Lecture Stat 302
Introduction to Probability - Slides 12

AD

March 2010

Hypergeometric Random Variable

- Consider a barrel or urn containing N balls of which m are white and $N - m$ are black. We take a simple random sample (i.e. without replacement) of size n and measure X, the number of white balls in the sample.
- The Hypergeometric distribution is the distribution of X under this sampling scheme and

$$P(X = i) = \frac{\begin{pmatrix} m \\ i \end{pmatrix} \begin{pmatrix} N - m \\ n - i \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

Example: Survey sampling

- Suppose that as part of a survey, 7 houses are sampled at random from a street of 40 houses in which 5 contain families whose family income puts them below the poverty line. What is the probability that: (a) None of the 5 families are sampled? (b) 4 of them are sampled? (c) No more than 2 are sampled? (d) At least 3 are sampled?
- Let X the number of families sampled which are below the poverty line. It follows an hypergeometric distribution with $N = 40$, $m = 5$ and $n = 7$. So (a) $P(X = 0)$ (b) $P(X = 4)$ (c) $P(X \leq 2)$ and (d) $P(X \geq 3)$

Mean and Variance of the Hypergeometric Distribution

- Let us introduce $p = m/N$ then

$$E(X) = np, \quad Var(X) = np(1-p)\left(1 - \frac{n-1}{N-1}\right).$$

- Suppose that m is very large compared to n, it seems reasonable that sampling without replacement is not too much different than sampling with replacement. It can indeed be shown that the hypergeometric distribution can be well approximated by the binomial of parameters $p = \frac{m}{N}$ and N.

Example: Capture-Recapture Experiments

- We are interested in estimating the population N of animals inhabiting a certain region. To achieve this, capture-recapture studies proceed as follows. First, you capture m individuals, mark them and release them in the nature. A few days later, you capture say n animals; among the n animals some of them are marked and some are not. Let X be the number of animals which have been recaptured, then X follows an hypergeometric distribution of parameters N, m and n.
- Assume you have recaptured $X = x$ animals, then you can *estimate* N by maximizing with respect to N the probability

$$P(X = x) = \binom{m}{x}\binom{N-m}{n-x} / \binom{N}{n}.$$

  This known as the Maximum Likelihood (ML) estimate of N.
- One can show that the ML estimate is the largest integer value not exceeding mn/x; i.e. $\widehat{N} = \lfloor mn/x \rfloor$.

Example: Counting Fishes

- Suppose that we have a lake containing N fishes where N is unknown. We capture and mark $m = 100$ fishes. A few days later, we capture $n = 50$ fishes, X of them are marked. What is the ML estimate of N if $X = 35$ and $X = 5$?
- If $X = 35$ then the ML estimate of N is $\widehat{N} \approx m \times n/35 = 142$.
- If $X = 5$ then the ML estimate of N is $\widehat{N} \approx m \times n/5 = 1000$.

Example: Estimating the size of an hidden population

- Capture-recapture ideas have been used to provide an estimate of the size of a population that cannot be directly counted. It is particularly useful for estimating the size of hidden populations for example it could be used to estimate criminal populations, victims of domestic violence or people with undiagnosed diseases.
- When applying this technique to human populations, the two random samples are usually replaced with lists of individuals constructed from two or more data sources such as police data, treatment data etc.
- The appearance of individuals in more than one dataset, or the overlap between lists, is equivalent to the 'recapture' of a fish already captured in a previous sample.

Example: Counting the number of Injecting Drug Users (IDU)

- Registry data were collected from the Drug Use Rehabilitation Center in Leshan Downtown district (Sichuan province) (J. Health Science, 2005).
- Capture data correspond to people entering the center and recapture data correspond to people reentering the center several months after having been released.
- We have $m = 71$ (Nov. 2001-Feb. 2002), $n = 191$ (July 2002-Oct. 2002) and $X = 3$ so $\widehat{N} \approx m \times n/3 = 4502$.

Midterm Practice: Exercise 1

- A carton contains twenty chocolate bars, three of which have been injected with a deadly poison. Unfortunately, you have eaten five bars, chosen at random from the carton.
- (a) What is the proba that you will suffer no ill effects?
- (b) What is the proba only one of the three bars you ate was poisoned?
- (c) What is the proba that you have eaten at least one poisoned bar?

