BUSINESS INTELLIGENCE AND DATA ANALYTICS

University of Petra

Institute of Analytics Recognized Programme

University of Petra

307102
Descriptive Statistics for Business

Introduction to Estimation in Statistics

Content

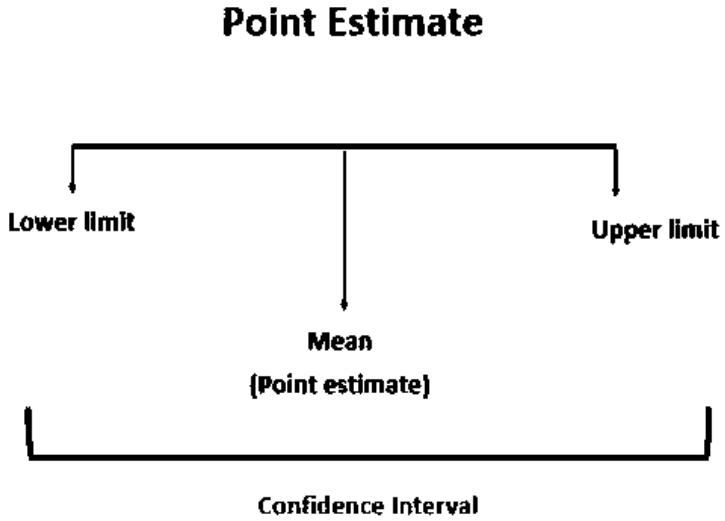
- Introduction to Inferential Statistics
- Estimation in Statistics (Point Estimate and Interval Estimate)
- The Central Limit Theorem (CLT)
- Confidence Intervals – Using Z-Tables
- Confidence Intervals – Using T-Tables

Estimation in Statistics

- Often in statistics we're interested in measuring population parameters.
- Two of the most common population parameters are:
  1. Population mean: the mean value of some variable in a population (e.g. the mean height of males in the U.S.)
  2. Population proportion: the proportion of some variable in a population (e.g. the proportion of residents in a county who support a certain law)
- Although we're interested in measuring these parameters, it's usually too costly and time-consuming to go around and collect data on every individual in a population in order to calculate the population parameter.
- Instead, **we typically take a random sample** from the overall population and use data from the sample to **estimate** the population parameter.

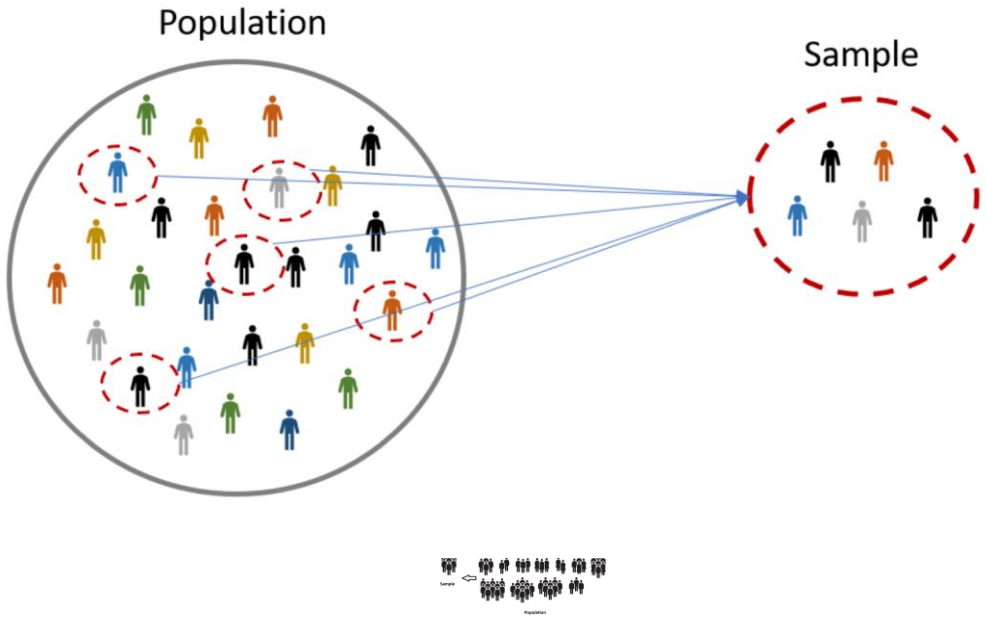
Point vs. Interval

- A point estimate is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- An interval estimate gives you a range of values where the parameter is expected to lie.
- Both types of estimates are important for gathering a clear idea of where a parameter is likely to lie



