ZNOTES // A-LEVEL SERIES

visit www.znotes.org

CIE A-LEVEL FURTHER MATHS 9231 (FS)

FORMULAE & SOLVED QUESTIONS FOR FURTHER STATISTICS

TABLE OF CONTENTS

1. Expectation Algebra

1.1 Expectation & Variance of a Function of X

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

(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:

$$Var(cT+d)=c^2Var(T)=144$$
$$16c^2=144$$
$$c=\pm 3$$

Use first equation to find two pairs:

$$c=3,\qquad d=85c=-3,\qquad d=115$$

1.2 Combinations of Random Variables

- Expectations of combinations of random variables:

$$E(aX+bY)=aE(X)+bE(Y)$$

- Variance of combinations of independent random variables:

$$Var(aX+bY+c)=a^2Var(X)+b^2Var(Y)$$
$$Var(X\pm Y)=Var(X)+Var(Y)$$

- 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 $\overline{X}=\frac{1}{2}(X_1+X_2)$

Solution:

Split the expectation into individual components

$$E\left(\frac{1}{2}(X_1+X_2)\right)=\frac{1}{2}E(X_1)+\frac{1}{2}E(X_2)$$

Substitute given values, hence

$$E\left(\frac{1}{2}(X_1+X_2)\right)=\frac{1}{2}\mu+\frac{1}{2}\mu=\mu$$

Split the variance into individual components

$$Var\left(\frac{1}{2}(X_1+X_2)\right)=\left(\frac{1}{2}\right)^2Var(X_1)+\left(\frac{1}{2}\right)^2Var(X_2)$$

Substitute given values, hence

$$Var\left(\frac{1}{2}(X_1+X_2)\right)=\frac{1}{4}\sigma^2+\frac{1}{4}\sigma^2=\frac{1}{2}\sigma^2$$

1.3 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 $Var(X)$

$$Var(X_1)\ldots+Var(X_{250})=250Var(X)=2500\text{kg}$$
$$Var(X)=\sigma^2=25000$$
$$\therefore \sigma=\sqrt{25000}=158.1\text{kg}$$

2. 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

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

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

- A Poisson distribution with mean λ can be noted as

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

2.1 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

2.2 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

2.3 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

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

$$= 0.819$$

2.4 Relationship of Inequalities

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

2.5 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

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

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

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

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

2.6 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 information from question form the parameters of Poisson distribution

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

$$\therefore \lambda = 40 > 15$$

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

$$\therefore 40 = E(X^2) - 40^2$$

$$E(X^2) = 1640 \text{ pence}$$

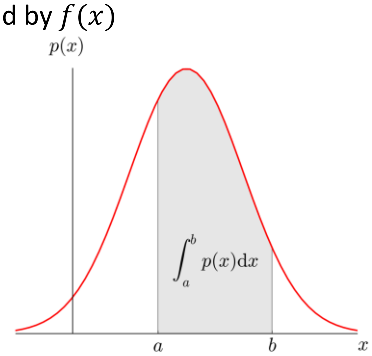
$$E(X^2) = £16.40$$

Expected cost for rectifying cloth is £16.40

3. Continuous Random Variable

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 –ve probabilities ∴ 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 Cumulative Distribution Function (cdf)

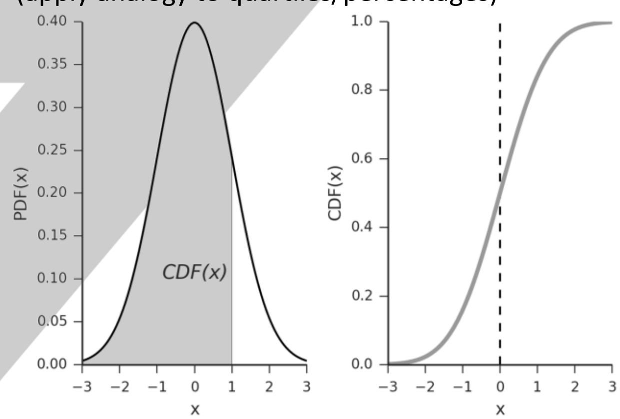
- Gives the probability that the value is less than x

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

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

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

- Median: the value of x for which $F(x) = 0.5$ (apply analogy to quartiles/percentages)



