How to calculate Confidence interval?

If your sample size > 30 and you know the population standard deviation, a z score is appropriate. If your sample size is small and your population variance is unknown then you can use a t-score. Let's check an example of Confidence interval calculating with Z score.

Example

For example we want to estimate the distribution of heights in the population. Let's import our datasets with heights.

```
#Import libraries
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import matplotlib.pyplot as plt


#import dataset
heightWeight = pd.read_csv("../input/heights-and-weights-dataset/SOCR-HeightWeight.csv")


#check data
heightWeight.head()


#cut only 'Height' column for our data
data = heightWeight.iloc[:, 1]
print(data)
```

$$\overline{x} \pm z \frac{s}{\sqrt{n}}$$

Here is a formula of Confidence Interval:

```
# 1. X - mean of population
mean = data.mean()
print(mean)
```

Let's figure it out. The Z score gives us an estimate of the number of standard deviations that an observation lies from the mean. The exact Z score depends on the selected confidence interval. Confidence level is equal 95% chosen most often.

Percentage Confidence	*z**-Value
80	1.28
90	1.645
95	1.96
98	2.33
99	2.58

In case of 95% we have Z-score with value 1.96.

```
# 2. Z-score
Z = 1.96
```

```
# 3. standard deviation of the entire population
std = data.std()
print(std)


# 4. n - the number of observations. It is sample. I am going to select for sample first 100 values.
n = len(data[:100])
print(n)


# let's collect all these values in the formula above.
import math

#left limit
left_limit = mean - (Z * (std / math.sqrt(n)))
print(left_limit)


#right limit
right_limit = mean + (Z * (std / math.sqrt(n)))
print(right_limit)
```

Our result says that the true mean height of the entire population (if we could measure all their heights) would be between 67.62 icnhes and 68.36 in 95% of cases. Let's plot this data.

```
plt.figure(figsize=(12,8))
plt.hist(data, bins=25, density=True, alpha=0.6, color='#e06666')

plt.axvline(left_limit, color='purple',lw=3, label=f'Left limit: ${left_limit}')
plt.axvline(right_limit, color='purple',lw=3, label=f'Right limit: ${right_limit}')

plt.xlabel('Heights')
plt.ylabel('Probability Density')
```

```
plt.title('Confidence interval', fontsize=16)
plt.legend()
```

Conclusion

We have considered with an example what a confidence interval is and why it is needed. We built a graph that clearly shows that the mean of the population will lie in a given interval.

If you enjoyed my notebook, please hit the upvote button! Thank you!