Exercise 1

- (a) To suffer from no ill effects, you need to have eaten no poisoned bar. There are
$$\binom{20}{5}$$
ways to pick 5 bars out of 20 bars and there are
$$\binom{3}{0}\binom{17}{5}$$
ways to pick no poisoned bar. So the probability of no ill effects is
$$\frac{\binom{3}{0}\binom{17}{5}}{\binom{20}{5}} = 0.3991$$

Exercise 1

- (b) There are

$$\begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 17 \\ 4 \end{pmatrix}$$

ways to pick exactly one poisoned bar. So the proba is

$$\frac{\begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 17 \\ 4 \end{pmatrix}}{\begin{pmatrix} 20 \\ 5 \end{pmatrix}} = 0.4605$$

- (c) We have

$$\begin{aligned} & P\left(\text{at least one poisoned bar}\right) \\ = \; & 1 - P\left(\text{no poisoned bar}\right) = 1 - 0.3991 = 0.6009 \end{aligned}$$

Exercise 2: Free throws at basketball

- Kevin is shooting 10 free throws, each throw is independent of the others and has a success probability $p = 0.7$.
- (a) What is the proba that Kevin is going to score more than 8 times?
- (b) What is the proba that he is going to score only one time during his first 3 trials and 7 times overall?
- (c) What is the conditional proba that he is going to score only 2 times in his last 7 trials after having scored 3 times in his first three throws?

Exercise 2: Free throws at basketball

- This is a typical Binomial distribution problem where you have independent Bernoulli trials, each with success proba p.
- (a) Let X be the number of times Kevin is going to score in 10 trials then X follows a Binomial distribution of parameters $n = 10$, $p = 0.7$ so the answer is

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10)$$

where

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$$

so

$$P(X \geq 8) = 0.2335 + 0.1211 + 0.0282 = 0.3828$$

Exercise 2: Free throws at basketball

- (b) Let X_1 be the number of successful throws during the first 3 trials and X_2 the number of successful throws during the remaining 7 trials. Then X_1 follows a Binomial distribution of parameter $n_1 = 3$, $p = 0.7$ and X_2 follows a Binomial distribution of parameter $n_2 = 7$, $p = 0.7$.
- We are interested in the proba of scoring only one time during the first 3 trials and 7 times overall so this is

$$\begin{aligned} P\left(X_1 = 1 \ \cap \ X_2 = 6\right) &= P\left(X_1 = 1\right) P(X_2 = 6) \\ &= \begin{pmatrix} 3 \\ 1 \end{pmatrix} p^1 \left(1-p\right)^2 \times \begin{pmatrix} 7 \\ 6 \end{pmatrix} p^6 \left(1-p\right)^1 \\ &= 0.0467 \end{aligned}$$

Exercise 2: Free throws at basketball

- We are interested in conditional proba that he is going to score only 2 times in his last 7 trials after having scored 3 times in his first three throws.
- This is

$$
\begin{aligned}
P\left(X_2 = 2 \middle| X_1 = 3\right) &= \frac{P\left(X_2 = 2 \cap X_1 = 3\right)}{P\left(X_1 = 3\right)} \\
&= \frac{P\left(X_2 = 2\right) P\left(X_1 = 3\right)}{P\left(X_1 = 3\right)} \\
&= P\left(X_2 = 2\right) \\
&= \binom{7}{2} p^2 \left(1 - p\right)^5 \\
&= 0.0250
\end{aligned}
$$

Exercise 3: Newspaper subscription

- If 50 percent of families in a certain city subscribe to the morning newspaper, 65 percent subscribe to the afternoon newspaper, and 85 percent of the families subscribe to at least one of the two newspapers, what proportion of families subscribe to both newspapers?
- Let M ="subscribe morning newspaper", A ="subscribe afternoon newspaper" and $A \cup M$ ="subscribe morning and/or afternoon newspaper", we know that

$$P(A \cup M) = P(A) + P(M) - P(A \cap M)$$

so

$$\begin{aligned} P(A \cap M) &= P(A) + P(M) - P(A \cup M) \\ &= 0.5 + 0.65 - 0.85 = 0.30 \end{aligned}$$

Exercise 4: Two cameras

- The flash mechanism on camera A fails on 10% of the shots, while that on camera B fails on 5% of the shots. These failures are independent. The cameras are used simultaneously for an indoor shot.
- (a) What is the probability that at least one of the camera flash machine fails?
- (b) What is the probability that camera A works given at least one of the camera flash machine fails?

