R Summary By Groups, One Variable All Statistics

Fan Wang

2020-04-01

Contents

One Variable Group Summary

Go to the **RMD**, **R**, **PDF**, or **HTML** version of this file. Go back to fan's REconTools Package, R4Econ Repository (bookdown site), or Intro Stats with R Repository.

There is a categorical variable (based on one or the interaction of multiple variables), there is a continuous variable, obtain statistics for the continuous variable conditional on the categorical variable, but also unconditionally.

Store results in a matrix, but also flatten results wide to row with appropriate keys/variable-names for all group statistics.

Pick which statistics to be included in final wide row

```
# Single Variable Group Statistics (also generate overall statistics)
ff_summ_by_group_summ_one <- function(
  df, vars.group, var.numeric, str.stats.group = 'main',
  str.stats.specify = NULL, boo.overall.stats = TRUE){

    # List of statistics
    # https://rdrr.io/cran/dplyr/man/summarise.html
    strs.center <- c('mean', 'median')
    strs.spread <- c('sd', 'IQR', 'mad')
    strs.range <- c('min', 'max')
    strs.pos <- c('first', 'last')
    strs.count <- c('n_distinct')

    # Grouping of Statistics
    if (missing(str.stats.specify)) {
        if (str.stats.group == 'main') {
            strs.all <- c('mean', 'min', 'max', 'sd')
        }
        if (str.stats.group == 'all') {
            strs.all <- c(strs.center, strs.spread, strs.range, strs.pos, strs.count)
        }
    } else {
        strs.all <- str.stats.specify
    }
```

```
# Start Transform
df <- df %>% drop_na() %>% mutate(!!(var.numeric) := as.numeric(!!sym(var.numeric)))

# Overall Statistics
if (boo.overall.stats) {
    df.overall.stats <- df %>% summarize_at(vars(var.numeric), funs(!!!strs.all))
    if (length(strs.all) == 1) {
        # give it a name, otherwise if only one stat, name of stat not saved
        df.overall.stats <- df.overall.stats %>% rename(!!strs.all := !!sym(var.numeric))
    }
    names(df.overall.stats) <- paste0(var.numeric, '.', names(df.overall.stats))
}

# Group Sort
df.select <- df %>%
             group_by(!!!syms(vars.group)) %>%
             arrange(!!!syms(c(vars.group, var.numeric)))

# Table of Statistics
df.table.grp.stats <- df.select %>% summarize_at(vars(var.numeric), funs(!!!strs.all))

# Add Stat Name
if (length(strs.all) == 1) {
    # give it a name, otherwise if only one stat, name of stat not saved
    df.table.grp.stats <- df.table.grp.stats %>% rename(!!strs.all := !!sym(var.numeric))
}

# Row of Statistics
str.vars.group.combine <- paste0(vars.group, collapse='_')
if (length(vars.group) == 1) {
    df.row.grp.stats <- df.table.grp.stats %>%
            mutate(!!(str.vars.group.combine) := paste0(var.numeric, '.',
                                        vars.group, '.g',
                                        (!!!syms(vars.group)))) %>%
            gather(variable, value, -one_of(vars.group)) %>%
            unite(str.vars.group.combine, c(str.vars.group.combine, 'variable')) %>%
            spread(str.vars.group.combine, value)
} else {
    df.row.grp.stats <- df.table.grp.stats %>%
      mutate(vars.groups.combine := paste0(paste0(vars.group, collapse='.')),
             !!(str.vars.group.combine) := paste0(interaction(!!!(syms(vars.group))))) %>%
      mutate(!!(str.vars.group.combine) := paste0(var.numeric, '.', vars.groups.combine, '.',
                                                  (!!sym(str.vars.group.combine)))) %>%
      ungroup() %>%
      select(-vars.groups.combine, -one_of(vars.group)) %>%
      gather(variable, value, -one_of(str.vars.group.combine)) %>%
      unite(str.vars.group.combine, c(str.vars.group.combine, 'variable')) %>%
      spread(str.vars.group.combine, value)
}

# Clean up name strings
names(df.table.grp.stats) <-
  gsub(x = names(df.table.grp.stats),pattern = "_", replacement = "\\.")
```

```
    names(df.row.grp.stats) <-
      gsub(x = names(df.row.grp.stats),pattern = "_", replacement = "\\.")

    # Return
    list.return <-
      list(df_table_grp_stats = df.table.grp.stats, df_row_grp_stats = df.row.grp.stats)

    # Overall Statistics, without grouping
    if (boo.overall.stats) {
        df.row.stats.all <- c(df.row.grp.stats, df.overall.stats)
        list.return <- append(list.return, list(df_overall_stats = df.overall.stats,
                                                df_row_stats_all = df.row.stats.all))
    }

    # Return
    return(list.return)

}
```

