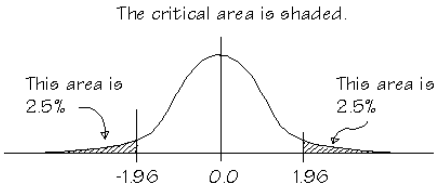


Z - Test

Please note that any topics that are not covered in today's lecture will be covered in the next lecture.

Content

1. Recap of Hypothesis Testing Framework
2. Recap CLT
3. One sample Z - test
4. Critical Value
5. Confidence Intervals
6. Power of Test

Recap of Hypothesis Testing Framework

We start any Hypothesis-testing problem with 2 things:

- Assumption
- Data

There exists a framework to compute a quantifiable metric (pvalue) that will help us decide if we should accept or reject our null hypothesis.

Let's summarise it into steps:-

1. Setup Null and Alternate Hypothesis
2. Choose the **test statistic**.
3. Select the Left vs Right vs Two-Tailed test, as per the hypothesis
4. Compute the P-Value
5. Compare the P-Value to the Significance Level (α) and Fail to reject/reject the Null Hypothesis accordingly.
    - Another term closely related to Significance Level is **Confidence Level**
    - If we're using $\alpha = 0.05$, this means that 5% significance
    - This can also be said, 95% confidence

Recap Central Limit Theorem

Suppose we have a population with the following parameters:

- Average height: 65 inches (μ)
- Std dev height: 2.5 inches (σ)

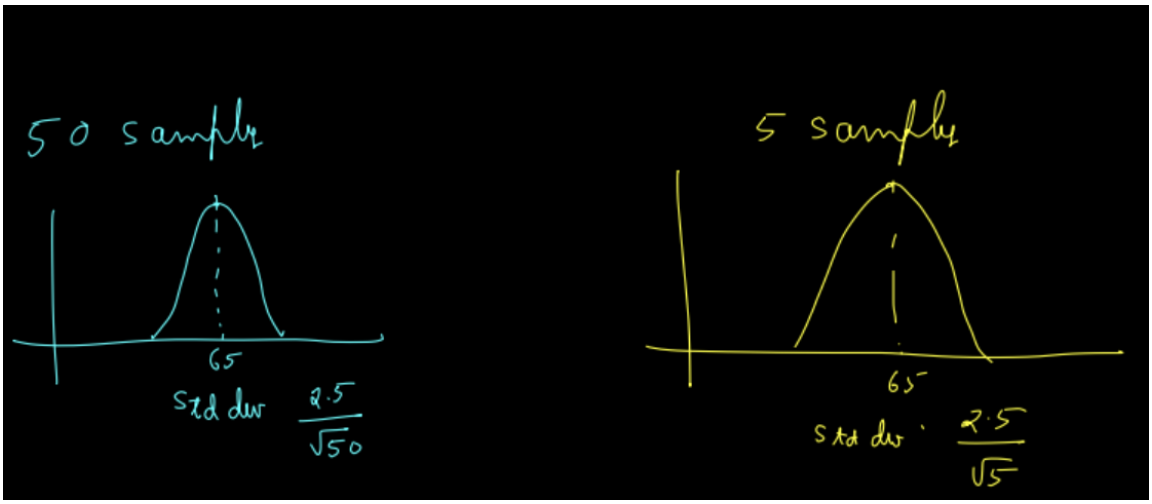
Suppose we take a sample of **randomly selected 50 people**. We can use CLT to find the following about this sample:

- Expected value of the mean height of this sample
  - It would be the same as the popluation mean: $E(m_{avg}) = \mu = 65$
- Expected value of the standard deviation of this sample
  - aka **Standard Error**
  - From CLT, we know that this can be found as: $\sigma_{sample} = \frac{\sigma_{population}}{\sqrt{n}}$
    - where, n: Size of sample
  - Therefore, $\sigma_{sample} = \frac{2.5}{\sqrt{50}}$

> **Quick comparison between cases having different sample sizes**

- Blue figure shows distribution with 50 samples
  - $\mu_{n=50} = 65$
  - $\sigma_{n=50} = \frac{2.5}{\sqrt{50}}$
- Yellow figure shows distribution with 5 samples
  - $\mu_{n=5} = 65$
  - $\sigma_{n=5} = \frac{2.5}{\sqrt{5}}$

Yellow distribution will be thicker, as it has a bigger standard deviation.



One sample Z - test

Marketing Case Study

Suppose there is a Retail Store Chain that sells Shampoo bottles

This chain has **2000 stores** across India.

The parameters for weekly sales of the shampoo bottle were reported as:

- Mean: 1800
- Standard deviation: 100

This was calculated by analyzing a lot of historical data

As a Manager / Owner / Data Scientist, you want to increase these sales, to generate more revenue.

> **Q1. What are the techniques at your disposal?**
>
> - Hire a marketing team

But there is an important factor to consider. These marketing teams/firms are not cheap, and would add a significant cost.

It stands to reason that you would not straightaway hand over all 2000 stores to them.

You would want an assurance that their work actually does impact the sales, and generate enough revenue that it is feasible to hire them.

