

Engineering Analytics & Machine Learning (ECSE202)

Seminar 2 Essential Statistics



Introduction to Probability

Our lives consists of uncertain events. For example, what bus will arrive at the bus stop when you reach it after school today ? How long will the queue be at your favorite stall during lunch tomorrow ?

The field of statistics includes the study of probabilities, which is an attempt to quantify how likely an event can happen. If an event is deemed highly probable, then we may need to prepare for it. If it is highly likely to rain in the afternoon, you may want to make plans so that you are affected as little as possible by the wet weather should it happen.

“Mama always said life was like a box of chocolates. You never know what you're gonna get.” –Forrest Gump (played by Tom Hanks)

Introduction to Probability

- Random experiment such as tossing a coin, a probability measure is a population quantity that summarizes the randomness
- Discrete outcome (x) of the coin is either 1 (Head) or 0 (Tail)
- The probability of the union of any two sets of outcomes that have nothing in common (mutually exclusive) is the sum of their respective probabilities



Introduction to Probability

- We measure probability by first establishing the total number of possible outcomes. For example, in an unbiased 6-sided dice, the total number of outcomes is 6.
- What then is the probability of getting a '5' on a roll ?
- Seeing that '5' is only one of six possible outcomes, the probability is $\frac{1}{6}$. This is the same with all other numbers. Each number has a $\frac{1}{6}$ probability of showing up.

Introduction to Probability

Since the dice roll results in random numbers, it is a random event and can be described by a random variable. This variable is discrete because the outcome of the dice roll is discrete (not floating point). Therefore, we use a random variable X to describe the outcome of a dice roll, and we say

$$P(X = 5) = \frac{1}{6}$$



That is, the probability of getting '5' on a dice roll is $\frac{1}{6}$.

Introduction to Probability

Building on the example of the dice roll previously, what is the probability then of getting a either a '1' or a '2' ?

Let the sample space S define all possible outcomes from the dice roll :

$$S = \{1, 2, 3, 4, 5, 6\}$$

Let A be the event that '1' shows up and B be the event that '2' shows up.

Introduction to Probability

We know that probability of obtaining a '1' is $\frac{1}{6}$, similarly the probability of obtaining '2' is also $\frac{1}{6}$. The probability of obtaining either a '1' or a '2' is written as :

$$P(A \cup B) = P(A) + P(B)$$

$A \cup B$ is read "A union B"

$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$



In this example, it is impossible to get both '1' and '2' in a single throw of the dice. The outcomes are said to be *mutually exclusive*, we get either a '1' or a '2', but not both. If the outcomes are not mutually exclusive, combining probabilities is calculated differently.

Axioms of Probability

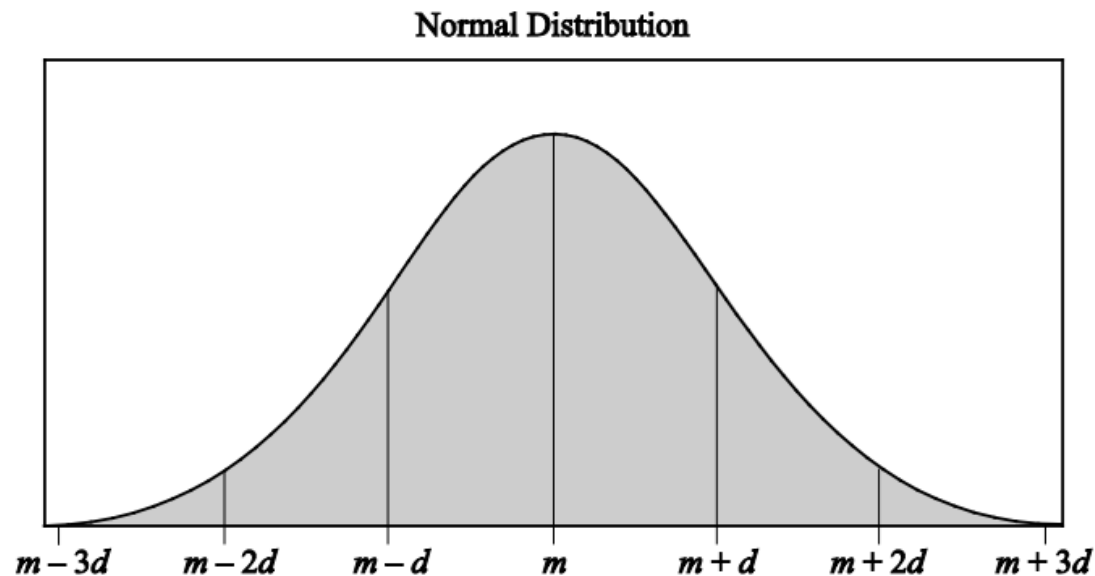
1. Let S be a sample space. Then $P(S) = 1$.
2. For any event A , $0 \leq P(A) \leq 1$
3. If event A and event B are mutually exclusive, then
$$P(A \cup B) = P(A) + P(B)$$

