

Data Project

Part I

Out of 194 countries from the World Health Organization (WHO) 10 were randomly chosen. Those 10 members include:

- Angola
- Ireland
- Timor-Leste
- Suriname
- Czechia
- Brazil
- Lesotho
- Slovakia
- Guyana
- Albania

The sampling method used was a simple random sampling. A random number generator function was used to generate 10 values between 0 and 193. Each value generated mapping to a country in WHO in alphabetical order. That is to say, where $WHO = (\text{Afghanistan, Albania, Algeria, ..., Zimbabwe})$, a value of 1 would correspond to $WHO_1 = \text{Albania}$.

Data regarding tuberculosis for each of these 10 countries is laid out in the following table:

Country	Total TB Incidence	Success Rate	Cohort Size
Angola	325	53%	64,859
Ireland	4.8	6%	223
Timor-Leste	486	91%	3,225
Suriname	29	75%	103
Czechia	3.9	69%	347
Brazil	48	67%	72,825
Lesotho	614	76%	4,478
Slovakia	2.8	86%	152
Guyana	83	67%	358
Albania	17	89%	240

We consider Brazil with regards to notified cases by age group and sex using Graph 1 as a reference. An estimation for the sum of all the values for females from top to bottom would be:

$$Total = 2800 + 2700 + 3000 + 4000 + 5000 + 3900 + 500 + 200 = 22100$$

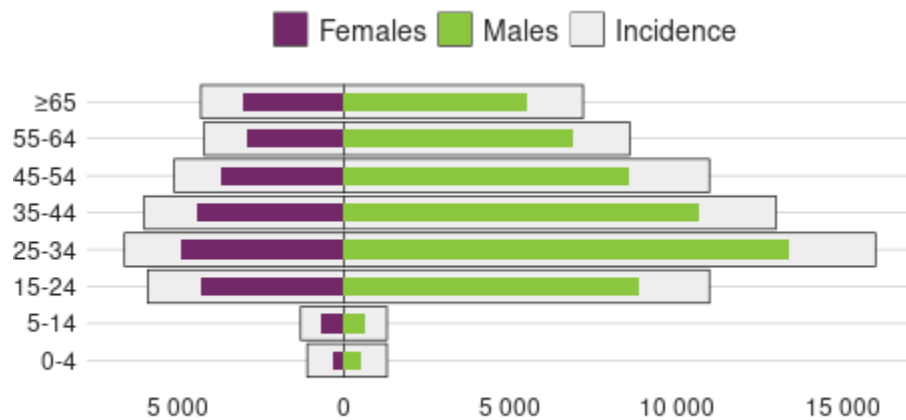
The age group for females that had the most notified cases (5000) would be ages 25-34. The relative frequency would then be $5000/22100 \approx 0.226$.

An estimation for the sum of all the values for males from top to bottom would be:

$$Total = 6000 + 7000 + 8000 + 10000 + 13000 + 8100 + 500 + 300 = 52900$$

The age group for males that had the least notified cases (300) would be ages 0-4. The relative frequency would then be $300/52900 \approx 0.006$.

If there were 5000 total notified cases for each age group, females ages 25-34 would amount to $5000 \times 0.226 = 1130$. As for males ages 0-4, they would amount to $5000 \times 0.006 = 30$.



