T_test

Qiuyi Feng

2025-02-20

Since U-test and T-test only focus on binary data, so some category data like 'Age Arrive USA', continous data 'quiz time(categorized by 10 minutes)' don't work.

1. quiz time(categorized by 10 minutes)
2. English Proficiency
3. Race
4. Home Language

Mann-Whitney U Test

1. Accomodations
2. Age Arrive USA
3. Sex
4. Born in USA

```
library(readxl)
df <- read_excel("MAMS and Dental Combined for Analysis.xlsx", sheet = "Sheet1")
```

```
## New names:
## * `` -> `...1`
## * `E.` -> `E....10`
## * `E.` -> `E....12`
```

```
df
```

```
## # A tibble: 36 x 30
##    ...1     Central diabetes ins~1 Decreased conduction~2 Mitral valve stenosi~3
##    <chr>    <chr>                  <chr>                  <chr>
##  1 Student  Q1                     Q2                     Q3
##  2 Student~ 1.0                    0.0                    0.0
##  3 Student~ 1.0                    1.0                    0.0
##  4 Student~ 0.0                    1.0                    0.0
##  5 Student~ 0.0                    0.0                    0.0
##  6 Student~ 1.0                    0.0                    1.0
##  7 Student~ 1.0                    1.0                    0.0
##  8 Student~ 1.0                    1.0                    0.0
```

```
##  9 Student~ 1.0                   0.0                   0.0
## 10 Student~ 0.0                   1.0                   0.0
## # i 26 more rows
## # i abbreviated names: 1: 'Central diabetes insipidus',
## #   2: 'Decreased conduction rate along the bundle branches',
## #   3: 'Mitral valve stenosis'
## # i 26 more variables:
## #   'Decreased pulmonary capillary hydrostatic fluid pressure' <chr>,
## #   Oxytocin <chr>, 'Graves' disease' <chr>, B. <chr>, ...
```

```
summary(df)
```

```
##      ...1           Central diabetes insipidus
##  Length:36          Length:36
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##  Decreased conduction rate along the bundle branches Mitral valve stenosis
##  Length:36                                           Length:36
##  Class :character                                    Class :character
##  Mode  :character                                    Mode  :character
##  Decreased pulmonary capillary hydrostatic fluid pressure   Oxytocin
##  Length:36                                                  Length:36
##  Class :character                                           Class :character
##  Mode  :character                                           Mode  :character
##  Graves' disease         B.
##  Length:36          Length:36
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##  Increased serum aldosterone concentration   E....10
##  Length:36                                  Length:36
##  Class :character                           Class :character
##  Mode  :character                           Mode  :character
##  Arterial O2 concentration   E....12          Blocked urethra
##  Length:36                  Length:36          Length:36
##  Class :character           Class :character   Class :character
##  Mode  :character           Mode  :character   Mode  :character
##  Excess maternal androgens
##  Length:36
##  Class :character
##  Mode  :character
##  The elastic recoil of the stretched arterial walls provides the force to continue blood flow in the
##  Length:36
##  Class :character
##  Mode  :character
##  Mutations that result in inactive IGF-1 receptors
##  Length:36
##  Class :character
##  Mode  :character
##  A decrease in Ca2+ resorption from bone Absence of a Y chromosome
##  Length:36                               Length:36
##  Class :character                        Class :character
##  Mode  :character                        Mode  :character
##  Testosterone stimulates GnRH from the hypothalamus
##  Length:36
```

```
##  Class :character
##  Mode  :character
##  Plasma angiotensin II concentration increases
##  Length:36
##  Class :character
##  Mode  :character
##  Its production is enhanced by cortisol. Total Score         Quiz Time
##  Length:36                               Length:36          Length:36
##  Class :character                        Class :character   Class :character
##  Mode  :character                        Mode  :character   Mode  :character
##  Accomodations          Sex            Race/Ethnicity     English Proficiency
##  Length:36          Length:36          Length:36          Length:36
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##    Born USA         Home Language      Age arrive USA
##  Length:36          Length:36          Length:36
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
df_students <- df[2:32, ]

head(df_students)
```

```
## # A tibble: 6 x 30
##   ...1      Central diabetes ins~1 Decreased conduction~2 Mitral valve stenosi~3
##   <chr>     <chr>                  <chr>                  <chr>
## 1 Student 1 1.0                    0.0                    0.0
## 2 Student 2 1.0                    1.0                    0.0
## 3 Student 3 0.0                    1.0                    0.0
## 4 Student 4 0.0                    0.0                    0.0
## 5 Student 5 1.0                    0.0                    1.0
## 6 Student 6 1.0                    1.0                    0.0
## # i abbreviated names: 1: ‘Central diabetes insipidus‘,
## #   2: ‘Decreased conduction rate along the bundle branches‘,
## #   3: ‘Mitral valve stenosis‘
## # i 26 more variables:
## #   ‘Decreased pulmonary capillary hydrostatic fluid pressure‘ <chr>,
## #   Oxytocin <chr>, ‘Graves’ disease‘ <chr>, B. <chr>,
## #   ‘Increased serum aldosterone concentration‘ <chr>, E....10 <chr>, ...
```

