Lecture 01: Introduction

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Statistical Review

Let Y be a discrete random variable with $f(y) = P(Y = y) = p_y$.

Then, the expectation of Y is defined as

$$E(Y) = \sum_y y f(y)$$

Similarly, the Variance of Y is defined as

$$\begin{aligned} Var(Y) &= E[(Y - E(Y))^2] \\ &= E(Y^2) - [E(Y)]^2 \end{aligned}$$

Conditional probabilities

- Let A denote the event that a randomly selected individual from the “population” has heart disease.
- Then, $P(A)$ is the probability of heart disease in the “population”.
- Let B denote the event that a randomly selected individual from the population has a defining characteristics such as smoking
- Then, $P(B)$ is the probability of smoking in the population
- Denote

$$P(A|B) = \begin{array}{l}\text{probability that a randomly selected individual} \\ \text{has characteristic A, given that he has characteristic B}\end{array}$$

- Then by definition,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(AB)}{P(B)}$$

provided that $P(B) \neq 0$

- $P(A|B)$ could be interpreted as the probability of that a smoker has heart disease

Associations

- The two characteristics, A and B are associated if

$$P(A|B) \neq P(A)$$

- Or, in the context of our example–the rate of heart disease depends on smoking status
- If $P(A|B) = P(A)$ then A and B are said to be independent

Bayes' theorem

- Note that

$$P(A|B) = \frac{P(AB)}{P(B)}$$

  and

$$P(B|A) = \frac{P(BA)}{P(A)}$$

- So

$$P(A|B)P(B) = P(B|A)P(A)$$

- and

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- which is known as Bayes' theorem

Law of Total Probability

- Suppose event B is made up of k mutually exclusive and exhaustive events/strata, identified by $B_1, B_2, \ldots B_k$
- If event A occurs at all, it must occur along with one (and only one) of the k exhaustive categories of B.
- Since $B_1, B_2, \ldots B_k$ are mutually exclusive

$$\begin{aligned} P(A) &= P[(A \text{ and } B_1) \text{ or } (A \text{ and } B_2) \text{ or } \ldots (A \text{ and } B_k)] \\ &= P(AB_1) + P(AB_2) + \ldots + P(AB_k) \\ &= \sum_{i=1}^{k} P(A|B_i)P(B_i) \end{aligned}$$

- This is known as the Law of Total Probability
- A special case when $k = 2$ is

$$P(A) = P(A|B)P(B) + P(A|B')P(B')$$

where B' is read “not B” – also view this as a weighted average

Application to screening tests

- A frequent application of Bayes' theorem is in evaluating the performance of a diagnostic test used to screen for diseases
- Let D^+ be the event that a person does have the disease;
- D^- be the event that a person does NOT have the disease;
- T^+ be the event that a person has a POSITIVE test; and
- T^- be the event that a person has a NEGATIVE test
- There are 4 quantities of interest:
  1. Sensitivity
  2. Specificity
  3. Positive Predictive Value (PPV)
  4. Negative Predictive Value (NPV)

Sensitivity and Specificity

- Sensitivity is defined as the probability a test is positive given disease

$$\text{Sensitivity} = P(T^+|D^+)$$

- Specificity is defined as the probability of a test being negative given the absence of disease

$$\text{Specificity} = P(T^-|D^-)$$

- In practice, you want to know disease status given a test result

PPV and NPV

- PPV is defined as the proportion of people with a positive test result that actually have the disease, which is $P(D^+|T^+)$
- By Bayes' theorem,

$$\text{PPV} = P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+)}$$

- NPV is defined as the proportion of people among those with a negative test who truly do not have the disease ($P(D^-|T^-)$)
- Which by Bayes' theorem is

$$\begin{aligned} \text{NPV} &= P(D^-|T^-) \\ &= \frac{P(T^-|D^-)\cdot P(D^-)}{P(T^-)} \\ &= \frac{P(T^-|D^-)\cdot(1-P(D^+))}{1-P(T^+)} \end{aligned}$$

