BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 10

Why parametric models?

- *Parametric models* or *distribution families* have a specific form but can change according to a fixed number of parameters.
- The objective is to model a population. Parametric models are often appropriate in common situations with similar mechanisms.
- Parametric models have many known and useful properties and are easy to work with. When fitting a population, only a few parameters need to be estimated: *parametric inference*.
- Sometimes one does not want to make parametric assumptions and would rather work with non-parametric models. But non-parametric models can be infinite dimensional. E.g. $f_X(x),\ x = 0, 1, 2, \ldots$ or $F_X(x),\ x \in \mathbb{R}$.
- In this course we emphasize parametric models.

Discrete uniform

X has the *discrete uniform*$(1, N)$ distribution if X is equally likely to be one of $\{1, 2, \ldots, N\}$.

sample space: $\{1, 2, \ldots, N\}$

pmf:

$$f_X(x) = \frac{1}{N}, \qquad x = 1, 2, \ldots, N$$

cdf:

$$F_X(x) = P(X \leq x) = \frac{x}{N}, \qquad x = 1, 2, \ldots, N$$

moments:

$$\mathsf{E}X = \frac{N+1}{2}$$

This definition can be extended to the range $N_0, \ldots, N_1$ (consecutive intergers starting at any integer N_0 and ending with N_1) with $f_X(x) = 1/(N_1 - N_0 + 1)$.

Bernoulli distribution

Consider an experiment where outcomes are binary (say, Success or Failure) and the probability of success is p.
Define the following random variable

$$Y = \begin{cases} 1 & \text{outcome is success} \\ 0 & \text{outcome is failure} \end{cases}$$

Then, Y has a Bernoulli Distribution.

sample space: $\{0, 1\}$

pmf: $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$. We can write this as:

$$f(y) = P(Y = y) = \begin{cases} p^y(1-p)^{(1-y)} & y = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

What are the cdf, mean and variance?

Binomial Distribution

Now consider a *series of n Bernoulli trials* where

1. trials are independent
2. Prob of success or failure is the same for each trial, i.e.

$$P(S_i) = p \ \text{ and } \ P(F_i) = q = 1 - p \text{ for the } \ i^{th} \ \text{ trial.}$$

More concisely, consider n *iid* (independent, identically distributed) Bernoulli rvs Y_i.

A *binomial(n,p)* random variable X is defined as the number of successes in n iid Bernoulli trials, each with probability p of success:

$$X = \sum_{i=1}^{n} Y_i$$

Example: Coin Tossing

Toss 3 coins ($n = 3$) $P(H) = p, \quad P(T) = q$

$$\begin{array}{rcll} P(HHH) & = & p^3 & P(TTT) = q^3 \\ P(THH) & = & p^2q & P(HTT) = pq^2 \\ P(HTH) & = & p^2q & P(THT) = pq^2 \\ P(HHT) & = & p^2q & P(TTH) = pq^2 \end{array}$$

The binomial distribution is concerned with the distribution of the **number** of successes (heads):

$$\begin{array}{rcrcl} P(0H) & = & P(3T) & = & q^3 \\ P(1H) & = & P(2T) & = & 3pq^2 \\ P(2H) & = & P(1T) & = & 3p^2q \\ P(3H) & = & P(0T) & = & p^3 \end{array}$$

Binomial distribution

For any particular sequence of s successes and $n - s$ failures

$$P(SFS \dots F) = p^s q^{n-s}$$

However there are $\binom{n}{s}$ ways to get s successes from n trials.
Formally, the *binomial distribution* has:

sample space: $\{0, 1, ...n\}$

pmf:

$$f_Y(s) = \begin{cases} \binom{n}{s} p^s q^{n-s} & s = 0, 1, ...n \\ 0 & \text{otherwise} \end{cases}$$

Is it easy to check that the pmf sums to 1?

