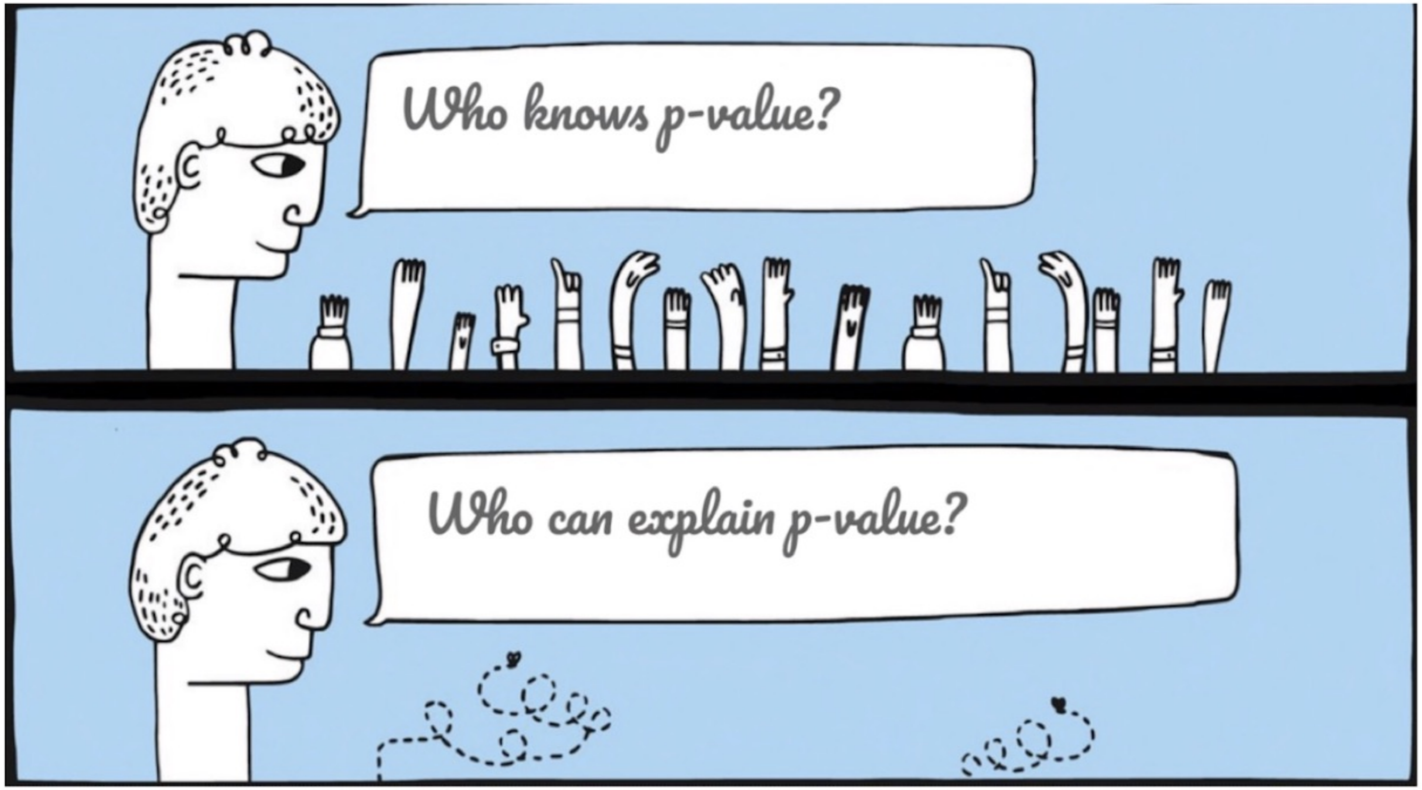
17th March 2023



Let's start @ 9:05 pm

Agenda

- Quick recap of Imp. terminologies
- Hypothesis Testing Framework
- Hypothesis Testing various approach
- Examples of Left, Right, Two Tailed Tests
- Hypothesis Testing in Python

  (Coding with real data)

