Lecture -- 4 -- Start

Outline

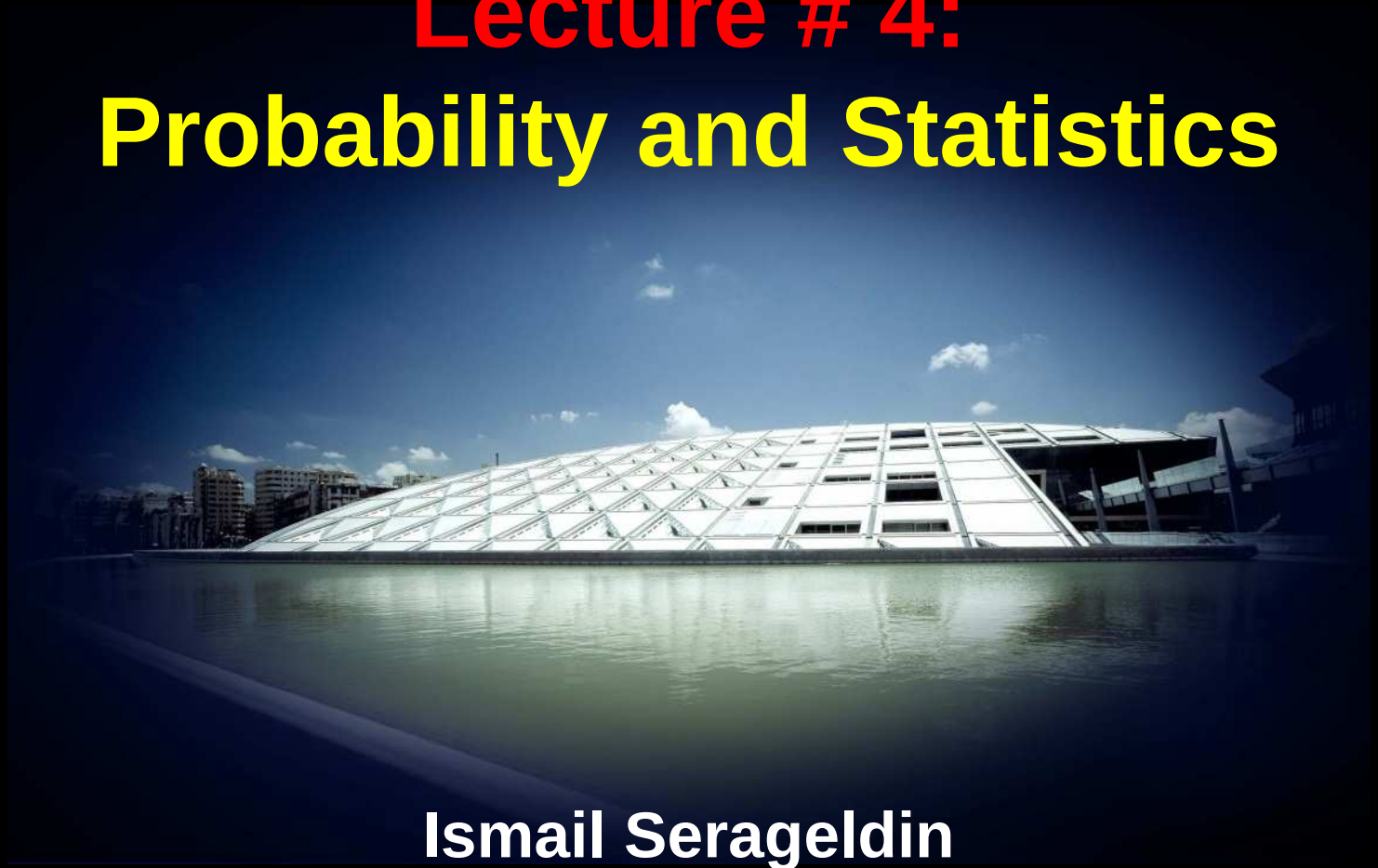
1. Science, Method & Measurement
2. On Building An Index
3. Correlation & Causality
4. Probability & Statistics
5. Samples & Surveys
6. Experimental & Quasi-experimental Designs
7. Conceptual Models
8. Quantitative Models
9. Complexity & Chaos
10. Recapitulation - Envoi

Outline

1. Science, Method & Measurement
2. On Building An Index
3. Correlation & Causality
4. **Probability & Statistics**
5. Samples & Surveys
6. Experimental & Quasi-experimental Designs
7. Conceptual Models
8. Quantitative Models
9. Complexity & Chaos
10. Recapitulation - Envoi

Quantitative Techniques for Social Science Research

Lecture # 4:
Probability and Statistics



Ismail Serageldin
Alexandria
2012

On Probabilities

Recall

Random events

Random events/outcomes require a probabilistic treatment

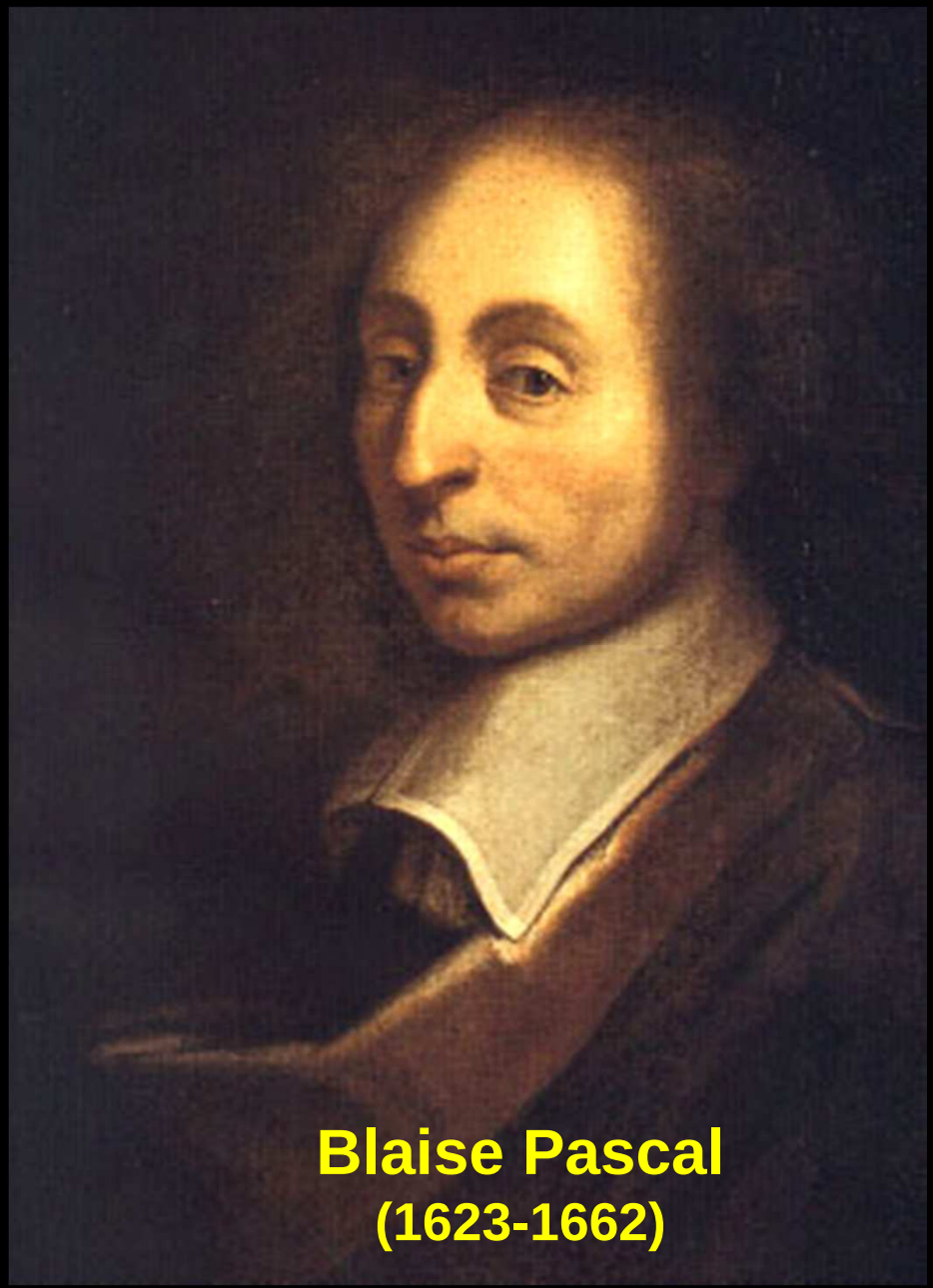
Social Science studies of events/outcomes usually require a statistical probabilistic treatment

Here multiple measurements and probabilistic techniques are used

Probability became a science in the 17^{th} century

A Genius: Blaise Pascal
(1623-1662)

- As a child he rediscovered much of geometry
- He wrote the most important study on conic sections in 1500 years – Descartes could not believe that a child of 16 could write such a treatise
- He invented one of the first calculating machines
- He established the rules of hydraulics



Blaise Pascal
(1623-1662)

His friends asked him if he could find the way to beat chance in gambling


17 3 4 9 9 6 27


00
0
3
6
9
12
2
5
8
11
15
1
4
7
10
14
18
21
13
17
20
24
16
19
23
27
30
22
26
29
33
25
28
32
36
31
35
34
1 - 18
Even
Odd
19 - 36
2 to 1
2 to 1
2 to 1

Pascal developed probability theory , corresponding with another genius: Pierre de Fermat

Pierre de Fermat
(1601-1665)

The Science of Probability was born

In general, for independent events:

Probability of an outcome =

number of ways that outcome can happen / the number of all possible outcomes

There are of course, a lot of other things, but this is a good place to start

A standard deck has 52 cards:
13 cards (A,K,Q,J,10,9,….,3,2) in each of 4 suits (Spades, Hearts, Clubs and Diamonds)

So, what is the probability of drawing any particular card or combination of cards?

