Mathematical Methods Units 3 & 4

Chapter 15 — Statistical inference and sample proportions

Aboriginal and Torres Strait Islander content has been excluded as accuracy cannot be guaranteed, and it is better to be respectful than cover the entire curriculum inaccurately.

Significant effort has been made to ensure this content is legal. Please send any areas of concern to admin@vwx.au with appropriate document/legal references so errors can be verified and changes be made.

Contents

Chapter 15 Statistical inference and sample proportions — 15A Populations and samples

Theory

Statistical inference uses information from a **sample** to draw conclusions about a whole **population**.

Population and parameter. The **population** is the entire group of interest. A **parameter** is a fixed number that describes the population — for a categorical characteristic, the key parameter is the **population proportion**

$$p = \frac{\text{number in the population with the characteristic}}{\text{population size}}.$$

The value of p is usually **unknown** and fixed.

Sample and statistic. A **sample** is a subset of the population. A **statistic** is a number computed from the sample. For a characteristic, the **sample proportion** is

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

where n is the sample size and X is the number in the sample with the characteristic. We use the statistic $\hat{p}$ to **estimate** the parameter p.

Sampling variation. Different random samples give different values of $\hat{p}$, so $\hat{p}$ is itself a **random variable** — its value varies from sample to sample. This spread of values is called **sampling variation**.

Random samples. In a **random sample** every member of the population is equally likely to be chosen. When the population is large (or sampling is with replacement) the selections behave like independent trials, so $X \sim \mathrm{Bi}(n, p)$ and the possible values of $\hat{p}$ are

$$0, 1/n, 2/n, \ldots, 1.$$

The distribution of $\hat{p}$ is centred on p, and as the sample size n increases the values of $\hat{p}$ cluster **more tightly** around p (less sampling variation).

Simulating repeated random sampling. When the population proportion p is known (or assumed for a model), we can **simulate** the sampling process using technology — a calculator or spreadsheet random-number generator, or a command such as `randBin(n, p)` — instead of physically drawing samples. One simulated sample of size n is produced by generating n random "success/failure" outcomes with success probability p, counting the successes X, and recording $\hat{p} = X/n$. Repeating this many times gives a list of simulated values of $\hat{p}$, which we display in a dot plot or histogram. In pseudocode the procedure is a repetition:

```
for trial from 1 to M do
    X ← number of successes in n trials, each with probability p
    record  X / n
end for
```

The simulated distribution illustrates the key facts about $\hat{p}$: it is **centred on** p, it is **spread** by sampling variation (with standard deviation about $\sqrt{p(1-p)/n}$), and the spread **narrows as** n **increases**. Increasing the *number of simulations* M does not change the spread — it simply fills in the shape of the distribution more clearly.

Worked examples

Example 1 — scaffolded

In a school of 500 students, 150 are left-handed. A random sample of 20 students contains 5 left-handed students. State the population proportion p and the sample proportion $\hat{p}$.

Working	Comment
Population proportion: $p=\dfrac{150}{500}=\dfrac{3}{10}=0.3$.	p is a **parameter** — it describes the whole population.
Sample proportion: $\hat{p}=\dfrac{X}{n}=\dfrac{5}{20}=\dfrac{1}{4}=0.25$.	$\hat{p}$ is a **statistic** — computed from the sample of $n=20$.
$\hat{p}=0.25$ estimates $p=0.3$; they differ because of sampling variation.	A different sample would generally give a different $\hat{p}$.

Example 2 — scaffolded

A market researcher asks a random sample of 120 shoppers whether they prefer a new product; 42 say yes. Identify the parameter and the statistic, and find the sample proportion.

Working	Comment
Parameter: p, the proportion of **all** shoppers who prefer the product (unknown).	The parameter is the population value we want to estimate.
Statistic: the sample proportion $\hat{p}$.	Computed from the $n=120$ shoppers surveyed.
$\hat{p}=\dfrac{42}{120}=\dfrac{7}{20}=0.35$.	So 0.35 is our estimate of p.

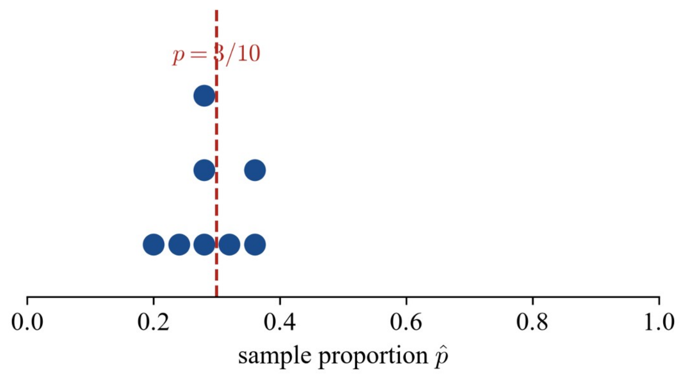
Example 3 — unscaffolded

Three separate random samples of 25 people each are taken; the numbers preferring brand A are 7, 9 and 6. Find the sample proportion for each and comment.

$$\hat{p}_1=\frac{7}{25}=0.28,\ \hat{p}_2=\frac{9}{25}=0.36,\ \hat{p}_3=\frac{6}{25}=0.24.$$

The three values differ even though each sample is drawn from the **same** population — this is sampling variation. Each is an estimate of the (fixed, unknown) population proportion p.

$\therefore$ the sample proportion varies from sample to sample, clustering around p (here near 0.3):



Example 4 — unscaffolded

A random sample of size 5 is taken from a population in which the proportion of successes is $p=\dfrac{2}{5}$. List the possible values of $\hat{p}$, and find $Pr\left(\hat{p}=\dfrac{2}{5}\right)$.

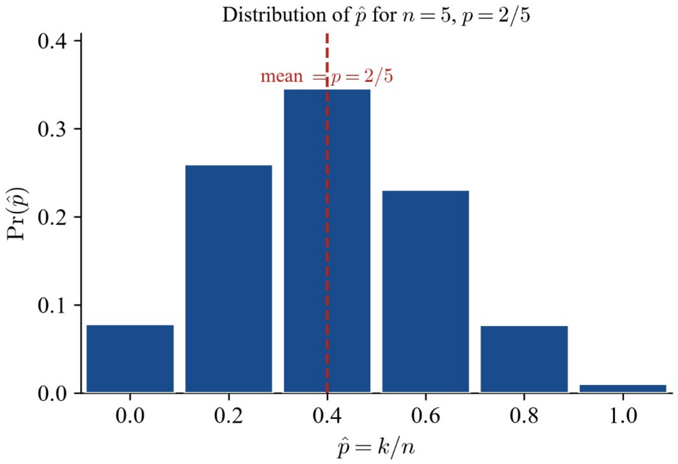
Possible values of $\hat{p}=\dfrac{X}{5}$ for $X=0,1,\ldots,5$ are

$$0,1/5,2/5,3/5,4/5,1.$$

Here $X \sim \mathrm{Bi}\left(5, \frac{2}{5}\right)$, so

$$Pr(\hat{p}=2/5) = Pr(X=2) = \binom{5}{2}(2/5)^2(3/5)^3 = \frac{216}{625} \approx 0.3456.$$

$\therefore$ the distribution of $\hat{p}$ is the binomial distribution of X, rescaled onto $\{0, 1/5, \ldots, 1\}$:



Example 5 — unscaffolded

A simulation draws 20 random samples, each of size $n=25$, from a population with $p=0.4$, and records the sample proportion $\hat{p} = X/25$ for each. The 20 results were

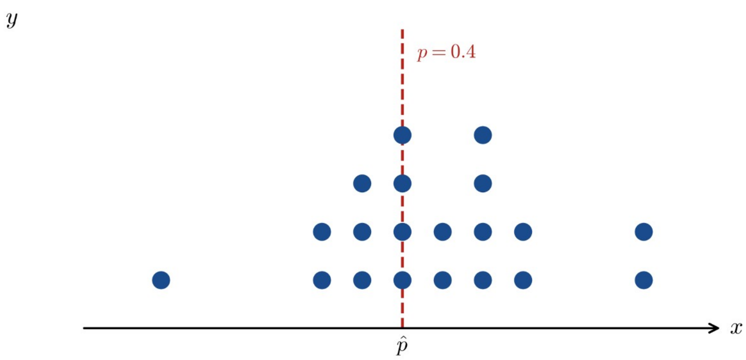
$$0.44, 0.52, 0.48, 0.32, 0.36, 0.52, 0.16, 0.48, 0.48, 0.40, 0.36, 0.36, 0.32, 0.40, 0.40, 0.40, 0.64, 0.48, 0.44, 0.64.$$

Find the mean of the simulated proportions, and comment on the centre and spread in relation to p.

The sum of the 20 values is 8.6, so the mean is $\frac{8.6}{20} = 0.43$.

The simulated proportions range from 0.16 to 0.64, and their mean 0.43 is close to the population proportion $p=0.4$: the distribution is **centred near** p. The spread agrees with the theoretical standard deviation $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4 \times 0.6}{25}} \approx 0.098$, so almost all values lie within about $2 \times 0.098 \approx 0.2$ of p.

$\therefore$ the simulation illustrates that $\hat{p}$ varies from sample to sample but clusters about p; with more simulated samples the dot plot would fill out the full distribution of $\hat{p}$:



Dot plot of 20 simulated sample proportions ($n=25$, $p=0.4$)

Practice questions

Scaffolded

Q1. In a random sample of 40 light globes, 18 are the energy-saving type. (a) State the sample size n and the count X ; (b) find the sample proportion $\hat{p}$; (c) is $\hat{p}$ a parameter or a statistic?

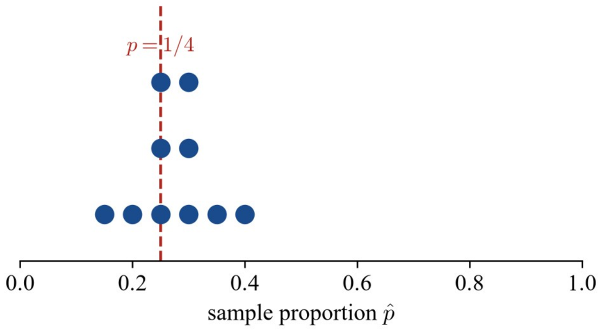
Q2. A factory produces bolts; 30 of a random sample of 200 are undersized. (a) find $\hat{p}$; (b) explain what $\hat{p}$ estimates; (c) another sample of 200 gives $\hat{p}=0.18$ — explain why the two values differ.

Q3. A random sample of size 8 is taken from a population. (a) list all possible values of $\hat{p}$; (b) if the sample contains 2 successes, find $\hat{p}$; (c) how many successes give $\hat{p}=0.75$?

Q4. In a large population the proportion of a certain type is $p=\frac{1}{4}$. A random sample of 8 is taken, so $X \sim \text{Bi}\left(8, \frac{1}{4}\right)$.

(a) find $Pr\left(\hat{p}=\frac{1}{4}\right)$; (b) find $Pr\left(\hat{p}=0\right)$; (c) state the most likely value of $\hat{p}$.

Q5. The dot plot shows the sample proportion from ten random samples of size 20. (a) state the smallest and largest values of $\hat{p}$ shown; (b) which value of $\hat{p}$ occurred most often; (c) the population proportion is $p=0.25$ — comment on how the sample proportions are spread about p.



Q6. State whether each quantity is a **parameter** or a **statistic**: (a) the proportion of all Australian voters who support a policy; (b) the proportion of a survey of 1000 voters who support the policy; (c) the mean height of every student in a school; (d) the proportion left-handed in a sample of 50 people.

Unscaffolded

Q7. A random sample of 60 households contains 24 with a pet. Find the sample proportion $\hat{p}$.

Q8. In a random sample of 52 cards drawn (with replacement) from a shuffled deck, 13 are hearts. Find $\hat{p}$ and state the population proportion p of hearts.

Q9. Of a random sample of 300 people, 45 are colour-blind. Find $\hat{p}$.

Q10. A random sample of 240 eggs contains 84 that are large. Find $\hat{p}$.

Q11. A random sample of size 10 is taken from a large population. List all the possible values of $\hat{p}$.

Q12. A random sample of size 4 is taken from a population with $p=\frac{1}{2}$, so $X \sim \text{Bi}\left(4, \frac{1}{2}\right)$. Find the probability distribution of $\hat{p}$ (a table of $\hat{p}$ against $Pr\left(\hat{p}\right)$).

Q13. For a random sample of size 5 from a population with $p=\frac{3}{10}$, find $Pr\left(\hat{p}=\frac{1}{5}\right)$, giving your answer as an exact fraction and correct to four decimal places.

Q14. Explain why the sample proportion $\hat{p}$ is a random variable but the population proportion p is not.

Q15. Two random samples are taken from the same population: one of size 50 and one of size 500. Explain which sample's proportion $\hat{p}$ is likely to be closer to the population proportion p, and why.

Q16. In a random sample of 45 trees in a forest, 27 are eucalypts. Find $\hat{p}$, and state whether it is a parameter or a statistic.

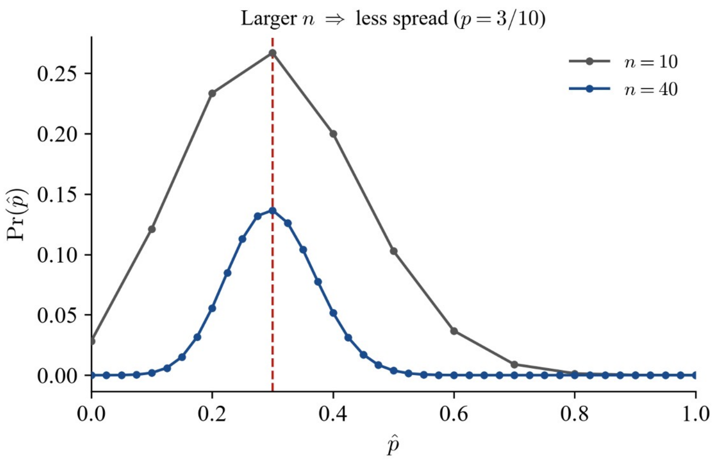
Q17. A quality inspector examines a random sample of 50 items and finds 8 imperfect. Find the sample proportion of imperfect items.

Q18. A population contains a proportion $p=\dfrac{2}{5}$ of a certain characteristic. For a random sample of size 5, use $X\sim\text{Bi}\left(5,\dfrac{2}{5}\right)$ to find $Pr\left(\hat{p}\geq\dfrac{3}{5}\right)$.

Q19. In an election, a random exit poll of 1000 voters finds 646 voted for the winning party. State the sample proportion $\hat{p}$, and explain why it is only an estimate of the true proportion p of all voters.

