

Data Analytics

Mrinal Das

Example Scenario



Two options: choose between them

How will you do ?

Data Driven Decision Making

One can take a decision by checking the specifications and using expertise

However, it is not possible in every case

If the machine is a new product, or the task is new

Often encountered in technology replacements such as automation

Data driven approach is safe and scientific

Process of Data Driven Decision Making

Try and observe: run both machines for some time and note down observations

1. Collect data
2. Create tables and plots/charts
3. Descriptive analytics – summarize the observations
4. Diagnostic analytics – analyse the causes
5. Predictive analytics – forecast the pros and cons
6. Prescriptive analytics – recommend

Pilot Trial

Run both machines for 100 times each

Machine-1 completes the task within time 70 times

Machine-2 completes the task within time 65 times

$P(\text{machine-1 is successful}) = 0.7$

$P(\text{machine-2 is successful}) = 0.65$

$P(\text{machine-1 is successful}) > P(\text{machine-2 is successful})$

Recommend machine-1

Uncertainty Factor

1. Collect data
2. Create tables and plots/charts
3. Descriptive analytics – summarize the observations
4. Diagnostic analytics – analyse the causes
5. Predictive analytics – forecast the pros and cons
6. Prescriptive analytics – recommend

Suppose the machine starts to behave differently after procurement

Whatever we observed during trial was by chance

Repeat Pilot Trial

Run both machines for 100 times each

Machine-1 completes the task within time 63 times

Machine-2 completes the task within time 65 times

$P(\text{machine-1 is successful}) = 0.63$

$P(\text{machine-2 is successful}) = 0.65$

$P(\text{machine-1 is successful}) < P(\text{machine-2 is successful})$

Recommend machine-2

Conclusion is different from Trial 1

Uncertainty Factor

Clearly one trial is not enough

Machines are not stable

How can we make a decision ?

Data analytics has failed ?

How many trials do we have to do ?

Multiple Trials

Let's repeat the trials 4 times and note the observations

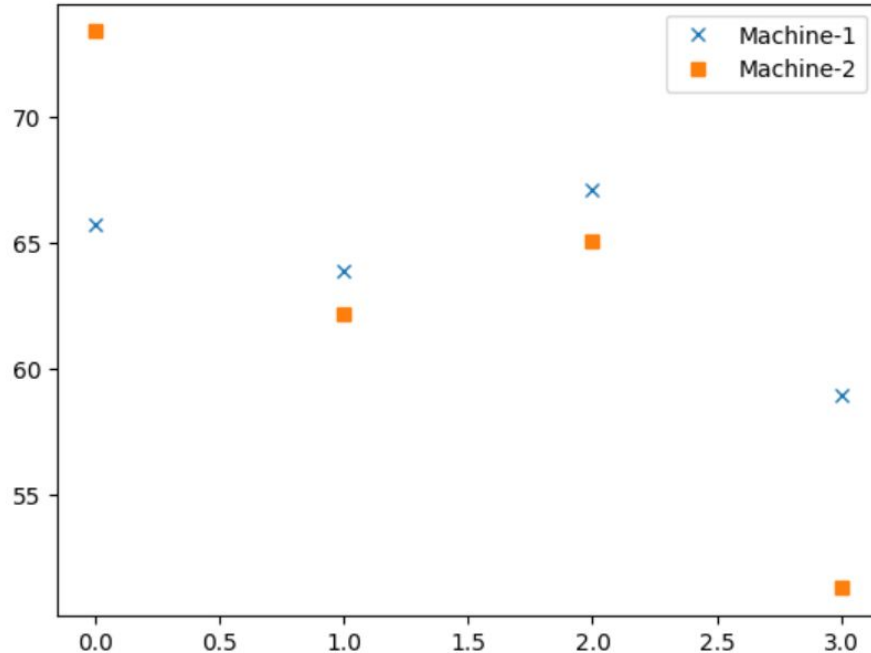
Machine-1 success {70, 63, 72, 68} – X

Machine-2 success {65, 67, 63, 70} – Y

Average success rate of machine-1: 68.25

Average success rate of machine-2: 66.25

Descriptive Analytics



Can we conclude that machine-1 is better ?

Visualizing Spread



How many Trials

Let's repeat the trials 4 times and note the observations

Machine-1 success {70, 63, 72, 68} – X

Machine-2 success {65, 67, 63, 70} – Y

Average success rate of machine-1: 68.25

Average success rate of machine-2: 66.25

Can we conclude that machine-1 is better ?

We may not because we do not know if in the 5th trial values will be 63 and 75

Concern of Uncertainty

Average success rate of machine-1: 68.25

Average success rate of machine-2: 66.25

Can we conclude that machine-1 is better ?

We may not because we do not know if in the 5th trail values will be 63 and 75

Have we been able to nullify the uncertainty effect yet ?

Concern of Uncertainty

Average success rate of machine-1: 68.25

Average success rate of machine-2: 66.25

Can we conclude that machine-1 is better ?

Have we been able to nullify the uncertainty effect yet ?

