Understanding Confidence Intervals

Confidence Interval (CI) is essential in statistics and very important for data scientists. As it sounds, the confidence interval is a range of values. In the ideal condition, it should contain the best estimate of a statistical parameter. It is expressed as a percentage. 95% confidence interval is the most common. You can use other values like 97%, 90%, 75%, or even 99% confidence interval if your research demands. Let's understand it by an example:

Here is a statement:

"In a sample of 659 parents with toddlers, about 85%, stated they use a car seat for all travel with their toddler. From these results, a 95% confidence interval was provided, going from about 82.3% up to 87.7%."

This statement means, we are 95% certain that the population proportion who use a car seat for all travel with their toddler will fall between 82.3% and 87.7%. If we take a different sample or a subsample of these 659 people, 95% of the time, the percentage of the population who use a car seat in all travel with their toddlers will be in between 82.3% and 87.7%.

Remember, 95% confidence interval does not mean 95% probability

The reason confidence interval is so popular and useful is, we cannot take data from all populations. Like the example above, we could not get the information from all the parents with toddlers. We had to calculate the result from 659 parents. From that result, we tried to get an estimate of the overall population. So, it is reasonable to consider a margin of error and take a range. That's why we take a confidence interval which is a range.

We want a simple random sample and a normal distribution to construct a confidence interval. But if the sample size is large enough (30 or more) normal distribution is not necessary.

How to Calculate the Confidence Interval

The calculation of the confidence interval involves the best estimate which is obtained by the sample and a margin of error. So, we take the best estimate and add a margin of error to it. Here is the formula for the confidence interval and the margin of error:

$$Best\ Estimate \pm Margin\ of\ Error$$

$$Margin\ of\ Error = z * Estimated\ SE$$

Here, SE is the standard error.

Normally, CI is calculated for two statistical parameters: the proportion and the mean.

Combining these two formulas above, we can elaborate the formula for CI as follows:

$$Population\ Proportion\ or\ Mean \pm z - score * Stadard\ Error$$

Population proportion or the mean is calculated from the sample. In the example of "the parents with toddlers", the best estimate or the population proportion of parents that uses car seats in all travel with their toddlers is 85%. So, the best estimate (population proportion) is 85. z-score is fixed for the confidence level (CL).

A z-score for a 95% confidence interval for a large enough sample size(30 or more) is 1.96.

Here are the z-scores for some commonly used confidence levels:

Confidence %	Z
75%	1.15
90%	1.64
95%	1.96
97%	2.17
99%	2.57
99.90%	3.29

The method to calculate the standard error is different for population proportion and mean. The formula to calculate standard error of population proportion is:

$$Standard\ Error\ For\ Population\ Proportion = \sqrt{(Population\ Proportion * \frac{(1 - Population\ Proportion)}{Number\ Of\ Observations})}$$

The formula to calculate the standard error of the sample mean is:

$$Standard\ Error\ For\ Mean = \frac{Standard\ Deviation}{\sqrt{Number Of\ Observations}}$$

As per the statement, the population proportion that uses a car seat for all travel with their toddlers is 85%. So, this is our best estimate. We need to add the margin of error to it. To calculate the margin of error we need the z-score and the standard error. I am going to calculate a 95% CI. The z-score should be 1.96 and I already mentioned the formula of standard error for the population proportion. $0.85 \pm 1.96 * \sqrt{(0.85(1 - 0.85)/659)}$ Plugging in all the values:

The confidence interval is 82.3% and 87.7% as we saw in the statement before.

Assumptions for a Single Population Proportion Confidence Interval:

When constructing confidence intervals the assumptions and conditions of the central limit theorem must be met in order to use the normal model.

- **Randomization Condition:** The data must be sampled randomly. Is one of the good sampling methodologies discussed in the Sampling and Data chapter being used?
- **Independence Assumption**: The sample values must be independent of each other. This means that the occurrence of one event has no influence on the next event. Usually, if we know that people or items were selected randomly we can assume that the independence assumption is met.
- **10% Condition:** When the sample is drawn without replacement (usually the case), the sample size, *n*, should be no more than 10% of the population.

