Foundation in Statistical Analysis

Labeeb Khan

Actuarial Analyst *Co-op*, Intact

Summary

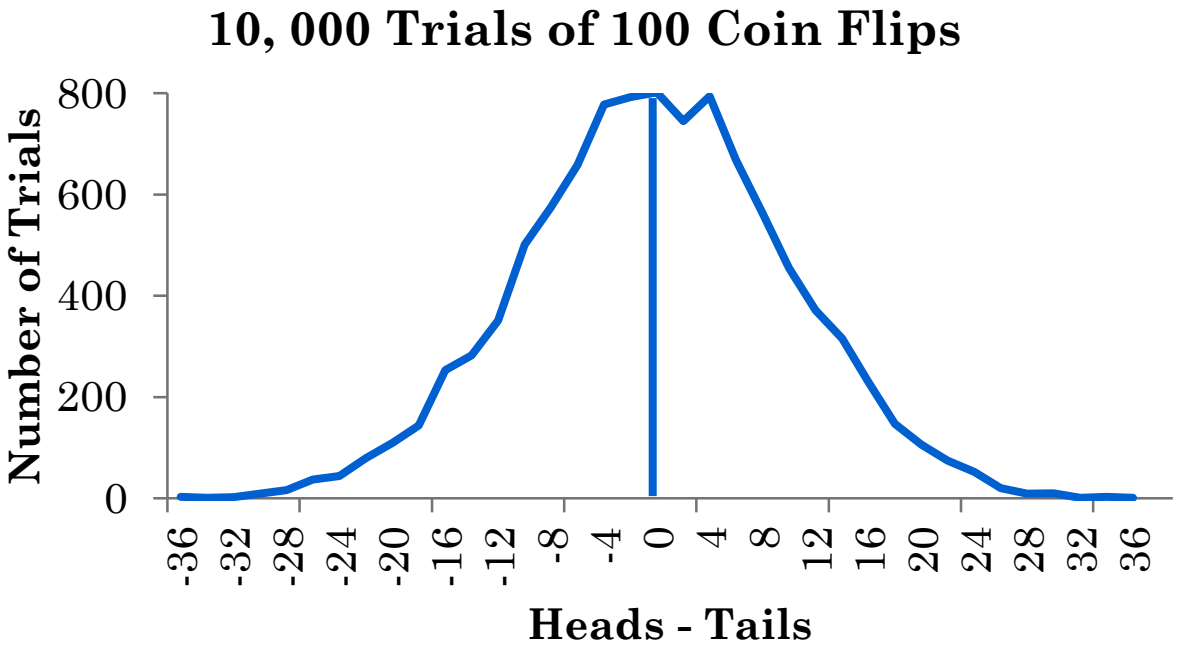
- Average & Expectation
- Variance and Standard Deviation
- Normal Distribution
  - Confidence Intervals
- Stochastic Processes
  - Deterministic vs Stochastic
  - Credibility vs Volume Trade-off

Average and Variance

The Behaviour of Data

Average

- The central value in a dataset
  - Helps us better understand the data
  - What is the average salary for Engineers?
  - What is the average age for someone starting a PhD?



$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

Expectation

- With randomness, what value can we expect?
- The average over a frequency or odds

Flip a coin. If you receive $1 for Heads, and -$1 for Tails, what is the expectation?

Outcome	Probability	Reward
Heads	50%	$1
Tails	50%	-$1

Expectation: $\mu = \left(\frac{1}{2}\%\right) * (\$1) + \left(\frac{1}{2}\%\right) * (-\$1) = 0\ Dollars$

