Probabilistic Modeling and Statistical Computing Fall 2015

November 17, 2015

What is a Confidence Interval?

- Given data from a distribution with an unknown parameter θ
- Compute an interval $[L, U]$ from a sample $x_1, \ldots, x_n$ to enclose θ
- Note: L and U are random
- Desired **coverage probability**

$$\mathcal{P}(L \leq \theta \leq U) = \alpha$$

- After the sample has been observed, L, U are fixed.
- We say "with confidence α" that $L \leq \theta \leq U$.

What is a Confidence Interval NOT?

- We don't know whether $L \leq \theta \leq U$ or not.
- We cannot say that $L \leq \theta \leq U$ "with probability α". *It is either true or not.*
- We cannot say about a new observation X that $\mathcal{P}(L \leq X \leq U) = \alpha$.

Exercise 1

Observe milk production in a random sample of 50 cows. A 95% confidence interval for the mean milk production is found to be (22, 30) kg/day. **Correct or not?**

- There is a 95% chance that a cow produces, on average, between 22 and 30 kg/day.
- We are 95% confident that the sample mean is between 22 and 30 kg/day.
- The mean milk production will be 22 - 30 kg/day 95% of the time.

Exercise 1

Observe milk production in a random sample of 50 cows. A 95% confidence interval for the mean milk production is found to be (22, 30) kg/day. **Correct or not?**

- We are 95% confident that cows produce, on average, between 22 and 30 kg/day.
- In 95% of samples, the mean milk production will be 22 to 30 kg/day.

Ex.: Confidence Interval for a Mean

Consider samples of size *n* from a normal distribution $N(\mu, \sigma^2)$ with known σ^2 and unknown μ. Fix $\alpha =$ coverage probability.

- Simulate a large number of samples.
- Make a $100\alpha\%$ bootstrap CI for μ for each sample and plot it.
- Compute the fraction of cases where μ is inside the interval.
- *This should be about α - not more, not less!*

Ex.: Confidence Interval for a Mean II

Knowing the distribution of the sample, a confidence interval can be found faster.

Sample mean $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$. Therefore,

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

Let z^* be such that $\mathcal{P}(-z^* \leq Z \leq z^*) = \alpha$, if $Z \sim N(0, 1)$.

```
zstar <- qnorm((1+alpha)/2)
```

Confidence Interval for a Mean

We know that

$$\mathcal{P}\left(-z^* \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq z^*\right) = \alpha$$

since $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Rework the expression inside $\mathcal{P}()$ to obtain inequalities for μ.

$-z^* \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ becomes $\mu \leq \bar{x} + z^*\sigma/\sqrt{n}$

$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq z^*$ becomes $\mu \geq \bar{x} - z^*\sigma/\sqrt{n}$

Confidence Interval for a Mean

Confidence interval for unknown mean μ

...with given variance σ^2:

$$\bar{x} - z^*\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z^*\sigma/\sqrt{n}$$

where $\bar{x} =$ sample mean, n = sample size, and z^* is such that $\mathcal{P}(-z^* \leq Z \leq z^*) = \alpha$ for the desired coverage probability α.
$ME = z^*\sigma/\sqrt{n} =$ "margin of error"

The confidence interval becomes larger if σ increases or n decreases or α increases.

Overview of Confidence Intervals

- **Parameter of interest: $\mu =$ mean**
- Variance can be known or unknown
- Distribution can be normal or not
- **Parameter of interest: $\sigma^2 =$ variance**
- Mean can be known or unknown
- Distribution can be normal or not
- **Parameter of interest:**
  $\mu_1 - \mu_2 =$ difference of means
- **Parameter of interest: $p =$ proportion**
- **Parameter of interest:**
  $p_1 - p_2 =$ difference of proportions

CI for Mean - Unknown Variance

ME for known variance: $ME = z^* \sigma / \sqrt{n}$

If σ is unknown, replace it with the sample standard deviation s.

However, $T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ no longer has a $N(0, 1)$ distribution.

T-Distribution

The distribution of $T_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ depends only on n.

Number of degrees of freedom: $n - 1$.

