CHAPTER 27 – THE T-TEST

Small sample hypothesis testing

We are still hypothesis testing in chapter 27.

The difference is, we are now testing hypothesis under conditions that may not warrant using a Z-score. Therefore, we do the same work as before but under a different distribution than the Z; we now look at t-scores.

page 432 of your book is a z-table (`stats.ppf()`)
page 433 of your book is a t-table. (`t.pdf()` and `t.cdf()` and `t.ppf()`)

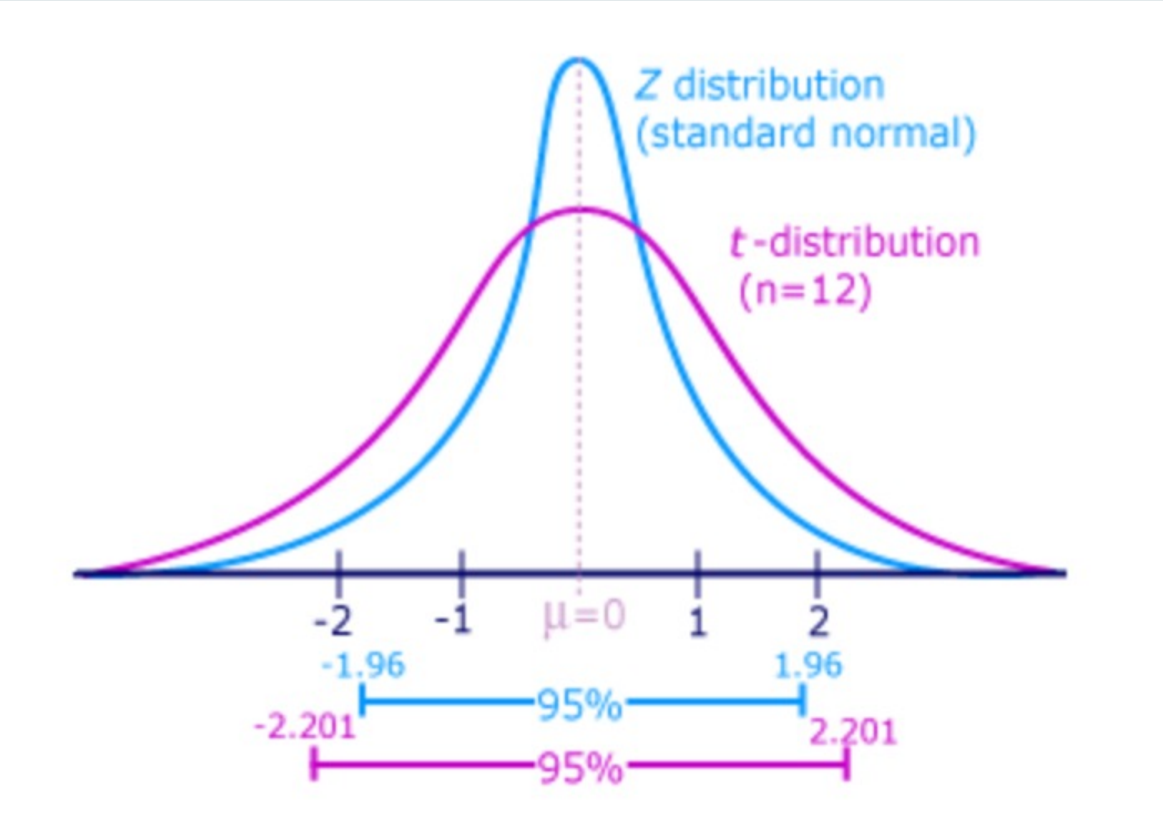
CI with Z-Score

$$CI = Expected \pm Margin\, of\, error = statistic \pm Z \cdot Standard\, error$$

CI with t-score

$$CI = Expected \pm Margin\, of\, error = statistic \pm t \cdot Standard\, error$$

The t-distribution looks different than the Z-distribution



Notice that 95% of the central values occur in different locations for each distribution.

First, let's just review the general concept of hypothesis testing.

We want to know if observed changes (from a sample) are statistically significant so, we construct a hypothesis test. A hypothesis test will require a Z-score or a t-score.

$$Z = \frac{statistic - parameter}{standard\ error\ of\ statistic} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$$t = \frac{statistic - parameter}{standard\ error\ of\ statistic} = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

$$CI = Expected\ \pm Margin\ of\ error = statistic\ \pm Z\ \cdot Standard\ error$$

$$CI = Expected\ \pm Margin\ of\ error = statistic\ \pm t\ \cdot Standard\ error$$

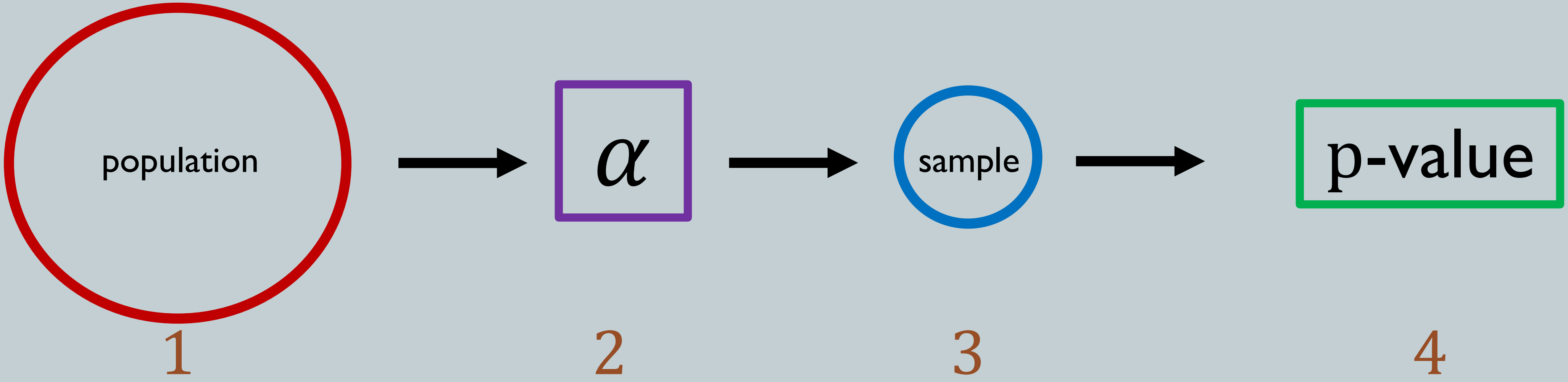
For either CI, there are essentially 4 steps to hypothesis testing:

1] Set the hypothesis

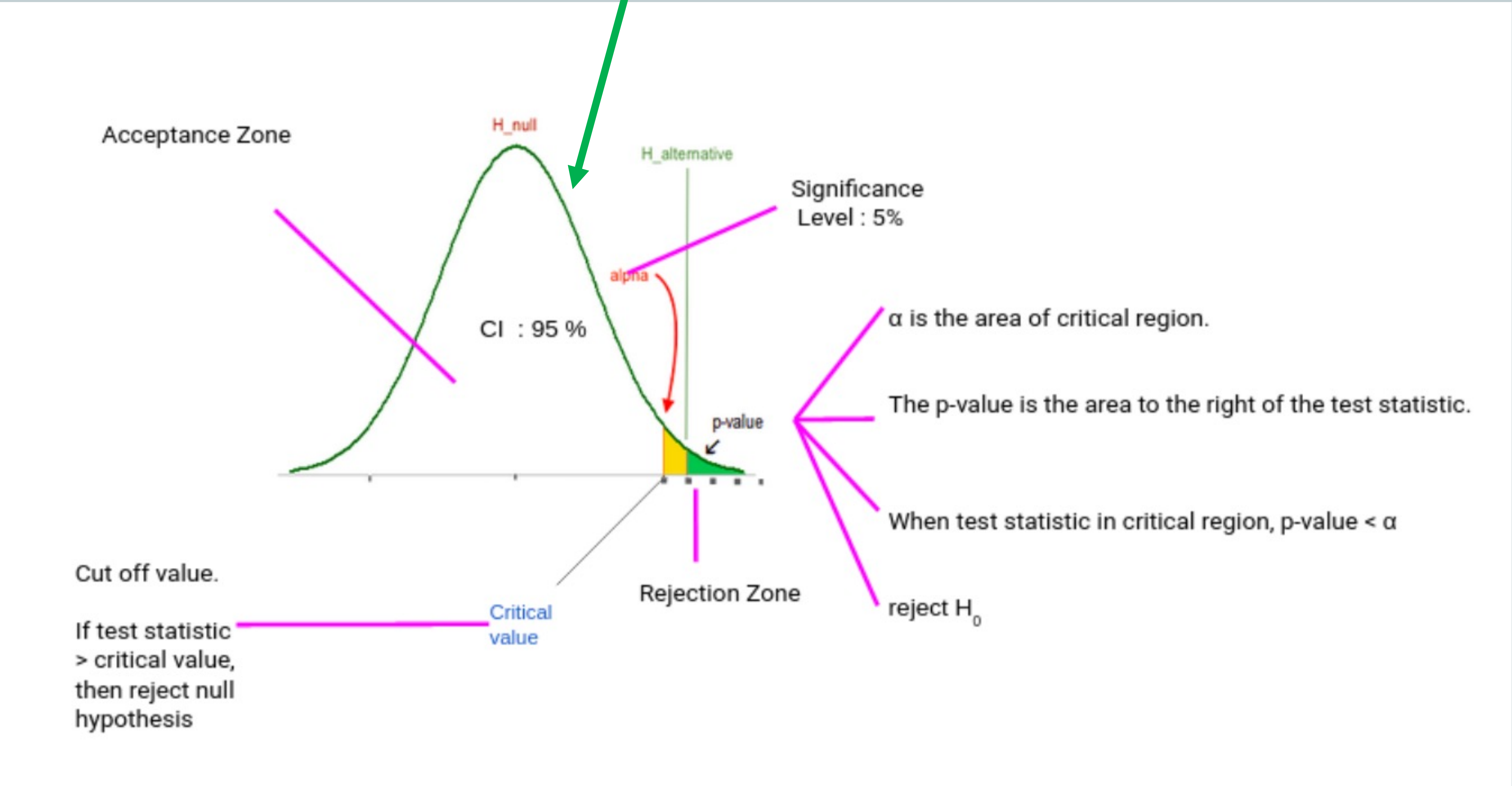
2] Set the significance level; the criteria for a decision.

