ITCS 121 Statistics

Lecture 6

Some Discrete Distributions

Discrete distributions

- A **discrete distribution** is another way of talking about the probability mass function (pmf) of a discrete random variable.
- In this class we will look at three important discrete random variables and study their distributions.
  - Binomial random variables.
  - Poisson random variables.
  - Hypergeometric random variables.

The binomial random variable

- Imagine an experiment that can either succeed or fail.
- The probability of success is p, and the probability of failure is $q = 1 - p$.
- You do this experiment n times.
- Each of these repeats is independent from the others, so the results of the experiments so far do not affect the next one.
- The number of successes you see during these n experiments is a **binomial random variable** .
- In this case the variable depends on n and p.
- If you specify n and p then you can work out the pmf for the variable.
- We say the binomial random variable is **parameterized** by n and p.

Example – coin tossing

- You toss a fair coin 10 times and count the number of times you get heads.

- This is a binomial random variable with the parameters $n = 10$ and $p = 0.5$.

- If the coin is not fair then p might be different.

- If you toss the coin a different number of times then n will be different.

The pmf of a binomial random variable

- Remember that if X is a discrete random variable then the pmf of X is the function telling us the probability that X will take different values.

- E.g. $P(X = 4)$.

- If X is a binomial random variable with parameters n and p then obviously $P(X = k) = 0$ whenever k is less than 0 or bigger than n.

- What about the other values of k?

Calculating binomial probabilities

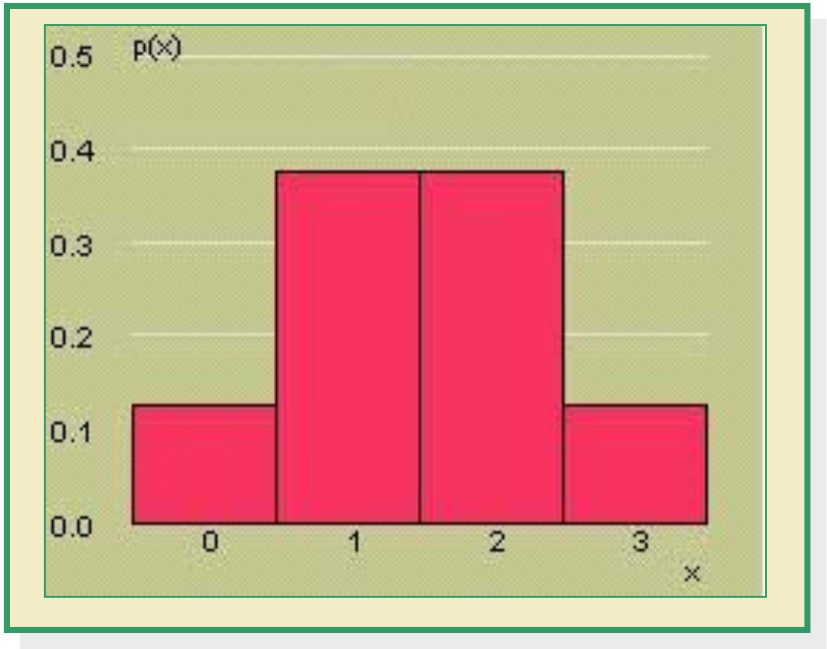
- Let X be a binomial random variable with parameters n and p.
- Let $0 \leq x \leq n$.
- What is $P(X = x)$?
- Well, we have $X = x$ if we get exactly x successes. How many ways are there to do this?
- This is choosing x out of n to be successes, and letting the rest be failures. I.e. $C(n, x)$.
- Each of the different ways we can get x successes has probability $p^x q^{n-x}$, because each success has probability p, each failure has probability q, and they are all independent.
- Since different sequences of coin tosses are mutually exclusive events, we get

$$P(X = x) = C(n, x) p^x q^{n-x} = \frac{n!}{(n-x)!\, x!} p^x q^{n-x}$$

Example – 3 coins

- Suppose we toss three fair coins and count the number of heads.
- This is a $binomial(3, 0.5)$ variable.
- This table shows the probabilities $p(x) = P(X = x)$
- We can represent this with a probability histogram.

x	$p(x)$
0	1/8
1	3/8
2	3/8
3	1/8



Notice that the sum of these probabilities is 1

Example - genes

- A geneticist samples 10 people and counts the number who have a gene linked to Alzheimer's disease.
- Suppose 15% of the population has this gene.
- We can treat this as a $binomial(10, 0.15)$ variable.
- $p = 0.15$ because if 15% of the population has the gene, a person picked at random should have a probability of 0.15 of having the gene.
  - Technically the trials are not independent, because if the first person picked has the gene it makes it very slightly less likely that the next person does, as the remaining population is smaller by one.
  - We ignore this though because we assume the population is large.

