Definitions for S1

Statistical Experiment
A text/investigation/process adopted for collecting data to provide evidence for or against a hypothesis.

"Explain briefly why mathematical models can help to improve our understanding of real world problems"
Simplifies a real world problem; enables us to gain a quicker / cheaper understanding of a real world problem

Advantage and disadvantage of statistical model
Advantage : cheaper and quicker
Disadvantage : not fully accurate

"Statistical models can be used to describe real world problems. Explain the process involved in the formulation of a statistical model."

- Observe real-world problem
- Devise a statistical model and collect data
- Compare and observe against expected outcomes and test model;
- Refine model if necessary.

A sample space
A list of all possible outcomes of an experiment

Event
Sub-set of possible outcomes of an experiment.

Normal Distribution

- Bell shaped curve
- symmetrical about mean; mean = mode = median
- 95% of data lies within 2 standard deviations of mean

2 conditions for skewness
Positive skew if $(Q_3 - Q_2) - (Q_2 - Q_1) > 0$ and if Mean – Median > 0.

Negative skew if $(Q_3 - Q_2) - (Q_2 - Q_1) < 0$ and if Mean – Median < 0.

Independent Events
$P(A \cap B) = P(A) \times P(B)$

Mutually Exclusive Events
$P(A \cap B) = 0$

Explanatory and response variables
The response variable is the dependent variable. It depends on the explanatory variable (also called the independent variable). So in a graph of length of life versus number of cigarettes smoked per week, the dependent variable would be length of life. It depends (or may do) on the number of cigarettes smoked per week.

Data

Discrete
Discrete data can only take certain values in any given range. Number of cars in a household is an example of discrete data. The values do not have to be whole numbers (e.g. shoe size is discrete).

Continuous
Continuous data can take any value in a given range. So a person's height is continuous since it could be any value within set limits.

Categorical
Categorical data is data which is not numerical, such as choice of breakfast cereal etc.

Data may be displayed as grouped data or ungrouped data.

We say that data is "grouped" when we present it in the following way:

Weight (w)	Frequency
65-	3
70-	7

Or

Score (s)	Frequency
5-9	2
10-14	5

NB: We can group discrete data or continuous data.

We must know how to interpret these groups,
So that

Weight (w)	
65-	$65 \leq w < 70$
70-	$70 \leq w < 75$

Or

Score (s)	
5-9	$4.5 \leq s < 9.5$
10-14	$9.5 \leq s < 14.5$

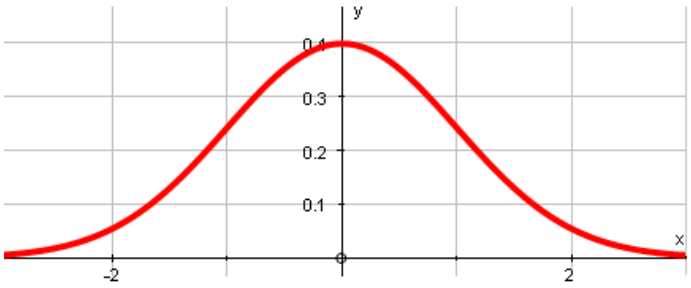
$$P(18 < X < 23) = P\left(\frac{18-20}{2} < \frac{X-20}{2} < \frac{23-20}{2}\right) = P(-1 < Z < 1.5).$$

If we now had a set of tables showing us all possible values for $P(Z < z)$ then we could calculate this since $P(-1 < Z < 1.5) = P(Z < 1.5) - P(Z < -1)$.

So we have two curves

(1) $N(\mu, \sigma^2)$ is the general normal distribution with parameters μ (the mean) and σ^2 (the variance). We use the variable X. For example,

(2) $N(0, 1)$ is the standard normal distribution with parameter 0 (the mean) and. 1 (the variance). We use the variable Z to distinguish it form the general normal case.

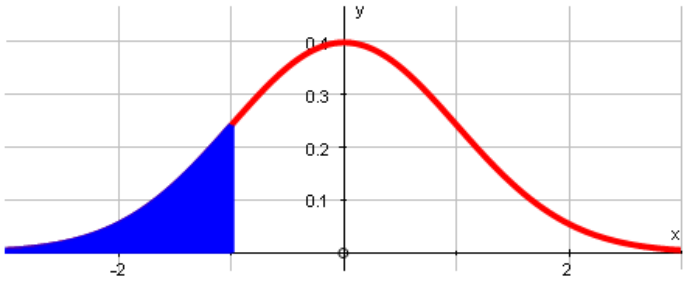


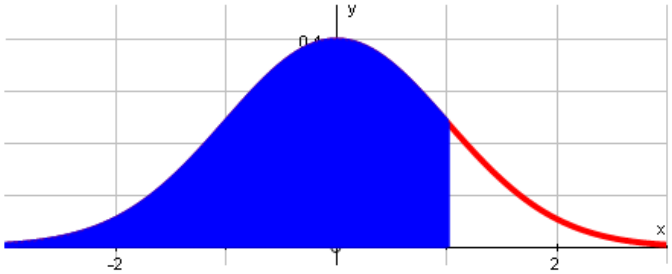
Standard Normal Distribution

The *cumulative distribution function* for the random variable Z is written as $\Phi(z)$.

