STA4030: Categorical Data Analysis
Preliminaries: Part I

Instructor: Bojun Lu

School of Science and Engineering
CUHK(SZ)

September 8, 2020

Agenda

1.1 Categorical Response Data

Definition 1 (Categorical Variable)

A variable has a measurement scale consisting of a set of categories is called a categorical variable.

Examples:

- x_1 = Grade received in a class
  Five categories: A, B, C, D, E
- x_2 = Social class
  Three categories: upper, middle, lower
- x_3 = Gender of a patient
  Two categories: male, female
- x_4 = Mode of transportation to work
  Five categories: automobile, bicycle, bus, subway, walk

1.1 Categorical Response Data

Definition 2 (Data Set)

A data set of categorical variables consists of frequency counts for the categories.

Example 3

Observations of X_1 *in a class with* $N = 50$ *students:*

Grade received	A	B	C	D	E
Frequency counts	15	25	7	2	1

1.1 Categorical Response Data

Categorical variables can be classified into some basic classes.

- Nominal variables: variables having categories without a natural ordering.

  For example,
  $\sharp$ x_1 = Gender of a patient
  $\sharp$ x_2 = Mode of transportation to work

  For a nominal variable, the order of listing the categories is irrelevant.

- Ordinal variables: variables having ordered categories.

  For example,
  $\sharp$ x_3 = Grade received in a class
  $\sharp$ x_4 = Social economic status

  Ordinal variables have ordered categories, but distances between categories are unknown.

1.1 Categorical Response Data

- Interval variables: variables having numerical distances between any two values.

  For example,

  ♯ blood pressure level

  ♯ annual income

- Continuous interval variables can be grouped into a number of categories.

  For example,

  ♯ blood pressure level x: $x < 80$ is normal; $80 < x < 89$ is prehypertension; $90 < x < 99$ is Stage 1 hypertension; $x > 100$ is Stage 2 hypertension.

  ♯ annual income x: $x < \$4000$, $\$4000 < x < \$10,000$, $\$10,000 < x < \$15,000$, etc.

1.1 Categorical Response Data

- The levels of categorical variables depend on the amount of information they include:

  nominal variables -> ordinal variables ->interval variables
  (lowest level) (highest level)

- Note: Tests designed for lower level variables can be applied to higher level variables, but tests for higher level variables should not be applied to lower level variables.

1.2 Some Important Distributions

1.2.1 Bernoulli Distribution

- This is the most basic one of all discrete random variables and it is also a building block of several other distributions.
- Let Y be a random variable with two possible values: $Y = 1$ with probability π and $Y = 0$ with probability $1 - \pi$.
- The probability mass function (pmf) or distribution of Y, $Bern(\pi)$, can therefore be written as,

$$p(Y = y) = \pi^y (1 - \pi)^{1-y}, \quad y = 0, 1,$$

with mean and variance,

$$\mu = E(Y) = \pi,$$

$$\sigma^2 = Var(Y) = \pi(1 - \pi).$$

1.2 Some Important Distributions

1.2.2 Binomial Distribution

- Let $Y_1, Y_2, \ldots, Y_n$ denote responses for n independent and identical trials such that

$$p(Y_i = 1) = \pi, \ \text{ and } \ p(Y_i = 0) = 1 - \pi.$$

- Then, $Y := \sum_{i=1}^{n} Y_i$ has the binomial distribution $B(n, \pi)$ with pmf,

$$p(Y = y) = p(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}, \ \ y = 0, 1, \ldots, n.$$

- Mean and variance can be calculated as,

$$\mu = E(Y) = n\pi,$$

$$\sigma^2 = Var(Y) = n\pi(1-\pi).$$

1.2 Some Important Distributions

- Note that, $B(1, \pi)$ is the Bernoulli distribution with probability π.
- If $Y_1, Y_2, \ldots, Y_n$ are independent, identically distributed (i.i.d.) $Bern(\pi)$ random variables, then $\sum_{i=1}^{n} Y_i$ has the binomial $B(n, \pi)$ distribution.
- For a fixed π, the distribution approaches the normal distribution,

  $$N(n\pi, n\pi(1-\pi)),$$

  as n grows large.

