Handout 4

Inference for a binomial

In handout 2, we have discussed how to obtain an approximate confidence interval for a binomial proportion π. If Y is the number of college graduate in a random sample of size n, then the maximum likelihood estimator (MLE) of π, the population proportion of college graduates, is given by the sample proportion $p = Y/n$. If both $n\pi$ and $n(1-\pi)$ are large, an approximate 95% confidence interval for π is given by $p \pm 1.96SE(p)$, where $SE(p) = \sqrt{p(1-p)/n}$.

Maximum Likelihood estimation.

Recall that the pdf (probability density function) of a binomial distribution is

$$f(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}.$$

Suppose we have observed $y = 71$ college graduates in a random sample of $n = 200$, then the likelihood function is

$$l(\pi) = f(71;\pi) = \binom{n}{71}\pi^{71}(1-\pi)^{n-71} = \binom{200}{71}\pi^{71}(1-\pi)^{200-71},$$

ie, the likelihood function $l(\pi)$ is a function of π treating y as fixed. We may plot $l(\pi)$ against π, and find the value of π at which $l(\pi)$ attains its maximum. If $\hat{\pi}$ is the value of π at which $l(\pi)$ attains its maximum. then $\hat{\pi}$ is called the maximum likelihood estimator of π. A plot of this function is given below when $n = 200$ and $y = 71$.

The maximum of $l(\pi)$ (and $\log(l(\pi))$) is attained at $\hat{\pi} = y/n = 71/200 = 0.3550$.

Instead of plotting $l(\pi)$ against π, we can use calculus to obtain the value of π at which $l(\pi)$ or equivalently $\log(l(\pi))$ attains its maximum. We differentiate $\log(l(\pi))$ with respect to π, equate the derivative to zero and solve for π.

[**Technical Note**: Note that

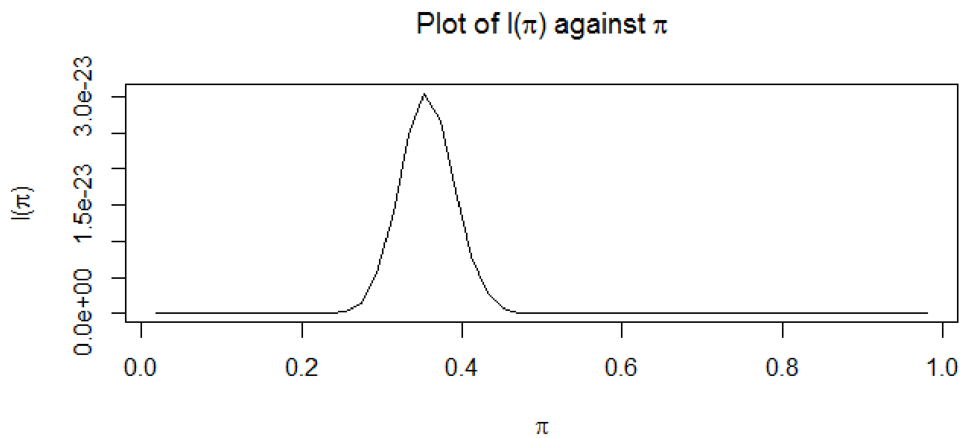
$$\begin{aligned}
\log(\pi)) &= \log\left(\frac{n!}{y!(n-y)!}\right) + y\log(\pi) + (n-y)\log(1-\pi), \text{ and} \\
\frac{d}{d\pi}\log(l(\pi)) &= y/\pi - (n-y)/(1-\pi), \\
0 &= \frac{d}{d\pi}\log(l(\pi)) \Longrightarrow \hat{\pi} = y/n.]
\end{aligned}$$