> **Q2. How would you get that assurance?**

Perhaps you can allot them a few stores, and analyze the sale parameters (Mean and Standard deviation).

If results are good in a couple of weeks, then hire for all 2000 stores.

You decide to do this experiment with 2 competing marketing firms

- **Firm A**
  - Worked on **50 stores**
  - Sold an **average 1850** bottles of shampoo
- **Firm B**
  - Worked on **5 stores**
  - Sold an **average 1900** bottles of shampoo

> **Q1. Which firm gave better results?**

Clearly the sales are more for Firm B, but it seems that the number of stores under them were siginficantly less than Firm A.

It is possible that this increase by Firm B is just a chance factor because the standard deviation of the population was 100.

> **Q2. How do we quantify this and determine if it is just by chance or if it is actually statistically significant?**

When we talk about statistical significance, the word significance level pops into mind.

Since this is a big decision that would affect revenue, you want to be very very sure (99% confidence) about your decision, i.e. $\alpha = 0.01$

So, we need to employ the framework we saw and conduct hypothesis testing to see which firm's results are more significant.

Our historical data parameters for weekly sales of the shampoo bottle were reported as:

- Mean: 1800
- Standard deviation: 100

Deep diving into Firm A's result

- We will follow the steps of the Framework that we discussed earlier.

STEP 1:

> **Q1. What should be the null and alternate hypothesis?**

- H_0: Marketing firms have no effect on sales, i.e. $\mu = 1800$
- H_a: Marketing firms have positive effect on sales, i.e. $\mu > 1800$

STEP 2:

Since **Firm A** worked on **50** stores, it's average can be calculated as:

$m = \frac{x_1 + x_2 + ... + x_{50}}{50}$

> **Q2. What is the distribution of m?**

- Gaussian.

From CLT, We know that the expected value of m, under the assumption of the Null hypothesis is:

- $E(m_{avg}) = \mu_m = 1800$
- $\sigma_m = \frac{100}{\sqrt{50}}$

Let's plot this distribution of m, under the null hypothesis

We will use the z-score as the test statistic.

STEP 3:

> **Q3. Is the marketing team looking for an effect towards the left side or right side or either?**

Since the marketing team wants to increase sales, it is looking for an effect towards the **right side**

STEP 4:

We have the details about the distribution of m, but

> **Q4. What is the observed value of m?**

- 1850

Since we wish to perform right right-tailed test, we will calculate the probability of the weekly sales being greater than 1850 under the assumption that the null hypothesis is true i.e. P(m > 1850 | H_0 is true)

> **Q5. What is this value known as?**

- P-value.

> **How to calculate this?**

- By using the z-score of 1850 in this distribution, and using `1 - norm.cdf(z)`
  - $z = \frac{1850-1800}{\frac{100}{\sqrt{50}}}$

This is also known as **Z Statistic** or **Test Statistic**.

In [ ]:
```
import numpy as np
from scipy.stats import norm
```

In [ ]:
```
z = (1850 - 1800) / (100 / np.sqrt(50))
pval = 1 - norm.cdf(z)

print("P-Value:",pval)
```

```
P-Value: 0.00020347600872250293
```

STEP 5:

We defined $\alpha = 0.01$ for confidence level 99%

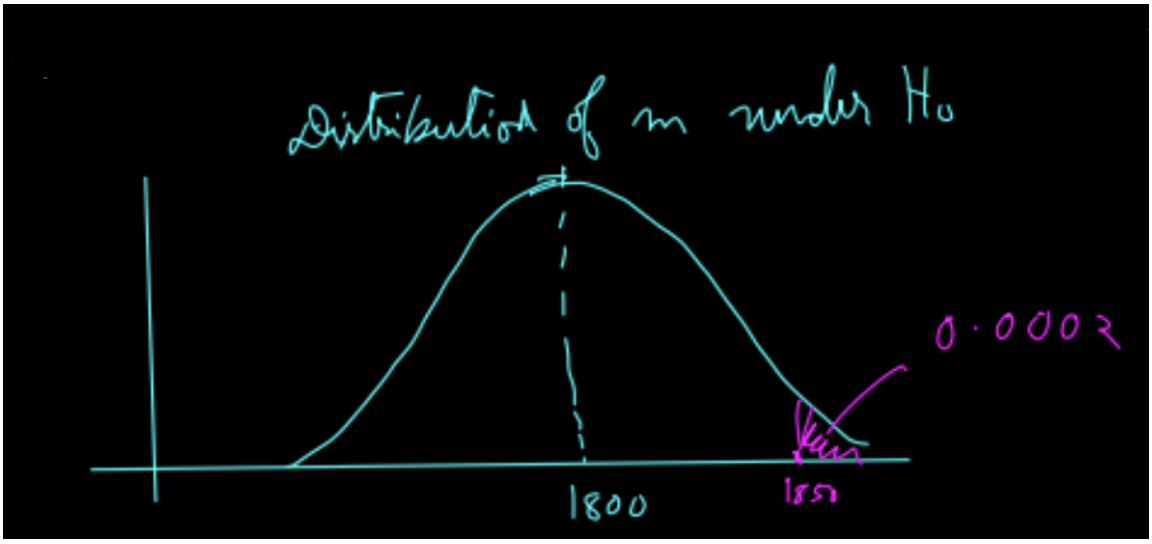
In [ ]:
```
alpha = 0.01
if pval < alpha:
  print("Reject the null hypothesis")
else:
  print("Fail to reject the null hypothesis")
```

```
Reject the null hypothesis
```

Since $0.00020347600872250293 < \alpha$

This means that we can **reject the null hypothesis**

So, we can report that the marketing team did have a positive effect.



