Applied Probability II

Section 3: Getting to grips with uncertainty

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 3: Getting to grips with uncertainty

Section 3.1: Probability versus Statistics

Uncertainty

If there is one certainty in life - it is uncertainty!

Think about a situation in your own life recently, where an outcome was uncertain. How did you try to make sense of it?

Some examples:

- car breaks down and waiting on road-side assistance.
- waiting on a bus or DART to arrive
- planning a cycle, will it be windy?
- choosing a treatment for a disease that has side-effects

How do we make sense of uncertainty?

We can try to quantify it, but how do we do that formally?

Probability

In Applied Probability I, you learned about distributions and probability rules.

If you make an assumption about the distribution of some outcome, you can then assign probabilities to the possible outcomes.

What else can you do?

Collect data!

Collecting data

Let's go back to one of the examples we looked at earlier:

Car has broken down and you call roadside assistance to come and need to know if they will arrive in the timeframe they said they would. This happened to me recently!

- Observation 1. How long did they take to arrive and was it in the timeframe they said?
- Observation 2. Again, how long did they take to arrive and was it in the timeframe they said?
- What will happen the third time? How high are the stakes?

Probability versus Statistics

- PROBABILITY: We start with a probability model, i.e., we consider outcomes that could occur and assign likelihood of those outcomes occurring. Outcomes are random, therefore we don't know what will happen so we 'quantify our degree of surprise'.
- STATISTICS: We assume that data at hand were generated by some probability model and that our task is to approximate what that model was.

Probability	**Statistics**
Predicting probabilities of events	Modelling data to understand events.
Rules -> data	Data -> rules

"All models are wrong, but some are useful", George Box

What are the stakes?

How high stakes are the various outcomes?

There may be times when quantifying uncertainty is more important than others.

For many of the uncertain phenomena that surround us, quantifying uncertainty is a necessity. It can, for example:

- Protect us (healcare, public safety, pandemic)
- Make money from us: business (decision making)

Decision making.

"If it was an easily solvable problem, or even a modestly difficult but solvable problem, it would not reach me, because, by definition, somebody else would have solved it." - Barack Obama

Some people see the world as black and white, however, better decision making can arise from moving away from black and white thinking, and considering the probabilities associated with the possible outcomes.

Section 3.2: Introduction to estimation

Random variables and realisations

Suppose we wish to describe a random process and have collected data

$$y_1, y_2, ..., y_n$$

- These values $y_1, y_2, ..., y_n$ are realisations of random variables $Y_1, Y_2, ..., Y_n$.
- In a random sample, these random variables are independent and follow the same probability distribution.

For example, commute to college on the DART.

Random sample of days' journey times: $Y_1, Y_2, ..., Y_n$ with realisations (in minutes):

$$y_1 = 20, y_2 = 19.6, ..., y_n = 21.5$$

A model for the mean

We may want to investigate the "centre" of the process.

- The true (population) mean is μ.
- As a first step, we compute the sample mean and say this is a fair estimate.

$$\bar{y} = \frac{\Sigma_{i=1}^{n} y_i}{n} = \hat{\mu}$$

- But how close or how far away will $\bar{y}$ be from μ?

Before we can go any further, we need to make some assumptions about the distribution of $Y_1, Y_2, ..., Y_n$.

For example, we might assume that $Y_i \sim N(\mu, \sigma^2)$, IID (independent and identically distributed).

Important note on terminology:

- What does the $\hat{}$ on $\hat{\mu}$ mean?
- What is the difference between μ and $\hat{\mu}$?

Assumptions in Statistics

Has anyone ever fit a statistical model to data before?

When we fit any statistical model to data, or perform any statistical analysis, we must:

1. Know what (if any) assumptions are being made by the model or analysis.
2. Verify that those assumptions are reasonable.

This is crucial!

Section 3.3: Sampling distributions

What is a sampling distribution?

