```r
library(readr)
```

```
Warning: package 'readr' was built under R version 4.4.3
```

```r
library(dplyr)
library(lubridate)
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```r
library(here)
library(tidyr)
```

```
Warning: package 'tidyr' was built under R version 4.4.3
```

```r
library(scales)
library(sf)
```

```
Warning: package 'sf' was built under R version 4.4.3
```

```r
here::i_am("hypothesis2.qmd")
```

```r
# Load data and define the analysis window.
taxi <- read_csv(here("data", "taxidata_clean.csv"), show_col_types = FALSE) %>%
  mutate(
    year = year(trip_start_timestamp),
    pickup_community_area = as.integer(pickup_community_area)
  )

late_hours <- c(23, 0, 1, 2, 3, 4)

taxi_clean <- taxi %>%
  filter(year %in% c(2014, 2019), !is.na(pickup_community_area)) %>%
  mutate(
    hour  = hour(trip_start_timestamp),
    late  = hour %in% late_hours,
    dow   = wday(trip_start_timestamp),
    month = month(trip_start_timestamp)
  )
```

### Per-area proportion tests with BH and BY correction

```r
# Official City of Chicago community area names via Socrata open-data API.
# This avoids hard-coding names and ensures alignment with the city's numbering.
ca_lookup <- read_sf("https://data.cityofchicago.org/resource/igwz-8jzy.geojson") %>%
  st_drop_geometry() %>%
  transmute(
    pickup_community_area = as.integer(area_num_1),
    area_name = tools::toTitleCase(tolower(community))
  ) %>%
  arrange(pickup_community_area)

# Area-year summary table.
area_year <- taxi_clean %>%
  group_by(pickup_community_area, year) %>%
  summarise(
    n_total = n(),
    n_late  = sum(late),
    .groups = "drop"
  )

area_wide <- area_year %>%
  pivot_wider(
    names_from  = year,
    values_from = c(n_total, n_late),
    values_fill = list(n_total = 0, n_late = 0)
  )

# One proportion test per area, then BH and BY correction.
test_results <- area_wide %>%
  filter(n_total_2014 >= 200, n_total_2019 >= 200) %>%
  rowwise() %>%
  mutate(
    p_value = prop.test(
      c(n_late_2014, n_late_2019),
      c(n_total_2014, n_total_2019),
      correct = FALSE
    )$p.value
  ) %>%
  ungroup() %>%
  left_join(ca_lookup, by = "pickup_community_area") %>%
  mutate(
    late_share_2014 = n_late_2014 / n_total_2014,
    late_share_2019 = n_late_2019 / n_total_2019,
    share_change = late_share_2019 - late_share_2014,
    p_adj_bh = p.adjust(p_value, method = "BH"),
    p_adj_by = p.adjust(p_value, method = "BY"),
    bh_reject = p_adj_bh < 0.05,
    by_reject = p_adj_by < 0.05
  ) %>%
  arrange(p_adj_bh)

test_results %>%
  select(
    pickup_community_area, area_name,
    n_total_2014, n_total_2019,
    late_share_2014, late_share_2019, share_change,
    p_value, p_adj_bh, p_adj_by, bh_reject
  ) %>%
  slice_head(n = 15)
```

```
# A tibble: 15 x 11
   pickup_community_area area_name     n_total_2014 n_total_2019 late_share_2014
                   <int> <chr>                <int>        <int>           <dbl>
 1                     8 Near North S~        64253        64355          0.224
 2                    32 Loop                 40253        58501          0.0911
 3                    28 Near West Si~        17825        23118          0.153
 4                     7 Lincoln Park         13440         4384          0.338
 5                     6 Lake View            16430         6751          0.345
 6                    76 Ohare                 9373        17374          0.146
 7                    33 Near South S~         5217         5927          0.0947
 8                     3 Uptown                3757         1922          0.259
 9                    77 Edgewater             2199         1556          0.225
10                     4 Lincoln Squa~         1095          665          0.327
11                    24 West Town             9447         1967          0.381
12                     5 North Center          1760          428          0.313
13                    22 Logan Square          3270          732          0.454
14                    41 Hyde Park             1155          705          0.0632
15                    21 Avondale               627          259          0.319
# i 6 more variables: late_share_2019 <dbl>, share_change <dbl>, p_value <dbl>,
#   p_adj_bh <dbl>, p_adj_by <dbl>, bh_reject <lgl>
```

