ZNOTES // A-LEVEL SERIES

visit www.znotes.org

Updated to 2019 Syllabus

CIE A-LEVEL MATHS 9709 (S2)

FORMULAE AND SOLVED QUESTIONS FOR STATISTICS 2 (S2)

TABLE OF CONTENTS

1. The Poisson Distribution

- The **Poisson distribution** is used as a model for the number, X, of events in a given interval of space or times. It has the probability formula

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, 2, \dots$$

Where λ is equal to the mean number of events in the given interval

- A Poisson distribution with mean λ can be noted as

$$X \sim Po(\lambda)$$

1.1 Suitability of a Poisson Distribution

- Occur randomly in space or time
- Occur singly – events cannot occur simultaneously
- Occur independently
- Occur at a constant rate – mean no. of events in given time interval proportional to size of interval

1.2 Expectation & Variance

- For a Poisson distribution $X \sim Po\ (\lambda)$
- Mean $= \mu = E(X) = \lambda$
- Variance $= \sigma^2 = Var(X) = \lambda$
- The mean & variance of a Poisson distribution are equal

1.3 Addition of Poisson Distributions

- If X and Y are independent Poisson random variables, with parameters λ and μ respectively, then $X + Y$ has a Poisson distribution with parameter $\lambda + \mu$

(IS) Ex 8d: **Question 1:**

The numbers of emissions per minute from two radioactive objects A and B are independent Poisson variables with mean 0.65 and 0.45 respectively.

Find the probabilities that:

i. In a period of three minutes there are at least three emissions from A.

ii. In a period of two minutes there is a total of less than four emissions from A and B together.

Solution:

Part (i):

Write the distribution using the correct notation

$$A \sim Po(0.65 \times 3) = A \sim Po(1.95)$$

Use the limits given in the question to find probability

$$P(A \geq 3) = 1 - P(A < 3)$$

$$= 1 - \left(\frac{1.95^2 e^{-1.95}}{2!} + \frac{1.95^1 e^{-1.95}}{1!} + \frac{1.95^0 e^{-1.95}}{0!}\right)$$

$$= 1 - 0.690 = 0.310$$

Part (ii):

Write the distribution using the correct notation

$$(A + B) \sim Po\big(2(0.65 + 0.45)\big) = (A + B) \sim Po(2.2)$$

Use the limits given in the question to find probability

$$P(A < 4) = e^{-2.2}\left(\frac{(2.2)^3}{3!} + \frac{(2.2)^2}{2!} + \frac{(2.2)^1}{1!} + \frac{(2.2)^0}{0!}\right)$$

$$= 0.819$$

1.4 Relationship of Inequalities

- $P(X < r) = P(X \leq r - 1)$
- $P(X = r) = P(X \leq r) - P(X \leq r - 1)$
- $P(X > r) = 1 - P(X \leq r)$
- $P(X \geq r) = 1 - P(X \leq r - 1)$

1.5 Poisson Approximation of a Binomial Distribution

- To approximate a binomial distribution given by:

$$X \sim B(n, p)$$

- If $n > 50$ and $np > 5$
- Then we can use a Poisson distribution given by:

$$X \sim Po(np)$$

(IS) Ex 8d: **Question 8:**

A randomly chosen doctor in general practice sees, on average, one case of a broken nose per year and each case is independent of the other similar cases.

i. Regarding a month as a twelfth part of a year,

a. Show that the probability that, between them, three such doctors see no cases of a broken nose in a period of one month is 0.779

b. Find the variance of the number of cases seen by three such doctors in a period of six months

ii. Find the probability that, between them, three such doctors see at least three cases in one year.

iii. Find the probability that, of three such doctors, one sees three cases and the other two see no cases in one year.