In addition, for any event A ,

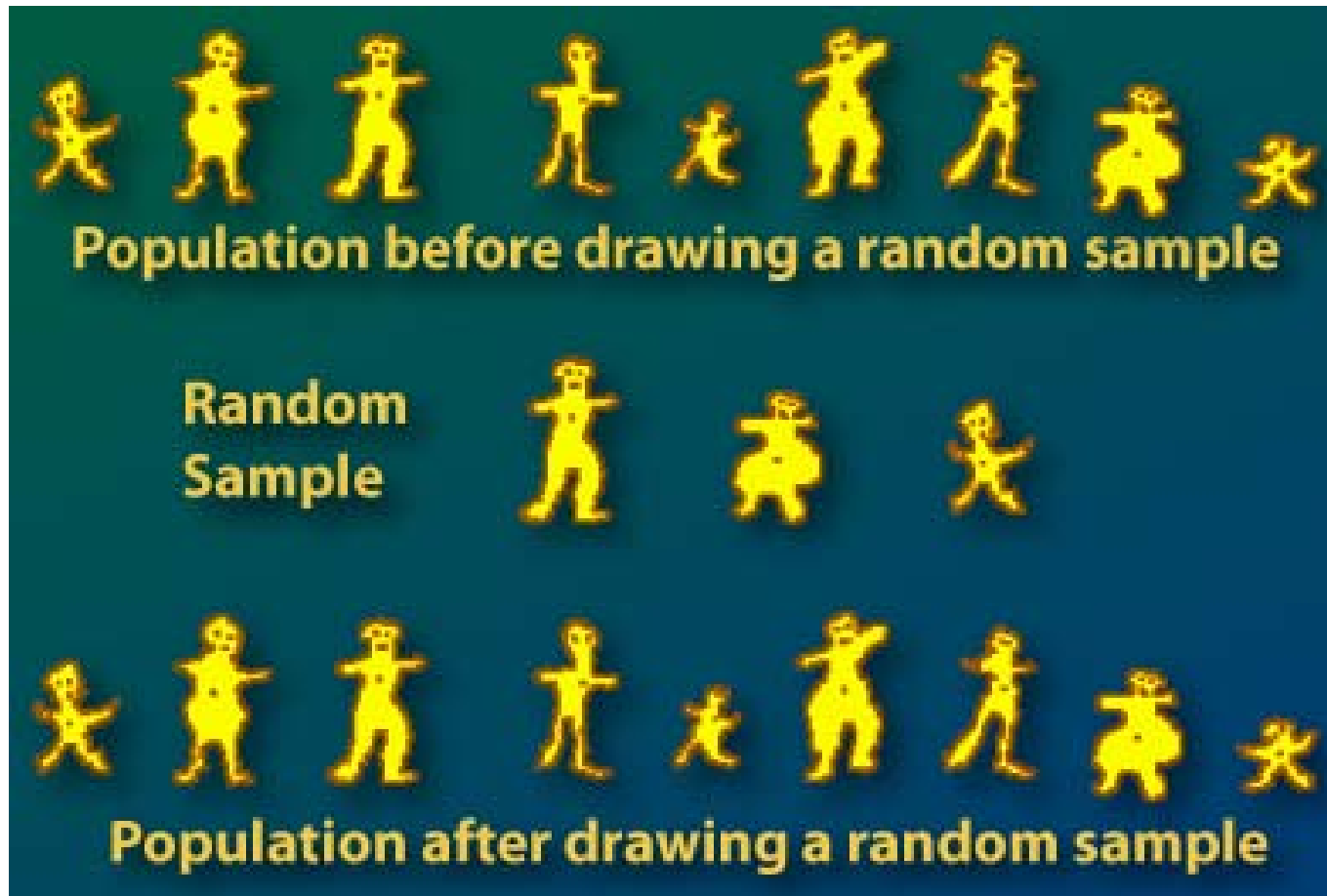
The probability that an event A does not happen is given by $P(A^c) = 1 - P(A)$

