Statistics week 2: The Normal Distribution

Rabail Tahir

The Normal Distribution

- Normal Distribution is also known as Gaussian Distribution.
- It is a type of continuous probability distribution, because it is related to continuous variables for example height, weight, temperature etc.
- It describes the probability of any variable taking a particular range of values.
- It occurs very frequently in the real world. It is seen a lot in nature, science, engineering and many other fields.
- It provides a good approximation to reality.

The Normal Distribution

The idealised normal distribution has the following properties:

- It is symmetric.
- It is infinite in both directions.
- It has a single peak at the centre.
- It is continuous.
- 95% of values lie within approximately 2 **standard deviations** of the **mean**.
- 99% lie within approximately 3 standard deviations of the mean.

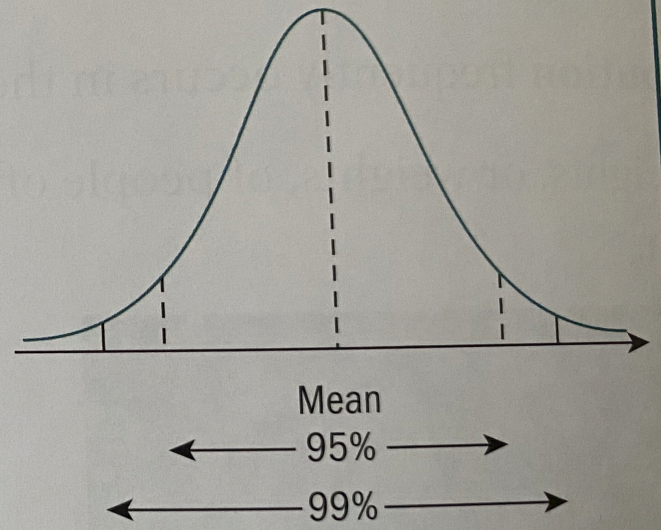


Image taken from “Complete probability & Statistics , James Nicholson”

The Normal Distribution

- The Probability Density Function (PDF), that creates the Normal Distribution is given by the following formula:

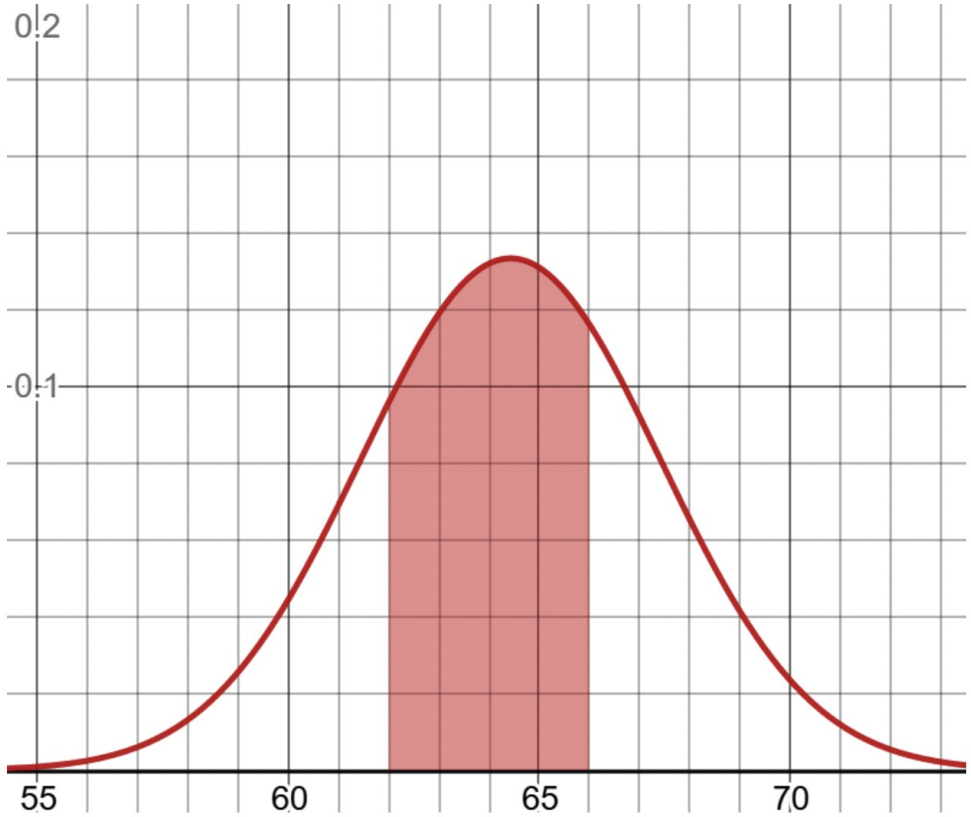
$$f(x) = \frac{1}{\sigma} * \sqrt{2\pi} * e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}^2\right)}$$