Point Estimates

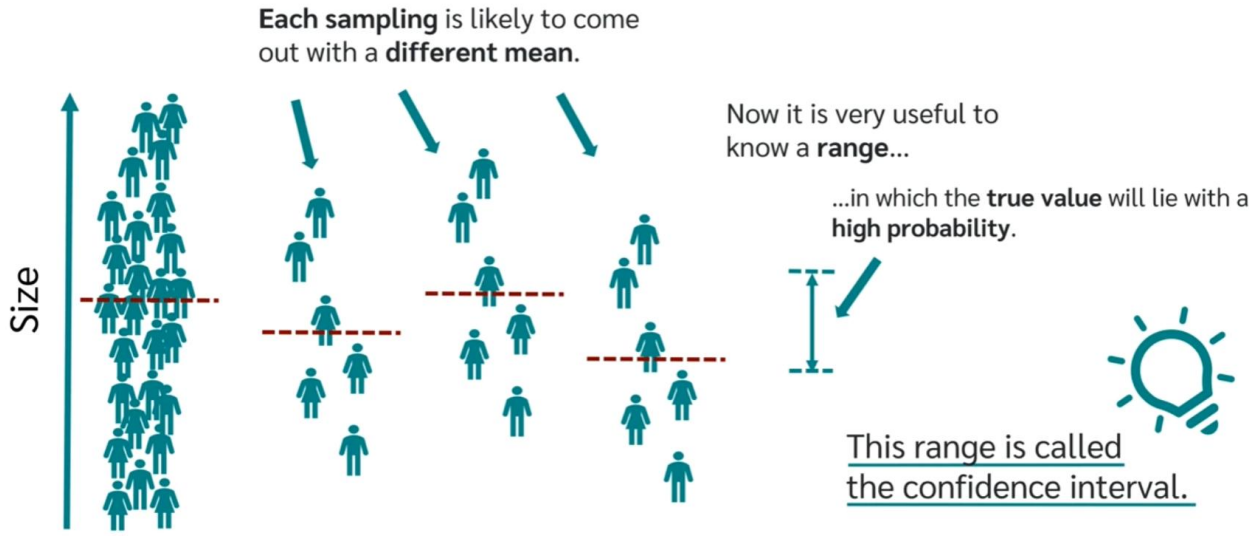
- Suppose we want to estimate the mean weight of a students in the university.
- Since there are thousands of students in the university, it would be extremely time-consuming and costly to go around and weigh each individual student.
- Instead, we might take a random sample of 50 students and use the mean weight of the students in this sample to estimate the true population mean.



- In this case, the mean weight of the sample is called a Point Estimate for the true mean weight of the population.

Interval Estimates - Confidence Intervals

- The problem with the previous example is that the mean weight of students in the sample **is not guaranteed** to exactly match the mean weight of students in the whole population.
- For example, we might just happen to pick a sample full of low-weight students or perhaps a sample full of heavy students.
- To capture this uncertainty, we can create an Interval Estimate or a confidence interval around our point estimate.
- The confidence interval gives us a range of values that are likely to contain the true population parameter.



Computing Confidence Intervals

To compute the confidence interval for the mean of a normally distributed data we use the formula below:

Confidence Interval = (point estimate) ± Margin of Error

Confidence Interval = (point estimate) ± [(critical value)*(standard error)]

