Categorical Data Analysis Note

Edgard Maboudou

Chapter 1: Review

Some Basic Probability

- Suppose X is a random variable (r.v.) and i denotes a possible value of X. We are interested in the events of the form $X = i$, i.e. $P(X = i)$.
- Two conditions are needed for $P(X = i)$: $0 \leq P(X = i) \leq 1$ and $\sum_i P(X = i) = 1$.
- If Y is another r.v. and j is a possible value of Y. $P(X = i, Y = j) := P(\{X = i\} \cap \{Y = j\})$ is called the joint probability of X and Y.
- $P(X = i)$ and $P(Y = j)$ are marginal probabilities.
- Law of Total Probability: $P(X = i) = \sum_j P(X = i, Y = j)$.
- Conditional probability:

$$P(X = i|Y = j) := \frac{P(X = i, Y = j)}{P(Y = j)}.$$

- Independence: X and Y are independent if $\forall\ i, j$,
- Multiplicative Law of Probability follows from the definition of conditional probability:
- Bayes Theorem:

Ex. Disease present $D = 1$ and absent $D = 2$. Symptoms present $S = 1$ and absent $S = 2$ with probability:

$S \setminus D$	1	2
1	.005	.095
2	.005	.895

Find the following probabilities

1. $P(D = 2, S = 1) =$
2. $P(D = 1) =$
3. $P(D = 2) =$
4. $P(S = 1) =$
5. $P(S = 2) =$
6. $P(D = 2|S = 1) =$
7. $P(D = 1|S = 1) =$
8. Are D and S independent?

Example

If a person has a particular disease, a diagnostic test will give a positive result with probability .995. If the person doesn't have the disease, the test is positive with probability .05. Suppose 1% of the population has the disease. If a person is chosen at random and test is positive, what is the probability that he has the disease?

1 Some Important Discrete Distributions

1.1 Binomial Distribution

- Based on Bernoulli trials with two outcomes, Success (S) and Failure (F), e.g. coin toss.
- Y is called a Bernoulii random variable when Y has two possible values: $Y = 1$ (for Success) or 0 (for failure)
- The probability that $Y = 1$ is π
- The probability that $Y = 0$ is $1 - \pi$
- The mean of a Bernoulli random variable, Y, is
- The variance of Y is
- A binomial random variable is the sum of n independent Bernoulli random variables, so n Independent Bernoulli trials.
- $\pi = P(S)$ on any one trial.
- n identical trials, $P(S)$ cannot change
- Let $X =$ number of successes in n trials. X is binomial with parameters n and π.
- Its associated probability function is for $x = 0, 1, \ldots,$ n

$$p(x) = P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}.$$

- The mean of a Binomial random variable, X, is

- and the variance of X is

- So, for a binomial random variable, once you know π and n, you know the mean and variance of the Binomial distribution.

$$E(X) = \mu = n\pi, \quad var(X) = \sigma^2 = n\pi(1-\pi).$$

Example

It is known that only 10% of the people that contract a particular disease recover. If 8 people have the disease, what is the probability that at least 2 recover?

1.2 Multinomial Distribution

- Similar to Binomial and is based on multinomial trials.

- Each trial has k possible outcomes with $k > 2$.
- π_i is the probability that the ith outcome occurs on any one trial.
- So, we have n independent and identical trials. Let X_i = number of trials resulting in outcome i with $X_1 + X_2 + \cdots X_k = n$.
- $(X_1, X_2, \cdots, X_k)$ has a multinomial distribution with parameters $n, \pi_1, \ldots, \pi_k$.
- Its associated probability function is

$$P(X_1 = x_1, \ldots X_k = x_k) = \frac{n!}{x_1!x_2!\cdots x_k!}\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}.$$

- $E(X_i) = n\pi_i$.

Example

A balanced die is tossed six times. Find the probability that each number occurs once.

1.3 Poisson Distribution

- The Poisson distribution...
  - Is used for counts of events that occur randomly over time (or space).
  - Is often useful when the probability of event on any particular trial is very small while the number of trials is very large.
  - Arises naturally from counting random events during a fixed period of time.
- Suppose events occurs at random time.
- We are going to count the number of these events for a fixed time period of length t.
- Let X = number of events that occur in period of t time units.

