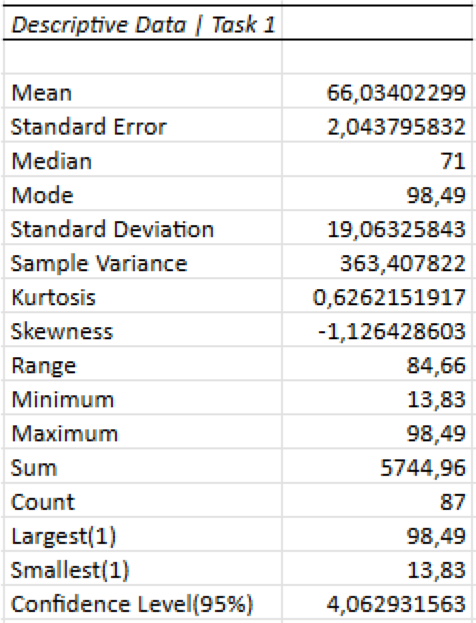
Project
Analysis of YouTube by Country
Shvets Ivan

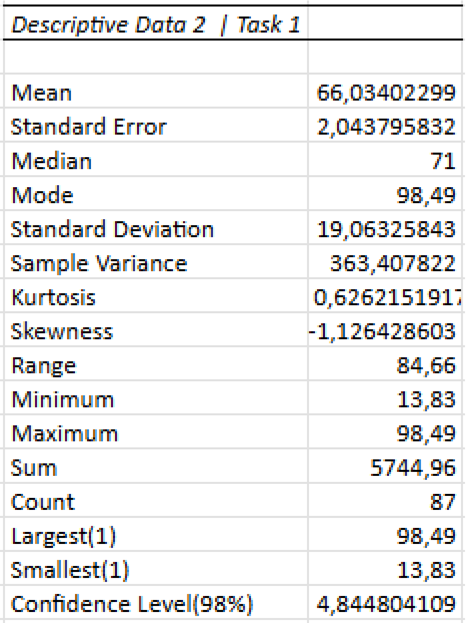
For this project I have chosen the number of Youtube users per country. The dataset includes a number of users per country. So in total the Youtube platform is used in 87 countries. I think this dataset is rather interesting for me since I am really invested into Youtube myself and I am curious how popular it is looking at each state. Data is useful especially for content creators, since they can consider it for determining their main audience. It is also worth mentioning that payment for views in each country differs and it influences the creators choice of audience.

Data Analysis

I have done a Descriptive Data analysis in Excel and got these results from my dataset for 95% and 98% Confidence Levels



Descriptive Data | Task 1	
Mean	66,03402299
Standard Error	2,043795832
Median	71
Mode	98,49
Standard Deviation	19,06325843
Sample Variance	363,407822
Kurtosis	0,6262151917
Skewness	-1,126428603
Range	84,66
Minimum	13,83
Maximum	98,49
Sum	5744,96
Count	87
Largest(1)	98,49
Smallest(1)	13,83
Confidence Level(95%)	4,062931563



Descriptive Data 2 | Task 1	
Mean	66,03402299
Standard Error	2,043795832
Median	71
Mode	98,49
Standard Deviation	19,06325843
Sample Variance	363,407822
Kurtosis	0,6262151917
Skewness	-1,126428603
Range	84,66
Minimum	13,83
Maximum	98,49
Sum	5744,96
Count	87
Largest(1)	98,49
Smallest(1)	13,83
Confidence Level(98%)	4,844804109

