Print

Demo: Examining the Distribution of Categorical Variables Using PROC FREQ and PROC UNIVARIATE

Filename: **st107d01.sas**

Let's examine the distribution of Bonus at each value of the predictors, Fireplaces and Lot_Shape_2. We'll create one-way frequency tables to view the frequency of the levels of each categorical variable. We'll then create two-way frequency tables, also known as crosstabulation tables, to look for a possible association between two categorical variables. The crosstabulation table shows frequency statistics for the combinations of levels of two variables.

PROC FREQ DATA=*SAS-data-set***;**
TABLES *table-request(s)* < */ options*>;
< additional statements >
RUN;

PROC UNIVARIATE DATA=*SAS-data-set <options>***;**
VAR *variables*;
HISTOGRAM *variables* < */ options*>;
INSET *keywords* < */options*>;
RUN;

1. Open program st107d01.sas.

```
/*st107d01.sas*/
title;
proc format;
    value bonusfmt 1 = "Bonus Eligible"
                   0 = "Not Bonus Eligible"
                  ;
run;

proc freq data=STAT1.ameshousing3;
    tables Bonus Fireplaces Lot_Shape_2
           Fireplaces*Bonus Lot_Shape_2*Bonus/
           plots(only)=freqplot(scale=percent);
    format Bonus bonusfmt.;
run;

proc univariate data=STAT1.ameshousing3 noprint;
    class Bonus;
    var Basement_Area ;
    histogram Basement_Area;
    inset mean std median min max / format=5.2 position=nw;
    format Bonus bonusfmt.;
run;
```

To start, we use PROC FORMAT to format the values of Bonus. If the value is 1, SAS displays Bonus

Eligible, and if the value is 0, SAS displays Not Bonus Eligible.

The PROC FREQ step uses the ameshousing3 data set to generate frequency tables and plots summarizing the categorical variables. The TABLES statement requests individual tables for the three categorical variables Bonus, Fireplaces, and Lot_Shape_2. To request crosstabulation tables, we specify an asterisk between the names of the variables that we want to appear in the table. The first variable represents the rows, and the second variable represents the columns. This TABLES statement requests a crosstabulation of Fireplaces by Bonus, and a table of Lot_Shape_2 by Bonus.

The PLOTS= option requests a frequency plot for each frequency table, and SCALE=PERCENT displays percentages, or relative frequencies.

The FORMAT statement applies the bonusfmt. format to the variable Bonus.

Because we also want to look at the distribution of the continuous variable Basement_Area by Bonus status (eligible or not eligible), well use PROC UNIVARIATE to create histograms for each level of Bonus. The CLASS statement indicates our categorical predictor variable. The VAR and HISTOGRAM statements specify Basement_Area, and we use the INSET statement to create a box of summary statistics in the northwest corner of the graph. Again, well format the values of Bonus.

2. Submit the code.

3. Review the output.

The first table is a one-way frequency table for Bonus. By default, four types of frequency measures are included in the table for each level of the variable. We see the frequency and percent of each level, as well as the cumulative frequency and cumulative percent. The last row always displays 100; 100% of the observations contain the last value and all other values listed above it. From this table, you can see that most homes, 85% of the ones in our sample, are not Bonus Eligible, meaning they didn't sell for more than $175,000.

The second table, a one-way frequency table for fireplaces, shows that most homes do not have a fireplace. Only 31% of homes in our sample have a single fireplace, and only 12 homes have 2 fireplaces.

The third one-way frequency table analyzes Lot_Shape_2. Approximately two-thirds of homes in our sample have a regular lot shape, and the other third have an irregular lot shape. Notice there's one missing value for the Lot_Shape_2 variable.

Tables 4 and 5 are the requested crosstabulation tables. By default, a crosstabulation table has four measures in each cell, indicated in the legend. Frequency indicates the number of observations that contain both the row variable value and the column variable value. We'll use Table 4, Fireplaces by Bonus as an example. It shows that 25 homes with 1 fireplace are bonus eligible. Percent indicates the number of observations in each cell as a percentage of the total number of observations. For example, about 3% of homes with 2 fireplaces are not bonus eligible. Row Pct indicates the number of observations in each cell as a percentage of the total number of observations in that row. The total number of observations for each row appears in the Total column for the row. In this table, the first row indicates that, of the 195 homes that do not have a fireplace, about 91% are not bonus eligible. And Col Pct indicates the number of observations in each cell as a percentage of the total number of observations in that column. The total number of observations for each column appears at the bottom. Of the 255 homes that are not bonus eligible, about 27% have one fireplace.

It seems there's some association between the variables Bonus and number of Fireplaces. For example, homes that are not bonus eligible are much more likely to have 0 fireplaces, at about 69%. Whereas bonus eligible homes are more likely to have 1 fireplace, at about 55%. With the unequal group sizes, the row percentages might not easily display if Fireplaces is associated with Bonus.

The cross-tabular frequency plot displays the simple frequencies from the crosstabulations. For example, the largest bar shows that 59% of all homes in the sample have 0 fireplaces and are not bonus eligible.

Now consider Table 5, the Bonus by Lot_Shape_2 crosstabulation table. When you compare the row percentages, there's a much larger probability of the home being not bonus eligible if the lot shape is regular, at about 94%, as opposed to only 67% for irregular lot shapes. The distribution of Bonus changes when the value of Lot_Shape_2 changes. There seems to be an association between the two variables. Later, we'll investigate this further to find whether the association is statistically significant.

The PROC UNIVARIATE histogram plot shows the distribution of the continuous variable, Basement_Area, by Bonus status. The distribution of homes that are not bonus eligible appears to be more variable and has a larger standard deviation. There certainly appears to be an association between Bonus and Basement_Area. The larger the basement area, the more likely the home is to be bonus eligible. The histograms are different and centered in different locations. The median of bonus eligible homes is almost 500-square feet larger than homes that are not bonus eligible.

Close