Genes - continued

- What is the probability that exactly 3 people out of the 10 have the gene?

- This is

$$\frac{10!}{(10-3)!\,3!}0.15^{3}0.85^{7} = 0.13\ to\ 2\ decimal\ places$$

Expectation and variance for binomial variables

- Let X be a binomial random variable with parameters n and p.
- We have formulas for the expectation and variance of X.

$$E(X) = np$$

$$Var(X) = npq$$

Deriving binomial expectation

- A **Bernoulli trial** is a random variable that takes the value 1 with probability p, and takes value 0 otherwise.
- The expected value of a Bernoulli trial B with parameter p is easy to find.

$$E(B) = (1 \times p) + (0 \times q) = p$$

- A binomial variable X with parameters n and p is like doing n Bernoulli trials with parameter p and adding up the results.
- So $X = B_1 + \cdots + B_n$.
- From linearity of expectation we get

$$E(X) = E(B_1) + \cdots + E(B_n) = np$$

Deriving binomial variance

- What is the variance of a Bernoulli trial B?
- This is $E(B^2) - E(B)^2$.
- We know $E(B) = p$, and we can work out $E(B^2) = p$ in exactly the same way.
- So $Var(B) = p - p^2 = p(1-p) = pq$.
- Again binomial $X = B_1 + \cdots + B_n$. Since each Bernoulli trial is independent, we have

$$Var(X) = Var(B_1) + \cdots + Var(B_n) = npq$$

Cumulative probability tables for binomial variables

- Let X be a binomial random variable with parameters n and p.

- To save us effort, people have made tables to help us find $P(X \leq x)$ for different values of x.

- You can find this on MyCourses.

Example – target shooting

- A target shooter hits a target 80% of the time. She fires five shots at the target. What is the probability that exactly 3 shots hit the target?
- Here $n = 5$ and $p = 0.8$.
- So $P(X = 3) = \frac{5!}{2!3!} 0.8^3 0.2^2 = 0.2048$.
- We also have $P(X = 3) = P(X \leq 3) - P(X \leq 2)$.
- So, we see from the table that

$$P(X = 3) = 0.2627 - 0.0579 = 0.2048$$

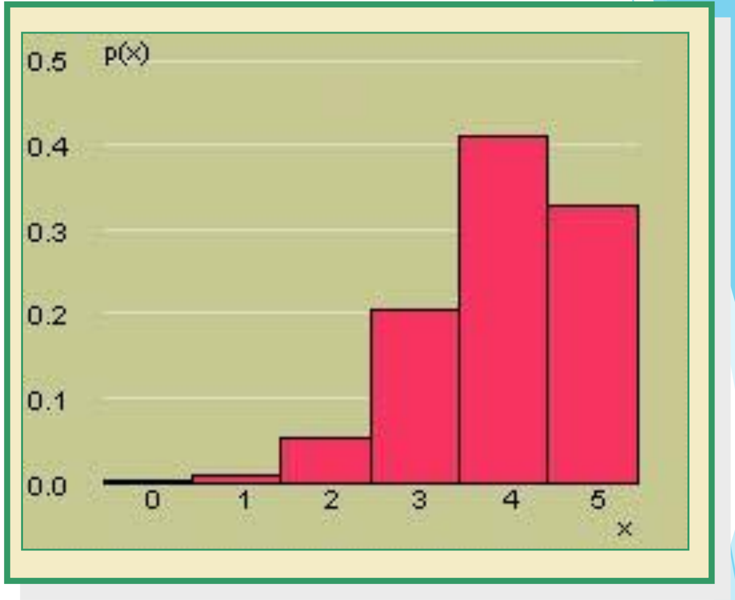
- This is the same answer, which it must be.
- The table also lets us easily work out e.g.

$$P(X > 3) = 1 - P(X \leq 3) = 1 - 0.2627 = 0.7373$$

More target shooting

- Here’s the probability distribution for the shooting example.

- We can also work out the expected value and the variance.



$$E(X) = 5 \times 0.8 = 4 \qquad Var(X) = 5 \times 0.8 \times 0.2 = 0.8$$

Class activity 1

A home security system is designed to have a 99% reliability rate. Suppose that 9 homes equipped with this system experience an attempted burglary.

1. Find the probabilities of these events:
    a) At least one of the alarms is triggered.
    b) More than seven of the alarms are triggered.
    c) Eight or fewer alarms are triggered.
2. How many alarms would you expect to be triggered?
3. What is the variance of the number of alarms triggered?

Poisson random variables

- The **Poisson random variable** is used to measure the number of times a random event occurs in a fixed time interval.
- For example:
  - How many line messages do you get in 1 hour?
  - How many car accidents in Thailand in one year?
  - Etc.