- Ex. X = number of accidents that occurs at a particular intersection during this year.
- If certain conditions are satisfied, i.e.
  - Events must be independent.
  - Time period (or space) must be fixed.
- then X will have a Poisson distribution.
- This distribution depends on one parameter μ: average number of events to occur in t time units.
- Its probability function for $x = 0, 1, \ldots,$

$$p(x) = P(X = x) = \frac{\mu^x e^{-\mu}}{x!}.$$

- Assume conditions are satisfied, $E(X) = \mu$ and $var(X) = \sigma^2 = \mu$.
- In practice, the variance is usually larger than the mean. This is called overdispersion.

Example 1: 1994 World Cup Soccer

Event = frequency of various number of goals scored by a team during the 1st round of play (out of 35 matches).

Suppose $\mu = 1.143$, the Poisson probabilities are:

The expected frequencies are:

Example 2

The mean number of car entering a mountain tunnel per 2 min period is 1. An excessive number of car entering the tunnel during a brief period is hazardous.

1. Find $P(X > 3)$ in 2 min.

1.4 Hypergeometric Distribution

- Suppose we have a population made up of N items.
- There are r successes and $N - r$ failures.
- Select at random n of these items.
- Let $X =$ number of successes out of these n items.
- $\frac{r}{N}$ is probability of success. This is dependent from trial to trial
- X is hypergeometric random variable with parameters $N,\ r,\ n$.
- Its probability function is

$$p(x) = P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}.$$

Example

An office has 10 employees, 3 males and 7 females. The manager randomly chooses 4 people for promotion. Find the probability that more females are chosen.

2 Types of Data

Variables are classified according to the values they can take on.

a Dichotomous-binary: success or failure, 2 possibilities.

b Unordered groups: brand of detergent used.

c Ordered groups: groups differ in magnitude; e.g. Age: young, middle, old.

d Integer valued: e.g. number of cars owned in each household.

e Continuous on entire interval– e.g. weight.

Alternative Classification of Data Type

1. Discrete variable: can take on a finite or countably infinite number of values. [a – d]
2. Continuous variable: uncountably infinite, possible values make up an entire interval. [e]
3. Qualitative variable: levels of variable differ only in quality, not quantity. [a, b]
4. Quantitative variable: levels differ in magnitude. [c, d, e]

5. Ordinal variable: values of variable are ordered but the magnitude of differences between different levels is unknown. [c]
6. Nominal variable: qualitative. [a, b]

In this class, we are mostly interested in ordinal and nominal data. [a, b, c]

3 Univariate Analysis

3.1 Maximum Likelihood Estimation

- The Likelihood Function for Binomial is

  where x is known and π is unknown.
- $P(\pi|x)$ is now how likely the population proportion equals π given the data
- The value of π that is the most likely given the data is the “maximum likelihood estimate” of π.
- Denote MLE's by “ˆ”, for example $\hat{\pi}$.

3.2 Sampling Distribution of Proportions

Let π be the proportion of success in a population

- Random samples: $Y_1, Y_2, \ldots, Y_n$

$$Y_i = \begin{cases} 0 & \text{if } i^{th}\text{sample item is failure;} \\ 1 & \text{if } i^{th}\text{sample item is success.} \end{cases}$$

- Recall that Y is a Bernoulli random variable and so $X = \sum_{i=1}^{n} Y_i$ is Binomial
- π can be estimated by using the sample proportion p: If $X \sim Bin(n, \pi)$,
- then the observed proportion, p, equals $\hat{\pi}$

- Proportions are sample means of the Bernoulli random variables,

$$p = \bar{X} = \frac{\sum Y_i}{n} = \frac{X}{n}$$

- The mean of the sampling distribution of p:

- The variance of p is

- Shape of the sampling distribution of p?

- By the central limit theorem, if n is "large enough", then p is approximately distributed as a normal random variable with mean π and variance $\pi(1-\pi)/n$

- $p \sim \mathcal{AN}(\pi, \pi(1-\pi)/n)$

- In other words, $\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim Z$, standard normal distribution $\mathcal{AN}(0,1)$.

