Estimation, Inference, and Hypothesis Testing

Learning Objectives

By the end of this chapter, you should be able to:

- Explain sampling distributions and their role in inference, including Type I and Type II errors.
- Construct and interpret confidence intervals.
- Formulate and test hypotheses using t-tests and F-tests.
- Distinguish between statistical and practical significance.
- Apply inference to business decision-making contexts.

Introduction: Why Inference Matters

In Chapter @ref(ch-framework), we established that predictive analytics aims to transform data into decisions. In Chapter @ref(ch-prob-stats), we learned the probability foundations that describe uncertainty. Now, we bridge these concepts: **how do we make confident decisions from uncertain data?** This chapter introduces **statistical inference** – the process of drawing conclusions about populations from samples and quantifying the uncertainty in those conclusions.

Consider a real-world situation. Imagine you are an operations manager at Ember, a coach service based in Scotland aiming to offer a more environmentally sustainable mode of transport. The company advertises that the Dundee to Braemar route takes 2 hours 15 minutes (135 minutes). However, over the past month you have noticed journey times seem more variable than expected. A sample of 100 recent trips shows an average of 142 minutes with complaints about late arrivals.

The critical question is not just "Did we see a difference?" Rather, you need to know:

1. How confident can we be that this 7-minute difference reflects a real schedule problem, not just random variation?
2. What range of journey times should we expect, and should we adjust our published schedule?
3. Is a 7-minute average delay practically meaningful?

These questions require **statistical inference** – the tools we develop in this chapter.

Why inference matters for this example: Inference gives us the tools to test whether route times genuinely deviate from published schedules, to quantify journey time variability for planning purposes, to compare different routes, times of day, or vehicle types, and to prioritise improvements based on statistical evidence rather than anecdote.

Sampling Distributions

Before diving into hypothesis tests, we need to understand the theoretical foundation: **sampling distributions**.

The Sampling Distribution Concept

When we calculate a statistic from a sample (like the mean journey time), that statistic is itself a random variable. It varies from sample to sample. The **sampling distribution** describes how that statistic behaves across all possible samples.

Example: If we repeatedly sampled 30 Ember journeys and calculated the mean each time, those means would form a distribution – the sampling distribution of the sample mean. Below, we vary the sample size (e.g., 30 vs. 100) to illustrate how sampling distributions tighten as n increases.

Key Result: Distribution of the Sample Mean

From Chapter @ref(ch-prob-stats), the Central Limit Theorem tells us that for independent observations $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{approximately, for large } n$$

More precisely (by the Central Limit Theorem, under i.i.d. sampling):

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Implications: First, sampling variability is predictable — even though individual samples vary, we know *how* they vary. Second, the standard error decreases with $\sqrt{n}$, so doubling precision requires about four times the sample size. Third, this predictable behaviour is what makes confidence intervals possible.

Application: Ember Journey Time Uncertainty

Suppose the true population of Dundee–Braemar journey times has mean $\mu = 140$ minutes and standard deviation $\sigma = 15$ minutes.

If we sample $n = 100$ journeys, the sample mean $\bar{X}$ is approximately:

$$\bar{X} \sim N\left(140, \frac{15^2}{100}\right) = N(140, 2.25)$$

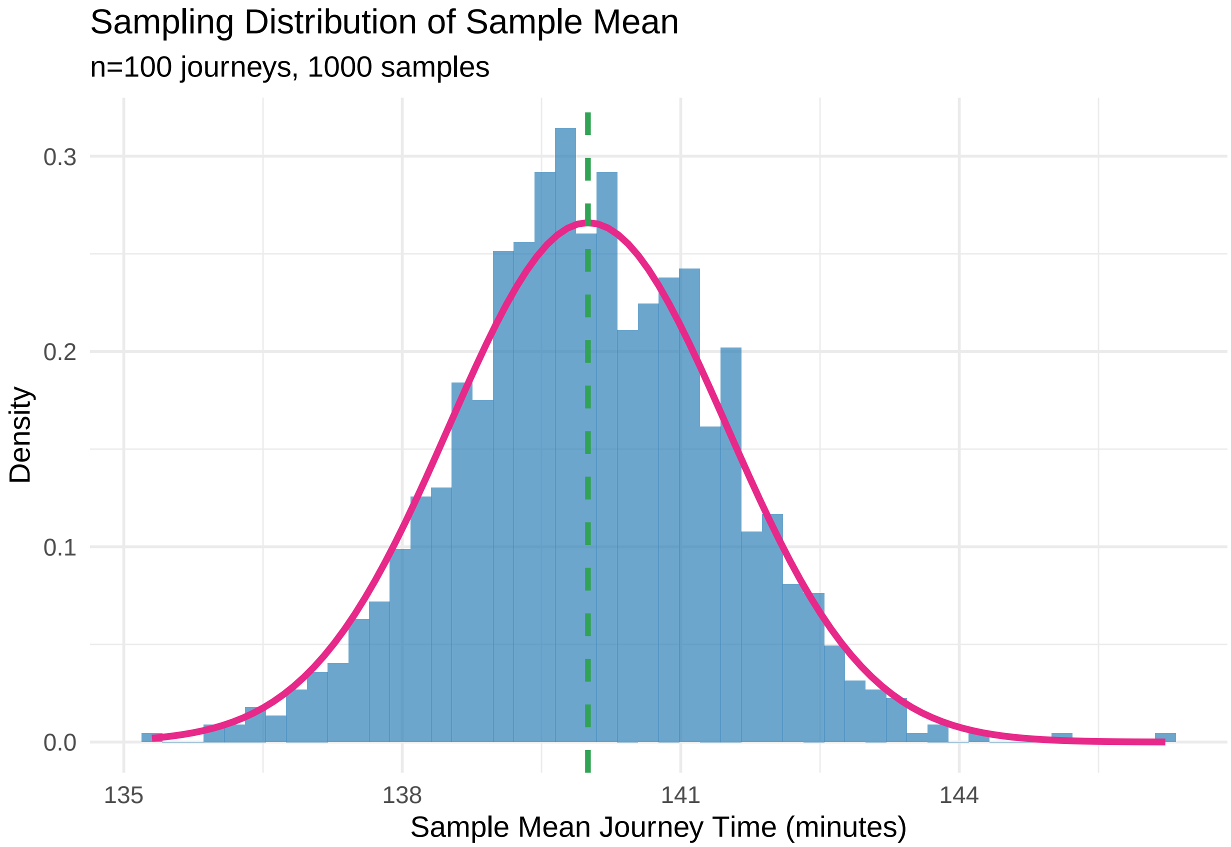
So $SE(\bar{X}) = 15/\sqrt{100} = 1.5$ minutes.

Insight: With 100 observations, our estimate of average journey time will typically be within ±3 minutes (2 standard errors) of the true average. This quantifies our precision for operational planning.

```
library(ggplot2)
library(dplyr)
# Demonstrate sampling distribution
set.seed(123)
true_mu <- 140
true_sigma <- 15
n_journeys <- 100
n_samples <- 1000

# Generate 1000 sample means
sample_means <- replicate(n_samples, {
  journeys <- rnorm(n_journeys, mean = true_mu, sd = true_sigma)
  mean(journeys)
})

# Compare empirical vs theoretical
data.frame(sample_mean = sample_means) %>%
  ggplot(aes(x = sample_mean)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50,
                 fill = "#2C7FB8", alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = true_mu, sd = true_sigma/sqrt(n_journeys)),
                color = "#E7298A", linewidth = 1.2) +
  geom_vline(xintercept = true_mu, linetype = "dashed",
             color = "#31a354", linewidth = 1) +
  labs(title = "Sampling Distribution of Sample Mean",
       subtitle = paste0("n=", n_journeys, " journeys, ", n_samples, " samples"),
       x = "Sample Mean Journey Time (minutes)",
       y = "Density") +
  theme_minimal()
```



```
data.frame(
  Measure = c("Theoretical SE", "Empirical SE"),
  Value = round(c(true_sigma / sqrt(n_journeys), sd(sample_means)), 3)
)
```

```
##          Measure Value
## 1 Theoretical SE 1.500
## 2   Empirical SE 1.426
```

The t-Distribution: When is Unknown

In practice, we rarely know the population standard deviation σ. Instead, we estimate it with the sample standard deviation s.

