STAT chapter 5: Discrete Random Variables and Probability Models

5.1 Random Variables and Probability Functions

defn A random variable is a function that assigns a real number to each point in a sample space S.

Discrete random variables take integer values or, more generally, values in a countable set.

Continuous random variables take values in some interval of real numbers like $(0,1)$ or $(0,\infty)$ or $(-\infty,\infty)$

defn Let X be a discrete random variable with range$(X) = A$.
The probability function (p.f.) of X is the function

$$f(x) = P(X = x) \quad \text{defined for all } x \in A$$

The set of pairs $\{(x; f(x)) : x \in A\}$ is called the probability distribution of X.

properties: $f(x) > 0$ for all $x \in A$ and $\sum f(x) = 1$

defn the cumulative distribution function (cdf) of X is the function denoted by $F(x) = P(X \le x)$, defined for all x

in general $F(x)$ can be obtained from $f(x)$ using

$$F(x) = P(X \le x) = \sum_{u \le x} f(u)$$

properties of cdf:
- $F(x)$ is a non-decreasing function of x
- $0 \le F(x) \le 1$
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

if X takes on integer values then for values x such that $x \in A$, $x-1 \in A$

$$f(x) = F(x) - F(x-1)$$

ex. Suppose that N balls labelled $1, 2 \ldots N$ are placed in a box, and n balls $(n \le N)$ are randomly selected without replacement. Define the random variable

$$X = \text{largest number selected.}$$

find the probability function for X

soln 1: if $X=x$ then we must select the number x plus $n-1$ numbers from the set $\{1,2\ldots x-1\}$, this gives

$$f(x) = P(X=x) = \frac{\binom{1}{1}\binom{x-1}{n-1}}{\binom{N}{n}} = \frac{\binom{x-1}{n-1}}{\binom{N}{n}} \quad \text{for } x = n, n+1 \cdots N$$

soln 2: first find $F(x) = P(X \le x)$. Noting that $X \le x$ iff $\forall$ n balls selected are from the set $\{1,2\cdots x\}$, we get

$$F(x) = \frac{\binom{x}{n}}{\binom{N}{n}} \quad \text{for } x = n, n+1, \cdots N$$

then
$$f(x) = F(x) - F(x-1) = \frac{\binom{x}{n} - \binom{x-1}{n}}{\binom{N}{n}} = \frac{\binom{x-1}{n-1}}{\binom{N}{n}}$$

5.2 Discrete Uniform Distribution

setup: Suppose X takes values $a, a+1, a+2 \cdots b$ with all values being equally likely. Then X has a discrete Uniform distribution, on the set $\{a, a+1 \ldots b\}$

illustrate:
- if X is the number obtained when a die is rolled, then X has a discrete uniform distribution with $a=1$, $b=6$.
- computer random number generator gives Uniform $[1, N]$ variables.

pf: there are $b-a+1$ values X can take so the probability at each of these values must be $\frac{1}{b-a+1}$ in order that $\sum f(x) = 1$, therefore:

$$f(x) = P(X=x) = \begin{cases} \frac{1}{b-a+1} & \text{for } x = a, a+1, \cdots b \\ 0 & \text{otherwise.} \end{cases}$$

exe Suppose a fair die is thrown once and let X be the number on the face. First find the cumulative distribution function, $F(x)$ of X

soln there is an example of a discrete uniform distribution on the set $\{1,2,3,4,5,6\}$ having $a=1$ $b=6$ and pf

$$f(x) = P(X=x) = \begin{cases} \frac{1}{6} & \text{for } x \in [1,6] \\ 0 & \text{otherwise.} \end{cases}$$

the cdf $F(x)$:

$$F(x) = P(X \le x) = \begin{cases} 0 & x < 1 \\ \frac{\lfloor x \rfloor}{6} & 1 \le x < 6 \\ 1 & x \ge 6 \end{cases}$$

5.3 Hypergeometric Distribution

Setup: • We have a collection of N objects which can be classified into two distinct

• types. Call one type success (S) and the other type failure (F). There are r successes and $N-r$ failures.

• Pick n objects at random without replacement.

• Let X be the number of successes obtained.

• Then X has a geometric distribution

illustration: • The number of aces X is a bridge hand has a hypergeometric distribution with $N=52$, $r=4$, and $n=13$.

