$ or $\frac{8}{20} = 0.40$.

$$P\left(\hat{p} = \tfrac{5}{20} \text{ or } \hat{p} = \tfrac{8}{20}\right) = {}^{20}\text{C}_5 \left(\tfrac{1}{3}\right)^5 \left(\tfrac{2}{3}\right)^{15} + {}^{20}\text{C}_8 \left(\tfrac{1}{3}\right)^8 \left(\tfrac{2}{3}\right)^{12}$$
$$= 0.29368\ldots,$$

Adding the answer to part **a**, the probability that the estimate is within 0.1 of the actual value is about 0.658.

8 USING THE SAMPLE PROPORTION

The sample proportion $\hat{p}$ can often be used to estimate the probability of obtaining an acceptable estimate to the probability p of a Bernoulli experiment.

Using the normal approximation to the sample proportion

We discussed in Section 17C how to use normal approximations to a binomial distribution to obtain reasonable approximations to the binomial distribution far more quickly. The probabilities of the sample proportions are just the probabilities of the corresponding values of the binomial variable, so the same approximation methods can be used here. We cannot prove in this course that the normal distribution approximates $\hat{p}$ — it is another example of the central limit theorem discussed at the start of Section 16G in the last chapter.

The normal approximation to the sample proportion thus has the same mean p as the sample proportion distribution, and the same variance $\frac{pq}{n}$, so we can approximate it using the normal distribution $N\left(p, \frac{pq}{n}\right)$.

As an example, let us increase the numbers in the previous worked example back up to realistic levels, and use the normal approximation to obtain the result. Political surveys before elections attempt to reduce the margin of error in their estimates to 1%–2%, that is, they attempt to estimate p correct to within 0.01–0.02.

Example 17 17D

A survey of 2000 Australian voters is being used to estimate the probability p that a random voter will vote for the Working Together Party. Assuming that the actual value of p is $\frac{1}{3}$, use the normal approximation to the sample proportion distribution to find the probability that the experiment will yield an estimate:

a within 0.01 of the actual value, **b** within 0.02 of the actual value.

SOLUTION

The numbers in the sample are large enough for us to ignore continuity corrections.
The normal approximation to $\hat{p}$ has

$$\text{mean} = \tfrac{1}{3} \doteqdot 0.3333 \quad \text{and} \quad \text{variance} = \frac{pq}{n} = \frac{2}{9} \times \frac{1}{2000} = \frac{1}{9000},$$

so taking the square root, the standard deviation is about 0.01054.

ISBN 978-1-108-76630-2

a We need to find $P(0.3233 \leq \hat{p} \leq 0.3433)$, so we find z-scores.

For 0.3233, $z = \dfrac{-0.01}{0.01054}$ $\doteqdot -0.9487$.

For 0.3433, $z = \dfrac{0.01}{0.01054}$ $\doteqdot 0.9487$.

Hence the probability that the estimate is within 0.01 of the actual value is about

$$P(-0.9487 \leq Z \leq 0.9487) \doteqdot 0.66$$

b We need to find $P(0.3133 \leq \hat{p} \leq 0.3533)$. The z-scores are now;

For 0.3233, $z = \dfrac{-0.02}{0.01054}$ $\doteqdot -1.8974$.

For 0.3433, $z = \dfrac{0.02}{1.01054}$ $\doteqdot 1.8974$.

Hence the probability that the estimate is within 0.02 of the actual value is about

$$P(-1.8974 \leq Z \leq 1.8974) \doteqdot 0.94$$

Example 18 17D

Election surveys usually claim that their margin of error is about 2%, and the last worked example has addressed this. What are some other issues that pollsters need to take into account?

SOLUTION

- Can I be sure that the sample is unbiased?
- Are people answering honestly?
- Are opinions changing as the election approaches?
- What is to be done with people who refuse to answer or 'don't know'?
- Were there language or cultural problems that interfered with the interview?

The data, the sample proportion, and the normal approximation

The situation has now become quite complicated because we have:

1. values of the sample proportion distribution $\hat{p}$ obtained from surveys,
2. the sample proportion distribution, if we assume a value for the mean p,
3. the normal approximation to the sample proportion distribution, also if we assume a value for the mean p.

We conclude this chapter by drawing one more picture, using the dataset that we gained by simulation of a binomial experiment in worked Example 9 of Section 16B. We are interested in the distribution of the sample proportion $\hat{p}$, that is, the distribution of the estimate of the probability, and we will use the cumulative distributions in our comparisons.

Notice that the density function of the normal approximation to $\hat{p}$ is obtained from the density function of the approximation to X by stretching horizontally by a factor of $\dfrac{1}{n}$, and then stretching vertically by a factor of n so that the area under the curve remains 1. Therefore we can't superpose the graph of that normal density function on the graph of the frequency polygon because the heights don't match. Instead, we will compare the cumulative graphs to see the agreement.