3] Compute the test statistic.

4] make a decision



However, our decision to reject/fail to reject the null is based on our understanding of our sample and the assumed shape of the distribution



α is the area of critical region.

The p-value is the area to the right of the test statistic.

When test statistic in critical region, p-value < α

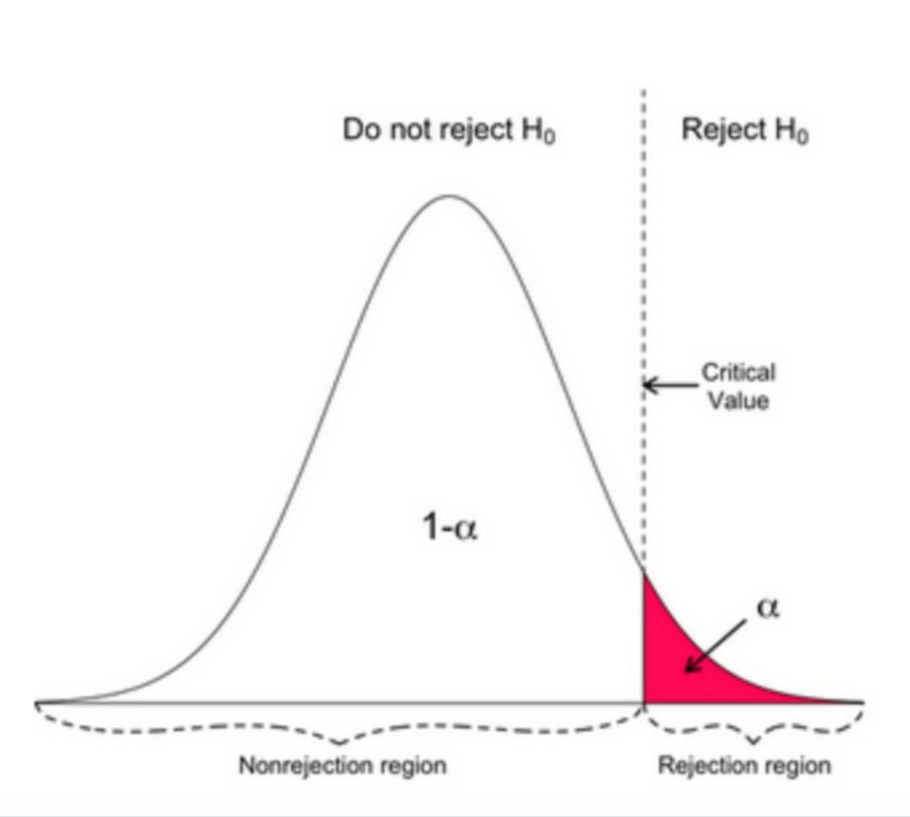
reject H_0

Cut off value.

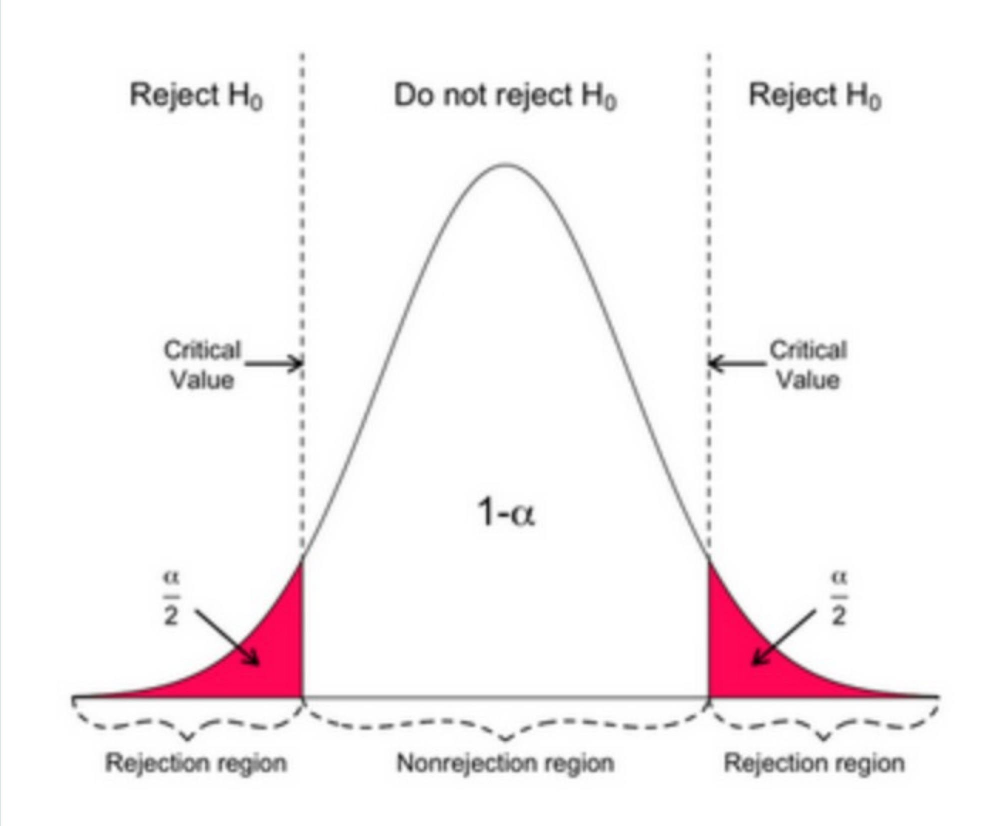
If test statistic > critical value, then reject null hypothesis

The ‘rejection region’ is found in the tail(s).
These regions are created depending upon exactly what we are trying to determine.

For instance: Person A will score significantly higher than Person B



Or perhaps Person A will score significantly different than B.



but again, the shape of the distribution controls our results

In step 3 of this process
(compute a test statistic)
we are making assumptions about the shape of the population distribution.

Hypothesis testing:

1] Set the hypothesis

2] Set the significance level; the criteria for a decision.

3] Compute the test statistic.

4] make a decision

For a Z-test we either know the population variance,
or we do not know the population variance BUT our sample size is $n \geq 30$.

So, if sample size is less than 30 <u>**AND**</u> we do not know the population variance:

then we must use a t-test instead of a Z-test.

Both the z-score and the t-score are used in hypothesis testing.

In class we will use the z-score more often while in application you would probably use the t-score more often.

There is no universal rule as to when to use a t-score.

You will hear/read a lot of ‘should’, ‘can’, ‘might’, ‘usually’, etc.

Different textbooks have different basic rules.

