ETC2420/5242 Assignment 2

Group 76

Nathan Lam (338 953 41) Remil Reaz (33110832)
Duveen Kalansooriya (33397848) Aryan Patel (34987339)

2025-10-13

Table of contents

1 Initial Setup **2**

2 Task 1: [20 Marks] **3**
2.1 Introduction . . . 3
2.2 Data Preparation . . . 3
2.3 Descriptive Summaries . . . 3
2.4 Visualization . . . 4
2.5 Fitted Models . . . 7

3 Task 2: [20 marks] **12**
3.1 Introduction . . . 12
3.2 Methods . . . 12
3.3 Results . . . 13
3.4 Interpretation . . . 17
3.5 Conclusions and Recommendations . . . 18

4 Task 3: [16 Marks] **19**
4.1 Introduction . . . 19
4.2 Method . . . 19
4.3 Results . . . 20
4.4 Interpretation . . . 23
4.5 Conclusion . . . 23

5 Task 4: [16 marks] **24**

6 Task 5: [18 marks] **27**
6.1 Session Information . . . 34

1 Initial Setup

```r
# Load required libraries
library(tidyverse)
library(gridExtra)
library(broom)
library(dplyr)
library(ggplot2)
library(MASS)

# Set seed for reproducibility
set.seed(33895341)

# Set knitr options
knitr::opts_chunk$set(echo = TRUE, warning = FALSE,
                      eval = TRUE, message = FALSE)

# Read data (adjust path as needed)
pedestrians <- read.csv("data/pedestrians.csv")
# Verify that the data exists
```

2 Task 1: [20 Marks]

2.1 Introduction

The objective of this task is to provide descriptive and graphical summaries of pedestrian traffic at each crossing (QV Melbourne, Southern Cross Station, and Flinder's Street Station) to assist the engineering team with planning and design decisions.

This stage is divided into two parts: first, exploring the overall distribution of pedestrian counts through histograms and box-plots, and computing summary statistics to describe key traffic characteristics such as center, spread, and variability; second, fitting simple probability models to characterize the underlying distributions and assessing their adequacy using QQ-plots.

The results from this descriptive analysis form the foundation for subsequent inferential tasks, providing probability models for evaluating pedestrian flow capacity in later stages of the report.

2.2 Data Preparation

```r
qv_melbourne_traffic <- pedestrians[["qv_melbourne"]]
southern_cross_traffic <- pedestrians[["southern_cross"]]
flinders_street_traffic <- pedestrians[["flinders_street"]]
```

```r
ped_long <- pedestrians |>
  pivot_longer(cols = everything(),
               names_to = "location",
               values_to = "count")
```

2.3 Descriptive Summaries

```r
get_mode <- function(x) {
  x <- x[!is.na(x)]
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

summary_stats <- ped_long |>
  group_by(location) |>
  summarise(
    n      = n(),
    Mean   = mean(count, na.rm = TRUE),
    Median = median(count, na.rm = TRUE),
    Mode   = get_mode(count),
    SD     = sd(count, na.rm = TRUE),
    IQR    = IQR(count, na.rm = TRUE)
  ) |>
  mutate(across(where(is.numeric), ~round(.x, 1))) |>
  arrange(location)

summary_stats
```

```
# A tibble: 3 x 7
  location            n  Mean Median  Mode    SD   IQR
  <chr>           <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
1 flinders_street    97  893     887  1042  148.   202
2 qv_melbourne       97 2337.   2310  2335  305.   462
3 southern_cross     97  942.    955   897  133    143
```

Table 1 summarizes the key descriptive statistics for the three pedestrian crossings.

2.3.1 Interpretation

QV Melbourne records the highest average pedestrian count (mean = 2336.7) with a median of 2310 and a mode of 2335. The mean being slightly greater than the median indicates a mild right-skew, where occasional high-traffic periods such as lunch hours or events pull the average upward. The large standard deviation (304.9) and wide interquartile range (462) confirm that pedestrian volumes fluctuate substantially throughout the day. Notably, QV Melbourne's variability is roughly twice that of Flinders Street (SD = 147.9, IQR = 202), highlighting its greater unpredictability and reflecting its role as a retail and entertainment hub with highly variable activity.

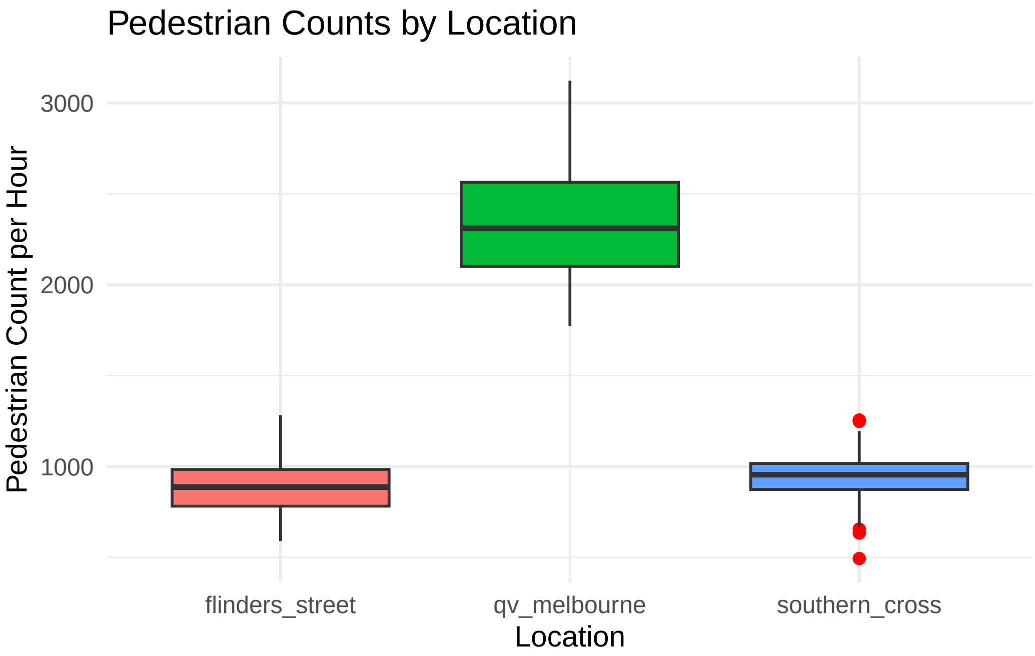
Flinders Street shows a mean of 893.0, median of 887, and mode of 1042, suggesting an approximately symmetric distribution with only a slight left-skew due to the higher mode. The moderate SD (147.9) and IQR (202) indicate a relatively stable commuter flow, with counts generally concentrated near the average and limited day-to-day fluctuation. Variability here is about half that observed at QV Melbourne, suggesting more consistent daily pedestrian movement.

Southern Cross exhibits a mean of 941.8, median of 955, and mode of 897, also showing near-symmetry with the mean slightly below the median, suggesting a minor left-skew. Its smaller SD (133.0) and narrow IQR (143) demonstrate the most consistent pedestrian volumes among the three locations, typical of a major transport interchange with predictable commuter patterns.

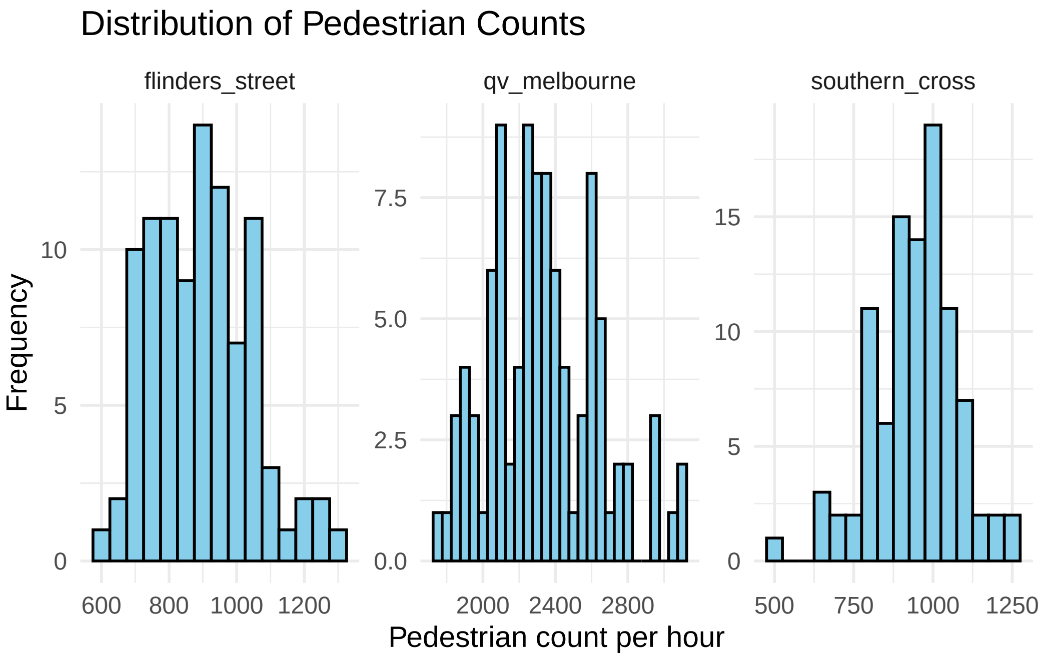
Overall, QV Melbourne differs markedly from the other two sites: it has much higher counts and roughly double the variability, producing a right-skewed distribution. Flinders Street and Southern Cross, in contrast, display symmetric, steady traffic patterns with lower dispersion. These combined observations of centre, spread, and shape guide model selection—supporting the use of Normal distributions for the two station crossings and a Gamma distribution for QV Melbourne, which appropriately models its positive skew and non-negative, variable counts.