### Supplementary global-null tests

```r
m <- nrow(test_results)

fisher_stat <- -2 * sum(log(test_results$p_value))
fisher_p <- pchisq(fisher_stat, df = 2 * m, lower.tail = FALSE)

p_sorted <- sort(test_results$p_value)
simes_p <- min(1, min(p_sorted * m / seq_len(m)))

tibble(
  test    = c("Fisher combination", "Simes"),
  p_value = c(fisher_p, simes_p)
)
```

```
# A tibble: 2 x 2
  test               p_value
  <chr>                <dbl>
1 Fisher combination       0
2 Simes                    0
```

### Discovery counts

```r
tibble(
  method      = c("BH (FDR control)", "BY (FDR under dependence)"),
  discoveries = c(sum(test_results$bh_reject), sum(test_results$by_reject)),
  tested      = c(m, m)
)
```

```
# A tibble: 2 x 3
  method                    discoveries tested
  <chr>                           <int>  <int>
1 BH (FDR control)                   20     24
2 BY (FDR under dependence)          18     24
```

```r
# Verify the direction of all BH discoveries.
n_decline <- sum(test_results$bh_reject & test_results$share_change < 0)
n_increase <- sum(test_results$bh_reject & test_results$share_change > 0)
tibble(
  direction = c("Decline (share_change < 0)", "Increase (share_change > 0)"),
  n_discoveries = c(n_decline, n_increase)
)
```

```
# A tibble: 2 x 2
  direction                  n_discoveries
  <chr>                              <int>
1 Decline (share_change < 0)            20
2 Increase (share_change > 0)            0
```

### Visualization

```
plot_df <- test_results %>%
  slice_max(order_by = abs(share_change), n = 20) %>%
  mutate(area_label = paste0(pickup_community_area, " - ", area_name))

ggplot(plot_df, aes(x = reorder(area_label, share_change), y = share_change, fill = bh_rejec
  geom_col() +
  coord_flip() +
  scale_y_continuous(labels = label_percent(accuracy = 0.1)) +
  scale_fill_manual(
    values = c("FALSE" = "grey70", "TRUE" = "#1f78b4"),
    labels = c("FALSE" = "Not rejected", "TRUE" = "Rejected")
  ) +
  labs(
    title = "Largest Changes in Late-Night Taxi Share",
    subtitle = "Blue = BH-significant at FDR 5%",
    x = NULL,
    y = "Change in late-night share (pp)",
    fill = "BH (alpha = 0.05)"
  ) +
  theme_minimal(base_size = 11) +
  theme(legend.position = "bottom")
```

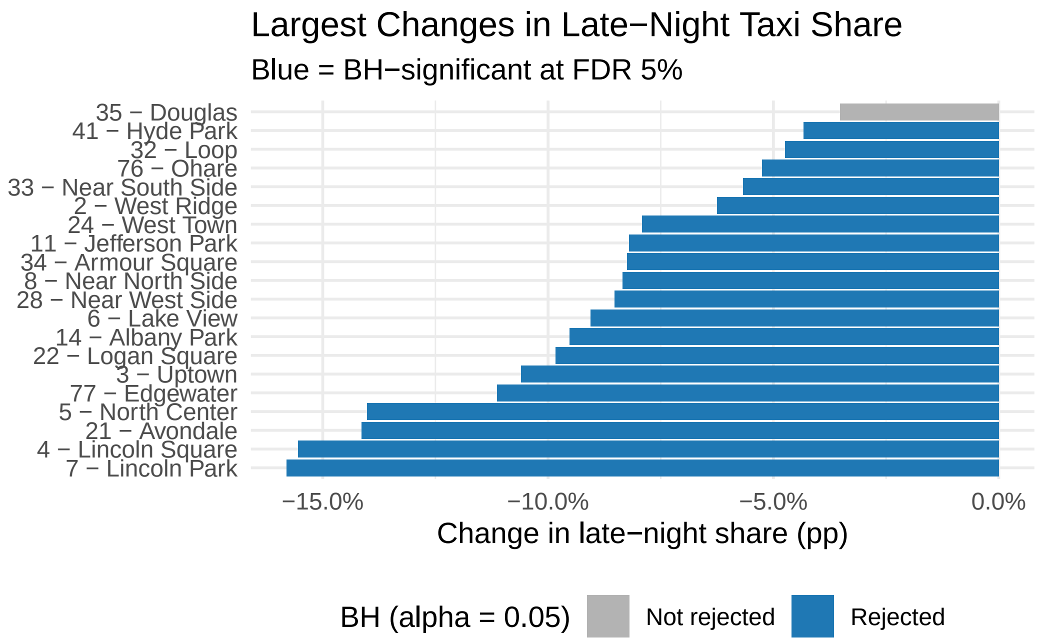

Figure 1: Change in late-night taxi share by community area (2019 minus 2014). Blue bars are BH-significant at FDR 5%; grey bars are not.

