Statistics 2

Revision Notes

June 2012

1 The Binomial distribution

Factorials

n objects in a row can be arranged in $n!$ ways.

$n! = n(n-1)(n-2)(n-3) \times \ldots. \times 4 \times 3 \times 2 \times 1$

Note that $0!$ is defined to be 1. This fits in with formulae for combinations.

Combinations

The number of ways we can choose r objects from a total of n objects, where the order does not matter, is called the *number of combinations* of r objects from n and is written as

$${}^nC_r = \frac{n(n-1)(n-2)\ldots \text{ as far as } r \text{ terms}}{r!} = \frac{n(n-1)(n-2)\ldots (n-r+1)}{r!}$$

or $${}^nC_r = \frac{n!}{(n-r)!\,r!}.$$

We can think of this as n choose r.

Example: Find the number of hands of 4 cards which can be dealt from a pack of 10.

Solution: In a hand of cards the order does not matter, so this is just the number of combinations of 4 from 10, or 10 choose 4

$${}^{10}C_4 = \frac{10\times9\times8\times7}{4!} = 210$$ notice 4 terms on top of the fraction

Properties of ${}^nC_r$

1. ${}^nC_0 = {}^nC_n = 1$
2. ${}^nC_r = {}^nC_{n-r}$ The number of ways of choosing r is the same as the number of ways of rejecting $n - r$.

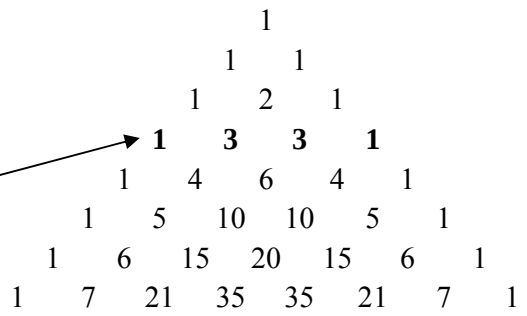
Binomial Theorem

Binomial coefficients

We can show that

$(p + q)^3 = \mathbf{1}p^3 + \mathbf{3}p^2q + \mathbf{3}pq^2 + \mathbf{1}q^3$.

The numbers 1, 3, 3, 1 are called the binomial coefficients and are the numbers in the 'third' row of Pascal's Triangle.



To write down the expansion of $(p + q)^6$

We write down the terms in a logical order $p^6 \quad p^5q \quad p^4q^2 \quad p^3q^3 \quad p^2q^4 \quad pq^5 \quad q^6$

then use the numbers in the '6th' row of the triangle $\mathbf{1} \quad \mathbf{6} \quad \mathbf{15} \quad \mathbf{20} \quad \mathbf{15} \quad \mathbf{6} \quad \mathbf{1}$

to give $(p + q)^6 = \mathbf{1}p^6 + \mathbf{6}p^5q + \mathbf{15}p^4q^2 + \mathbf{20}p^3q^3 + \mathbf{15}p^2q^4 + \mathbf{6}pq^5 + \mathbf{1}q^6$.

These binomial coefficients are often written $\binom{n}{r}$.

and we have $\binom{4}{0} = 1 \quad \binom{4}{1} = 4 \quad \binom{4}{2} = 6 \quad \binom{4}{3} = 4 \quad \binom{4}{4} = 1$

Binomial coefficients and combinations

The binomial coefficients $\binom{n}{r}$ are equal to the number of combinations ${}^nC_r$

$\Rightarrow \binom{n}{r} = {}^nC_r = \frac{n!}{(n-r)!\,r!}$

Binomial distribution B(n, p)

Conditions for a Binomial Distribution

A single trial has exactly two possible outcomes – *success* and *failure*.
This trial is repeated a *fixed number*, n, times.
The n trials are *independent*.
The *probability* of success, p, *remains the same* for each trial.

The probability of success in a single trial is usually taken as p and the probability of failure as q.
Note that $p + q = 1$.

Example: 10 dice are rolled. Find the probability that there are 4 sixes.

Solution: If X is the number of sixes then $X \sim \text{B}\left(10, \frac{1}{6}\right)$

We could have $6,6,6,6,\times,\times,\times,\times,\times,\times,$ in that order with probability $\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6$ $\times$ is 'not 6'

but the 4 sixes could appear on the 10 dice in a total of ${}^{10}C_4$ ways, each way having the same probability, giving

$P(X = 4) = {}^{10}C_4 \times \left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6 = 0{\cdot}54265875851 = 0{\cdot}543$ to 3 S.F. using calculator

In general, for $X \sim \text{B}(n, p)$

the probability of r successes is

$P(X = r) = {}^nC_r \times p^r\, q^{n-r}$, where $q = 1 - p$.

Binomial distribution

The binomial distribution $X \sim \text{B}(n, p)$ is shown below.

x	0	1	2	...	r	...	n
$P(X = x)$	${}^nC_0\, q^n$	${}^nC_1\, pq^{n-1}$	${}^nC_2\, p^2q^{n-2}$	...	${}^nC_r\, p^rq^{n-r}$	...	${}^nC_n\, p^n$

N.B. The term *probability distribution* means

the set of all possible outcomes (in this case the values of $x = 0, 1, 2, \ldots, n$)

together with their probabilities.

Example: A game of chance has probability of winning 0·73 and losing 0·27. Find the probability of winning more than 7 games in 10 games.

Solution: The number of successes is a random variable $X \sim \text{B}(10, 0{\cdot}73)$, assuming independence of trials.

$\text{P}(X > 7) = \text{P}(X = 8) + \text{P}(X = 9) + \text{P}(X = 10)$

$= {}^{10}C_8 \times 0{\cdot}73^8 \times 0{\cdot}27^2 + {}^{10}C_9 \times 0{\cdot}73^9 \times 0{\cdot}27^1 + 0{\cdot}73^{10}$

$= 0{\cdot}34709235895$

$= 0{\cdot}347$ to 3 S.F. using calculator

P(more than 7 wins in 10 games) = 0·347

Cumulative binomial probability tables

Example: For $X \sim B(30, 0{\cdot}35)$, find the probability that $7 < X \leq 12$.

Solution: A moment's thought shows that we need $P(X = 8,\ 9,\ 10,\ 11 \text{ or } 12)$

$= P(X \leq 12) - P(X \leq 7) = 0{\cdot}7802 - 0{\cdot}1238,$ using tables for $n = 30,\ p = 0{\cdot}35$

$= 0{\cdot}6564$ to 4 D.P. from tables

Example: A bag contains a large number of red and white discs, of which 85% are red. 20 discs are taken from the bag; find the probability that the number of red discs lies between 12 and 17 inclusive.

Solution: As there is a *large* number of discs in the bag, we can assume that the probability of a red disc remains the same for each trial, $p = 0{\cdot}85$.

Let X be the number of red discs $\Rightarrow X \sim B(20, 0{\cdot}85)$

We now want $P(12 \leq X \leq 17)$.

At first glance this looks simple until we realise that the tables stop at probabilities of $0{\cdot}5$.

We need to consider the number of white discs, $Y \sim B(20, 0{\cdot}15)$, where $0{\cdot}15 = 1 - 0{\cdot}85$,

For $12 \leq X \leq 17$ we have $X = 12, 13, 14, 15, 16 \text{ or } 17$ it is worth writing out the numbers

for which values $Y = 8,\ 7,\ 6,\ 5,\ 4 \text{ or } 3$

$\Rightarrow P(12 \leq X < 17) = P(3 \leq Y \leq 8)$ for $Y \sim (20, 0{\cdot}15)$

$= P(Y \leq 8) - P(Y \leq \mathbf{2}) = 0{\cdot}9987 - 0{\cdot}4049$

$= 0{\cdot}5938$ to 4 D.P. from tables

Mean and variance of the binomial distribution.