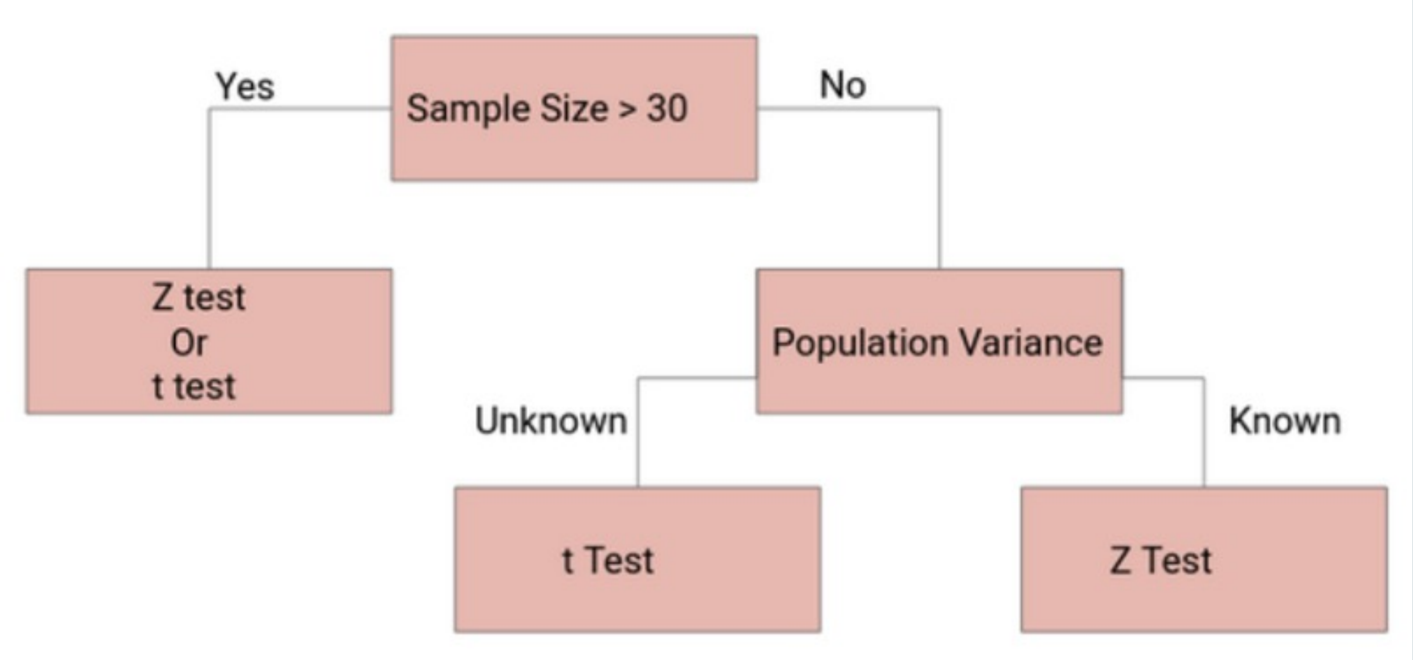
Outside of textbook problems we rarely know sigma, therefore use a t-score.

Technically, you could always use a t-score, as it works for all sample sizes.

The t-score approaches the z-score as sample size increases.

The z-score is only applicable for sample sizes greater than 30.

Z-scores or t-scores ?



If the sample size is large enough, then the Z-test and the t-test will conclude the same results.

For a large sample size, sample variance will be a better estimate of population variance.

So, even if population variance is unknown, we can use the Z-test using sample variance.

Similarly, for a large sample, we have a high degree of freedom.

And since the t-distribution approaches the normal distribution, the difference between the z score and t score is negligible.

Comparing the Z test and the t test

Sample Size

As the sample size differs from analysis to analysis, a suitable test for hypothesis testing can be adopted for any sample size.

For example, the z-test is used for it when sample size is large, generally $n > 30$.

Whereas t-test is used for hypothesis testing when sample size is small, usually $n < 30$ where n is used to quantify the sample size.

Comparing the Z test and the t test

Use of test

The t-test is the statistical test that can be deployed to measure and analyze whether the means of two different populations are different or not when the standard deviation is not known.

The z-test is the parametric test, implemented to determine if the means of two different datasets are different from each other, when the standard deviation is known.

Comparing the Z test and the t test

Types of distribution

Both t-test and z-test employ the use of distribution to correlate values and make conclusions in terms of hypothesis testing.

Notably, t-test is based on the Student's t-distribution, and the z-test counts on Normal Distribution.

Comparing the Z test and the t test

Population Variance

Implementing both tests in testing of hypothesis, population variance is significant in obtaining the t-score and z-score

While the population variance in the z-test is known, it is unknown in the t-test.

Comparing the Z test and the t test

Conclusion I

The t-test and the z-test are the substantive tests in determining the significance difference between sample and population. While the formulas are similar, the selection of a particular test relies on sample size and the standard deviation of population.

The t-test and z-test are relatively similar, but their applicability is different. The fundamental difference is that the t-test is applicable when sample size is less than 30, and the z-test is practically conducted when size exceeds 30.

Comparing the Z test and the t test
Conclusion II

The decision to use a z-score or a t-score boils down to four things:

1] Are you working with a mean (25 people) or a proportion (25% of people) ?
Proportion problems are never t-test problems.

2] Do you know the population SD ?
In real life we usually don’t.

3] Is the population normally distributed ? Important when sample size is small.

4] Size.”Small” samples are considered less than 30.
When sample size is large, the central limit theorem informs us that we need not worry about whether or not the population is normally distributed.

When we use a t-test we also need to be concerned about 'degrees of freedom.'

Many calculations in statistics have degrees of freedom.
What are degrees of freedom?
"The number of independent pieces of information that go into calculating that estimate."

We are going to see about 8 slides explaining what 'degrees of freedom' means, however the bottom line is: degrees of freedom for a one sample test is $n - 1$.

So, if the sample size is $n = 4$, then $df = 4 - 1 = 3$.
And if the sample size is $n = 99$ then $df = 99 - 1 = 98$.

Degrees of freedom for a 2-sample test are $(N_1 + N_2) - 2$.
If you have two samples and want to find a parameter, like the mean, you have two "n's" to consider (sample 1 and sample 2)

Why is $df = n - 1$? Why do we care?

A distributions shape is dependent upon how many variables we are using.
$n - 1$ is the number of values that are free to vary (i.e., "are variable").

What does 'free to vary' mean?
Suppose we want three numbers with mean of 10.
These three numbers could be $\{9, 10, 11\}$ or $\{8, 10, 12\}$ or $\{5, 10, 15\}$.

But once you have chosen the first two numbers, the third is fixed.
i.e., there are only $3 - 1 = 2$ 'variables' in the calculation of the mean.

You CANNOT choose the third item in the set.
$\frac{9+10+x}{3} = 10$ implies that the x is fixed. It is not variable.

9 and 10 could have been anything, they were free to vary.
But we have no freedom to pick the third number.

So, for a set of 3 items, the 3^{rd} number is removed from the degrees of freedom.

Degrees of freedom are an integral part of inferential statistical analyses, which estimate or make inferences about population parameters based on sample data.

As an illustration, think of people filling up a 30-seat classroom. The first 29 people have a choice of where they sit, but the 30th person to enter can only sit in the one remaining seat.

Similarly, if you calculated the mean of a sample of 30 numbers, the first 29 are free to vary but 30th number would be determined as the value needed to achieve the given sample mean. Therefore, when estimating the mean of a single population, the degrees of freedom is 29.

Degrees of freedom are important for finding critical cutoff values for inferential statistical tests.

Depending on the type of the analysis you run, degrees of freedom typically (but not always) relate to the size of the sample.

Because higher degrees of freedom generally mean larger sample sizes, a higher degree of freedom means more power to reject a false null hypothesis and find a significant result.

degrees of freedom

Take a look at the t-score formula in a hypothesis test:

$$t = \frac{\bar{x} - \mu_0}{{}^{s}/_{\sqrt{n}}}$$

When n increases, the t-score goes up. This is because of the square root in the denominator: as it gets larger, the fraction s/√n gets smaller and the t-score (the result of another fraction) gets bigger.

As the degrees of freedom are defined as $n - 1$, you would think that the t-critical value should also get bigger, but they don't: **they get *smaller***.

This seems counter-intuitive.

However, **think about what a t-test is actually for**.

You're using the t-test because you don't know the standard deviation of your population and therefore you don't know the shape of your graph. It could have short, fat tails. It could have long skinny tails. You just have no idea.

The degrees of freedom affect the shape of the graph in the t-distribution; as the df get larger, the area in the tails of the distribution get smaller.

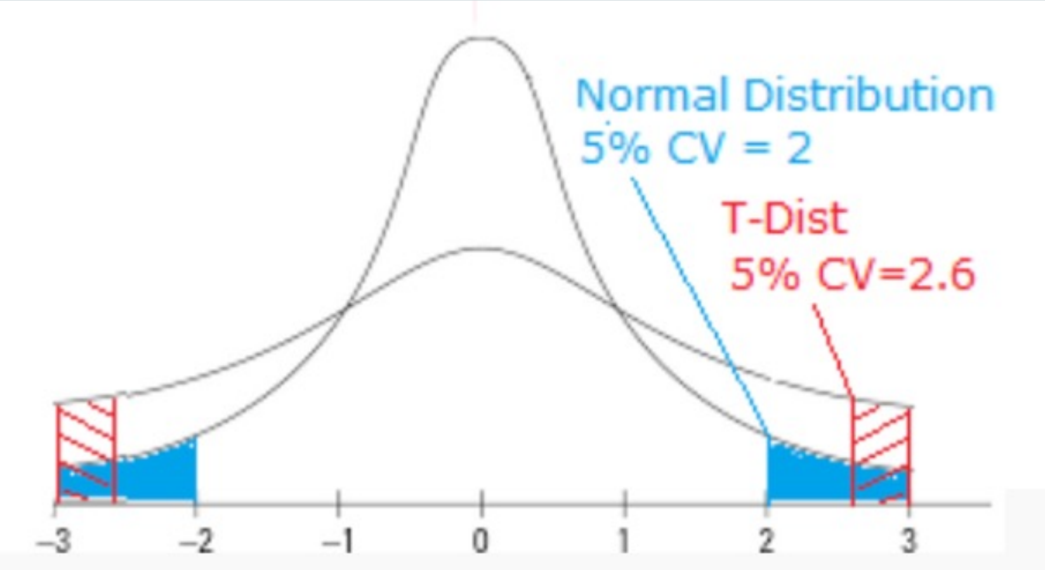
As df approaches infinity, the t-distribution will look like a normal distribution.