- **Sample Size Condition:** The sample size must be sufficiently large. Although the Central Limit Theorem tells us that we can use a Normal model to think about the behavior of sample means when the sample size is large enough, it does not tell us how large that should be. If the population is very skewed, you will need a pretty large sample size to use the CLT, however if the population is unimodal and symmetric, even small samples are ok. So think about your sample size in terms of what you know about the population and decide whether the sample is large enough. In general a sample size of 30 is considered sufficient.
- When working with **numerical data and σ is unknown** the assumptions of randomization, independence and the 10% condition must be met. In addition, with small sample sizes we cannot assume that that data follows a normal distribution so we need to check the **nearly normal condition**. To check the nearly normal condition start by making a histogram or stemplot of the data, it is a good idea to make an outlier boxplot, too. If the sample is small, less than 15 then the data must be normally distributed. If the sample size is moderate, between 15 and 40, then a little skewing in the data will can be tolerated. With large sample sizes, more than 40, we are concerned about multiple peaks (modes) in the data and outliers. The data might not be approximately normal with either of these conditions and you may want to run the test both with and without the outliers to determine the extent of their effect. If there are multiple modes in the data if could be that there are two groups in the data that need to be separated.
- When working with **binomial or categorical data** the assumptions of randomization, independence and the 10% condition must be met. In addition, a new assumption, the **success/ failure condition**, must be checked. When working with proportions we need to be especially concerned about sample size when the proportion is close to zero or one. To check that the sample size is large enough calculate the success by multiplying the sample percentage by the sample size and calculate failure by multiplying one minus the sample percentage by the sample size. If both of these products are larger than ten then the condition is met.

Interpretations & Assumptions for Two Population Proportion Intervals:

- We need to assume that we have two independent random samples.
- We also need large enough sample sizes to assume that the distribution of our estimate is normal. That is, we need $n1\hat{p}1$, $n1(1-\hat{p}1)$, $n2\hat{p}2$, and $n2(1-\hat{p}2)$ to all be at least 10.

Confidence interval in Python

I am assuming that you are already a python user. But even if you are not a python user you should be able to get the concept of the calculation and use your own tools to calculate the same. The tools I used for this exercise are:

1. Numpy Library

2. Pandas Library
3. Statsmodels Library
4. Jupyter Notebook environment.

If you install an anaconda package, you will get a Jupyter Notebook and the other tools as well. There are some good youtube videos to demonstrate how to install anaconda package if you do not have that already.

CI for the population Proportion in Python

I am going to use the Heart dataset from Kaggle. Please click on the link to download the dataset. First, I imported the packages and the dataset:

```
import pandas as pd
import numpy as np
df = pd.read_csv('Heart.csv')
df
```

	Unnamed: 0	Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	AHD
0	1	63	1	typical	145	233	1	2	150	0	2.3	3	0.0	fixed	No
1	2	67	1	asymptomatic	160	286	0	2	108	1	1.5	2	3.0	normal	Yes
2	3	67	1	asymptomatic	120	229	0	2	129	1	2.6	2	2.0	reversable	Yes
3	4	37	1	nonanginal	130	250	0	0	187	0	3.5	3	0.0	normal	No
4	5	41	0	nontypical	130	204	0	2	172	0	1.4	1	0.0	normal	No

The last column of the data is ‘AHD’. It says if a person has heart disease or not. In the beginning, we have a ‘Sex’ column as well.

We are going to construct a CI for the female population proportion that has heart disease.

First, replace 1 and 0 with ‘Male’ and ‘Female’ in a new column ‘Sex1’.

```
df['Sex1'] = df.Sex.replace({1: "Male", 0: "Female"})
```

We do not need all the columns in the dataset. We will only use the ‘AHD’ column as that contains if a person has heart disease or not and the Sex1 column we just created. *Make a DataFrame with only these two columns and drop all the null values.*

```
dx = df[["AHD", "Sex1"]].dropna()
```

Here is the output table:

Sex1	Female	Male
AHD		
No	72	92
Yes	25	114

The number of females who have heart disease is 25. *Calculate the female population proportion with heart disease.*

```
p_fm = 25/(72+25)
```

The ‘p_fm’ is 0.26. *The size of the female population:*

```
n = 72+25
```

The size of the female population is 97. *Calculate the standard error*

```
se_female = np.sqrt(p_fm * (1 - p_fm) / n)
```

The standard error is 0.044.

Now construct the CI using the formulas above. The z-score is 1.96 for a 95% confidence interval.

```
z_score = 1.96
lcb = p_fm - z_score* se_female #lower limit of the CI
ucb = p_fm + z_score* se_female #upper limit of the CI
```

The confidence interval is 0.17 and 0.344.

You can calculate it using the library ‘statsmodels’.

```
import statsmodels.api as sm
sm.stats.proportion_confint(n * p_fm, n)
```

The confidence interval comes out to be the same as above.

Confidence Intervals for Differences between Population Parameters

1. CI for the Difference in Population Proportion

Is the population proportion of females with heart disease the same as the population proportion of males with heart disease? If they are the same, then the difference in both the population proportions will be zero.

We will calculate a confidence interval of the difference in the population proportion of females and males with heart disease.

