Summarize a Quantitative/Continuous Variable with Categorical Groups

This page is from fan's R4Econ Repository. Go back to the R4Econ Repository or the Intro Stats with R Repository.

- r generate text string as csv
- r tibble matrix hand input

```
options(knitr.duplicate.label = 'allow')
```

```
rm(list = ls(all.names = TRUE))
library(tidyverse)
library(tidyr)
library(knitr)
library(kableExtra)
library(REconTools)
# file name
st_file_name = 'fst_hist_onevar'
# Generate R File
try(purl(paste0(st_file_name, ".Rmd"), output=paste0(st_file_name, ".R"), documentation = 2))
# Generate PDF and HTML
# rmarkdown::render("C:/Users/fan/R4Econ/summarize/dist/fst_hist_onevar.Rmd", "pdf_document")
# rmarkdown::render("C:/Users/fan/R4Econ/summarize/dist/fst_hist_onevar.Rmd", "html_document")
```

Generate Datasets

First, we will generate a test score dataset, directly from string. Below we type line by line a dataset with four variables in comma separated (csv) format, where the first row includes the variables names. These texts could be stored in a separate file, or they could be directly included in code and read in as csv

A Dataset with only Two Continuous Variable

```
ar_test_scores_ec3 <- c(107.72,101.28,105.92,109.31,104.27,110.27,91.92846154,81.8,109.0071429,103.07,98
ar_test_scores_ec1 <- c(101.72,101.28,99.92,103.31,100.27,104.27,90.23615385,77.8,103.4357143,97.07,93.
mt_test_scores <- cbind(ar_test_scores_ec1, ar_test_scores_ec3)
ar_st_varnames <- c('course_total_ec1p','course_total_ec3p')
tb_final_twovar <- as_tibble(mt_test_scores) %>% rename_all(~c(ar_st_varnames))
summary(tb_final_twovar)
```

```
##  course_total_ec1p course_total_ec3p
##  Min.   : 40.48    Min.   : 44.23
##  1st Qu.: 76.46    1st Qu.: 79.91
##  Median : 86.35    Median : 89.28
##  Mean   : 83.88    Mean   : 87.90
##  3rd Qu.: 95.89    3rd Qu.:100.75
##  Max.   :104.27    Max.   :112.22
```

```
ff_summ_percentiles(df = tb_final_twovar, bl_statsasrows = TRUE, col2varname = FALSE)
```

```
## # A tibble: 17 x 3
##    stats   course.total.ec1p course.total.ec3p
##    <chr>   <chr>             <chr>
##  1 n       46                46
##  2 NAobs   0                 0
##  3 ZERObs  0                 0
##  4 mean    83.87572          87.90239
##  5 sd      15.87272          16.76041
##  6 cv      0.1892409         0.1906706
##  7 min     40.475            44.225
##  8 p01     42.14434          45.82202
##  9 p05     56.9650           57.1575
## 10 p10     63.05462          66.07500
## 11 p25     76.45616          79.90500
## 12 p50     86.35236          89.27923
## 13 p75     " 95.89054"       100.75250
## 14 p90     100.8137          106.8200
## 15 p95     102.9125          109.2343
## 16 p99     103.8946          111.3439
## 17 max     104.2700          112.2225
```