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

the expected mean is $\quad E[X] = \mu = np,$

the expected variance is $\quad \text{Var}[X] = \sigma^2 = npq = np(1-p).$

This means that if the set of n trials were to be repeated a large number, n, times and the number of successes recorded each time, $x_1, x_2, x_3, x_4 \ \ldots \ x_n$

then the mean of $x_1, x_2, x_3, x_4 \ \ldots\ x_n$ would be $\mu = np$

and the variance of $x_1, x_2, x_3, x_4 \ \ldots\ x_n$ would be $\sigma^2 = npq$

Example: A coin is spun 100 times. Find the expected mean and variance of the number of heads.

Solution: $X \sim \text{B}(100, \frac{1}{2})$

$\Rightarrow \quad \mu = np = 100 \times \frac{1}{2} = 50$

and $\sigma^2 = npq = 100 \times \frac{1}{2} \times \frac{3}{4} = 37{\cdot}5$

$\Rightarrow \quad \sigma = \sqrt{37 \cdot 5} = 6{\cdot}1237\ldots$

$\mu - 2\sigma = 50 - 12{\cdot}24\ldots = 38,$
$\mu + 2\sigma = 50 + 12{\cdot}24\ldots = 62$

So we would expect that a probability of about 0·95 that the number of heads lies between 38 and 62, inclusive. This assumes that the probability distribution is approximately Normal.

Example: It is believed that 35% of people like fish and chips. A survey is conducted to verify this. Find the minimum number of people who should be surveyed if the expected number of people who like fish and chips is to exceed 60.

Solution: If X is the number of people who like fish and chips in a sample of size n,

then $X \sim$ B(n, 0·35).

$E[x] = \mu = np = 0.35n > 60$

$\Rightarrow \quad n > \frac{60}{0{\cdot}35} = 171.4285714$

$\Rightarrow \quad n = 172,$ since the expected mean has to **exceed** 60.

2 The Poisson distribution

Conditions for a Poisson distribution

Events must occur

singly	–	two cannot occur simultaneously
uniformly	–	at a constant rate
independently and randomly	–	the occurrence of one event does not influence the occurrence of another

Examples:

a) scintillations on a geiger counter placed near a radio-active source

singly – in a very short time interval the probability of one scintillation is small and the probability of two is negligible.
uniformly – over a 'longer' period of time we expect the scintillations to occur at a constant rate
independently – one scintillation does not affect another.

b) distribution of chocolate chips in a chocolate chip ice cream

singly – in a very small piece of ice cream the probability of one chocolate chip is small and the probability of two is negligible.
uniformly – in 'larger' equal size pieces of ice cream, we expect the number of chocolate chips to be roughly constant (provided that the mixture has been well mixed).
independently – the presence of one chocolate chip does not affect the presence of another.

c) defects in the production of glass rods

singly – in a very small length of glass rod the probability of one defect is small and the probability of two is negligible.
uniformly – in 'larger' equal length sections of glass rod of the same size, we expect the number of defects to be roughly constant (provided that the molten glass has been well mixed).
independently – the presence of one defect does not affect the presence of another.

Poisson distribution

In a Poisson distribution with mean λ in an interval, the probability of r occurrences in a similar interval is

$$P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

The Poisson distribution of $X \sim \text{P}_\text{O}(\lambda)$ is shown below.

x	0	1	2	...	r	...
$P(X = x)$	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2 e^{-\lambda}}{2!}$	...	$\frac{\lambda^r e^{-\lambda}}{r!}$	...

As before, the term *probability distribution* means

the set of all possible outcomes (in this case the values of $x = 0, 1, 2, \ldots, n$)

together with their probabilities.

Notice that in a Poisson distribution, x can take any positive or zero integral value, no matter how large. In practice, the probabilities of the 'larger' values will be very, very small.

Finding probabilities for the Poisson distribution is very similar to finding probabilities for the Binomial –

you just use $\frac{\lambda^r e^{-\lambda}}{r!}$ instead of ${}^nC_r \times p^r\, q^{n-r}$,

and cumulative tables for Poisson are used in a similar way to the Binomial.

Example: Cars pass a particular point at a rate of 5 cars per minute.

(*a*) Find the probability that exactly 4 cars pass the point in a minute.

(*b*) Find the probability that between at least 3 but fewer than 8 cars pass in a particular minute.

(*c*) Find the probability that more than 8 cars pass in 2 minutes.

(*d*) Find the probability that more than 3 cars pass in each of two separate minutes.

Solution: Let X be the number of cars passing in a minute, then $X \sim \text{P}_\text{O}(5)$

(*a*) $X \sim \text{P}_\text{O}(5)$

$\Rightarrow$ $P(X = 4) = \frac{5^4 \times e^{-5}}{4!} = 0{\cdot}175467369768 = 0{\cdot}175$ to 3 S.F. using calculator

or $P(X = 4) = P(X \leq 4) - P(X \leq 3) = 0{\cdot}4405 - 0{\cdot}2650 = 0{\cdot}1755$ to 4 D.P. using tables

(*b*) P(at least 3 but fewer than 8) $= P(3 \leq X < 8)$

$= P(X \leq 7) - P(X \leq 2) = 0{\cdot}8666 - 0{\cdot}1247 = 0{\cdot}7419$ to 4 D.P. using tables

(*c*) We know that the Poisson distribution is uniform, so if a mean of 5 cars pass each minute, it means that a *mean* of 10 cars pass in a *two* minute period.

Thus, if Y is the number of cars passing in two minutes

$Y \sim \text{P}_\text{O}(10)$, and we need

$P(Y > 8) = 1 - P(Y \leq 8) = 1 - 0{\cdot}3328 = 0{\cdot}6672$ to 4 D.P. using tables

(*d*) For probability of more than 3 in one 1 minute period, we have $X \sim \text{P}_\text{O}(5)$

$\Rightarrow$ $P(X > 3) = 1 - P(X \leq 2) = 1 - 0{\cdot}1247 = 0{\cdot}8753$ from tables

$\Rightarrow$ probability of this happening in two separate minutes

is $0{\cdot}8753^2 = 0{\cdot}76615009 = 0{\cdot}766$ to 3 S.F. using calculator

Notice the difference in the parts (*c*) and (*d*). Make sure that you read every question carefully!!

Mean and variance of the Poisson distribution.

If $X \sim \text{P}_\text{O}(\lambda)$ then it can be shown that

the expected mean is $\text{E}[X] = \mu = \lambda$

and the expected variance is $\text{Var}[X] = \sigma^2 = \lambda$.

This means that if the set of n trials were to be repeated a large number, n, times and the number of occurrences recorded each time, $x_1, x_2, x_3, x_4 \ \ldots\ldots\ x_n$

then the mean of $x_1, x_2, x_3, x_4 \ \ldots\ldots\ x_n$ would be $\mu = \lambda$

and the variance of $x_1, x_2, x_3, x_4 \ \ldots\ldots\ x_n$ would be $\sigma^2 = \lambda$

Note that the *mean is equal to the variance* in a Poisson distribution.

Example: In producing rolls of cloth there are on average 4 flaws in every 10 metres of cloth.

(*a*) Find the mean number of flaws in a 30 metre length.

(*b*) Find the probability of fewer than 3 flaws in a 6 metre length.

(*c*) Find the variance of the number of flaws in a 15 metre length.

Solution: Assuming a Poisson distribution – flaws in the cloth occur singly, independently, uniformly and randomly.

(*a*) If the mean number of flaws in 10 metres is 4,

then the mean number of flaws in 30 metre lengths is $3 \times 4 = 12$.

(b) If there are 4 flaws on average in a 10 metre length there will be $\frac{6}{10} \times 4 = 2{\cdot}4$ flaws on average in a 6 metre length.