Deep Dive into Firm B's result

STEP 1:

Same null and alternate hypothesis as before

- H_0: Marketing firms have no effect on sales, i.e. $\mu = 1800$
- H_a: Marketing firms have positive effect on sales, i.e. $\mu > 1800$

STEP 2:

Here, **Firm B** only had **5** stores.

Though it is a small number of stores, we will make an approximation that `m` here also follows **Gaussian Distribution**.

It's average can be calculated as: $m = \frac{x_1+x_2+x_3+x_4+x_5}{5}$

- Expected value of m, under assumption of hypothesis: $\mu_m = 1800$
- std dev: $\sigma_m = \frac{100}{\sqrt{5}}$

We will use the z-score as the test statistic.

STEP 3:

> **Is the marketing team looking for an effect towards the left side or right side or either?**

Since the marketing team wants to increase sales, it is looking for an effect towards the right side.

STEP 4:

We have the details about the distribution of m, but

> **What is the observed value of m?**

- 1900

Therefore p-value or **Z Statistic** or **Test Statistic** for this case:

- P(m > 1900 | H_0 is true)

In [ ]:
```
z = (1900 - 1800) / (100 / np.sqrt(5))

pval = 1 - norm.cdf(z)

print("P-Value:",pval)
```

```
P-Value: 0.0126736593387341
```

STEP 5:

We defined $\alpha = 0.01$ for confidence level 99%

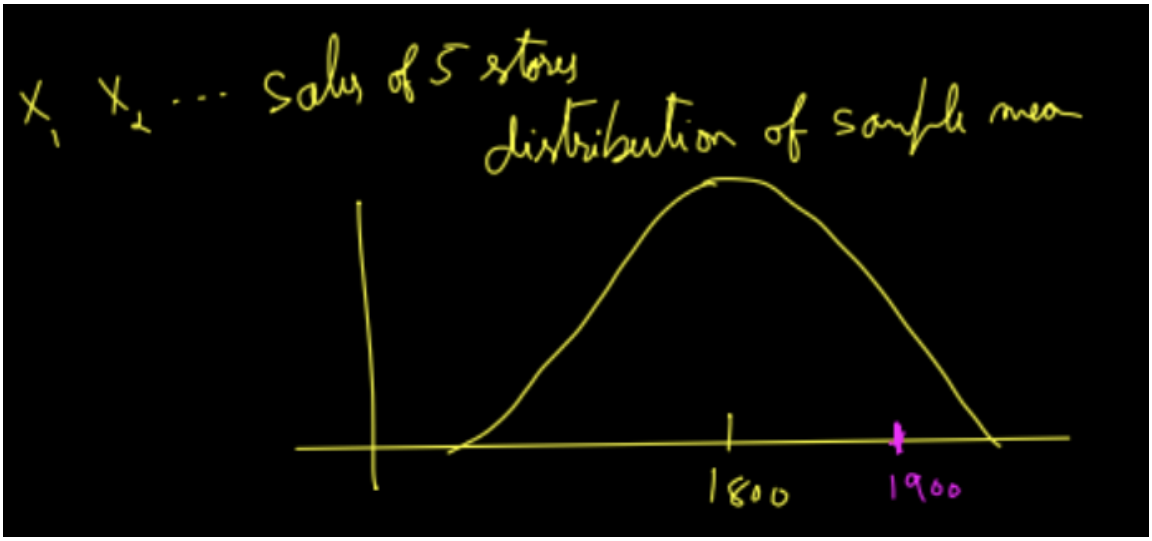
In [ ]:
```
alpha = 0.01
if pval < alpha:
  print("Reject the null hypothesis")
else:
  print("Fail to reject the null hypothesis")
```

```
Fail to reject the null hypothesis
```

Since $0.0126736593387341 > \alpha$

This means that we can **fail to reject the null hypothesis**

So, we can report that the marketing team did not have a positive effect (i.e. No Effect).



Conclusion

This way, you can now report that it is a good decision to hire Firm A, based on the given data.

As a layman also you can imagine

- It is tougher to maintain an average of 1850 across 50 stores
- Comparatively easier / luck to get 1900 for merely 5 stores.