As a function of disease prevalence

- For both PPV and NPV, the disease prevalence ($P(D^+)$) influences the value of the screening test.
- Consider the following data

	Test result		
Disease status	Positive	Negative	Total
Present	950	50	1000
Absent	10	990	1000

- Sensitivity and Specificity for this test are

$$\text{Sen} = P(T^+|D^+) = 950/1000 = 0.95$$

and

$$\text{Spec} = P(T^-|D^-) = 990/1000 = 0.99$$

- However, the real question is what is the probability that an individual has the disease given a positive test result.

- With some easy algebra (substituting definitions into the previous equations), it can be shown that

$$\text{PPV} = \frac{Sens \cdot \Pi}{Sens \cdot \Pi + (1 - Spec)(1 - \Pi)}$$

- and

$$\text{NPV} = \frac{Spec \cdot (1 - \Pi)}{Spec \cdot (1 - \Pi) + (1 - Sens) \cdot \Pi}$$

where Π is the disease prevalence ($P(D^+)$)

- Thus, the PPV and NPV for rare to common disease could be calculated as follows:

Π	PPV	NPV
1/1,000,000	0.0001	1.0
1/500	0.16	0.99990
1/100	0.49	0.99949

Interpretation?

- For a rare disease that affects only 1 in a million,
  1. A negative test result almost guarantees the individual is free from disease (NOTE: this is a different conclusion of a 99% specificity)
  2. A positive test result still only indicates that you have a probability of 0.0001 of having the disease (still unlikely–which is why most screening tests indicate that “additional verification may be necessary”)
- However, if the disease is common (say 1 in 100 have it)
  1. A negative test result would correctly classify 9995 out of 10,000 as negative, but 5 of 10,000 would be wrongly classified (i.e., they are truly positive and could go untreated)
  2. However, of 100 people that do have a positive test, only 49 would actually have the disease (51 would be wrongly screened)
- Does the test “work”
- It “depends”

Application to Pregnancy Tests

- Most home pregnancy tests claims to be “over 99% **accurate**”
- By accurate, the manufactures mean that 99% of samples are “correctly” classified (i.e., pregnant mothers have a positive test, non-pregnant mothers have a negative test)
- This measure is flawed in that it is highly dependent on the number of cases (i.e., pregnant mothers) and controls (i.e., non-pregnant mothers) – FYI: we’ll revisit this concept again in future lectures
- However, for sake of illustration, lets consider a sample of 250 pregnant mothers and 250 non-pregnant mothers

Example Data–Based on at home pregnancy tests

Suppose we have the following data observed in a clinical trial:

	Truth		
	Pregnant	Not Pregnant	
Test +	N_{++}	b	
Test -	a	N_{--}	
	250	250	500

We know that we have 99% accuracy (because the manufactures tell us so), we have a constraint

$$\frac{N_{++} + N_{--}}{500} \geq 0.99$$

so

$$N_{++} + N_{--} \geq 495$$

and for illustrative purposes, let $a = 3$ and $b = 2$ so that the following table results.

	Truth		
	Pregnant	Not Pregnant	
Test +	247	2	249
Test -	3	248	251
	250	250	500

Then

$$Sens = P(T^+|D^+) = 247/250 = 0.988$$

and

$$Spec = P(T^-|D^-) = 248/250 = 0.992$$

Using these values and simplifying the previous equations for PPV and NPV,

$$PPV = \frac{0.988\Pi}{0.980\Pi + 0.008}$$

$$NPV = \frac{0.992 - 0.992\Pi}{0.992 - 0.98\Pi}$$

where Π is again the “disease rate” (or in this case, the probability of being pregnant)

Π	PPV	NPV
0.001	0.110022	0.999988
0.01	0.555056	0.999878
0.1	0.932075	0.998658
0.5	0.991968	0.988048

- Here, the "population" at risk is those females, of childbearing age, who engaged in sexual activity during the previous menstrual cycle, and are at least 2 days late in the new cycle.
- The success rate of birth control may be in the range of 99%.
- How do you feel about the marketing claim that the product is "over 99% accurate"?

