L13.

We wish to model how many patients will be visiting the hospital from 10:am ~11:00 am during a day.

Suppose the average number of patients that would arrive is 2.

- possible values: $0, 1, 2 \cdots, \infty$

  (Binomial (n, p), $1, 2 \cdots, n$ (upper limit)).

- average number is known and fixed.

  (Bin (n, p), n known, p-pcss).

- # of patients within a specified time interval 10:00 ~11:00.

  (Bin (n, p), repeat n times).

Relations:

time interval = 1 hr.

cut the 1 hr interval into 6-sub intervals of 10 mins.

- each interval: the number of patients. 1-0. prob $\frac{2}{6}$.
- independent.

the total # of patients Binomial $(6, \frac{2}{6})$. $p = \frac{2}{6}$.

cut into 60-sub intervals Bin $(60, \frac{2}{60})$

⋮

n-sub intervals. Bin $(n, \frac{2}{n})$

$n \to \infty$ → poisson distribution

L14 10/24

More example of Poisson Distribution.

- The number of accident claims per month by an auto insurance company.

- The number of flaws on a loom cable
  length interval

- The # of typing errors on a page
  "area interval"

- The # of raindrops per square inch.
  "area interval"

Def. Poisson Experiment.

Suppose we are given an interval (this could be time / length / area / volum) and we are interested in the number of "successes" in that interval.
(a certain event occurs.)

Assume that the interval can be divided into very small subintervals.

s.t. ① the probability of more than one success in any subinterval is 0

② the probability of success in a subinterval is constant for all subintervals and is proportional to the length of the interval.

③ the subintervals are independent

This is called a Poisson experiment.

Def. Possion Distribution

We assume the following:

- The random variable X denotes the # of successes in the whole interval.
- λ is the mean number of successes in the whole interval.

Then X has a Possion Distribution with parameter λ, denoted $X \sim \text{Possion}(\lambda)$.

with pmf

$$P(X=k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}, \quad k = 0, 1, 2, \cdots$$

E.g. 1. The number X of customers arriving at a service desk follows a Possion Distribution with rate 2 per minute. Find the probability that in a given minute the # of customers will exceed 4.

- $X \sim \text{Possion}(2)$.

$$P(X>4) = P(X=5) + P(X=6) + \cdots$$

$$= 1 - P(X=0) - P(X=1) - P(X=2) - P(X=3) - P(X=4)$$

$$= 1 - \frac{e^{-2} \cdot 2^0}{0!} - \frac{e^{-2} \cdot 2^1}{1} - \frac{e^{-2} \cdot 2^2}{2 \times 1} - \frac{e^{-2} \cdot 2^3}{3 \times 2 \times 1} - \frac{e^{-2} \cdot 2^4}{4 \times 3 \times 2 \times 1}$$

$$\approx 0.942 = 9.42\%.$$

E.g. 2. The number of flaws in a fibre optic cable follows a Possion distribution. The average number of flaws in 10m of cables is 1.2.

(a) what is the probability of exactly 3 flaws in 150cm of cable.

- X: # of flaws in 150m of cable.

$X \sim \text{Possion}(3.6)$.

$$P(X=3) = \frac{e^{-3.6} \cdot 3.6^3}{3 \times 2 \times 1} \approx 0.212$$

(b) what is the probability of at least 2 flaws in 100m of cable.

- X : # of flaws in 100m of cable

$X \sim$ Possion (2.4)

$$P(X \geq 2) = 1 - P(X=0) - P(X=1)$$

$$= 1 - \frac{e^{-2.4} \cdot 2.4^0}{0!} - \frac{e^{-2.4} \cdot 2.4^1}{1}$$

$$\approx 0.691$$

(c) what is the probability of exactly one flaw in the first 50m of cable and exactly one flaw in the second 50m of cable.

- X : # of flaws in 50m of cable.