Build Program

Test Load data and test

```
# Library
library(tidyverse)

# Load Sample Data
setwd('C:/Users/fan/R4Econ/_data/')
df <- read_csv('height_weight.csv')
```

```
## Parsed with column specification:
## cols(
##   S.country = col_character(),
##   vil.id = col_double(),
##   indi.id = col_double(),
##   sex = col_character(),
##   svymthRound = col_double(),
##   momEdu = col_double(),
##   wealthIdx = col_double(),
##   hgt = col_double(),
##   wgt = col_double(),
##   hgt0 = col_double(),
##   wgt0 = col_double(),
##   prot = col_double(),
##   cal = col_double(),
##   p.A.prot = col_double(),
##   p.A.nProt = col_double()
## )
```

Function Testing By Gender Groups Need two variables, a group variable that is a factor, and a numeric

```
vars.group <- 'sex'
var.numeric <- 'hgt'
```

```
df.select <- df %>% select(one_of(vars.group, var.numeric)) %>% drop_na()
```

Main Statistics:

```
# Single Variable Group Statistics
ff_summ_by_group_summ_one(
  df.select, vars.group = vars.group, var.numeric = var.numeric,
  str.stats.group = 'main')
```

```
## $df_table_grp_stats
## # A tibble: 2 x 5
##   sex     mean   min   max    sd
##   <chr>  <dbl> <dbl> <dbl> <dbl>
## 1 Female  82.8  41.2  171.  29.8
## 2 Male    84.7  41.3  183.  31.8
##
## $df_row_grp_stats
## # A tibble: 1 x 8
##   hgt.sex.gFemale.max hgt.sex.gFemale.mean hgt.sex.gFemale.min hgt.sex.gFemale.sd hgt.sex.gMale.max h
##                 <dbl>                <dbl>               <dbl>              <dbl>             <dbl>
## 1                171.                 82.8                41.2               29.8              183.
##
## $df_overall_stats
## # A tibble: 1 x 4
##   hgt.mean hgt.min hgt.max hgt.sd
##      <dbl>   <dbl>   <dbl>  <dbl>
## 1     83.8    41.2    183.   30.9
##
## $df_row_stats_all
## $df_row_stats_all$hgt.sex.gFemale.max
## [1] 170.6
##
## $df_row_stats_all$hgt.sex.gFemale.mean
## [1] 82.81198
##
## $df_row_stats_all$hgt.sex.gFemale.min
## [1] 41.2
##
## $df_row_stats_all$hgt.sex.gFemale.sd
## [1] 29.79351
##
## $df_row_stats_all$hgt.sex.gMale.max
## [1] 182.9
##
## $df_row_stats_all$hgt.sex.gMale.mean
## [1] 84.68152
##
## $df_row_stats_all$hgt.sex.gMale.min
## [1] 41.3
##
## $df_row_stats_all$hgt.sex.gMale.sd
## [1] 31.75037
##
## $df_row_stats_all$hgt.mean
```

```
## [1] 83.80921
##
## $df_row_stats_all$hgt.min
## [1] 41.2
##
## $df_row_stats_all$hgt.max
## [1] 182.9
##
## $df_row_stats_all$hgt.sd
## [1] 30.86631
```

