Inferential Statistics

Nidhi Gangrade
FactualAI

Descriptive Statistics is concerned with Data Summarization, Graphs/Charts, and Tables

Inferential Statistics is a method used to talk about a Population Parameter from a Sample.

Data Exploration

To Check What has Already happened in Data

To Check What might happen in Data

To Explain What We have Found in Data !!!

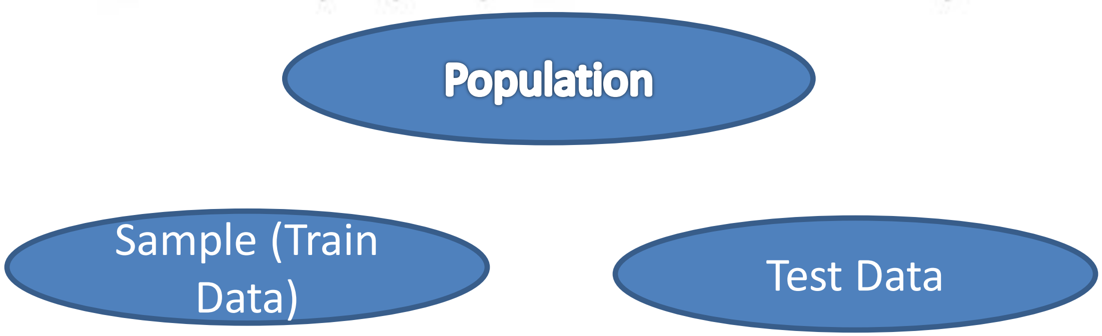
Population, Parameter, Sample, Statistic

A Population is the universe of possible data for a specified object. Example: People who have visited or will visit a website.

A Parameter is a numerical value associated with a population. Example: The average amount of time people spend on a website.

A Sample is a selection of observations from a population. Example: People (or IP addresses) who visited a website on a specific day.

A Statistic is a numerical value associated with an observed sample. Example: The average amount of time people spent on a website on a specific day.



Aim: Estimate a statistical property (mean) of the population

Will need to do so from a sample

Use properties of sample to estimate property of population

Mean and Variance

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

These statistics only apply to the sample of data, and so are known as **sample statistics**

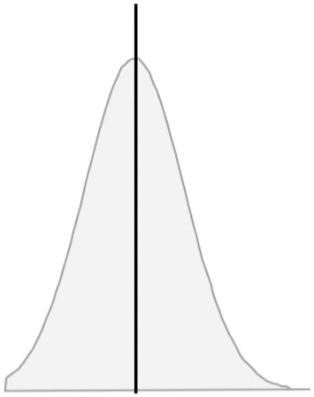
The corresponding figures for all possible data points out there are called **population statistics**

Distribution

All values are equally likely

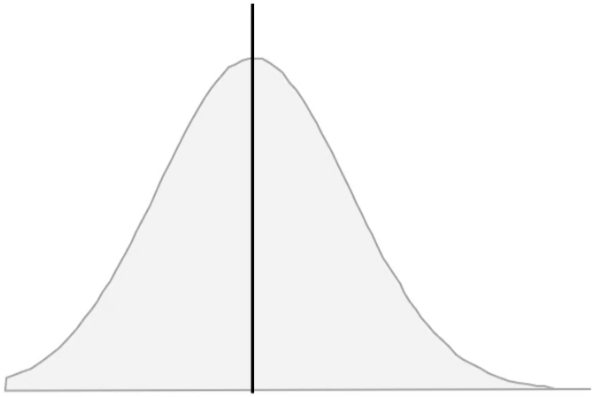
Values close to the mean are more likely

