Statistical Computing
Week 10 — Confidence Intervals
6 December 2019

Confidence Intervals: Contents

- Practical: Computing a confidence Interval
- Confidence level
- t-Distribution
- Properties of a confidence Interval
- Interpretation of a confidence Interval

Interval estimation

Suppose we have an iid random sample $X_1, X_2, \ldots, X_n$, sampled from a population with expectation $E(X) = \mu$ and variance $Var(X) = \sigma^2$.

We do not know μ or σ^2, but we can estimate them.

We can estimate μ using $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and the central limit theorem tells us that this estimate centres in on the true expectation.
This is called a **point estimate**.

A sensible question is: "How good is this point estimate?"
We don't expect that our estimate is exactly right, but how near can we expect it to be?

We answer this by defining an "**interval estimate**". We have confidence that the true value of μ lies in this interval, but we cannot be certain that it does.

Practical

You will each collect a random sample of 9 Body Mass Index (BMI) values and calculate your own confidence interval (CI) for the expected BMI value. We will then compare and discuss the results.

To save time: you will use R-Studio to *simulate* the BMI values!

Download the R Data file `my.random.sample.Rda` which contains the function `my.rs`.

Load the file into R using

```
load("my.random.sample.Rda")
```


1) Obtain the random sample and assign it to the object called `x`

   ```
   > x<-my.rs()
   ```

2) Use RStudio to calculate the BMI sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
3) Calculate the standard deviation of the sample s_x (tip: `sd()`).
4) Calculate the lower limit of the confidence interval. $C_1 = \overline{x} - 1.86\frac{s_x}{\sqrt{n}}$.
5) Calculate the upper limit of the confidence interval. $C_2 = \overline{x} + 1.86\frac{s_x}{\sqrt{n}}$.
6) Express your confidence interval in the form $[C_1; C_2]$.

The lecturer has already done this and has calculated his confidence interval (next slide).

An explanation of where 1.86 comes from is explained later in the lecture.

The lecturer's random sample is:

$x_1 = 19.73$, $x_2 = 21.73$, $x_3 = 25.97$, $x_4 = 15.87$, $x_5 = 26.4$

$x_6 = 22.59$, $x_7 = 27.71$, $x_8 = 27.58$, $x_9 = 22.75$

My random sample is:

$x_1 =$ $x_2 =$ $x_3 =$ $x_4 =$ $x_5 =$

$x_6 =$ $x_7 =$ $x_8 =$ $x_9 =$

	Lecturer's	Mine
Sample mean $\overline{x}$	23.37	
Sample standard deviation s_x	3.96	
Lower limit: $C_1 = \overline{x} - 1.86\frac{s_x}{\sqrt{n}}$	20.91	
Upper limit: $C_2 = \overline{x} + 1.86\frac{s_x}{\sqrt{n}}$	25.83	

The Lecturer's confidence interval is $[20.91; 25.83]$

My confidence interval is

$$\left[\quad ; \quad \right].$$

I'm confident that the mean BMI of the population (expectation, μ) lies between

_____ and _____.

The true value of μ

Because this data was simulated, we know the true value of μ is equal to 24.0.
This would not be the case, had we actually collected real data. Simulated data allow us to assess our confidence intervals.

Each CI which contains 24.0 is a “hit” and
Each CI which does not contain 24.0 is a “miss”

In the practical:
the number of hits was

and the number of misses was

The coefficient 1.86 was chosen so that approximately 90% of the CIs are hits.

Confidence level

The 90% value on the previous slide is called the **confidence level** and is written as $1 - \alpha$.
$\alpha = 0.1$ and is called the significance level and is relevant for hypothesis testing.

The confidence level measures how “confident” we are that the CI is a hit.
The probability that the CI contains the true expectation μ.

The standard value for the confidence level is 95%, but confidence levels of 90% and 99% are also common.

If the confidence interval is a “hit”, it has the property

$$\begin{aligned} & \overline{X} - c_\alpha \leqslant \mu \leqslant \overline{X} + c_\alpha \\ \Leftrightarrow \quad & c_\alpha \geqslant \overline{X} - \mu \geqslant -c_\alpha && \mid \times (-1) \\ \Leftrightarrow \quad & \mu - c_\alpha \leqslant \overline{X} \leqslant \mu + c_\alpha && \mid + \mu \end{aligned}$$

c_α is chosen so that

$$1 - \alpha = P(\mu - c_\alpha \leqslant \overline{X} \leqslant \mu + c_\alpha)$$

$$1 - \alpha = P(-c_\alpha \leqslant \overline{X} - \mu \leqslant c_\alpha)$$

t-Distribution

Let $X_1, X_2, \ldots, X_n$ be $N(\mu, \sigma^2)$ **iid** random variables, then $\overline{X}$ is also normally distributed with mean μ and variance σ^2.