Recap of CLT

- Avg. height = 65 inches, σ = 2.5 inches

① Sample 50 people, $n=50$

↓

m → samplemean

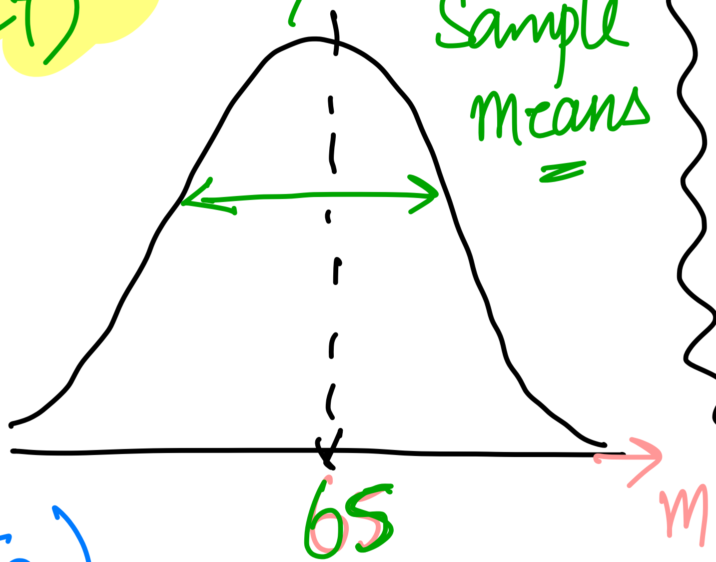
↓

random variable

↓

follow normal distbⁿ (from CLT)

distbⁿ of Sample means



- $E(m) = 65$
- std. dev. $= \dfrac{\sigma}{\sqrt{50}}$

$= (2.5/\sqrt{50})$

② Sample 5 people

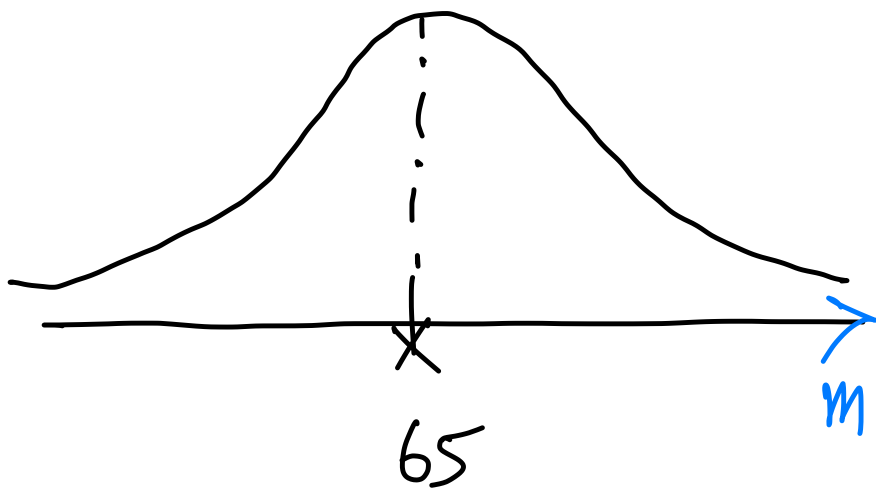
↓

m → Sample mean.

↓

$\sim N$

- $E(m) = 65$
- std. dev. $= (2.5/\sqrt{5})$



Supply Chain Example:

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

A retailer has 2000 stores in the country
Historical data tells us that weekly sales of shampoo bottles has an average of 1800, with a standard deviation of 100

- Sales team → improve sales.

Team A

on 50 stores

weekly sales = 1850

One week

Team B.

on 5 stores

1900

One week

⇒ 99% Confident → hire ✓

$1 - 0.99 = \alpha$.

$\boxed{\alpha = 0.01}$

Significance level

Team A

$\mu = 1800$, $\sigma = 100$, $\alpha = 0.01$

$n = 50$ stores. → $x_1, x_2, \dots, x_{50}$ → $\left(\frac{x_1 + \dots + x_{50}}{50}\right)$

$= 1850$

$\left(\text{Z score} = \frac{x - \mu}{\text{std dev.}}\right)$

$H_0 : \mu = 1800$

$H_A : \mu > 1800$

$m = 1850$ → test statistic.

Random Variable ~ Normal dbn.