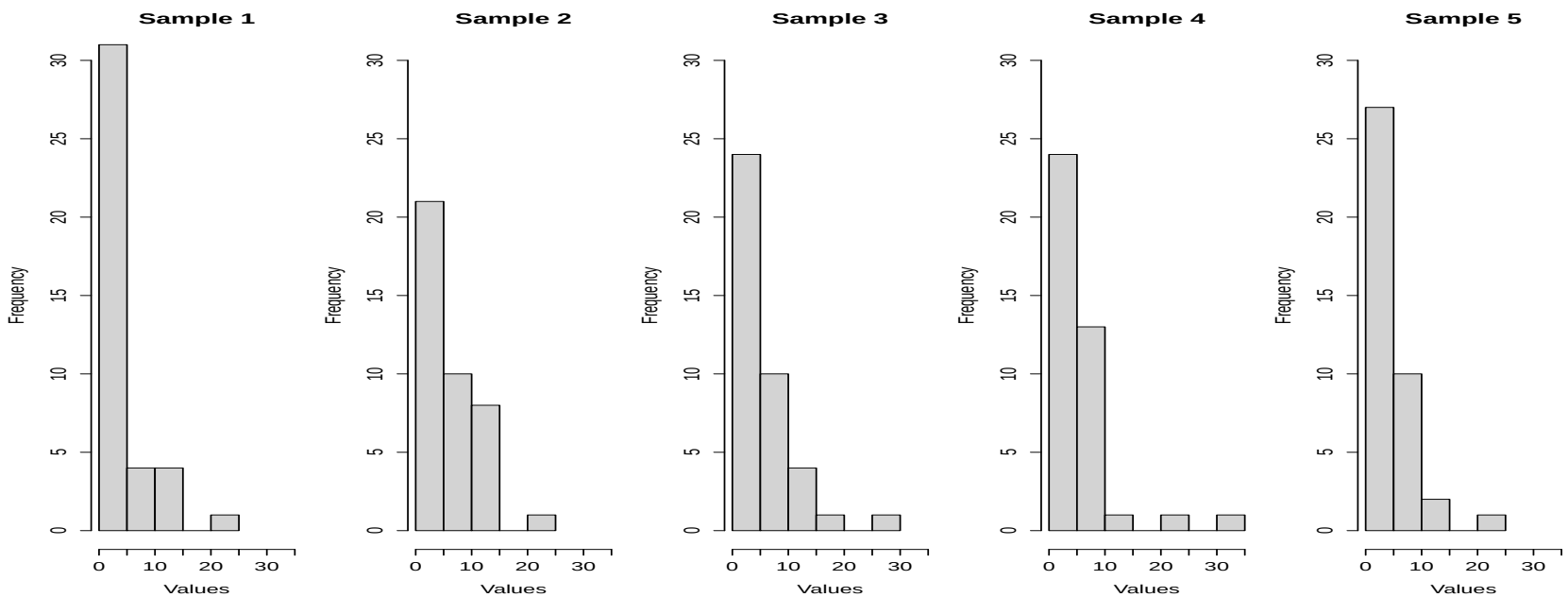
- **Parameter**: Population characteristic.
  - For example, μ, σ, σ^2, π.
- **Sample statistic**: Any quantity computed from values in a sample.
  - For example $\bar{x}$, s, s^2, p.
- The value of a population characteristic is fixed, but if you take, for example, ten samples from a population and compute $\bar{x}$ for each sample, would you expect each $\bar{x}$ to be the same?
- A sample statistic (considered in the context of any possible sample from the population) is a random variable and it has a probability distribution called the 'sampling distribution'. For example, we may talk about the sampling distribution of $\bar{X}$.
- **NB: The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample.**

Illustration of the sampling distribution of the mean

From previous slide: The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample.

Let's simulate from the exponential distribution with $\lambda = 0.2$. (Aside: what is the mean of this distribution? And the variance?)

