UPCSE Biology – Basic Statistic course

The normal distribution

Empirical vs theoretical distributions

- One of the points of statistics is to make confident prediction about the nature or behaviour of a population (e.g. snails, plants, breeding patterns of animals, cell growth or decay, breaking strength of materials, etc.)
- To do this we need to get probabilities about the nature or behaviour of our population of snails, plants, breeding patterns of animals, cell growth or decay, etc.

Empirical vs theoretical distributions

- Where do the probabilities comes from? By doing experiments, collecting data and normalising that data.
- **Example**: Suppose you measure the weight of 7000 hamsters. You decide to collect weight into groups of 1g, and you end up with the following type of data:

Table 1

Data range (in grams)	Frequency (i.e. number of data within each range)
$48 - 49^{-}$	y_1
$49 - 50^{-}$	y_2
$50 - 51^{-}$	y_3
.	.
.	.
$55 - 56^{-}$	y_8
$56 - 57^{-}$	y_9
$57 - 58$	y_{10}

Empirical vs theoretical distributions

- **Example**: Suppose you now group the weights into groups of 0.5g. You would end up with the following table:

Table 2

Data range (in cm)	Frequency (i.e. number of data within each range)
48 – 48.5^{-}	y_1
48.5 – 49^{-}	y_2
49 – 49.5^{-}	y_3
49.5 – 50^{-}	y_4
50 – 50.5^{-}	y_5
50.5 – 21^{-}	y_6
.	.
.	.
56.5 – 57^{-}	y_{18}
57 – 57.5^{-}	y_{19}
57.5 – 58	y_{20}

Empirical vs theoretical distributions

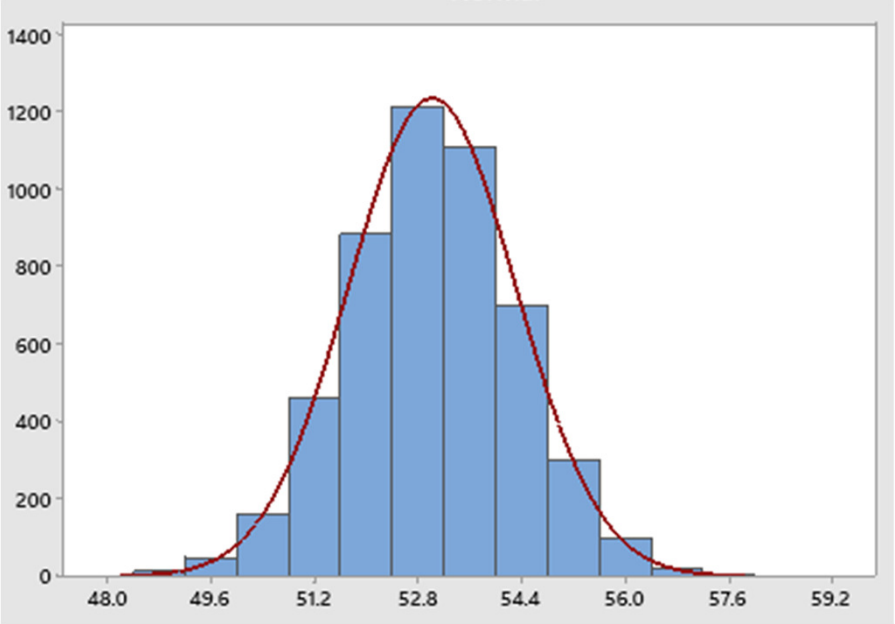
- **Example**: Now suppose you group the weights into groups of 0.1g. You would end up with the following table:
- Finally suppose you grouped weights by every 0.01g. You would end up with another more refined count for weights (not shown here)

