# REVISION SHEET – STATISTICS 2 (AQA)

# CONTINUOUS RANDOM VARIABLES

**The main ideas are:**

- Properties of Continuous Random Variables
- Mean, Median and Mode
- Normal approximations to other distributions

***Before the exam you should know:***

- The properties of continuous random variables, including the p.d.f. function.
- How to calculate the mean, variance, median and mode.
- And be able to use the cumulative distribution function.
- How to approximate to the normal distribution from other distributions.

## Continuous Random Variables

A continuous random variable is a random variable that can take any value within a range, i.e. height or weight. It is described by a *probability density function* (p.d.f.). A probability density function may be found from the results of an experiment, or it may be given as an algebraic expression. For a continuous random variable, the total area under the curve of the probability density function must be 1.

The expectation $E(X) = \mu = \int xf(x)dx$ and

$\text{Var}(X) = \int (x - \mu)^2 f(x)\,dx$ *where* $\mu = E(X)$

$= \int (x^2 f(x) - 2\mu x f(x) + \mu^2 f(x))\,dx$

$= \int x^2 f(x)\,dx - 2\mu \int xf(x)\,dx + \mu^2 \int f(x)\,dx$

$= \int x^2 f(x)\,dx - 2\mu^2 + \mu^2$ *since* $\int xf(x)dx = E(X) = \mu$

*and* $\int f(x)dx = 1$

$= \int x^2 f(x)\,dx - \mu^2$

$\text{Var}(X) = \int x^2 f(x)\,dx - [E(X)]^2$

**Example**

A continuous random variable $X$ has p.d.f f(x), where:

$$f(x) = \begin{cases} \frac{1}{3}(x-1) & \text{for } 1 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the expectation and variance of $X$.

**Solution**

$$E(X) = \int_1^3 \tfrac{1}{3}x(x-1)dx = \tfrac{1}{3}\int_1^3 (x^2 - x)dx = \tfrac{1}{3}\left[\tfrac{1}{3}x^3 - \tfrac{1}{2}x^2\right]_1^3 = \tfrac{14}{9}$$

$$\text{Var}(X) = \int_1^3 \tfrac{1}{3}x^2(x-1)dx - [E(X)]^2 = \tfrac{1}{3}\int_1^3 (x^3 - x^2)dx - \left[\tfrac{14}{9}\right]^2$$

$$= \tfrac{1}{3}\left[\tfrac{1}{4}x^4 - \tfrac{1}{3}x^3\right]_1^3 - \left[\tfrac{14}{9}\right]^2$$

$$= \frac{100}{81}$$

**Extension of Mean and Variance**

In other questions you will need to use the following properties:

$E(aX+b) = a\,E(X)+b$ e.g. $E(3X+2) = 3\,E(X)+2$

and

$Var(aX+b) = a^2\,Var(X)$ e.g. $Var(3X) = 9\,Var(X)$

**Rectangular distribution**

A distribution with for which f($x$) is a constant within a particular range and zero elsewhere.

Its p.d.f. is given by:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

The expectation is $\dfrac{a+b}{2}$ and the variance is $\dfrac{1}{12}(b-a)^2$.

**Cumulative distribution function (c.d.f.)**

The cumulative distribution function $F(x) = P(X \le x)$.
It can be found from the p.d.f. f($x$) as follows:

$F(x) = 0 \quad x < a$

$= \int_a^x f(u)du \quad a \le x \le b$

$= 1 \quad x > b$

where $a$ is the lower limit of f($x$), and $b$ is the upper limit of f($x$).

**Normal Distribution as an approximation to the Binomial Distribution**

Suppose $X \sim \text{Binomial}(n, p)$.

If $n$ is large and $p$ is not too close to 0 or 1 (i.e. the distribution is reasonably symmetrical), then using the mean ($np$) and variance ($npq$) of a binomial distribution we can approximate using the normal distribution.

$$X \sim N(np, npq)$$

**Normal Distribution as an approximation to the Poisson Distribution**

Suppose $X \sim Poisson(\lambda)$

If λ is large, then the Poisson distribution is reasonably symmetrical.
Then using the mean (λ) and variance (λ) of a Poisson distribution we can approximate using the normal distribution.

$$X \sim N(\lambda, \lambda)$$

**Important:** In both cases above we are using a continuous distribution to approximate a discrete one and as such we must use continuity correcting when calculating a probability. Make sure you understand how to do this.

