

TB Data Project - Part 1

- A. In order to get 10 random countries for this project, I decided to use a simple random sampling method. First, I created a spreadsheet with a list of all 194 WHO countries. Then I added a second column with a unique number for each row (I used 1-194) and I used the following formula in 10 additional cells:

=VLOOKUP(RANDBETWEEN(1,194),A\$2:B\$195,2)

This formula chooses a random number within the given range and then finds that number's corresponding country. This method is not perfect because it might select duplicate countries, but in my case I did not get any duplicates:

D2 fx =VLOOKUP(RANDBETWEEN(1,194),A\$2:B\$195,2)

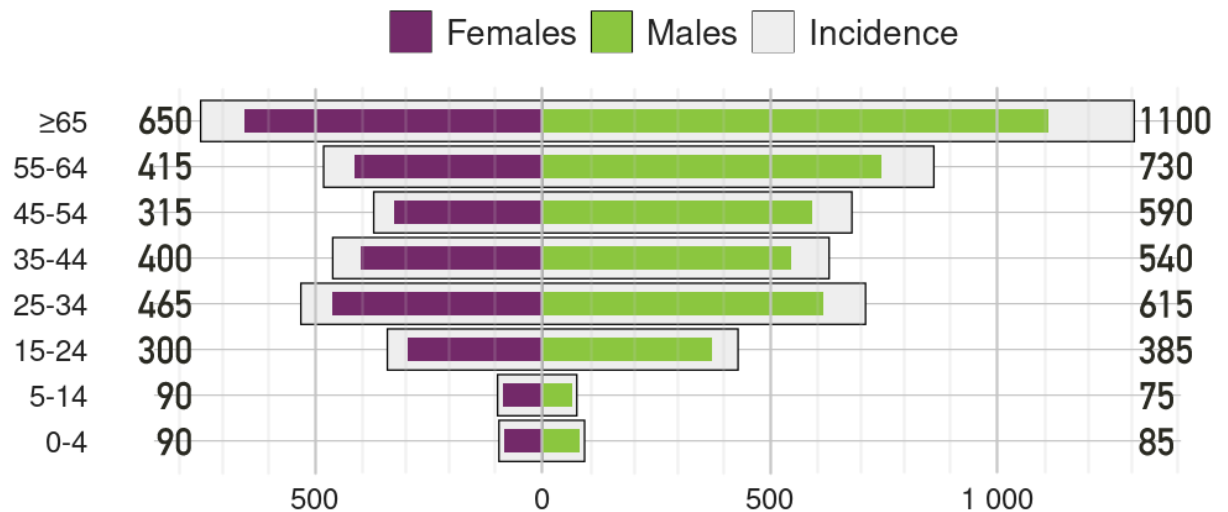
	A	B	C	D
1		WHO Countries		10 Random Entries
2	1	Afghanistan	1	Saint Vincent and the Grenadines
3	2	Albania	2	Gabon
4	3	Algeria	3	Bhutan
5	4	Andorra	4	Denmark
6	5	Angola	5	Sri Lanka
7	6	Antigua and Barbuda	6	Cabo Verde
8	7	Argentina	7	Kiribati
9	8	Armenia	8	Russian Federation
10	9	Australia	9	Luxembourg
11	10	Austria	10	United States of America
12	11	Azerbaijan		
13	12	Bahamas		
14	13	Bahrain		
15	14	Bangladesh		

- B. I have listed total TB incidence rate per 100k population, success rate of treatment, and cohort size for my 10 countries:

Country	Total TB Incidence / 100k population	Success Rate	Cohort Size
Grenadines	6.7	100%	4
Gabon	527	67%	5,399
Bhutan	165	94%	937
Denmark	4.9	45%	259
Sri Lanka	64	85%	8,186

Country	Total TB Incidence / 100k population	Success Rate	Cohort Size
Cabo Verde	39	89%	203
Kiribati	425	92%	410
Russia	46	68%	54,589
Luxembourg	5.9	36%	47
United States	2.4	75%	8,406

- C. I modified the Tuberculosis profile chart for the United States of America from the WHO:TB page by adding grid marks for every 100 cases and estimating totals of female/male notified cases per age group:



Incidence, Notified cases by age group and sex, 2020, United States of America

1. The most-notified female cases belong to the ≥ 65 years age group. The estimated relative frequency is $650 / 2725$, or about **0.239** (23.9%) of all female notified cases.
2. The least-notified male cases belong to the 5-14 years age group, with an estimated relative frequency of only $75 / 4120$, or about **0.018** (1.8%) of all male notified cases.
3. If there were 5000 total notified cases, we could estimate how many individuals would be in each of the above categories with the following calculations:
 - a. The incidence of female notified cases is roughly 39.8% of the total notified cases ($2725 / 6845$), which makes the incidence of male notified cases roughly 60.2%.

- b. 39.8% of 5000 notified cases is 1990 female notified cases, of which the ≥ 65 age group's 23.9% would be **475.61**.
- c. 60.2% of 5000 notified cases is 3010 male notified cases, of which the 5.14 age group's 1.8% would be **54.18**.

D. The success rate of treatment for Saint Vincent and the Grenadines is listed as 100%, making for an interesting simulation:

Describe process:

Probability of success (π):

Sample size (n):

Number of samples:

☐ Show animation

Total Samples = 50

Choose statistic:

- ☒ Number of successes
☐ Proportion of successes

Count samples

As extreme as

Proportion of samples:
 $(0 + 0) / 50 = 0$

Options:

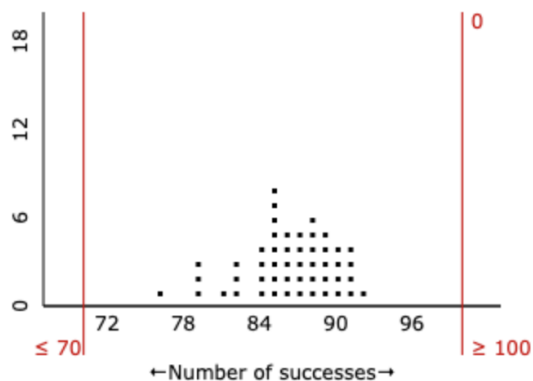
☒ Two-sided (between: ☐)

Most recent results

Number of Successes = 79

Number of Failures = 21

☐ Summary Statistics



☐ Show previous results

☐ Show sliders

I'll choose a conservative alpha value of 0.01, which is greater than my p-value of 0.00.

My hypotheses are the following:

$$H_0 : p = 0.85$$

$$H_A : p \neq 0.85$$

The p-value is less than alpha, which means we will reject our null hypothesis: We have sufficient evidence to conclude that the successful treatment rate of TB in Saint Vincent and the Grenadines is not 85%.

Part 2





A. I would like to find the true proportion of the success rate of TB treatment for both Luxembourg and the United States. I'll check each country's data against the three conditions to compute a 95% confidence interval:

1. The data is from a random sample or randomized experiment
2. The sample size is independent (5% or less of the whole population)
3. There are at least 10 successes and 10 failures





I will use the below table to calculate conditions 2 and 3 for each country:

Country	Total TB Incidence / 100k population	Success Rate (\hat{p})	Cohort Size (n)
Luxembourg	5.9	36%	47
United States	2.4	75%	8,406

Luxembourg

1. The data is from a random sample 
2. $n \leq 0.05N$
 $47 \leq 0.05(620001^*)$
 $47 \leq 31000.05$ 
3. $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$
 - a. $n\hat{p} \geq 10$
 $47(0.36) \geq 10$
 $16.92 \geq 10$ 
 - b. $n(1 - \hat{p}) \geq 10$
 $47(1 - 0.36) \geq 10$
 $47(0.64) \geq 10$
 $30.08 \geq 10$ 

United States

1. The data is from a random sample 
2. $n \leq 0.05N$
 $8406 \leq 0.05(328329953^*)$
 $8406 \leq 16416497.65$ 
3. $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$
 - a. $n\hat{p} \geq 10$
 $8406(0.75) \geq 10$
 $6304.5 \geq 10$ 
 - b. $n(1 - \hat{p}) \geq 10$
 $8406(1 - 0.75) \geq 10$
 $8406(0.25) \geq 10$
 $2101.5 \geq 10$ 

(* Population data source: <https://data.worldbank.org/indicator/>)

I believe that this will lead to valid confidence intervals because my data passes the above conditions.