- $E(m) = 1800$
- std. dev. of $m = 100/\sqrt{50}$
- Right Tailed
- P-value $= P[m \geq 1850 \mid H_0 \text{ is True}]$

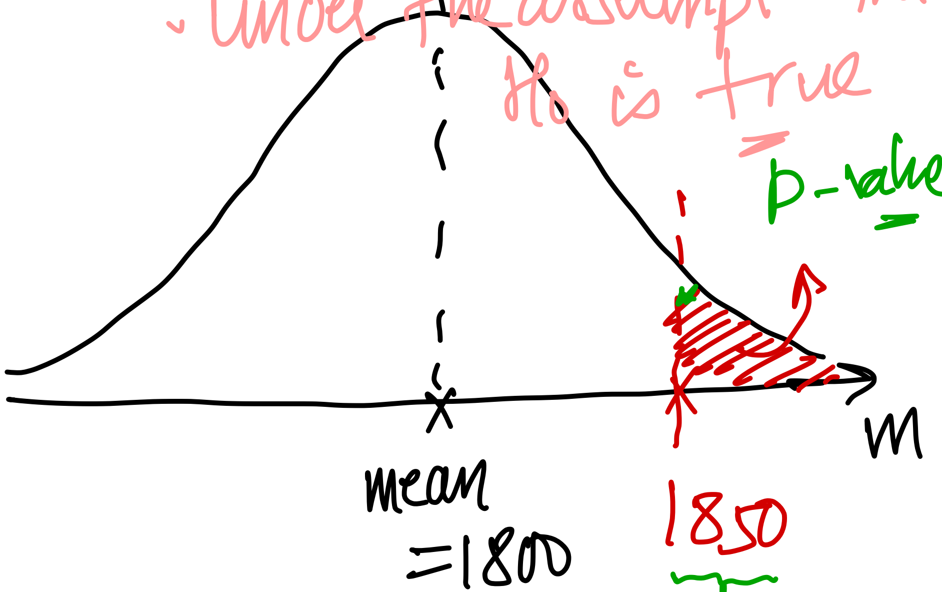
$= 1 - \text{norm.cdf}(3.53)$

$= 0.0002$

- $\alpha = 0.01$

pvalue < α → $\boxed{\text{Reject } H_0}$

under the assumpt^n that H_0 is true



$\text{Z score} = \left(\frac{1850 - 1800}{100/\sqrt{50}}\right)$

$\boxed{\text{Z score} = 3.53}$

* Team A has true impact on Sales.

Team B

$\mu = 1800$, $\sigma = 100$, $\alpha = 0.01$

$\downarrow$

$n = 5$ stores. $\rightarrow$ $x_1, x_2, ----, x_5$ $\rightarrow$ $\left(\frac{x_1 + --- + x_5}{5}\right)$

$= 1900$

$H_0: \mu = 1800$

$H_A: \mu > 1800$

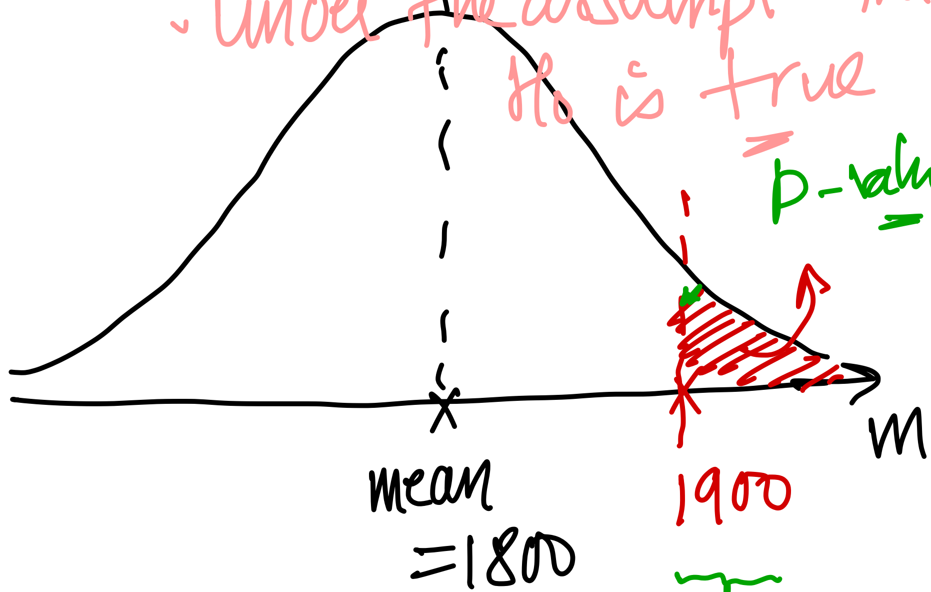
$m = 1900 \rightarrow$ test statistic.

$\downarrow$ Random Variable $\sim$ Normal db^n.