The Normal Distribution

- The Y-axis in a Normal distribution curve, does not represent probability, it represents the likelihood of ***Data.***
- Hence to calculate the probability, we need to look at the given range and then calculate the area under the curve for that particular range.
- That is why we use, Cumulative Density Function

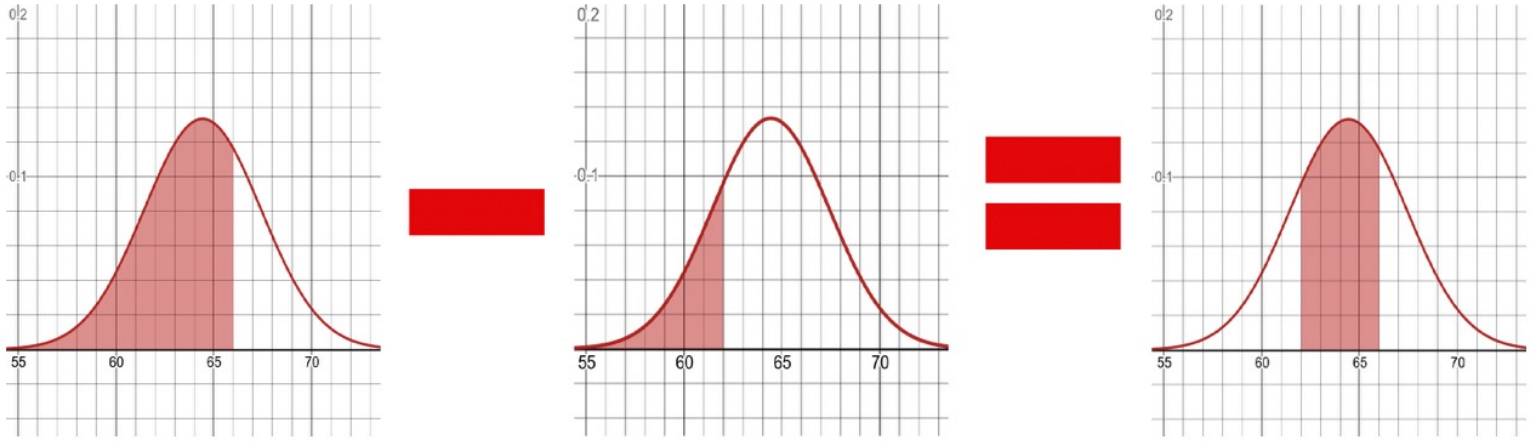
Example

- For example, we want to find the probability of a golden retriever weighing between 62 and 66 pounds



Example continued

- If we wanted to find the probability of observing a golden retriever between 62 and 66 pounds, we would calculate the area up to 66 and subtract the area up to 62.