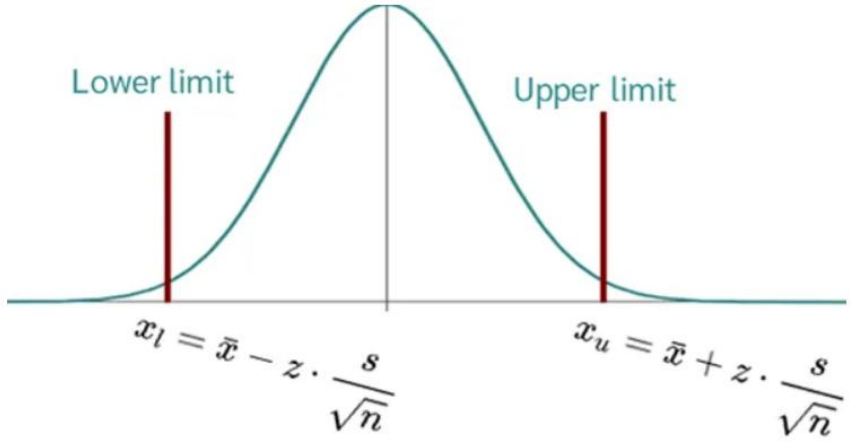
The critical value is the z-score and the standard error = $s/\sqrt{n}$

Confidence Interval is calculated as:

Point Estimate ± Margin of Error

Point Estimate ± (Critical Value) (Standard Error)

$$\text{C.I.} = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



https://medium.com/@ashisharora2204/hypothesis-testing-confidence-interval-level-margin-of-error-39aa7c7ddcd2

The Significance Level Alpha (α) and The Confidence Level

- To compute the confidence interval, we need to define the Significance Level Alpha (*α)* or the Confidence Level.
- The two terms are related to each other according to the following equation:

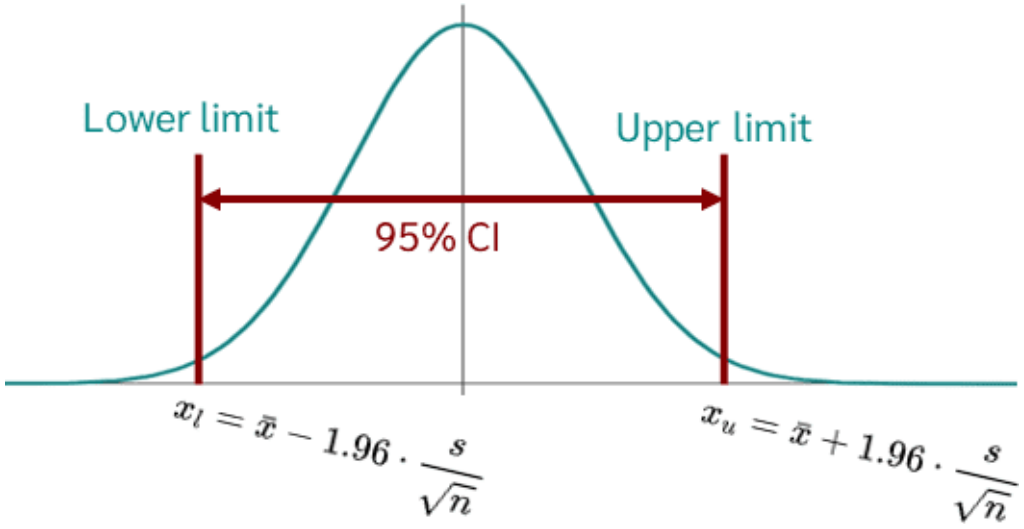
  Confidence level= 1 – *α*

- Common values for these notations are listed in the table

Alpha - α	**5%**	**1%**
Confidence level	**95%**	**99%**
z-Value	**1.96**	**2.58**

The Significance Level Alpha (α) and The Confidence Level

- The confidence level (e.g. 95%) means that if we collected 100 samples and created confidence intervals for each of these samples, we would expect that 95% of these confidence intervals will contain the true population parameter.
- It's important to note that this is a theoretical long-term proportion—it's about what we'd expect to happen over many repetitions of the same process, not a guarantee for any individual interval.



Relationship Between Confidence Level and Significance Level in Statistics

The **confidence level** and the **significance level (α)** are inverse and complementary concepts in statistical inference and hypothesis testing.

