Interval estimation: Part 2

(Module 4)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2020

Contents

1 Confidence intervals **1**

1.1 Less common scenarios . . . 1
1.2 General techniques . . . 3
1.3 Properties . . . 4
1.4 Choice of confidence level . . . 5
1.5 Interpretation . . . 6
1.6 Summary . . . 6

2 Prediction intervals **7**

3 Sample size determination **8**

Aims of this module

- Explain some less common scenarios where confidence intervals are used
- Describe some general aspects of confidence intervals
- Introduce **prediction intervals**, an interval estimator in the context of predicting a future value
- Explain how to calculate the sample size required for a study

1 Confidence intervals

1.1 Less common scenarios

Less common scenarios: overview

- One-sided CIs
- CIs based on discrete statistics

One-sided confidence intervals

We can construct one-sided confidence intervals, e.g. just an upper or lower bound.

For example, if we sample from $\mathrm{N}(\mu, \sigma^2)$ with known σ:

$$\Pr\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < c\right) = 1-\alpha$$

where $c = \Phi^{-1}(1-\alpha)$. Rearranging gives

$$\Pr\left(\bar{X} - c\frac{\sigma}{\sqrt{n}} < \mu\right) = 1-\alpha$$

and therefore a *one-sided* $100 \cdot (1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{x} - c\frac{\sigma}{\sqrt{n}}, \infty\right).$$

Remarks

- The main thing to remember is to start with a one-sided probability statement about the pivot.
- In this example, we obtained a lower bound.
- To get an upper bound, start with an inequality in the other direction.
- Other scenarios are analogous. For example, if σ is unknown then replace σ with s and let c be a quantile from t_{n-1} rather than $N(0,1)$.
- Since we only need one tail probability, we don't need to separate α into two parts. That's why we use the $1-\alpha$ quantile here rather than the $1-\alpha/2$ quantile.

Example (one-sided interval)

A winemaker requires a minimum concentration of 10 g/L of sugar in the grapes used to make a certain wine. In a sample of 30 units she finds an average concentration of 11.9 g/L and a standard deviation of 0.96. Is that high enough?

She calculates a 95% lower bound (one-sided CI) as follows:

$$\bar{x} - c\frac{s}{\sqrt{n}} = 11.9 - 1.699 \times \frac{0.96}{\sqrt{30}} = 11.60$$

where $c = 1.699$ is the 0.95 quantile from t_{29}.

On that basis, she is confident that the average sugar content is adequately high.

Example using R

Recall the butterfat example from the previous module. Now re-doing using one-sided CIs...

```
> t.test(butterfat,
+        conf.level  = 0.90,
+        alternative = "less")

...

90 percent confidence interval:
    -Inf 534.146
```

```
> t.test(butterfat,
+        conf.level  = 0.90,
+        alternative = "greater")

...

90 percent confidence interval:
 480.854     Inf
```

Confidence intervals based on discrete statistics*

Our starting point has been probability intervals like:

$$\Pr\left(a(\theta) < T < b(\theta)\right) = 0.95$$

What if T is discrete? For example, $T \sim \text{Bi}(n, \theta)$

Limitation: $a()$ and $b()$ can only take specific (discrete) values.

$\Rightarrow$ Cannot guarantee an exact probability (confidence level).

$\Rightarrow$ Inversion is messy.

Usually aim for something close, with 'at least' probability. For example,

$$\Pr\left(a(\theta) \leqslant T \leqslant b(\theta)\right) \geqslant 0.95$$

where:

- $a(\theta)$ is the largest value of x such that $\Pr(x \leqslant T \mid \theta) \geqslant 0.975$
- $b(\theta)$ is the smallest value of x such that $\Pr(T \leqslant x \mid \theta) \geqslant 0.975$

How do we invert these?

For an observed value t_{obs} (of T), we have:

- c is such that $\Pr(t_{\text{obs}} \leqslant T \mid \theta = c) = 0.025$
- d is such that $\Pr(T \leqslant t_{\text{obs}} \mid \theta = d) = 0.025$

Then, the 'at least' 95% confidence interval is (c, d).

1.2 General techniques

CIs from MLEs

Maximum likelihood estimators have many convenient properties. We will cover some of the theory later in the semester. For now, it is useful to know the following...

Let,

$$V(\theta) = -\frac{\partial^2 \ln L}{\partial \theta^2}$$

This is known as the *observed information function*. It can be used to estimate the standard deviation of the MLE:

$$\text{se}(\hat{\theta}) = \frac{1}{\sqrt{V(\hat{\theta})}}$$

Moreover, the MLE is **asymptotically unbiased** and **asymptotically normally distributed**.