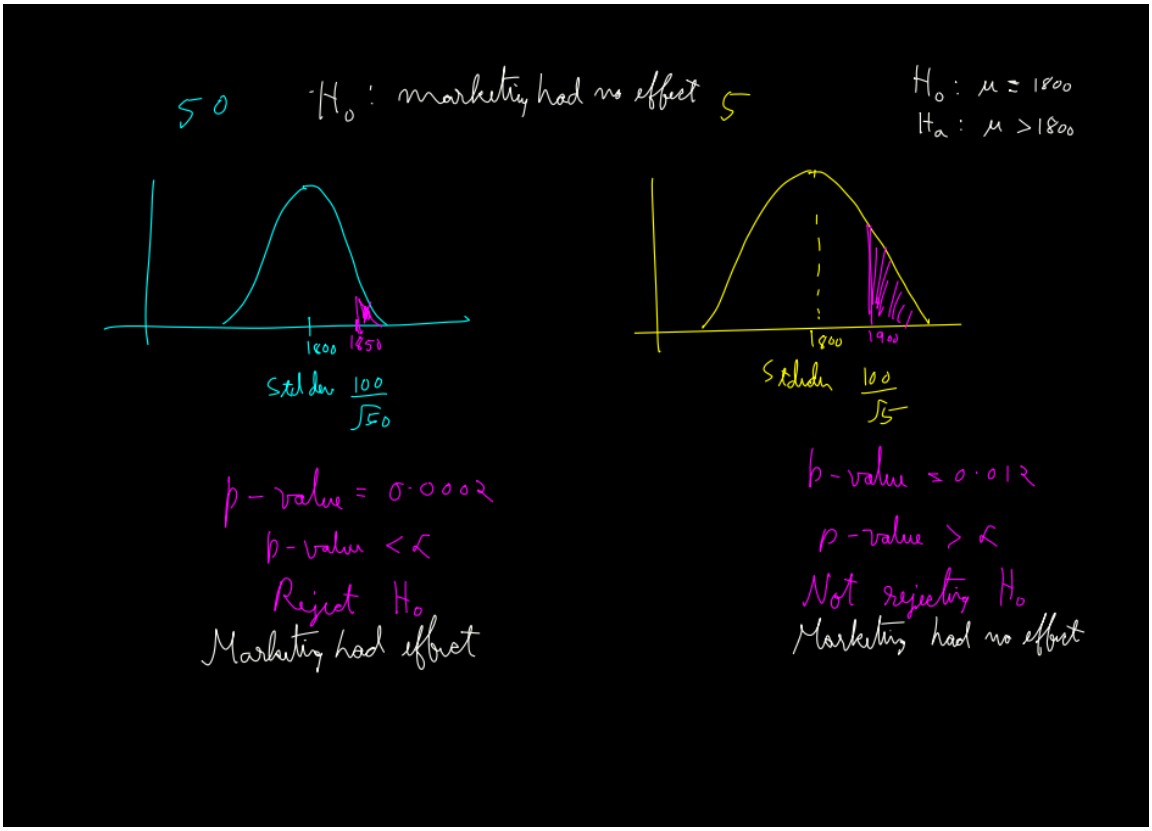
From a statistical standpoint, we saw the obvious difference in the p-value, but also, under the assumption of null hypothesis,

- The sample mean for both cases is 1800
- standard deviation for Firm A: $\sigma_A = \frac{100}{\sqrt{50}}$
- standard deviation for Firm B: $\sigma_B = \frac{100}{\sqrt{5}}$

This means $\sigma_B > \sigma_A$

Even when we plot their distributions, we can see that it'll be fatter/thicker for Firm B, thereby enclosing more area when performing right tailed test

Hence, a higher test statistic value (p-value)



> **Conceptual Clarity Note:**

- In the case of Firm B, we **fail to reject the null hypothesis**
- This does not necessarily mean that H_0 is TRUE, but rather, it means that there is not enough evidence to suggest it is FALSE.
- In fact, as a data scientist you may even suggest to upper management to experiment by handing over more stores to Firm B, and analyze that, to give your final opinion.

- On the other hand, we had enough data / evidence to conclude that Firm A did drive up the sales (reject the null hypothesis).

Let's understand this by re-considering the court judge example.

If we have the following evidence:

- Carries knife
- Blood on knife
- Fingerprints
- CCTV
- motive

If he is innocent, that implies that getting this much evidence is impossible (highly unlikely)

- The counter-positive of this statement is that if we observe this data, most likely he is guilty.
- So, only after seeing this much evidence did the judge finally reject the null hypothesis and call him guilty. Otherwise, without adequate data, he cannot.

Critical Value

What should be the minimum **weekly average sales** for the **50** stores under Firm A, to convince us that their marketing had a positive effect with a confidence level of 99% ?

We want to determine the point,

- after which (to the right) we say that marketing had an effect
- below which (to the left) we say that marketing did not have any effect.