$X \sim$ Possion (1.2).

$$P(C) = P(X=1) \cdot P(X=1)$$

$$= \left(\frac{e^{-1.2} \cdot 1.2^1}{1}\right)^2$$

$$\approx 0.361$$

Remark (a) 150m Possion (3.6) [Possion (1.2)]

$$P(X=3) = \frac{e^{-3.6} \cdot 2.6^3}{3!} \neq 3 \times \frac{e^{-1.2} \cdot 1.2^3}{3!}$$

(c) 1st 50m and 2nd 50m 100m

Possion (1.2) Possion (1.2) Possion (2.4)

$$P(X_1=1, X_2=1) \neq P(Y=2).$$

Mean and Variance

If $X \sim \text{Poisson}(\lambda)$, then

① $E(X) = \lambda$

② $\text{Var}(X) = \lambda$,

③ $SD(X) = \sqrt{\lambda}$.

Proof: ① $E(X) \overset{\text{def}}{=} \sum_{k=0}^{\infty} k \cdot P(X=k)$

$$= \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot k.$$

$$= e^{-\lambda} \cdot \lambda \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}$$

$$= \frac{\lambda^0}{0!} + \frac{\lambda^1}{1} + \frac{\lambda^2}{2!} \cdots\cdots$$

$$= e^{\lambda}$$

$$= e^{-\lambda} \cdot \lambda \cdot e^{\lambda}$$

$$= \lambda$$

E.g. It's believed that the number of bookings taken per day at an online travel agency follows a Poisson Distribution. Past records indicate the daily mean number of bookings is 20 and the s.d is 6.
Comment on the suitability of the Poisson Distribution for this example.

- Poisson distribution ?

$\lambda = \text{mean} = 6 \neq sd^2 = 36$

Sum of independent Poisson distribution r.v.s.

If $X_1, X_2, \cdots, X_j$ are independent Poisson r.vs with parameters $\lambda_1, \lambda_2, \cdots \lambda_j$.
then $X_1 + X_2 + \cdots + X_j$ also follows Poisson distribution with parameter

equals $\lambda_1 + \lambda_2 + \cdots + \lambda_j$.

Proof: textbook §3.5 P227.

Eg. Suppose X and Y are independent Poisson r.v.s each w/ mean 1.5.

Find ① $P(X+Y=2)$

② $E[(X+Y)^2]$

① $X+Y \sim$ Possion (3).

(i) $P(X+Y=2) = \frac{\lambda^{-3} \cdot 3^2}{2!} \approx 0.2240$.

(ii) $P(X=0, Y=2) + P(X=1, Y=1) + P(X=2, Y=0)$.

② (i) $E[(X+Y)^2] \overset{def}{=} \sum (x+y)^2 P(X=x, Y=y)$.

(ii) $Var(X+Y) = E[(X+Y)^2] - [E(X+Y)]^2$

$3 = E[(X+Y)^2] - 3^2$

$E[(X+Y)^2] = 12$

(iii) $E[(X+Y)]^2 = E(X^2 + Y^2 + 2XY)$.

$= EX^2 + 2E(X)E(Y) + EY^2$.

$= V(X) + [E(X)]^2 + 2 \times 1.5^2 + Var(Y) + [E(Y)]^2$

$= 12$

LIS 10/29

Poisson Approximation (to Binomial dist(n, p)) §2.4.

Suppose

$n \to \infty$

$p \to 0$, with np staying constant

Thus, writing $\lambda = np$, it can be shown that the binomial probabilities can be approximated by poisson probabilities.

$$P(X=k) = \binom{n}{k} p^k q^{n-k} \approx \frac{e^{-\lambda} \cdot \lambda^k}{k!}, \text{ where } \lambda = np. \quad \forall k = 1, 2, \cdots, n.$$

Rule of Thumb.

- $n \geq 20$ and $p \leq .05$ → acceptable approximation.
- $n \geq 100$, and $np \leq 10$ → excellent approximation.