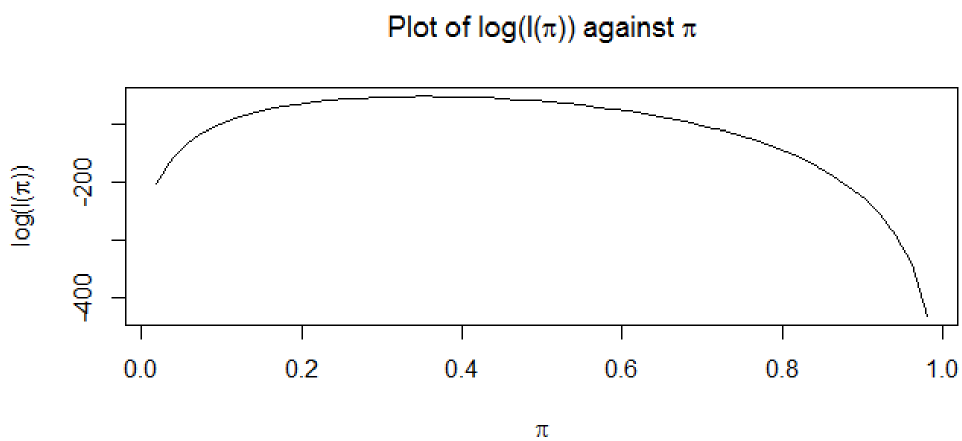
Hypothesis testing (Wald and score statistics)

Suppose we wish to test if the percentage of college graduates is 30% or it is higher. In the statistical jargon we state this as: test $H_0 : \pi = 0.3$ against $H_1 : \pi > 0.3$ at some given level of significance, say $\alpha = 0.05$. The test statistic is

$$z = \frac{p - 0.3}{SE},$$

where SE is an estimate of $\sqrt{Var(p)} = \sqrt{\pi(1-\pi)/n}$. If we estimate SE as $\sqrt{p(1-p)/n}$, then the corresponding z-statistic is called the Wald statistic. It SE is estimated at the null (ie, $\pi = 0.3$), then the


Plot of l(π) against π
l(π)
0.0e+00
1.5e-23
3.0e-23
0.0
0.2
0.4
0.6
0.8
1.0
π


Plot of log(l(π)) against π
log(l(π))
-400
-200
0.0
0.2
0.4
0.6
0.8
1.0
π

corresponding z-statistic is called the score statistic. Thus if we want to test $H_0 : \pi = \pi_0$ vs $H_1 : \pi > \pi_0$, where π_0 is some prespecified value (say $\pi_0 = 0.3$), the Wald and the score statistics are

$$\begin{aligned} z &= \frac{p - \pi_0}{\sqrt{p(1-p)/n}}, \text{ Wald statistic,} \\ z &= \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}, \text{ score statistic.} \end{aligned}$$

If $\alpha = 0.05$, then we reject H_0 if the calculated z value is larger than 1.645 (since area to the left of 1.645 under the standard normal curve is 0.95).

For our data $y = 71, n = 200$, thus we have $p = y/n = 0.355$, and

$$\begin{aligned} &\text{Wald statistic:} && z = \frac{0.355 - 0.3}{\sqrt{(0.355)(1-0.355)/200}} = 1.625 \\ &\text{Score statistic:} && z = \frac{0.355 - 0.3}{\sqrt{(0.3)(1-0.3)/200}} = 1.697. \end{aligned}$$

Thus according to the test using Wald statistic, we fail to reject the null. Whereas we reject the null if we use the score statistic. Normally, we do not expect different conclusions for these two different statistics. However, in this case they turn out to be different.

We now turn to the calculation of the p-value. If we use the score statistic, then the p-value is area to the right of 1.697 under the standard normal curve. Thus the p-value is 0.0450. Similarly, the p-value when we use the Wald statistic is 0.052.