If X is the number of flaws in a 6 metre length then $X \sim \text{P}_\text{O}(2.4)$.

$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$

$= e^{-2.4} + 2{\cdot}4 \times e^{-2.4} + \frac{2{\cdot}4^2 \times e^{-2{\cdot}4}}{2!}$

$= (1 + 2{\cdot}4 + 2{\cdot}4 \times 1{\cdot}2)\, e^{-2.4}$

$= 0{\cdot}569708746658 = 0{\cdot}570$ to 3 S.F. using calculator

(c) If the mean number of flaws in 10 metre lengths is 4, then the mean number of flaws in 15 metre lengths will be

$$\lambda = \frac{15}{10} \times 4 = 6$$

Since, in a Poison distribution, the variance equals the mean

the variance is 6.

The Poisson as an approximation to the binomial

Binomial B(n, p) for small p

If in the Binomial distribution B(n, p) p is *'small'* and n is *'large'*, then we can approximate by a Poisson distribution with mean $\lambda = np$, $\text{P}_\text{O}(np)$.

Notice that when p is *'small'* and n is *'large'*,

the expected variance of B(n, p) is $npq \approx np$ since $q = 1 - p \approx 1$

and so expected mean ≈ the variance.

In a Poisson distribution, the expected mean = the variance, so the approximation is suitable.

The approximation is also suitable if q is *'small'* and n is *'large'*.

In practice we use this approximation when p is small and $np \leq 10$,

and when $np > 10$ we use the Normal approximation (see later).

Example: If the probability of hitting the bull in a game of darts is $\frac{1}{20}$, find the probability of hitting at least 3 bulls in 50 throws using

(*a*) the Binomial distribution

(*b*) the Poisson approximation.

Solution: P(at least three bulls) = $1 - P(0, 1 \text{ or } 2) = 1 - P(\leq 2)$.

(*a*) $X \sim$ B(50, 0·05) the cumulative binomial tables give

$P(X \leq 2) = 0{\cdot}5405$

$\Rightarrow$ P(at least three bulls) = $1 - 0{\cdot}5405 = 0{\cdot}4595$ to 4 D.P. using tables

(*b*) $X \sim$ B(50, 0·05) the expected mean $\lambda = np$

$\Rightarrow \lambda = 50 \times 0{\cdot}5 = 2.5 \quad (< 10)$

We use the approximation $Y \sim P_O(2.5)$.

The cumulative Poisson tables for $\lambda = 1$ give

$P(Y \leq 2) = 0{\cdot}5438$

P(at least three bulls) = $1 - 0{\cdot}5438 = 0{\cdot}4562$ to 4 D.P. using tables

Not surprisingly the answers to parts (*a*) and (*b*) are different but not very different.

Selecting the appropriate distribution

Sometimes you will need to use a mixture of distributions to solve one problem.

Example: On average I make 7 typing errors on a page (and that is on a good day!).

(*a*) Find the probability that I make more than 10 mistakes on a page.

(*b*) In typing 5 pages find the probability that I make more than 10 mistakes on exactly 3 pages.

Solution:

(*a*) Assuming single, uniform and independent we can use the Poisson distribution $P_O(7)$ and from the cumulative Poisson tables, taking X as the number of typing errors

$X \sim P_O(7) \quad \Rightarrow \quad P(\leq 10) = 0{\cdot}9015$

$\Rightarrow \quad P(> 10) = 1 - 0{\cdot}9015 = 0{\cdot}0985$ to 4 D.P. (using tables)

(*b*) From (*a*) we know that the probability of one page with more than 10 errors is 0.0985 and taking Y as the number of pages with more than 10 errors

$Y \sim$ B(5, 0·0985)

$\Rightarrow$ P(3 pages with more than 10 errors) $= {}^5C_3 \times (0{\cdot}0985)^3 \times (0{\cdot}9015)^2$

$= 0{\cdot}0117$ to 3 S.F. using calculator

---

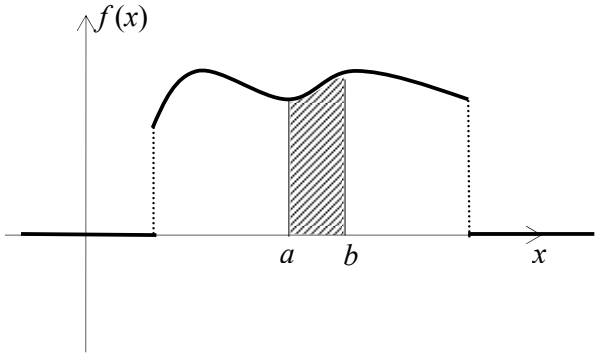
3 Continuous random variables

Probability density functions

For a continuous random variable we use a *probability density function* instead of a probability distribution for discrete values.

Conditions

A continuous random variable, X, has probability density function $f(x)$, as shown



where

1. total area is 1 $\Rightarrow \quad \int f(x)dx = 1$
2. the curve never goes below the x-axis
   $\Rightarrow \; f(x) \geq 0$ for all values of x
3. probability that X lies between a and b is the area from a to b

   $\Rightarrow \quad P(a < X < b) = \int_a^b f(x)\,dx.$

4. Outside the interval shown, $f(x) = 0$ and this **must** be shown on any sketch.
5. Notice that $P(X < b) = P(X \leq b)$ as no extra area is added.
6. $P(X = b)$ **always equals 0** (as there is no area)
   but this does **not** mean that X can never equal b.

Example: X is a random variable with probability density function

$$f(x) = kx(4 - x^2) \qquad \text{for } 0 \leq x \leq 2$$
$$f(x) = 0 \qquad \text{for all other values of } x.$$

(*a*) Find the value of k.

(*b*) Find the probability that $\frac{1}{2} < x \leq 1$.

(*c*) Sketch the probability density function.

Solution: (*a*) The total area between 0 and 2 must be 1

$\Rightarrow \quad \int_0^2 kx(4 - x^2) = 1$

$\Rightarrow \quad k\left[2x^2 - \frac{1}{4}x^4\right]_0^2 = 1$

$\Rightarrow \quad k \times [8 - 4] = 1$

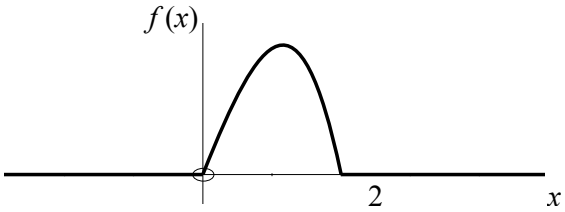
$\Rightarrow \quad k = \frac{1}{4}$

(*b*) The probability that $\frac{1}{2} < x \leq 1$ is the area between $\frac{1}{2}$ and 1

$= \int_{0.5}^{1} \frac{1}{4}x(4 - x^2)\ dx = \frac{1}{4}\left[2x^2 - \frac{x^4}{4}\right]_{0.5}^{1}$

$= \frac{1}{4}\left[\left(2 - \frac{1}{4}\right) - \left(\frac{1}{2} - \frac{1}{64}\right)\right] = \frac{81}{256} = 0{\cdot}316$ to 3 S.F. using calculator

(*c*) Note that $f(x)$ is zero outside the interval [0, 2] and this must be shown on your sketch to gain full marks in the exam.



Cumulative probability density function

This is like cumulative frequency;

the cumulative probability density function $F(X) = P(x < X)$ or $P(x \leq X)$

Note that there is no difference between the two expressions for a continuous distribution.

So for a probability density function $f(x)$

$F(X) = P(x < X) = \int_{-\infty}^{X} f(x)\, dx$

$\Rightarrow \quad f(x) = \frac{d(F(x))}{dx}$

Notice that for a cumulative probability density function $F(X)$, $0 \leq F(X) \leq 1$.