Relationship:

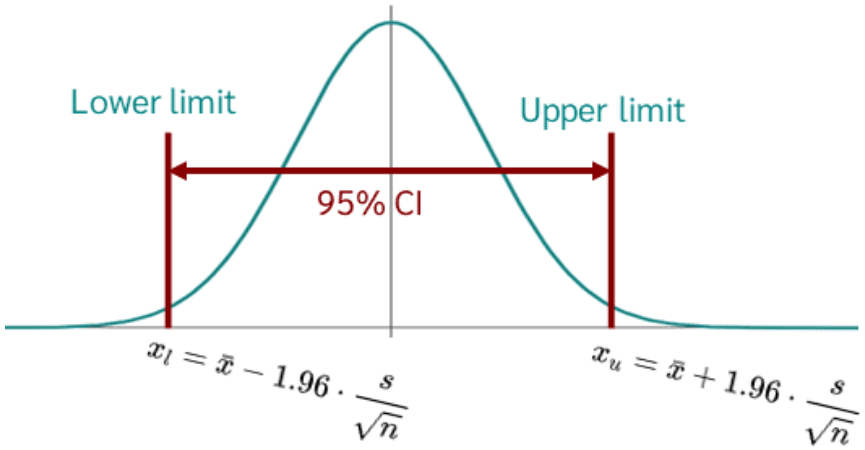
- The two are **complementary**, meaning that if you increase one, you decrease the other, and vice versa.
- A **higher confidence level** corresponds to a **lower significance level**, while a **lower confidence level** corresponds to a **higher significance level**.

Formula:

- Confidence Level=1−α

Example:

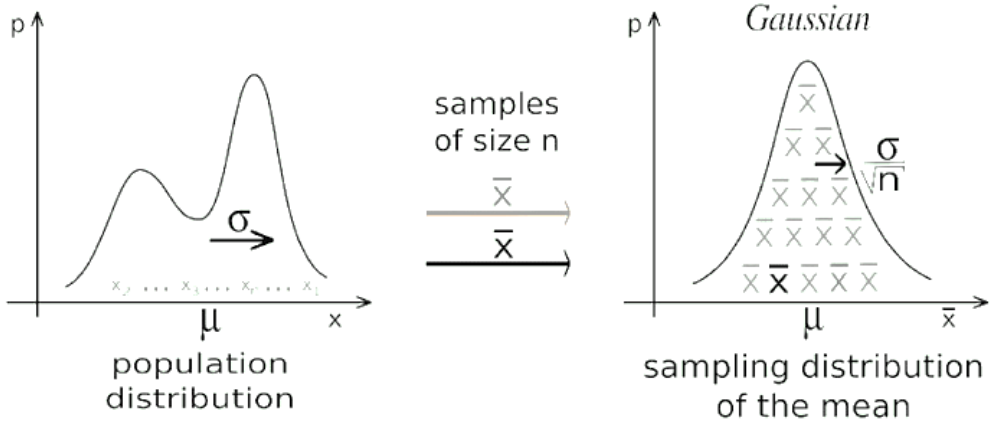
- A 95% confidence level means the significance level is 0.05 (α = 5%).
- A 99% confidence level means the significance level is 0.01 (α = 1%).



Where did the Standard Error Come From?
The Central Limit Theorem

The Central Limit Theorem

- The Central Limit Theorem (CLT) states that samples means created from large number of samples will be approximately normally distribution although the distribution we sample from might not be normal.
- When we collect large number of samples and we compute the average for each of these samples, we call the generated means as the Sampling Distribution of the means.
- The distribution of samples means becomes more normal when the samples size increases.



According to CLT:

1- The mean of the sampling distribution would be equal to the mean of the original distribution.

$$\mu_{\bar{x}} = \mu$$

2- The variance of the sampling distribution of the sample mean would be equal to the variance of the original distribution divided by n, where n is the size of the samples.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