Note that in order to make a decision in any hypotheses testing problem, we use the cutoff value (also called the critical value) which is obtained from an appropriate statistical table (Normal, t, chi-square, F etc.). For instance, in the hypothesis testing problem above $H_0 : \pi = 0.3$ against $H_1 : \pi > 0.3$ at level $\alpha = 0.05$, the cutoff value is 1.645, and we reject H_0 if the z-statistic is larger than 1.645. We will use the terms 'cutoff value' and 'critical value' interchangeably. Also note that in any hypothesis testing problem, we may calculate the p-value instead of finding the cutoff value. Then the decision rule is: reject H_0 if the p-value is smaller than the given level of significance α.

Remark 1

(a) Note that the method for testing $H_0 : \pi = 0.3$ vs $H_1 : \pi > 0.3$ is identical to the testing for $H_0 : \pi \leq 0.3$ vs $H_1 : \pi > 0.3$. Thus if $\alpha = 0.05$ then we reject the null if the value of the z statistic is larger than 1.645.

(b) For the two-sided alternative, we test $H_0 : \pi = 0.3$ vs $H_1 : \pi \neq 0.3$ at level $\alpha = 0.05$, then we reject H_0 if $|z| > 1.96$ (note that area between -1.96 and 1.96 under the standard normal curve is 0.95). The p-value is two times the p-value of the test $H_0 : \pi = 0.3$ vs $H_1 : \pi > 0.3$. If we use the Wald statistic, then the p-value is $(2)(0.052) = 0.104$.

(c) For the test with two-sided alternative, we may also use an equivalent test called the chi-square test using the square of the z statistic . Under H_0, $z^2 \overset{approx}{\sim} \chi^2_1$ [χ^2_1 stands for chi-square with 1 df]. Thus if $\alpha = 0.05$, then we reject H_0 if $z^2 > 3.841$ [area to the left of 3.841 under the chi-square curve with 1 df is 0.95]. Note that $|z| > 1.96$ is equivalent to $z^2 > 1.96^2 = 3.841$. [**You cannot use the chi-square test when the alternative is one-sided. The chi-square test is valid only when the alternative hypothesis is two-sided**.]

Hypothesis testing (likelihood-ratio statistic)

When the alternative is two-sided, there is another method of testing which uses what is known as the likelihood ratio. Suppose that we are testing $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$, where $\pi_0 = 0.3$. When the observed number of college graduates is $y = 71$, then the likelihood function is

$$l(\pi) = f(y; \pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y} = \binom{200}{y} \pi^y (1-\pi)^{200-y}.$$

The maximum likelihood estimate $\hat{\pi} = p = y/n$ of π is obtained by maximizing $l(\pi)$ with respect to π. Thus we expect $l(\hat{\pi})$ to be larger than $l(\pi)$ for any $\pi \neq \hat{\pi}$, and hence $l(\hat{\pi}) > l(\pi_0)$. Denote

$$l_1 = l(\hat{\pi}), \;\; l_0 = l(\pi_0).$$

The ratio l_0/l_1 is called the likelihood ratio. Note that the likelihood ratio is smaller than 1. If the null hypothesis $H_0 : \pi = \pi_0$ is true, then $\hat{\pi}$ is close to π_0 and hence the likelihood ratio should be close to 1. If however, H_0 is false, then the likelihood ratio should be quite a bit smaller than 1. Thus a reasonable test would be: reject H_0 if the likelihood ratio is quite a bit smaller than 1. We need a cutoff point to decide how small is small. For technical reasons, we look at $-2\log(l_0/l_1)$ instead of the likelihood ratio itself. **The likelihood-ratio (LR) test statistc is defined as** $-2\log(l_0/l_1)$. Note that if H_0 is indeed correct (ie, $\pi = \pi_0$), then the LR statistic should be small. If H_0 is false, the LR statistic should be quite a bit larger than 0. This forms the basic idea of the test: reject H_0 if the calculated LR statistic is quite a bit larger than 0. Fortunately, we can find a cutoff point for the LR statistic using the chi-square table. When $\alpha = 0.05$, we reject the null hypothesis if the value of the LR statistic is larger than 3.841 [area to the left of 3.841 under the chi-square curve with 1 df is 0.95]. The justification comes from the following result.