For the 'smallest' value of x, $F(x) = 0$, and

for the 'largest' value of x, $F(x) = 1$.

Example: The random variable X has probability density function

$f(x) = \frac{x}{8}$ for $0 \leq x \leq 4$,

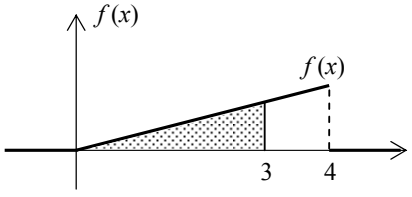
$f(x) = 0$ otherwise.

Find the cumulative probability that $X \leq 3$, i.e. find $F(3)$.

Solution: We want $F(3) = P(x \leq 3) = \int_0^3 \frac{x}{8}\ dx = \left[\frac{x^2}{16}\right]_0^3 = \frac{9}{16}$.

Notice that we could have drawn a sketch and found the area of the triangle

$P(x \leq 3) = \frac{1}{2} \times 3 \times \frac{3}{8} = \frac{3}{16}$



Example: A random variable X has a probability density function

$$f(x) = \begin{cases} \frac{1}{5}x & 0 \le x < 2 \\ \frac{2}{3} - \frac{2}{15}x & 2 \le x < 5 \\ 0 & \text{otherwise} \end{cases}$$

(*a*) Find $F(x)$.

(*b*) Sketch the graph of $F(x)$.

Solution:

(*a*) $F(x) = \int f(x)\ dx$

$0 \le x < 2$

$$F(x) = \int \tfrac{1}{5}x\ dx = \tfrac{1}{10}x^2 + c$$

$F(0) = 0 \quad \Rightarrow \quad c = 0$ — F(*smallest value*) = 0

$$\Rightarrow \quad F(x) = \tfrac{1}{10}x^2$$

$2 \le x < 5$

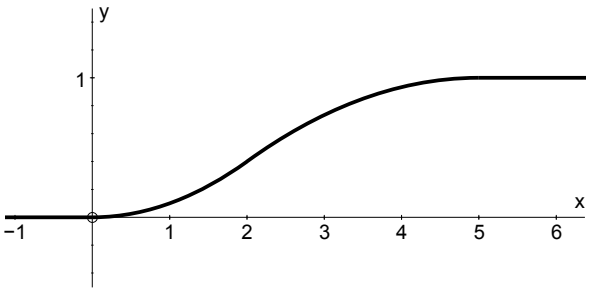
$$F(x) = \int \tfrac{2}{3} - \tfrac{2}{15}x\ dx = \tfrac{2}{3}x - \tfrac{1}{15}x^2 + c'$$

$F(5) = 1 \quad \Rightarrow \quad c' = 1 + \frac{25}{15} - \frac{10}{3} = -\frac{2}{3}$ — F(*largest value*) = 1

$$\Rightarrow \quad F(x) = \tfrac{2}{3}x - \tfrac{1}{15}x^2 - \tfrac{2}{3}$$

$$\Rightarrow \quad F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{10}x^2 & 0 \le x < 2 \\ \frac{2}{3}x - \frac{1}{15}x^2 - \frac{2}{3} & 2 \le x < 5 \\ 1 & x \ge 5 \end{cases}$$

(*b*)



Example: A dart is thrown at a dartboard of radius 25 cm. Let X be the distance from the centre to the point where the dart lands.

Assuming that the dart is equally likely to hit any point of the board find

(*a*) the cumulative probability density function for X.

(*b*) the probability density function for X.

Solution:

(*a*) $F(x) = P(X < x)$ = P(the dart lands a distance of less than x from the centre)

$$= \frac{\text{area of circle of radius } x}{\text{total area of the board}}$$

$$= \frac{\pi x^2}{\pi 25^2} = \frac{x^2}{625}.$$

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

$F(x) = \frac{x^2}{625}$ $\quad 0 \le x \le 25$

$F(x) = 1$ $\quad x > 25$

(*b*) $f(x) = \frac{d(F(x))}{dx} = \frac{d}{dx}\left(\frac{x^2}{625}\right) = \frac{2x}{625}$

$\Rightarrow$ $f(x) = \frac{2x}{625}$ $\quad 0 \le x \le 25$

$f(x) = 0$ $\quad$ otherwise

Expected mean and variance

Frequency, discrete and continuous probability distributions

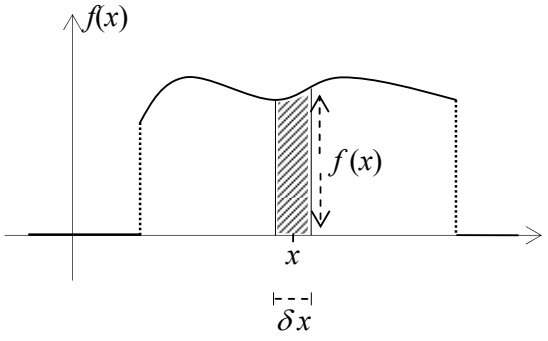
To change from a frequency distribution to a discrete probability distribution think of each probability p_i as $\frac{f_i}{N}$;

and to change from a discrete probability distribution think of the probability of x as the area of a narrow strip around x.

$\Rightarrow$ $p_i \approx f(x)\,\delta x$

then the formula for mean and variance etc are 'the same'.

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



Frequency distribution	Discrete probability distribution	Continuous probability distribution
$x_1, x_2, \ \ldots\ , x_n$	$x_1, x_2, \ \ldots\ , x_n$	$-\infty < x < \infty$
$f_1, f_2, \ \ldots\ , f_n$	$p_1, p_2, \ \ldots\ , p_n$	$f(x)$
$\sum f_i = N$	$\sum p_i = 1$	$\int_{-\infty}^{\infty} f(x)\ dx = 1$
$m = \frac{1}{N}\sum x_i f_i$	$\mu = \sum x_i p_i$	$\mu = \int_{-\infty}^{\infty} x\, f(x)\ dx$
$s^2 = \frac{1}{N}\sum x_i^2 f_i - m^2$ $= \frac{1}{N}\sum (x_i - m)^2$	$\sigma^2 = \sum x_i^2 p_i - \mu^2$ $= \sum (x_i - \mu)^2 p_i$	$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\ dx - \mu^2$ $= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\ dx$

Mode, median & quartiles for a continuous random variable

Mode

The *mode* is the 'most popular' and so will be at the greatest value of $f(x)$ in the interval.

It is best to sketch a graph, using calculus to find the stationary points.

Remember that the mode might be at one end of the interval, not in the middle.

Example: Find the mode for a random variable with probability density function

$f(x) = \frac{3}{40}(x^2 - 2x + 2)$ for $0 \le x \le 4$,

$f(x) = 0$ otherwise

Solution: $\frac{df}{dx} = \frac{3}{40}(2x - 2) = 0$ when $x = 1$.

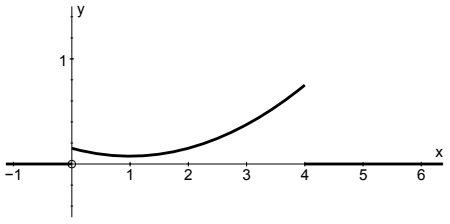
$\Rightarrow$ $\frac{d^2f}{dx^2} = \frac{6}{40}$, positive for all values of x $\Rightarrow$ **min**imum when $x = 1$

We now look at the whole graph in the interval

$f(0) = \frac{6}{40}$, $f(1) = \frac{3}{40}$, $f(4) = \frac{30}{40}$

$\Rightarrow$ graph has the largest value when $x = 4$

$\Rightarrow$ mode is $x = 4$.