To find out the probability of drawing any particular 5-card hand (without replacement)

- Given all combinations of 5 cards randomly drawn from a full deck of 52 without replacement. Wild cards are not considered.
- The probability of drawing a given hand is calculated by dividing the number of ways of drawing the hand by the total number of 5-card hands (the sample space, five-card hands).

Without replacement is an important point

- The first card is to be drawn is 1/52
- The second card to be drawn (given the outcome of the first draw) will be drawn out of 51 cards not 52.
- The third will be drawn from 50 cards.
- The combined probability will take into account how many ways you can draw the hand (the sequence of the cards does not matter)

The total number of possible 5-card hands is: 2,598,960

To calculate the probability of a particular 5-card hand

- requires finding out how many ways we can get that hand.
- Poker hands are combinations of cards (when the order does not matter, but each object can be chosen only once.)
- The total number of possible 5 card hands is 2,598,960.

Four drawing four Aces

- The number of hands which contain 4 aces is 48 (the fifth card can be any of 48 other cards.)
- So there is 1 chance in (2,598,960 / 48) = 54,145 of being dealt 4 aces in a 5 card hand.
- probability is 1 / 54145 = 0.0018469%.

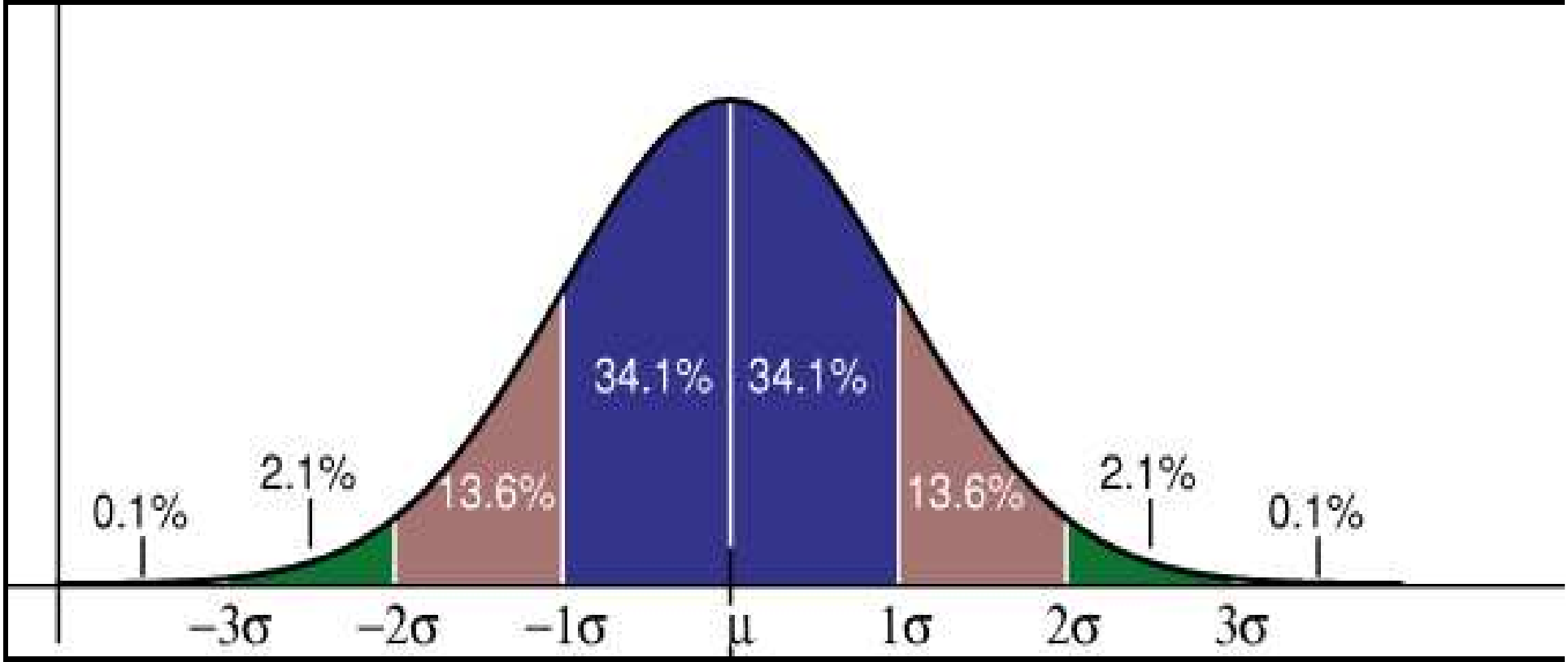
Probability of Four Aces:
1: 54145 = 0.0018469%.

Probability of a Royal Flush:
1 : 649, 739 = 0.000154%

Thus was probability theory born!

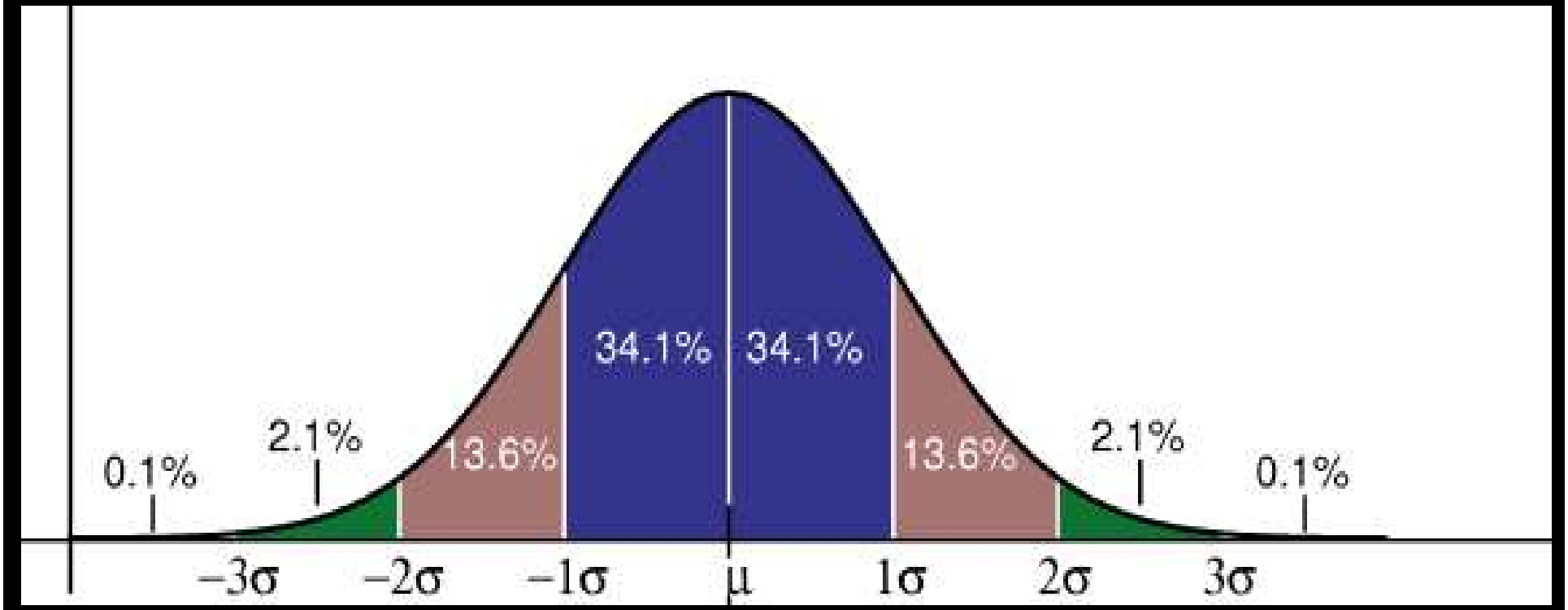
If you map a lot of independent observations you get a bell-shaped curve