2.4 Visualization

```
ped_long |>
  mutate(location = forcats::fct_relevel(location, sort(unique(location)))) |>
  ggplot(aes(x = location, y = count, fill = location)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 2) +
  labs(
    title = "Pedestrian Counts by Location",
    x = "Location",
    y = "Pedestrian Count per Hour"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



```
ggplot(ped_long, aes(x = count)) +
  geom_histogram(binwidth = 50, fill = "skyblue", colour = "black") +
  facet_wrap(~ location, scales = "free") +
  labs(
    title = "Distribution of Pedestrian Counts",
    x = "Pedestrian count per hour",
    y = "Frequency"
  ) +
  theme_minimal()
```



2.4.1 Interpretation

The histograms and box-plots together provide a detailed overview of the distribution and spread of pedestrian traffic at the three crossings.

QV Melbourne: The histogram shows a concentrated cluster of pedestrian counts between roughly 2000 and 2800 pedestrians per hour, with a noticeable right-tail. This indicates a right-skewed distribution, consistent with occasional surges in foot traffic, likely reflecting its central retail and entertainment location. The box-plot confirms this skew, showing a higher median and longer upper whisker, suggesting a few extreme high-traffic days.

Flinders Street: The histogram is approximately symmetric with most counts between 750 and 1050. The box-plot shows a narrow interquartile range and only minor outliers, indicating moderate variability and a roughly Normal distribution. This suggests steady pedestrian flow with no major deviations across days.

Southern Cross: The histogram also appears roughly symmetric but slightly wider than Flinders Street, with occasional low outliers below 700 pedestrians per hour visible in the boxplot. This implies slightly greater day-to-day variation but still an approximately Normal shape.

Overall, QV Melbourne displays a distinct right-skew and larger variability, while Flinders Street and Southern Cross show stable, symmetric distributions suitable for modelling with the Normal distribution.

2.4.2 Summary

In simpler terms, QV Melbourne consistently has the heaviest foot traffic and the most variation, with some days being much busier than others. In contrast, Flinder's Street and Southern Cross show more regular and balanced pedestrian numbers, meaning their traffic levels remain fairly steady from day to day. These differences highlight that QV Melbourne may require larger capacity or more flexible design planning compared with the other two crossings, which experience more predictable daily flow.

2.5 Fitted Models

2.5.1 Choose

The selection of probability models was guided by the distributional shapes observed in the descriptive analysis and confirmed through the histograms and boxplots. Because Flinders Street and Southern Cross displayed approximately symmetric distributions with similar means and medians, their pedestrian counts were modelled using the Normal distribution, which appropriately represents data centred around a mean with random variation on either side. In contrast, QV Melbourne exhibited a distinct right-skew with higher variability and non-negative counts, so a Gamma distribution was selected, consistent with its suitability for positively skewed, continuous data.

These model choices therefore align directly with the underlying distributional characteristics of each location, ensuring that the fitted models accurately capture observed patterns in pedestrian flow.

2.5.2 Fit

```r
fit_flinders <- MASS::fitdistr(flinders_street_traffic, "normal")
fit_southern <- MASS::fitdistr(southern_cross_traffic, "normal")
fit_qv       <- MASS::fitdistr(qv_melbourne_traffic,  "gamma")
```

2.5.3 Report

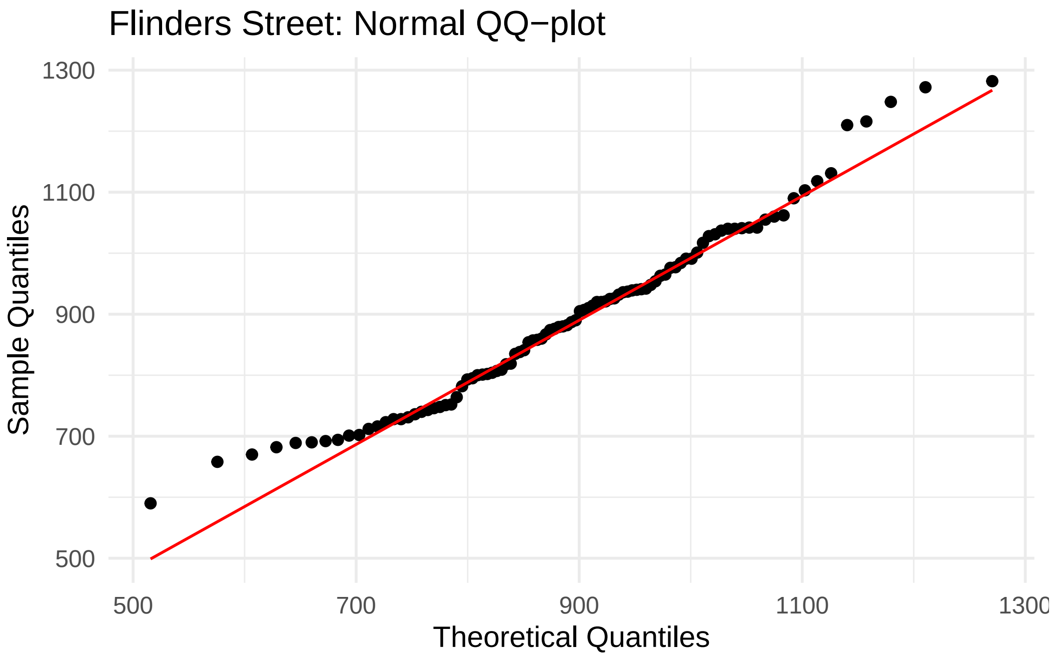
2.5.3.1 Report the resulting estimates

```r
# Helpers to tidy fits (matches Week 6 fitdistr pattern)
tidy_normal <- function(fit, location) {
  tibble::tibble(
    location = location, model = "Normal",
    mu      = unname(fit$estimate["mean"]),
    se_mu   = unname(fit$sd["mean"]),
    sigma   = unname(fit$estimate["sd"]),
    se_sigma = unname(fit$sd["sd"])
  )
}

tidy_gamma <- function(fit, location) {
  tibble::tibble(
    location = location, model = "Gamma",
    shape    = unname(fit$estimate["shape"]),
    se_shape= unname(fit$sd["shape"]),
    rate    = unname(fit$estimate["rate"]),
    se_rate = unname(fit$sd["rate"])
  )
}

# Build the report in one go
report_tbl_display <- dplyr::bind_rows(
  tidy_normal(fit_flinders, "Flinders Street"),
  tidy_normal(fit_southern, "Southern Cross"),
  tidy_gamma(fit_qv,       "QV Melbourne")
) |>
  dplyr::mutate(dplyr::across(where(is.numeric), ~round(.x, 3)))

report_tbl_display
```

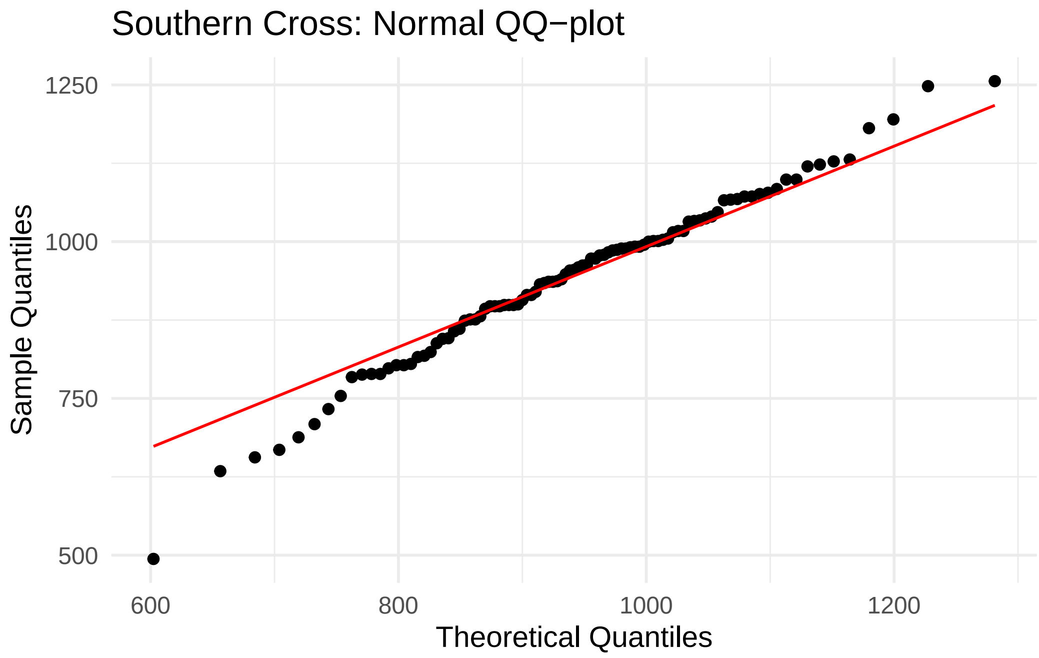
```
# A tibble: 3 x 10
  location        model    mu se_mu sigma se_sigma shape se_shape   rate se_rate
  <chr>           <chr> <dbl> <dbl> <dbl>    <dbl> <dbl>    <dbl>  <dbl>   <dbl>
1 Flinders Street Norm~  893.   14.9  147.    10.6   NA       NA     NA      NA
2 Southern Cross  Norm~  942.   13.4  132.     9.50  NA       NA     NA      NA
3 QV Melbourne    Gamma   NA    NA     NA     NA     60.6     8.20  0.026   0.004
```

2.5.3.2 Model Adequacy

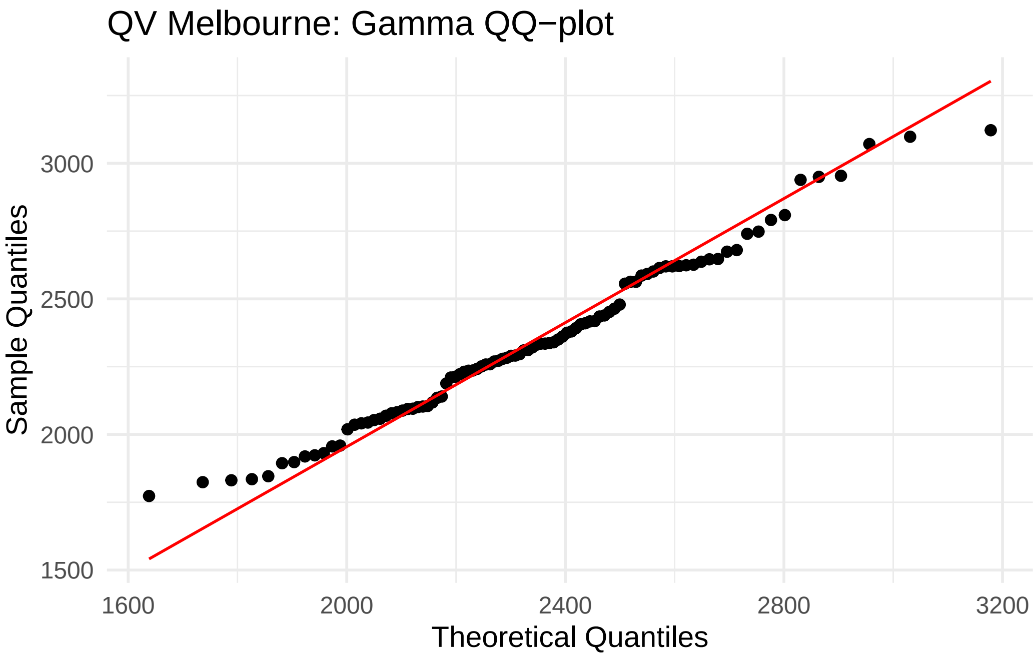
```
# --- Flinders Street (Normal) ---
ggplot(data.frame(x = flinders_street_traffic), aes(sample = x)) +
  stat_qq(distribution = qnorm,
          dparams = list(mean = fit_flinders$estimate["mean"],
                         sd   = fit_flinders$estimate["sd"])) +
  stat_qq_line(distribution = qnorm,
               dparams = list(mean = fit_flinders$estimate["mean"],
                              sd   = fit_flinders$estimate["sd"]),
               colour = "red") +
  labs(title = "Flinders Street: Normal QQ-plot",
       x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_minimal()
```