Median

The median is the middle value and so the probability of being less than the median is ½ ;

so find M such that $P(X < M) = \frac{1}{2}$.

$\Rightarrow \quad \int_{-\infty}^{M} f(x)\ dx = \frac{1}{2}$.

Example: Find the median for a random variable with probability density function

$f(x) = \frac{20}{x^2}$ for $4 \le x \le 5$,

$f(x) = 0$ otherwise.

Solution: The median M is given by $\int_4^M \frac{20}{x^2}\ dx = \frac{1}{2}$

$\Rightarrow \quad \left[-\frac{20}{x}\right]_4^M = \frac{1}{2} \qquad \Rightarrow \quad -\frac{20}{M} + 5 = \frac{1}{2}$

$\Rightarrow \quad M = 4\frac{4}{9}$

Quartiles

Quartiles are found in the same way as the median.

$P(X < Q_1) = \frac{1}{4} \Rightarrow \int_{-\infty}^{Q_1} f(x)\ dx = \frac{1}{4}$.

$P(X < Q_3) = \frac{3}{4} \Rightarrow \int_{-\infty}^{Q_3} f(x)\ dx = \frac{3}{4}$.

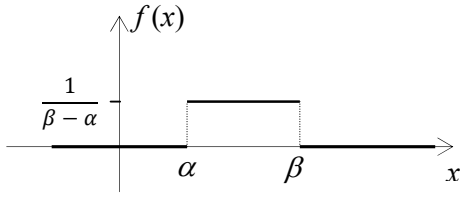
---

4 Continuous uniform (rectangular) distribution

Definition

A continuous uniform distribution has **constant** probability density over a fixed interval.

Thus $f(x) = \dfrac{1}{\beta - \alpha}$ is the continuous uniform p.d.f. over the interval $[\alpha, \beta]$ and has a rectangular shape.



Median

By symmetry the median is $\dfrac{\alpha + \beta}{2}$

Mean and Variance

The expected mean is $\mathrm{E}[X] = \mu = \dfrac{\alpha + \beta}{2}$, which is the same as the median.

and the expected variance is $\mathrm{Var}[X] = \sigma^2 = \dfrac{(\beta - \alpha)^2}{12}$.

Proofs

(*a*) Expected mean

$$\mathrm{E}[X] = \int x f(x)\,dx = \int_\alpha^\beta x \frac{1}{\beta - \alpha}\,dx = \left[\frac{x^2}{2(\beta - \alpha)}\right]_\alpha^\beta$$

$$= \frac{\beta^2 - a^2}{2(\beta - \alpha)} = \frac{(\beta - \alpha)(\beta + \alpha)}{2(\beta - \alpha)} = \frac{(\beta + \alpha)}{2}$$

or by symmetry.

(*b*) Expected variance

$$\mathrm{Var}[X] = \int x^2 f(x)\,dx - \mu^2 = \int_\alpha^\beta x^2 \frac{1}{\beta - \alpha}\,dx - \left(\frac{\beta + \alpha}{2}\right)^2$$

$$= \left[\frac{x^3}{3(\beta - \alpha)}\right]_\alpha^\beta - \left(\frac{\beta + \alpha}{2}\right)^2 = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} - \left(\frac{\beta + \alpha}{2}\right)^2$$

$$= \frac{(\beta - \alpha)(\beta^2 + \alpha\beta + \alpha^2)}{3(\beta - \alpha)} - \frac{(\beta^2 + 2\alpha\beta + \alpha^2)}{4}$$

$$= \frac{(\beta^2 - 2\alpha\beta + \alpha^2)}{12} = \frac{(\beta - \alpha)^2}{12}$$

5 Normal Approximations

The normal approximation to the binomial distribution

Conditions for approximation

For a binomial distribution $B(n, p)$ we know that the mean is $\mu = np$ and the variance is $\sigma^2 = npq$.

If p is 'near' $\frac{1}{2}$ and **if *n*** is **large**, $np > 10$, then the normal distribution $N(np, npq)$ can be used as an approximation to the binomial.

This is usually used when using the binomial would give awkward or tedious arithmetic.

Continuity correction

A continuity correction **must** always be used when approximating the binomial with the normal (this means that 47 must be taken as 46·5 or 47·5 depending on the sense of the question).

Example: Find the probability of more than 20 sixes in 90 rolls of a fair die.

Solution: The exact distribution is binomial $X \sim B(90, \frac{1}{6})$, where X is the number of sixes; but finding the exact probability would involve much tedious arithmetic.

Note that n is large, $np = 15 > 10$ so we can use the normal $N(np, npq)$ as an approximation.

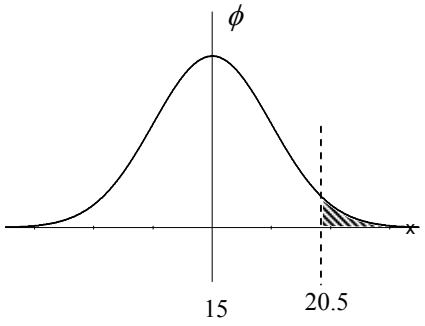
$\mu = np = 90 \times \frac{1}{6} = 15$

$\sigma^2 = npq = 90 \times \frac{1}{6} \times \frac{5}{6} = 12\frac{1}{2}$

$\Rightarrow \quad \mu = 15$ and $\sigma = \sqrt{12{\cdot}5} = 3{\cdot}53553$

So we use $Y \sim N(12, 12{\cdot}5)$, where Y is the number of sixes

To find P(more than 20 sixes) we must include 21 but not 20 so, using a continuity correction, we find the area to the right of 20·5,



$\Rightarrow \quad P(X > 20) = P(Y > 20{\cdot}5)$

$= 1 - \Phi\left(\frac{20{\cdot}5 - 15}{3{\cdot}53553}\right) = 1 - \Phi(1{\cdot}5556)$

$= 1 - \Phi(1{\cdot}56) = 1 - 0{\cdot}9406 = 0{\cdot}0594$ to 4 D.P. using tables

The normal approximation to the Poisson distribution

Conditions for approximation

For a Poisson distribution $P_O(\lambda)$ we know that the mean is $\mu = \lambda$ and the variance is $\sigma^2 = \lambda$ and

if *n* is **large and** $\lambda > 10$ then the normal distribution $N(\lambda, \lambda)$ can be used as an approximation to the Poisson distribution $P_O(\lambda)$.

This is usually used when using the Poisson would give awkward or tedious arithmetic.

Continuity correction

As with the normal approximation to the binomial a continuity correction **must** always be used when approximating the Poisson with the normal.

Example: Cars arrive at a motorway filling station at a rate of 18 every quarter of an hour. Find the probability that at least 23 cars arrive in a quarter of an hour period.

Solution:

The exact distribution is Poisson

X ~ $X \sim P_O(18)$, where X is the number of cars arriving in a $\frac{1}{4}$ hour period;

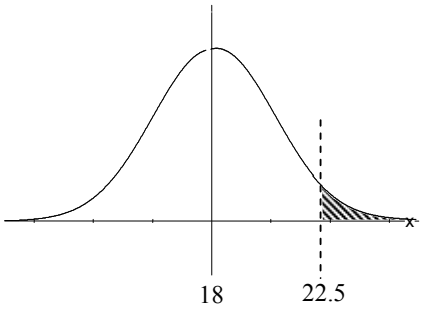
but finding the exact probability would involve much tedious arithmetic.

Note that n is large, $\lambda = 18 > 10$ so we can use the normal $Y \sim N(18, 18)$, where Y is the number of cars arriving in a ¼ hour period, as an approximation.