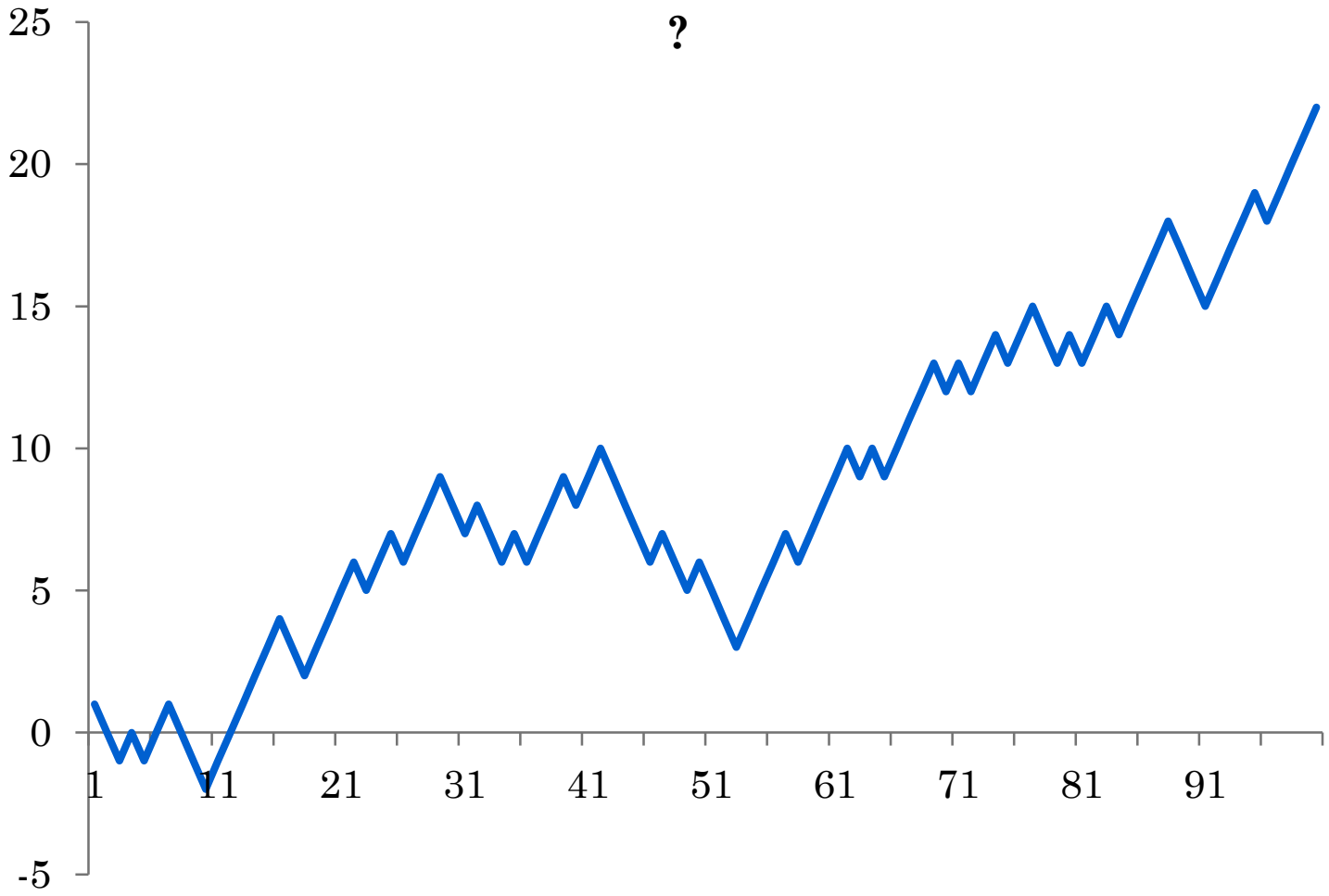
Lottery Example

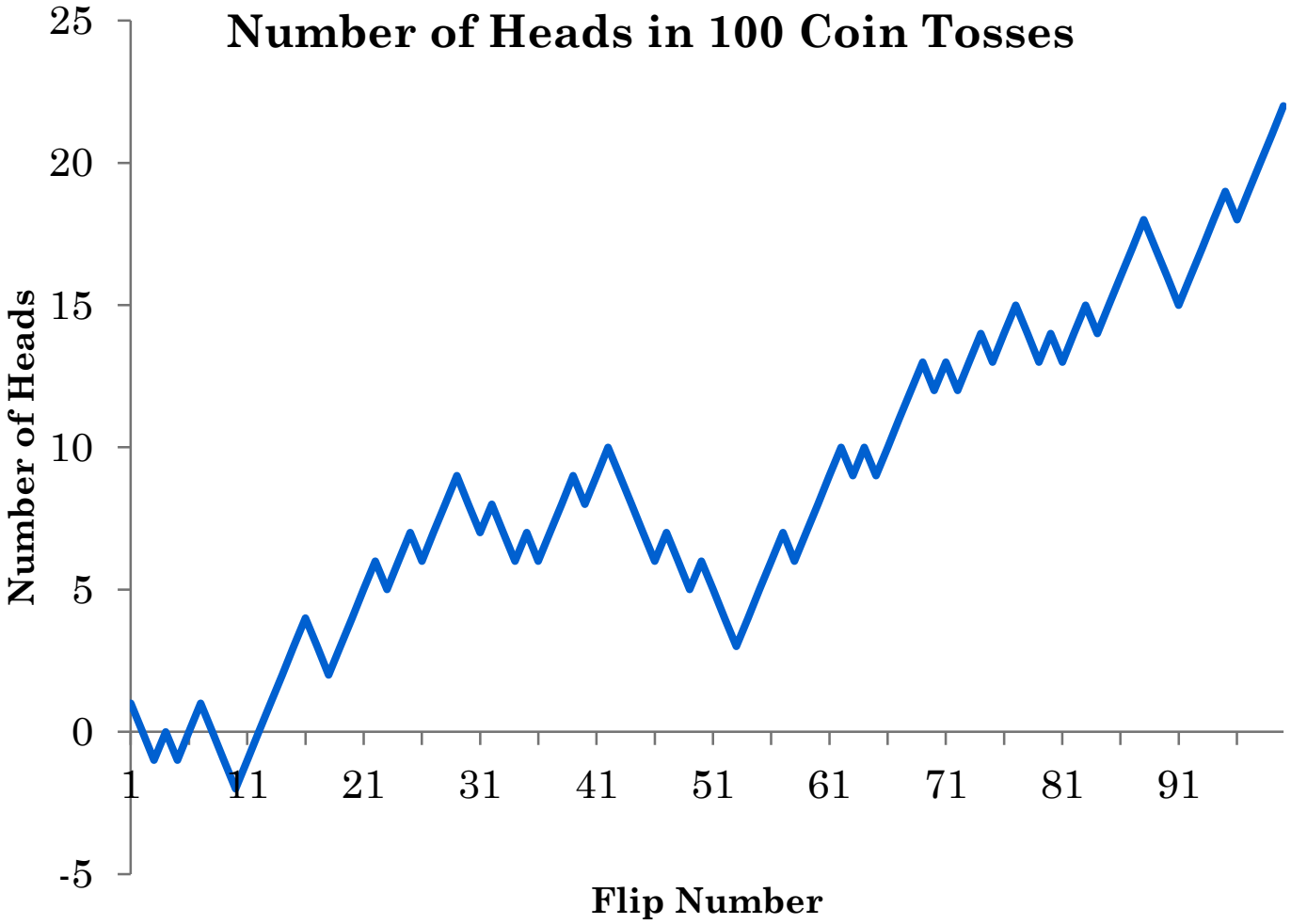
- The chance of winning the lottery (lotto max) is
  1 in 26 million
- The reward is $50 million
- The cost is $5
- What is our expectation?