- How large must n be? For this class, When the "parent" distribution is Binomial, "large enough" usually means that $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

3.3 Inferences about a population proportion, π

3.3.1 Confidence Interval for π

- The usual method for $(1-\alpha)\times$ 100% CI for π is

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

3.4 Test of Hypothesis for π

- Want to test: $H_0: \ \pi = \pi_0$ v.s. $H_a: \ \pi \neq \pi_0$
- We can use Z and the standard normal distribution:
- Let p be the observed proportion of occurrences of the event.
- and π_0 be the null hypothesized probability.
- then $\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$ is the standard error of the sampling distribution of p (under H_0).

$$\boxed{\begin{array}{c} H_0: \ \pi = \pi_0 \text{ v.s. } H_a: \ \pi \neq \pi_0 \\ \text{T.S. } z = \dfrac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \end{array}}$$

Rejection Region (R.R.) depends on significant level α given (otherwise assume $\alpha = .05$).
Reject H_0 for the following cases:

1. $H_a: \pi > \pi_0$, if $z > z_\alpha$.
2. $H_a: \pi < \pi_0$, if $z < -z_\alpha$.
3. $H_a: \pi \neq \pi_0$, if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$.

Example

A politician claimed that 55% of the Americans support the president foreign policy. A newspaper editor randomly selected 1000 people finding out that 510 of them supported the policy. Does the editor has sufficient evidence to disbelieve the politician with $\alpha = .02$?

3.5 P-Value

What is the observed significant level?
P-value =

- P-value $\leq \alpha$, reject H_0.
- P-value $> \alpha$, fail to reject H_0.

Find 98% CI for π:

Example

The 2000 US presidential election came down to votes in Florida. The official results from the Florida Department of State, Division of Elections for the two top candidates as of Sunday November 28, 2000 are

This gave George W. Bush a 537 vote lead.

A 99% CI of π is

3.6 Small Sample Test for π

If $n\pi < 5$ or $n(1-\pi) < 5$, use the exact test because of the fact that if H_0 is true, then $X = \sum_i Y_i \sim Bin(n, \pi_0)$.

Example

A manufacturing process is supposed to produce items with no more than 5% of them defective. If in a sample of 20 items, 2 are defective, is the process out of control?

Note: Due to the discreteness of X, there are not many choices of α. So R.R. will not be of the form: Reject H_0 if $x \geq c$ for some integer c.

$c = 2$	$\alpha = .2642$
$c = 3$	$\alpha = P(X \geq 3) = .075$
$c = 4$	$\alpha = P(X \geq 4) = .016$

So .05 has a c^* between 3 and 4.

4 Inference About Multinomial Proportions

- Suppose a sample yields $(X_1, X_2, \ldots, X_k)$ which has a multinomial distribution with parameters $n, \pi_1, \ldots, \pi_k$, but $\pi_1, \ldots, \pi_k$ are unknown.
- Suppose we wish to test $H_0 : \pi_1 = \pi_{1_0}, \ldots, \pi_k = \pi_{k_0}$ with $\pi_{1_0} + \cdots + \pi_{k_0} = 1$.

- Large sample test compares observed and expected number for each of the k outcomes.

outcomes	1	2	$\cdots$	k
observed #	X_1	X_2	$\cdots$	X_k
expected # (under H_0)	$E(X_1) = n\pi_{1_0}$	$n\pi_{2_0}$	$\cdots$	$n\pi_{k_0}$

- Goodness-of-Fit test. T.S. $\chi^2 = \sum_{i=1}^{k} \frac{(X_i - n\pi_{i_0})^2}{n\pi_{i_0}}$.
- If H_0 is true, then approximately $\chi^2 \sim \chi^2_{k-1}$.
- For large sample reject H_0 if $\chi^2 > \chi^2_{\alpha,k-1}$.
- We have large sample if for all i, $n\pi_{i_0} \geq 5$.

Example

In a particular community, 4 brands of coffee A, B, C, D are on the market. A random sample of 800 drinkers gives:

Brands	A	B	C	D
X_i	190	198	187	225

use $\alpha = .05$, is there reason to say that brands are equally popular?