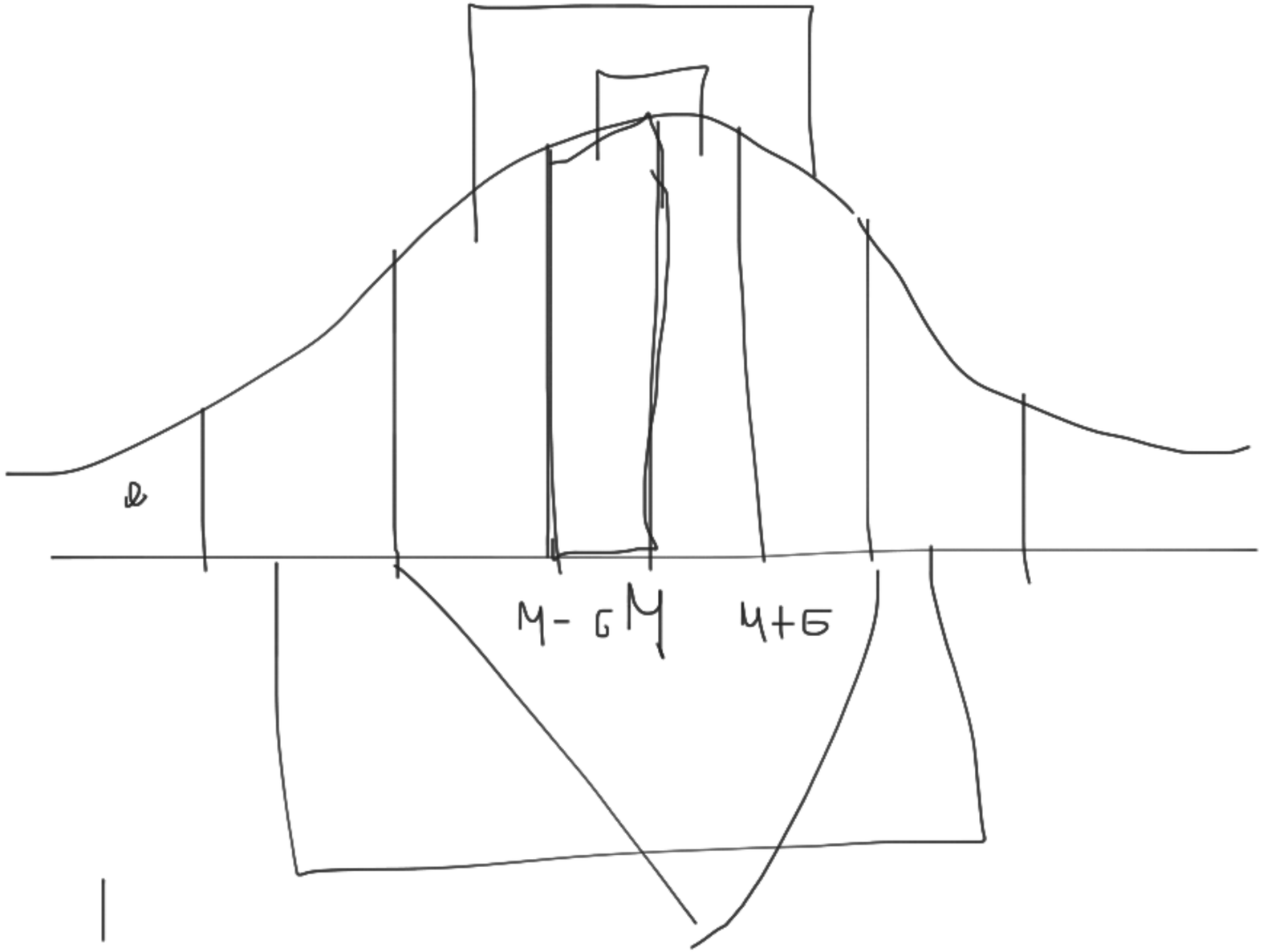

μ- σ
μ
μ+σ

Binomial Distribution

Success

Failure

Experiments

x -> No. of success
n -> No. of trails/observations
p -> Prob. of SUCCESS

Properties:

An experiment is said to be a Binomial experiment if:

1. Each of the event should be independent
2. There should be only two possible outcomes in an event/trail - SUCCESS or FAILURE
3. These kind of events is repeated only a fixed number of times
4. The probability of SUCCESS & FAILURE is same for all the events/trails

Mean = n*p

n -> No. of trails
p -> Prob. of SUCCESS
q -> Prob. of FAILURE

Variance = n*p*q OR n*p*(1-p)

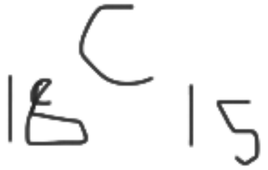
p + q = 1
q = 1-p

$P(X) = {}^{n}C_{x} p^{x} (1-p)^{n-x}$

Q. In a recent survey, it was found that 85 households in the US have high speed internet. If you take a sample of 18 household, what is the probability that exactly 15 will have the high speed internet?

Q. Is this experiment being repeated for a fix no. of times?
YES

Q. Are the events independent?
YES



Q. Are there two mutually exclusive events?
YES

$$P(15) = 18C15*(0.85)^{15}(1 - 0.85)^{18 - 15}$$

$$P(15) = 0.24$$

x = 15
n = 18
p = 0.85

atleast 15 will have High Speed Internet
P(x>=15) = P(15) + P(16) + P(17) + P(18)

Poisson Distribution with interval

This distribution is used when computing the probabiloty of a certain number of success within a specified interval of time.

Properties:

1. The probability of two success in a small enough interval is 0%.
2. The probability of a success is the same for any two intervals which share the same length
3. Successes are independent of success in other event

Mean: λt

SD: $\sqrt{\lambda_t}$

$$P(x) = \frac{(\lambda * t)^{x} e^{-\lambda t}}{x!}$$

x -> No. of success

t -> a length of time

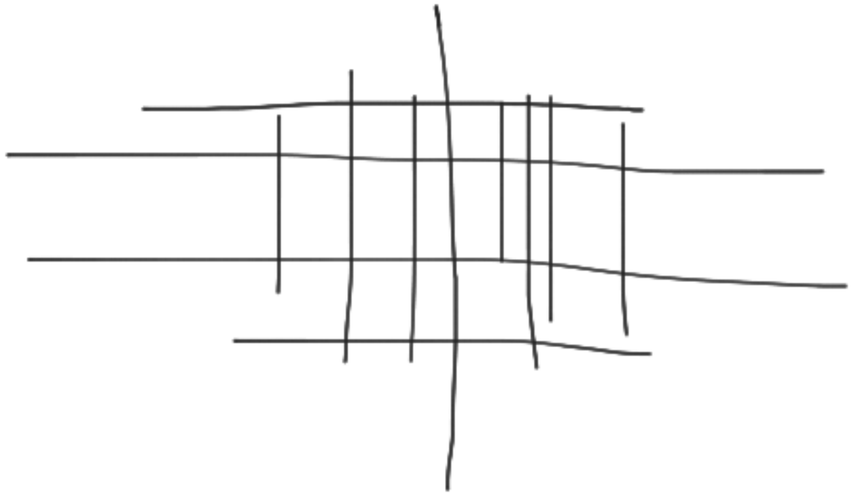
λ -> average no. of success in an interval

At a theme park, there is a roll,er coaster that sends an avg. of 3 cars through its circuit every minute b/w 6PM to 7PM. A random variable 'A' represents the number times the roller coaster allows car to pass through the circuit b/w 6PM to 6:10PM. What is the probability that 35 cars will pass through the circuit b/w 6PM to 6:10PM.

$x = 35$

$t = 10$ (min)