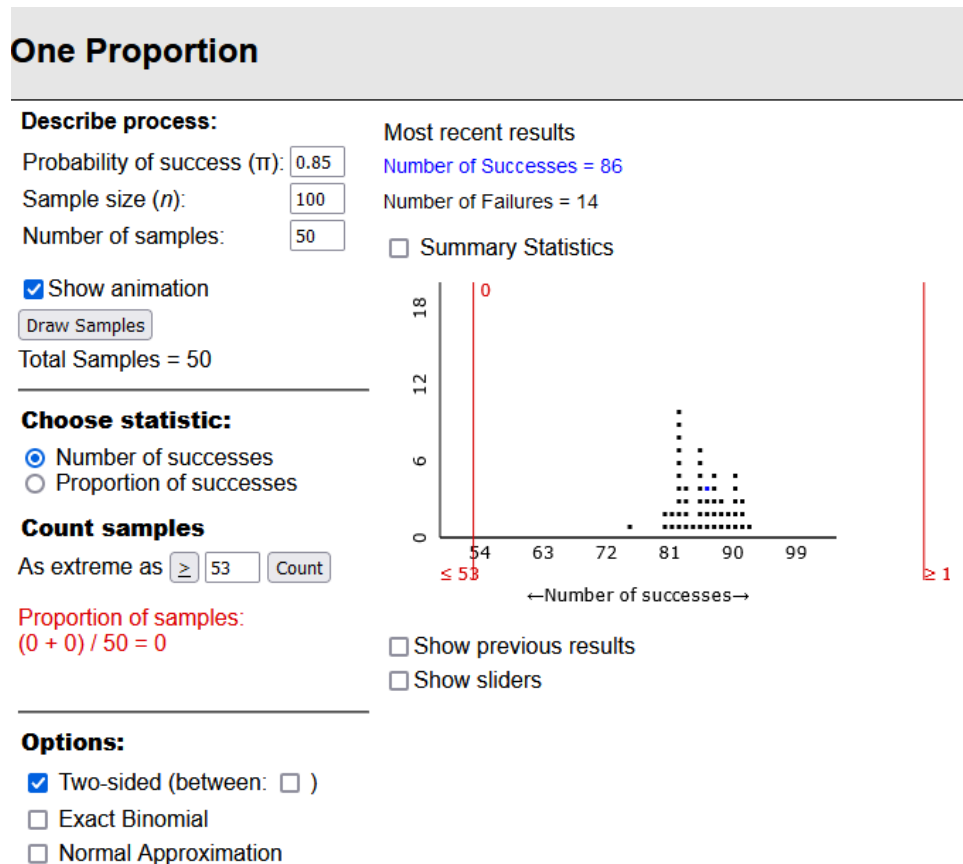
Graph 1. Incidence, Notified cases by age group and sex, 2021

We would like to consider the first country in our sample of WHO, Angola. The purpose being to gauge how Angola treats tuberculosis with a relation to a threshold of 85%. To do this, we lay out the null and alternate hypothesis formally:

$$H_0 : p = 0.85$$

$$H_A : p \neq 0.85$$

With $\alpha = 0.05$ as our alpha value, all we need is the two-sided p-value which can be obtained from the following simulation:



Based on this simulation, the two-sided p-value is found to be 0. Since 0 is less than our alpha value of 0.05 we conclude by saying Angola does not meet the threshold and therefore the alternative hypothesis is accepted, which in this particular case would mean that Angola has TB rate that is different from the 85% global threshold.

Part II

The following discussion will involve two countries chosen from the list in Part I, the two countries being Lesotho and Czechia. First, we'll verify if the three conditions are met in order to proceed with computing a confidence interval.

Lesotho	Czechia
Random Sample	Random Sample
$4478 \leq 0.05N \Rightarrow N \geq 89,560$	$347 \leq 0.05N \Rightarrow N \geq 6,940$
$4478(0.76) \approx 3403 \geq 10$ $4478(1 - 0.76) \approx 1075 \geq 10$	$347(0.69) \approx 239 \geq 10$ $347(1 - 0.69) \approx 108 \geq 10$

We cannot fully trust the confidence interval since the total TB incidence for both Lesotho and Czechia are less than what N must be for each. For Lesotho the total TB incidence is around 14,000 and for Czechia 400.

Computing for the confidence interval:

Z Estimate of a Proportion ▾

Confidence Level 0.95

Sample

Successes 3403

N 4478

Result


Z Estimate of a Proportion

Successes	3403
N	4478
SE	0.0064
Lower Limit	0.7474
Upper Limit	0.7724
Interval	0.7599 ± 0.0125

Z Estimate of a Proportion ▾

Confidence Level 0.95

Sample

Successes 239 

N 347

Result

Z Estimate of a Proportion

Successes	239
N	347
SE	0.0249
Lower Limit	0.64
Upper Limit	0.7375
Interval	0.6888 ± 0.0487

The one on the left relates to Lesotho and the one right to Czechia. What we can interpret from these computations is that we can say with 95% confidence that the interval between (0.7474, 0.7724) captures the true proportion of success in treating tuberculosis in Lesotho. Likewise, we can say with 95% confidence that the interval between (0.64, 0.7375) captures the true proportion of success in treating tuberculosis in Czechia.

