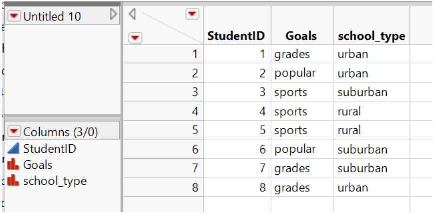
DSA 8010 - Categorical data analysis

chi-square test for association

Method	types of variables
inference on one proportion	binary
inference on two proportions	binary & binary
one sample t test	numeric (one variable)
two-sample t test	numeric & binary
simple linear regression	numeric & numeric
one-way ANOVA	numeric & categorical
chi-square test for homogeneity (not covered)	categorical (one variable)
chi-square test for association	categorical & categorical

Categorical data

Spreadsheet format.



	StudentID	Goals	school_type
1	1	grades	urban
2	2	popular	urban
3	3	sports	suburban
4	4	sports	rural
5	5	sports	rural
6	6	popular	suburban
7	7	grades	suburban
8	8	grades	urban

Table format.

Goals	School Area Rural	Suburban	Urban	Total
Grades	57	87	24	168
Popular	50	42	6	98
Sports	42	22	5	69
Total	149	151	35	335

Source: `http://www.stat.yale.edu/Courses/1997-98/101/chisq.htm`

Multinomial experiments

The multinomial experiment satisfies the following conditions:

1. A fixed number (n) of independent trials.
2. Each trial results in one of k outcomes.
3. There is a fixed probability π_i of a single trial resulting in outcome i.
4. The expected count for outcome i is $n\pi_i$.

Multinomial experiments

Examples:

- Randomly select 100 graduating seniors for an exit survey. Record if they have a job in their field, a job outside of their field, plans to attend graduate school, or none of the above.
- Randomly sample 500 of your customers and record their satisfaction with their most recent transaction as "Highly dissatsified," "somewhat dissatisfied," "somewhat satisfied," "highly satisfied." Also record whether their interaction with the company representatives was primarily over phone or email.

Example (skin disease)

The CDC wants to know if there is an association between severity of a skin disease and patient's age group. The severity is classified into three categories and there are four age groups. The following table summarizes age group and severity level for a sample of patients:

	I	II	III	IV	All ages
Moderate	15	32	18	5	70
Mildly severe	8	29	23	18	78
Severe	1	20	25	22	68
All severities	24	81	66	45	216

We can view these data as arising from a multinomial experiment with 216 trials and 12 possible outcomes.

Notation in two way tables

Designate one variable as the "row variable" and another as the "column variable." Let r denote the number of possible outcomes from the "row" variable and let c denote the number of possible outcomes from the "column" variable.

The table has $r \times c$ cells in total, with each cell representing one possible outcome.

- n_{ij} denotes the observed count in row i and column j for $i = 1, \ldots, r; 1, \ldots, c$.
- $n_{i.}$ denotes the observed total in row i, added across columns, for $i = 1, \ldots, r$.
- $n_{.j}$ denotes the observed total in column j, added across rows, for $j = 1, \ldots, c$.
- $n_{..}$ (or sometimes just n) denotes the total sample size.

Notation: two way tables

Example: Variable A has three possible outcomes ($r = 3$). Variable B has four possible outcomes ($c = 4$). There are a total of 12 possible outcomes from the multinomial experiment and the results can be summarized in a 3×4 table, the observed two-way table is

	B1	B2	B3	B4	Total
A1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
A2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
A3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$

Notation: two way tables

Let π_{ij} denote the probability of observing the outcome in the i, jth cell of the table.

	B1	B2	B3	B4	Total
A1	π_{11}	π_{12}	π_{13}	π_{14}	$\pi_{1.}$
A2	π_{21}	π_{22}	π_{23}	π_{24}	$\pi_{2.}$
A3	π_{31}	π_{32}	π_{33}	π_{34}	$\pi_{3.}$
Total	$\pi_{.1}$	$\pi_{.2}$	$\pi_{.3}$	$\pi_{.4}$	1

Inference on two categorical variables: data and notation

Data. Two categorical variables, with frequencies organized in a two-way table. One is the row variable and one is the column variable.