Q20. A random sample of size n from a large population gives $\hat{p}=0.32$ and it is known that $X=16$. Find the sample size n.

Q21. The two graphs show the distribution of $\hat{p}$ for samples of size 10 and size 40, both from a population with $p=0.3$. Describe how increasing the sample size changes the distribution of $\hat{p}$.



Q22. A student uses a calculator to simulate 50 random samples of size 30 from a population with $p=0.5$, recording $\hat{p}$ for each and drawing a dot plot. (a) Around what value should the dot plot be centred? (b) Will most values lie above, below, or on both sides of that centre? (c) State the approximate standard deviation of $\hat{p}$, using $\sqrt{p(1-p)/n}$.

Q23. In a simulation, ten random samples of size 20 were taken from a population with $p=0.3$; the recorded sample proportions were

$$0.25, 0.35, 0.30, 0.20, 0.40, 0.30, 0.15, 0.35, 0.30, 0.40.$$

(a) Find the mean of these simulated proportions; (b) state the range; (c) comment on how well the simulation reflects $p=0.3$.

Q24. A simulation of the sampling of $\hat{p}$ can be changed in two ways. Explain the effect on the dot plot of $\hat{p}$ of (a) increasing the size n of each sample, and (b) increasing the number M of simulated samples.

Answers

Q1. (a) $n=40$, $X=18$; (b) $\hat{p}=\frac{9}{20}=0.45$; (c) statistic. **Q2.** (a) $\hat{p}=\frac{3}{20}=0.15$; (b) the population proportion p of undersized bolts; (c) sampling variation. **Q3.** (a) $0, 1/8, 1/4, 3/8, 1/2, 5/8, 3/4, 7/8, 1$; (b) $\hat{p}=\frac{1}{4}$; (c) 6 successes.

Q4. (a) $\frac{5103}{16384}\approx 0.3115$; (b) $\frac{6561}{65536}\approx 0.1001$; (c) $\hat{p}=\frac{1}{4}$. **Q5.** (a) smallest 0.15, largest 0.40; (b) 0.30; (c) spread roughly symmetrically about $p=0.25$, most within a little of it. **Q6.** (a) parameter; (b) statistic; (c) parameter; (d) statistic. **Q7.** $\hat{p}=\frac{2}{5}=0.4$. **Q8.** $\hat{p}=\frac{1}{4}=0.25$; $p=\frac{13}{52}=\frac{1}{4}$. **Q9.** $\hat{p}=\frac{3}{20}=0.15$. **Q10.** $\hat{p}=\frac{7}{20}=0.35$. **Q11.** $0, 1/10, 1/5, 3/10, 2/5, 1/2, 3/5, 7/10, 4/5, 9/10, 1$. **Q12.** $\hat{p}$: $0, 1/4, 1/2, 3/4, 1$ with Pr: $\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}$.

Q13. $\frac{7203}{20000}\approx 0.3602$. **Q14.** $\hat{p}$ depends on the random sample chosen (so it varies); p is a fixed feature of the population. **Q15.** The sample of size 500 — larger n gives less sampling variation, so $\hat{p}$ is likely closer to p. **Q16.** $\hat{p}=\frac{3}{5}=0.6$; statistic. **Q17.** $\hat{p}=\frac{4}{25}=0.16$. **Q18.** $Pr(\hat{p}\geq 3/5)=Pr(X\geq 3)=\frac{144+48+32}{3125}\cdot\ldots=\frac{992}{3125}\approx 0.3174$.

Q19. $\hat{p}=\frac{646}{1000}=\frac{323}{500}=0.646$; it comes from one random sample, so sampling variation means it differs from p.

Q20. $n=\frac{16}{0.32}=50$. **Q21.** Both are centred on $p=0.3$; the $n=40$ distribution is **narrower and taller** (less spread), so $\hat{p}$ is more likely to be close to p. **Q22.** (a) 0.5; (b) both sides (roughly symmetric about 0.5); (c) $\sqrt{0.5\times 0.5/30}\approx 0.091$. **Q23.** (a) 0.30; (b) $0.40-0.15=0.25$; (c) the mean equals $p=0.3$ and the values spread on both sides of it — a good reflection of the distribution of $\hat{p}$. **Q24.** (a) larger $n \Rightarrow$ smaller spread, so $\hat{p}$ clusters more tightly about p; (b) larger $M \Rightarrow$ same centre and spread, but a clearer, smoother picture of the distribution.

Fully-worked solutions

Q1. $n=40$, $X=18$. $\hat{p}=\frac{18}{40}=\frac{9}{20}=0.45$. It is computed from a sample, so it is a **statistic**.

Q2. $\hat{p}=\frac{30}{200}=\frac{3}{20}=0.15$. It estimates p, the proportion of **all** bolts that are undersized. A second sample of 200 contains different bolts, so its count (and hence $\hat{p}=0.18$) differs — this is sampling variation.

Q3. For $n=8$, $\hat{p}=\frac{X}{8}$ for $X=0,\ldots,8$, giving $0, 1/8, 1/4, 3/8, 1/2, 5/8, 3/4, 7/8, 1$. With $X=2$: $\hat{p}=\frac{2}{8}=\frac{1}{4}$. For $\hat{p}=\frac{3}{4}=\frac{6}{8}$ we need $X=6$ successes.

Q4. $X\sim \text{Bi}(8, 1/4)$. (a) $Pr(\hat{p}=1/4)=Pr(X=2)=\binom{8}{2}(1/4)^2(3/4)^6=\frac{5103}{16384}\approx 0.3115$. (b) $Pr(X=0)=(3/4)^8=\frac{6561}{65536}\approx 0.1001$. (c) The largest probability is at $X=2$, so the most likely value is $\hat{p}=\frac{1}{4}$.

Q5. Reading the dot plot: (a) smallest $\hat{p}=0.15$, largest $\hat{p}=0.40$; (b) $\hat{p}=0.30$ occurs most often (three dots); (c) the values are spread on both sides of $p=0.25$, most of them within about 0.1 of it — typical sampling variation.

Q6. (a) parameter (all voters); (b) statistic (a survey); (c) parameter (every student); (d) statistic (a sample).

Q7. $\hat{p}=\frac{24}{60}=\frac{2}{5}=0.4$.

Q8. $\hat{p}=\frac{13}{52}=\frac{1}{4}=0.25$. There are 13 hearts in 52 cards, so $p=\frac{13}{52}=\frac{1}{4}$ (here $\hat{p}=p$ by coincidence of this sample).

Q9. $\hat{p}=\frac{45}{300}=\frac{3}{20}=0.15$.

Q10. $\hat{p}=\frac{84}{240}=\frac{7}{20}=0.35$.

Q11. $\hat{p}=\frac{X}{10}$ for $X=0,\ldots,10$: $0,1/10,1/5,3/10,2/5,1/2,3/5,7/10,4/5,9/10,1$.

Q12. $X\sim\mathrm{Bi}(4,1/2)$, so $Pr(X=k)=\binom{4}{k}(1/2)^4=\frac{1}{16}\binom{4}{k}$:

$\hat{p}$	0	$1/4$	$1/2$	$3/4$	1
$Pr(\hat{p})$	$\frac{1}{16}$	$\frac{4}{16}=\frac{1}{4}$	$\frac{6}{16}=\frac{3}{8}$	$\frac{4}{16}=\frac{1}{4}$	$\frac{1}{16}$

Q13. $X\sim\mathrm{Bi}(5,3/10)$; $\hat{p}=\frac{1}{5}\Rightarrow X=1$. $Pr(X=1)=\binom{5}{1}(3/10)^1(7/10)^4=\frac{7203}{20000}\approx 0.3602$.

Q14. The population proportion p is a fixed number determined by the whole population, so it does not change. The sample proportion $\hat{p}=X/n$ depends on which members happen to fall in the random sample, and X is a random variable; hence $\hat{p}$ is also a random variable.

Q15. The sample of size 500. The standard deviation of $\hat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$, which decreases as n increases, so a larger sample has less sampling variation and $\hat{p}$ is more likely to be close to p.

Q16. $\hat{p}=\frac{27}{45}=\frac{3}{5}=0.6$; it is computed from a sample, so it is a **statistic**.

Q17. $\hat{p}=\frac{8}{50}=\frac{4}{25}=0.16$.

Q18. $X\sim\mathrm{Bi}(5,2/5)$; $\hat{p}\geq 3/5\Leftrightarrow X\geq 3$. $Pr(X\geq 3)=Pr(X=3)+Pr(X=4)+Pr(X=5)=\frac{144}{625}+\frac{48}{625}+\frac{32}{3125}$.
Converting to a common denominator: $\frac{720+240+32}{3125}=\frac{992}{3125}\approx 0.3174$.

Q19. $\hat{p}=\frac{646}{1000}=\frac{323}{500}=0.646$. The poll is one random sample of 1000 from all voters, so sampling variation means $\hat{p}$ generally differs from the true population proportion p; it is a point estimate of p.

Q20. $\hat{p}=\frac{X}{n}\Rightarrow 0.32=\frac{16}{n}\Rightarrow n=\frac{16}{0.32}=50$.

Q21. Both distributions are centred on the population proportion $p=0.3$. Increasing n from 10 to 40 makes the distribution of $\hat{p}$ **narrower and more peaked** (its standard deviation $\sqrt{p(1-p)/n}$ is halved), so a larger sample's proportion is more tightly clustered about p.

Q22. The distribution of $\hat{p}$ is centred on p, so the dot plot should be centred on 0.5, with values roughly symmetric on **both sides** of 0.5. Its standard deviation is $\sqrt{\frac{0.5\times 0.5}{30}}=\sqrt{0.008\overline{3}}\approx 0.091$.

Q23. Mean $= \dfrac{0.25+0.35+0.30+0.20+0.40+0.30+0.15+0.35+0.30+0.40}{10} = \dfrac{3.00}{10} = 0.30$. Range $= 0.40 - 0.15 = 0.25$. The mean of the simulated proportions equals the population proportion $p = 0.3$, and the values are spread on both sides of 0.3 — the simulation reflects that $\hat{p}$ is centred on p with sampling variation about it.

Q24. Increasing the sample size n reduces the standard deviation $\sqrt{p(1-p)/n}$ of $\hat{p}$, so the dot plot becomes **narrower** — the proportions cluster more tightly about p. Increasing the number of simulations M adds more dots but does **not** change where they are centred or how far they spread; it simply produces a fuller, smoother picture of the (fixed) distribution of $\hat{p}$.

Theory

To estimate an unknown **population proportion** p (the fraction of a population with some characteristic), we take a random sample of size n and record X, the number in the sample with the characteristic. Provided the sample is drawn at random, $X \sim Bi(n, p)$.

The sample proportion. The **sample proportion** is the random variable

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

It varies from sample to sample, so it is itself a random variable. Because X can equal $0, 1, 2, \ldots, n$, the sample proportion $\hat{P}$ takes the values

$$0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} = 1.$$

Its exact distribution. Since $\hat{P} = \frac{k}{n}$ exactly when $X = k$,

$$Pr\left(\hat{P} = \frac{k}{n}\right) = Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

So the exact distribution of $\hat{P}$ is just the binomial distribution of X, read on the rescaled values $\frac{k}{n}$. To find a cumulative probability, convert back to X: for example $Pr(\hat{P} \le c) = Pr(X \le nc)$.

Mean and standard deviation. Using $E(aX) = aE(X)$ and $Var(aX) = a^2 Var(X)$ with $a = \frac{1}{n}$:

$$E(\hat{P}) = \frac{1}{n} E(X) = \frac{np}{n} = p,$$

$$Var(\hat{P}) = \frac{1}{n^2} Var(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}, sd(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}.$$

So the distribution of $\hat{P}$ is **centred on the population proportion** p, and its spread **decreases as the sample size** n **increases** (larger samples give proportions clustered more tightly about p).

Worked examples

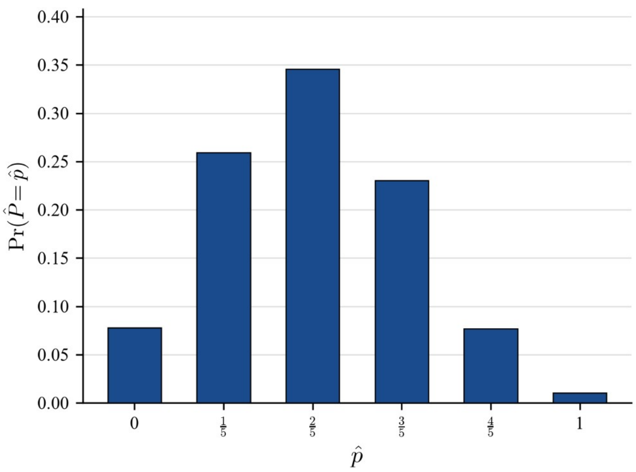
Example 1 — scaffolded

A population has proportion $p = \frac{2}{5}$. A random sample of size $n = 5$ is taken. Construct the exact distribution of $\hat{P} = \frac{X}{5}$, and find $E(\hat{P})$ and $sd(\hat{P})$.

Working	Comment
$X \sim Bi(5, 2/5)$, so $\hat{P} = \frac{X}{5}$ takes $0, 1/5, 2/5, 3/5, 4/5, 1$.	The values are $\frac{k}{n}$ for $k = 0, \ldots, 5$.

Working	Comment
$Pr\left(\widehat{P}=\frac{k}{5}\right)=\binom{5}{k}\left(\frac{2}{5}\right)^{k}\left(\frac{3}{5}\right)^{5-k}.$	Each probability is $Pr(X=k)$.
$Pr(X=0)=\frac{243}{3125}$, $Pr(X=1)=\frac{162}{625}$, $Pr(X=2)=\frac{216}{625}$.	Compute term by term.
$Pr(X=3)=\frac{144}{625}$, $Pr(X=4)=\frac{48}{625}$, $Pr(X=5)=\frac{32}{3125}$.	The six probabilities sum to 1.
$E(\widehat{P})=p=\frac{2}{5}$.	The mean of $\widehat{P}$ is the population proportion.
$sd(\widehat{P})=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{\frac{2}{5}\cdot\frac{3}{5}}{5}}=\sqrt{\frac{6}{125}}=\frac{\sqrt{30}}{25}.$	Substitute into the standard-deviation formula.

