STAT 4006: Categorical Data Analysis

Lecturer: Dr. John Wright

Department of Statistics

The Chinese University of Hong Kong

Chapter 1. Preliminaries

1.1.1 Definition

Categorical variable: A variable has a measurement scale consisting of a set of categories.

Examples:

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

1.1.2 Data set

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

e.g. 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

1.1.3 Classifying categorical variables

Nominal variables: variables having categories without a natural ordering.

e.g. x_3 – Gender of a patient
x_4 – Mode of transportation to work

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

Ordinal variables: variables having ordered categories.

e.g. x_1 – Grade received in a class
x_2 – Social economic status

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

Interval variables: variables having numerical distances between any two values

e.g. blood pressure level, annual income

Continuous interval variables can be **grouped** into a number of categories

e.g. 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

e.g. annual income x: $x < \$4000$, $\$4000 < x < \$10000$, 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)

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 of all discrete random variables and a building block of several others.
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

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

Mean:

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

Variance:

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

1.2.2 Binomial distribution

Let $Y_1, Y_2, \cdots, Y_n$ denote responses for n independent and identical trials such that $p(Y_i = 1) = \pi$ and $p(Y_i = 0) = 1 - \pi$. Then, $Y = \sum_{i=1}^{n} Y_i$ has the binomial distribution $B(n, \pi)$:

$$p(Y = y) = p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, 2, \cdots, n.$$

Mean:

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

Variance:

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

1.2 Some Important Distributions

Clearly $B(1, \pi)$ is the Bernoulli distribution with probability π.

If $Y_1, Y_2, \ldots, Y_n$ are independent, identically distributed (IID) $Bern(\pi)$ 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.

1.2.3 Multinomial distribution

The multinomial distribution extends the binomial distribution:

- a binomial random variable can take one of **2** possible outcomes on each trial
- 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 outcome E_j occurs is π_j. The probabilities satisfy $\sum_{j=1}^{c} \pi_j = 1$.

Consider the random variables
$N_j =$ "# trials in which E_j occurs", for $j = 1, \ldots, c$. Then $\boldsymbol{N} = (N_1, \ldots, N_c)$ has the multinomial distribution with parameters n and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$.

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 $\boldsymbol{N}$ takes the value $(n_1, \ldots, n_c)$ is

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

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:

$$\mu_j = E(N_j) = n\pi_j,$$

Variance:

$$Var(N_j) = n\pi_j(1 - \pi_j),$$

Covariance:

$$Cov(N_j, N_h) = -n\pi_j\pi_h.$$

(Note the N_j are negatively correlated - they must be, as their sum is fixed.)

The probabilities $\pi_j, j = 1, \ldots, c$ are constrained to lie inside the simplex (region of c-dimensional space) defined by $0 \leq \pi_1, \ldots, \pi_c \leq 1$ and $\sum_{j=1}^{c} \pi_j = 1$. 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_c by $1 - \pi_1 - \cdots - \pi_{c-1}$.

1.2 Some Important Distributions

Example: $c = 5$,

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)

$\vdots$

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

Clearly $Mult(n, \boldsymbol{\pi})$ with $c = 2$ is equivalent to the binomial distribution.

The marginal distribution of each N_j is binomial. That is, $N_j \sim B(n, \pi)$ for each $j = 1, \ldots, c$.

We can decompose into n IID random variables, Y_i for $i = 1, \ldots, n$, say:

- $\boldsymbol{N} = \sum_{i=1}^{n} Y_i$
- $Y_i \sim Mult(1, \boldsymbol{\pi})$, for $i = 1, \ldots, n$.
- Y_i is the 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 trial i, and all other entries are zero.
- The entries of Y_i are correlated Bernoulli random variables.

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 distribution

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

and moments Mean: $E(Y) = \mu$, Variance: $Var(Y) = \mu$.

Note the 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 $Y \sim B(n, \pi)$ and $n \to \infty$ and $\pi \to 0$ such that $n\pi \to \mu$, 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. Hence, $Po(n\pi)$ can be used to approximate $B(n, \pi)$ when n is large and π is small.

If $Y_j \sim Po(\mu_j), j = 1, \ldots, c$, are independent, then $\sum_{j=1}^{c} Y_j \sim Po(\sum_{j=1}^{c} \mu_j)$.