$\Rightarrow$ $\mu = 18$ and $\sigma = \sqrt{18} = 4{\cdot}2426$

To find P(at least 23 cars) we must include 23 but not 22 so, using a continuity correction,

we find the area to the right of 22·5,



$\Rightarrow$ $P(X \geq 23) = P(Y > 22{\cdot}5)$

$= 1 - \Phi\left(\frac{22{\cdot}5 - 18}{4{\cdot}2426}\right) = 1 - \Phi(1{\cdot}06)$

$= 1 - 0.8554 = 0.1446$ to 4 D.P. using tables

6 Populations and sampling

Words and their meanings

Population A collection of items.

Finite population A population is one in which **each** individual member can be given a number (a population might be so large that it is difficult or impossible to give each member a number – e.g. grains of sand on the beach).

Infinite population
A population is one in which **each** individual member can**not** be given a number.

Census
An investigation in which every member of the population is evaluated.

Sampling unit
A single member of the population which could be included in a sample **.**

Sampling frame
A list of **all** sampling units, by name or number, from which samples are to be drawn (usually the whole population but not necessarily).

Sample
A selection of sampling units from the sampling frame.

Simple random sample
A simple random sample of size n, is one taken so that every possible sample of size n has an equal chance of being selected.
The members of the sample are independent random variables, $X_1, X_2, \ldots, X_n$, and each X_i has the same distribution as the population.

Sample survey
An investigation using a sample.

Statistic
A quantity calculated only from the data in the sample, using no *unknown* parameters (for example, μ and σ).

Sampling distribution of a statistic
This is the set of all possible values of the statistic together with their individual probabilities; this is sometimes better described by giving the relevant probability density function.

Advantages and disadvantages of taking a census

Advantages

Every member of the population is used.
It is unbiased.
It gives an accurate answer.

Disadvantages

It takes a long time.
It is costly.
It is often difficult to ensure that the whole population is surveyed.

Advantages and disadvantages of sampling

Advantages

Sample will be representative if population large and well mixed.
Usually cheaper.
Essential if testing involves destruction (life of a light bulb, etc.).
Data usually more easily available.

Disadvantages:

Uncertainty, due to the natural variation – two samples are unlikely to give the same result.
Uncertainty due to **bias** prevents the sample from giving a representative picture of the population and can occur through:

sampling from an incomplete sampling frame – e.g. using a telephone directory for people living in Bangkok

influence of *subjective choice* where supposedly random selection is affected by personal preferences - e.g. interviewing only the pretty women!

non-response where questionnaires about a particular mobile phone service is not answered by many who do not use that service

substituting convenient sampling units when those required are not readily available – e.g. visiting neighbours when sampling unit is out!

NOTE: Bias cannot be removed by increasing the size of the sample.

Sampling distributions

To find the *sampling distribution* of the ******

*We need all possible values of ******, together with their probabilities*

Write down all possible samples together with their probabilities
Calculate the value of ****** for each sample
The sampling distribution of ****** is a list of all possible values of ****** together with their probabilities.

Example: A large bag contains £1 and £2 coins in the ratio 3 : 1.

A random sample of three coins is taken and their values X_1, X_2 and X_3 are recorded.

Find the sampling distribution for the mean.

Solution: We must first find **each** sample, its mean and probability

Sample	Mean	Probability
(1, 1, 1)	1	$(^3/_4)^3 = {}^{27}/_{64}$
(1, 1, 2), (1, 2, 1), (2, 1, 1)	$1\frac{1}{3}$	$3 \times (^3/_4)^2 \times (^1/_4) = {}^{27}/_{64}$
(1, 2, 2), (2, 1, 2), (2, 2, 1)	$1\frac{2}{3}$	$3 \times (^3/_4) \times (^1/_4)^2 = {}^{9}/_{64}$
(2, 2, 2)	2	$(^1/_4)^3 = {}^{1}/_{64}$

and so the sampling distribution of the mean is

Mean	1	$1\frac{1}{3}$	$1\frac{2}{3}$	2
Probability	$^{27}/_{64}$	$^{27}/_{64}$	$^{9}/_{64}$	$^{1}/_{64}$

OR, you may be able to use a standard probability distribution

Example: A disease is present in a 23% of a population. A random sample of 30 people is taken and the number with the disease, D, is recorded. What is the sampling distribution of D?

Solution: The possible outcomes (values of D) are

D	0	1	…	r	…	30
with probabilities	$0{\cdot}77^{30}$	$0{\cdot}77^{29} \times 0.23$	…	${}^{30}C_r \times 0{\cdot}77^{30-r} \times 0.23^r$	…	0.23^{30}

which we recognise as the Binomial distribution,

so the sampling distribution of D is $D \sim B(30, 0.23)$

7 Hypothesis tests

Null hypothesis, H_0
: The hypothesis which is assumed to be correct unless shown otherwise.

Alternative hypothesis, H_1
: This is the conclusion that should be made if H_0 is rejected

Hypothesis test
: A mathematical procedure to examine a value of a population parameter proposed by the null hypothesis, H_0, compared to the alternative hypothesis, H_1.

Test statistic
: This is the statistic (calculated from the sample) which is tested (in cumulative probability tables, or with the normal distribution etc.) as the last part of the test.

Critical region
: The range of values which would lead you to reject the null hypothesis, H_0

Significance level
: The **actual** significance level is the probability of rejecting H_0 when it is in fact true.

Null and alternative hypotheses, H_0 and H_1

Both null and alternative hypotheses **must** be stated in symbols only.

The null hypothesis, H_0, is the 'working hypothesis', i.e. what you assume to be true for the purpose of the test.

The alternative hypothesis, H_1, is what you conclude if you reject the null hypothesis: it also determines whether you use a one-tail or a two-tail test.

Your conclusion **must** be stated in full – both in statistical language **and in the context of the question**.

Critical region and significance level

From your observed result (test statistic) you decide whether to reject or not to reject the null hypothesis.

There will be an 'acceptable' region and if your observed result (test statistic) lies within this region then you will not reject the null hypothesis, H_0.

The region outside this 'acceptable' region is called the *critical region* and if your observed result (test statistic) lies in this critical region then you will reject the null hypothesis, H_0, and your conclusion will be based on the alternative hypothesis, H_1.

If the *significance level* is 5% then the 'acceptable region' is 95% of all possible outcomes and the *critical region* is the remaining 5% of all possible outcomes.

(**Note** that 'acceptable region' is not official jargon so do not use it!!)

One-tail and two-tail tests

The alternative hypothesis, H_1, will indicate whether you should use a one-tail or a two-tail test.

For example:

H_0: $a = b$

H_1: $a > b$

You reject H_0 **only** if b is significantly bigger than a.
Thus you are **only** looking at **one end** of the population and a *one-tail test* is suitable.

H_0: $a = b$

H_1: $a \neq b$

You reject H_0 **either** if b is significantly bigger than a **or** if b is significantly less than a.
Thus you are looking at **both ends** of the population and a *two-tail test* is suitable.

The points above are best illustrated by worked examples. Note that we always find the probability of the observed result **or worse**: this enables us to see easily whether the observed result lies in the critical region or not.

Worked examples (binomial, one-tail test)

Example: A tetrahedral die (one with four faces! – each equally likely) is rolled 40 times and 6 'ones' are observed. Is there any evidence at the 10% level that the probability of a score of 1 is less than a quarter?

Notice that the expected mean is 10 $\left(= 40 \times \frac{1}{4}\right)$, and we are really asking if the observed result (test statistic) 6 is 'surprisingly low'.

Solution: H_0: $p = 0.25$

H_1: $p < 0.25$.

From H_1 we see that a one-tail test is required, at 10% significance level.