In other words $\Phi(z) = P(Z < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\, dt$.

From the tables we have $\Phi(0) = 0.5, \Phi(1) = 0.8413, \Phi(2) = 0.9772, \Phi(3) = 0.9987$ etc.





We see from the above that to calculate $\Phi(z)$ when $z < 0$ we use symmetry, i.e. we use the fact that $\Phi(z) = 1 - \Phi(-z)$

So, for example $\Phi(-1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$

General Normal Distribution

So we have seen that if the random variable X has the distribution $N(\mu, \sigma^2)$ then we need to transform this into standard normal distribution $N(0, 1)$ (with random variable Z) using the coding of $Z = \frac{X - \mu}{\sigma}$ and use the tables for z to find the answers.

Find mean or standard deviation

Five stages:

- Write it out using the random variable X
- Turn it into a "greater than problem"
- Turn it into a "Z problem"
- Ensure that the probability on the right hand side is less than ½.
- Use tables (REVERSE SIDE –SEE BELOW)

PERCENTAGE POINTS OF THE NORMAL DISTRIBUTION

The values z in the table are those which a random variable $Z \sim N(0, 1)$ exceeds with probability p; that is, $P(Z > z) = 1 - \Phi(z) = p$.

p	z	p	z
0.5000	0.0000	0.0500	1.6449
0.4000	0.2533	0.0250	1.9600
0.3000	0.5244	0.0100	2.3263
0.2000	0.8416	0.0050	2.5758
0.1500	1.0364	0.0010	3.0902
0.1000	1.2816	0.0005	3.2905

e.g. The amount of jam in a jar is normally distributed with mean μ and standard deviation 5g. Find the mean given that the probability that a jam chosen at random contains:

(a) more than 142g is 0.05

(b) less than 142g is 0.01

(a)

$$P(X > 142) = 0.05$$
$$\Rightarrow P\left(Z > \frac{142-\mu}{5}\right) = 0.05$$
$$\Rightarrow \frac{142-\mu}{5} = 1.6449$$
$$\Rightarrow \mu = 133.8\text{g (to 1dp)}$$

(b)

$$P(X < 142) = 0.01$$
$$\Rightarrow P(X > 142) = 0.99$$
$$\Rightarrow P\left(Z > \frac{142-\mu}{5}\right) = 0.99$$
$$\Rightarrow P\left(Z > \frac{\mu-142}{5}\right) = 0.01$$
$$\Rightarrow \frac{\mu-142}{5} = 2.3263$$
$$\Rightarrow \mu = 153.6\text{g (to 1dp)}$$

X is normally distributed with mean 140g and standard deviation σ. Find σ given that the probability of being more than 150g is 0.2.

$$P(X > 150) = 0.2$$
$$\Rightarrow P\left(Z > \frac{150-140}{\sigma}\right) = 0.2$$
$$\Rightarrow \frac{10}{\sigma} = 0.8416$$
$$\Rightarrow \sigma = 11.9\text{g (to 1dp)}$$

Find mean and standard deviation

X is normally distributed with mean μ and standard deviation σ. Find μ and σ given that the probability of being more than 70 is 0.1 and the probability of being less than 60 is 0.2.

$P(X > 70) = 0.1$

$\Rightarrow P\left(Z > \frac{70-\mu}{\sigma}\right) = 0.1$

$\Rightarrow \frac{70-\mu}{\sigma} = 1.2816$

$\Rightarrow 70 = \mu + 1.2816\sigma$

$P(X < 60) = 0.2$

$\Rightarrow P(Z > 60) = 0.8$

$\Rightarrow P\left(Z > \frac{60-\mu}{\sigma}\right) = 0.8$

$\Rightarrow \frac{\mu-60}{\sigma} = 0.8416$

$\Rightarrow 60 = \mu - 0.8416\sigma$

We now have two simultaneous equations to solve

$70 = \mu + 1.2816\sigma$ (1)

$60 = \mu - 0.8416\sigma$ (2)

Subtract (2) from (1) gives us

$10 = (1.2816 + 0.8416)\sigma$

$10 = 2.1232\sigma$

$\sigma = 4.71$ (to 3sf)

Plugging this back into (1) gives us $\mu = 64.0$ (to 3sf)