Specify Two Specific Statistics:

```
ff_summ_by_group_summ_one(
  df.select, vars.group = vars.group, var.numeric = var.numeric,
  str.stats.specify = c('mean', 'sd'))
```

```
## $df_table_grp_stats
## # A tibble: 2 x 3
##   sex     mean    sd
##   <chr>  <dbl> <dbl>
## 1 Female  82.8  29.8
## 2 Male    84.7  31.8
##
## $df_row_grp_stats
## # A tibble: 1 x 4
##   hgt.sex.gFemale.mean hgt.sex.gFemale.sd hgt.sex.gMale.mean hgt.sex.gMale.sd
##                  <dbl>              <dbl>              <dbl>            <dbl>
## 1                 82.8               29.8               84.7             31.8
##
## $df_overall_stats
## # A tibble: 1 x 2
##   hgt.mean hgt.sd
##      <dbl>  <dbl>
## 1     83.8   30.9
##
## $df_row_stats_all
## $df_row_stats_all$hgt.sex.gFemale.mean
## [1] 82.81198
##
## $df_row_stats_all$hgt.sex.gFemale.sd
## [1] 29.79351
##
## $df_row_stats_all$hgt.sex.gMale.mean
## [1] 84.68152
##
## $df_row_stats_all$hgt.sex.gMale.sd
## [1] 31.75037
##
## $df_row_stats_all$hgt.mean
## [1] 83.80921
##
## $df_row_stats_all$hgt.sd
## [1] 30.86631
```

Specify One Specific Statistics:

```
ff_summ_by_group_summ_one(
  df.select, vars.group = vars.group, var.numeric = var.numeric,
  str.stats.specify = c('mean'))
```

```
## $df_table_grp_stats
## # A tibble: 2 x 2
##   sex     mean
##   <chr>  <dbl>
## 1 Female  82.8
## 2 Male    84.7
##
## $df_row_grp_stats
## # A tibble: 1 x 2
##   hgt.sex.gFemale.mean hgt.sex.gMale.mean
##                  <dbl>              <dbl>
## 1                 82.8               84.7
##
## $df_overall_stats
## # A tibble: 1 x 1
##   hgt.mean
##      <dbl>
## 1     83.8
##
## $df_row_stats_all
## $df_row_stats_all$hgt.sex.gFemale.mean
## [1] 82.81198
##
## $df_row_stats_all$hgt.sex.gMale.mean
## [1] 84.68152
##
## $df_row_stats_all$hgt.mean
## [1] 83.80921
```

Function Testing By Country and Gender Groups Need two variables, a group variable that is a factor, and a numeric. Now joint grouping variables.

```
vars.group <- c('S.country', 'sex')
var.numeric <- 'hgt'
```

```
df.select <- df %>% select(one_of(vars.group, var.numeric)) %>% drop_na()
```

Main Statistics:

```
ff_summ_by_group_summ_one(
  df.select, vars.group = vars.group, var.numeric = var.numeric,
  str.stats.group = 'main')
```

```
## $df_table_grp_stats
## # A tibble: 4 x 6
## # Groups:   S.country [2]
##   S.country sex     mean   min   max    sd
##   <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>
## 1 Cebu      Female  84.6  41.3  171.  32.5
## 2 Cebu      Male    87.0  41.3  183.  35.0
## 3 Guatemala Female  76.6  41.2  120.  15.7
```

```
## 4 Guatemala Male    77.0  41.5  125.  15.1
##
## $df_row_grp_stats
## # A tibble: 1 x 16
##   hgt.S.country.s~ hgt.S.country.s~ hgt.S.country.s~ hgt.S.country.s~ hgt.S.country.s~ hgt.S.country
##              <dbl>            <dbl>            <dbl>            <dbl>            <dbl>            <dl
## 1             171.             84.6             41.3             32.5             183.             87
## # ... with 7 more variables: hgt.S.country.sex.Guatemala.Female.mean <dbl>, hgt.S.country.sex.Guatema
## #   hgt.S.country.sex.Guatemala.Female.sd <dbl>, hgt.S.country.sex.Guatemala.Male.max <dbl>, hgt.S.co
## #   hgt.S.country.sex.Guatemala.Male.min <dbl>, hgt.S.country.sex.Guatemala.Male.sd <dbl>
##
## $df_overall_stats
## # A tibble: 1 x 4
##   hgt.mean hgt.min hgt.max hgt.sd
##      <dbl>   <dbl>   <dbl>  <dbl>
## 1     83.8    41.2    183.   30.9
##
## $df_row_stats_all
## $df_row_stats_all$hgt.S.country.sex.Cebu.Female.max
## [1] 170.6
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Female.mean
## [1] 84.61326
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Female.min
## [1] 41.3
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Female.sd
## [1] 32.53651
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Male.max
## [1] 182.9
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Male.mean
## [1] 87.02836
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Male.min
## [1] 41.3
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Male.sd
## [1] 34.9909
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Female.max
## [1] 119.9
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Female.mean
## [1] 76.58771
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Female.min
## [1] 41.2
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Female.sd
## [1] 15.71801
##
```

```
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Male.max
## [1] 124.7
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Male.mean
## [1] 77.0471
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Male.min
## [1] 41.5
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Male.sd
## [1] 15.11444
##
## $df_row_stats_all$hgt.mean
## [1] 83.80921
##
## $df_row_stats_all$hgt.min
## [1] 41.2
##
## $df_row_stats_all$hgt.max
## [1] 182.9
##
## $df_row_stats_all$hgt.sd
## [1] 30.86631
```

