The Relationship Between Categorical Variables

Example: Art Exhibition

Artists often submit slides of their work to be reviewed by judges who decide which artists' work will be selected for an exhibition. In the 1980 Marietta College Crafts National Exhibition, a total of 1099 artists applied to be included in a national exhibit of modern crafts. If we classify each artist according to two categorical variables:

- Was the artist selected or not?
- Where does the artist live?

then we arrive at the following two-way contingency table.

	Number of Artists		
	Selected	Rejected	TOTAL
North Central	63	299	362
Northeast	55	207	262
South	44	208	252
West	54	169	223
TOTAL	216	883	1,099

Based on Siegel and Morgan, pp. 482–494.

Column Percentages

- Divide each entry in the table of counts by the total for its column.

Top left column percent in this table is

$$\frac{63}{216} = 0.292, \text{ or } 29.2\%$$

Tells us

> *"Of the applicants selected to appear in the exhibition, 29.2% came from the North Central region."*

Answers the question

> *"If the artist was selected (column 1), how likely was he/she to come from the North Central Region (row 1)?"*

Column percents are more appropriate if we are thinking of

- The column variable as explanatory and
- the row variable as response.

The complete table of column percents

	Column percentages of artists		
	Selected	Rejected	Overall
North Central	29.2%	33.9%	32.9%
Northeast	25.5%	23.4%	23.8%
South	20.4%	23.6%	22.9%
West	25.0%	19.1%	20.3%
Overall	100.0%	100.0%	100.0%

This is basically a table of geographical distributions:

- The first column gives the *conditional distribution* of artists' geographical area, among artists selected for the exhibition: 29.2% for North Central, 25.5% for Northeast, etc.

- The second column gives the *conditional distribution* of geographical area among artists not selected.

- The third column gives the *marginal* distribution of geographcal area among all artists.

The 100%'s at the bottom of each column remind us that these are column percents and not row percents.

Row Percentages

- Divide each entry in the table of counts by the total for its column.

The top left row percent in the original table of counts would be

$$\frac{63}{362} = 0.174, \text{ or } 17.4\%$$

Answers a different question:

> *"If the artist was from the North Central region (row 1), how likely was he/she to be selected for the exhibition (column 1)?"*

Row percents are more appropriate if we are thinking of the row variable as explanatory and the column variable as response.

The complete table of row percents

	Row percentages of artists		
	Selected	Rejected	Overall
North Central	17.4%	82.6%	100.0%
Northeast	21.0%	79.0%	100.0%
South	17.5%	82.5%	100.0%
West	24.2%	75.8%	100.0%
Overall	19.7%	80.3%	100.0%

For this data, row percents answer

> *Was there discrimination in accepting artists for the competition, based on the region of the country they came from?*

- The first row gives the conditional distribution of selection for artists from the North Central region: 17.4% selected, 82.6% rejected.

- The next row gives the conditional distribution of selection for artists from the Northeast region, and so forth.

- The last row gives the *marginal* distribution of selection: among all artists, 19.7% were selected and 80.3% were rejected.

It is important to identify

- *which variable is the explanatory variable, and*
- *which is the response variable*

so that the appropriate conditional distribution is calculated:

- *Column percents if the column variable is explanatory;*
- *Row percents if the row variable is explanatory.*

Association and Independence

- Two categorical variables are *independent* if the conditional distribution of the response variable *does not change*, as we switch from one value to another of the explanatory variable.

- Two categorical variables are *associated* if the conditional distribution of the response variable *does change*, as we switch from one value to another of the explanatory variable.

When two variables are independent, knowledge of the values of one variable does not help us predict the outcome of the other variable.

- You can use column percents *<u>or</u>* row percents to check independence.

For the art exhibit data,

  - The column percents are not exactly the same, as we move from the “selected” column to the “rejected” column.

  - The row percents are also not exactly the same as we move from region to region (row to row).

Strictly speaking, it does not appear that the variables are independent; they appear to be associated.

Assessing the Strength of Association

In the art exhibit example, are the differences in percent selected...

- ...just due to minor fluctuations in the quality of art submitted to the exhibition judges?

  Then we would not want to claim regional discrimination.

  We would say the differences are *not statistically significant*.

- ...too large to explain as "chance fluctuations"?

  Then the differences are *statistically significant*. Whether they are also significant "in the real world" depends on the problem!

We can assess both

- The "significance" of deviations from independence; and
- The pattern of deviations

by comparing the table of counts *actually observed* with what would have been *expected* if independence held exactly.

The Expected Table

- The *expected table* is a hypothetical table, that shows how the counts would have been if the hypothesis of independence had been exactly true.
- Each count in the expected table is computed by multiplying the row total by the column total, and then dividing by the grand total:

$$\text{Expected} = \frac{(\text{row total}) \times (\text{column total})}{(\text{grand total})} = \frac{r \times c}{n}$$

From the art exhibit data, the expected number of artists selected from the North Central region would be

$$\begin{aligned}\frac{362 \times 216}{1099} &= 362 \times \frac{216}{1099} \\ &= 362 \times 0.196542 \\ &= 71.148\end{aligned}$$

As another sample calculation, the expected number of artists from the South who were rejected is

$$\begin{aligned}\frac{252 \times 883}{1099} &= 252 \times \frac{883}{1099} \\ &= 252 \times 0.803458 \\ &= 202.471\end{aligned}$$

This formula is applied to each and every entry in the orginal table of counts. Gathering together all of these expected counts, we find the expected table.

	Observed and Expected Counts			
	Selected		Rejected	
	Obs	Exp	Obs	Exp
N. Cent.	63	71.148	299	290.852
Northeast	55	51.494	207	210.506
South	44	49.529	208	202.471
West	54	43.829	169	179.171

How do we know if the difference between the observed count and the expected count is too big?

We use a tool called *standardized residuals*, which obey the 68-95-99.7% rule.

Computing Standardized Residuals

1. First, compute the expected table (we've done that).

2. Compute the differences "observed − expected" in each cell of the table:

	(Observed − Expected) Counts	
	Selected	Rejected
N. Cent.	-8.148	8.148
Northeast	3.506	-3.506
South	-5.494	5.529
West	10.171	-10.171

3. Finally, we divide each number in this table by the square-root of the expected count.

	$(O - E)/\sqrt{E}$	
	Selected	Rejected
North Central	-0.966	0.478
Northeast	0.489	-0.242
South	-0.781	0.389
West	1.536	-0.760

This is the table of *table of standardized residuals*.

Interpreting Standardized Residuals

$$\text{Std Res} \;=\; \frac{(O-E)}{\sqrt{E}} \quad \text{is like} \quad Z \;=\; \frac{x - \text{Mean}}{SD}$$

so that standardized residuals roughly obey the 68-95-99.7% rule. They tell us how "far from independence" the corresponding observed count is:

- If the standardized residual is positive, the observed count was bigger than the expected count;
- If the standardized residual is negative, the observed count was smaller than the expected count;
- Standardized residuals between about -1.5 and 1.5 indicate cells that agree with independence;
- Standardized residuals between about -2.0 and -1.5 or between 1.5 and 2.0 indicate cells that give *mild* evidence against independence;
- Standardized residuals less than -2.0 or greater than 2.0 indicate cells that give *strong* evidence against independence.

In this example, none of the cells are contributing much evidence against independence.