### Grouped comparison: pre-specified area sets

```
nightlife_residential <- c(6, 7, 8, 22, 24)
airport_downtown_business <- c(28, 32, 33, 76)

group_year <- area_year %>%
  mutate(
    area_group = case_when(
      pickup_community_area %in% nightlife_residential ~ "Nightlife/Residential",
      pickup_community_area %in% airport_downtown_business ~ "Airport/Downtown/Business",
      TRUE ~ "Other"
    )
  ) %>%
  filter(area_group != "Other") %>%
  group_by(area_group, year) %>%
  summarise(
    n_total = sum(n_total),
    n_late  = sum(n_late),
    .groups = "drop"
  )

group_results <- group_year %>%
  pivot_wider(
    names_from  = year,
    values_from = c(n_total, n_late)
  ) %>%
  mutate(
    late_share_2014 = n_late_2014 / n_total_2014,
    late_share_2019 = n_late_2019 / n_total_2019,
    share_change    = late_share_2019 - late_share_2014
  ) %>%
  rowwise() %>%
  mutate(
    p_value = prop.test(
      c(n_late_2014, n_late_2019),
      c(n_total_2014, n_total_2019),
      correct = FALSE
    )$p.value
  ) %>%
  ungroup()

group_results %>%
  select(
    area_group, n_total_2014, n_total_2019,
    late_share_2014, late_share_2019, share_change, p_value
  )
```

```
# A tibble: 2 x 7
  area_group           n_total_2014 n_total_2019 late_share_2014 late_share_2019
  <chr>                       <int>        <int>           <dbl>           <dbl>
1 Airport/Downtown/Bu~        72668       104920           0.114          0.0569
2 Nightlife/Residenti~       106840        78189           0.278          0.158
# i 2 more variables: share_change <dbl>, p_value <dbl>
```

```
# Difference-in-differences: year x group interaction test.
interaction_data <- taxi_clean %>%
  filter(
    pickup_community_area %in% c(nightlife_residential, airport_downtown_business)
  ) %>%
  mutate(
    nightlife = as.integer(pickup_community_area %in% nightlife_residential),
    post      = as.integer(year == 2019)
  )

fit <- glm(late ~ post * nightlife, data = interaction_data, family = binomial)
summary(fit)
```

```

Call:
glm(formula = late ~ post * nightlife, family = binomial, data = interaction_data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -2.05409    0.01169 -175.74   <2e-16 ***
post           -0.75360    0.01773  -42.52   <2e-16 ***
nightlife       1.09854    0.01354   81.14   <2e-16 ***
post:nightlife  0.03953    0.02137    1.85   0.0644 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 313092  on 362616  degrees of freedom
Residual deviance: 291890  on 362613  degrees of freedom
AIC: 291898

Number of Fisher Scoring iterations: 5
```

```
# Extract and interpret the interaction coefficient.
beta3 <- coef(fit)["post:nightlife"]
se3 <- summary(fit)$coefficients["post:nightlife", "Std. Error"]
p3 <- summary(fit)$coefficients["post:nightlife", "Pr(>|z|)"]
tibble(
  term       = "year2019 x nightlife (interaction)",
  estimate   = beta3,
  std_error  = se3,
  p_value    = p3
)
```

```
# A tibble: 1 x 4
  term                               estimate std_error p_value
  <chr>                                 <dbl>     <dbl>   <dbl>
1 year2019 x nightlife (interaction)   0.0395    0.0214  0.0644
```