$\hat{p}$	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$Pr(\widehat{P}=\hat{p})$	$\frac{243}{3125}$	$\frac{162}{625}$	$\frac{216}{625}$	$\frac{144}{625}$	$\frac{48}{625}$	$\frac{32}{3125}$

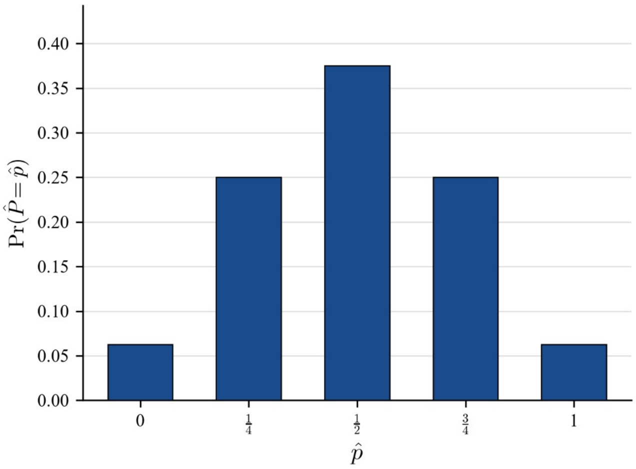


Example 2 — scaffolded

A fair coin is tossed $n=4$ times and $\widehat{P}$ is the proportion of heads. Construct the distribution of $\widehat{P}$, find $Pr\left(\widehat{P}\geq\frac{3}{4}\right)$, and comment on the shape.

Working	Comment
$X\sim Bi(4,1/2)$; $\widehat{P}=\frac{X}{4}$ takes $0,1/4,1/2,3/4,1$.	Here $p=\frac{1}{2}$.
Probabilities $\frac{1}{16},\frac{1}{4},\frac{3}{8},\frac{1}{4},\frac{1}{16}$.	$Pr(X=k)=\binom{4}{k}\left(\frac{1}{2}\right)^{4}$.
$Pr\left(\widehat{P}\geq\frac{3}{4}\right)=Pr\left(\widehat{P}=\frac{3}{4}\right)+Pr(\widehat{P}=1)=\frac{1}{4}+\frac{1}{16}=\frac{5}{16}.$	Add the top two values.

Working	Comment
The distribution is **symmetric** about $\frac{1}{2}$, with $E\left(\widehat{P}\right)=\frac{1}{2}$ and $sd\left(\widehat{P}\right)=\sqrt{\frac{\frac{1}{2}\cdot\frac{1}{2}}{4}}=\frac{1}{4}$.	Symmetry follows from $p=\frac{1}{2}$.



Example 3 — unscaffolded

For a population with $p=\frac{1}{3}$ and a sample of size $n=6$, find $Pr\left(\widehat{P}=\frac{1}{3}\right)$, $Pr\left(\widehat{P}<\frac{1}{2}\right)$, $E\left(\widehat{P}\right)$ and $sd\left(\widehat{P}\right)$.

$X\sim Bi(6,1/3)$, $\widehat{P}=\frac{X}{6}$.

$$Pr\left(\widehat{P}=\frac{1}{3}\right)=Pr(X=2)=\binom{6}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^4=15\cdot\frac{16}{729}=\frac{80}{243}.$$

$$Pr\left(\widehat{P}<\frac{1}{2}\right)=Pr(X\leq 2)=\frac{64}{729}+\frac{64}{243}+\frac{80}{243}=\frac{496}{729}.$$

$$E\left(\widehat{P}\right)=p=\frac{1}{3};\ sd\left(\widehat{P}\right)=\sqrt{\frac{\frac{1}{3}\cdot\frac{2}{3}}{6}}=\sqrt{\frac{1}{27}}=\frac{\sqrt{3}}{9}.$$

$\therefore \widehat{P}$ is centred on $\frac{1}{3}$ with standard deviation $\frac{\sqrt{3}}{9}\approx 0.192$.

Example 4 — unscaffolded

In a large population, $30\,\%$ support a proposal. A random sample of $n=10$ is taken. Find $E\left(\widehat{P}\right)$, $sd\left(\widehat{P}\right)$, and $Pr\left(\widehat{P}\leq\frac{1}{5}\right)$, correct to four decimal places.

Here $p=\frac{3}{10}$, $n=10$, $X\sim Bi(10,3/10)$.

$$E\left(\widehat{P}\right)=p=\frac{3}{10};\ sd\left(\widehat{P}\right)=\sqrt{\frac{\frac{3}{10}\cdot\frac{7}{10}}{10}}=\sqrt{\frac{21}{1000}}=\frac{\sqrt{210}}{100}\approx 0.1449.$$

$Pr\left(\hat{P}\le\frac{1}{5}\right)=Pr(X\le 2)=0.3828$ (to 4 d.p.).

$\therefore E(\hat{P})=0.3$, $sd(\hat{P})\approx 0.1449$, and about 38% of samples have $\hat{P}\le 0.2$.

Practice questions

Scaffolded

Q1. A population has $p=\frac{1}{4}$; a sample of size $n=4$ is taken. (a) List the possible values of $\hat{P}$. (b) Find $Pr\left(\hat{P}=\frac{1}{2}\right)$. (c) Find $E(\hat{P})$ and $sd(\hat{P})$.

Q2. A population has $p=\frac{3}{5}$; a sample of size $n=5$ is taken. (a) Construct the exact distribution of $\hat{P}$. (b) Find $Pr\left(\hat{P}\ge\frac{3}{5}\right)$. (c) Find $E(\hat{P})$ and $sd(\hat{P})$.

Q3. A fair die is rolled $n=6$ times; $\hat{P}$ is the proportion of even numbers (so $p=\frac{1}{2}$). (a) Find $Pr\left(\hat{P}=\frac{1}{2}\right)$. (b) Find $E(\hat{P})$ and $sd(\hat{P})$. (c) Explain why the distribution of $\hat{P}$ is symmetric.

Q4. A population has $p=\frac{1}{4}$; a sample of size $n=8$ is taken. (a) State $E(\hat{P})$. (b) Find $sd(\hat{P})$ in exact form. (c) Find $Pr\left(\hat{P}\le\frac{1}{4}\right)$, correct to four decimal places.

Q5. A population has $p=\frac{1}{5}$; a sample of size $n=5$ is taken. (a) Construct the exact distribution of $\hat{P}$. (b) Find $Pr\left(\hat{P}\le\frac{2}{5}\right)$. (c) Find $E(\hat{P})$ and $sd(\hat{P})$.

Q6. A population has $p=\frac{3}{4}$; a sample of size $n=4$ is taken. (a) Find $Pr(\hat{P}=1)$. (b) Find $Pr\left(\hat{P}\ge\frac{1}{2}\right)$. (c) Find $E(\hat{P})$ and $sd(\hat{P})$.

Unscaffolded

Q7. For $p=\frac{1}{2}$, $n=5$: construct the exact distribution of $\hat{P}$ and state $E(\hat{P})$ and $sd(\hat{P})$.

Q8. For $p=\frac{2}{3}$, $n=6$: find $Pr\left(\hat{P}\ge\frac{2}{3}\right)$, $E(\hat{P})$ and $sd(\hat{P})$.

Q9. For $p=\frac{2}{5}$, $n=4$: construct the distribution of $\hat{P}$, find $Pr\left(\hat{P}\le\frac{1}{2}\right)$, and state $E(\hat{P})$ and $sd(\hat{P})$.

Q10. For $p=\frac{1}{2}$, $n=8$: find $E(\hat{P})$, $sd(\hat{P})$ (exact), and $Pr\left(\hat{P}\le\frac{1}{4}\right)$.

Q11. For $p=\frac{4}{5}$, $n=5$: construct the distribution of $\hat{P}$, find $Pr(\hat{P}=1)$, and state $E(\hat{P})$ and $sd(\hat{P})$.

Q12. For $p=\frac{1}{5}$, $n=10$: find $E(\hat{P})$, $sd(\hat{P})$ (exact), and $Pr\left(\hat{P}\le\frac{1}{10}\right)$ correct to four decimal places.

Q13. For $p=\frac{1}{6}$, $n=6$: find $E(\hat{P})$, $sd(\hat{P})$ (exact), and $Pr\left(\hat{P}=\frac{1}{6}\right)$.

Q14. For $p=\frac{1}{3}$, $n=4$: construct the distribution of $\hat{P}$, find $Pr\left(\hat{P}\ge\frac{1}{2}\right)$, and state $E(\hat{P})$ and $sd(\hat{P})$.

Q15. For $p=\frac{3}{10}$, $n=5$: find $E(\hat{P})$, $sd(\hat{P})$ (exact), and $Pr\left(\hat{P}\le\frac{2}{5}\right)$ correct to four decimal places.

Q16. For $p=\frac{1}{3}$, $n=9$: find $E(\hat{P})$, $sd(\hat{P})$ (exact), and $Pr\left(\hat{P}\le\frac{1}{3}\right)$ correct to four decimal places.

Q17. For $p=\frac{3}{5}$, $n=6$: find $E(\hat{P})$, $sd(\hat{P})$ (exact), and $Pr\left(\hat{P}\ge\frac{1}{2}\right)$.

Q18. For $p=\frac{2}{5}$, $n=10$: find $E(\hat{P})$, $sd(\hat{P})$ (exact), and $Pr\left(\hat{P}\ge\frac{1}{2}\right)$ correct to four decimal places.

Q19. For $p=\frac{1}{5}$, $n=4$: construct the distribution of $\hat{P}$, find $Pr(\hat{P}=0)$, and state $E(\hat{P})$ and $sd(\hat{P})$.

Q20. For $p=\frac{2}{3}$, $n=5$: construct the distribution of $\hat{P}$, find $Pr\left(\hat{P}\ge\frac{3}{5}\right)$, and state $E(\hat{P})$ and $sd(\hat{P})$.

Q21. The sample proportion is $\hat{P}=\frac{X}{n}$, where $X\sim Bi(n,p)$. **Show that** $E(\hat{P})=p$ and $sd(\hat{P})=\sqrt{\frac{p(1-p)}{n}}$.

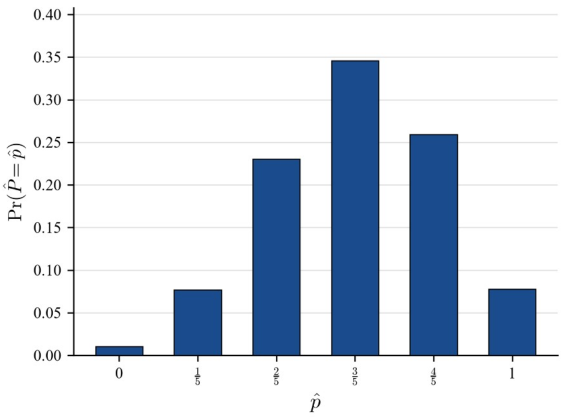
Answers

Q1. (a) $0, 1/4, 1/2, 3/4, 1$; (b) $\frac{27}{128}$; (c) $E=\frac{1}{4}$, $sd=\frac{\sqrt{3}}{8}$. **Q2.** (b) $\frac{2133}{3125}$; (c) $E=\frac{3}{5}$, $sd=\frac{\sqrt{30}}{25}$. **Q3.** (a) $\frac{5}{16}$; (b) $E=\frac{1}{2}$, $sd=\frac{\sqrt{6}}{12}$; (c) $p=1/2$. **Q4.** (a) $\frac{1}{4}$; (b) $\frac{\sqrt{6}}{16}$; (c) 0.6785. **Q5.** (b) $\frac{2944}{3125}$; (c) $E=\frac{1}{5}$, $sd=\frac{2\sqrt{5}}{25}$. **Q6.** (a) $\frac{81}{256}$; (b) $\frac{243}{256}$; (c) $E=\frac{3}{4}$, $sd=\frac{\sqrt{3}}{8}$. **Q7.** $E=\frac{1}{2}$, $sd=\frac{\sqrt{5}}{10}$. **Q8.** $Pr=\frac{496}{729}$; $E=\frac{2}{3}$, $sd=\frac{\sqrt{3}}{9}$. **Q9.** $Pr(\hat{P}\le 1/2)=\frac{513}{625}$; $E=\frac{2}{5}$, $sd=\frac{\sqrt{6}}{10}$. **Q10.** $E=\frac{1}{2}$, $sd=\frac{\sqrt{2}}{8}$, $Pr=\frac{37}{256}\approx 0.1445$. **Q11.** $Pr(\hat{P}=1)=\frac{1024}{3125}$; $E=\frac{4}{5}$, $sd=\frac{2\sqrt{5}}{25}$. **Q12.** $E=\frac{1}{5}$, $sd=\frac{\sqrt{10}}{25}$, $Pr=0.3758$. **Q13.** $E=\frac{1}{6}$, $sd=\frac{\sqrt{30}}{36}$, $Pr(\hat{P}=1/6)=\frac{3125}{7776}$. **Q14.** $Pr(\hat{P}\ge 1/2)=\frac{11}{27}$; $E=\frac{1}{3}$, $sd=\frac{\sqrt{2}}{6}$. **Q15.** $E=\frac{3}{10}$, $sd=\frac{\sqrt{105}}{50}$, $Pr=0.8369$. **Q16.** $E=\frac{1}{3}$, $sd=\frac{\sqrt{2}}{9}$, $Pr=0.6503$. **Q17.** $E=\frac{3}{5}$, $sd=\frac{1}{5}$, $Pr(\hat{P}\ge 1/2)=\frac{513}{625}$. **Q18.** $E=\frac{2}{5}$, $sd=\frac{\sqrt{15}}{25}$, $Pr=0.3669$. **Q19.** $Pr(\hat{P}=0)=\frac{256}{625}$; $E=\frac{1}{5}$, $sd=\frac{1}{5}$. **Q20.** $Pr(\hat{P}\ge 3/5)=\frac{64}{81}$; $E=\frac{2}{3}$, $sd=\frac{\sqrt{10}}{15}$. **Q21.** Proof — see solutions.

Fully-worked solutions

Q1. $X \sim Bi(4, 1/4)$. (a) $\hat{P}$ takes $0, 1/4, 1/2, 3/4, 1$. (b) $Pr(\hat{P}=1/2)=Pr(X=2)=\binom{4}{2}(1/4)^2(3/4)^2=6\cdot\frac{9}{256}=\frac{27}{128}$. (c) $E(\hat{P})=p=\frac{1}{4}$; $sd(\hat{P})=\sqrt{\frac{\frac{1}{4}\cdot\frac{3}{4}}{4}}=\sqrt{\frac{3}{64}}=\frac{\sqrt{3}}{8}$.

Q2. $X \sim Bi(5, 3/5)$. (a) $\hat{p}$: $0, 1/5, 2/5, 3/5, 4/5, 1$ with $Pr=\frac{32}{3125}, \frac{48}{625}, \frac{144}{625}, \frac{216}{625}, \frac{162}{625}, \frac{243}{3125}$.