1.2 Some Important Distributions

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

- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$ and $\pi_j = \frac{\mu_j}{\sum_{j=1}^{c} \mu_j}$ for $j = 1, \ldots, c$.

This means we can "split" the unconditional distribution of $\boldsymbol{Y}$ into two parts

- a Poisson part for the overall total, n
- a multinomial part for the distribution of $\boldsymbol{Y}$ given n

and crucially, 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.

1.2.5 Negative Binomial distribution

Duality between Binomial and Negative Binomial:

- Binomial:

n — Number of Bernoulli trials (fixed)
Y — Number of successes among n Bernoulli trials (random)

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

- Negative Binomial:

r — Number of successes (fixed)
Y — Number of Bernoulli trials until r successes (random)

$$P(Y=y)=\binom{y-1}{r-1}\pi^{r}(1-\pi)^{y-r},\quad y=r,r+1,\cdots$$

1.2 Some Important Distributions

Be careful: there are several different formulations of the Negative Binomial distribution

- $r =$ # successes (fixed); $Y =$ # trials until r successes (random)
- $r =$ # failures (fixed); $Y =$ # successes until r failures (random)
  - $p(Y = y) = \binom{y+r-1}{y}\pi^y(1-\pi)^r$, for $y = 0, 1, \ldots$
- $r =$ # successes (fixed); $Y =$ # failures until r successes (random)
  - $p(Y = y) = \frac{\Gamma(y+r)}{\Gamma(r)\Gamma(y+1)}\left(\frac{r}{\mu+r}\right)^r\left(1-\frac{r}{\mu+r}\right)^y$ for $y = 0, 1, \ldots,$ $\mu \geq 0$ and $\pi = \frac{r}{\mu+r}$. $\Gamma(\cdot)$ is the Gamma function: $\Gamma(n) = (n-1)!$ for positive integer n.

1.2 Some Important Distributions

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

$$E(Y) = \mu; \; Var(Y) = \mu + \frac{\mu^2}{r}$$

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.

1.3.1 Likelihood functions

The distributions we have seen depend on parameters, many of which (e.g. π, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_2)$, μ, etc.) are unknown. Much of this course will involve making inferences about these unknown parameters. Our principal tool for doing so is likelihood.

Take a probability distribution (a PMF or PDF) $p(y)$. This depends on some unknown parameter(s), θ. So let's make that dependence explicit by writing $p(y) = p(y; \theta)$.

e.g. $Y \sim Po(\mu)$, hence $p(y) = p(y; \mu) = \frac{e^{-\mu}\mu^y}{y!}$

e.g. $Y \sim Bern(\pi)$, hence $p(y) = p(y; \pi) = \pi^y(1-\pi)^{1-y}$

1.3 Likelihood and Maximum-likelihood Estimation

If we plug an observed value into $p(y;\theta)$, we end up with a function of the unknown parameter(s) θ only

e.g. $Y \sim Po(\mu)$, we observe value of 3 hence

$$p(3) = p(3;\mu) = \frac{e^{-\mu}\mu^3}{6} := L(\mu)$$

e.g. $Y \sim Bern(\pi)$, we observe value of 0 hence

$$p(0) = p(0;\pi) = (1-\pi) := L(\pi)$$

Thus $L(\theta)$, called the likelihood function, is the result of plugging in an observed value into distribution function $p(y;\theta)$. To make the influence of the observed data y on the likelihood function explicit, we can use the notation $L(\theta) := L(\theta;y)$.

1.3 Likelihood and Maximum-likelihood Estimation

Of course, in practice we don't just observe one single value - we usually have an independent sample of size n, e.g. $y = (y_1, \ldots, y_n)$. In that case, the overall likelihood is a product of the individual likelihoods

$$L(\theta; y) = L(\theta; y_1) \times \cdots \times L(\theta; y_n) = \prod_{i=1}^{n} L(\theta; y_i) = \prod_{i=1}^{n} p(y_i; \theta).$$

e.g. $Y \sim Po(\mu)$, we observe $y = (y_1, \ldots, y_n)$ hence
$L(\mu; y) = e^{-n\mu} \frac{\mu^{\sum_{i=1}^{n} y_i}}{y_1! \cdots y_n!}$.

e.g. $Y \sim Bern(\pi)$, we observe $y = (y_1, \ldots, y_n)$ hence
$L(\pi; y) = \pi^{\sum_{i=1}^{n} y_i} (1 - \pi)^{n - \sum_{i=1}^{n} y_i}$.

