Confidence Intervals II

18.05 Spring 2018

R Quiz

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
- Use R help and google.

Agenda

- Confidence intervals using order statistics.
- CLT $\Rightarrow$ large sample confidence intervals for the mean.
- Three views of confidence intervals.
- Constructing a confidence interval without normality: the exact binomial confidence interval for θ

Some order statistics

Won't define order statistics in general, but here's an example.

- Suppose data $\{x_1, \ldots, x_n\}$ consists of real numbers.
- Define $x_{(k)} = k$th largest datum ($1 \leq k \leq n$).
- $x_{(1)} =$ smallest datum, $x_{(n)} =$ largest datum.
- $x_{((n+1)/2)} =$ median (n odd).
- Each $x_{(k)}$ is a statistic, since it's computable from the data.
- To do NHST using these statistics, we need to know how they're distributed. Of course that depends on the distribution from which the data is drawn.

Beta and order

Fact from class prep notes: If $\{x_1, \ldots, x_n\}$ are independent draws from a uniform$(0, 1)$ distribution, then the kth smallest datum $x_{(k)}$ follows a beta$(k, n-k+1)$ distribution.

Formal consequence: If $\{x_1, \ldots, x_n\}$ are independent draws from a uniform(a, b) distribution, then $(x_{(k)} - a)/(b - a)$ follows a beta$(k, n-k+1)$ distribution.

Beta-izing: The process

$$x_{(k)} \to (x_{(k)} - a)/(b - a),$$

making the order statistic $x_{(k)}$ follow beta$(k, n-k+1)$, is just like

$$\overline{x} \to z = (\overline{x} - \mu)/(\sigma\sqrt{n})$$

for making the sample mean follow a normal distribution.

Rejection regions

Under the null hypothesis that data comes from a uniform(a, b) distribution, $(x_{(k)} - a)/(b - a) \sim \text{beta}(k, n - k + 1)$.

To do a two-sided NHST, we use the critical values

$$\begin{aligned} c_{1-\alpha/2} &= \texttt{qbeta}(\alpha/2, \texttt{k}, \texttt{n} - \texttt{k} + 1), \\ c_{\alpha/2} &= \texttt{qbeta}(1 - \alpha/2, \texttt{k}, \texttt{n} - \texttt{k} + 1). \end{aligned}$$

We reject the null hypothesis if

$$(x_{(k)} - a)/(b - a) < c_{1-\alpha/2} \quad \text{or} \quad (x_{(k)} - a)/(b - a) > c_\alpha/2.$$

While there are two parameters a and b to worry about, it's complicated to talk about confidence intervals.

One parameter and a confidence interval

So suppose a **is unknown but the interval width** $w = b - a$ **is known**; that is, that our data comes from uniform$(a, a + w)$ with unknown a.

We fail to reject the null hypothesis $a = a_0$ if

$$c_{1-\alpha/2} \le (x_{(k)} - a_0)/w \le c_{\alpha/2}.$$

By pivoting as in the notes, these conditions become

$$x_{(k)} - wc_{\alpha/2} \le a_0 \le x_{(k)} - wc_{1-\alpha/2}.$$

This is our $1 - \alpha$ confidence interval for a, computed using the kth-smallest datum:

$$[x_{(k)} - wc_{\alpha/2}, x_{(k)} - wc_{1-\alpha/2}].$$

Board question: confidence interval using median

You're given seven independent random samples from uniform$(a, a+10)$, with a unknown:

$$7.08, \quad 9.48, \quad 6.13, \quad 15.93, \quad 14.39, \quad 7.52, \quad 12.87.$$

- Calculate the fourth smallest datum $x_{(4)}$.
- What estimate does $x_{(4)}$ suggest for a? (Hint: $x_{(4)} \sim a + 10 * \text{beta}(4,4)$, which has mean $a+5$.)
- Find a 90% confidence interval for a using just $x_{(4)}$.
- Some relevant values from R are

$$\texttt{qbeta}(0.05, 4, 4) = 0.225, \quad \texttt{qbeta}(0.1, 4, 4) = 0.279,$$
$$\texttt{qbeta}(0.9, 4, 4) = 0.721, \quad \texttt{qbeta}(0.95, 4, 4) = 0.775.$$

