Newdata

Chang Lu

2025-03-27

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: ’lubridate’

## The following objects are masked from ’package:base’:
##
##     date, intersect, setdiff, union
```

```
library(stringr)
library(purrr)
library(readxl)
```

```
# Computing Total_Quit_Time_Computed
submit_time_cols <- names(combined_df)[str_detect(names(combined_df), "-T_Page Submit")]

combined_df$Total_Quit_Time_Computed <- combined_df %>%
  select(all_of(submit_time_cols)) %>%
  mutate_all(as.numeric) %>%
  rowSums(na.rm = TRUE)


combined_df %>% select(...1, Total_Quit_Time_Computed) %>% head()
```

```
## # A tibble: 6 x 2
##   ...1      Total_Quit_Time_Computed
##   <chr>                        <dbl>
## 1 Student 1                     38.9
## 2 Student 2                   1243. 
## 3 Student 3                   1209. 
## 4 Student 4                    172. 
## 5 Student 5                   1495. 
## 6 Student 6                   2465. 
```

```
length(combined_df$Total_Quit_Time_Computed)
```

```
## [1] 32
```

```
# Rename demographic groups
combined_df <- combined_df %>%
  rename(
    Accommodations       = Q21,
    Gender               = Q22,
    Ethnicity            = Q23,
    English_Proficiency  = Q24,
    Country_You          = Q25_1,
    Country_Mother       = Q25_2,
    Country_Father       = Q25_3,
    Home_Language        = Q69
  ) %>%
  select(-Q68)  # Drop Q68 for a few people answered
```

```
combined_df
```

```
## # A tibble: 32 x 153
##    ...1       ‘Duration (in seconds)‘ Q1    ‘Q1-T_First Click‘ ‘Q1-T_Last Click‘
##    <chr>      <chr>                   <chr> <chr>              <chr>
##  1 Student 1  99.0                    Neph~ 1.358              1.358
##  2 Student 2  1463.0                  Cent~ 28.552             79.346
##  3 Student 3  1798.0                  Cent~ 2.957              42.543
##  4 Student 4  265.0                   Type~ 28.61              28.61
##  5 Student 5  2574.0                  Cent~ 43.518             43.518
##  6 Student 6  2744.0                  Neph~ 44.038             139.353
##  7 Student 7  2861.0                  Cent~ 2.081              3.231
##  8 Student 8  10090.0                 Cent~ 49.556             123.64
##  9 Student 9  1036.0                  Cent~ 15.864             40.063
## 10 Student 10 93675.0                 Neph~ 56.798             56.798
## # i 22 more rows
## # i 148 more variables: ‘Q1-T_Page Submit‘ <chr>, ‘Q1-T_Click Count‘ <chr>,
## #   ‘Q1-I‘ <chr>, ‘Q1-I_5_TEXT‘ <chr>, Q2 <chr>, ‘Q2-T_First Click‘ <chr>,
## #   ‘Q2-T_Last Click‘ <chr>, ‘Q2-T_Page Submit‘ <chr>,
## #   ‘Q2-T_Click Count‘ <chr>, ‘Q2-I‘ <chr>, ‘Q2-I_5_TEXT‘ <chr>, Q3 <chr>,
## #   ‘Q3-T_First Click‘ <chr>, ‘Q3-T_Last Click‘ <chr>,
## #   ‘Q3-T_Page Submit‘ <chr>, ‘Q3-T_Click Count‘ <chr>, ‘Q3-I‘ <chr>, ...
```

Normality Check

```
# # Accomodations
# group_yes <- combined_df %>% filter(Accommodations == "Yes") %>% pull(Total_Quit_Time_Computed)
# group_no  <- combined_df %>% filter(Accommodations == "No") %>% pull(Total_Quit_Time_Computed)
#
# # Shapiro-Wilk test
# shapiro.test(group_yes)
# shapiro.test(group_no)
```

```
# par(mfrow = c(2, 2))  # Plot layout
#
# # Histogram
# hist(group_yes, main = "Accommodations: YES", xlab = "Total Time", col = "skyblue")
# hist(group_no, main = "Accommodations: NO", xlab = "Total Time", col = "salmon")
#
# # Q-Q plots
# qqnorm(group_yes); qqline(group_yes, col = "blue")
# qqnorm(group_no); qqline(group_no, col = "red")
```

Accomodations Error in shapiro.test(group_yes) : sample size must be between 3 and 5000 -> Mann–Whitney U test

```
# Group sizes
table(combined_df$Gender)
```

Gender

```
##
## Female   Male
##     23      8
```

```
# Normality check
by(combined_df$Total_Quit_Time_Computed, combined_df$Gender, shapiro.test)
```

```
## combined_df$Gender: Female
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.75886, p-value = 9.016e-05
##
## ------------------------------------------------------------
## combined_df$Gender: Male
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.95465, p-value = 0.7579
```

We tested whether total quit time differed by gender. The Female group violated the assumption of normality (Shapiro-Wilk $p < 0.001$), so we used a non-parametric Mann–Whitney U test.

