Checklist

1. Would you help me? If yes, unmute yourself, and let me know.
   1. Be my co-host, and admit people from the waiting room.
   2. If there is a question in the chat, please remind me.
   3. If you notice somebody raises their hand during lecture, please remind me.
2. Type your questions, if any, in the chat now.
3. I will set an alarm at 12:45pm.
4. I will press the record button now.

Unit 1: Data Summary and Visualization I

4. Categorical variables II

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

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

1. A revisit to the admission example

2. Today: Relationship between two categorical variables

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

4. Summary

Admission results

A table of counts, based on data from UC Berkely's graduate admission process in 1973.

	Men	Women
Accepted	533	113
Denied	663	336
Total	1198	449

Admission results

A table of counts, based on data from UC Berkely's graduate admission process in 1973.

	Men	Women
Accepted	533	113
Denied	663	336
Total	1198	449

Proportion of male applicants who were accepted out of the total male applicants is $533/(533 + 663) \approx .445 \approx 45\%$.

Proportion of female applicants who were accepted out of the total male applicants is $113/(113 + 336) \approx .25 \approx 25\%$.

Admission results

A table of counts, based on data from UC Berkely's graduate admission process in 1973.

	Men	Women
Accepted	533	113
Denied	663	336
Total	1198	449

Proportion of male applicants who were accepted out of the total male applicants is $533/(533+663) \approx .445 \approx 45\%$.

Proportion of female applicants who were accepted out of the total male applicants is $113/(113+336) \approx .25 \approx 25\%$.

We observed that the acceptance rate for men is almost 20 percentage points higher than the acceptance rate for women (44.5% vs. 25.2%).

Admission results

Let's proceed to dig a little deeper. The data actually came from two different programs, let's call them A and F. This table show the counts for these two programs separately.

	M(accpt)	M(denied)	W(accpt)	W(denied)
ProgA	511	314	89	19
ProgF	22	351	24	317
Total	533	665	113	336

Let's proceed to dig a little deeper. The data actually came from two different programs, let's call them A and F. This table show the counts for these two programs separately.

	M(accpt)	M(denied)	W(accpt)	W(denied)
ProgA	511	314	89	19
ProgF	22	351	24	317
Total	533	665	113	336

Within A, the proportion of male applicants who were accepted out of the total male applicants who applied for A is $511/(511+314) = 511/825 \approx .619 \approx 62\%$

Admission results

Let's proceed to dig a little deeper. The data actually came from two different programs, let's call them A and F. This table show the counts for these two programs separately.

	M(accpt)	M(denied)	W(accpt)	W(denied)
ProgA	511	314	89	19
ProgF	22	351	24	317
Total	533	665	113	336

In summary,
Program A, men: $511/(511+314) = 511/825 \approx .619 \approx 62\%$
Program A, women: $89/(89+19) = 89/108 \approx .824 \approx 82\%$
Program F, men: $22/(22+351) = 22/373 \approx .059 \approx 6\%$
Program F, women: $24/(24+317) = 24/341 \approx .070 \approx 7\%$

This is very strange!

Admission results

Edited from Biana, Tiffany and Zoe

“In Part I when we look at the macro level on the overall acceptance trend, it showed that men’s acceptances were almost twice that of women. In Part II where the data was broke down admissions between men and women into two programs, the data shows the reverse; more women were accepted into either program than men. In Part I, women seem to be at a disadvantage; but in Part II, the opposite seems true.”

Admission results

Edited from Biana, Tiffany and Zoe

“In Part I when we look at the macro level on the overall acceptance trend, it showed that men’s acceptances were almost twice that of women. In Part II where the data was broke down admissions between men and women into two programs, the data shows the reverse; more women were accepted into either program than men. In Part I, women seem to be at a disadvantage; but in Part II, the opposite seems true.”

Why so?

Why so?

Hint: think about what's the difference between program A and program F?

Why so?

Hint: think about what's the difference between program A and program F?

Let' look at the data table more carefully.

	M(accpt)	M(denied)	W(accpt)	W(denied)
ProgA	511	314	89	19
ProgF	22	351	24	317
Total	533	665	113	336

What's the acceptance rate for program A:

$(511 + 89)/(511 + 314 + 89 + 19) \approx 64\%$

What's the acceptance rate for program F:

$(22 + 24)/(22 + 351 + 24 + 317) \approx 5\%$

Why so?

Hint: think about what's the difference between program A and program F?

Let' look at the data table more carefully.

	M(accpt)	M(denied)	W(accpt)	W(denied)
ProgA	511	314	89	19
ProgF	22	351	24	317
Total	533	665	113	336

How many men applied for program A? $511 + 314 = 825$
How many women applied for program A? $89 + 19 = 108$

How many men applied for program F? $22 + 351 = 373$
How man women applied for program F? $24 + 317 = 341$

Why so?

