# Chi-Square ~~Goodness of Fit~~ Tests

Suppose I walk around asking people to tell me a random (integer) number from 1 to 10.

I am interested in whether people are more likely to tell me one number than others. I can phase this as a hypothesis testing problem:

$$H_0 : p_1 = p_2 = \cdots = p_{10} = \frac{1}{10}$$

$$H_1 : \text{otherwise}$$

where $p_i = \text{Prob}(i)$.

More generally, suppose an experiment has $k$ outcomes. Let $X$ denote the outcome. $(1, 2, \ldots, 10)$. Set

$$p_i = P(X = i).$$

We are interested in testing

$$H_0: \quad p_i = p_{i0} \quad \text{for } i = 1, \ldots, k$$

vs.

$$H_1: \quad \text{otherwise}$$

where $p_{10}, p_{20}, \ldots, p_{k0}$ are fixed, known, and satisfy

① $0 < p_{i0} < 1$ for all $i$

② $\sum_{i=1}^{k} p_{i0} = 1$.

In the example, $k = 10$ and $p_{i0} = \frac{1}{10}$ for each $i$.

What can we do? Well, if $k = 2$ we know how to proceed:

Let $X_1, \ldots, X_n$ be a random sample, and let $Y_1$ denote the number of times $X_j = 1$ occurs in the sample.

Then $Y_1 \sim \text{binom}(n, p_1)$.

By the CLT, $$\frac{Y_1 - np_{10}}{\sqrt{np_{10}(1-p_{10})}} \approx N(0,1).$$

if $H_0$ is true.

Therefore, if we accept $H_0$ iff

$$\frac{|Y_1 - np_{10}|}{\sqrt{np_{10}(1-p_{10})}} \leq z_{\alpha/2},$$

we have a test w/ level of significance approximately $\alpha$.

To generalize this to $k > 2$, it is helpful to consider a different test, also based on the CLT:

Let $$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

and $$Q_1 = Z^2 .$$

Then $$Q_1 \approx \chi^2(1).$$

Thus another test of $H_0$ vs. $H_1$ is to accept $H_0$ iff

$$Q_1 \le \chi^2_\alpha(1).$$

We may rewrite $Q_1$ as follows:

$$Q_1 = Z^2 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)}$$

$$= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_2)}.$$

It is true that

- $p_2 = 1 - p_1$
- $(Y_2 - np_2)^2 = (Y_1 - np_1)^2$ $\boxed{\text{simple algebra}}$

$\boxed{Y_2 = n - Y_1}$

Therefore

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}$$

Return to the general case:

- $X$ has $k$ outcomes
- $X_1, \dots, X_n$ is a random sample
- $Y_k = \#\{X_j = k\}$
- $p_i = P(X = i)$

<u>FACT</u> If $Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}$

then $Q_{k-1} \approx \chi^2(k-1)$.

(we will not prove this).

Therefore, to test $H_0$ vs. $H_1$ at a level of significance $\alpha$, we accept $H_0$ when

$$q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}} \leq \chi^2_\alpha(k-1).$$

Example : Primary colors are red, blue, yellow.

Every pick which is your favorite.

Identify
$$\text{Red} \Longleftrightarrow 1$$
$$\text{Blue} \Longleftrightarrow 2$$
$$\text{Yellow} \Longleftrightarrow 3.$$

Random sample from class:

$Y_1 =$ _____

$Y_2 =$ _____

$Y_3 =$ _____

Let's test

$$H_0 : p_1 = p_2 = p_3 = \frac{1}{3}.$$

at $\alpha = 0.05$.

$n =$ ________ (# of people in class)

$$q_2 = \frac{(Y_1 - np_{10})^2}{np_{10}} + \frac{(Y_2 - np_{20})^2}{np_{20}} + \frac{(Y_3 - np_{30})^2}{np_{30}}$$

$$= \qquad\qquad + \qquad\qquad +$$

$$= \qquad\qquad + \qquad\qquad +$$

$$=$$

From the table,

$$\chi^2_{0.05}(2) = 5.991.$$

Therefore, we ________ the null hypothesis.

The kind of $\chi^2$ test we studied is sometimes called a "$\chi^2$ goodness of fit" test.

There are other useful tests based on a $\chi^2$ statistic.

## Tests for Homogeneity

Let $X^1$ = primary color chosen by a female
$X^2$ = " " " " " male.

Denote $p_{i1} = P(X^1 = i)$, $i = 1, 2, 3$
$p_{i2} = P(X^2 = i)$, $i = 1, 2, 3$.

Lets consider

$$H_0 : p_{11} = p_{12} \text{ and } p_{21} = p_{22} \text{ and } p_{31} = p_{32}$$

"Do males' preferences tend to differ from those of females?"