We simulate 1000 samples from this distribution, each containing 40 values. Here are histograms of the first five samples:



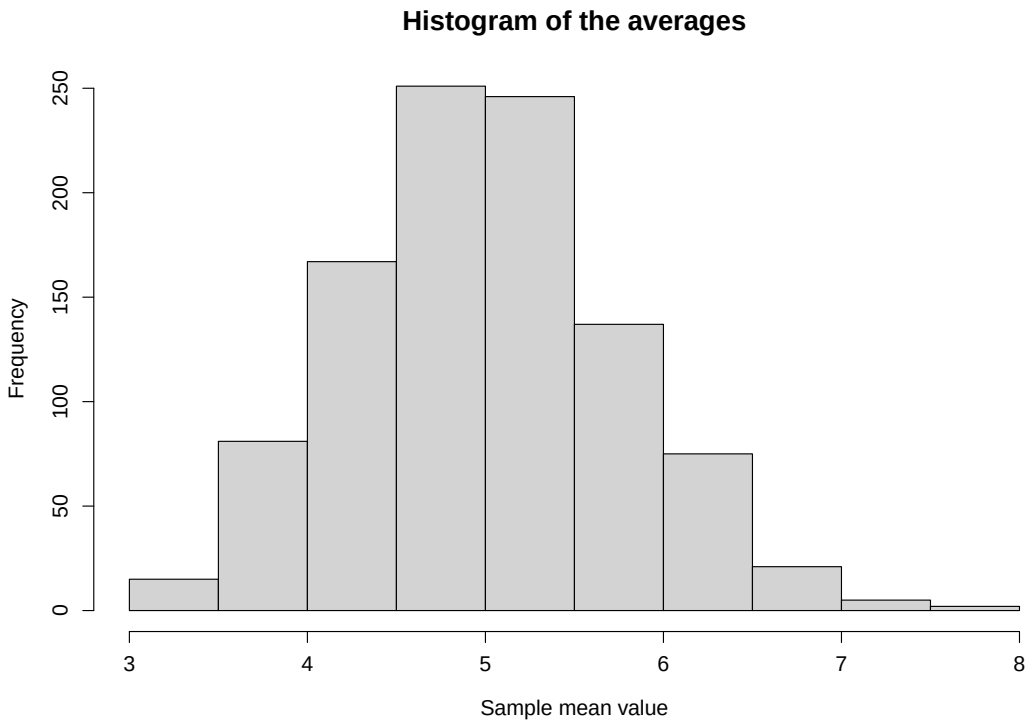
We could look at the histogram for any of the 1000 simulated samples.

Examine the means from the 1000 samples

We saw the exponential shape in the histograms of the first five of 1000 samples simulated.

The averages from the 40 values from each of the 1000 samples was computed. The first five are: 4.38, 5.87, 5.44, 5.54, 4.15.

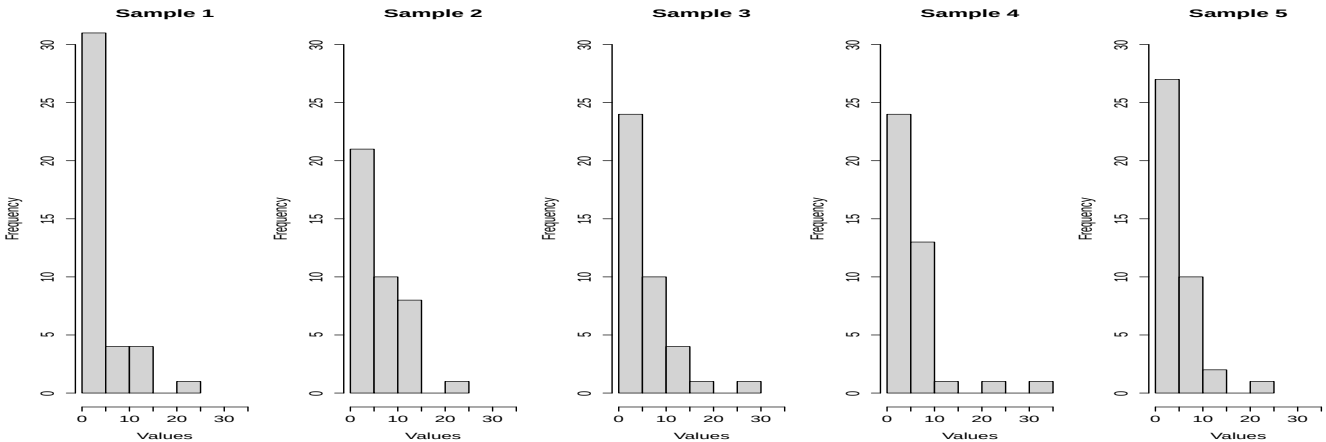
Let's plot the averages from all 1000 samples generated.



