Chapter 4: Discrete Distributions

- A <u>random variable</u> assigns numbers to outcomes in the sample space.
- Eg. # of children with retinitis pigmentosa
  # of individuals with leukemia
  etc...
- Basically, an event that is a number
- <u>Discrete</u> random variable: Can count them (but may be infinite)
  - typically 0, 1, 2, 3, ...
- <u>Continuous</u> random variable: Cannot count them
  - typically some interval $(-\infty, \infty)$, $[0,1]$, etc...
- Discrete example: $0, \frac{1}{2}, 1, \frac{3}{2}, \ldots$

  Continuous example: $[0,1] \cup (3,4]$
- Denote random variables with capital X, Y, Z, etc...
- A <u>probability mass function</u> (PMF) assigns a probability to a possible value r. Denote this probability by $Pr(X = r)$.
  - function of r, not X. X is used to denote the random variable.

Hypergeo

Ex.) Let X = # of patients in a trial of 4 wh. have improved blood pressure.

$P_r(X=0) = 0.008$
$P_r(X=1) = 0.076$
$P_r(X=2) = 0.265$
$P_r(X=3) = 0.411$
$P_r(X=4) = 0.240$

- $0 \leq P_r(X=r) \leq 1$ for all r
- $\sum_r P_r(X=r) = 1$
  ↑ all possible r
- Expected Value: Measure of center of a PMF (aka mean)

$$E[X] = \sum_r r\, P_r(X=r) = \mu$$

↑ all possible r's

hypergeo

- Ex) $E[X] = 0 \cdot 0.008 + 1 \cdot 0.076 + 2 \cdot 0.265 + 3 \cdot 0.411 + 4 \cdot 0.240$

$= 2.8$

↑ Average value of X across many trials

- Note: μ is a population parameter, not a statistic, which is a function of observed data

E.g. Across many trials, we might see

x	freq
0	0
1	9
2	24
3	48
4	19

↑ times seen

$$\bar{X} = 0 \cdot 0 + 1 \cdot \frac{9}{100} + 2 \cdot \frac{24}{100} + 3 \cdot \frac{48}{100} + 4 \cdot \frac{19}{100}$$

$$= 2.77$$

- Variance: Measure of spread

$$Var(X) = \sum_{r} (r - \mu)^2 \Pr(X = r) = \sigma^2$$

↑ all r

$$SD(X) = \sqrt{\sigma^2} = \sigma$$

- Note: $Var(X) = E(X^2) - E(X)^2$

$$= \sum_{r} r^2 \Pr(X = r) - \sum_{r} r \Pr(X = r)$$

↑ larger means more variable

- Cumulative Distribution Function (CDF)

$F(x) = Pr(X \leq x)$ = Probability X is less than or equal to x

- Ex.) Hypertension

$F(0) = 0.008$
$F(1) = 0.008 + 0.076$
$F(2) = 0.008 + 0.076 + 0.265$
$F(3) = 0.008 + 0.076 + 0.265 + 0.441$
$F(4) = 0.008 + 0.076 + 0.265 + 0.441 + 0.240$

- Useful for probability calculation

$Pr(1 \leq x \leq 3) = Pr(X \leq 3) - Pr(X \leq 0)$

- Exercise: X = # boys in a family of 4

r	$Pr(X=r)$
0	1/16
1	1/4
2	3/8
3	1/4
4	1/16

1) Calculate $E(X)$, $SD(X)$, $F(X)$

- Some distributions are seen in real data over and over again.
  - Binomial = # out of n
  - Poisson = # during some time interval

- these have specific PDF's / CDF's,
- we need to know about permutations / combinations to understand them.