$\left(Z\text{score} = \frac{x - \mu}{\text{std dev.}}\right)$

$\frac{\sigma}{\sqrt{n}}$

· under the assumpt^n that H_0 is true



· $E(m) = 1800$

· std. dev. of $m = 100/\sqrt{5}$

· Right Tailed

· P-value $= P[m \geq 1900 \mid H_0 \text{ is True}]$

$= 1 - \text{norm.cdf}(2.23)$

$= 0.0126$

$\downarrow$

$Z\text{score} = \left(\frac{1900 - 1800}{100/\sqrt{5}}\right)$

$\boxed{Z\text{score} = 2.23}$

· $\alpha = 0.01$

pvalue $> \alpha$ $\rightarrow$ $\boxed{\text{Failed to Reject } H_0}$

* Team B impact on marketing is not SS

Supply chain example

$\alpha = 0.01$

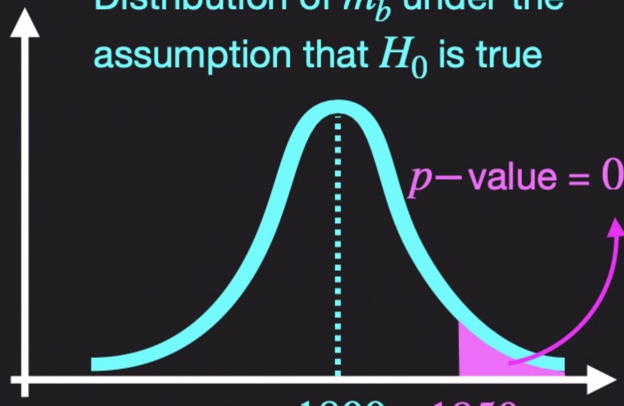
$\mu = 1800$
$\sigma = 100$

50 stores with average of 1850

$H_0 : \mu_b = 1800$

$H_a : \mu_b > 1800$

Distribution of m_b under the assumption that H_0 is true



