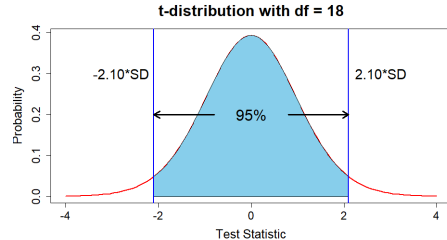
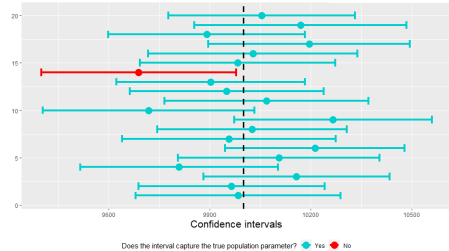


Introduction to Confidence Interval and its Calculations in R



Outline

- 1 Confidence Interval
- 2 Calculating Confidence Intervals in R
- 3 Summary

Learning Objectives

In this video, we will

- Introduce another approach to statistical inference: ***Confidence interval***.
- Use some examples to teach you how to compute the confidence intervals in R.

Confidence Interval

Confidence Interval

An Example

What is the population average monthly household income of Singaporeans?

- The exact answer to the question is referred to as the **population parameter**.
- Suppose we have randomly selected 100 households from a list of all households in Singapore.
- The sample mean, 8000 dollars, is a **point estimate** of the population parameter.
- We would like to know how far the estimate is from the true, unknown parameter.



What is a confidence interval?

Definition

Confidence Interval is a range of plausible values such that we are reasonably certain the interval contains the true population parameter, with an associated confidence level.

- The default confidence level is often chosen as 95%.
- Intuitively speaking, a point estimate is like fishing with a spear, while a confidence interval is like fishing with a net.
- A confidence interval shortlists the plausible values that most likely capture the true parameter.

How do we construct a confidence interval?

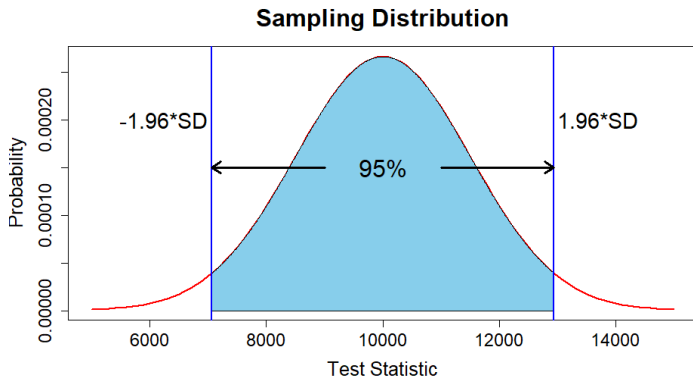
A Template Formula:

$$\begin{aligned}\text{Confidence Interval} &= \text{Point Estimate} \pm \text{Margin of Error} \\ &= \text{Point Estimate} \pm \text{Critical Value} * \text{Standard Error}\end{aligned}$$

- Confidence intervals are closely related to classical hypothesis tests.
- For both p-values and confidence intervals, it is important to know the sampling distribution of the test statistic.
- The critical value is a quantile of the sampling distribution, which also corresponds to the confidence level that we desire.

Construct the Confidence Interval for the Example

- The point estimate is the sample mean, 8,000.
- The critical value comes from a normal distribution, and it is typically taken as 1.96.
- The standard error is equal to the standard deviation of the sampling distribution. Suppose the standard error is 1500.
- $95\%CI = 8000 \pm 1.96 * 1500 = 8000 \pm 2940 = (5060, 10940)$.



How do we interpret a 95% confidence interval?

Correct Interpretation:

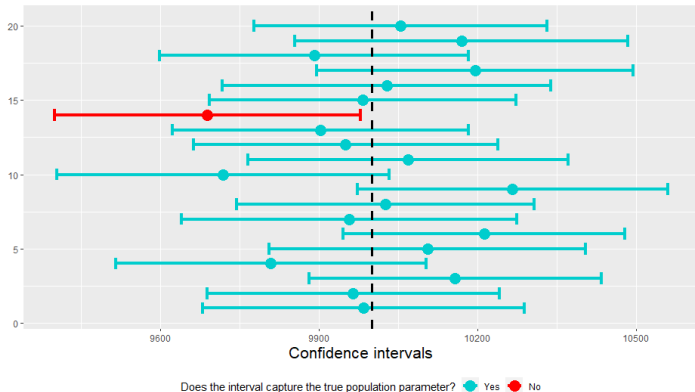
We are 95% confident that the interval contains the population parameter.