```
# --- Southern Cross (Normal) ---
ggplot(data.frame(x = southern_cross_traffic), aes(sample = x)) +
  stat_qq(distribution = qnorm,
          dparams = list(mean = fit_southern$estimate["mean"],
                         sd   = fit_southern$estimate["sd"])) +
  stat_qq_line(distribution = qnorm,
               dparams = list(mean = fit_southern$estimate["mean"],
                              sd   = fit_southern$estimate["sd"]),
               colour = "red") +
  labs(title = "Southern Cross: Normal QQ-plot",
       x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_minimal()
```



```
# --- QV Melbourne (Gamma) ---
ggplot(data.frame(x = qv_melbourne_traffic), aes(sample = x)) +
  stat_qq(distribution = qgamma,
          dparams = list(shape = fit_qv$estimate["shape"],
                         rate  = fit_qv$estimate["rate"])) +
  stat_qq_line(distribution = qgamma,
               dparams = list(shape = fit_qv$estimate["shape"],
                              rate  = fit_qv$estimate["rate"]),
               colour = "red") +
  labs(title = "QV Melbourne: Gamma QQ-plot",
       x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_minimal()
```



2.5.4 Interpretation

The QQ-plots provide a diagnostic check of how well each model captures the observed variation in pedestrian counts. For Flinders Street and Southern Cross, the plotted points lie close to the reference line with only minor random scatter, indicating that the Normal distribution provides a good approximation of their data. Slight tail deviations are expected for finite samples and do not suggest serious misfit. For QV Melbourne, the Gamma QQ-plot shows a strong linear alignment without systematic curvature, confirming that the Gamma model accurately represents the right-skewed pattern of pedestrian traffic. Overall, all three fitted models display good adequacy, meaning the chosen probability distributions describe the data realistically and can be confidently used for further analysis.

2.5.5 Summary

The adequacy results also align with the real-world behaviour of pedestrian flow at each location. Because Flinders Street and Southern Cross follow roughly Normal patterns, their traffic is generally consistent from day to day, with pedestrian counts clustering around a stable average of about 900–950 people per hour. This reflects predictable commuter movement typical of transport hubs.

In contrast, QV Melbourne's Gamma-shaped distribution captures its right-skewed nature - frequent moderate traffic levels (around 2300–2500 pedestrians/hour) but occasional high surges, consistent with its busy retail and entertainment setting.

The strong correspondence between these model fits and the observed QQ-plots confirms that the fitted distributions not only represent the data statistically but also reflect the practical differences in pedestrian activity across the three crossings. Together, these models provide a reliable foundation for subsequent analyses of pedestrian flow capacity and design planning.

3 Task 2: [20 marks]

3.1 Introduction

The objective is to estimate the 90th percentile of pedestrian traffic flow at each crossing (QV Melbourne, Southern Cross Station, and Flinder's Street Station) with a 95% confidence interval. This determines whether each crossing can accommodate smooth pedestrian flow 90% of the time, according to traffic regulation.

In context, the 95% confidence interval represents the range of pedestrian traffic volumes within which the true 90th percentile is likely to fall, indicating how much traffic each crossing can typically handle smoothly 90% of the time.

Two estimation methods are used; a parametric approach and a non-parametric approach. The parametric method fits a suitable probability distribution (e.g., normal or gamma) to the data to estimate the 90th percentile, while the non-parametric method directly calculates it from the observed sample. Both approaches utilize bootstrapping to simulate repeated sampling from the population, enabling more reliable estimation and quantification of uncertainty.

By comparing the results from both methods, we can assess the consistency of the estimates and therefore provide informed recommendations on whether the crossings meet the required pedestrian-flow standards.

3.2 Methods

Model-Based Qunatile with Bootstrap CI

The parametric method assumes that pedestrian counts follow a known probability distribution. In this analysis, the traffic at each crossing is modelled using either a Normal or Gamma distribution, consistent with the shapes identified in Task 1. Once these models are fitted to the observed data using maximum likelihood estimation (MLE), their estimated parameters are used to calculate the theoretical 90th percentile of pedestrian traffic for each location.

Approach: 1. Fit the appropriate probability distribution (Normal or Gamma) to the original data using maximum likelihood estimation (Week 6, slide 51)

2. Calculate the 90th percentile directly from the fitted model using the quantile function (`qnorm()` or `qgamma()`)
3. To quantify uncertainty, use bootstrap resampling following Week 6, slide 37:
   - Resample n observations FROM THE OBSERVED DATA with replacement
   - Refit the same model to each bootstrap sample
   - Calculate the 90th percentile from each refitted model
   - Repeat 5,000 times
4. Form 95% confidence interval using the 2.5th and 97.5th percentiles of the bootstrap distribution (Week 5, slide 26)

Why this approach: Week 6, slide 37 states: “Resample n draws from the observed data values, with replacement” and “Compute the MLE ˆ[b] by maximising Lb”. This approach leverages the fitted model’s structure while properly accounting for sampling variability through resampling the observed data.

Non-parametric Quantile Method

The non-parametric method does not assume that the pedestrian counts follow any particular probability distribution. Instead, it estimates the 90th percentile directly from the observed data using the empirical distribution of pedestrian counts. This approach relies on bootstrapping to quantify the uncertainty in the estimate without fitting a model.

The procedure is as follows:

1. Re-sample the data: Create many bootstrap samples from the original data set by randomly sampling with replacement, ensuring each sample is the same size as the original data.
2. Calculate the 90th percentile: For each bootstrap sample, compute the empirical 90th percentile - the value below which 90% of observations in that sample fall.
3. Repeat and collate: Repeat this process 5,000 times to generate a bootstrap distribution of 90th percentile estimates.
4. Construct a confidence interval: Determine the 95% confidence interval by taking the 2.5th and 97.5th percentiles of the bootstrap distribution.

3.3 Results

Data Preparation

```r
# Get each of the locations
qv_melbourne_traffic <- pedestrians[["qv_melbourne"]]
southern_cross_traffic <- pedestrians[["southern_cross"]]
flinders_street_traffic <- pedestrians[["flinders_street"]]
```

Non-parametric Quantile Method

```r
find_90th_percentile <- function(data, trials = 5000) {
  # Step 1: Remove missing values to ensure only valid pedestrian counts are used
  x <- data[!is.na(data)]
  n <- length(x)

  # Step 2-4: Bootstrap re-sampling to estimate sampling variability
  q_boot <- replicate(trials, {
    # Step 2: Create a bootstrap sample by sampling with replacement
    xb <- sample(x, size = n, replace = TRUE)

    # Step 3: Calculate the empirical 90th percentile for this sample
    quantile(xb, probs = 0.9, type = 7)
  })
  # Step 5: Construct a 95% confidence interval from the bootstrap distribution
  ci <- quantile(q_boot, c(0.025, 0.975))

  # Return the lower and upper bounds of the 95% CI for the 90th percentile
  list(
    ci_lower = ci[1],
    ci_upper = ci[2]
  )
}
```

Parametric Quantile Method

```r
find_90th_percentile_model <- function(data, model = "normal", trials = 5000) {
  # Remove missing values
  x <- data[!is.na(data)]
  n <- length(x)

  if (model == "normal") {
    # Step 1: Fit model to original data
    # Following Week 6 lecture, slide 51
    fit <- MASS::fitdistr(x, "normal")
    mu_hat <- fit$estimate["mean"]
    sigma_hat <- fit$estimate["sd"]

