Inferential Statistics

Rabail Tahir

Learning Outcomes

- Define inferential statistics.
- Explain the difference between a sample and a population.
- Explain how confidence intervals work.
- List and explain the steps in hypothesis testing.
- Explain the difference between the null hypothesis and the alternative hypothesis.
- Explain the difference between a probability value and the significance level.

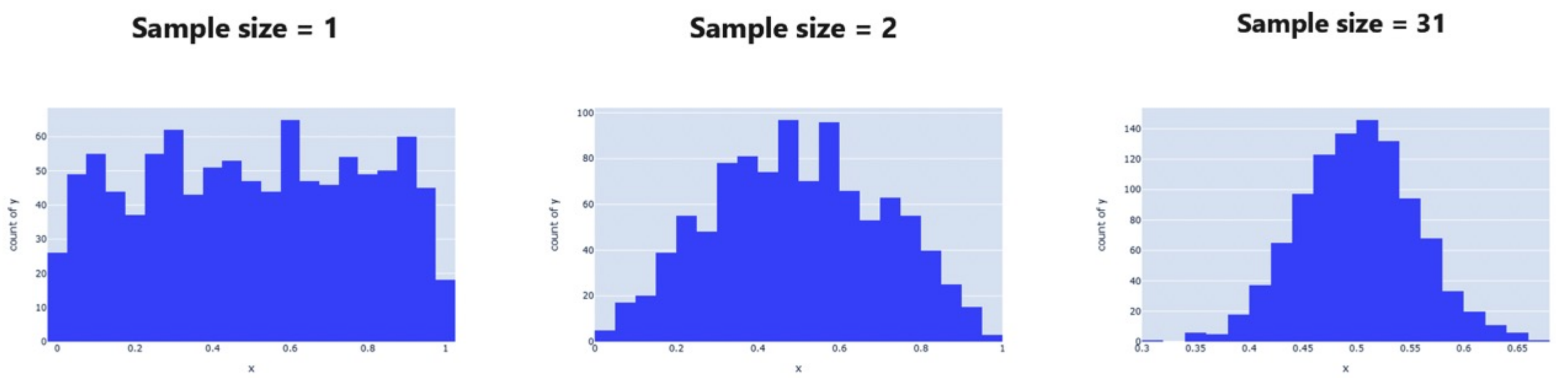
What is Inferential statistics?

- It is the study of the abstract relationship between sample and population.
- As a Data Scientist , you will need to keep all your bias aside and do not jump to conclusions.
- Inferential Statistics helps us do the following:
  - Estimate parameters
  - Construct confidence intervals
  - Test hypothesis
  - Make decisions
  - Develop strategies
  - Forecasting

The Central Limit theorem

- The Central Limit Theorem states that when large samples are taken from populations, and are plotted after calculating their means, the distribution that results from that is the same as Normal Distribution.
- Some important points:
  - The mean of the sample means is equal to the population mean.
  - If the population is normal, then the sample means will be normal.
  - If the population is not normal, but the sample size is greater than 30, the sample means will still roughly form a normal distribution.
  - The standard deviation of the sample means equals the population standard deviation divided by the square root of n:

$$\text{sample standard deviation} = \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$



Larger sample sizes approach Normal Distribution

Confidence Interval

- "A ***confidence interval*** is a range calculation showing how confidently we believe a sample mean (or other parameter) falls in a range for the population mean."
- For example, "*Based on a sample of 31 golden retrievers with a sample mean of 64.408 and a sample standard deviation of 2.05, I am 95% confident that the population mean lies between 63.686 and 65.1296*" .
- Confidence interval helps you determine how much uncertainty there is with any statistical measurement.
- It helps us to know how much confident we can be for any statistical result.
- Confidence Interval uses the concept of Margin of Error.

Margin of Error

- *Margin of error (E)*, which is the range around the sample mean that contains the population mean at that level of confidence.
- Steps to calculate Margin of Error:
  - Calculate the critical value, which can be either z-score or t-score. (T-score is used when sample size is under 30 and population standard deviation is also unknown)
  - Find the standard deviation
  - Apply these values in the following formula

$$E = \pm z_c \frac{s}{\sqrt{n}}$$

How to calculate Critical Z value for a given confidence level

- **Step 1**: Determine the confidence level, denoted C, where C is a number (decimal) between 0 and 100.
- **Step 2:** Obtain the confidence level, denoted α by evaluating $\alpha = 1-(c/100)$
- **Step 3:** Use the z-table (or a calculator) to obtain the z-score $Z\alpha/2$.