- Notes:
  - Since it is always impossible to have a value of X smaller than $-\infty$ or larger than ∞:

$$F(-\infty) = 0 \qquad F(\infty) = 1$$

  - As x increase, $F(x)$ either increase or remains constant, but never decreases.
  - F is a continuous function even if f is discontinuous
- Useful relations:
  - $P(c < X < d) = F(d) - F(c)$
  - $P(X > x) = 1 - F(x)$

(IS) Ex 9b: **Question 9:**

Given that:

$$f(x) = \begin{cases} k & 0 < x < 1 \\ 4k & 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

i. Find the value of k

ii. Find $F(x)$

iii. Find the difference between the median and the fifth percentile of X

Solution:

Part (i):

Total area must equal 1 hence

$$\int_0^1 k + \int_1^3 4k = [kx]_0^1 + [4kx]_1^3 = 1$$
$$= (k-0) + (12k - 4k) = 9k = 1$$
$$\therefore k = \frac{1}{9}$$

Part (ii):

Integrate each case separately from its $-\infty$ to x

For the first interval $0 \le x \le 1$

$$F(x) = \int_0^x \frac{1}{9} = \left[\frac{1}{9}x\right]_0^x = \frac{1}{9}x$$

We must split next interval $0 \le x \le 3$ as

$$F(x) = P(X \le 3) = P(X \le 1) + P(1 \le x \le 3)$$

and $P(X \le 1) = F(1) = {}^{1}/_{9}$

$$\therefore F(x) = \frac{1}{9} + \int_1^x 4 \times \frac{1}{9}$$
$$= \frac{1}{9} + \left[4 \times \frac{1}{9}x\right]_1^x = \frac{4}{9}x - \frac{3}{9}$$

Writing in correct notation and fixing intervals (adding equal sign to inequalities)

$$F(x) = \begin{cases} 0 & x \le 0 \\ \frac{1}{9}x & 0 \le x \le 1 \\ \frac{4}{9}x - \frac{3}{9} & 1 \le x \le 3 \\ 1 & x \ge 3 \end{cases}$$

Part (iii):

Finding the median, you must check in which interval it lies. Do this by substituting the maximum value for x in the first case

$$\frac{1}{9} \times 1 = \frac{1}{9} < \frac{1}{2}$$

This means the median does not lie in this interval $\therefore$

$$\frac{4}{9}x - \frac{3}{9} = 0.5$$
$$x = \frac{15}{8}$$

The fifth percentile lies in the first interval as $\frac{1}{20} < \frac{1}{9}$ so

$$\frac{1}{9}x = \frac{1}{20}$$
$$x = \frac{9}{20}$$

Find the difference

$$\frac{15}{8} - \frac{9}{20} = \frac{57}{40}$$

3.3 Expectation and Variance

- To calculate expectation

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

- To calculate variance:
  - First calculate $E(X)$ as above
  - The calculate $E(X^2)$ by

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

  - Substitute information and calculate using

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

3.4 Obtaining f(x) from F(x)

- As F is obtained by integrating f, then f can be obtained by differentiating F

(IS) Ex 9d: **Example 13:**

The random variable has cdf given by

$$F(x) = \begin{cases} 0 & x \le 1 \\ \dfrac{(x-1)^3}{8} & 1 \le x \le 3 \\ 1 & x \ge 3 \end{cases}$$

Find the form of the pdf of X

Solution:

$F(x)$ is unchanging for $x < 1$ and for $x > 3$, therefore $f(x)$ is equal to 0. Hence we must find differentiate in the interval $1 < x < 3$

$$f(x) = F'(x)$$
$$f(x) = \frac{d}{dx}\left(\frac{(x-1)^3}{8}\right) = \frac{3}{8}(x-1)^2$$

Hence:

$$f(x) = \begin{cases} \frac{3}{8}(x-1)^2 & 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

3.5 Distribution of a Function of a Random Variable

- We can deduce the distribution of a simple function of X either increasing or decreasing with this procedure:

$$\mathrm{f_X} \to \mathrm{F_X} \to \mathrm{F_Y} \to \mathrm{f_Y}$$

(IS) Ex 9e: **Example 15:**

The random variable X has pdf $f_X(x)$ given by,

$$f_X(x) = \begin{cases} 1 & 2 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

The random variable Y is given by $Y = 2X + 3$.

Determine the pdf and cdf of Y.