If X is number of 'ones' then assuming binomial $X \sim B(40, 0{\cdot}25)$ and using the cumulative binomial tables

The test statistic (observed value) is $X = 6$

$P(X \leq 6$ 'ones' in 40 rolls$) = 0{\cdot}0962 = 9{\cdot}62\%$ from tables

Since $9{\cdot}62\% < 10\%$ the test statistic (observed result) lies in the critical region.

We reject H_0 and conclude that

there is evidence to show that the probability of a score of 1 is lower than $\frac{1}{4}$.

Example: The probability that a footballer scores from a penalty is 0·8. In twenty penalties he scores only 13 times. Is there any evidence at the 5% level that the footballer is losing his form?

Solution: If X is the number of scores from 20 penalties, then $X \sim B(20, 0{\cdot}8)$

The cumulative binomial tables do not deal with $p > 0.5$ so we must 'turn the problem round' and consider Y, the number of misses in 20 penalties, where $Y \sim B(20, 0{\cdot}2)$.

H_0: $p = 0{\cdot}8$ (or $p = 0{\cdot}2$).

H_1: $p < 0{\cdot}8$ (or $p > 0{\cdot}2$)

Observed value is $X = 13$

We consider the values $X \leq 13$

$\Rightarrow$ $X =$ 13 12 11 10 …

$\Rightarrow$ $Y =$ 7 8 9 10 …

$\Rightarrow$ $Y \geq 7$

Using the cumulative binomial tables, $Y \sim B(20, 0{\cdot}2)$

$P(X \leq 13) = P(Y \geq 7)$

$= 1 - P(Y \leq 6) = 1 - 0{\cdot}9133 = 0{\cdot}0867 = 8{\cdot}67\%$

$8{\cdot}67\% > 5\%$ (significance level)

$\Rightarrow$ the test statistic (observed result) 7 is not significant (does not lie in the critical region),

Do not reject H_0.

Conclude that there is evidence that the *player has not lost his form*, or that there is evidence that the *probability of scoring from a penalty is still 0·8*.

Worked example (binomial test, two-tail critical region)

Example: A tetrahedral die is manufactured with numbers 1, 2, 3 and 4 on its faces. The manufacturer claims that the die is fair.

All dice are tested by rolling 30 times and recording the number of times a 'four' is scored.

(*a*) Using a 5% significance level, find the critical region for a two-tailed test that the probability of a 'four' is $\frac{1}{4}$.
Find critical values which give a probability which is *closest* to 0·025.

(*b*) Find the **actual** significance level for this test.

(*c*) Explain how a die could pass the manufacturer's test when it is in fact biased.

Solution:

(*a*) H_0: $p = 0{\cdot}25$.

H_1: $p \neq 0{\cdot}25$ the die is not fair if there are *too many* or *too few* 'fours'

From H_1 we can see that a two-tailed test is needed, significance level 2·5% at each end.

Let X be number of 'four's in 30 rolls then we assume a binomial distribution, $X \sim B(30, 0{\cdot}25)$.

We shall reject the hypothesis if the observed result lies in either half of the critical region each half having a significance level of 2·5%;

For a two tail test, find the values of X which give a probability closest to 2·5% at each end.

Using cumulative binomial tables for $X \sim B\,(30, 0{\cdot}25)$:

for the lower critical value from tables

$P(X \leq 2) = 0{\cdot}0106$ (2.5% – 1.06% = 1.44%)
$P(X \leq 3) = 0{\cdot}0374$ (3.74% – 2.5% = 1.24%)

$X \leq 3$ gives the value closest to 2·5%, so $X = 3$ is lower critical value

and for the higher critical value from tables

$P(X \geq 13) = 1 - P(X \leq 12) = 1 - 0.9784 = 0.0216$ (2.5% – 2.16% = 0.34%)

$P(X \geq 12) = 1 - P(X \leq 11) = 1 - 0.9493 = 5.07\%$ (5.07% – 2.5% = 2.57%)

$X \geq 13$ gives the value closest to 2·5%, so $X = 13$ is higher critical value

Thus the critical region is $X \leq 3$ **or** $X \geq 13$.

(*b*) The **actual** significance level is $0{\cdot}0374 + 0{\cdot}0216 = 0{\cdot}0590 = 5{\cdot}90\%$. to 3 S.F.

(*c*) The die could still be biased in favour of, or against, one of the other numbers.

Worked example (Poisson)

Example: Cars usually arrive at a motorway filling station at a rate of 3 per minute. On a Tuesday morning cars are observed to arrive at the filling station at a rate of 5 per minute. Is there any evidence at the 10% level that this is an unusually busy morning?

Solution: H_0: $\lambda = 3$

H_1: $\lambda > 3$ unusually busy would mean that λ would increase

From H_1 we see that a one-tail test is needed, significance level 10%.

It seems sensible to assume that the numbers of cars arriving per minute is independent, uniform and single so a Poisson distribution is suitable.

Let X be the number of cars arriving per minute, then $X \sim \text{Po}(3)$

The test statistic (observed value) is $X = 5$

$P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0{\cdot}8153 = 0{\cdot}1847 = 18{\cdot}47\% > 10\%$ using tables

which is not significant at the 10% level. Do not reject H_0.

Conclude that there is evidence that *Tuesday morning* is *not unusually busy*,

or there is evidence that *cars are still arriving* at *a rate of 3 cars per minute*.

Worked example (Poisson, critical region)

Example: Over a long period in the production of glass rods the mean number of flaws per 5 metres is 4. A length of 10 metres is to be examined. Find the critical region to show that the machine is producing too many flaws at the 5% level.

Find the lowest value which gives a probability of less than 5%.

Solution:

H_0: $\lambda = 8$ flaws occur uniformly, so if mean per 5 metres is 4, then mean per 10 metres is 8

H_1: $\lambda > 8$ machine producing too many flaws would mean that λ would increase

We can see from H_1 that we need a one- tail test, significance level 5%.

For a one tail test, find the first value of X for which the probability is less than 5%
It seems sensible to assume that the number of flaws per 10 metre lengths is independent, uniform and single so a Poisson distribution is suitable.

$X \sim \text{Po}(8)$ where X is the number of flaws in a 10 metre length

$P(X \geq 13) = 1 - P(X \leq 12) = 1 - 0{\cdot}9362 = 0{\cdot}0638 = 6{\cdot}38\% > 5\%$ from tables

$P(X \geq 14) = 1 - P(X \leq 13) = 1 - 0{\cdot}9784 = 0{\cdot}0216 = 2{\cdot}16\% < 5\%$

$\Rightarrow$ $X = 14$ is the smallest value for which the probability is less than 5%

$\Rightarrow$ the critical region is $X \geq 14$.

Hypothesis testing using approximations.

Example: With current drug treatment, 9% of cases of a certain disease result in total recovery. A new treatment is tried out on a random sample of 100 patients, and it is found that 16 cases result in total recovery. Does this indicate that the new treatment is better at a 5% level of significance?

Solution: Let X be the number of cases resulting in total recovery.

H_0: $p = 0.09$

H_1: $p > 0.09$

$X \sim B(100, 0.09)$, which is not in the tables and is awkward arithmetic, so we use an approximation.

λ or $\mu = np = 100 \times 0.09 = 9 < 10$,

***n* large and *p* small** $\Rightarrow$ we should use the Poisson approximation $Y \sim P_O(9)$

The test statistic is $Y = 16$.

We want $P(X \geq 16) \approx P(Y \geq 16)$

$= 1 - P(Y \leq 15) = 1 - 0{\cdot}9780 = 0{\cdot}0220 < 5\%$ from tables

$\Rightarrow$ the test statistic is significant at 5%

$\Rightarrow$ Reject H_0.

Conclude that there is some evidence that the *proportion of cases of total recovery* has *increased* from 0.09 under the *new treatment*.