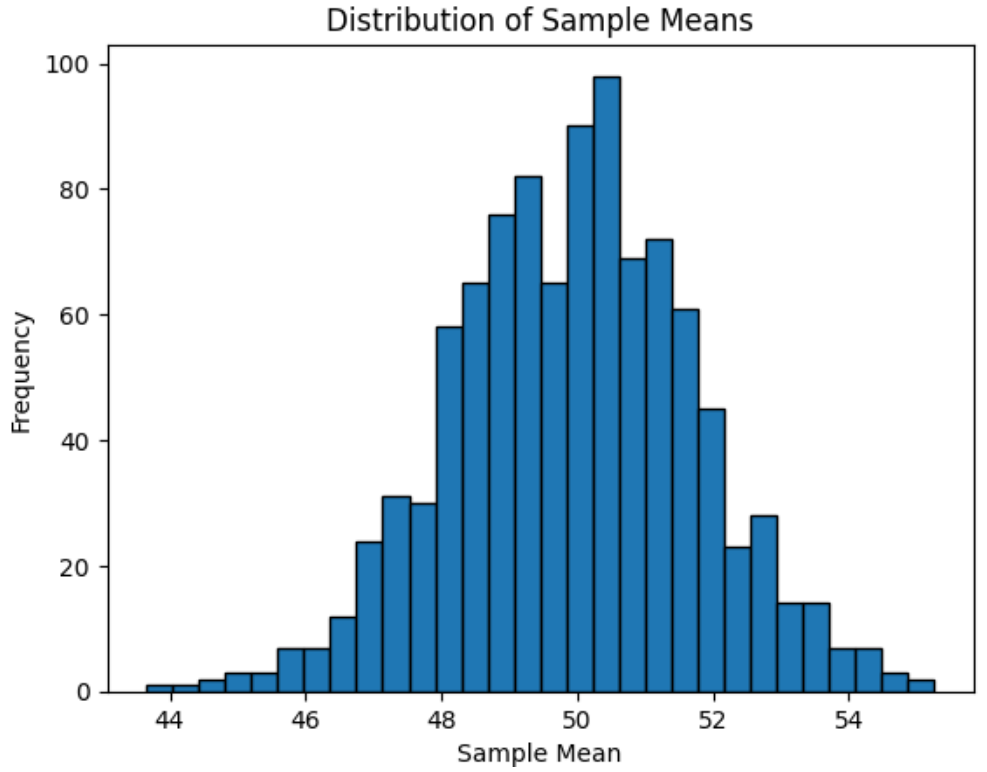
Click here for a CLT simulator.

VIDEO TUTORIAL

Empirical Proof that Standard Error is an Approximate for the True SD of Sample Means

```python
import numpy as np
import matplotlib.pyplot as plt

# Set population parameters
population = np.random.normal(loc=50, scale=10, size=1000000)
true_sigma = np.std(population)

# Simulation setup
sample_size = 30
num_samples = 1000
sample_means = []
sample_sds = []

for _ in range(num_samples):
    sample = np.random.choice(population, sample_size)
    sample_means.append(np.mean(sample))
    sample_sds.append(np.std(sample, ddof=1))

# Calculate actual SD of sample means
sd_of_sample_means = np.std(sample_means)

# Calculate average standard error estimate
average_se = np.mean([s / np.sqrt(sample_size) for s in sample_sds])

print(f"True SD of sample means: {sd_of_sample_means:.4f}")
print(f"Average estimated SE:     {average_se:.4f}")

# Optional: plot histogram
plt.hist(sample_means, bins=30, edgecolor='black')
plt.title('Distribution of Sample Means')
plt.xlabel('Sample Mean')
plt.ylabel('Frequency')
plt.show()
```

```
True SD of sample means: 1.7942
Average estimated SE:     1.8023
```

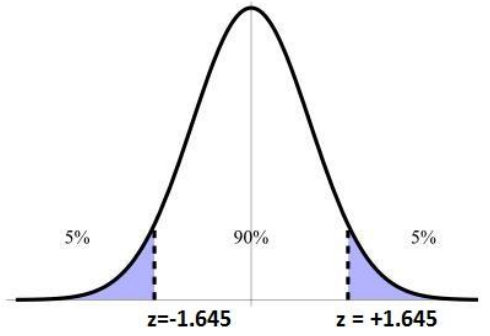


Confidence Intervals Example

Suppose we randomly collected a sample of students with the following information:

- Sample size n = 50
- Sample mean weight x = 73.16 kg
- Population standard deviation s = 8.63

What is the 90% confidence interval for the true population mean weight?



z	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608

- To find the upper bound z-score for the 90% CI, we need to look for area 95% in the z table i.e. 90% + 5% (Look at the Figure)
- Therefore, according to the z-table, the upper bound z-score for the 90% Confidence Interval = 1.64 i.e. lower bound is -1.64.
- The Margin of Error = 1.64 * (8.63/√50) = ~2
- Therefore, the 90% CI = 73.16 ± 2 = [71.16, 75.16]