B. I have used Geogebra to calculate a 95% confidence interval for both countries:

Z Estimate of a Proportion ▼

Confidence Level

Sample

Successes

N

Z Estimate of a Proportion

Successes	17
N	47
<u>Result</u> SE	0.0701
Lower Limit	0.2243
Upper Limit	0.4991
Interval	0.3617 ± 0.1374

I am 95% confident that the interval 0.223 to 0.499 contains the true proportion of the success rate of TB treatment for Luxembourg.

Z Estimate of a Proportion ▼

Confidence Level

Sample

Successes

N

Z Estimate of a Proportion

Successes	6305
N	8406
<u>Result</u> SE	0.0047
Lower Limit	0.7408
Upper Limit	0.7593
Interval	0.7501 ± 0.0093

I am 95% confident that the interval 0.741 to 0.759 contains the true proportion of the success rate of TB treatment for the United States.

C. The global success rate of TB treatment is defined to be 85%. Neither of the confidence intervals of my two countries include this number, with the 95% confidence interval of Luxembourg calculated to be 0.2243 to 0.4991 and the 95% confidence interval of the

United States calculated to be 0.7408 to 0.7593. This leads me to believe that the true proportion of successful TB treatment for both of these countries is not 85%.

D. Now we will check if Saint Vincent and the Grenadines meets the conditions for carrying out a two-sided test of whether $p = 0.85$ or not:

1. The data is from a random sample ☒

2. $n \leq 0.05N$

$$4 \leq 0.05(110593^*)$$

$$4 \leq 5529.65 \quad \checkmark$$

3. $np_0 \geq 10$ and $n(1 - p_0) \geq 10$

a. $np_0 \geq 10$

$$4(0.85) \geq 10$$

$$3.4 \geq 10 \quad \times$$

b. $n(1 - p_0) \geq 10$

$$4(1 - 0.85) \geq 10$$

$$4(0.15) \geq 10$$

$$0.6 \geq 10 \quad \times$$

I believe that the results I obtain in the hypothesis test will be invalid because my sample size is too small, leading both successes and failures to be below 10.

E. I'll conduct a hypothesis test by first plugging in information about my hypotheses, successes, and sample size into Geogebra:

Z Test of a Proportion

Null Hypothesis $p =$

Alternative Hypothesis ☐ $<$ ☐ $>$ ☒ \neq

Sample

Successes

N

Z Test of a Proportion

Successes	4
Result N	4
Z	0.8401680504168
P	0.4008141693829

Two-sided hypothesis test of whether $p = 0.85$ or not

Then I will evaluate my P-value against my chosen alpha value of 0.05. Because 0.40 is greater than 0.05, I do not have sufficient evidence that the success rate of TB treatment in the Grenadines is not 85%, and I fail to reject my null hypothesis. In context, I can make the conclusion that there is not enough evidence to say that the successful treatment rate of TB in the Grenadines is different than the successful treatment rate

worldwide.

F. In Part I, I simulated a hypothesis test using the following experiment:

Describe process:

Probability of success (π):

Sample size (n):

Number of samples:

☐ Show animation

Total Samples = 50

Choose statistic:

- ☒ Number of successes
☐ Proportion of successes

Count samples

As extreme as

Proportion of samples:
 $(0 + 0) / 50 = 0$

Options:

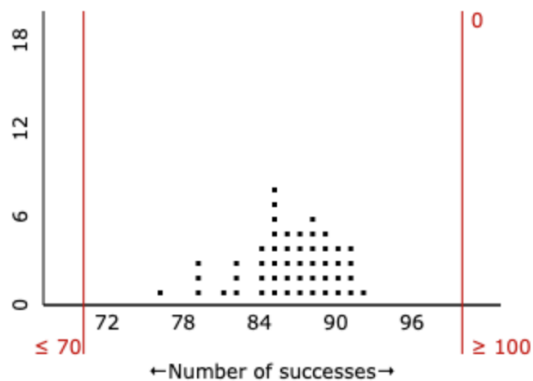
☒ Two-sided (between:)

Most recent results

Number of Successes = 79

Number of Failures = 21

☐ Summary Statistics



☐ Show previous results

☐ Show sliders

In the previous experiment I had a sample size of 100 and a P-value of 0, while my new experiment with a sample size of 4 resulted in a P-value of 0.40. These results are wildly different but not wholly unexpected. We went into both experiments knowing that our original success rate and sample size were likely not a good estimate of the whole population (mostly due to the small sample size), which we know can lead to erratic and likely incorrect results.