Note. You will also be expected to build upon knowledge gained in S1 on Confidence Intervals, only using the t distribution. Please check that you can do this.

# REVISION SHEET – STATISTICS 2 (AQA)

# DISCRETE RANDOM VARIABLES

**The main ideas are:**

- Discrete random variables
- Expectation (mean) of a discrete random variable
- Variance of a discrete random variable

***Before the exam you should know:***

- Discrete random variables are used to create mathematical models to describe and explain data you might find in the real world.
- You must understand the notation that is used.
- You must know that a discrete random variable X takes values $r_1$, $r_2$, $r_3$, $r_4$, …, $r_n$ with corresponding probabilities: $p_1$, $p_2$, $p_3$, $p_4$, …, $p_n$.
- Remember that the sum of these probabilities will be 1 so $p_1+ p_2+ p_3+ p_4, \dots + p_n = \Sigma P(X=r_k) = 1$.
- You should understand that the expectation (mean) of a discrete random variable is defined by

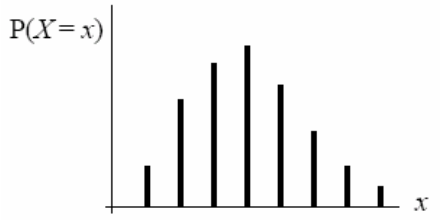
$$E(X) = \mu = \Sigma r P(X=r_k)$$

- You should understand that the variance of a discrete random variable is defined by:

$$\text{Var}(X) = \sigma^2 = E(X-\mu)^2 = \Sigma(r-\mu)^2 P(X=r)$$

$$\text{Var}(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

**Discrete random variables** with probabilities $p_1$, $p_2$, $p_3$, $p_4$, …, $p_n$ can be Illustrated using a vertical line chart:




**Notation**

- A discrete random variable is usually denoted by a capital letter (X, Y etc).
- Particular values of the variable are denoted by small letters (r, x etc)
- $P(X=r_1)$ means the probability that the discrete random variable X takes the value $r_1$
- $\Sigma P(X=r_k)$ means the sum of the probabilities for all values of r, in other words $\Sigma P(X=r_k) = 1$

**Example:** A child throws two fair dice and adds the numbers on the faces. Find the probability that

(i) P(X=4) (the probability that the total is 4)

(ii) P(X<7) (the probability that the total is less than 7)

**Answer:**

(i) $P(X=4) = \frac{3}{36} = \frac{1}{12}$ (ii) $P(X<7) = \frac{15}{36} = \frac{5}{12}$

**Example:** $X$ is a discrete random variable given by $P(X = r) = \frac{k}{r}$ for $r$ = 1, 2, 3, 4 Find the value of k and illustrate the distribution.

**Answer:**

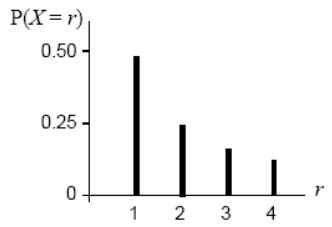
To find the value of $k$, use $\Sigma P(X = x_i) = 1$

$\Sigma P(X = x_i) = \frac{k}{1}+\frac{k}{2}+\frac{k}{3}+\frac{k}{4} = 1$

$\Rightarrow \frac{25}{12}k = 1$

$\Rightarrow k = \frac{12}{25} =$ **0.48**

Illustrate with a vertical line chart:




**Example**

Calculate the expectation and variance of the distribution

**Answer:**

Expectation is

$E(X) = \mu$

$= \Sigma r P(X = r) = 1\times0.48 + 2\times0.24 + 3\times0.16 + 4\times0.12$

$=$ **1.92**

$E(X^2) = \Sigma r^2 P(X = r)$

$= 1^2\times0.48 + 2^2\times0.24 + 3^2\times0.16 + 4^2\times0.12$

$= 4.5$

Variance is $\text{Var}(X) = E(X^2) - [E(X)]^2$

$= 4.5 - 1.92^2$ = **0.8136**

**Using tables:**

For a small set of values it is often convenient to list the probabilities for each value in a table

| $r_i$ | $r_1$ | $r_2$ | $r_3$ | …. | $r_{n-1}$ | $r_n$ |
|---|---|---|---|---|---|---|
| $P(X = r_i)$ | $p_1$ | $p_2$ | $p_3$ | …. | $p_{n-1}$ | $p_n$ |

**Using formulae**:

Sometimes it is possible to define the probability function as a formula, as a function of $r$, $P(X = r) = f(r)$

**Calculating probabilities**:

Sometimes you need to be able to calculate the probability of some compound event, given the values from the table *or* function.

**Explanation of probabilities**:

Often you need to explain how the probability $P(X = r_k)$, for some value of $k$, is derived from first principles.

**Example**:

The discrete random variable $X$ has the distribution shown in the table

| $r$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X = r)$ | 0.15 | 0.2 | 0.35 | 0.3 |

(i) Find $E(X)$.
(ii) Find $E(X^2)$.
(iii) Find $Var(X)$ using (*a*) $E(X^2) - \mu^2$ and (*b*) $E(X - \mu)^2$.
(iv) Hence calculate the standard deviation.

| $r$ | 0 | 1 | 2 | 3 | totals |
|---|---|---|---|---|---|
| $P(X = r)$ | 0.15 | 0.2 | 0.35 | 0.3 | **1** |
| $rP(X = r)$ | 0 | 0.2 | 0.7 | 0.9 | **1.8** |
| $r^2P(X = r)$ | 0 | 0.2 | 1.4 | 2.7 | **4.3** |
| $(r - \mu)^2$ | 3.24 | 0.64 | 0.04 | 1.44 | **5.36** |
| $(r-\mu)^2P(X=r)$ | 0.486 | 0.128 | 0.014 | 0.432 | **1.06** |

*This is the expectation (μ)*

*This is $E(X^2)$*

*This is $Var(X) = \Sigma(r-\mu)^2P(X=r)$*

(i) $E(X) = \mu = \Sigma r\, P(X = r) = 0\times0.15 + 1\times0.2 + 2\times0.35 + 3\times0.3$
$= 0 + 0.2 + 0.7 + 0.9$
$=$ **1.8**

(ii) $E(X^2) = \Sigma r^2\, P(X = r) = 0^2x0.15 + 1^2x0.2 + 2^2x0.35 + 3^2x0.3$
$= 0 + 0.2 + 1.4 + 2.7$
$=$ **4.3**

(iii) (a) $Var(X) = E(X^2) - \mu^2 = 4.3 - 1.8^2 =$ **1.06**

(b) $Var(X) = E(X - \mu)^2 = 0.15(0-1.8)^2 + 0.2(1-1.8)^2 + 0.35(2-1.8)^2 + 0.3(3-1.8)^2$
$= 0.486 + 0.128 + 0.014 + 0.432$
**= 1.06**

(iv) $s = \sqrt{1.06} = 1.02956 =$ **1.030** (3d.p.)

*Notice that the two methods give the same result since the formulae are just rearrangements of each other.*

*standard deviation (s) is the square root of the variance*

# REVISION SHEET – STATISTICS 2 (AQA)

# HYPOTHESIS TESTING & CONTINGENCY TABLES

## The main ideas are:

- Hypothesis Testing Normal Distribution
- $\chi^2$ and Contingency Tables

## *Before the exam you should know:*

- About hypothesis testing for the mean using the Normal Distribution.
- About using a known and an estimated standard deviation.
- The $\chi^2$ test for independence in a contingency table.