E.g. Suppose we sample 100 items from a production line, on average, 2% defectives. Use Poisson approximation to estimate the probability of exactly 3 defectives.

X - # of defectives among 100 items.

$X \sim$ Binomial $(100, 0.2)$, $\lambda = np = 100 \times 0.2 = 2$.

$$P(X=3) = \binom{100}{3} \times .02^3 \times .98^{97} \approx 0.182259$$

$$\approx \frac{e^{-2} \cdot 2^3}{3!} = 0.180477.$$

Geometric Distribution

Motivating example.

Suppose we toss a fair dice until the first six appear. Let x denote

the number of tosses. Find pmf of x.

- possible value of x: $1, 2, \cdots$
- $p(x=k) = ?$

$p(x=1) = \frac{1}{6}$

$p(x=2) = \frac{5}{6} \times \frac{1}{6}$

$\vdots$

$p(x=k) = \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6}$ $k=1, 2, \cdots$ $\Rightarrow$ Geometric $\left(\frac{1}{6}\right)$.

Def: the experiment consists of a sequence of independent trials

- Each trial results in either S or F. ("event of interests occurs")
- $p(S) = p$, constant from trail to trail
- the trail are performed until first sucess is observed.

Suppose X denote the total number of trials required, then X has the following pmf.

$p(x=k) = q^{k-1} \cdot p$, $k=1, 2, 3 \cdots$

and x is called a geometric distributed random variable

$X \sim$ Geometric(p).

$\sum_{k=1}^{\infty} p(x=k) = 1$ $\quad (0<p<1)$.

proof: $\sum_{k=1}^{\infty} q^{k-1} \cdot p = p \sum_{k=1}^{\infty} q^{k-1}$

$= p(q^0 + q^1 + \cdots q^{\infty})$

$= p\left(\frac{1}{1-q}\right)$

$= 1$ $\square$

Mean and Variance.

If $X \sim$ Geometric (p), then

- $E(X) = \frac{1}{p}$ ⟶ on average, the # of trials required to achieve the first success.
- $Var(X) = \frac{1-p}{p^2}$

proof. $LHS = E(X) = \sum_{k=1}^{\infty} k \cdot p(X=k) = \sum_{k=1}^{\infty} k \cdot q^{k-1} \cdot p$

$= p \sum_{k=1}^{\infty} k q^{k-1}$

$= p \cdot \Sigma.$

(1) $\Sigma = 1 + 2q + 3q^2 + \cdots$

(2) $q\Sigma = 0 + q + 2q^2 + \cdots$

$(1-q)\Sigma = 1 + q + q^2 + \cdots q^{\infty} = \frac{1}{1-q}.$

$\Sigma = \frac{1}{(1-q)^2} = \frac{1}{p^2}.$

$E(X) = \frac{p}{p^2} = \frac{1}{p} = RHS$ □.

P213 Proof for Variance.

Now, suppose a fair dice is tossed until r sixes are observed.
Let X denote the # of tosses required until the r-th six.
Find the distribution of X.

→ say $r=2$.

Possible value of X $\quad \overset{r}{2}, \overset{r+1}{3}, \cdots$

$P(X=k) = \binom{k-1}{r-1} \left(\frac{1}{6}\right)^{r-1} \cdot \left(\frac{5}{6}\right)^{k-r} \cdot \frac{1}{6} = \binom{k-1}{r-1} \cdot \left(\frac{1}{6}\right)^{r} \cdot \left(\frac{5}{6}\right)^{k-r}.$

↳ k tosses, r sixes.

↓ a negative binomial dist $(r, \frac{1}{6})$

Def: the experiment consists of a sequence of independent trials

- each trial - S or F
- $P(S) = p$ - constant.
- trails are performed until r successes have been observed.