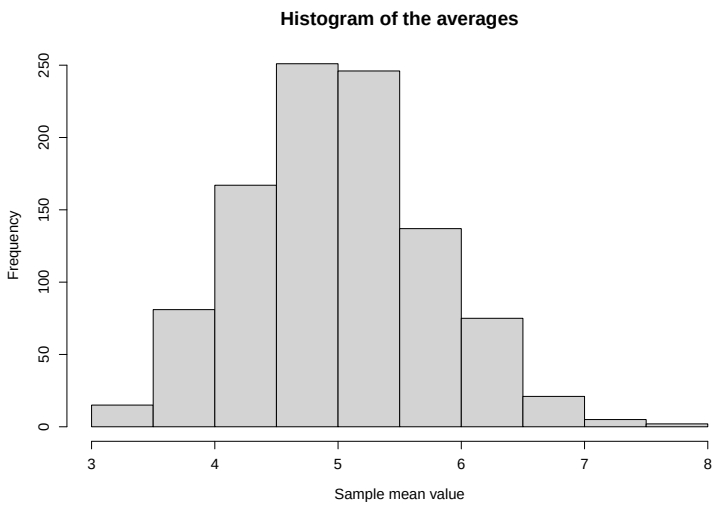
Does this look to be exponential? How would you describe this distribution?

Sample distribution versus sampling distribution

Sample distributions:



Sampling distribution:



Theory underlying sampling distributions

Recap:

The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample.

In later sections in this module, we will examine the theory behind the observation that we have just seen.

Section 3.4: In-class experiment

What were the questions?

- During the first lecture, many of you participated in an in-class expeirment. Thank you again for this!
- What did we test during this experiment?
- Each person was asked just one question, but there were six different questions.
- Each question required you to put a list of words in alphabetical order; they were:
  - bouncing handle kitchen tracksuit university
  - washing weird which wisdom wonderful
  - mobile model moment mountain movie
  - stack stake standard stapler statistics
  - starboard stardom starfish starry startle
  - crossbar crossfire crossing crossroad crossword

What was I trying to prove?

What hypothesis do you think I was trying to prove?

Here are the questions again.

- Each question required you to put a list of words in alphabetical order; they were:
  - bouncing handle kitchen tracksuit university
  - **w**ashing **w**eird **w**hich **w**isdom **w**onderful
  - **mo**bile **mo**del **mo**ment **mo**untain **mo**vie
  - **sta**ck **sta**ke **sta**ndard **sta**pler **sta**tistics
  - **star**board **star**dom **star**fish **star**ry **star**tle
  - **cross**bar **cross**fire **cross**ing **cross**road **cross**word

How did things work out?

Here is a quick look at the first few rows of data:

```
##   Question Seconds Correct
## 1        1      16       1
## 2        1      18       1
## 3        1      23       1
## 4        1      24       1
## 5        1      25       1
## 6        1      26       1
```

What do we need to think about to continue making sense of this data?

Group level information?