Solution

- The fourth smallest datum is $x_{(4)} = 9.48$.
- The mean of its distribution is $a + 5$, so it suggests the estimate $a \approx 9.48 - 5 = 4.48$.
- The previous slides say that $(x_{(4)} - a)/10 \sim \text{beta}(4, 7 - 4 + 1) = \text{beta}(4, 4)$. For this distribution, 5% of the probability is larger than

$$c_{0.05} = \texttt{qbeta}(0.95, 4, 4) = 0.775,$$

and 5% is smaller than

$$c_{0.95} = \texttt{qbeta}(0.05, 4, 4) = 0.225.$$

The formula for the confidence interval from the previous slides is

$$= [9.48 - 10 * (0.775), 9.48 - 10 * (0.225)]$$
$$= [1.73, 7.23].$$

Was this a clever approach?

The confidence interval for a

$$[1.73, 7.23]$$

is just what the median $x_{(4)}$ tells you.

Since the smallest datum is 6.13, and the data comes from $[a, a+10]$, you know separately that $a \leq 6.13$.

Similarly, the largest datum 15.93 tells you that $a \geq 5.93$.

So just looking at the numbers tells you for certain (under the null hypothesis) that a is in [5.93,6.13].

So this problem was a lousy way to analyze the data. The point was to work hard with confidence intervals, to try to understand them better.

Large sample confidence interval

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

A version of the central limit theorem says that large n,

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

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

$$\left[\frac{n-1}{c_{\alpha/2}} s^2,\ \frac{n-1}{c_{1-\alpha/2}} s^2\right]; \qquad \text{not symmetric around } s^2$$

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

What's wrong with this table?

n	nominal conf. $1-\alpha$	simulated conf.
20	0.95	0.936
20	0.90	0.885
50	0.95	0.944
50	0.90	0.894
100	0.95	0.947
100	0.900	0.896
400	0.950	0.949
400	0.900	0.898

Simulations for N(0, 1).

In R we (many times) drew n samples from $N(0,1)$, calculated

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

and recorded how often this interval contained zero ("simulated confidence").

Why are all simulated confidence levels smaller than calculated "nominal" ones?

Three views of confidence intervals

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

This is the cookbook mathematics we all love!

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

This is a thoughtful approach that will always work.

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

Brought to you by your friendly neighborhood formal mathematicians!

View 1: Using a standardized point statistic

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

The standardized sample mean follows a standard normal distribution.

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

Therefore:

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

Pivot to:

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

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

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

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

View 1: Other standardized statistics

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

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

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

View 2: Using hypothesis tests

Set up: Unknown parameter θ. Test statistic x.

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

$$H_0: \theta = \theta_0$$

at significance level α.

Definition. Given x, the $(1-\alpha)$ confidence interval consists of 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: exact binomial confidence interval

Use this table of binomial(8,θ) probabilities to:

1. Color the (two-sided) rejection region with significance level 0.10 for each value of θ.
2. Given $x = 7$, find the 90% confidence interval for θ.
3. Repeat for $x = 4$.

$\theta \backslash x$	0	1	2	3	4	5	6	7	8
.1	0.430	0.383	0.149	0.033	0.005	0.000	0.000	0.000	0.000
.3	0.058	0.198	0.296	0.254	0.136	0.047	0.010	0.001	0.000
.5	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
.7	0.000	0.001	0.010	0.047	0.136	0.254	0.296	0.198	0.058
.9	0.000	0.000	0.000	0.000	0.005	0.033	0.149	0.383	0.430

Solution

For each θ, the non-rejection region is blue, the rejection region is red. In each row, the rejection region has probability at most $\alpha = 0.10$.

θ/x	0	1	2	3	4	5	6	7	8
.1	0.430	0.383	0.149	0.033	0.005	0.000	0.000	0.000	0.000
.3	0.058	0.198	0.296	0.254	0.136	0.047	0.010	0.001	0.000
.5	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
.7	0.000	0.001	0.010	0.047	0.136	0.254	0.296	0.198	0.058
.9	0.000	0.000	0.000	0.000	0.005	0.033	0.149	0.383	0.430

For $x = 7$ the 90% confidence interval for p is $[0.7, 0.9]$.
These are the values of θ we wouldn't reject as null hypotheses. They are the blue entries in the $x = 7$ column.

For $x = 4$ the 90% confidence interval for p is $[0.3, 0.7]$.

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