Role of Sigma



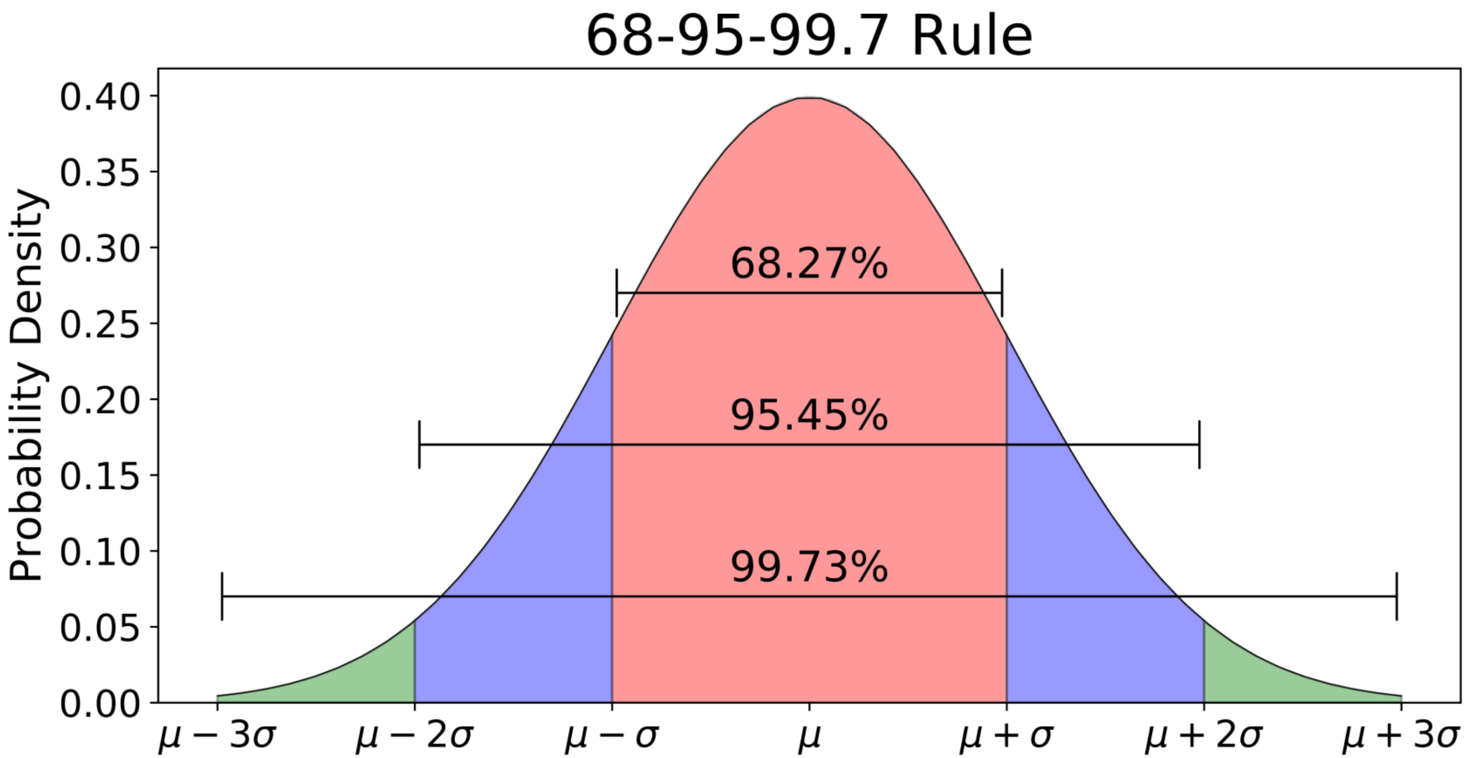
Small Standard Deviation

Few points far from the mean

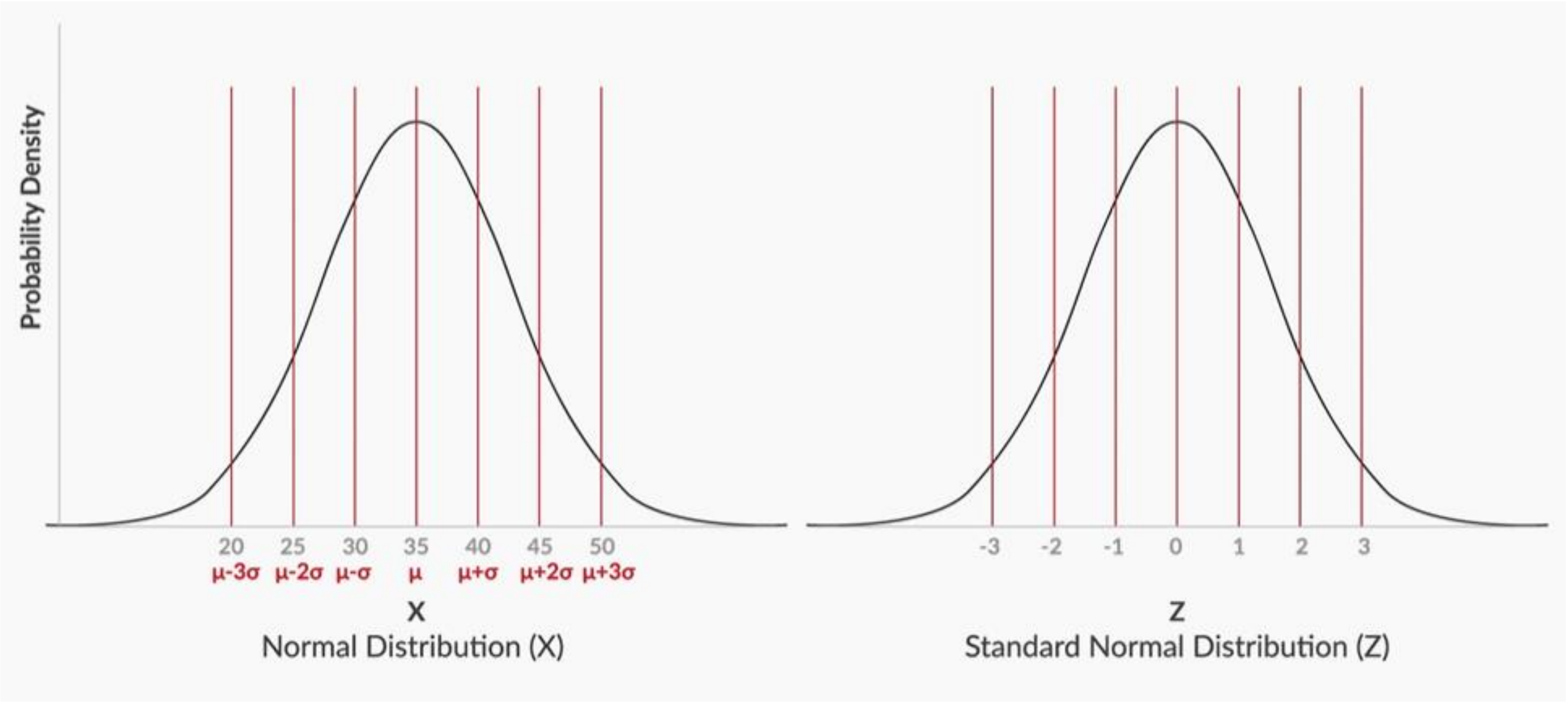


Large Standard Deviation

Many points far from the mean



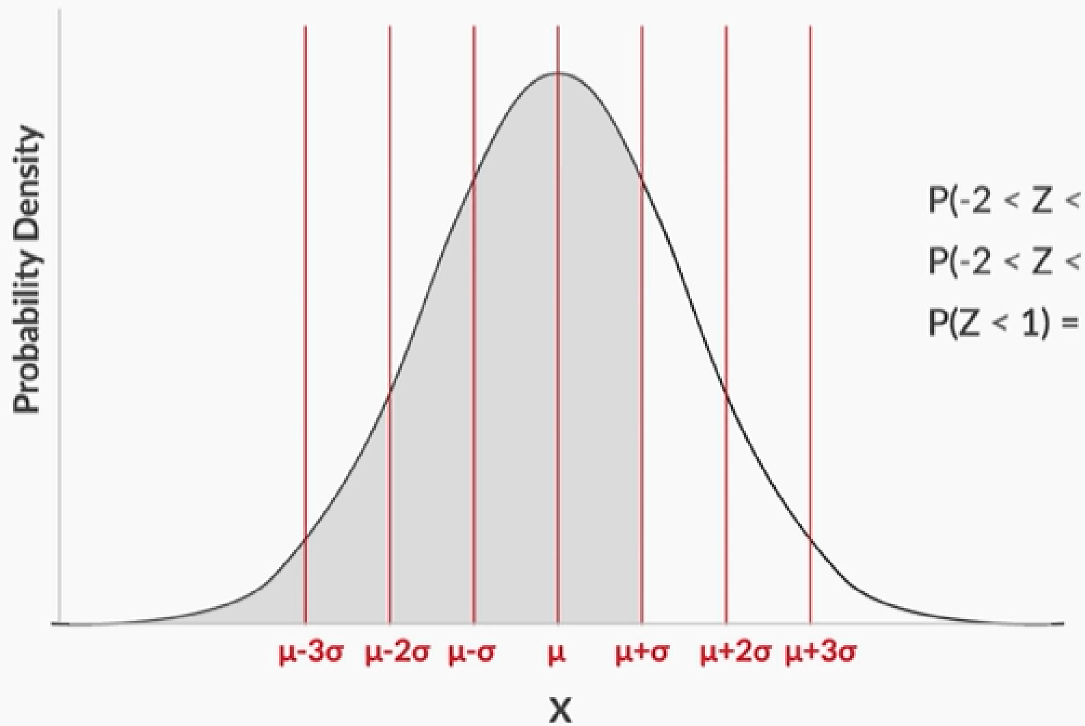
Normal Distribution 1-2-3 Rules



The **standardized random variable** is an important parameter. It is given by:

$$Z = X - \mu / \sigma$$

FINDING THE PROBABILITY FOR STANDARDISED NORMAL VARIABLE Z



$P(-2 < Z < 2) = P(\mu-2\sigma < X < \mu+2\sigma) \approx 95\%$

$P(-2 < Z < 3) = P(\mu-2\sigma < X < \mu+3\sigma) \approx 97.35\%$

$P(Z < 1) = P(X < \mu+\sigma) \approx 84\%$

Extrapolating Sample Statistic to Population

For example, for an office of 10,000 employees, we wanted to find the average commute time. So, instead of asking all employees, we asked only 100 of them and collected the data. The mean = 30.5 minutes and the standard deviation = 10 minutes.

