Negative Binomial Regression with R

Introduction Negative binomial regression is for modeling count variables, usually for over-dispersed count outcome variables.

Examples of negative binomial regression

- **Example 1** School administrators study the attendance behavior of high school juniors at two schools.
  Predictors of the number of days of absence include the type of program in which the student is enrolled and a standardized test in math.
- **Example 2** A health-related researcher is studying the number of hospital visits in past 12 months by senior citizens in a community based on the characteristics of the individuals and the types of health plans under which each one is covered.

Negative Binomial Regression with R

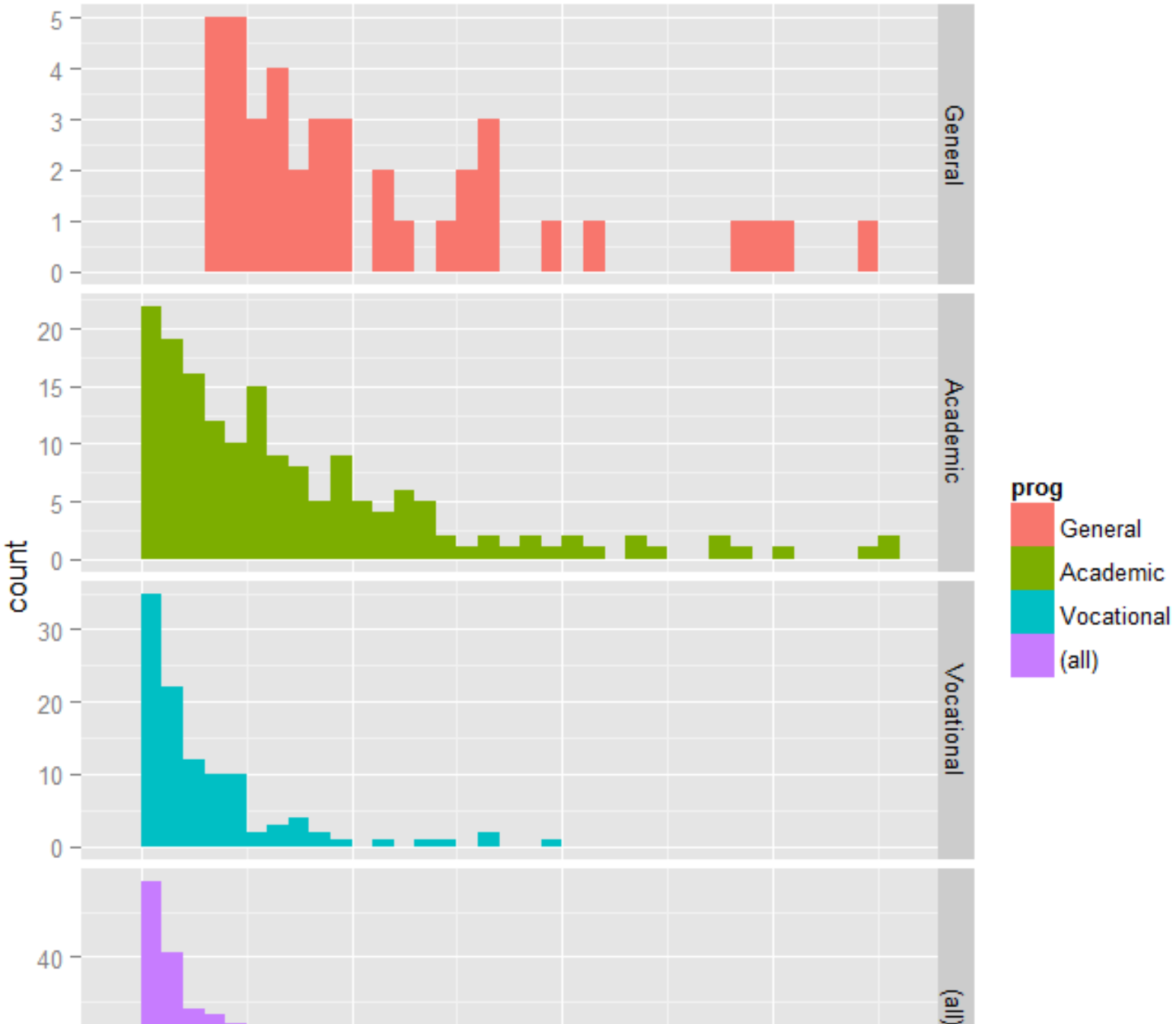
Description of the data Let's pursue Example 1 from above.

- We have attendance data on 314 high school juniors from two urban high schools in the file **negbin.csv** .
- The response variable of interest is days absent, **daysabs**.
- The variable **math** gives the standardized math score for each student.
- The variable **prog** is a three-level nominal variable indicating the type of instructional program in which the student is enrolled.

Negative Binomial Regression with R

Exploratory Data Analysis

```
summary(dat)
       id          gender         math          daysabs
  1001   :  1   female:160   Min.   : 1.0   Min.   : 0.00
  1002   :  1   male  :154   1st Qu.:28.0   1st Qu.: 1.00
  1003   :  1                Median :48.0   Median : 4.00
  1004   :  1                Mean   :48.3   Mean   : 5.96
  1005   :  1                3rd Qu.:70.0   3rd Qu.: 8.00
  1006   :  1                Max.   :99.0   Max.   :35.00
  (Other):308
         prog
  General   : 40
  Academic  :167
  Vocational:107
```


count
General
Academic
Vocational
(all)
prog
General
Academic
Vocational
(all)

Negative Binomial Regression with R

```
ggplot(dat, aes(daysabs, fill = prog)) + geom_
., margins = TRUE, scales = "free")
```

Histogram plots showing distribution of the data
Each variable has 314 valid observations and their distributions seem quite reasonable. The unconditional mean of our outcome variable is much lower than its variance.