We know the popluation paramters:

- $\mu = 1800$
- $\sigma = 100$

We know the sample parameters (weekly data for stores under Firm A)

- $\mu_m = 1800$
- $\sigma_m = \frac{100}{\sqrt{50}}$

Let's call the value of minimum weekly average sales needed to show that marketing had a effect be x

We know that we can calculate the z-score for this point as:

$$z = \frac{x - 1800}{\frac{100}{\sqrt{50}}}$$

> **If we know the z-score, can we calculate the p-value for that point ?**

We know that we need to perform Right tailed test, so pvalue can be given as:

`p-value = 1 - norm.cdf(z)`

Also, at this point x, we know that the p-value is same as α, i.e. 0.01, since we're looking for the minimum point

This means that `p-value = 1 - norm.cdf(z) = 0.01`

And, `z = norm.ppf(0.99)`

In [ ]:
```
from scipy.stats import norm

z = norm.ppf(0.99)
z
```

Out[ ]:
```
2.3263478740408408
```

Now we can evaluate x using the relation:

- $z = \frac{x - 1800}{\frac{100}{\sqrt{50}}}$

In [ ]:
```
import numpy as np

x = (z*(100/np.sqrt(50))) + 1800
x
```

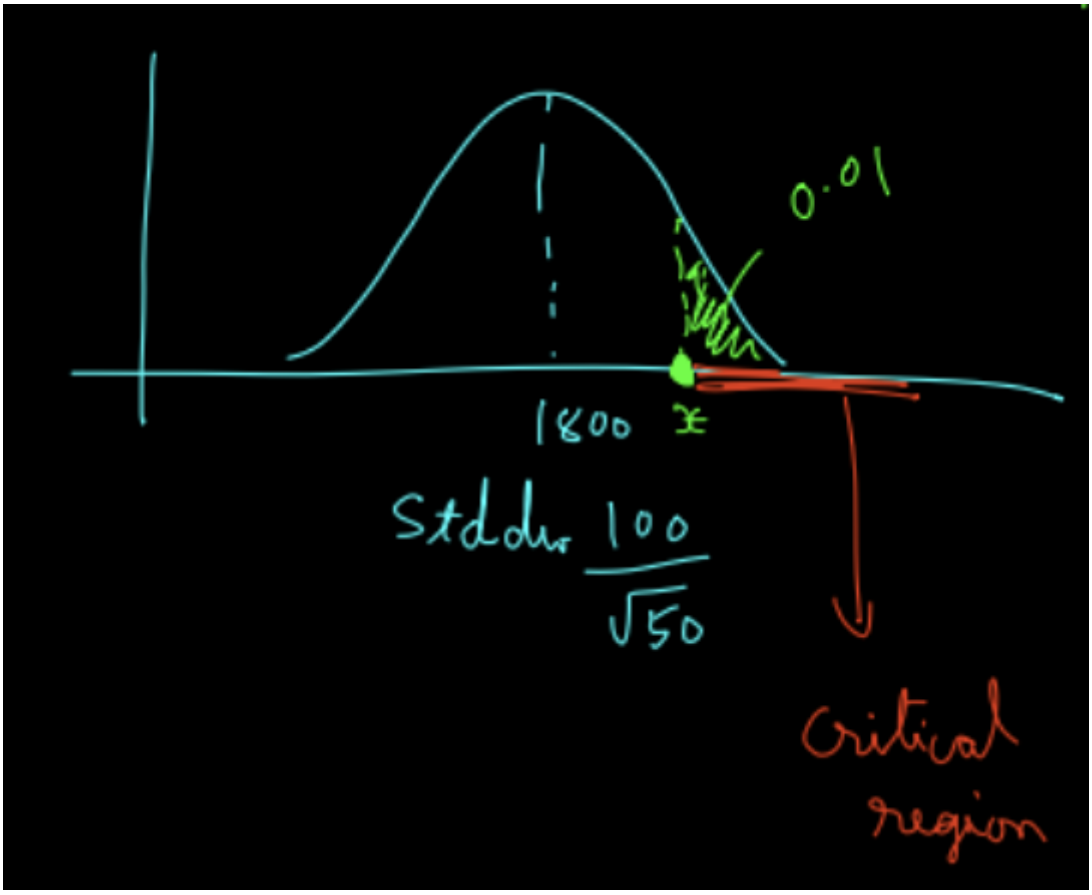
Out[ ]:
```
1832.8995271426638
```

This value is also known as the **Critical Value**, and for Firm A, it is equal to 1832.89

The region shown in red is known as **Critical Region**.

This means that **Firm A** must deliver a minimum of **1832**.

- 89 bottles of shampoo, as the weekly average for the 50 stores it looked after, in order for us to say that marketing drives the sale with 99% confidence.