Confidence Intervals Example

We interpret this 90% confidence interval as follows:

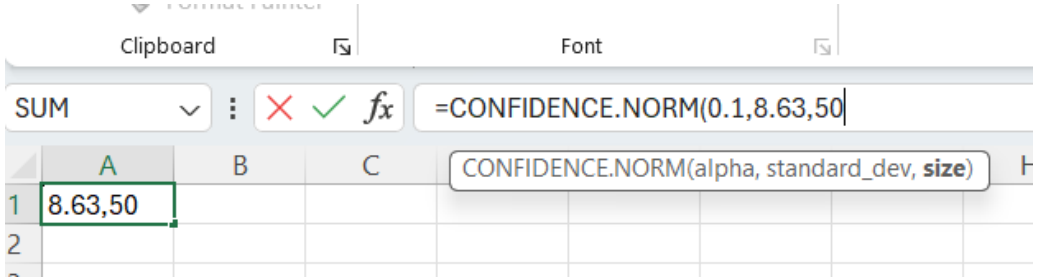
- The interval [71.16 kgs to 75.16 kgs] gives us a good estimate of where the mean lies, based on the collective information from our samples.
- The 90% means that if we estimate the population parameter and computed the confidence intervals 100 times, we expect that 90% of these intervals will contain our parameter.

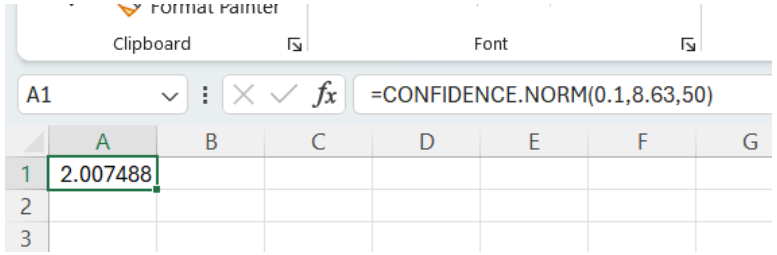
Solving Confidence Intervals using Excel Functions

Suppose we randomly collected a sample of students with the following information:

- Sample size n = 50
- Sample mean weight x = 73.16 kg
- Sample standard deviation s = 8.63

What is the 90% confidence interval for the true population mean weight?





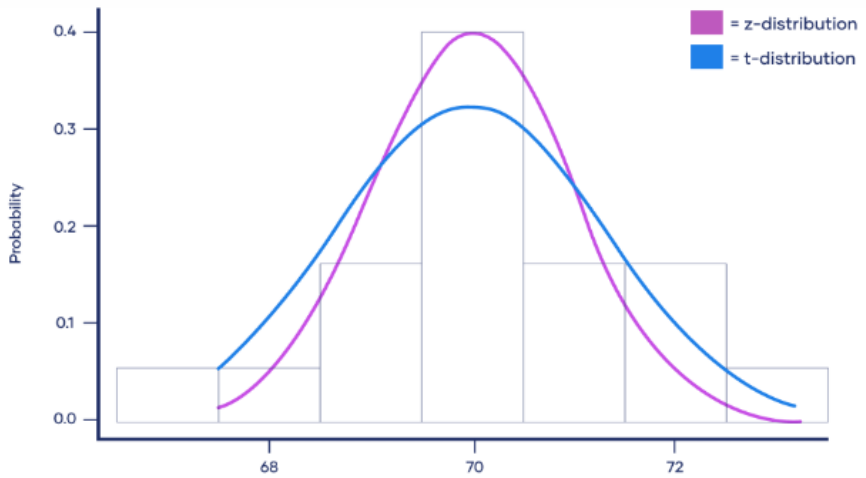
- We can compute the confidence interval using Excel CONFIDENCE.NORM function.
- This function computes the margin of error.
- CONFIDENCE.NORM(.1 , 8.63 , 50) = 2.0074 ~ 2
- Therefore, the 90% CI = 73.16 ± 2 = [71.16, 75.16]