Different Case-Control Ratio

	Truth		
	Pregnant	Not Pregnant	
Test +	397	2	399
Test -	3	98	101
	400	100	500

Then

$$Sens = P(T^+|D^+) = 397/400 = 0.9925$$

and

$$Spec = P(T^-|D^-) = 98/100 = 0.98$$

*Note: Sensitivity is now higher and specificity is lower than previously assumed

Π	PPV	NPV
0.001	0.047324	0.999992
0.01	0.333894	0.999923
0.1	0.846482	0.99915
0.5	0.980247	0.992405

What are categorical data

- What are categorical data?
- Agresti's answer: a variable with a measurement scale consisting of a set of categories
- In this class, we will examine categorical variables as an outcome (ie., dependent variable) and as a predictor (ie., covariate, independent variable)

Quantitative vs. Qualitative Variable Distinctions

Qualitative Variables: Distinct categories differ in quality, not in quantity

Quantitative Variables: Distinct levels have differing amounts of the characteristic of interest.

Clearly, a qualitative variable is synonymous with "nominal" (black, white, green, blue). Also, an interval variable is clearly quantitative (weight in pounds).

However, ordinal variables are a hybrid of both a quantitative and qualitative features. For example, "small, medium and large" can be viewed as a quantitative variable.

At this point, the utility in the variable descriptions may appear unnecessary. However, as the course progresses, the statistical methods presented will be appropriate for a specific classification of data.

Core Discrete Distributions for Categorical Data Analysis

There are three core discrete distributions for categorical data analysis

1. Binomial (with the related Bernoulli distribution)
2. Multinomial
3. Poisson

We will explore each of these in more detail.

Bernoulli Trials

Consider the following,

- n independent patients are enrolled in a single arm (only one treatment) oncology study.
- The outcome of interest is whether or not the experimental treatment can shrink the tumor.
- Then, the outcome for patient i is

$$Y_i = \begin{cases} 1 \text{ if new treatment shrinks tumor (success)} \\ 0 \text{ if new treatment does not shrinks tumor (failure)} \end{cases},$$

$i = 1, ..., n$

Each Y_i is assumed to be independently, identically distributed as a Bernoulli random variables with the probability of success as

$$P(Y_i = 1) = p$$

and the probability of failure is

$$P(Y_i = 0) = 1 - p$$

Then, the probability function is Bernoulli

$$P(Y_i = y) = p^y(1-p)^{1-y} \qquad \text{for } y = 0, 1$$

and is denoted by

$$Y_i \sim Bern(p)$$

Properties of Bernoulli

- MEAN

$$\begin{aligned} E(Y_i) &= 0 \cdot P(Y_i = 0) + 1 \cdot P(Y_i = 1) \\ &= 0(1-p) + 1p \\ &= p \end{aligned}$$

- VARIANCE

$$\begin{aligned} Var(Y_i) &= E(Y_i^2) - [E(Y_i)]^2 \\ &= E(Y_i) - [E(Y_i)]^2 \;\; ; \;\; \text{since } Y_i^2 = Y_i \\ &= E(Y_i)[1 - E(Y_i)] \\ &= p(1-p) \end{aligned}$$

Binomial Distribution

Let Y be defined as

$$Y = \sum_{i=1}^{n} Y_i,$$

where n is the number of bernoulli trials. We will use Y (the number of successes) to form test statistics and confidence intervals for p, the probability of success.

Example 2,
Suppose you take a sample of n independent biostatistics professors to determine how many of them are nerds (or geeks).

We want to estimate the probability of being a nerd given you are a biostatistics professor.