```r
table(combined_df$Ethnicity)
```

Race

```
##
##                      Asian Black or African American        Hispanic or Latino
##                         10                         1                         5
##                      Other                     White
##                          2                        13
```

```r
# Kruskal-Wallis test (non-parametric ANOVA)
kruskal.test(Total_Quit_Time_Computed ~ Ethnicity, data = combined_df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Total_Quit_Time_Computed by Ethnicity
## Kruskal-Wallis chi-squared = 1.0582, df = 4, p-value = 0.9008
```

A Kruskal-Wallis test revealed no significant difference in total quit time across ethnic groups ($\chi^2(4) = 1.06$, $p = 0.901$).

```r
# Clean and recode English proficiency
combined_df <- combined_df %>%
  mutate(English_Proficiency = case_when(
    str_trim(English_Proficiency) %in% c("Native", "Native Speaker") ~ "Native",
    str_trim(English_Proficiency) == "Fluent" ~ "Fluent",
    TRUE ~ English_Proficiency  # keep as-is in case there's other unexpected values
  ))

table(combined_df$English_Proficiency)
```

English Proficiency

```
##
## Fluent Native
##     13     18
```

```r
wilcox.test(Total_Quit_Time_Computed ~ English_Proficiency, data = combined_df)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by English_Proficiency
## W = 113, p-value = 0.8902
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test found no significant difference in total quit time between Fluent and Native English speakers (W = 113, p = 0.89)

```
# You
combined_df <- combined_df %>%
  mutate(Born_in_US = ifelse(Country_You == "United States", "Yes", "No"))

wilcox.test(Total_Quit_Time_Computed ~ Born_in_US, data = combined_df)
```

Born in US

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by Born_in_US
## W = 92, p-value = 0.729
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test found no significant difference in total quit time between students born in the U.S. and those born elsewhere (W = 92, p = 0.73).

```
# Mother
combined_df <- combined_df %>%
  mutate(Born_in_US_M = ifelse(Country_Mother == "United States", "Yes", "No"))

wilcox.test(Total_Quit_Time_Computed ~ Born_in_US_M, data = combined_df)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by Born_in_US_M
## W = 156, p-value = 0.09266
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test comparing total quit time by mother's country of birth showed a trend toward significance (W = 156, p = 0.093), but did not reach conventional significance levels. It may need larger sample size to support.

```
# Father
combined_df <- combined_df %>%
  mutate(Born_in_US_F = ifelse(Country_Father == "United States", "Yes", "No"))

wilcox.test(Total_Quit_Time_Computed ~ Born_in_US_F, data = combined_df)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by Born_in_US_F
## W = 149, p-value = 0.2107
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test showed no significant difference in total quit time based on father's country of birth (W = 149, p = 0.211).

```
combined_df <- combined_df %>%
  mutate(English_Home = ifelse(Home_Language == "English", "Yes", "No"))

wilcox.test(Total_Quit_Time_Computed ~ English_Home, data = combined_df)
```

Home Language

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by English_Home
## W = 152, p-value = 0.1297
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test showed no significant difference in total quit time between students who spoke English at home and those who did not (W = 152, p = 0.130).

To qiuyi

Hi Qiuyi, when comparing total quit time across demographic groups, here's how to decide between using a t-test or the Mann–Whitney U test:

Use t-test if: Both groups have at least 10 observations and the distribution of total quit time is approximately normal in both groups (check via Shapiro-Wilk or Q-Q plot)(I didnt draw plots) \

Use Mann–Whitney U test if: 1.One or both groups violate normality

2.Sample size is small (especially < 10)

T-test and U-test

Due to the sample size, we decide to use both T-test and U-test to check each variable. The reason is that even the specific variable corresponding to the normality check but the small sample size leads to a large variance so we cannot believe the simple result.

1.English Proficiency

```
group1 <- combined_df$`Total_Quit_Time_Computed`[combined_df$`English_Proficiency` == 'Fluent']
group2 <- combined_df$`Total_Quit_Time_Computed`[combined_df$`English_Proficiency` == 'Native']
t_test_result <- t.test(group1, group2, var.equal = TRUE)
print(t_test_result)
```

```
##
##  Two Sample t-test
##
```

```
## data:  group1 and group2
## t = 0.78701, df = 29, p-value = 0.4377
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -360.4643  811.4053
## sample estimates:
## mean of x mean of y
## 1132.1058  906.6353
```

```
u_test_english <- wilcox.test(combined_df$Total_Quit_Time_Computed ~ combined_df$English_Proficiency, e

print(u_test_english)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  combined_df$Total_Quit_Time_Computed by combined_df$English_Proficiency
## W = 113, p-value = 0.8886
## alternative hypothesis: true location shift is not equal to 0
```

2.Race

```
group1 <- combined_df$Total_Quit_Time_Computed[combined_df$Ethnicity == 'White']
group2 <- combined_df$Total_Quit_Time_Computed[combined_df$Ethnicity != 'White']

t_Race <- t.test(group1, group2, var.equal = TRUE)

print(t_Race)
```

