Chi-square tests
Chi-square Goodness of Fit Test

Set up: Count data on a **categorical variable**.

- There are k categories, labeled as $i = 1, \ldots, k$.
- Observed data: O_i, $i = 1, \ldots, k$, where $O_i = \#$ of observations in the i-th category.
- $n = \sum_{i=1}^{k} O_i$ is the total number of observations.

Hypotheses: H_0 : The data follow a given model, versus, H_1 : The data don't follow the given model.

- Let $p_i =$ proportion of observations in the population that fall in the i-th category, $i = 1, \ldots, k$. We can also think of p_i as the probability that a randomly selected observation from the population falls in the i-th category.
- $H_0 : p_i = p_{i,0}, i = 1, \ldots, k$, where $p_{i,0}$ are known proportions that add up to 1.
- Model is **completely known** under H_0.

Case 1: Two categories

Case 2: More than two categories

Basic idea: Compare O_i's with E_i's — counts expected assuming H_0 is true — using a **chi-square statistic**

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

- Large χ^2 :
- Reject H_0 when
- $E_i =$
- Null distribution:
- Rule of thumb: All $E_i \geq 5$. Collapse adjacent categories if this is not the case.

Ex: Suppose 60 independent rolls of a die lead to the following data.

category	1	2	3	4	5	6	total
observed count (O_i)	4	6	17	16	8	9	60
expected count (E_i)							

Is the die fair? Answer this question by performing an appropriate test of hypothesis at 5% level of significance.

R code:

```
> x <- c(4, 6, 17, 16, 8, 9)
> sum(x)
[1] 60
> sum( (x-10)^2/10)
[1] 14.2
> 1-pchisq(14.2, 5)
[1] 0.01438768
>
```

Chi-square test of independence

Set up: Count data on two categorical variables (or factors) A and B obtained from a sample of n subjects. Suppose the categories of A are $i = 1, \ldots, k$, and the categories of B are $j = 1, \ldots, m$. The data are arranged in a $k \times m$ table. Let O_{ij} = observed count in (i, j)-th cell.

Hypotheses: H_0 : A and B are independent (i.e., are not associated), vs., H_1 : A and B are not independent (i.e., are associated). If there is an association, the value one variable depends (at least to some extent) on the value of the other.

Example: The table below shows 695 children under 15 years of age are cross-classified according to ethnic group and hemoglobin level. Is hemoglobin level associated (related) to ethnicity?

	Hemoglobin Level (g/100 ml)				
Ethnic Group	≥ 10	9.0 - 9.9	< 9.0	Total	Proportion
A	80	100	20	200	
B	99	190	96	385	
C	70	30	10	110	
Total	249	320	126	695	
Proportion					

- If He level is not associated to ethnicity, then the proportion of subjects in population that fall a He group does not depend on ethnicity, i.e., it is the same for each ethnicity group, and vice versa.

To do a chi-square test, we need the expected counts E_{ij} assuming that H_0 is true. Let X and Y indicate respective categories of A and B in which a randomly selected subject from the population falls. When A and B are independent,

$$P(X = i, Y = j) = P(X = i)P(Y = j) \text{ for all } i, j.$$

- $P(X = i)$ is estimated as
- $P(Y = j)$ is estimated as
- Assuming independence, $P(X = i, Y = j)$ is estimated as
- Assuming independence, E_{ij} is estimated as

Test statistic:

Degrees of freedom:

Example (continued): The expected counts for all cells (in parenthesis below next to the observed counts) are:

Ethnic Group	Hemoglobin Level (g/100 ml)			Total
	≥ 10	9.0 - 9.9	< 9.0	
A	80 ( )	100 (92.09)	20 (36.26)	200
B	99 (137.94)	190 (177.27)	96 (69.80)	385
C	70 (39.41)	30 (50.65)	10 (19.94)	110
Total	249	320	126	695

R code:

```
> x <- c(80, 100, 20, 99, 190, 96, 70, 30, 10)
> xmat <- matrix(x, byrow=T, ncol=3)
> xmat
     [,1] [,2] [,3]
[1,]   80  100   20
[2,]   99  190   96
[3,]   70   30   10
> chisq.test(xmat)

        Pearson's Chi-squared test

data:  xmat
X-squared = 67.8015, df = 4, p-value = 6.606e-14

>
```

Chi-Square test of Homogeneity

Often we are interested in comparing different populations with respect to a variable of interest, e.g., are the populations of carriers and non-carriers of a certain antigen *homogeneous* with respect to blood type?

Example: A sample of 150 carriers of a certain antigen and a sample of 500 non-carriers showed the following blood group distributions:

Blood Group	Carriers	Non-Carriers	Total
O	72	230	302
A	54	192	246
B	16	63	79
AB	8	15	23
Total	150	500	650

Are carriers and non-carriers similar with respect to blood group distributions?

Test of Homogeneity vs. Test of Independence

Comparing the layout of this table with the table for the test of independence, we see that the two layouts are
Thus, mathematically the tests of homogeneity and independence are exactly the same. So, the same formulas apply. However, there are some key conceptual differences.

Sampling procedure:

- *Test of independence*: one overall sample is collected first and then each observation is classified by levels of the two variables. So, neither row nor column totals are fixed in advance.
- *Test of homogeneity*: several samples are collected from several populations with each sample size fixed in advance. After collecting these pre-determined # of observations, each is classified by various levels of one variable. So, in the above example, ... totals are fixed.

Number of variables:

- *Test of independence*: **two** variables.
- *Test of homogeneity*: **one** variable. The column/row representing "population" is fixed due to the sampling process.

Hypotheses:

- *Test of independence*: H_0:
- *Test of homogeneity*: H_0:

R code:

```
x <- c(72, 230, 54, 192, 16, 63, 8, 15)
xmat <- matrix(x, byrow=T, ncol=2)
# > xmat
     # [,1] [,2]
# [1,]   72  230
# [2,]   54  192
# [3,]   16   63
# [4,]    8   1
# >
# > chisq.test(xmat)
# Pearson's Chi-squared test

# data:  xmat
# X-squared = 2.4052, df = 3, p-value = 0.4927
# >
```