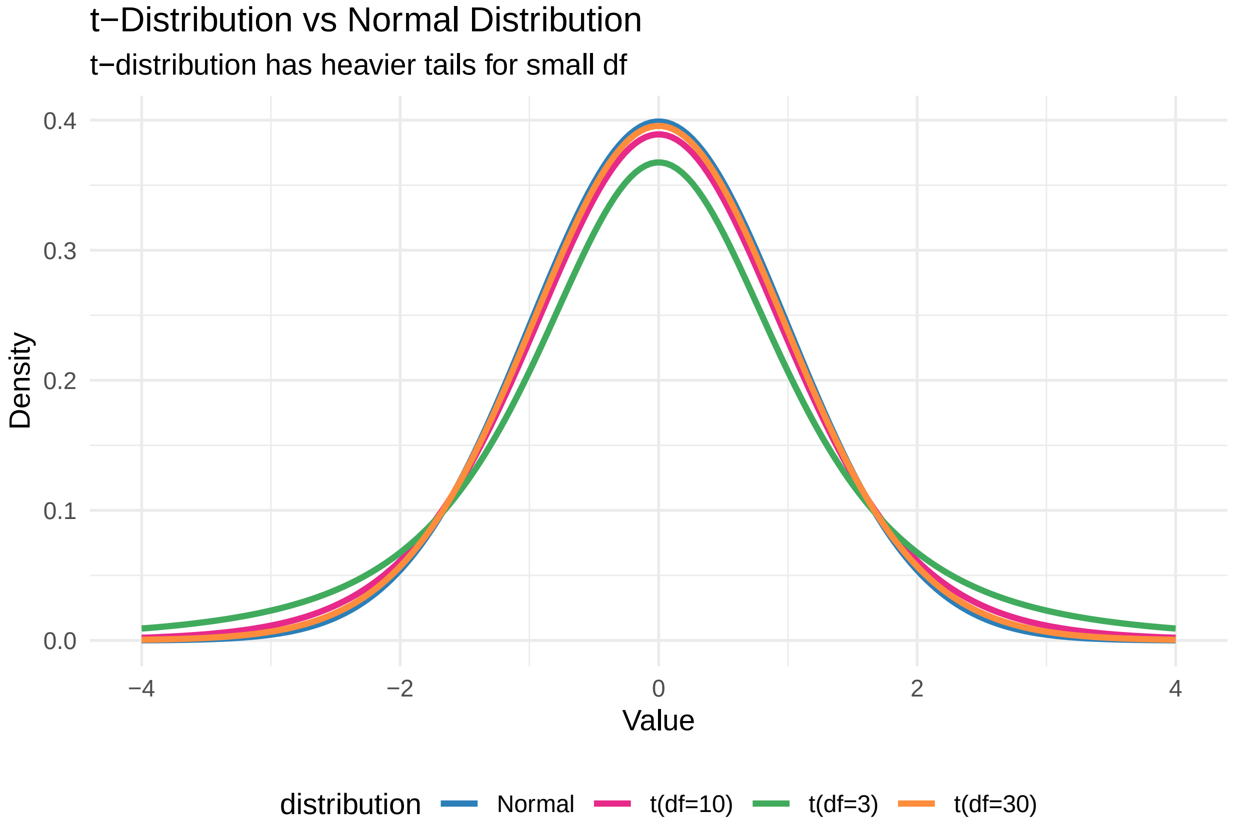
When we replace σ with s, the distribution changes from normal to **t-distribution**:

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Because s is itself random, the standardised statistic has extra uncertainty compared with the case where σ is known.

The t-distribution has heavier tails than the normal, placing more probability in the extremes. It depends on the **degrees of freedom** ($df = n - 1$) and approaches the normal as $n \to \infty$.

For small samples, this difference is substantial; for large samples ($n > 30$), the t-distribution is close enough to the normal that the distinction is minor.



Point and Interval Estimation

The Problem: Population Parameters from Samples

We rarely have access to complete populations. Instead, we work with samples. In this context, the **population** is all possible journey times on the Dundee–Braemar route — thousands of past and future trips — while the **sample** is the 100 observed journeys from the past three months.

Whether the sample is representative is crucial, but outside the scope of this book. For example, was the journey data collected across a full year or only during one season (e.g., winter)?

If the sample is representative, we can use it to estimate population parameters such as the average journey time (μ), the proportion of journeys completing within schedule (p), and the standard deviation of journey times (σ).

Point Estimation

A **point estimator** is a function of sample data used to approximate an unknown population parameter.

For example, the **sample mean** estimates the population mean:

$$\hat{\mu} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example: Average Journey Time

Suppose Ember samples 100 recent Dundee-Braemar trips and finds $\bar{x} = 142.3$ minutes. This is our point estimate for average journey time across all trips on this route.

```r
# Simulated journey time data (in minutes)
set.seed(42)
journey_times <- rnorm(100, mean = 142, sd = 15)
mean_time <- mean(journey_times)
round(mean_time, 1)
```

```
## [1] 142.5
```

The limitation: A point estimate provides no information about uncertainty. Did we get lucky with this sample, or is 142.3 minutes representative?

Interval Estimation: Quantifying Uncertainty

A **confidence interval** provides a range of plausible values for the population parameter, constructed so that it contains the true parameter with specified probability.

For a mean with known variance, a $100(1-\alpha)\%$ confidence interval is:

$$\bar{X}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

When σ is unknown (the usual case), we use the sample standard deviation s and a t-distribution:

$$\bar{X}_n \pm t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Interpretation:

> If we repeated this sampling procedure many times, about 95% of the resulting intervals would contain the true mean. The parameter is fixed; the interval is random.

This does **not** mean "there is a 95% probability the true mean is in this interval." The true mean is a fixed quantity; it is the interval that is random.