Calculating it using Python

```
from scipy.stats import norm

mean = 64.43
std_dev = 2.99

x = norm.cdf(66, mean, std_dev) - norm.cdf(62, mean, std_dev)

print(x) # prints 0.4920450147062894
```

The inverse CDF

- It is used to calculate the corresponding X-value of a given probability.
- For example, we want to find the weight that 95% of golden retrievers fall under.
- You can use ppf() function in Python.

```
from scipy.stats import norm

x = norm.ppf(.95, loc=64.43, scale=2.99)
print(x) # 69.3481123445849
```

Standardised Scores

- The Normal Distribution allows us to compare individuals in a normal population.
- But sometimes it becomes harder with different normal populations, for example we use a different criteria to judge a “tall man” than we use for a ”tall woman”.
- We rescale the Normal Distribution, and convert it to Standard Normal Distribution , where mean = 0 and Standard Deviation = 1.
- This makes it simpler to compare different distributions even if they have different means and variances.

Z-Score

- We use z-score to compare different Normal distributions.
- A *z*-score (or *standard score*) represents the number of standard deviations a given value x falls from the mean, μ.
- It is given as follows:

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

Z-score

Z-score Calculation:

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

- value being examined x
- population mean μ
- population standard deviation σ

Example 1

- In her exams, Alexandra scores 75 in History and 87 in Maths. For the year group as a whole, History has a mean score of 63 in the examination with a standard deviation of 8, while Maths has a mean of 69 with a standard deviation of 15. Compare Alexandra's performance in these 2 subjects.

For History z = (75-63)/ 8 = 1.5

For Maths z = (87 – 69) / 15 = 1.2

Since Alexandra's z score is higher in History than in Maths, so her performance is better in History than in Maths.

Example 2

- We have two homes from two different neighborhoods. Neighborhood A has a mean home value of $140,000 and standard deviation of $3,000. Neighborhood B has a mean home value of $800,000 and standard deviation of $10,000.
- Now we have two homes from each neighborhood. House A from neighborhood A is worth $150,000 and house B from neighborhood B is worth $815,000. Which home is more expensive relative to the average home in its neighborhood?