Table 3

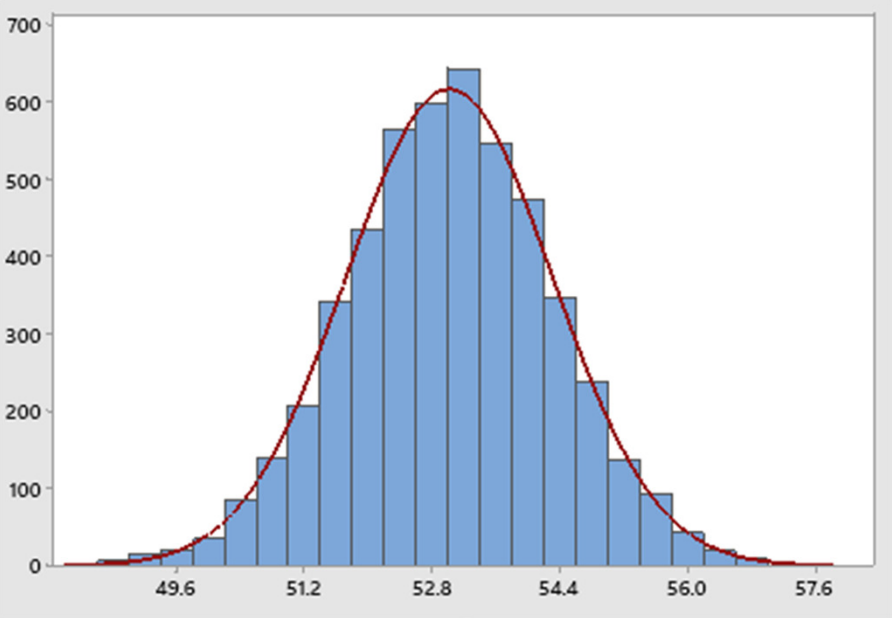
Data range (in cm)	Frequency (i.e. number of data within each range)
48 – 48.1$^-$	y_1
48.1 – 48.2$^-$	y_2
48.2 – 48.3$^-$	y_3
48.3 – 48.4$^-$	y_4
48.4 – 48.5$^-$	y_5
48.5 – 48.6$^-$	y_6
.	.
.	.
57.7 – 57.8$^-$	y_{98}
57.8 – 57.9$^-$	y_{99}
57.9 – 58	y_{100}

Empirical vs theoretical distributions

- Now plot a histogram for each table. You might get something looking as below (plotted using artificially generated data):



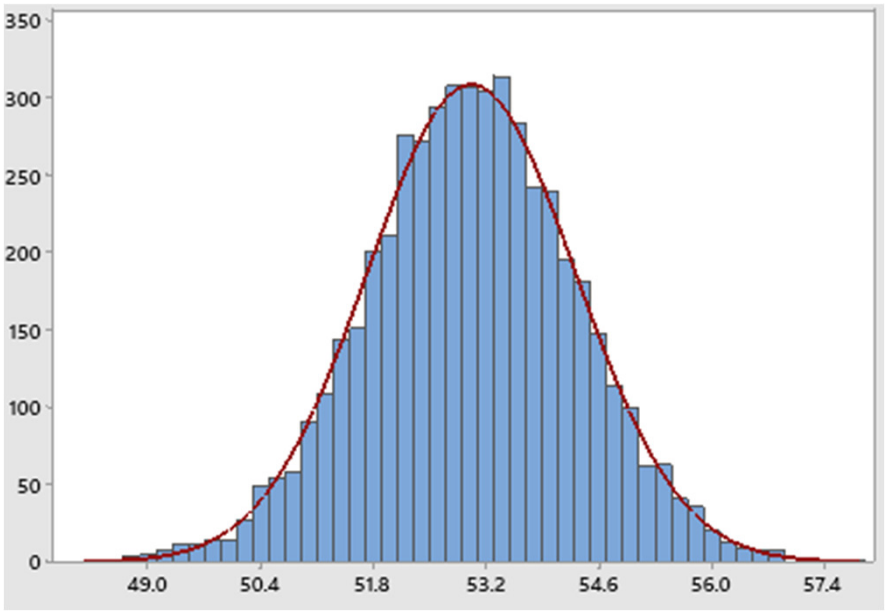
Histogram for table 1



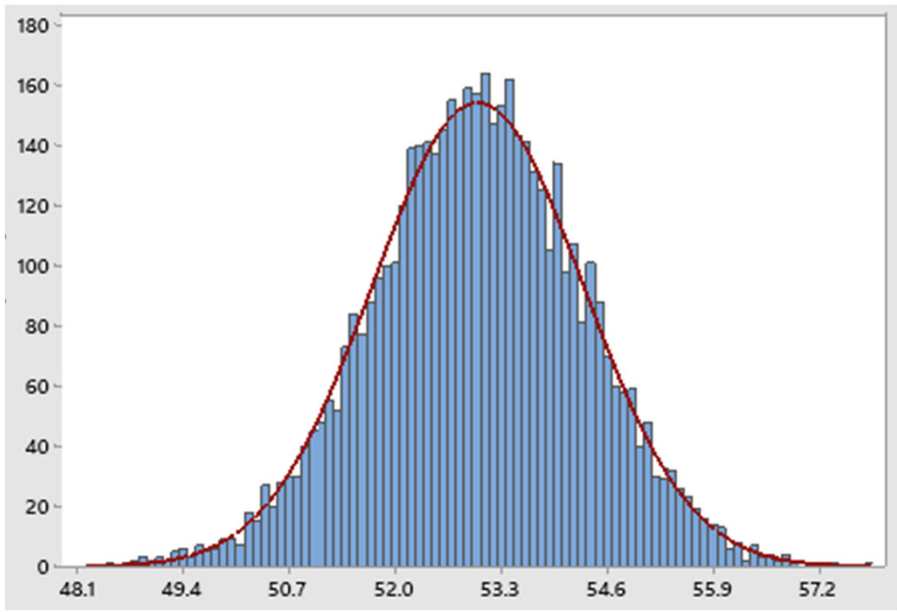
Histogram for table 2

Empirical vs theoretical distributions

- As we choose to plot using narrower and narrower rectangles, what do we see about how close the trend of the rectangles is compared to the red curve?