Continuing the Journey Time Example:

```r
ci_result <- t.test(journey_times, conf.level = 0.95)
round(ci_result$conf.int, 1)
```

```
## [1] 139.4 145.6
## attr(,"conf.level")
## [1] 0.95
```

Interpretation: The 95% confidence interval for the true average journey time lies entirely above the published schedule of 135 minutes. Even the lower bound of the interval exceeds 135, suggesting we need to investigate delays and potentially adjust the timetable.

Sample Size and Precision

Notice the width of the confidence interval depends on $\frac{s}{\sqrt{n}}$ – the **standard error**.

Points to remember: Larger samples produce narrower intervals and more precise estimates, but doubling precision requires quadrupling the sample size. Conversely, very large samples may yield statistically precise differences that are practically meaningless.

```
# Compare confidence intervals for different sample sizes
set.seed(123)
true_mean_time <- 142
true_sd_time <- 15

sample_sizes <- c(50, 100, 200, 500)
ci_widths <- sapply(sample_sizes, function(n) {
  sample_data <- rnorm(n, true_mean_time, true_sd_time)
  ci <- t.test(sample_data)$conf.int
  diff(ci)
})

data.frame(
  SampleSize = sample_sizes,
  CI_Width_Minutes = round(ci_widths, 2),
  Improvement = c(NA, round(ci_widths[1] / ci_widths[-1], 2))
)
```

```
##   SampleSize CI_Width_Minutes Improvement
## 1         50             7.89          NA
## 2        100             5.74        1.37
## 3        200             4.03        1.96
## 4        500             2.64        2.99
```

Hypothesis Testing: The Framework

Hypothesis testing formalises the process of making decisions under uncertainty and forms the cornerstone of frequentist statistics. The framework begins with a **null hypothesis** (H_0), which represents the "status quo" or "no effect" position, and an **alternative hypothesis** (H_1 or H_a), which is the claim we wish to evaluate. We then compute a **test statistic** that measures how far the data deviate from what H_0 would predict, and from it derive a **p-value** — the probability of observing data at least as extreme as ours, assuming H_0 is true. The **decision rule** is to reject H_0 if the p-value falls below a pre-specified significance level (typically 0.05).

We assess whether the data are sufficiently inconsistent with H_0 to reject it. In frequentist statistics, we do not determine which hypothesis is "true"; instead, we evaluate whether the observed data provide sufficient evidence against the null hypothesis. The p-value is used as a measure of evidence against H_0, and should be interpreted alongside effect sizes and confidence intervals.

The t-Test: Mathematical Foundation

The **t-statistic** is the workhorse of hypothesis testing in regression and comparison of means.

Formula:

$$t = \frac{\hat{\theta} – \theta_0}{SE(\hat{\theta})}$$

where $\hat{\theta}$ is the estimated parameter (e.g., the sample mean), θ_0 is the hypothesised value under H_0 (often 0), and $SE(\hat{\theta})$ is the standard error of the estimate.

For testing a population mean:

$$t = \frac{\bar{X} – \mu_0}{s/\sqrt{n}}$$

Distribution Under H_0:

Under the null hypothesis $H_0 : \mu = \mu_0$, this statistic follows a t-distribution with $n - 1$ degrees of freedom:

$$t \sim t_{n-1}$$

Why t-distribution? Because we're using the sample standard deviation s to estimate the unknown σ. This introduces additional uncertainty captured by the fatter tails of the t-distribution.

As $n \to \infty$: $t_{n-1} \to N(0, 1)$, so for large samples, t-tests and z-tests give similar results.

Application: Testing Ember Route Profitability

Scenario: Route profitability target is £500/day. A sample of 25 days shows a mean of £520 and an SD of £80. Is this significantly above target?

Hypotheses:

- $H_0 : \mu = 500$ (meeting target, not exceeding)
- $H_1 : \mu > 500$ (exceeding target)

Test Statistic:

$$t = \frac{520–500}{80/\sqrt{25}} = \frac{20}{16} = 1.25$$

Critical value for t_{24} at $\alpha = 0.05$ (one-tailed): $t_{0.95,24} \approx 1.711$

Since $t = 1.25 < 1.711$, we fail to reject H_0.

Managerial Insight: While the sample mean exceeds the target, the evidence isn't strong enough to conclude the route consistently exceeds £500/day profitability. We might need more data or accept that we're meeting (not beating) the target.

```r
# Simulate route profitability data
set.seed(2026)
daily_profit <- rnorm(25, mean = 520, sd = 80) # Here we assume the observed data have a mean close to

# Test against target of £500
profit_test <- t.test(daily_profit, mu = 500, alternative = "greater")
profit_test
```

```
##
##  One Sample t-test
##
## data:  daily_profit
## t = -0.41387, df = 24, p-value = 0.6587
## alternative hypothesis: true mean is greater than 500
## 95 percent confidence interval:
##  467.2918      Inf
## sample estimates:
## mean of x
##  493.6289
```

Testing a Route Optimisation

Scenario: Ember has historically averaged 138 minutes for the Dundee–Braemar route. After implementing a new routing system that bypasses Blairgowrie to avoid peak traffic, you observe journey times over 10 trips. Has the new route improved journey times?