Expected Values

- **Expected values** are useful for characterizing a distribution
- The **mean** is a characterization of its center
- The **variance** and **standard deviation** are characterizations of how spread out it is
- Sample expected values (the sample mean and variance) will estimate the population mean and variance



Sampling



Central Limit Theorem

- The Central Limit Theorem (CLT) is one of the most important theorems in statistics.
- The central limit theorem (CLT) is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.
- Furthermore, all of the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size.

Read more: Central Limit Theorem (CLT)

https://www.investopedia.com/terms/c/central_limit_theorem.asp#ixzz5CwasElcJ

Sample vs. Population

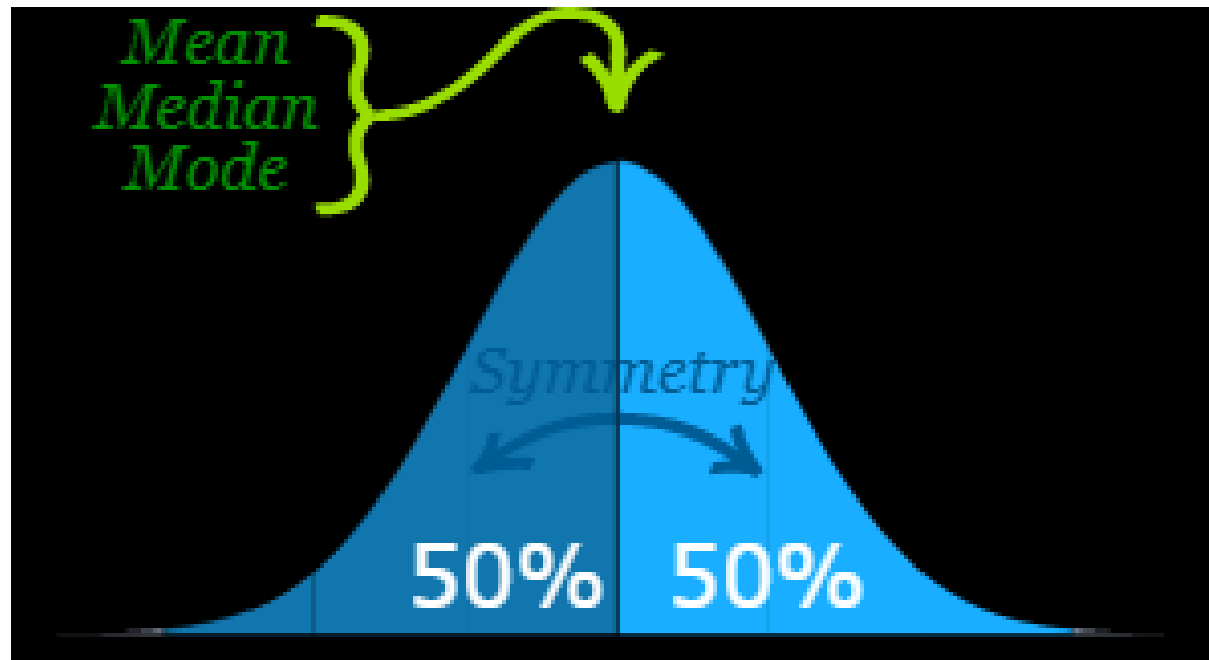
- The mean, variance and standard deviation of a random sample is referred to as the sample mean, sample variance and sample standard deviation
- The sample mean, variance, standard deviation can only give us an approximation to the population mean, the population variance and population standard deviation
- When the value of the sample, n , is large enough, the sample mean, variance and standard deviation gives a good estimation of the population statistic parameters.