(b) $Pr(\hat{P}\ge 3/5)=\frac{216}{625}+\frac{162}{625}+\frac{243}{3125}=\frac{2133}{3125}$.

(c) $E=\frac{3}{5}$; $sd=\sqrt{\frac{\frac{3}{5}\cdot\frac{2}{5}}{5}}=\sqrt{\frac{6}{125}}=\frac{\sqrt{30}}{25}$.

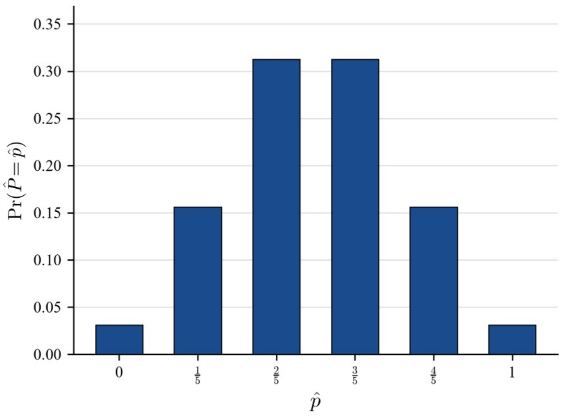
Q3. $X \sim Bi(6, 1/2)$. (a) $Pr(\hat{P}=1/2)=Pr(X=3)=\binom{6}{3}(1/2)^6=\frac{20}{64}=\frac{5}{16}$. (b) $E=\frac{1}{2}$; $sd=\sqrt{\frac{\frac{1}{2}\cdot\frac{1}{2}}{6}}=\sqrt{\frac{1}{24}}=\frac{\sqrt{6}}{12}$. (c) Because $p=\frac{1}{2}$, $Pr(X=k)=Pr(X=6-k)$, so the distribution is symmetric about $\hat{p}=\frac{1}{2}$.

Q4. $X \sim Bi(8, 1/4)$. (a) $E(\hat{P}) = \frac{1}{4}$. (b) $sd(\hat{P}) = \sqrt{\frac{\frac{1}{4} \cdot \frac{3}{4}}{8}} = \sqrt{\frac{3}{128}} = \frac{\sqrt{6}}{16}$. (c) $Pr(\hat{P} \le 1/4) = Pr(X \le 2) = 0.6785$ (4 d.p.).

Q5. $X \sim Bi(5, 1/5)$. (a) $\hat{p}$: $0, 1/5, 2/5, 3/5, 4/5, 1$ with $Pr = \frac{1024}{3125}, \frac{256}{625}, \frac{128}{625}, \frac{32}{625}, \frac{4}{625}, \frac{1}{3125}$. (b) $Pr(\hat{P} \le 2/5) = \frac{1024}{3125} + \frac{256}{625} + \frac{128}{625} = \frac{2944}{3125}$. (c) $E = \frac{1}{5}$; $sd = \sqrt{\frac{\frac{1}{5} \cdot \frac{4}{5}}{5}} = \sqrt{\frac{4}{125}} = \frac{2\sqrt{5}}{25}$.

Q6. $X \sim Bi(4, 3/4)$. (a) $Pr(\hat{P} = 1) = Pr(X = 4) = (3/4)^4 = \frac{81}{256}$. (b) $Pr(\hat{P} \ge 1/2) = \frac{27}{128} + \frac{27}{64} + \frac{81}{256} = \frac{243}{256}$. (c) $E = \frac{3}{4}$; $sd = \sqrt{\frac{\frac{3}{4} \cdot \frac{1}{4}}{4}} = \frac{\sqrt{3}}{8}$.

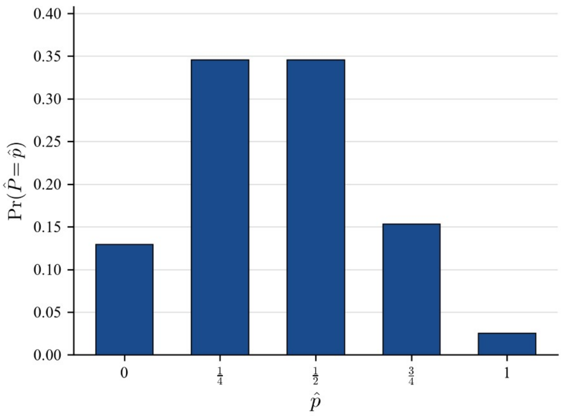
Q7. $X \sim Bi(5, 1/2)$. $\hat{p}$: $0, 1/5, 2/5, 3/5, 4/5, 1$ with $Pr = \frac{1}{32}, \frac{5}{32}, \frac{5}{16}, \frac{5}{16}, \frac{5}{32}, \frac{1}{32}$. $E = \frac{1}{2}$; $sd = \sqrt{\frac{1}{20}} = \frac{\sqrt{5}}{10}$.



Q8. $X \sim Bi(6, 2/3)$. $Pr(\hat{P} \ge 2/3) = Pr(X \ge 4) = \frac{80}{243} + \frac{64}{243} + \frac{64}{729} = \frac{496}{729}$. $E = \frac{2}{3}$; $sd = \sqrt{\frac{\frac{2}{3} \cdot \frac{1}{3}}{6}} = \sqrt{\frac{1}{27}} = \frac{\sqrt{3}}{9}$.

Q9. $X \sim Bi(4, 2/5)$. $\hat{p}$: $0, 1/4, 1/2, 3/4, 1$ with $Pr = \frac{81}{625}, \frac{216}{625}, \frac{216}{625}, \frac{96}{625}, \frac{16}{625}$.

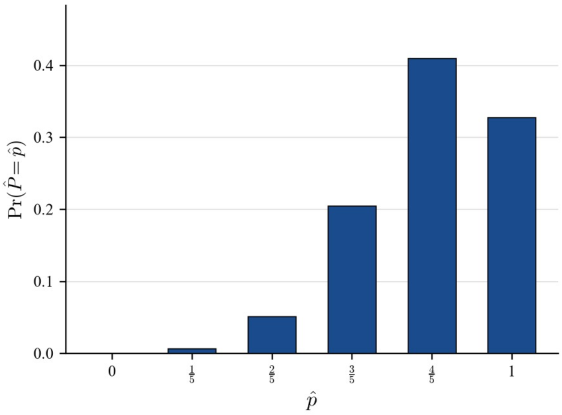
$Pr(\hat{P} \le 1/2) = \frac{81 + 216 + 216}{625} = \frac{513}{625}$. $E = \frac{2}{5}$; $sd = \sqrt{\frac{\frac{2}{5} \cdot \frac{3}{5}}{4}} = \sqrt{\frac{6}{100}} = \frac{\sqrt{6}}{10}$.



Q10. $X \sim Bi(8, 1/2)$. $E = \frac{1}{2}$; $sd = \sqrt{\frac{1}{32}} = \frac{\sqrt{2}}{8}$. $Pr(\hat{P} \le 1/4) = Pr(X \le 2) = \frac{37}{256} \approx 0.1445$.

Q11. $X \sim Bi(5, 4/5)$. $\hat{p}$: $0, 1/5, 2/5, 3/5, 4/5, 1$ with $Pr = \frac{1}{3125}, \frac{4}{625}, \frac{32}{625}, \frac{128}{625}, \frac{256}{625}, \frac{1024}{3125}$.

$Pr(\hat{P} = 1) = \frac{1024}{3125}$. $E = \frac{4}{5}$; $sd = \sqrt{\frac{\frac{4}{5} \cdot \frac{1}{5}}{5}} = \frac{2\sqrt{5}}{25}$.

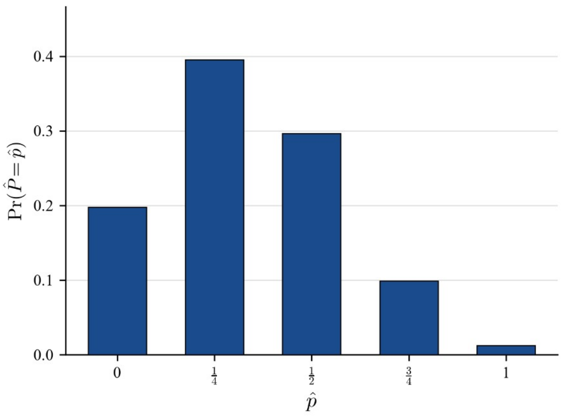


Q12. $X \sim Bi(10, 1/5)$. $E = \frac{1}{5}$; $sd = \sqrt{\frac{\frac{1}{5} \cdot \frac{4}{5}}{10}} = \sqrt{\frac{2}{125}} = \frac{\sqrt{10}}{25}$. $Pr(\hat{P} \le 1/10) = Pr(X \le 1) = 0.3758$ (4 d.p.).

Q13. $X \sim Bi(6, 1/6)$. $E = \frac{1}{6}$; $sd = \sqrt{\frac{\frac{1}{6} \cdot \frac{5}{6}}{6}} = \sqrt{\frac{5}{216}} = \frac{\sqrt{30}}{36}$. $Pr(\hat{P} = 1/6) = Pr(X = 1) = \binom{6}{1}(1/6)(5/6)^5 = \frac{3125}{7776}$.

Q14. $X \sim Bi(4, 1/3)$. $\hat{p}$: $0, 1/4, 1/2, 3/4, 1$ with $Pr = \frac{16}{81}, \frac{32}{81}, \frac{8}{27}, \frac{8}{81}, \frac{1}{81}$. $Pr(\hat{P} \ge 1/2) = \frac{8}{27} + \frac{8}{81} + \frac{1}{81} = \frac{11}{27}$.

$E = \frac{1}{3}$; $sd = \sqrt{\frac{\frac{1}{3} \cdot \frac{2}{3}}{4}} = \sqrt{\frac{1}{18}} = \frac{\sqrt{2}}{6}$.



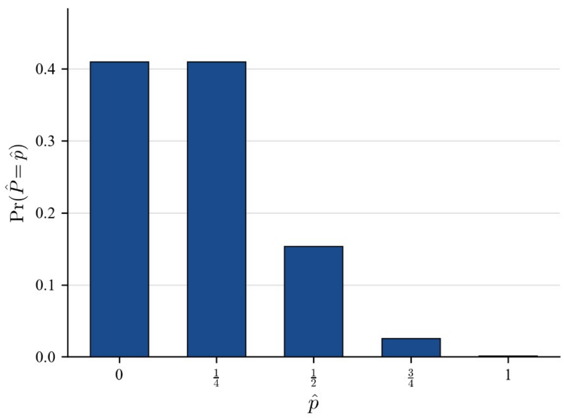
Q15. $X \sim Bi(5, 3/10)$. $E = \frac{3}{10}$; $sd = \sqrt{\frac{\frac{3}{10} \cdot \frac{7}{10}}{5}} = \sqrt{\frac{21}{500}} = \frac{\sqrt{105}}{50}$. $Pr(\hat{P} \le 2/5) = Pr(X \le 2) = 0.8369$ (4 d.p.).

Q16. $X \sim Bi(9, 1/3)$. $E = \frac{1}{3}$; $sd = \sqrt{\frac{\frac{1}{3} \cdot \frac{2}{3}}{9}} = \sqrt{\frac{2}{81}} = \frac{\sqrt{2}}{9}$. $Pr(\hat{P} \le 1/3) = Pr(X \le 3) = 0.6503$ (4 d.p.).

Q17. $X \sim Bi(6, 3/5)$. $E = \frac{3}{5}$; $sd = \sqrt{\frac{\frac{3}{5} \cdot \frac{2}{5}}{6}} = \sqrt{\frac{1}{25}} = \frac{1}{5}$. $Pr(\hat{P} \ge 1/2) = Pr(X \ge 3) = \frac{513}{625}$.

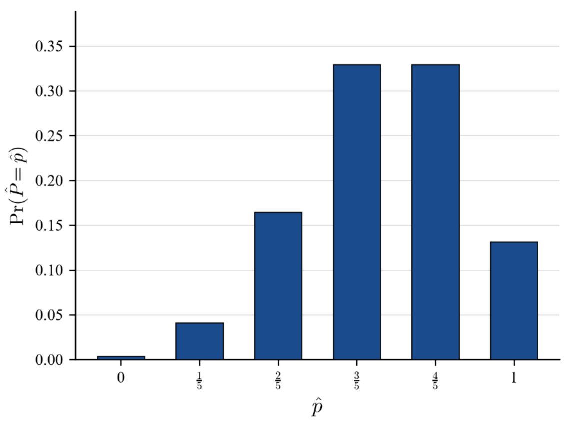
Q18. $X \sim Bi(10, 2/5)$. $E = \frac{2}{5}$; $sd = \sqrt{\frac{\frac{2}{5} \cdot \frac{3}{5}}{10}} = \sqrt{\frac{3}{125}} = \frac{\sqrt{15}}{25}$. $Pr(\hat{P} \ge 1/2) = Pr(X \ge 5) = 0.3669$ (4 d.p.).

Q19. $X \sim Bi(4, 1/5)$. $\hat{p}$: $0, 1/4, 1/2, 3/4, 1$ with $Pr = \frac{256}{625}, \frac{256}{625}, \frac{96}{625}, \frac{16}{625}, \frac{1}{625}$. $Pr(\hat{P} = 0) = \frac{256}{625}$. $E = \frac{1}{5}$; $sd = \sqrt{\frac{\frac{1}{5} \cdot \frac{4}{5}}{4}} = \sqrt{\frac{1}{25}} = \frac{1}{5}$.



Q20. $X \sim Bi(5, 2/3)$. $\hat{p}$: $0, 1/5, 2/5, 3/5, 4/5, 1$ with $Pr = \frac{1}{243}, \frac{10}{243}, \frac{40}{243}, \frac{80}{243}, \frac{80}{243}, \frac{32}{243}$.

$Pr(\hat{P} \ge 3/5) = \frac{80+80+32}{243} = \frac{192}{243} = \frac{64}{81}$. $E = \frac{2}{3}$; $sd = \sqrt{\frac{\frac{2}{3} \cdot \frac{1}{3}}{5}} = \sqrt{\frac{2}{45}} = \frac{\sqrt{10}}{15}$.



Q21. $\hat{P} = \frac{X}{n}$ with $X \sim Bi(n, p)$, so $E(X) = np$ and $Var(X) = np(1-p)$. Using $E(aX) = aE(X)$ with $a = \frac{1}{n}$:

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} \cdot np = p.$$

Using $Var(aX) = a^2 Var(X)$:

$$Var(\hat{P}) = \frac{1}{n^2} Var(X) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n},$$

so $sd(\hat{P}) = \sqrt{Var(\hat{P})} = \sqrt{\frac{p(1-p)}{n}}$. $\therefore E(\hat{P}) = p$ and $sd(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$, as required.