```
# Define the new column names for the questions
question_names <- paste0("Q", 1:20)  # Generates Q1, Q2, ..., Q20

# Rename only the question columns
colnames(df_students)[2:21] <- question_names

head(df_students)
```

```
## # A tibble: 6 x 30
##   ...1   Q1    Q2    Q3    Q4    Q5    Q6    Q7    Q8    Q9    Q10   Q11   Q12
##   <chr>  <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Stude~ 1.0   0.0   0.0   0.0   0.0   1.0   1.0   0.0   1.0   0.0   1.0   1.0
## 2 Stude~ 1.0   1.0   0.0   0.0   1.0   1.0   1.0   1.0   1.0   0.0   0.0   1.0
```

```
## 3 Stude~ 0.0   1.0   0.0   1.0   1.0   1.0   1.0   1.0   0.0   0.0   0.0   0.0
## 4 Stude~ 0.0   0.0   0.0   1.0   1.0   1.0   1.0   1.0   1.0   0.0   0.0   0.0
## 5 Stude~ 1.0   0.0   1.0   1.0   1.0   0.0   1.0   1.0   1.0   0.0   0.0   1.0
## 6 Stude~ 1.0   1.0   0.0   0.0   0.0   1.0   1.0   1.0   1.0   0.0   1.0   0.0
## # i 17 more variables: Q13 <chr>, Q14 <chr>, Q15 <chr>, Q16 <chr>, Q17 <chr>,
## #   Q18 <chr>, Q19 <chr>, Q20 <chr>, `Total Score` <chr>, `Quiz Time` <chr>,
## #   Accomodations <chr>, Sex <chr>, `Race/Ethnicity` <chr>,
## #   `English Proficiency` <chr>, `Born USA` <chr>, `Home Language` <chr>,
## #   `Age arrive USA` <chr>
```

```
colnames(df_students)
```

```
##  [1] "...1"                "Q1"                  "Q2"
##  [4] "Q3"                  "Q4"                  "Q5"
##  [7] "Q6"                  "Q7"                  "Q8"
## [10] "Q9"                  "Q10"                 "Q11"
## [13] "Q12"                 "Q13"                 "Q14"
## [16] "Q15"                 "Q16"                 "Q17"
## [19] "Q18"                 "Q19"                 "Q20"
## [22] "Total Score"         "Quiz Time"           "Accomodations"
## [25] "Sex"                 "Race/Ethnicity"      "English Proficiency"
## [28] "Born USA"            "Home Language"       "Age arrive USA"
```

```
# Convert value fromat to numeric
df_students$`Total Score` <- as.numeric(df_students$`Total Score`)
df_students$`Quiz Time` <- as.numeric(df_students$`Quiz Time`)
df_students$Accomodations <- as.numeric(df_students$Accomodations)
df_students$Sex <- as.numeric(df_students$Sex)
df_students$`Race/Ethnicity` <- as.numeric(df_students$`Race/Ethnicity`)
df_students$`English Proficiency` <- as.numeric(df_students$`English Proficiency`)
df_students$`Born USA` <- as.numeric(df_students$`Born USA`)
df_students$`Home Language` <- as.numeric(df_students$`Home Language`)
df_students$`Age arrive USA`<- as.numeric(df_students$`Age arrive USA`)


# Convert Quiz Time from seconds to minutes
df_students$`Quiz Time Minutes` <- df_students$`Quiz Time` / 60 # for later test
# Create Quiz Time Bins
df_students$`Quiz Time Group` <- cut(df_students$`Quiz Time Minutes`,
                                     breaks = c(0, 10, 20, Inf),
                                     labels = c("0-10 min", "10-20 min", "20+ min"),
                                     include.lowest = TRUE)

# Check new groups
table(df_students$`Quiz Time Group`)
```

```
##
##  0-10 min 10-20 min   20+ min
##         7         7        17
```

```
# Verify changes
str(df_students)
```

```
## tibble [31 x 32] (S3: tbl_df/tbl/data.frame)
##  $ ...1                : chr [1:31] "Student 1" "Student 2" "Student 3" "Student 4" ...
##  $ Q1                  : chr [1:31] "1.0" "1.0" "0.0" "0.0" ...
##  $ Q2                  : chr [1:31] "0.0" "1.0" "1.0" "0.0" ...
##  $ Q3                  : chr [1:31] "0.0" "0.0" "0.0" "0.0" ...
##  $ Q4                  : chr [1:31] "0.0" "0.0" "1.0" "1.0" ...
##  $ Q5                  : chr [1:31] "0.0" "1.0" "1.0" "1.0" ...
##  $ Q6                  : chr [1:31] "1.0" "1.0" "1.0" "1.0" ...
##  $ Q7                  : chr [1:31] "1.0" "1.0" "1.0" "1.0" ...
##  $ Q8                  : chr [1:31] "0.0" "1.0" "1.0" "1.0" ...
##  $ Q9                  : chr [1:31] "1.0" "1.0" "0.0" "1.0" ...
##  $ Q10                 : chr [1:31] "0.0" "0.0" "0.0" "0.0" ...
##  $ Q11                 : chr [1:31] "1.0" "0.0" "0.0" "0.0" ...
##  $ Q12                 : chr [1:31] "1.0" "1.0" "0.0" "0.0" ...
##  $ Q13                 : chr [1:31] "1.0" "1.0" "1.0" "1.0" ...
##  $ Q14                 : chr [1:31] "1.0" "1.0" "1.0" "1.0" ...
##  $ Q15                 : chr [1:31] "1.0" "1.0" "1.0" "1.0" ...
##  $ Q16                 : chr [1:31] "1.0" "1.0" "1.0" "1.0" ...
##  $ Q17                 : chr [1:31] "1.0" "0.0" "1.0" "1.0" ...
##  $ Q18                 : chr [1:31] "1.0" "0.0" "1.0" "1.0" ...
##  $ Q19                 : chr [1:31] "1.0" "1.0" "1.0" "1.0" ...
##  $ Q20                 : chr [1:31] "1.0" "1.0" "1.0" "1.0" ...
##  $ Total Score         : num [1:31] 0.7 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 ...
##  $ Quiz Time           : num [1:31] 1463 2574 2744 93675 38755 ...
##  $ Accomodations       : num [1:31] 2 2 2 1 2 2 2 2 2 2 ...
##  $ Sex                 : num [1:31] 0 1 0 1 1 0 1 1 1 1 ...
##  $ Race/Ethnicity      : num [1:31] 1 1 1 0 0 1 0 1 1 0 ...
##  $ English Proficiency: num [1:31] 0 1 1 0 0 0 0 0 1 1 ...
##  $ Born USA            : num [1:31] 0 0 1 0 1 0 0 0 0 0 ...
##  $ Home Language       : num [1:31] 0 1 1 0 1 0 0 0 1 0 ...
##  $ Age arrive USA      : num [1:31] 0 0 4 0 3 0 0 0 0 0 ...
##  $ Quiz Time Minutes   : num [1:31] 24.4 42.9 45.7 1561.2 645.9 ...
##  $ Quiz Time Group     : Factor w/ 3 levels "0-10 min","10-20 min",..: 3 3 3 3 3 2 3 2 3 2 ...
```