$$E(X) = \sum_{i=1}^{n} P(X_i) * X_i$$

Outcome	Probability	Reward
Win	$\frac{1}{26{,}000{,}000} = 0.00001\%$	$50, 000, 000
Lose	$\frac{25{,}999{,}999}{26{,}000{,}000} = 99.99999\%$	-$5

$$\mu = (0.00001\%) * (\$50{,}000{,}000) + (0.9999999\%) * (-\$5)$$
$$\mu = -\$3.08$$


?
25
20
15
10
5
0
-5
1
11
21
31
41
51
61
71
81
91

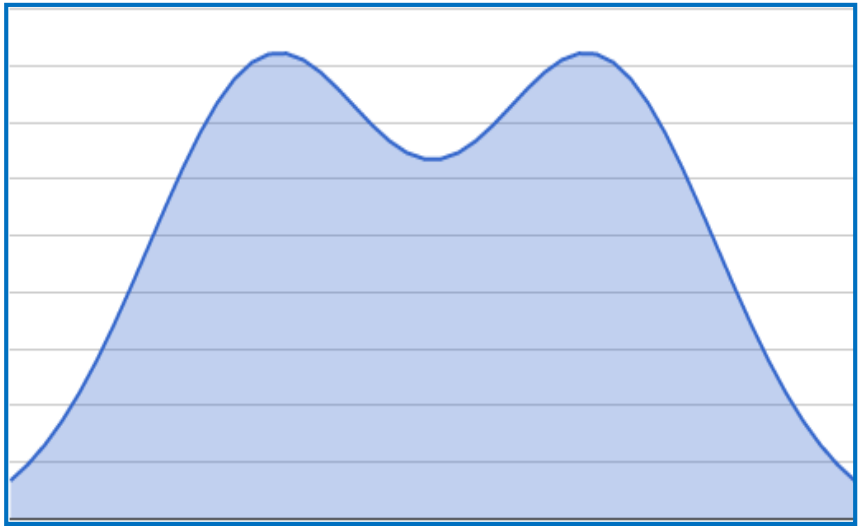


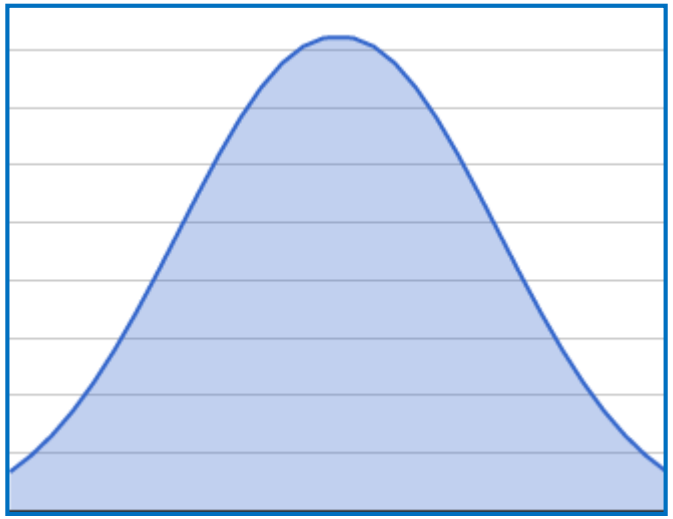
1. With randomness present, you cannot entirely depend on the history
2. The average alone is not always a strong indicator