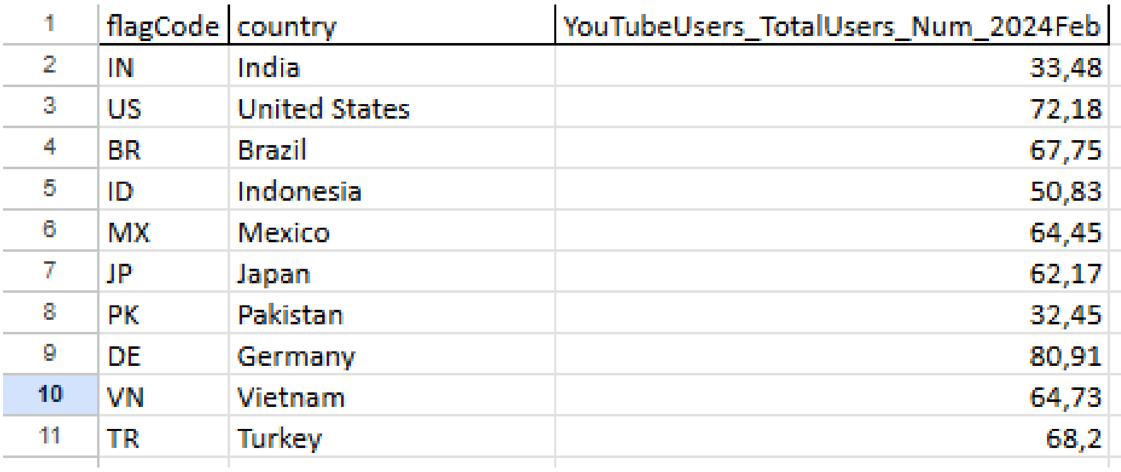
Based on these variables, I can do my further analysis.

1. In order to calculate 95% and 98% Confidence Intervals for the mean, I need to use formula: **Mean +- Confidence Level (95%, 98%)**
   We determine, that:

95% CI = (61.97; 70.09)
98% CI = (61,19; 70,87)
These 2 Confidence Intervals are NOT identical because:

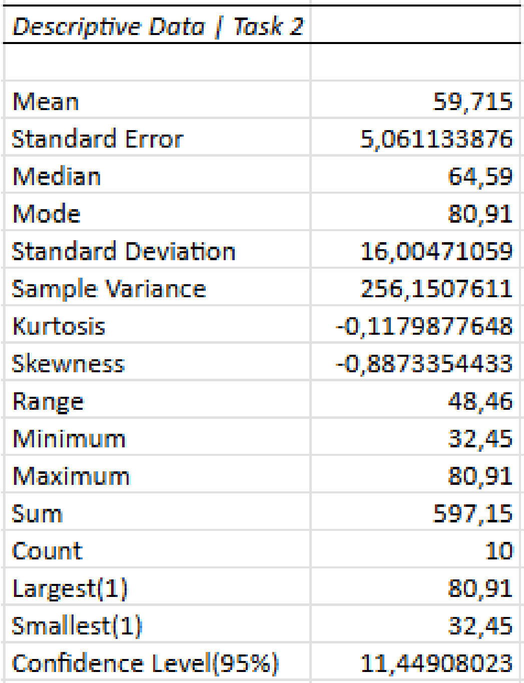
1) **Different confidence levels**: A 98% confidence interval requires greater certainty than a 95% interval, meaning we need to capture a larger portion of the sampling distribution.
2) **Different critical values**: The z-value (or t-value) increases as the confidence level increases. The 98% interval uses a larger z-value (2.33) compared to the 95% interval (1.96).
3) **Width difference**: The 98% confidence interval is wider than the 95% interval. This illustrates the fundamental trade-off in statistics: higher confidence requires wider intervals.
4) **Interpretation difference**: A 95% CI means that if we were to take 100 different samples and compute a confidence interval for each sample, about 95 of the 100 intervals would contain the true population mean. For a 98% CI, about 98 of 100 intervals would contain the true population mean.