Example: With current drug treatment, 20% of cases of a certain disease result in total recovery. A new treatment is tried out on a random sample of 100 patients, and it is found that 26 cases result in total recovery. Does this indicate that the new treatment is better at a 5% level of significance?

Solution: Let X be the number of cases resulting in total recovery.

H_0: $p = 0.2$

H_1: $p > 0.2$

$X \sim B(100, 0.2)$, which is not in the tables and is awkward arithmetic,.

$\mu = np = 100 \times 0.2 = 20 > 10,$ ***n* large and *p* is 'near $\frac{1}{2}$'**

so we use a normal approximation.

$\sigma^2 = np(1-p) = 100 \times 0.2 \times 0.8 = 16$

Use the approximation $Y \sim N(20, 16)$

The test statistic is $X = 26$.

We want $P(X \geq 26) \approx P(Y \geq 25{\cdot}5)$ **you must use a continuity correction**

$= P\left(Z \geq \frac{25{\cdot}5-20}{4}\right) = 1 - P(Z \geq 1 \cdot 375)$ (use $Z = 1{\cdot}38$)

$= 1 - 0{\cdot}9162 = 0{\cdot}0838 > 5\%$

$\Rightarrow$ the test statistic is **not** significant at 5%

$\Rightarrow$ Do not reject H_0.

Conclude that there is evidence that the *proportion of cases of total recovery* has *not increased* from 0·2 under the *new treatment*,

or conclude that there is evidence that the *new treatment* is *not better*.

---

8 Context questions and answers

Accuracy

You are required to *give your answers to an appropriate degree of accuracy*.

There is no hard and fast rule for this, but the following guidelines should never let you down.

1. If stated in the question give the required degree of accuracy.
2. When using a calculator, give 3 S.F.
   unless finding $S_{xx,}$ S_{xy} etc. in which case you can give more figures – you should use *all* figures when finding the PMCC or the regression line coefficients.
3. Sometimes it is appropriate to give a mean to 1 or 2 D.P. rather than 3 S.F.
4. When using the tables and doing simple calculations (which do not *need* a calculator), you should give 4 D.P.

General vocabulary

Question 1

Define a statistic.

Answer

A *number calculated* from known observations (from a *sample*) – *no unknown parameters*.

Question 2

Explain what you understand by the statistic *Y*.

Answer

A statistic is a *calculation* from the values in the *sample*, $X_1, X_2, ...X_n$ that does not contain any *unknown parameters.*

Question 3

Define the critical region of a test statistic.

Answer

The set of *values* of the *test statistic* for which the *null hypothesis is rejected* in a *hypothesis test.*

Question 4

Explain what you understand by

(*a*) a population,

(*b*) a statistic.

A researcher took a sample of 100 voters from a certain town and asked them who they would vote for in an election. The proportion who said they would vote for Dr Smith was 35%.

(*c*) State the population and the statistic in this case.

(*d*) Explain what you understand by the sampling distribution of this statistic.

Answer

(*a*) A population is *collection* of *all items*

(*b*) A *calculation* from the *sample* which contains no *unknown quantities/parameters*.

(*c*) The population is '*voters in the town*'. The statistic is '*percentage/proportion voting for Dr Smith*'.

(*d*) *List* of *all possible samples* (of size 100) of those *voting for Dr Smith* together with the *probability of each sample*.

Skew

Question 1

Explain why the distribution is negatively skewed.

Answer

Mean < median < mode ($\Rightarrow$ negative skew): *give values*.

Question 2

Given that the median is 1.40, describe the skewness of the distribution. Give a reason for your answer.

Answer

Positive skew

$Q_3 - Q_2 > Q_2 - Q_1$ *give values* (box plot with Q_1, Q_2, Q_3 values marked on – optional)

Or mode = 1 and mode < median = *give value*.

Binomial distribution

Question 1

A company claims that a quarter of the bolts sent to them are faulty. To test this claim the number of faulty bolts in a random sample of 50 is recorded. Give two reasons why a binomial distribution may be a suitable model for the number of faulty bolts in the sample

Answer

Two from 2 outcomes / faulty or not faulty / success or fail.
A constant probability of a faulty bolt
Trials are independent.
Fixed number of trials (fixed n).
Poisson distribution

Question 2

State two conditions under which a Poisson distribution is a suitable model for X.

Answer

*Misprint*s are *random / independent*, occur *singly* in space and at a *constant rate*

Question 3

An estate agent sells properties at a mean rate of 7 per week.

Suggest a suitable model to represent the number of properties sold in a randomly chosen week. Give two reasons to support your model.

Answer

$P_O(7)$ *Sales* occur *independently/randomly*, *singly*, at a *constant rate* (*context needed once*)

Question 4

A call centre agent handles telephone calls at a rate of 18 per hour.

(*a*) Give two reasons to support the use of a Poisson distribution as a suitable model for the number of calls per hour handled by the agent.

Answer

Calls occur *singly* any two of the 3

Calls occur at a *constant rate*

Calls occur *independently or randomly*.

Question 5

Explain how the answers from part (*c*) support the choice of a Poisson distribution as a model.

Answer

For a *Poisson* model , *Mean* = *Variance* ; For *these data* $3.69 \approx 3.73 \Rightarrow$ Poisson

---

Approximations to Poisson and Binomial

Question 1

Giving a justification for your choice, use a suitable approximation to estimate the probability that there are exactly 5 defective articles.

Answer

For the binomial distribution $X \sim$ B(200 , 0·02)
n is *large*, *p* is *small* so use the Poisson approximation $Y \sim P_O(np) =$ Po (4).

Question 2

(*a*) State the condition under which the normal distribution may be used as an approximation to the Poisson distribution.

(*b*) Explain why a continuity correction must be incorporated when using the normal distribution as an approximation to the Poisson distribution.

Answer

(*a*) $\lambda > 10$ or *large* (use μ instead of λ is OK).

(*b*) The *Poisson* distribution is *discrete* and the *normal* distribution is *continuous.*

Question 3

Write down the conditions under which the Poisson distribution may be used as an approximation to the Binomial distribution.

Answer

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

n is *large*, $n > 50$

p is *small*, $p < 0{\cdot}2$ (or $q = 1 - p$ is small)
but I would not worry too much about the 0·2 and the 50.

then X can be approximated by Po(np)

Question 4

Write down two conditions for $X \sim \text{B}(n, p)$ to be approximated by a normal distribution $Y \sim \text{N}(\mu, \sigma^2)$.

Answer

If $X \sim \text{B}(n,p)$ and

n is *large* or $n > 10$ or $np > 5$ o**r** $nq > 5$

p is *close to* 0.5 or $nq > 5$ and $np > 5$

then X can be approximated by $Y \sim \text{N}(np, npq)$.

Sampling

Question 1

Explain what you understand by

(*a*) a sampling unit.

(*b*) a sampling frame.

(*c*) a sampling distribution.

Answer

(*a*) *Individual member* or *element* of the *population* or *sampling frame*.

(*b*) A *list* of *all* sampling units *or all* the population.

(*c*) *All* possible *samples* are chosen from a population; the *values* of a *statistic* together with the associated *probabilities* is a sampling distribution.

Question 2

Before introducing a new rule, the secretary of a golf club decided to find out how members might react to this rule.

(*a*) Explain why the secretary decided to take a random sample of club members rather than ask all the members.

(*b*) Suggest a suitable sampling frame.

(*c*) Identify the sampling units.

Answer

(*a*) Saves time / cheaper / easier or

a census/asking all members takes a long time or is expensive or difficult to carry out

(*b*) *List, register or database* of *all* club *members/golfers*

(*c*) Club *member*(*s*)

Question 3

Explain what you understand by the sampling distribution of Y.

Answer

The probability distribution of Y, or the distribution of all *possible values* of Y together with their *probabilities*.

Index