Z-table (https://users.stat.ufl.edu/~mripol/STA2023/Z-table.pdf)

Margin of Error Example

- Recall that our sample of 31 golden retrievers has a mean of 64.408 and standard deviation of 2.05. The formula to get this margin of error is:

$$E = \pm z_c \frac{s}{\sqrt{n}}$$

$$E = \pm 1.95996 * \frac{2.05}{\sqrt{31}}$$

$$E = \pm 0.72164$$

- If we apply that margin of error against the sample mean, we finally get the confidence interval!

$$95\% \text{ confidence interval} = 64.408 \pm 0.72164$$

Example

- A biologist wants to estimate the mean height of all of the trees in a park. In a sample of 455 of the park's trees, the mean height is 39.2 feet. The biologist wants a 99% confidence interval for the true mean height of all of the trees in the park. What z-score should be used when constructing the interval?

Solution

- Step 1 : Determine confidence level which is 99%
- Step 2: Calculate significance level ,

$$\alpha = 1 - \frac{99}{100} = 1 - 0.99 = 0.01.$$

Step 3 : Use the z-table to obtain the z-score for z($\alpha/2$)
The z score z($\alpha/2$)=z(0.01/2)=z(0.005) is the number on the horizontal axis that has the area 0.005 to its right, so it has area 1 – 0.005 = 0.995 to its left. Using the z table gives z(0.005) = 2.575

Understanding P- values

- *P-value*, the probability of something occurring by chance rather than because of a hypothesized explanation.
- When we set up any experiment, we have to consider the possibility of random luck influencing the results of the experiment.
- This helps us define our null hypothesis, that any particular variable had no influence on the results, and that any positive results are due to random luck.
- On the other hand, it can also help us form the alternative hypothesis that the particular variable did indeed influence the results.
- The threshold for statistical significance is a p-value of 5% or less.

P-value

- P-Value is used in Hypothesis testing to accept or reject the null hypothesis.
- A small P-value means that null hypothesis should be rejected.
- P values are expressed as decimals.
- A smaller P-value implies that your results are “significant” and are not a result of just random luck and there is a variable influencing it.

Alpha values

- Alpha values are values that are controlled by the researcher and are related to confidence levels.

Alpha value = 100% - Confidence level

You calculate Alpha value by subtracting the confidence level from 100%.

- For example, if you want to be 98% confident, the alpha value will be equal to 2%.
- When you run the test or the experiment, it will give you the p value, which you then compare to the alpha value.
- If p – value < = alpha value, hypothesis is rejected, else if its greater than the alpha value, the alternative hypothesis is weak.

How are P-values calculated?

- Can be calculated using P – value table or any statistical software.
- It also depends on the statistical test you are using to test your hypothesis.
- P – value is always between 0 and 1.

P-value	Decision
Less than 0.05*	**Reject Null** (H_0) Hypothesis Statistical difference between groups
Greater than 0.05*	**Fail to Reject** Null (H_0) Hypothesis No statistical difference between groups, or not enough evidence (data) to find a difference

* Assuming $\alpha = 0.05$

Formulas for Test Statistics

Test Statistic	Formula	Finding
T-value for 1-sample t-test	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	Take the sample mean, subtract the hypothesized mean, and divide by the standard error of the mean.
T-value for 2-sample t-test	$t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	Take one sample mean, subtract the other, and divide by the pooled standard deviation.
F-value for F-tests and ANOVA	$F = \dfrac{s_1^2}{s_2^2}$	Calculate the ratio of two variances.
Chi-squared value (χ^2) for a Chi-squared test	$\chi^2 = \sum \dfrac{(Observed - Expected)^2}{Expected}$	Sum the squared differences between observed and expected values divided by the expected values.

Hypothesis Testing

- Hypothesis testing are all the methods and practices used by a researcher to prove or reject a hypothesis by using the sample and population data.
- Null hypothesis is assuming that the particular event will not occur. It is represented by H0 (H-naught).
- Alternative hypothesis is the logical opposite of the null hypothesis and is represented by H1.