However, it would not be fair to infer that the population mean is exactly equal to the sample mean. This is because the flaws of the sampling process must have led to some error. Hence, the sample mean's value has to be reported with some **margin of error**.

For example, the mean commute time for the office of 30,000 employees would be equal to 30.5 $\pm$ 5 minutes, 30.5 $\pm$ 2 minutes, or 30.5 $\pm$ 10 minutes, i.e.
30.5 minutes $\pm$ some margin of error.

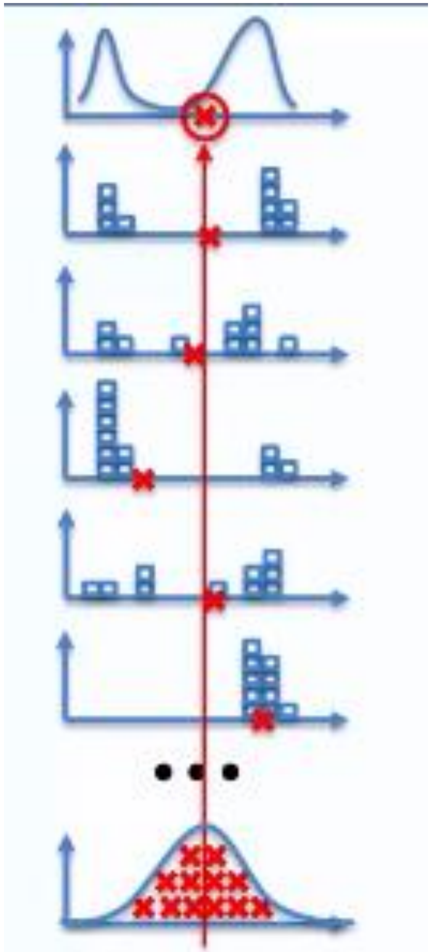
However, at this point in time, you do not exactly know how to find what this margin of error .Hence, we move on to sampling distributions, which have some properties that would help you find this margin of error.

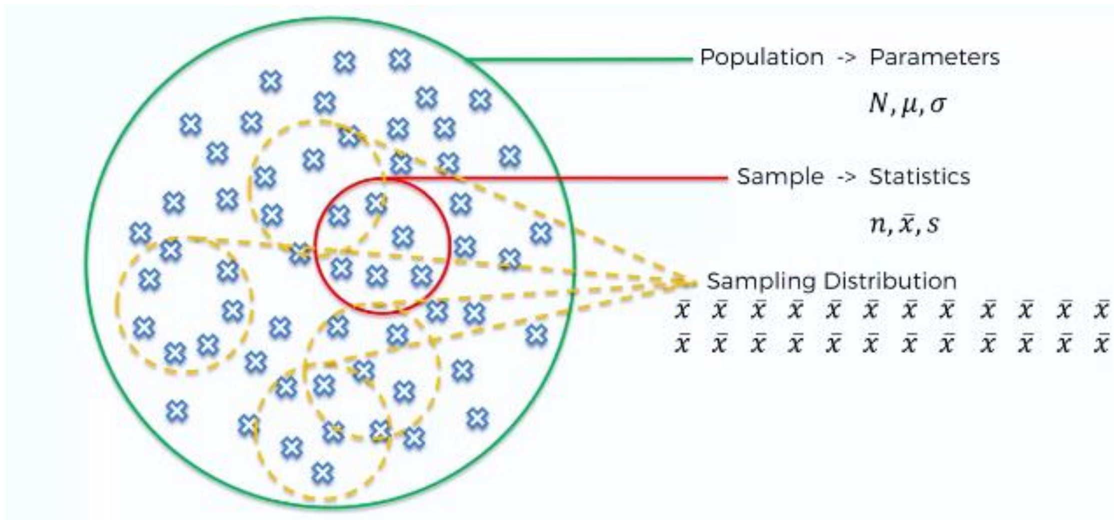
Sampling distributions

A vital concept in inferential statistics is that the particular random sample that you draw for a study is just one of a large number of possible samples that you could have pulled from your population of interest. Understanding this broader context of all possible samples and how your study's sample fits within it provides valuable information.

Suppose we draw a substantial number of random samples of the same size from the same population and calculate the sample mean for each sample. During this process, we'd observe a broad spectrum of sample means, and we can graph their distribution.

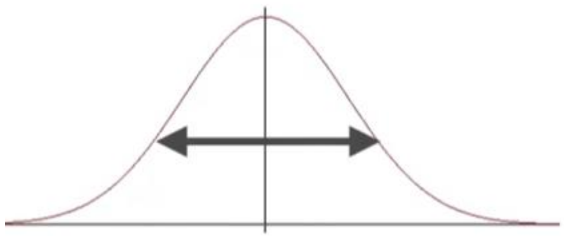
This type of distribution is called a **sampling distribution**. Sampling distributions allow you to determine the likelihood of obtaining different sample values, which makes them crucial for performing hypothesis tests.





Population -> Parameters
N, μ, σ
Sample -> Statistics
$n, \bar{x}, s$
Sampling Distribution
$\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$
$\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$

Sampling Distribution



Population mean μ has a distribution called the sampling distribution

This is a normal distribution

- Mean = Sample mean
- Variance ≈ Sample variance / n
- Std dev. = Sample std dev. / sqrt(n)

Central Limit Theorem (CLT)

Mathematically, confidence in our estimate depends upon

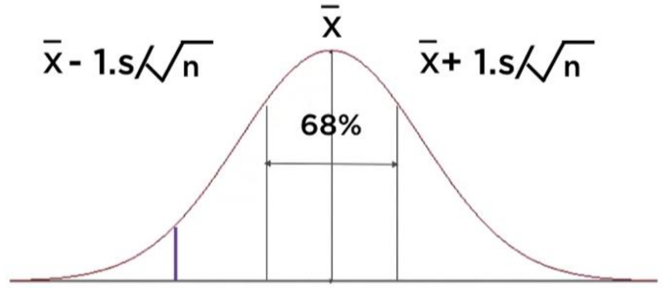
- Sample variance
- Sample size

What exactly is ϭ / Sqrt(n)---Standard Error

This value of $\sigma/\sqrt{n}$ isn't the standard deviation of the sample or the population but rather it is the standard deviation of the sampling distribution for the given sample. We conduct our hypothesis test on the sampling distribution only. Thus when the standard deviation of the population isn't given, we take the sample's standard deviation to calculate the sampling distribution's standard deviation.