Histogram for table 3



Histogram for what would have been table 4

Empirical vs theoretical distributions

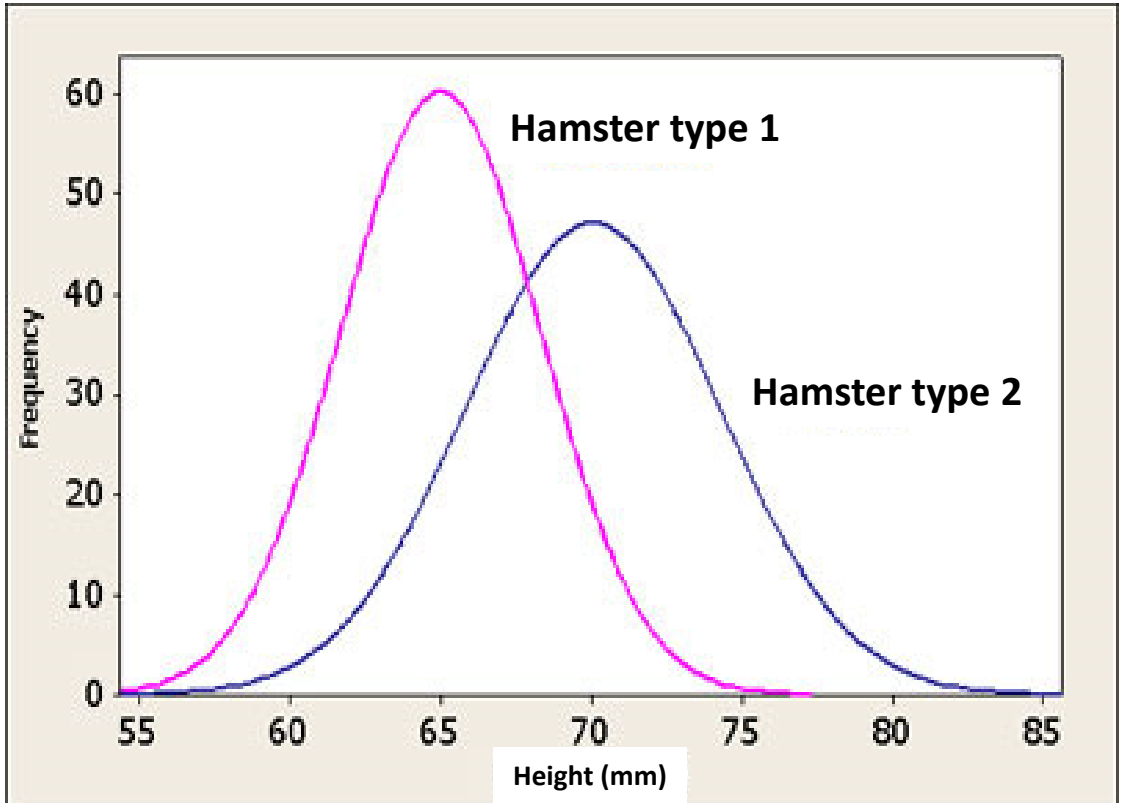
- Ultimately, if we used an infinite number of infinitely thin rectangles the tops of our rectangles would fit the red curve perfectly.
- Such a situation would represent the theoretically perfect distribution of data.
- So this red curve is the theoretical model we use to compare our experimental results against.

Empirical vs theoretical distributions

- Data shaped as shown in the histograms above is known as normally distributed data.
- The red curve is the ideal/perfect normal distribution of data.
- But different types of hamster have different weight ranges, these being: < 25g, 25g – 50g, 51g – 100g, 101g – 130g, > 130g.

Empirical vs theoretical distributions: long version

- So for two groups of hamsters we may have two different normal distributions:



Empirical vs theoretical distributions

- Notice that the heights and frequency counts are different for both types of hamsters.
- This means that we need separate normal distributions for each type or breed of hamster.
- And what about other animals!? We would need separate a normal distributions for each type or breed of each type of animal!

Empirical vs theoretical distributions

- This is clearly impossible.
- So we standardise all data to one and the same form.
- We do this in two ways. **Firstly** we convert the frequency, or count of the number of data, into a percentage.
- For example, in the histogram for table 1 the frequency count of the largest rectangle is about 1200 hamsters.
- Therefore there are 1200/7000*100=17.14% of hamster whose weights are in that range of weights.