Chapter 15 Statistical inference and sample proportions — 15C The approximate normality of the sample proportion

Theory

The **sample proportion** $\hat{p} = \frac{X}{n}$, where X is the number of successes in a sample of size n, is a random variable: different samples give different values of $\hat{p}$. Because $X \sim \mathrm{Bi}(n, p)$, the exact distribution of $\hat{p}$ is binomial in form (Section 15B). For a **large** sample this exact distribution is well approximated by a normal distribution.

The normal approximation. For large n,

$$\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right),$$

that is, $\hat{P}$ is approximately normal with

- **mean** $\mu_{\hat{P}} = p$, and
- **standard deviation** $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$.

The mean equals the population proportion, so $\hat{P}$ is centred on p; the standard deviation shrinks like $\frac{1}{\sqrt{n}}$, so a larger sample gives a $\hat{p}$ that clusters more tightly about p.

When is the approximation reasonable? The larger n is, the better; a common guideline is that **both** $np \geq 10$ and $n(1-p) \geq 10$ (some texts use ≥ 5). If p is very close to 0 or 1, or n is small, the approximation is poor.

Finding probabilities. To estimate a probability about $\hat{P}$, use the normal model with the mean and standard deviation above (with technology, e.g. a CAS `normCdf`), or standardise:

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}.$$

Standard error (when p is unknown). In practice p is usually unknown. We then estimate the standard deviation of $\hat{P}$ using the observed sample proportion $\hat{p}$ in place of p:

$$\text{standard error} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

This is used in Section 15D to build confidence intervals.

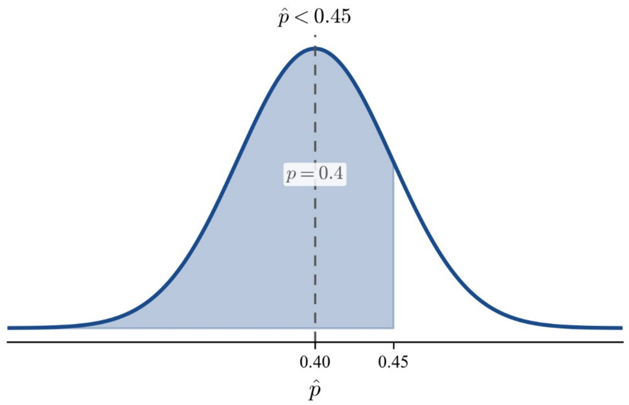
Worked examples

Example 1 — scaffolded

In a large population, 40% support a proposal, so $p = 0.4$. A random sample of $n = 100$ is taken. (a) State the approximate distribution of $\hat{P}$. (b) Find $Pr(\hat{P} < 0.45)$. (c) Find $Pr(\hat{P} > 0.35)$.

Working	Comment
$np = 40$ and $n(1-p) = 60$, both ≥ 10, so the normal approximation is reasonable.	Check the condition first.
Mean $p = 0.4$; sd $= \sqrt{\frac{0.4 \times 0.6}{100}} = \frac{\sqrt{6}}{50} \approx 0.0490$.	Compute $\mu_{\hat{P}}$ and $\sigma_{\hat{P}}$.

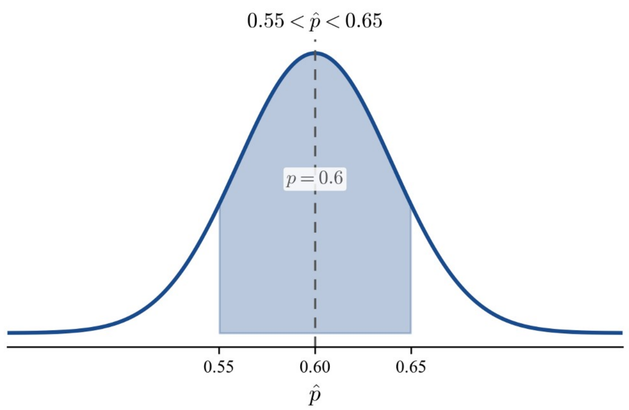
Working	Comment
$\hat{P} \approx N(0.4, 0.0490^2)$.	State the model (variance $= \sigma^2$).
(b) $Pr(\hat{P} < 0.45) = 0.8463$ (to 4 d.p.).	By technology / standardising $z = \dfrac{0.45 - 0.4}{0.0490} \approx 1.02$.
(c) $Pr(\hat{P} > 0.35) = 0.8463$.	By symmetry 0.35 and 0.45 are equidistant from 0.4, so the two probabilities are equal.



Example 2 — scaffolded

A machine fills bottles; the proportion that are underweight is $p = 0.6$ in a certain batch. A sample of $n = 150$ is inspected. Find $Pr(0.55 < \hat{P} < 0.65)$.

Working	Comment
$n\,p = 90$, $n(1 - p) = 60$, both ≥ 10: approximation reasonable.	Condition check.
sd $= \sqrt{\dfrac{0.6 \times 0.4}{150}} = \dfrac{1}{25} = 0.04$; $\hat{P} \approx N(0.6, 0.04^2)$.	$\dfrac{0.24}{150} = 0.0016 = 0.04^2$.
$Pr(0.55 < \hat{P} < 0.65) = 0.8904 - 0.1013 = 0.7887$.	0.55 and 0.65 are each 1.25 sd from the mean.
$\therefore$ about $78.87\,\%$ of such samples give $\hat{p}$ between 0.55 and 0.65.	Interpret in context.



Example 3 — unscaffolded

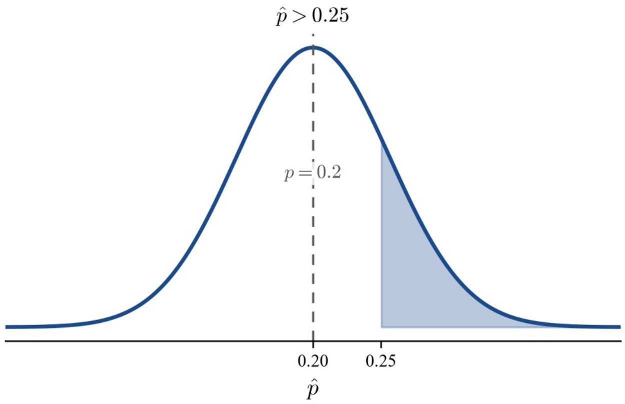
A fair spinner lands on red with probability $p = 0.2$. In $n = 50$ spins, find $Pr(\hat{P} > 0.25)$, where $\hat{P}$ is the proportion of reds.

$n\,p = 10$, $n(1 - p) = 40$, both ≥ 10: the approximation is reasonable.

Mean 0.2; sd $= \sqrt{\frac{0.2 \times 0.8}{50}} = \frac{\sqrt{2}}{25} \approx 0.0566$, so $\hat{P} \approx N(0.2, 0.0566^2)$.

$Pr(\hat{P} > 0.25) = 0.1884$ (to 4 d.p.).

$\therefore$ about $18.84\,\%$ of such samples give a red proportion above 0.25.



Example 4 — unscaffolded

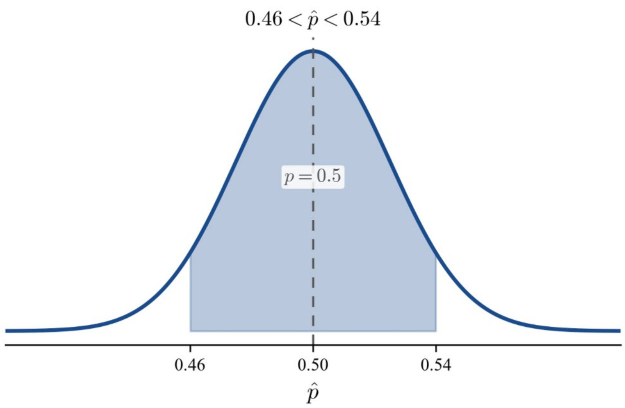
A coin is fair, $p = 0.5$. In $n = 400$ tosses, find the probability that the sample proportion of heads is within 0.04 of 0.5; that is, $Pr(0.46 < \hat{P} < 0.54)$.

$n\,p = n(1 - p) = 200 \geq 10$: reasonable.

Mean 0.5; sd $= \sqrt{\frac{0.5 \times 0.5}{400}} = \frac{1}{40} = 0.025$, so $\hat{P} \approx N(0.5, 0.025^2)$.

$Pr(0.46 < \hat{P} < 0.54) = 0.8904$ (to 4 d.p.).

$\therefore$ there is about an $89.04\,\%$ chance the sample proportion lies within 0.04 of 0.5.



Practice questions

Scaffolded

Q1. In a population $p = 0.3$; a sample of $n = 100$ is taken. (a) State the approximate distribution of $\hat{P}$ (give the standard deviation exactly and to 4 d.p.). (b) Find $Pr(\hat{P} < 0.35)$. (c) Find $Pr(\hat{P} > 0.25)$.

Q2. For $p = 0.7$ and $n = 200$: (a) state the mean and standard deviation of $\hat{P}$; (b) find $Pr(0.66 < \hat{P} < 0.74)$.

Q3. For $p = 0.15$ and $n = 120$: (a) verify the approximation is reasonable; (b) find the standard deviation of $\hat{P}$; (c) find $Pr(\hat{P} > 0.2)$.

Q4. For $p = 0.55$ and $n = 250$: (a) state the distribution of $\hat{P}$; (b) find $Pr(\hat{P} < 0.5)$.

Q5. For $p = 0.8$ and $n = 160$: (a) state the mean and standard deviation of $\hat{P}$; (b) find $Pr(0.75 < \hat{P} < 0.85)$.

Q6. A fair coin ($p = 0.5$) is tossed $n = 100$ times. (a) State the distribution of $\hat{P}$. (b) Find $Pr(\hat{P} > 0.6)$. (c) Relate your answer to the 68–95–99.7 % rule.

Unscaffolded

In each of Q7–Q18, state the approximate distribution of $\hat{P}$ and find the required probability (to 4 d.p.).

Q7. $p = 0.35$, $n = 140$: find $Pr(\hat{P} < 0.4)$.

Q8. $p = 0.25$, $n = 300$: find $Pr(\hat{P} > 0.28)$.

Q9. $p = 0.6$, $n = 90$: find $Pr(0.5 < \hat{P} < 0.7)$.

Q10. $p = 0.45$, $n = 200$: find $Pr(\hat{P} > 0.5)$.

Q11. $p = 0.9$, $n = 100$: find $Pr(\hat{P} < 0.85)$.

Q12. $p = 0.1$, $n = 250$: find $Pr(\hat{P} > 0.12)$.

Q13. $p = 0.5$, $n = 64$: find $Pr(0.4 < \hat{P} < 0.6)$.

Q14. $p = 0.4$, $n = 600$: find $Pr(\hat{P} < 0.38)$.

Q15. $p = 0.7$, $n = 350$: find $Pr(0.68 < \hat{P} < 0.73)$.

Q16. A random sample of $n = 400$ items has sample proportion $\hat{p} = 0.6$ defective. The population proportion is unknown. Find the **standard error** of $\hat{P}$.

Q17. In a survey of $n = 500$ people, 300 said yes. Find $\hat{p}$ and the standard error of $\hat{P}$.

Q18. $p = 0.3$, $n = 1000$: find $Pr(\hat{P} > 0.32)$.

Q19. A rare condition affects $p = 0.02$ of a population. Explain why the normal approximation to the distribution of $\hat{P}$ is **not** reasonable for a sample of size $n = 100$.

Q20. A poll aims to estimate a proportion believed to be about $p = 0.5$. Find the smallest sample size n for which the standard deviation of $\hat{P}$ is at most 0.02.

Q21. In a large population, 25 % are left-handed ($p = 0.25$). A random sample of $n = 200$ is taken. (a) State the approximate distribution of $\hat{P}$. (b) Find $Pr(\hat{P} > 0.3)$. (c) Find the probability that more than 55 people in the sample are left-handed.

Answers

Q1. $\hat{P} \approx N(0.3, \sigma^2)$, $\sigma = \frac{\sqrt{21}}{100} \approx 0.0458$; (b) 0.8624; (c) 0.8624. **Q2.** $\mu = 0.7$, $\sigma = \frac{\sqrt{42}}{200} \approx 0.0324$; (b) 0.7830. **Q3.** reasonable ($np = 18$, $n(1-p) = 102$); $\sigma = \frac{\sqrt{170}}{400} \approx 0.0326$; 0.0625. **Q4.** $N(0.55, 0.0315^2)$, $\sigma = \frac{3\sqrt{110}}{1000}$; (b) 0.0560. **Q5.** $\mu = 0.8$, $\sigma = \frac{\sqrt{10}}{100} \approx 0.0316$; (b) 0.8862. **Q6.** $N(0.5, 0.05^2)$; (b) 0.0228; (c) 0.6 is 2 sd above the mean, and $1/2(1-0.95) = 0.025 \approx 0.0228$. **Q7.** $\sigma = \frac{\sqrt{65}}{200} \approx 0.0403$; $Pr(\hat{P} < 0.4) = 0.8926$. **Q8.** $\sigma = 0.025$; $Pr(\hat{P} > 0.28) = 0.1151$. **Q9.** $\sigma = \frac{\sqrt{15}}{75} \approx 0.0516$; $Pr(0.5 < \hat{P} < 0.7) = 0.9472$. **Q10.** $\sigma = \frac{3\sqrt{22}}{400} \approx 0.0352$; $Pr(\hat{P} > 0.5) = 0.0776$. **Q11.** $\sigma = 0.03$; $Pr(\hat{P} < 0.85) = 0.0478$. **Q12.** $\sigma = \frac{3\sqrt{10}}{500} \approx 0.0190$; $Pr(\hat{P} > 0.12) = 0.1459$. **Q13.** $\sigma = 0.0625$; $Pr(0.4 < \hat{P} < 0.6) = 0.8904$. **Q14.** $\sigma = 0.02$; $Pr(\hat{P} < 0.38) = 0.1587$. **Q15.** $\sigma = \frac{\sqrt{6}}{100} \approx 0.0245$; $Pr(0.68 < \hat{P} < 0.73) = 0.6826$. **Q16.** standard error $= \sqrt{\frac{0.6 \times 0.4}{400}} = \frac{\sqrt{6}}{100} \approx 0.0245$. **Q17.** $\hat{p} = 0.6$; standard error $= \sqrt{\frac{0.6 \times 0.4}{500}} = \frac{\sqrt{30}}{250} \approx 0.0219$. **Q18.** $\sigma = \frac{\sqrt{210}}{1000} \approx 0.0145$; $Pr(\hat{P} > 0.32) = 0.0838$. **Q19.** $np = 100 \times 0.02 = 2 < 10$, so the approximation is not reasonable. **Q20.** $n = 625$. **Q21.** $N(0.25, 0.0306^2)$, $\sigma = \frac{\sqrt{6}}{80}$; (b) 0.0512; (c) 0.2071.