Solution

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$

$$z_A = \frac{150000 - 140000}{3000} = 3.\overline{333}$$

$$z_B = \frac{815000 - 800000}{10000} = 1.5$$

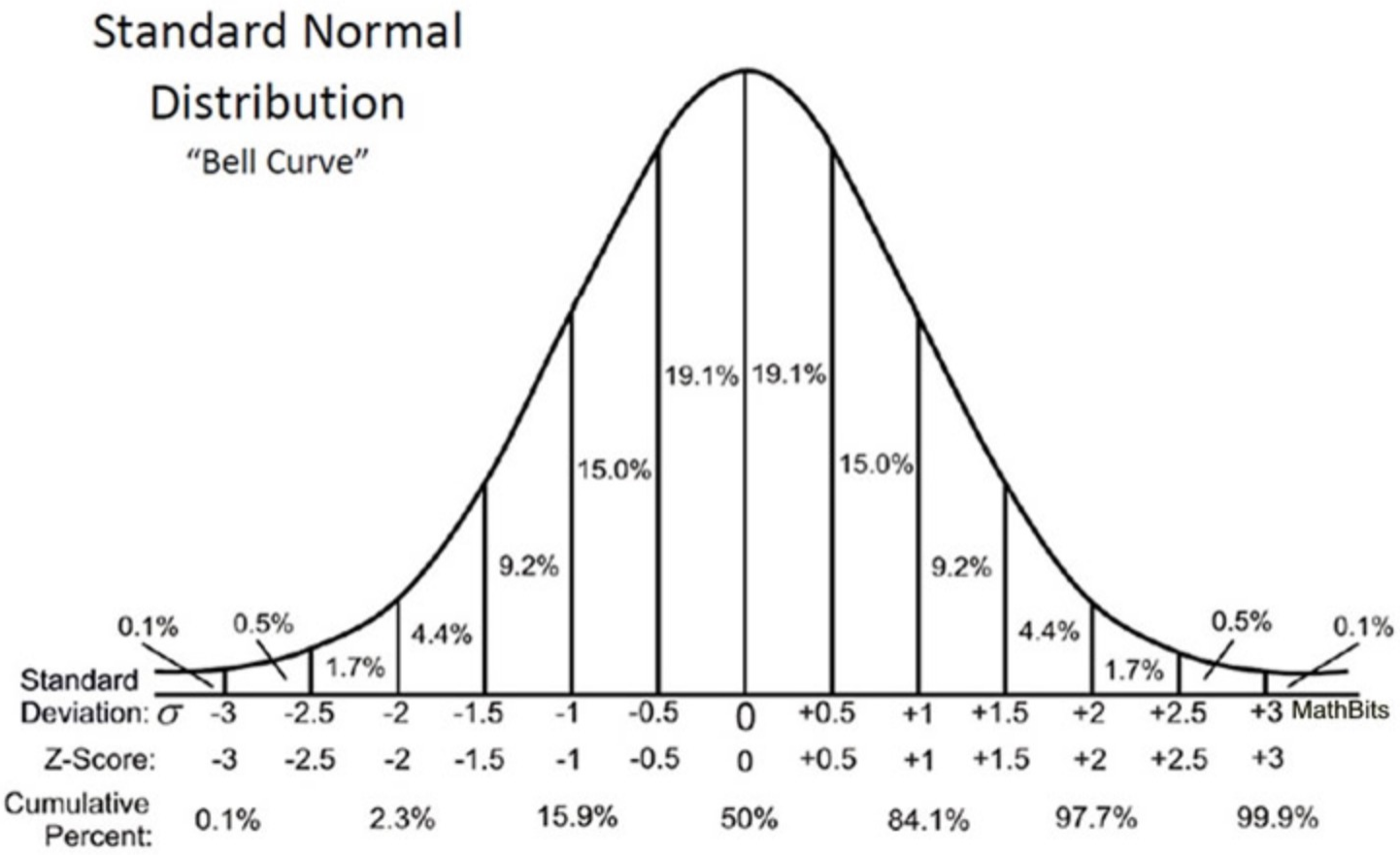
Standard normal distribution

- “There are many different normal distributions, with each one depending on two parameters: the population mean, μ, and the population standard deviation, σ. Rather than performing computations on each new set of parameters for a variety of normal curves, it is easier to work in reference to the "simplest case" of the normal curves, called the standard normal distribution.”

- For standard normal distribution, mean is 0 and standard deviation is 1.
- In Standard Normal Distribution, X is distributed as a normal random variable with mean μ and variance σ^2
- Since all normal distributions are the same basic shape, we only need to have probabilities for one particular case to allow us to calculate probabilities for all cases.
- We use a Z-table or a standard normal table to calculate Z-scores and Probability percentages.

Understanding z-scores https://mathbitsnotebook.com/Algebra2/Statistics/STzScores.html

Standard Normal Distribution



Z-scores are rounded to 2- decimal places

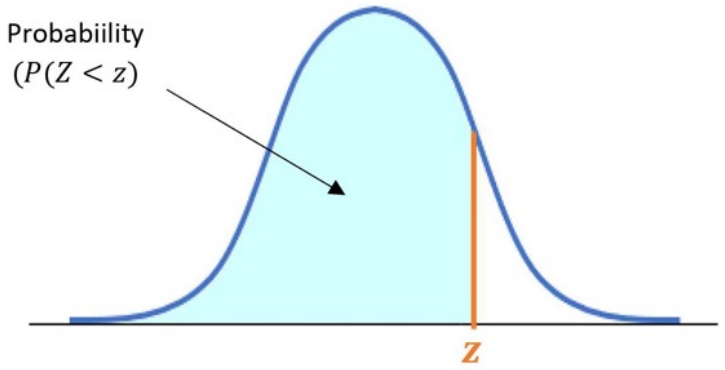
Z-score

- If $z=0$, the value equals the mean.
- If $z>0$, the value is to the right of the mean.
- If $z<0$, the value is to the left of the mean