The Gaussian Distribution



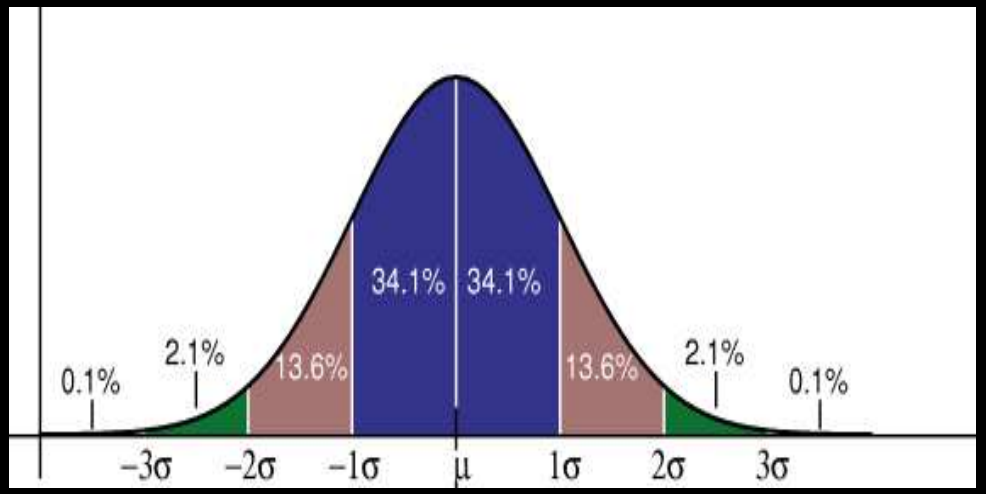
- **As the figure above illustrates, 68% of the values lie within 1 standard deviation of the mean; 95% lie within 2 standard deviations; and 99.7% lie within 3 standard deviations.**

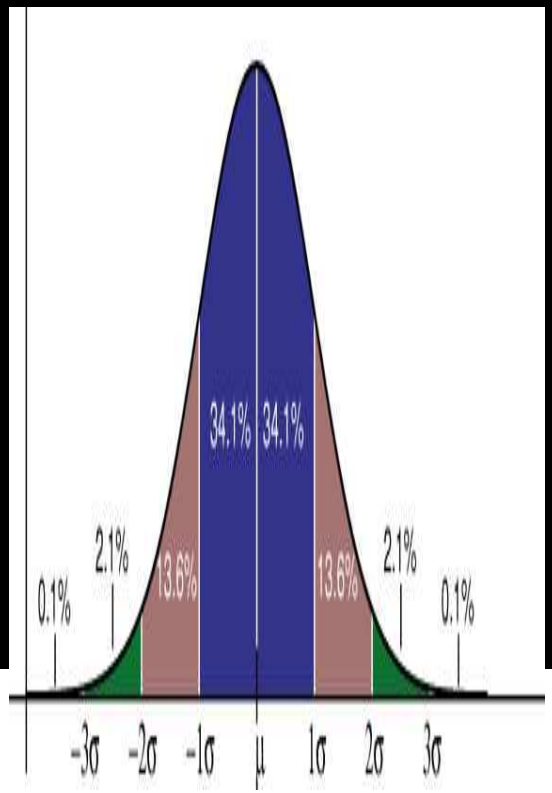
The Properties of the Gaussian Distribution

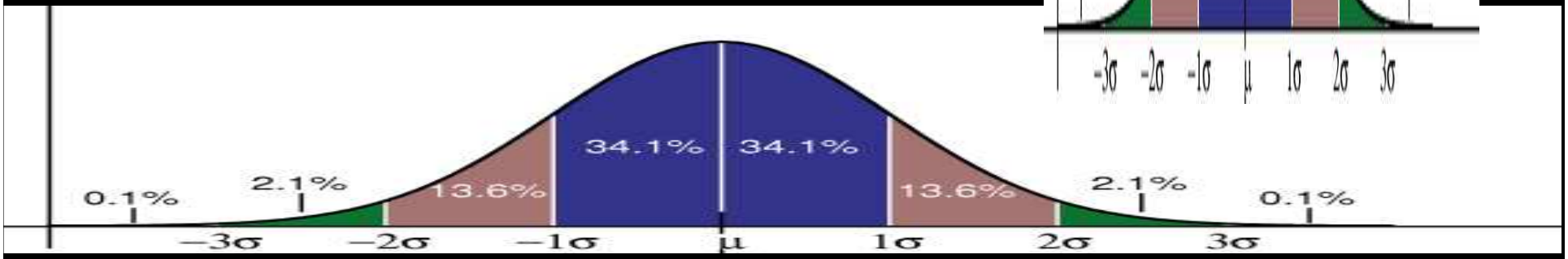


- 68% of the values lie within 1 standard deviation of the mean;
- 95% lie within 2 standard deviations; and
- 99.7% lie within 3 standard deviations.

The Properties remain the same whatever the values of the mean and the standard deviation of the Gaussian Distribution







The Gaussian (normal) distribution

- The Gaussian (normal) distribution was historically called the *law of errors*.
- It was used by Gauss to model errors in astronomical observations, which is why it is usually referred to as the Gaussian distribution.

The Gaussian (normal) distribution

- **The *probability density function* for the standard Gaussian distribution (mean 0 and standard deviation 1) and the Gaussian distribution with mean μ and standard deviation σ is given by the following formulas.**

$$\varphi(z)=\frac{1}{\sqrt{2n}}\exp\left(-\frac{x^2}{2}\right)$$

$$\varphi(z;\ \mu;\sigma\ )=\frac{1}{\varphi\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\varphi^2}\right)$$

The Gaussian (normal) distribution

- The *cumulative distribution function* for the standard Gaussian distribution and the Gaussian distribution with mean μ and standard deviation σ is given by the following formulas:

$$\varphi(z)=\int_{-\infty}^{x} \varphi(x)\mathbf{dx}$$

$$\varphi(z;\ \mu;\sigma\ )=\int_{-\infty}^{x} \varphi(\mathrm{x};\ \mu;\sigma)\mathbf{dx}$$

Carl Friedrich Gauss

(1777-1855)

A parenthesis:
An example of the genius of Gauss

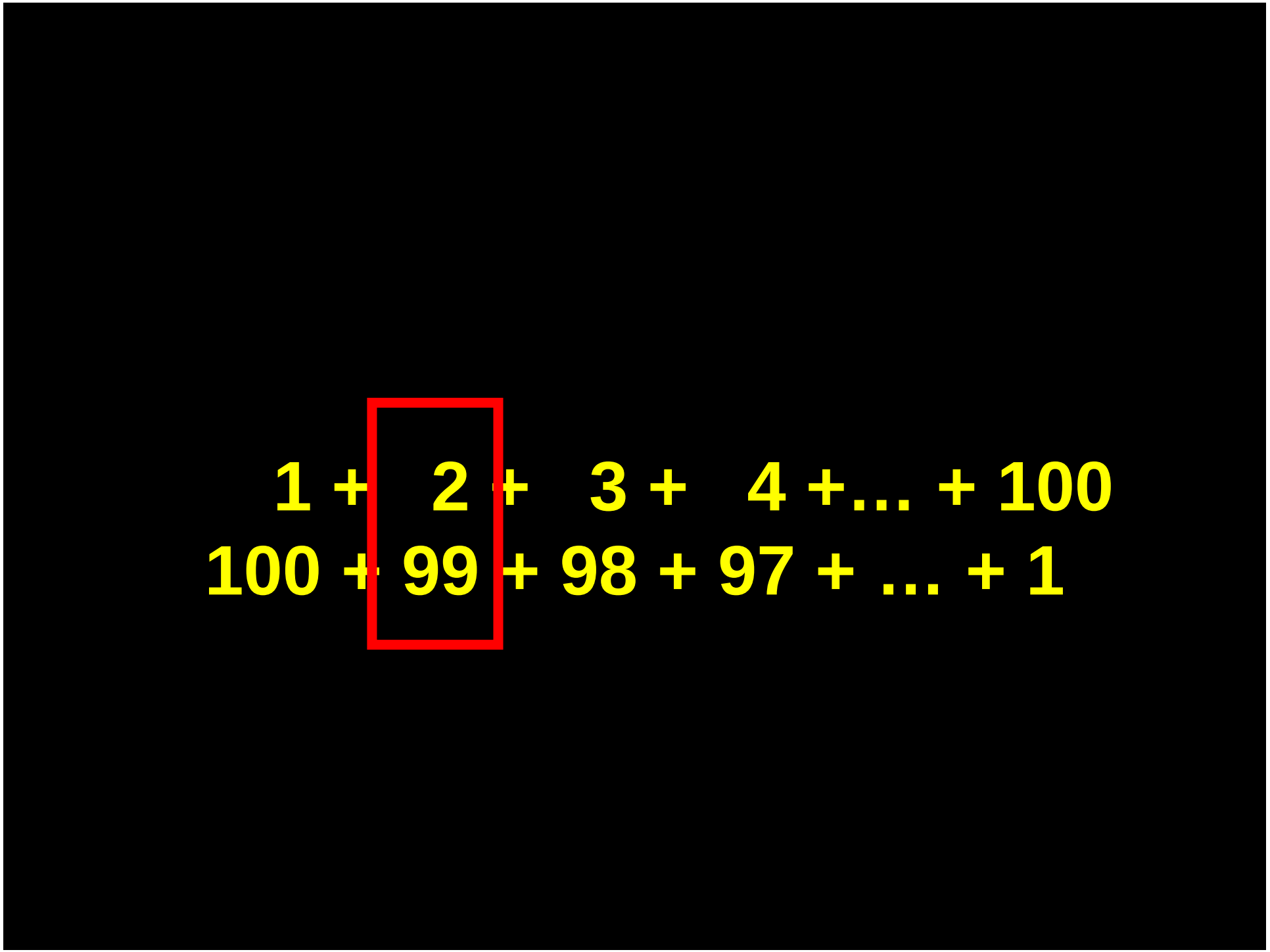
$$1 + 2 + 3 + 4 + 5 + \ldots + 100 = ?$$

