Math 4501 - Probability and Statistics II

7.3 - Confidence intervals for proportions

Overview

We will construct approximate confidence intervals for:

- a proportion ⇝ like Sec. 7.1
- the difference of two proportions ⇝ like Sec. 7.2

Single proportion

random sample $X_1, \ldots, X_n$ Bernoulli r.v.s

$Y = \sum_{i=1}^{n} X_i$ = # of successes out of n trials $\sim b(n, p)$

Let Y be the number of observed successes in n Bernoulli trials with (unknown) probability p of success on each trial.

our random sample $X_1, \ldots, X_n$

GOAL: Find an interval estimate for p.

Recall that:

- Y is $b(n, p)$
- $\dfrac{Y - np}{\sqrt{np(1-p)}} = \dfrac{(Y/n) - p}{\sqrt{p(1-p)/n}}$ is approximately $N(0,1)$.

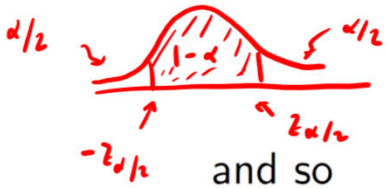
for n large enough

Using CLT: { $X_1, \ldots, X_n$ iid r.v.s with mean μ and variance σ^2 then

$\sum X_i = n\bar{X}$

$\dfrac{Y - np}{\sqrt{n}\sigma} = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$

We find that



$$P\left[-z_{\alpha/2} \le \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right] \approx 1 - \alpha$$

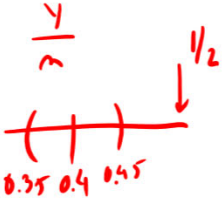
and so

solving for p in the numerator

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \approx 1 - \alpha\ .$$

Note: the unknown parameter p appears in the endpoints of this inequality.

Approach 1 (easier and most practical work around)



Make an additional approximation, by replacing p with its estimate Y/n in $p(1-p)/n$ in the endpoints.

p is probability of success

$\hat{p} = \bar{X} = \frac{Y}{n}$

point estimate for p

If n is large enough, it is still true that

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}}\right] \approx 1-\alpha\,.$$

unknown $p \in ?$

$\frac{Y}{n}$

$\hat{p}$ gives the proportion of successes $\frac{Y}{n}$ in n trials

Thus, for large enough n, if the observed Y equals y, then the interval

$$\rightsquigarrow \quad \left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}, \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}\right]$$

serves as an approximate $100(1-\alpha)\%$ confidence interval for p.

$X_1, \ldots, X_n \sim$ Bernoulli pmf $p^x(1-p)^{1-x}$, $x = 0, 1$

$Y \sim$ Binomial pmf $\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n$

Approach 2 (more precise than approach 1, but requires more work)

Note that the inequality

$$\frac{|Y/n - p|}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}$$

in

$$P\left[-z_{\alpha/2} \leq \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

is equivalent to

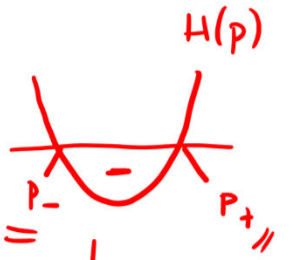
$$H(p) = \left(\frac{Y}{n} - p\right)^2 - \frac{z_{\alpha/2}^2 p(1-p)}{n} \leq 0$$

- Since $H(p)$ is quadratic in p, we can find those values of p for which $H(p) \leq 0$ by finding the two zeros of $H(p)$.

square both sides

$$\frac{(Y/n - p)^2}{p(1-p)/n} \leq (z_{\alpha/2})^2$$

$$\left(\frac{Y}{n} - p\right)^2 \leq (z_{\alpha/2})^2 \frac{p(1-p)}{n}$$



gives confidence interval (p_-, p_+)

and graph of $H(p)$ opens up because the coefficient of p^2 is $1 + \frac{(z_{\alpha/2})^2}{n} > 0$

Set $\hat{p} = \dfrac{Y}{n}$ and $z_0 = z_{\alpha/2}$ in

$$H(p) = \left(\frac{Y}{n} - p\right)^2 - \frac{z_{\alpha/2}^2\, p(1-p)}{n} \leq 0$$

to get

$$H(p) = \left(1 + \frac{z_0^2}{n}\right) p^2 - \left(2\hat{p} + \frac{z_0^2}{n}\right) p + \hat{p}^2 \,,$$

and then use the quadratic formula to obtain the zeros of $H(p)$ as

$$p_{\pm} = \frac{\hat{p} + z_0^2/(2n) \pm z_0 \sqrt{\hat{p}(1-\hat{p})/n + z_0^2/(4n^2)}}{1 + z_0^2/n} \,.$$

These zeros give the endpoints for an approximate $100(1-\alpha)\%$ confidence interval for p.