Notation. n_{ij} is the count in the i, jth cell of the two-way table.

Inference on two categorical variables: data and notation

Data. Two categorical variables, with frequencies organized in a two-way table. One is the row variable and one is the column variable.

Notation. n_{ij} is the count in the i, jth cell of the two-way table.

Statistical model. The set of n_{ij} have a multinomial distribution, with π_{ij} as the probability of the i, jth outcome.

Inferential question. Is there a significant association between the two variables?

Equivalently, is there evidence that the variables are not independent?

Example (iPod defects)

Open Intro Statistics, 4th edition, Diez et al Section 6.4.

We all buy used products – cars, computers, textbooks, and so on – and we sometimes assume the sellers of those products will be forthright about any underlying problems with what they're selling. This is not something we should take for granted. Researchers recruited 219 participants in a study where they would sell a used iPod[40] that was known to have frozen twice in the past. The participants were incentivized to get as much money as they could for the iPod since they would receive a 5% cut of the sale on top of $10 for participating. The researchers wanted to understand what types of questions would elicit the seller to disclose the freezing issue.

Unbeknownst to the participants who were the sellers in the study, the buyers were collaborating with the researchers to evaluate the influence of different questions on the likelihood of getting the sellers to disclose the past issues with the iPod. The scripted buyers started with "Okay, I guess I'm supposed to go first. So you've had the iPod for 2 years ..." and ended with one of three questions:

- General: What can you tell me about it?
- Positive Assumption: It doesn't have any problems, does it?
- Negative Assumption: What problems does it have?

Example (iPod defects)

Open Intro Statistics, 4th edition, Diez et al Section 6.4.

The question is the treatment given to the sellers, and the response is whether the question prompted them to disclose the freezing issue with the iPod. The results are shown in Figure 6.14, and the data suggest that asking the, *What problems does it have?*, was the most effective at getting the seller to disclose the past freezing issues. However, you should also be asking yourself: could we see these results due to chance alone, or is this in fact evidence that some questions are more effective for getting at the truth?

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2	23	36	61
Hide Problem	71	50	37	158
Total	73	73	73	219

Figure 6.14: Summary of the iPod study, where a question was posed to the study participant who acted

Example (iPod defects)

Is the decision to disclose the problem associated with initial assumptions?

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2	23	36	61
Hide Problem	71	50	37	158
Total	73	73	73	219

Example (iPod defects)

Is the decision to disclose the problem associated with initial assumptions?

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2	23	36	61
Hide Problem	71	50	37	158
Total	73	73	73	219

A descriptive analysis of a contingency table might include calculating row proportions and comparing across the rows. If the row proportions are very different, the variables might be associated.

Example (iPod defects)

What would the table look like if there were no association?

- Those that disclose the problem would be evenly distributed across the three initial assumptions.
- Those that did not disclose would also be evenly distributed across the three initial assumptions.

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2	23	36	61
Hide Problem	71	50	37	158
Total	73	73	73	219

Chi-square test for association

Hypotheses. H_0 : variables are independent (no association).
H_A : variable are not independent (they are associated).

Test statistic.

$$X_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\widehat{E}_{ij})^2}{\widehat{E}_{ij}}, \text{ where } \widehat{E}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

.

Chi-square test for association

Hypotheses. H_0 : variables are independent (no association).
H_A : variable are not independent (they are associated).

Test statistic.

$$X_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - \widehat{E}_{ij})^2}{\widehat{E}_{ij}}, \text{ where } \widehat{E}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

$$X_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(observed - expected)^2}{expected}$$

The expected counts come from the contingency table that would be observed if there were independence (no association) between the variables.

Observed and expected tables

Main idea: create an "ideal" table that we would observe under perfect independence. Then measure how far the observed table is from the perfectly independent table.