Fully-worked solutions

Q1. $np = 30$, $n(1-p) = 70$ (both ≥ 10). Mean 0.3; $\sigma = \sqrt{\frac{0.3 \times 0.7}{100}} = \frac{\sqrt{21}}{100} \approx 0.0458$, so $\hat{P} \approx N(0.3, 0.0458^2)$. (b) $Pr(\hat{P} < 0.35) = 0.8624$. (c) $Pr(\hat{P} > 0.25) = 0.8624$ (equal to (b) by symmetry: 0.25 and 0.35 are each 0.05 from 0.3).

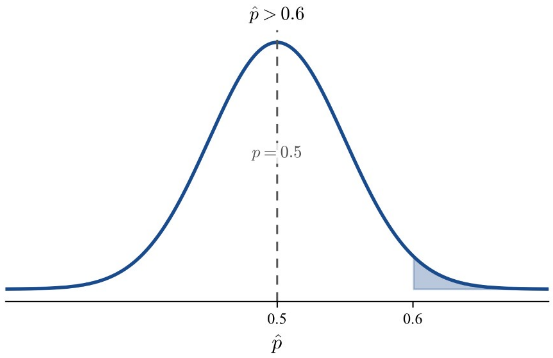
Q2. Mean 0.7; $\sigma = \sqrt{\frac{0.7 \times 0.3}{200}} = \frac{\sqrt{42}}{200} \approx 0.0324$. $Pr(0.66 < \hat{P} < 0.74) = 0.7830$ (each bound is ≈ 1.23 sd from 0.7).

Q3. (a) $np = 120 \times 0.15 = 18 \geq 10$ and $n(1-p) = 102 \geq 10$, so reasonable. (b) $\sigma = \sqrt{\frac{0.15 \times 0.85}{120}} = \frac{\sqrt{170}}{400} \approx 0.0326$. (c) $Pr(\hat{P} > 0.2) = 0.0625$.

Q4. Mean 0.55; $\sigma = \sqrt{\frac{0.55 \times 0.45}{250}} = \frac{3\sqrt{110}}{1000} \approx 0.0315$, so $\hat{P} \approx N(0.55, 0.0315^2)$. (b) $Pr(\hat{P} < 0.5) = 0.0560$.

Q5. Mean 0.8; $\sigma = \sqrt{\frac{0.8 \times 0.2}{160}} = \frac{\sqrt{10}}{100} \approx 0.0316$. (b) $Pr(0.75 < \hat{P} < 0.85) = 0.8862$.

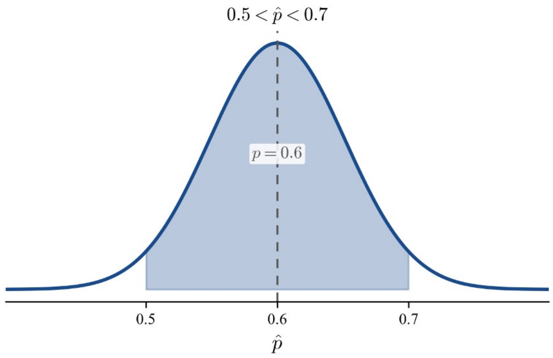
Q6. (a) $\sigma = \sqrt{\frac{0.5 \times 0.5}{100}} = 0.05$, so $\hat{P} \approx N(0.5, 0.05^2)$. (b) $Pr(\hat{P} > 0.6) = 0.0228$. (c) $0.6 = 0.5 + 2(0.05)$ is exactly two standard deviations above the mean; the 95% rule gives $Pr(\hat{P} > \mu + 2\sigma) \approx 1/2(1 - 0.95) = 0.025$, close to 0.0228.



Q7. $\sigma = \sqrt{\dfrac{0.35 \times 0.65}{140}} = \dfrac{\sqrt{65}}{200} \approx 0.0403$; $\hat{P} \approx N\left(0.35, 0.0403^2\right)$. $Pr\left(\hat{P} < 0.4\right) = 0.8926$.

Q8. $\sigma = \sqrt{\dfrac{0.25 \times 0.75}{300}} = 0.025$; $Pr\left(\hat{P} > 0.28\right) = 0.1151$.

Q9. $\sigma = \sqrt{\dfrac{0.6 \times 0.4}{90}} = \dfrac{\sqrt{15}}{75} \approx 0.0516$; $Pr\left(0.5 < \hat{P} < 0.7\right) = 0.9472$.

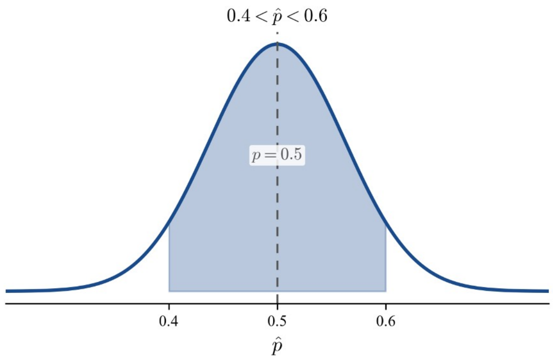


Q10. $\sigma = \sqrt{\dfrac{0.45 \times 0.55}{200}} = \dfrac{3\sqrt{22}}{400} \approx 0.0352$; $Pr\left(\hat{P} > 0.5\right) = 0.0776$.

Q11. $\sigma = \sqrt{\dfrac{0.9 \times 0.1}{100}} = 0.03$; $\hat{P} \approx N\left(0.9, 0.03^2\right)$. $Pr\left(\hat{P} < 0.85\right) = 0.0478$. (Here $n\left(1 - p\right) = 10$, at the edge of the guideline, so the approximation is only fair.)

Q12. $\sigma = \sqrt{\dfrac{0.1 \times 0.9}{250}} = \dfrac{3\sqrt{10}}{500} \approx 0.0190$; $Pr\left(\hat{P} > 0.12\right) = 0.1459$.

Q13. $\sigma = \sqrt{\dfrac{0.5 \times 0.5}{64}} = \dfrac{1}{16} = 0.0625$; $Pr\left(0.4 < \hat{P} < 0.6\right) = 0.8904$ (each bound is 1.6 sd from 0.5).



Q14. $\sigma = \sqrt{\dfrac{0.4 \times 0.6}{600}} = 0.02$; $Pr(\hat{P} < 0.38) = 0.1587$ (0.38 is one sd below 0.4, and $1/2(1 - 0.68) = 0.16$).

Q15. $\sigma = \sqrt{\dfrac{0.7 \times 0.3}{350}} = \dfrac{\sqrt{6}}{100} \approx 0.0245$; $Pr(0.68 < \hat{P} < 0.73) = 0.6826$.

Q16. Since p is unknown, use $\hat{p} = 0.6$: standard error $= \sqrt{\dfrac{0.6 \times 0.4}{400}} = \sqrt{0.0006} = \dfrac{\sqrt{6}}{100} \approx 0.0245$.

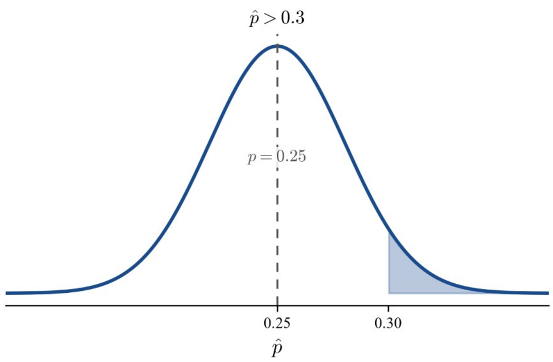
Q17. $\hat{p} = \dfrac{300}{500} = 0.6$; standard error $= \sqrt{\dfrac{0.6 \times 0.4}{500}} = \dfrac{\sqrt{30}}{250} \approx 0.0219$.

Q18. $\sigma = \sqrt{\dfrac{0.3 \times 0.7}{1000}} = \dfrac{\sqrt{210}}{1000} \approx 0.0145$; $Pr(\hat{P} > 0.32) = 0.0838$.

Q19. For the normal approximation we need (roughly) $np \geq 10$. Here $np = 100 \times 0.02 = 2$, which is far below 10 (only about 2 affected people are expected in the sample), so the distribution of $\hat{P}$ is strongly right-skewed and the symmetric normal curve is a poor model. $\therefore$ the approximation is not reasonable.

Q20. Require $\sigma_{\hat{P}} = \sqrt{\dfrac{0.5 \times 0.5}{n}} \leq 0.02$. Squaring: $\dfrac{0.25}{n} \leq 0.0004$, so $n \geq \dfrac{0.25}{0.0004} = 625$. $\therefore$ the smallest sample size is $n = 625$.

Q21. (a) $np = 50$, $n(1-p) = 150$ (both ≥ 10). $\sigma = \sqrt{\dfrac{0.25 \times 0.75}{200}} = \dfrac{\sqrt{6}}{80} \approx 0.0306$, so $\hat{P} \approx N(0.25, 0.0306^2)$. (b) $Pr(\hat{P} > 0.3) = 0.0512$. (c) “More than 55 left-handed” means $\hat{P} > \dfrac{55}{200} = 0.275$, and $Pr(\hat{P} > 0.275) = 0.2071$.



Theory

A sample proportion $\hat{p} = \frac{x}{n}$ (where x successes are observed in a sample of size n) is a **point estimate** of the unknown population proportion p. A single number gives no sense of how reliable the estimate is, so instead we construct an **interval estimate** — a range of plausible values for p.

The approximate confidence interval. For a large sample, $\hat{P}$ is approximately normally distributed with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. Since p is unknown we use $\hat{p}$ in the standard deviation, giving an approximate $C\,\%$ **confidence interval** for p:

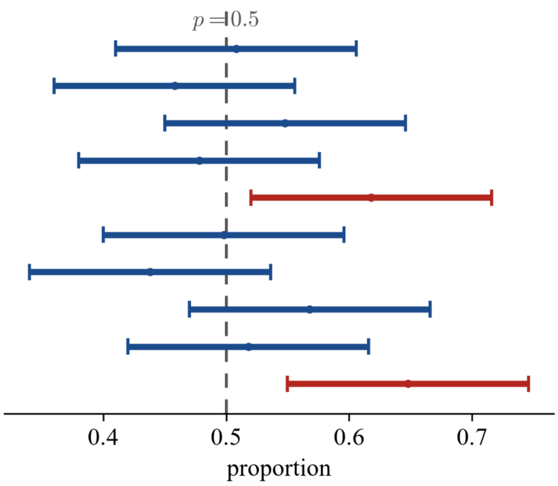
$$\left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$

where z is the value with $C\,\%$ of the standard-normal area between $-z$ and z:

Confidence level	z
90 %	1.645
95 %	1.96
99 %	2.576

Margin of error. The interval is $\hat{p} \pm M$, where the **margin of error** is $M = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. The interval is centred on $\hat{p}$ and has width $2M$.

Interpretation. We are $C\,\%$ *confident* that this interval contains p. More precisely: if many samples were taken and an interval computed from each, about $C\,\%$ of those intervals would contain the true value p. A particular interval either contains p or it does not — the confidence refers to the long-run success rate of the method, not to one interval.



Ten 95% confidence intervals from repeated samples: about 95% contain the true p (blue); the rest miss it (red).

What changes the width?

- A **larger sample size** n makes M smaller, so the interval is **narrower** (more precise).
- A **higher confidence level** uses a larger z, so the interval is **wider**.

Choosing a sample size. To guarantee a margin of error of at most M, we need $z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le M$, i.e.

$$n \ge \left(\frac{z}{M}\right)^2 \hat{p}(1-\hat{p}),$$

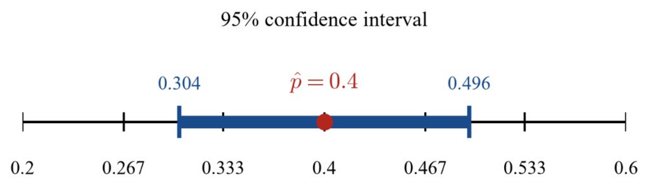
then **round up** to the next whole number. If no estimate of p is available, use $\hat{p}=0.5$ (which makes $\hat{p}(1-\hat{p})$ largest, giving the safest, largest n).

Worked examples

Example 1 — scaffolded

In a sample of 100 voters, 40 support a proposal. Find an approximate $95\,\%$ confidence interval for the population proportion p who support it.

Working	Comment
$\hat{p}=\frac{40}{100}=0.4$	The sample proportion is the point estimate.
$95\,\%$: $z=1.96$; $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=\sqrt{\frac{0.4\times 0.6}{100}}=0.048990$	Read z from the table; compute the standard deviation.
$M=1.96\times 0.048990=0.096$	The margin of error.
CI: $0.4\pm 0.096=(0.304,0.496)$	The interval is $\hat{p}\pm M$.
$\therefore$ we are $95\,\%$ confident that p lies between 0.304 and 0.496.	Interpret in context.



Example 2 — scaffolded

A random sample of 400 items contains 100 defectives. Find an approximate $90\,\%$ confidence interval for the proportion of defectives in the population.

Working	Comment
$\hat{p}=\frac{100}{400}=0.25$	The sample proportion.
$90\,\%$: $z=1.645$; $\sqrt{\frac{0.25\times 0.75}{400}}=0.021651$	For $90\,\%$ confidence use $z=1.645$.
$M=1.645\times 0.021651=0.0356$	Margin of error (to 4 decimal places).
CI: $0.25\pm 0.0356=(0.2144,0.2856)$	$\hat{p}\pm M$.

$\therefore$ an approximate $90\,\%$ confidence interval is $(0.2144,0.2856)$.

Example 3 — unscaffolded

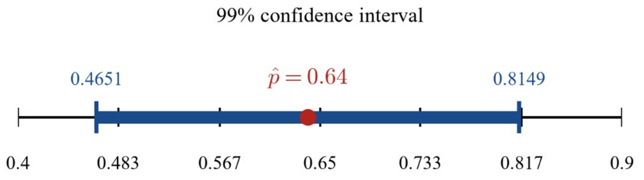
In a sample of 50 people, 32 own a car. Find an approximate $99\,\%$ confidence interval for the population proportion p.

$\hat{p}=\dfrac{32}{50}=0.64$. For 99% confidence, $z=2.576$.

$M=2.576\sqrt{\dfrac{0.64\times0.36}{50}}=2.576\times0.067882=0.1749$.

CI: $0.64\pm0.1749=(0.4651,0.8149)$.