Specify Two Specific Statistics:

```
ff_summ_by_group_summ_one(
  df.select, vars.group = vars.group, var.numeric = var.numeric,
  str.stats.specify = c('mean', 'sd'))
```

```
## $df_table_grp_stats
## # A tibble: 4 x 4
## # Groups:   S.country [2]
##   S.country sex     mean    sd
##   <chr>     <chr>  <dbl> <dbl>
## 1 Cebu      Female  84.6  32.5
## 2 Cebu      Male    87.0  35.0
## 3 Guatemala Female  76.6  15.7
## 4 Guatemala Male    77.0  15.1
##
## $df_row_grp_stats
## # A tibble: 1 x 8
##   hgt.S.country.sex.~ hgt.S.country.sex.~ hgt.S.country.sex.~ hgt.S.country.sex~ hgt.S.country.sex~ h
##                 <dbl>               <dbl>               <dbl>             <dbl>              <dbl>
## 1                84.6                32.5                87.0              35.0               76.6
##
## $df_overall_stats
## # A tibble: 1 x 2
##   hgt.mean hgt.sd
##      <dbl>  <dbl>
## 1     83.8   30.9
##
## $df_row_stats_all
## $df_row_stats_all$hgt.S.country.sex.Cebu.Female.mean
## [1] 84.61326
```

```
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Female.sd
## [1] 32.53651
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Male.mean
## [1] 87.02836
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Male.sd
## [1] 34.9909
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Female.mean
## [1] 76.58771
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Female.sd
## [1] 15.71801
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Male.mean
## [1] 77.0471
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Male.sd
## [1] 15.11444
##
## $df_row_stats_all$hgt.mean
## [1] 83.80921
##
## $df_row_stats_all$hgt.sd
## [1] 30.86631
```

Specify One Specific Statistics:

```
ff_summ_by_group_summ_one(
  df.select, vars.group = vars.group, var.numeric = var.numeric, str.stats.specify = c('mean'))
```

```
## $df_table_grp_stats
## # A tibble: 4 x 3
## # Groups:   S.country [2]
##   S.country sex     mean
##   <chr>     <chr>  <dbl>
## 1 Cebu      Female  84.6
## 2 Cebu      Male    87.0
## 3 Guatemala Female  76.6
## 4 Guatemala Male    77.0
##
## $df_row_grp_stats
## # A tibble: 1 x 4
##   hgt.S.country.sex.Cebu.Female.mean hgt.S.country.sex.Cebu.Male.mean hgt.S.country.sex.Guatemala.Fe
##                                <dbl>                            <dbl>
## 1                               84.6                             87.0
##
## $df_overall_stats
## # A tibble: 1 x 1
##   hgt.mean
##      <dbl>
## 1     83.8
##
```

```
## $df_row_stats_all
## $df_row_stats_all$hgt.S.country.sex.Cebu.Female.mean
## [1] 84.61326
##
## $df_row_stats_all$hgt.S.country.sex.Cebu.Male.mean
## [1] 87.02836
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Female.mean
## [1] 76.58771
##
## $df_row_stats_all$hgt.S.country.sex.Guatemala.Male.mean
## [1] 77.0471
##
## $df_row_stats_all$hgt.mean
## [1] 83.80921
```