## Hypothesis Testing

A ***null hypothesis*** ($H_0$) is tested against an ***alternative hypothesis*** ($H_1$) at a particular ***significance level***.
According to given criteria, the null hypothesis is either rejected is not rejected.
The hypothesis test can be either 1-tailed or 2-tailed.

Sample data, drawn from the parent population, may be used to carry out a hypothesis test on the null hypothesis that the population mean has some particular value, $\mu_0$.

***Hypothesis testing procedure***

**(1)** Establish null and alternative hypotheses:
$H_0$: $\mu = \mu_0$
$H_1$: $\mu < \mu_0$ *or* $\mu > \mu_0$ (1-tail test);
*or* $\mu \neq \mu_0$ (2-tail test)

**(2)** Decide on the significance level: *s%*

**(3)** Collect data (independent and at random): obtain sample of size *n* from the parent population and calculate mean $\bar{x}$.

**(4)** Conduct test:
Calculate test statistic: $z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$
[if $\sigma$ unknown, use *s* – provided *n* is large]
1-tail: $H_1$: $\mu < \mu_0$ – compare *z* with $\Phi^{-1}(s\%)$
1-tail: $H_1$: $\mu > \mu_0$
– compare *z* with $\Phi^{-1}((100 - s)\%)$
2-tail: $H_1$: $\mu \neq \mu_0$
if $\bar{x} < \mu_0$ compare *z* with $\Phi^{-1}(\frac{1}{2}s\%)$
if $\bar{x} > \mu_0$ compare *z* with $\Phi^{-1}((100 - \frac{1}{2}s)\%)$

**(5)** Interpret result in terms of the original claim:
1-tail: if $z < \Phi^{-1}(s\%)$ reject $H_0$
1-tail: if $z > \Phi^{-1}((100 - s)\%)$ reject $H_0$
2-tail: if $z < \Phi^{-1}(\frac{1}{2}s\%)$ *or* $z > \Phi^{-1}((100 - \frac{1}{2}s)\%)$ reject $H_0$
*Finally present conclusion in context of problem.*

## Distribution of Sample Means

For samples of size *n* drawn from a Normal distribution with mean $\mu$ and finite variance $\sigma^2$, [ $X \sim N(\mu, \sigma^2)$ ] the distribution of sample means, $\bar{X}$, is Normal with mean $\mu$ and variance $\dfrac{\sigma^2}{n}$, i.e. $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

*The* standard error of the mean *(i.e. the standard deviation of the sample means)* is given by $\dfrac{\sigma}{\sqrt{n}}$.

**Example 1**

The packaging on a type of electric light bulb states that the average lifetime of bulbs is 1000 hours. A consumer association thinks that this is an overestimate and tests a sample of 100 bulbs, recording the life-time, *x* hours, of each bulb.

Assuming the distribution of lifetimes is Normal, test the consumer association's claim at the 5% level.

***Hypothesis testing procedure***

**(1)** $H_0$: $\mu = 1000$; $H_1$: $\mu < 1000$ (1-tail test)

**(2)** Significance level: 5%

**(3)** Summary statistics for the lifetimes, *x*, of a random sample of 100 bulbs:
$n = 100$, $\Sigma x = 99860$, $\Sigma x^2 = 99725047$
$\bar{x} = 99860 \div 100 = 998.6$;
$s = \sqrt{\dfrac{99725047 - 100 \times 998.6^2}{99}} = 7$

**(4)** Test statistic: $z = \dfrac{998.6 - 1000}{\dfrac{7}{\sqrt{100}}} = -2$
Critical value in lower tail: $\Phi^{-1}(0.05) = -1.645$, which is greater than –2.

**(5)** Since $-2 < -1.645$, there is sufficient evidence to reject $H_0$, i.e. the consumer association's claim that the average life-time is less than 1000 hours is upheld at the 5% significance level.

# Contingency Tables

An $m \times n$ ***contingency table*** results when two variables are measured on a sample, with the first variable having $m$ possible categories of results and the second variable having $n$ possible categories.

Each cell contains an *observed frequency* ($f_o$), with which that pair of categories of values of the two variables occurs in the sample.