- It is inaccurate to say that there is a 95% chance that the population parameter lies within the confidence interval.
- In fact, since the population parameter is fixed and usually unknown, there are only two possibilities:
 - 1 the interval includes the population parameter;
 - 2 the interval does not include the population parameter.

How do we interpret a 95% confidence interval?

Cont'd

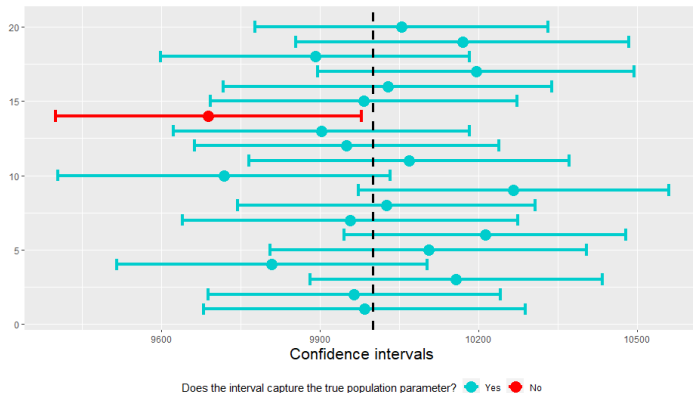
- If we have repeatedly taken the samples 100 times, using the same procedure, around 95 of the calculated confidence intervals should contain the true population parameter.
- In other words, approximately 5% of the intervals may fail to include the parameter.



How do we interpret a 95% confidence interval?

Cont'd

- We have simulated 20 random samples of size 100, and plot the 95% confidence intervals.
- Most of the intervals include the parameter, while the only exception case is highlighted in red.
- Hence, 1 out of 20 confidence intervals fails to include the true parameter value, which is consistent with 95% confidence level.



Why 95%?

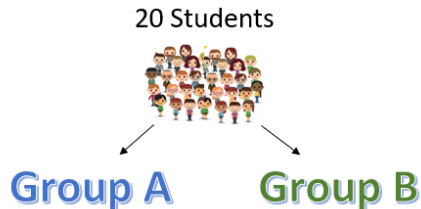
- Note that a 95% confidence interval will be wider than an 80% confidence interval computed on the same dataset.
- So a 95% confidence interval is just like a larger net, which is more likely to capture the true parameter than the 80% one.
- Though a 100% confidence interval will definitely contain the parameter, it is not helpful, as it includes all possible values, say, from 0 to infinity.
- In practice, we often make a trade-off such that we live with a small risk of missing the parameter, while we benefit from a narrower interval of values to work with.

Calculating Confidence Intervals in R

Apply it

Aim of the Study

To evaluate whether the population mean scores of the two exam papers are different.



- The point estimate is chosen as the difference of the two samples' mean scores.

```
x1 <- mean(df1[df1$treatment == "a",2]) # Group A's mean
x2 <- mean(df1[df1$treatment == "b",2]) # Group B's mean
# Point estimate
x1 - x2

[1] -9.4
```

Standard Error

Standard Error

$$SE = s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

is an estimator of the pooled standard deviation; n_1 and n_2 are the two samples' sizes; s_1 and s_2 are the two samples' standard deviations.

- We only need to know that the formula utilises four values: the sizes and the standard deviations for both sample groups.

Standard Error

Cont'd

- We can calculate the standard error in R, as follows.

