Ch. 28: The Chi-Square Test

The χ^2 – test (chi-square, pronounced "ki-square") helps us answer questions such as:

- Are the data consistent with a given chance model, or are they far off?

- Has someone manipulated the data to make them fit the chance model?

- Are two things independent in the population from which the sample was drawn?

Ex: A gambler is accused of using a "loaded" die. A record has been kept of the last 60 draws:

4	3	3	1	2	3	4	6	5	6
2	4	1	3	3	5	3	4	3	4
3	3	4	5	4	5	6	4	5	1
5	4	6	3	3	3	5	3	1	4

Summarized, the results are:

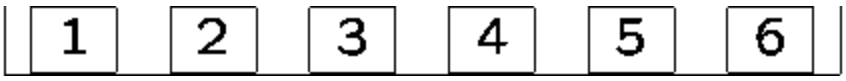
Value	Observed frequency	Expected frequency
1	4	
2	6	
3	17	
4	16	
5	8	
6	9	

Are these results consistent with what we would expect from a fair die? Or are there too many 3's and 4's?

The χ^2–test

Step 1: State hypotheses:

Null hypothesis:
The die is fair. Rolling this die is like drawing at random with replacement from the box:



Alternative hypothesis:
The die is loaded. Rolling this die is not like drawing at random from the above box.

Step 2: Calculate test statistic:

We combine all our frequency data into one value that measures how well the model is doing.

This value is called the _____.

Calculating χ^2–statistic

$$\chi^2 = \text{sum of } \frac{(\text{observed freq. - expected freq.})^2}{\text{expected frequency}}$$

In our die example:

Value	Obs.	Exp.	(Obs.– Exp.)^2	$\frac{(Obs.-Exp.)^2}{Exp.}$

$\chi^2 =$ _____

Step 3: Obtain P-value:

The P-value is obtained from the _____ with

$$\#\text{ degrees of freedom} = \#\text{ of categories} - 1$$

The P-value is the area to the right of the calculated χ^2.

To find the approximate P-value, we use the _____ which is similar to the t–table.

In our die example:

d.f. = _____

P-value: _____

Step 4: State conclusions:

Note: The χ^2–test is valid only if all expected frequencies are 5 or more!

Applications in Genetics

Ex: According to a certain genetic theory, inheritance of flower color in *Mirabilis jalapa* (four o'clock) plants shows the following pattern: when a plant with red flowers is crossed with a plant with white flowers, all of the offspring have pink flowers. However, when the plants of this second generation are crossed with each other, 1/4 of the plants in the resulting third generation have red flowers, 1/2 of the plants have pink flowers, and 1/4 of the plants have white flowers.

You bought some third generation "four o'clock" seeds and got 88 plants with red flowers, 201 with pink flowers, and 111 with white flowers. Is there evidence that your seeds were "manipulated"?

χ^2–test for "Four o'clock" plants

Testing Independence

Ex: Independent random samples of workers in three parts of the country were asked whether they considered unemployment or inflation the more serious problem. We can sum up their responses in a *2×3 table* (or more generally, an *m×n table*).

	Northeast	Midwest	Southwest	Total
Unemp.	87	73	66	226
Infl.	113	77	84	274
Total	200	150	150	500

Do we have reason to believe that workers in different parts of the country feel differently about these issues?

χ^2–test for independence

Step 1: State hypotheses:

Our model looks like:

NE	??×[U] ??×[I]	200 draws
MW	??×[U] ??×[I]	150 draws
SW	??×[U] ??×[I]	150 draws.

Null hypothesis:

i.e., the percentage of [U] tickets is the same in each box.

Alternative hypothesis:

i.e., at least one box is different.

Step 2: Calculate test statistic:

$$\chi^2 = \text{sum of } \frac{(\text{observed freq. - expected freq.})^2}{\text{expected frequency}}$$

The observed frequencies are the entries in our 2×3 table.

The expected frequencies are obtained assuming independence.

45% of the workers surveyed (_____) consider unimployment to be more important, whereas 55% of the workers (_____) consider inflation to be more important.

If the null hypothesis is true, we expect that the same percentage of workers in all three regions feel the same way.

Region	Total	U (45%)	I (55 %)
NE	200	...	...
MW	150	...	...
SW	150	...	...

We usually use the short–cut formula:

Expected frequencies for an $m \times n$ table can be found for each cell by multiplying the row total by the column total and dividing by the total for the entire table.

When we finish, we get a table with both observed and expected frequencies.

Obs.	Exp.	Northeast		Midwest		Southwest	
Unemp.		87	90	73	68	66	68
Infl.		113	110	77	82	84	82

Now we can calculate the χ^2-statistic as we did before:

$\chi^2 =$ _____

d. f. $= (m-1)\times(n-1) =$ _____

Step 3: Obtain P-value:

Step 4: State conclusions:

Ex: A simple random sample of 200 Utah schoolchildren were asked whether or not they like math. 102 kids were boys, 41 of whom said they liked math. The other 98 were girls, 29 of whom said they like math. Is liking math independent of gender for Utah schoolchildren?