1.3.2 Loglikelihood function

For computational reasons, we will usually work with the loglikelihood function $l(\theta; y)$, which is just the natural logarithm of the likelihood function

$$l(\theta; y) := \log(L(\theta; y)) = \log(\prod_{i=1}^{n} L(\theta; y_i)) = \sum_{i=1}^{n} l(\theta; y_i)$$

e.g. $Y \sim Po(\mu)$, we observe $y = (y_1, \ldots, y_n)$ hence
$l(\mu; y) = -n\mu + \sum_{i=1}^{n} y_i \log(\mu) - \sum_{i=1}^{n} \log(y!)$.

e.g. $Y \sim Bern(\pi)$, we observe $y = (y_1, \ldots, y_n)$ hence
$l(\pi; y) = \sum_{i=1}^{n} y_i \log(\pi) + (n - \sum_{i=1}^{n} y_i) \log(1 - \pi)$.

1.3.3 Maximum likelihood estimation

As the sample size grows, two things happen for "nice" loglikelihood functions:

- they become more and more peaked around a maximum value
- their shape becomes more and more quadratic

Clearly this maximum value is important. We know how to find it: differentiate $l(\theta; x)$ with respect to θ; equate this to zero and solve. What results is a numerical value for θ which maximizes the loglikelihood and therefore the likelihood. Intuitively, it seems a good guess for what the true value of θ might be.

We usually label this number $\hat{\theta}$. It is called the **maximum likelihood estimate (MLE) of θ.**

e.g. $Y \sim Po(\mu)$, we observe $y = (y_1, \ldots, y_n)$ hence $l(\mu; y) = -n\mu + \sum_{i=1}^n y_i \log(\mu) - \sum_{i=1}^n \log(y!)$.

- Thus $\frac{\partial l(\mu;y)}{\partial \mu} = -n + \frac{1}{\mu}\sum_{i=1}^n y_i$.
- MLE $\hat{\mu}$ satisfies $0 = -n + \frac{1}{\hat{\mu}}\sum_{i=1}^n y_i$.
- Solving, we find that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n y_i$, the sample mean.

e.g. $Y \sim Bern(\pi)$, we observe $y = (y_1, \ldots, y_n)$ hence $l(\pi; y) = \sum_{i=1}^n y_i \log(\pi) + (n - \sum_{i=1}^n y_i)\log(1-\pi)$.

- Thus $\frac{\partial l(\pi;y)}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^n y_i - \frac{1}{1-\pi}(n - \sum_{i=1}^n y_i)$.
- MLE $\hat{\pi}$ satisfies $0 = \frac{1}{\hat{\pi}}\sum_{i=1}^n y_i - \frac{1}{1-\hat{\pi}}(n - \sum_{i=1}^n y_i)$.
- Solving, we find $\hat{\pi} = \frac{1}{n}\sum_{i=1}^n y_i$, the sample mean.

Note that for finding MLEs, only the part of $l(\theta; y)$ involving θ is relevant. This part is called the **kernel.**

1.3 Likelihood and Maximum-likelihood Estimation

MLEs (usually) have several nice properties:

- they are unbiased - $E(\hat{\theta}) = \theta$
- they are consistent - $\hat{\theta} \to \theta$ as the sample size $n \to \infty$
- they are asymptotically normal - $\hat{\theta} \sim N(\theta, \sigma^2_{MLE})$ where $\sigma^2_{MLE} = \frac{1}{I(\theta)}$ and $I(\theta) = -E\left(\frac{\partial^2 l(\theta;Y)}{\partial \theta^2}\right)$, the Fisher Information.

e.g. $Y \sim Bern(\pi)$, we observe $y = (y_1, \ldots, y_n)$. We have shown $l(\pi; y) = \sum_{i=1}^n y_i \log(\pi) + (n - \sum_{i=1}^n y_i) \log(1 - \pi)$ and $\hat{\pi} = y/n$.