More generally, consider two experiments E1 and E2, each with the same set of possible outcomes, $1, \ldots, k$.

Define $p_{ij} = P(\text{experiment } j \text{ results in outcome } i)$

We want to test for

$$H_0 : p_{i1} = p_{i2} \quad \text{for all } i = 1, \ldots, k$$

Assume we have 2 random samples

$$X_1^1, X_2^1, \ldots, X_{n_1}^1 \quad \text{from E1}$$

$$X_1^2, X_2^2, \ldots, X_{n_2}^2 \quad \text{from E2.}$$

Define $\quad Y_{i1} = \#$ of times $i$ occurs in sample 1

$\quad\quad\quad\; Y_{i2} = \#$ " " " " " " 2.

Then
$$\sum_{i=1}^{k} \frac{(Y_{i1} - n_1 p_{i1})^2}{n_1 p_{i1}} \approx \chi^2(k-1)$$

$$\sum_{i=1}^{k} \frac{(Y_{i2} - n_2 p_{i2})^2}{n_2 p_{i2}} \approx \chi^2(k-1)$$

and so
$$\sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}} \approx \chi^2(2k-2)$$

Under $H_0$, $p_{i1} = p_{i2}$ for each $i$.

In applications, however, these frequencies are unknown.

Thus it is customary to estimate $p_{i1} = p_{i2}$

with $\dfrac{Y_{i1} + Y_{i2}}{n_1 + n_2}$.

It can be shown that under $H_0$,

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{\left( Y_{ij} - n_j \cdot \frac{(Y_{i1} + Y_{i2})}{n_1+n_2} \right)^2}{n_j \left( \frac{Y_{i1} + Y_{i2}}{n_1+n_2} \right)}$$

$$\approx \chi^2(k-1)$$

( k-1 degrees of freedom are lost in estimating the outcome frequencies ).

Thus $Q \geq \chi^2_\alpha(k-1)$ $\boxed{\text{Reject } H_0}$

has level of significance approximately $\alpha$.

Back to our example (preferrences for primary colors)

E1 : ask a female her favorite primary color

E2 : " " male his " " "

We observed

$n_1 =$ ______ $n_2 =$ ______

$Y_{11} =$ ______ $Y_{21} =$ ______

$Y_{12} =$ ______ $Y_{22} =$ ______

$Y_{13} =$ ______ $Y_{23} =$ ______

$$\hat{p}_1 = \frac{Y_{11} + Y_{12}}{n_1 + n_2} =$$

$$\hat{p}_2 = \frac{Y_{21} + Y_{22}}{n_1 + n_2} =$$

$$\hat{p}_3 = \frac{Y_{31} + Y_{32}}{n_1 + n_2} =$$

Then $$q = \sum_{j=1}^{2} \sum_{i=1}^{3} \frac{(Y_{ij} - n_j \hat{p}_i)^2}{n_j \hat{p}_i}$$

$$= \frac{(Y_{11} - n_1 \hat{p}_1)^2}{n_1 \hat{p}_1} + \frac{(Y_{21} - n_1 \hat{p}_2)^2}{n_1 \hat{p}_2} + \frac{(Y_{31} - n_1 \hat{p}_3)^2}{n_1 \hat{p}_3}$$

$$+ \frac{(Y_{12} - n_2 \hat{p}_1)^2}{n_2 \hat{p}_1} + \frac{(Y_{22} - n_2 \hat{p}_2)^2}{n_2 \hat{p}_2} + \frac{(Y_{32} - n_2 \hat{p}_3)^2}{n_2 \hat{p}_3}$$

$$= \underline{\qquad\qquad} + \underline{\qquad\qquad} + \underline{\qquad\qquad}$$

$$+ \underline{\qquad\qquad} + \underline{\qquad\qquad} + \underline{\qquad\qquad}$$

$$= \underline{\qquad\qquad} \qquad \chi^2_{0.05}(2) = 5.991$$

Thus, we ______ $H_0$.

## Contingency Tables

A <u>contingency table</u> is a concise way of summarizing data for testing homogeneity:

Suppose 500 people were polled on their opinion of smoking in public places:

| Sex | Favor | Oppose | Indifferent | Totals |
|---|---|---|---|---|
| Male | 151 | 73 | 12 | 236 |
| Female | 101 | 136 | 27 | 264 |
| Totals | 252 | 209 | 39 | 500 |

We can use this table to apply a $\chi^2$ test for homogeneity.

Sometimes an investigator will ask

"are the responses/outcomes dependent on the group?"

For example, "is preference for smoking in public places dependent on sex?"

It turns out that testing

$$H_0 : p_i \cdot p_j = p_{ij} \qquad \text{(independence)}$$

results in the same test as testing for homogeneity.