A Dataset with one Continuous Variable and Histogram

```
ar_final_scores <- c(94.28442509,95.68817475,97.25219512,77.89268293,95.08795497,93.27380863,92.3,84.253
mt_test_scores <- cbind(seq(1,length(ar_final_scores)), ar_final_scores)
ar_st_varnames <- c('index', 'course_final')
tb_onevar <- as_tibble(mt_test_scores) %>% rename_all(~c(ar_st_varnames))
summary(tb_onevar)
```

```
##      index       course_final
##  Min.   : 1.0   Min.   :  2.293
##  1st Qu.:12.5   1st Qu.: 76.372
##  Median :24.0   Median : 86.959
##  Mean   :24.0   Mean   : 82.415
##  3rd Qu.:35.5   3rd Qu.: 94.686
##  Max.   :47.0   Max.   :100.898
```

```
ff_summ_percentiles(df = tb_onevar, bl_statsasrows = TRUE, col2varname = FALSE)
```

```
## # A tibble: 17 x 3
##    stats   course.final index
##    <chr>   <chr>        <chr>
##  1 n       47           47
##  2 NAobs   0            0
##  3 ZERObs  0            0
##  4 mean    82.41501     24.00000
##  5 sd      18.35476     13.71131
##  6 cv      0.2227113    0.5713046
##  7 min     2.292683     1.000000
##  8 p01     18.67401     " 1.46000"
##  9 p05     49.72075     " 3.30000"
## 10 p10     66.28051     " 5.60000"
```

```
## 11 p25     76.37177     12.50000
## 12 p50     86.95932     24.00000
## 13 p75     94.68619     35.50000
## 14 p90     97.52332     42.40000
## 15 p95     99.47459     44.70000
## 16 p99     100.5244     " 46.5400"
## 17 max     100.898      " 47.000"
```

