Lecture Stat 302
Introduction to Probability - Slides 11

AD

March 2010

Discrete Random Variables

- A discrete r.v. X takes at most a countable number of possible values $\{x_1, x_2, ...\}$ with *p.m.f.*

$$p(x_i) = P(X = x_i)$$

where

$$p(x_i) \geq 0 \text{ and } \sum_{i=1}^{\infty} p(x_i) = 1.$$

- *Expected value/mean*

$$\mu = E(X) = \sum_{i=1}^{\infty} x_i \; p(x_i) \, .$$

- *Variance*

$$Var(X) = E\left((X-\mu)^2\right) = E\left(X^2\right) - \mu^2.$$

Poisson Random Variable

- A discrete r.v. X taking values $0, 1, 2, ...$ is said to be a Poisson r.v. with parameter λ, $\lambda > 0$, if

$$p(i) = P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}.$$

- This expresses the probability of a number of events occurring in a fixed period of time if these events occur with a *known average rate* λ.
- If we consider a binomial r.v. X of parameters (n, p) such that n is large and p is small enough so that np is moderate then the binomial distribution can be well-approximated by the Poisson distribution of parameter $\lambda = np$.

Example: Waiting at the bus stop

- *Example*: Assume that 6 buses per hour stop at your bus stop. If the buses were arriving exactly every 10 minutes, then your average waiting time would be 5 minutes; i.e. you wait between 0 and 10 minutes. What is the probability that you are going to wait at least 5 minutes without seeing any bus if the buses follow a Poisson distribution? What is the proba to wait at least 10 minutes? What is the proba of seeing two buses in 10 minutes?
- *Answer*: If we let X_1 be the number of buses arriving in 5 minutes, it is a Poisson r.v. with parameter 0.5 (average rate 6 per hour). So we have

  $$P\left(X_1 = 0\right) = e^{-0.5} = 0.60.$$

  If we let X_2 be the number of buses arriving in 10 minutes, it is a Poisson r.v. with parameter 1. So we have

  $$P\left(X_2 = 0\right) = e^{-1} = 0.368, \; P\left(X_2 = 2\right) = e^{-1}\frac{1^2}{2!} = 0.184.$$

Mean of a Poisson Random Variable

- We have $E(X) = \lambda$.
- We have

$$\begin{aligned} E(X) &= \sum_{i=0}^{\infty} i\ P(X = i) = \sum_{i=0}^{\infty} i\ e^{-\lambda} \frac{\lambda^{i}}{i!} \\ &= e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda\ \lambda^{i-1}}{(i-1)!} \\ &= e^{-\lambda} \lambda \sum_{j=0}^{\infty} \frac{\lambda^{j}}{j!} \quad (\text{change } j \leftarrow i-1) \\ &= e^{-\lambda} \lambda e^{\lambda} = \lambda. \end{aligned}$$

- Note that this is in agreement with our approximation of Binomial by Poisson. A Binomial has mean np and we approximate it by a Poisson of parameter $\lambda = np$ which is also the mean of the Poisson distribution.

Variance of a Poisson Random Variable

- We have $Var\left(X\right)=E\left(X^{2}\right)-E\left(X\right)^{2}=\lambda$.
- Proof: We have

$$\begin{aligned} E\left(X^{2}\right) &= \sum_{i=0}^{\infty} i^{2}\ e^{-\lambda}\frac{\lambda^{i}}{i!}=\sum_{i=1}^{\infty} i^{2}\ e^{-\lambda}\frac{\lambda^{i}}{i!} \\ &= e^{-\lambda}\sum_{i=1}^{\infty} i\ \frac{\lambda\ \lambda^{i-1}}{(i-1)!}=\lambda\sum_{j=0}^{\infty}(j+1)\ \ e^{-\lambda}\frac{\lambda^{j}}{j!} \\ &= \lambda E\left(Y+1\right) \end{aligned}$$

  where Y is a Poisson random variable of parameter λ, hence $E\left(Y+1\right)=\lambda+1$.
- We can now conclude

$$Var\left(X\right)=\lambda\left(\lambda+1\right)-\lambda^{2}=\lambda.$$

- Note that for Poisson random variable, we have $E\left(X\right)=Var\left(X\right)=\lambda$. For binomial we have $E\left(X\right)=np$ and $Var\left(X\right)=np\left(1-p\right)$ and so $Var\left(X\right)\approx E\left(X\right)$ only if $p<<1$.

Example: Brain Cancer

- Brain cancer is a rare disease. In any year there are about 3.1 cases per 100,000 of population (US figures from TIME). Suppose a small medical insurance company has 150,000 people on their books. How many claims stemming from brain cancer should the company expect in any year? What is the probability of getting more than 2 claims related to brain cancer in a year?
- Assume you use the Poisson approximation to the Binomial, then $\lambda = 3.1\frac{150000}{100000} = 4.65$. So if we denote X the number of claims stemming from brain cancer then

$$E(X) = \lambda.$$

- The probability of getting more than 2 claims is

$$\begin{aligned} P(X > 2) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2}{2} e^{-\lambda} \\ &= 0.8426 \end{aligned}$$

Example: Alpha Particle Emissions

- In a 1910 study of the emission of alpha-particles from a Polonium source, Rutherford and Geiger counted the number of particles striking a screen in each of $n = 2608$ time intervals of length one eighth of a minute. Rutherford and Geiger's observations are recorded in the following repeated-data frequency table form giving the number of time intervals (out of the n) in which 0; 1; 2; 3 etc particles had been observed.

