STAT 426

1.2 Distributions for Categorical Data

Distribution for categorical data

What are the random mechanisms generating categorical data?
We will make assumptions about the probability distributions where data observations arise. The most important distributions are:

- Bernoulli
- Binomial
- Multinomial
- Poisson

Bernoulli Distribution

Assume n independent binary (taking values 0 or 1) observations arising from independent and identical trials: $y_1, y_2, \ldots, y_n$ such that:

$$P(Y_i = 1) = \pi \quad \text{and} \quad P(Y_i = 0) = 1 - \pi$$

Random variables Y_i are normally called Bernoulli trials.

$$Y_i \sim \text{Bernoulli}(\pi)$$

$$p(y) = \begin{cases} \pi & y = 1 \\ 1 - \pi & y = 0 \end{cases}$$

$$E(Y_i) = \pi \qquad \text{var}(Y_i) = \pi(1 - \pi)$$

Binomial distribution

The random variable $Y = \sum_{i=1}^{n} Y_i$ has the Binomial distribution with index n and parameter π denoted as $Y \sim \text{bin}(n, \pi)$.

Mass probability function for Y:

$$P(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \quad y = 0, 1, 2, \ldots, n$$

with $\binom{n}{y} = n!/[y!(n-y)!]$

Binomial distribution

Mean and Variance:

$$E(Y) = \mu = n\pi \quad \text{var}(Y) = \sigma^2 = n\pi(1-\pi)$$

Skewness:

$$E(Y-\mu)^3/\sigma^3 = (1-2\pi)/\sqrt{n\pi(1-\pi)}$$

If the independence assumption is violated, the Binomial distribution does not apply.

$$\frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} \quad \xrightarrow[n\to\infty]{d} \quad N(0,1)$$

(Normal approximation)

Multinomial Distribution

Assume n independent trials have outcomes in $c > 2$ categories.

- Let $y_{ij} = 1$ if trial i has outcome in category j; otherwise $y_{ij} = 0$.

  For example, if $c = 5$, a possible outcome is $(0, 1, 0, 0, 0)$.

- Multinomial trial with binary vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ic})$.

- $\sum_j y_{ij} = 1$ whereas $\sum_i y_{ij} = n_j$ is the number of outcomes for category j. Note that y_{ic} is redundant because it is dependent on the remaining outcomes: $y_{ic} = 1 - \sum_{j=1}^{c-1} y_{ij}$.

Multinomial Distribution

- The vector of counts $(n_1, n_2, \ldots, n_c)$ has a multinomial distribution, with mass probability function:

$$p(n_1, n_2, \ldots, n_{c-1}) = \frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

  where $\pi_j = P(Y_{ij} = 1)$

- Marginal distribution of each n_j is a binomial distribution.

- Binomial distribution is a special case of the multinomial distribution when $c = 2$

- $E(n_j) = n\pi_j$, $\text{var}(n_j) = n\pi_j(1 - \pi_j)$, $\text{cov}(n_j, n_k) = -n\pi_j\pi_k$

Exercise: : Derive the expression for the covariance equation.

Poisson distribution

- Assume $Y = \#$ of events (counts) occurring randomly in a given period of time or space.

  For example (i) number of earthquakes of magnitude greater than 6, in the next 10 years; (ii) number of typographical errors in a the first 100 pages of a book; and so on.
- Assume independence in disjoint periods or regions.
- There is not a fixed number of trials.

Poisson probability mass function (pmf):

$$P(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, 3 \ldots$$

It satisfies $E[Y] = \text{var}[Y] = \mu$

Poisson distribution

- The Poisson pmf is unimodal with mode equal to the integer part of μ.
- Skewness: $E(y-\mu)^3/\sigma^3 = 1/\sqrt{\mu}$
- It is an approximation to the binomial distribution when n is large and π is small, such that $\mu = n\pi$.
- For some applications it is difficult to assume a mean equal to the variance. There might be a higher variability than the mean. This phenomenon is called overdispersion.

Overdispersion definition

In some cases, a count random variable can have a higher variance than the predicted by the binomial or the Poisson distribution.

For example, assume $Y =$ number of car break-ins in San Francisco at any given day:
Any parked car might have the same probability of suffering a break-in, but the expected number of break-ins $E[Y] = \mu$ might vary with neighbourhood, type of parking, car condition, and so on.

If $Y|\mu$ is a Poisson random variable for a given value of μ, and μ itself varies, such that $E[\mu] = \theta$, we can calculate the unconditional $E[Y]$ and $var[Y]$ as:

$$E[Y] = E[E[Y|\mu]] = E[\mu] = \theta$$

$$\begin{aligned} var[Y] &= E[var[Y|\mu]] + var[E[Y|\mu]] \\ &= E[\mu] + var[\mu] = \theta + var[\mu] > \theta \end{aligned}$$

Poisson and multinomial connection

Consider a sum of independent Poisson random variables Y_i with parameters μ_i.

- $\sum_i Y_i$ has a Poisson distribution with parameter $\mu = \sum_i \mu_i$.
- If $\sum_i Y_i = n$ and n is fixed, the random variables $Y_i|n$ are no longer independent nor have a Poisson distribution.
- For a c number of Poisson random variables, we can calculate the joint probability distribution of a set of counts $\{n_i\}$ conditioned on $\sum_i Y_i = n$ as:

$$P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c | \sum_i Y_i = n)$$

Poisson and multinomial connection

$$P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c | \sum_i Y_i = n)$$

$$\begin{aligned}
&= \frac{P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c)}{P(\sum_i Y_i = n)} \\
&= \frac{\prod_{i=1}^{c} \exp^{-\mu_i} \mu_i^{n_i} / n_i!}{\exp(-\sum \mu_i)(\sum \mu_i)^n / n!} \\
&= \frac{n!}{\prod_i n_i!} \prod_i \pi_i^{n_i}
\end{aligned}$$

with:

$$\pi_i = \mu_i / \sum_i \mu_i$$

This results in a **multinomial** $(n, \{\pi_i\})$ **distribution**.