Solution:

Part (i)(a):

Write down the information we know and need

1 doctor $=$ 1 nose per year $= \frac{1}{12}$ noses per month

3 doctors$= \frac{3}{12} = \frac{1}{4}$ noses per month

Write the distribution using the correct notation

$$X \sim Po(0.25)$$

Use the limits given in the question to find probability

$$P(X=0)=\frac{0.25^0 e^{-0.25}}{0!}=0.779$$

Part (i)(b):

Use the rules of a Poisson distribution

$$Var(X)=\mu=\lambda$$

Calculate λ in this scenario:

$$\lambda=6\times\mu\ (in\ one\ month)=6\times 0.25=1.5$$
$$\therefore Var(X)=1.5$$

Part (ii):

Calculate λ in this scenario:

$$\lambda=12\times\mu\ (in\ one\ month)=12\times 0.25=3$$

Use the limits given in the question to find probability

$$P(X\geq 3)=1-P(X\leq 2)$$
$$=1-e^{-3}\left(\frac{3^2}{2!}+\frac{3^1}{1!}+\frac{3^0}{0!}\right)=1-0.423=0.577$$

Part (iii):

We will need two different λs in this scenario:

$$\lambda\ for\ one\ doctor\ in\ one\ year=1$$
$$\lambda\ for\ other\ two\ doctors\ in\ one\ year=2\times 1=2$$

For the first doctor:

$$P(X=3)=e^{-1}\left(\frac{1^3}{3!}\right)$$

For the two other doctors:

$$P(X=0)=e^{-1}\left(\frac{1^0}{0!}\right)$$

Considering that any of the three could be the first

$$P(X)=e^{-1}\left(\frac{1^3}{3!}\right)\times e^{-1}\left(\frac{1^0}{0!}\right)\times {}^3C_2=0.025$$

1.6 Normal Approximation of a Poisson Distribution

- To approximate a Poisson distribution given by:

$$X\sim P(\lambda)$$

- If $\lambda>15$
- Then we can use a normal distribution given by:

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

Apply continuity correction to limits:

Poisson	Normal
$x=6$	$5.5\leq x\leq 6.5$
$x>6$	$x\geq 6.5$
$x\geq 6$	$x\geq 5.5$
$x<6$	$x\leq 5.5$
$x\leq 6$	$x\leq 6.5$

(IS) Ex 10h: **Question 11:**

The no. of flaws in a length of cloth, lm long has a Poisson distribution with mean $0.04l$

i. Find the probability that a 10m length of cloth has fewer than 2 flaws.

ii. Find an approximate value for the probability that a 1000m length of cloth has at least 46 flaws.

iii. Given that the cost of rectifying X flaws in a 1000m length of cloth is X^2 pence, find the expected cost.

Solution:

Part (i):

Form the parameters of Poisson distribution

$$l=10 \text{ and } \lambda=0.04l$$
$$\therefore \lambda=0.4$$

Write down our distribution using correct notation

$$X\sim Po(0.4)$$

Write the probability required by the question

$$P(X<2)$$

From earlier equations:

$$P(X<2)=e^{-0.4}\left(\frac{0.4^0}{0!}+\frac{0.4^1}{1!}\right)=0.938$$

Part (ii):

Using question to form the parameters

$$l=10 \text{ and } \lambda=0.04l$$
$$\therefore \lambda=40>15$$

Thus we can use the normal approximation

Write down our distribution using correct notation

$$X\sim Po(40)\rightarrow Y\sim N(40,40)$$

Write the probability required by the question

$$P(X\geq 46)$$

Apply continuity correction for the normal distribution

$$P(Y\geq 45.5)$$

Evaluate the probability

$$P(Y\geq 45.5)=1-\Phi\left(\frac{45.5-40}{\sqrt{40}}\right)=0.192$$

Part (iii):

Using the variance formula

$$Var(X)=E(X^2)-\left(E(X)\right)^2$$

For a Poisson distribution

$$E(X)=Var(X)=\lambda \text{ and } \lambda=40$$

Substitute into equation and solve for the unknown

$$\therefore 40=E(X^2)-40^2$$
$$E(X^2)=1640 \text{ pence}$$
$$E(X^2)=£16.40$$

Expected cost for rectifying cloth is £16.40

2. Linear Combinations of Random Variables

2.1 Expectation & Variance of a Function of X

$$E(aX+b) = aE(X)+b$$
$$Var(aX+b) = a^2 Var(X)$$

(IS) Ex 6a: **Question 12:**
The random variable T has mean 5 and variance 16. Find two pairs of values for the constants c and d such that $E(cT+d) = 100$ and $Var(cT+d) = 144$

