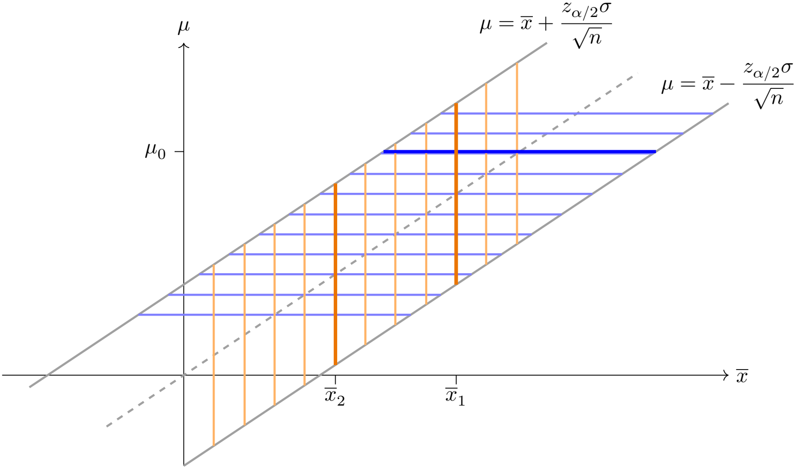
Confidence Intervals II
18.05 Spring 2022



Announcements/Agenda

Announcements

- None

Agenda

- R Quiz info.
- Polling: estimating θ in Bernoulli(θ).
- CLT $\Rightarrow$ large sample confidence intervals for the mean.
- Three views of confidence intervals.
- Pivoting: constructing confidence intervals and non-rejection regions.

R Quiz info

R Quiz, Friday May 6 in 4-149, 3-4 PM

- Formatted like the R Studios
- Open internet, open notes (no communication with other sentient beings).
- Simple calculation
- Simple plotting
- Standard statistics: mean, variance, quantiles, etc.
- Standard distributions: `dnorm()`, `pnorm()`, `dexp()`, `...`
- Simulation: `sample()`, `rnorm()`, `...`
- Standard tests
- Bayesian updating
- At least one problem that requires R help and/or Google.

Polling confidence interval

Also called a binomial proportion confidence interval

Polling means sampling from a Bernoulli(θ) distribution,

i.e. data $x_1, \ldots, x_n \sim$ Bernoulli(θ).

- Conservative normal confidence interval for θ:

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}$$

  Proof uses the CLT and the observation $\sigma = \sqrt{\theta(1-\theta)} \leq 1/2$.

- Rule-of-thumb 95% confidence interval for θ:

$$\overline{x} \pm \frac{1}{\sqrt{n}}$$

  (Reason: $z_{0.025} \approx 2$.)

Binomial proportion confidence intervals

Political polls often give a margin-of-error of $\pm 1/\sqrt{n}$, i.e. they use the rule-of-thumb 95% confidence interval.

For example, if a poll reports a margin of error of ± 0.05 this means

$$\frac{1}{\sqrt{n}} = \frac{1}{20} \;\; \Rightarrow \;\; n = 400 \text{ people polled}$$

There are many types of binomial proportion confidence intervals:
https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

Board question: Confidence intervals for binomial proportion

For a poll to find the proportion θ of people supporting X we know that a $(1-\alpha)$ confidence interval for θ is given by

$$\left[\bar{x} - \frac{z_{\alpha/2}}{2\sqrt{n}},\ \bar{x} + \frac{z_{\alpha/2}}{2\sqrt{n}} \right].$$

(a) How many people would you have to poll to have a margin of error of 0.01 with 95% confidence? (You can do this in your head.)

(b) How many people would you have to poll to have a margin of error of 0.01 with 80% confidence. (You'll want R or other calculator here.)

(c) If $n = 900$, compute the 95% and 80% confidence intervals for θ.

Concept question: overnight polling

During the presidential election season, pollsters often do 'overnight polls' and report a 'margin of error' of about $\pm 4\%$.

The number of people polled is in which of the following ranges?

(a) 0 – 50
(b) 50 – 100
(c) 100 – 500
(d) 300 – 600
(e) 600 – 1000

Problems with overnight election polling

What are some of the methodological problems with overnight polls?

Problems with overnight election polling

What are some of the methodological problems with overnight polls?

- Nonrepresentative sample: who can be reached and who chooses to respond.
- Pollsters try to adjust results to take account of the difference between the sample group and the group of actual voters.
- "At the end of the day, it is important to place preelection results in the proper context. **It is easy to confuse quantification with precision** – especially nowadays, given the prominence of data and data-driven stories. But it is always important to understand what was done to collect and produce the results so poll consumers can be better informed."
  https://www.pewresearch.org/fact-tank/2021/07/21/a-conversation-about-u-s-election-polling-problems-in-2020/

Large sample confidence interval

Data $x_1, \dots, x_n$ independently drawn from a distribution that may not be normal but has finite mean and variance.

The central limit theorem says that for large n,

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \approx \quad \mathsf{N}(0, 1)$$

i.e. the sampling distribution of the studentized mean is approximately standard normal:

So for large n the $(1 - \alpha)$ confidence interval for μ is approximately

$$\left[\bar{x} - \frac{s}{\sqrt{n}} \cdot z_{\alpha/2},\ \ \bar{x} + \frac{s}{\sqrt{n}} \cdot z_{\alpha/2}\right]$$

This is called the large sample confidence interval.