When this happens, you can be certain of your standard deviation (which is 1 on a normal distribution). And of course, SD controls the shape; the tails.

Let's say you took repeated sample weights from four people, drawn from a population with an unknown standard deviation. You measure their weights, calculate the mean difference between the sample pairs and repeat the process over and over.

The tiny sample size of 4 will result in a t-distribution with fat tails. The fat tails tell you that you're more likely to have extreme values in your sample.

You test your hypothesis at $\alpha = .05$, which **cuts off the last 5% of your distribution**. But what are the tails shaped like?

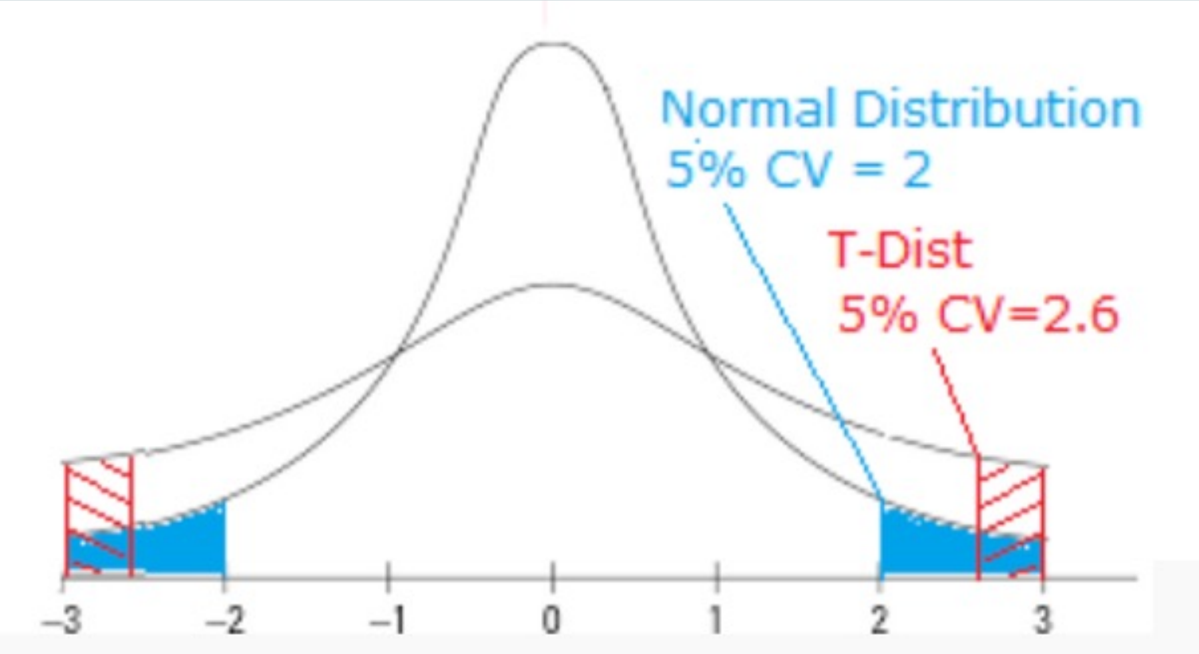


t-distribution, William Gosset,
(Chemistry, Mathematics)
Small samples, Guiness Brewery

https://www.physoc.org/magazine-articles/the-strange-origins-of-the-students-t-test/

This graph shows the t-distribution with a 5% cut off.
This gives a critical value of 2.6.
(**Note**: This is a hypothetical, so the CV is not exact).

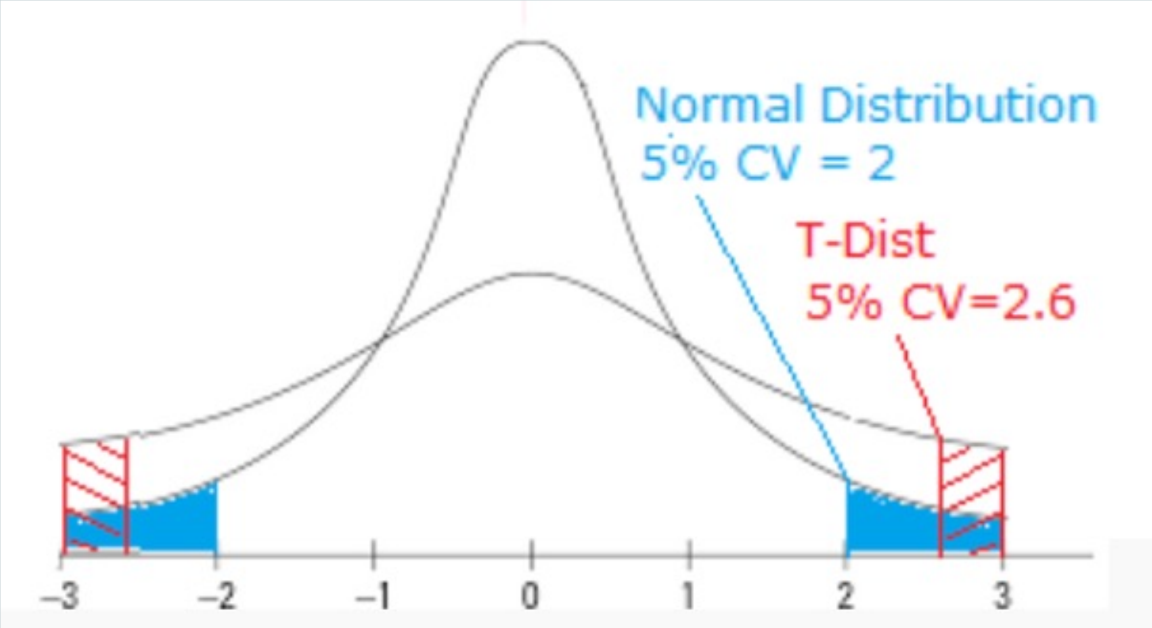
Now look at the normal distribution.
We have less chance of extreme values with the normal distribution.
Our 5% alpha level cuts off at a CV of 2.



Back to the original question:

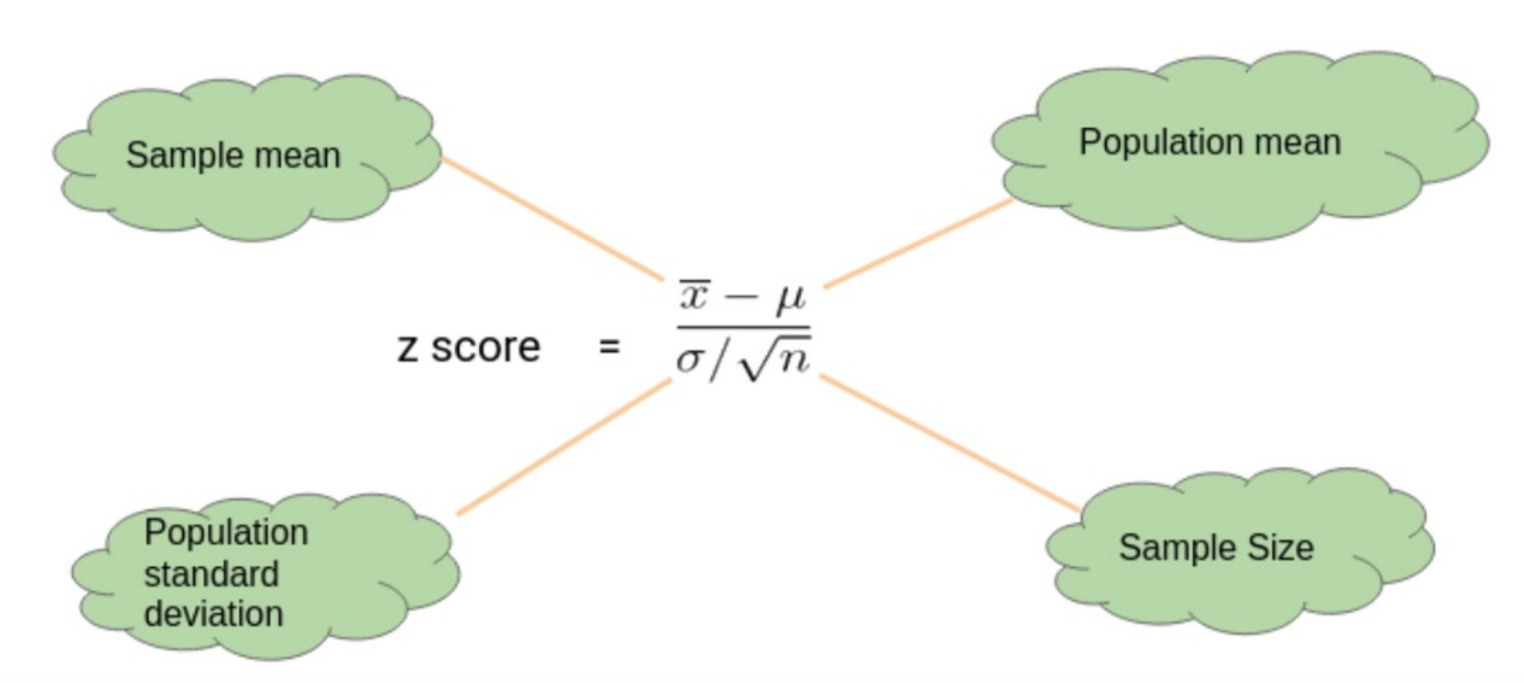
“Why do critical values decrease while DF increases?”