What is the distribution of the number of successes,

$$Y = \sum_{i=1}^{n} Y_i,$$

resulting from n identically distributed, independent trials with

$$Y_i = \begin{cases} \text{1 if professor } i \text{ is a nerd (success)} \\ \text{0 if professor } i \text{ is not a nerd (failure)} \end{cases} .$$

and

$$P(Y_i = 1) = p; \qquad P(Y_i = 0) = (1 - p)$$

for all $i = 1, ..., n$.

The probability function can be shown to be binomial:

$$P(Y = y) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y} = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y},$$

where

$$y = 0, 1, 2, ..., n$$

and
the number

$$\left(\begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{(n-y)!y!}$$

is the number of ways of partitioning n objects into two groups; one group of size y, and the other of size $(n-y)$.
The distribution is denoted by

$$Y \sim Bin(n, p)$$

Properties of the Binomial

- MEAN

$$
\begin{aligned}
E(Y) &= E\left(\sum_{i=1}^{n} Y_i\right) \\
&= \sum_{i=1}^{n} E(Y_i) \\
&= \sum_{i=1}^{n} p \\
&= np
\end{aligned}
$$

(Recall the expectation of a sum is the sum of the expectations)

- VARIANCE

$$
\begin{aligned}
Var(Y) &= Var\left(\sum_{i=1}^{n} Y_i\right) \\
&= \sum_{i=1}^{n} Var(Y_i) \\
&= \sum_{i=1}^{n} p(1-p) \\
&= np(1-p)
\end{aligned}
$$

(Variance of a sum is the sum of the variances if observations are independent)

Multinomial

Often, a categorical may have more than one outcome of interest. Recall the previous oncology trial where Y_i was defined as

$$Y_i = \begin{cases} \text{1 if new treatment shrinks tumor (success)} \\ \text{0 if new treatment does not shrinks tumor (failure)} \end{cases}$$

However, sometimes is may be more beneficial to describe the outcome in terms of

$$Y_i = \begin{cases} \text{1 Tumor progresses in size} \\ \text{2 Tumor remains as is} \\ \text{3 Tumor decreases in size} \end{cases}$$

Multinomial

y_{ij} is the realization of Y_{ij}. Let $y_{ij} = 1$ if subject i has outcome j and $y_{ij} = 0$ else. Then

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{ic})$$

represents a multinomial trial, with $\sum_j y_{ij} = 1$ and c representing the number of potential levels of Y.

For each trial, let $\pi_j = P(Y_{ij} = 1)$ denote the probability of outcome in category j and $n_j = \sum_i y_{ij}$ denote the number of trials having outcome in category j. The counts $(n_1, n_2, \cdots, n_c)$ have the multinomial distribution.

$$P(n_1, n_2, \cdots, n_{c-1}) = \left(\frac{n!}{n_1!n_2!\cdots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

This is $c-1$ dimensional because $n_c = n - (n_1 + n_2 + \ldots + n_{c-1})$ and $\pi_c = 1 - (\pi_1 + \pi_2 + \ldots + \pi_{c-1})$

Special Case of a Multinomial

When $c = 2$, then there is Binomial distribution

$$P(n_1) = \left(\frac{n!}{n_1!n_2!} \right) \pi_1^{n_1} \pi_2^{n_2}$$

Due to the constraints $\sum_c n_c = n$ and $\sum_c \pi = 1$, $n_2 = n - n_1$ and $\pi_2 = 1 - \pi_1$.

Therefore,

$$P(n_1) = \left(\frac{n!}{n_1!(n - n_1!)} \right) \pi_1^{n_1} (1 - \pi_1)^{n - n_1}$$

Note: For most of the class, I will use p for probability, Agresti tends to use π

Poisson

Sometimes, count data does not arrive from a fixed number of trials. For example,
Let $Y =$ number of babies born at MUSC in a given week.

Y does not have a predefined maximum and a key feature of the Poisson distribution is that the variance equals its mean.

The probability that $Y = 0, 1, 2, \cdots$ is written as

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}$$

where $\mu = E(Y) = Var(Y)$.