Sample vs. Population

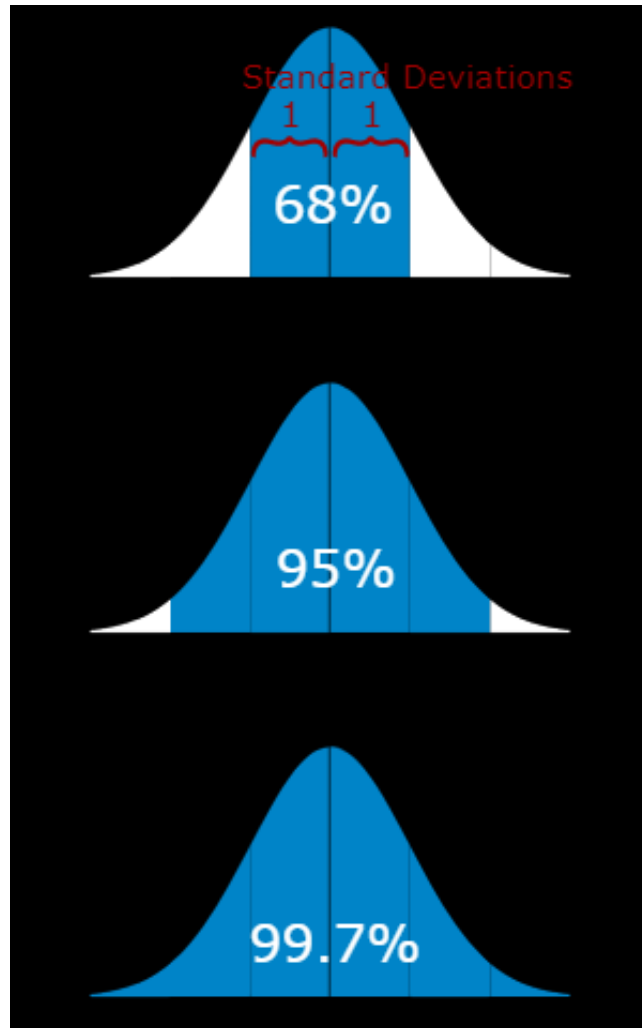
Table 1: Mean and Standard Deviation				
Measure Name	Symbol for Population	Symbol for Sample	Computation for Population	Computation for Sample
Mean	μ	\bar{x}	$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
Standard Deviation	σ	s_x	$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$ <p>or the equivalent form</p> $\sigma = \sqrt{\frac{\sum_{i=1}^N (x^2) - \frac{(\sum_{i=1}^N x_i)^2}{N}}{N}}$	$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$ <p>or the equivalent form</p> $s_x = \sqrt{\frac{\sum_{i=1}^n (x^2) - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n - 1}}$

Normal Distribution

- Mean=median=mode
- Symmetry about the center
- 50% of values less than the mean and 50% greater than the mean