$\lambda = 3$



$$P(35) = \frac{(3 * 10)^{35}}{35!} e^{-30}$$

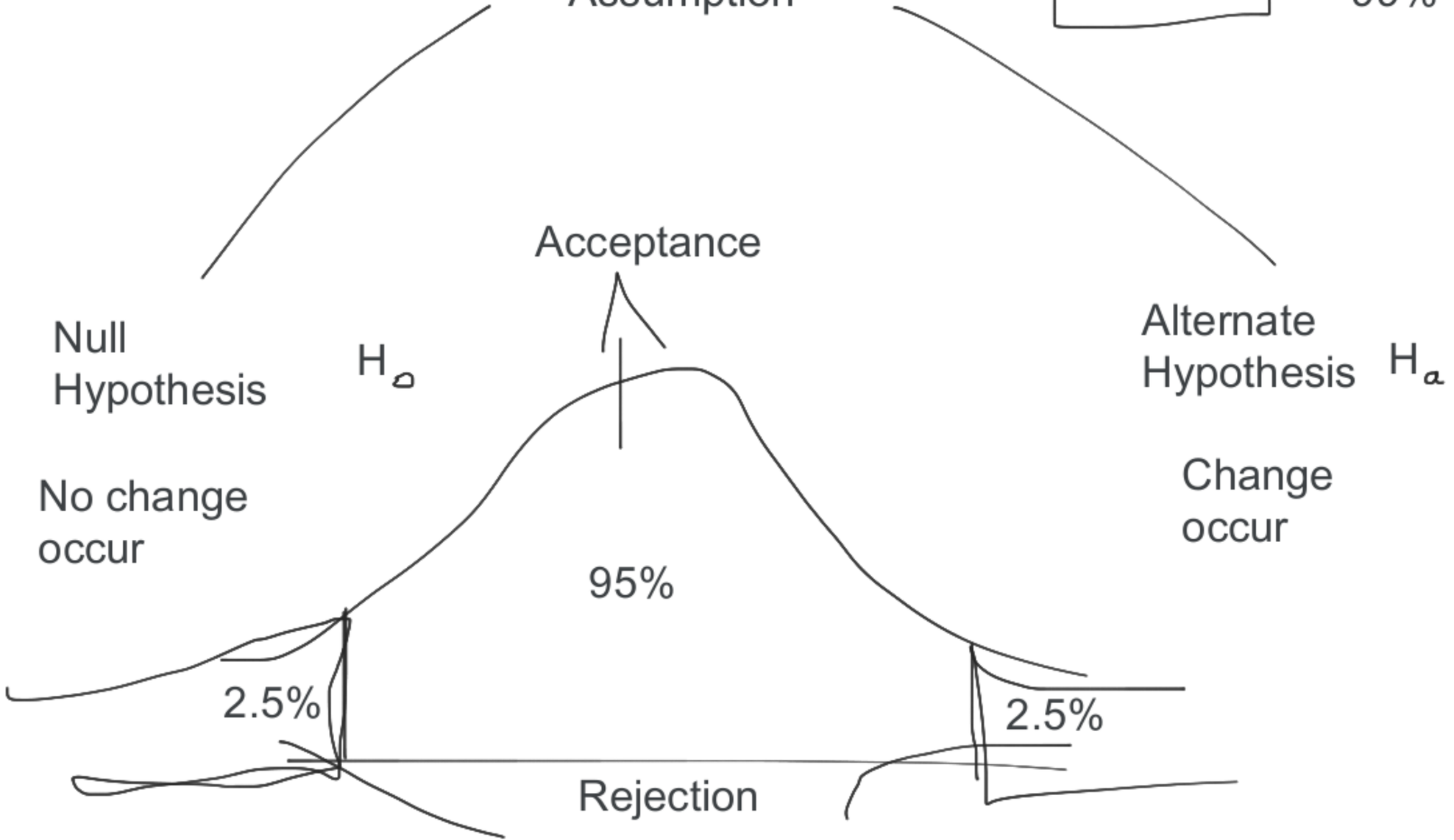
$P(35) = 0.045$

Population
Sample

Hypothesis Testing

Assumption

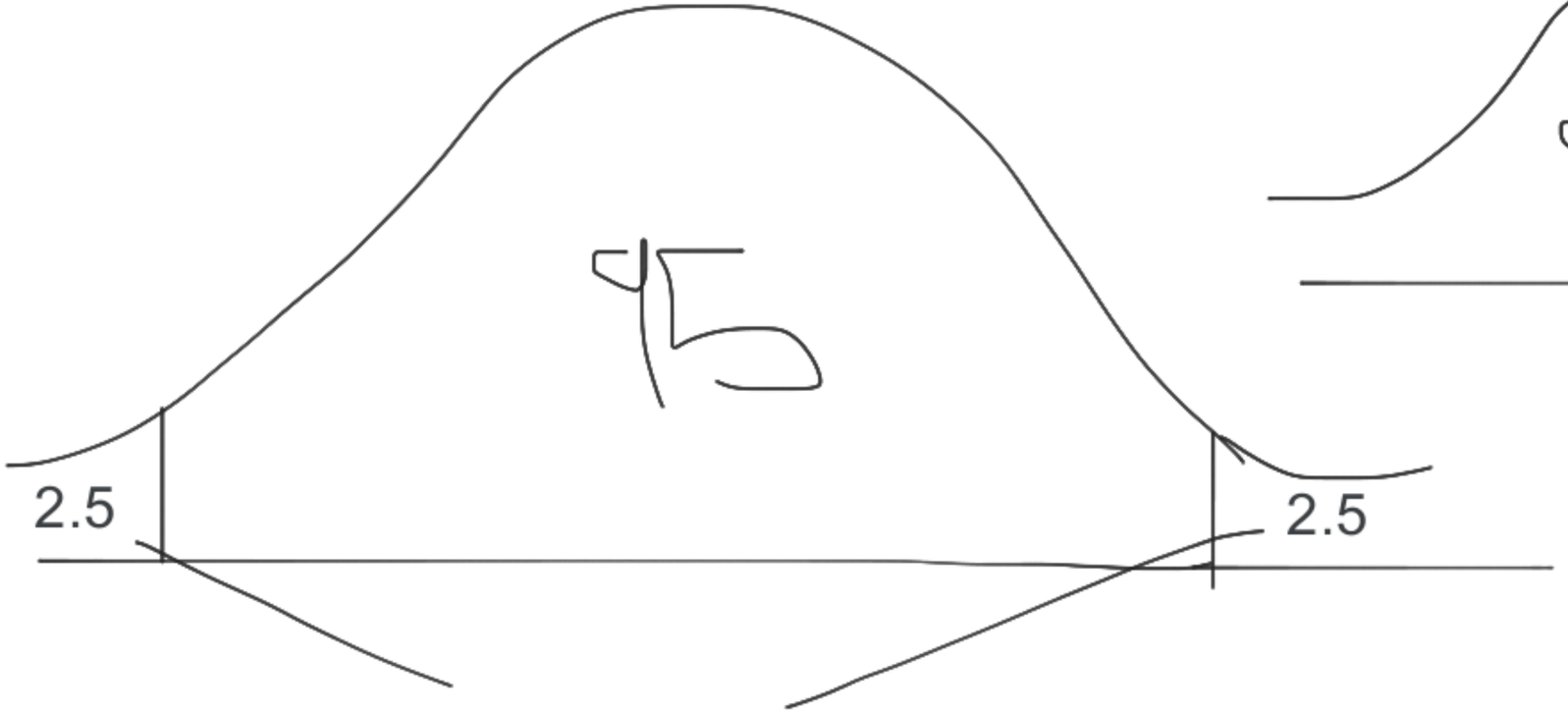
95%
99%



1. Null & Alternate Hypothesis
2. Perform the test: t-test, z-test, ANOVA
3. Level of significance ( )
4. Level of confidence ( )
5. Accept/Reject the NULL Hypothesis

$\alpha + c = 1$

One tailed & Two Tailed test



z-test

$$\frac{x_i - x_{mean}}{SD}$$

Use case scenario:

1. Population variance shold be known, or
2. If we don't know the population variance, but the sample size is large (n>30)
3. SD & Mean of the population should be known

*Note:. if we have a sample size less than 30 and we don't know the population variance, then we must use the Student t-test.

One sample z-test

When you want to compare a sample mean with the population mean.

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$\bar{x}$ - Sample Mean
μ - Population Mean
σ - Population SD
n - Sample Size