• In a fleet of 200 trucks there are 12 which have defective brakes. In a safety check 10 trucks are picked at random for inspection. The number of trucks X with defective breaks choosen for inspective has a hypergeometric distribution. with $N=200$ $r=12$, $n=10$

pf. using counting techniques we note there are $\binom{N}{n}$ points in the sample space S if we don't consider order of selection. There are $\binom{r}{n}$ ways to choose the x success objects from the r available and $\binom{N-r}{n-r}$ ways to choose the remaining $(n-x)$ objects from the $(N-r)$ failures. Hence

$$f(x) = P(X=x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

range of x: $x \geq \max(0, n-N+r)$
$x \leq \min(r, n)$

ex. In Lotto 6/49 a player selects a set of six numbers (with no repeats) from the set $\{1, 2, \cdots, 49\}$. in the lottery draw six numbers are select at random. Find the pf for X, the number from your set which are drawn.

Soln: X has a Hypergeometric distribution with $N=49$ $r=6$ and $n=6$, so

$$f(x) = P(X=x) = \frac{\binom{6}{x}\binom{43}{6-x}}{\binom{49}{6}} \quad \text{for } x = 0 \cdots 6$$

5.4 Binomial Distribution

setup: • Suppose an experiment has two types of distinct outcomes. Call these types Success (S) and failure (F)

- let their prob. be p (for S) and $(1-p)$ (for F)
- Repeat the experiment n independent times
- let X = # of sucesses obtained.
- X has a Binomial distri $X \sim Bin(n, p)$

The n individual experiments in the process are called "Bernoulli trials", the process is called the Bernoulli process.

illustration: • Toss a fair die 10 times and let X be the number of sixes occur. $X \sim Bin(10, \frac{1}{6})$

- In a microcircuit manufacturing process. 90% of the chips produces work. Suppose we selected 25 chips independently and let X be the number that work then $X \sim Bin(25, 0.9)$

note: • the prob. p of Success is constant over the n trials.

- outcome (S or F) on any trial is indep. of the outcome on other trials.

pf • There are $\frac{n!}{x!(n-x)!} = \binom{n}{x}$ different arrangement of x S and $(n-x)$F over the n trials.

- the prob for each of these arrangements has p multiplied together x times and $(1-p)$ multiplied $(n-x)$ times.
- Since the trial are indep. each arrangement has prob. $p^x(1-p)^{n-x}$

$$f(x) = P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2 \ldots n$$

ex. Suppose that in a weekly lottery you have prob. 0.02 of winning a prize with a single ticket. If you buy 1 ticket per week for 52 weeks what is prob. (a) that you win no prize and (b) win 3 or more prizes.?

soln Let X be the number of weeks that you win, then $X \sim Bin(52, 0.02)$ we find.

a) $P(X=0) = f(0) = \binom{52}{0}(0.02)^0(0.98)^{52} = 0.350$

b) $P(X \geq 3) = 1 - P(X \leq 2) = 1 - f(0) - f(1) - f(2) = 0.0859$

notes: • Binomial requires indep repetitions with the same prob. of S
• Hypergeometric are made from fixed collection of objets. without replacements.

• for small proportion of a large collection of objects, Bin is use as an approx to Hypergeometric $(N >> n)$

ex. Suppose we have 15 cans of soup with no labels, but 6 are tomato and 9 are pea soup. We randomly pick 8 cans and open them. Find prob. 3 are tomato

soln X = number of tomato soups picked

$$f(3) = P(X=3) = \frac{\binom{6}{3}\binom{9}{5}}{\binom{15}{8}} = 0.396$$

5.5 Negative Binomial Distribution

setup: • Bernoulli trials
• repeated independently with prob. p
• continue doing the experiment until a specified number, k, of success have been obtained.
• let X be the number of failures obtained before the k^{th} success
• X has a negative binomial distribution $X \sim NB(k, p)$

illustration: • if a fair coin is tossed until we get our 5^{th} head, the number of tails we obtained has a Negative Bin $k=5$ $p=\frac{1}{2}$

pf: • in all there will be $x+k$ trials (x F and k S)
• last trial must be a success.
• In the first $x+k-1$ trials need x fails and $(k-1)$ successes
• each order has prob. $p^k(1-p)^x$

$$f(x) = P(X=x) = \binom{x+k-1}{x} p^k (1-p)^x \quad \text{for } x = 0, 1, 2 \cdots$$

note: Bin: we know the number n of trials in advance but we don't know the number of success we will obtain after the experiment
NB: we know the number k S in advance but do not know the number of trials that will be needed to obtain k S

ex. The fraction of a large population that has a specific blood type T is 0.08. for blood donation purposes it is neccessary to find 5 people with type T. If randomly selected individuals from the population are tested one after another. then (a) what is the prob. y persons have to be tested to get 5 type T persons, and (b) what is the prob. that over 80 people have to be tested

soln. • Think a type T person as S and non-type T a F

• let Y = # of person who have to be tested.