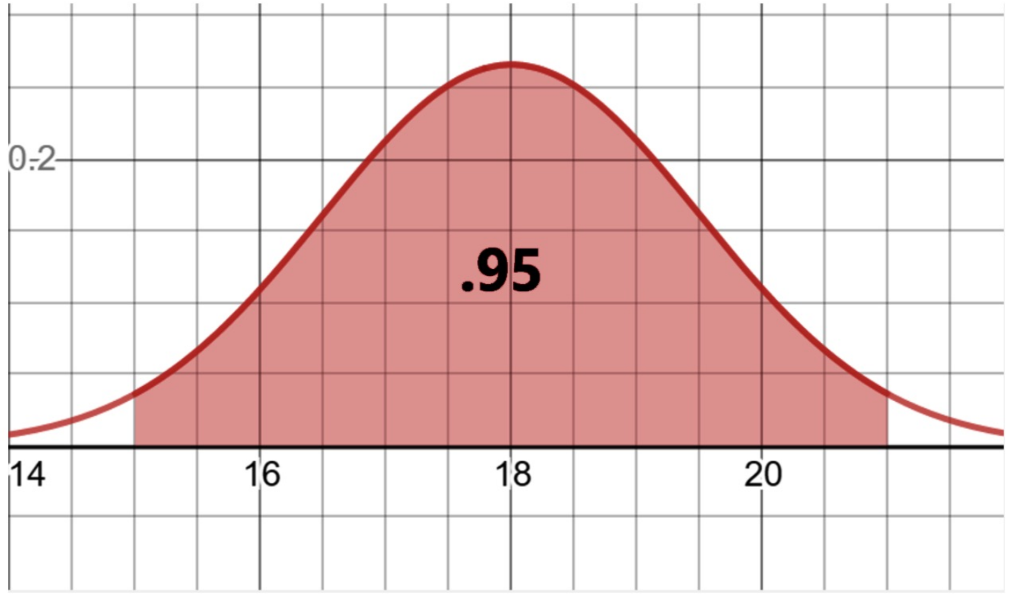
Alternative Hypothesis H_1:			**Null Hypothesis H_0:**
Symbol	Clue words	Type of test	**Symbol**
<	Less than, decreased, faster	Left tailed Test	**=**
>	More than, increased, slower	Right tailed Test	**=**
≠	Not equal to, has changed	Two Tailed Test	**=**

Examples

- We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:
  H_0: $\mu = 2.0$
  H_a: $\mu \neq 2.0$
- We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:
  H_0: $\mu \geq 5$
  H_a: $\mu < 5$

Understanding Hypothesis testing

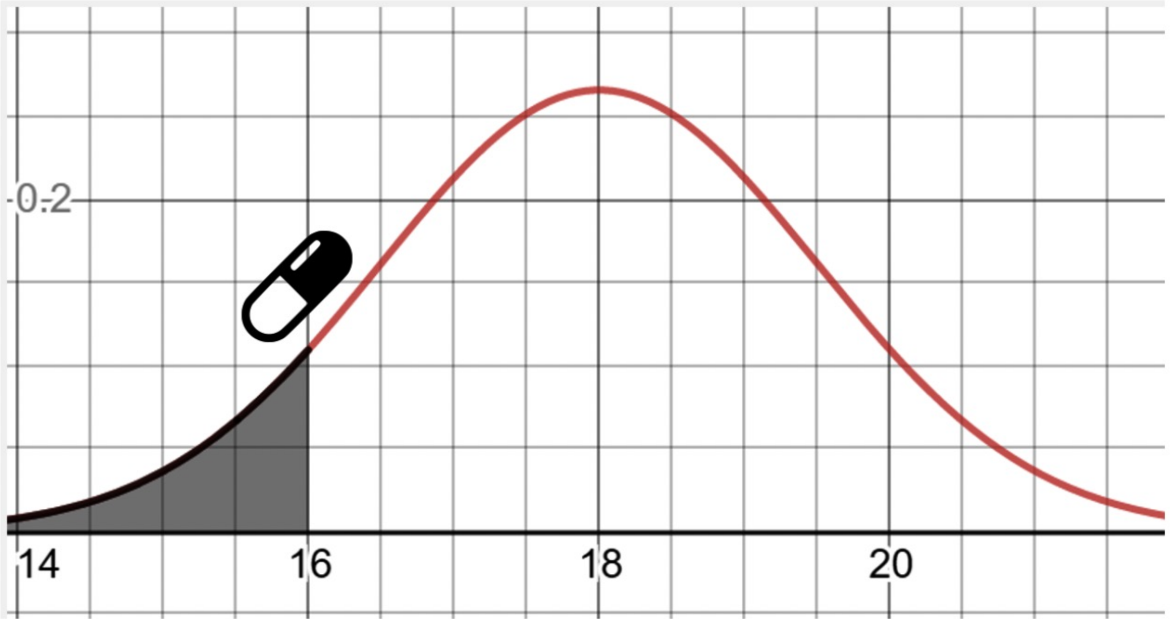
- Lets take an example of recovery time and drug testing for cold.
- Studies suggest that the mean recovery time for common cold is 18 days, with a standard deviation of 1.5 days.