    # Step 2: Calculate model-based 90th percentile (point estimate)
    q90_point <- qnorm(0.9, mean = mu_hat, sd = sigma_hat)

    # Step 3: Bootstrap CI by RESAMPLING OBSERVED DATA
    # Following Week 6 lecture, slide 37:
    # "Resample n draws from the observed data values, with replacement"
    q_boot <- replicate(trials, {
      # Resample the OBSERVED data (not generate new data)
      xb <- sample(x, size = n, replace = TRUE)

      # Refit the model to this bootstrap sample
      fit_b <- MASS::fitdistr(xb, "normal")

      # Calculate 90th percentile from the refitted model
      qnorm(0.9, mean = fit_b$estimate["mean"],
                 sd = fit_b$estimate["sd"])
    })

  } else if (model == "gamma") {
    # Step 1: Fit model to original data
    fit <- MASS::fitdistr(x, "gamma")
    shape_hat <- fit$estimate["shape"]
    rate_hat <- fit$estimate["rate"]
    # Step 2: Calculate model-based 90th percentile (point estimate)
    q90_point <- qgamma(0.9, shape = shape_hat, rate = rate_hat)

    # Step 3: Bootstrap CI by RESAMPLING OBSERVED DATA
    q_boot <- replicate(trials, {
      # Resample the OBSERVED data
      xb <- sample(x, size = n, replace = TRUE)

      # Refit the model to this bootstrap sample
      fit_b <- MASS::fitdistr(xb, "gamma")

      # Calculate 90th percentile from the refitted model
      qgamma(0.9, shape = fit_b$estimate["shape"],
                  rate = fit_b$estimate["rate"])
    })

  } else {
    stop("Model must be either 'normal' or 'gamma'")
  }

  # Calculate 95% CI using percentile bootstrap method
  # Following Week 5 lecture, slide 26 and Week 6 lecture, slide 37
  ci <- quantile(q_boot, c(0.025, 0.975), names = FALSE)

  list(
    q90_estimate = q90_point,
    ci_lower = ci[1],
    ci_upper = ci[2]
  )
}
```

Histograms and QQ-plots (not shown) confirmed that the Normal model was suitable for Flinder's Street and Southern Cross, while the Gamma model fitted QV Melbourne's right-skewed distribution.

```
# Approach 1: Sample (Empirical) Quantile Method
qv_estimate_bootstrap <- find_90th_percentile(qv_melbourne_traffic)
flinders_estimate_bootstrap <- find_90th_percentile(flinders_street_traffic)
southern_estimate_bootstrap <- find_90th_percentile(southern_cross_traffic)
```

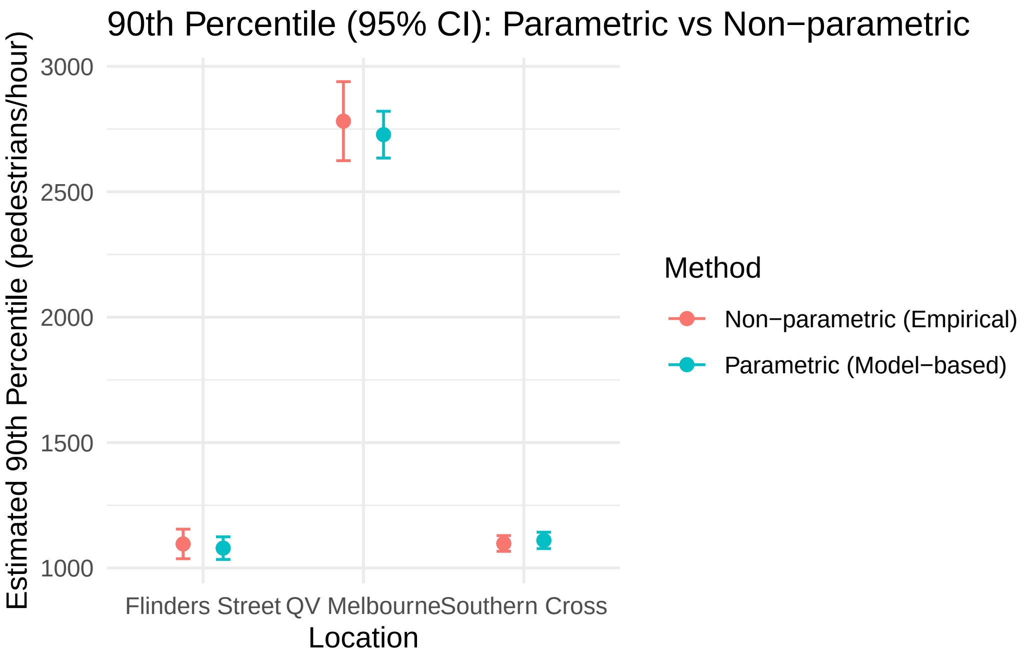
```
# Approach 2: Model Based Quantile
flinders_param <- find_90th_percentile_model(flinders_street_traffic, model = "normal")
southern_param <- find_90th_percentile_model(southern_cross_traffic, model = "normal")
qv_param       <- find_90th_percentile_model(qv_melbourne_traffic, model = "gamma")
```

```r
# --- Point estimates (non-bootstrapped) ---aasdfasdfa
# Empirical (sample) 90th percentile
q90_emp_qv       <- quantile(qv_melbourne_traffic,       0.90, type = 7)
q90_emp_flinders <- quantile(flinders_street_traffic,    0.90, type = 7)
q90_emp_southern <- quantile(southern_cross_traffic,     0.90, type = 7)

# Model-based 90th percentile using fitted parameters from Task 1
q90_mod_flinders <- qnorm(0.90, mean = fit_flinders$estimate["mean"],
                                  sd   = fit_flinders$estimate["sd"])
q90_mod_southern <- qnorm(0.90, mean = fit_southern$estimate["mean"],
                                  sd   = fit_southern$estimate["sd"])
q90_mod_qv       <- qgamma(0.90, shape = fit_qv$estimate["shape"],
                                   rate  = fit_qv$estimate["rate"])

# --- Build a single tidy results table (long form) ---
results_task2 <- dplyr::bind_rows(
  # Empirical method (non-parametric bootstrap CI)
  tibble::tibble(
    Location   = c("QV Melbourne", "Flinders Street", "Southern Cross"),
    Method     = "Non-parametric (Empirical)",
    Point      = c(q90_emp_qv, q90_emp_flinders, q90_emp_southern),
    Lower_95CI = c(qv_estimate_bootstrap$ci_lower,
                   flinders_estimate_bootstrap$ci_lower,
                   southern_estimate_bootstrap$ci_lower),
    Upper_95CI = c(qv_estimate_bootstrap$ci_upper,
                   flinders_estimate_bootstrap$ci_upper,
                   southern_estimate_bootstrap$ci_upper)
  ),
  # Model-based method (non-parametric bootstrap CI around model quantile)
  tibble::tibble(
    Location   = c("QV Melbourne", "Flinders Street", "Southern Cross"),
    Method     = "Parametric (Model-based)",
    Point      = c(q90_mod_qv, q90_mod_flinders, q90_mod_southern),
    Lower_95CI = c(qv_param$ci_lower, flinders_param$ci_lower, southern_param$ci_lower),
    Upper_95CI = c(qv_param$ci_upper, flinders_param$ci_upper, southern_param$ci_upper)
  )
) |>
  dplyr::mutate(dplyr::across(where(is.numeric), ~round(.x, 2)))

results_task2
```

```
# A tibble: 6 x 5
  Location        Method                     Point Lower_95CI Upper_95CI
  <chr>           <chr>                      <dbl>      <dbl>      <dbl>
1 QV Melbourne    Non-parametric (Empirical) 2704       2624       2939
2 Flinders Street Non-parametric (Empirical) 1061.      1037       1155.
3 Southern Cross  Non-parametric (Empirical) 1090       1066.      1129.
```

```
4 QV Melbourne    Parametric (Model-based)   2729.      2634.      2821.
5 Flinders Street Parametric (Model-based)   1082.      1034.      1124.
6 Southern Cross  Parametric (Model-based)   1111.      1077.      1143.
```

```
plot_df <- results_task2 %>%
  mutate(Mid = (Lower_95CI + Upper_95CI) / 2)