std dev = $\dfrac{100}{\sqrt{50}}$

$$z = \frac{1850 - 1800}{100/\sqrt{50}} = 3.53$$

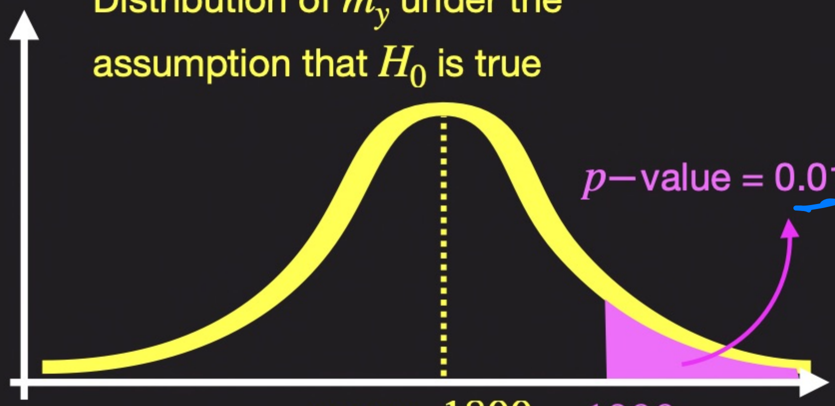
Reject H_0

5 stores with average of 1900

$H_0 : \mu_y = 1800$

$H_a : \mu_y > 1800$

Distribution of m_y under the assumption that H_0 is true



std dev = $\dfrac{100}{\sqrt{5}}$

$$z = \frac{1900 - 1800}{100/\sqrt{5}} = 2.23$$

Fail to reject H_0

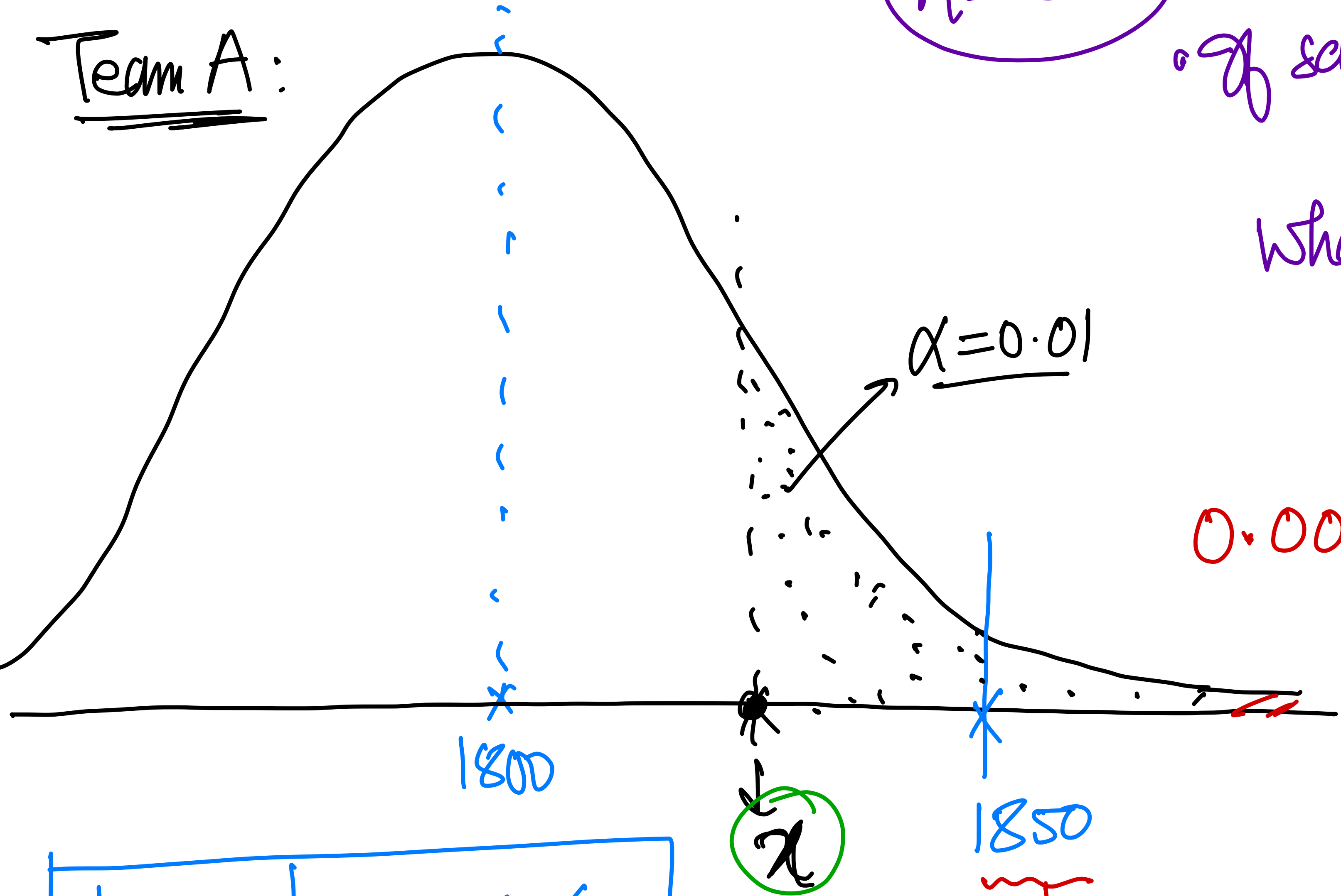
pvalue 0.0128 little higher than 0.01

α

CRITICAL VALUE Appr: $\alpha = 0.01$

$n = 50$

Team A:



% of sales higher than 1850

what happens to p value

decreases.

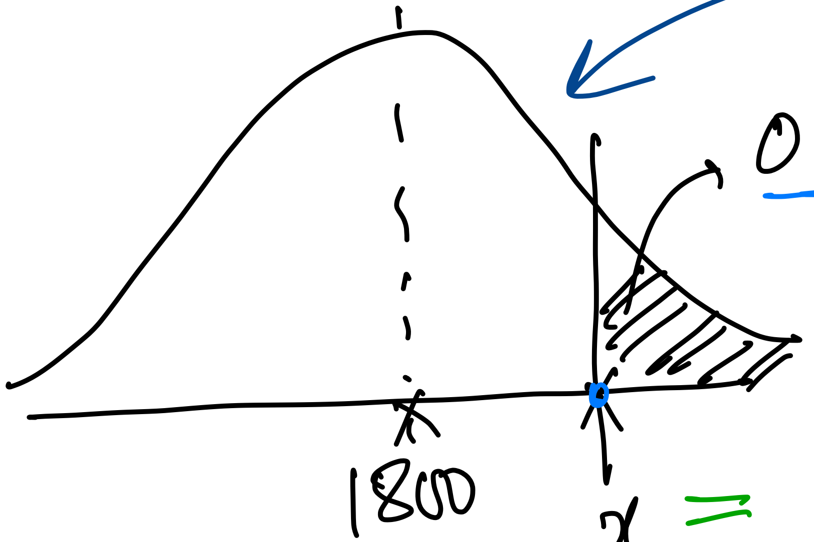
0.0002 → p value

x → Critical value at which pvalue = α.

z scor = 3.53.

$$\boxed{p\text{-value} < \alpha}$$

max p-value is $\alpha = 0.01$



$0.01 \to \alpha$.