95% chance of recovery between 15 to 21 days

Understanding Hypothesis testing

- 2.5% chances that the recovery will be less than 15 days and 2.5% chances that the recovery will be more than 21 days.
- A new drug was tested on a group of 40 people, and it took them an average of 16 days to recover.

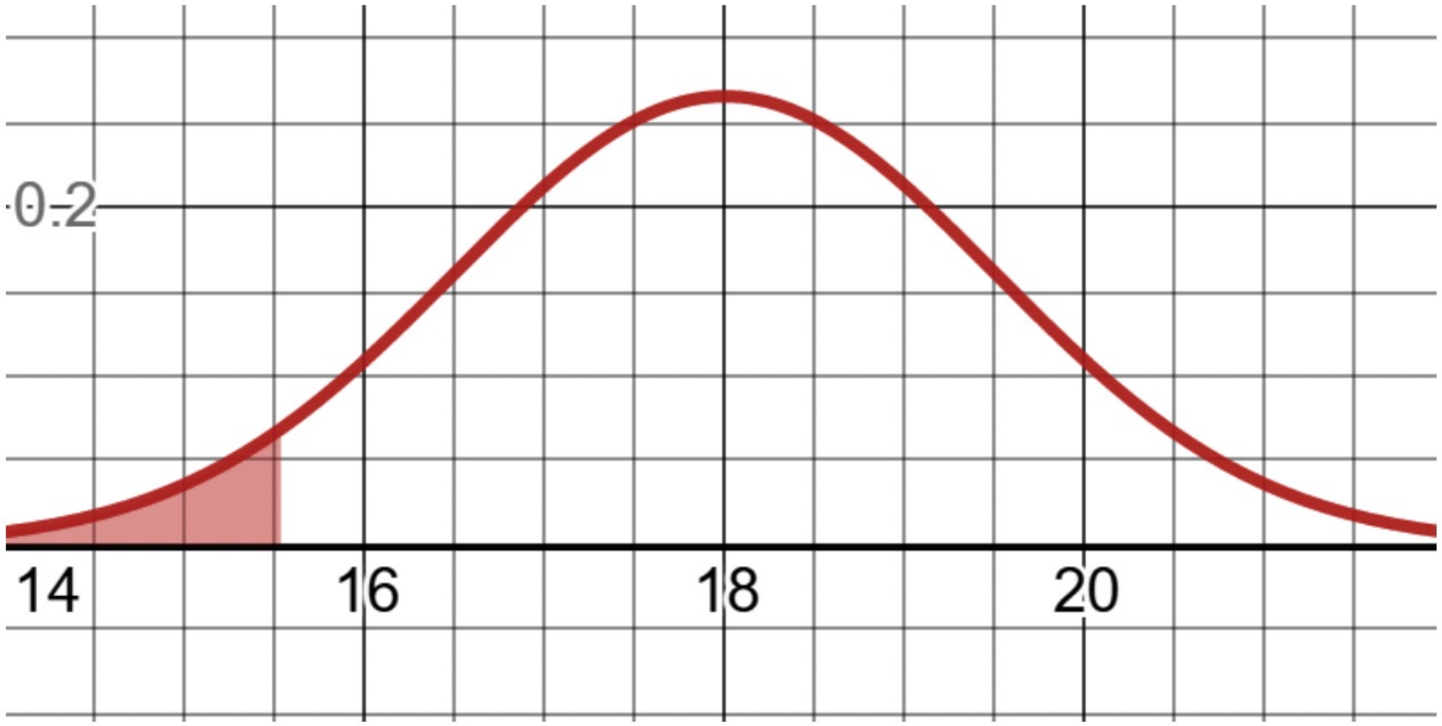


- Null hypothesis: Did the drug not work and the short recovery period was just a coincidence?
- Alternate hypothesis: Does the drug show a significant result and works?
- There are 2 types of tests to calculate this, one tailed test and two tailed test.