Marginal row and column totals are used to calculate *expected frequencies* ($f_e$).

# Hypothesis Testing

A ***null hypothesis*** ($H_0$) is tested against an ***alternative hypothesis*** ($H_1$) at a particular ***significance level*** with a number of ***degrees of freedom***.

According to given criteria, the null hypothesis is either rejected is not rejected.
*The hypothesis test is always 1-tailed.*

Sample data, drawn from the parent population, may be used to carry out a hypothesis test on the null hypothesis that there is *no association* between the two variables, i.e. are *independen*t.

***Hypothesis testing procedure***

**(1)** Establish null and alternative hypotheses:
$H_0$: no association between the variables,
$H_1$: the variables are *not* independent**.**

**(2)** Decide on the significance level: *s%*

**(3)** Collect data in the form of an $m$ by $n$ contingency table of observed frequencies ($f_o$)

**(4)** Conduct test:

Calculate marginal row and column totals.

Calculate expected frequencies ($f_e$):

$$f_e = \frac{\text{row total} \times \text{column total}}{\text{total sample size}}$$

Calculate test statistic $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

Find degrees of freedom $v = (m-1)(n-1)$

Compare with critical value from tables, dependant on significance level $s$% and d.o.f. $v$. Note. for a 2x2 table Yates' Correction is required. (plese see your notes)

**(5)** Interpret result in terms of the original claim:

If test statistic $\chi^2$ > critical value, then reject $H_0$ (i.e. accept $H_1$)

If test statistic $\chi^2$ < critical value, then do not reject $H_0$ (i.e. accept $H_0$)

*Present conclusion in context of problem.*

**(6)** Discuss conclusions in terms of which cells make the greatest contribution to the total value of the test statistic.

## Example 2

A personnel manager is investigating whether there is any association between the length of service of the employees and the type of training they receive.

Carry out a hypothesis at the (a) 5% and (b) 1% significance level, to determine if there is any association between length of service and type of training.

***Hypothesis testing procedure***

**(1)** $H_0$: the variables are independent,
$H_1$: the variables are *not* independent**.**

**(2)** Significance level = 5% (and 1%)

**(3)** Records of a random sample of 200 employees are shown in the following ***contingency table*** of observed frequencies ($f_o$):

| *Type of training* | *Length of service* | | | |
|---|---|---|---|---|
| | Short | Medium | Long | *Totals* |
| Induction course | 14 | 23 | 13 | 50 |
| Initial on-the-job | 12 | 7 | 13 | 32 |
| Continuous | 28 | 32 | 58 | 118 |
| *Totals* | 54 | 62 | 84 | 200 |

**(4)** Marginal row and column totals are shown above.

Expected frequencies ($f_e$):

| *Type of training* | *Length of service* | | | |
|---|---|---|---|---|
| | Short | Medium | Long | *Totals* |
| Induction course | 13.5 | 15.5 | 21 | 50 |
| Initial on-the-job | 8.64 | 9.92 | 13.44 | 32 |
| Continuous | 31.86 | 36.58 | 49.56 | 118 |
| *Totals* | 54 | 62 | 84 | 200 |

Contributions to $\chi^2$ : $\frac{(f_o - f_e)^2}{f_e}$

| | | |
|---|---|---|
| 0.019 | 3.629 | 3.048 |
| 1.307 | 0.86 | 0.014 |
| 0.468 | 0.573 | 1.437 |

Test statistic $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

$= 0.019 + 3.629 + \ldots + 0.573 + 1.437$
$= 11.354$

Degrees of freedom $v = (3-1)(3-1) = 4$

Critical values: (a) 5%: 9.488, (b) 1%: 13.28

**(5)** At 5% level of significance:

**(a)** Since 11.354 > 9.488, reject $H_0$ (i.e. accept $H_1$), there is an association between the length of service and the type of training.

**(b)** Since 11.354 < 13.28, reject $H_1$ (i.e. accept $H_0$), there is no association between the length of service and the type of training.

**(6)** The cells with the largest values are medium/ induction and long/ induction, so medium and long service may seem to be associated, respectively, with more than and fewer than expected employees with induction-only training.