Fact 1. Under H_0, the LR statistic is approximately distributed as chi-square with 1 df if n is large.

We want to test $H_0 : \pi = 0.3$ vs $H_1 : \pi \neq 0.3$ at level $\alpha = 0.05$. We have $n = 200, y = 71, \hat{\pi} = 0.355$, $\pi_0 = 0.3$, and thus

$$\begin{aligned}
\frac{l_0}{l_1} &= \frac{l(\pi_0)}{l(\hat{\pi})} = \frac{\binom{n}{y}\pi_0^y(1-\pi_0)^{n-y}}{\binom{n}{y}\hat{\pi}^y(1-\hat{\pi})^{n-y}} = \left(\frac{\pi_0}{\hat{\pi}}\right)^y \left(\frac{1-\pi_0}{1-\hat{\pi}}\right)^{n-y}, \\
-2\log(l_0/l_1) &= -2\log\left(\frac{\pi_0}{\hat{\pi}}\right)^y - 2\log\left(\frac{1-\pi_0}{1-\hat{\pi}}\right)^{n-y} \\
&= 2y\log\left(\frac{\hat{\pi}}{\pi_0}\right) + 2(n-y)\log\left(\frac{1-\hat{\pi}}{1-\pi_0}\right) \\
&= (2)(71)\log\left(\frac{0.355}{0.3}\right) + 2(200-71)\log\left(\frac{1-0.355}{1-0.3}\right) \\
&= 2.791.
\end{aligned}$$

Area to the right of 3.841 under the chi-square curve with 1 df is 0.05, and hence the cutoff point is 3.841 when $\alpha = 0.05$. Here we cannot reject H_0 as the LR statistic is smaller than the cutoff point.

Calculation of p-value for moderate sample sizes.

Suppose that in random sample of size $n = 20$, we have 8 college graduates. We want to test $H_0 : \pi = 0.3$ vs $H_1 : \pi > 0.3$ using the score statistic. We would like to calculate the p-value using the normal approximation. When ni is of moderate size, the accuracy of the approximation can be improved by using what is known as the 'continuity correction. Note that estimated value of π and z-statistic are

$$\begin{aligned}
\hat{\pi} &= 8/20 = 0.4, \\
z &= \frac{\hat{\pi} - 0.3}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.1}{\sqrt{(0.3)(1-0.3)/20}} = 0.9759.
\end{aligned}$$

The p-value can be calculated as the area to the right of 0.9759 under the standard normal curve, which equals 0.1646.

Now let us look at the original definition of the p-value. The p-value is the probability of getting 8 or more graduates if the null were true (ie, when $\pi = 0.3$). Thus the p-value is (exact calculation using binomial probabilities)

$$P(Y \geq 8) = 1 - P(Y \leq 7) = 0.2277.$$

This value can be obtained using the R command: 1-pbinom(7,size=16,prob=0.3). Note that pbinom(7,size=16,prob=0.3) gives you $P(Y \leq 7)$.

Central Limit Theorem tells us under H_0 the random variable Y, the number of college graduate, is approximately normally distributed with mean $n\pi_0 = (20)(0.3) = 6$ and $SD = \sqrt{n\pi_0(1-\pi_0)} = \sqrt{(20)(0.3)(1-0.7)} = 2.0494$. Without continuity correction, we approximate $P(Y \geq 8)$ by area to the right of

$$(8-6)/2.0494 = 0.9759$$

under the standard normal curve, and this area is 0.1646. Thus the approximate p-value using **normal approximation without continuity correction** is 0.1646.

If we use continuity correction, then $P(Y \geq 8)$ is approximated by the area under the standard normal curve to the right of

$$(7.5-6)/2.0494 = 0.7319.$$

Area to the right of 0.7319 under the normal curve is 0.2321. **With continuity correction**, the p-value is approximately equal to 0.2321. Note that the p-value using the continuity correction is much closer to the correct value than the normal approximation without continuity correction.