Here are some summary statistics for each 'group'.

```
##   Question Count Av_Seconds Sd_Seconds
## 1        1    13      31.85     20.663
## 2        2    25      34.48     11.034
## 3        3    25      46.12     27.213
## 4        4    27      47.00     19.213
## 5        5    18      39.11     10.272
## 6        6    31      42.97     14.775
```

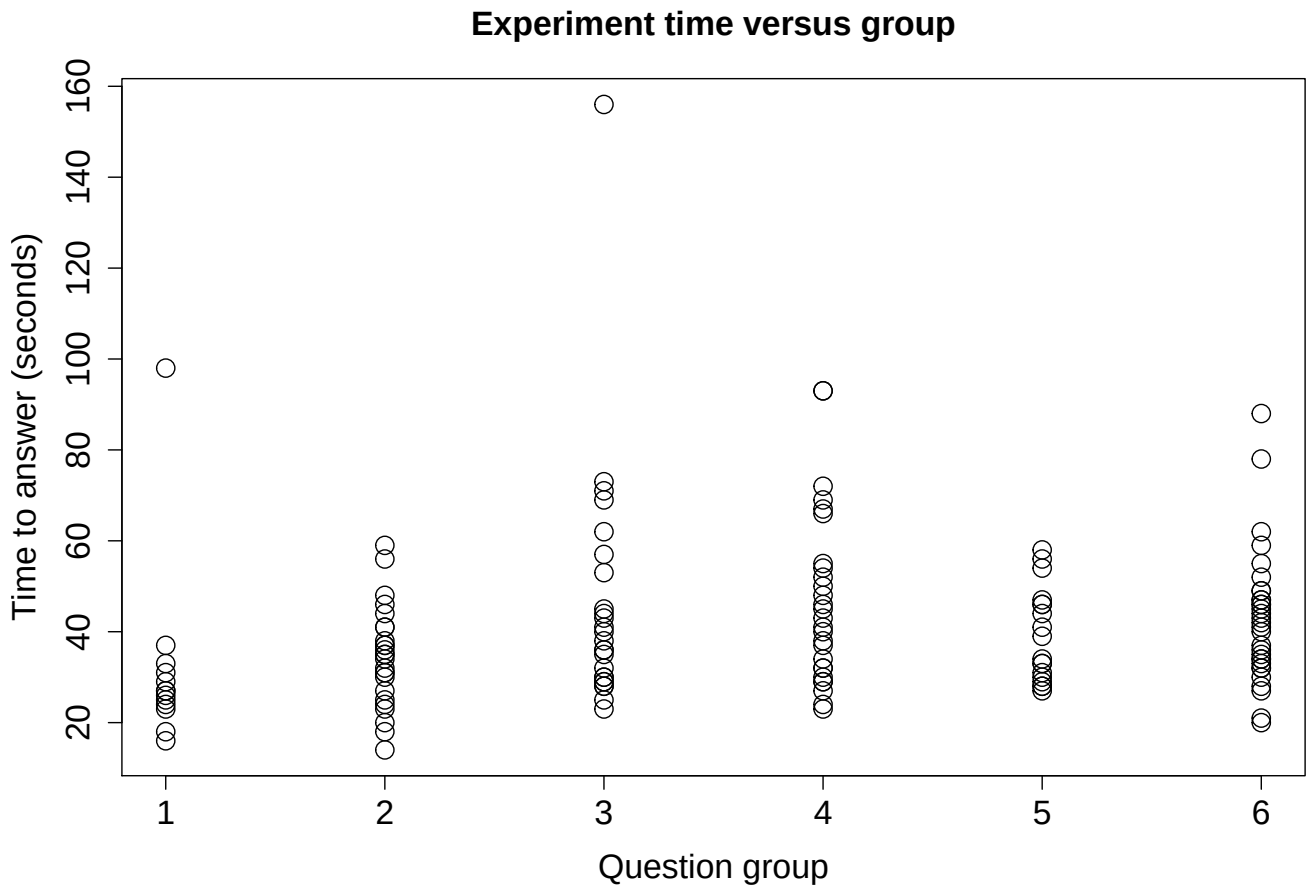
What do we need to think about to continue making sense of this data?

How many got the question wrong?

- One person in each group, except the first one.
- Does that matter?

Visualise the data

Here is a quick look at the first few rows of data and a scatter plot:



What might be the next steps to analyse this data?