Let $\underline{X}$ is average success rate of machine-1

Let $\underline{Y}$ is average success rate of machine-2

We need to check $\underline{X} > \underline{Y}$ hold for most of the cases

Concern of Uncertainty

Let $\underline{X}$ is average success rate of machine-1

Let $\underline{Y}$ is average success rate of machine-2

We need to check $\underline{X} > \underline{Y}$ hold for most of the cases

$P(\underline{X} > \underline{Y})$ is high or $\underline{X} > \underline{Y}$ is statistically significant

Concern of Uncertainty

Let $\underline{X}$ is average success rate of machine-1

Let $\underline{Y}$ is average success rate of machine-2

We need to check $\underline{X} > \underline{Y}$ hold for most of the cases

$P(\underline{X} > \underline{Y})$ is high or $\underline{X} > \underline{Y}$ is statistically significant

Statistical significance test

Statistical hypothesis testing

Hypothesis

There are two hypotheses

Null hypothesis: based on intuition or experience

Alternative hypothesis: alternative to null hypothesis

Null hypothesis: machine-1 is better

Alternative hypothesis: machine-2 is better

We have to check which hypothesis is correct.

Hypothesis Testing

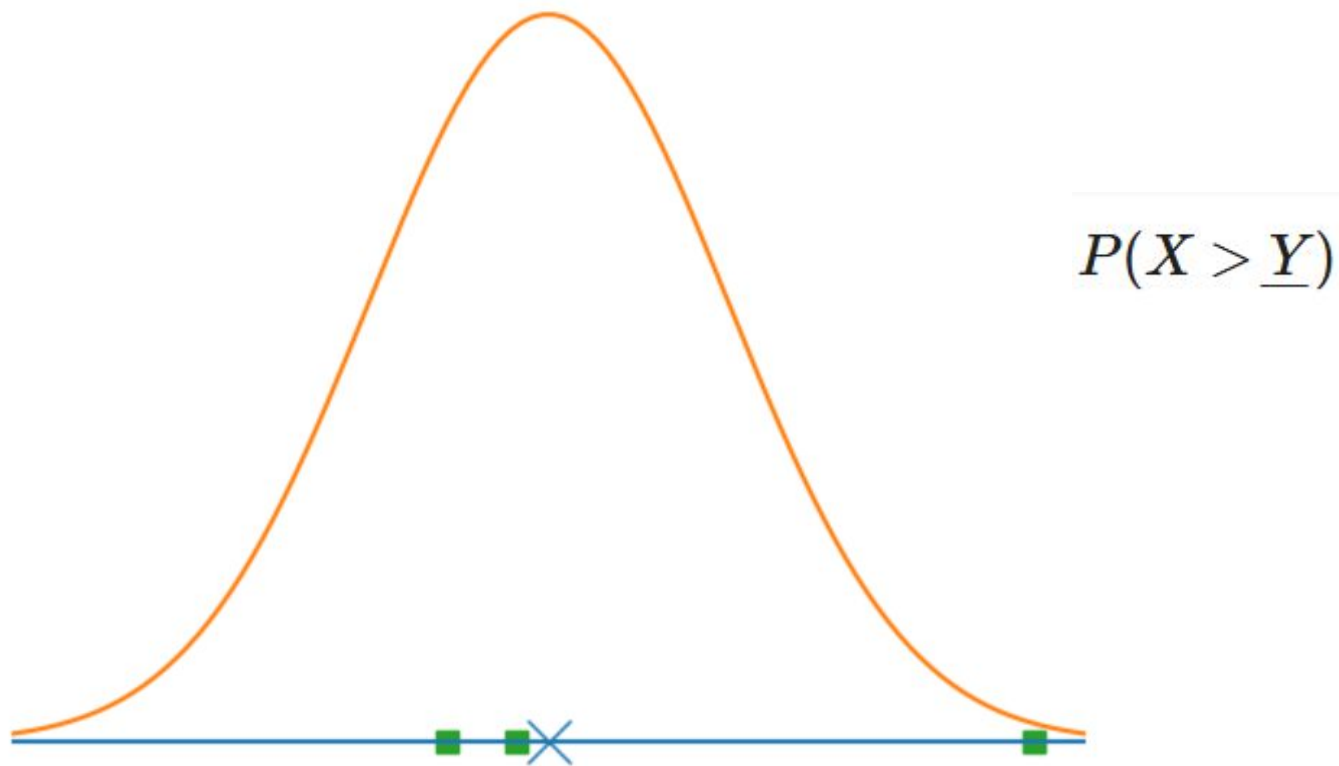
Let $\underline{X}$ is average success rate of machine-1

Let $\underline{Y}$ is average success rate of machine-2

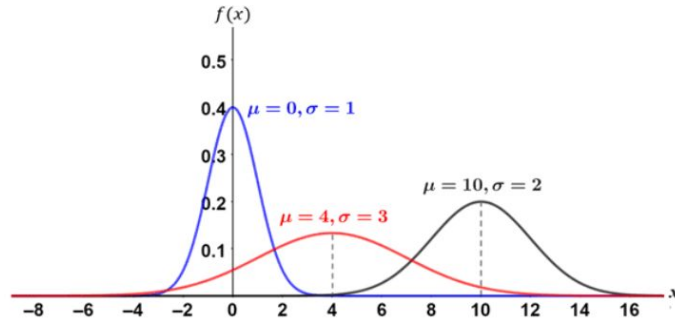
We need to check $\underline{X} > \underline{Y}$ hold for most of the cases