The odds is because more women applied for the program that is difficult to get into.

This explains how it happens that women have a higher acceptance rate than men in both programs but a lower acceptance rate than men when the programs are combined.

Outline

1. A revisit to the admission example

2. Today: Relationship between two categorical variables

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

4. Summary

Outline

1. A revisit to the admission example

2. Today: Relationship between two categorical variables

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

4. Summary

Outline

1. A revisit to the admission example

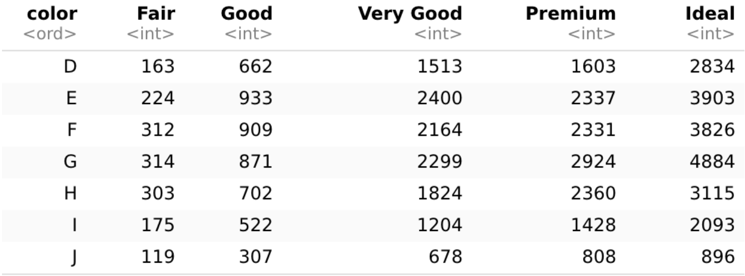
2. Today: Relationship between two categorical variables

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

4. Summary

From yesterday

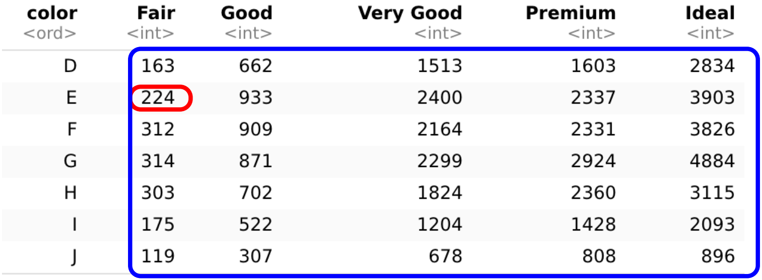
Below is the contingency table of the two variables ‘cut‘ and ‘color‘ in the diamonds dataset from yesterday.



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

From yesterday

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



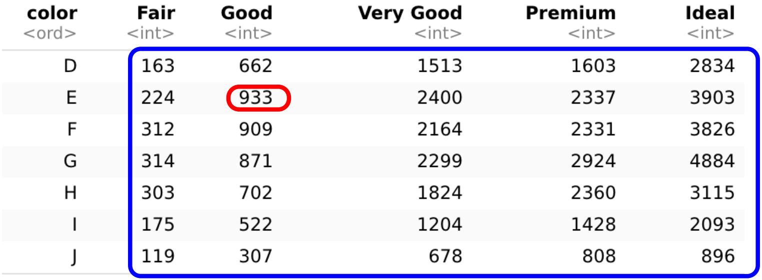
color	Fair	Good	Very Good	Premium	Ideal
<ord>	<int>	<int>	<int>	<int>	<int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

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

From yesterday

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



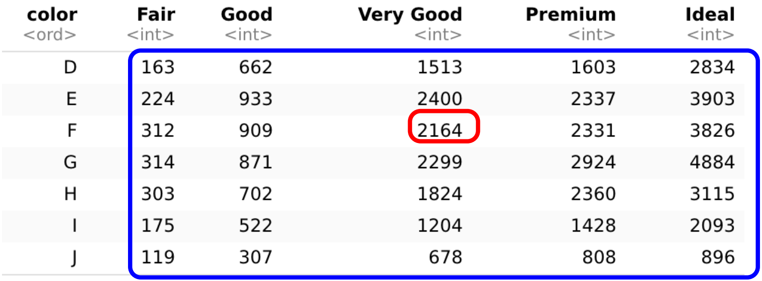
color	Fair	Good	Very Good	Premium	Ideal
<ord>	<int>	<int>	<int>	<int>	<int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

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

From yesterday

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



color	Fair	Good	Very Good	Premium	Ideal
<ord>	<int>	<int>	<int>	<int>	<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

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

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

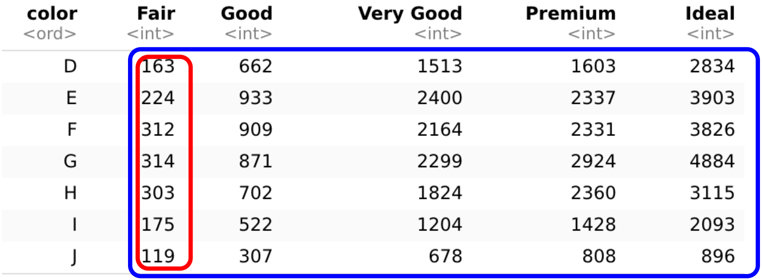
$$
\begin{gathered}
P(\text{cut} = \text{Very Good, color} = \text{F}) \\
P(\text{cut} = \text{Good, color} = \text{D}) \\
P(\text{cut} = \text{Fair, color} = \text{E}) \\
\vdots
\end{gathered}
$$

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?