Review: confidence intervals for normal data

Suppose the data $x_1, \ldots, x_n$ is drawn from $\mathrm{N}(\mu, \sigma^2)$

Confidence level $= 1 - \alpha$

- z confidence interval for the mean (σ known)

$$\left[\bar{x} - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}},\ \bar{x} + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right] \qquad \text{or} \qquad \bar{x} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$$

- t confidence interval for the mean (σ unknown)

$$\left[\bar{x} - \frac{t_{\alpha/2} \cdot s}{\sqrt{n}},\ \bar{x} + \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}\right] \qquad \text{or} \qquad \bar{x} \pm \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}$$

- χ^2 confidence interval for σ^2

$$\left[\frac{n-1}{c_{\alpha/2}} s^2,\ \frac{n-1}{c_{1-\alpha/2}} s^2\right];$$

- t and χ^2 have $n - 1$ degrees of freedom.

Three views of confidence intervals

View 1: Define/construct CI using a standardized point statistic.

View 2: Define/construct CI based on hypothesis tests.

View 3: Define CI as any interval statistic satisfying a formal mathematical property.

View 1: Using a standardized point statistic

Example. $x_1 \ldots, x_n \sim \mathsf{N}(\mu, \sigma^2)$, where σ is known.

The standardized sample mean follows a standard normal distribution.

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1)$$

Therefore:

$$P(-z_{\alpha/2} < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \mid \mu) = 1 - \alpha$$

Pivot to:

$$P(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \mid \mu) = 1 - \alpha$$

This is the $(1-\alpha)$ confidence interval for μ:

$$\left[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = \overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Think of it as $\overline{x} \pm$ error

View 1: Other standardized statistics

The t and χ^2 statistics fit this paradigm as well:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \ \sim \ t(n-1)$$

$$X^2 \ = \ \frac{(n-1)s^2}{\sigma^2} \ \sim \ \chi^2(n-1)$$

So, $-t_{\alpha/2} < \dfrac{\overline{x} - \mu}{s/\sqrt{n}} < t_{\alpha/2}$ becomes

$$\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \ < \ \mu \ < \ \overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

Likewise, $c_{1-\alpha/2} \leq \dfrac{(n-1)s^2}{\sigma^2} \leq c_{\alpha/2}$ becomes

$$\frac{n-1}{c_{\alpha/2}}\, s^2 \leq \sigma^2 \leq \frac{n-1}{c_{1-\alpha/2}}\, s^2.$$

View 2: Using hypothesis tests

Set up: Unknown parameter θ. Test statistic x.

For any value θ_0, we can run an NHST with null hypothesis

$$H_0: \ \theta \ = \ \theta_0$$

at significance level α.

Definition. Given x, the $(1-\alpha)$ confidence interval contains all θ_0 which are not rejected when they are the null hypothesis.

Definition. A type 1 CI error occurs when the confidence interval does not contain the true value of θ.

For a $1-\alpha$ confidence interval, the type 1 CI error rate is α.

Board question: pivoting: confidence intervals and non-rejection regions

This question gets at the relationship between confidence intervals and non-rejection regions.

Main point: For a sample with sample mean $\overline{x}$, the confidence interval consists of all values μ for which a NHST with null hypothesis mean $= \mu$ would not reject on seeing $\overline{x}$.

Assume we have independent data $x_1, \dots, x_n \sim N(\mu, \sigma^2)$, where μ is unknown and σ is known.

Answer the questions on the next slide.

Board question: continued

(a) For null hypothesis $\mu = \mu_0$ give the two-sided non-rejection region for significance level α.

(b) Call the data average $\overline{x}$. Give the $1-\alpha$ confidence interval for μ.

(c) Use the $\overline{x}, \mu$-plane on the next slide. Note the conveniently included guides.

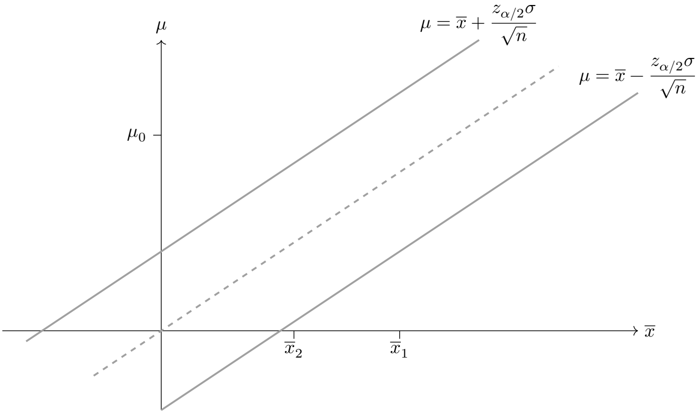
(i) Plot the horizontal line segment at height μ_0 showing the non-rejection region for $H_0 : \mu = \mu_0$ (significance level $= \alpha$).

(ii) Plot the horizontal line segment at other heights showing the non-rejection region for the corresponding μ.

(iii) Plot the vertical line segments showing the $1-\alpha$ confidence intervals around $\overline{x}_1$ and $\overline{x}_2$

(iv) Plot the vertical line segment at other values of $\overline{x}$ showing the corresponding confidence interval.

Board question axes



Understand how the main point connects with your graph.

View 3: Formal

Recall: An interval statistic is an interval I_x computed from data x.

This is a random interval because x is random.

Suppose x is drawn from $f(x|\theta)$ with unknown parameter θ.

Definition:
A $(1-\alpha)$ confidence interval for θ is an interval statistic I_x such that

$$P(I_x \text{ contains } \theta \mid \theta) \ = \ 1-\alpha$$

for all possible values of θ (and hence for the true value of θ).

Note: equality in this definition is often relaxed to $\geq$ or $\approx$.
$=$: z, t, χ^2
$\geq$: rule-of-thumb and exact binomial (polling)
$\approx$: large sample confidence interval

(See the Class 23 reading for more on this view.)