

## Statistics and Probability

### Statistics:

Mean = average =  $\mu = \text{sum\_of\_items} / \text{number\_of\_items} = E(X) = \text{estimated value of } X$

Standard deviation =  $\sigma$

Variance:  $\text{var}(X) = \sigma^2 = E((X - \mu)^2) = E(X^2) - \mu^2$

Median = middle value in an ordered data set.

If the number of data points is even use the mean of the 2 central data points.

Mode = maximum (or most common) value in a data set.

Range = largest value – smallest value.

Mean absolute deviation (MAD) from mean =  $E(|X - \mu|) = (\sum |X - \mu|)/n$  where n is number of terms.

Arithmetic mean of a and b is  $(a + b)/2$ .

Geometric mean of a and b is  $\sqrt{(ab)}$ .

### Frequency distributions:

A frequency distribution is a list, table or graph that displays the frequency of outcomes in a sample set.

A cumulative frequency distribution is a list, table or graph that displays the frequency of outcomes below a particular value in a sample set.

### Time series:

Can have:

Increasing/decreasing trend

Repetitive seasonal variations within a year

Long term cycles

Random variations

Seasonal index =  $\text{actual\_figure} / \text{deseasonalised\_figure}$

Seasonal indices have an average value of 1.

A seasonal index of 1.2 for summer indicates a result 20% greater than for an average season.

### Sampling types:

Random

Stratified: divide into subgroups, then do random sampling proportionally on subgroups.

Systematic (e.g. every 50<sup>th</sup>)

Quantitative: involving numerical measurement.

Qualitative: not involving numerical measurement.

### Sample and population:

A sample is a (usually) small part of a population.

Population proportion  $p = (\text{number in population with attribute}) / \text{total\_population}$

Sample proportion  $p_s = (\text{number in sample with attribute})/\text{number\_in\_sample}$

### Data set types:

Univariate data: one variable, bivariate data: two variables.

Discrete: numeric and countable. E.g. age in integer years, pupils, sheep.

Continuous: real value. E.g. age, height, income.

Categorical: named, but might have a numerical order. E.g. month, colour, surname.

Nominal: no specific order. Ordinal data: specific order.

### Skewed data:

Left or negative skewed: mean < mode. (i.e. max to right of centre).

No skew = normal distribution (for example).

Bimodal data: 2 peaks.

### Normal distribution

$$P(x) = (1/(\sigma \sqrt{2\pi})) e^{-((x - \mu)^2/(2 \sigma^2))}$$

Symmetric and bell shaped.

$\mu$  = Mean = mode = median

$\mu$  and  $\sigma$  can have any value.

### Standard Normal distribution

$$P(z) = (1/(\sqrt{2\pi})) e^{-(z^2/2)}$$

$\mu = 0$ , and  $\sigma = 1$ .

To convert to standard normal distribution use Z-score for any score x:

$$z = \text{standardised score} = (x - \mu)/\sigma = (x - \text{mean})/\text{standard\_deviation}$$

Approximate % for Pr(Z) within 1, 2 or 3 standard deviations of mean: 68, 95, 99.7 .

So: approximately 68% of scores have z-scores between -1 and 1 and thus lie within 1 standard deviation of the mean.

This can be written:  $\Pr(z < \mu - \sigma) + \Pr(z > \mu + \sigma) = 0.32$

Here 0.32 is the p-value for the statistical hypothesis above on the left.

### Least squares line of best fit:

Least squares regression line : line of best fit to data.

$$y = mx + c, \text{ where } m = r (\sigma_y/\sigma_x) \text{ and } c = \mu_y - m \mu_x$$

$m$  = gradient

$c$  = y-intercept

$r$  = correlation coefficient

residual value = actual value – predicted value

### Boxplots

IQR = interquartile range =  $Q_3 - Q_1 = Q_U - Q_L = \text{upper\_quartile} - \text{lower\_quartile}$

25% of values lie below  $Q_1$ . 75% of values lie below  $Q_3$ .

Lower fence or whisker =  $Q_1 - 1.5 \text{ IQR}$

Upper fence or whisker =  $Q_3 + 1.5 \text{ IQR}$

Outliers lie outside the fences.

$Q_L$  = median of lower half of data set not counting a central data point.

$Q_U$  = median of upper half of data set not counting a central data point.

For median of an even number of points use interpolation to get average at centre.

e.g.: for six ordered points use position 3.5 for median, 2 for  $Q_1$  and 5 for  $Q_3$ .

e.g.: for seven ordered points use position 4 for median, 2 for  $Q_1$  and 6 for  $Q_3$ .

For box and whisker plot:

Ends of box are lower and upper quartiles. Median is shown.

Ends of lines are lowest and highest observed data.

**Five Number Summary.** For a **set of data**, the minimum, first quartile, median, third quartile, and maximum. Note: A boxplot is a visual display of the **five-number summary**.

### Group data frequency table

Use midpoint of each interval.

### Correlation coefficient examples

0 for horizontal straight line

1 for  $y = x$

-1 for  $y = -x$ .

### Correlation:

Correlation =  $r$  = Pearsons correlaton = Pearsons product-moment correlation. (Range -1 to 1)

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Coefficient of determination =  $r^2$

### Statistics (AISSE syllabus):

Mode calculation from within range  $r_1$  in a frequency distribution.