Variance

- The average fails to understand the stretching of the data

Case 1	Case 2
$100, 000	$50, 000
$0	$50, 000
Average $50, 000	Average $50, 000





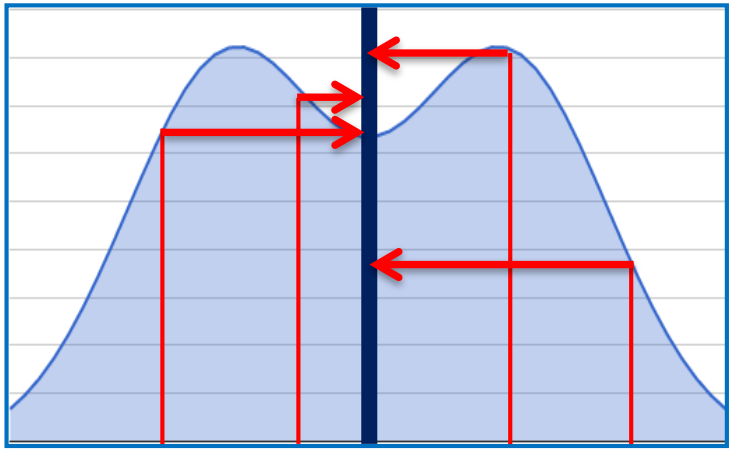
Variance

Variance

- $\frac{(100{,}000 - 50{,}000)^2 + (0 - 50{,}000)^2}{1}$
- $50,000,000,000$

- A way to measure the spread of the data
  - 0 Variance says the numbers are all identical
  - The stretch around the average

$$Var(X) = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n - 1}$$



Standard Deviation

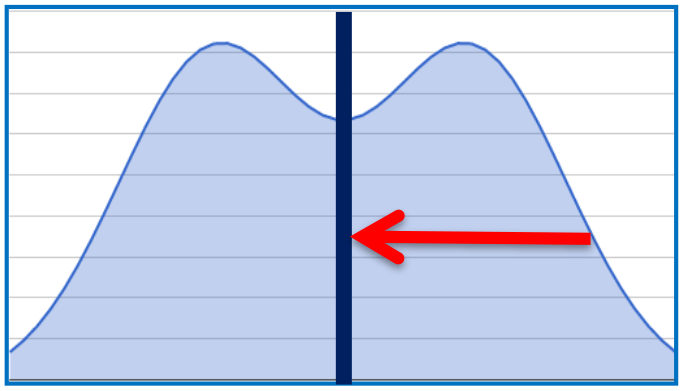
- $Sqrt(Var) = 70,700$

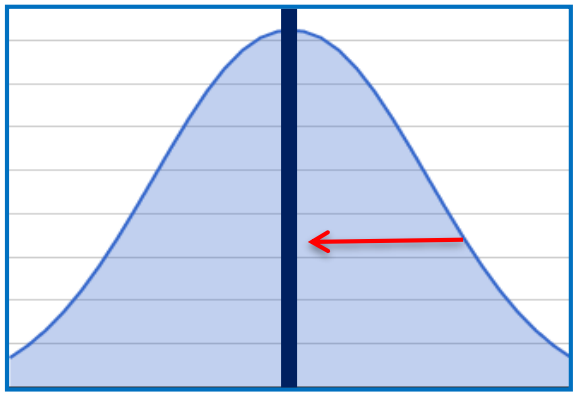
Standard Deviation

- Dispersion around the average

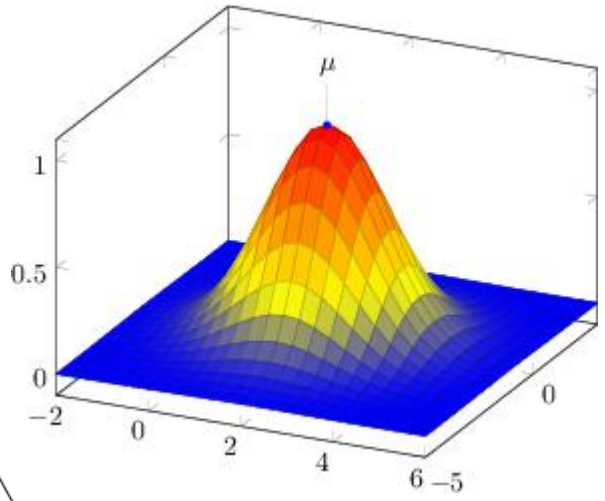
$$\sigma = \sqrt{Var(X)}$$

- low σ : Values are very close to the mean
- High σ : Values are very spread out from the mean