A Dataset with Multiple Variables

```
#load in data empirically by hand
txt_test_data <- "init_prof, later_prof, class_id, exam_score
 'SW', 'SW', 1, 102
 'SW', 'SW', 1, 102
 'SW', 'SW', 1, 101
 'SW', 'SW', 1, 100
 'SW', 'SW', 1, 100
 'SW', 'SW', 1, 99
 'SW', 'SW', 1, 98.5
 'SW', 'SW', 1, 98.5
 'SW', 'SW', 1, 97
 'SW', 'SW', 1, 95
 'SW', 'SW', 1, 94
 'SW', 'SW', 1, 91
 'SW', 'SW', 1, 91
 'SW', 'SW', 1, 90
 'SW', 'SW', 1, 89
 'SW', 'SW', 1, 88.5
 'SW', 'SW', 1, 88
 'SW', 'SW', 1, 87
 'SW', 'SW', 1, 87
 'SW', 'SW', 1, 87
 'SW', 'SW', 1, 86
 'SW', 'SW', 1, 86
 'SW', 'SW', 1, 84
 'SW', 'SW', 1, 82
 'SW', 'SW', 1, 78.5
 'SW', 'SW', 1, 76
 'SW', 'SW', 1, 72
 'SW', 'SW', 1, 70.5
 'SW', 'SW', 1, 67.5
 'SW', 'SW', 1, 67.5
 'SW', 'SW', 1, 67
 'SW', 'SW', 1, 63.5
 'SW', 'SW', 1, 60
 'SW', 'SW', 1, 59
 'SW', 'SW', 1, 44.5
 'SW', 'SW', 1, 44
 'SW', 'SW', 1, 42.5
 'SW', 'SW', 1, 40.5
 'SW', 'SW', 1, 40.5
 'SW', 'SW', 1, 36.5
 'SW', 'SW', 1, 35.5
 'SW', 'SW', 1, 21.5
```

```
'SW', 'SW', 1, 4
'MP', 'MP', 2, 105
'MP', 'MP', 2, 103
'MP', 'MP', 2, 102
'MP', 'MP', 2, 101
'MP', 'MP', 2, 101
'MP', 'MP', 2, 100.5
'MP', 'MP', 2, 100
'MP', 'MP', 2, 99
'MP', 'MP', 2, 97
'MP', 'MP', 2, 97
'MP', 'MP', 2, 97
'MP', 'MP', 2, 97
'MP', 'MP', 2, 96
'MP', 'MP', 2, 95
'MP', 'MP', 2, 91
'MP', 'MP', 2, 89
'MP', 'MP', 2, 85
'MP', 'MP', 2, 84
'MP', 'MP', 2, 84
'MP', 'MP', 2, 84
'MP', 'MP', 2, 83.5
'MP', 'MP', 2, 82.5
'MP', 'MP', 2, 81.5
'MP', 'MP', 2, 80.5
'MP', 'MP', 2, 80
'MP', 'MP', 2, 77
'MP', 'MP', 2, 77
'MP', 'MP', 2, 75
'MP', 'MP', 2, 75
'MP', 'MP', 2, 71
'MP', 'MP', 2, 70
'MP', 'MP', 2, 68
'MP', 'MP', 2, 63
'MP', 'MP', 2, 56
'MP', 'MP', 2, 56
'MP', 'MP', 2, 55.5
'MP', 'MP', 2, 49.5
'MP', 'MP', 2, 48.5
'MP', 'MP', 2, 47.5
'MP', 'MP', 2, 44.5
'MP', 'MP', 2, 34.5
'MP', 'MP', 2, 29.5
'CA', 'MP', 3, 103
'CA', 'MP', 3, 103
'CA', 'MP', 3, 101
'CA', 'MP', 3, 96.5
'CA', 'MP', 3, 93.5
'CA', 'MP', 3, 93
'CA', 'MP', 3, 93
'CA', 'MP', 3, 92
'CA', 'MP', 3, 90
'CA', 'MP', 3, 90
```

```
'CA', 'MP', 3, 89
'CA', 'MP', 3, 86.5
'CA', 'MP', 3, 84.5
'CA', 'MP', 3, 83
'CA', 'MP', 3, 83
'CA', 'MP', 3, 82
'CA', 'MP', 3, 78
'CA', 'MP', 3, 75
'CA', 'MP', 3, 74.5
'CA', 'MP', 3, 70
'CA', 'MP', 3, 54.5
'CA', 'MP', 3, 52
'CA', 'MP', 3, 50
'CA', 'MP', 3, 42
'CA', 'MP', 3, 36.5
'CA', 'MP', 3, 28
'CA', 'MP', 3, 26
'CA', 'MP', 3, 11
'CA', 'SN', 4, 103
'CA', 'SN', 4, 103
'CA', 'SN', 4, 102
'CA', 'SN', 4, 102
'CA', 'SN', 4, 101
'CA', 'SN', 4, 100
'CA', 'SN', 4, 98
'CA', 'SN', 4, 98
'CA', 'SN', 4, 98
'CA', 'SN', 4, 95
'CA', 'SN', 4, 95
'CA', 'SN', 4, 92.5
'CA', 'SN', 4, 92
'CA', 'SN', 4, 91
'CA', 'SN', 4, 90
'CA', 'SN', 4, 85.5
'CA', 'SN', 4, 84
'CA', 'SN', 4, 82.5
'CA', 'SN', 4, 81
'CA', 'SN', 4, 77.5
'CA', 'SN', 4, 77
'CA', 'SN', 4, 72
'CA', 'SN', 4, 71.5
'CA', 'SN', 4, 69
'CA', 'SN', 4, 68.5
'CA', 'SN', 4, 68
'CA', 'SN', 4, 67
'CA', 'SN', 4, 65.5
'CA', 'SN', 4, 62.5
'CA', 'SN', 4, 62
'CA', 'SN', 4, 61.5
'CA', 'SN', 4, 61
'CA', 'SN', 4, 57.5
'CA', 'SN', 4, 54
'CA', 'SN', 4, 52.5
```