- $\frac{\partial^2 l(\pi;y)}{\partial \pi^2} = -\frac{\sum_{i=1}^n y_i}{\pi^2} - \frac{n - \sum_{i=1}^n y_i}{(1-\pi)^2}$.
- Observations $y_1, \ldots, y_n$ only appear in the sum $\sum_{i=1}^n y_i$. The random version of this sum, $\sum_{i=1}^n Y_i$, follows the $B(n, \pi)$ distribution. Hence $E(\sum_{i=1}^n Y_i) = n\pi$.
- $-E\left(\frac{\partial^2 l(\pi; Y=(Y_1,\ldots,Y_n))}{\partial \pi^2}\right) = \frac{n\pi}{\pi^2} + \frac{n - n\pi}{(1-\pi)^2}$
- We find $I(\pi) = \frac{n}{\pi(1-\pi)}$ and $\hat{\pi} \sim N(\pi, \frac{\pi(1-\pi)}{n})$, or $\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1)$ for n large.

1.4 Large Sample Inference

The asymptotic properties of maximum likelihood estimators provide ways for us to make large sample inference on the parameters of discrete distributions.

We shall consider three significance tests of a null hypothesis $H_0 : \theta = \theta_0$.

1.4.1 Wald test and CI

1.4.2 Score test and CI

1.4.3 Likelihood Ratio test and CI

1.4.1 Wald test and CI

The asymptotic variance of the MLE $\hat{\theta}$ derived from the Fisher Information $I(\theta)$ is a function of θ, the unknown parameter. If we plug in the unrestricted MLE $\hat{\theta}$, we obtain an estimated variance/standard error of $\hat{\theta}$. Let $\iota(\hat{\theta})$ be the Fisher Information evaluated at $\hat{\theta}$. Then the statistic

$$z = (\hat{\theta} - \theta_0)/SE, \text{ where } SE = 1/\sqrt{\iota(\hat{\theta})}$$

has an approximate standard normal distribution when $\theta = \theta_0$. Alternatively, the statistic z^2 has an approximate chi-squared distribution with df = 1, under $\theta = \theta_0$.

This kind of statistic which uses the non-null estimated standard error, is called a **Wald statistic.**

e.g. We have a sample of n IID Bernoulli random variables with probability of success π (equivalently, we observe a binomially distributed random variable with parameters n and π).

Consider $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$

The Wald test statistic $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$ can be used to obtain one- or two-sided P-values.

The related $100(1-\alpha)\%$ confidence interval for π is given by $|z| < z_{\alpha/2}$, or

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

1.4.2 Score test and CI

The **score function** $u(\theta)$ is the first derivative of the loglikelihood, i.e. $u(\theta) := \frac{\partial l(\theta;y)}{\partial \theta}$.

Evaluated at the MLE $\hat{\theta}$, the score function is zero. Evaluated at the null value of θ, θ_0, the score function tends to be larger in absolute value the farther $\hat{\theta}$ is from θ_0. Hence, roughly speaking, the larger the absolute value of $u(\theta_0)$, the less the data supports the null hypothesis H_0.

The test statistic $z = u(\theta_0)/\sqrt{\iota(\theta_0)}$ has an approximate standard normal distribution. Alternatively, the statistic z^2 has an approximate chi-squared distribution with df = 1.

Note the score statistic z (or z^2) uses the null SE and does not require the computation of $\hat{\theta}$, the MLE.

e.g. Again consider a sample of n IID Bernoulli random variables with probability of success π. We wish to test $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$.

The score function is
$u(\pi) := \frac{\partial l(\pi;y)}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^{n} y_i - \frac{1}{1-\pi}(n - \sum_{i=1}^{n} y_i)$.

Thus the score test statistic is

$$z = \frac{u(\pi_0)}{\sqrt{\iota(\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

The related $100(1-\alpha)\%$ confidence interval for π is given by all possible values for π_0 for which $|z| < z_{\alpha/2}$.

1.4.3 Likelihood Ratio test and CI

The Likelihood Ration (LR) test takes two maximizations of the likelihood function: one maximum over the possible parameter values under the null hypothesis H_0; the other is the maximum over the larger set of possible parameter values under H_0 *or* H_1, the alternate hypothesis.

Let ℓ_0 be the maximized likelihood under H_0; let ℓ_1 be the maximized likelihood under H_1. The ratio $\Lambda := \ell_0/\ell_1$ cannot be greater than 1.

It is known that the **LR test statistic** $-2\log(\Lambda)$ has a chi-squared distribution in the limit as $n \to \infty$. The df is the difference between the difference in the dimensions of the parameter spaces under $H_0 \cup H_1$ and H_0.

e.g. Again consider a sample of n IID Bernoulli random variables with probability of success π. We wish to test $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$.