2. For this task, I selected the following 10 points:

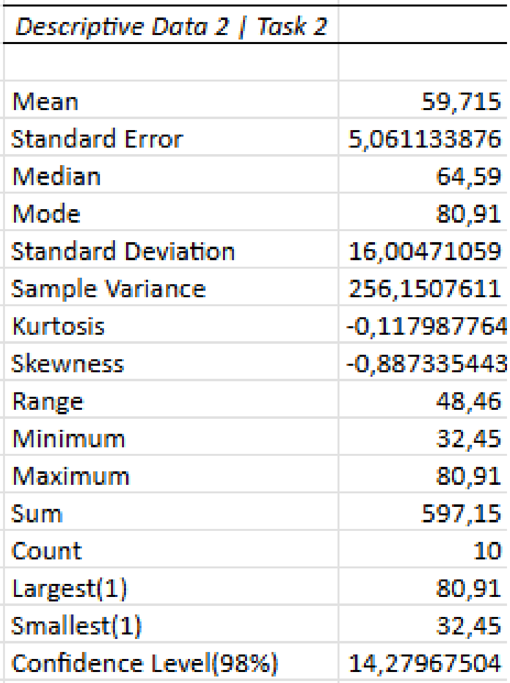


1	flagCode	country	YouTubeUsers_TotalUsers_Num_2024Feb
2	IN	India	33,48
3	US	United States	72,18
4	BR	Brazil	67,75
5	ID	Indonesia	50,83
6	MX	Mexico	64,45
7	JP	Japan	62,17
8	PK	Pakistan	32,45
9	DE	Germany	80,91
10	VN	Vietnam	64,73
11	TR	Turkey	68,2

Using XLMiner Extension, I have calculated 95% and 98% **Confidence Level** for the specific interval:



Descriptive Data | Task 2	
Mean	59,715
Standard Error	5,061133876
Median	64,59
Mode	80,91
Standard Deviation	16,00471059
Sample Variance	256,1507611
Kurtosis	-0,1179877648
Skewness	-0,8873354433
Range	48,46
Minimum	32,45
Maximum	80,91
Sum	597,15
Count	10
Largest(1)	80,91
Smallest(1)	32,45
Confidence Level(95%)	11,44908023



Descriptive Data 2 | Task 2	
Mean	59,715
Standard Error	5,061133876
Median	64,59
Mode	80,91
Standard Deviation	16,00471059
Sample Variance	256,1507611
Kurtosis	-0,117987764
Skewness	-0,887335443
Range	48,46
Minimum	32,45
Maximum	80,91
Sum	597,15
Count	10
Largest(1)	80,91
Smallest(1)	32,45
Confidence Level(98%)	14,27967504

Using the formula: **Mean +- Confidence Level (95%, 98%)**, I have calculated the 95% and 98% **Confidence Intervals:**
95% CI: (48.27; 71,15)
98% CI: (45,44; 73,98)

Why are they different?

1) **Different datasets**: The most fundamental reason is that we're using completely different data. The YouTube user counts (with mean ~59,7) are less than the original dataset (with mean ~66,03).
2) **Sample size effect**:
    - Original data had n=87 data points
    - Current data has only n=10 data points
    - With fewer data points, we have less precision (wider intervals)

Assumptions Check

Normality:

- Sample skewness: -0,88 (original: -1,12)
- Sample kurtosis: -0,11 (original: 0,62)

- The YouTube data is still left-skewed but less extremely than the original dataset

Outliers:

After calculating lower and upper bounds:
Lower bound = Q1 - 1.5 × IQR = 56.09 - 33.09 ≈ **23.00**

Upper bound = Q3 + 1.5 × IQR = 78.15 + 33.09 ≈ **111.24**

We can clearly detect the outliers below the lower bound

- Nigeria – 13.83
- Philippines – 19.49
- Bangladesh – 20.42
- Ghana – 20.13
- Kenya – 18.2
- Senegal – 20.65

But, there are non above upper bound

Independence:

- The data points represent different countries, so they can be considered independent

3. To calculate 95% and 98% CI for the standard deviation and variance, we need to first consider Descriptive Data for both 95% and 98%.
   We know, that:
   Sample Variance: 363,407822
   Sample Size: 87
   Degrees of Freedom: Sample Size - 1 = 86
   Confidence Level: 95%, 98%