Similarly,

Calculate the critical point for Firm B

For this sample, we have

- $\mu_m = 1800$
- $\sigma_m = \frac{100}{\sqrt{5}}$

Here also,

`p-value = 1 - norm.cdf(z) = 0.01`

Hence, `z = norm.ppf(0.99) = 2.32`

Now we can evaluate x using the relation

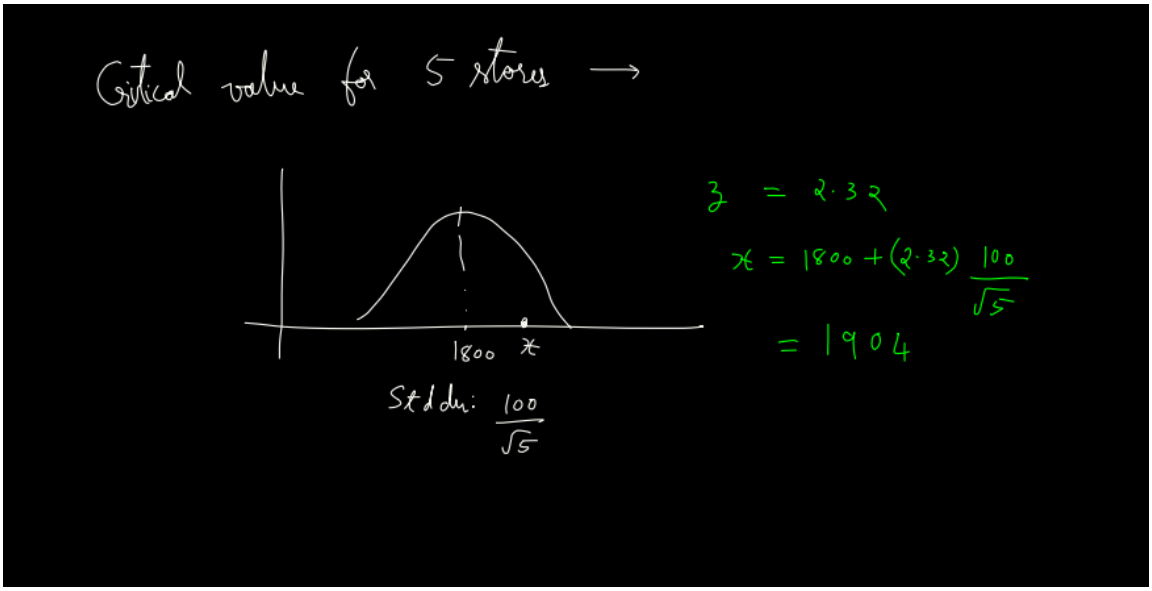
- $z = \frac{x - 1800}{\frac{100}{\sqrt{5}}}$

Note:

- The only thing that changed here is the value of std dev
- This is because the z-score will be the same for a point of 99% confidence.

In [ ]:
```
x = (z*(100/np.sqrt(5))) + 1800
x
```

Out[ ]: 1904.0374397133487



Confidence Intervals

Understanding Confidence Intervals:

- A confidence interval is a range of values that provides an estimate of a population parameter, such as the population mean, with a specified level of confidence.
- It is expressed as **(point estimate ± margin of error)**,
  - Where the margin of error depends on the chosen confidence level.

Example Scenario:

- Consider the same marketing example of Shampoo sales
- Firm A
  - Worked on **50** stores
  - Sold an average of **1850** bottles of shampoo

Setting the Stage:

- **Significance Level (α):** $\alpha = 0.01$, which represents the probability of making a Type I error is 0.01
- **Confidence Level (1 - α):** The complement of the significance level, i.e 0.99 or 99 %, represents our level of confidence.

Constructing a Confidence Interval:

- We calculate a critical value (Z) that corresponds to the chosen confidence level. In this case, Z for a 99% confidence level is approximately **2.576.**
- **Margin of Error (ME):** The margin of error is determined by:

    - The Z-score (critical value).
    - The population standard deviation (σ).
    - The sample size (n), usually in the form of the square root of n.
- **Confidence Interval (CI):** We construct the confidence interval as follows:
    - **CI = Sample Mean ± (Z * (σ / √n))**

- H_0: Marketing firms have no effect on sales, i.e. $\mu = 1800$
- H_a: Marketing firms have positive effect on sales, i.e. $\mu > 1800$

In [ ]:
```python
import numpy as np
from scipy.stats import norm

# Given data
population_mean = 1800
sample_mean = 1850
population_stddev = 100
sample_size = np.sqrt(50)
alpha = 0.01  # Significance level (1 - alpha will give us the confidence le

# Calculate the critical value (Z) for a right-tailed test at the given alph
z_critical = norm.ppf(1 - alpha)

# Calculate the margin of error
margin_of_error = z_critical * (population_stddev / sample_size)

# Calculate the confidence interval
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_o

print("Confidence Interval:", confidence_interval)


# Check if the population mean (1800) falls within the confidence interval

if confidence_interval[0] <= population_mean <= confidence_interval[1]:
    print("The population mean falls within the confidence interval. Then we
else:
    print("The population mean does not fall within the confidence interval.
```

```
Confidence Interval: (1817.1004728573362, 1882.8995271426638)
The population mean does not fall within the confidence interval. Then we r
eject the null hypothesis
```

- Similarly, Firm B
    - Worked on **5** stores
    - Sold an average of **1900** bottles of shampoo

In [ ]:
```python
import numpy as np
from scipy.stats import norm

# Given data
population_mean = 1800
sample_mean = 1900
population_stddev = 100
sample_size = np.sqrt(5)
alpha = 0.01  # Significance level (1 - alpha will give us the confidence le

# Calculate the critical value (Z) for a right-tailed test at the given alph
```

```
z_critical = norm.ppf(1 - alpha)

# Calculate the margin of error
margin_of_error = z_critical * (population_stddev / sample_size)

# Calculate the confidence interval
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_

print("Confidence Interval:", confidence_interval)
===

# Check if the population mean (1800) falls within the confidence interval

if confidence_interval[0] <= population_mean <= confidence_interval[1]:
    print("The population mean falls within the confidence interval. Then we
else:
    print("The population mean does not fall within the confidence interval.
```

```
Confidence Interval: (1795.9625602866513, 2004.0374397133487)
The population mean falls within the confidence interval. Then we fail to r
eject the null hypothesis
```