Empirical vs theoretical distributions

- We repeat this calculation for all frequencies/counts (i.e. all rectangles).
- Note that the total count is 7000 hamsters, so the total percentage of hamsters should be 100%.
- **Secondly** we do a transformation. It is a simple piece of arithmetic given by

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

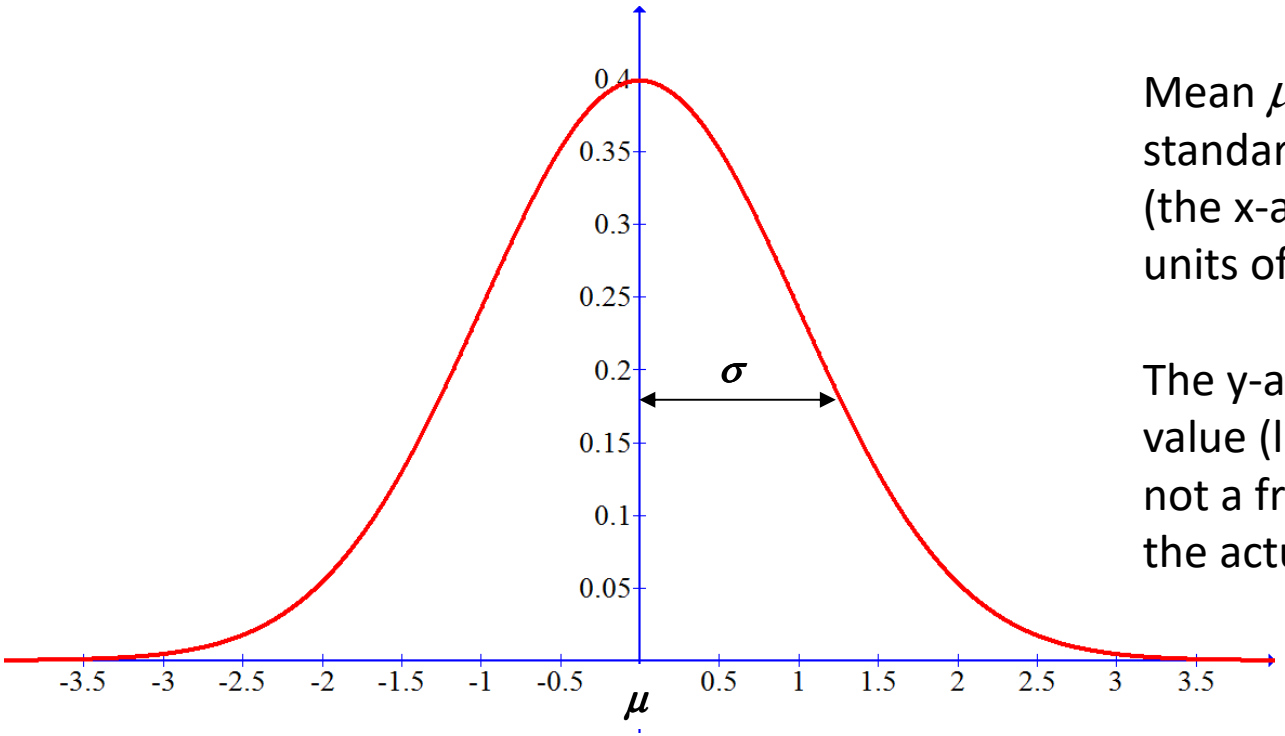
where x is any one of your data values, μ is the mean of your data, and σ is the standard deviation of your data.

Empirical vs theoretical distributions

- Converting to percentages and doing the transformation then converts all possible different normal distributions, based on raw data, into one and the same normal distribution based on percentages.
- Except that instead of using percentage values 0 – 100 we simply use values 0 – 1 (so that 53.7% would now be 0.537).
- These values relate to probabilities and the values we see in probability tables when doing significance testing.

Empirical vs theoretical distributions

- We now plot z instead of x. This then gives the graph below which is the theoretical model of the standard normal distribution.



Mean μ is now 0, and the standard deviation / spread σ (the x-axis) is now measured in units of 1, 2, 3, 4, …

The y-axis is now a probability value (like a percentage value) not a frequency or raw count of the actual data.

Specific distributions

- The normal distribution is a distribution of probabilities. It is a theoretical distribution which represent what would be the case if an experiment were conducted an "infinite" number of times, in a perfect world.

- In your biology course you will use two probability distributions:
  - t - distribution,
  - χ^2 (chi-squared) distribution.