Observed:

	B 1	B 2	B 3	B 4	Total
A 1	16	26	34	45	121
A 2	23	52	19	13	107
A 3	11	31	7	9	58
Total	50	109	60	67	286

Expected:

	B 1	B 2	B 3	B 4	Total
A 1					121
A 2					107
A 3					58
Total	50	109	60	67	286

Expected cell counts

The estimated cell count for outcome ij when the two variables are independent, denoted by $\widehat{E}_{ij}$, is calculated as

$$\widehat{E}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}.$$

$$\frac{\text{total for row i} * \text{total for column j}}{\text{total sample size}}$$

Expected cell counts

The estimated cell count for outcome ij when the two variables are independent, denoted by $\widehat{E}_{ij}$, is calculated as

$$\widehat{E}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}.$$

Expected

	B 1	B 2	B 3	B 4	Total
A 1					121
A 2					107
A 3					58
Total	50	109	60	67	286

Find the expected counts for (A1, B1) and (A3, B2).

Observed and expected tables

Main idea: create an "ideal" table that we would observe under perfect independence. Then measure how far the observed table is from the perfectly independent table.

Observed:

	B 1	B 2	B 3	B 4	Total
A 1	16	26	34	45	121
A 2	23	52	19	13	107
A 3	11	31	7	9	58
Total	50	109	60	67	286

Expected:

	B1	B2	B3	B4	Total
A1	21.15	46.12	25.38	28.35	121
A2	18.71	40.78	22.45	25.07	107
A3	10.14	22.10	12.17	13.59	58
Total	50	109	60	67	286

Chi-square test for association

Hypotheses. H_0 : variables are independent (no association).
H_A : variable are not independent (they are associated.)

Test statistic.

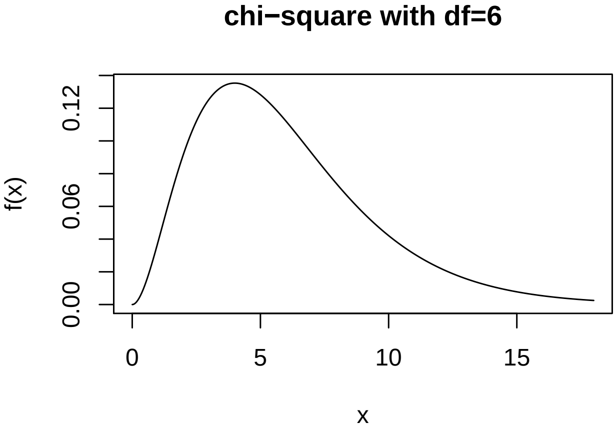
$$X_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\widehat{E}_{ij})^2}{\widehat{E}_{ij}}$$

P-value and decision. Use the χ^2 (chi-square) distribution with $df = (r-1)\times(c-1)$ to find the p-value. Reject H_0 if the p-value is $< \alpha$.

```
pchisq(x0.squared,df=(r-1)*(c-1),lower.tail=FALSE).
```

The chi-square distribution

Chi-square distribution with 6 degrees of freedom:



Chi-square test for association

- The term

$$\frac{(n_{ij} - \widehat{E}_{ij})^2}{\widehat{E}_{ij}}$$

  is called the *ij*th *contribution* to the χ^2 statistic.

  The contribution is large for cells where the observed is very different from the expected. The contribution is small in cells where the observed and expected values are similar.

- If H_0 is rejected, we conclude that there is some association between the two variables. We do not necessarily know whether the association is strong or weak.
- The chi-square test is most accurate when the expected count in each cell is at least 5.

Recap: independent events

If events A and B are independent, then

$$P(A \cap B) = P(A)P(B).$$

Recap: independent events

If events A and B are independent, then

$$P(A \cap B) = P(A)P(B).$$

	B	B'	Total
A			$P(A)$
A'			$P(A')$
Total	$P(B)$	$P(B')$	1

In other words, if the A and B are independent, then we can obtain probabilities of intersections using only the row and column totals.

Recap: independent events

If events A and B are independent, then

$$P(A \cap B) = P(A)P(B).$$

	B	B'	Total
A			$P(A)$
A'			$P(A')$
Total	$P(B)$	$P(B')$	1

In other words, if the A and B are independent, then we can obtain probabilities of intersections using only the row and column totals.