Binomial cont.

cdf:

$$F_Y(y) = \sum_{s=0}^{y} \binom{n}{s} p^s q^{n-s}$$

i.e.

$$\begin{aligned} F(y) &= 0 \quad \text{for} \quad y < 0 \\ F(0) &= q^n \\ F(1) &= q^n + npq^{n-1} \\ F(2) &= q^n + npq^{n-1} + \frac{n(n-1)}{2} p^2 q^{n-2} \\ &\vdots \\ F(y) &= 1 \quad \forall y \geq n \end{aligned}$$

Binomial example

A physician treats $n = 10$ people having a particular disease where P(success)$= 1/4$. What are probabilities of outcomes?

$$f(s) = \binom{10}{s} p^s q^{10-s}, \quad p = 1/4$$

s	$f(s)$	$F(s)$	s	$f(s)$	$F(s)$
0	.056		6	.016	
1	.188		7	.003	
2	.282		8	.0004	
3	.250		9	.00003	
4	.146		10	.00000009	
5	.059				

Poisson Distribution

The Poisson distribution was derived by the French mathematician Poisson in 1837 as a limiting version of the binomial distribution.

Suppose Y has a binomial(n, p) distribution, but consider what happens when n becomes large, but p is small enough so that np stays constant, and equal to a fixed value λ.

$$\lim_{n\to\infty} \binom{n}{y} p^y q^{n-y} = \frac{e^{-\lambda}\lambda^y}{y!}$$

Proof:

$$\begin{aligned}
f_Y(y) &= \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \\
&= \frac{n!}{y!(n-y)!} \left(\frac{\lambda}{n}\right)^y \left(1-\frac{\lambda}{n}\right)^{n-y} \\
&= \frac{\lambda^y}{y!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-y} \frac{(n-y+1)\ldots n}{n^y}
\end{aligned}$$

Binomial vs Poisson

Comparison of binomial and Poisson *pmf's*

$n = 5,\ \ p = 1/5\ \ (\lambda = 1)$

y	binomial	poisson
0	.328	.368
1	.410	.368
2	.205	.184
3	.051	.061
4	.006	.015
5	.000	.003
6+	0	.001

Binomial to Poisson via cfs

Another way to see the convergence of a binomial to a Poisson is via convergence of the cf.

The cf of $X \sim Binomial(n,p)$ is

$$\phi_X(t) = (pe^{it} + 1 - p)^n$$

Let $p = \lambda/n$:

$$\phi_X(t) = \left(\frac{\lambda}{n}e^{it} + 1 - \frac{\lambda}{n}\right)^n = \left(1 + \frac{\lambda(e^{it}-1)}{n}\right)^n$$

As $n \to \infty$,

$$\phi_X(t) \to \exp[\lambda(e^{it}-1)]$$

which is the cf of a Poisson. Since cfs characterize distributions, the distribution fo the rv X converges to a Poisson.

Poisson example

The death rate among females (Age 15-44) from pulmonary embolism is 4 per million. In a city of one million females (Age 15-44), what is the probability distribution of number of cases; i.e. $\lambda = 4$.

Note that $\lambda = np$ when $p = \frac{4}{1{,}000{,}000}$ is probability for a woman $\lambda = np = 10^6 \cdot 4/10^6 = 4$.

s	$e^{-4}4^s/s!$	s	$e^{-4}4^s/s!$	s	$e^{-4}4^s/s!$
0	.018	4	.20	8	.03
1	.07	5	.16	9	.01
2	.15	6	.10		
3	.20	7	.06		

Uses of the Poisson

The Poisson distribution is often used to describe:

- incidence of rare events in time or space
- failure of equipment
- incidence of rare diseases
- mortality
- pixel intensity in CT and PET imaging
- readings of molecular binding experiments (e.g. gene transcription)

Also arises in queueing theory (cashiers and internet), survival analysis, etc.

Hypergeometric Distribution

Example: Capture-Recapture Method

Suppose we wish to estimate how many fish there are in a lake.
Consider the following technique:
Capture a certain number of fish, 100 say, mark them, and return them to the lake.
Wait a sufficient amount of time for them to intermix again with the other fish.
Capture a new batch, 120 say, and see how many of these were previously marked.
Suppose there are 10 marked. Then the argument goes: the proportion marked that were recaptured should be in the same proportion as those amongst the non-captured. And so we can estimate the total number of fish in the lake to be the solution N to:

$$\frac{100-10}{N-120} = \frac{10}{120} \qquad \Rightarrow \qquad N =$$

This is a situation where we might apply a hypergeometric distribution.

