Handout 4

REVIEW.

- A binomial is a sum of *independent* binaries: if independence is not plausible, the hypergeometric distribution may be the correct choice:

$$P(x; N, m, n) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \tag{1}$$

  The parameters are: N = Population size, m = Number of defective, n = Sample size. In words, expression (1) says that the probability of having x defective parts what taking a sample from a population of N parts of which m are defective is the total number of way in which we can pick x defectives in a sample of n divided by the total number of samples of size n from a population of N.

- Tables for the hypergeometric do exist, but they cannot possibly cover every combination ob N, n and m. You can use functions `{d,p,q,r}hyper` in R or you can compute the necessary values with applications such as PROBABILITY DISTRIBUTIONS by M. Bognar[1] among many others.

- Continuity correction can be neglected for very large numbers, otherwise it needs to be taken into account. The basic rule is that the probability of a discrete value k in a binomial is close to the probability of the interval $(k - \frac{1}{2}, k + \frac{1}{2})$ in the approximating normal.

PROBLEMS.

1. Using normal approximations compute the following probabilities:

   (a) $P(x \leq 50)$ if $X \sim b(p = 0.3, n = 100)$

   (b) $P(x = 20)$ if $X \sim b(p = 0.25, n = 100)$

   (c) $P(x > 20)$ if $X \sim b(p = 0.25, n = 100)$

   (d) $P(20 \leq x \leq 30)$ if $X \sim b(p = 0.25, n = 100)$

2. In the following cases, say which distribution you would use to compute the required probabilities (and give their values).

   (a) Assume you have a regular coin that you throw 10 times. What is the probability of 6 heads?

[1] See `http://homepage.divms.uiowa.edu/~mbognar/` and the usual repositories for Android and iOS devices.

(b) Assume you have a box with names of football teams, 10 from Spain and 10 from other countries. You have to play a set of matches with four teams taken at random. What is the probability that they are all from Spain?

(c) You have 100 fuses of which 5 are defective. What is the probability that you discover 3 when a sample of 20 are inspected?

(d) Same as previous, but 100.000 fuses and 5.000 defective, and you still pick a sample of 20.

(e) The probability of a person in a given population on size $N = 300000$ contracting a certain disease is $p = 0.01$. What would be the probability of $x = 2000$ cases of more? What assumptions are you making? How would your answer change, if at all, if you were told that the disease is highly contagious?

3. You are charged with auditing a company in which 10 clerks are in charge of making invoices. You have to check that invoices are correct (tax assessed correctly, no numeric or pricing errors, etc.). The total number of invoices in a given year is 13000, roughly one tenth of which are the work of each clerk.

   (a) If the true proportion of errors is 2%, what is the distribution of the number of errors in a sample of $n = 100$ randomly chosen invoices?

   (b) Do you think a normal approximation would be good in this case?

   (c) Each clerk keeps copies of the invoices he/she has made. You reason that rather than disturbing all 10 clerks you might as well pick at random two clerks and check 50 invoices from each[2]. Would the distribution of the number of errors you uncover be the same as before if some clerks are notoriously less careful than the rest?

[2]We may come back to this under the name of *conglomerate sampling* about the end of the course.

Pautas docentes

1. Problemas un poco reiterativos, que no necesitan ser hechos todos, y pueden ser sustituidos por otros.

2. Hay que enfatizar especialmente que la distribución binomial es un buen modelo en unos casos y no en otros: el ejercicio 2 plantea diversos supuestos para ayudarles a distinguir.