The adjacent ranges considered are  $r_0$ ,  $r_1$  and  $r_2$ .

$f_i$  = frequency within range  $r_i$ .

$L$  = minimum within range  $r_1$ .

$h$  = range width.

Mode =  $L + (h(f_1 - f_0))/(2f_1 - f_0 - f_2)$

Empirical relationship between three measures of central tendencies:

3 x median = mode + 2 x mean

### Probability:

Probability of an event with equally likely outcomes:

$P(\text{event}) = \text{number\_of\_favourable\_outcomes} / \text{total\_number\_of\_outcomes}$

e.g. throwing a 6 sided die 2 times and adding the result. Each die can give 1-6 so the total can be 2 - 12.  $\Pr(X = 3)$  is the probability that the total result is 3. Dice is the plural of die but a single die is sometimes called a dice.

$0 \leq \Pr(X = x) \leq 1$

The sum of all the possible  $\Pr(X)$  must be 1.

$X$  is a discrete random variable that represents the outcome of an event.

$$N! = N(N - 1)(N - 2) \dots 3 \times 2 \times 1$$

$C_n^N = (N!) / (n!(N - n)!) =$  number of ways or combinations of choosing  $n$  from  $N$  when order is unimportant such as cards.

$P_n^N = (n!) C_n^N = (N!) / (N - n)! =$  number of ways or permutations of choosing  $n$  from  $N$  when order is important such as the first  $n$  in a race.

$$\Pr(A) = 1 - \Pr(A^c) = 1 - \Pr(\text{not } A)$$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \Pr(\text{either } A \text{ or } B \text{ or both})$$

$$\Pr(A \cap B) = \Pr(\text{both } A \text{ and } B) \\ = 0 \text{ if } A \text{ and } B \text{ are mutually exclusive}$$

$$\Pr(A|B) = \Pr(A \cap B) / \Pr(B) = \Pr(A \text{ given } B) \\ \text{This is a conditional probability.}$$

**Mean**       $X$  and  $Y$  are random variables.  
 $\mu = E(X) =$  expected value of  $X$   
 $E(aX + b) = aE(X) + b$   
 $E(aX + bY) = aE(X) + bE(Y)$

**Variance**       $X$  and  $Y$  are independent random variables.  
 $\text{Var}(X) = \sigma^2 = E((X - \mu)^2) = E(X^2) - \mu^2$   
 $\text{Var}(aX + b) = a^2 \text{Var}(X)$   
 $\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$   
 $\sigma =$  standard deviation of  $X = \sqrt{\text{Var}(X)}$

### Independent events:

$A$  and  $B$  are independent if  $B$  occurring has no effect on  $A$ .

$$\Pr(A|B) = \Pr(A)$$

$$\text{And } \Pr(A \cap B) = \Pr(A) \Pr(B)$$

### Combinations:

$b$  factorial =  $b! = b \times (b-1) \times (b-2)$  etc to  $1$ .

e.g.  $4! = 4 \times 3 \times 2 \times 1 = 24$

$C_a^b = b! / (a!(b-a)!) =$  number of ways to select  $a$  objects from  $b$  objects when order does not matter.

e.g.  $C_2^5 = 5! / (2! 3!) = (5 \times 4) / (2!) = 20 / 2 = 10$

### Probability distributions

#### Discrete random variable:

$$0 \leq p(x) \leq 1$$

The sum of all the possible  $p(x)$  must be  $1$ .

$$\Pr(X = x) = p(x) \qquad \mu = E(X) = \sum x p(x) \qquad \sigma^2 = \sum (x - \mu)^2 p(x)$$

#### Continuous:

$f(x) \geq 0$ , for all  $x$ .

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$\Pr(a < X < b) = \int_a^b f(x) dx \quad \mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad \sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$
$$E(g(x)) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

For the median  $m$ :  $\int_{-\infty}^m f(x) dx = 0.5$

**Sample proportions:**

Mean of distribution X	$p$
Standard deviation of X	$s = \sqrt{(p(1 - p)) / n}$
Standard deviation of mean = standard error =	$s / \sqrt{n}$
Variance of mean	$s^2 / n$
Approximate confidence interval for p	$(p - z (s / \sqrt{n}), p + z (s / \sqrt{n}))$
For a 95% confidence interval, standardized score $z = 1.96$ or approximately 2.	
Margin of error =	$z (s / \sqrt{n})$

**Binomial distribution:**

Let  $n$  be the number of successes in  $N$  trials with the probability of each success being  $p$ , then:

$$\Pr(X = n) = C_n^N p^n (1 - p)^{N-n}, \text{ for } n = 0, 1 \text{ etc to } N.$$

$$\mu = n p \quad \text{var} = n p (1 - p) = \sigma^2$$

The sum of the binomial probabilities = 1.

The results of tossing a coin  $N$  times follows the binomial distribution.

If  $N = 1$  this is called a Bernoulli distribution.

**Example calculation:**

Let  $p$  be probability of a successful event.

What is probability of  $n$  failures before success:

$$\Pr(X = n) = (1 - p)^n p$$