Exercise 4: Two cameras

- (a) Let us introduce the events $E = \{\text{camera flash } A \text{ fails}\}$ and $F = \{\text{camera flash } B \text{ fails}\}$ then the event that at least one of the camera flash machines fails is $E \cup F$ and

$$
\begin{aligned}
P(E \cup F) &= P(E) + P(F) - P(E \cap F) \\
&= P(E) + P(F) - P(E)P(F) \\
&= 0.1 + 0.05 - 0.1 \times 0.05 = 0.1450
\end{aligned}
$$

Exercise 4: Two cameras

- (b) We are interested in

$$P\left(E^c \mid E\cup F\right) = 1 - P\left(E \mid E\cup F\right)$$

where

$$\begin{aligned} P\left(E \mid E\cup F\right) &= \frac{P\left(E\cap\left(E\cup F\right)\right)}{P\left(E\cup F\right)} \\ &= \frac{P\left(E\right)}{P\left(E\cup F\right)} = \frac{0.1}{0.1450} = 0.69 \end{aligned}$$

so

$$P\left(E^c \mid E\cup F\right) = 0.31$$

Exercise 5: Watering plants

- You ask your neighbour to water a plant while you are on vacation. Without water, it will die with proba. 0.8. With water, it will die with proba 0.15. You are 90% certain that your neigbour will remember to water the plant.
- (a) What is the proba that the plant will be alive when you return?
- (b) If it is dead what is the proba that your neighbour forgot to water it?
- (c) If it is alive, what is the proba that your neighbour watered it?

Exercise 5: Watering plants

- (a) Let A ="plant alive" and W ="plant watered", then the proba the plant will be alive is

$$\begin{aligned} P(A) &= P(A \cap W) + P(A \cap W^c) \\ &= P(A|W)P(W) + P(A|W^c)P(W^c) \end{aligned}$$

where $P(W) = 1 - P(W^c) = 0.9$, $P(A|W) = 1 - 0.15 = 0.85$ and $P(A|W^c) = 1 - 0.8 = 0.2$ so

$$P(A) = 0.85 \times 0.9 + 0.2 \times 0.1 = 0.785$$

- (b) We are now interested in

$$P(W^c|A^c) = \frac{P(W^c \cap A^c)}{P(A^c)} = \frac{P(A^c|W^c)P(W^c)}{P(A^c)}$$

where $P(A^c) = 1 - P(A) = 0.215$, $P(A^c|W^c) = 0.8$ so

$$P(W^c|A^c) = 0.3721$$

Exercise 5: Watering plants

- (c) We are interested in

$$\begin{aligned} P(W|A) &= \frac{P(A|W)\,P(W)}{P(A)} \\ &= \frac{0.85 \times 0.9}{0.785} = 0.9745 \end{aligned}$$

so you can be confident he/she watered the plant....

Exercise 6: Hospital Beds

- Each day a hospital makes available two beds to each of two surgeons (four beds in total). The demand each surgeon has for these beds is assumed to be independently Poisson with expectation 1.
- (a) Calculate the probability that the demand for beds for a particular surgeon exceeds those available
- (b) Calculate the probability that the demand for beds for at least one of the two surgeons exceeds the number available.

Exercise 6: Hospital Beds

.

- (a) Let X_i the demand for beds for surgeon i; $i = 1, 2$. We know that X_i is Poisson with $\lambda = 1$ so

$$\begin{aligned} P(X_i > 2) &= 1 - P(X_i \leq 2) \\ &= 1 - e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right) = 0.0803 \end{aligned}$$

- (b) The demand for beds for at least one of the two surgeons exceeds the number available is the event $\{X_1 > 2\} \cup \{X_2 > 2\}$ and

$$\begin{aligned} & P(\{X_1 > 2\} \cup \{X_2 > 2\}) \\ = \;& P(X_1 > 2) + P(X_2 > 2) - P(\{X_1 > 2\} \cap \{X_2 > 2\}) \\ = \;& P(X_1 > 2) + P(X_2 > 2) - P(X_1 > 2) P(X_2 > 2) \\ = \;& 0.1542 \end{aligned}$$