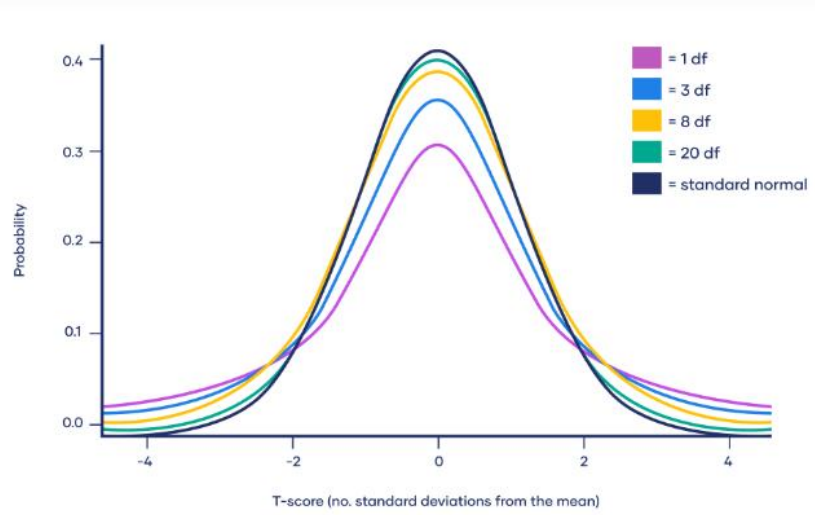
Transitioning to T-Distribution

- We can use the Z-distribution when the population variance is known or when our data is large enough (n > 30).
- However, real-world data often doesn't come with this information.
- In practice, the population variance is unknown, and we usually estimate the population variance using the sample variance, which introduces more variability into our statistics.
- The T-distribution accommodates this additional uncertainty and is especially useful for small sample sizes.
- The T-distribution often serves as a safer alternative because it accounts for the additional uncertainty from estimating the population variance.

Characteristics of T-Distribution

- The T-distribution, like the normal distribution, is bell-shaped and symmetric but has heavier tails.
- This means there is more probability in the tails and less in the center compared to a normal distribution.
- The exact shape of the T-distribution depends on the degrees of freedom (df), which are related to the sample size.
- Degrees of freedom in the context of the T-distribution refer to the number of independent values in a calculation of a statistic that are free to vary.
- For a confidence interval, df is equal to the sample size (n) minus 1 (df = n - 1).
- As the sample size increases, the T-distribution approaches the normal Z-distribution.





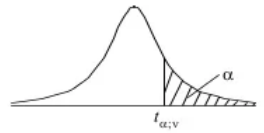
https://www.scribbr.com/statistics/t-distribution/

The T Table

- The first column, denoted "v," lists the degrees of freedom.
- Degrees of freedom typically equal the sample size minus one (n-1) and relate to the number of independent values in a set of observations.
- The top row lists different significance levels (α), like the probabilities in z-table.
- The intersection of a row and column gives the t-value or t score, which is the cutoff point on the t-distribution.
- The table lists the areas (probability) on the right side, of the t-scores (opposite to the z-tables), therefore, to find a t scores for confident intervals we split α by 2 and look out the resulting value in the table.
- For example, to find the 90% confidence interval for a sample of 20 observations, we find α which is 1 - .9 = .1
- We split α by 2 = .1 / 2 = 0.05 and we lookout that value in the table = 1.729

Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where $\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Confidence Intervals Using T-Distribution

- The formula for a confidence interval using the T-distribution is like the one with the Z-distribution:

$$CI = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

- $\bar{x}$ is the sample mean
- s is the sample standard deviation
- n is the sample size
- $t_{\alpha/2}$ is the t-score from the T-distribution that corresponds to the desired confidence level.

Interval Estimates - Confidence Intervals

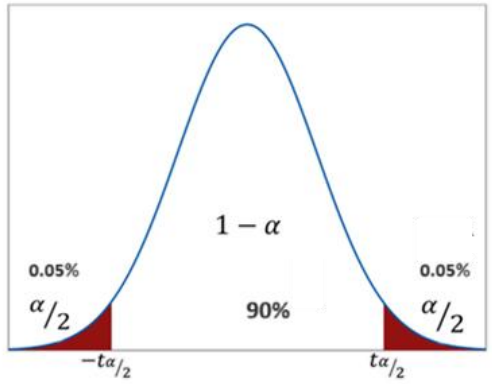
Suppose we randomly collected a sample of students with the following information:

- Sample size n = 20
- Sample mean weight x = 72.94 kg
- Samples standard deviation s = 8.3946

What is the 90% confidence interval for the true population mean weight?

