Activity #5
How Typical are our Households' Ages?
The Chi–Square Test*

Introduction

Consider a sample made up of the people in the households of each pearson in our class. Would their ages be representative of the ages of all households in the United States? Probably not. After all, it would not be a random sample. But how unrepresentative would the sample be with regard to ages of people in the United States?

Question

The latest census in the United States yielded the following age distribution†:

Age Range	Proportion in US Population
0–18	0.26
18–64	0.62
64+	0.12

Table 1: US 2000 Census: Household Age Distribution

Is the age distribution of the people from the households in the class typical of that of all residents in the United States?

Discussion

1. Would you expect the age distribution from the households in your class to be typical of the age distribution of all US residents? Why or why not?

*Adapted from *Activity–Based Statistics,* Second Edition, by Scheaffer, Watkins, Witmer and Gnanadesikan. Key College Publishing, 2004

†Source: `http://www.census.gov/prod/2001pubs/c2kbr01-12.pdf`

2. If there is any discrepancy between the two distributions, how would you measure that discrepancy?

Data Collection

In the following table record the number of people in households from this class for each age range. In the last column, assuming that the class distribution will follow the national distribution, record the number of people you would expect to see in each age range.

Age Range	Number Observed in Households from this Class	Expected Number in the Class
0–18		
18–64		
64+		
Total		

Table 2: Class Household Age Data

Analysis

If we call the observed numbers in the second column O_1, O_2 and O_3, for each age range respectively, and E_1, E_2 and E_3 the corresponding expected values, then one way to measure the discrepancy between the two distributions is by computing the statistic

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3}.$$

This is usually denoted by χ^2 and is known as the *Chi–square* statistic.

Is the χ^2 statistic you computed large enough to convince you that the ages of the households from this class are not similar to those of a typical random sample of US residents?

Simulations

Simulate drawing random samples of the ages from the population of US residents to see how often we get a value of the χ^2 statistic that is as large or larger than the one from the class.

Discuss how you would use R to run this simulation.

Conclusion

Based on your simulations, estimate the probability that a random sample of US residents (of the same size as the sample in the class) would have a χ^2 statistic as large or larger than the one obtained in class.

Is the distribution of ages of the households from the class similar to those of a typical random sample of US residents? What do you conclude?