Introduction to Statistics

Exercises: Part 1

Sophie Lee

Exercise 1: Missing data

A mobile phone app collects user information. Some of this data are missing as users chose to opt out of location-based services.

Give one scenario where the missing data would lead to **biased** analysis results.

Give another scenario where the missing data leads to a reduced sample but could still be used to produce **unbiased** results.

Exercise 2: Summary statistics

Question 1

The following line graphs are taken from the Criminal Courts statistics quarterly report, showing the average number of days from offense to completion for defendants at the Crown Court. Which graph is the most appropriate to display the average time and why?

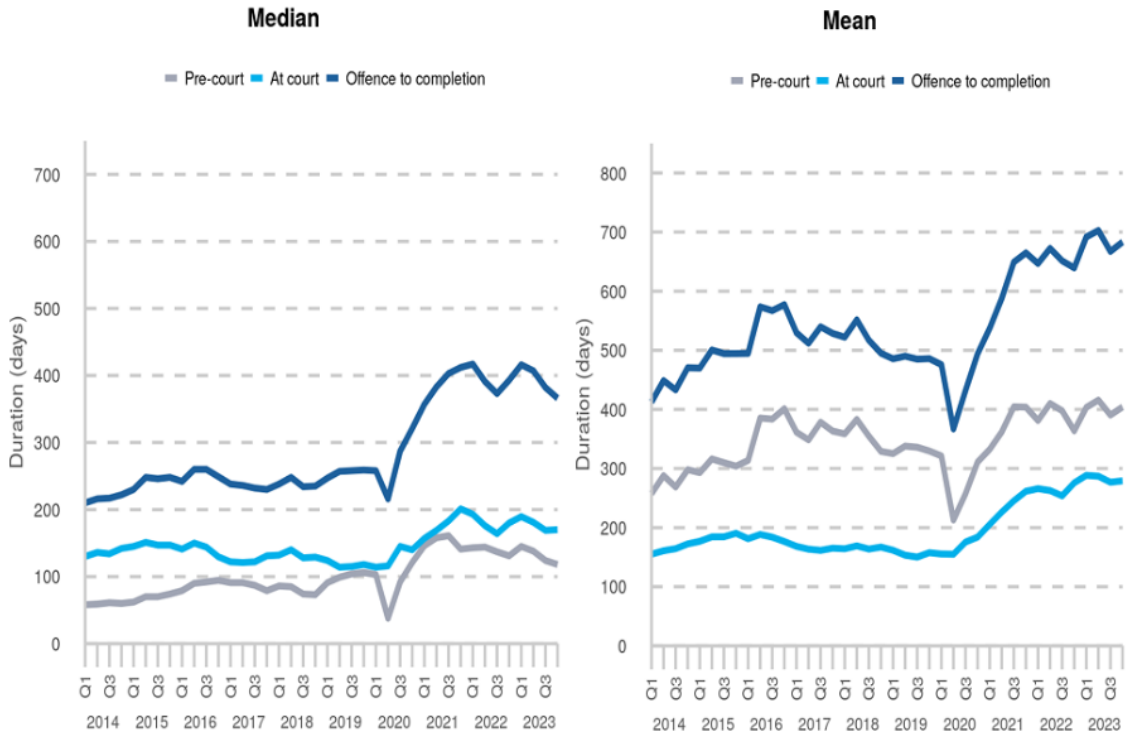


Figure 1: Average number of days from offence to completion for defendants dealt with at the Crown Court, Q1 2014 – Q4 2023

Question 2

The mean and standard deviation of waiting times in weeks for hearings in the Crown Court between 2020 and 2023 are given in the table below. Using this information, what can you tell about the distribution of these times?

Year	Mean wait (weeks)	SD wait (weeks)
2020	13.5	10.2
2021	19.1	15.6
2022	20.9	13.9
2023	22.8	14.7

Exercise 3: Inferential statistics

The output below is taken from an analysis that compared reoffending behaviour of 249 men participating with the Keyworking programme from Only Connect (OC) with those receiving standard support. More information about the findings of this report and the intervention itself can be found on the report webpage.