*Acknowledgement: Material in this revision sheet was originally created by Bob Francis and we acknowledge his permission to reproduce it here.*

REVISION SHEET – STATISTICS 2 (AQA)

# HYPOTHESIS TESTING USING THE BINOMIAL DISTRIBUTION

## The main ideas are:

- Establishing the null and alternative hypotheses
- Conducting the test, doing the necessary calculations
- Interpreting the results

## *Before the exam you should know:*

- The vocabulary associated with hypothesis testing.
- How to write the null and alternative hypotheses.
- How to decide whether the hypothesis test is one or two tailed.
- How to compare a value to the significance level.
- How to find critical values/regions.
- How to decided whether to reject $H_0$ or not and how to write a conclusion based on the situation.
- How to carry out a 2-tail test.

## Vocabulary

You should be familiar with the following terms/notation for binomial hypothesis tests

**Probability of success:** $p$ **Number of trials:** $n$ **Number of successes:** $X$

| | |
|---|---|
| **Null Hypothesis ($H_0$):** | The statement that the probability of success is equal to a certain value. |
| **Alternative Hypothesis ($H_1$):** | The statement that the probability of success is actually $<$, $>$ or $\neq$ to the value in given in $H_0$. |
| **Significance level:** | The probability at which you make the decision that an observed outcome hasn't happened by chance, given the probability of success in $H_0$. |
| **1-tail test:** | A test based on the probability in $H_0$ being either too high or too low (but not both). |
| **2-tail test:** | A test based on the probability in $H_0$ being incorrect (too high or too low). |
| **Critical value:** | The maximum (for $<$) or minimum (for $>$) value, $X$, for the number of successes that would result in rejecting $H_0$. |
| **Critical region:** | The set of values of $X$ for the number of successes that would result in rejecting $H_0$. |
| **Acceptance region:** | The set of values of $X$ for the number of successes that would result in accepting $H_0$. |
| **Errors** | P(type I error) = P(reject $H_0$ \| $H_0$ true), P(type II error) = P(accept $H_0$ \| $H_0$ false). |

## Hypothesis Tests

Hypothesis testing is based on assuming that the probability of success, $p$, takes a certain value, then conducting an experiment to test it. Given this assumption, if the result of the experiment is sufficiently rare (i.e. unlikely to have happened by chance) you can conclude that the probability, $p$, is likely to be incorrect.

### Setting up

The statement of the value of this assumed probability, $p$, is known as the Null Hypothesis ($H_0$) (this is what you are testing). You must then decide if the situation leads you to think this value is too high, too low or, in the case of a 2-tailed test, incorrect.

### Conducting

The probability of obtaining the value recorded in the experiment, $x$, or something more extreme is compared to the significance level to see if it is sufficiently rare to reject the null hypothesis. You must use $P(X \leq x)$ or $P(X \geq x)$ as opposed to $P(X = x)$.

### Drawing conclusions

If the probability is smaller than the significance level then reject $H_0$ in favour of $H_1$, otherwise you accept $H_0$ at the stated significance level.

**Example**

The makers of the drink Fizzicola claim that three-quarters of people prefer their drink to any other brand of cola. A rival company suspects that the claim by Fizzicola is exaggerated. They wish to carry out a hypothesis test to test this claim.

(i) State suitable Null and Alternative Hypotheses.

The rival company take a sample of 15 cola drinkers of whom 9 say they prefer Fizzicola to any other brand.

(ii) Using these data, carry out a hypothesis test at the 5% level stating your conclusion carefully.