Solution:

First step is to find $F_X(x)$ and suppose we do,

$$F_X(x) = P(X \le x) = \begin{cases} 0 & x \le 2 \\ x-2 & 2 \le x \le 3 \\ 1 & x \ge 3 \end{cases}$$

Find the ranges for Y

$$(2\times 2) + 3 \le y \le (3\times 2) + 3$$
$$7 \le y \le 9$$

Convert cdf from X to Y using relationship given

$$F_Y(y) = P(Y \le y) = P(2X + 3 \le y)$$
$$= P\left(X \le {}^1/_2\,(y-3)\right)$$

Now substitute ${}^1/_2\,(y-3)$ for x in cdf function

$$\left({}^1/_2\,(y-3)\right) - 2 \Rightarrow {}^1/_2\,(y-7)$$

Expressing cdf of Y with ranges worked out

$$F_Y(y) = P(Y \le y) = \begin{cases} 0 & x \le 7 \\ {}^1/_2\,(y-7) & 7 \le y \le 9 \\ 1 & x \ge 9 \end{cases}$$

Differentiate function to find pdf

$$f_Y(y) = \begin{cases} {}^1/_2 & 7 < y < 9 \\ 0 & \text{otherwise} \end{cases}$$

- Method can be used for both increasing and decreasing functions as well functions with powers (e.g. $W = X^2$)

4. Geometric & Exponential Distribution

4.1 Geometric Distribution

Conditions for a Geometric Distribution:

- Only two possible outcomes: success or failure
- Probability of success, p, is constant
- Each event is independent

- The geometric distribution is used to find the number of trials required to obtain the first success

  $$P(X = n) = (1-p)^{n-1}p \qquad n = 1, 2, 3, \ldots$$

  Where p is the probability of success, $(1-p)$ is the probability of failure and n is the number of trials
- A geometric distribution with probability of success p can be noted as

  $$X \sim Geo(p)$$
- The distribution is called geometric because successive probabilities, p, $(1-p)p$, $(1-p)^2p$… form a geometric progression with first term p and common ratio $(1-p)$

4.2 Cumulative Probabilities

- Calculating cumulative probabilities

$$P(X \le x) = 1 - (1-p)^x \qquad P(X \ge x) = (1-p)^{x-1}$$
$$P(X < x) = 1 - (1-p)^{x-1} \qquad P(X > x) = (1-p)^x$$

Example:

In the village of Nanakuli, about 80% of the residents are of Hawaiian ancestry. Suppose you fly to Hawaii and visit Nanakuli.

i. What is the probability that the fifth villager you meet is Hawaiian?
ii. What is the probability that you do not meet a Hawaiian until the third villager?

Solution:

Part (i):

Using the formula

$$P(X = 5) = (1 - 0.80)^{5-1}(0.80) = 0.00128$$

Part (ii):

Not meeting until third means the probability

$$P(X > 3)$$

Using relationships above

$$P(X > 3) = (1 - 0.80)^3 = 0.008$$

4.3 Mean & Variance of a Geometric Distribution

- The expectation (mean) of a geometric distribution:

  $$E(X) = \frac{1}{p}$$
- The variance of a geometric distribution:

  $$Var(X) = \frac{1-p}{p^2}$$

4.4 Exponential Distribution

- Used for modeling duration of events

  $$P(X < x) = 1 - e^{-\lambda x}$$
  $$P(X > x) = e^{-\lambda x}$$
  $$P(a < X < b) = e^{-\lambda a} - e^{-\lambda b}$$

  Where λ is the average no. of events in 1 unit of time and x is the duration
- An exponential distribution with average λ can be noted:

  $$X \sim Exp(\lambda)$$
- The exponential distribution is memory-less

  $$P[X > (a+b) | X > a] = P(X > b)$$
  - e.g. if a motor has been running for 3 hours and you are asked to calculate the probability of it running for more than 4 hours, you only need to find the probability of it running for the next hour as the previous condition does not affect the probability

4.5 Mean & Variance of an Exponential Distribution

- The expectation (mean) of an exponential distribution:

$$E(X) = \frac{1}{\lambda}$$

- The variance of an exponential distribution:

$$Var(X) = \frac{1}{\lambda^2}$$

Example:

Calls arrive at an average rate of 12 per hour. Find the probability that a call will occur in the next 5 minutes given that you have already waited 10 minutes.

