

## Recap



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

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 wants to improve sales by hiring a marketing team

Hiring a marketing team can be expensive, so we need to be very sure that they will improve sales

Before deploying their strategy for all 2000 stores, they are tested in 50 stores

On the 50 stores, their average sales for that week was 1850

You are the data scientist who should tell your sales team whether this is statistically significant

Sales team has said that we will hire only if we are 99 % confident  $\alpha = 0.01$

Another marketing team is also being considered

They are tested on 5 stores

On the 5 stores, their average sales for that week was 1900

Would you say this team is better than the first one?

Between the “blue team” and the “yellow team”, whom will you choose?

Recap

$\alpha = 0.01$

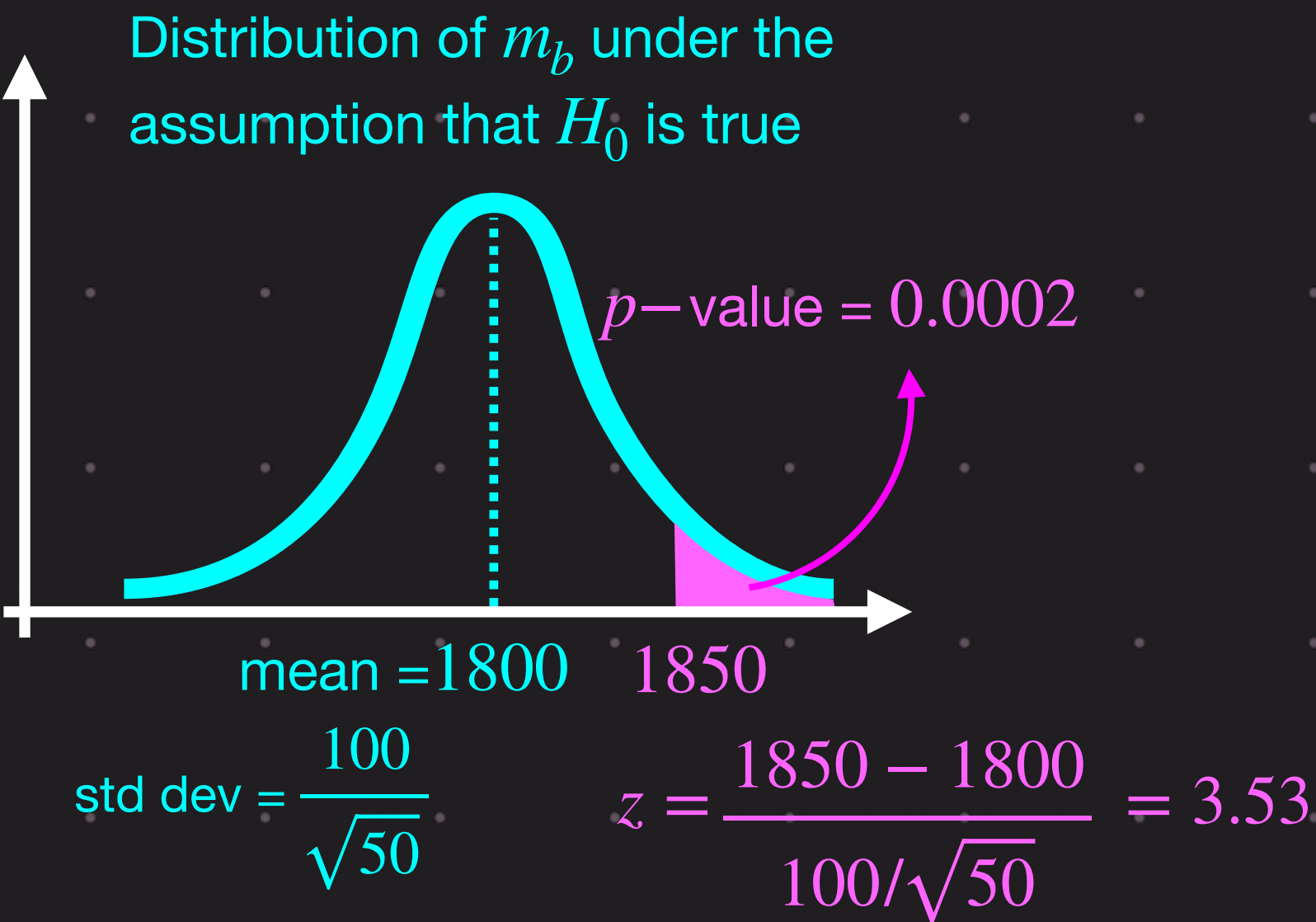


$\mu = 1800$   
 $\sigma = 100$

50 stores with average of 1850

$H_0 : \mu_b = 1800$

$H_a : \mu_b > 1800$

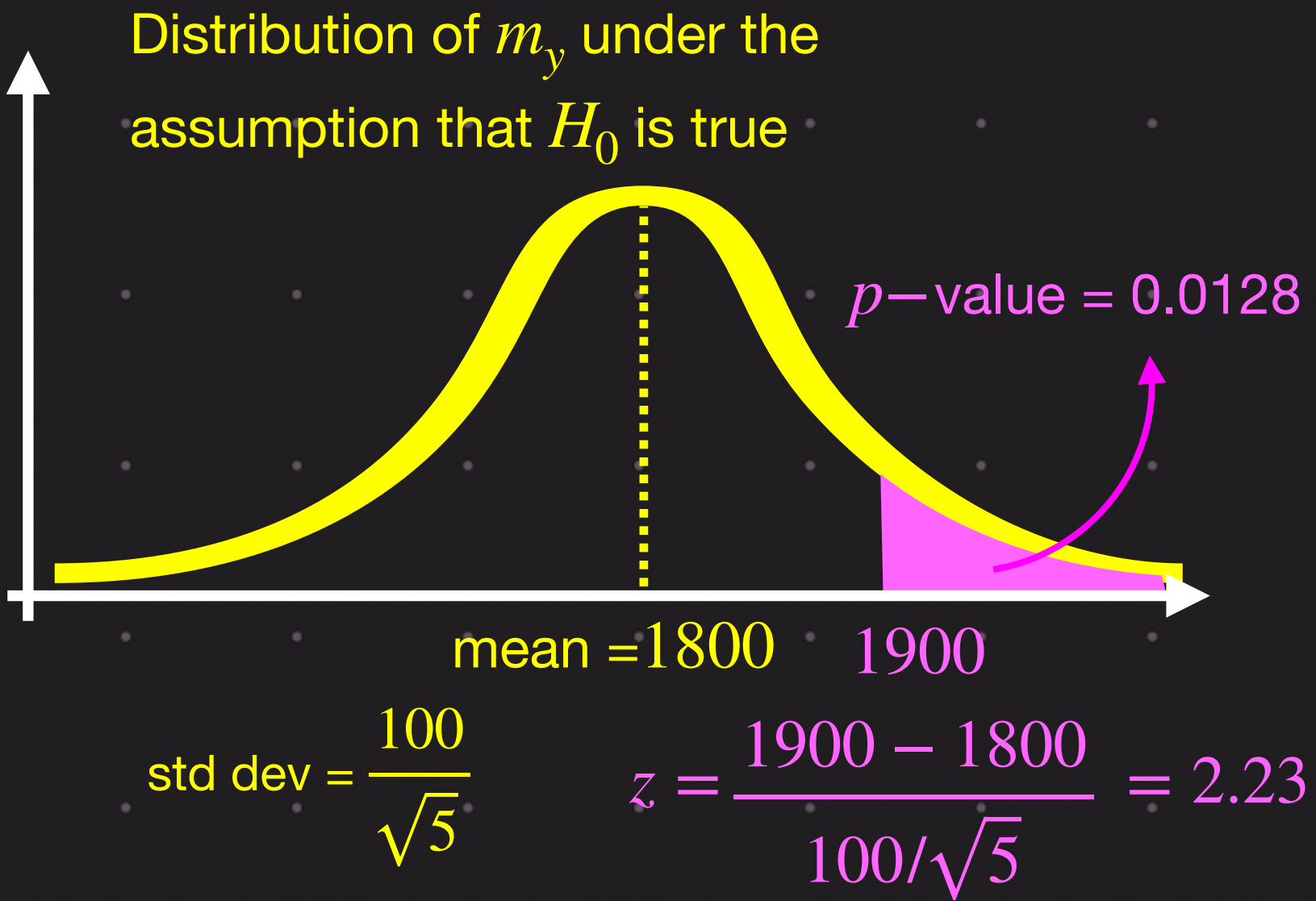


Reject  $H_0$

5 stores with average of 1900

$H_0 : \mu_y = 1800$

$H_a : \mu_y > 1800$



Fail to reject  $H_0$



# Drug Recovery

Suppose two companies develop a drug for a disease.

Drug 1 was tested on 100 people, and the recovery days look like this

[8, 5, 9, 10, ..., 16]

The mean recovery time was 7.1 days

Drug 2 was tested on 120 people, and the recovery days look like this

[12, 4, 7, 13, ..., 8]

The mean recovery time was 8.07 days

Can we say one drug was better than the other?

Or was this small difference a coincidence?

For such cases we use the two-sample t-test