Power of Test

Let's recap from the previous lecture. Consider the judge example:

		The Person is	
		Actually Innocent	Actually Guilty
Judge Declares the accused as	Innocent	- True Negative (TN) - Probability of correctly not rejecting a true null hypothesis (H0) - Represented by (1 - α) - i.e. Level of Significance	- False Negative (FN) - Type-II Error - Probability of failing to reject a false null hypothesis. - Represented by β
	Guilty	- False Positive (FP) - Type-I Error - Probability of rejecting a true null hypothesis. - Represented by α	- True Positive (TP) - Probability of correctly rejecting a false null hypothesis (H0) - Represented by (1 - β) - i.e. Power of Test

There are 2 possible results to a Hypothesis Test:

- We reject H0
- We fail to reject H0

Indicating that the ultimate objective is to try and check if we can reject the H0. Hence, we regard

- Rejecting H0 (calling them guilty) as **Positive**, and
- Failing to reject H0 (calling them innocent) as **Negative**.

Based on this, recall that we end up with 4 cases:

- **False Positive:** It is not True that this person is Positive (guilty).
    - The person is actually innocent, but the judge says they are guilty.
    - This is **Type I error (α).**

- **False Negative:** It is not True that this person is Negative (innocent).
    - The person is actually guilty, but the judge says they are innocent. This is a miscarriage of justice.
    - This is the **Type II error (β)**.
- **True Negative:** It is True that the person is innocent.
    - The person is actually innocent, and the judge says they are innocent (1 - α).
    - This represents the cases where the person is truly innocent (H0) and the test correctly identifies them as innocent.
    - Since it captures the cases where the null hypothesis (H0) is correctly accepted.
    - It can be regarded as the **complement of the Type 1 error rate.**
    - Therefore, $TN = 1 - \alpha$
- **True Positive:** It is True that the person is guilty.
    - The person is actually guilty, and the judge says they are guilty.
    - Since, this captures the cases where the alternative hypothesis (Ha) is correctly accepted.
    - Hence, TP can be regarded as the **complement of the Type 2 error** rate
    - Therefore, $TP = 1 - \beta$

Type II Error

- This kind of error occurs when we fail to reject the null hypothesis even if it is wrong.
- It is equivalent to a `false negative` conclusion.
- The probability of type II error is given by the value of β.

When conducting hypothesis tests, we often focus on controlling Type I error (α), which is the probability of incorrectly rejecting a true null hypothesis.

- However, it is equally important to consider **Type II error (β)**, which is the probability of failing to reject a false null hypothesis.
- Now, we will explore the concept of statistical power, denoted as "Power" and its critical role in hypothesis testing.

What is Statistical Power?

- **Power** (often represented as 1 - β) is the probability of correctly rejecting a false null hypothesis.

Factors Influencing Power:

- **Sample Size (n):**

    - Increasing the sample size generally increases power.
    - Larger samples provide more information and make it easier to detect effects.

- **Significance Level (α):**
    - Reducing the significance level (e.g., from 0.05 to 0.01) increases the risk of Type II error, decreasing power.
- **Variability (σ):**
    - Reducing the variability in your data increases the power of your test.

- **Effect Size (d):**

  Let's explore this with an example scenario. Imagine you're following a specific diet plan:

  Diet Plan A:

- Individuals on Diet Plan A may experience weight changes, but the degree of weight loss or gain varies widely among participants.
- It's like having a group where some individuals may lose a small amount of weight, while others may experience a significant change.

  Now, let's relate this analogy to effect size and Cohen's d:

- Effect size is related to, **"Is there a noticeable difference in weight changes with Diet Plan A?"**
    - If there's a big difference in the average weight change, we say there's a large effect size.
    - If the difference is small, we say there's a small effect size.
    - The effect size is often denoted as "d". Cohen's d is a way to measure and quantify the size of the difference in weight changes with Diet Plan A.

  ### Cohen's d for One-Sample Test:
    - For a one-sample test comparing a sample mean to a known population mean, the effect size can be calculated using:

      d = (Sample Mean – Population Mean)/Sample Standard Deviation

    - If Cohen's d is a large number, it means the difference is substantial. If it's a small number, the difference is more modest.

Question:

```
Imagine you are a quality control manager at a chocolate
factory.
You're responsible for ensuring that the average weight of
chocolate bars produced in your factory meets a certain
standard.

The **standard weight** of a chocolate bar is **50 grams**,
```

```
and it rarely deviates, with a known population **standard
deviation** of **2 grams**.

To maintain the quality of your chocolate bars, you collect a
**sample** of **30 bars** every day and weigh them.

You want to know if your production process is still on track
and that the average weight of the chocolate bars is 50
grams.

You set the significance level (α) to 0.05, and you want to
calculate the power of your quality control test.

data = [55, 45, 52, 48, 55, 52, 52, 53, 48, 52, 53, 47, 54,
51, 52, 51, 48, 52, 53, 54, 51, 51, 52, 54, 47, 52, 53, 48,
51, 54]
```

> Defining Hypothesis
>
> - H_0: Null Hypothesis $\mu = 50$ grams
> - H_a: Alternative Hypothesis $\mu <> 50$ grams

In the context of a quality control test for chocolate bars:

False Positive (Type I Error): This occurs if the test incorrectly concludes that the average weight of chocolate bars is different from the standard (50 grams) when, in reality, it is not.