The pmf of a Poisson random variable

- Let X be a Poisson random variable.
- Then X is parametrized by a number λ (the Greek letter lambda).
- λ represents the expected number of times the event occurs in the given time interval.
- The pmf of X is then given by

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Example - traffic

- Suppose the average number of car accidents on a particular road is two per week.
- What is the probability there will be exactly one accident this week?
- This is $Poisson(2)$.
- So $P(X = 1) = \frac{2^1 e^{-2}}{1!} = 0.2707$.
- We can also work out the probability that there will be exactly 3 accidents in the next 4 weeks.
- This is $Poisson(8)$, because if there are on average 2 accidents per week, there are on average 8 accidents in 4 weeks.
- So $P(X = 3) = \frac{8^3 e^{-8}}{3!} = 0.0286$.

Deriving the Poisson pmf - part 1

- We can explain why the Poisson random variable has the pmf it does.
- Imagine that the time interval is broken down into n discrete parts.
- Since the random events have the same chance of occurring in each of these parts, and since the average number of events is λ, the probability of the event occurring at each part is $\frac{\lambda}{n}$.
- Assume that n is bigger than λ, so $0 \leq \frac{\lambda}{n} < 1$.
- This approximates the Poisson variable with a $Binomial(n, \frac{\lambda}{n})$ variable.

Deriving the Poisson pmf - part 2

- We approximated the Poisson variable with a $Binomial(n, \frac{\lambda}{n})$ variable.
- So the pmf of the Poisson variable is approximated by

$$P(X = k) = \frac{n!}{(n-k)!\,k!}\left(\frac{\lambda}{n}\right)^{k}\left(1 - \frac{\lambda}{n}\right)^{n-k}$$

- This approximation gets better as n gets bigger.
- To represent the continuous nature of time we have to take the limit as n tends to infinity.
- We can rewrite the above formula to get

$$\frac{\lambda^k}{k!}\underbrace{\frac{n!}{(n-k)!\,n^k}}_{lim\,=\,1}\underbrace{\left(1 - \frac{\lambda}{n}\right)^{n}}_{lim\,=\,e^{-\lambda}}\underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{lim\,=\,1}$$

- So the limit for the whole formula is $\frac{\lambda^k e^{-\lambda}}{k!}$ as claimed.

Tables for Poisson variables

- People have made cumulative probability tables for Poisson variables too.
- You can find one on MyCourse.
- E.g. If X is $Poisson(2)$, what is $P(X = 1)$?
- We have $P(X = 1) = P(X \leq 1) - P(X \leq 0)$
- The table says this is $0.406 - 0.1353 = 0.2707$.
- This agrees with the formula.
  - $P(X = 1) = \frac{2^1 e^{-2}}{1!} = 0.2707.$
- Similarly, we can work out e.g.
  - $P(X \geq 8) = 1 - P(X \leq 7) = 1 - 0.9989 = 0.0011$

Class activity 2

The number x of people entering the intensive care unit at a particular hospital on any one day has a Poisson probability distribution with mean equal to five people per day.

1. What is the probability that the number of people entering the intensive care unit on a particular day is two?
2. Less than or equal to two?
3. Is it likely that x will be more than 12? Explain.

The Poisson approximation

- We can use the Poisson distribution to approximate binomial distributions when n is large.
- This makes sense because the Poisson distribution is defined as the limit of binomial distributions as n tends to infinity.
- By the logic of the derivation of the Poisson pmf formula, we can approximate $Binomial(n, p)$ using $Poisson(np)$.
- There is no fixed rule that says when this approximation will be good.
- From experience many statisticians say it can safely be used in both the following situations:
  1. If $n \geq 20$ and $p \leq 0.05$.
  2. If $n \geq 100$ and $np \leq 10$.

Why is this useful?

- The binomial formula is simple, but $\frac{n!}{(n-k)!k!}$ can be very large and difficult to calculate for large values of n.
- On the other hand, $\frac{np^k e^{-np}}{k!}$ can be easier (the Poisson pmf with $\lambda = np$).
- In these situations the Poisson approximation can be useful.

Example – Poisson approximation

- Suppose a life insurance company insures the lives of 5000 men aged 42. If actuarial studies show the probability that any 42 year-old man will die in a given year to be 0.001, find the exact probability that the company will have to pay 4 claims during a given year.

- This is $Binomial(5000{,}0.001)$, and we want to find $P(X = 4)$.
- We can calculate this with $P(X = 4) = \frac{5000!}{4996!4!} 0.001^4 0.999^{4996}$.
- This is difficult to compute (try it), but because $n \geq 100$ and $p \leq 0.05$ we can use the Poisson approximation.
- This gives $P(X = 4) = \frac{(5000\times 0.001)^4 e^{-5000\times 0.001}}{4!} = \frac{5^4 e^{-5}}{4!} = 0.17547$.
- This is almost the same as the true value (0.17552).