The PLOS One report identified the global threshold for successful treatment to be 85%, with this in mind we can say that this success rate is not a likely value for either Lesotho or Czechia since the upper limit of success rates for both falls below 0.85.

Going back to our discussion of Angola, we would like to conduct a hypothesis test but first we check to see if Angola meets the three conditions:

Angola
Random Sample
$64,859 \leq 0.05N \Rightarrow N \geq 1,297,180$
$64,859(0.85) \approx 55,130 \geq 10$ $64,859(1 - 0.85) \approx 9,879 \geq 10$

The requirement that $N \geq 1,297,180$ means that hypothesis test will not be valid since the total tuberculosis incidents in Angola is 112,000.

We state the hypothesis:

$$H_0: p = 0.85$$

$$H_A: p \neq 0.85$$


Using Geogebra to find the relevant values:

Z Test of a Proportion ▼

Null Hypothesis $p =$ 0.85

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

Sample

Successes 34755 

N 64859

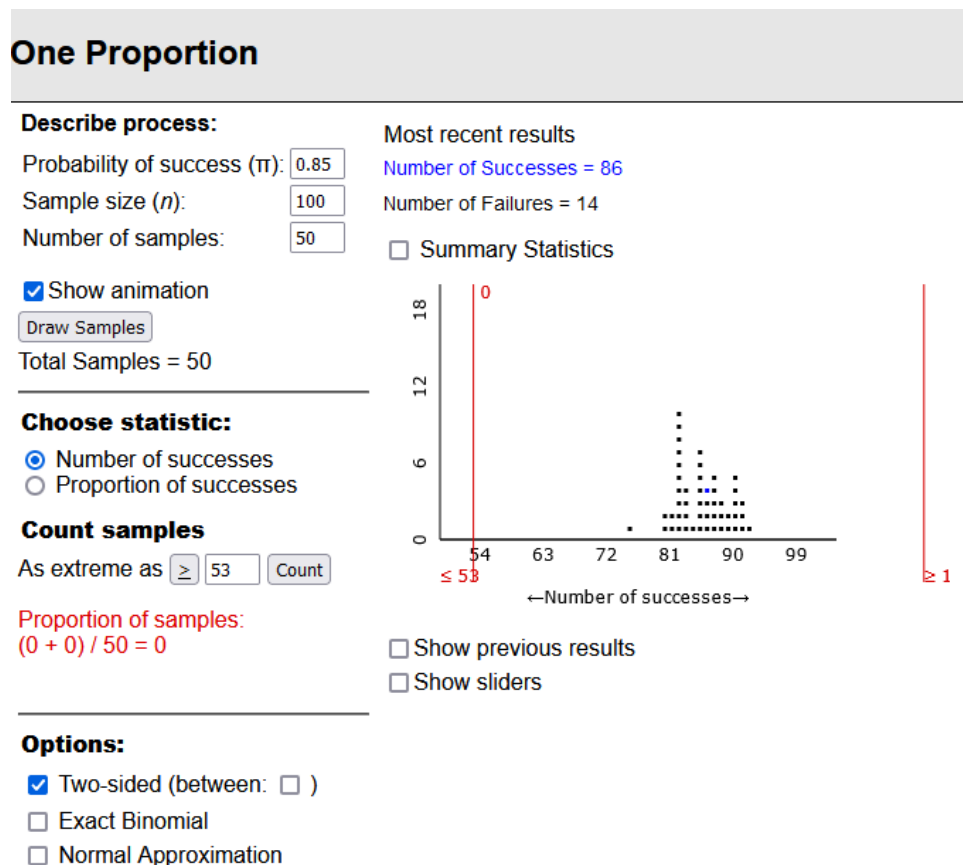
[Result](#)

Z Test of a Proportion

Successes	34755
N	64859
Z	-224.0581
P	0

From the calculations we can see that test statistic is roughly -224.058 and the p-value is 0. In order to move forward, we can choose an alpha value of 0.05 to test our p-value with. Since $0 < 0.05$ we can conclude that there is sufficient evidence that shows that Angola does not meet the threshold of 85% with respect to treating tuberculosis.