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = -0.014514, df = 29, p-value = 0.9885
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -596.3593  587.9545
## sample estimates:
## mean of x mean of y
##  998.7474 1002.9498
```

```
u_test_race <- wilcox.test(combined_df$Total_Quit_Time_Computed ~ (combined_df$Ethnicity == 'White'), e

print(u_test_race)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  combined_df$Total_Quit_Time_Computed by combined_df$Ethnicity == "White"
## W = 142, p-value = 0.3267
## alternative hypothesis: true location shift is not equal to 0
```

3.Country_you

```
group1 <- combined_df$Total_Quit_Time_Computed[combined_df$Country_You == 'United States']
group2 <- combined_df$Total_Quit_Time_Computed[combined_df$Country_You != 'United States']

t_Country <- t.test(group1, group2, var.equal = TRUE)

print(t_Country)
```

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = -0.40245, df = 29, p-value = 0.6903
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -834.0544  559.7835
## sample estimates:
## mean of x mean of y
##  970.2214 1107.3569
```

```
u_test_country <- wilcox.test(group1, group2, exact = FALSE)

print(u_test_country)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  group1 and group2
## W = 76, p-value = 0.7231
## alternative hypothesis: true location shift is not equal to 0
```

4.Country Mother

```
group1 <- combined_df$Total_Quit_Time_Computed[combined_df$Country_Mother == 'United States']
group2 <- combined_df$Total_Quit_Time_Computed[combined_df$Country_Mother != 'United States']

t_Mother <- t.test(group1, group2, var.equal = TRUE)

print(t_Mother)
```

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = -0.49639, df = 29, p-value = 0.6234
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -742.3533  452.3818
## sample estimates:
## mean of x mean of y
##  912.3253 1057.3110
```

```
u_test_mother <- wilcox.test(group1, group2, exact = FALSE)

print(u_test_mother)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  group1 and group2
## W = 72, p-value = 0.09237
## alternative hypothesis: true location shift is not equal to 0
```

5.Country Father

```
group1 <- combined_df$Total_Quit_Time_Computed[combined_df$Country_Father == 'United States']
group2 <- combined_df$Total_Quit_Time_Computed[combined_df$Country_Father != 'United States']

t_Father <- t.test(group1, group2, var.equal = TRUE)

print(t_Father)
```

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = -0.29547, df = 29, p-value = 0.7697
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -676.6882  505.8513
## sample estimates:
## mean of x mean of y
##  951.5897 1037.0081
```

```
u_test_father <- wilcox.test(group1, group2, exact = FALSE)

print(u_test_father)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  group1 and group2
## W = 85, p-value = 0.2073
## alternative hypothesis: true location shift is not equal to 0
```

6.Home Language

```r
group1 <- combined_df$Total_Quit_Time_Computed[combined_df$Home_Language == 'English']
group2 <- combined_df$Total_Quit_Time_Computed[combined_df$Home_Language != 'English']

t_HomeLang <- t.test(group1, group2, var.equal = TRUE)

print(t_HomeLang)
```

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = -0.92596, df = 29, p-value = 0.3621
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -858.8964  323.5509
## sample estimates:
## mean of x mean of y
##  897.5722 1165.2450
```

```r
u_test_homeLang <- wilcox.test(group1, group2, exact = FALSE)

print(u_test_homeLang)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  group1 and group2
## W = 76, p-value = 0.1283
## alternative hypothesis: true location shift is not equal to 0
```

Gender

```r
group1 <- combined_df$Total_Quit_Time_Computed[combined_df$Gender == 'Female']
group2 <- combined_df$Total_Quit_Time_Computed[combined_df$Gender != 'Female']

t_Gender <- t.test(group1, group2, var.equal = TRUE)

print(t_Gender)
```

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = -0.59719, df = 29, p-value = 0.555
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -857.5164  469.9186
## sample estimates:
## mean of x mean of y
##  951.1749 1144.9737
```

```r
u_test_Gender <- wilcox.test(group1, group2, exact = FALSE)

print(u_test_Gender)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  group1 and group2
## W = 66, p-value = 0.2497
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion

So the final conclusion is:

```r
library(knitr)
data1 <- data.frame(
  Variable = c("English Proficiency", "Race", "Country You",
               "Country Mother", "Country Father", "Home Language", "Gender"),
  `T-test P value` = c(0.4377, 0.9885, 0.6903, 0.6234, 0.7697, 1, 0.555),
  `U-test P value` = c(0.8886, 0.3267, 0.7231, 0.09237, 0.2073, 0.2073, 0.2497)
)

kable(data1, caption = "P-values from T-test and U-test for different variables")
```

Table 1: P-values from T-test and U-test for different variables

Variable	T.test.P.value	U.test.P.value
English Proficiency	0.4377	0.88860
Race	0.9885	0.32670
Country You	0.6903	0.72310
Country Mother	0.6234	0.09237
Country Father	0.7697	0.20730
Home Language	1.0000	0.20730
Gender	0.5550	0.24970

The conclusion is that almost all variables they don’t have such a huge significant difference in the aspect of quiz_time.