1.2 Some Important Distributions

1.2.3 Multinomial Distribution

- The multinomial distribution extends the binomial distribution:
  $\sharp$ a binomial random variable can take one of 2 possible outcomes on each trial;
  $\sharp$ a multinomial random variable can take one of c possible outcomes on each trial.
- Take n independent trials. Each trial has the same c possible outcomes, $E_1, E_2, \ldots, E_c$. On each trial, the probability of the outcome E_j occurs is π_j. The probabilities satisfy,

$$\sum_{j=1}^{c} \pi_j = 1.$$

- Then $\boldsymbol{N} = (N_1, \ldots, N_c)$ has the multinomial distribution with parameters n and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$, where N_j denotes the $\#$ of trials in which E_j occurs, $j = 1, 2, \ldots, c$.

1.2 Some Important Distributions

- For a multinomial random variable $\boldsymbol{N}$ with n trials and c possible outcomes with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$, we may write

$$\boldsymbol{N} \sim Mult(n, \boldsymbol{\pi}).$$

- The probability of $\boldsymbol{N}$ taking the value $(n_1, \ldots, n_c)$ is,

$$\begin{aligned} p(N_1 = n_1, \ldots, N_c = n_c) &= p(n_1, \ldots, n_c) \\ &= \frac{n!}{n_1!n_2!\cdots n_c!}\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_c^{n_c}, \end{aligned}$$

  for all possible $(n_1, \ldots, n_c)$ such that each $n_j \in \{0, 1, \ldots, n\}$ and $\sum_{j=1}^{c} n_j = n$.

1.2 Some Important Distributions

- Mean:

$$\mu_j = E(N_j) = n\pi_j, \quad j = 1, 2, \ldots, c.$$

- Variance:

$$Var(N_j) = n\pi_j(1 - \pi_j), \quad j = 1, 2, \ldots, c.$$

- Covariance:

$$Cov(N_j, N_h) = -n\pi_j\pi_h, \quad j, h = 1, 2, \ldots, c.$$

- Note: according to the expression, the N_js are negatively correlated. Meanwhile, intuitively, as their sum $\sum_{j=1}^{c} N_j$ is fixed, they should be negatively correlated as well.
- The probabilities π_j, $j = 1, 2, \ldots, c$ are constrained to lie inside the simplex (a region in the c-dimensional space) defined by,

$$0 \leq \pi_1, \ldots, \pi_c \leq 1, \quad \text{and} \quad \sum_{j=1}^{c} \pi_j = 1.$$

- As such, only $c - 1$ of them are “free": any of them must equal one minus the sum of the others. For example, we could replace π_1 by $1 - \pi_2 - \cdots - \pi_c$.

1.2 Some Important Distributions

Examples: $c = 5$,

y	1	2	3	4	5
p	π_1	π_2	π_3	π_4	π_5

3 — (0,0,1,0,0) , 2 — (0,1,0,0,0)

5 — (0,0,0,0,1) , 3 — (0,0,1,0,0)

1 — (1,0,0,0,0) , 4 — (0,0,0,1,0)

⋮

Repeat n multinomial trials ($\sum_{j=1}^{5} n_j = n, \ \sum_{j=1}^{5} \pi_j = 1$):

$$P(n_1, n_2, n_3, n_4) = \left(\frac{n!}{n_1!\, n_2!\, n_3!\, n_4!\, n_5!} \right) \pi_1^{n_1} \pi_2^{n_2} \pi_3^{n_3} \pi_4^{n_4} \pi_5^{n_5}.$$

1.2 Some Important Distributions

- It is easy to identify that $Mult(n, \boldsymbol{\pi})$ with $c = 2$ is equivalent to the binomial distribution.
- The marginal distribution of each N_i is binomial. That is,

$$N_j \sim B(n, \pi), \ \ j = 1, 2, \ldots, c.$$

- We can decompose $\boldsymbol{N}$ into n i.i.d. random variables, Y_i, $i = 1, 2, \ldots, n$, that is,

  $\sharp$ $\boldsymbol{N} = \sum_{i=1}^{n} Y_i$.
  $\sharp$ $Y_i \sim Mult(1, \boldsymbol{\pi})$, i=1,2,. . . ,n.
  $\sharp$ Y_i denotes that outcome of the ith trial. We can think of it as a vector of length c that takes a value 1 in entry j if outcome E_j occurs on the ith trial, and all other entries are zero.
  $\sharp$ The entries of Y_i are correlated Bernoulli random variables.