Let's compare this hypothesis test to the one using the simulation from Part I. From Part I the results were:

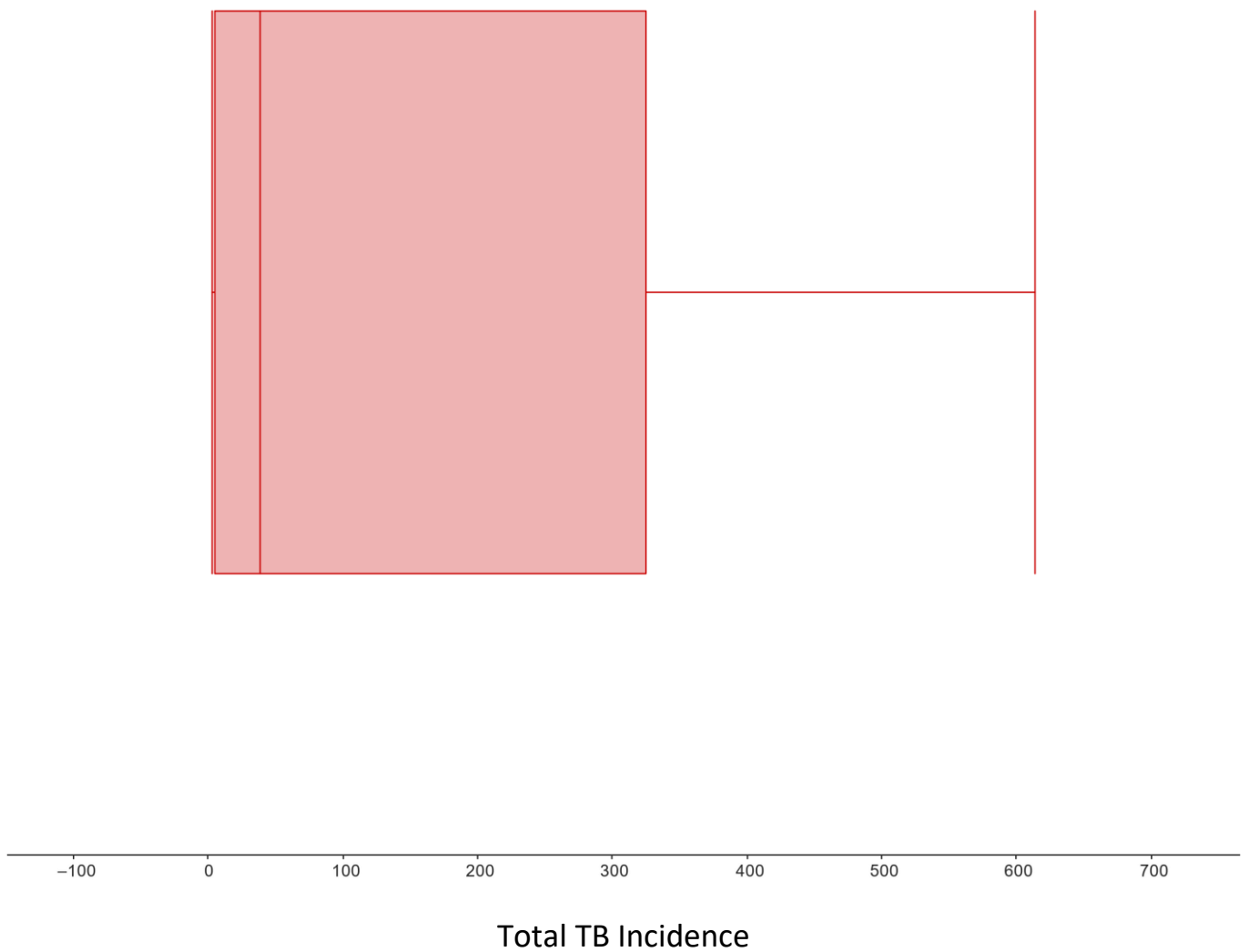


One of the most glaring differences is in the sample sizes. The sample size of the simulation from Part I was 100 vs 64,859 in the hypothesis just performed. The p-value for both tests results in the same value of 0, this of course implies that they both result in the same conclusion. I think the hypothesis we just performed is more valid than the one performed in Part I due to a significantly larger sample size.