Experimental Data

# particles u_j	Observed frequency f_j	Observed proportion f_j / n
0	57	0.022
1	203	0.078
2	383	0.147
3	525	0.201
4	532	0.204
5	408	0.156
6	273	0.105
7	139	0.053
8	45	0.017
9	27	0.010
10	10	0.004
11	6	0.002

Example: Alpha Particle Emissions

- Could it be that the emission of alpha-particles occurs randomly in a way that obeys the conditions for a Poisson process? Let's try to find out. Let X be the number hitting the screen in a single time interval. We want to check whether X follow a Poisson distribution of parameter λ where λ is the underlying average number of particles striking per unit time.
- We don't know λ, but will use the observed average number from the data as an estimate

$$\overline{X} = \frac{\sum_{j=1}^{n} u_j f_j}{n} = 3.87$$

- Now let us consider $\lambda = 3.87$ and we compare the observed proportion f_j / n to the Poisson distribution $P(X = i) = e^{-\lambda} \lambda^i / i!$, we have

	0	1	2	3	4	5	6	7
Obs. prop.	0.022	0.078	0.147	0.201	0.204	0.156	0.105	0.053
Poisson dist.	0.021	0.081	0.156	0.201	0.195	0.151	0.097	0.054

Geometric Random Variable

- Consider yet again independent trials, each with success proba p. These trials are performed until a success occurs.
- Let X the number of trials required then $X \in \{1, 2, ...\}$ then it follows a geometric p.m.f.

$$P(X = n) = (1 - p)^{n-1} p.$$

- It is indeed a valid p.m.f. as

$$\sum_{k=1}^{\infty} (1 - p)^{k-1} p = p \sum_{l=0}^{\infty} (1 - p)^{l} = p \frac{1}{1 - (1 - p)} = 1.$$

Mean and Variance of a Geometric Random Variable

- We have
$$E(X) = \frac{1}{p} \text{ and } Var(X) = \frac{1-p}{p^2}.$$
- Proof of expression of $E(X)$: If we set $q = 1 - p$, we have
$$E(X) = \sum_{k=1}^{\infty} kq^{k-1}p = p \sum_{k=1}^{\infty} kq^{k-1}.$$
Now we know that
$$J(q) = \sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$$
so by taking the derivative with respect to q
$$J'(q) = \sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2} = \frac{1}{p^2}$$
so
$$E(X) = \frac{1}{p}.$$

Example: Collecting pictures

- **Question**: Assume that, every time your little brother buys a box of Wheaties, he receives a picture of one of the n players of the Canadian Hockey team. Let X_k be the number of additional boxes he has to buy, after he has obtained $k-1$ different pictures, in order to obtain the next new picture. Thus $X_1 = 1$, X_2 is the number of boxes bought after this to obtain a picture different from the 1st pictured obtained, and so forth. a) What is the distribution of X_k? (We assume that Wheaties does not favour any player, i.e. Proba(finding player i)=Proba(finding player j)=$1/n$). Let Y the total number of boxes you must buy to get the n different players. b) What is the expectation of Y?

Example: Collecting pictures

- **Answer a)**: After having obtained $k-1$ different pictures, a successful event corresponds to buying a box of Wheaties with a picture of one of the remaining $(n-(k-1)) = (n-k+1)$ player. The probability of this event is

$$p_k := (n-k+1) \times \frac{1}{n} = 1 - \frac{(k-1)}{n}.$$

  Hence X_k follows a geometric distribution with p_k.

- **Answer b)**: We have $Y = X_1 + X_2 + \cdots + X_n$ so

$$\begin{aligned} E(Y) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= \frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_n} \\ &= 1 + \frac{1}{1-\frac{1}{n}} + \cdots + \frac{1}{1-\frac{n-1}{n}} \end{aligned}$$

  For $n = 6$, we have $E(Y) = 14.7$ and for $n = 14$, we have $E(Y) = 45.5$.

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

$$P(X = i) = \frac{\left(\begin{array}{c} m \\ i \end{array}\right)\left(\begin{array}{c} N-m \\ n-i \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}$$

Applications of the Hypergeometric distribution

- The two color urn model gives a physical analogy (or model) for any situation in which we take a simple random sample of size n (i.e. without replacement) from a finite population of size N and count X, the number of individuals (or objects) in the sample who have a characteristic of interest.
- With a sample survey, white balls and black balls may correspond variously to people who do (white balls) or don't (black balls) have leukemia, people who do or don't smoke, people who do or don't favor the death penalty, or people who will or won't vote for a particular political party.
- Here N is the size of the population, m is the number of individuals in the population with the characteristic of interest, while X measures the number with that characteristic in a sample of size n.
- The reason for conducting surveys as above is to estimate m, or more often the proportion of white balls $p = m/N$, from an observed value of X.

Example: Company fleet

- Suppose a company fleet of 20 cars contains 7 cars that do not meet government exhaust emissions standards and are therefore releasing excessive pollution. Moreover, suppose that a traffic policeman randomly inspects 5 cars. How many cars is he likely to find that exceed pollution control standards?
- This is like sampling from an urn. The $N = 20$ balls in the urn correspond to the 20 cars, of which $m = 7$ are white (i.e. polluting). When $n = 5$, the distribution of X, the number in the sample exceeding pollution control standards has a Hypergeometric distribution with $N = 20$, $m = 7$ and $n = 5$.
- For example, the probability of no more than 2 polluting cars being selected is

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.0830 + 0.3228 + 0.3874 = 0.7932. \end{aligned}$$

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) $P(X \geq 3)$

Example: Quality inspection

- In industrial quality control, lots of size N are subjected to sampling inspection. The defective items in the lot play the role of "white" elements and their number m is typically unknown.
- A sample size n is taken, and the number X of defective items in it is determine.
- We know that X follow a hypergeometric distribution of parameter N, m and n.
- Having observed $X = x$, we can estimate m by finding the value of m which maximizes $P(X = x)$; this is called the maximum likelihood estimate of m.