• Let X = # of non-type T persons in order to get 5 S's

• Then, $X \sim NB(k=5, p=0.08)$

$$P(X=x) = f(x) = \binom{x+4}{x}(0.08)^5(0.92)^x$$

• $Y = X+5$, $P(Y=y) = P(X=y-5)$

$$= f(y-5)$$

$$= \binom{y-1}{y-5} 0.08^5 (0.92)^{y-5} \text{ for } y = 5, 6, 7 \ldots$$

• $P(Y>80) = P(X>75) = 1 - P(X \le 75)$

$$= 1 - \sum^{75} f(x)$$

$$= 0.2235$$

5.6 Geometric Distribution

setup. • Consider NB with $k=1$

• repeat Bernoulli trials with two types of out come (S&F) with prob. p

• let X be the number of failures obtained before the first success.

$X \sim Geo(p)$

illustration: the prob. you win a lottery prize in any given week is a constant p, the number of weeks before you win a prize for the first time is Geo.

pf. $f(x) = P(X=x) = (1-p)^x p$ for $x = 0, 1, 2 \ldots$

Bernoulli Trials

1) independent
2) have 2 distinct types of outcome (S & F)
3) have the same probability p of "success" each time.

5.7 Poisson Distribution from Binomial

- the poisson distribution has prob. function of the form

$$f(x) = P(X=x) = e^{-\mu}\frac{\mu^x}{x!} \quad \text{for } x = 0, 1, 2 \ldots$$

where $\mu > 0$ is a parameter whose value depends on the setting for the model.

Setup: • r.v. X represents the number of events of some type.
- limiting case of Binomial distribution as $n \to \infty$ & $p \to 0$
- we keep product np fixed at constant value μ
- and let $n \to \infty$, then $p \to 0$
- use Poi dis. with $\mu = np$ as approx to Bin dis. for which n is large and p is small

ex. There are 200 people at a party. what is the proba that 2 of them were born Jan 1

soln assume all day of the year are equally likely (ignore Feb 29)
use Bin dis. $n = 200$ $p = 1/365$ for X = # people born Jan 1.

$$f(2) = \binom{200}{2}\left(\frac{1}{365}\right)^2\left(1-\frac{1}{365}\right)^{198} = 0.087$$

since n is large and p is close to 0, use Poi to approx:

$\mu = np = \frac{200}{365}$

$$f(2) = \frac{\left(\frac{200}{365}\right)^2 e^{-\left(\frac{200}{365}\right)}}{2!} = 0.087$$

5.8 Poisson Distribution from Poisson Process

set up: a situation in which a certain type of event occurs at random points in time (or space) according to the following conditions:

1. Independence: the number of occurrences in non-overlapping intervals are independent.
2. Individuality: for sufficiently short time periods of length Δt, the probability of 2 or more events occuring in the interval is close to zero, that is, events occur singly not in clusters.
   As $\Delta t \to 0$, the proba of two or more events in the interval of length Δt must go to zero faster that $\Delta t \to 0$

3. homogeneity or uniformity: events occur at a uniform or homogeneous rate λ over time so that the prob. of one occurrence in an interval $(t, t+\Delta t)$ is about $\lambda\Delta t$ for small Δt for any value of t, or:
$P(\text{one event in } (t, t+\Delta t)) = \lambda\Delta t + o(\Delta t)$

note: we use "order" notation: $g(\Delta t) = o(\Delta t)$ as $\Delta t \to 0$ to mean that the function g approaches 0 faster that $\Delta t \to 0$.

these three conditions together define a Poisson Process.

illustration: the emission of radioactive particles from a substance follows Poi process

pf: $f(x) = \dfrac{(\lambda t)^x e^{-\lambda t}}{x!}$ for $x = 0, 1, 2 \cdots$

In a Poi process with rate λ, the number of event occur X in a time interval of length t has Poi dis $\mu = \lambda t$

ex. Suppose earthquakes recorded in Ontario each year follows a Poi process with avg of 6 per year. What is prob that 7 will recorded in 2 year.

soln $t=2$ $\lambda = 6$, let $X = \#$ of earthquake in 2 year.

$\mu = t\lambda = 12$

$f(7) = \dfrac{12^7 e^{-12}}{7!} = 0.0437$

ex. At a nuclear power station, avg 8 leaks of heavy water are reported each year. Find the prob. 2 or more leaks in 1 month.

soln Let $X = \#$ of leaks in one month.