One tailed test

- In one tailed test we hypothesize around the mean using inequalities.
- To reject our null hypothesis, we need to show that our sample mean of the patients who took the drug is not likely to have been coincidental. Since a p-value of .05 or less is traditionally considered statistically signficant, we will use that as our threshold.
- We only check one tail on the normal distribution curve in a one tailed test. In this example we will calculate the x – value with 5% of area behind it (since 0.05 is the p-value)

One tailed test



Python code for getting x-value with 5% of area behind it

```
from scipy.stats import norm

# Cold has 18 day mean recovery, 1.5 std dev
mean = 18
std_dev = 1.5

# What x-value has 5% of area behind it?
x = norm.ppf(.05, mean, std_dev)

print(x) # 15.53271955957279
```

One tailed test

- Therefore, if we achieve an average 15.53 or fewer days of recovery time in our sample group, our drug is considered statistically significant enough to have shown an impact. However, our sample mean of recovery time is actually 16 days and does not fall into this null hypothesis rejection zone. Therefore, the statistical significance test has failed.

- The area up to that 16-day mark is our p-value, which is .0912, and we calculate it in Python.
- Since the p-value of .0912 is greater than our statistical significance threshold of .05, we do not consider the drug trial a success and fail to reject our null hypothesis.

```python
from scipy.stats import norm

# Cold has 18 day mean recovery, 1.5 std dev
mean = 18
std_dev = 1.5

# Probability of 16 or less days
p_value = norm.cdf(16, mean, std_dev)

print(p_value) # 0.0912112197258678
```

Two – Tailed test

- In two tailed test we check at both tails of the Normal distribution curve.
- It is a better practice.
- We spread the p value statistical significance to both tails.
- Instead of using inequalities , we use equal and not equal.

$$H_0 : \text{population mean} = 18$$

$$H_1 : \text{population mean} \neq 18$$

Two Tailed test example

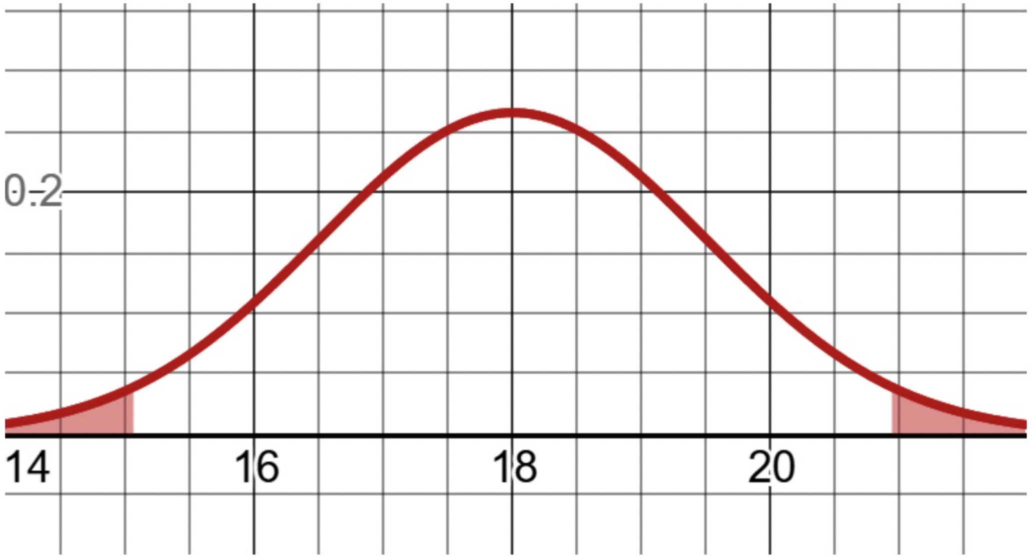
- Using the same example, as the one tailed test.
- Our hypothesis is now whether the drug had any impact on the cold recovery time or not.