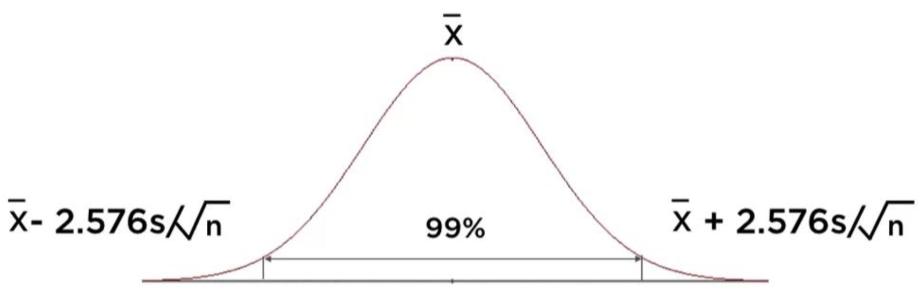
This is a very important concept to know and you should be pretty confident in how we are manipulating the values here. We used this same concept in calculating the confidence intervals. In that case we used $S/\sqrt{n}$ instead which denoted that we can only use the sample's standard deviation in that case.

68% Confidence That μ is within 1σ of $\bar{x}$

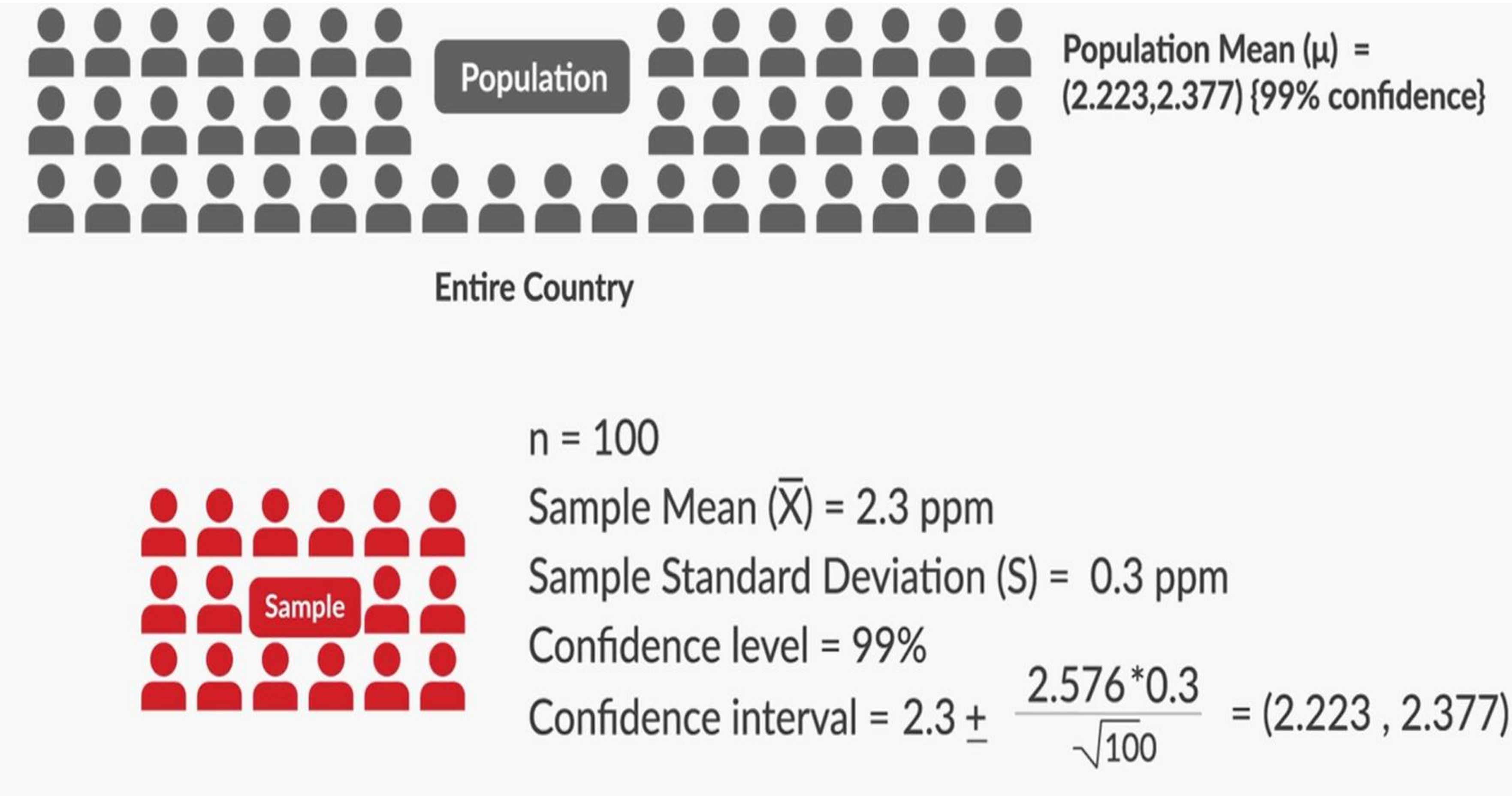


We can state with 68% confidence that the population mean μ lies in the range $\bar{x} - 1.s/\sqrt{n}$ to $\bar{x} + 1.s/\sqrt{n}$

99% Confidence That μ is within 2.57σ of $\bar{x}$



We can state with 99% confidence that the population mean μ lies in the range $\bar{x} - 2.576s/\sqrt{n}$ to $\bar{x} + 2.576s/\sqrt{n}$


Population
Population Mean (µ) = (2.223,2.377) {99% confidence}
Entire Country
n = 100
Sample Mean ($\bar{X}$) = 2.3 ppm
Sample Standard Deviation (S) = 0.3 ppm
Sample
Confidence level = 99%
Confidence interval = $2.3 \pm \frac{2.576*0.3}{\sqrt{100}}$ = (2.223 , 2.377)

Connecting the Dots

Descriptive Statistics

Explore the data

No points-of-view yet

Inferential Statistics

Frame hypotheses and test them

Tentatively evaluating many points-of-view

Machine Learning Models

Build models that change with the data

Full circle - back to no points-of-view

Hypothesis Testing

Sometimes, the research question is less ambitious in the sense that we are not interested in precise estimates of a parameter, but we only want to examine whether a statement about a parameter of interest or the research hypothesis is true or not.

Another related issue is that once an analyst estimates the parameters on the basis of a random sample, (s)he would like to infer something about the value of the parameter in the population. Statistical hypothesis tests facilitate the comparison of estimated values with hypothetical values.

As a simple example, consider the case where we want to find out whether the proportion of votes for a party P in an election will exceed 30% or not.

Typically, before the election, we will try to get representative data about the election proportions for different parties(e.g.by telephone interviews)and then make a statement like “yes”, we expect that P will get more than 30% of the votes or “no”, we do not have enough evidence that P will get more than 30% of the votes.

In such a case, we will only know after the election whether our statement was right or wrong. Note that the term representative data only means that the sample is similar to the population with respect to the distributions of some key variables, e.g. age, gender, and education.