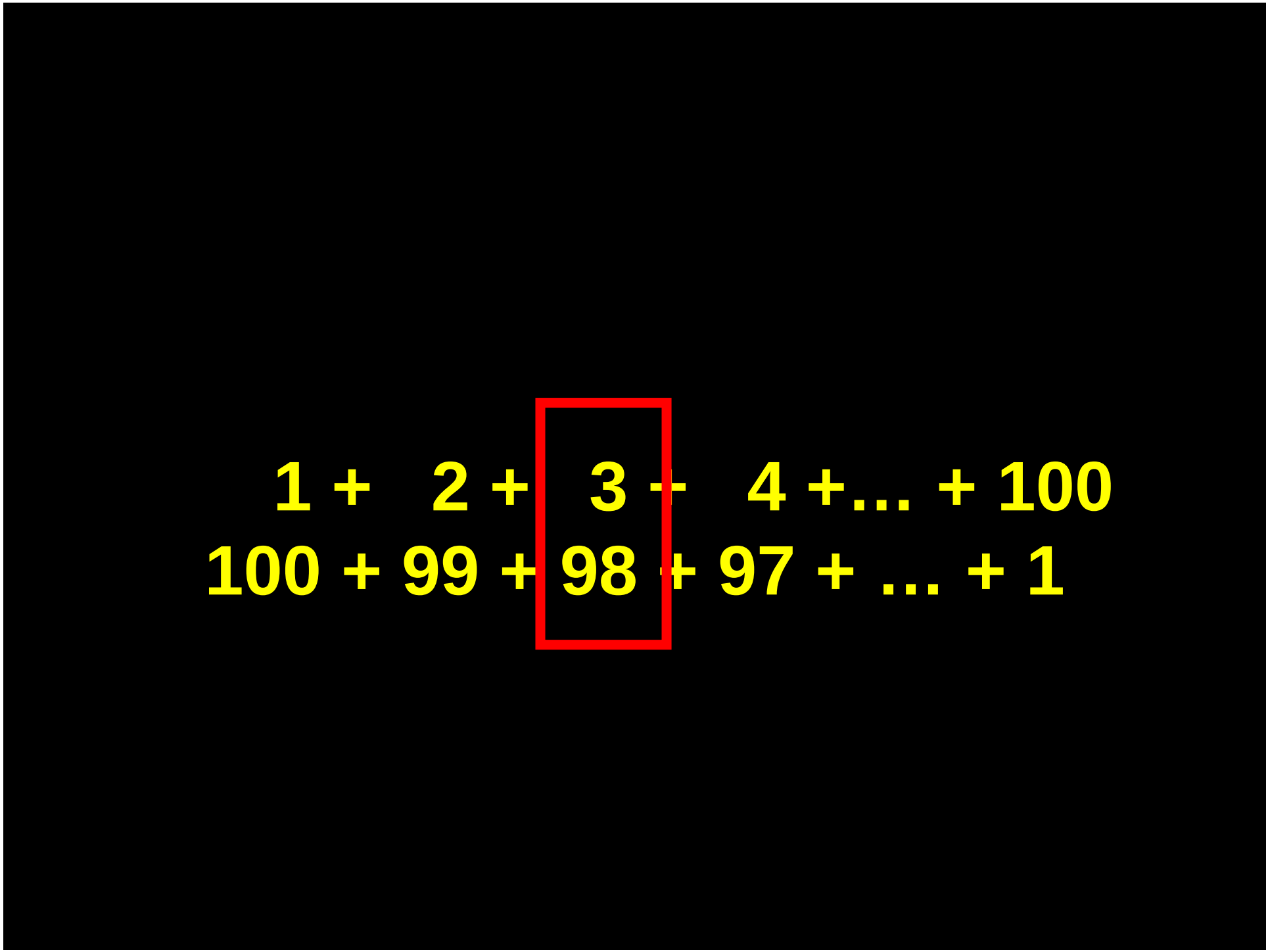
5050

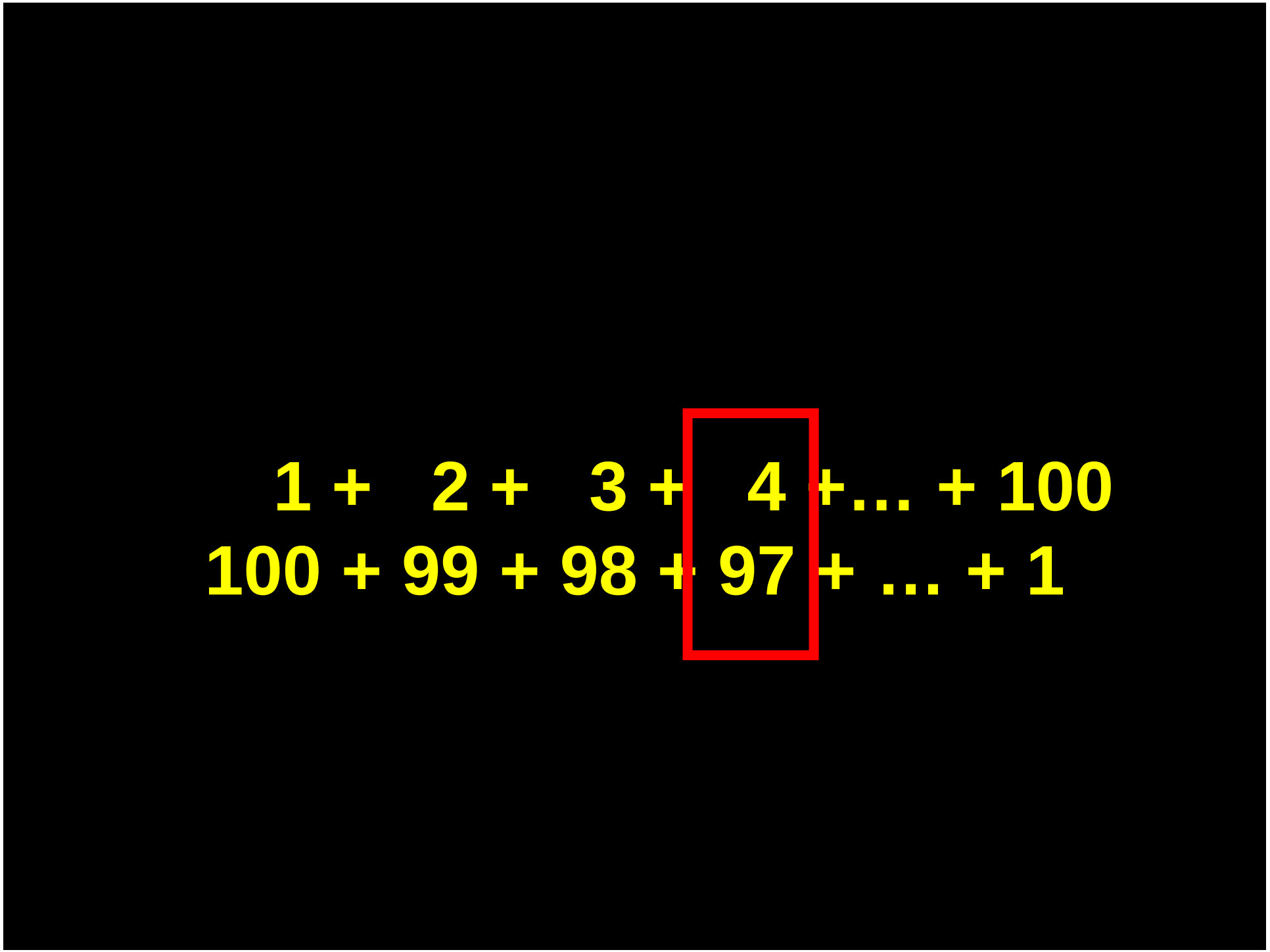
$$1 + 2 + 3 + 4 + \ldots + 100$$

$$100 + 99 + 98 + 97 + \ldots + 1$$

$$1 + 2 + 3 + 4 + \ldots + 100$$
$$100 + 99 + 98 + 97 + \ldots + 1$$


1 + 2 + 3 + 4 +… + 100
100 + 99 + 98 + 97 + … + 1


1 + 2 + 3 + 4 +… + 100
100 + 99 + 98 + 97 + … + 1


1 + 2 + 3 + 4 +... + 100
100 + 99 + 98 + 97 + ... + 1

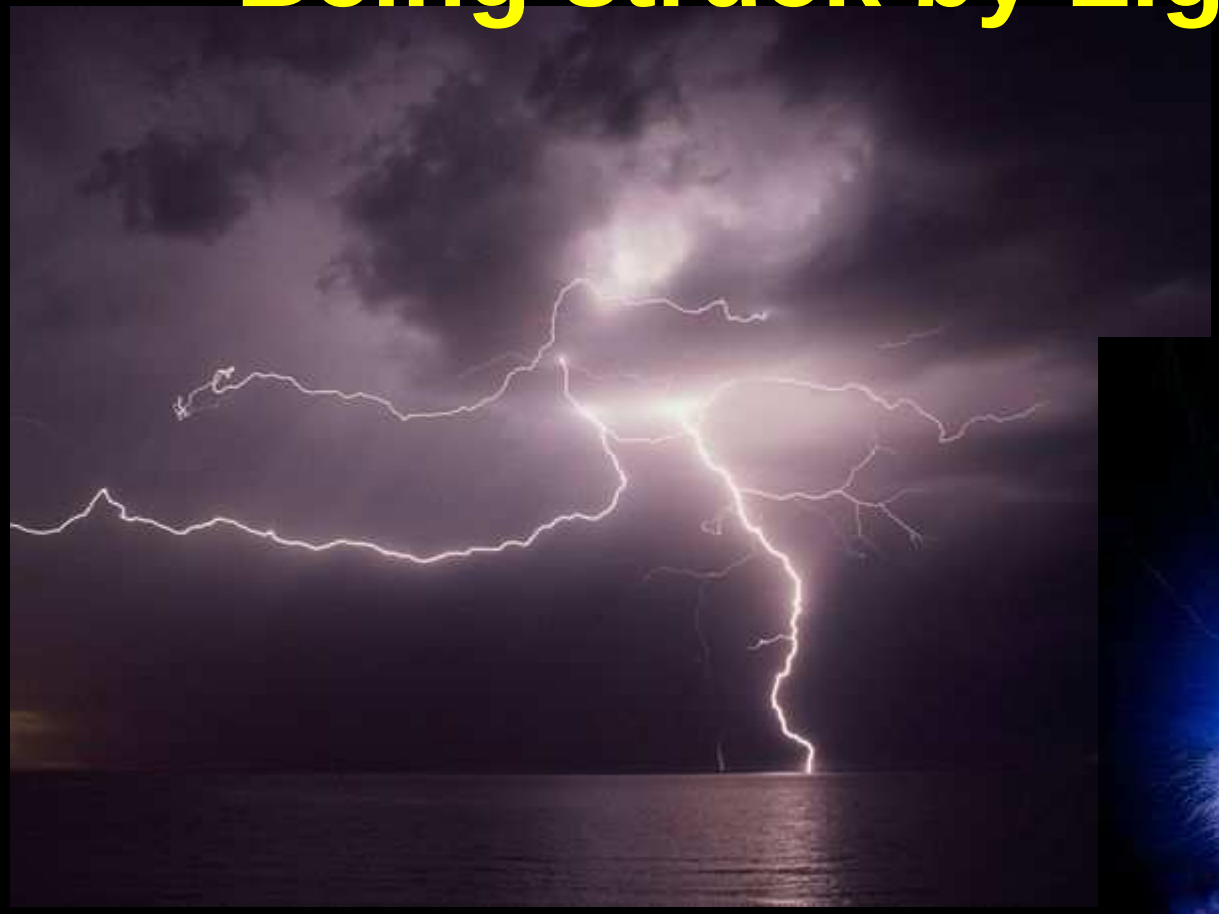
$$101 \times 100 \times \tfrac{1}{2} = 5050$$

$$1+ 2+ \ldots+n = (1+n) \times (n/2)$$

He was six years old!

Let’s look at an example…

Example:
Being struck by Lightening



US Data:

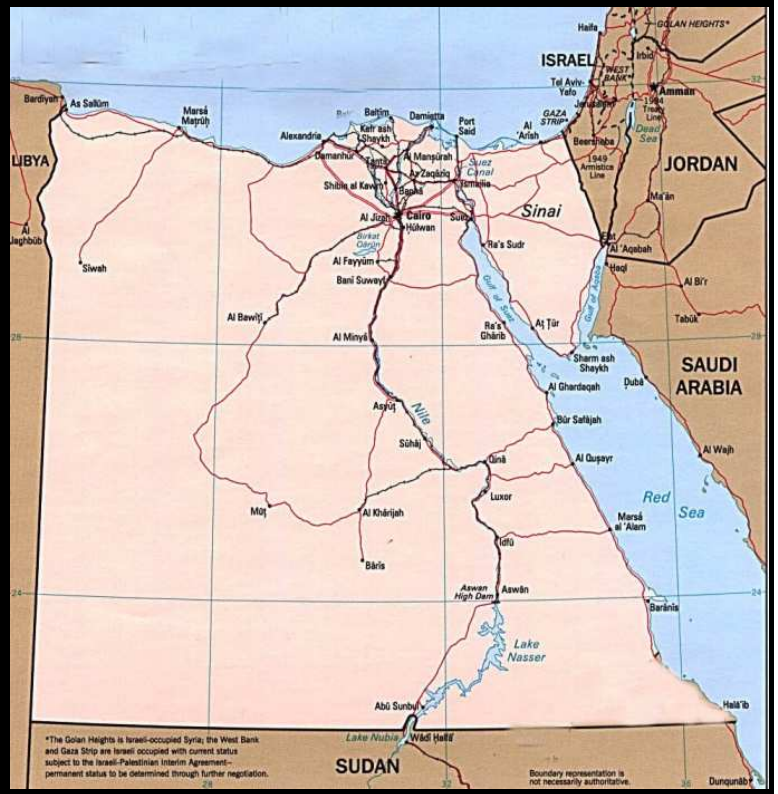
- **USA Population:**
  - **1961 183.7 Million**
  - **1999 272.7 Million**
  - **Average over the 38 year period: 228 Million**
- **Average deaths by being struck by lightening: 89 per year for the 38 years**
- **Average probability of dying by being struck by lightening: 1 in 2.5 million**

So I have a 1 in 2.5 million chance of being struck by lightening…

Is that correct?

Why?

Differs where you are:



In USA or Egypt

Differs where you are:

In Open country or in the City

Differs by time of year: e.g. for 1996

- In May -- 405 lightning strokes were recorded.
- In June -- 15,750
- In July -- 56,049
- In August -- 32,196 lightning strokes were recorded.
- In September-- 7,300
- In October -- 1,072 in October
- In November only 90 lightning strokes were recorded.

Remember:

You must be very careful how you generalize from any particular data set…

Lets think about some other probability problems

Three Coins Problem

Three coins are tossed simultaneously

- What is the probability that all three coins will come up heads?
- What is the probability of obtaining a head and two tails?

Answer

- Probability of getting 3 heads : 1/8
- i.e. p(3h) = 0.125

- Probability of 1 head and 2 tails : 3/8
- i.e. p(1h2t) =0.375

Three coins problem: Solution

- List all possible outcomes (call that A).
- Then ask: In how many ways can three heads appear? (call that B)
- Probability of that outcome is B/A
- Likewise: What is the probability of obtaining a head and two tails?
- Ask In how many ways can a head and two tails appear? (call that C)