Let X be the total number of trials required. X has the following pmf.

$$p(x=k) = \binom{k-1}{r-1} p^r \cdot q^{k-r}, \quad k = r, r+1, r+2 \cdots$$

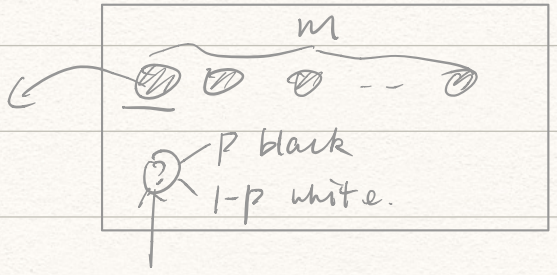
* If $r=1$, the negative bin distribution is geometric distribution.

$X \sim nb(r, p)$

Mean and Variance.

- $E(X) = \frac{r}{p}$
- $Var(X) = \frac{r(1-p)}{p^2}$

E.g. A box contains m black balls, At each step, a black ball is removed and a new ball, with probability p to be black and (1-p) to be white, is put into its place. Find the expected (number of steps) required until there are no black balls in the box



at least m steps : no black balls.

X : $m, m+1$ - - - -

"Success" : if a black ball is removed.

and a black ball is put into

$P(S) = 1-p$.

m successes are required.

$X \sim nb(m, 1-p)$

$E(X) = \frac{m}{1-p}$.

e.g. $m=4$ black

$P = 0.2$ black

0.8 white.

$E(X) = \frac{4}{0.8} = 5$.

E.g. the collector's problem.

Suppose each box of a particular brand of cereal contains one out of a set of n different coupons. Suppose that the coupon in each box is equally likely to be anyone of the set of n, independent of what coupon are in other boxes. Let X be the number of cereal boxes a collector must buy in order to obtain a complete set of a different coupons.

- Find $E(X)$ # of cereal boxes the collector must buy.
- Find the variance $Var(X)$.

X : possible value $n, n+1$ - - - -

Say $n=2$.

1 box is needed to obtain the 1st coupon.

X_2 boxes required to obtain the "2nd coupon" after the 1st box.

$P(\text{2nd}) = \frac{1}{2}$. $X_2 \sim$ Geometric $(\frac{1}{2})$.

$E(X) = 1 + \frac{1}{\frac{1}{2}} = 3$.

Lab 10/31

In general:

X_1 : # of boxes ~~to~~ buy to obtain 1st coupon. $X_1 = 1$.

X_2 : # of boxes to buy to obtain the 2nd coupon after the 1st one is obtained $p = \frac{n-1}{n}$ $X_2 \sim$ Geometric $(\frac{n-1}{n})$

⋮

X_n : # of boxes to buy until the n^{th} coupon is observed after the first $(n-1)$ coupons are obtained.

$P_n = \frac{1}{n}$

$X_n \sim$ Geometric $(\frac{1}{n})$.

$X_k \sim$ Geometric $\left(\frac{n-(k-1)}{n}\right)$, $k = 1, 2, \cdots n$. independent.

$X = X_1 + X_2 + \cdots X_n$.

$E(X) = E(X_1) + \cdots E(X_n)$

$= \frac{1}{1} + \frac{1}{\frac{n-1}{n}} + \frac{1}{\frac{n-2}{n}} + \cdots \frac{1}{\frac{1}{n}}$

$= 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots n$.

$= n(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots \frac{1}{2} + 1)$.

↳ Harmonic Series H_n.

For $n=2$

$E(X) = 2(\frac{1}{2} + 1) = 3$.

$Var(X) = Var(X_1) + \cdots Var(X_n) = \sum_{k=1}^{n} \frac{\frac{k-1}{n}}{(1-\frac{k-1}{n})^2}$

Remark $E(X) = nH_n$.

as $n \to \infty$, $H_n \approx \ln n + \gamma + \frac{1}{2n}$, where $\gamma = 0.57721$. (Euler's Constant)

Thus, $E(X) \approx n\ln n = n\gamma + \frac{1}{2}$ as $n \to \infty$. by using Euler's approximation.