Solution:
Expand expectation equation:
$$E(cT+d) = cE(T)+d = 100$$
$$\therefore 5c+d = 100$$
Expand variance equation:
$$Var(cT+d) = c^2 Var(T) = 144$$
$$16c^2 = 144$$
$$c = \pm 3$$
Use first equation to find two pairs:
$$c = 3, \quad d = 85c = -3, \quad d = 115$$

2.2 Combinations of Random Variables

- Expectations of combinations of random variables:
$$E(aX+bY) = aE(X)+bE(Y)$$

- Variance of combinations of independent random variables:
$$Var(aX+bY+c) = a^2 Var(X) + b^2 Var(Y)$$
$$Var(X \pm Y) = Var(X) + Var(Y)$$

- Combinations of identically distributed random variables having mean μ and variance σ^2

$E(2X) = 2\mu$ and $E(X_1)+E(X_2) = 2\mu$
$Var(2X) = 4\sigma^2$ but $Var(X_1+X_2) = 2\sigma^2$

(IS) Ex 6b: **Question 3:**
It is given that X_1 and X_2 are independent, and $E(X_1) = E(X_2) = \mu$, $Var(X_1) = Var(X_2) = \sigma^2$
Find $E(\overline{X})$ and $Var(\overline{X})$, where $\overline{X} = \frac{1}{2}(X_1+X_2)$

Split the variance into individual components
$$Var\left(\frac{1}{2}(X_1+X_2)\right) = \left(\frac{1}{2}\right)^2 Var(X_1) + \left(\frac{1}{2}\right)^2 Var(X_2)$$
Substitute given values, hence
$$Var\left(\frac{1}{2}(X_1+X_2)\right) = \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{1}{2}\sigma^2$$

2.3 Expectation & Variance of Sample Mean

$$E(\overline{X}) = \mu \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

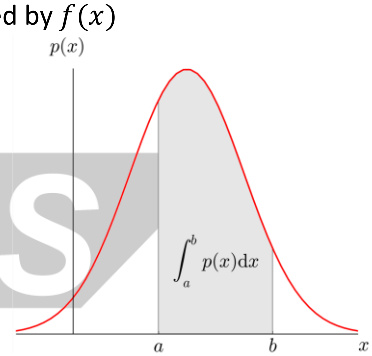
(IS) Ex 6c: **Question 5:**
The mean weight of a soldier may be taken to be 90kg, and $\sigma = 10$kg. 250 soldiers are on board an aircraft, find the expectation and variance of their weight. Hence find the μ and σ of the total weight of soldiers.

Solution:
Let X be the average weight, therefore
$$E(\overline{X}) = \mu = 90$$
$$Var(\overline{X}) = \frac{\sigma^2}{n} = \frac{10^2}{250} = 0.4 \text{ kg}^2$$
To find μ of total weight, you are calculating
$$E(X_1)+E(X_2) \ldots + E(X_{250}) = 250E(X) = 22\,500\text{kg}$$
To find σ, first find $Var(X)$
$$Var(X_1) \ldots + Var(X_{250}) = 250Var(X) = 2500\text{kg}$$
$$Var(X) = \sigma^2 = 25000$$
$$\therefore \sigma = \sqrt{25000} = 158.1\text{kg}$$

3. Continuous Random Variables

3.1 Probability Density Functions (pdf)

- Function whose area under its graph represents probability used for continuous random variables
- Represented by $f(x)$



Conditions:

- Total area always $= 1$
$$\int_c^d f(x)\ dx = 1$$
- Cannot have negative probabilities $\therefore$ graph cannot dip below x-axis; $f(x) \geq 0$
- Probability that X lies between a and b is the area from a to b
$$P(a < X < b) = \int_a^b f(x)\ dx$$
- Outside given interval $f(x) = 0$; show on a sketch
- $P(X = b)$ always equals 0 as there is no area

- Notes:
  - $P(X < b) = P(X \leq b)$ as no extra area added
  - The mode of a pdf is its maximum (stationary point)

(IS) Ex 9a: **Question 6:**

Given that:

$$f(x) = \begin{cases} kx(6-x) & 2 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$

i. Find the value of k

ii. Find the mode, m

iii. Find $P(X < m)$

Solution:

Part (i):

