Packet 4: Interval Estimation

Chap 7.3 Confidence interval for proportions

Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli trial with success probability p.

$X_i = 0$ (failure) or 1 (success) and $P(X_i = 1) = p$.

The point estimate $\hat{p} = \bar{X}$ is called the sample proportion.

Example 1: A researcher is interested in answering the following research question: What proportion of American Internet users is addicted to the Internet?

The researcher deems that a person is addicted to the Internet if the person exhibits at least five of ten possible characteristics, such as using the Internet to escape from his/her problems, trying unsuccessfully to cut back his/her usage, and finding himself/herself preoccupied with the Internet when no longer sitting at a computer.

What proportion of the 230,630,000 American Internet users (Source: World Bank, 2008) are addicted to the Internet? population

Solution: Clearly, the researcher can't evaluate the behavior of all of the 230 million American Internet users. Instead, suppose the researcher takes a random sample of 17,251 American Internet users and evaluates them for Internet addiction. The resulting data would be 17,251 data points, to be exact by yes, yes, no, yes, no, no, no, no, no, no, ...

is a sufficient statistic for the parameter p.

The data in such a raw format are not particularly helpful, so the researcher uses a "sample statistic" to summarize the data. We may summarize data as the total number of successes, $Y = \sum_{i=1}^{n} X_i \sim Binomial(n, p)$. n is known and fixed

$p(Y = y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}$, $E(Y) = np$, $Var(Y) = np(1-p)$.

p.m.f. $y = 990$ MLE and MOM chap 6.4

Based on his sample, the researcher got 990 'yes' and calculated the proportion in his sample that he deemed addicted to the Internet: $\hat{p} = 990/17521 = 0.057$. In this Chapter, we can use a confidence interval to help quantify the value of a population parameter and make a statement such as this: "We can be 95% confident that the interval between 0.0534 and 0.0606 (0.057 ± 0.0036) contains the true proportion of Americans addicted to the Internet."

This narrow C.I. is because of the large sample size n.

In a typical survey for proportion estimation, will have sample size 1000 ~ 1600 Chap 7.4 sample size calculation

$E(X_i) = p$, $Var(X_i) = p(1-p)$. By central limit theorem, we have

If n is large $\quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = Y/n \sim N(p, p(1-p)/n)$ 1st Approx.

The approximation is valid if $np \geq 5$ and $n(1-p) \geq 5$.

standarization → pivotal $\quad \frac{Y/n - p}{\sqrt{p(1-p)/n}} \sim N(0,1).$

BUT, p is unknown. We replace p with $\hat{p} = Y/n$. in the denominator

$$\frac{Y/n - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1).$$

p is unknown so we can use

$\hat{p} = y/n$

$n \times \hat{p} = y \geq 5$

$n \times (1-\hat{p}) = n - y \geq 5$

2nd Approx.

For a $100(1-\alpha)\%$ C.I., find $z_{\alpha/2}$, such that

$$P(-z_{\alpha/2} \leq \frac{Y/n - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{\alpha/2}) = 1 - \alpha.$$

$\sim N(0,1)$

$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \to \frac{\bar{X} - \mu}{\sqrt{\hat{S}^2/n}}$

Summary of confidence interval for p

- A $100(1-\alpha)\%$ two sided confidence interval for p is

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$

where $z_{\alpha/2}$ is a constant satisfying $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$ if $Z \sim N(0,1)$.

- An one sided $100(1-\alpha)\%$ confidence interval with lower bound for p is

$$\left(\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \ 1\right);$$

the parameter space is $0 \leq p \leq 1$

- An one sided $100(1-\alpha)\%$ confidence interval with upper bound for p is

$$\left(0,\ \ \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$

where z_{α} is a constant satisfying $P(Z > z_{\alpha}) = \alpha$ if $Z \sim N(0,1)$.

* What affects the width of C.I. ?

① $\alpha \uparrow$ $100(1-\alpha)\%$ ↓ narrow C.I.

② $n \uparrow$ narrow C.I.

③ $\hat{p} \to 0$ $\hat{p} \to 1$ $\hat{p}(1-\hat{p})$ ↓ narrow C.I.

Example 2: We are interested in the success rate p of Nittany Lion football team, say, in 10 games, Nittany Lion won 7 times.

Find the two sided 95% C.I. for p.