Hypergeometric Distribution.

Def: · A finite population with N elements.

· Each element is a "success" or a "failure".

Total M successes.

· A sample of size n is drawn in such a way that each subset of size n is equally likely to be chosen. [w/o replacement]

Let X = the number of successes in the sample.

X is a hypergeometric distributed r.v.

The probabilities $P(X=x) = h(x; n, M, N)$, will depend on n, M, N.

* Characteristics.

1) 2 possible outcomes
2) Probability of a success is not the same.
3) trials are dependent.
4) Population is finite ($N < \infty$).

Def: pmf of X is given by $P(X=x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$ for $x = 0, 1, 2, \cdots n$.

also impose $x \le M$.

$$n - x \le N - M \quad \to \quad x \ge n - (N-M)$$

Thus, $\max(0, n - N + M) \le x \le \min(n, M)$.

E.g. Randomly draw 5 balls from a box containing 4 blue balls and 6 black balls. Find the pmf of x, where x = # of black balls in the sample

$N = 10$.

→ select 5. $= n$.

$M = 6$

$$P(x = X) = \frac{\binom{6}{x}\binom{4}{5-x}}{\binom{10}{5}}, \quad X = 1, 2, 3, 4, 5.$$

$\max(0, 5-10+6) \le x \le \min(5, 6)$.

$$p(x=2) = \frac{\binom{6}{2}\binom{4}{3}}{\binom{10}{5}} = \frac{60}{252}.$$

Mean, Variance

If $X \sim h(n, M, N)$, then.

$$E(X) = n \cdot \frac{M}{N}.$$

$$Var(X) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \underbrace{\left(\frac{N-n}{N-1}\right)}_{\text{finite population correction factor.}}$$

($\frac{N-n}{N-1} \to 1$ as $N \to \infty$)

V.S. $bin(n, p)$ $\quad E(X) = np \quad Var(X) = npq$.

* Sampling with replacement $\quad X \sim Bin(n, p) \quad p = \frac{M}{N}$.

w/o $\quad X \sim hyper(n, M, N)$

- the expectation of binomial dist & hyper distribution the same.
- the variance of bin dist & hyperg dist is diff by a factor.
- if N is large, n is small. $\frac{n}{N}$ is very small. $p = \frac{M}{N}$ is among from 0 to 1. then, the 2 distributions are approximately equal.

Eg. A lake contains 5000 fish, 60 of which have been tagged by scientists. A researcher randomly catches 10 fish from the lake, find a formula for the probability mass function of X, where X is the number of tagged fish in the researcher's sample.

$$P(X=x) = \frac{\binom{60}{x}\cdot\binom{5000-60}{10-x}}{\binom{5000}{10}}, \quad x = 0, 1, 2, \cdots, 10$$

Summary.

1) $E(X_1 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots E(X_n)$ without assumption.

$E(X_1 \cdots X_n) = E(X_1)E(X_2)\cdots E(X_n)$ } independent.

$Var(X_1 + \cdots X_n) = Var(X_1) + \cdots Var(X_n)$ } independent.

$Var(X_1 X_2 \cdots X_n) \neq Var(X_1)Var(X_2)\cdots Var(X_n)$.

2). If $X_1 \cdots X_n$ are independent random variables with same distribution as r.v. X

- $E(X_i) = E(X)$, $i = 1, \cdots n$.
- $Var(X_i) = Var(X)$, $i = 1, \cdots n$.

Then $E(X_1 + X_2 + \cdots X_n) = nE(X)$

$Var(X_1 + \cdots X_n) = nVar(X)$

$SD(X_1 + \cdots + X_n) = \sqrt{n}\, SD(X)$

$X_1, \cdots X_n$ independent identically distributed r.v.s.

"iid" random variables