Total area must equal 1 hence

$$\int_2^5 kx(6-x) = \left[3kx^2 - \frac{kx^3}{3}\right]_2^5 = 1$$

$$= 75k - \frac{125}{3}k - 12k + \frac{8}{3}k = 24k = 1$$

$$\therefore k = \frac{1}{24}$$

Part (ii):

Mode is the value which has the greatest probability hence we are looking for the max point on the pdf

$$\frac{d}{dx}[kx(6-x)] = 6k - 2kx$$

Finding max point hence stationary point

$$6k - 2kx = 0$$

$$x = \frac{6\left(\frac{1}{24}\right)}{2\left(\frac{1}{24}\right)} = 3$$

$$\therefore \text{mode} = 3$$

Part (iii):

$P(X < m)$ can be interpreted as $P(-\infty < X < m)$

$$\int_{-\infty}^{m} kx(6-x) = \int_2^3 kx(6-x) = \left[3kx^2 - \frac{kx^3}{3}\right]_2^3$$

$$= \frac{1}{24}\left(3(3^2) - \frac{3^3}{3} - 3(2^2) + \frac{2^3}{3}\right) = \frac{13}{36}$$

3.2 Mean & Variance

- To calculate mean/expectation

$$E(X) = \int_{-\infty}^{\infty} xf(x)\ dx$$

- To calculate variance:
  - First calculate $E(X)$ then $E(X^2)$ by

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\ dx$$

  - Substitute information and calculate using

$$Var(X) = E(X^2) - E(X)^2$$

3.3 The Median

The Cumulative Distribution Function (cdf)

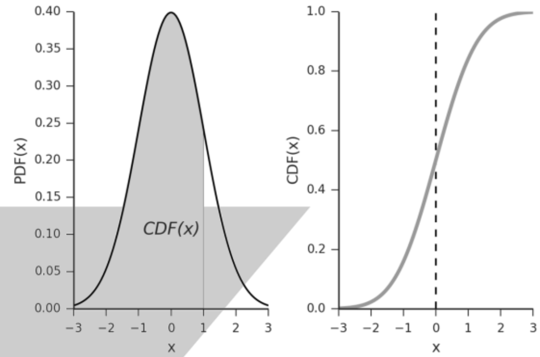
- Gives the probability that the value is less than b

$$P(X < b) \quad \text{or} \quad P(X \leq b)$$

- Represented by $F(b)$
- It is the integral of $f(x)$

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

- **Median:** the value of b for which $F(b) = 0.5$



4. Sampling & Estimation

4.1 Sample & Population

- **Population:** collection of all items
- **Sample:** subset of population used as a representation of the entire population

4.1 Central Limit Theorem

- If $(X_1, X_2, \dots, X_n)$ is a random sample of size n drawn from any population with mean μ and variance σ^2 then the sample has:

Expected mean, μ

Expected variance, $\frac{\sigma^2}{n}$

It forms a normal distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

(IS) Ex 10f: **Question 12:**

The weights of the trout at a trout farm are normally distributed with mean 1kg & standard deviation 0.25kg

a. Find, to 4 decimal places, the probability that a trout chosen at random weighs more than 1.25kg.

b. If $\bar{Y}$kg represents mean weight of a sample of 10 trout chosen at random, state the distribution of $\bar{Y}$: evaluate the mean and variance.
Find the probability that the mean weight of a sample of 10 trout will be less than 0.9kg

Solution:

Part (a):

Write down distribution

$$X \sim N(1, 0.25^2)$$

Write down the probability they want

$$P(X > 1.25) = 1 - P(X < 1.25)$$

Standardize and evaluate

$$1 - P\left(Z < \frac{1.25 - 1}{0.25}\right) = 0.1587$$

Part (b):

Write down initial distribution

$$X \sim N(1, 0.25^2)$$

For sample, mean remains equal but variance changes

Find new variance

$$\text{Variance of sample} = \frac{\sigma^2}{n} = \frac{0.25^2}{10} = 0.00625$$

Write down distribution of sample

$$\bar{Y} \sim N(1, 0.00625)$$

Write down the probability they want

$$P(\bar{Y} < 0.9)$$

Standardize and evaluate

Standardized probability is negative so do 1 minus

$$P\left(Z < \frac{0.9 - 1}{0.00625}\right) = 1 - P\left(Z < \frac{0.1}{0.00625}\right) = 0.103$$