$$H_0 : \text{population mean} = 18$$

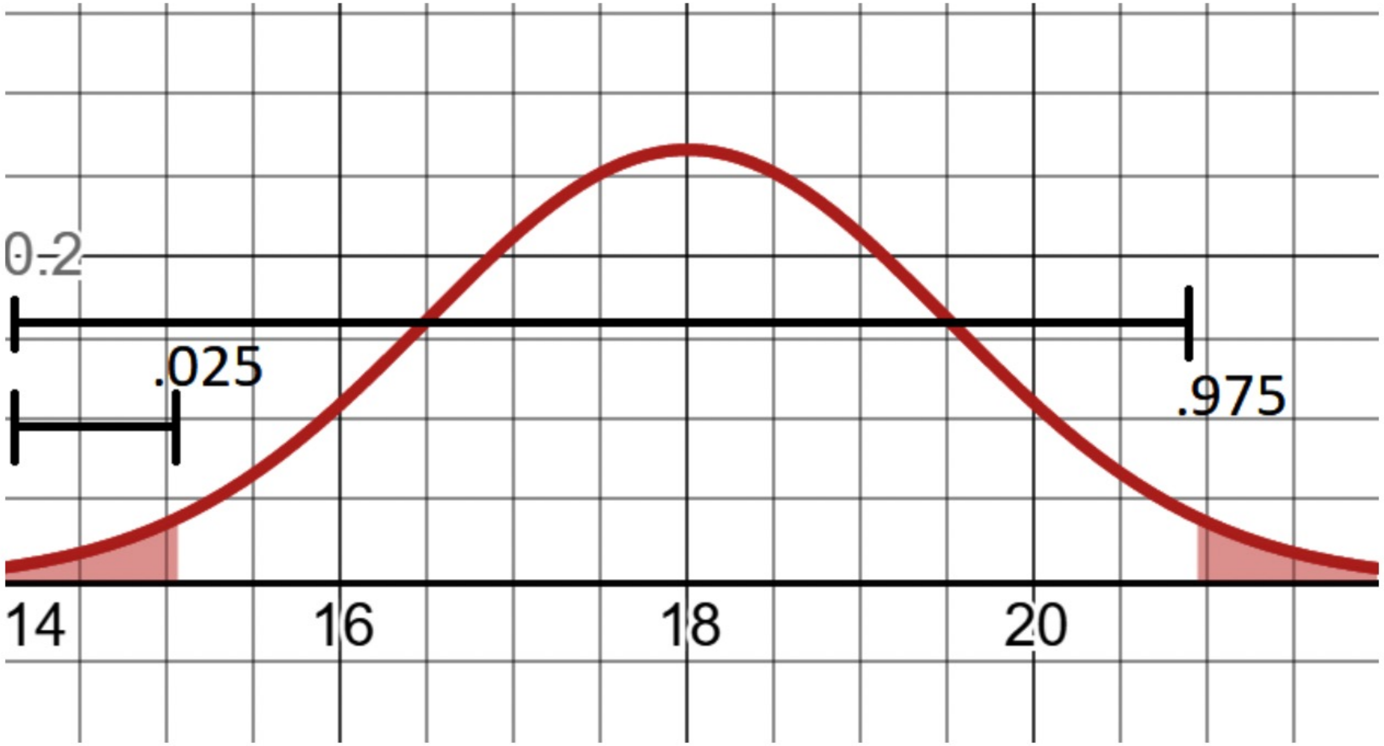
$$H_1 : \text{population mean} \neq 18$$

Two Tailed test

- “We spread our p-value statistical significance threshold to both tails, not just one. If we are testing for a statistical significance of 5%, then we split it and give each 2.5% half to each tail. If our drug’s mean recovery time falls in either region, our test is successful and we reject the null hypothesis.”