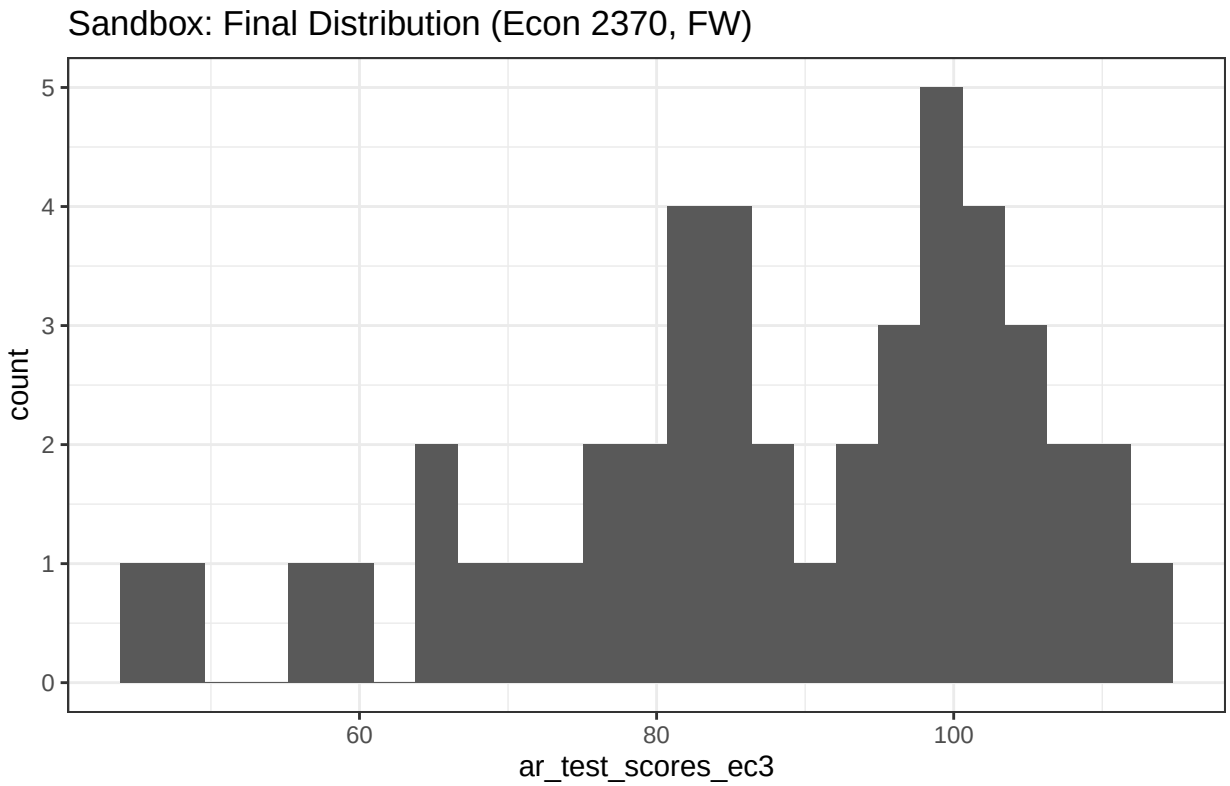
```
 'CA', 'SN', 4, 51
 'CA', 'SN', 4, 50.5
 'CA', 'SN', 4, 50
 'CA', 'SN', 4, 49
 'CA', 'SN', 4, 43
 'CA', 'SN', 4, 39.5
 'CA', 'SN', 4, 32.5
 'CA', 'SN', 4, 25.5
 'CA', 'SN', 4, 18"

csv_test_data = read.csv(text=txt_test_data, header=TRUE)
ar_st_varnames <- c('first_half_professor', 'second_half_professor', 'course_id', 'exam_score')
tb_test_data <- as_tibble(csv_test_data) %>% rename_all(~c(ar_st_varnames))
summary(tb_test_data)
```

```
##  first_half_professor second_half_professor   course_id       exam_score
##   'CA':72              'MP':70              Min.   :1.000   Min.   :  4.00
##   'MP':42              'SN':44              1st Qu.:1.000   1st Qu.: 60.00
##   'SW':43              'SW':43              Median :2.000   Median : 82.00
##                                             Mean   :2.465   Mean   : 75.08
##                                             3rd Qu.:4.000   3rd Qu.: 94.00
##                                             Max.   :4.000   Max.   :105.00
```