Step 1: State hypotheses

- H_0: $\mu = 138$ minutes (new route has no effect)
- H_1: $\mu < 138$ minutes (new route reduces journey time)

Step 2: Collect data and compute test statistic

Journey times (minutes) for 10 trips using the new route yield a sample mean of $\bar{x} = 132.4$ minutes and sample SD of $s = 8.5$ minutes.

```r
# Journey times (minutes) on new route via Blairgowrie
set.seed(22)
new_route_times <- c(135, 128, 140, 125, 138, 131, 134, 126, 132, 135)

# Test against historical average of 138 minutes
test_result <- t.test(new_route_times, mu = 138, alternative = "less")
test_result
```

```
##
##  One Sample t-test
##
## data:  new_route_times
## t = -3.5624, df = 9, p-value = 0.003048
## alternative hypothesis: true mean is less than 138
## 95 percent confidence interval:
##      -Inf 135.2816
## sample estimates:
## mean of x
##     132.4
```

Step 3: Interpret

With p-value = 0.003 $<$ 0.05, we reject H_0 and conclude that the new route bypassing Blairgowrie has reduced journey times. However, we must consider several caveats. Statistical significance does not imply practical significance — is saving 5.6 minutes worth the potential longer distance and fuel costs? The 95% confidence interval provides a lower bound on the average time saving, which helps quantify the effect. And other factors such as weather conditions, traffic patterns, and time-of-day effects would need investigation before making a permanent routing change.

Testing Multiple Hypotheses: F-Tests

Sometimes we want to test multiple restrictions simultaneously. For example, we might ask whether seasonal effects are jointly significant in predicting journey times, whether multiple route characteristics collectively impact profitability, or whether adding weather variables improves forecast accuracy.

The Problem with Multiple t-Tests

If we test 3 coefficients individually at = 0.05, our overall Type I error rate is actually higher than 5%.

Example: Testing whether winter, spring, and fall each significantly affect journey times using 3 separate t-tests at = 0.05:

Probability of at least one false positive = 1–$(1 - 0.05)^3 = 0.143$ (14.3%)!

The F-Test Solution

The **F-test** allows us to test multiple restrictions simultaneously while controlling the overall Type I error rate.

Test: H_0: Multiple coefficients simultaneously equal zero

F-Statistic Formula:

$$F = \frac{(SSR_r – SSR_u)/q}{SSR_u/(n-k)} \sim F_{q,n-k}$$

where SSR_r and SSR_u are the residual sums of squares from the restricted and unrestricted models respectively, q is the number of restrictions being tested, n is the sample size, and k is the number of parameters in the unrestricted model.

Under H_0, this follows an **F-distribution** with $(q, n-k)$ degrees of freedom.

Intuition: The F-test compares how much worse the fit becomes when we impose restrictions (numerator) against the natural variation in the data (denominator).

If the restrictions hurt the fit a lot relative to natural variation, we reject H_0.

Application: Seasonal Effects in Ember Routes

Question: Do winter, spring, and fall conditions jointly affect journey times on the Dundee-Braemar route?

Models: The restricted model includes only route length as a predictor (JourneyTime ~ RouteLength), while the unrestricted model adds seasonal dummies (JourneyTime ~ RouteLength + Winter + Spring + Fall). Summer serves as the reference category.

```r
# Simulate journey time data with seasonal effects
set.seed(789)
n_journeys <- 120
route_length <- rep(85, n_journeys)  # km
winter <- rep(c(1,0,0,0), each = 30)
spring <- rep(c(0,1,0,0), each = 30)
fall <- rep(c(0,0,1,0), each = 30)

# True model: winter adds 8 min, spring adds 3 min, fall adds 5 min
journey_time <- 60 + 0.8 * route_length +
                8 * winter + 3 * spring + 5 * fall +
                rnorm(n_journeys, sd = 6)

ember_data <- data.frame(
  journey_time = journey_time,
  route_length = route_length,
  winter = winter,
  spring = spring,
  fall = fall
)

# Restricted model (no seasonal effects)
model_restricted <- lm(journey_time ~ route_length, data = ember_data)

# Unrestricted model (with seasonal effects)
model_full <- lm(journey_time ~ route_length + winter + spring + fall,
                 data = ember_data)

# F-test for joint significance of seasonal dummies
anova_result <- anova(model_restricted, model_full)
anova_result
```

```
## Analysis of Variance Table
##
## Model 1: journey_time ~ route_length
## Model 2: journey_time ~ route_length + winter + spring + fall
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    119 4887.4
## 2    116 4017.1  3    870.29 8.377 4.354e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

Since p < 0.001, we reject H_0 and conclude that seasonal factors jointly impact journey times. This justifies including seasonal adjustments in published schedules, planning extra buffer time during winter, considering differential pricing across seasons, and adapting staffing and maintenance plans to seasonal patterns.

Important Point: Even if individual seasonal coefficients aren't all significant in separate t-tests, the F-test can detect that they're jointly important.

Common Hypothesis Tests

Two-Sample t-test: Comparing Groups

Example: Comparing journey times for morning departures (7:00 AM) vs afternoon departures (2:00 PM) to understand traffic pattern impacts.