Degrees of freedom are related to sample size $(n - 1)$.
If the df increases, it also stands that the sample size is increasing;
the graph of the t-distribution will have skinnier tails, pushing the critical value towards the mean.



Now let’s actually look at some examples of Z-tests and t-tests.

The one-sample Z-test is used when we want to compare a sample mean with the population mean



Sample mean
Population mean
z score = $\frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$
Population standard deviation
Sample Size

Suppose we want to determine if student end-of-course scores in CSCI3200 are higher than 600.

We have the information that the standard deviation for this class is 100.

We collect data from 20 students in CSCI 3200 by using a random sample.

Finally, we set our significance level to be $\alpha = .05$

Sample scores:

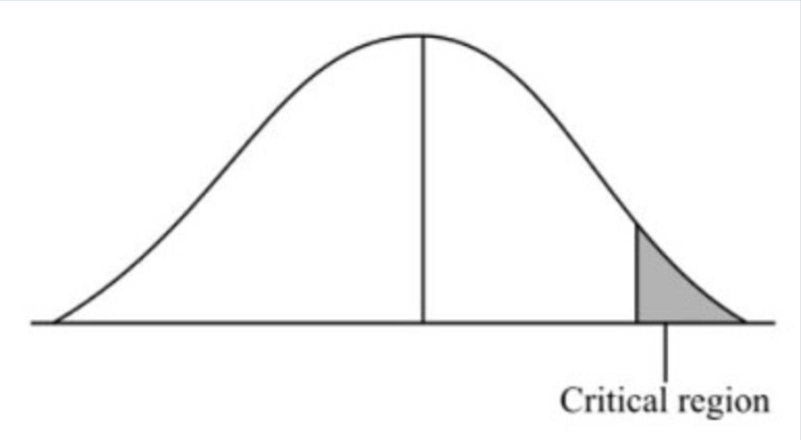
{650, 730, 510, 670, 480, 800, 690, 530, 590, 620, 710, 670, 640, 780, 650, 490, 800, 600, 510, 700}

Mean score for CSCI3200 students: $\bar{x} = 641$

Sample size: $n = 20$

Population mean: $\mu = 600$

The standard deviation for the population: $\sigma = 100$



$H_0 : \mu \leq 600$ and $H_1 : \mu > 600$

$$Z \text{ score} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{641 - 600}{100 / \sqrt{20}} = 1.8336$$

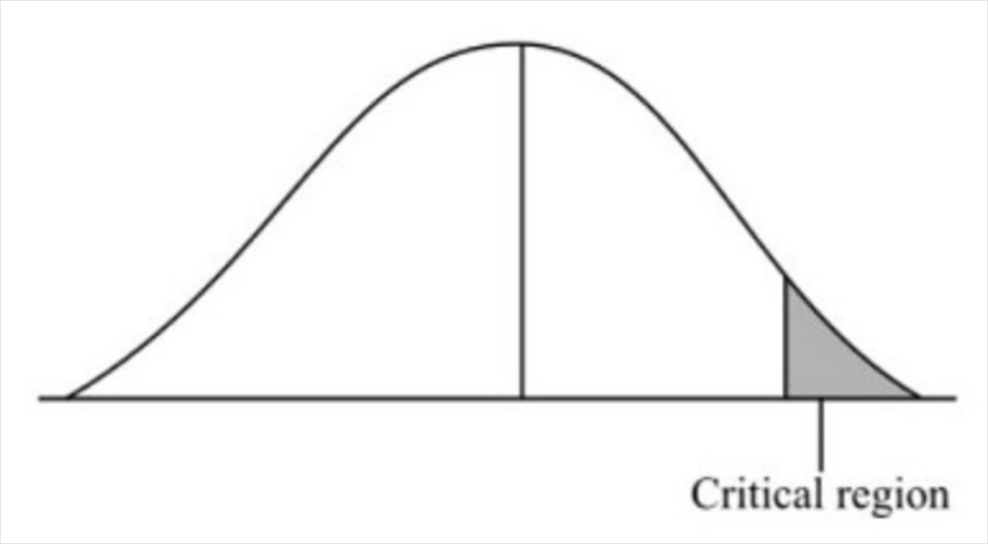
$P(Z \geq 1.8336) = 1 - \Phi(1.8336) = 0.033357$ or p value $= 0.033357$

$\alpha = .05$ implies Critical value $= 1.645$

Notice $\text{p value} < \alpha$ or $0.033357 < 0.05$

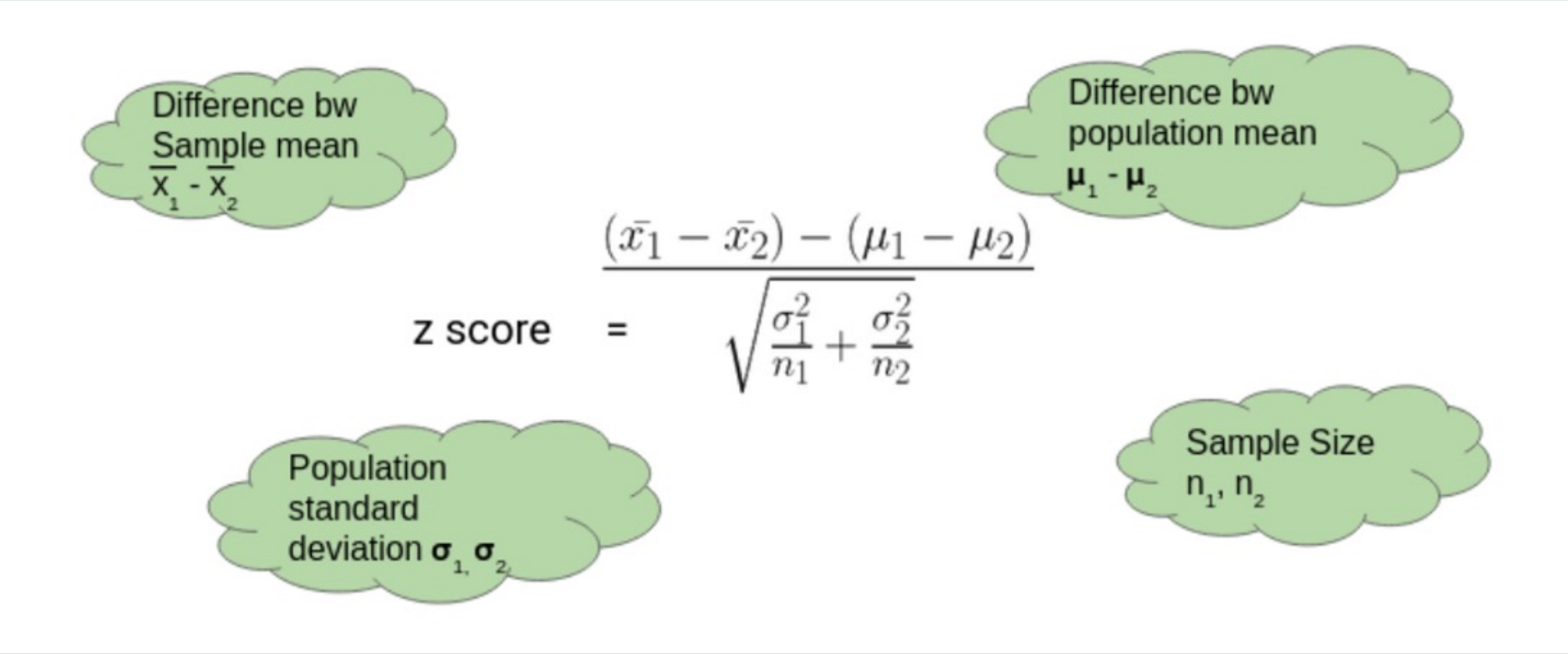
Also, notice $\text{Critical Value} < \text{Z Score}$ or $1.645 < 1.8336$

Either of these would indicate that we reject the null.



Since the p-value is less than .05, we can reject the null hypothesis and conclude, based on our result, that CSCI3200 students have an end-of-course average higher than 600.

The two sample Z test is used when we want to compare the mean of two samples:



Difference bw Sample mean $\bar{x}_1 - \bar{x}_2$

Difference bw population mean $\mu_1 - \mu_2$

$$\text{z score} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Population standard deviation σ_1, σ_2

Sample Size n_1, n_2

Suppose we want to know if CSCI 3200 students on average score 10 points more than APPM 1360 students on their end-of-course grade.