$\sigma = \frac{150}{\sqrt{50}}$

$x = 1832.8 \simeq 1833$ *

Critical value

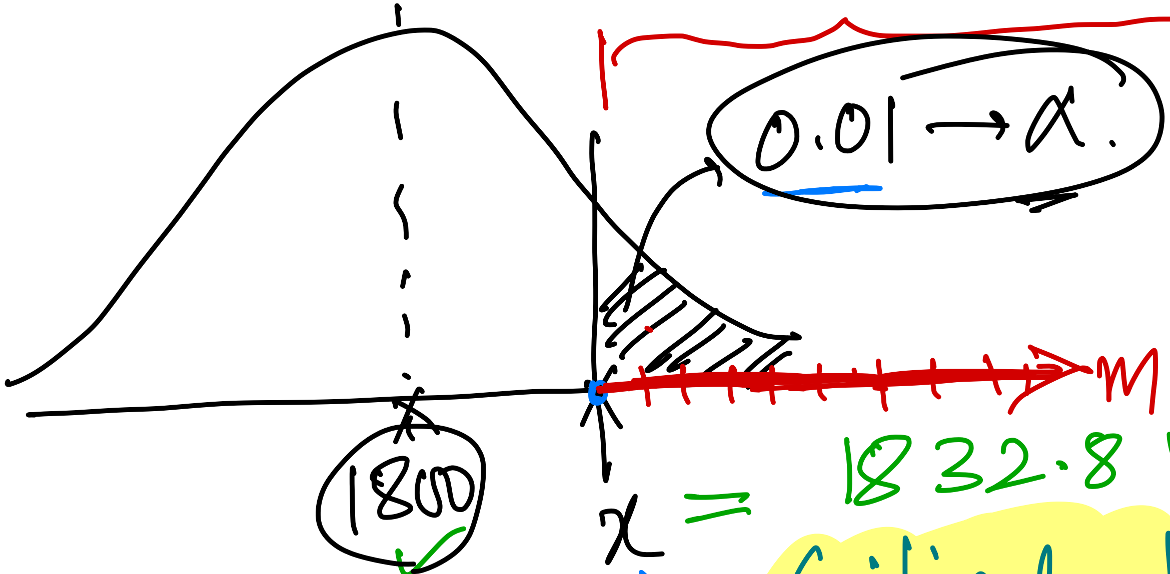
Zscore $\to$ norm.ppf(0.99) $\to$ 2.32

table pdf

Zscore of $x = 2.32$

$$Z = \left(\frac{x-\mu}{\sigma}\right) \Rightarrow 2.32 = \frac{x-1800}{(100/\sqrt{50})} \Rightarrow x = 1832.8$$

• Company has budget to measure sales impact on 50 stores → n → $\frac{\sigma}{\sqrt{n}}$, μ → $\frac{100}{\sqrt{50}}$



Critical Region

x = 1832.8 ≈ 1833

Critical value

• If sale > x (critical value) 99% confidence

* reject H_0 → i.e +ve impact on sales.

Supply chain example

$\alpha = 0.01$

Team B *

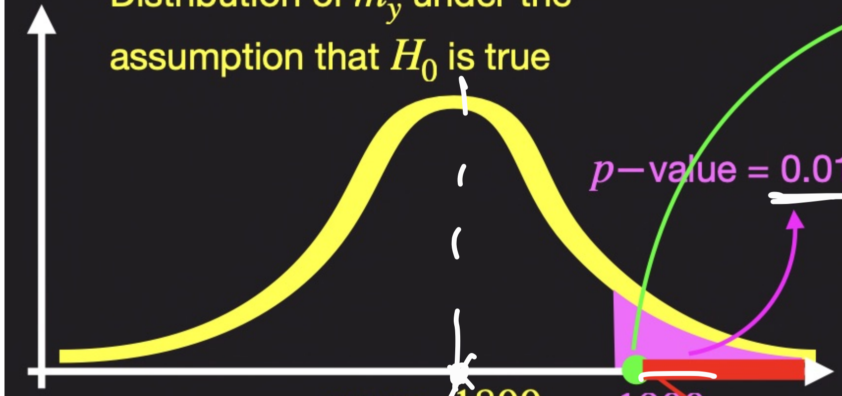
$\mu = 1800$
$\sigma = 100$

5 stores with average of 1900

$H_0 : \mu_y = 1800$

$H_a : \mu_y > 1800$

Distribution of m_y under the assumption that H_0 is true



std dev = $\dfrac{100}{\sqrt{5}}$

$$z = \frac{1900 - 1800}{100/\sqrt{5}} = 2.23$$

What should be the z-score such that we can reject if mean is larger, and accept if mean is lesser?