### Robustness: stratified permutation test

```r
set.seed(42)
B <- 2000

# Observed late-night share difference per area.
obs_diff <- taxi_clean %>%
  group_by(pickup_community_area, year) %>%
  summarise(late_share = mean(late), n = n(), .groups = "drop") %>%
  pivot_wider(
    names_from = year, values_from = c(late_share, n),
    values_fill = list(late_share = NA, n = 0)
  ) %>%
  filter(n_2014 >= 200, n_2019 >= 200) %>%
  mutate(obs_stat = late_share_2019 - late_share_2014)

# Keep only areas passing the sample-size filter.
areas_to_test <- obs_diff$pickup_community_area

perm_data <- taxi_clean %>%
  filter(pickup_community_area %in% areas_to_test) %>%
  mutate(stratum = interaction(month, dow, drop = TRUE)) %>%
  select(pickup_community_area, year, late, stratum)

# Function: one permutation round for all areas.
# We shuffle year labels within each (area, stratum) cell using base-R split-apply
# to avoid the dplyr grouped-mutate size constraint.
# Note: sample(x) when length(x)==1 is interpreted as sample.int(x,1), so we use
# x[sample.int(length(x))] instead, which is safe for any length.
one_perm <- function(df) {
  idx <- split(seq_len(nrow(df)), list(df$pickup_community_area, df$stratum))
  year_perm <- df$year # start as a copy; overwrite in place
  for (i in idx) {
    n_i <- length(i)
    if (n_i > 1L) {
      year_perm[i] <- df$year[i][sample.int(n_i)]
    }
    # n_i <= 1: nothing to permute
  }
  df$year_perm <- year_perm

  df %>%
    group_by(pickup_community_area, year_perm) %>%
    summarise(late_share = mean(late), .groups = "drop") %>%
    pivot_wider(names_from = year_perm, values_from = late_share) %>%
    mutate(perm_stat = `2019` - `2014`) %>%
    select(pickup_community_area, perm_stat)
}

# Run B permutations.
perm_stats <- bind_rows(lapply(seq_len(B), function(b) {
  one_perm(perm_data) %>% mutate(b = b)
}))

# Compute two-sided permutation p-values.
perm_pvals <- perm_stats %>%
  inner_join(obs_diff %>% select(pickup_community_area, obs_stat), by = "pickup_community_are
  group_by(pickup_community_area) %>%
  summarise(
    perm_p = (sum(abs(perm_stat) >= abs(obs_stat)) + 1) / (n() + 1),
    .groups = "drop"
  )

# Merge with test_results and apply BH.
perm_results <- obs_diff %>%
  select(pickup_community_area, obs_stat) %>%
  inner_join(perm_pvals, by = "pickup_community_area") %>%
  mutate(
    perm_p_bh    = p.adjust(perm_p, method = "BH"),
    perm_reject  = perm_p_bh < 0.05
  ) %>%
  arrange(perm_p_bh)

perm_results <- perm_results %>%
  left_join(ca_lookup, by = "pickup_community_area")

perm_results %>%
  select(pickup_community_area, area_name, obs_stat, perm_p, perm_p_bh, perm_reject) %>%
  slice_head(n = 15)
```

```
# A tibble: 15 x 6
   pickup_community_area area_name       obs_stat   perm_p perm_p_bh perm_reject
                   <int> <chr>              <dbl>    <dbl>     <dbl> <lgl>
 1                     3 Uptown          -0.106   0.000500  0.000750 TRUE
 2                      4 Lincoln Square   -0.156  0.000500  0.000750 TRUE
 3                      5 North Center     -0.140  0.000500  0.000750 TRUE
 4                      6 Lake View        -0.0906 0.000500  0.000750 TRUE
 5                      7 Lincoln Park     -0.158  0.000500  0.000750 TRUE
 6                      8 Near North Side  -0.0835 0.000500  0.000750 TRUE
 7                     21 Avondale         -0.141  0.000500  0.000750 TRUE
 8                     22 Logan Square     -0.0983 0.000500  0.000750 TRUE
 9                     24 West Town        -0.0792 0.000500  0.000750 TRUE
10                     28 Near West Side   -0.0853 0.000500  0.000750 TRUE
11                     32 Loop             -0.0474 0.000500  0.000750 TRUE
12                     33 Near South Side  -0.0567 0.000500  0.000750 TRUE
13                     41 Hyde Park        -0.0433 0.000500  0.000750 TRUE
14                     56 Garfield Ridge   -0.0277 0.000500  0.000750 TRUE
15                     76 Ohare            -0.0525 0.000500  0.000750 TRUE
```

```
cat("Permutation-based BH discoveries:", sum(perm_results$perm_reject), "of", nrow(perm_resul
```

```
Permutation-based BH discoveries: 20 of 24
```

```
cat("Parametric BH discoveries:       ", sum(test_results$bh_reject), "of", nrow(test_results
```

```
Parametric BH discoveries:        20 of 24
```