Class activity 3

Let X be a binomial random variable with $n = 20$ and $p = 0.01$, use the Poisson approximation to calculate $P(X \leq 2)$.

Expected value and variance of a Poisson variable

- If X is a $Poisson(\lambda)$ random variable then we have

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

- Now, from the definition of the Poisson distribution, we should have $E(X) = \lambda$, because λ is supposed to be average number of times the events we're counting occur in the time interval.
- But it's not obvious how we get $E(X) = \lambda$ from the expected value formula $E(X) = \sum_{k \in \mathbb{N}} k \frac{\lambda^k e^{-\lambda}}{k!}$
- We'll see why this works as predicted on the next slide.

Expected of a Poisson variable

- The expected value of a $Poisson(\lambda)$ variable X is given by the formula

$$E(X) = \sum_{k\in\mathbb{N}} k\frac{\lambda^k e^{-\lambda}}{k!}$$

- We can show that $\sum_{k\in\mathbb{N}}\frac{\lambda^k e^{-\lambda}}{k!} = 1$, which must be true as the probabilities for discrete random variables must sum to 1.
- So

$$E(X) = \sum_{k\in\mathbb{N}} k\frac{\lambda^k e^{-\lambda}}{k!} = \lambda \sum_{k\in\mathbb{N}\backslash\{0\}} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} = \lambda \sum_{k\in\mathbb{N}} \frac{\lambda^k e^{-\lambda}}{k!} = \lambda$$

What we expect!

Variance of a Poisson variable

- We know $E(X) = \lambda$.
- We can show that $E(X^2) = \lambda^2 + \lambda$ using similar arguments.
- So we have $Var(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Hypergeometric random variables

- Imagine you have a bag containing N balls.
- M of these balls are red, and the other $N - M$ are blue.
- We take n balls from the bag randomly, and we count the number of red balls we get.
- This defines a discrete random variable X.
- X is **hypergeometric** with parameters N, n and M.

Hypergeometric vs binomial

- The hypergeometric random variable is a bit like the binomial random variable, but it’s not exactly the same.
- The binomial random variable counts successes in a sequence of independent success/fail trials.
- A hypergeometric variable also counts successes in a sequence of success/fail trials, but here they are *not* independent as the probabilities change depending on which balls have been taken already.

The pmf of the hypergeometric random variable

- What is $P(X = k)$?
- We can work this out with a counting argument.
- There are $C(N, n)$ ways we can choose n balls out of N.
- The cases we are interested in are where exactly k are red, and the other $n - k$ are blue.
- There are $C(M, k)C(N - M, n - k)$ ways this can happen.
- So

$$P(X = k) = \frac{C(M, k)C(N - M, n - k)}{C(N, n)}$$

Example - hypergeometric variable

- A package of 8 AA batteries contains 2 batteries that are defective (i.e. don't work). A student randomly selects four batteries and replaces the batteries in his calculator. What is the probability that all four batteries work?
- This is hypergeometric with parameters 8, 4 and 6.

No. good batteries

Total no. batteries

No. batteries taken

Success = working battery

N = 8

M = 6

n = 4

$$P(x=4)=\frac{C_4^6 C_0^2}{C_4^8}$$

$$=\frac{6(5)/2(1)}{8(7)(6)(5)/4(3)(2)(1)}=\frac{15}{70}$$

Expected value and variance of hypergeometric random variable

- We have

$$E(X) = n\left(\frac{M}{N}\right)$$

- And

$$Var(X) = n\left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Expected value for hypergeometric variable

- We can derive the expected value for a hypergeometric random variable X using a neat trick.
- Imagine arranging all the balls in the bag in a random order.
- Each ball in this sequence has probability $\frac{M}{N}$ of being red.
- We draw n red balls in the bag at random. We can assume this is the first n balls in the ordering.
- We can think of X as being the sum of random variables $X_1, \dots, X_n$, where X_i is 1 if the ith ball is red, and 0 otherwise.
- I.e. $X = X_1 + \cdots + X_n$.
- We have $E(X_i) = \frac{M}{N}$ for all i.
- And so $E(X) = E(X_1) + \cdots + E(X_n) = n\left(\frac{M}{N}\right)$

Class activity 4

A piece of electronic equipment contains six computer chips, two of which are defective. Three computer chips are randomly chosen for inspection, and the number of defective chips is recorded. What is the probability that exactly one computer chip inspected is defective?

Class activity 5

If X is the random variable from previous exercise, what are $E(X)$ and $Var(X)$?