Note: if n is large, $z_0^2/(2n)$, $z_0^2/(4n^2)$, and z_0^2/n are all small and these two alternative confidence intervals are approximately equal when n is large.

Example

A certain Bernoulli experiment has unknown probability of success p.

A number $n = 40$ of Bernoulli trials were performed and $y = 8$ successes were observed.

Determine an approximate 90% confidence interval for p.

point estimate for population ch success $\hat{p} = \frac{y}{n} = \frac{8}{40}$

Note that $y/n = 8/40 = 0.20$ and $z_{\alpha/2} = 1.645$.

Using the first approach described above, the 90% confidence interval for p is

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}, \frac{y}{n} - z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}\right] =$$

$$= \left[0.20 - 1.645\sqrt{\frac{(0.20)(0.80)}{40}}, 0.20 + 1.645\sqrt{\frac{(0.20)(0.80)}{40}}\right] = [0.096, 0.304] .$$

Using the second approach, we look for the zeros of

$\hat{p} = \frac{y}{n} = \frac{8}{40} = 0.2$

$$H(p) = \left(1 + \frac{z_0^2}{n}\right) p^2 - \left(2\hat{p} + \frac{z_0^2}{n}\right) p + \hat{p}^2$$

$$= \left(1 + \frac{1.645^2}{40}\right) p^2 - \left(2(0.2) + \frac{1.645^2}{40}\right) p + 0.2^2$$

→ find zeros using quadratic formula

to get the 90% confidence interval $[0.117, 0.321]$. (compare with answer using Approach 1: $[0.096, 0.304]$)

For larger sample sizes n, the two solutions will be closer:

For instance, if the sample size had been $n = 400$ and $y = 80$ successes had been observed (so that $y/n = 80/400 = 0.20$), the two 90% confidence intervals would have been

$\hat{p} = \frac{80}{400} = 0.2$

$$[0.167, 0.233] \quad \text{and} \quad [0.169, 0.235],$$

respectively.

much less wide than the intervals obtained with sample of size 40

One-sided confidence intervals

The techniques discussed above allow us to determine one-sided confidence intervals for the unknown proportion p.

The one-sided confidence interval for p given by

$$\left[0, \frac{y}{n} + z_\alpha \sqrt{\frac{(y/n)[1-(y/n)]}{n}}\right]$$

provides an upper bound for p. while

$$\left[\frac{y}{n} - z_\alpha \sqrt{\frac{(y/n)[1-(y/n)]}{n}}, 1\right]$$

provides a lower bound for p.

Difference of proportions

Often, there are two (or more) independent ways of performing an experiment.

- Suppose these have probabilities of success p_1 and p_2, respectively.

For each method to perform the experiment, let:

- n_1 and n_2 be the respective number of independent trials
- Y_1 and Y_2 be the respective number of successes

Recall that, for each $i = 1, 2$:

- Y_i is $b(n_i, p_i)$
- $\dfrac{Y_i - n_i p_i}{\sqrt{n_i p_i (1 - p_i)}} = \dfrac{(Y_i/n_i) - p_i}{\sqrt{p_i(1 - p_i)/n_i}}$ is approximately $N(0, 1)$

Thus,

$$Z = \frac{(Y_1/n_1) - (Y_2/n_2) - (p_1 - p_2)}{\sqrt{p_1 (1 - p_1)/n_1 + p_2 (1 - p_2)/n_2}} \quad \text{is approximately} \quad N(0, 1) \, .$$