- Confidence Interval 90% → *α = 1 - 0.9 = 0.1* → *α/2 = 0.05*
- According to the t-table, the t-score for the 0.05 with df 20 – 1 = 19 → t score = 1.729
- The Margin of Error = 1.729 * (8.3946/√20) = 3.25
- Therefore, the 90% CI = 72.94 ± 3.25 = [69.69, 76.19]

ν \ α	0.1	0.05				
1	3.078	6.314				
2	1.886	2.920				
3	1.638	2.353				
4	1.533	2.132				
5	1.476	2.015				
6	1.440	1.943				
7	1.415	1.895				
8	1.397	1.860				
9	1.383	1.833				
10	1.372	1.812				
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450



What is the difference between the 90% Confidence Interval of the Z-Distribution and the 90% Confidence Interval of the T-Distribution?

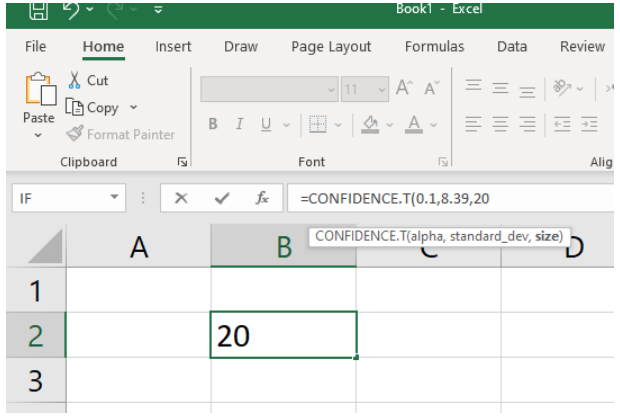
Solving Confidence Intervals using Excel Functions

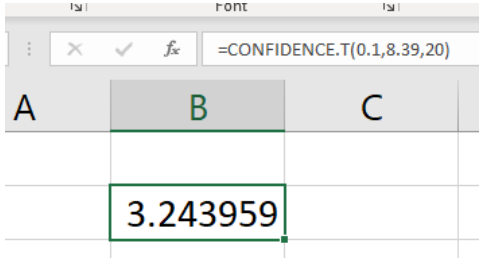
Suppose we randomly collected a sample of students with the following information:

- Sample size n = 20
- Sample mean weight x = 72.94 kg
- Samples standard deviation s = 8.3946

What is the 90% confidence interval for the true population mean weight?

- We can compute the confidence interval using Excel CONFIDENCE.T function.
- This function computes the margin of error.
- CONFIDENCE.T(.1 , 8. 39 , 20) = 3.24 ~ 2
- Therefore, the 90% CI = 72.94 ± 3.24 = [69.69, 76.19]





Steps to Calculate the Needed Sample Size

1. **Define your research objectives and questions.**
2. **Choose a significance level (α) and desired margin of error (E).**
3. **Estimate population variability (σ),** or use conservative estimates if unknown.
4. **Determine the population size (N).**
5. **Select the type of sampling** (random or stratified).
6. **Choose the statistical test or analysis** that will be applied.
7. **Use a sample size formula or software tool** to calculate the required sample size.
8. **Consider practical constraints** and adjust for potential non-response.
9. **Conduct the study, analyze data, and interpret results.**

1. For large populations (N very large or unknown):

$$n = \frac{Z^2 \cdot \sigma^2}{E^2}$$

Where:

- n = required sample size
- Z = Z-score (e.g., 1.96 for 95% confidence, 2.58 for 99%)
- σ^2 = population variance (if unknown, use 0.25 for proportions)
- E = margin of error

---

2. For proportions (most common):

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{E^2}$$

Where:

- p = estimated population proportion (use 0.5 if unknown for maximum sample size)

---

3. With finite population correction (when N is known):

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

Where:

- n_0 = initial sample size from formula (without considering N)
- N = population size