Hypergeometric Distribution

Suppose a population of N entities is made up of two types, and there are M of the first type; and so $N - M$ of the second type.
Suppose we take a sample of size K, and we wish to know X, the number in the sample of the first type.
The probability mass function of X is given by:

$$f_X(x) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$

for $x = \max(0, M - N + K), \ldots, \min(M, K)$.
The sample space is defined so that all binomial coefficients are valid.
We must have:

$$0 \le x \le K, \qquad 0 \le x \le M, \qquad , 0 \le K - x \le N - M$$

Often $K < M$ and $K < N - M$ so the range becomes $0 \le x \le K$.

Hypergeometric Moments

Mean:

$$EX = \sum_{x=0}^{K} x \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} = \sum_{x=1}^{K} \text{ ditto}$$

We need the identity

$$x\binom{M}{x} = x\frac{M!}{x!(M-x)!} = M\frac{(M-1)!}{(x-1)!(M-x)!} = M\binom{M-1}{x-1}$$

or

$$\binom{M}{x} = \frac{M}{x}\binom{M-1}{x-1}$$

assuming everything is legit, i.e. all the numbers are positive integers, etc.. So

$$EX = \sum_{x=1}^{K} \frac{M}{\frac{N}{K}} \frac{\binom{M-1}{x-1}\binom{N-M}{K-1-(x-1)}}{\binom{N-1}{K-1}} = \frac{MK}{N}$$

Random sampling

Example: Vaccines are manufactured in batches of size N. Suppose n vials are sampled.
Decision Rule: If no vials are defective, batch is accepted.
What is the probability that batch with M defective vials is accepted?

Example: Experience suggests that a treatment for liver cancer should be considered effective if 20% of treated patients respond. A hospital plans to run a trial of a new treatment in 12 patients, and will consider the drug ineffective (no better than standard) if less than 2 patients respond. What is the probability that a drug with a true efficacy rate of 30% is classified as ineffective?

cont.

If we think of N being very large compared with n, then it makes sense to approximate this probability by thinking of sampling **with** replacement, in which case, we have

$$P(\text{Accept batch}) = (1 - M/N)^n$$

Table of $P(A)$ (approximate)

		M/N		
		.001	**.01**	**.1**
	5	.995	.951	.59
n	10	.990	.904	.349
	50	.95	.605	.005

Hypergeometric vs. Binomial

We can show that the limiting form of the hypergeometric pmf is the binomial pmf

$$
\begin{aligned}
P(s) &= \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{n}} \\
&= \frac{\frac{M!}{s!(M-s)!}\frac{(N-M)!}{(n-s)!(N-M-n+s)!}}{\frac{N!}{n!(N-n)!}} \\
&= \frac{\frac{n!}{s!(n-s)!}\frac{M!}{(M-s)!}\frac{(N-M)!}{(N-M-n+s)!}}{\frac{N!}{(N-n)!}}
\end{aligned}
$$

cont.

Note

$$
\begin{aligned}
\frac{M!}{(M-s)!} &= \frac{M(M-1)(M-2)\ldots(M-s)!}{(M-s)!} \\
&= M^s\left[1(1-\frac{1}{M})\ldots(1-\frac{s-1}{M})\right] \\
\frac{N!}{(N-n)!} &= N^n\left[1(1-\frac{1}{N})\ldots(1-\frac{n-1}{N})\right] \\
\frac{(N-M)!}{[(N-M)-(n-s)]!} &= (N-M)^{n-s} \\
&\quad \left[1\cdot(1-\frac{1}{N-M})\ldots(1-\frac{n-s-1}{N-M})\right]
\end{aligned}
$$

cont.

Letting $N \to \infty$, $M \to \infty$, $\frac{M}{N} \to p$,

$$\begin{aligned} P(s) &= \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{n}} \\ &\sim \binom{n}{s}\frac{M^s(N-M)^{n-s}}{N^n} \\ &= \binom{n}{s}\left(\frac{M}{N}\right)^s\left(1-\frac{M}{N}\right)^{n-s} \\ &\to \binom{n}{s}p^s(1-p)^{n-s} \qquad \text{Binomial Distribution} \end{aligned}$$

Summary

Hypergeometric $\to$	Binomial $\to$	Poisson
$N \to \infty$,	$n \to \infty$	$\lambda = np$
$M \to \infty$	$p \to 0$	
$\frac{M}{N} \to p$	$np \to \lambda$	