Since we use one sample to compare it with a fixed value (30%), we call it a **one-sample problem.**

Consider another example in which a clinical study is conducted to compare the effectiveness of a new drug (B) to an established standard drug (A) for a specific disease, for example too high blood pressure.

Assume that, as a first step, we want to find out whether the new drug causes a higher reduction in blood pressure than the already established older drug.

A frequently used study design for this question is a randomized (i.e. patients are randomly allocated to one of the two treatments) controlled clinical trial (double blinded, i.e. neither the patient nor the doctor know which of the drugs a patient is receiving during the trial), conducted in a fixed time interval, say 3months. A possible hypothesis is that the average change in the blood pressure in group B is higher than in group A, i.e. $\delta B > \delta A$ where $\delta j = \mu j0 - \mu j3$, $j = A, B$ and $\mu j0$ is the average blood pressure at baseline before measuring the blood pressure again after 3months ($\mu j3$).

Note that we expect both the differences δA and δB to be positive, since otherwise we would have some doubt that either drug is effective at all. As a second step (after statistically proving our hypothesis), we are interested in whether the improvement of B compared to A is relevant in a medical or biological sense and is valid for the entire population or not. Since we are comparing two drugs, we need to have two samples from each of the drugs; hence, we have a two sample problem. Since the patients receiving A are different from those receiving B in this example, we refer to it as a **"two-independent-samples problem".**

In another example, we consider an experiment in which a group of students receives extra mathematical tuition. Their ability to solve mathematical problems is evaluated before and after the extra tuition.

We are interested in knowing whether the ability to solve mathematical problems increases after the tuition, or not. Since the same group of students is used in a pre–post experiment, this is called a **"two-dependent-samples problem" or a "paired data problem".**

- **Hypothesis: A Claim or Assumption you make on one or more Population Parameters.**

- **Formulation of Hypothesis: Null Hypothesis (Status-Quo(=, <= or >=)**
  Alternate Hypothesis (No Equality)

- **Types of Test: Two-Tailed , Left Tailed, Right Tailed Test**

- **Taking Decision Using: Critical Value Method**
  p-Value Method

- **Types of Error: Type I Error**
  Type II Error

Hypothesis Tests

Hypothesis Testing is a method of statistical inference.

A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. These two statements are called the null hypothesis and the alternative hypothesis.

Hypothesis tests are not 100% accurate because they use a random sample to draw conclusions about entire populations. When you perform a hypothesis test, there are two types of errors related to drawing an incorrect conclusion.

Type I error: The rejects a null hypothesis that is true. You can think of this as a false positive.
Type II error: The test fails to reject a null hypothesis that is false. You can think of this as a false negative.

A test result is statistically significant when the sample statistic is unusual enough relative to the null hypothesis that you can reject the null hypothesis for the entire population. "Unusual enough" in a hypothesis test is defined by how unlikely the effect observed in your sample is if the null hypothesis is true.
If your sample data provide sufficient evidence, you can reject the null hypothesis for the entire population. Your data favor the alternative hypothesis.

Type I and Type II Error

If we undertake a statistical test, two types of error can occur.

- The hypothesis H0 is true but is rejected; this error is called type I error.
- The hypothesis H0 is not rejected although it is wrong; this is called type II error.

The significance level is the probability of type I error, $P(H1|H0) = \alpha$, which is the probability of rejecting H0 (accepting H1) if H0 is true. If we construct a test, the significance level α is pre specified, e.g. $\alpha = 0.05$.

A significance test is constructed such that the probability of a type I error does not exceed α while the probability of a type II error depends on the true but unknown parameter values in the population(s) and the sample size. Therefore, the two errors are not symmetrically treated in a significance test. In fact, the type II error β, $P(H0|H1) = \beta$ is not controlled by the construction of the test and can become very high, sometimes up to $1-\alpha$. This is the reason why a test not rejecting H0 is not a (statistical) proof of H0.

In mathematical statistics, one searches for the best test which maintains α and minimizes β. Minimization of both α and β simultaneously is not possible. The reason is that when α increases then β decreases and vice versa. So one of the errors needs to be fixed and the other error is minimized. Consequently, the error which is considered more serious is fixed and then the other error is minimized. The tests discussed in the this sections are obtained based on the assumption that the type I error is more serious than the type II error.

So the test statistics are obtained by fixing α and then minimizing β. In fact, the null hypothesis is framed in such a way that it implies that the type I error is more serious than the type II error. The probability $1-\beta = P(H1|H1)$ is called the power of the test. It is the probability of making a decision in favor of the research hypothesis H1, if it is true, i.e. the probability of detecting a correct research hypothesis.

Relationship between confidence intervals and hypothesis tests

There is an interesting and useful relationship between confidence intervals and hypothesis tests. If the null hypothesis H0 is rejected at the significance level α, then there exists a $100(1-\alpha)\%$ confidence interval which yields the same conclusion as the test: if the appropriate confidence interval does not contain the value θ0 targeted in the hypothesis, then H0 is rejected. We call this duality.

The probability of a type I error: the most extreme values in the two tails together have 5% probability and are just the probability that the test statistic falls into the critical region although H0 is true. Also, these areas are those which have the least probability of occurring if H0 is true. For $\alpha = 0.05$, we get $z\,(1-\alpha/2) = 1.96$

For example, a factory where a six sigma quality control system has been implemented requires that errors never add up to more than the probability of being six standard deviations away from the mean (an incredibly rare event). Type I error is generally reported as the p-value.

Statistics derives its power from random sampling.

The argument is that random sampling will average out the differences between two populations and the differences between the populations seen post "treatment" could be easily traceable as a result of the treatment only.

Obviously, life isn't as simple. There is little chance that one will pick random samples that result in significantly same populations. Even if they are the same populations, we can't be sure whether the results that we are seeing are just one time (or rare) events or actually significant (regularly occurring) events.

The level of significance α of a hypothesis test is the same as the probability of a **type 1 error**.

Therefore, by setting it lower, it reduces the probability of a **type 1 error**. "Setting it lower" means you need stronger evidence against the null hypothesis H0 (via a lower p -value) before you will reject the null.