We have the information that the standard deviation for CSCI 3200 students is 100 and for APPM 1360 students the standard deviation is 90.

We collect data for both groups by using random samples ($n = 20$ for both).
Finally, we set $\alpha = .05$.

Sample scores for CSCI and for APPM:

{650, 730, 510, 670, 480, 800, 690, 530, 590, 620, 710, 670, 640, 780, 650, 490, 800, 600, 510, 700}

{630, 720, 462, 631, 440, 783, 673, 519, 543, 579, 677, 649, 632, 768, 615, 463, 781, 563, 488, 650}

Mean, SD, and sample size for CSCI: $\bar{x}_1 = 641$, $\sigma_1 = 100$, $n = 20$

Mean, SD, and sample size for APPM: $\bar{x}_2 = 613.3$, $\sigma_2 = 90$, $n = 20$

Difference between Mean of population is 10

$H_0\colon \mu_1 - \mu_2 \leq 10$ and $H_1\colon \mu_1 - \mu_2 > 10$

$$Z\ score = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(641 - 613.3) - (10)}{\sqrt{\frac{100^2}{20} + \frac{90^2}{20}}} = 0.588$$

$P(Z \geq 0.588) = 1 - \Phi(0.588) = 0.2783$ or p value $= 0.2783$

$\alpha = .05$ implies Critical Value $= 1.645$

Notice $p \text{ value} > \alpha$ or $0.2783 > 0.05$

Also notice $Z \text{ score} < \text{Critical value}$ or $0.588 < 1.645$

Either of these indicate that we cannot reject the null.

Based on the p-value, we fail to reject the null hypothesis.

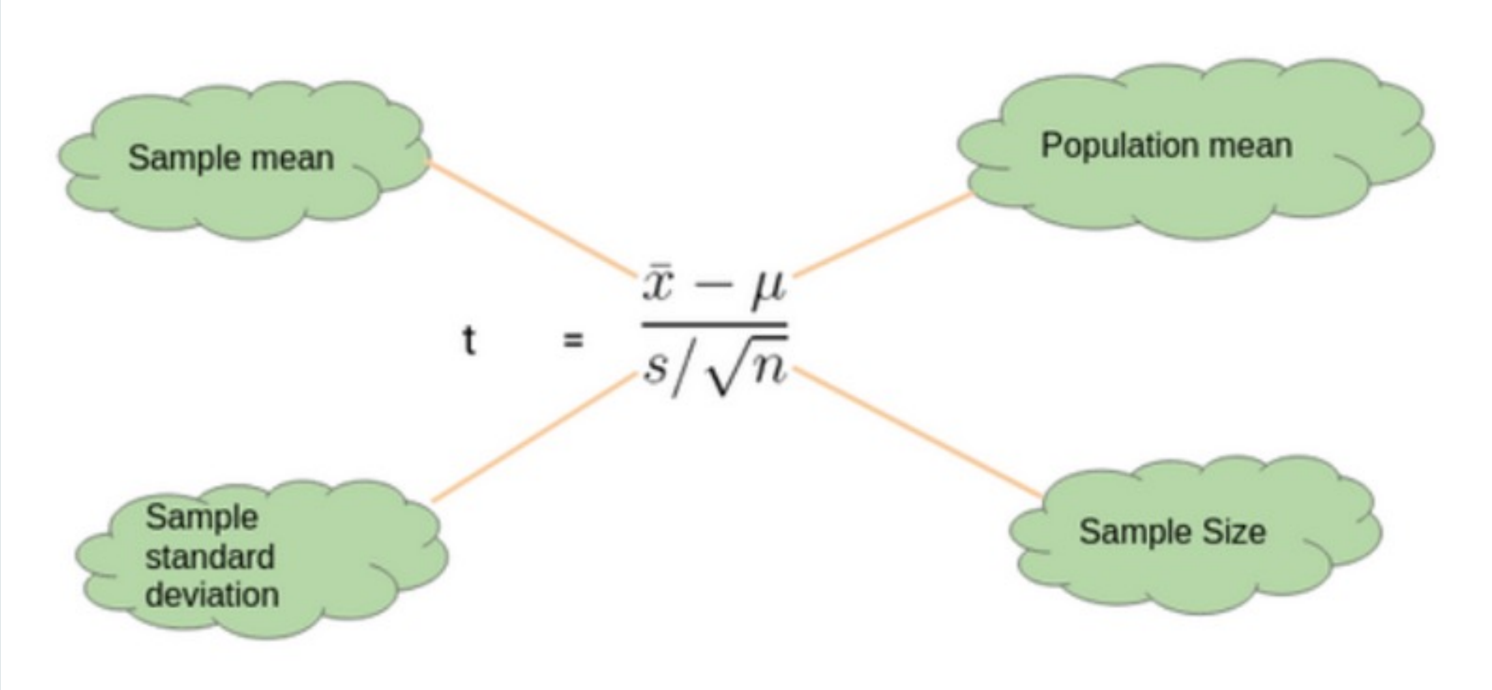
We do not have enough evidence to conclude that CSCI students score 10 more points than do APPM students.

So, what then are t-tests?

t-tests are a statistical way of testing a hypothesis when we do not know the population variance and our sample size is less than 30.

The one-sample t-test is used when we want to compare a sample mean with the population mean.

The difference from the Z-test is that we do not have the information on population variance. We use the sample standard deviation instead of population standard deviation.



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

Suppose we want to determine if on average CSCI 3200 students score more than 600 points for their end-of-course score.

We do not have the information related to variance (or SD) for these scores.

To perform a t-test, we design our experiment:
randomly collect the data of 10 of these students with their scores and choose $\alpha = .05$.

H_0: $\mu \leq 600$
H_1: $\mu > 600$

Scores: $\{587, 602, 627, 610, 619, 622, 605, 608, 596, 592\}$

Mean score for these 10 CSCI3200 students: $\bar{x} = 606.8$

Sample size: $n = 10$

Population mean: $\mu = 600$

Sample standard deviation: $s = 13.14$

Therefore, the t-score: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{606.8 - 600}{13.14/\sqrt{10}} = 1.64$

Now, we need to find critical values. But for a t distribution…

```
scipy.stats.t.ppf(q, df)
```

`q`: The significance level and `df`: the degrees of freedom

```
# find T critical value for a left tailed test
   scipy.stats.t.ppf(q=.05, df=9)
```

```
# find T critical value for a right tailed test
   scipy.stats.t.ppf(q=1-.05, df=9)
```

```
# find T critical value for a two tailed test
   scipy.stats.t.ppf(q=1-.05/2, df=9)
```

Or page 433 chart in your book

t-score: $t = \frac{\bar{x}-\mu}{{}^{s}/_{\sqrt{n}}} = \frac{606.8-600}{{}^{13.14}/_{\sqrt{10}}} = 1.64$

`scipy.stats.t.ppf(q=1-.05, df=9)` produces 1.8331129326536335

So, $\alpha = .05$ implies a critical value $= 1.833$ (9 degrees of freedom)

We now see that our $\mathrm{t\ score} < \mathrm{Critical\ value}$ or $1.64 < 1.833.$

`1 - stats.t.cdf(1.64, 9)` produces 0.0678

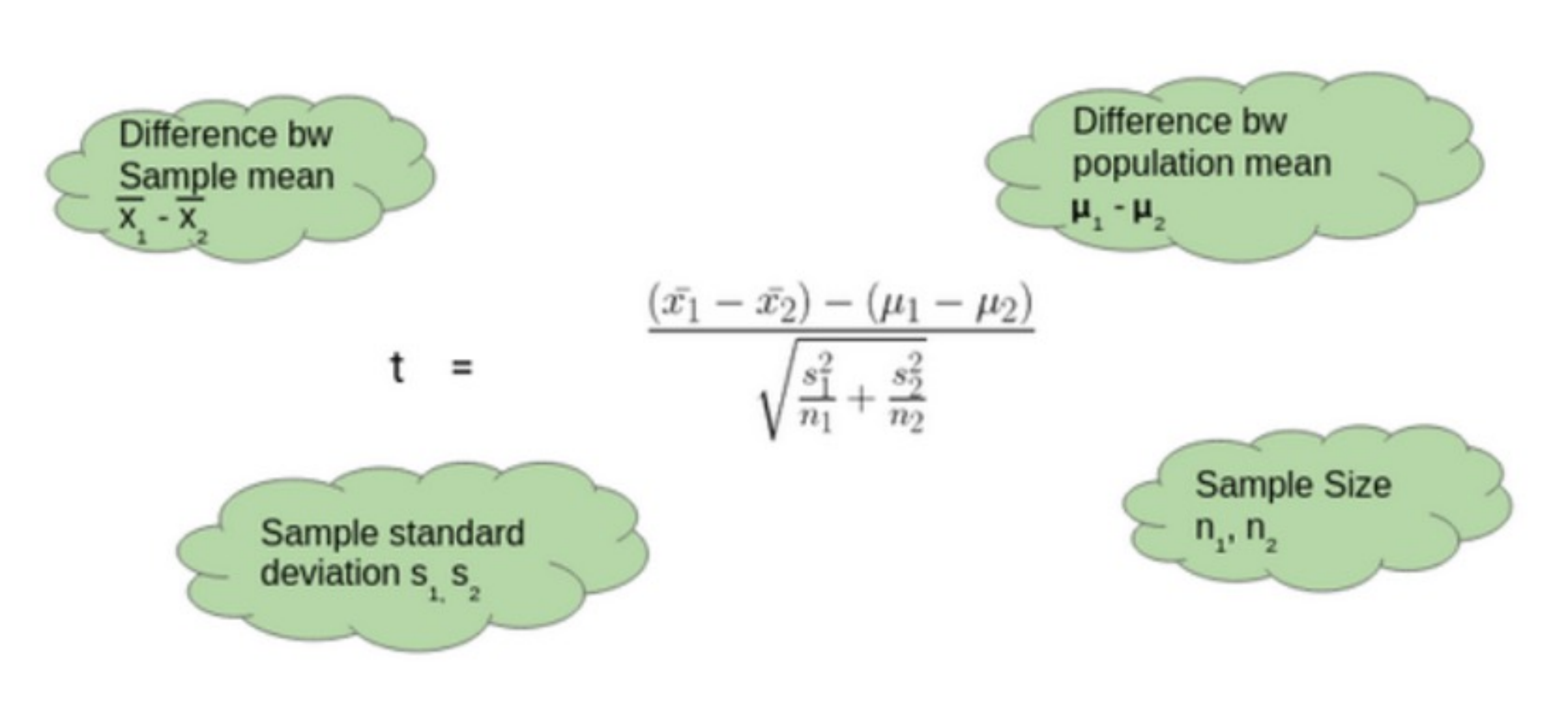
`stats.t.sf(1.64, 9)` also produces 0.0678