Part III

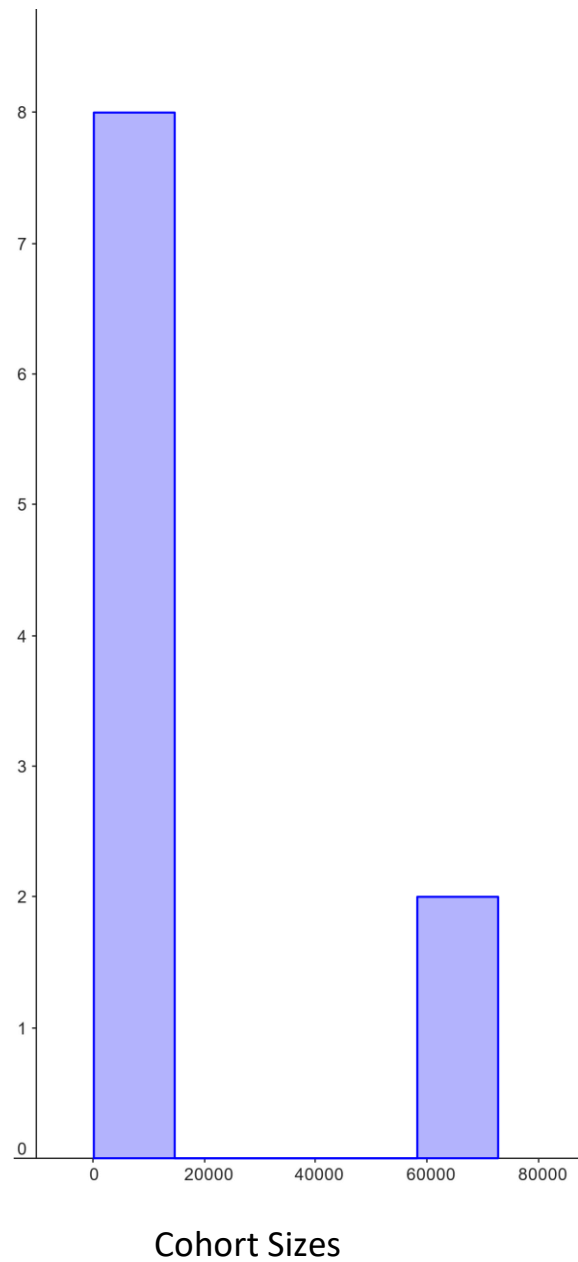
The graph below shows the total TB incidence for the sampled 10 members of WHO:

Total TB Incidence in 10 Sampled Countries



The graph below shows the cohort size for the sampled 10 members of WHO:

Cohort Sizes in 10 Sampled Countries



The shape of the graph of the Total TB incidence graph is right skewed for this distribution. The center or median in this case is 38.5 and the spread or IQR is 320.2.

In order to figure out if there are any outliers in this data set we first compute the upper and lower fences.

$$Upper Fence = 325 + 480.3 = 805.3$$

$$Lower Fence = 4.8 - 480.3 = -475.5$$

We note that we have no value in our data set that goes beyond our fences so there is no outliers in our data set.

Proceeding with calculating for the confidence interval will require checking the three conditions first. First, since we chose the 10 countries randomly, this can be considered a random sample. Next, we check to see if the observations are independent using the sum of the cohort values:

$$\begin{aligned} n &= 64,859 + 223 + 3,225 + 103 + 347 + 72,825 + 4,478 + 152 + 358 + 240 \\ &= 146810 \end{aligned}$$

$$146810 \leq 0.05N$$

$$2936200 \leq N$$

The observations are independent because the sample size is less than 5% of the population size, the population size N being the population of the whole world. The sample size is also large enough because $n \geq 30$. Therefore, the conditions are met and the proceeding results will be valid.


Moving on to the confidence interval for total TB incidence, we can use Geogebra to compute it:

T Estimate of a Mean ▼

Confidence Level 0.95

Sample

Mean 161.35

s 228.24 

N 10

Result

T Estimate of a Mean

Mean	161.35
s	228.24
SE	72.1758
N	10
df	9
Lower Limit	-1.9231
Upper Limit	324.6231
Interval	161.35 ± 163.2731

We can conclude by saying that we are 95% confident that the interval between (-1.92, 324.62) captures the population mean of total TB incidence.