4.2 Point Estimate & Confidence Interval

- A **point estimate** is a numerical value calculated from a set of data (sample) which is used as an estimate of an unknown parameter in a population
- Examples of point estimates are:

$$\text{Sample mean } \bar{x} \xrightarrow{estimates} \text{population mean } \mu$$

$$\text{Sample proportion } \frac{r}{n} \xrightarrow{estimates} \text{population proportion } p$$

$$\text{Sample variance } s^2 \xrightarrow{estimates} \text{population variance } \sigma^2$$

- Point estimate close to population value but not exact
- We can determine a **confidence interval** where the population value is likely to lie in $(\bar{x} - \delta,\ \bar{x} + \delta)$

4.3 The Variance

- Variance can be calculated/given for either a sample or a population and there is a difference between them

Using the divisor n

- This is appropriate to use when
  - data is given for the whole population and you are interested in the variance of the whole
  - data is given for the sample and you are interested in the variance of just the sample

$$\sigma^2 = \frac{1}{n}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$$

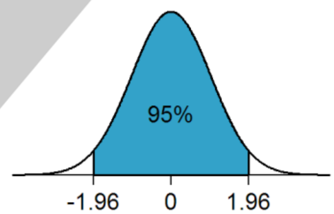
Using the divisor $(n - 1)$

- Appropriate to use when data is given for a sample and you are estimating variance of the whole population
- The quantity calculated s^2 is known as the **unbiased estimate of the population variance**

$$s^2 = \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$$

4.4 Percentage Points for a Normal Distribution

- The percentage points are determined by finding the z-value of specific percentages
- E.g. to find the z-value of a 95% confidence level, we can see that the 5% would be removed equally from both sides (2.5%) so the z-value we would actually be finding would be of $100\% - 2.5\% = 97.5\%$



Percentage Points Table

Confidence level	90%	95%	98%	99%
***z*-value**	1.645	1.960	2.326	2.576

4.5 Confidence Interval for a Population Mean

Sample taken from a normal population distribution with known population variance

$$\left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$$

- z is the value corresponding to the confidence level required and n is the sample size
- The confidence interval calculated is exact

Large sample taken from an unknown population distribution with known population variance

- By the Central Limit Theorem, the distribution of $\bar{X}$ will be approximately normal so same method as above

$$\left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$$

- The confidence interval calculated is an approximate

Large sample taken from an unknown population distribution with unknown population variance

- As the population variance is unknown, you must first estimate the population variance, s, using sample data

$$\left(\bar{x} - z\frac{s}{\sqrt{n}}, \bar{x} + z\frac{s}{\sqrt{n}}\right)$$

- The confidence interval calculated is an approximate

{W13-P71}: **Question 2:**

Heights of a certain species of animal are normally distributed with $\sigma = 0.17$m. Obtain a 99% confidence interval for the population mean, with total width less than 0.2m. Find the smallest sample size required.

Solution:

For a 99% confidence interval, find z where $\Phi(z) = 0.995$ (think of the 1% cut from both sides)

$$z = 2.576$$

Subtract the limits of the interval and equate to 0.2

$$\left(\bar{x} + z\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{x} - z\frac{\sigma}{\sqrt{n}}\right) = 0.2$$

$$2\left(z\frac{\sigma}{\sqrt{n}}\right) = 0.2$$

Substitute information given and find n

$$\sqrt{n} = \frac{0.2}{2 \times 2.576} \times 0.17$$

$$n = 4126.53 \approx 4130$$

4.6 Confidence Interval for a Population Proportion

- Calculating the confidence interval from a random sample of n observations from a population in which the proportion of successes is p and the proportion of failures is q
- The observed proportion of success $\hat{p}$ is $\frac{r}{n}$ where r represents the number of successes

$$\left(\hat{p} - z\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

{S10-P71}: **Question 2:**

A random sample of n people were questioned about their internet use. 87 of them had a high-speed internet connection. A confidence interval for the population proportion having a high-speed internet connection is $0.1129 < p < 0.1771$.

i. Write down the mid-point of this confidence interval and hence find the value of n.

ii. This interval is an α% confidence interval. Find α.