- Probability of that outcome is C/A

Three coins solution (cont'd)

- So : List all possible outcomes A = 8

  hhh, thh, hth, hht, tth, tht, htt, ttt
- Only one possible way in which we get 3 heads. So B=1
- So the probability that all three coins will come up heads is B/A = 1/8
- In how many ways can a head and two tails appear? So C=3
- So the probability of obtaining a head and two tails is C/A = 3/8

The Birthday problem

What is the probability that at least two persons here where born on the same date?

How many people to get a match of two who have the same birthday?

The Birthday Problem or The Birthday Paradox

- **Question:** What is the probability that, in a set of n randomly chosen people, some pair of them will have the same birthday.
- Clearly, the probability reaches 100% when the number of people reaches 366 (since there are 365 possible birthdays, excluding February 29th).
- But what is the number required to have >50% probability?

Answer:

- **>50% probability is reached with just 23 people.**
- **And, 99% probability is reached with just 57 people.**
- **How come the numbers are so low?**

Explanation

- These conclusions are based on the assumption that each day of the year (except February 29) is equally probable for a birthday.
- The key point is that the birthday problem asks whether any of the people in a given group has a birthday matching any of the others — not one in particular.

Remember:
Any Birthday Matched With Any Other

- In a list of 23 people:
  - Comparing the birthday of the first person on the list to the others allows 22 chances for a matching birthday
  - The second person on the list to the others allows 21 chances for a matching birthday,
  - The third person has 20 chances, and so on.
  - Hence total chances are: 22+21+20+....+1 = 253),

So now let's calculate the probabilities:

- In a group of 23 people there are 253 possible pairs (combinations of pairing possible)
- Assume that the events of having a match are independent
- When events are independent of each other, the probability of all of the events occurring is equal to a product of the probabilities of each of the events occurring.