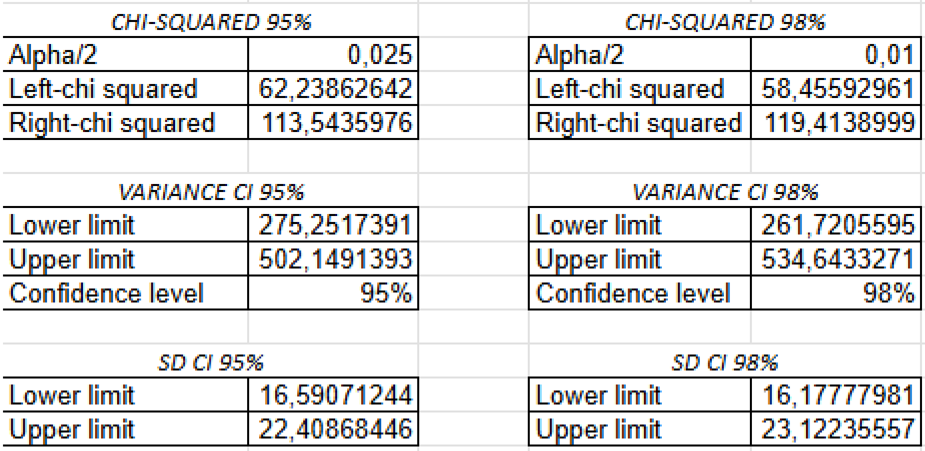
In order to calculate the confidence interval of variance I use chi-square values, which can be calculated by this formula in Excel: **CHISQ.INV.RT**

Left chi-squared value: 62,23862642
Right chi-squared value: 113,5435976

After calculating these values, we can now use the formula for CI for population Variance:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Using Excel, I have calculated the following results for 95% and 98%:



CHI-SQUARED 95%	
Alpha/2	0,025
Left-chi squared	62,23862642
Right-chi squared	113,5435976

CHI-SQUARED 98%	
Alpha/2	0,01
Left-chi squared	58,45592961
Right-chi squared	119,4138999

VARIANCE CI 95%	
Lower limit	275,2517391
Upper limit	502,1491393
Confidence level	95%

VARIANCE CI 98%	
Lower limit	261,7205595
Upper limit	534,6433271
Confidence level	98%

SD CI 95%	
Lower limit	16,59071244
Upper limit	22,40868446

SD CI 98%	
Lower limit	16,17777981
Upper limit	23,12235557

4. In order to test if the sample comes from a population where the mean = lower quartile (Q1), we firstly need to Identify sample values:

Sample mean (x) = 66,03402299
Standard deviation (s) = 19,06325843
Sample size (n) = 87
Lower quartile (Q1) = Lower Quartile 57,17 (Calculated in Excel)

Now, we can state our hypotheses:

Null (H_0): μ = Q1 (population mean equals lower quartile)

Alt (H_1): μ ≠ Q1

In order to perform the test, we use T-statistics formula:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

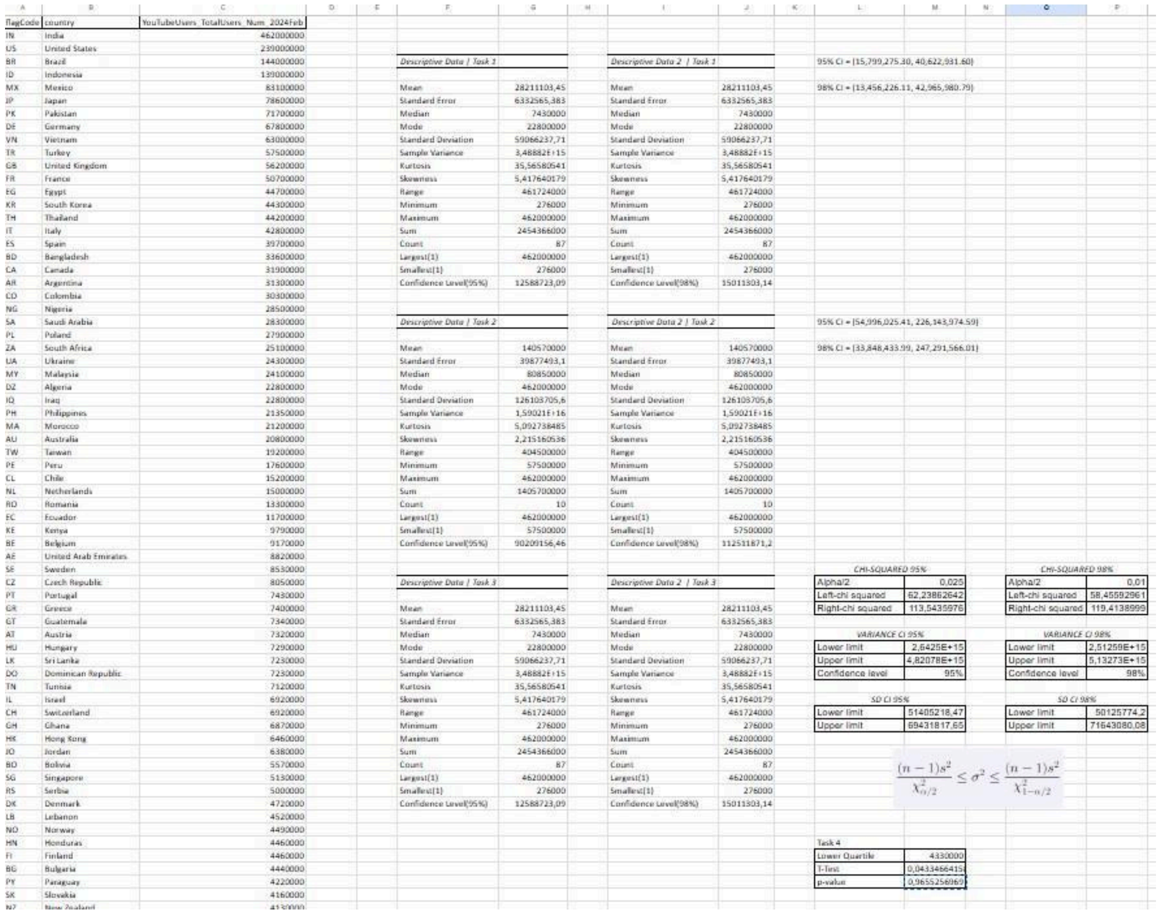
I calculated it using Excel, and found that **t = 0,04985102725**