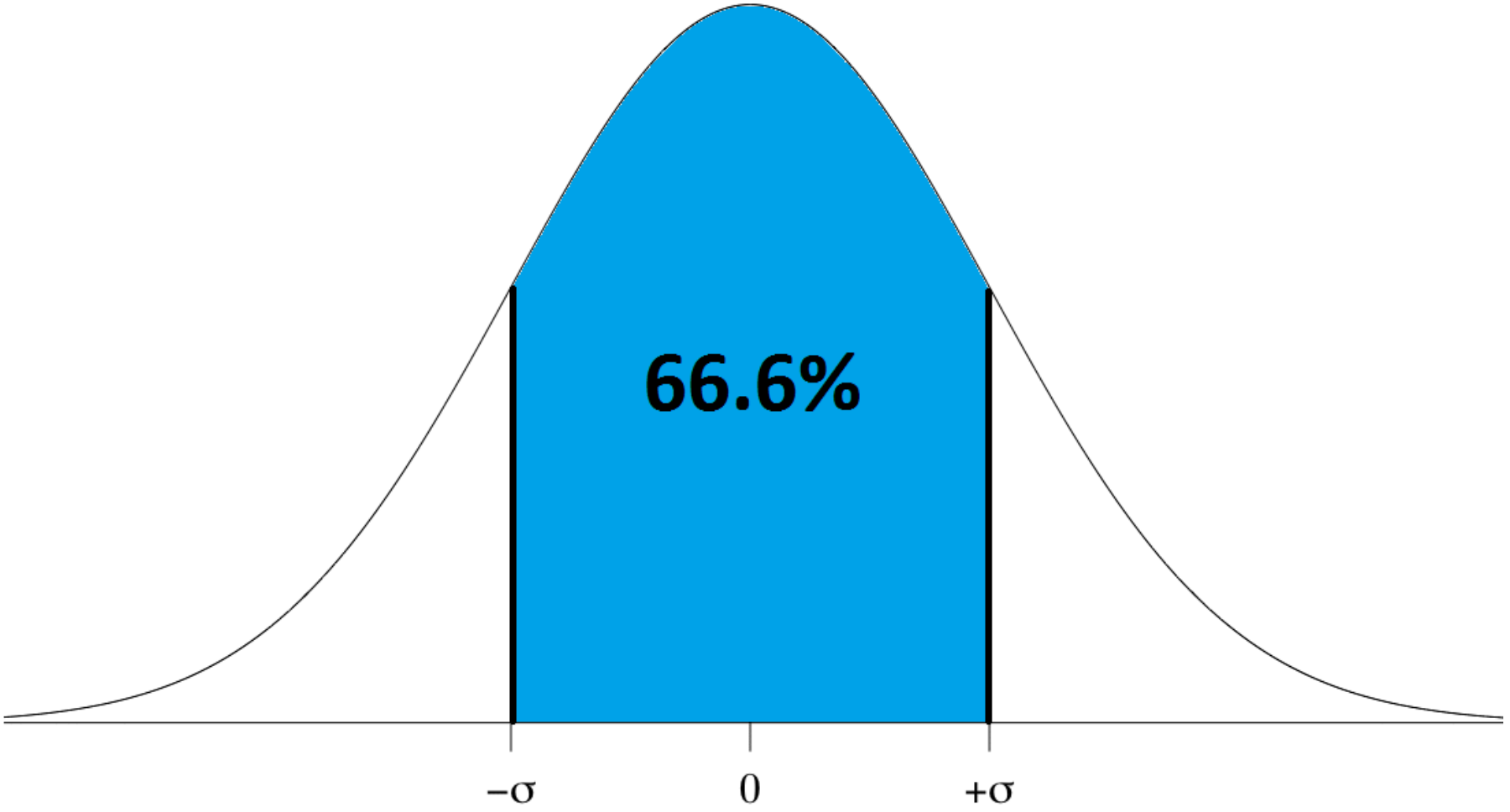


Normal Distribution


μ
1
0.5
0
−2
0
2
4
6
−5
0

−σ
0
+σ

$\mu - \sigma$ $\mu + \sigma$



Confidence Intervals

- Adding σ to μ covers 33.3% of the upper tail

- $\mu + \sigma$ and $\mu - \sigma$ cover 66.6% of the population
- $\mu + 2\sigma$ and $\mu - 2\sigma$ cover 95% of the population
- $\mu + 3\sigma$ and $\mu - 3\sigma$ cover 99.7% of the population