So, our t-score produced a $\mathrm{p\ value} = 0.0678$

and we now see that our $\mathrm{p\ value} > \alpha$ or $0.0678 > 0.05$

Our p-value is greater than 0.05 (or t score < critical value) thus we fail to reject the null hypothesis and don’t have enough evidence to support the hypothesis that on average CSCI32000 students score more than 600 for their end-of-course score.

Let’s perform a two-sample t-test when we want to compare the mean of 2 samples.



$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Difference bw Sample mean $\bar{x}_1 - \bar{x}_2$
Difference bw population mean $\mu_1 - \mu_2$
Sample standard deviation s_1, s_2
Sample Size n_1, n_2

We want to determine if, on average, APPM 1360 students score 15 points higher than do CSCI 3200 students.

We do not have the information related to variance (or SD) for the CSCI students or the APPM student's sores.

To perform a t-test we randomly collect the data of 10 students from each of CSCI and APPM.

$H_0: \mu_1 - \mu_2 \leq 15$ and $H_1: \mu_1 - \mu_2 > 15$
We choose our $\alpha = .05$ for our hypothesis test.

CSCI scores: {587, 602, 627, 610, 619, 622, 605, 608, 596, 592}
APPM scores: {626, 643, 647, 634, 630, 649, 625, 623, 617, 607}

CSCI: $\bar{x}_2 = 606.8, \ s_2 = 13.14$
APPM: $\bar{x}_1 = 630.1, \ s_1 = 13.42$

Difference between population means: 15

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(630.1 - 606.8) - (15)}{\sqrt{\frac{(13.42)^2}{10} + \frac{(13.14)^2}{10}}} = 1.3975$$

`1 - stats.t.cdf(1.3975, 8)` produces 0.0999
`stats.t.sf(1.3975, 8)` also produces 0.0999

So, our t-score produced a $\text{p value} = 0.0999$
and we now see that our $\text{p value} > \alpha$ or $0.0999 > 0.05$

`scipy.stats.t.ppf(q=1-.05, df=8)` produces 1.8595

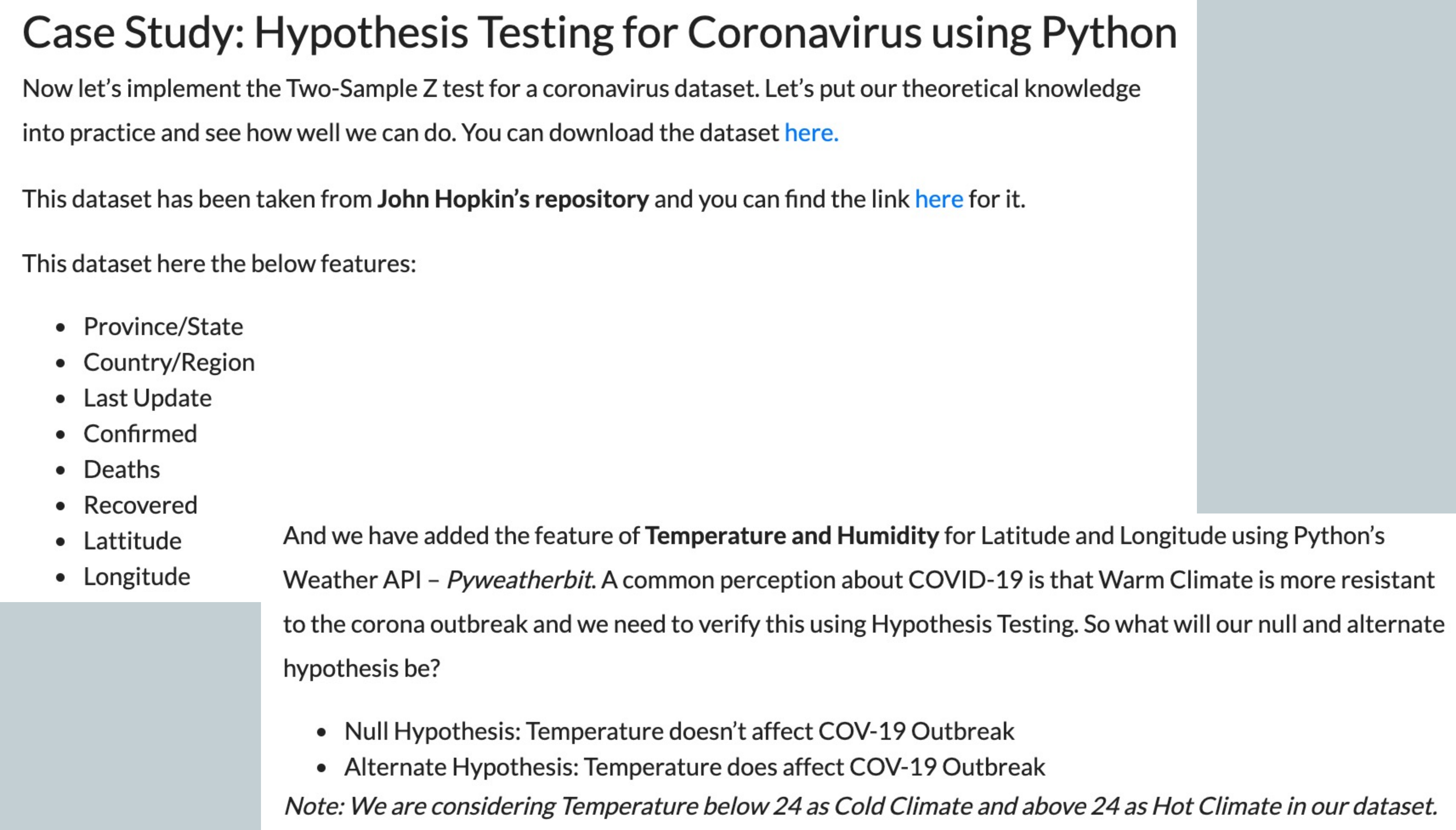
In a similar vein, we see the t-score < critical value.
That is $1.3975 < 1.8595$

Thus, by either considering the p-value or the t-score, we cannot reject the null hypothesis and conclude that on average the difference between CSCI scores and APPM scores is less than or equal to 15 points.

Next Time: Simple Linear Regression

k

Chi-square distribution ?



Case Study: Hypothesis Testing for Coronavirus using Python

Now let's implement the Two-Sample Z test for a coronavirus dataset. Let's put our theoretical knowledge into practice and see how well we can do. You can download the dataset here.

This dataset has been taken from **John Hopkin's repository** and you can find the link here for it.