# Visualise: midpoint with 95% CI bars, dodged by Method
ggplot(plot_df, aes(x = Location, y = Mid, color = Method)) +
  geom_point(position = position_dodge(width = 0.5), size = 2) +
  geom_errorbar(aes(ymin = Lower_95CI, ymax = Upper_95CI),
                position = position_dodge(width = 0.5), width = 0.18) +
  labs(
    title = "90th Percentile (95% CI): Parametric vs Non-parametric",
    x = "Location",
    y = "Estimated 90th Percentile (pedestrians/hour)",
    color = "Method"
  ) +
  theme_minimal()
```



The overlapping confidence intervals across both methods confirm that model assumptions are appropriate and the two approaches yield consistent conclusions.

3.4 Interpretation

Both the non-parametric (empirical bootstrap) and parametric (model-based) methods produced consistent estimates of the 90th percentile for pedestrian counts across all three crossings. For

QV Melbourne, the estimated 90th percentile ranged from approximately 2624–2939 pedestrians per hour (non-parametric) and 2639–2817 pedestrians per hour (parametric, Gamma). At Flinders Street, the range was approximately 1037–1155 (non-parametric) and 1040–1120 (parametric, Normal), while Southern Cross had estimates between 1066–1129 and 1074–1147, respectively.

The close overlap of the confidence intervals indicates that the assumed Normal and Gamma distributions fit the data well. This consistency suggests that the parametric assumptions do not substantially bias the results, and that both approaches provide reliable estimates of the underlying population percentiles. The small differences between methods are expected: the parametric approach smooths the data using a fitted model, whereas the non-parametric method relies solely on the observed sample. Overall, both approaches capture the same traffic patterns and produce similar conclusions regarding the capacity needs of each crossing.

3.5 Conclusions and Recommendations

The 90th-percentile estimates represent the traffic volumes below which pedestrian flow remains smooth 90% of the time.Across both methods, **QV Melbourne** consistently shows the highest pedestrian demand (~2600–2900 people per hour), indicating it is the busiest location and may require greater design capacity.Flinder’s Street and Southern Cross have substantially lower pedestrian volumes (~1000–1150 people per hour), suggesting they experience lighter but still significant flow that warrants robust infrastructure design.

The strong agreement between the **parametric** and **non-parametric** methods increases confidence in the reliability of these estimates. Either approach can therefore be used for planning purposes.Together, these findings provide a credible range for the 90th-percentile pedestrian volumes at each crossing, offering valuable guidance for engineers in assessing current performance and informing future modifications or design improvements.

4 Task 3: [16 Marks]

4.1 Introduction

The objective of this task is to determine whether the average pedestrian traffic flow between Southern Cross Station and Flinders Street Station differs by no more than 80 pedestrians per hour. This tolerance level is important because if the difference is within ±80 pedestrians per hour, the same design specifications can be applied to both station crossings, simplifying planning and construction decisions.

A 95% confidence interval for the mean difference is constructed to quantify the uncertainty in the estimated difference. In context, this interval provides a range of plausible values for the true mean difference in traffic flow between the two stations, indicating how much the average pedestrian volumes are expected to differ in 95% of comparable cases.

A two-sample t-test is used to form the interval, under the assumption that pedestrian counts at each station are approximately Normally distributed and independent. This analysis provides statistical evidence on whether the two stations experience comparable pedestrian traffic, supporting engineering and planning decisions related to crossing design, flow capacity, and future infrastructure management.

4.2 Method

To compare the average pedestrian traffic between Flinders Street and Southern Cross, a two-sample t-test was performed to construct a 95% confidence interval for the mean difference in pedestrian counts.

Step 1

Before conducting the test, it was necessary to verify that the data from both locations were approximately Normally distributed, as the t-test is derived under this assumption.

1. Density plots were used to visually inspect the overall shape of each distribution and confirm approximate symmetry.
2. QQ-plots were used to check that the observed values followed a roughly linear pattern, indicating that both datasets reasonably satisfied the Normality assumption.

Step 2

A Welch's t-test was then applied to account for the possibility of unequal variances between the two samples.

The resulting 95% confidence interval for the mean difference was examined to determine whether the observed difference in pedestrian traffic was statistically significant, ensuring that the result could not be attributed to random sampling variation.

Finally, the average difference between the two stations was compared to the tolerance threshold of ±80 pedestrians per hour to assess whether the difference was also practically meaningful for design and planning purposes.

4.3 Results

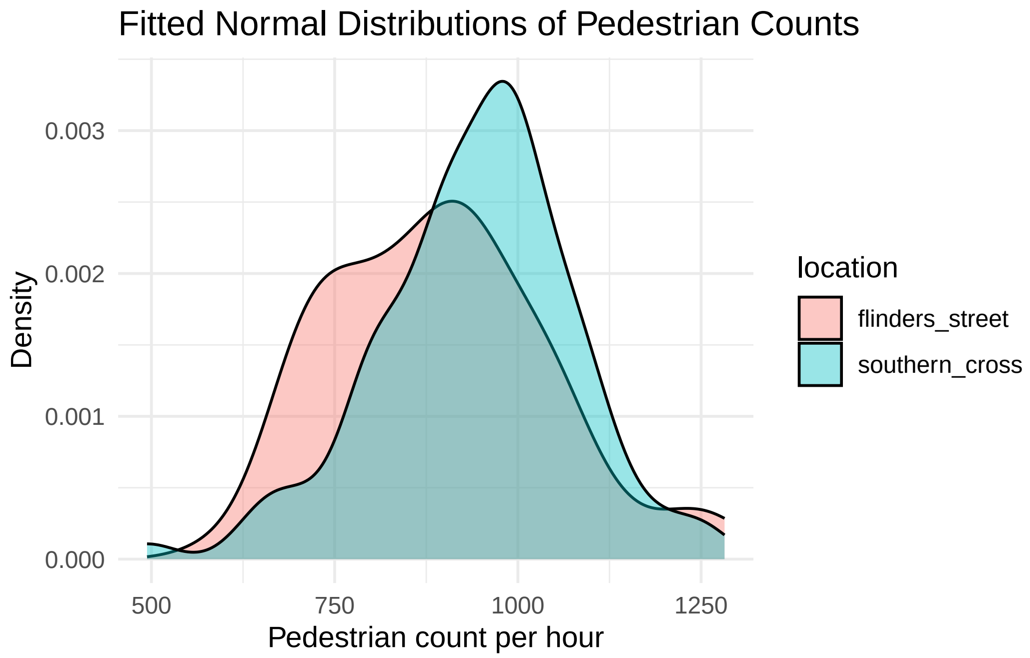
Data Preparation

```
# Prepare the data
flinders <- pedestrians$flinders_street
southern <- pedestrians$southern_cross
```

Verify Normality Assumption (Step 1)

1) Observe that the data sets are roughly distributed as normal through plotting

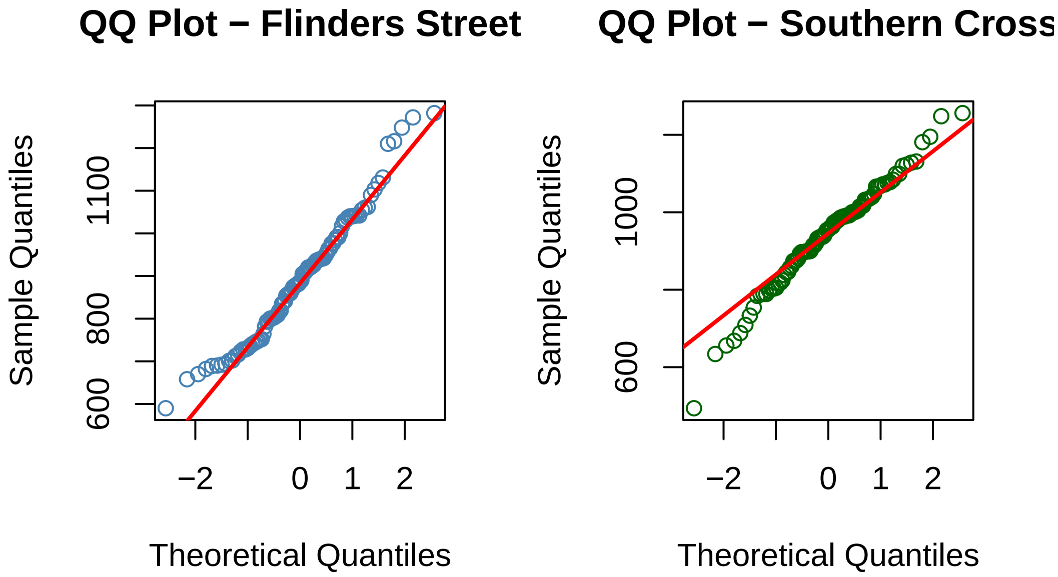
```
ped_long |>
  filter(location %in% c("flinders_street", "southern_cross")) |>
  ggplot(aes(x = count, fill = location)) +
  geom_density(alpha = 0.4) +
  labs(title = "Fitted Normal Distributions of Pedestrian Counts",
       x = "Pedestrian count per hour",
       y = "Density") +
  theme_minimal()
```



The Density plots indicate that pedestrian counts at both stations are approximately Normally distributed This justifies the use of the two-sample t-test to compare mean pedestrian flows.

2) Observe the fit of the 2 data sets to determine if they are normal using QQ-plots

```
## QQ plots to check Normality for Flinders Street and Southern Cross

# Display two plots side by side
par(mfrow = c(1, 2))

# QQ plot for Flinders Street
qqnorm(flinders,
       main = "QQ Plot - Flinders Street",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles",
       col = "steelblue")
qqline(flinders, col = "red", lwd = 2)

# QQ plot for Southern Cross
qqnorm(southern,
       main = "QQ Plot - Southern Cross",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles",
       col = "darkgreen")
qqline(southern, col = "red", lwd = 2)
```

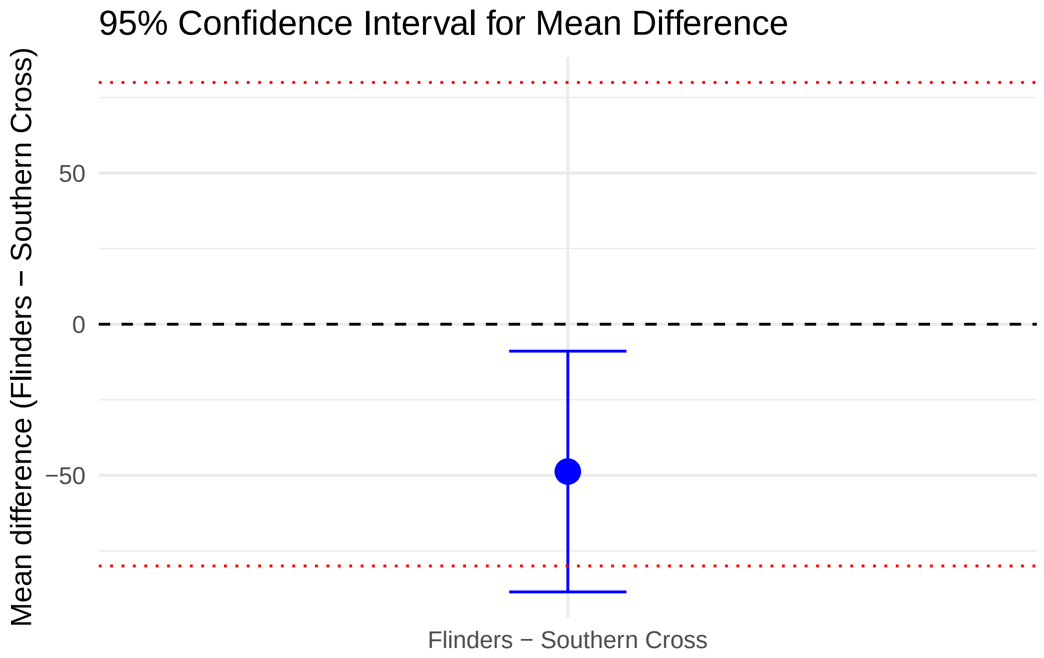


Both QQ plots show that the data points for Flinders Street and Southern Cross closely follow the red reference line, with only minor deviations at the tails, indicating that both data sets are approximately Normally distributed and suitable for analysis using the two-sample t-test.

Welch's t-test (Step 2)