Solution:

Interpreting the information,

$$\lambda = 12 \text{ per hour} = 0.2 \text{ per minute}$$

We are being asked to calculate

$$P(T \le 15 | T > 10)$$

As the exponential distribution is memory-less; the previous condition does not affect it hence we are simply being asked to find $P(T \le 5)$

$$P(T \le 5) = 1 - e^{-0.2\times 5} = 0.63$$

5. Sampling & Central Limit Theorem

5.1 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:

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

(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

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

6. Point and Interval Estimation

6.1 The Variance

- The 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 $(n-1)$

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

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

6.2 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:

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

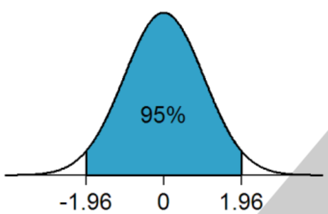
$$\text{Sample proportion } \frac{r}{n} \xrightarrow{estimates} \text{population proportion } p$$

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

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

6.3 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

6.4 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

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

$$n = 4126.53 \approx 4130$$

6.5 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 $\alpha\%$ 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

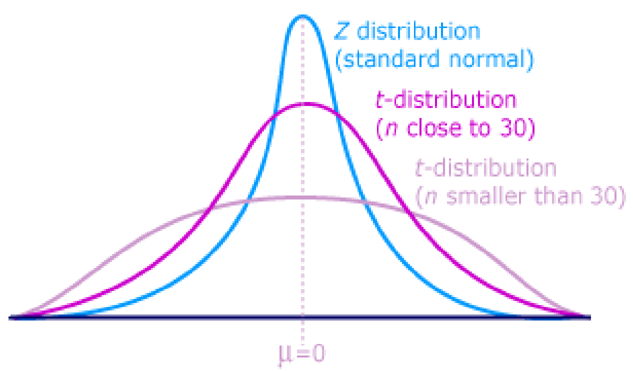
$$\Phi(z) = 0.9872$$

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

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

Hence the $\alpha\%$ confidence is $= 97.44\%$

6.6 Percentage Points for a t-Distribution



6.7 Confidence Interval for a Population Mean with a Small Sample

Small sample (<30) taken from a Normal population distribution with unknown population variance

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

- As sample is small, the normal distribution cannot be used and instead the t-distribution is used
- For a small sample n, its t-distribution is t_{n-1} (degree of freedom $v = n - 1$
- Use the tables to find the percentage point, c
- As the population variance is unknown, you must estimate the population variance, s, using sample data
- The confidence interval calculated is exact

7. Hypothesis Tests

7.1 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 the test depending on whether the value of the test statistic lies in the rejection or 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 the null hypothesis

- The critical values for some commonly used rejection regions:

Significance level	Two-tail $\mu \neq \mu_0$	One-tail $\mu > \mu_0$	One-tail $\mu < \mu_0$
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

7.2 Hypothesis Testing with 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
- **Test for mean, small sample, variance unknown**

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

  - Firstly, you must estimate the variance, calculate s
  - The distribution of the corresponding random variable, T, is t_{n-1} (i.e. one less than sample size n)

7.3 Hypothesis Tests and Confidence Interval

- If a $c\%$ symmetric confidence interval excludes the population value of interest, then the null hypothesis that the population parameter takes this value will be rejected at the $100(1-c)\%$ level

7.4 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

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

    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

7.5 Comparison of Two Means

- When testing the hypothesis that two population have the same mean
- Two cases when comparing two population means:
  - Population variances are known
  - Although population variances unknown, they can be assumed to have the same value

Known population variance

- For two random variables X and Y with unknown means μ_x and μ_y and known variances σ_x^2 and σ_y^2,
  - The null hypothesis is:

$$H_0: \mu_x = \mu_y$$

  - The alternate hypothesis can be one or two-tailed:

$$H_1: \mu_x \neq \mu_y \quad \text{or} \quad H_1: \mu_x > \mu_y$$

- When calculating the z value for the hypothesis test use the following formula:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

- Carry out hypothesis test as normal

Common unknown population variance

- We are assuming that $\sigma_x^2 = \sigma_y^2 = \sigma^2$
- To find a common variance, we calculate the **pooled estimated of the common variance** s_p^2 by:

$$s_p^2 = \frac{\sum(x-\bar{x})^2 + \sum(y-\bar{y})^2}{n_x + n_y - 2}$$

- The hypothesis are the same as above however as the variance is the same, the z value is given by:

$$z = \frac{\bar{x}-\bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}$$

- For a **small sample** size, you cannot continue to use the normal distribution and instead must use t-distribution with $n_x + n_y - 2$ degrees of freedom. The test statistic is calculated same as above.