Recall $L(\pi; y) = \pi^{\sum_{i=1}^n y_i}(1-\pi)^{n-\sum_{i=1}^n y_i}$. Under H_0, π can only take one possible value, π_0. Hence, under H_0, the maximum (and only) value $L(\pi; y)$ can take is $\ell_0 = L(\pi_0; y)$.

Alternatively, under $H_0 \cup H_1$, π can take any possible value in $[0, 1]$. We have already shown that $L(\pi; y)$ is maximized at $\pi = \hat{\pi} = \sum_{i=1}^n y_i/n$. Hence $\ell_1 = L(\hat{\pi}; y)$.

The LR test statistic is given by

$$
\begin{aligned}
-2\log(\ell_0/\ell_1) &= 2(\log(\ell_1) - \log(\ell_0)) \\
&= 2\log(\hat{\pi})\sum_{i=1}^{n} y_i + 2(n - \sum_{i=1}^{n} y_i)\log(1-\hat{\pi}) \\
&\quad - 2\log(\pi_0)\sum_{i=1}^{n} y_i + 2(n - \sum_{i=1}^{n} y_i)\log(1-\pi_0) \\
&= 2\log\left(\frac{\hat{\pi}}{\pi_0}\right)\sum_{i=1}^{n} y_i + 2(n - \sum_{i=1}^{n} y_i)\log\left(\frac{1-\hat{\pi}}{1-\pi_0}\right)
\end{aligned}
$$

No unknown parameters occur under H_0 but one occurs under $H_0 \cup H_1$. Thus the LR test statistic will have the χ^2_1 distribution.

The related $100(1-\alpha)\%$ confidence interval for π is given by the set pf π_0 for which the likelihood ratio test has a P-value exceeded α. That is, for all π_0 which satisfy

$$2\log\left(\frac{\hat{\pi}}{\pi_0}\right)\sum_{i=1}^{n} y_i + 2(n-\sum_{i=1}^{n} y_i)\log\left(\frac{1-\hat{\pi}}{1-\pi_0}\right) \leq \chi_1^2(\alpha)$$

In practice, the confidence interval will be found using numerical methods to iteratively solve for the values of π_0 which satisfy the above inequality.

1.4.4 Comparing the tests

- The three tests are asymptotically equivalent - in the limit, their (squared for Wald and Score tests) test statistics will follow a chi-squared distribution with the same df - if H_0 is true.
- If H_0 is not true, the test statistics may take very different values. But in such a situation, usually the test statistics will be large and so H_0 will be rejected nevertheless.

1.4 Large Sample Inference

- The Wald test uses $\hat{\theta}$ and the curvature of likelihood at $\hat{\theta}$. The Score test depends on the slope and curvature of likelihood at θ_0. The LR test uses the values of likelihood at $\hat{\theta}$ and θ_0.
- The Wald test is the most commonly used, because it is simplest. However, the other two are increasingly available in software.
- For small to moderate sample sizes, the LR and Score tests are usually more reliable than the Wald test.
- All three tests rely on "large" sample sizes. A rule of thumb for testing binomial parameter π is $n\pi \geq 5$ and $n(1-\pi) \geq 5$

1.4.5 Example: Eyesight of students and staff

We randomly selected 100 CUHK Statistics students. 53 of these wear glasses. We wish to test whether the (binomial) proportion of CUHK stats students who wear glasses is equal to 0.5 or not. The triumvirate of tests yields the following confidence intervals

- Wald CI: $(0.432, 0.627)$
- Score CI: $(0.433, 0.625)$
- LR CI: $(0.432, 0.626)$

1.4 Large Sample Inference

We randomly selected 10 CUHK Statistics staff. 6 of these wear glasses. We wish to test whether the (binomial) proportion of CUHK stats staff who wear glasses is equal to 0.5 or not. The triumvirate of tests yields the following confidence intervals

- Wald CI: $(0.296, 0.904)$
- Score CI: $(0.313, 0.832)$
- LR CI: $(0.300, 0.854)$

We randomly selected 10 CUHK Fine Arts students. 1 of these wears glasses. We wish to test whether the (binomial) proportion of CUHK Fine Arts students who wear glasses is equal to 0.5 or not. The triumvirate of tests yields the following confidence intervals

- Wald CI: $(-0.086, 0.286)$
- Score CI: $(0.018, 0.404)$
- LR CI: $(0.006, 0.372)$