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\overline{X} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

There’s a problem here: we don’t know σ!

In some exceptional circumstances σ is known, in which case you can use the *normal distribution* to find c_α, but we will concentrate on the more realistic scenario where σ is unknown.

We can estimate σ using S_X, as we did in the practical.

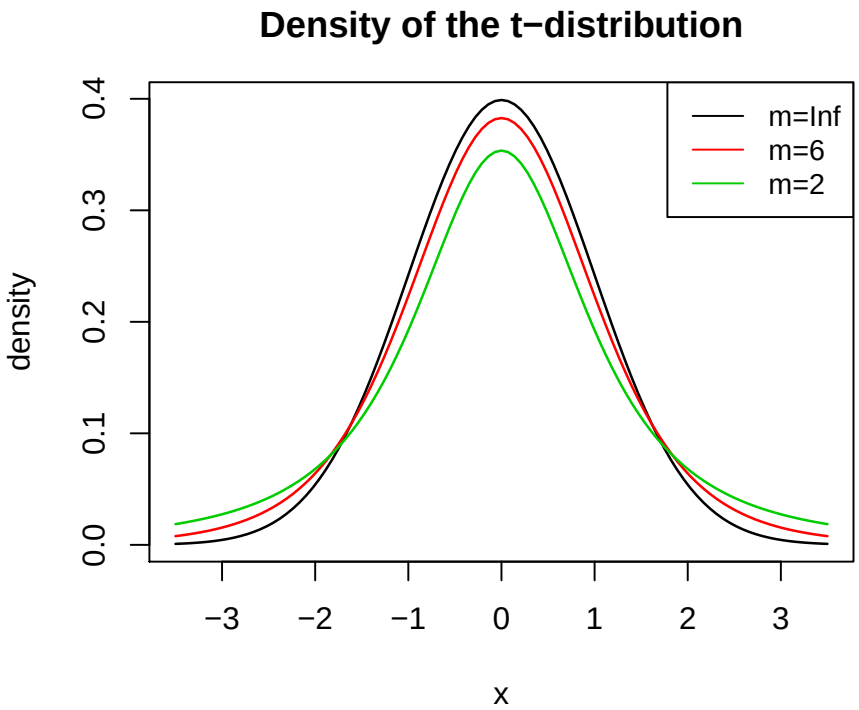
When we estimate σ, then the normal distribution property above is no longer true.

We use the theorem that

$$\frac{\overline{X} - \mu}{S_X}\sqrt{n} \sim t_{n-1},$$

where t_m is the t-distribution with parameter m (the parameter has the odd name “degrees of freedom”).

The density (pdf) of the t-distribution looks similar to the normal distribution but has heavier tails, especially for small m.



Approximation:
For “large” m (> 30) the t distribution is approximately a $N(0, 1)$ distribution.

When $1 - \alpha = 90\%$

$$0.9 = P\left(-c_\alpha \leqslant \frac{\overline{X} - \mu}{S_X}\sqrt{n} \leqslant c_\alpha\right)$$

The 0.95-Quantile of the t_{n-1} distribution gives the appropriate coefficient: $t_{n-1;0.95} = 1.86$

```
 > qt(0.95,8)
[1] 1.859548
```

Diagram

Summary: confidence interval for μ

A $1 - \alpha$ confidence interval for μ can be specified using:

$$\left[\overline{X} - t_{n-1,1-\alpha/2}\frac{S_X}{\sqrt{n}}\ ;\ \overline{X} + t_{n-1,1-\alpha/2}\frac{S_X}{\sqrt{n}}\right]$$

- n is the sample size
- $\overline{X}$ is the sample mean
- $t_{n-1,1-\alpha/2}$ is the relevant t-distribution quantile
- S_X is the sample standard deviation.

The confidence level $1 - \alpha$ should be specified by the data analyst before the interval is calculated.

Properties of a confidence interval

A $1-\alpha$ confidence interval for μ has the general form

$$\left[\overline{X} - c_\alpha \; ; \; \overline{X} + c_\alpha\right],$$

so the width (or length) of the confidence interval is $2c_\alpha$.