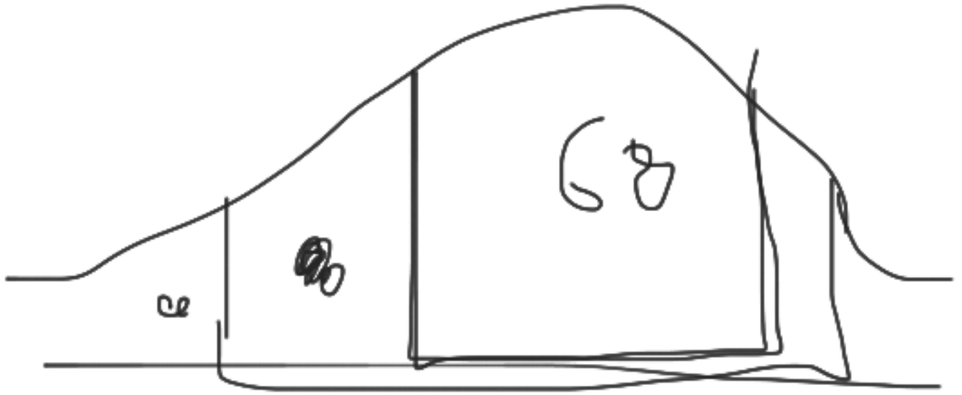
Two sample z-test

When you want to compare the mean of two samples.

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

```
def z_test(data):
    outlier = []
    mean_data = np.mean(data)
    SD_data = np.std(data)
    threshold = 3


    for i in data:
        z = (i - mean_data)/SD_data
        if z > threshold:
            outlier.append(i)
    return outlier
```



Student t-test

Use scenario:

1. Population variance is unknown
2. The same size is less than 30 (n<30)
3. Data points should be independent

One sample test

When you want to compare a sample mean with the population mean.

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

Two sample test

When you want to compare the mean of two samples.

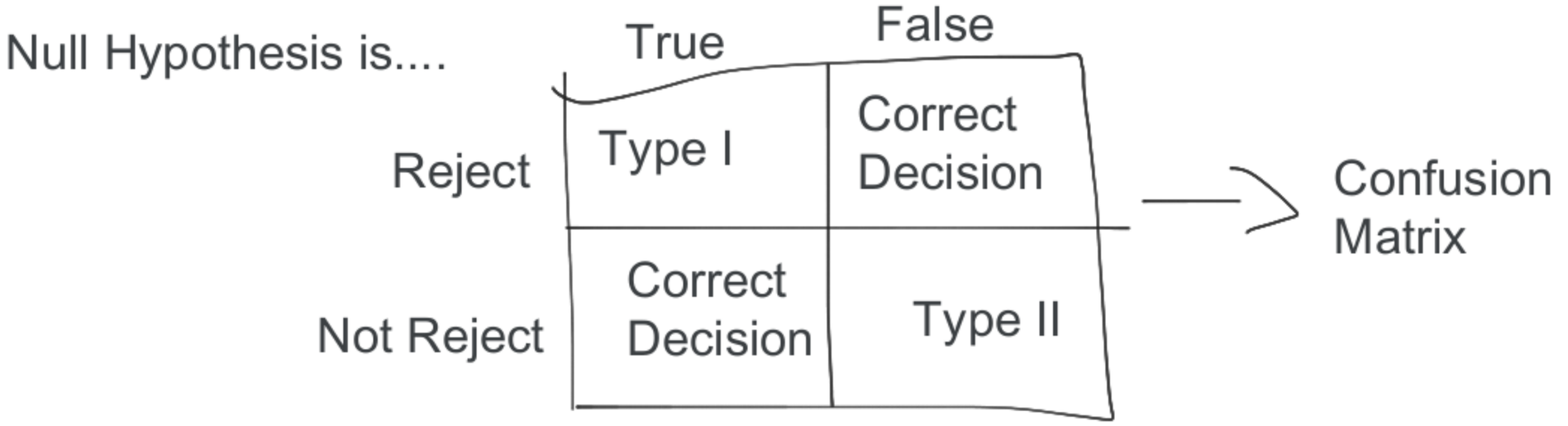
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}}$$

Gender	Age	Weight (kg)	Height(m)
M	Elderly	70	1.4
M	Adult	60	1.2
F	Adult	65	1.4
F	Chjild	45	1.0
F	Adult	78	1.3
M	Elderly	67	1.3
F	Adult	65	1.9
F	Adult	56	1.3

H_0	Most of the Elderly age group people is M
H_1	Most of the Elderly age group people is not M
Test	

One categorical feature -> One sample proportion test
Two categorical feature -> Chi Square test
One numerical feature -> Student t-test
Two numerical feature -> Correlation
Atleast one numerical & categorical variable - ANOVA

Type 1 & Type 2 Error



Type I Error - Rejecting the Null Hypothesis when in reality it is True.

Type II Error - Accepting the Null Hypothesis when in reality it is False.