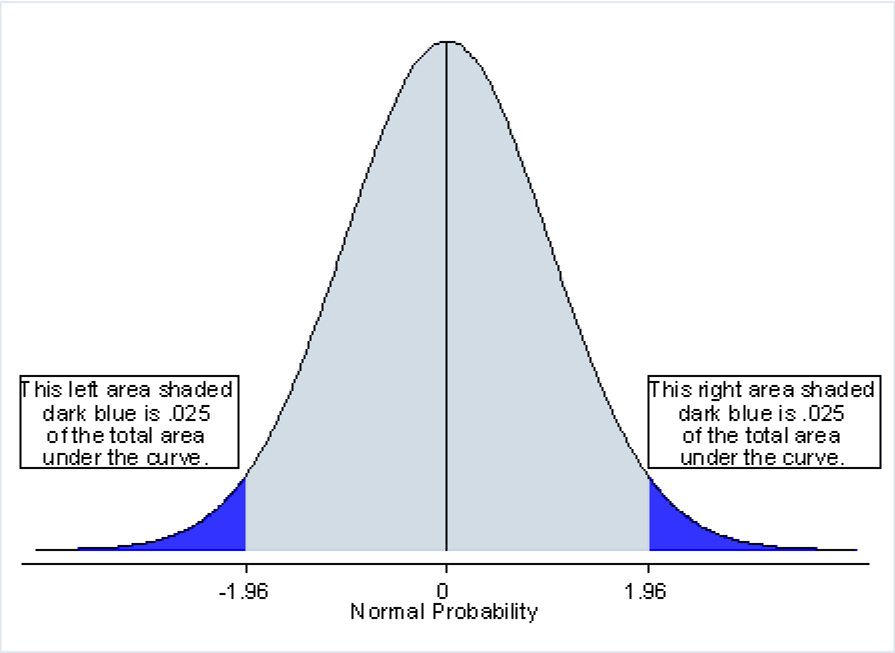
Type of the Test

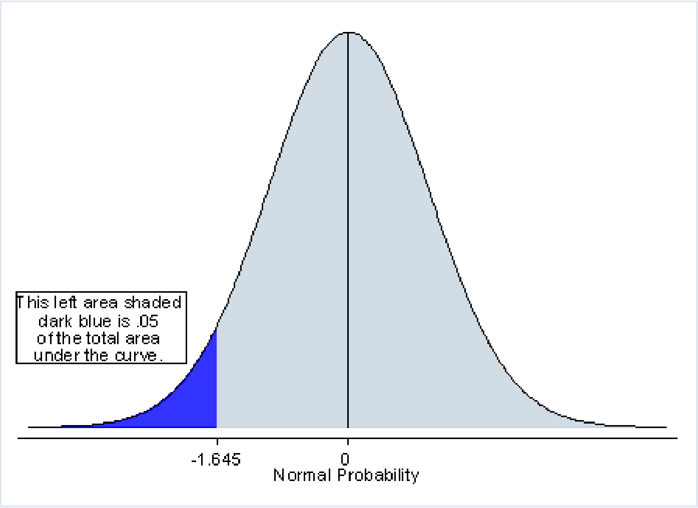
You can tell the type of the test and the position of the critical region on the basis of the **'sign' in the alternate hypothesis.**

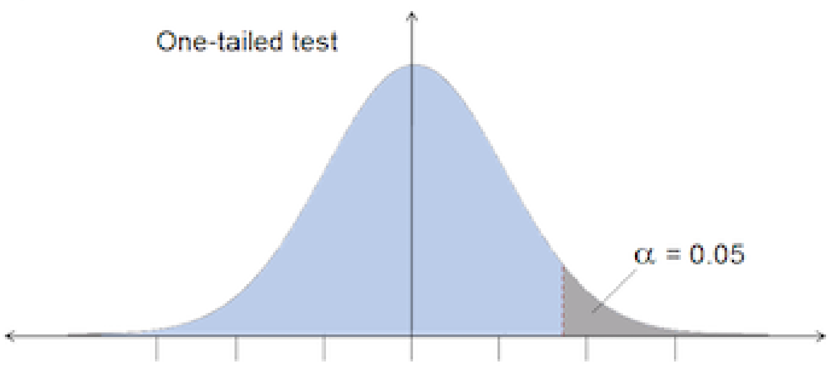
$\neq$ in H_1 → Two-tailed test → Rejection region on **both sides** of distribution

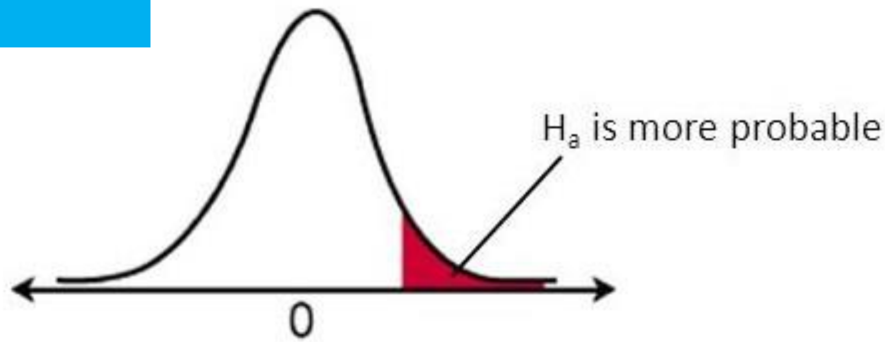
$<$ in H_1 → Lower-tailed test → Rejection region on **left side** of distribution

$>$ in H_1 → Upper-tailed test → Rejection region on **right side** of distribution


This left area shaded dark blue is .025 of the total area under the curve.
This right area shaded dark blue is .025 of the total area under the curve.
-1.96
0
1.96
Normal Probability

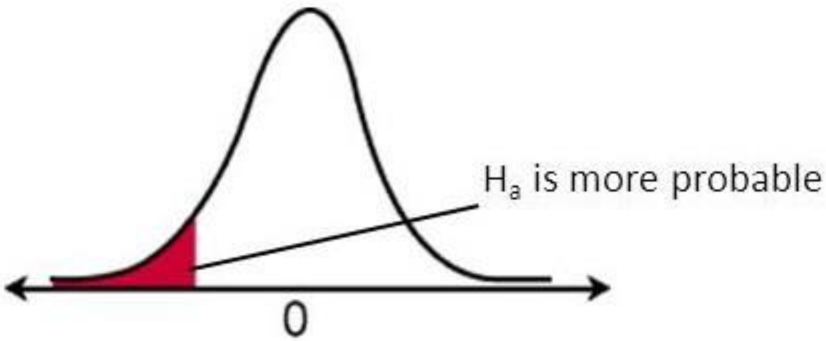

This left area shaded dark blue is .05 of the total area under the curve.
-1.645
0
Normal Probability


One-tailed test
α = 0.05



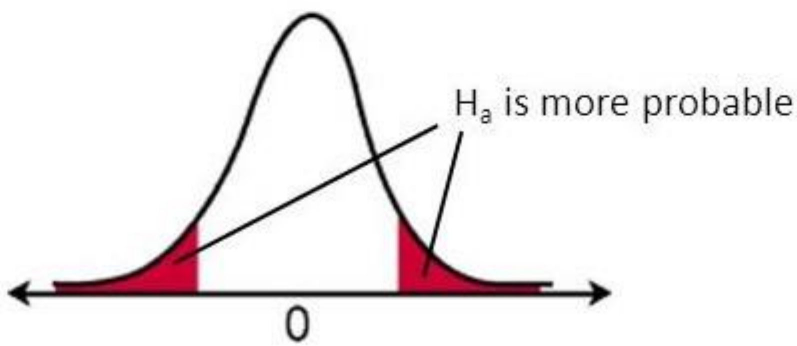
Right-tail test

$$H_a\colon \mu > \text{value}$$



Left-tail test

$$H_a\colon \mu < \text{value}$$



Two-tail test

$$H_a\colon \mu \neq \text{value}$$

Critical-value Method

After formulating the hypothesis, the steps you have to follow to **make a decision** using **the critical value method** are as follows:

- Calculate the value of Zċ from the given value of α (significance level). Take it a 5% if not specified in the problem.
- Calculate the critical values (UCV and LCV) from the value of Zċ.
- Make the decision on the basis of the value of the sample mean x with respect to the critical values (UCV AND LCV).