Specific distributions

- These are theoretical distributions we use to model different patterns in the real world. They also represent what would be the case if an experiment were conducted an “infinite” number of times, in a perfect world.

Specific distributions

- Then, when we do an experiment :
  - we always compare our experimental data against the theoretical model
  - In other words, we compare the results found from the t or χ^2 formulae against *t*-distributions or χ^2 distributions.

Specific distributions

- Then, when we do an experiment :
  - So we compare our calculated t-values or χ^2-values (calculated from formulae) with the values in tables (or from Minitab).
  - From the comparison we decide to accept or reject a particular hypothesis we will have set-up beforehand.

Specific distributions

- When we do an experiment :
  - More specifically we base our decisions about whether something (like the difference in mean weight of two groups of turtles on different islands) is significant or not by how much our data differs from the theoretical model.

Specific distributions

Example

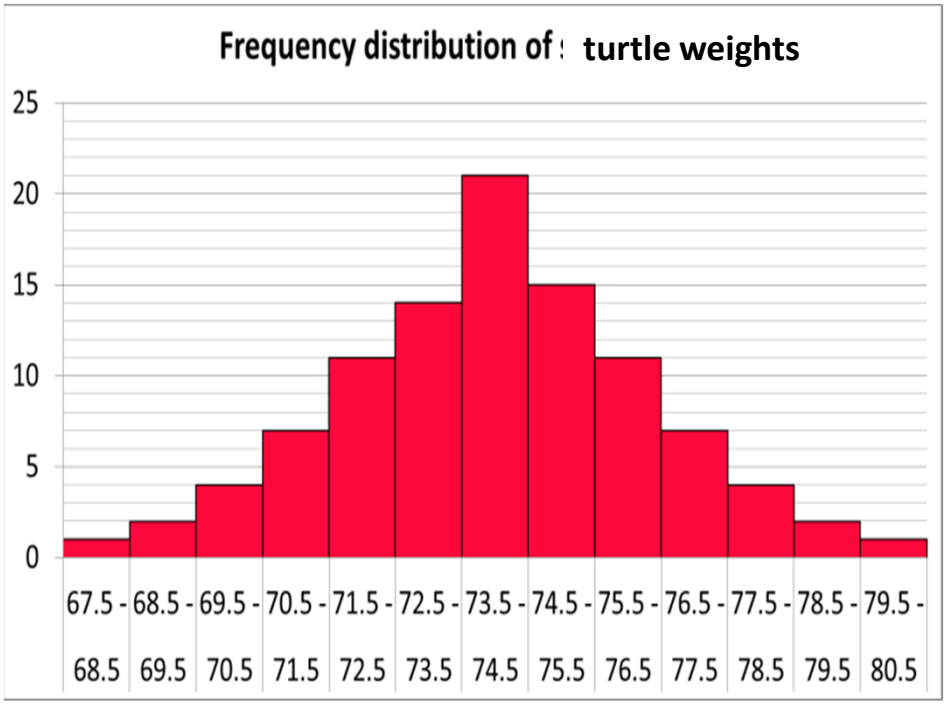
- For example, we do an experiment to study the average weight of turtles on an island of 2000 turtles.
- We take a sample of 200 turtles.
- We might then be interested in the probability of finding a turtle whose average weight lies between 7 kg and 7.3 kg.
- We would then do a calculation and compare the result against the theoretical distribution which matches the standard pattern of weight distributions.