Therefore, for large sample sizes, we can construct approximate CIs using:

$$\hat{\theta} \pm \frac{c}{\sqrt{V(\hat{\theta})}}$$

where $c = \Phi^{-1}(1 - \alpha/2)$.

Example (approximate CI from MLE)

Sampling (iid) from: $X \sim \text{Exp}(\theta)$. Previously we found that $\hat{\theta} = \bar{X}$ and

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2}.$$

Differentiate once more,

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{n}{\theta^2} - \frac{2\sum x_i}{\theta^3}$$

and so we have,

$$\text{se}(\hat{\theta}) = \left(-\frac{n}{\hat{\theta}^2} + \frac{2\sum x_i}{\hat{\theta}^3}\right)^{-\frac{1}{2}}$$

and an approximate 95% confidence interval is given by $\hat{\theta} \pm 1.96\,\text{se}(\hat{\theta})$.

Review of general methods for constructing CIs

Methods:

- Invert a probability interval based on a known sampling distribution (use a pivot)
- Use the asymptotic MLE result

Common approximations:

- Normality (based on the CLT or the asymptotic MLE)
- Substitute parameter estimates into the expression for the standard deviation of the estimator

1.3 Properties

CIs are random intervals

Recall: the CI estimator is a **random interval**.

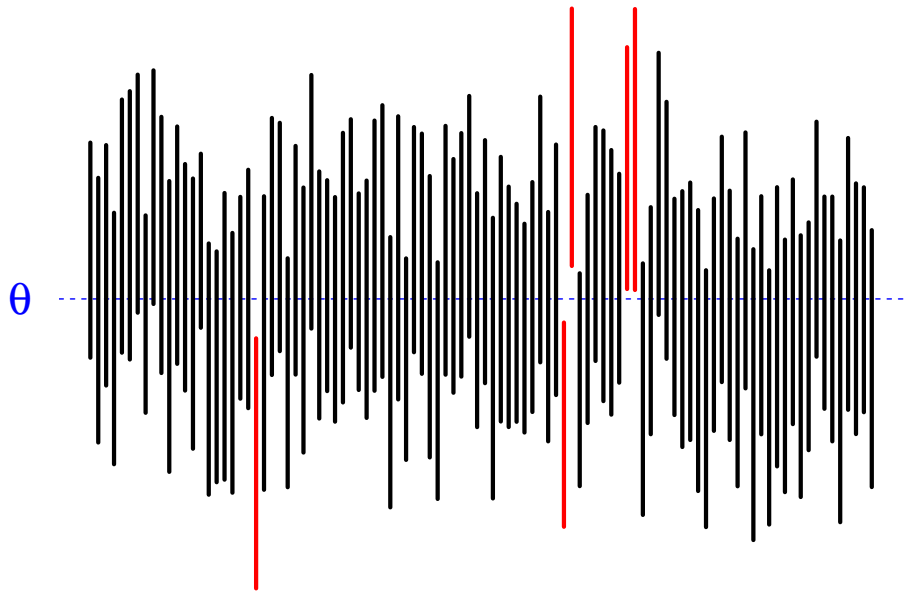
A CI consists of two statistics: the lower bound and the upper bound of the interval. They both have sampling distributions.

The random elements are therefore the endpoints, not the parameter:

$$\Pr(\mathbf{L} < \theta < \mathbf{U}) = 0.95$$

Contrast this with a probability statement for a statistic:

$$\Pr(l < \mathbf{T} < u) = 0.95$$



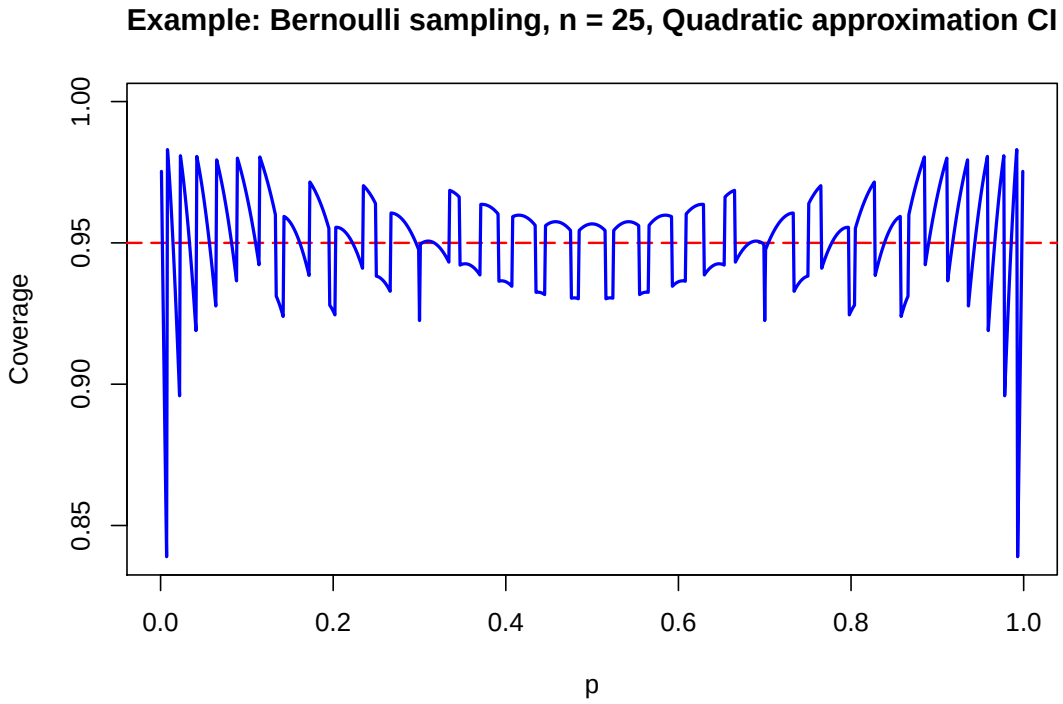
Coverage

The *coverage* or *coverage probability* of a confidence interval (estimator) is the probability it contains the true value of the parameter,

$$C = \Pr(L < \theta < U)$$

Usually this is equal to the confidence level, which is also known as the *nominal coverage probability*.

However, due to various approximations we use, the actual coverage achieved may vary from the confidence level.



More detail about the quadratic approximation will be shown in the tutorials and lab classes.

1.4 Choice of confidence level

Choice of confidence level

This is somewhat arbitrary.

If very high:

- More likely to capture the true value.
- Impractically wide: won't act as a useful guide for showing plausible values based on the data.
- It will place too much emphasis on tails of the sampling distribution, which aren't actually all that likely.

If very low:

- More 'useful' in the sense of being more selective about the possible values of the parameter.
- This comes at the expense of the loss of 'confidence', i.e. not as certain about whether the true value is captured inside the interval.

Choice of confidence level: some guidelines

- **95%** is a very common convention. If you follow this, it will rarely be questioned. Others may be expecting this, so always be clear if you deviate from it.
- **90%** can also be a reasonable choice.
- **50%** is sometimes useful, due to easy interpretation. A good use case: plotting a large number of overlapping intervals, to reduce visual clutter.
- The choice can **vary by application**, and you may even use different choices for the same problem (e.g. 50% for a particular plot, but 95% when reporting a headline result in text).
- Whatever you choose, remember that the true value is **never guaranteed** to be inside the interval. There is **always** a chance it will be outside.

1.5 Interpretation

Explaining CIs

The probability associated with a CI (i.e. the confidence level) relates to the **sampling procedure**. In particular, it refers to **hypothetical repeated samples**.

Once a specific sample is observed and a CI is calculated, the confidence level **cannot** be interpreted probabilistically in the context of the specific data at hand.

It is incorrect to say things like:

- This CI has a 95% chance of including the true value
- We can be '95% confident' that this CI includes the true value

Don't do it!

The probability only has a meaning when considering potential replications of the whole sampling and estimation procedure.

We can only say something like:

- If we were to repeat this experiment, then 95% of the time the CI we calculate will cover the true value.

(This is a bit of a mouthful...)

In practice:

- If you are reporting results to people who know what they are, you can just state that the "95% confidence interval is..."
- If people want to know what this means, use an intuitive notion like, "it is the set of plausible values of the parameter that are consistent with the data". (Note: this is not actually true in general, but will be accurate enough for all of the examples we cover this semester.)
- If you need to actually explain what a CI is precisely, you need to explain it in terms of repeated sampling. (No shortcuts!)

Communicating results: general tips

- Describe the extent of your uncertainty
- Emphasise a range a plausible values
- Phrase results in terms of the **degree of evidence** (e.g. 'strong/modest/weak evidence of...')

1.6 Summary

Confidence intervals: summary

- Interval estimates are the most common way to quantify uncertainty.
- Confidence intervals are the most common type of interval estimate.
- Confidence intervals are straightforward to construct if we know or can approximate the sampling distribution of the statistic and can construct a pivot.
- We have looked at some well known (and widely used) examples for means, variances and proportions.
- We can derive CIs, whether exact or approximate, for a variety of scenarios, and have techniques for constructing them in general.
- 95% CIs are the most common convention.

2 Prediction intervals

Prediction intervals

Suppose we want to estimate the value of a *future observation*, rather than a parameter of the distribution. We usually call this *prediction* rather than 'estimation'.