ISBN 978-1-108-76630-2

Example 19 17D

In worked Example 9 of Section 17B, we ran 100 simulations of the binomial experiment of tossing 10 coins and recording the number of heads. Now imagine that the purpose of each simulation was to produce an estimate of the probability of tossing heads, that is, a value of the random variable $\hat{p}$. Note that this random variable $\hat{p}$ has the 11 values $0 = \frac{0}{10}, \frac{1}{10}, \frac{2}{10}, \ldots, 1 = \frac{10}{10}$.

a From the data, produce a table showing the cumulative relative frequency of obtaining each value of $\hat{p}$, and find the mean and standard deviation.

b Use binomial probability to calculate the cumulative probabilities of obtaining the values of $\hat{p}$, and find the mean and standard deviation.

c Use the normal approximation to the sample proportion distribution to calculate the cumulative probabilities of values of $\hat{p}$, and find the mean and standard deviation. Apply the continuous correction, because $n = 10$ is small.

d On one graph, draw the cumulative relative frequency polygon of the data, the cumulative probability polygon of the distribution, and the CDF of the normal approximation.

SOLUTION

The calculations in parts **a** and **b** were done in worked Example 9, and we only need to divide the values through by 10.

a After running the experiment 100 times, each relative frequency is the estimated probability of obtaining each estimate of $\hat{p}$.

$\hat{p}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
f_r	0.00	0.02	0.06	0.15	0.2	0.23	0.17	0.13	0.01	0.03	0.00
cf_r	0.00	0.02	0.08	0.23	0.43	0.66	0.83	0.96	0.97	1.00	1.00

After calculation, $\hat{p}$ has mean 0.482 and standard deviation about 0.170.

b From previous calculations, $\mu = 0.5$ and $\sigma = \frac{1}{10}\sqrt{2.5} \doteq 0.158$.

$\hat{p}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(\hat{p})$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001
Cmlve	0.001	0.011	0.055	0.172	0.377	0.623	0.828	0.945	0.989	0.999	1.000

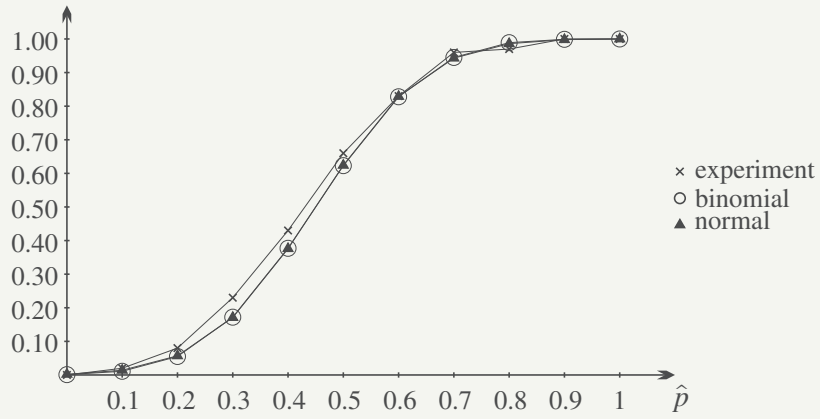
c The mean and standard deviation of the normal approximation are the same as for the sample proportion distribution. We need z-scores first, and we apply the continuity correction with 0.1 as the gaps in the values for $\hat{p}$. For example, for $\hat{p} = 0.7$ we find the z-score corresponding to $\hat{p} = 0.75$.

$\hat{p}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
z	−2.85	−2.21	−1.581	−0.949	−0.316	0.316	0.949	1.581	2.21	2.85	3.48
$\Phi(z)$	0.002	0.014	0.057	0.171	0.376	0.624	0.829	0.943	0.986	0.998	1.000

ISBN 978-1-108-76630-2

Cambridge University Press
Updated July 2022

d



At the resolution of this graph, the theoretical probability distribution curve and the normal approximation to it are not separated. Look at the tables.

Exercise 17D

FOUNDATION

Note: In the previous exercise we used continuity corrections when we approximated a discrete binomial distribution by a continuous normal distribution. For large values of n this correction is less necessary. In this section, however, using the continuity correction when using sample proportions leads to much messier calculations, and the numbers are mostly large, so it will seldom be applied.

1 **a** A student tosses five coins and the results are: H T H T T. What is the sample proportion $\hat{p}$ of heads in this experiment?

b A student selects ten cards from a standard pack with replacement, and records the suit: ♡♣♠◇◇♠♡♠♠◇. What is the sample proportion $\hat{p}$ of spades in this sample?

c A manufacturer takes a sample of 12 items from their recent batch of gizmos, testing each item to see if it passes quality control (P) or not (F). The results were: P P P F F P P P P F P P. What is the proportion $\hat{p}$ of items that pass?

2 A single fair coin is tossed five times, and the number x of heads is recorded.

a Copy and complete the upper table to the right, using the binomial probability formula.

x	0	1	2	3	4	5
$P(X = x)$						

b Copy and complete the lower table to the right to convert part **a** to a table of probabilities of the sample proportions. You will need to divide each score x by 5 to obtain the corresponding sample proportion.

$\hat{p}$	
$P(\hat{p})$	

c Calculate the mean of the second table.

d How would you interpret this mean?

3 Every Saturday for 20 weeks, a marketer surveyed five people chosen at random in a suburban shopping centre. The first question is, 'Do you live in the suburb?', and the weekly frequency of the number x of 'Yes' answers is given in the upper table.

x	0	1	2	3	4	5
f	1	1	3	2	6	7

ISBN 978-1-108-76630-2