Proof of Expectation

$$
\begin{aligned}
E[Y] &= \sum_{i=0}^{\infty} \frac{ie^{-\mu}\mu^i}{i!} \\
&= \frac{0 \cdot e^{-\mu}}{0!} + \sum_{i=1}^{\infty} \frac{ie^{-\mu}\mu^i}{i!} && \text{See Note 1} \\
&= 0 + \mu e^{-\mu} \sum_{i=1}^{\infty} \frac{\mu^{i-1}}{(i-1)!} \\
&= \mu e^{-\mu} \sum_{j=0}^{\infty} \frac{\mu^j}{j!} && \text{See Note 2} \\
&= \mu && \text{See Note 3}
\end{aligned}
$$

Notes:

1. 0! = 1 and we separated the 1^{st} term (i=0) of the summation out
2. Let $j = i - 1$, then if $i = 1, \ldots, \infty$, $j = 0, \ldots, \infty$
3. Since $\sum_{j=0}^{\infty} \frac{\mu^j}{j!} = e^{\mu}$ by McLaurin expansion of e^x

Try finding the variance of the Poisson

Likelihood-based large-sample inference

There are three primary likelihood-based methods for statistical inference.

- Wald Test
- Likelihood Ratio Test
- Score Test

They are called the Holy Trinity of Tests. All three methods exploit the large-sample normality of ML estimators.

But first, lets review what a likelihood is.

Maximum Likelihood Estimation (MLE)

The purpose of MLE is to choose, as estimates, those values of the parameters, Θ, that maximize the likelihood function

$$L(\Theta|y_1, y_2, ..., y_n),$$

where

$$\begin{aligned} L(\Theta|y_1, y_2, ..., y_n) &= f(y_1)f(y_2)\cdots f(y_n) \\ &= \prod_{i=1}^{n} f(y_i) \end{aligned}$$

The maximum likelihood estimator of $L(\Theta)$ is the function $\widehat{\Theta}$ that produces

$$L(\widehat{\Theta}|y_1, y_2, ..., y_n) \geqslant L(\Theta|y_1, y_2, ..., y_n) \forall \Theta \in \Omega$$

That is, given an observed sample and a specified distribution, $\widehat{\Theta}$ is the value that maximizes

the likelihood (or produces the largest probability of occurrence).

MLE Continued

Recall from Calculus, the maximum value for a function occurs when the following conditions hold

1. The derivative of the function equals zero
2. The second derivative is negative
3. The value of the likelihood at the "ends" (boundaries of the parameter space) is less than $L(\widehat{\Theta})$

Since $\log(\cdot)$ is a monotonic function, the value that maximizes

$$l(\Theta|y_1, y_2, ..., y_n) = \log(L(\Theta|y_1, y_2, ..., y_n))$$

also maximizes $L(\Theta|y_1, y_2, ..., y_n)$.

Example MLE

Let $y_1, y_2, \ldots, y_n$ be an independent, identically distributed random sample with the $P(Y = 1) = p$ and the $P(Y = 0) = (1 - p)$. We want to find the MLE of p.

The **Likelihood** function of p, given the data is written as

$$
\begin{aligned}
L(p) &= \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i} \\
&= p^{\sum_{i=1}^{n} y_i} (1-p)^{n-\sum_{i=1}^{n} y_i}
\end{aligned}
$$

and the

$$
\begin{aligned}
l(p) &= \log(L(p)) \\
&= \log\left[p^{\sum y_i} (1-p)^{n-\sum y_i}\right] \\
&= \sum y_i \cdot \log(p) + (n - \sum y_i) \cdot \log(1-p)
\end{aligned}
$$

Then

$$
\frac{d\ l(p)}{dp} = \sum_{i=1}^{n} y_i \cdot \frac{1}{p} + (n - \sum_{i=1}^{n} y_i) \cdot \frac{-1}{1-p}
$$

Example MLE continued

Setting $\frac{d\ l(p)}{dp} = 0$ and solving for p yields

$$(1-p)\sum_{i=1}^{n} y_i = p(n - \sum_{i=1}^{n} y_i) \Longrightarrow \widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

General Likelihood Terminology

- Kernel: The part of the likelihood function involving the parameters.
- Information Matrix: The inverse of the $cov(\widehat{\beta})$ with the (j,k) element equaling

$$-E\left(\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_k}\right)$$

Note: Agresti uses $l(.)$ to indicate the regular likelihood function and $L(.)$ to represent the log-likelihood. I'll use the more traditional notation of $l(.)$ to represent the log.