Solution:

Part (i):

Find the midpoint of the limits, finding p

$$0.1129 + \frac{0.1771 - 0.1129}{2} = 0.145$$

The midpoint is equal to the proportion of people with high-speed internet use so

$$\frac{87}{n} = 0.145 \qquad \therefore n = 600$$

Part (ii):

Using the upper limit, this was calculate by:

$$0.1771 = 0.145 + z\sqrt{\frac{pq}{n}}$$

Substituting values calculated ($q = 1 - p$), find z

$$0.0321 = z\sqrt{\frac{\frac{87}{600} \times \frac{513}{600}}{600}} \qquad \therefore z = 2.233$$

Use normal tables and find corresponding probability

$$\Phi(z) = 0.9872$$

Think of symmetry, the same area is chopped off from both sides of the graph so

$$1 - 2(1 - 09872) = 0.9744$$

Hence the α% confidence is = 97.44%

5. Hypothesis Tests

5.1 Null & Alternative Hypothesis

- For a hypothesis test on the population mean μ, the **null hypothesis H_0** proposes a value μ_0 for μ

$H_0: \mu = \mu_0$

- The **alternative hypothesis H_1** suggests the way in which μ might differ from μ_0. H_1 can take three forms:

$H_1: \mu < \mu_0$, a one-tail test for a decrease

$H_1: \mu > \mu_0$, a one-tail test for an increase

$H_1: \mu \neq \mu_0$, a two-tail test for a difference

- The **test statistic** is calculated from the sample. Its value is used to decide whether the null hypothesis should be rejected
- The **rejection** or **critical region** gives the values of the test statistic for which the null hypothesis is rejected
- The **acceptance region** gives the values of the test statistic for which the null hypothesis is accepted
- The **critical values** are the boundary values of the rejection region
- The **significance level** of a test gives the probability of the test statistic falling in the rejection region

To carry out a hypothesis test:

- Define the null and alternative hypotheses
- Decide on a significance level
- Determine the critical value(s)
- Calculate the test statistic
- Decide on the outcome of test depending on whether value of test statistic lies in rejection/acceptance region
- State the conclusion in words
- The test statistic Z can be used to test a hypothesis about a population

$$z = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

where μ is the population mean specified by H_0

- The critical values for some commonly used rejection regions:

Significance level	Two-tail $\mu \neq \mu_0$	One-tail $\mu > \mu_0$	$\mu < \mu_0$
10%	± 1.645	1.282	-1.282
5%	± 1.960	1.645	-1.645
2%	± 2.326	2.054	-2.054
1%	± 2.576	2.326	-2.326

5.2 Testing Different Distributions

- **Test for mean, known variance, normal distribution or large sample**

$$X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  - Use general procedure as outlined above
- **Test for mean, large sample, variance unknown**

$$X \sim N\left(\mu, \frac{s^2}{n}\right)$$

  - Use the same procedure however must use unbiased estimate of the population variance, s
- **Test for large Poisson mean**

$$X \sim N\left(\lambda, \frac{\lambda}{n}\right)$$

  - Use general procedure but must approximate normal distribution using the mean given
  - Must apply continuity correction
- **Test for proportion, large sample** (Binomial distribution)

$$X \sim N\left(p, \frac{pq}{n}\right)$$

  - Similar to Poisson approximation; using probability of success and applying continuity correction

5.3 Type I and Type II Errors

- A **Type I error** is made when a true null hypothesis is rejected
- A **Type II error** is made when a false null hypothesis is accepted

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

- **P(Type I error)** = significance level
- **Calculating P(Type II error):**
  - Firstly, calculate the acceptance region by leaving $\bar{x}$ as a variable and equating the test statistic to the significance level
  - Next, calculate the conditional probability that μ is now μ' and $\bar{x}$ is still in the acceptance region

$$\text{P}(\bar{x} \text{ is in acceptance region} \mid \mu = \mu')$$

  Calculate this by substituting the limit of the acceptance region as $\bar{x}$ (calculated previously) and the new, given μ' into the test statistic equation and find the probability

CIE A-LEVEL MATHEMATICS//9709

WWW.ZNOTES.ORG