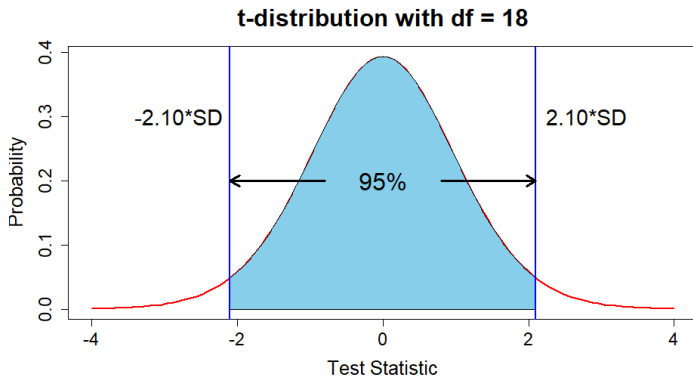
```
sp <- sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2))  
SE <- sp * sqrt(1/n1 + 1/n2)  
SE
```

```
[1] 5.008881
```

- The SE value is very close to 5.

t-distribution and Critical Value

- It is known that the test statistic follows a t-distribution, and the degree of freedom is 18.
- The t-critical value refers to the boundary t-statistic value, which is the 97.5% quantile of the t-distribution.

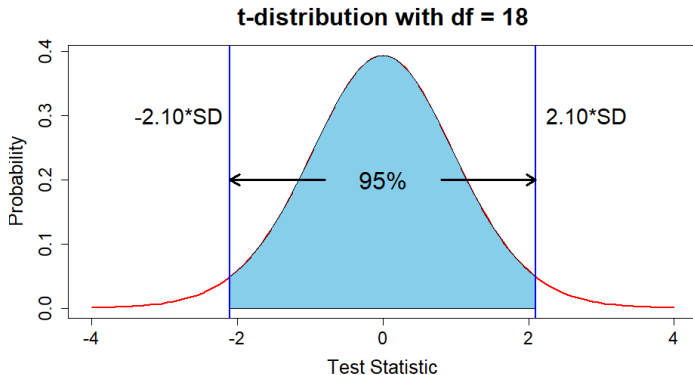


t-distribution and Critical Value

Cont'd

```
# Critical value (for t-distribution)  
qt(0.975,df=n1+n2-2)
```

```
[1] 2.100922
```



Construct the Confidence Interval

$$\begin{aligned} 95\% \text{Confidence Interval} &= \text{Point Estimate} \pm \text{Critical Value} * \text{Standard Error} \\ &= -9.4 \pm 2.1 * 5 \\ &= (-19.9, 1.1) \end{aligned}$$

```
t.test(outcome~treatment, alternative = "two.sided", paired=FALSE,  
       var.equal=TRUE, data = df1)
```

Two Sample t-test

data: outcome by treatment

t = -1.8767, df = 18, p-value = 0.07687

alternative hypothesis: true difference in means between group a and group b is not equal to 0

95 percent confidence interval:

-19.923268 1.123268

sample estimates:

mean in group a mean in group b

55.2

64.6

Construct the Confidence Interval

Cont'd

```
● result1 <- t.test(outcome~treatment, alternative = "two.sided",  
  paired=FALSE, var.equal=TRUE, data = df1)  
result1$conf.int
```

```
[1] -19.923268    1.123268
```

```
attr(,"conf.level")
```

```
[1] 0.95
```

- We are 95% confident that the difference of population mean scores of the two exam papers lies between -19.9 and 1.1 .
- As the interval includes 0, the population difference value might be 0, or some value below 0, or some value slightly above 0.
- We are not certain that the two exam papers must have different difficulty levels.
- Hence, we don't reject the null hypothesis.

Calculate 95% CI

Aim of the Study

To evaluate whether the population proportions of infected subjects among the two groups are different.

- The point estimate is chosen as the difference of the two samples' proportions, particularly, the proportion of the control group, minus that of the vaccine group.

```
table1 <- table(df2)    # 2 by 2 contingency table
# Sample size and sample proportion for the control group
n1 <- 24
p1 <- table1[1,1]/sum(table1[1,])
# Sample size and sample proportion for the vaccine group
n2 <- 24
p2 <- table1[2,1]/sum(table1[2,])
p1 - p2    # Point estimate
```

```
[1] 0.2916667
```

Standard Error

- The test statistic can be defined as the proportions' difference, 0.292, divided by the standard error.

Standard Error

$$SE = \sqrt{\frac{p_1 * (1 - p_1)}{n_1} + \frac{p_2 * (1 - p_2)}{n_2}}$$

where n_1 and n_2 are the two samples' sizes; p_1 and p_2 are the two samples' proportions.

- We only need to know that the formula utilises four values: the sizes and the proportions for both sample groups.
- We can calculate the standard error in R, as follows.