```
from scipy.stats import ttest_ind
```

# IQ across two schools

Suppose there are two schools competing against each other

Each school says their students have higher IQ

So we conduct a test

Say the first school had numbers like this

[101, 115, 90, ..., 112, 97]

Let us say average was 103.7

And yellow team has these numbers

[108, 105, 99, ..., 111, 98]

Let us say average was 102.9

Is there a statistical significance in this difference?

Or was it just chance?

For such cases we use the two-sample t-test

```
from scipy.stats import ttest_ind
```



## One sample Test

Supply chain:

We compared average sales in 50 stores against a fixed number: 1800

$$H_0 : \mu_b = 1800$$

$$H_a : \mu_b > 1800$$

Premature children IQ

We compared average IQ of 50 prematurely born children fixed number: 1800

$$H_0 : \mu = 100$$

$$H_a : \mu \neq 100$$

Drug Vs No drug

Patients recover on average 15 days without drug. Will it reduce after taking the drug?

$$H_0 : \mu = 15$$

$$H_a : \mu < 15$$

**We are comparing one set of samples against a fixed number (population mean)**

## Two sample test

Supply chain:

We directly compare two marketing teams

$$H_0 : \mu_b = \mu_y$$

$$H_a : \mu_b \neq \mu_y$$

Two states IQ

Two competing states, where each state says their students have higher IQ

Let  $\mu_1$  be the average IQ of first state, and  $\mu_2$  be the average IQ of another state

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

Two Drugs

Let  $\mu_1$  be the average recovery for drug 1, and  $\mu_2$  be the average for drug 2

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

**We are comparing two sets of samples against each other**