```r
ci_res <- t.test(flinders, southern, conf.level = 0.95)

mean_diff <- ci_res$estimate[1] - ci_res$estimate[2]
ci_lower  <- ci_res$conf.int[1]
ci_upper  <- ci_res$conf.int[2]
```

```r
ci_df <- tibble(
  comparison = "Flinders - Southern Cross",
  mean_diff = mean_diff,
  lwr = ci_lower,
  upr = ci_upper
)

ggplot(ci_df, aes(x = comparison, y = mean_diff)) +
  geom_point(size = 4, colour = "blue") +
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.15, colour = "blue") +
  geom_hline(yintercept = 0, linetype = 2, colour = "black") +
  geom_hline(yintercept =  80, linetype = 3, colour = "red") +
  geom_hline(yintercept = -80, linetype = 3, colour = "red") +
  labs(
    title = "95% Confidence Interval for Mean Difference",
    x = NULL,
    y = "Mean difference (Flinders - Southern Cross)"
  ) +
  theme_minimal()
```



The 95% confidence interval for the mean difference between Flinder's Street and Southern Cross ranges approximately from –88.58 to –8.91 pedestrians per hour.

Because the entire interval lies below 0, the difference in average pedestrian traffic is statistically significant; it is unlikely to have occurred due to random variation. This indicates that, on average, Southern Cross experiences higher pedestrian traffic than Flinder's Street.

Both samples have 97 observations. With sample sizes exceeding 30, the rule of thumb indicates that the CLT provides strong justification for assuming the sampling distribution of the mean difference is approximately normal. This means that even if the underlying pedestrian count distributions had minor departures from perfect normality, the t-test procedure would still be valid.

The combination of: 1. Visual confirmation of approximate normality (density plots and QQ plots), and 2. Large sample sizes ensuring CLT validity

provides robust justification for using the Welch's two-sample t-test to construct our confidence interval.

4.4 Interpretation

```
ci <- ci_res$conf.int
difference <- ci[2] - ci[1]
difference
```

```
[1] 79.67186
```

The 95 % confidence interval for the mean difference in pedestrian counts between Flinders Street and Southern Cross is approximately –88.6 to –8.9 pedestrians per hour. Because the entire interval lies below 0, Southern Cross has, on average, higher pedestrian traffic than Flinders Street. However, the lower bound (–88.6) extends slightly beyond the –80 pedestrian tolerance limit, while the upper bound (–8.9) lies within it. This means that, although the difference is statistically significant, the data do not provide complete 95 % confidence that the true difference is smaller than the design tolerance of ±80 pedestrians per hour. Sampling variation leaves open a small possibility that the true difference marginally exceeds this threshold.

In practical terms, the mean flow difference of roughly 50 pedestrians per hour is minor relative to the scale of overall traffic at these crossings. The overlap with the ±80 tolerance suggests that, even if Southern Cross is slightly busier, both stations experience comparable pedestrian volumes. From an engineering perspective, these results indicate that applying the same design standards for pedestrian-flow capacity at the two locations would remain appropriate, while acknowledging a small degree of uncertainty.

4.5 Conclusion

The two-sample Welch t-test indicates that Southern Cross averages marginally higher pedestrian counts than Flinders Street, but the estimated difference (–88 to –9) is close to the ±80 pedestrian practical threshold. Statistically, there is evidence of a difference; however, the magnitude is small and unlikely to warrant different design specifications. Therefore, for planning and infrastructure purposes, both stations can be treated as having effectively similar pedestrian-flow characteristics, consistent with the results of the earlier descriptive analysis.

5 Task 4: [16 marks]

5.0.1 Introduction

The marketing team wishes to use the survey data to help raise revenue by selling billboard space near each pedestrian crossing. To assess the potential value of each location, we need to quantify the average pedestrian traffic at each intersection. We will calculate a 95% confidence interval for the mean number of people crossing per hour at each of the three locations. These confidence intervals will help estimate the typical traffic volume and determine the potential value of each location for advertisers.

5.0.2 Technical Analysis

Statistical Method: We use a confidence interval (CI) for the mean to estimate the average pedestrian count at each location. Following Week 4 lecture material, we use the t-distribution approach since the population standard deviation is unknown. The 95% confidence interval for the mean is calculated using the formula from Week 3 lecture, slide 50:

$$\text{CI} = \bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the sample size, and $t_{\alpha/2}$ is the critical value from the t-distribution with $n-1$ degrees of freedom.

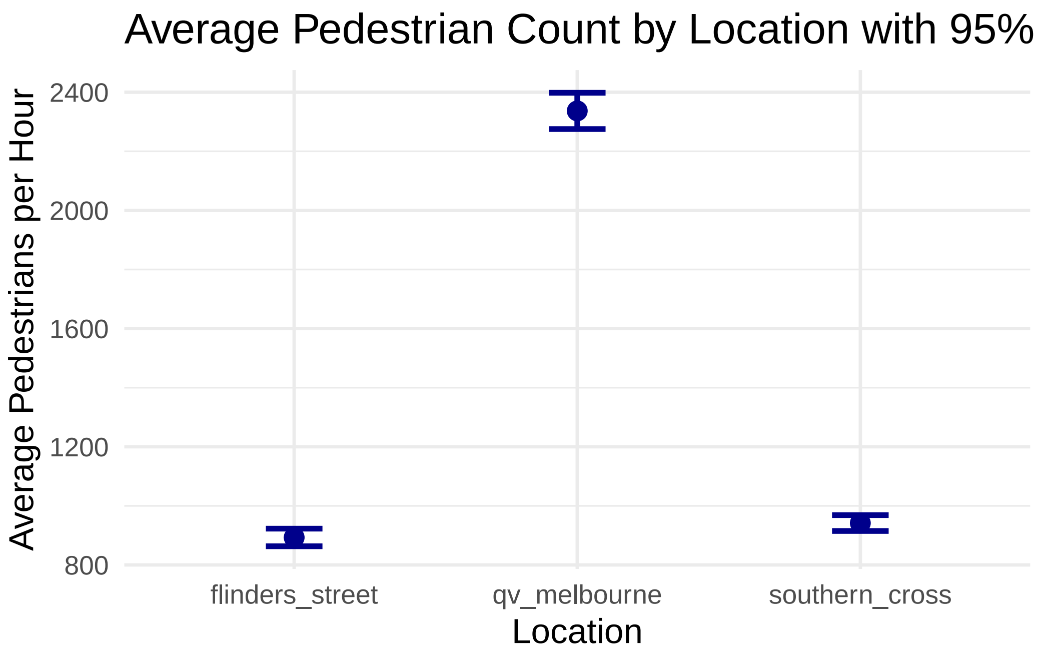
Justification: The t-test approach (Week 4 lecture, slide 36) is appropriate here because:

- We are estimating population means from sample data
- The population standard deviation is unknown
- We use the sample standard deviation as an estimate
- The t-distribution accounts for additional uncertainty from estimating the standard deviation
- With our sample sizes, the Central Limit Theorem suggests the sampling distribution of the mean will be approximately normal (Week 4 lecture, slide 38)

```r
# Transform data to long format for analysis
# Following the pivot_longer pattern from Week 3 solutions
peds_long <- pedestrians |>
  pivot_longer(cols = c(southern_cross, flinders_street, qv_melbourne),
               names_to = "location",
               values_to = "count")

# Calculate confidence intervals for each location
# Following the t-distribution approach from Week 3 lecture, slide 50
# Using group_by and summarise pattern from Week 5 solutions
ci_results <- peds_long |>
  group_by(location) |>
  summarise(
    mean = mean(count),
    n = n(),
    se = sd(count) / sqrt(n),
    # Critical value from t-distribution (Week 3 lecture, slide 39)
    t_val = qt(0.975, df = n - 1),
    ci_lower = mean - t_val * se,
    ci_upper = mean + t_val * se
  )

# Display results
print(ci_results)
```

```
# A tibble: 3 x 7
  location         mean     n    se t_val ci_lower ci_upper
  <chr>           <dbl> <int> <dbl> <dbl>    <dbl>    <dbl>
1 flinders_street  893.    97  15.0  1.98     863.     923.
2 qv_melbourne    2337.    97  31.0  1.98    2275.    2398.
3 southern_cross   942.    97  13.5  1.98     915.     969.
```