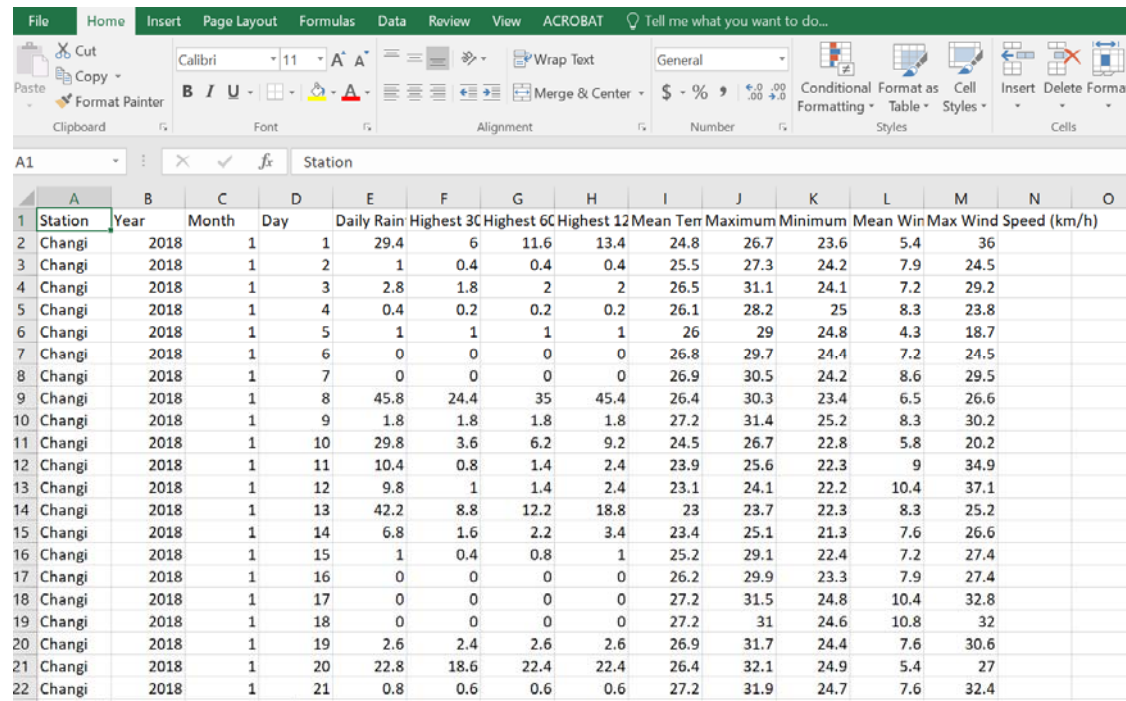
Normal Distribution



- 68% of values are within 1 standard deviation of the mean
- 95% of values are within 2 standard deviation of the mean
- 99.7% of values are within 3 standard deviation of the mean

Historical Weather Data

Still remember the Historical Weather Data?



Station	Year	Month	Day	Daily Rain	Highest 3C	Highest 6C	Highest 12C	Mean Ten	Maximum	Minimum	Mean Win	Max Wind Speed (km/h)
Changi	2018	1	1	29.4	6	11.6	13.4	24.8	26.7	23.6	5.4	36
Changi	2018	1	2	1	0.4	0.4	0.4	25.5	27.3	24.2	7.9	24.5
Changi	2018	1	3	2.8	1.8	2	2	26.5	31.1	24.1	7.2	29.2
Changi	2018	1	4	0.4	0.2	0.2	0.2	26.1	28.2	25	8.3	23.8
Changi	2018	1	5	1	1	1	1	26	29	24.8	4.3	18.7
Changi	2018	1	6	0	0	0	0	26.8	29.7	24.4	7.2	24.5
Changi	2018	1	7	0	0	0	0	26.9	30.5	24.2	8.6	29.5
Changi	2018	1	8	45.8	24.4	35	45.4	26.4	30.3	23.4	6.5	26.6
Changi	2018	1	9	1.8	1.8	1.8	1.8	27.2	31.4	25.2	8.3	30.2
Changi	2018	1	10	29.8	3.6	6.2	9.2	24.5	26.7	22.8	5.8	20.2
Changi	2018	1	11	10.4	0.8	1.4	2.4	23.9	25.6	22.3	9	34.9
Changi	2018	1	12	9.8	1	1.4	2.4	23.1	24.1	22.2	10.4	37.1
Changi	2018	1	13	42.2	8.8	12.2	18.8	23	23.7	22.3	8.3	25.2
Changi	2018	1	14	6.8	1.6	2.2	3.4	23.4	25.1	21.3	7.6	26.6
Changi	2018	1	15	1	0.4	0.8	1	25.2	29.1	22.4	7.2	27.4
Changi	2018	1	16	0	0	0	0	26.2	29.9	23.3	7.9	27.4
Changi	2018	1	17	0	0	0	0	27.2	31.5	24.8	10.4	32.8
Changi	2018	1	18	0	0	0	0	27.2	31	24.6	10.8	32
Changi	2018	1	19	2.6	2.4	2.6	2.6	26.9	31.7	24.4	7.6	30.6
Changi	2018	1	20	22.8	18.6	22.4	22.4	26.4	32.1	24.9	5.4	27
Changi	2018	1	21	0.8	0.6	0.6	0.6	27.2	31.9	24.7	7.6	32.4

We may want to know what is the probability that it would rain on any day in January.