T-test 1.English Proiciency

```
group1 <- df_students$`Total Score`[df_students$`English Proficiency` == 0]
group2 <- df_students$`Total Score`[df_students$`English Proficiency` == 1]

t_test_result <- t.test(group1, group2, var.equal = TRUE)
print(t_test_result)
```

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = 1.0918, df = 29, p-value = 0.2839
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06362701  0.20935350
## sample estimates:
## mean of x mean of y
## 0.4805556 0.4076923
```

2.Race

```
group1 <- df_students$`Total Score`[df_students$`Race/Ethnicity` == 0]
group2 <- df_students$`Total Score`[df_students$`Race/Ethnicity` == 1]

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

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = -0.097293, df = 29, p-value = 0.9232
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1458685  0.1326206
## sample estimates:
## mean of x mean of y
## 0.4461538 0.4527778
```

3.Home Language

```
group1 <- df_students$`Total Score`[df_students$`Home Language` == 0]
group2 <- df_students$`Total Score`[df_students$`Home Language` == 1]

t_HomeLanguage <- t.test(group1, group2, var.equal = TRUE)
print(t_HomeLanguage)
```

```
##
##  Two Sample t-test
##
## data:  group1 and group2
## t = -0.098565, df = 29, p-value = 0.9222
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1478624  0.1342659
## sample estimates:
## mean of x mean of y
## 0.4473684 0.4541667
```

U-test 1.Sex

```
u_test_sex <- wilcox.test(df_students$`Total Score` ~ df_students$Sex, exact = FALSE)

print(u_test_sex)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df_students$‘Total Score‘ by df_students$Sex
## W = 125, p-value = 0.139
## alternative hypothesis: true location shift is not equal to 0
```

Sex is 0.139, which is not verified to the U test from client.

2.Born USA

```
u_test_born <- wilcox.test(df_students$`Total Score` ~ df_students$`Born USA`, exact = FALSE)

print(u_test_born)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df_students$‘Total Score‘ by df_students$‘Born USA‘
## W = 96.5, p-value = 0.5675
## alternative hypothesis: true location shift is not equal to 0
```

The p-value of Born in USA is also different 3.Accomodations

```
u_test_accomodations <- wilcox.test(df_students$`Total Score` ~ df_students$Accomodations, exact = FALSE

print(u_test_accomodations)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df_students$‘Total Score‘ by df_students$Accomodations
## W = 29, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

We can find in all those different grouping method, the P-value is larger than 0.05, which means there is no such a huge difference between diffient groups.