I believe that with good data, both methods are valid in testing a hypothesis. I personally prefer the method that simply uses calculations instead of running simulations because I believe that there's a possibility that with a simulation you can get misleading results if you use too few samples.

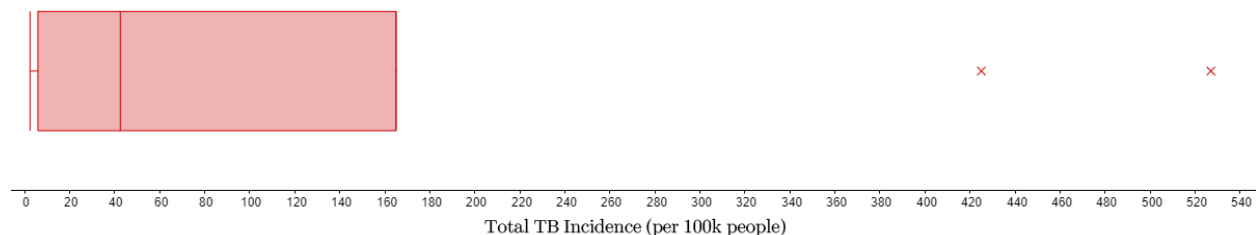
Part 3

A. Below is a table showing the total TB incidence for 10 randomly selected countries:

Country	Total TB Incidence / 100k population
Grenadines	6.7
Gabon	527
Bhutan	165
Denmark	4.9
Sri Lanka	64
Cabo Verde	39
Kiribati	425
Russia	46
Luxembourg	5.9
United States	2.4

I have used these values to generate a boxplot graph:

Total TB Incidence in 10 Sampled Countries



B. The graph representing the total TB incidence in 10 countries shows that a majority of the values are on the left side, making it right-skewed. Based on this information, we know that the median is going to be the best measure of center while the IQR will be the best measure of spread. These values are as follows:

Median: 42.5

IQR: 159.1


C. We can calculate the fences for our data by subtracting $(1.5 * \text{IQR})$ from Q1 and adding $(1.5 * \text{IQR})$ to Q3:

Lower fence: $5.9 - (1.5 * 159.1) = -232.75$

Upper fence: $165 + (1.5 * 159.1) = 403.65$

We do not have any values below our lower fence. However, we have two values higher than our upper fence of 403.65: 425 (Kiribati) and 527 (Gabon). These are our outliers.

D. We would like to use the normal model to compute a confidence interval or perform a hypothesis test for μ . Let's check the three conditions:

1. The data is from a random sample 
2. We will combine the cohort sizes of all of our countries to get n:


Country	Cohort Size
Grenadines	4
Gabon	5,399
Bhutan	937
Denmark	259
Sri Lanka	8,186
Cabo Verde	203
Kiribati	410
Russia	54,589
Luxembourg	47
United States	8,406
Total	78,440

The population value N is 10,000,000, or the estimated total number of worldwide TB incidences in 2019 (Source:

<https://www.sciencedirect.com/science/article/pii/S1201971221001934>).


$n \leq 0.05N$

$78440 \leq 0.05(10000000)$

$78440 \leq 500000$ 

3. The sample size n is either ≥ 30 or the population size N is normal. We don't know whether the population size is normal, so we'll check the size of n:

$n \geq 30$

$78440 \geq 30$ 

E. We will compute a 95% confidence interval for our data:

T Estimate of a Mean ▼

Confidence Level

Sample

Mean

s

N

T Estimate of a Mean

Mean	128.59
s	190.9239
SE	60.3754
<u>Result</u> N	10
df	9
Lower Limit	-7.9887
Upper Limit	265.1687
Interval	128.59 ± 136.5787

With these calculations, we are 95% confident that the interval -7.99 to 265.17 contains the true mean of TB incidence per 100k people.