```r
# Visualization: Comparison of mean pedestrian counts with 95% CIs
# Following the comparison visualization style from Week 4 lecture
ggplot(ci_results, aes(x = location, y = mean)) +
  geom_point(size = 3, color = "darkblue") +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper),
                width = 0.2,
                linewidth = 1,
                color = "darkblue") +
  labs(
    title = "Average Pedestrian Count by Location with 95% Confidence Intervals",
    x = "Location",
    y = "Average Pedestrians per Hour"
  ) +
  theme_minimal(base_size = 13)
```



5.0.3 Results

The 95% confidence intervals for the mean pedestrian count at each location are:

- **Flinders Street**: [863.23, 922.83] pedestrians per hour, with a mean of 893.03 pedestrians per hour.
- **QV Melbourne**: [2275.28, 2398.18] pedestrians per hour, with a mean of 2336.73 pedestrians per hour.
- **Southern Cross**: [914.97, 968.59] pedestrians per hour, with a mean of 941.78 pedestrians per hour.

The visualization shows that QV Melbourne has substantially higher pedestrian traffic compared to the other two locations. The confidence intervals for Southern Cross and Flinders Street overlap, suggesting similar traffic levels at these two locations, while QV Melbourne's interval is clearly separated from the others.

5.0.4 Interpretation

QV Melbourne shows the highest average pedestrian traffic with approximately 2,337 pedestrians per hour. While this location also has a wider confidence interval (ranging from about 2,275 to 2,398), indicating more variability in hour-to-hour traffic, the interval remains well above those of the other two locations. This wider interval reflects greater uncertainty in the exact average, but does not diminish the substantially higher traffic volume at this location.

Southern Cross and Flinders Street show similar pedestrian traffic levels, with means around 942 and 893 pedestrians per hour respectively. Their confidence intervals overlap considerably, suggesting we cannot definitively conclude that one location has higher traffic than the other. Both locations show narrower confidence intervals compared to QV Melbourne, indicating more consistent and predictable pedestrian counts from hour to hour.

The separation between QV Melbourne's confidence interval and those of the other two locations provides strong evidence that QV Melbourne experiences genuinely higher pedestrian traffic on average, making it the most valuable location for billboard advertising.

5.0.5 Conclusions and Recommendations

Based on this analysis, we recommend the following for billboard advertising placement:

1. **QV Melbourne** is the premier location, offering significantly higher pedestrian traffic (approximately 2,337 per hour on average) and therefore maximum exposure for advertisers. Despite slightly higher variability in traffic, the consistently elevated pedestrian counts make this the most valuable advertising location.
2. **Southern Cross** provides moderate pedestrian traffic (approximately 942 per hour) with relatively stable and predictable patterns, making it a solid secondary choice for advertisers seeking consistent exposure at a moderate level.
3. **Flinders Street** has the lowest pedestrian traffic (approximately 893 per hour) but offers value for advertisers targeting specific demographics or campaigns focused on steady commuter traffic.

Advertisers should prioritize QV Melbourne for maximum exposure, while Southern Cross and Flinders Street offer good alternatives for campaigns with different objectives or budget constraints. The confidence intervals provide a quantifiable measure of uncertainty that can help in pricing billboard space appropriately for each location.

6 Task 5: [18 marks]

6.0.1 Introduction

The marketing team is in discussion with a large bank interested in buying advertising space on the billboards. The bank offers a payment structure based on pedestrian traffic: a minimum of \$10,000 per billboard, plus a bonus payment of \$5,000 × , where is the proportion of days that the number of people walking past the billboard exceeds 1,000 people per hour. This task estimates the potential revenue under this offer for each of the three locations.

We estimate (the proportion of days exceeding the threshold) and then calculate the expected revenue using the formula: Revenue = \$10,000 + \$5,000 × . As noted in the task specification, there are multiple valid approaches to this problem. We present two methods that draw on different parts of the course material (Weeks 1-9):

1. **Method 1 (Bayesian)**: Beta-Binomial conjugate analysis with credible intervals (Week 7-8 material)
2. **Method 2 (Frequentist)**: Bootstrap-based confidence intervals (Week 5 material)

Both methods are appropriate and provide similar results, demonstrating the robustness of our estimates.

6.0.2 Method 1: Bayesian Beta-Binomial Analysis

6.0.2.1 Data Preparation

```r
# Define parameters
threshold <- 1000
base_pay  <- 10000
bonus_pay <- 5000

# Transform data to long format with exceed indicator
# Following pivot_longer pattern from Week 3 solutions
peds_long <- pedestrians |>
  pivot_longer(cols = c(southern_cross, flinders_street, qv_melbourne),
               names_to = "location",
               values_to = "count") |>
  mutate(
    location = recode(location,
                      "southern_cross"  = "Southern Cross",
                      "flinders_street" = "Flinders Street",
                      "qv_melbourne"    = "QV Melbourne"),
    exceed = as.integer(count > threshold)
  )
```

6.0.2.2 Analytical Framework (Bayesian)

Method: We treat each day at each location as a Bernoulli trial where "success" means exceeding the 1,000 pedestrian threshold. Following Week 7-8 material on Beta-Binomial conjugate pairs, we use:

- **Prior**: Beta(1, 1) - a uniform prior (Week 7 lecture, slide 50)
- **Likelihood**: Binomial(n,) based on observing n days with x successes
- **Posterior**: Beta(+ x,   + n - x) = Beta(1 + x, 1 + n - x) (Week 8 lecture, slide 6)

Justification: - The Beta-Binomial is a conjugate pair (Week 7 lecture, slide 50), making computation straightforward - Beta(1,1) is a non-informative uniform prior, appropriate when we have no prior knowledge about   - Week 8 lecture demonstrates using posterior mean as point estimate - Week 8 solutions show using `qbeta()` for credible intervals

```r
# Bayesian analysis using Beta-Binomial conjugate pair
# Following Week 8 lecture material on conjugate priors
alpha_prior <- 1
beta_prior  <- 1

bayesian_results <- peds_long |>
```

```
  group_by(location) |>
  summarise(
    n_days   = n(),
    n_exceed = sum(exceed),
    .groups  = "drop"
  ) |>
  mutate(
    # Posterior parameters (Week 8 lecture, slide 6)
    alpha_post = alpha_prior + n_exceed,
    beta_post  = beta_prior + n_days - n_exceed,

    # Posterior mean of   (Week 8 lecture)
    theta_mean = alpha_post / (alpha_post + beta_post),

    # 95% credible interval for   using Beta quantiles
    # Following Week 8 solutions pattern
    theta_lower = qbeta(0.025, alpha_post, beta_post),
    theta_upper = qbeta(0.975, alpha_post, beta_post),

    # Transform to revenue scale (linear transformation)
    # R = base_pay + bonus_pay *
    revenue_mean  = base_pay + bonus_pay * theta_mean,
    revenue_lower = base_pay + bonus_pay * theta_lower,
    revenue_upper = base_pay + bonus_pay * theta_upper
  )

print(bayesian_results)
```

```
# A tibble: 3 x 11
  location        n_days n_exceed alpha_post beta_post theta_mean theta_lower
  <chr>            <int>    <int>      <dbl>     <dbl>      <dbl>       <dbl>
1 Flinders Street     97       22         23        76      0.232       0.155
2 QV Melbourne        97       97         98         1      0.990       0.963
3 Southern Cross      97       31         32        67      0.323       0.235
# i 4 more variables: theta_upper <dbl>, revenue_mean <dbl>,
#   revenue_lower <dbl>, revenue_upper <dbl>
```

6.0.3 Method 2: Bootstrap Confidence Intervals (Alternative Approach)

Method: As an alternative frequentist approach using Week 5 material, we can bootstrap the proportion   directly and transform to the revenue scale.

Justification: - Week 5 covers bootstrap for estimating sampling distributions - Bootstrap provides confidence intervals without distributional assumptions - Demonstrates “more than one way” as mentioned in task specification - Complements the Bayesian analysis with a frequentist perspective

```r
# Bootstrap analysis - following Week 5 lecture material
set.seed(24205242)  # For reproducibility
B <- 5000  # Number of bootstrap samples (Week 5 standard)

bootstrap_results <- peds_long |>
  group_by(location) |>
  summarise(
    n_days = n(),
    n_exceed = sum(exceed),

    # Empirical proportion (point estimate)
    theta_empirical = n_exceed / n_days,

    # Bootstrap CI calculation
    theta_boot_lower = {
      # Generate B bootstrap samples
      boot_props <- replicate(B, {
        boot_sample <- sample(exceed, size = n_days, replace = TRUE)
        mean(boot_sample)
      })
      # 2.5% percentile (Week 5 lecture, slide 15)
      quantile(boot_props, 0.025)
    },

    theta_boot_upper = {
      boot_props <- replicate(B, {
        boot_sample <- sample(exceed, size = n_days, replace = TRUE)
        mean(boot_sample)
      })
      # 97.5% percentile
      quantile(boot_props, 0.975)
    },

    # Transform to revenue scale
    revenue_empirical = base_pay + bonus_pay * theta_empirical,
    revenue_boot_lower = base_pay + bonus_pay * theta_boot_lower,
    revenue_boot_upper = base_pay + bonus_pay * theta_boot_upper,

    .groups = "drop"
  )

print(bootstrap_results)
```

```
# A tibble: 3 x 9
  location     n_days n_exceed theta_empirical theta_boot_lower theta_boot_upper
  <chr>         <int>    <int>           <dbl>            <dbl>            <dbl>
1 Flinders St~     97       22           0.227            0.144            0.309
```