**Solution**

(i) $H_0$: $p = 0.75$; The probability of a person chosen at random preferring Fizzicola is 0.75.
$H_1$: $p < 0.75$; The claim is exaggerated, the probability of a person chosen at random preferring Fizzicola is less than 0.75.
*The alternative hypothesis is based on the rival branding thinking the claim is exaggerated, i.e. the proportion stated is too high.*

(ii) From the tables: $P(X \leq 9) = 0.1484$. This value is not significant at the 5% level, therefore we accept $H_0$. There is not sufficient evidence to suggest Fizzicola are overestimating the proportion.
*The probability of 9 or fewer is used, as opposed to exactly 9, as if you would accept 9 as evidence of overestimating then you would have also accepted 8, 7, 6, …*
*The significance level tells you the value at which a probability is considered so rare that is unlikely to have happened by chance. In this example case 5% is used, so an event with probability smaller than 0.05 is considered rare: 0.1484 is not smaller than 0.05 so the event is not rare.*
*As the event is not rare, it is likely that it occurred by chance, so there is no evidence to suggest that the makers of Fizzicola were overestimating. Note that you are not saying that they are correct, just that you don't have strong enough evidence to contradict them.*

**Alternative solution using critical value/critical region**

(ii) From the tables: $P(X \leq 7) = 0.0173$, $P(X \leq 8) = 0.0566$. The critical value is 7, (the critical region is 0-7). 9 is not in the critical region (it is in the acceptance region), therefore we do not reject $H_0$. There is not sufficient evidence to suggest Fizzicola are overestimating the proportion.
*The critical value is the largest (because $H_1$ is <) value of x such that $P(X < x)$ is smaller than the significance level.*

This example used an alternative hypothesis of the form $H_1$: $p < 0.75$ (because the rival firm thought the company was overestimating). This made it easy to read the values for $P(X \leq 7)$, $P(X \leq 8)$ and $P(X \leq 9)$ from the tables. If the alternative hypothesis had been of the form $H_1$: $p > 0.75$ (e.g. if the firm thought 0.75 was an underestimate), you would need to work with ≥ probabilities, using the complement of the values in the table.

**e.g.** If the alternative hypothesis had been $H_1$: $p > 0.75$ you would have calculated $P(X \geq 9)$.
$P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.0566 = 0.9434$: this is not smaller than 0.05 so you do not reject $H_0$.

## 1-tail vs 2-tail tests.

If there is no indication in the situation as to whether the probability used in $H_0$ is too high or too low you use a 2-tailed test, splitting the significance level in half and using half at each end.

**Example**

A teacher is forming a 12-person committee of students. She does not want the selection system to unfairly favour either boys or girls. Construct a hypothesis test at the 5% level to test this.

**Solution**

$H_0$: $p = 0.5$, There is an equal chance of a boy or girls being chosen.
$H_1$: $p \neq 0.5$, The selection system favours one gender.
You then split the significance level in half forming two critical regions of 2.5% at the top and bottom, totalling 5%.
Critical regions: 0 – 2 and 10 – 12.

# REVISION SHEET – STATISTICS 2 (AQA)

# POISSON DISTRIBUTION

## The main ideas are:

- Calculations using the Poisson Distribution
- Modelling the Binomial distribution with the Poisson distribution

## *Before the exam you should know:*

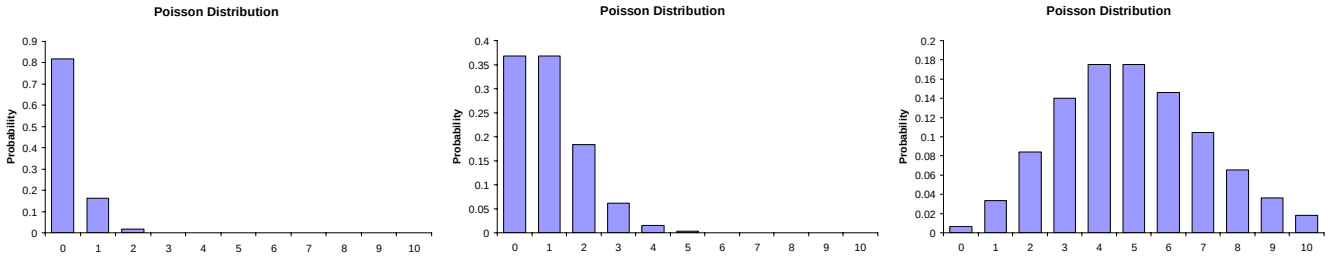
- When the Poisson distribution is an appropriate model for a given situation.
- The relationship $e^y = x \Leftrightarrow y = \ln x$, this is sometimes useful in questions.
- How to use the formula $P(X = r) = e^{-\lambda}\frac{\lambda^r}{r!}$ (without getting confused between $\lambda$ and $r$).
- How to look up $P(X \le r)$ in the tables given.