A narrower interval implies a more precise estimate for μ.
What are the factors that can influence the interval width?

Consider the half length of the interval, $c_\alpha = t_{n-1,1-\alpha/2}\dfrac{S_X}{\sqrt{n}}$.

- When the sample size n is larger then the interval width is smaller and the estimate more precise. It can be costly (in time or money) to increase the sample size.
- A smaller value of α means that we have more confidence that the interval contains the correct parameter, but the interval is wider.
  A 100% CI is $(-\infty; \infty)$, which is uninformative.
  A 0% CI is $[\overline{X}; \overline{X}]$: very precise but we have no confidence in this at all.
  We do not believe that the population mean is *exactly* the same as the sample mean.
- The smaller the sample standard deviation S_X, the narrower the interval so the estimate is more precise. However S_X is an estimate for σ, a value over which we have no influence, as σ is a naturally occurring parameter.

Confidence interval: interpretation

The formal interpretation of a CI is not easy.

The informal interpretation is: “We believe from the given data that the true parameter value lies in the confidence interval.”

We need be more exact about the meaning of a CI. We will assume the CI is for an expectation μ and has confidence level 95%.

Wrong!!!

- The confidence interval contains μ
- μ lies between C_1 and C_2.

Good: a day-to-day interpretation

- The probability that the CI contains μ is 95%.

  The problem with this explanation is that after we have collected the sample, the data are no longer random!
  The FIFA world cup winners for 2022 is random.
  The FIFA world cup winners for 2018 is no longer random.

Perfect: The exact interpretation

- *The probability that a future CI contains μ is 95%*.
- If we were to repeat this random sampling method infinitely often then 95% of the resulting CIs would contain μ.

To avoid using the inelegant *exact* wording, statisticians use the word *confidence*:

“We are 95 % confident that CI contains μ.”

But this is no more than statistical jargon for the exact interpretation.

Example

A fruit juice producer has a machine which fills litre bottles with apple juice. The machine does not fill each bottle with exactly 1000 ml but a random amount that is approximately 1 litre.

The manager knows that the machine, when operating correctly, fills each bottle with an expected volume of 1000 ml and standard deviation of 10 ml.

A sample of 20 bottles is taken and the apple juice volume is measured.

Exercise

The sample mean of the 20 bottles is 998 ml and the sample standard deviation is 11 ml. Construct a confidence interval with confidence level 95%, i.e., $\alpha = 0.05$.

Hint: $\left[\overline{x} \pm t\frac{s_x}{\sqrt{n}}\right]$ and use the R command `qt(1-alpha/2,n-1)` to obtain the t coefficient.

$$t = 2.09$$

The expected volume lies between ml and ml with 95% confidence.

Exercise: Interpretation

Suppose the manager takes a sample of 20 bottles each week, to check that the machine is working correctly.

Each week there is a 95% chance that the CI is a hit, contains the correct expectated volume.

After 20 weeks, on average 19 out of 20 samples will be hits and one will be a miss.

After 100 weeks (approx 2 years), on average 95 out of 100 samples will be hits and five will be a miss.

etc. usw.

Next week the probability of a hit is 95%

In R

The command in R to find the limits of a CI for an expectation μ is called `t.test`. You will learn what a *t*-test is next week, this function does both because the are related.

```
> t.test(x,conf.level=0.9)

One Sample t-test

data:  x
t = 17.688, df = 8, p-value = 1.067e-07
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 20.91315 25.82685
sample estimates:
mean of x
    23.37
```

► Type this now to check your CI from earlier.
The default value of `conf.level` is 0.95.

Other types of confidence interval

We have focussed on just one type of confidence interval: a CI for an expectation μ.

There are many other types of confidence intervals which have similar underlying theory.

Some examples which you will probably come across soon are:

- A confidence interval for a proportion.
- A confidence interval for the population variance.
- A confidence interval for the difference between the mean in two similar populations.
- A confidence interval for a regression parameter.

Exercises

There is no Worksheet this week.
Work through these two exercises either at the end of the lecture, or at home.

Exercise 1 The systolic blood pressure of 21 women participating in a keep fit class gave the following statistics: $\overline{x} = 128.52$ and $s_x = 14.31$.
Specify a 99% CI for the expected population systolic blood pressure.

Exercise 2 Write an R function which takes a vector of values and outputs a vector of length 2 with the two limits of a 95% CI. Check your code by comparing the output with `t.test`.

Adapt the function so that the user can provide the confidence level as an argument with default value 0.95.