We want only 1% area to the right

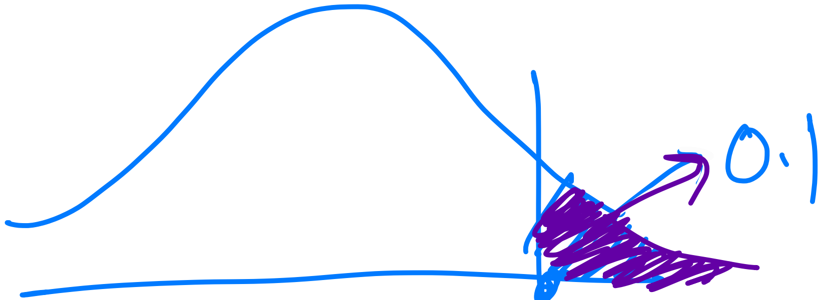
Upper critical value = `norm.ppf(0.99)` $= 2.32$

If z = 2.32, then $x = 1800 + 2.32 * \dfrac{100}{\sqrt{5}} = 1903.7$

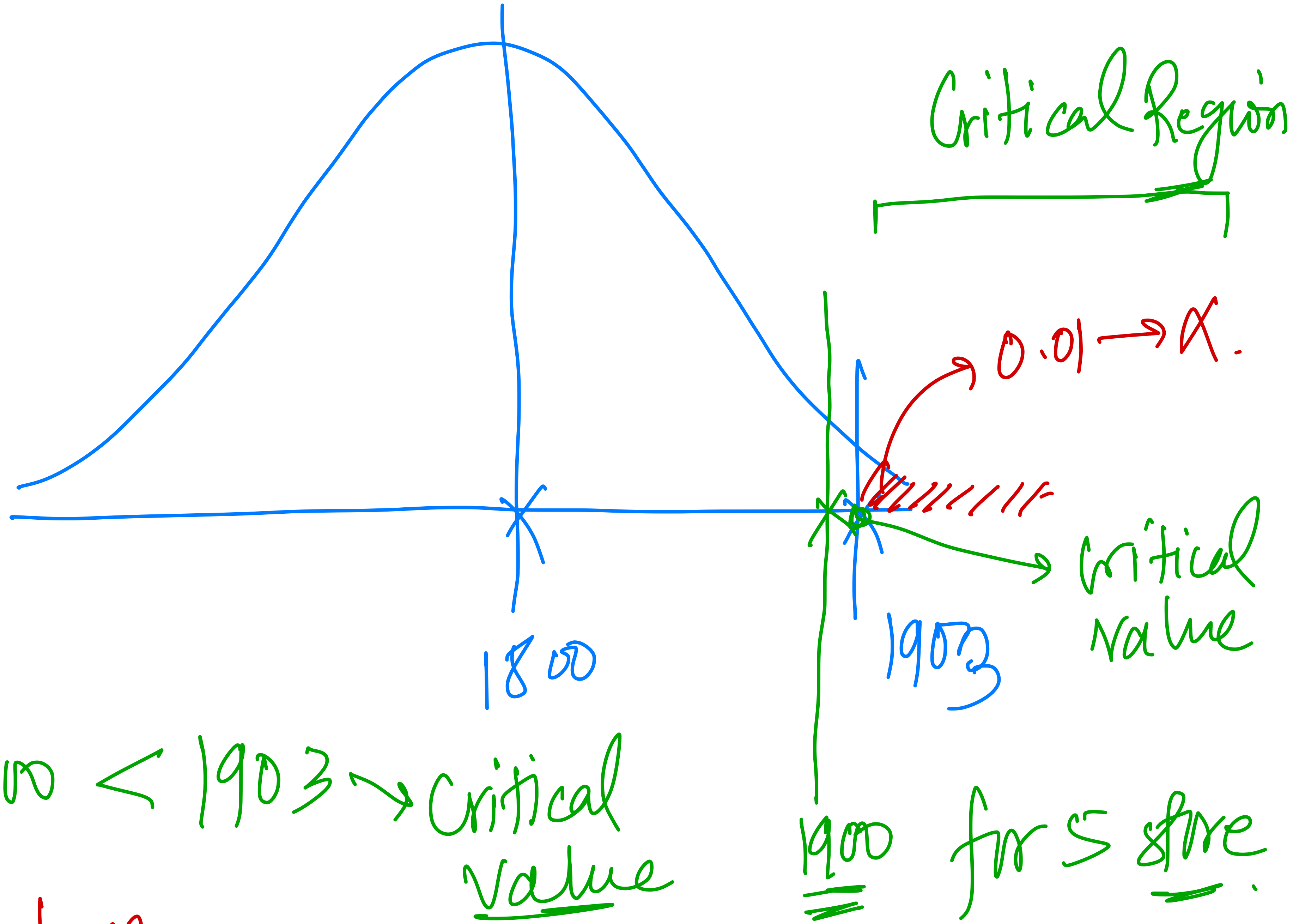
To summarise, if we are testing for 5 samples, we can reject the null hypothesis only if the average sales is greater than 1903.7