8. Goodness of Fit

8.1 χ^2 Test

- Used to test whether a particular type of distribution is appropriate for the data given
- Test statistic involves squares – only interested in upper limit critical values
- The χ^2 test can only be used to test two lists of frequencies – the observed and the expected frequencies calculated from the hypothesis.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and expected frequencies

- When calculating, set up a table as follows

Variable	Probability	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
⋮	⋮	⋮	⋮	⋮
Total				

- If the expected frequency for a class is less than 5, then you must group this class with the next class (or two …)
- Hypothesis when testing:
  - H_0: the … distribution is a suitable model
  - H_1: the … distribution is not a suitable model

8.2 Comparing the χ^2 Value

- Once you have calculated the χ^2 value of the data given, you must then compare it to the critical values of the χ^2 distribution
- To test 5 classes at a 5% significance level, find the critical value of the χ^2 distribution at 95% with 4 degrees of freedom
- If the distribution fits, the calculate value should be less than the critical value, accepting H_0

8.3 Goodness of Fit to Prescribed Distribution Type

- This is the case where the null hypothesis states that the data has a 'particular named distribution' but does not specify all the parameters of the distribution
- You must then calculate the parameter in order to carry out the test e.g.
  - Normal: mean and estimated sample variance
  - Poisson: mean
  - Binomial: probability of success
- For k parameters calculated from the observed data, you must subtract k from the degrees of freedom v
- Hence, with m different outcomes,

$$v = m - 1 - k$$

8.4 Contingency Table

- This is a table which contains the frequencies for two or more variables.
- You may then assess whether the variables are associated or independent.
- Hypothesis when testing:
  - H_0: the variables are independent
  - H_1: the variables are associated
- For example:

	A	***B***	***C***	
X				$\sum R_1$
Y				$\sum R_2$
Z				$\sum R_3$
	$\sum C_1$	$\sum C_2$	$\sum C_3$	$\sum$

- The expectation of each variable is calculated by

$$\frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

- List each variable and set up table as before
- The degree of independence for an r by c table is

$$v = (r-1)(c-1)$$

9. Regression and Correlation

9.1 Regression

- This is finding a linear relationship between two variables where one variable is dependent on the other e.g. y on x
- The **regression line** is the line summarizing the relation between x and y
- The line must pass through the mean values i.e. $\bar{x}$ and $\bar{y}$ hence the line of the equation can be written as

$$\bar{y} = a + b\bar{x}$$

  where b is the **regression coefficient**
- Rearranging equation, the value of a can be calculated

$$a = \bar{y} - b\bar{x}$$

$$a = \frac{1}{n}(\sum y - b\sum x)$$

9.2 Calculating the Regression Coefficient

- The value of b can be calculated using the method of least squares where

$$b = \frac{S_{xy}}{S_{xx}}$$

- Where the quantities S_{xy} and S_{xx} are given by

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} \qquad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

9.3 Two Regression Lines

- When both X and Y are random variables, there are two regression models:

y on x	x on y
$y = a + bx$	$x = c + dy$

- The two regression lines both pass through the point $(\bar{x}, \bar{y})$ which is therefore the point of intersection
- To predict a value of x when, for the given data, the x values are fixed (as opposed to being observations of a random variable), then it is appropriate to use the regression line of y on x 'in reverse' rather than using the regression line of x on y

9.4 Correlation

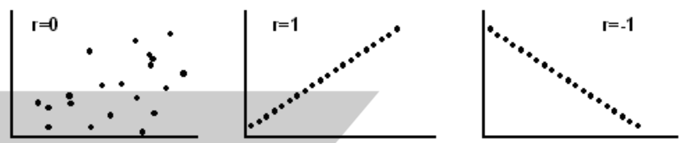
- Used when both X and Y are random variables
- The **correlation coefficient** is a number between −1 and +1 calculated so as to represent the linear dependence of two variables or sets of data
- **Positive correlation;** correlation coefficient > 0; regression lines of Y on X and X on Y have +ve gradients
- **Negative correlation;** correlation coefficient < 0; regression lines of Y on X and X on Y have -ve gradients
- **Zero correlation:** no linear relationship, does not mean X and Y are unrelated (e.g. parabolic relationship)

9.5 Product-Moment Correlation Coefficient

- Is the measurement of scatter that lies between -1 and 1

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

- Correlation graphs:



- Relationship between r and the regression coefficients:

$$r^2 = b_1 b_2$$

CIE A-LEVEL FURTHER MATHEMATICS//9231