```r
# Journey times (minutes) for different departure times
set.seed(789)
morning_departures <- rnorm(30, mean = 136, sd = 12)
afternoon_departures <- rnorm(30, mean = 144, sd = 12)

comparison <- t.test(morning_departures, afternoon_departures)
comparison
```

```
##
##  Welch Two Sample t-test
##
## data:  morning_departures and afternoon_departures
## t = -4.4349, df = 50.08, p-value = 5.048e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.229826  -6.864939
## sample estimates:
## mean of x mean of y
##  132.6286  145.1760
```

Interpretation: "Afternoon departures take significantly longer (mean difference = 8.0 minutes, $p < 0.001$). This suggests we should adjust afternoon schedules or recommend morning travel to time-sensitive passengers."

Paired t-test: Before/After Comparisons

Example: Testing whether a driver efficiency training program reduces journey times. Same 25 drivers measured before and after the training.

```r
# Journey times (minutes): before and after driver training
set.seed(321)
before_training <- rnorm(25, mean = 142, sd = 10)
after_training <- before_training - rnorm(25, mean = 3, sd = 4)

paired_test <- t.test(before_training, after_training, paired = TRUE)
paired_test
```

```
##
##  Paired t-test
##
## data:  before_training and after_training
## t = 3.2685, df = 24, p-value = 0.003252
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.9946351 4.4028782
## sample estimates:
## mean difference
##        2.698757
```

The t-test results show a statistically significant difference between the “before_training” and “after_training” means (t(24) = 3.2685, p = 0.003). Since the p-value is well below the conventional 0.05 threshold, we can reject the null hypothesis that the true mean difference is zero. Indeed the training program is making drivers more efficient.

Proportion Tests

Example: Testing whether a new timetable (with 5 extra minutes built in) improves on-time arrival rates.

```r
# Old timetable: 782 on-time arrivals out of 1000 trips (78.2%)
# New timetable: 847 on-time arrivals out of 1000 trips (84.7%)

prop_test <- prop.test(x = c(782, 847), n = c(1000, 1000),
                       alternative = "less")
prop_test
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(782, 847) out of c(1000, 1000)
## X-squared = 13.555, df = 1, p-value = 0.0001158
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.03550712
## sample estimates:
## prop 1 prop 2
##  0.782  0.847
```

Interpretation: “The new timetable with built-in buffer time significantly improves on-time performance ($p < 0.001$). The 6.5 percentage point improvement is both statistically significant and practically meaningful for customer satisfaction.”

Confidence Regions and Simultaneous Inference

Beyond Individual Confidence Intervals

When we have multiple parameters (e.g., multiple regression coefficients), we might want to make statements about them jointly, not just individually.

Individual 95% CIs: Each interval has 95% probability of containing its true parameter.

Joint confidence: What’s the probability that ALL intervals simultaneously contain their true parameters?

If we have k independent 95% CIs:

$$P(\text{all } k \text{ intervals correct}) = 0.95^k$$

For $k = 3$: $0.95^3 = 0.857$ (only 85.7%)!

Bonferroni Correction

To achieve overall confidence level $1 - \alpha$ for k tests:

Use significance level $\alpha' = \alpha/k$ for each individual test.

Example: Testing 5 route characteristics at overall $\alpha = 0.05$: use $\alpha' = 0.05/5 = 0.01$ for each test. This controls the family-wise error rate.

Trade-off: More conservative (harder to reject), but protects against false discoveries.

Application: Monitoring

Ember monitors 20 route KPIs monthly. At $\alpha = 0.05$ per test, the expected false alarms per month are 20 × 0.05 = 1. With Bonferroni correction ($\alpha' = 0.05/20 = 0.0025$), the expected false alarms drop to 20 × 0.0025 = 0.05.

Managerial Decision: Accept more false alarms (cheaper) vs. fewer false alarms (more expensive but fewer unnecessary investigations)

Type I and Type II Errors

Every hypothesis test involves two types of potential errors:

	H_0 is True	H_0 is False
Reject H_0	Type I Error ( )	Correct Decision (Power)
Fail to Reject	Correct Decision (1 –)	Type II Error ( )

Practical Implications

A **Type I error** (false positive) means concluding that a new route is faster when it isn't, which could lead to implementing a more expensive route unnecessarily. The typical cost is increased operational expenditure with no corresponding benefit.

A **Type II error** (false negative) means concluding that a new route isn't faster when it actually is, missing an opportunity to improve service and reduce costs. The typical cost is lost efficiency, competitive disadvantage, and customer dissatisfaction.

The fundamental trade-off is that lowering (being more conservative about false positives) increases (the chance of missing true effects). The choice of should therefore reflect the relative business costs. When the cost of a false positive is high — for instance, major infrastructure investments or vehicle purchases — a stricter threshold such as = 0.01 is appropriate. When the cost of a false negative is high — route testing or schedule adjustments that are cheap to reverse — a more liberal = 0.10 may be justified. The conventional = 0.05 represents a balanced default.

Statistical Power

Power = 1 – = Probability of detecting an effect when it exists

Power increases with sample size and effect size, and decreases when we use a stricter significance level or when the population variance is larger. Understanding these relationships is essential for study design.