F. Now, we want to determine if our data shows that the global incidence rate of TB in the world is different from the reported value of 132. We will set up our hypotheses as follows:

$$H_0 : \mu = 132$$

$$H_A : \mu \neq 132$$

We will carry out our test using Geogebra's T Test of a Mean function:

T Test of a Mean ▼

Null Hypothesis $\mu =$

Alternative Hypothesis ☐ $<$ ☐ $>$ ☒ \neq

Sample

Mean

s

N

T Test of a Mean

Mean	128.59
s	190.9239
SE	60.3754
N	10
df	9
t	-0.0565
P	0.9562

Result

Our test statistic is 128.59 and the resulting p-value is 0.9562. If we choose a common alpha value of 0.05 and compare it to our p-value, we will fail to reject the null hypothesis: We do not have sufficient evidence to show that the mean global TB incidence is not 132 per 100k people.

Part 4

- A. Below is a contingency table showing the treatment success rate for 10 randomly selected WHO countries:

		Treatment		TOTAL
		Success	Failure	
Member of WHO	Grenadines	100	0	100
	Gabon	67	33	100
	Bhutan	94	6	100
	Denmark	45	55	100
	Sri Lanka	85	15	100
	Cabo Verde	89	11	100
	Kiribati	92	8	100
	Russia	68	32	100
	Luxembourg	36	64	100
	United States	75	25	100
TOTAL		751	249	1000

- B. Using this table, we will calculate the following probabilities:

1. Probability that a randomly selected case is from Sri Lanka or Cabo Verde:

$$\frac{200}{1000} = 0.2$$

2. Probability that a randomly selected case is from Sri Lanka or is a failure:

$$\frac{(100 + 249 - 15)}{1000} = \frac{334}{1000} = 0.334$$

3. Probability that a randomly selected case is from Sri Lanka and is a failure:

$$\frac{15}{1000} = 0.015$$

4. Probability that a randomly selected case is from Cabo Verde, given that it is a failure:

$$\frac{11}{249} = 0.044$$

5. Probability that three randomly selected cases are all successes from Russia:

$$\frac{68}{1000} \cdot \frac{67}{999} \cdot \frac{66}{998} = \frac{300696}{997002000} = 0.0003$$

- C. We will now compute a confidence interval for our data:

Z Estimate, Difference of Proportions v

Confidence Level

Sample 1	Sample 2
Successes <input type="text" value="85"/>	Successes <input type="text" value="89"/>
N <input type="text" value="100"/>	N <input type="text" value="100"/>

Z Estimate, Difference of Proportions

	Sample 1	Sample 2
Successes	85	89
N	100	100
SE	0.0474763098819	
Lower Limit	-0.1330518574873	
Upper Limit	0.0530518574873	
Interval	-0.04 ± 0.0930518574873	

We are 95% confident that the interval -0.133 to 0.053 contains the true difference of proportions of successful treatment for TB in Sri Lanka and Cabo Verde.

- D. Now we will conduct a hypothesis test to determine if there is a difference in the proportions of successful treatment for Sri Lanka and Cabo Verde. Our test parameters are as follows:

$$H_0 : P_{Sri-Lanka} = P_{Cabo-Verde}$$

$$H_A : P_{Sri-Lanka} \neq P_{Cabo-Verde}$$

Our test shows the following:

1. Test statistic: -0.841
2. P-value: 0.4

3. Alpha: 0.05
4. Conclusion: Our p-value is larger than alpha, so we will fail to reject the null hypothesis. We have insufficient evidence to show that there is a difference in the successful treatment for TB in Sri Lanka and Cabo Verde.

Z Test, Difference of Proportions v

Null Hypothesis $p_1 - p_2 =$

Alternative Hypothesis ☐ $<$ ☐ $>$ ☒ \neq

Sample 1

Sample 2

Successes

Successes

N

N

Z Test, Difference of Proportions

	Sample 1	Sample 2
Successes	85	89
N	100	100
SE	0.0475604878024	
Z	-0.8410342670624	
P	0.4003287377162	

Result