Z-Table

- Standard normal distributions are very useful because we have a z-table (standard normal table) that gives us the area (i.e. probability) from the far left of the curve up to the z-value.
- Z-Table helps us to calculate the z-score or the area under the curve(probability,percentage)

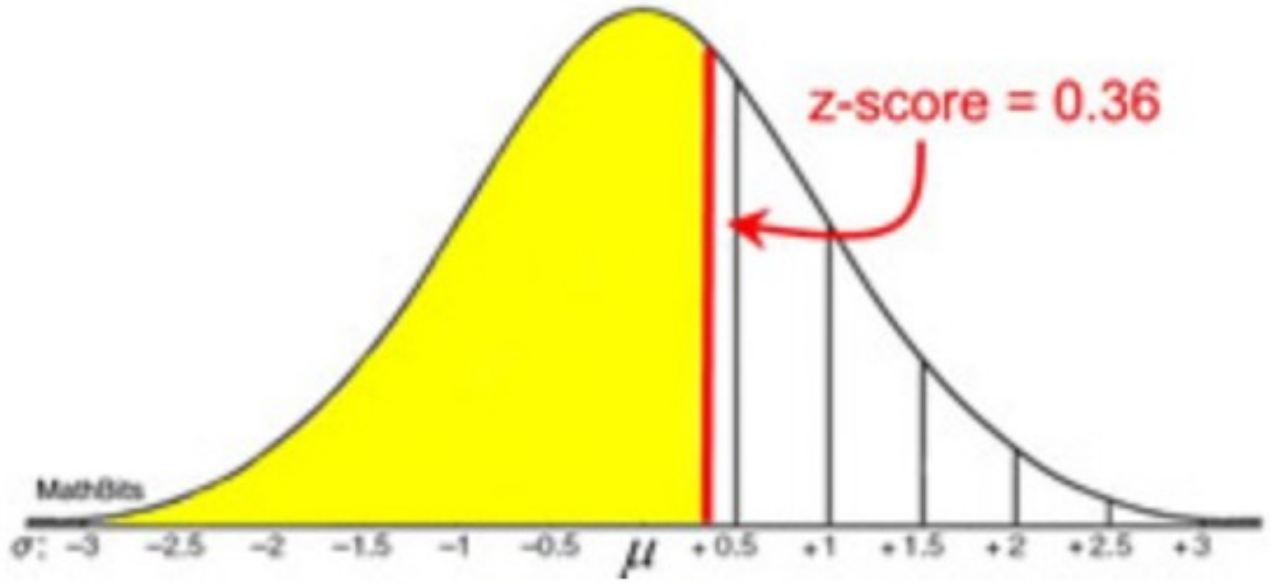
Z-Table



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5754
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7258	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7996	0.8023	0.8051	0.8079	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890

How to use Z-table
Example 1

- Find the probability that a variable has a *z*-score of less than 0.36.
- Solution:
  - First visualize the given question on a standard normal curve



- Second step is to use the Z-table.
- The intersection shows 0.6406. The probability is 64.06% (or the area percentage of the yellow region is 0.6406).

Positive z-scores: (Cumulative Areas from the Left)										
Z	**0.0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879

Example 2

- A normally distributed population of test scores has a mean of 80 and a standard deviation of 5.2.