$\therefore$ we are 99% confident that p lies between 0.4651 and 0.8149. (The interval is wide because the sample is small and the confidence level is high.)



Example 4 — unscaffolded

A pollster wants to estimate a population proportion to within a margin of error of 0.04, with 95% confidence. No prior estimate of p is available. Find the smallest sample size required.

With no estimate of p, use $\hat{p}=0.5$ (the most conservative choice). For 95%, $z=1.96$.

Require $M\le0.04$: $n\ge\left(\dfrac{z}{M}\right)^2\hat{p}(1-\hat{p})=\left(\dfrac{1.96}{0.04}\right)^2(0.5)(0.5)=49^2\times0.25=600.25$.

$\therefore$ round up: the smallest sample size is $n=601$.

Practice questions

Use $z=1.645$ for 90%, $z=1.96$ for 95%, $z=2.576$ for 99% confidence. Give margins and interval endpoints correct to four decimal places.

Scaffolded

Q1. In a sample of 200, there are 120 successes. (a) Find $\hat{p}$. (b) Find the margin of error for a 95% confidence interval. (c) Write down the 95% confidence interval.

Q2. A sample of 250 contains 90 successes. (a) Find $\hat{p}$; (b) find the 95% margin of error; (c) state the 95% confidence interval.

Q3. In a sample of 500, 300 have a certain feature. (a) Find $\hat{p}$; (b) find a 90% confidence interval for p; (c) is this interval narrower or wider than the 95% interval from the same data? Explain.

Q4. A sample of 80 contains 20 successes. (a) Find $\hat{p}$; (b) find a 95% confidence interval for p; (c) comment on the width of the interval, given the sample size.

Q5. In a sample of 1000, 550 agree with a statement. (a) Find $\hat{p}$; (b) find a 99% confidence interval for p.

Q6. A proportion is to be estimated to within a margin of 0.03 with 95% confidence, with no prior estimate of p. (a) Which value of $\hat{p}$ should be used, and why? (b) Find the smallest required sample size.

Unscaffolded

Find the required confidence interval (or sample size), and interpret where asked.

Q7. $n=150$, $x=96$; 95% confidence interval.

Q8. $n = 300$, $x = 210$; 90% confidence interval.

Q9. $n = 64$, $x = 48$; 95% confidence interval.

Q10. $n = 400$, $x = 160$; 99% confidence interval.

Q11. $n = 225$, $x = 135$; 95% confidence interval, and interpret it in context.

Q12. $n = 500$, $x = 125$; 90% confidence interval.

Q13. $n = 1200$, $x = 900$; 95% confidence interval. Comment on why the interval is narrow.

Q14. $n = 40$, $x = 26$; 95% confidence interval.

Q15. $n = 600$, $x = 372$; 99% confidence interval.

Q16. A proportion known to be about 0.6 is to be estimated to within 0.05 with 95% confidence. Find the smallest required sample size.

Q17. $n = 900$, $x = 315$; 95% confidence interval.

Q18. $n = 160$, $x = 100$; 90% confidence interval. Interpret the interval in context.

Q19. $n = 2000$, $x = 1100$; 95% confidence interval. Explain why it is narrower than the interval in Q11.

Q20. A proportion is to be estimated to within 0.02 with 95% confidence, with no prior estimate. Find the smallest required sample size.

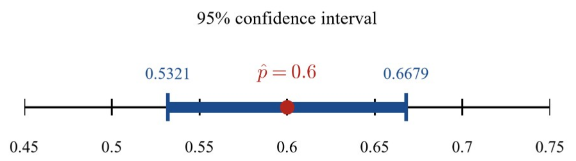
Q21. $n = 350$, $x = 224$; 99% confidence interval, and state what happens to the interval if a 95% level were used instead.

Answers

Q1. $\hat{p}=0.6$; $M=0.0679$; $(0.5321, 0.6679)$. **Q2.** $\hat{p}=0.36$; $M=0.0595$; $(0.3005, 0.4195)$. **Q3.** $\hat{p}=0.6$; $(0.564, 0.636)$; narrower than the 95% interval (smaller z). **Q4.** $\hat{p}=0.25$; $(0.1551, 0.3449)$; wide, because n is small. **Q5.** $\hat{p}=0.55$; $(0.5095, 0.5905)$. **Q6.** $\hat{p}=0.5$ (maximises $\hat{p}(1-\hat{p})$); $n=1068$. **Q7.** $(0.5632, 0.7168)$. **Q8.** $(0.6565, 0.7435)$. **Q9.** $(0.6439, 0.8561)$. **Q10.** $(0.3369, 0.4631)$. **Q11.** $(0.536, 0.664)$; 95% confident p is between 0.536 and 0.664. **Q12.** $(0.2181, 0.2819)$. **Q13.** $(0.7255, 0.7745)$; narrow because n is large. **Q14.** $(0.5022, 0.7978)$. **Q15.** $(0.569, 0.671)$. **Q16.** $n=369$. **Q17.** $(0.3188, 0.3812)$. **Q18.** $(0.562, 0.688)$; 90% confident p is between 0.562 and 0.688. **Q19.** $(0.5282, 0.5718)$; narrower than Q11 because n is much larger. **Q20.** $n=2401$. **Q21.** $(0.5739, 0.7061)$; a 95% interval would be narrower (smaller z).

Fully-worked solutions

Q1. $\hat{p}=\dfrac{120}{200}=0.6$. $M=1.96\sqrt{\dfrac{0.6\times0.4}{200}}=1.96\times0.034641=0.0679$. CI: $0.6\pm0.0679=(0.5321, 0.6679)$.



Q2. $\hat{p}=\dfrac{90}{250}=0.36$. $M=1.96\sqrt{\dfrac{0.36\times0.64}{250}}=1.96\times0.030358=0.0595$. CI: $(0.3005, 0.4195)$.

Q3. $\hat{p}=\dfrac{300}{500}=0.6$. $M=1.645\sqrt{\dfrac{0.6\times0.4}{500}}=1.645\times0.021909=0.036$. CI: $(0.564, 0.636)$. This is **narrower** than a 95% interval from the same data, because 90% uses the smaller value $z=1.645$ (vs 1.96).

Q4. $\hat{p}=\dfrac{20}{80}=0.25$. $M=1.96\sqrt{\dfrac{0.25\times0.75}{80}}=1.96\times0.048412=0.0949$. CI: $(0.1551, 0.3449)$. The interval is **wide** because $n=80$ is small, so the standard error is large.

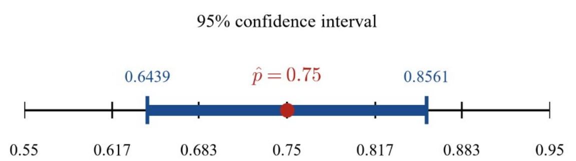
Q5. $\hat{p}=\dfrac{550}{1000}=0.55$. $M=2.576\sqrt{\dfrac{0.55\times0.45}{1000}}=2.576\times0.015732=0.0405$. CI: $(0.5095, 0.5905)$.

Q6. (a) Use $\hat{p}=0.5$, since $\hat{p}(1-\hat{p})$ is greatest there, giving the largest (safest) n. (b) $n\geq\left(\dfrac{1.96}{0.03}\right)^2(0.5)(0.5)=1067.11$, so $n=1068$.

Q7. $\hat{p}=\dfrac{96}{150}=0.64$. $M=1.96\sqrt{\dfrac{0.64\times0.36}{150}}=1.96\times0.039192=0.0768$. CI: $(0.5632, 0.7168)$.

Q8. $\hat{p}=\dfrac{210}{300}=0.7$. $M=1.645\sqrt{\dfrac{0.7\times0.3}{300}}=1.645\times0.026458=0.0435$. CI: $(0.6565, 0.7435)$.

Q9. $\hat{p}=\dfrac{48}{64}=0.75$. $M=1.96\sqrt{\dfrac{0.75\times0.25}{64}}=1.96\times0.054127=0.1061$. CI: $(0.6439, 0.8561)$.



Q10. $\hat{p}=\dfrac{160}{400}=0.4$. $M=2.576\sqrt{\dfrac{0.4\times 0.6}{400}}=2.576\times 0.024495=0.0631$. CI: $\left(0.3369,0.4631\right)$.

Q11. $\hat{p}=\dfrac{135}{225}=0.6$. $M=1.96\sqrt{\dfrac{0.6\times 0.4}{225}}=1.96\times 0.032660=0.064$. CI: $\left(0.536,0.664\right)$. We are 95% confident that the population proportion p lies between 0.536 and 0.664.

Q12. $\hat{p}=\dfrac{125}{500}=0.25$. $M=1.645\sqrt{\dfrac{0.25\times 0.75}{500}}=1.645\times 0.019365=0.0319$. CI: $\left(0.2181,0.2819\right)$.

Q13. $\hat{p}=\dfrac{900}{1200}=0.75$. $M=1.96\sqrt{\dfrac{0.75\times 0.25}{1200}}=1.96\times 0.012500=0.0245$. CI: $\left(0.7255,0.7745\right)$. The interval is narrow because the large sample size $n=1200$ makes the standard error small.

Q14. $\hat{p}=\dfrac{26}{40}=0.65$. $M=1.96\sqrt{\dfrac{0.65\times 0.35}{40}}=1.96\times 0.075415=0.1478$. CI: $\left(0.5022,0.7978\right)$.

Q15. $\hat{p}=\dfrac{372}{600}=0.62$. $M=2.576\sqrt{\dfrac{0.62\times 0.38}{600}}=2.576\times 0.019818=0.051$. CI: $\left(0.569,0.671\right)$.

Q16. $n\geq\left(\dfrac{1.96}{0.05}\right)^2\left(0.6\right)\left(0.4\right)=368.79$, so $n=369$.

Q17. $\hat{p}=\dfrac{315}{900}=0.35$. $M=1.96\sqrt{\dfrac{0.35\times 0.65}{900}}=1.96\times 0.015899=0.0312$. CI: $\left(0.3188,0.3812\right)$.

Q18. $\hat{p}=\dfrac{100}{160}=0.625$. $M=1.645\sqrt{\dfrac{0.625\times 0.375}{160}}=1.645\times 0.038273=0.063$. CI: $\left(0.562,0.688\right)$. We are 90% confident the proportion in the population lies between 0.562 and 0.688.

Q19. $\hat{p}=\dfrac{1100}{2000}=0.55$. $M=1.96\sqrt{\dfrac{0.55\times 0.45}{2000}}=1.96\times 0.011124=0.0218$. CI: $\left(0.5282,0.5718\right)$. This is much narrower than Q11's interval because $n=2000$ is far larger than 225, so the standard error (and hence the margin of error) is much smaller.

Q20. $n\geq\left(\dfrac{1.96}{0.02}\right)^2\left(0.5\right)\left(0.5\right)=98^2\times 0.25=2401$. Since this is exact, $n=2401$.

Q21. $\hat{p}=\dfrac{224}{350}=0.64$. $M=2.576\sqrt{\dfrac{0.64\times 0.36}{350}}=2.576\times 0.025652=0.0661$. CI: $\left(0.5739,0.7061\right)$. If a 95% level were used, z would drop from 2.576 to 1.96, so the interval would be **narrower**.

Topic 3 Test A — Technology-active

One bound reference and an approved CAS calculator are permitted. Reading time 5 minutes; writing time 50 minutes. Total: 50 marks. In all questions where a numerical answer is required, an exact value must be given unless otherwise specified; where more than one mark is available, appropriate working must be shown.

Question 1 (7 marks)

The discrete random variable X has the probability distribution shown.

x	0	1	2	3
$Pr(X=x)$	0.1	0.3	k	0.2

a. Find the value of k. 1 mark

b. Find $Pr(X \geq 2)$. 1 mark

c. Find $\mathrm{E}(X)$. 2 marks

d. Find $\mathrm{Var}(X)$ and the standard deviation $\mathrm{sd}(X)$. 3 marks

Question 2 (8 marks)

A biased coin shows heads with probability 0.4. It is tossed 10 times; let X be the number of heads.

a. Write down the distribution of X. 1 mark

b. Find $Pr(X=4)$, correct to four decimal places. 2 marks

c. Find $Pr(X \geq 4)$, correct to four decimal places. 2 marks

d. Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. 2 marks

e. The coin is instead tossed n times. Find the smallest value of n for which the probability of obtaining at least one head exceeds 0.99. 1 mark

Question 3 (9 marks)

The continuous random variable X has probability density function

$$f(x)=\begin{cases}\frac{3}{8}x^2, & 0 \leq x \leq 2\\ 0, & \text{elsewhere.}\end{cases}$$

a. Verify that f is a valid probability density function. 2 marks

b. Find $Pr(X \leq 1)$. 2 marks

c. Find the mean $\mathrm{E}(X)$. 2 marks

d. Find the exact value of the median m. 3 marks

Question 4 (10 marks)

The heights (in cm) of a large population are normally distributed with mean $\mu=170$ and standard deviation $\sigma=8$. Let X be the height of a randomly chosen individual.

a. Find $Pr(X<180)$, correct to four decimal places. 2 marks

b. Find $Pr(160<X<180)$, correct to four decimal places. 2 marks

c. Find, correct to one decimal place, the height that is exceeded by only 5% of the population. 3 marks

d. Twenty people are chosen at random. Find, correct to four decimal places, the probability that at least 3 of them are taller than 180 cm. 3 marks

Question 5 (8 marks)

In a random sample of 200 voters, 90 support a proposal. Let $\hat{p}$ be the sample proportion.

a. Find $\hat{p}$. 1 mark

b. Determine an approximate 95% confidence interval for the population proportion p, giving the endpoints correct to three decimal places. 3 marks

c. State the margin of error for this interval, correct to three decimal places. 1 mark

d. A second survey is to be conducted so that the margin of error is half that of part (c), keeping the same sample proportion and confidence level. How large a sample is required? 3 marks

Question 6 (8 marks)

The waiting time T minutes at a service counter has probability density function $f(t)=\frac{1}{4}e^{-t/4}$ for $t\geq 0$, and $f(t)=0$ otherwise.

a. Find $Pr(T>4)$, correct to four decimal places. 2 marks

b. Find the exact value of the median waiting time. 2 marks

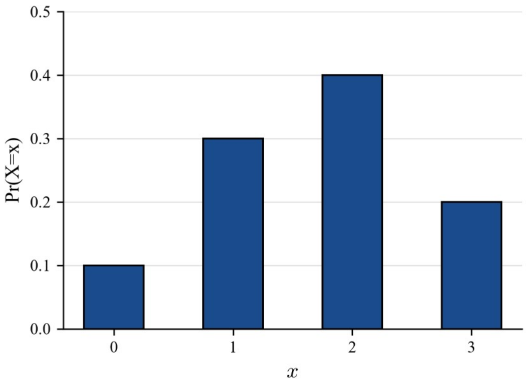
c. Find $Pr(T>6 / T>2)$, correct to four decimal places. 2 marks

d. Find the mean waiting time $\mathrm{E}(T)$. 2 marks

End of Test A

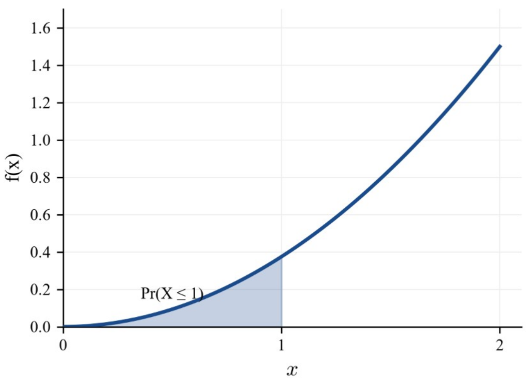
Test A — solutions

Q1. (a) $k=1-(0.1+0.3+0.2)=0.4$. (b) $Pr(X\geq 2)=0.4+0.2=0.6$. (c) $\mathrm{E}(X)=0(0.1)+1(0.3)+2(0.4)+3(0.2)=1.7$. (d) $\mathrm{E}(X^2)=0+0.3+4(0.4)+9(0.2)=3.7$, so $\mathrm{Var}(X)=3.7-1.7^2=0.81$ and $\mathrm{sd}(X)=\sqrt{0.81}=0.9$.