Some useful R commands.

If $Y \sim binomial(20, 0.4)$, then

(a) $P(Y = 6)$ can be obtained by the R command dbinom(6,size=20,prob=4),

(b)$P(Y \leq 6)$ can be obtained by the R command pbinom(6,size=20,prob=4).

If $Y \sim Poisson(4)$, then

(a) $P(Y = 6)$ can be obtained by the R command dbpois(6,lambda=4),

(b)$P(Y \leq 6)$ can be obtained by the R command ppois(6,lambda=4).

If $Y \sim N(3, 2^2)$, then

(a) $P(Y \leq 6)$ can be obtained by the R command pnorm(6,mean=3,sd=2),

(b) 0.9 quantile can be obtained by the R command pnorm(0.9,mean=3,sd=2). [0.9 quantile of the distribution of Y is that value y such that $0.9 = P(Y \leq y)$.]

A note on Normal approximation for the binomial.

For large n, the binomial distribution may be approximated by the normal distribution. However, when the sample size is moderate, this approximation can be improved by using what is known as the 'continuity correction'. Let $Y \sim binomial(16, 0.6)$. Then

$$\begin{aligned} E(Y) &= n\pi = (16)(0.6) = 9.6, \\ Var(Y) &= n\pi(1-\pi) = (16)(0.6)(0.4) = 3.84, \\ SD(Y) &= \sqrt{3.84} = 1.9596. \end{aligned}$$

Suppose that we want to calculate $P(Y \leq 8)$. Then

$$P(Y \leq 8) = \sum_{y=0}^{8} P(Y = y) = \sum_{y=0}^{8} \binom{16}{y} (0.6)^y (0.4)^{16-y} = 0.284.$$

If we use the normal approximation for the binomial **without continuity correction**, then $P(Y \leq 8)$ is approximately equal to the area under the normal curve to the left of

$$(8 - E(Y))/SD(Y) = (8 - 9.6)/1.9596 = -0.8165.$$

So using the normal table, we have $P(Y \leq 8) \approx 0.207$. Note that this approximation is not that accurate.

The method of continuity correction suggests approximating $P(Y \leq 8)$ by the area under the normal curve to the left of

$$(8.5 - E(Y))/SD(Y) = (8.5 - 9.6)/1.9596 = -0.5613,$$

and this area is about 0.287 which is close to the correct value.

Suppose that we want to calculate $P(Y = 8)$ using the normal approximation. First note that $P(Y = 8) = 0.142$ using the binomial formula. We may try to approximate this probability by finding the area under the normal curve between

$$\begin{aligned} (7.5 - 9.6)/1.9596 &= -1.0717 \text{ and} \\ (8.5 - 9.6)/1.9596 &= -0.5613, \end{aligned}$$

and this area is about 0.145. Once again, note that this approximation is close to the true value.

Hypothesis testing for the multinomial.

A recruiting company believes that the highest educational levels of the adults in a certain large city are: 40% college graduates, 40% high school graduates, and 20% none. In order to verify this claim, a sample of $n = 200$ adults are taken, and let n_1, n_2 and n_3 be the sample counts of college graduates, high school graduates, and no degree, respectively. Then $(n_1, n_2, n_3) \sim multinomial(n; \pi_1, \pi_2, \pi_3)$, where π_1, π_2 and π_3 are the true population proportions. Based on the data we would like to test $H_0 : \pi_1 = \pi_{10}, \pi_2 = \pi_{20}, \pi_3 = \pi_{30}$ against H_1 : at least one $\pi_j \neq \pi_{j0}$, where $\pi_{10} = 0.4, \pi_{20} = 0.4$, and $\pi_{30} = 0.2$.