Data Variable as Random Process

	A	B	C	D	E	N	O
1	Station	Year	Month	Day	Daily Rain	Rain	
2	Changi	2018	1	1	29.4	1	
3	Changi	2018	1	2	1	1	
4	Changi	2018	1	3	2.8	1	
5	Changi	2018	1	4	0.4	1	
6	Changi	2018	1	5	1	0	
7	Changi	2018	1	6	0	0	
8	Changi	2018	1	7	0	1	
9	Changi	2018	1	8	45.8	1	
10	Changi	2018	1	9	1.8	1	
11	Changi	2018	1	10	29.8	1	
12	Changi	2018	1	11	10.4	1	
13	Changi	2018	1	12	9.8	1	
14	Changi	2018	1	13	42.2	1	
15	Changi	2018	1	14	6.8	1	
16	Changi	2018	1	15	1	0	
17	Changi	2018	1	16	0	0	
18	Changi	2018	1	17	0	0	

1. Create a new column call Rain.
2. Rain would have value of '1' if Daily Rain > 0
3. Rain would have value '0' if Daily rain ≤

Data Variable as Random Process

	A	B	C	D	E	N	O	P	Q
1	Station	Year	Month	Day	Daily Rain	Rain			
2	Changi	2018	1	1	29.4	1			
3	Changi	2018	1	2	1	1			
4	Changi	2018	1	3	2.8	1		Total sample =	31
5	Changi	2018	1	4	0.4	1		Rain day =	25
6	Changi	2018	1	5	1	0		Non Rain day=	6
7	Changi	2018	1	6	0	0		Probability of rain =	0.806452
8	Changi	2018	1	7	0	1			
9	Changi	2018	1	8	45.8	1			
10	Changi	2018	1	9	1.8	1			
11	Changi	2018	1	10	29.8	1			
12	Changi	2018	1	11	10.4	1			
13	Changi	2018	1	12	9.8	1			
14	Changi	2018	1	13	42.2	1			
15	Changi	2018	1	14	6.8	1			
16	Changi	2018	1	15	1	0			
17	Changi	2018	1	16	0	0			
18	Changi	2018	1	17	0	0			

Worked Example 1

A sensor collected temperature data of a boiler. 95% of the temperature values are between 100 degrees to 120 degrees. Assume the data is normally distributed, what is the mean and standard deviation?

Solution:

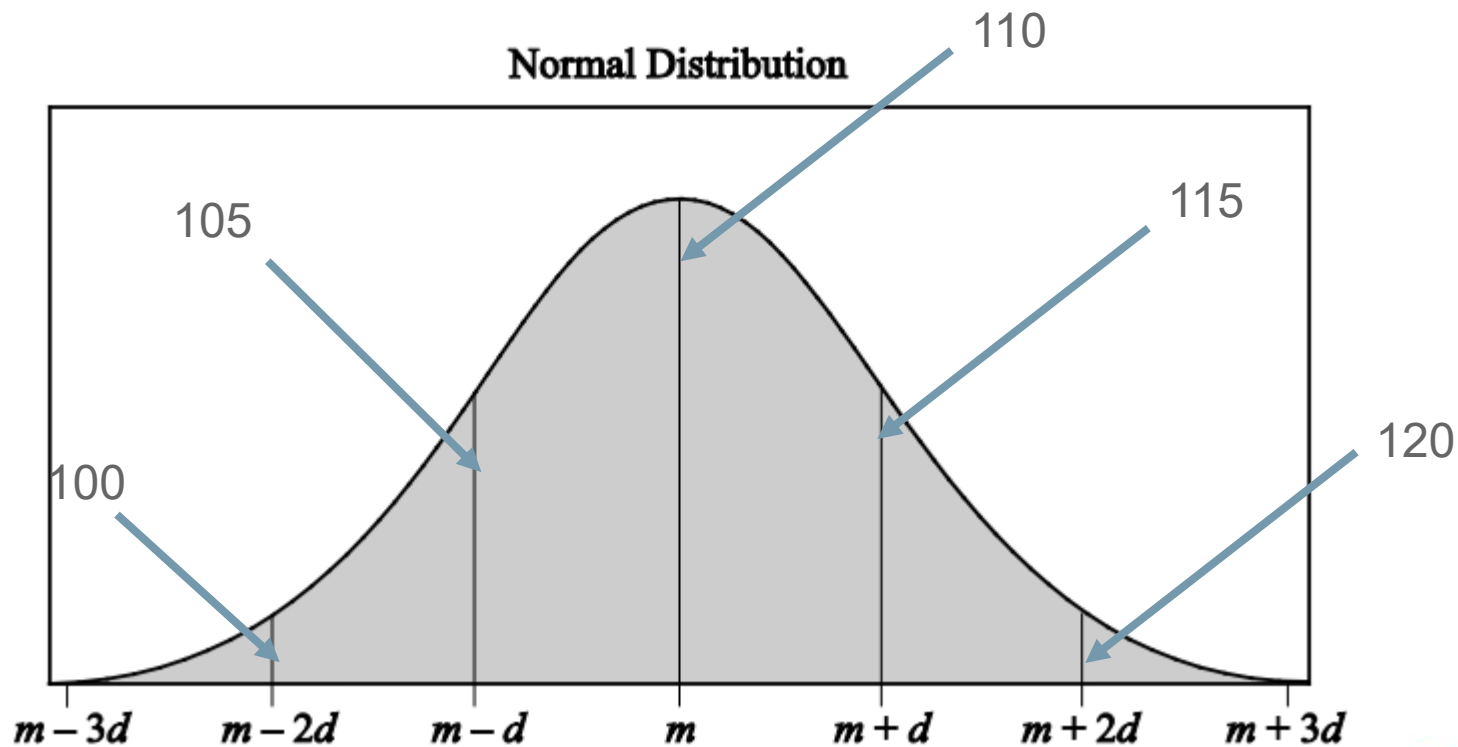
The mean is halfway between 100 and 120 degrees:

$$\text{mean} = (100+120)/2 = 110 \text{ degrees}$$

95% is 2 standard deviations (std) either side of the mean (total 4 standard deviation):

$$\text{std} = (120-100)/4 = 5 \text{ degrees}$$

Worked Example 1



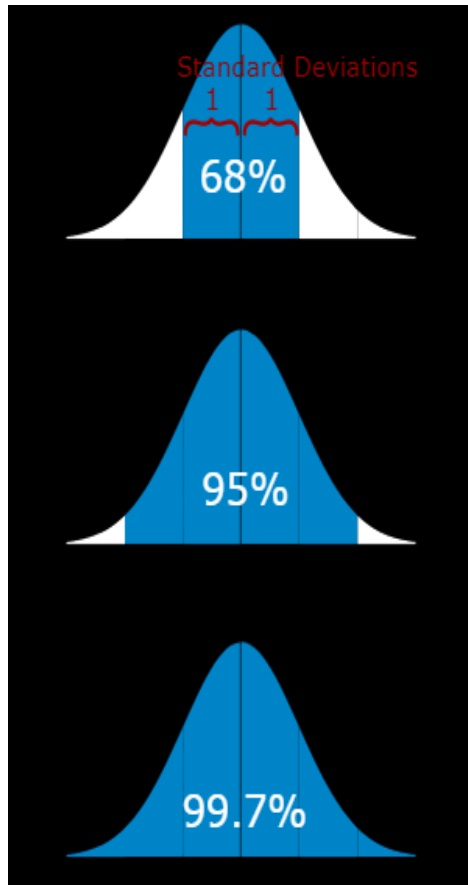
Worked Example 2

Company XYZ wants to buy a laser cutter to cut metal bar to length of 10mm with a tolerance of $\pm 0.1\text{mm}$ and can only accept a failure rate of 5%. The technician tested four models of cutter and produce the following report:

Model	Mean (mm)	Standard Deviation
A	9.99	0.04
B	10.03	0.05
C	10.01	0.07
D	9.98	0.06

Assuming the data is normally distributed. Which model of the laser cutter should the company buy? Show workings to justify your recommendation.

Worked Example 2



Error rate of 5% mean 95% of the product must be within $\pm 0.1\text{mm}$

Require cutter to be within tolerance 95% of the time, the accuracy from 9.90mm to 10.10mm

Model A: 95% of the time, the accuracy will be from 9.91mm to 10.07mm

Model B: 95% of the time, the accuracy will be from 9.93mm to 10.13mm

Model C: 95% of the time, the accuracy will be from 9.87mm to 10.15mm

Model D: 95% of the time, the accuracy will be from 9.86mm to 10.1mm

Model A is well within the range required.