$P(\underline{X} > \underline{Y})$ is high or $\underline{X} > \underline{Y}$ is statistically significant

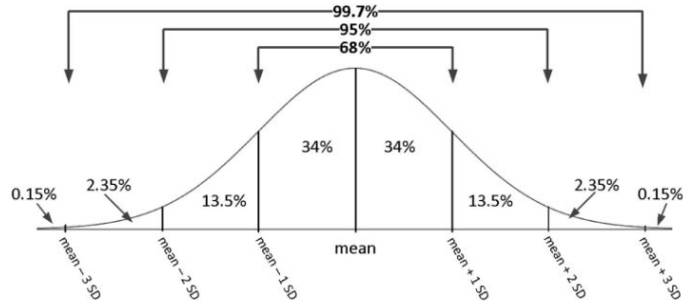
Visualizing Uncertainty



Normal Distribution



- $P(8|10, 2) > P(8|4, 3)$
- $P(6|10, 2) < P(6|4, 3)$



Estimating Uncertainty

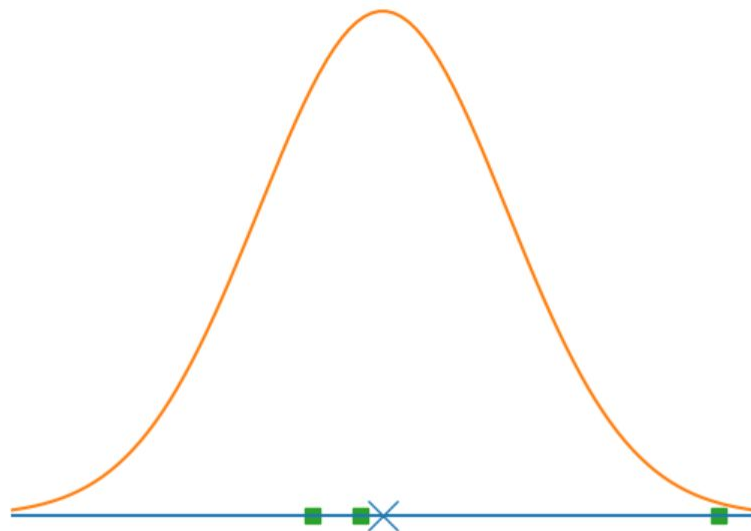
$$P(X > \underline{Y}) \quad \text{Objective}$$

$$X > \underline{Y} \implies \frac{X - \underline{X}}{s} > \frac{\underline{Y} - \underline{X}}{s}$$

$$P(X > \underline{Y}) = P\left(\frac{X - \underline{X}}{s} > \frac{\underline{Y} - \underline{X}}{s}\right)$$

$$Z = \frac{X - \underline{X}}{s}, \quad \delta = \frac{\underline{Y} - \underline{X}}{s}$$

$$P(Z > \delta) \quad \text{Objective}$$



Estimating the Uncertainty

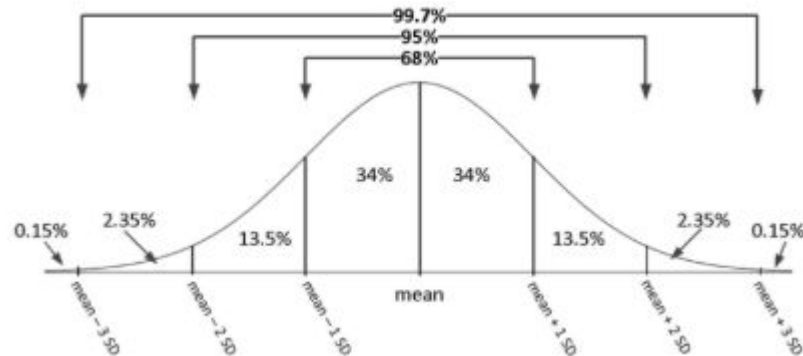
$$P(Z > \delta) \quad \text{Objective}$$

$$Z \sim \mathcal{N}(0, 1) \quad \text{Standard Normal}$$

$$P(Z > \delta) = 1 - \Phi(\delta)$$

$$p = 1 - \Phi(\delta)$$

$$p < \alpha$$



Examples

```
X = np.array([70, 63, 72, 68]) # success rate of machine-1  
Y = np.array([65, 67, 63, 70]) # success rate of machine-2
```

Can not conclude that machine-1 is better

```
X = np.array([70, 72, 72, 68]) # success rate of machine-1
```

Machine-1 is better

```
X = np.array([70, 72, 72, 68, 68, 68])
```

Machine-1 is better

Quiz-7

1. Write 2 things that you liked/understood today.
2. Write 2 things that you did not like/understand today.
3. When should you use statistical significance test and when you may not ?
4. Consider the case study you picked in the last session.
 - a. Identify two hypotheses
 - b. Describe how will you accept one