- The number of students arriving at the SUB during lunch averages 6.2 people per minute. (a) For a given two minutes during lunch, what is the proba that no more than 5 students arrive in that two minutes? (b) What is the proba that no more than 5 students arrive in that two minutes given that you know that at least two of your friends arrive in that two minutes? Assume that the Poisson distribution holds.
- (a) Let X be the number of students arriving in a given two minutes, then X follows a Poisson distribution of parameter $\lambda = 2 \times 6.2 = 12.4$. So we have

$$\begin{aligned} P(X \leq 5) &= e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\frac{\lambda^4}{4!}+\frac{\lambda^5}{4!}\right) \\ &= 0.0158 \end{aligned}$$

Exercise 7: Arrivals to the SUB

- If 2 of your friends have already arrived then you want to compute

$$
\begin{aligned}
P\left(X \leq 5 \mid X \geq 2\right) &= \frac{P\left(X \leq 5 \cap X \geq 2\right)}{P\left(X \geq 2\right)} \\
&= \frac{P\left(2 \leq X \leq 5\right)}{P\left(X \geq 2\right)}
\end{aligned}
$$

  where

$$
\begin{aligned}
P\left(2 \leq X \leq 5\right) &= e^{-\lambda}\left(\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\frac{\lambda^4}{4!}+\frac{\lambda^5}{4!}\right) = 0.0157, \\
P\left(X \geq 2\right) &= 1 - e^{-\lambda}\left(1+\lambda\right) = 0.9999
\end{aligned}
$$

- So we have

$$
P\left(X \leq 5 \mid X \geq 2\right) = 0.0157
$$

Exercise 8: Great Expectations

- In the Virginia State Lottery game "Cash 5," you pick 5 numbers from a set of 34 and try to match the 5 numbers chosen at random by the state. For a 1$ ticket, the payoff is

nb. correct numbers	payoff in \$
5	100, 000
4	100
3	5
2	0
1	0

- What is the expected payoff of this game?

Exercise 8: Great Expectations

- Let's call Y the gain then

  $$E(Y) = 100,000 \times P(Y = 100,000) + 100 \times P(Y = 100)$$
  $$+5 \times P(Y = 5)$$
  $$= 100,000 \times P(X = 5) + 100 \times P(X = 4) + 5 \times P(X = 3)$$

  where X is the number of correct numbers.
- We have for $i = 0, 1, 2, ..., 5$

  $$P(X = i) = \frac{\binom{5}{i}\binom{29}{5-i}}{\binom{34}{5}}$$

  so

  $$E(Y) = 0.48\$$$
- Hence the expected payoff is

  $$1\$ - E(Y) = -0.52\$$$

Exercise 9: Uniform discrete distribution

- An important discrete probability distribution not covered in the lectures is the discrete uniform distribution. In this scenario X can take the values $\{1, 2, 3, ..., n\}$ with equal probability. Compute $E(X)$.
- We have

$$
\begin{aligned}
E(X) &= \sum_{i=1}^{n} i \times P(X = i) \\
&= \frac{1}{n}\sum_{i=1}^{n} i \quad \text{as} \quad P(X = i) = \frac{1}{n} \\
&= \frac{1}{n}\frac{n(n+1)}{2} = \frac{(n+1)}{2}
\end{aligned}
$$

Exercise 10: Local gig

- You have organized a gig in your garage. The admission charge is 2$ per person. To convince people to come, you have organized a lottery; the first and unique prize costs you 20$. The number of spectators that will turn up is assumed to have a Poisson distribution of mean 10.
- (a) What is the expected profit?
- (b) What is the standard deviation of the profit?
- (c) What is the proba of making a loss?

Exercise 10: Local gig

- Let's call Y the profit and X the number of persons who will come. Then we have

$$Y = 2X - 20$$

and

$$\begin{aligned} E(Y) &= E(2X-20) = 2E(X) - 20 \\ &= 20 - 20 = 0 \end{aligned}$$

- The standard deviation is given by

$$Std(Y) = \sqrt{Var(Y)}$$

where

$$Var(Y) = Var(2X-20) = 2^2 Var(X) = 40$$

so $Std(X) = 6.3246\$$

Exercise 10: Local gig

- The probability of making a loss is

$$\begin{aligned} P(Y<0) &= P(2X-20<0) \\ &= P(X<10)=\sum_{i=0}^{9} P(X=i) \\ &= 0.4768 \end{aligned}$$