Unit 5 · Data, Evidence and Decisions

WORKSHEET

Full marks 31

Name _____ Class _____ Roll no. ____ Date _____ Marks _____ / 31

1 Key words [4]

Fill in the blanks with words from the box. Two words in the box are not needed. Each blank carries 0.5 mark.

biased	**bootstrap**	**confounder**	**histogram**	**median**	**null hypothesis**	**overfitting**
p-hacking	**population**	**residual**				

The whole group you want to know about is the _____. A sample whose way of being chosen makes its estimates miss the truth in the same direction, again and again, is _____. Taking many new samples, with replacement, from the sample itself to see how much an estimate could vary is the _____. A data point's real value minus the line's prediction is its _____. A model that fits its own data closely, chance wobbles and all, but predicts new data badly has been _____. A factor that affects both things being compared, and so can make them go together, is a _____. The idea that there is no real difference and any difference seen is chance is the _____. Testing many things and reporting only what crossed the line is _____.

2 Multiple choice [5]

Choose the correct option for each of the following questions.

i. A news website asks its readers, "Should the school year be longer?" and 15,000 people click Yes or No. Why can't the result tell you what Nepal's people think? [1]

a. 15,000 is too few for a country of millions.

b. The people who answered chose themselves: readers of that site, with internet, who cared enough to click.

c. Online answers are always false.

d. Polls can only ask questions about elections.

ii. A survey of 100 households chosen at random gives an estimate that is typically about 5 percentage points from the truth. About how many households would bring that to about 2.5 points? [1]

a. 150 **b.** 200

c. 400 **d.** 10,000

iii. A line fitted to the prices of second-hand motorbikes is: price (Rs.) = 180,000 − 12,000 × age (years). What does the slope say, in words? [1]

a. A motorbike costs Rs. 12,000 a year to run.

b. On average, each extra year of age goes with a price about Rs. 12,000 lower.

c. Every motorbike loses exactly Rs. 12,000 each year.

d. A new motorbike costs Rs. 12,000.

iv. In Nepal, districts with more mobile towers have more road accidents. What is the most likely explanation? **[1]**

a. Mobile signals distract drivers.

b. Accidents make companies build towers.

c. A third factor, more people and more traffic, brings both more towers and more accidents.

d. It must be chance.

v. A news report says: "People who drink green tea every day live longer, a study of 20,000 adults finds." What is the best first question to ask? **[1]**

a. Was the tea grown in Nepal?

b. Were the green tea drinkers different in other ways, such as being richer or healthier already?

c. Is 20,000 enough people?

d. Did the study use a computer?

3 Fit a line by hand **[5]**

Fill in the table, then answer.

Five students said how many hours a week they study at home (x), and their marks in a test (y). The mean of x is 6 hours and the mean of y is 63 marks.

Hours x	Marks y	x − 6	y − 63	(x − 6)(y − 63)	$(x - 6)^2$
2	50				
4	58				
6	61				
8	70				
10	76				
Total					

i. Work out the slope and the intercept of the least-squares line: slope = total of (x − 6)(y − 63) ÷ total of $(x - 6)^2$, and intercept = 63 − slope × 6.

ii. What mark does the line predict for 7 hours a week?

iii. Say what the slope means, in words. Does the line show that studying an extra hour raises a student's marks?

4 Reading a fitted line [3]

Answer on the lines.

The least-squares line for 15 of Nepal's weather stations, from 72 m to 2,363 m high, is temperature (°C) = 25.45 − 0.00487 × height (m).

i. Pokhara is at 827 m and averages 21.5 °C. What does the line predict for it, and what is Pokhara's residual?

ii. What does the line say for a place at 6,000 m? Why should this number not be trusted?

5 How sure is the estimate? [3]

Answer on the lines.

A ward office asked 40 households, chosen at random, what they spend on mobile data in a month. The average was Rs. 612, and the bootstrap's middle 95% of resampled averages was Rs. 518 to Rs. 710.

i. Write the result as an honest one-sentence report.

ii. About how many households would make the interval about half as wide?

iii. A volunteer suggests asking 4,000 people on the ward's Facebook page instead. Would that be better? Why?

6 What does it print? [4]

Write exactly what the program prints, in the box under it. Then answer.

A hospital brought in a new booking system and compared the waits in its two shifts before and after.

```
# each shift: (patients, average wait in minutes)
before = {"morning": (40, 30), "afternoon": (60, 10)}
after = {"morning": (80, 25), "afternoon": (20, 8)}

for name, shifts in [("before", before), ("after", after)]:
    patients = sum(n for n, wait in shifts.values())
    minutes = sum(n * wait for n, wait in shifts.values())
    print(name, patients, minutes / patients)
```

OUTPUT

i. In both shifts, did the waits go up or down? Why did the average wait over the whole day go the other way?

7 Find and fix the bug [3]

Read the program, then answer.

The program should find a 95% interval for the average by the bootstrap. It prints `Average: 585`, then `The middle 95%: 585 to 585`.

```
import random
import statistics

spend = [200, 400, 800, 600, 400, 250, 550, 450, 1550, 650]

random.seed(1)
means = []
for i in range(1000):
    again = random.sample(spend, k=len(spend))
    means.append(statistics.mean(again))

means.sort()
print("Average:", statistics.mean(spend))
print("The middle 95%:", means[25], "to", means[974])
```

i. Why is every resampled average 585?

ii. Write the corrected line.

8 Could chance do it? [2]

Answer on the lines.

An A/B test of two sign-up buttons, each shown to 400 people, finds 6 more sign-ups with the new button. A shuffle test gives a p-value of about 60%. Riya says: "So there is a 60% chance that the two buttons are just as good." Correct her, and say what the test does show.

9 Think [2]

Think, then answer.

A school wants to know whether a new mathematics app raises marks. The head teacher suggests comparing the students who choose to use it with those who don't. Explain why that could mislead, and describe a fairer test.

10 Activity: A survey of one ward (not marked)

At a computer or on a phone, open the unit's page (the QR code) and find **A survey of one ward.** Take 20 samples of each kind and size, and write the lowest and highest estimate you get. The true share with internet is 40%.

Sample	Lowest estimate	Highest estimate
10 households at random		
40 households at random		
160 households at random		
40 houses along the road		
160 houses along the road		