Week 2: Data Summary and Visualization

2. Categorical variables II

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

1. Today: Relationship between two categorical variables

2. Main ideas
    1. Joint, marginal and conditional distribution
    2. Independence of two categorical variables

3. Summary

Below is the contingency table of the two variables ‘cut‘ and ‘color‘ in the diamonds dataset from yesterday.

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Calculate the proportion of diamonds that have
'cut' *Fair* **and** 'color' *E*

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

$$P(\text{cut} = \text{Fair, color} = \text{E}) = \frac{224}{53940} = 0.004$$

Calculate the proportion of diamonds that have
'cut' *Good* **and** 'color' *E*

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

$$P(\text{cut} = \text{Good},\ \text{color} = \text{E}) = \frac{933}{53940} = 0.017$$

Calculate the proportion of diamonds that have 'cut' *Very Good* **and** 'color' *F*

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

$$P(\text{cut} = \text{Very Good, color} = \text{F}) = \frac{2164}{53940} = 0.04$$

If move the red box around over all possible entries in the table, we get the **joint distribution** of the ‘cut‘ and ‘color‘ variables.

Mathematically, for all possible combination of levels of ‘cut‘ and ‘color‘, compute

$$P(\text{cut} = \text{Very Good, color} = \text{F})$$
$$P(\text{cut} = \text{Good, color} = \text{D})$$
$$P(\text{cut} = \text{Fair, color} = \text{E})$$
$$\vdots$$

In other words, when we compute the joint distribution, we are really asking

What proportion of the data fall in each combination of levels of the ‘cut‘ and ‘color‘ variables?

Calculate the proportion of diamonds fall in 'cut' *Fair* (aggregating across all values of 'color')

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

$$P(\text{cut} = \text{Fair}) = \frac{1509}{53940} = 0.028$$

Calculate the proportion of diamonds fall in 'cut' *Very Good* (aggregating across all values of 'color')

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

$$P(\text{cut} = \text{Very Good}) = \frac{12082}{53940} = 0.224$$

If move the red box around over all possible columns in the table, we get the **marginal distribution** of the 'cut' variable.

Mathematically, for all possible levels of 'cut', compute

$$
\begin{gathered}
P(\text{cut} = \text{Very Good}) \\
P(\text{cut} = \text{Good}) \\
P(\text{cut} = \text{Fair}) \\
\vdots
\end{gathered}
$$

In other words, when we compute the marginal distribution, we are really asking

What proportion of the observational units fall into each level of the 'cut' variable (aggregating across all values of 'color')?

Among those cases where the diamonds have ‘cut‘ *Fair* , calculate the proportion of diamonds that have ‘color‘ *E*

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

$$P(\text{color} = \text{E} \mid \text{cut} = \text{Fair}) = \frac{224}{1509} = 0.148$$

Among those cases where the diamonds have ‘cut‘ *Fair* , calculate the proportion of diamonds that have ‘color‘ *F*

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

$$P(\text{color} = \text{F} \mid \text{cut} = \text{Fair}) = \frac{312}{1509} = 0.207$$

If move the red box around, over all possible entries in the blue column in the table, we get the **conditional distribution** of the ‘color‘ variable given that the diamonds have the ‘cut‘ Fair.

Mathematically, for all possible levels of ‘cut‘, compute

$$
\begin{gathered}
P(\text{color} = \text{D} \mid \text{cut} = \text{Fair}) \\
P(\text{color} = \text{E} \mid \text{cut} = \text{Fair}) \\
P(\text{color} = \text{F} \mid \text{cut} = \text{Fair}) \\
\vdots
\end{gathered}
$$

In other words, when we compute the conditional distribution over the variable ‘cut‘ equals Fair, we are really asking

Among those cases where the ‘cut‘ is ’Fair’, what proportion of the observational units fall in each level of the ‘color‘ variable?

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are not.

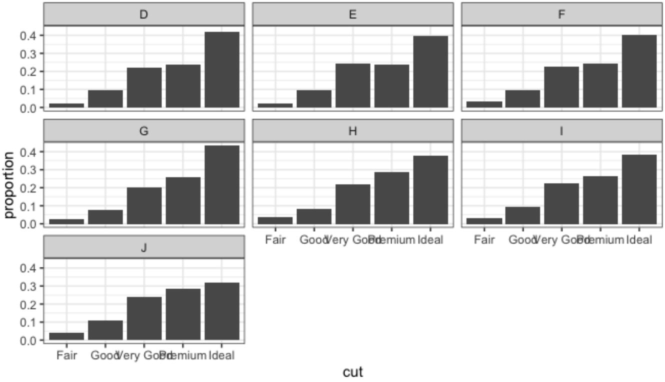
Independence (there's no association between these variables)

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are not.

Independence (there's no association between these variables)

Two variables are independent when the distribution of one does not depend on the the other.

For each color, we construct a barplot that shows the distribution of ‘Cut‘.



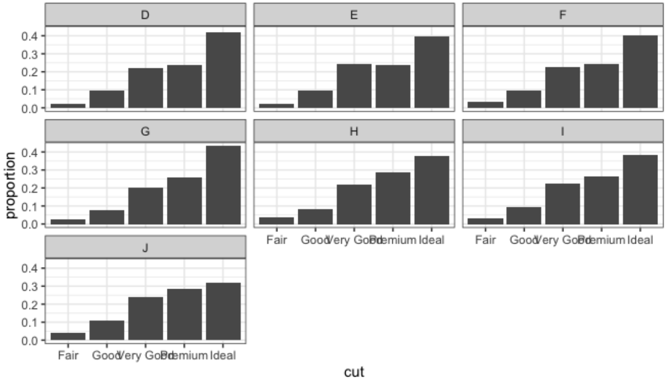
What did you notice?

The barplots seem to be roughly the same for each color. It does not seem like the distribution of Cut changes significantly depending on the color; that is, it **appears** that the variables ‘Color‘ and ‘Cut‘ are independent.

Summary

1. for each level of variable A, construct a bar plot of variable B using the data that fall into that level of A.
2. if the boxplots are roughly the same, we can expect that two variables are independent.

For each color, we construct a barplot that shows the distribution of ‘Cut‘.



What else did you notice?

If the variables Cut and Color are independent, then the distribution of Cut will not change if a diamond's Color is known. This could be written as

$$P(\text{Cut} \mid \text{Color}) = P(\text{Cut})$$

1. Today: Relationship between two categorical variables

2. Main ideas
    1. Joint, marginal and conditional distribution
    2. Independence of two categorical variables

3. Summary

1. Joint, marginal and conditional distribution
2. Independence of two categorical variables