Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from some distribution.

Consider testing

H_0 : The sample comes from a particular, specified distribution

versus

H_1 : "Not H_0"

Examples of H_0:

- The sample comes from a binomial distribution with parameters 8 and 0.2.

- The sample comes from a $N(0,1)$ distribution.

- The sample comes from this distribution:

x	0	1	2	3
$P(X = x)$	0.2	0.4	0.1	0.3

Here are some sampled values:

1, 1, 2, 0, 1, 1, 1, 1, 3, 3,

1, 2, 0, 3, 1, 1, 3, 3, 1, 2

Here is a distribution:

x	0	1	2	3
$P(X = x)$	0.2	0.4	0.1	0.3

Test H_0 : The sample comes from this distribution.

vs H_1 : The sample does not come from this distribution.

Here are some sampled values:

1, 1, 2, 0, 1, 1, 1, 1, 3, 3,

1, 2, 0, 3, 1, 1, 3, 3, 1, 2

Collect the observed counts:

$O_0 = 2$, $O_1 = 10$, $O_2 = 3$, $O_3 = 5$

(total is $n=20$)

Here is the distribution for H_0:

x	0	1	2	3
$P(X = x)$	0.2	0.4	0.1	0.3

When H_0 is true, the expected counts are:

$$E_0 = (20)(0.2) = 4$$

$$E_1 = (20)(0.4) = 8$$

$$E_2 = (20)(0.1) = 2$$

$$E_3 = (20)(0.3) = 6$$

Consider the test statistic:

$$W := \sum_{i=0}^{3} \frac{(O_i - E_i)^2}{E_i}$$

Claim: Under H_0, W has roughly a

$\chi^2(3)$ distribution!

number of
categories
minus 1

Claim: Under H_0, W has roughly a $\chi^2(3)$ distribution!

This is a result of

- The Central Limit Theorem which says that

$$O_i = \sum_{i=j}^{n} I_{\{X_j = i\}}$$

gets normal in the limit.

Claim: Under H_0, W has roughly a $\chi^2(3)$ distribution!

This is a result of

- The fact that a $N(0, 1)$ random variable squared has a $\chi^2(1)$ distribution.
- The fact that a $N(0, 1)$ random variable squared has a $\chi^2(1)$ distribution.

Claim: Under H_0, W has roughly a $\chi^2(3)$ distribution!

This is a result of

- The fact that a sum of k independent $\chi^2(1)$ random variables has a $\chi^2(k)$ distribution

Claim: Under H_0, W has roughly a $\chi^2(3)$ distribution!

However, it is complicated by the fact that $O_0 + O_1 + O_2 + O_3 = 20$, so these 4 random variables are not independent.

In general, for k categories and n observations,

$$W := \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{\text{approx}}{\sim} \chi^2(k-1)$$

for “large” sample sizes.

- “Large” n is not quite enough.
- If you have a large sample but one of the true probabilities is very small

x	0	1	2	3
$P(X = x)$	0.2	0.4	1×10^{-12}	**the rest**

then you still will have a difficult time getting observations of the outcome 2.

Rule of Thumb: Want the expected number (under H_0) in each category to be at least 5.

Back to the Original Example:

x	0	1	2	3
$P(X = x)$	0.2	0.4	0.1	0.3

When H_0 is true, the expected counts are:

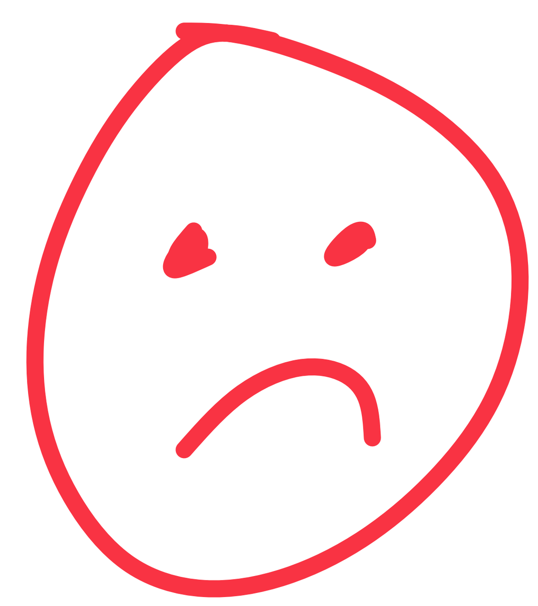
$$E_0 = (20)(0.2) = 4$$

$$E_1 = (20)(0.4) = 8$$

$$E_2 = (20)(0.1) = 2$$

$$E_3 = (20)(0.3) = 6$$

Need more data!



Increased sample size to 100 and observed

$O_0 = 18, \quad O_1=33, \quad O_2=12, \quad O_3= 37$

$$W := \sum_{i=0}^{3} \frac{(O_i - E_i)^2}{E_i} \approx 3.458$$

Reject H_0 if W is “large”.

Increased sample size to 100 and observed

$O_0 = 18$, $O_1 = 33$, $O_2 = 12$, $O_3 = 37$

$$W := \sum_{i=0}^{3} \frac{(O_i - E_i)^2}{E_i} \approx 3.458$$

For a test of size $\alpha = 0.05$, "large" means $W > \chi^2_{0.05,3} \approx 7.8147$.

Conclusion:

We fail to reject H_0 at level 0.05.

It appears that the data did come from the distribution

x	0	1	2	3
$P(X = x)$	0.2	0.4	0.1	0.3

```
> mysample<-sample(c(0,1,2,3),100,
  replace=T,prob=c(0.2,0.4,0.1,0.3))
> table(mysample)
mysample
 0  1  2  3
18 33 12 37
> obs<-c(18,33,12,37)
> exp<-100*c(0.2,0.4,0.1,0.3)
> W<-sum(((obs-exp)^2)/exp)
> W
[1] 3.458333
> qchisq(0.95,3)
[1] 7.814728
> 
```

H_0 : The sample comes from a binomial distribution with parameters 8 and 0.2.

You have n observations of

0's, 1's, …, 8's.

Count up observations:

$$O_0, O_1, \ldots, O_8$$

Expected numbers:

$E_i = np_i$ where

$$p_i = P(X = i) = \binom{8}{i} 0.2^i (1 - 0.2)^{n-i}$$

H_0 : The sample comes from the N(0,1) distribution.

- Continuous data!
- Group data values into bins and do the test on this finite number of categories.
- Test can be sensitive to your choice of bins!
- Try a few different bin widths. Be leery of results if they are highly variable.