Summary of general statistical inference

- We will make distinctions of the NULL and NON-NULL standard errors
- A non-null standard error is based on what you assume before you collect the data. I.e., in H_0, you may assumed $X \sim N(\mu, \sigma^2)$. Then, the non-null standard error would be based on σ^2
- However, when you take a random sample, you observe a mean and estimate the standard error of the mean
- This estimate could be (and is commonly) used in hypothesis testing
- Here we want to test the null hypothesis $H_0 : \beta = \beta_0$ vs some alternative hypothesis H_a.

Wald Test

With the **nonnull** standard error (SE) of $\widehat{\beta}$, the test statistic

$$z = (\widehat{\beta} - \beta_0)/\text{SE}$$

and its related transformation, z^2,

have a $N(0,1)$ distribution and χ^2 distribution with 1 degrees of freedom, respectively.

$$\begin{aligned} z^2 &= (\widehat{\beta} - \beta_0)^2/\text{SE}^2 \\ &= (\widehat{\beta} - \beta_0)'[Var(\widehat{\beta})]^{-1}(\widehat{\beta} - \beta_0) \end{aligned}$$

or in vector notation for more than one parameter

$$W = (\widehat{\vec{\beta}} - \vec{\beta_0})'[cov(\widehat{\beta})]^{-1}(\widehat{\vec{\beta}} - \vec{\beta_0})$$

Note: This is the typically hypothesis testing and is know as a WALD Test

Score Test

The score function, $u(\beta)$, is written as

$$u(\beta) = \frac{\partial l(\beta)}{\partial \beta}$$

Let $u(\beta_0)$ be the score value evaluated β_0 and $\imath(\beta_0) = -E[\partial^2 l(\beta)/\partial\beta^2]^2$ evaluated at β_0 (i.e., the information).

$u(\beta_0)$ tends to increase in value as $\widehat{\beta}$ is farther from β_0.

The statistic

$$\frac{[u(\beta_0)]^2}{\imath(\beta_0)} = \frac{[\partial l(\beta)/\partial\beta_0]^2}{-E[\partial^2 l(\beta)/\partial\beta_0^2]}$$

is distributed approximately χ^2 with 1 df and is known as a SCORE TEST.

Likelihood Ratio Test

Let L_0 be the likelihood value obtained by substituting in the null hypothesis value.
Let L_1 be the maximum likelihood value obtained from the data.

If L_1 is close to L_0, then you would expect that the null hypothesis would be true.

Note: $L_1 \geqslant L_0$ since $\widehat{\beta}$ and β_0 come from the same parameter space and $\widehat{\beta}$ was chosen as the maximum.

Let $\Lambda = \frac{L_0}{L_1}$. Then $-2\text{l}og\Lambda$ is distributed approximately as a χ^2 with the degrees of freedom equal the difference in the dimensions of the parameter spaces under $H_0 \bigcup H_a$ and under H_0.

The likelihood-ratio test statistics equals
$-2\text{l}og\Lambda = -2(l_0 - l_1)$

Comparison of the 3 methods

- All three methods are likelihood based
- The Wald test uses the NONNULL standard error
- The Score test uses the NULL standard error (information evaluated at the null)
- The likelihood ratio test combines information from the null and observed likelihoods
- For small to medium samples, the likelihood ratio test is better

In general, most of what we will discuss will be the likelihood ratio based tests.