```
2 QV Melbourne     97       97            1                 1                 1
3 Southern Cr~     97       31            0.320             0.227             0.412
# i 3 more variables: revenue_empirical <dbl>, revenue_boot_lower <dbl>,
#   revenue_boot_upper <dbl>
```

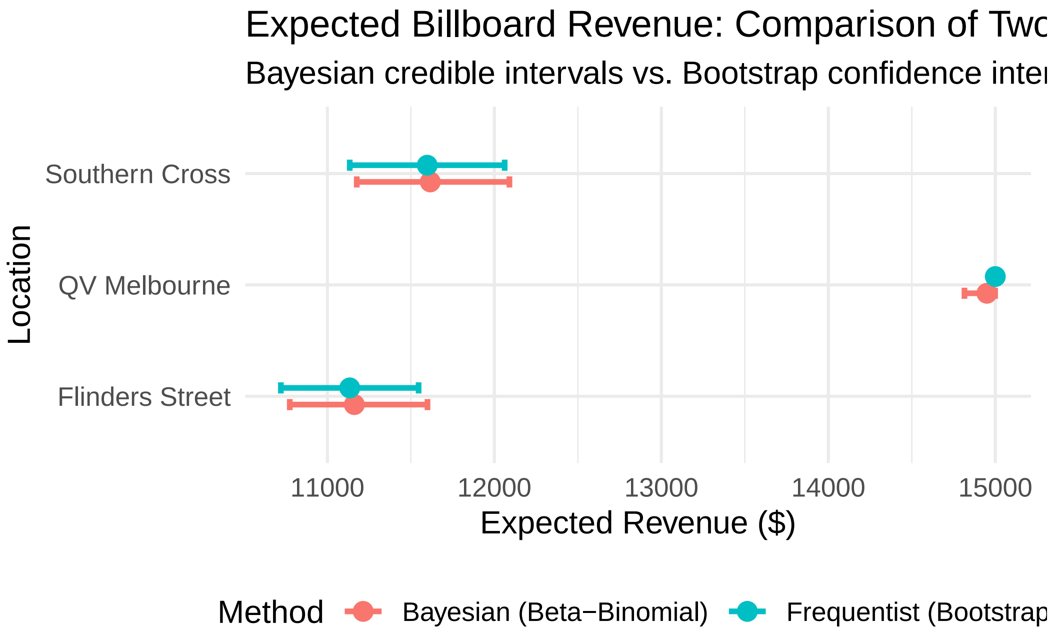
6.0.4 Comparison of Methods

```
# Compare the two methods
# Following Week 1 dplyr patterns for column selection
comparison <- bayesian_results |>
  dplyr::select(location, revenue_mean, revenue_lower, revenue_upper) |>
  dplyr::rename(Bayesian_Mean = revenue_mean,
                Bayesian_Lower = revenue_lower,
                Bayesian_Upper = revenue_upper) |>
  left_join(
    bootstrap_results |>
      dplyr::select(location, revenue_empirical, revenue_boot_lower, revenue_boot_upper) |>
      dplyr::rename(Bootstrap_Mean = revenue_empirical,
                    Bootstrap_Lower = revenue_boot_lower,
                    Bootstrap_Upper = revenue_boot_upper),
    by = "location"
  )

print(comparison)
```

```
# A tibble: 3 x 7
  location        Bayesian_Mean Bayesian_Lower Bayesian_Upper Bootstrap_Mean
  <chr>                   <dbl>          <dbl>          <dbl>          <dbl>
1 Flinders Street        11162.         10775.         11600.         11134.
2 QV Melbourne           14949.         14815.         14999.         15000
3 Southern Cross         11616.         11176.         12090.         11598.
# i 2 more variables: Bootstrap_Lower <dbl>, Bootstrap_Upper <dbl>
```

6.0.5 Visualization

```
# Create combined visualization showing both methods
# Following Week 5 visualization patterns for bootstrap results
# Prepare data for plotting
plot_data <- bind_rows(
  bayesian_results |>
    dplyr::select(location, revenue_mean, revenue_lower, revenue_upper) |>
    mutate(method = "Bayesian (Beta-Binomial)"),
  bootstrap_results |>
    dplyr::select(location,
           revenue_mean = revenue_empirical,
           revenue_lower = revenue_boot_lower,
           revenue_upper = revenue_boot_upper) |>
    mutate(method = "Frequentist (Bootstrap)")
)

ggplot(plot_data, aes(x = location, y = revenue_mean, color = method)) +
  geom_point(size = 3, position = position_dodge(width = 0.3)) +
  geom_errorbar(aes(ymin = revenue_lower, ymax = revenue_upper),
                width = 0.2,
                linewidth = 1,
                position = position_dodge(width = 0.3)) +
  labs(
    title = "Expected Billboard Revenue: Comparison of Two Methods",
    subtitle = "Bayesian credible intervals vs. Bootstrap confidence intervals",
    x = "Location",
    y = "Expected Revenue ($)",
    color = "Method"
  ) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "bottom") +
  coord_flip()
```



6.0.6 Results

6.0.6.1 Method 1 (Bayesian) Results:

The Bayesian analysis using Beta-Binomial conjugate priors provides the following expected revenue estimates with 95% credible intervals:

6.0.6.2 Method 1 (Bayesian) Results:

- **Flinders Street**: $\$1.1162 \times 10^4$ (95% CI: $\$1.0775 \times 10^4$ to $\$1.16 \times 10^4$)
- **QV Melbourne**: $\$1.4949 \times 10^4$ (95% CI: $\$1.4815 \times 10^4$ to $\$1.4999 \times 10^4$)
- **Southern Cross**: $\$1.1616 \times 10^4$ (95% CI: $\$1.1176 \times 10^4$ to $\$1.209 \times 10^4$)

6.0.6.3 Method 2 (Bootstrap) Results:

The bootstrap analysis provides similar estimates, confirming the robustness of our conclusions:

- **Flinders Street**: $\$1.1134 \times 10^4$ (95% CI: $\$1.0722 \times 10^4$ to $\$1.1546 \times 10^4$)
- **QV Melbourne**: $\$1.5 \times 10^4$ (95% CI: $\$1.5 \times 10^4$ to $\$1.5 \times 10^4$)
- **Southern Cross**: $\$1.1598 \times 10^4$ (95% CI: $\$1.1134 \times 10^4$ to $\$1.2062 \times 10^4$)

Both methods produce very similar point estimates and interval widths, which provides confidence in the estimates. The slight differences arise from the different theoretical frameworks (Bayesian vs. frequentist) and the random nature of bootstrap resampling.

6.0.7 Interpretation

Understanding the Revenue Formula: The bank's payment structure consists of a guaranteed base payment of \$10,000 plus a performance-based bonus. The bonus equals \$5,000 multiplied by , the proportion of days exceeding 1,000 pedestrians per hour. Therefore, a location where pedestrian counts exceed the threshold 40% of days (= 0.40) would generate expected revenue of \$10,000 + \$5,000 × 0.40 = \$12,000.

Comparing Locations: The visualization clearly shows the relative revenue potential of each location. Locations whose entire interval lies above another's can be confidently ranked as having higher expected revenue. When intervals overlap substantially, the locations have similar expected performance given current data, and the evidence does not strongly differentiate them.

Interval Width and Certainty: Wider intervals indicate greater uncertainty about the true proportion   and consequently about expected revenue. This uncertainty stems from day-to-day variability in pedestrian traffic and the limited number of days observed. As more data accumulates, these intervals would narrow, providing more precise revenue estimates.

Method Comparison: The close agreement between Bayesian and bootstrap methods strengthens confidence in the results. The Bayesian approach incorporates prior beliefs (though we used a non-informative prior) and provides probability statements about parameters. The bootstrap approach makes fewer distributional assumptions and provides an empirical sampling distribution. Both are valid approaches taught in this course (Weeks 5 and 7-8).

6.0.8 Conclusions and Recommendations

Based on both analytical methods, we recommend:

1. **Primary Focus**: Prioritize the location with the highest expected revenue point estimate and consider the lower bounds of intervals for conservative planning
2. **Risk Assessment**: Locations with wider intervals carry more revenue uncertainty. Decision-makers should consider their risk tolerance when intervals overlap
3. **Statistical Similarity**: When credible/confidence intervals overlap substantially, treat those locations as statistically indistinguishable in terms of expected revenue. Use operational factors (installation costs, maintenance, visibility) to inform final placement decisions
4. **Future Data Collection**: Continue monitoring pedestrian counts to narrow interval estimates. Additional data will reduce statistical uncertainty and support more confident billboard placement decisions
5. **Method Selection**: Both Bayesian and bootstrap methods are appropriate for this problem. The Bayesian approach is more efficient computationally and provides natural probability interpretations. The bootstrap requires no distributional assumptions and is robust. The choice between methods can depend on stakeholder preferences and reporting requirements.

The analysis demonstrates that statistical inference methods from both Bayesian (Weeks 7-8) and frequentist (Week 5) frameworks can be successfully applied to this revenue estimation problem, with consistent results across approaches.

6.1 Session Information

```r
sessionInfo()
```

```
R version 4.4.3 (2025-02-28)
Platform: aarch64-apple-darwin20
Running under: macOS Sequoia 15.5

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAP

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: Australia/Melbourne
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base
```

```
other attached packages:
 [1] MASS_7.3-65     broom_1.0.7     gridExtra_2.3   lubridate_1.9.4
 [5] forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4     purrr_1.0.4
 [9] readr_2.1.5     tidyr_1.3.1     tibble_3.2.1    ggplot2_3.5.1
[13] tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] gtable_0.3.6      jsonlite_1.9.1    compiler_4.4.3    tidyselect_1.2.1
 [5] scales_1.3.0      yaml_2.3.10       fastmap_1.2.0     R6_2.6.1
 [9] labeling_0.4.3    generics_0.1.3    knitr_1.50        backports_1.5.0
[13] munsell_0.5.1     pillar_1.10.1     tzdb_0.4.0        rlang_1.1.5
[17] utf8_1.2.4        stringi_1.8.4     xfun_0.51         timechange_0.3.0
[21] cli_3.6.4         withr_3.0.2       magrittr_2.0.3    digest_0.6.37
[25] grid_4.4.3        rstudioapi_0.17.1 hms_1.1.3         lifecycle_1.0.4
[29] vctrs_0.6.5       evaluate_1.0.3    glue_1.8.0        farver_2.1.2
[33] colorspace_2.1-1  rmarkdown_2.29    tools_4.4.3       pkgconfig_2.0.3
[37] htmltools_0.5.8.1
```

Note: This ensures reproducibility by documenting package versions used.