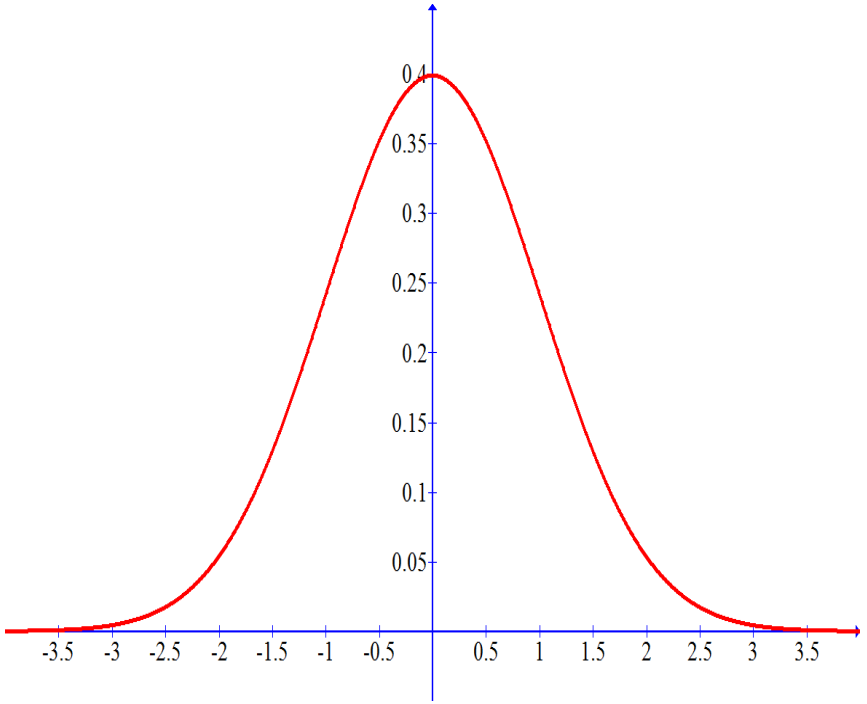
The normal distribution

- The normal distribution would be the appropriate theoretical distribution to use.
- The data we have collected classifies as *experimental data*.
- The normal distribution curve classifies as the *theoretical distribution* of data.

The normal distribution

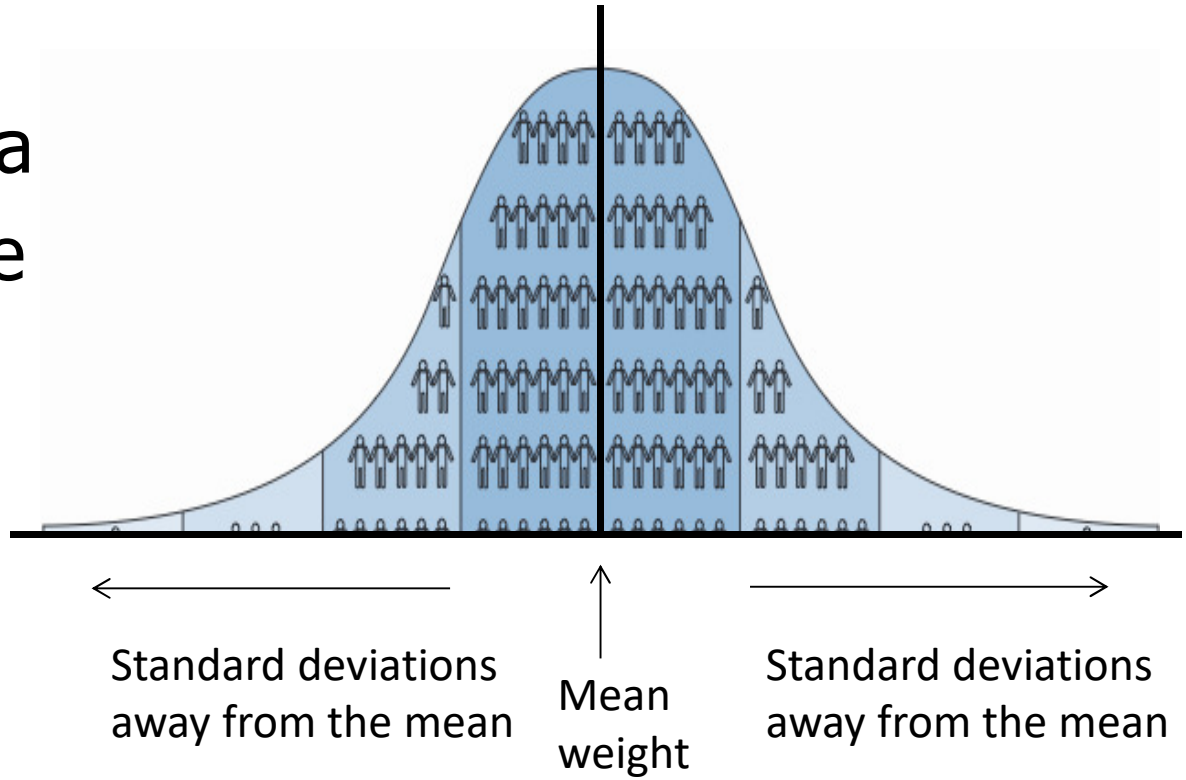
- Actual distribution of data on turtle weights.
- Theoretical distribution of all possible data on turtle weights that could ever be taken.