We have available data that arose from the same probability distribution. Can we use this to come up with an interval estimate?

Yes. Easiest to see with an example...

Example (prediction interval)

Random sample (iid): $X_1, \ldots, X_n$ on $X \sim \mathrm{N}(\mu, 1)$

Let X^* be a future observation on X, independent of those currently observed.

By independence, we have:

$$\begin{aligned} \bar{X} &\sim \mathrm{N}(\mu, \frac{1}{n}) \\ X^* &\sim \mathrm{N}(\mu, 1) \\ \bar{X} - X^* &\sim \mathrm{N}(0, 1 + \frac{1}{n}) \end{aligned}$$

Therefore we can write,

$$\Pr\left(-1.96\sqrt{1+\frac{1}{n}} < \bar{X} - X^* < 1.96\sqrt{1+\frac{1}{n}}\right) = 0.95$$

$$\Pr\left(\bar{X} - 1.96\sqrt{1+\frac{1}{n}} < X^* < \bar{X} + 1.96\sqrt{1+\frac{1}{n}}\right) = 0.95$$

From this we get a *95% prediction interval* for X^*,

$$\bar{x} \pm 1.96\sqrt{1+\frac{1}{n}}$$

Compare this with the 95% confidence interval for μ,

$$\bar{x} \pm 1.96\sqrt{\frac{1}{n}}$$

Remarks

- The prediction interval for X is much wider than the confidence interval for μ.
- As $n \to \infty$, the width of the confidence interval shrinks to zero, but the width of the prediction interval tends to the width of the corresponding population probability interval ($\mu \pm 1.96$).
- This makes sense: we get complete certainty about μ, but each observation on X has inherent variability (in this case, a variance of 1).
- In the prediction interval estimator, all quantities are random variables:

$$\Pr(\mathbf{L} < \mathbf{X} < \mathbf{U}) = 0.95$$

3 Sample size determination

Sample size determination: overview

We are planning a study. How much data do we need?

It depends on how much precision is required, often measured by the desired width of a confidence interval.

We will go through two estimation scenarios:

- Means
- Proportions

Example: sample size for means

Random sample (iid): $X_1, \ldots, X_n \sim \mathrm{N}(\mu, 15^2)$

Want a 95% confidence interval of width 2 (i.e. $\bar{x} \pm 1$).

The confidence interval will be given by $\bar{x} \pm 1.96\frac{15}{\sqrt{n}}$.

So we need,

$$1.96\frac{15}{\sqrt{n}} = 1$$

which gives

$$\sqrt{n} = 29.4, \quad \text{or} \quad n \approx 864.36$$

and so for our study we need sample size of at least **865**.

Sample size for means

Confidence interval of the form:

$$\bar{x} \pm c\frac{\sigma}{\sqrt{n}} = \bar{x} \pm \epsilon$$

where $c = \Phi^{-1}(1 - \alpha/2)$.

For a prespecified ϵ, we have:

$$\epsilon = c\frac{\sigma}{\sqrt{n}}, \quad \text{or} \quad n = \left(\frac{c\sigma}{\epsilon}\right)^2$$

Example 2: sample size for means

A researcher plans to select a sample of first-grade girls in order to estimate their mean height μ. The sample is required to be large enough to get an estimate to within 0.5 cm. From previous studies we know $\sigma \approx 2.8$ cm.

$$n = \left(\frac{c\sigma}{\epsilon}\right)^2 = \left(\frac{1.96 \times 2.8}{0.5}\right)^2 = 120.47$$

The researcher selects 121 girls.

Sample size for proportions

Confidence interval is of the form:

$$\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $c = \Phi^{-1}(1 - \alpha/2)$. In order to have $\hat{p} \pm \epsilon$ for a given ϵ, we need:

$$\epsilon = c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \text{or} \quad n = \frac{c^2\,\hat{p}(1-\hat{p})}{\epsilon^2}$$

Can use a preliminary estimate of $\hat{p}$ if this is available. Otherwise, note that $\hat{p}(1-\hat{p}) \leqslant \frac{1}{4}$, which means we can use $n = c^2/(4\epsilon^2)$ as a conservative choice.

Example: Sample size for proportions

The unemployment rate has been 8% for a while. A researcher wishes to take new sample to estimate it and wants to be 'very certain', by using a 99% CI, that the new estimate is within 0.001 of true proportion.

$$n = \frac{c^2\,\hat{p}(1-\hat{p})}{\epsilon^2} = \frac{2.576^2 \times 0.08 \times (1-0.08)}{0.001^2} \approx 488394$$

At this stage the researcher panics and says they don't really need to be that sure!

Try again... a 98% CI and a difference of 0.01 gives $n = 3982$, which is more practical, although possibly still a bit large.