A Place where the Binomial Distribution is Not Applicable.

The goal is to estimate the proportion of people in a large city who consulted a medical doctor in the past year. One approach is to select a random sample of a fixed number of persons is selected from those persons living in the city, and to ask each person selected

Did you or did you not consult a medical doctor in the past year?

Then, the count of the total number of persons who saw a medical doctor in the past year will follow a binomial distribution.

However, in many surveys, a random sample of households is selected instead of a random sample of persons. Still, each person in a selected household is asked whether or not they consulted a medical doctor in the past year. The sample total number of persons who saw a doctor no longer follows a binomial distribution.

When one person in a household goes to a doctor, typically the other members of the household are likely to go to see a doctor as well. Moreover, either all members of a family have health insurance or none of them have health insurance hence all would go for checkups or none would go. This clumping or clustering effect violates the independence of trials where each trial is a person having seen or not seen a doctor in the past year.

To estimate the proportion of persons in the city who consulted a doctor in the past year, we could consider the reasonable estimate

Unfortunately, this deviates even further from results based on the binomial Formula. The denominator is no longer the fixed number of trials. Because the number of persons in a single household is a random variable so is the total sample size or total number of persons in the sample is also a random variable. The confidence interval or error margin for the proportion, based on the binomial distribution, can not be applied. when the sample size is not fixed..

For simplicity, consider households of size 4. For an individual household, if all persons have the same response, their answer will be either 0 or 4 who saw a doctor in the past year. These two extreme answers cause the variance to be larger than for a binomial random variable with $n = 4$. The variance of the total from all households, which is the sum of the variances for each house, will be larger than the corresponding variance for the total number of persons in the sample who saw a doctor in the past year.

Exercise. Consider a household of size 4 selected from a large households of size 4 in a large city. If half of the persons in the city saw a doctor in the last year, and persons made decisions independently, the binomial distribution

with $n = 4$ and $p = .5$ would apply to X =number who consulted a doctor in the past year. If everyone in the household responded alike, the distribution of X would be $P[X = 0] = P[X = 4] = .5$. Calculate the mean and variance for the binomial distribution and this latter distribution. Compare the means and variances.