Set $p_1 = Y_1/n_1$ and $p_2 = Y_2/n_2$. For large enough n_1 and n_2

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

Thus, for a given $1-\alpha$, we can find $z_{\alpha/2}$ so that

$$P\left[-z_{\alpha/2} \le \frac{(Y_1/n_1) - (Y_2/n_2) - (p_1 - p_2)}{\sqrt{(Y_1/n_1)(1 - Y_1/n_1)/n_1 + (Y_2/n_2)(1 - Y_2/n_2)/n_2}} \le z_{\alpha/2}\right] \approx 1-\alpha$$

Once Y_1 and Y_2 are observed to be y_1 and y_2, respectively, we obtain an approximate $100(1-\alpha)\%$ confidence interval

$$\frac{y_1}{n_1} - \frac{y_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{(y_1/n_1)(1 - y_1/n_1)}{n_1} + \frac{(y_2/n_2)(1 - y_2/n_2)}{n_2}}$$

for the unknown difference $p_1 - p_2$.

Example

Two detergents were tested for their ability to remove stains of a certain type. An inspector judged the first one to be successful on 63 out of 91 independent trials and the second one to be successful on 42 out of 79 independent trials

Determine an approximate 90% confidence interval for the difference between the rate of success of the two detergents.

point estimates for p_1 and p_2

$\alpha = 0.1 \quad (1-\alpha = 0.9)$

Note that $y_1/n_1 = 63/91 = 0.692$, $y_2/n_2 = 42/79 = 0.532$. and $z_{\alpha/2} = 1.645$.

The endpoints for an approximate 90% confidence interval for $p_1 - p_2$ are

$$\frac{y_1}{n_1} - \frac{y_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{(y_1/n_1)\,(1-y_1/n_1)}{n_1} + \frac{(y_2/n_2)\,(1-y_2/n_2)}{n_2}} =$$

$$= \left(\frac{63}{91} - \frac{42}{79}\right) \pm 1.645\sqrt{\frac{(63/91)(28/91)}{91} + \frac{(42/79)(37/79)}{79}},$$

yielding the confidence interval $[0.039, 0.283]$.

$p_1 - p_2$

$p_1 - p_2 \in [\ ,\ ]$

Since this interval does not include zero, it seems that the first detergent is probably better than the second one for removing the type of stains in question.

Summary 1

Probability of success p of a Bernoulli experiment

- y is the number of success observed in n Bernoulli trials
- confidence level $1 - \alpha$
- two-sided confidence interval: $\frac{y}{n} \pm z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}$
- upper bound: $\frac{y}{n} + z_{\alpha}\sqrt{\frac{(y/n)(1-y/n)}{n}}$
- lower-bound: $\frac{y}{n} - z_{\alpha}\sqrt{\frac{(y/n)(1-y/n)}{n}}$

Summary 2

Difference $p_1 - p_2$ between the probability of success of two Bernoulli experiments

- y_1 is the number of success observed in n_1 Bernoulli trials of type 1
- y_2 is the number of success observed in n_2 Bernoulli trials of type 2
- confidence level $1 - \alpha$
- two-sided confidence interval:

$$\frac{y_1}{n_1} - \frac{y_2}{n_2} \pm z_{\alpha/2} \sqrt{\frac{(y_1/n_1)\,(1 - y_1/n_1)}{n_1} + \frac{(y_2/n_2)\,(1 - y_2/n_2)}{n_2}}$$

- upper bound:

$$\frac{y_1}{n_1} - \frac{y_2}{n_2} + z_{\alpha} \sqrt{\frac{(y_1/n_1)\,(1 - y_1/n_1)}{n_1} + \frac{(y_2/n_2)\,(1 - y_2/n_2)}{n_2}}$$

- lower-bound:

$$\frac{y_1}{n_1} - \frac{y_2}{n_2} - z_{\alpha} \sqrt{\frac{(y_1/n_1)\,(1 - y_1/n_1)}{n_1} + \frac{(y_2/n_2)\,(1 - y_2/n_2)}{n_2}}$$

Math 4501 - Probability and Statistics II

7.4 - Sample size

Overview: confidence intervals for the mean → Sec 7.1

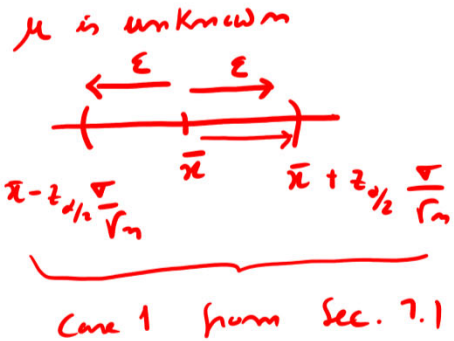
We will see how to select the sample size in order to estimate the unknown mean of a normal distribution or population to a given degree of accuracy.