This dataset here the below features:

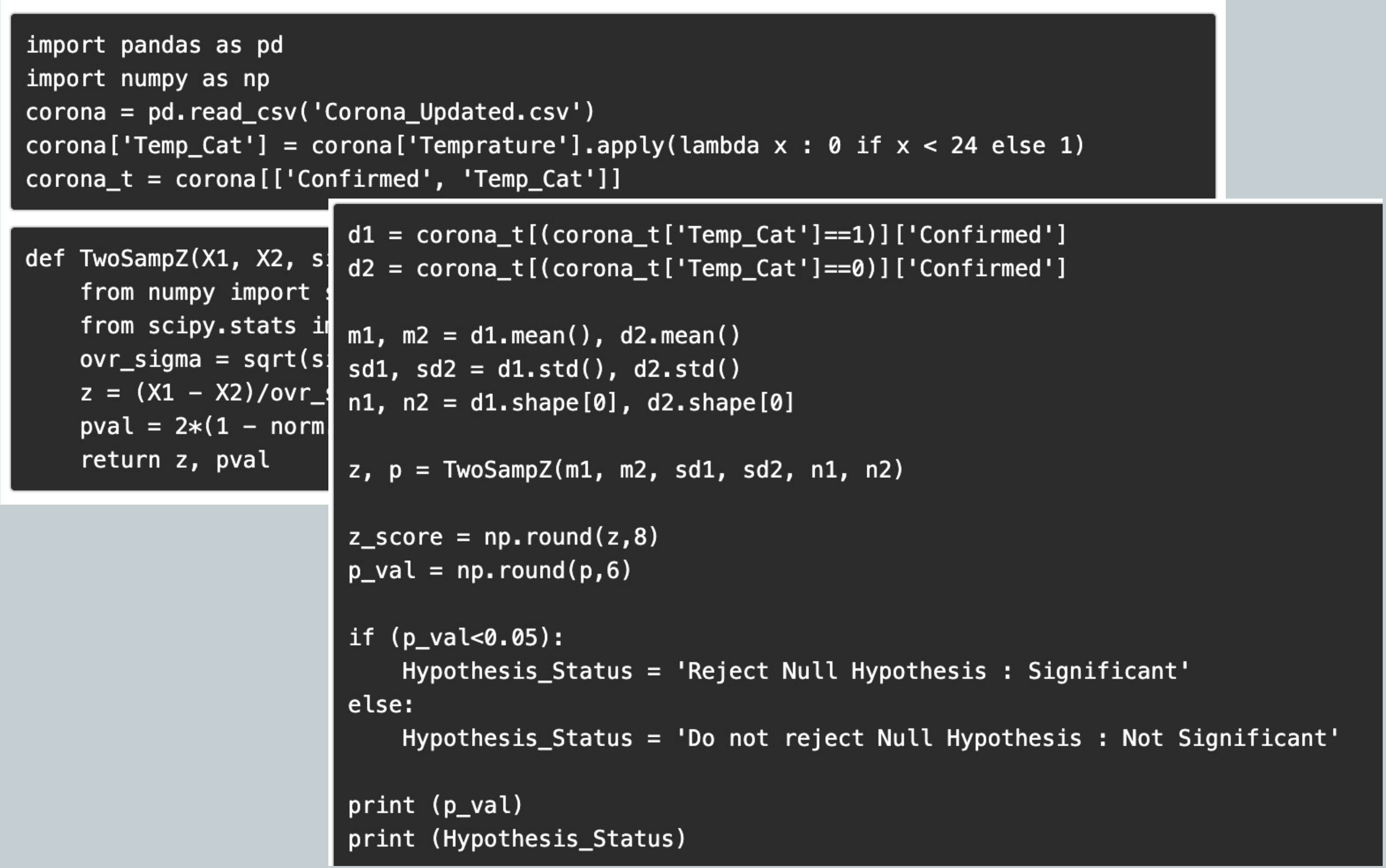
- Province/State
- Country/Region
- Last Update
- Confirmed
- Deaths
- Recovered
- Lattitude
- Longitude

And we have added the feature of **Temperature and Humidity** for Latitude and Longitude using Python's Weather API – *Pyweatherbit*. A common perception about COVID-19 is that Warm Climate is more resistant to the corona outbreak and we need to verify this using Hypothesis Testing. So what will our null and alternate hypothesis be?

- Null Hypothesis: Temperature doesn't affect COV-19 Outbreak
- Alternate Hypothesis: Temperature does affect COV-19 Outbreak

Note: We are considering Temperature below 24 as Cold Climate and above 24 as Hot Climate in our dataset.

k

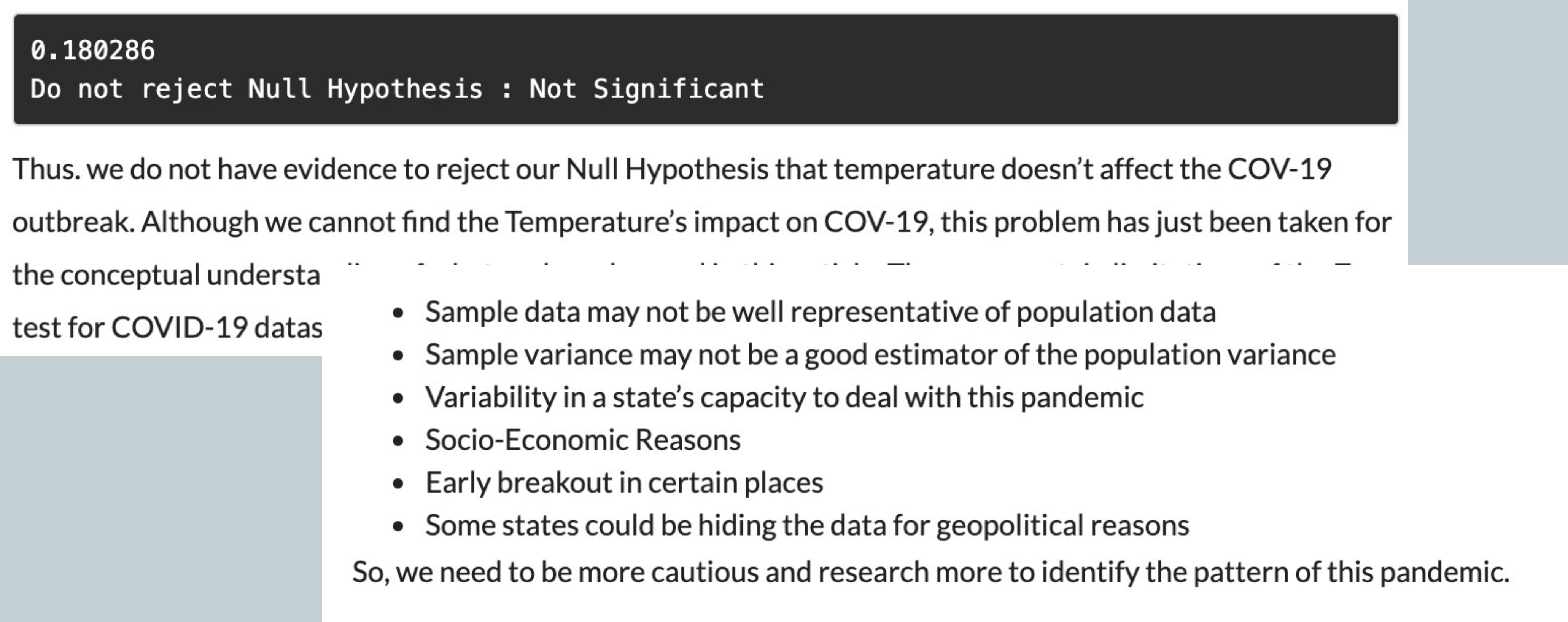


```
import pandas as pd
import numpy as np
corona = pd.read_csv('Corona_Updated.csv')
corona['Temp_Cat'] = corona['Temprature'].apply(lambda x : 0 if x < 24 else 1)
corona_t = corona[['Confirmed', 'Temp_Cat']]
```

```
def TwoSampZ(X1, X2, s
    from numpy import s
    from scipy.stats i
    ovr_sigma = sqrt(s
    z = (X1 - X2)/ovr_
    pval = 2*(1 - norm
    return z, pval
```

```
d1 = corona_t[(corona_t['Temp_Cat']==1)]['Confirmed']
d2 = corona_t[(corona_t['Temp_Cat']==0)]['Confirmed']

m1, m2 = d1.mean(), d2.mean()
sd1, sd2 = d1.std(), d2.std()
n1, n2 = d1.shape[0], d2.shape[0]

z, p = TwoSampZ(m1, m2, sd1, sd2, n1, n2)

z_score = np.round(z,8)
p_val = np.round(p,6)

if (p_val<0.05):
    Hypothesis_Status = 'Reject Null Hypothesis : Significant'
else:
    Hypothesis_Status = 'Do not reject Null Hypothesis : Not Significant'

print (p_val)
print (Hypothesis_Status)
```

k



```
0.180286
Do not reject Null Hypothesis : Not Significant
```

Thus. we do not have evidence to reject our Null Hypothesis that temperature doesn’t affect the COV-19 outbreak. Although we cannot find the Temperature’s impact on COV-19, this problem has just been taken for the conceptual understa

test for COVID-19 datas

- Sample data may not be well representative of population data
- Sample variance may not be a good estimator of the population variance
- Variability in a state’s capacity to deal with this pandemic
- Socio-Economic Reasons
- Early breakout in certain places
- Some states could be hiding the data for geopolitical reasons

So, we need to be more cautious and research more to identify the pattern of this pandemic.

k

Now let’s implement the Two-Sample Z test for a coronavirus dataset. Let’s put our theoretical knowledge into practice and see how well we can do. You can download the dataset here.

This dataset has been taken from **John Hopkin’s repository** and you can find the link here for it.