CI for Mean - Unknown Variance

We know that

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T_{n-1}$$

Let t^* be such that $\mathcal{P}(-t^* \leq T_{n-1} \leq t^*) = \alpha$, if T_{n-1} has T-distribution with $n-1$ d.o.f..

```
tstar <- qt((1+alpha)/2,n-1)
```

Rework the expression inside $\mathcal{P}()$ to obtain inequalities for μ.

Confidence Interval for a Mean

Confidence interval for unknown mean μ

. . . with unknown variance:

$$\bar{x} - t^*_{n-1} s/\sqrt{n} \leq \mu \leq \bar{x} + t^*_{n-1} s/\sqrt{n}$$

where $\bar{x} =$ sample mean, n = sample size, and t^*_{n-1} is such that $\mathcal{P}(-t^*_{n-1} \leq T_{n-1} \leq t^*_{n-1}) = \alpha$ for the desired coverage probability α.
$SE = s/\sqrt{n} =$ "standard error"

The confidence interval becomes larger if σ increases or n decreases or α increases.

Exercise 3

Suppose that 20 years ago, the mean cholesterol level of adult men in A-Town was 185 mg/dL with a standard deviation of 50 mg/dL.

- Assume that $\bar{x} = 190$ for a sample size $n = 100$. Assume that σ^2 has not changed. Make a 90% CI for the mean cholesterol level.
- *You are making an additional unstated assumption. What is it?*
- What should be the sample size for a 95% CI that has ME = 8?

Non-Normal Distribution

Want to make a confidence interval for the population mean. What if the distribution is not normal?

Example: Exponential distribution, with small sample size

We could just use the formula for the CI for the mean from a normal distribution with unknown variance.

Justification: Central Limit Theorem

Problematic for highly skewed distributions and small sample size.

CI for Mean, Non-Normal Case

Confidence interval for unknown mean μ

. . . from non-normal distribution:

$$\bar{x} - t^*_{n-1} s/\sqrt{n} \leq \mu \leq \bar{x} + t^*_{n-1} s/\sqrt{n}$$

where $\bar{x} =$ sample mean, n = sample size, and $t^*_{n-1} = t^*_{n-1}(\alpha)$ is as before.

- Symmetric distribution, no multiple peaks: OK for $n \gtrsim 10$
- Highly skewed distribution: OK for $n \gtrsim 100$

One-sided Confidence Interval

Upper Confidence Bound for unknown mean μ

. . . from non-normal distribution:

$$\mu \leq \bar{x} + t^*_{n-1} s/\sqrt{n}$$

where $\bar{x} =$ sample mean, n = sample size, and $t^*_{n-1} = t^*_{n-1}(\alpha)$ is such that $\mathcal{P}(T_{n-1} \leq t^*_{n-1}) = \alpha$ for the desired coverage probability α.

```
tstar = qt(alpha, n-1)
```

Difference of Two Means

Given samples from normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, sample sizes n_1 and n_2. The test statistic is the difference of the sample means. **Here's how to work out the confidence interval:**

We know $\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ and therefore

$$\boxed{Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)}$$

Work backwards from $\mathcal{P}(-z^* \leq Z \leq z^*) = \alpha$.

CI for Difference of Two Means

CI for $\mu_1 - \mu_2$

...from normal distribution with known variances:

$$\bar{x}_1 - \bar{x}_2 - ME \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + ME$$

where $ME = z^* \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ and $z^* = z^*(\alpha)$ is such that $\mathcal{P}(-z^* \leq Z \leq z^*) = \alpha$ for the desired coverage probability α.

CI for Difference of Two Means

CI for $\mu_1 - \mu_2$

...from non-normal distribution with unknown variances:

$$\bar{x}_1 - \bar{x}_2 - ME \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + ME$$

where $ME = t^* \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, the s_i are standard deviations and t^* is such that $\mathcal{P}(-t^* \leq T \leq t^*) = \alpha$ for the desired coverage probability α.

Which T shall we use?

Proportion of Successes

Observe n trials of a Bernoulli random variable.
$X =$ Number of successes.

- Opinion poll: Should marijuana be legal?
- Opinion poll: Is the country headed in the wrong direction?
- School testing: Does a sixth-grader pass math proficiency tests?
- Online marketing: Does a Web user click through on my ad?

Maximum likelihood estimate for success probability p is $\hat{p} = \frac{X}{n}$. Make a CI for p.