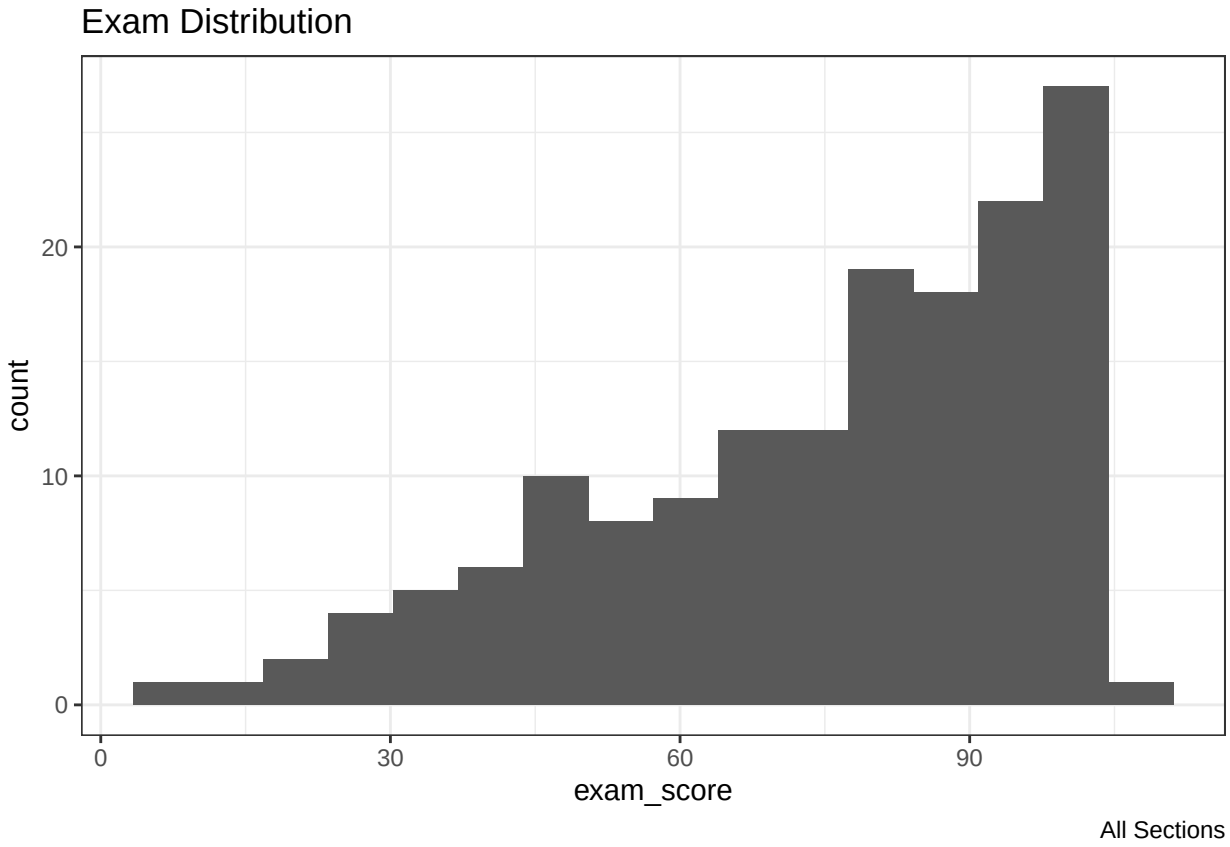
Analyze Test Scores Distribution

Histograms

```
ggplot(tb_final_twovar, aes(x=ar_test_scores_ec3)) +
  geom_histogram(bins=25) +
  labs(title = paste0('Sandbox: Final Distribution (Econ 2370, FW)'),
       caption = 'FW Section, formula: 0.3*exam1Perc + 0.3*exam2Perc + 0.42*HWtotalPerc + 0.03*Attendan
  theme_bw()
```




```
ggplot(tb_test_data, aes(x=exam_score)) +
  geom_histogram(bins=16) +
  labs(title = paste0('Exam Distribution'),
       caption = 'All Sections') +
  theme_bw()
```


Exam Distribution
count
20
10
0
0
30
60
90
exam_score
All Sections