Here is the step by step process:

Calculate the male population proportion with heart disease and standard error using the same procedure.

```
p_male = 114/(114+92) #male population proportion
n = 114+92 #total male population
```

The male population proportion with heart disease is 0.55 and the male population size is 206. *Calculate the standard error for the male population proportion.*

```
se_male = np.sqrt(p_male * (1 - p_male) / n)
```

The standard error for the male population is 0.034. *Calculate the difference in standard error.*

Here is the formula to calculate the difference in two standard errors:

$$\sqrt{SE_1^2 + SE_2^2}$$

Let's use this formula to calculate the difference in the standard error of male and female population with heart disease.

```
se_diff = np.sqrt(se_female**2 + se_male**2)
```

Use this standard error to calculate the difference in the population proportion of males and females with heart disease and construct the CI of the difference.

```
d = 0.55 - 0.26
lcb = d - 1.96 * se_diff #lower limit of the CI
ucb = d + 1.96 * se_diff #upper limit of the CI
```

The CI is 0.18 and 0.4. This range does not have 0 in it. Both the numbers are above zero. So, We cannot make any conclusion that the population proportion of females with heart disease is the same as the population proportion of males with heart disease. If the CI would be -0.12 and 0.1, we could say that the male and female population proportion with heart disease is the same. Calculation of CI of mean

We will use the same heart disease dataset. The dataset has a 'chol' column that contains the cholesterol level. For this demonstration,

we will calculate the confidence interval of the mean cholesterol level of the female population.

Let's find the mean, standard deviation, and population size for the female population. I want to get the same parameters for the male population as well. Because it will be useful for our next exercise. Use pandas groupby and aggregate methods for this purpose. If you need a refresher on pandas groupby and aggregate method, please check out this article:

Here is the code to get the mean, standard deviation, and population size of the male and female population:

```
df.groupby("Sex1").agg({"Chol": [np.mean, np.std, np.size]})
```

	Chol		
	mean	std	size
Sex1			
Female	261.752577	64.900891	97
Male	239.601942	42.649757	206

If we extract the necessary parameters for the female population only:

```
mean_fe = 261.75 #mean cholesterol of female
sd = 64.9 #standard deviation for female population
n = 97 #Total number of female
z = 1.96 #z-score from the z table mentioned before
```

Here 1.96 is the z-score for a 95% confidence level.

Calculate the standard error using the formula for the standard error of the mean

```
se = sd /np.sqrt(n)
```

Now we have everything to construct a CI for mean cholesterol in the female population.

Construct the CI

```
lcb = mean_fe - z* se #lower limit of the CI
ucb = mean_fe + z* se #upper limit of the CI
(lcb, ucb)
```

The CI came out to be 248.83 and 274.67.

That means the true mean of the cholesterol of the female population will fall between 248.83 and 274.67

2. Calculation of CI of The Difference in Mean

There are two approaches to calculate the CI for the difference in the mean of two populations.

Pooled approach and unpooled approach

As mentioned earlier, we need a simple random sample and a normal distribution. If the sample is large, a normal distribution is not necessary.

There is one more assumption for a pooled approach. That is, the variance of the two populations is the same or almost the same.

If the variance is not the same, the unpooled approach is more appropriate.

$$Mean_{diffrence} \pm z * \sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}$$

The formula of the standard error for the pooled approach is:

Here, s1 and s2 are the standard error for the population1 and population2. In the same way, n1 and n2 are the population size of population1 and population2.

The formula of the standard error for the unpooled approach is:

Here, we will construct the CI for the difference in mean of the cholesterol level of the male and female

$$Mean_{diffrence} \pm z * \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} * \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}$$

population.

We already derived all the necessary parameters from the dataset in the previous example. Here they are:

```
n1 = 97
n2 = 206
mean_female = 261.75
mean_male = 239.6
sd_female = 64.9
sd_male = 42.65
```

As we can see, the standard deviation of the two target populations is different. So. the variance must be different as well.

So, for this example, the unpooled approach will be more appropriate.

Calculate the standard error for male and female population using the formula we used in the previous example

```
sem_female = sd_female / np.sqrt(97)
sem_male = sd_male / np.sqrt(206)
```

The difference in mean of the two samples

```
mean_d = mean_female - mean_male
```

The difference in mean ‘mean_d’ is 22.15.

Using the formula for the unpooled approach, calculate the difference in standard error:

```
sem_d = (np.sqrt((n1-1)*se_female**2 + (n2-1)*se_male**2)/(n1+n2-2))*(np.sqrt(1/n1 + 1/n2))
```

Finally, construct the CI for the difference in mean

```
lcb = mean_d - 1.96*sem_d #lower limit of the CI
ucb = mean_d + 1.96*sem_d #upper limit of the CI
(lcb, ucb)
```

The lower and upper limit of the confidence interval came out to be 22.1494 and 22.15. They are almost the same. That means the mean cholesterol of the female population is not different than the mean cholesterol of the male population.