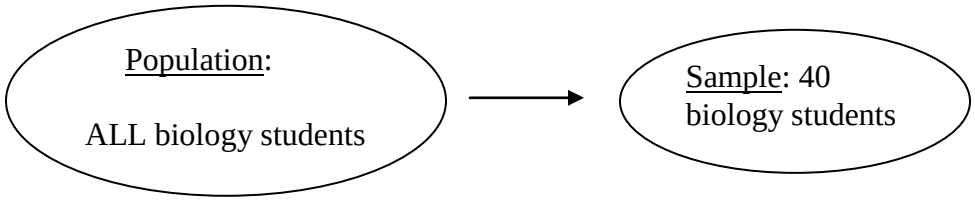
Categorical Data Analysis

A 2x2 contingency table is a common way to summarize data on two categorical variables, each with two possible outcomes. This kind of data can arise from different sampling frameworks, and the type of hypotheses that we can pose depends on the framework that is used in the study. These sampling frameworks and tests are as follows:

(1) Test for Association

A simple random sample, drawn from one population of interest, in which we obtain two binary responses from each sampling unit.

Example: We take a simple random sample of 40 students from the department of Biology. We record two responses for each student: Gender, and whether or not they have a pet.



The dataset will look something like this:

Student	Gender	Pet Owner?
1	M	Yes
2	M	No
3	F	No
4	F	No
etc.		

The hypothesis of interest here is association:
Is there an association between gender and whether or not a student has a pet? In other words, are gender and having a pet independent?

H_0: Gender and pet ownership are independent (there is no association).
H_A: Gender and pet ownership are not independent (there is an association).

For our sample of 40 students, the 2x2 contingency table of outcomes to the two questions of interest will look like the following:

	Have a pet?		
Gender	Yes	No	Total
Male			$r1$
Female			$r2$
Total	$c1$	$c2$	40

Note that the only fixed number at the start of the study in this contingency table is the overall total (= 40 students in the sample). The rest of the marginal totals ($c1$, $c2$, $r1$, $r2$) will vary depending on the responses in the sample, and we will know what they are only after all the sample responses have been recorded.

Suppose:

$c1$ = the total # students out of 40 who have pets
$c2$ = the total # students out of 40 who don't have pets
$r1$ = the total # students out of 40 who are male
$r2$ = the total # students out of 40 who are female

Then $c1+c2 = 40$ and $r1+r2 = 40$

Once we have all the data, we would have the above table filled in, and this would be the "Observed table".

Assuming the null hypothesis were true, we can calculate the expected cell count for each cell in this table. Let's see how this works for the first cell (Male and Has a pet).

If H0 were true, then Gender and Pet Ownership would be independent.

So Pr(Male AND Has a pet) = Pr(Male) * Pr(Has a pet)

$$= \frac{r1}{40} \times \frac{c1}{40}$$

The *Expected* number of people who are Male AND Have a pet =

(Total # of people in the sample) * Pr(Male AND Has a pet)

$$= 40 \times \left[\frac{r1}{40} \times \frac{c1}{40} \right] \qquad = \frac{r1 \times c1}{40}$$

= (Row Total * Column Total)/Overall Total.

We can similarly obtain the expected cell count for each cell using the same formula:

Expected cell count = (Row Total * Column Total)/Overall Total

The test statistic for this test would then be, summing over all the cells:

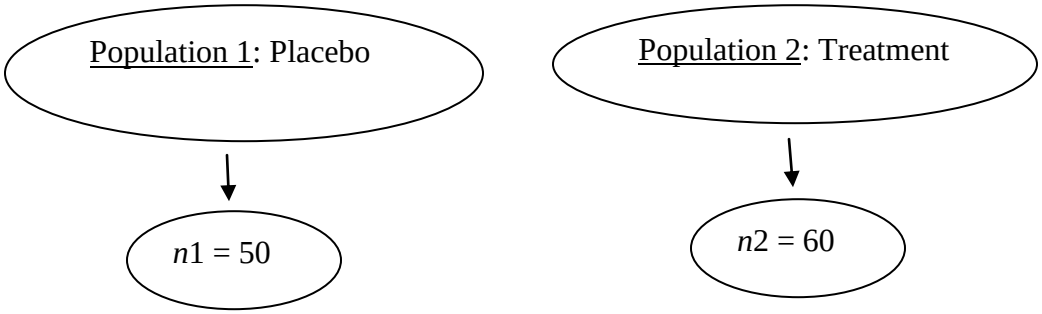
$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

where Obs = observed cell count, and
Exp = expected cell count

(2) Test for Homogeneity

Two independent samples from two populations, where we obtain a binary response from each individual in each group.

Example: Patients randomized to Placebo ($n1 = 50$) or Treatment group ($n2 = 60$) and the outcome measured on each patient is whether or not they were cured.



The hypothesis of interest here is homogeneity, and the number of successes in each group (number of people cured) follows an independent binomial distribution.

If $p1$ = the true proportion of people on placebo who are cured
$p2$ = the true proportion of people on treatment who are cured

H0: $p1 = p2$
HA: $p1 \neq p2$

	Cured?		
Group	Yes	No	Total
Placebo			$n1$=50
Treatment			$n2$=60
Total	$c1$	$c2$	110

Note that in this design, we know the marginal totals (at least for the rows) as well as the overall total before the data are collected for each individual. The row marginal totals are the sample sizes in the two groups and these are fixed at the start of the study (recall that for the binomial distribution to hold, we need fixed sample size and a constant probability of success).

Here $c1+c2 = 110$ and $n1+n2 = 110$

Once we have the data, we would have the above table filled in, and this would be the "Observed table".

Assuming the null hypothesis were true, we can calculate the expected cell count for each cell in this table. Let's see how this works for the first cell (Placebo group and Cured).

If H0 were true, the proportion of people cured would be the same in both groups. So we can estimate this common proportion using information pooled across both groups.

If p = the common proportion of people cured,

p = [(# cured in placebo group) + (# cured in treatment group)] divided by overall total.

Here, $p = \frac{c1}{110}$

The expected number of people cured in the placebo group would then be = (total # people in placebo group) * (proportion of people cured) = $n1 \times \left(\frac{c1}{110} \right)$

= (Row Total * Column Total)/Overall Total.

We can similarly obtain the expected cell count for each cell using the same formula:
Expected cell count = (Row Total * Column Total)/Overall Total

The test statistic for this test would then be, summing over all the cells:

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

where Obs = observed cell count, and
Exp = expected cell count

Summary

As we can see from the above, both the Test for Association and the Test for Homogeneity have the same test statistic. So *functionally,* the procedure for both the tests is the same. The difference is in the sampling framework, and the way the data are collected. Each sampling framework allows us to pose a different type of question in the null and the alternate hypotheses. Therefore, at the end of the test, the results are interpreted according to the kind of hypothesis that was originally posed.