To simplify

- Lets calculate the probability of NOT having a match p(NM)
- The probability of having a match p(M) is complementary
- Therefore : p(M) = 1-p(NM)
- Calculating p(NM) for 23 people should =< 50%

So let’s see…

Consider each “Non Match” an independent Event

- For **Event 1**, the first person, there are no previously analyzed people. Therefore, the probability, P(NM1), that person number 1 does not share his/her birthday with previously analyzed people is **1**, or **100%.**
- Ignoring leap years for this analysis, the probability of 1 can also be written as **365/365**, for reasons that will become clear below.

Continuing

- the probability, P(NM2), that Person 2 has a different birthday than Person 1 is 364/365.
- This is because, if Person 2 was born on any of the other 364 days of the year, Persons 1 and 2 will not share the same birthday.
- P(NM3) = 363/365
- P(NM4) = 362/365 …. And so on…

Bringing this all together…

- P(23NM) = 343/365
- And these independent events all together …having No Match in the 23 persons … is equal to:
- P(NM) = 365/365 × 364/365 × 363/365 × 362/365 × ... × 343/365 = X
- P(NM) for 23 persons = 0.492703
- P(M) = 1- p(NM)= 1- 0.492703
- P(M) = 0.507297

So…

- The probability of having a match with someone's birthday in a group of :
- just 23 people is over 50% !!!
- For 57 people it is 99%
- There are variants to this problem statement. Let's discuss those

Can 23 really be enough to have >50% chance of a match?

Yes!
Here are some informal examples:

- Of the 73 male actors to win the Academy Award for Best Actor, there are six pairs of actors who share the same birthday.
- Of the 67 actresses to win the Academy Award for Best Actress, there are three pairs of actresses who share the same birthday.
- Of the 61 directors to win the Academy Award for Best Director, there are five pairs of directors who share the same birthday.
- Of the 52 people to serve as Prime Minister of the United Kingdom, there are two pairs of men who share the same birthday.

Now, let’s test a variant…

Variant: Same birthday as you

- Now we want to find the probability q(n) that someone in a room of n other people has the same birthday as you.
- Note that in the birthday problem, neither of the two people is chosen in advance.
- Now, this is different we want to find the probability q(n) that someone in a room of n other people has the same birthday as you.

Same birthday as you (cont'd.)

- To find the probability q(n) that someone in a room of n other people has the same birthday as you.

- The general form of the equation is given by:

$$\mathbf{q}(n; d) = 1 - \left(\frac{d-1}{d}\right)^{\mathbf{n}}$$

- And for the same birthday as you (d=365):

$$\mathbf{q}(n) = 1 - \left(\frac{365-1}{365}\right)^{\mathbf{n}}$$

Same birthday as you

- *So:* for the same birthday as you:
  - For n = 23 gives about 6.1%, which is less than 1 chance in 16.
  - You need at least 253 people in the room to have a greater than 50% chance that one person has the same birthday as you.

Same birthday as you

- *So:* for the same birthday as you:
  - For n = 23 gives about 6.1%, which is less than 1 chance in 16.
  - You need at least 253 people in the room to have a greater than 50% chance that one person has the same birthday as you.
- Note that this 253 number is significantly higher than 365/2 = 182.5. Why?

Same birthday as you

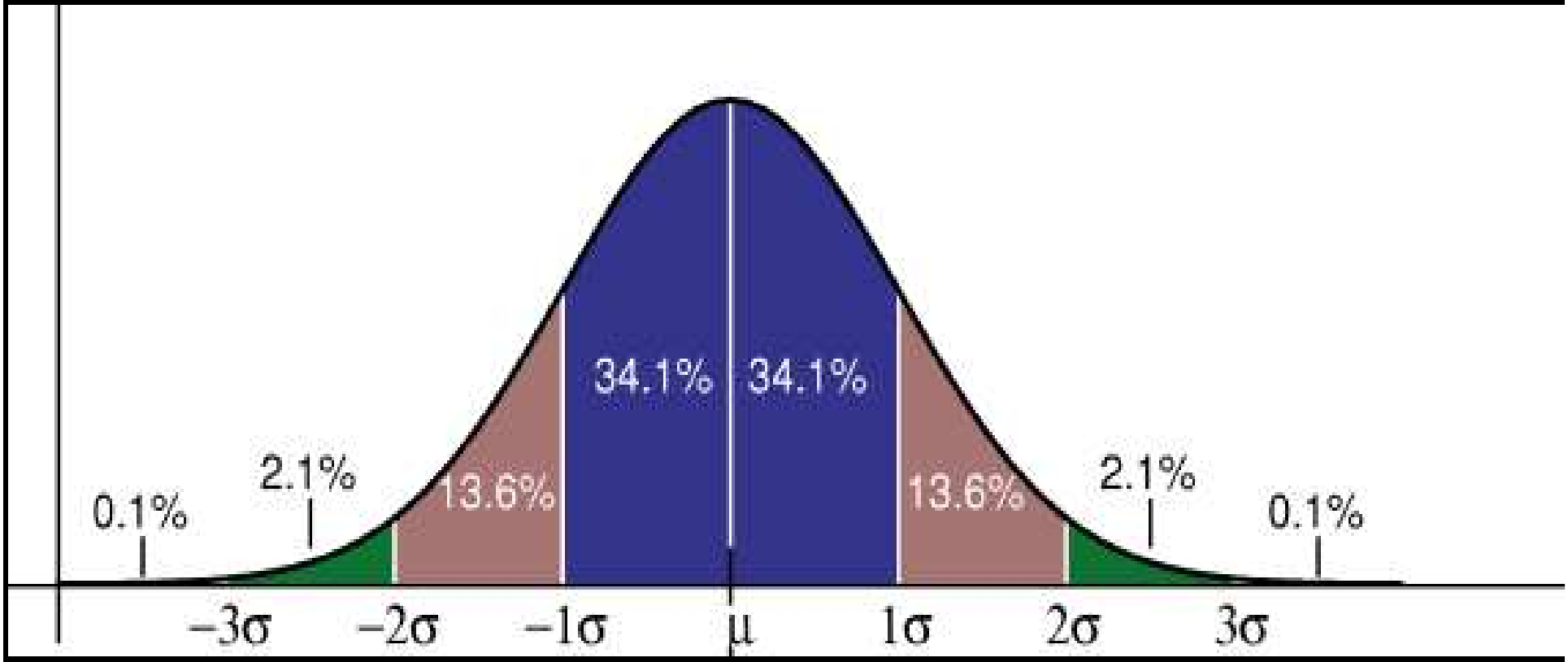
- *So:* for the same birthday as you:
  - For n = 23 gives about 6.1%, which is less than 1 chance in 16.
  - You need at least 253 people in the room to have a greater than 50% chance that one person has the same birthday as you.
- Note that this 253 number is significantly higher than 365/2 = 182.5. Why?
- The reason is that it is likely that there are some birthday matches among the other people in the room.

Probability is a science.
Its results can often be counter-intuitive.

FYI

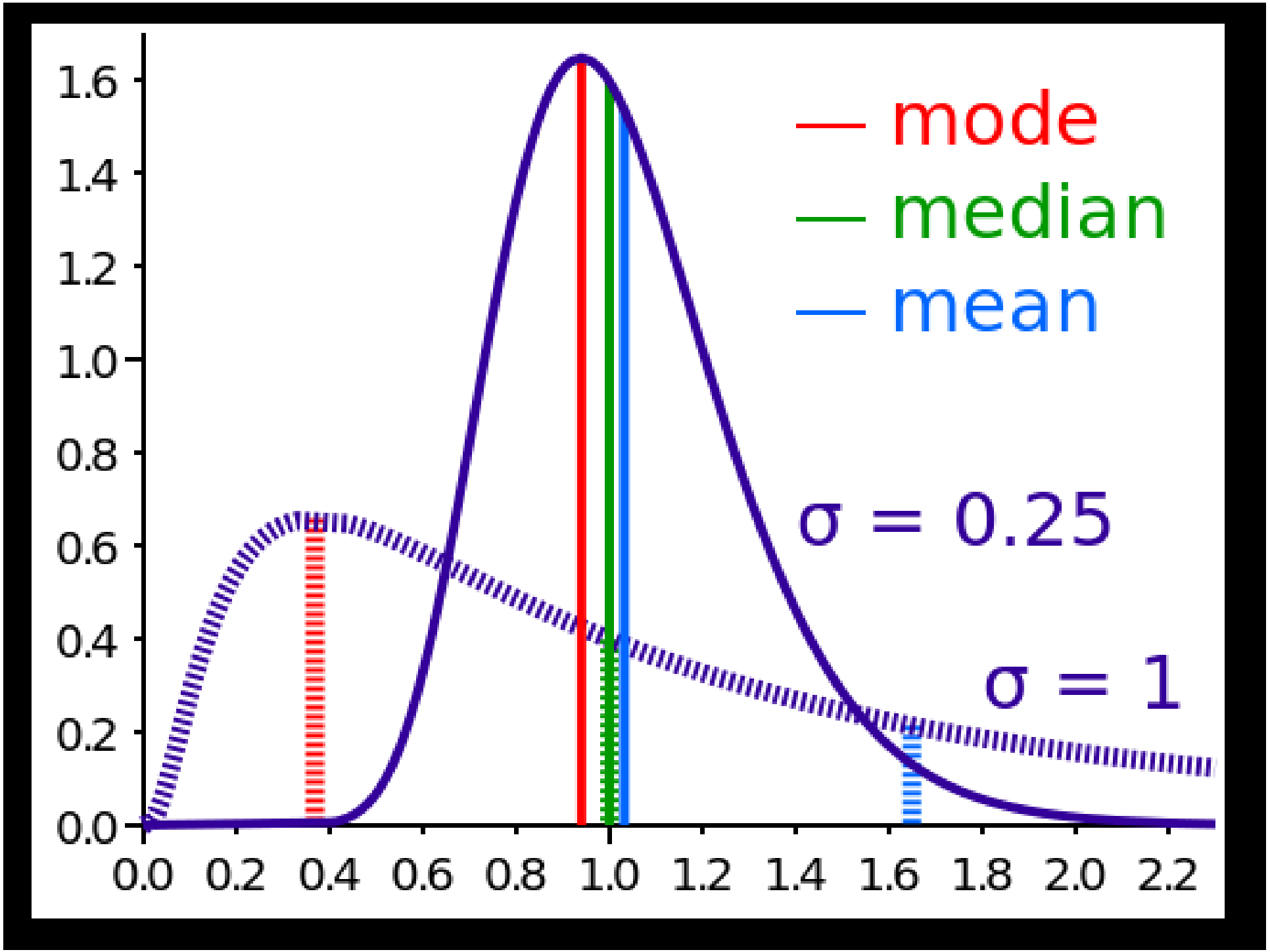
- **The probability of large number of observations of independent events will generally map out as a normal distribution (the bell curve, the Gaussian distribution).**
- **The hump or high point will always be the mode**
- **If and only if the curve is symmetrical, that will also be the mean and the median.**