Independent events and two way tables

If the row variable and column variable are independent of each other, then

$$\pi_{ij} = \pi_{i.}\pi_{.j},$$

	B 1	B 2	B 3	B 4	Total
A 1	$\pi_{1.}\pi_{.1}$	$\pi_{1.}\pi_{.2}$	$\pi_{1.}\pi_{.3}$	$\pi_{1.}\pi_{.4}$	$\pi_{1.}$
A 2	$\pi_{2.}\pi_{.1}$	$\pi_{2.}\pi_{.2}$	$\pi_{2.}\pi_{.3}$	$\pi_{2.}\pi_{.4}$	$\pi_{2.}$
A 3	$\pi_{3.}\pi_{.1}$	$\pi_{3.}\pi_{.2}$	$\pi_{3.}\pi_{.3}$	$\pi_{3.}\pi_{.4}$	$\pi_{3.}$
Total	$\pi_{.1}$	$\pi_{.2}$	$\pi_{.3}$	$\pi_{.4}$	1

i.e., the multinomial probabilities can be calculated using only the row and column probabilities.

Expected cell counts

The fourth characteristic of the multinomial experiment (from the slide at the beginning of lecture) says that the expected count in the *ij*th cell of the table is

$$E_{ij} = n\pi_{ij}.$$

The formula for the expected count corresponds to our assumption of independence:

$$\widehat{E}_{ij} = n_{..}\widehat{\pi}_{i.}\,\widehat{\pi}_{.j} = n_{..}\frac{n_{i.}}{n_{..}}\,\frac{n_{.j}}{n_{..}} = \frac{n_{i.}n_{.j}}{n_{..}}$$

Example (skin disease)

The CDC wants to know if there is an association between severity of a skin disease and patient's age group. The severity is classified into three categories and there are four age groups. The following table summarizes age group and severity level for a sample of patients:

Observed.

	I	II	III	IV	All ages
Moderate	15	32	18	5	70
Mildly severe	8	29	23	18	78
Severe	1	20	25	22	68
All severities	24	81	66	45	216

Find the expected cell counts under independence.

Example (skin disease)

Expected.

	I	II	III	IV	All ages
Moderate					70
Mildly severe					78
Severe					68
All severities	24	81	66	45	216

Example (skin disease)

Is there evidence of a significant association between age group and disease severity? (use $\alpha = 0.05$.)

Observed.

	I	II	III	IV
Moderate	15	32	18	5
Mildly severe	8	29	23	18
Severe	1	20	25	22

Expected.

	I	II	III	IV
Moderate	7.778	26.250	21.389	14.583
Mildly severe	8.667	29.250	23.833	16.250
Severe	7.556	25.500	20.778	14.167

Example (skin disease)

Observed.

	I	II	III	IV
Moderate	15	32	18	5
Mildly severe	8	29	23	18
Severe	1	20	25	22

Expected.

	I	II	III	IV
Moderate	7.778	26.250	21.389	14.583
Mildly severe	8.667	29.250	23.833	16.250
Severe	7.556	25.500	20.778	14.167

Test statistic:

$$X_0^2 = \frac{(15-7.778)^2}{7.778} + \frac{(32-26.250)^2}{26.250} + \frac{(18-21.389)^2}{21.389} + \frac{(5-14.583)^2}{14.583} + \ldots + \frac{(1-7.556)^2}{7.556} + \frac{(20-25.500)^2}{25.500} + \frac{(25-20.778)^2}{20.778} + \frac{(22-14.167)^2}{14.167} = 27.135$$

Example (skin disease)

Is there evidence of a significant association between age group and disease severity? (use $\alpha = 0.05$.)

	I	II	III	IV	All ages
Moderate	15	32	18	5	70
Mildly severe	8	29	23	18	78
Severe	1	20	25	22	68
All severities	24	81	66	45	216

Example (iPod defects)

Is the decision to disclose the problem associated with initial assumptions? Perform a test using $\alpha = 0.01$.

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2	23	36	61
Hide Problem	71	50	37	158
Total	73	73	73	219