## Poisson Distribution

This models events which are random, independent, which occur singly and with a uniform likelihood.

If $X \sim$ Poisson($\lambda$) then: $P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}$, where E($X$) = $\mu = \lambda$ and Var($X$) = $\sigma^2 = \lambda$.

The Poisson Distribution for various values of $\lambda$ is shown below.




$\lambda = 0.2$ $\lambda = 1$ $\lambda = 5$

## Calculations using the Poisson Distribution

You should be able to use the formula $P(X = r) = e^{-\lambda}\frac{\lambda^r}{r!}$ and the cumulative Poisson tables (which give $P(X \le r)$ for various values of $\lambda$) to find simple probabilities.

**Example 1**

The number of goals, $X$, scored by a team playing at home in the Premier League is modelled by a Poisson distribution with a mean of 1.6. What is the probability that the team scores

*a) 3 goals in a game*

*b) More than 4 goals in a game*

**Solution**

*a*) The probability of the team scoring 3 goals in a game is: $P(X = 3) = e^{-1.6}\frac{1.6^3}{3!} = 0.138$ (to 3 d.p.)

*b*) The probability of the team scoring more than 4 goals in a game is:

$$P(X > 4) = 1 - P(X \le 4) = 1 - 0.9763 = 0.0237$$

## More Complicated Questions

In other questions you will need to use the following properties of the Poisson Distribution:

If $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ then: $nX \sim \text{Poisson}(n\lambda_1)$ and $X+Y \sim \text{Poisson}(\lambda_1+\lambda_2)$

**Example 2**
The mean number of burgers sold per minute at a snack bar is 0.7. The mean number of hotdogs sold per minute is 0.5. Using a Poisson distribution model calculate the probability that the burger bar sells:

*a) 5 burgers in a 5 minute period.*
*b) No hot dogs or burgers in a 1 minute period.*

**Solution**

a) The mean number of burgers sold in one minute is 0.7. Therefore the mean number of burgers sold in five minutes is $0.7 \times 5 = 3.5$. So, $5X$ is the number of burgers sold in 5 minutes and we have that $5X \sim \text{Poisson}(3.5)$. Therefore,

$$P(5X = 5) = e^{-3.5}\frac{3.5^5}{5!} = 0.132 \text{ (to 3 d.p.)}$$

b) If $X$ is the number of burgers sold in a minute then $X \sim \text{Poisson}(0.7)$. If $Y$ is the number of burgers sold in a minute then $Y \sim \text{Poisson}(0.5)$. So the total number of hotdogs and burgers sold in a minute is $X + Y$ and $X+Y \sim \text{Poisson}(0.7 + 0.5 = 1.2)$. Therefore,

$$P(X + Y = 0) = e^{-1.2}\frac{1.2^0}{0!} = 0.301$$

## Approximating the Binomial Distribution with the Poisson Distribution

If $X \sim \text{Binomial}(n, p)$ a Poisson approximation of $X \sim \text{Poisson}(np)$ can be used when

- $n$ is large
- $p$ is small (i.e. it is a rare event)

but it is only useful if $np$ is not too large.

For example if $n = 1000$, $p = 0.002$, then $np = 2$. Under the binomial distribution $X \sim \text{Binomial}(1000, 0.02)$

$$P(X = 10) = {}^{1000}C_{10} \times 0.002^{10} \times 0.998^{990} = 0.000037 \text{ to (6 d.p)}$$

With the Poisson Distribution $X \sim \text{Poisson}(2)$

$$P(X = 10) = e^{-2}\frac{2^{10}}{10!} = 0.000038 \text{ (to 6 d.p)}$$

The difference between these two values is only 0.000001