- This is called the 66, 95, and 99 rule
  - Have a confidence region over what to expect
    - Much better than just an average

Stochastic Processes

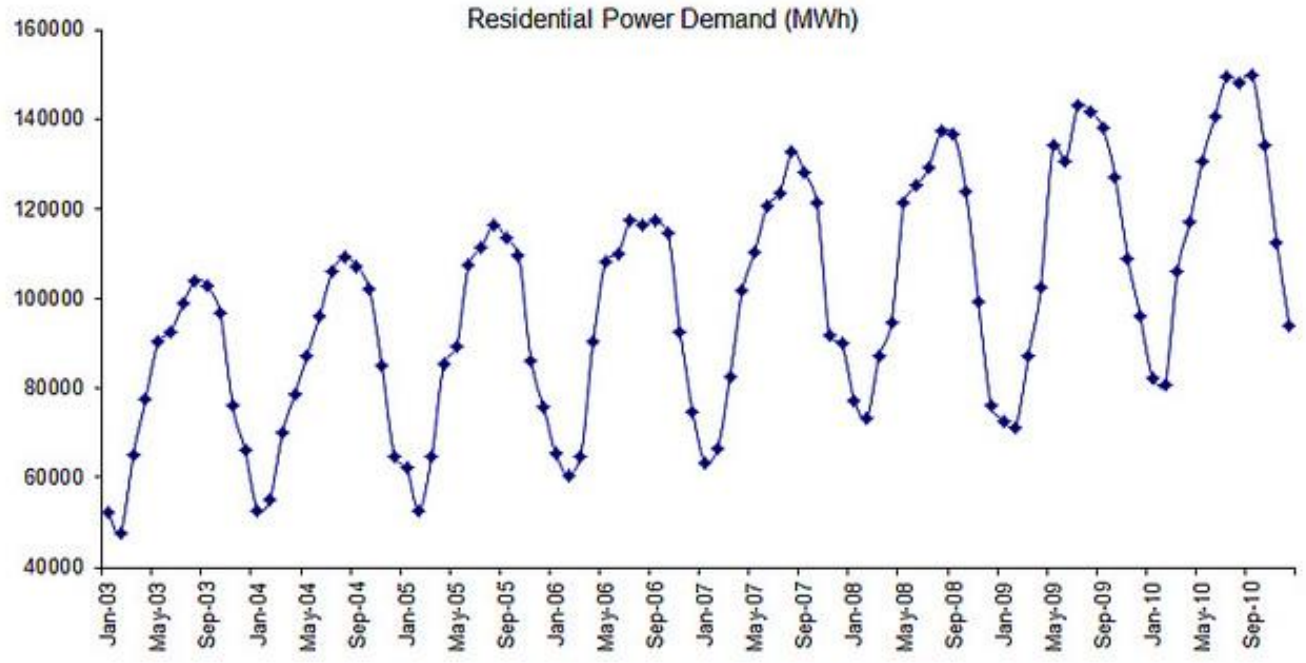
The Behaviour of Random Data

Stochastic Process

Future = Past + Randomness

$$X_{n+1} = X_n + \epsilon$$

- Less Randomness → More Deterministic
- More Randomness → Less Deterministic



Stochastic Process

- Deterministic
  - The farther you go into the past, the more predictive
- Stochastic
  - The farther you go into the past, the less predictive

If the data behaves both stochastic and deterministic:

There exists a trade-off between volume and credibility

- More history → less predictive
- Less history → lower volume → less credible