From yesterday

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



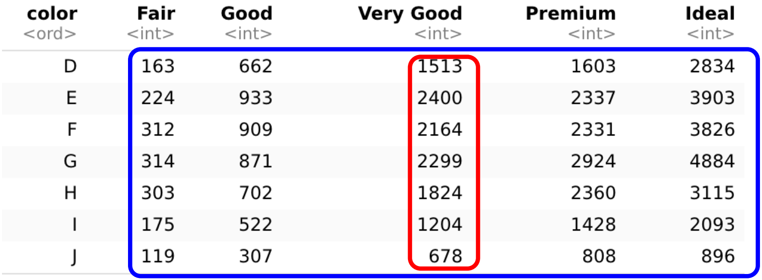
color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

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

From yesterday

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



color	Fair	Good	Very Good	Premium	Ideal
<ord>	<int>	<int>	<int>	<int>	<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

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

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

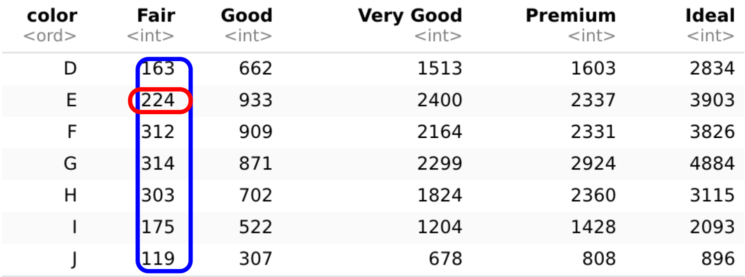
$$
\begin{gathered}
P(\text{cut} = \text{Very Good}) \\
P(\text{cut} = \text{Good}) \\
P(\text{cut} = \text{Fair}) \\
\vdots
\end{gathered}
$$

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‘)?

From yesterday

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



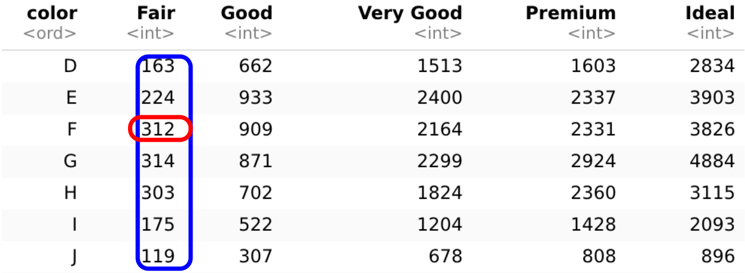
color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Mathematically, we write

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

From yesterday

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

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

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?

Outline

1. A revisit to the admission example

2. Today: Relationship between two categorical variables

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

4. Summary

(2) Independence of two categorical 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.

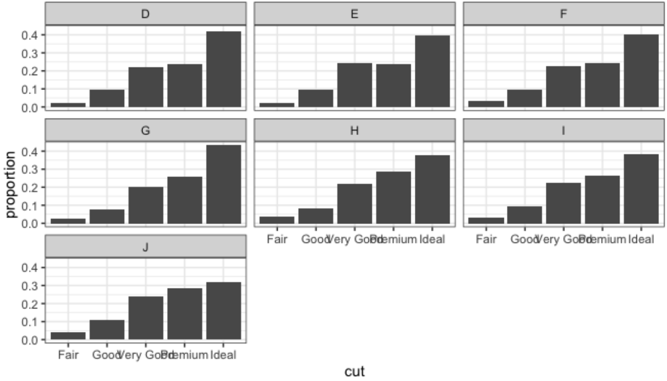
Two ways to check independence:

1. compare barplots
2. compare conditional distribution

Compare barplots

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, the two variables are independent.

Eg: Barplot of cut against color



For each level in the variable ‘Color‘, we construct a barplot that shows the distribution of ‘Cut‘.
The variables ‘Color‘ and ‘Cut‘ are independent.

Compare conditional distribution

Comparing conditional distributions of one variable across categories of another tells us about the association between variables. If the conditional distributions of one variable are (roughly) the same for every category of the other, the variables are independent.

Miki asked a question on Piazza...

" How do you do the 'Just Checking ' problem in the reading? "

Tutorial exercise: For the rest of the class

Finish 03 Exercise
See link in zoom chat or on daily schedule
Goal:
(1) practice computing joint, marginal, conditional distribution on a real example.
(2) investigate the relationship between two variables using conditional distribution.
If you could not finish this in class, try to find a time to work with your group after class.
Submit your answer on Piazza. No need to submit on Moodle.

Outline

1. A revisit to the admission example

2. Today: Relationship between two categorical variables

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

4. Summary

Summary of main ideas

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

Important message

Simpson's paradox: A phenomenon where a trend appears in several different groups of data, but disappears or reverses when they are combined.