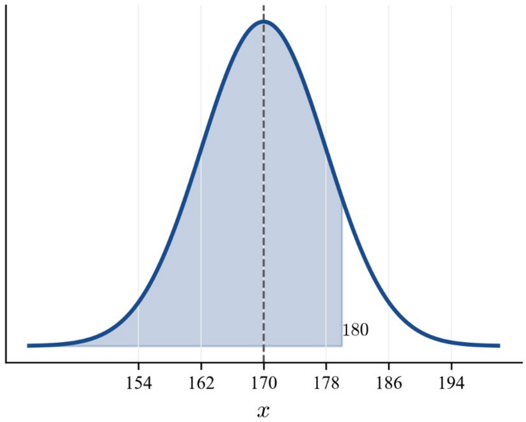


Q2. (a) $X\sim \mathrm{Bi}(10,0.4)$. (b) $Pr(X=4)=\binom{10}{4}(0.4)^4(0.6)^6=0.2508$. (c) $Pr(X\geq 4)=1-Pr(X\leq 3)=0.6177$. (d) $\mathrm{E}(X)=np=10(0.4)=4$; $\mathrm{Var}(X)=np(1-p)=10(0.4)(0.6)=2.4$. (e) $Pr(\text{at least one head})=1-0.6^n>0.99\Rightarrow 0.6^n<0.01\Rightarrow n>\dfrac{\log_e 0.01}{\log_e 0.6}\approx 9.02$, so the smallest n is 10.

Q3. (a) $f(x)\geq 0$ on $[0,2]$, and $\int_0^2 \frac{3}{8}x^2\,dx=\frac{3}{8}\left[\frac{x^3}{3}\right]_0^2=\frac{3}{8}\cdot\frac{8}{3}=1$, so f is a valid pdf. (b) $Pr(X\leq 1)=\int_0^1 \frac{3}{8}x^2\,dx=\frac{3}{8}\cdot\frac{1}{3}=\frac{1}{8}$. (c) $\mathrm{E}(X)=\int_0^2 x\cdot\frac{3}{8}x^2\,dx=\frac{3}{8}\left[\frac{x^4}{4}\right]_0^2=\frac{3}{8}\cdot 4=\frac{3}{2}$. (d) $\int_0^m \frac{3}{8}x^2\,dx=\frac{1}{2}\Rightarrow\frac{m^3}{8}=\frac{1}{2}\Rightarrow m^3=4$, so $m=\sqrt[3]{4}\approx 1.587$.

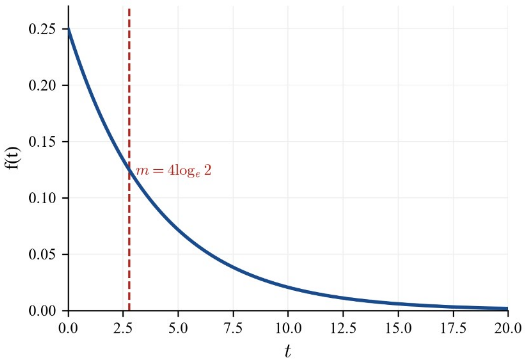


Q4. (a) $Pr(X<180)=0.8944$ (standardising, $z=180-170/8=1.25$). (b) $Pr(160<X<180)=0.7887$. (c) Require a with $Pr(X>a)=0.05$, i.e. $Pr(X<a)=0.95$; the inverse normal gives $a\approx 183.2$ cm. (d) $p=Pr(X>180)=1-0.8944=0.1057$ (using 0.105650). Let Y be the number taller than 180; $Y\sim\mathrm{Bi}(20,0.1057)$, so $Pr(Y\geq 3)=1-Pr(Y\leq 2)=0.3551$.



Q5. (a) $\hat{p}=\frac{90}{200}=0.45$. (b) $\hat{p}\pm1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=0.45\pm1.96\sqrt{\frac{0.45\times0.55}{200}}=0.45\pm0.069$, i.e. $(0.381, 0.519)$. (c) Margin of error $=1.96\sqrt{\frac{0.45\times0.55}{200}}\approx0.069$. (d) The margin of error is proportional to $\frac{1}{\sqrt{n}}$, so halving it requires 4 times the sample: $n=4\times200=800$.

Q6. (a) $Pr(T>4)=\int_4^\infty \frac{1}{4}e^{-t/4}\,dt=[-e^{-t/4}]_4^\infty=e^{-1}=0.3679$. (b) $\int_0^m \frac{1}{4}e^{-t/4}\,dt=\frac{1}{2}\Rightarrow 1-e^{-m/4}=\frac{1}{2}\Rightarrow e^{-m/4}=\frac{1}{2}\Rightarrow m=4\log_e 2$. (c) $Pr(T>6/T>2)=\frac{Pr(T>6)}{Pr(T>2)}=\frac{e^{-3/2}}{e^{-1/2}}=e^{-1}=0.3679$ (the exponential is memoryless). (d) $\mathrm{E}(T)=\int_0^\infty t\cdot\frac{1}{4}e^{-t/4}\,dt=4$.



Topic 3 Test B — Technology-free

No calculator and no notes. 45 minutes. Total: 40 marks. All numerical answers must be exact; where more than one mark is available, appropriate working must be shown.

Question 1 (7 marks)

The discrete random variable X has the probability distribution shown.

x	1	2	3	4
$Pr(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

a. Show that the probabilities sum to 1. 1 mark

b. Find $\mathrm{E}(X)$. 2 marks

c. Find $\mathrm{Var}(X)$. 3 marks

d. Find $\mathrm{E}(2X-1)$. 1 mark

Question 2 (7 marks)

Let $X \sim \mathrm{Bi}\left(4, \frac{1}{3}\right)$.

a. Find $Pr(X=2)$ as an exact fraction. 2 marks

b. Find $Pr(X \geq 3)$. 2 marks

c. Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. 2 marks

d. State the most likely value of X. 1 mark

Question 3 (9 marks)

The continuous random variable X has probability density function

$$f(x)=\begin{cases} k(2-x), & 0 \leq x \leq 2 \\ 0, & \text{elsewhere.} \end{cases}$$

a. Show that $k=\frac{1}{2}$. 2 marks

b. Find $Pr(X \leq 1)$. 2 marks

c. Find $\mathrm{E}(X)$. 3 marks

d. Find the exact value of the median m. 2 marks

Question 4 (9 marks)

A quantity X is normally distributed with mean $\mu=60$ and standard deviation $\sigma=5$. Use the 68–95–99.7 % rule.

a. Find the standardised value (z-score) corresponding to $X=70$. 1 mark

b. Find $Pr(55<X<65)$. 2 marks

c. Find $Pr(X>70)$. 2 marks

d. Find $Pr(50<X<65)$. 2 marks

e. Given that $Pr(X < a) = 0.16$, find the value of a. 2 marks

Question 5 (5 marks)

In a large population, the proportion with a certain characteristic is $p = \frac{1}{4}$. A random sample of size $n = 48$ is taken; let $\hat{p} = \frac{X}{n}$ be the sample proportion, where X is the number in the sample with the characteristic.

a. State $\mathrm{E}(\hat{p})$. 1 mark

b. Find the exact value of $\mathrm{sd}(\hat{p})$. 2 marks

c. State $\mathrm{E}(X)$ and $\mathrm{sd}(X)$. 2 marks

Question 6 (3 marks)

A random variable X has a binomial distribution with $\mathrm{E}(X) = 6$ and $\mathrm{Var}(X) = 4.2$. Find the values of n and p.

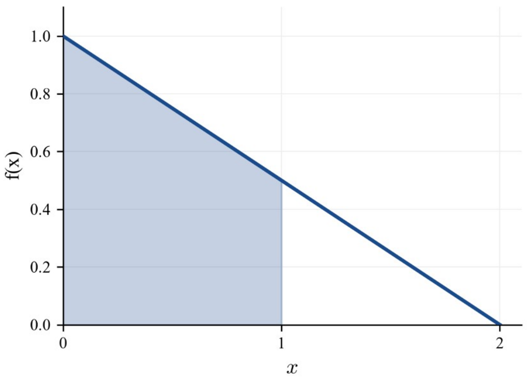
End of Test B

Test B — solutions

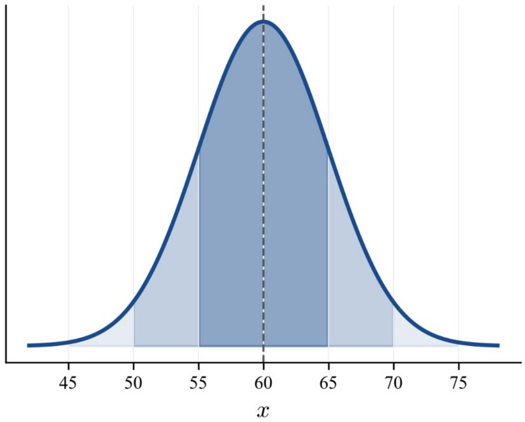
Q1. (a) $\frac{1}{8}+\frac{3}{8}+\frac{3}{8}+\frac{1}{8}=\frac{8}{8}=1$. (b) $\mathrm{E}(X)=\frac{1}{8}(1)+\frac{3}{8}(2)+\frac{3}{8}(3)+\frac{1}{8}(4)=\frac{20}{8}=\frac{5}{2}$. (c) $\mathrm{E}(X^2)=\frac{1}{8}(1)+\frac{3}{8}(4)+\frac{3}{8}(9)+\frac{1}{8}(16)=\frac{56}{8}=7$, so $\mathrm{Var}(X)=7-\left(\frac{5}{2}\right)^2=7-\frac{25}{4}=\frac{3}{4}$. (d) $\mathrm{E}(2X-1)=2\mathrm{E}(X)-1=2\cdot\frac{5}{2}-1=4$.

Q2. (a) $Pr(X=2)=\binom{4}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^2=6\cdot\frac{1}{9}\cdot\frac{4}{9}=\frac{8}{27}$. (b) $Pr(X\geq 3)=\binom{4}{3}\left(\frac{1}{3}\right)^3\frac{2}{3}+\left(\frac{1}{3}\right)^4=\frac{8}{81}+\frac{1}{81}=\frac{9}{81}=\frac{1}{9}$. (c) $\mathrm{E}(X)=np=\frac{4}{3}$; $\mathrm{Var}(X)=np(1-p)=4\cdot\frac{1}{3}\cdot\frac{2}{3}=\frac{8}{9}$. (d) $Pr(X=0)=\frac{16}{81}$, $Pr(X=1)=\frac{32}{81}$, $Pr(X=2)=\frac{24}{81}$; the greatest is at $X=1$, so the most likely value is 1.

Q3. (a) $\int_0^2 k(2-x)\,dx=k\left[2x-\frac{x^2}{2}\right]_0^2=k(4-2)=2k$. Setting $2k=1$ gives $k=\frac{1}{2}$. (b) $Pr(X\leq 1)=\int_0^1 \frac{1}{2}(2-x)\,dx=\frac{1}{2}\left[2x-\frac{x^2}{2}\right]_0^1=\frac{1}{2}\cdot\frac{3}{2}=\frac{3}{4}$. (c) $\mathrm{E}(X)=\int_0^2 x\cdot\frac{1}{2}(2-x)\,dx=\frac{1}{2}\int_0^2(2x-x^2)\,dx=\frac{1}{2}\left[x^2-\frac{x^3}{3}\right]_0^2=\frac{1}{2}\cdot\frac{4}{3}=\frac{2}{3}$. (d) $\int_0^m \frac{1}{2}(2-x)\,dx=\frac{1}{2}\Rightarrow 2m-\frac{m^2}{2}=1\Rightarrow m^2-4m+2=0\Rightarrow m=2-\sqrt{2}$ (the root in $[0,2]$).



Q4. (a) $z=\frac{70-60}{5}=2$. (b) 55 and 65 are one standard deviation either side of the mean, so $Pr(55<X<65)=0.68$. (c) $70=\mu+2\sigma$, so $Pr(X>70)=\frac{1-0.95}{2}=0.025$. (d) $50=\mu-2\sigma$ and $65=\mu+\sigma$, so $Pr(50<X<65)=\frac{0.95}{2}+\frac{0.68}{2}=0.475+0.34=0.815$. (e) $Pr(X<\mu-\sigma)=\frac{1-0.68}{2}=0.16$, so $a=\mu-\sigma=55$.



Q5. (a) $\mathrm{E}(\hat{p}) = p = \frac{1}{4}$. (b) $\mathrm{sd}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{1/4 \cdot 3/4}{48}} = \sqrt{\frac{3}{768}} = \sqrt{\frac{1}{256}} = \frac{1}{16}$. (c) $\mathrm{E}(X) = np = 48 \cdot \frac{1}{4} = 12$; $\mathrm{sd}(X) = \sqrt{np(1-p)} = \sqrt{48 \cdot 1/4 \cdot 3/4} = \sqrt{9} = 3$.

Q6. $\mathrm{E}(X) = np = 6$ and $\mathrm{Var}(X) = np(1-p) = 4.2$. Dividing, $1 - p = \frac{4.2}{6} = 0.7$, so $p = 0.3$ and $n = \frac{6}{0.3} = 20$.