```r
# Power analysis: detecting 5-minute reduction in journey time
if (!requireNamespace("pwr", quietly=TRUE)) install.packages("pwr")
library(pwr)

# Assuming SD = 12 minutes, want 80% power to detect 5-minute reduction
effect_size <- 5 / 12  # Cohen's d
power_result <- pwr.t.test(d = effect_size, sig.level = 0.05, power = 0.80)
ceiling(power_result$n)
```

```
## [1] 92
```

Application: Before investing in a new route trial, use power analysis to determine how many test journeys you need to reliably detect meaningful time savings. This prevents underpowered studies that waste resources.

Statistical vs. Practical Significance

One of the most important distinctions in applied statistics: A result can be statistically significant but practically meaningless.

Example: The Large Sample Problem

```r
set.seed(999)
# Large sample: n = 10,000 journeys each
large_sample_route_A <- rnorm(10000, mean = 140.0, sd = 12)
large_sample_route_B <- rnorm(10000, mean = 140.4, sd = 12)

large_test <- t.test(large_sample_route_A, large_sample_route_B)
large_test
```

```
##
##  Welch Two Sample t-test
##
## data:  large_sample_route_A and large_sample_route_B
## t = -2.8858, df = 19994, p-value = 0.003909
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8261553 -0.1578161
## sample estimates:
## mean of x mean of y
##  139.8642  140.3562
```

The 0.4-minute (24 second) difference is statistically significant ($p < 0.05$) but **may be completely irrelevant** depending on context. A 24-second difference is negligible for scheduling, produces minimal fuel savings over a 140-minute journey, and is imperceptible to passengers. However, multiplied across 10,000 journeys per year, it amounts to 66.7 hours — which starts to look meaningful at scale.

Best practice: Always report both the statistical significance (p-value and confidence interval) and the effect size (the actual magnitude and its practical importance).

Multiple Testing and the Problem of P-Hacking

The Multiple Comparisons Problem

Testing multiple hypotheses inflates the Type I error rate. Testing a single hypothesis at = 0.05 gives a 5% chance of a false positive, but testing 20 independent hypotheses gives a 1–(0.95)^{20} = 64% chance of at least one false positive.

Example: Route Testing

Imagine Ember's operations dashboard tracking 30 KPIs monthly across different routes — journey times, on-time performance, and fuel efficiency for each of 10 routes.

Under the null hypothesis of "no real changes," you'd expect:

30 × 0.05 = 1.5 "significant" changes per month just by chance!

Real scenario: In January, your dashboard flags the Dundee–Braemar journey time as having increased (p = 0.04) and Perth–Inverness fuel efficiency as having decreased (p = 0.03). Are these real problems requiring investigation, or statistical noise?

Solutions

Several approaches address this problem. The **Bonferroni correction** is the simplest: use $\alpha' = \alpha/m$ for m tests. Testing 20 hypotheses? Use $\alpha = 0.05/20 = 0.0025$ for each. This is conservative but straightforward.

The **False Discovery Rate (FDR)** controls the proportion of false positives among rejections rather than the overall error rate. It is less conservative than Bonferroni and appropriate when some false positives are acceptable.

Finally, **pre-registration** — specifying hypotheses before seeing data — is the gold standard for confirmatory research. It prevents "p-hacking" (testing until something is significant) by committing the analyst to a fixed set of tests in advance.

From Inference to Regression Modelling

The hypothesis testing framework extends naturally to regression analysis, which we explore in Chapter @ref(ch-linear):

Coefficient Tests (Preview)

In regression $Y = \beta_0 + \beta_1 X + \varepsilon$:

Null hypothesis: $H_0 : \beta_1 = 0$ (no relationship)

Test statistic:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

This is exactly the t-test framework we developed here, applied to regression coefficients.

Interpretation: "Does traffic volume significantly predict journey time on the Dundee-Braemar route?"

Confidence vs. Prediction Intervals

Confidence interval for $E(Y|X)$: Where do we expect the *average* journey time for given conditions? This interval is narrower and more precise – for example, "What is the average journey time on a Tuesday afternoon in good weather?"

Prediction interval for $Y|X$: Where might an *individual* journey fall? This interval is wider because it accounts for individual variation – for example, "What time range should we give a passenger for their specific Tuesday afternoon trip?"

Summary: Inference in the Analytics Workflow