When we are familiar with the t value, we need to find a two-tailed **p-value**. I found it using the following formula in Excel:
= 2 * T.DIST.RT(ABS(T); 86), which is = **0,9603567018**

Conclusion:
Since my p-value (0,96) is much greater than the common significance level (0.05), we fail to reject the null hypothesis. This means that there is no statistically significant difference between the population mean and the lower quartile. Therefore, we can conclude that the mean of the population is not significantly different from the lower quartile based on the sample data.

ATTACHMENTS

link to google sheets:
youtube-users-by-country-2025, Ivan Shvets



flagCode	country	YouTubeUsers_TotalUsers_Num_2024Feb
IN	India	462000000
US	United States	239000000
BR	Brazil	144000000
ID	Indonesia	139000000
MX	Mexico	83100000
JP	Japan	78600000
PK	Pakistan	71700000
DE	Germany	67800000
VN	Vietnam	63000000
TR	Turkey	57500000
GB	United Kingdom	56200000
FR	France	50700000
EG	Egypt	44700000
KR	South Korea	44300000
TH	Thailand	44200000
IT	Italy	42800000
ES	Spain	39700000
BD	Bangladesh	33600000
CA	Canada	31900000
AR	Argentina	31300000
CO	Colombia	30300000
NG	Nigeria	28500000
SA	Saudi Arabia	28300000
PL	Poland	27900000
ZA	South Africa	25100000
UA	Ukraine	24300000
MY	Malaysia	24100000
DZ	Algeria	22800000
IQ	Iraq	22800000
PH	Philippines	21350000
MA	Morocco	21200000
AU	Australia	20800000
TW	Taiwan	19200000
PE	Peru	17600000
CL	Chile	15200000
NL	Netherlands	15000000
RO	Romania	13300000
EC	Ecuador	11700000
KE	Kenya	9790000
BE	Belgium	9170000
AE	United Arab Emirates	8820000
SE	Sweden	8530000
CZ	Czech Republic	8050000
PT	Portugal	7430000
GR	Greece	7400000
GT	Guatemala	7340000
AT	Austria	7320000
HU	Hungary	7290000
LK	Sri Lanka	7230000
DO	Dominican Republic	7230000
TN	Tunisia	7120000
IL	Israel	6920000
CH	Switzerland	6920000
GH	Ghana	6870000
HK	Hong Kong	6460000
JO	Jordan	6380000
BO	Bolivia	5570000
SG	Singapore	5130000
RS	Serbia	5000000
DK	Denmark	4720000
LB	Lebanon	4520000
NO	Norway	4490000
HN	Honduras	4460000
FI	Finland	4460000
BG	Bulgaria	4440000
PY	Paraguay	4220000
SK	Slovakia	4160000
NZ	New Zealand	4130000