Comment on the way in which these results have been presented. Are the results clear? What has been done well? Is there anything you think could be improved? Do the results shown appear to be valid (from the information we are given)?

Table 1: Proportion of men who committed a proven reoffence in a one-year period (reoffending rate) after support from Only Connect compared with a matched comparison group

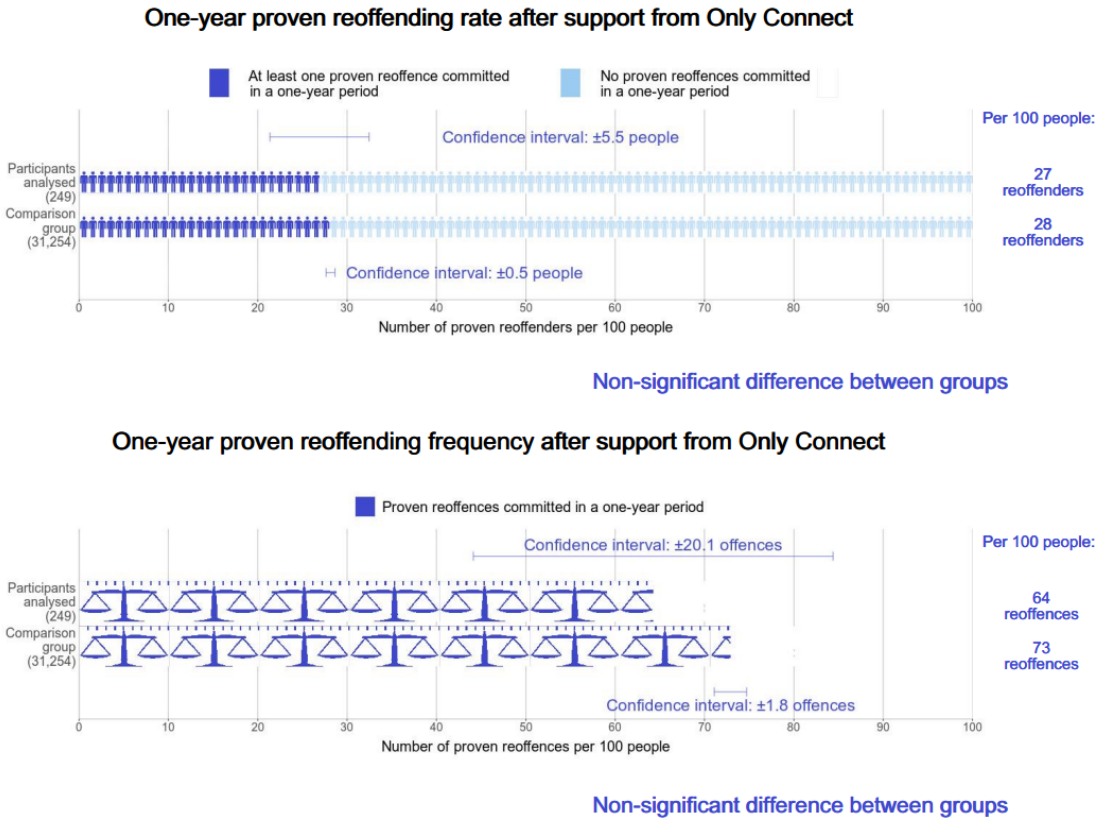
Number in treatment group	Number in comparison group	Treatment group rate (%)	Comparison group rate (%)	Estimated difference (% points)	Significant difference?	p-value
249	31,254	27	28	-7 to 4	No	0.66

Table 2: Number of proven reoffences committed in a one-year period (reoffending frequency - offences per person) by men who received support from Only Connect compared with a matched comparison group

Number in treatment group	Number in comparison group	Treatment group frequency	Comparison group frequency	Estimated difference	Significant difference?	p-value
249	31,254	0.64	0.73	-0.29 to 0.12	No	0.40

Table 3: Average time (days) to first proven reoffence in a one-year period for men who received support from Only Connect, compared with a matched comparison group (reoffenders only)

Number in treatment group	Number in comparison group	Treatment group time (days)	Comparison group time (days)	Estimated difference	Significant difference?	p-value
67	7,503	184	166	-6 to 43	No	0.14



Exercise 4

Below are outputs from 4 models aiming to answer the research question: was body mass of penguins related to flipper length.