Confidence intervals for the mean: known variance

Problem: Choose the sample size n so that the $100(1-\alpha)\%$ confidence interval for the mean μ of a normal distribution, given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad (\sigma \text{ is known}),$$

is not longer than $\bar{x} \pm \varepsilon$, for some fixed $\varepsilon > 0$.



Case 1 from Sec. 7.1

Solution: Solve

$$\frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \varepsilon$$

for n to get

$$n \geq \frac{z_{\alpha/2}^2\sigma^2}{\varepsilon^2}.$$

and then take n at least as large as the smallest integer greater than $\frac{z_{\alpha/2}^2\sigma^2}{\varepsilon^2}$.

Note: We call $\varepsilon = z_{\alpha/2}(\sigma/\sqrt{n})$ the *maximum error of the estimate*.

Example

A sample is to be taken from a $N(\mu, 15^2)$ population to construct a 95% confidence interval for the mean μ.

How large should the sample be so that the maximum error of the estimate provided by such confidence interval is at most 1.

Using that $z_{0.025} = 1.96$, we find that the interval is given by $\bar{x} \pm 1.96(15/\sqrt{n})$.

Solve

$$1.96\left(\frac{15}{\sqrt{n}}\right) \leq 1$$

for n to get

$$\sqrt{n} \geq 29.4, \quad \text{and thus} \quad n \geq 864.36$$

Take $n = 865$ because n must be an integer.

If a sample of size $n = 865$ is deemed too large for practical purposes, recalling that

$$n = \frac{z^2_{\alpha/2}\sigma^2}{\varepsilon^2}$$

we see that the sample size may be reduced by

- lowering the confidence level $1 - \alpha$
- increasing the maximum error of the estimate ε

Set, for instance, $1 - \alpha = 0.8$ and $\varepsilon = 2$ in the previous example.

Solving

$$1.282\left(\frac{15}{\sqrt{n}}\right) \leq 2$$

for n, we get

$$\sqrt{n} \geq 9.615, \quad \text{so that} \quad n \geq 92.4$$

Take n to be at least 93.

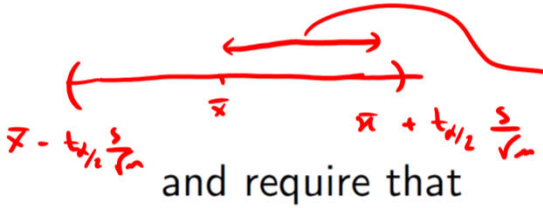
The case of unknown variance (a bit more realistic!)

If the variance σ^2 is not known, we first take a preliminary sample to obtain the point estimate s^2.

↑ with a sample size of our choice!

We then determine the $100(1-\alpha)\%$ confidence interval for μ, given by

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\,, \qquad \leftarrow \text{Sec. 7.1}$$



and require that

$$t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \leq \varepsilon\,,$$

ε : maximum error

that is, n must be such that

solve for n

$$\frac{n}{(t_{\alpha/2}(n-1))^2} \geq \frac{s^2}{\varepsilon^2}\,.$$

To find the least value of n satisfying the inequality

$$\frac{n}{(t_{\alpha/2}(n-1))^2} \geq \frac{s^2}{\varepsilon^2}$$

we start by observing that

$$t_{\alpha/2}(n-1) > z_{\alpha/2}$$

because t-distribution has larger tails than standard norm.

for all $n \in N$.

Hence, the set of values of n satisfying the desired condition is bounded below by

$$n_{\text{low}} = \frac{z_{\alpha/2}^2 s^2}{\varepsilon^2} .$$

tells us where we should start our trial and error to solve

Upon determining the value n_{low} we may gradually increase it until finding (by trial and error) the least value of $n \geq n_{\text{low}}$ for which

$$\frac{n}{(t_{\alpha/2}(n-1))^2} \geq \frac{s^2}{\varepsilon^2}$$

holds.