- Number of permutations of n things taken k times

$$_nP_k = n(n-1)\cdots(n-k+1)$$

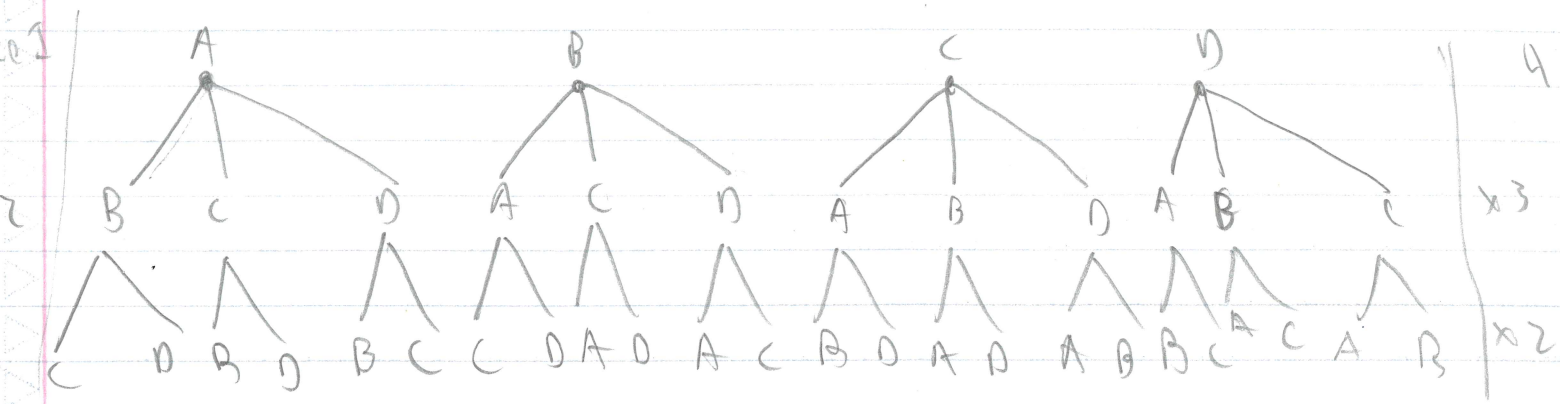
$$= \frac{n!}{(n-k)!}$$

Ex.) Individuals = A, B, C, D

$$_4P_2 = \frac{4!}{2!} = 4 \cdot 3 = 12$$

A, B	C, A
A, C	C, B
A, D	C, D
B, A	D, A
B, C	D, B
B, D	D, C

$_4P_3$ Example



- What if order does not matter? E.g. $\{A,D\} = \{D,A\}$
- The number of <u>combinations</u> of n things taken k at a time

$$_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k!)}$$

- There are $\frac{n!}{(n-k)!}$ permutations of size k

Each of those stores elements with $k! = {}_kP_k$

Eg. A, B, C, D $\quad {}_4C_3$

$A,B,C = A,C,B = B,A,C = B,C,A = C,A,B = C,B,A = 3!$

↳ Divide by $k!$ to get # combinations.

- **Binomial Distribution**

1.) n trials
2.) Outcome of each trial is "success" of "failure"
3.) $Pr(\text{"success"}) = p$ for each trial
4.) trials independent

X = # successes

then $X \sim Bin(n, p)$
↑ Distributed as

- Ex) White blood count

X = # neutrophiles out of 100 white blood cells

$Pr(\text{neutrophile}) = 0.6$

$\Rightarrow X \sim Bin(100, 0.6)$

- If $X \sim Bin(n, p)$ then

$$Pr(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$$

Ex.) Suppose $X \sim Bin(3, 0.3)$

$Pr(X=2) = Pr(2 \text{ succes and } 1 \text{ false})$

$Pr(SSF) = Pr(SFS) = Pr(FSS) = p^2(1-p)$

So $Pr(X=2) = 3p^2(1-p) = \binom{3}{2} p^2(1-p)$

- Claim: # of ways to order r successes and $n-r$ failures is $\binom{n}{r}$

proof:

Positions $1, 2, \ldots, n$

Choose r of these to be S, rest are F //

- Ex.) X = # boys out of 5 children, $p = 0.51$

$Pr(X=2) = \binom{5}{2} 0.51^2 \, 0.49^3 = 0.306$

$\binom{5}{2} = \frac{5 \cdot 4}{2 \cdot 1} = 10$

- CDF:

$$Pr(X \le x) = \sum_{r=0}^{x} Pr(X=r) = \sum_{r=0}^{x} \binom{n}{r} p^r (1-p)^{n-r}$$

↑ No simpler form

- Mean:

$$E(x) = \sum_{r=0}^{n} r \binom{n}{r} p^r (1-p)^{n-r} = np$$

↑ Expected # success = # trials · Pr(success)

- Variance

$$Var(x) = np(1-p)$$

↑ $n \uparrow \Rightarrow$ var ↑

↑ Variance highest @ $p = 0.5$, smallest at $p = 0$ or 1

- **Binomial functions in R**

dbinom() = $Pr(X=r)$
pbinom() = $Pr(X \le x)$
qbinom() = quantile
rbinom() = random generation.