In general if the multinomial has c categories, we may want to test $H_0 : \pi_j = \pi_{j0}$, $j = 1, \ldots, c$ against H_1 : at least one $\pi_j \neq \pi_{j0}$, where the values of π_{j0} are prespecified. We know that $E(n_j) = n\pi_j$. If H_0 were true, then $E(n_j) = n\pi_{j0}$ is the expected count (frequency) under H_0. Denote the expected frequency $n\pi_{j0}$ under H_0 by μ_j. Since $\hat{\pi}_j = n_j/n$ estimates π_j, any reasonable hypothesis testing method should compare $\hat{\pi}_j$'s to π_{j0}'s, or equivalently compare n_j's to μ_j's.. Two well known test statistics are

$$\begin{aligned} X^2 &= \sum_{j=1}^{c} \frac{(n_j - \mu_j)^2}{\mu_j} \text{ [Pearson's chi square statistic]}, \\ G^2 &= 2\sum n_j \log(n_j/\mu_j) \text{ [Likelihood Ratio (LR) statistic]}. \end{aligned}$$

Under H_0, both X^2 and G^2 are approximately distributed as χ^2_{c-1} (chi-square with $c-1$ df) if $\mu_j = n\pi_{j0}$ is large for all j.

If we use any of these statistics, then we can reject H_0 if the value of the statistic is larger than the cutoff value obtained from the chi-square table.

Example. We have the following counts based on a random sample of $n = 200$ adults in a large city: 93 (college graduate), 75 (high school graduates), 32 (none). Let $\pi_{10} = 0.4, \pi_{20} = 0.4, \pi_{30} = 0.2$. We would like to test $H_0 : \pi_j = \pi_{j0}$, $j = 1, \ldots, c = 3$ against H_1 : at least one $\pi_j \neq \pi_{j0}$, at level $\alpha = 0.05$.

In this example $c = 3$. The expected counts under H_0 are:

$$\begin{aligned} \mu_1 &= n\pi_{10} = (200)(0.4) = 80, \\ \mu_2 &= n\pi_{20} = (200)(0.4) = 80, \\ \mu_3 &= n\pi_{30} = (200)(20) = 40. \end{aligned}$$

Thus we have

$$\begin{aligned} X^2 &= \sum_{j=1}^{c} \frac{(n_j - \mu_j)^2}{\mu_j} \\ &= \frac{(93-80)^2}{80} + \frac{(75-80)^2}{80} + \frac{(32-40)^2}{40} = 4.0250, \\ G^2 &= 2\sum n_j \log(n_j/\mu_j) \\ &= 2\left[(93)\log(93/80) + (75)\log(75/80) + (32)\log(32/40)\right] = 4.0446 \end{aligned}$$

Area to the right of 5.9915 under the chi-square curve with $c-1 = 2$ df is 0.05 (the R command qchisq(0.95,2) yields 5.9915). Note that both X^2 and G^2 are smaller than 5.9915. Thus the null hypothesis cannot be rejected by any of the two-tests.

If we use Pearson's chi-square, then the p-value is area to the right of 4.0250 under the chi-square curve with 2 df, ie, p-value=$P(\chi^2_2 \geq 4.0250)$. This p-value is 0.1334 (using the R command 1-pchisq(4.0250,2)).

Remark 2.

(a) Both Pearson's chi-square and LR tests assume that $\mu_j = n\pi_{j0}$ is large for all j. A rule of thumb: μ_j should be 5 or larger for all j.

(b) If the null hypothesis were true (or if the true π_j's are close to π_{j0}'s), then the values of X^2 and G^2 are usually not all that different. However, if the true π_j's are quite different from π_{j0}'s, then the values of X^2 and G^2 may be quite different.

(c) Pearson's chi-square and LR statistics are the two most widely used test statistics in the analysis of categorical data. Some researchers recommend the use of X^2 (instead of G^2) if some of the observed counts are small.