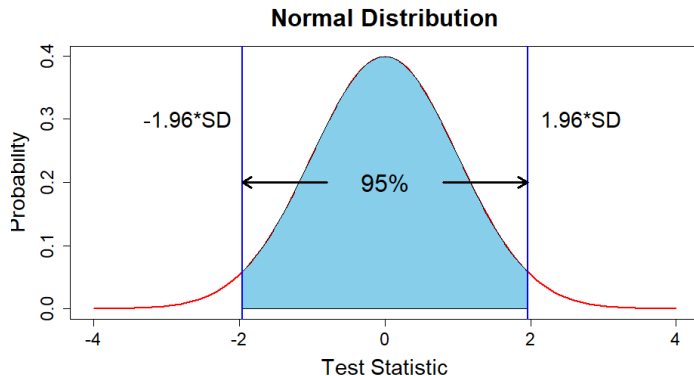
```
SE <- sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
SE
[1] 0.1211801
```

Normal Distribution and Critical Value

- Under some assumptions, the test statistic follows a normal distribution.
- The critical value refers to the 97.5% quantile of the normal distribution.

```
qnorm(0.975) # Critical value (for normal distribution)
```

```
[1] 1.959964
```



Construct the Confidence Interval

$$\begin{aligned} 95\% \text{ Confidence Interval} &= \text{Point Estimate} \pm \text{Critical Value} * \text{Standard Error} \\ &= 0.292 \pm 1.96 * 0.1212 \\ &= (0.054, 0.529) \end{aligned}$$

```
# Besides p-value, the prop.test function also returns the confidence interval.
```

```
prop.test(table1, correct = FALSE)
```

```
2-sample test for equality of proportions without continuity correction
```

```
data: table1      X-squared = 5.1692, df = 1, p-value = 0.02299
```

```
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
```

```
0.05415812 0.52917522
```

```
sample estimates:
```

```
prop 1    prop 2
```

```
0.4166667 0.1250000
```


Construct the Confidence Interval

Cont'd

- We are 95% confident that the difference of two population proportions, lies between 5.4% and 52.9%.
- As the interval does not includes 0, the plausible differences are all positive values.
- Hence, we are certain that the infection rate among the control group is higher than that of the vaccine group.
- So, we reject the null hypothesis, and we confidently claim the association, between vaccination and infection status, does exist.

Check Assumptions

Assumptions of two sample t-tests

- ① **Numerical scale:** The dependent (outcome) variable is numerical.
 - ② **Independence:** The observations are independent within and between the two groups.
 - ③ **Normality:** The population data follow normal distribution.
-
- As the key of constructing a confidence interval is the variability of the test statistic and the sampling distribution, the assumptions of confidence intervals naturally follow from those of hypothesis tests.
 - When comparing two samples' mean scores, the assumptions exactly coincide with those of the two sample t-tests, including normality.
 - The normality assumption may not be satisfied; hence, the validity and the accuracy of the confidence interval could be affected.

Check Assumptions

Cont'd

- When comparing two samples' proportions, we assume the sampling distribution is normal.
- So, we need one particular assumption that all four numbers in the contingency table must be larger than 10.

```
table1
```

	outcome	
treatment	Infected	Not Infected
Control	10	14
Vaccine	3	21

- However, it is not satisfied.
- In such case, it is recommended to turn on the continuity correction, as follows.

```
prop.test(table1, correct = TRUE)
```

Summary

Summary

- Both p-values and confidence intervals of classical approaches depend on the distributional assumption of the population data.
- When the population data do not follow normal distribution, or when the sample size is small, the validity of the classical p-values and confidence intervals are questionable.
- To overcome the above limitation, we have introduced another approach to estimating p-values: permutation testing.
 - ▶ The basic idea is to generate the sampling distribution by simulating multiple permutations.
 - ▶ It utilises the idea of re-sampling from the original sample, without replacement.


Summary

Cont'd

- Another drawback of classical confidence intervals is that they are always symmetric, due to the plus-minus part in the template formula.
 - ▶ When the sampling distribution is skewed, say, right skewed, this is not appropriate.
 - ▶ We should expect an asymmetric interval, to adjust for the fact that the true population mean is more likely a larger value.

Bootstrapping

References

-  Mine Çetinkaya-Rundel and Johanna Hardin (2021)
Introduction to Modern Statistics