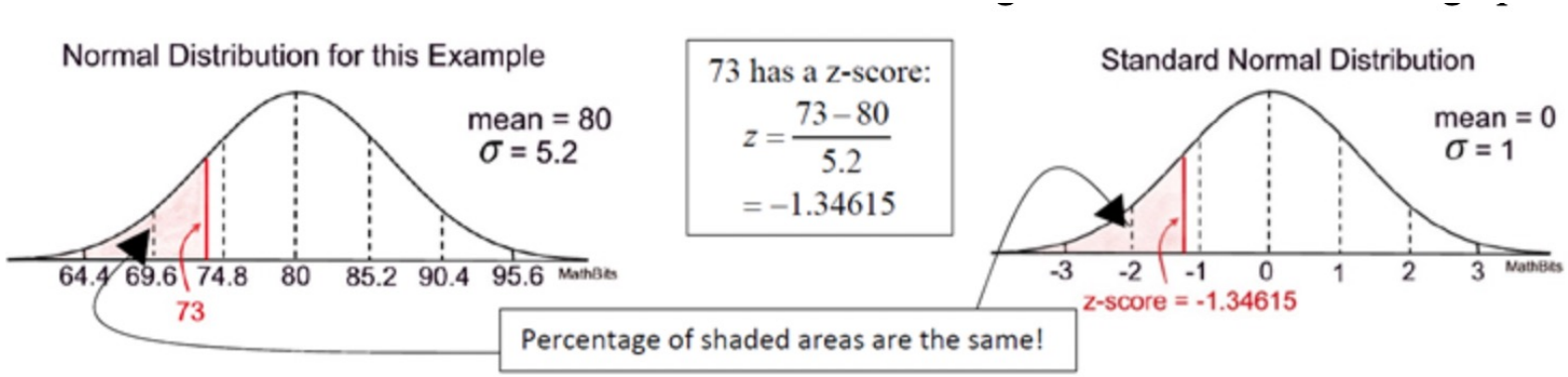
a) Find the percentage of scores that lies below 73.

b) Find the percentage of scores that lies between 82 and 86.

c) Find the percentage of scores that lies above 73.

a) Find the percentage of scores that lies below 73.

- First visualize the problem



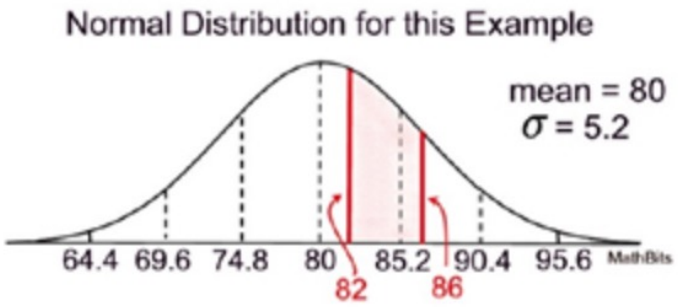
Now, that the z-score is known, the Z-score Chart will show the percentage of the shaded area to the left of this z-score of **-1.34.**

	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.0
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968

Percentage of scores below 73 is 9.01%.

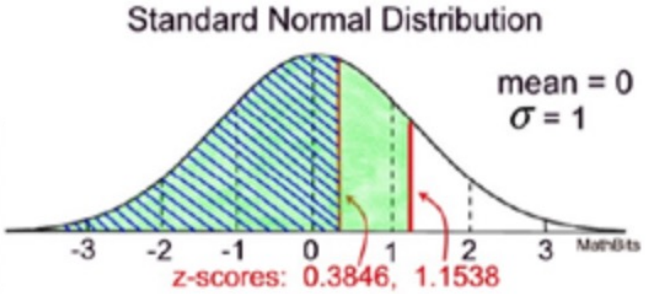
b) Find the percentage of scores that lies between 82 and 86.

To find the percentage of the area between two values, such as 82 and 86, find the z-scores for each value. From the Z-Score Table, find the area to the left of 86 and subtract the area to the left of 82.



86: z-score of 1.1538
82: z-score of 0.3846

Area to left of 86 = 0.8749
Area to left of 82 = 0.6480
Area between: .2269=22.69%



c) Find the percentage of scores that lies above 73.

To find the percentage of the area that lies "above" the z-score, take the total area under a normal curve (which is 1) and subtract the cumulative area to the left of the z-score.
In part a, 73 had a z-score of -1.34615 with a cumulative area to the left of 0.0901 or 9.01%.
The area to the right of this z-score will be 1 - 0.0901 = 0.9099 or 90.99%.

Any questions?