Descriptive Data / Task 1		*Descriptive Data 2 / Task 1*	
Mean	28211103,45	Mean	28211103,45
Standard Error	6332565,383	Standard Error	6332565,383
Median	7430000	Median	7430000
Mode	22800000	Mode	22800000
Standard Deviation	59066237,71	Standard Deviation	59066237,71
Sample Variance	3,48882E+15	Sample Variance	3,48882E+15
Kurtosis	35,56580541	Kurtosis	35,56580541
Skewness	5,417640179	Skewness	5,417640179
Range	461724000	Range	461724000
Minimum	276000	Minimum	276000
Maximum	462000000	Maximum	462000000
Sum	2454366000	Sum	2454366000
Count	87	Count	87
Largest(1)	462000000	Largest(1)	462000000
Smallest(1)	276000	Smallest(1)	276000
Confidence Level(95%)	12588723,09	Confidence Level(98%)	15011303,14

95% CI = [15,799,275.30, 40,622,931.60]

98% CI = [13,456,226.11, 42,965,980.79]

Descriptive Data / Task 2		*Descriptive Data 2 / Task 2*	
Mean	140570000	Mean	140570000
Standard Error	39877493,1	Standard Error	39877493,1
Median	80850000	Median	80850000
Mode	462000000	Mode	462000000
Standard Deviation	126103705,6	Standard Deviation	126103705,6
Sample Variance	1,59021E+16	Sample Variance	1,59021E+16
Kurtosis	5,092738485	Kurtosis	5,092738485
Skewness	2,215160536	Skewness	2,215160536
Range	404500000	Range	404500000
Minimum	57500000	Minimum	57500000
Maximum	462000000	Maximum	462000000
Sum	1405700000	Sum	1405700000
Count	10	Count	10
Largest(1)	462000000	Largest(1)	462000000
Smallest(1)	57500000	Smallest(1)	57500000
Confidence Level(95%)	90209156,46	Confidence Level(98%)	112511871,2

95% CI = [54,996,025.41, 226,143,974.59]

98% CI = [33,848,433.99, 247,291,566.01]

Descriptive Data / Task 3		*Descriptive Data 2 / Task 3*	
Mean	28211103,45	Mean	28211103,45
Standard Error	6332565,383	Standard Error	6332565,383
Median	7430000	Median	7430000
Mode	22800000	Mode	22800000
Standard Deviation	59066237,71	Standard Deviation	59066237,71
Sample Variance	3,48882E+15	Sample Variance	3,48882E+15
Kurtosis	35,56580541	Kurtosis	35,56580541
Skewness	5,417640179	Skewness	5,417640179
Range	461724000	Range	461724000
Minimum	276000	Minimum	276000
Maximum	462000000	Maximum	462000000
Sum	2454366000	Sum	2454366000
Count	87	Count	87
Largest(1)	462000000	Largest(1)	462000000
Smallest(1)	276000	Smallest(1)	276000
Confidence Level(95%)	12588723,09	Confidence Level(98%)	15011303,14

CHI-SQUARED 95%

Alpha/2	0,025
Left-chi squared	62,23862642
Right-chi squared	113,5435976

CHI-SQUARED 98%

Alpha/2	0,01
Left-chi squared	58,45592961
Right-chi squared	119,4138999

VARIANCE CI 95%

Lower limit	2,6425E+15
Upper limit	4,82078E+15
Confidence level	95%

VARIANCE CI 98%

Lower limit	2,51259E+15
Upper limit	5,13273E+15
Confidence level	98%

SD CI 95%

Lower limit	51405218,47
Upper limit	69431817,65

SD CI 98%

Lower limit	50125774,2
Upper limit	71643080,08

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Task 4

Lower Quartile	4330000
T-Test	0,0433466415
p-value	0,9655256969