- In other words, it falsely indicates a problem with the production process.

False Negative (Type II Error): This occurs if the test fails to detect a significant difference when the average weight of chocolate bars is indeed different from the standard.

- It means the test misses an actual issue with the production process.

We calculate the power of a test to essentially assess its ability to detect a true effect or difference when it exists. **A higher power indicates a lower probability of making a Type II error.**

- This is especially important in quality control or any testing situation because it ensures that the test can reliably spot problems in the production process, making the quality control measures more trustworthy.

Reference for power : Documentation

In [ ]:
```python
import numpy as np
from scipy import stats
from statsmodels.stats import power

# Given data
alpha = 0.05  # Significance level (for a two-tailed test)
confidence_level = 1 - (alpha / 2)  # 95% confidence level
sample_size = 30  # Number of chocolate bars in the sample
```

```
# Calculate the z-critical value for a 5% significance level (as you did pre
z_critical = np.abs(round(stats.norm.ppf(1 - alpha/2), 4))

# Calculate the sample mean (average weight of the chocolate bars)
data = [55, 45, 52, 48, 55, 52, 52, 53, 48, 52, 53, 47, 54, 51, 52, 51, 48,
samp_mean = np.mean(data)
samp_std = np.std(data)

# Null hypothesis value (standard weight)
hypo_mean = 50

# Calculate the effect size (difference between sample mean and hypothesized
effect_size = (samp_mean - hypo_mean) / samp_std
print("Effect size:",effect_size)

# Use 'zt_ind_solve_power()' to calculate the power of the z-test
# ratio=0 it implies that the function assumes equal sample sizes in both gr
# In other words, it assumes that the number of observations in the two grou
power = power.zt_ind_solve_power(effect_size=effect_size,
                                 nobs1=sample_size,
                                 alpha=alpha,
                                 ratio=0,
                                 alternative='two-sided')

print('Power of the test:', power)
```

```
Effect size: 0.5261336417646574
Power of the test: 0.8216812302268112
```

- **Effect Size:** The effect size, around 0.53, tells us how much the average weight of the sampled chocolate bars differs from the standard weight. In this case, it suggests a noticeable difference.
- **Power of the Test:** With a power of about 83%, there's a good chance that our quality control test will correctly spot any significant difference in the average weight. Essentially, it indicates how well our test can catch deviations from the standard weight, making our quality control process quite effective in maintaining chocolate bar quality.
- **Type 2 Error:** The 17% Type 2 error means that there's a chance (17 out of 100 times) our quality control might miss a real issue with the chocolate bar weights. So, even though there could be a difference in the average weight, our test might not catch it every time.

The goal is to minimize both Type 1 and Type 2 errors to ensure effective quality monitoring.

Interactive Tool

Tool Link: https://rpsychologist.com/d3/nhst/

Solve for? Power Alpha n d

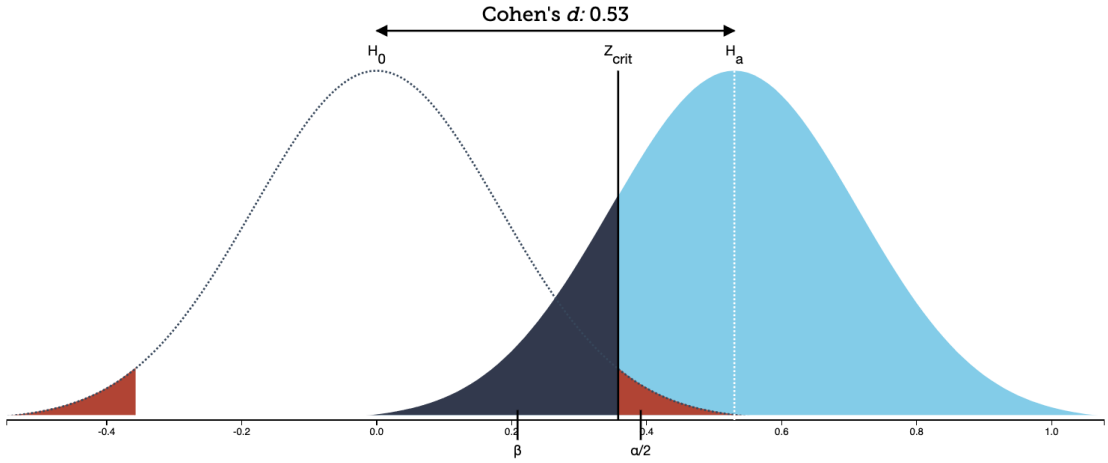
Significance level (α = 0.05)

Sample size (n = 30)

Effect size (d = 0.53)

One-tailed Two-tailed

Reset zoom



5 %

Type I error

17 %

Type II error

83 %

Power

30

Sample size