- Upper tail, we take 0.975 and lower tail, 0.25



Calculating a range for a statistical significance of 5%

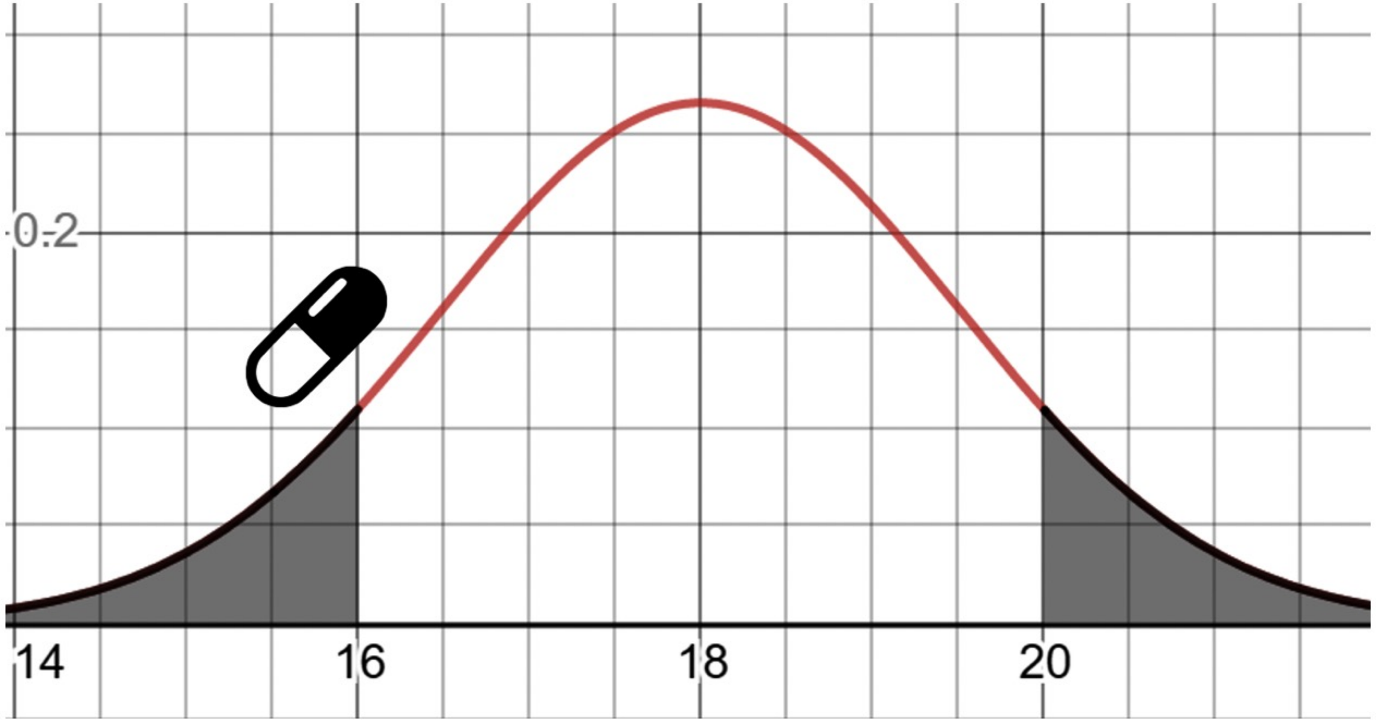
```python
from scipy.stats import norm

# Cold has 18 day mean recovery, 1.5 std dev
mean = 18
std_dev = 1.5

# What x-value has 2.5% of area behind it?
x1 = norm.ppf(.025, mean, std_dev)

# What x-value has 97.5% of area behind it
x2 = norm.ppf(.975, mean, std_dev)

print(x1) # 15.060054023189918
print(x2) # 20.93994597681008
```

Results

- The sample mean value for the drug test group is 16, and 16 is not less than 15.06 nor greater than 20.9399.
- So like the one-tailed test, we still fail to reject the null hypothesis.
- Our drug still has not shown any statistical significance to have any impact

What about the P value?

- When we sum both those areas, we get a p-value of .1824. This is a lot greater than .05, so it definitely does not pass our p-value threshold of 0.05



Errors

- When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	**H_0 IS ACTUALLY**	**...**
	True	False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

Errors

- Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.
- α = probability of a Type I error = ***P*(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.
- β = probability of a Type II error = ***P*(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.
- α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

Example

- Suppose the null hypothesis, H_0, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.
- **Type I error**: The emergency crew thinks that the victim is dead when, in fact, the victim is alive. **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.
- α = **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(Type I error). β = **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = P(Type II error).
- The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Any questions