P-value Method

After formulating the null and alternate hypotheses, the steps to follow in order to **make a decision** using the **p-value method** are as follows:

- Calculate the value of z-score for the sample mean point on the distribution

- Calculate the p-value from the cumulative probability for the given z-score using the z-table

- Make a decision on the basis of the p-value (multiply it by 2 for a two-tailed test) with respect to the given value of α (significance value).

Suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it's more than that. So you conduct a hypothesis test and randomly sample some delivery times to test the claim:

Null hypothesis — The mean delivery time is 30 minutes or less
Alternative hypothesis — The mean delivery time is greater than 30 minutes
The goal here is to determine which claim — the null or alternative — is better supported by the evidence found from our sample data.

We'll use one-tailed test in this case since we **only care about if the mean delivery time is greater than 30 minutes**. We'll disregard the possibility in the other direction since the consequences of having a mean delivery time lower or equal to 30 minutes are even more preferable. **What we want to test here is to see if there is a chance that the mean delivery time is greater than 30 minutes. In other words, we want to see if the pizza place lied to us somehow.** ☺
One of the common ways to do the hypothesis testing is to use Z-test.

If the pizza delivery time is 30 minutes or less (null hypothesis is true), how surprising is my evidence ?

P-value answers this question with a number — **probability**.

The lower the p-value than a significance level –alpha , the more surprising the evidence is, the more ridiculous our null hypothesis looks & we can reject the null hypothesis.

"How does the normal distribution apply to our previous hypothesis testing?"

Since we used Z-test to conduct our hypothesis testing, we need to calculate **Z-Scores** which is the number of standard deviations from the mean a data point is.

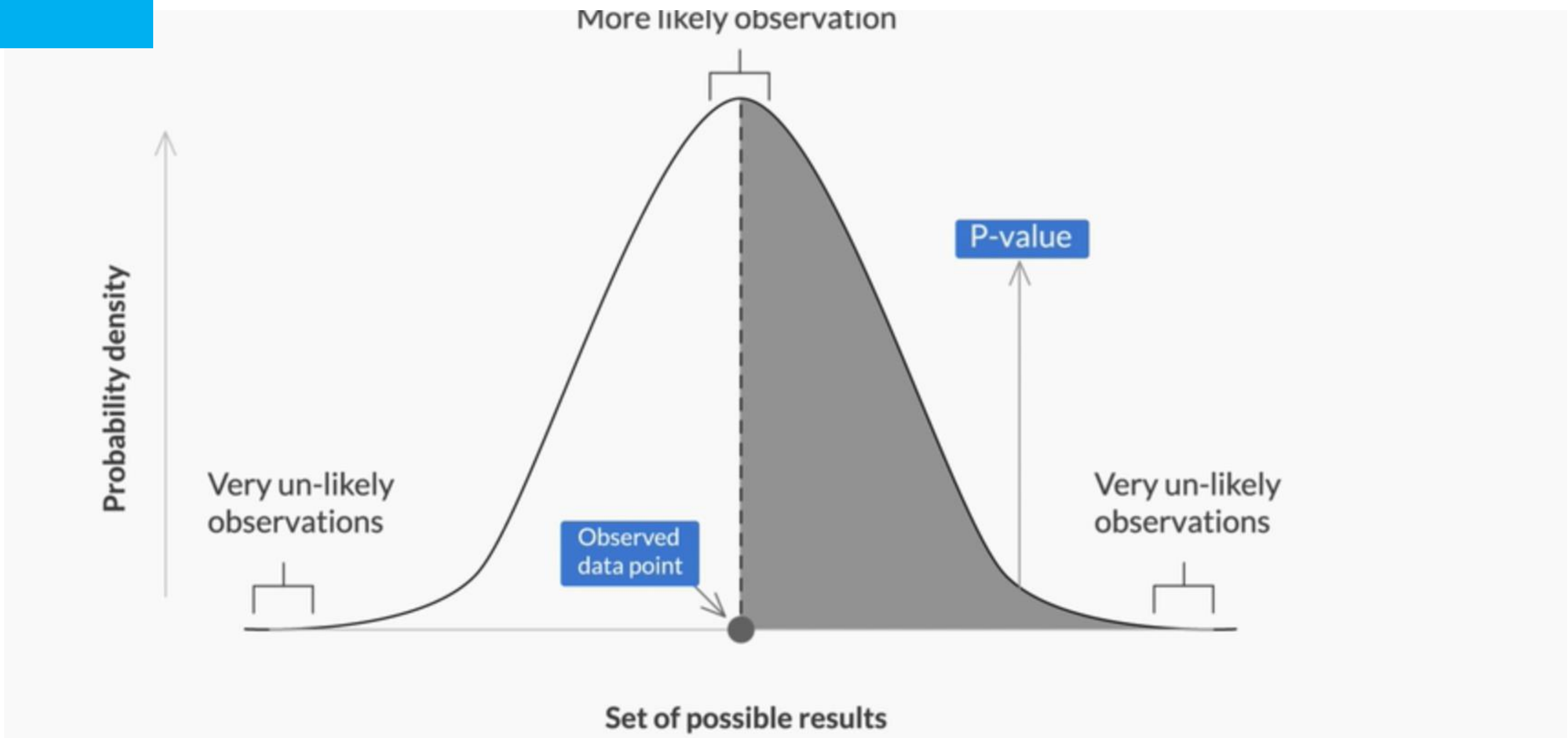
In this case, **each data point is the pizza delivery time that we collected.**

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

The higher or lower the Z-score, the more unlikely the result is to happen by chance and the more likely the result is meaningful.

Here we are talking about p-value!



Higher p-value implies the higher probability of Null Hypo to be True. Lower p-value denotes that chances are more for Null Hypo to be Rejected.

Exercise Problem : Government regulatory bodies have specified that the maximum permissible amount of lead in any food product is 2.5 parts per million or 2.5 ppm.

Let’s say you are an analyst working at the food regulatory body of India FSSAI. Suppose you take 100 random samples of Sunshine from the market and have them tested for the amount of lead. The mean lead content turns out to be 2.6 ppm with a standard deviation of 0.6.

One thing you can notice here is that the standard deviation of the sample is given as 0.6, instead of the population’s standard deviation. In such a case, you can approximate the population’s standard deviation to the sample’s standard deviation, which is 0.6 in this case.