Using the CDC's claim that the global incidence rate of TB is 132, we want to determine if the global incidence rate of TB in the world is different than the reported value of 132. To do this we first start by stating the null and alternate hypothesis:

$$H_0 : \mu = 132$$

$$H_A : \mu \neq 132$$


The results can be computed using Geogebra:

T Test of a Mean ▼

Null Hypothesis $\mu =$ 132

Alternative Hypothesis ☐ < ☐ > ☒ \neq

Sample

Mean 161.35 

s 228.24

N 10

[Result](#)

T Test of a Mean

Mean	161.35
s	228.24
SE	72.1758
N	10
df	9
t	0.4066
P	0.6938

We can see from the results that the test statistic = 0.4066 and the p-value = 0.6938. An alpha value of 0.05 will suffice for this hypothesis test. Since our p-value of $0.6938 > \alpha = 0.05$, we fail to reject the null hypothesis and say that there is not sufficient evidence to accept the claim the global incidence rate of TB is different than 132.

Part IV

Contingency table for the success rate data can be summarized using the following table:

Member of WHO	Treatment Success/Failure		
Angola	53	47	100
Ireland	6	94	100
Timor-Leste	91	9	100
Suriname	75	25	100
Czechia	69	31	100
Brazil	67	33	100
Lesotho	76	24	100
Slovakia	86	14	100
Guyana	67	33	100
Albania	89	11	100
Total	679	321	1000

We would like to compute some probabilities using the contingency table from above.

A randomly selected case is from the 5th or 6th member of WHO in your table:

$$P(5^{\text{th}} \text{ or } 6^{\text{th}}) = 1/10 + 1/10 = 0.2$$

A randomly selected case is from the 5th member of WHO in your table or is a failure:

$$P(5^{\text{th}} \text{ or Failure}) = 1/10 + 321/1000 - 31/1000 = 0.39$$

A randomly selected case is from the 5th member of WHO in your table and is a failure:

$$P(5^{\text{th}} \text{ and Failure}) = 1/10 * 31/100 = 0.031$$

A randomly selected case is from the 6th member of WHO in your table, given it is a failure:

$$P(6^{\text{th}} \mid \text{Failure}) = 33/321 \approx 0.1028$$


Three randomly selected cases (without replacement) are all successes from the 8th member of WHO in your table:

$$86/100 * 85/99 * 84/98 \approx 0.6329$$

We can use Geogebra to compute a 95% confidence interval, let's consider the 5th and 6th members of WHO (Czechia and Brazil) for this exercise:

Z Estimate, Difference of Proportions ▾

Confidence Level 0.95

Sample 1		Sample 2	
Successes	<u>69</u>	Successes	<u>67</u>
n	<u>100</u>	n	<u>100</u> 

[Result](#)

Z Estimate, Difference of Proportions

	Sample 1	Sample 2
Successes	69	67
n	100	100
SE	0.066	
Lower Limit	-0.1093	
Upper Limit	0.1493	
Interval	0.02 ± 0.1293	

Based on the outputs we can say that we are 95% confident that the difference of proportions for successes considered in the two countries (Czechia and Brazil), is captured by the interval (-0.1093, 0.1493).

Using the same two countries as above, we now turn our attention towards a hypothesis test for whether there is a difference in the proportions of successful treatment for the two members of WHO. We begin by stating the two hypotheses:

$$H_0: p_{Czechia} = p_{Brazil}$$

$$H_A: p_{Czechia} \neq p_{Brazil}$$

We use Geogebra for the computations:

Z Test, Difference of Proportions ▼

Null Hypothesis $p_1 - p_2 =$

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

Sample 1		Sample 2	
Successes	<input type="text" value="69"/>	Successes	<input type="text" value="67"/>
n	<input type="text" value="100"/>	n	<input type="text" value="100"/>

Result

Z Test, Difference of Proportions

	Sample 1	Sample 2
Successes	69	67
n	100	100
SE	0.066	
z	0.3032	
p	0.7618	

For this test an alpha value of 0.05 is sufficient. We note that the test statistic is 0.3032 and the p-value is 0.7618. Since our p-value is greater than our alpha value, we fail to reject the null hypothesis and conclude by saying that there is not sufficient evidence to support the claim that there is a difference in the proportions of successful treatment between the two members of WHO.