The Gaussian Distribution



- **As the figure above illustrates, 68% of the values lie within 1 standard deviation of the mean; 95% lie within 2 standard deviations; and 99.7% lie within 3 standard deviations.**

If and only if the curve is symmetrical, that will also be the mean and the median.


mode
median
mean
σ = 0.25
σ = 1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2

Let’s review some things about probability

Rules of Probability

Probability

$$P(A) = \frac{\#\ \text{successes}}{\text{total outcomes}}$$

Joint occurrence of independent events: $P(AB) = P(A)P(B)$

Simultaneous occurrence: $P(A + B) = P(A) + P(B) - P(AB)$

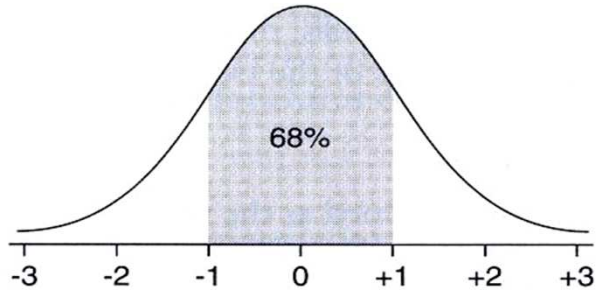
Conditional probability: $P(A|B) = \frac{P(AB)}{P(B)}$

Binomial distribution: $P(x) = \frac{n!}{x!(n-x)!}\pi^2(1-\pi)^{n-x}$ with $\mu = \pi$

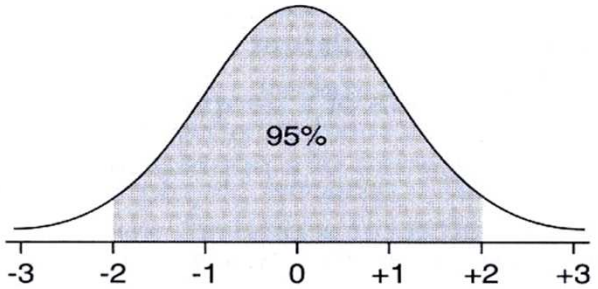
$$\text{and } \sigma = \sqrt{n\pi(1-\pi)}$$

The Gaussian, Normal or Bell Curve

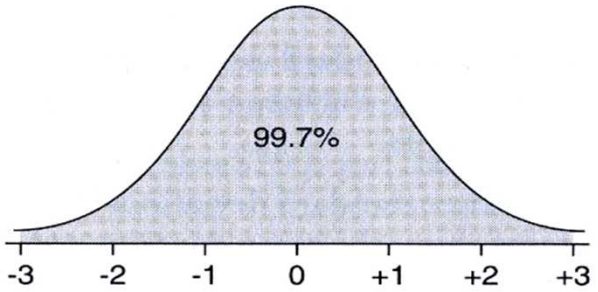
The interval ± σ from the mean contains 68% of the measurements.



The interval ± 2σ from the mean contains 95% of the measurements.



The interval ± 3σ from the mean contains 99.7% of the measurements.



This is a very useful curve and we will use it a lot in various analyses

Statistics, Standard Scores And Normalization

Statistics & Standard Score

- In statistics, a standard score indicates by how many standard deviations an observation or datum is above or below the mean.

- It is a dimensionless quantity.

Standardizing, Normalizing

- The Standard Score is derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation:

- This conversion process is called standardizing or normalizing.

The Standard Score

- The standard score of a raw score x is:

$$z = \frac{x - \mu}{\sigma}$$

- where:
  - μ is the mean of the population;
  - σ is the standard deviation of the population.

The quantity is in terms of the standard deviation of the population

- The quantity z represents the distance between the raw score and the population mean in units of the standard deviation.

- z is negative when the raw score is below the mean, positive when above.

You must know the population parameters, not sample statistics

- A key point is that calculating z requires the population mean and the population standard deviation, not the sample mean or sample deviation. It requires knowing the population parameters, not the statistics of a sample drawn from the population of interest.

Statistics & Standard Score

- Standard scores are also called z-values, z-scores, normal scores, and standardized variables.
- The use of "Z" is because the normal distribution is also known as the "Z distribution".

Z - Score

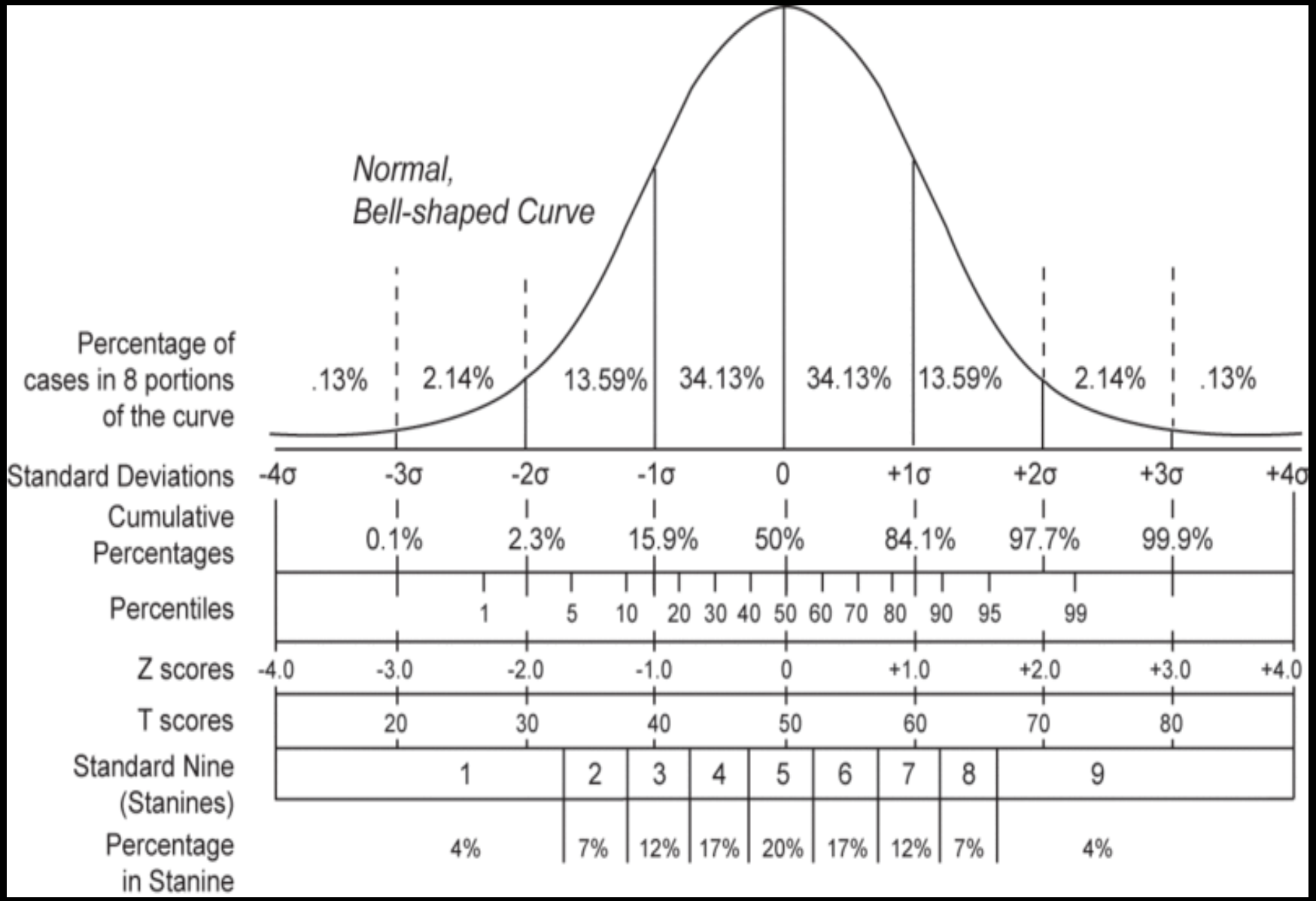
- **Z-scores** are most frequently used to compare a sample to a standard normal deviate (standard normal distribution, with $\mu = 0$ and $\sigma = 1$), though they can be defined without assumptions of normality.