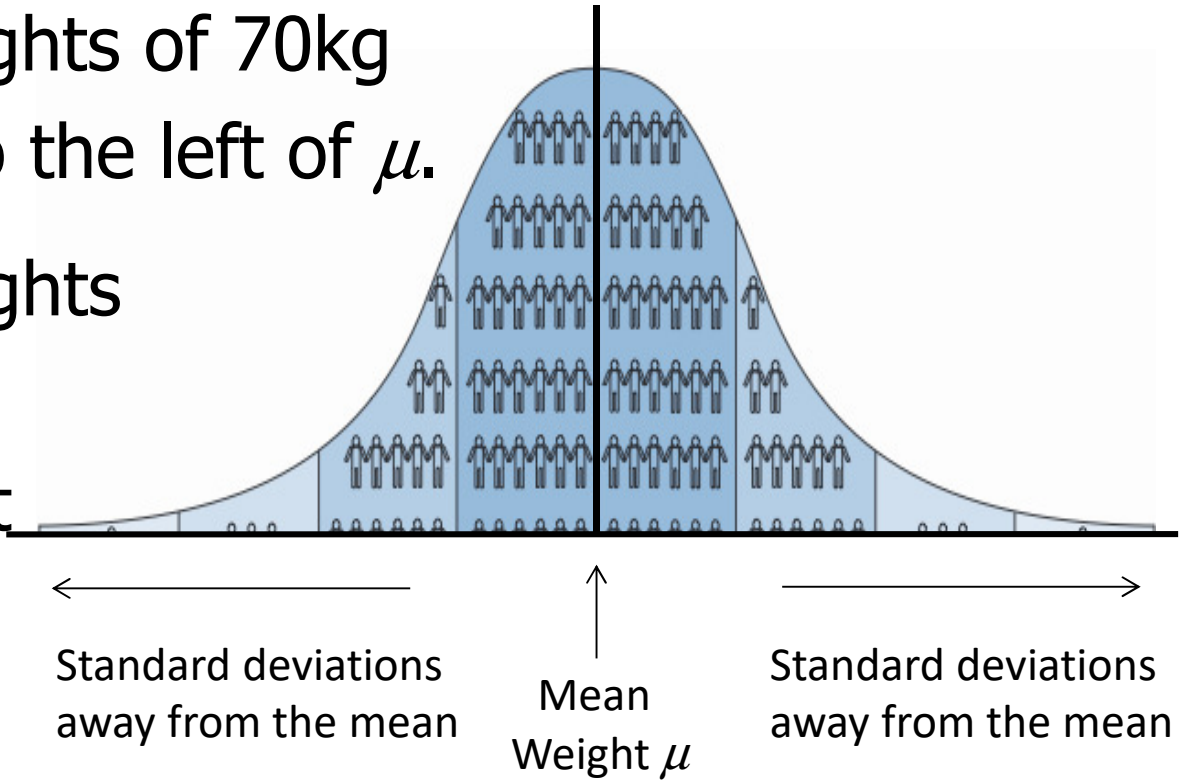
More on the normal distribution

- Normal distributions are based on means and SDs only.
- In other words, all you need to draw a normal curve is the mean of your data and the standard deviation of your data.