1.2 Some Important Distributions

1.2.4 Poisson Distribution

- The Poisson distribution is used for describing the counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent.
- If random variable Y follows the Poisson distribution with parameter μ (i.e. $Y \sim Po(\mu)$), then it has pmf,

$$p(Y = y) = p(y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

  with mean and variance,

$$E(Y) = \mu, \quad \text{and} \quad Var(Y) = \mu.$$

- Note: support for the Poisson distribution is infinite, unlike any of the other distributions we have seen so far.

1.2 Some Important Distributions

- The Poisson distribution approaches the normal distribution $N(\mu; \mu)$ as μ grows large.
- If $Y \sim B(n; \pi)$, $n \to \infty$ and $\pi \to 0$ with $n\pi \to \mu$, where μ is a constant, then the distribution of Y will tend towards $Po(\mu)$. (This is the so called "Law of Rare Events".)
- That is, the Poisson distribution is a limiting case of the Binomial distribution. Therefore, $Po(n\pi)$ can be used to approximate $B(n; \pi)$ when n is large and π is small.
- If $Y_i \sim Po(\mu_i)$, $i = 1, 2, \ldots, c$, are independent, then,

$$\sum_{i=1}^{c} Y_i \sim Po(\sum_{i=1}^{c} \mu_i).$$

1.2 Some Important Distributions

- Consider c independent Poisson variables, $Y_1, Y_2, \ldots, Y_c$, with parameters $\mu_1, \mu_2, \ldots, \mu_c$. Then the distribution of $\boldsymbol{Y} := (Y_1, Y_2, \ldots, Y_c)$ conditioned on the event $\sum_{i=1}^{c} Y_i = n$ is $Mult(n, \boldsymbol{\pi})$, where,

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c), \quad \text{and} \quad \pi_i = \frac{\mu_i}{\sum_{i=1}^{c} \mu_i}, \quad i = 1, \ldots, c.$$

- This means that it is possible to “split" the unconditional distribution of Y into two parts,
  - $\sharp$ a Poisson part for the overall total;
  - $\sharp$ a Multinomial part for the distribution of Y given n.
- Note: n and $\boldsymbol{\pi}$ are completely independent of each other. This is very important for drawing inference about $\boldsymbol{\pi}$, as we shall see later.

1.2 Some Important Distributions

1.2.5 Negative Binomial Distribution

Duality between Binomial and Negative Binomial:

- Binomial:
  - $\sharp$ n: Number of Bernoulli trials (fixed)
  - $\sharp$ Y: Number of successes among *n* Bernoulli trials (random)

$$p(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \quad y = 0, 1, \ldots, n.$$

- Negative Binomial:
  - $\sharp$ r: Number of successes (fixed)
  - $\sharp$ Y: Number of Bernoulli trials until r successes (random)

$$p(Y = y) = \binom{y-1}{r-1} \pi^r (1 - \pi)^{y-r}, \quad y = r, r + 1, \ldots.$$

1.2 Some Important Distributions

- Be careful: there are several different formulations of the Negative Binomial distribution.
- r = # of successes (fixed); Y = # of trials until r successes (random).
- r = # of failures (fixed); Y = # of successes until r failures (random),

$$p(Y=y)=\binom{y+r-1}{y}\pi^y(1-\pi)^r, \quad y=0,1,\ldots.$$

- r = # of successes (fixed); Y = # of failures until r successes (random),

$$p(Y=y)=\frac{\Gamma(y+r)}{\Gamma(r)\Gamma(y+1)}\left(\frac{r}{\mu+r}\right)^r\left(1-\frac{r}{\mu+r}\right)^y, \quad y=0,1,\ldots$$

with $\mu \geq 0$, $\pi=\frac{r}{\mu+r}$, and $\Gamma(\cdot)$ is the Gamma function.

1.2 Some Important Distributions

- In the last formulation, the distribution was parameterized using mean μ rather than probability of success π.
- Let us denote that distribution by $NB(r;\mu)$. Then if $Y \sim NB(r;\mu)$,

$$E(Y) = \mu, \text{ and } Var(Y) = \mu + \frac{\mu^2}{r}.$$

- Compare this with the Poisson distribution: both have infinite support; the means are the same; but the Negative Binomial distribution has a larger variance.
- This is the major motivation for knowing about the Negative Binomial distribution: when the observed variance is too large for the Poisson distribution (this is called overdispersion), then perhaps the Negative Binomial distribution can be used.