Statistical inference serves as a reality check for our data. It helps us determine whether the patterns we observe reflect genuine signals or merely random variation that is unlikely to recur. Rather than relying on intuition, inference provides tools to quantify uncertainty in our estimates, so that we can judge which results are worth acting on.

When reporting predictions, we should always accompany point estimates with confidence intervals that make the margin of error explicit. We must also assess whether a result matters in practice – a statistically significant difference may still be too small to justify action. When testing many hypotheses simultaneously, corrections such as Bonferroni are needed to guard against false discoveries. Setting up studies with adequate power from the outset, and being transparent about assumptions, helps ensure that the conclusions we draw are reliable.

In Chapter @ref(ch-linear), we apply these inferential principles to regression models, where we simultaneously test multiple predictors and build sophisticated forecasting systems – all built on the foundation established here.

Exercises

Question 1: Confidence Intervals and Decision-Making

Ember is considering adjusting the published schedule for the Dundee-Braemar route from 135 minutes to 140 minutes. You collect journey time data for 50 trips during peak season:

- Sample mean: 138.5 minutes
- Sample standard deviation: 14.2 minutes

a) Calculate a 95% confidence interval for the true mean journey time. Show your working.

b) Based on this interval, would you recommend changing the published schedule to 140 minutes? Explain your reasoning, considering both the statistical evidence and practical implications.

c) How would your confidence interval change if you had collected data from 200 trips instead of 50 (assuming the same sample mean and standard deviation)? What does this tell you about the value of larger samples?

Question 2: Hypothesis Testing with Sales Data

A retail company claims that a new sales training program increases average weekly sales per employee by at least £500. To test this claim, you track the weekly sales performance for 30 employees during the 4 weeks before and 4 weeks after completing the training program.

Results:

- Before training: Mean = £3,240, SD = £680
- After training: Mean = £3,680, SD = £720
- Mean difference (After − Before) = £440

a) State appropriate null and alternative hypotheses for this test. Should you use a one-tailed or two-tailed test? Justify your choice.

b) Suppose the p-value for this paired t-test is 0.06. What conclusion would you draw at = 0.05?

c) The training manager is frustrated, arguing "We DID increase sales by £440 per employee per week! That's £22,880 extra annual revenue per employee. Why isn't this significant?" Write a brief explanation (2-3 sentences) that addresses:

- Why the result is not statistically significant at $\alpha = 0.05$
- What the p-value of 0.06 actually means
- What practical actions you might recommend given these results

---

Question 3: Statistical vs. Practical Significance

You are analysing on-time performance data for three Ember routes. Each route has 2,000 recorded journeys. An arrival is "on-time" if it arrives within 5 minutes of the scheduled time.

Results:

- **Route A** (Dundee-Aberdeen): 1,580/2,000 on-time (79.0%)
- **Route B** (Dundee-Edinburgh): 1,620/2,000 on-time (81.0%)
- **Route C** (Dundee-Glasgow): 1,660/2,000 on-time (83.0%)

A proportion test comparing Route A vs. Route C gives $p < 0.001$ (highly significant).

a) Calculate the absolute difference in on-time rates between Route A and Route C (in percentage points).

b) Despite the highly significant p-value, the operations manager questions whether this difference is practically meaningful. List three factors you would consider when deciding whether a 4 percentage point difference in on-time performance justifies operational changes.

c) Suppose instead of 2,000 journeys per route, you had only 50 journeys per route. How would this affect:

- The statistical significance (p-value)?
- The practical significance?

Explain your reasoning.

Question 4: F-Tests and Model Comparison

You are building a model to predict Ember journey times. You have two candidate models:

- **Model 1**: JourneyTime ~ RouteLength
- **Model 2**: JourneyTime ~ RouteLength + DayOfWeek + Weather

An F-test comparing these models gives $F = 8.42$ (df = 2, 115) with $p = 0.0004$.

a) What is the null hypothesis being tested by this F-test?

b) What does the small p-value tell you about adding DayOfWeek and Weather?

c) Explain why we use an F-test rather than two separate t-tests for DayOfWeek and Weather.