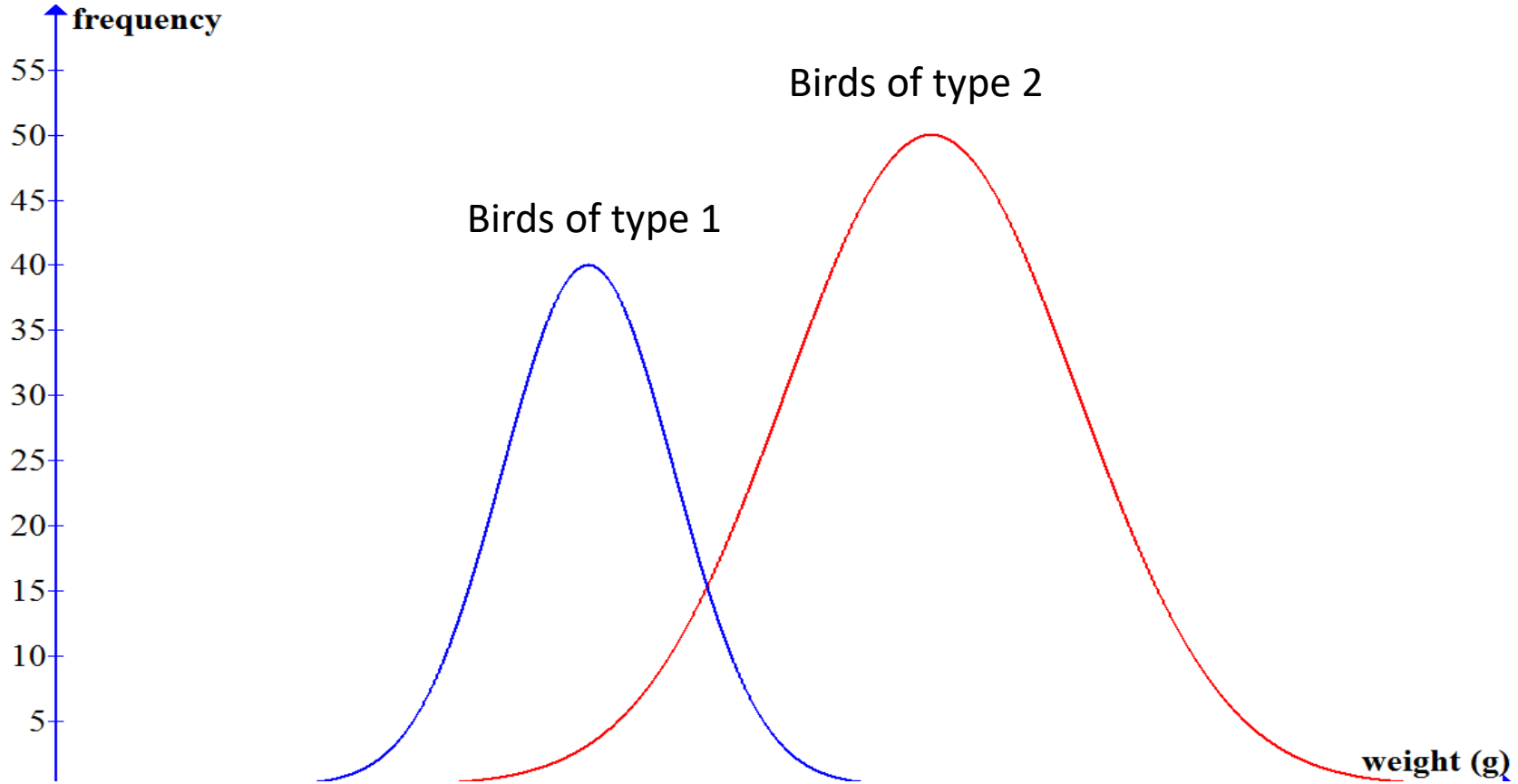
More on the normal distribution

- Let the mean weight be μ = 80kg. This is located at the centre of the curve.
- Species having weights of 70kg would be located to the left of μ.
- Species having weights of 90kg would be located to the right of μ.



More on the normal distribution

- Different normal distributions have different means and SDs.



More on the normal distribution

- The weights of eggs of a birds of type 1 would have its own different normal distribution.
- The weights of eggs of a birds of type 2 would have its own different normal distribution.
- The weights of eggs of a birds of type 3 would have its own different normal distribution.

etc. for 100 different bird types.

More on the normal distribution

- We would then want to see what chance/likelihood there was of an egg chosen at random, for each bird type, having a specific weight.
- **The issue is:** We would need 100 normal distributions, one for each type of bird.
- So, each of these normal distributions will need its own table of values of probabilities.

More on the normal distribution

- But there are hundreds or thousands more situations which have their own normal distribution:
  - shoot length, weight of turtle eggs, adult height, etc.
- We would need separate probability tables for each one, but we can’t have 1000s of tables.

More on the normal distribution

- **Question:** What do we do about this?
- **Answer:** We transform our data.

- **Question:** How?
- **Answer:** By using the formula

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

- This gives us what is known as a **z-score**.

Standard normal distribution

- This formula has the effect of "normalising" our data, i.e. converting it into a standard or normal form.
- Specifically, it transforms our data with mean $\mu = \mu_1$ and standard deviation $\sigma = \sigma_1$ into mean $\mu = 0$ and standard deviation $\sigma = 1$.
- So it doesn't matter what your real (raw) data is, the formula will transform all of it so that $\mu = 0$ & $\sigma = 1$.

Standard normal distribution

- So we can transform any normal distribution into something called a *standard normal distribution*.
- This means that we only need one table of probability values instead of hundreds of separate tables for hundreds of different normal distributions.

Standard normal distribution

- So, we can now answer question such as,
  - If a population of adults has height μ = 177cm and σ = 16cm, what is the probability of choosing a person at random having a height of at most 183cm?

  Here x = 183, so our z-score is z = 0.375.

- We now look up this value in a table of z-scores prepared for us, and get our probability result.

Standard normal distribution

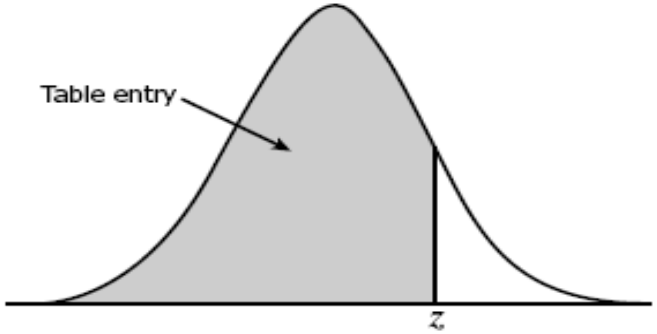


Table entry for z is the area under the standard normal curve to the left of z.

Find the average of these two values.
So the probability is 0.64615

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

The end