Solution: Estimate the success rate p by $\hat{p} = 7/10 = 0.7$. $z_{\alpha/2} = z_{0.025} = 1.96$.

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

leads to

$$\left(0.7 - 1.96\sqrt{\frac{0.7(1-0.7)}{10}}, \quad 0.7 + 1.96\sqrt{\frac{0.7(1-0.7)}{10}}\right)$$

which is

$$(0.416, \quad 0.984)\,.$$

Find the one sided 95% C.I. for p with the lower bound.

Solution: Estimate the success rate p by $\hat{p} = 7/10 = 0.7$. $z_{\alpha} = z_{0.05} = 1.645$.

$$\left(\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad 1\right)$$

leads to

$$\left(0.7 - 1.645\sqrt{\frac{0.7(1-0.7)}{10}}, \quad 1\right)$$

which is

$$(0.462, \quad 1)\,.$$

C.I. for the difference between two proportions $p_1 - p_2$

We are interested in comparing the success rates between two populations p_1 and p_2.

Example 3: we have two strategies for the football game, to know which way is more effective, we perform the experiments

- n_1 times using the first strategy, and got Y_1 of successes;
- n_2 times using the second strategy, and got Y_2 of successes.

We estimate

- $\hat{p}_1 = Y_1/n$ an estimator for success rate p_1 of the first strategy;
- $\hat{p}_2 = Y_2/n$ an estimator for success rate p_2 of the first strategy.

By central limit theorem, we have

$$Y_1/n_1 \sim N(p_1, \frac{p_1(1-p_1)}{n_1}),$$
$$Y_2/n_2 \sim N(p_2, \frac{p_2(1-p_2)}{n_2}),$$

1st Approx. CLT

and thus

$$Y_1/n_1 - Y_2/n_2 \sim N(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}).$$

parameter of interest

standarization

$$\frac{(Y_1/n_1 - Y_2/n_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1).$$

Replace the unknown p_1 and p_2 in the variance components by $\hat{p}_1$ and $\hat{p}_2$. 2nd Approx.

square root term

Summary of confidence interval for $p_1 - p_2$

- A $100(1-\alpha)\%$ two sided confidence interval for $p_1 - p_2$ is

$$\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right),$$

- An one sided $100(1-\alpha)\%$ confidence interval with lower bound for $p_1 - p_2$ is

$$\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ 1\right);$$

- An one sided $100(1-\alpha)\%$ confidence interval with upper bound for $p_1 - p_2$ is

$$\left(-1,\ \hat{p}_1 - \hat{p}_2 + z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right).$$

$0 \le p_1 \le 1 \quad 0 \le p_2 \le 1$

$-1 \le p_1 - p_2 \le 1$

Never use t distn for proportion parameters

Example 3: we have two strategies for the football game, to know which way is more effective, we perform the experiments

- 25 times using the first strategy, and got 17 of successes;
- 25 times using the second strategy, and got 23 of successes.

We estimate

Step ①

- $\hat{p}_1 = 17/25 = 0.68$;
- $\hat{p}_2 = 23/25 = 0.92$.

If we are interested in whether two strategies make no difference: $p_1 = p_2$ or $p_1 - p_2 = 0$. A 95% two sided confidence interval for $p_1 - p_2$ is

Step ②

$$0.68 - 0.92 \pm 1.96\sqrt{\frac{0.68(1-0.68)}{25} + \frac{0.92(1-0.92)}{25}},$$

Chap 8 Hypothesis Test

which is (-0.451, -0.028).

$z_{\alpha/2} = z_{0.025}$

C.I. does not cover 0

We believe there is a difference

If we are interested in whether the second strategy is better than the first strategy: $p_1 < p_2$ or $p_1 - p_2 < 0$. A 95% one sided confidence interval for $p_1 - p_2$ with upper bound is

look at the highest possible value of $P_1 - P_2$ = upper bound

$$\left(-1,\ \ 0.68 - 0.92 + 1.645\sqrt{\frac{0.68(1-0.68)}{25} + \frac{0.92(1-0.92)}{25}}\right),$$

which is (-1, -0.062).

$z_\alpha = z_{0.05}$

b/c upper bound is still below 0

We believe 2nd Strategy is better.

More Examples: 7.3-1, 7.3-2, 7.3-3, 7.3-4

HW 4 Chap 7.2 Chap 7.3