$\lambda = 8$, $t = 1/12$ $\mu = \lambda t = 8/12$

$\therefore P(X \geq 2) = 1 - P(X < 2)$

$= 1 - (f(0) + f(1))$

$= 1 - \left(\dfrac{(8/12)^0 e^{-8/12}}{0!} + \dfrac{(8/12)^1 e^{-8/12}}{1!} \right)$

$= 0.1443.$

note: Random Occurrence of Events in Space: The Poi process also applies when "events" occur randomly in space (2 or 3 dimensional)

ex. Caliform bacteria occur in river water with avg intensity 1 backeria per 10 cubic cm of water. Find (a) the probability there are no backeria in a 20cc sample of water which is tested, (b) the prob. there are 5 or more backeria in a 50 cc sample.

soln Let X = # of becteria in a sample of volume v cc

$\lambda = 0.1$ bacteria per cc,

$\mu = 0.1v$ and $f(x) = e^{-0.1v}\frac{(0.1v)^x}{x!}$

a) $v=20$, $\mu=2$ $\therefore P(X=0) = f(0) = e^{-2} = 0.135$

b) $v=50$, $\mu=5$ $P(X\geq 5) = 1-P(X\leq 4) = 0.440$

note: Distinguishing Poisson from Bin and Other Distributions:

1. Can we specify in advance the maximum value which X can take? if we can, it is <u>not</u> Poi
2. Does it make sense to ask how often the events did <u>not</u> occur? if it makes sense, it is <u>not</u> Poi

<u>5.9 Combining Other Models with Poisson Process</u>

ex. A very large (essentially infinite) number of ladybugs is released in a large orchard. They scatter randomly so that on avg a tree has 6 ladybugs.

a) find the prob. a tree has >3 ladybugs.

b) when 10 trees are picked at random, what is prob. 8 of these has >3 ladybugs.

c) trees are checked until 5 with >3 lady bugs are found. Let X be the total number of trees checked. find $f(x)$

d) find the prob. a tree with >3 ladybug has exactly 6

e) on 2 trees there are a total t lady bugs. Find prob. that x of these are on the first of these 2 trees

soln a) if the ladybugs are randomly scattered the most suitable model is Poi

$\lambda = 6$ $v=1$ $\therefore \mu = \lambda v = 6$

$P(X>3) = 1-P(X\leq 3) = 1-[f(0)+f(1)+f(2)+f(3)]$

$= 1-\left[\frac{6^0e^{-6}}{0!}+\frac{6^1e^{-6}}{1!}+\frac{6^2e^{-6}}{2!}+\frac{6^3e^{-6}}{3!}\right] = 0.8488$

b) use Bin where S means >3 lady bugs, we have $n=10$ $p=0.8488$

$$f(8)=\binom{10}{8}(0.8488)^8(1-0.8488)^2=0.2772$$

c) use NB, S: $k=5$ F: $(x-5)$.

$$f(x)=\binom{x-5+5-1}{x-5}(0.8488)^5(1-0.8488)^{x-5}$$

$$=\binom{x-1}{4}(0.8488)^5(0.1512)^{x-5} \quad x=5,6,\cdots$$

d) this is conditional prob. let $A=\{6 \text{ ladybugs}\}$ and $B=\{>3\}$.

$$P(A|B)=\frac{P(AB)}{P(B)}=\frac{P(6 \text{ ladybug})}{P(>3)}=\frac{6^6e^{-6}}{6!}\Big/0.8488=0.1892$$

e) conditional prob:

$$P(x \text{ on } 1^{st} | \text{total of } t)=\frac{P(x \text{ on } 1^{st} \text{ and total of } t)}{P(\text{total of } t)}$$

$$=\frac{P(x \text{ on } 1^{st} \text{ and } t-x \text{ on } 2^{nd})}{P(\text{total of } t)}$$

$$=\frac{P(x \text{ on } 1^{st})\,P(t-x \text{ on } 2^{nd})}{P(\text{total } t)}$$

use Poi :

$$=\frac{\left(\frac{6^xe^{-6}}{x!}\right)\left(\frac{6^{t-x}e^{-6}}{(t-x)!}\right)}{\frac{12^te^{-12}}{t!}}$$

$$=\binom{t}{x}\left(\tfrac{1}{2}\right)^x\left(1-\tfrac{1}{2}\right)^{t-x} \quad x=0,1,\cdots t$$

5.10 Summary of Single Variable Discrete Models.

Name	Pf	range
DU	$f(x)=\frac{1}{b-a+1}$	$x=a,a+1,\cdots,b$
HyperGeo	$f(x)=\binom{r}{n}\binom{N-r}{n-x}/\binom{N}{n}$	$x=\max(0,n-(N-r)),\cdots\min(n,r)$
Bin	$f(x)=\binom{n}{x}p^x(1-p)^{n-x}$	$x=0,1,2,\cdots,n$
NB	$f(x)=\binom{x+k-1}{x}p^k(1-p)^x$	$x=0,1,2,\cdots$
Geo	$f(x)=p(1-p)^x$	$x=0,1,2,\cdots$
Poi	$f(x)=\frac{e^{-\mu}\cdot\mu^x}{x!}$	$x=0,1,2,\cdots$