From Z-Score to t-Statistic

- The z-score is only defined if one knows the population parameters, as in standardized testing; if one only has a sample set, then the analogous computation with sample mean and sample standard deviation yields the Student's t-statistic.

Anyway, the S, Z, t or F statistic is not important for now… just understand the underlying distribution..

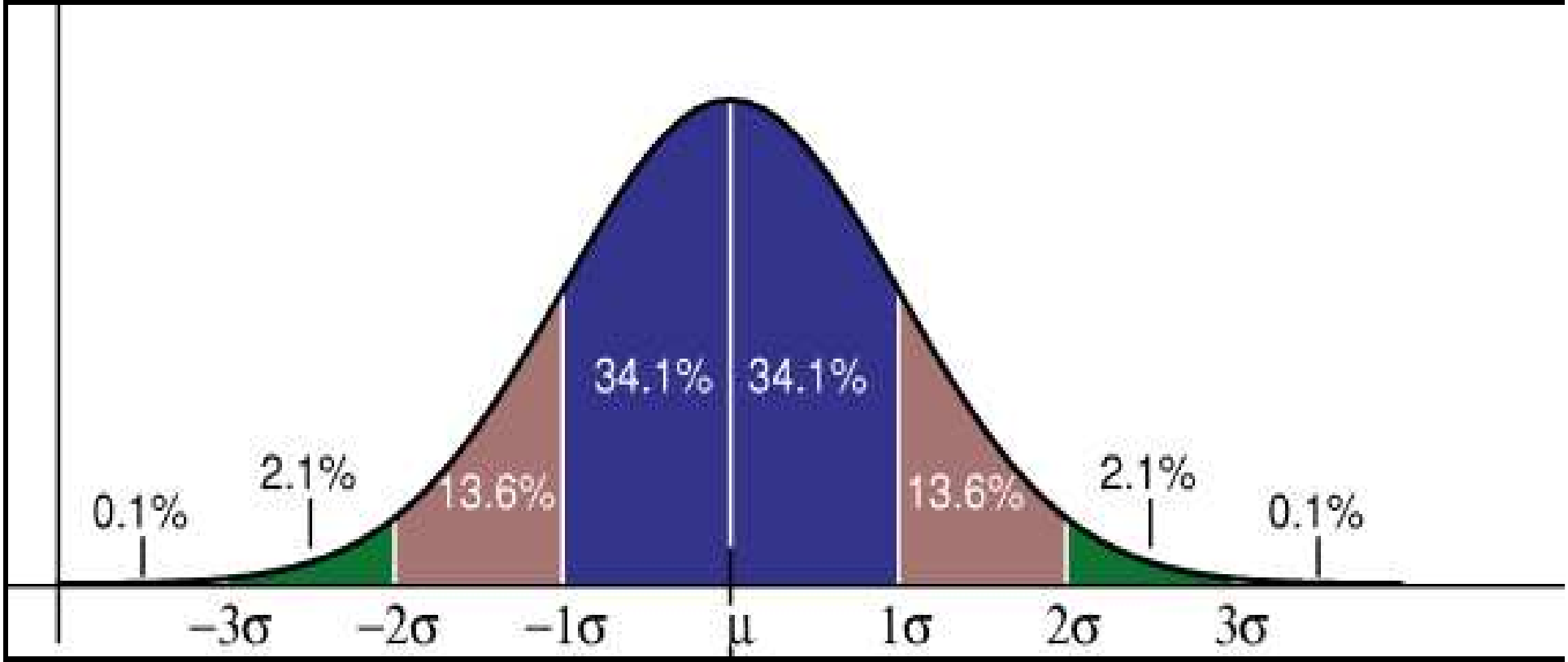
Back to the Normal Bell-shaped Curve



All this to show how much we will use the Gaussian Distribution, Normal Curve, bell Curve, Z-curve… Whatever you call it…

It is at the heart of many of our quantitative analyses…

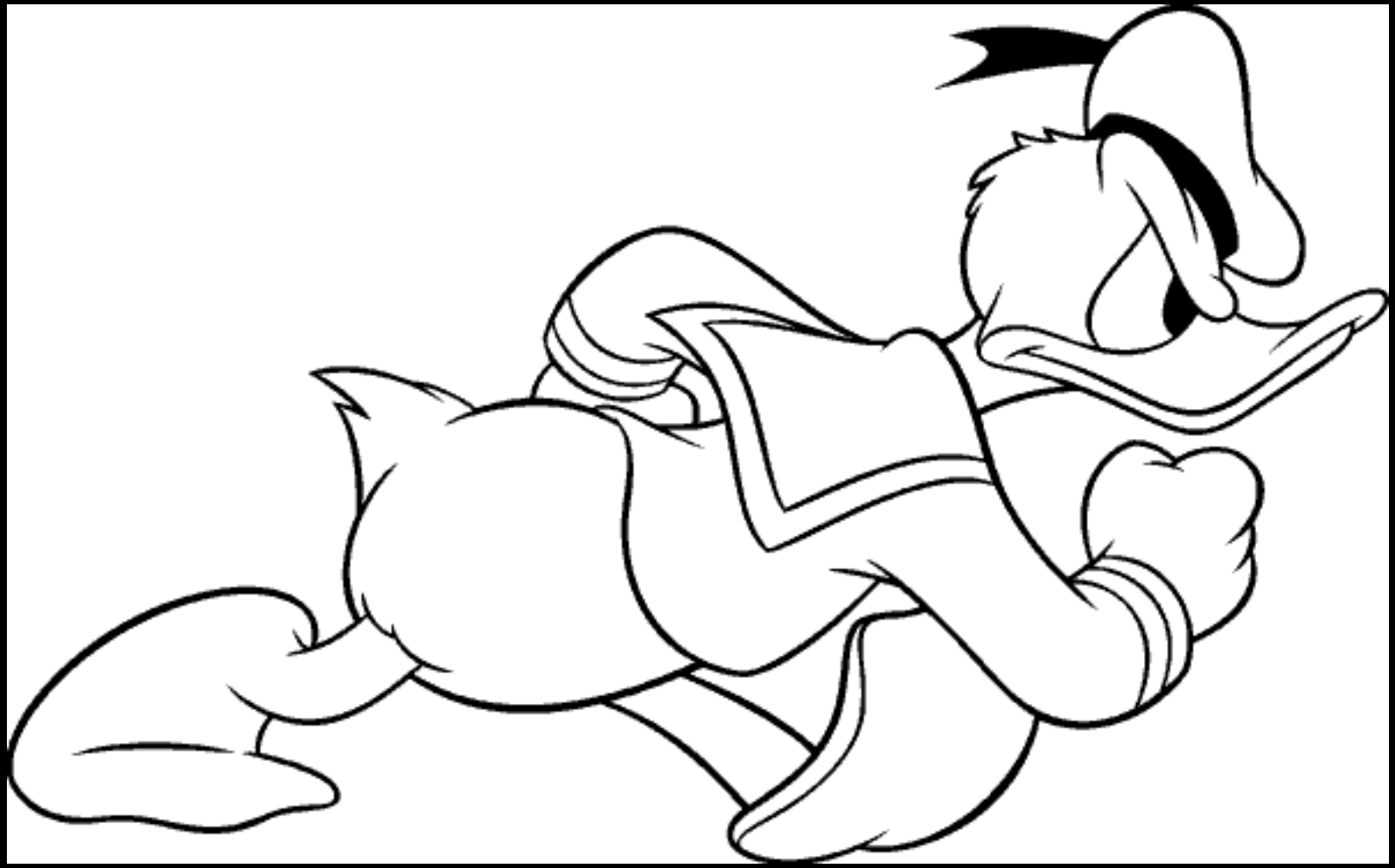
And it is easy to understand…



- **As the figure above illustrates, 68% of the values lie within 1 standard deviation of the mean; 95% lie within 2 standard deviations; and 99.7% lie within 3 standard deviations.**

Are there things you did not understand?


Stay Happy… Don’t Explode!



Don’t Get Angry… Ask

Make sure you understand before we move on…

Thank You