This region is called the "critical region"

$$2.23 = \frac{x - 1800}{(100/\sqrt{5})}$$



zcroe of x → remains same ?



1900 < 1903 → Critical value

We fail to Reject H_0 ✓.

Premature Children

Average IQ of all people is 100, with a standard deviation of 15
Medical researches want to know if prematurelyborn children have similar IQ or not
They sampled 50 such children and did an IQ test → Sample mean

$\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{50}}$

In what range should the sample mean be to say they have normal IQ with 95% confidence?

Sol^n:

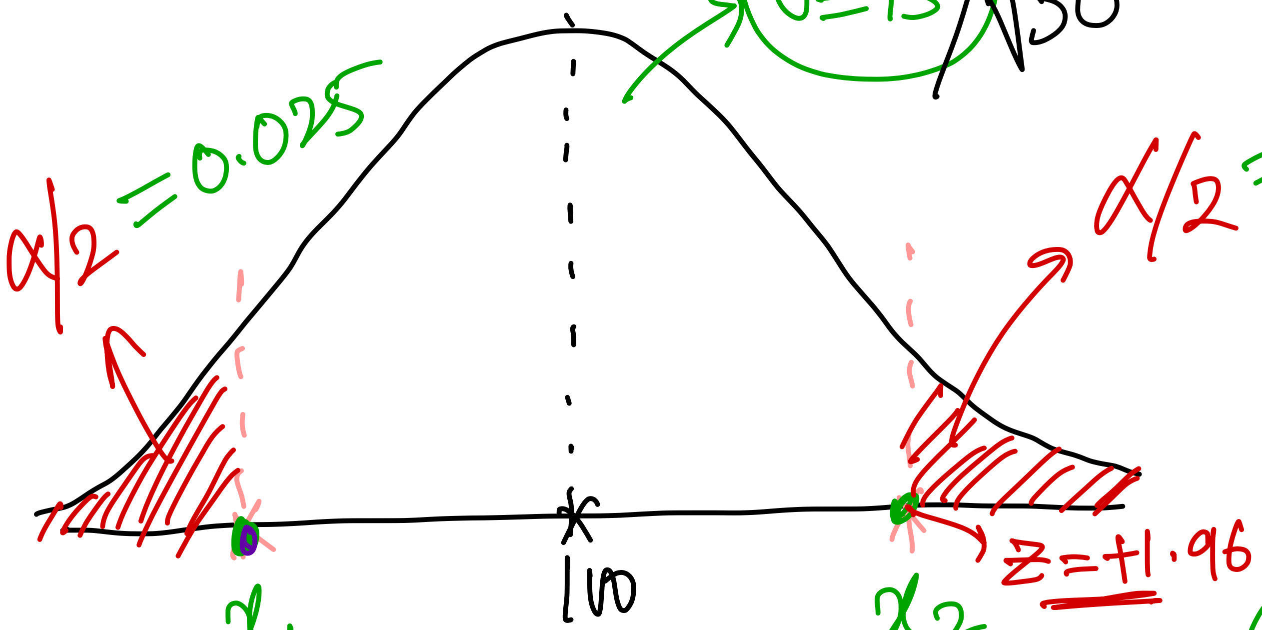
$H_0 : \mu = 100$

$H_A : \mu \neq 100$

Confidence → 95%

$\alpha = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025$



$(z)_{x_1} = \text{norm.ppf}(0.025)$
$= -1.96$

x_1 is -1.96 std. dev^n away frm mean $(=100)$

$x_1 \rightarrow (z)_{x_1}$ $\quad (z)_{x_2}$

$[x_1, x_2]$
$= [95.84, 104.15]$

$x_1 = 100 - 1.96 * 15/\sqrt{50}$