Poisson:

↕ Counts of rare events over some period of time or space

Eg.) # typhoid cases in a year

Eg.) # bacterial colonies on an agar plate

- Assume
  1.) For small time interval Δt, $Pr(\text{"success"} \text{ in } \Delta t)$ is about $\lambda \Delta t$ (for some λ)
  2.) $Pr(\text{more than 1 "successes" in } \Delta t) \approx 0$
  3.) Stationarity: $Pr(\text{"success"})$ about the same for all time intervals
  4.) Independence: One success has no bearing on any other success

      ↳ violated, e.g., in epidemic

Then $X = \#$ "successes" in time t

$$X \sim Pois(\mu) \quad \text{s.t.} \quad \mu = \lambda t$$

$$Pr(X = k) = \frac{e^{-\mu} \mu^k}{k!}$$

Note

If $X \sim \text{Pois}(\mu)$ over time t

then $X \sim \text{Pois}(c\mu)$ over time ct

Ex) X = # typhoid deaths in 1 year

$X \sim \text{Pois}(4.6)$

Let Y = # typhoid deaths in half a year

$Y \sim \text{Pois}(4.6/2) = \text{Pois}(2.3)$

Ex.) X = # bacteria colonies in $100\ cm^2$

$X \sim \text{Pois}(2)$

Y = # bacteria colonies in $1000\ cm^2$

$Y \sim \text{Pois}(20)$

- Mean: $X \sim \text{Pois}(\mu)$, $E(X) = \mu$
- Variance: $\text{Var}(X) = \mu$
- If $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$, then $X+Y \sim \text{Pois}(\lambda_1 + \lambda_2)$

  [Not generally true for other distributions (Eg. Not for binomial).

  [Independent!

- Relation to binomial:
- If $X \sim Bin(n, p)$

  n large (> 100)

  p small (< 0.01)

  np intermediate

  then $X \approx Pois(np)$

  ↑ approximate

- This is used to justify Poisson in cases where we know n is large, but we don't know it exactly.

Eg.) X = # RNA molecules of a gene observed (on the order of 100)

- We don't know n, but know it is large
- we don't know p, but we know it is small (because X is ≈ 100)
- use poisson to model X!

R functions

- Exercises (4.24 - 4.29)

# of episodes for 1 child to have URI is media (ear disease) in 1 year is Pois(1.6)

4.24.) What is Prob of getting 3 or more episodes in the first 2 years of life?

Sol: X = # in 2 years ~ Pois(3.2)

1 - ppois(q = 2, lambda = 3.2)

= 0.6201

4.25.) What is Prob of not getting any in 1st year?

X = # in first year ~ Pois(1.6)

dpois(x = 0, lambda = 1.6)

= 0.2019

4.26.) Prob 2 siblings will both have 3 or more episodes in first year of life?

= Pr(sib1 has 2 or more) · Pr(sib2 has 2 or more) ← assumes independence

= 0.6201 · 0.6201 = 0.3845

out of 2

4.27.) Prob exactly 1 sibling will have 3 or more.

Y = # sibs

$\sim Bin(2, 0.6201)$

$P_i(Y=2) = dbinom(1, size=2, prob=0.6201)$

$= 2 \cdot 0.6201 \cdot (1-0.6201)$

$= 0.4712$

4.28.) Prob neither will have 3 or more episodes in first 2 years?

$(1-0.6201)^2 = 0.1443$

4.29.) Expected number of siblings in 2-sibling household who will have 3 or more episodes in first two years?

$E[Y] = 2 \cdot 0.6201 = 1.24$