Using model comparisons and checking the diagnostic tests, which model would you choose to use to answer this question and what would your answer be?

```
lm_flipper <- lm(body_mass_g ~ flipper_length_mm, data = penguins)
lm_flipper_sex <- lm(body_mass_g ~ flipper_length_mm + sex, data = penguins)
lm_flipper_sex_bill <- lm(body_mass_g ~ flipper_length_mm + sex + bill_depth_mm, data = pengu
lm_bill_sex<- lm(body_mass_g ~ sex + bill_depth_mm, data = penguins)

model_output <- function(model){
  broom::tidy(model, conf.int = T) %>%
    mutate(across(is.numeric, ~round(., 2)),
           con.int = paste0("[", conf.low, ", ", conf.high, "]"),
           pvalue = ifelse(p.value == 0, "<0.01", p.value)) %>%
```

```
    select(term, estimate, con.int, pvalue) %>%
    kable(col.names = c("", "Coefficient", "95% CI", "p"))
}

models <- list(lm_flipper, lm_flipper_sex, lm_flipper_sex_bill, lm_bill_sex)

map(models, model_output)
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `across(is.numeric, ~round(., 2))`.
Caused by warning:
! Use of bare predicate functions was deprecated in tidyselect 1.1.0.
i Please use wrap predicates in `where()` instead.
  # Was:
  data %>% select(is.numeric)

  # Now:
  data %>% select(where(is.numeric))
```

```
[[1]]
```

	Coefficient	95% CI	p
(Intercept)	-5780.83	[-6382.36, -5179.3]	<0.01
flipper_length_mm	49.69	[46.7, 52.67]	<0.01

```
[[2]]
```

	Coefficient	95% CI	p
(Intercept)	-5410.30	[-5972.52, -4848.09]	<0.01
flipper_length_mm	46.98	[44.15, 49.82]	<0.01
sexmale	347.85	[268.49, 427.21]	<0.01

```
[[3]]
```

	Coefficient	95% CI	p

```
|(Intercept)       |    -2246.83|[-3476.89, -1016.77] |<0.01 |
|flipper_length_mm |       38.19|[34.09, 42.29]       |<0.01 |
|sexmale           |      538.08|[437.14, 639.02]     |<0.01 |
|bill_depth_mm     |      -86.95|[-117.35, -56.54]    |<0.01 |

[[4]]


|              | Coefficient|95% CI             |p     |
|:-------------|-----------:|:------------------|:-----|
|(Intercept)   |     8779.10|[8304.42, 9253.79] |<0.01 |
|sexmale       |     1122.13|[1009.87, 1234.4]  |<0.01 |
|bill_depth_mm |     -299.34|[-327.89, -270.8]  |<0.01 |
```

```r
evaluations <- function(model) {
  tibble(terms = attr(model$model, "term.labels"),
         adj.r = summary(model)$adj.r.squared,
         rmse = rmse(model$model$body_mass_g, predict(model)),
         mae = mae(model$model$body_mass_g, predict(model))) %>%
    kable()
}
map(models, evaluations)
```

```
[[1]]


|     adj.r|     rmse|      mae|
|---------:|--------:|--------:|
| 0.7582837| 393.1236| 313.0018|

[[2]]


|     adj.r|     rmse|     mae|
|---------:|--------:|-------:|
| 0.8046607| 354.2762| 283.245|

[[3]]


|     adj.r|     rmse|      mae|
|---------:|--------:|--------:|
```

```
| 0.8212593| 338.3763| 276.4011|

[[4]]


|     adj.r|     rmse|      mae|
|---------:|--------:|--------:|
| 0.6399402| 480.9883| 383.8665|
```