Find out if a regulatory alarm should be raised against Sunshine or not, at 3% significance level.

H_0 - Average lead content ≤ 2.5 ppm

H_1 - Average lead content > 2.5 ppm

Significant Level = 3%

= (1-0.03)= 0.97

Cumulative Probability of critical point = 0.97

Zc = 1.88

Critical Value = 2.61 ppm = μ + Zc (Ϭ / sqrt(n)

2.5 + 1.88(0.6 /√100) = 2.61

Sample mean ($\bar{x}$) = 2.6 ppm

Decision = Fail to reject the null hypothesis

Example : The painkiller drug needs to have a time of effect of at most 200 seconds to be considered as having done a satisfactory job.

Given the same sample data (size, mean, and standard deviation), test the claim that the newer batch produces a satisfactory result and passes the quality assurance test.

Utilize 2 hypothesis testing methods to make your decision. Take the significance level at 5 %. Clearly specify the hypotheses, the calculated test statistics, and the final decision that should be made for each method.

For the effectiveness test, a sample of 100 drugs was taken. The mean time of effect was 207 seconds, with the standard deviation coming to 65 seconds. Using this information, you are required to estimate the range in which the population mean might lie — with a 95% confidence level.

Find the required range.

Given: Sample size (n): 100 drugs, Sample Mean = 207 sec

Sample Standard Deviation (σ Xbar) = 65 Sec & Confidence Level = 95% = 0.95

Thus significance level (alpha)= 0.05 (0.025 both the sides of distribution) & (1-0.025)= 0.975 is the cumulative probability corresponds to which the z score or z critical using the Z table is found as : Zc= 1.96

Critical Values : From Formula: Xbar +- Zc (Ϭ / sqrt(n))

Upper Critical value= UCV= 207 + 1.96(65/10) = 219.74

Lower Critical value= LCV= 207- 1.96(65/10) = 194.26

Range= (194.26 - 219.74)

To solve this problem, we need to calculate the margin of error for the given sample mean. Defining the Null (H0) & Alternate(H1) hypothesis

H0 : Mean μ <= 200
H1: Mean μ > 200

Since the Alternate hypothesis have '>' sign. The Critical region lies on the right side of the distribution i.e it will give Upper Critical Value (UCV) after which the critical region (rejection region) would be depicted in the distribution curve.

Given: Sample size (n)=100 drugs, Sample Mean X Bar = 207 sec Sample Standard Deviation (σXBar) = 65 Sec Significance Level = α= 5% = 0.05 We have 2 hypothesis test methods: Critical Value Method and p- value Method.

1.**Critical Value Method**: This hypothesis test would follow a one tailed test, with the critical region lying on the right of the Z critical value.

2. **p-value method** - p-value defines the probability of a Null Hypothesis to be True. Higher the value of p (compare to alpha) higher will be chances for H0 to be true. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis

Z-test

Sol. 3 a) (i) critical-value method:

Given $\mu = 200$, $\sigma_{\bar{x}} = 65$, $n = 100$

$\alpha = 0.05$ (significance level)

Confidence level = 0.95

Z-value corresponding to 0.95 from Z-table is $\frac{1.64 + 1.65}{2} = 1.645$

i.e $Z_c = 1.645$ (Z-score)

critical value can be calculated by

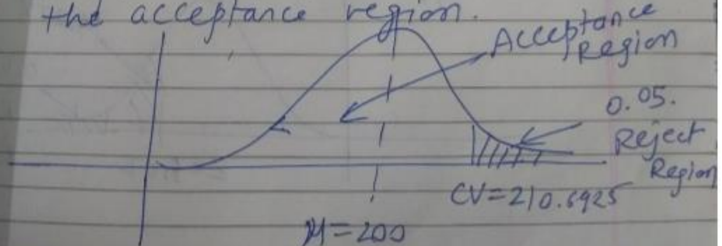
$$CV = \mu + Z_c \frac{\sigma_{\bar{x}}}{\sqrt{n}}$$

$$CV = 200 + 1.645 \times \frac{65}{\sqrt{100}}$$

$$CV = 200 + 10.6925$$

$$CV = 210.6925$$

Sample mean $\bar{x} = 207$ lies inside the acceptance region.



Thus we fail to reject the Null Hypothesis H_0.

a) p-value method:

we know that

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}} / \sqrt{n}}$$

~~As per~~ the given values substituting

$$z = \frac{207 - 200}{65 / \sqrt{100}} = 7/6.5 = 1.0769$$

given $\alpha = 0.05$.

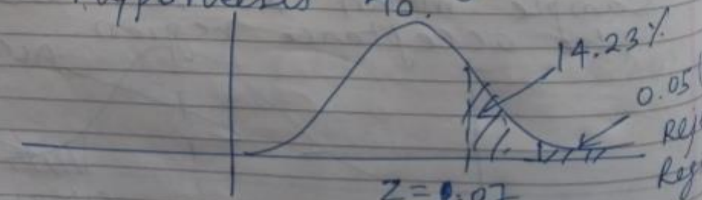
from Z-table for $z = 1.0769$

$$p = (1 - 0.8577)$$

$$p = 0.1423 \quad (14.23\ \%)$$

As p-value (14.23 %) > 5 %

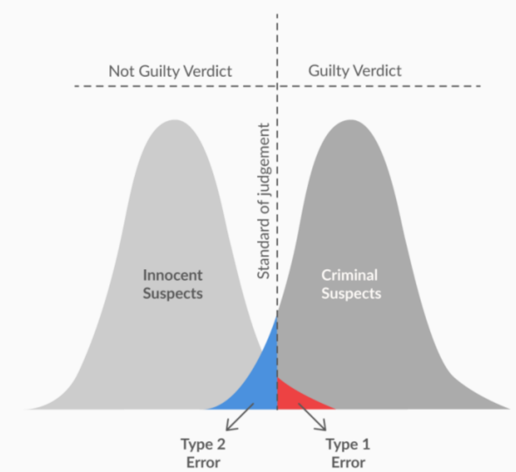
Thus we fail to reject Null Hypothesis H_0.



We fail to reject Null Hypothesis as p-value is 14.23% which is greater than 5% falling in Acceptance Region.

Types of Errors

	The null hypothesis is true	The null hypothesis is false
We decide to reject the null hypothesis	Type I error (rejecting a true null hypothesis) α	Correct decision
We fail to reject the null hypothesis	Correct decision	Type II error (failing to reject a false null hypothesis) β



Hypothesis Testing

Null Hypothesis H_0
True until proven false
Usually posits no relationship

Alternative Hypothesis
Negation of null hypothesis
Usually asserts specific relationship

Select Test
Pick from vast library
Know which one to choose

Test Statistic
Convert to p-value
How likely it was just luck?

Significance Level
Usually 1% or 5%
What threshold for luck?

Accept or Reject
Small p-value? Reject H_0
Small: Below significance level

Happy Learning

Thank You !

References: Multiple E-Sources/Books