Statistics 2

Revision Notes

June 2016

Statistics 2

1 The Binomial distribution

Factorials

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

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

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

$$^nC_r = \frac{n(n-1)(n-2)\ldots \text{ up to } r \text{ numbers}}{r!} = \frac{n(n-1)(n-2)\ldots\ (n-r+1)}{r!}$$

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 numbers 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.

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1

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

We write down the terms in a logical order	p^6	p^5q	p^4q^2	p^3q^3	p^2q^4	pq^5	q^6
then use the numbers in the '6th' row of the triangle	**1**	**6**	**15**	**20**	**15**	**6**	**1**

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

Binomial coefficients and combinations

${}^nC_r$ is often written $\binom{n}{r}$, and gives the numbers in the '*n*th row' of Pascal's Triangle..

	1	4	6	4	1
and we have	${}^4C_0 = \binom{4}{0} = 1$	${}^4C_1 = \binom{4}{1} = 4$	${}^4C_2 = \binom{4}{2} = 6$	${}^4C_3 = \binom{4}{3} = 4$	${}^4C_4 = \binom{4}{4} = 1$

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

1) A single trial has exactly two possible outcomes – *success* and *failure*.
2) This trial is repeated a *fixed number*, *n*, times.
3) The *n* trials are *independent* of each other.
4) 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 $X \sim B\left(10, \frac{1}{6}\right)$

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

or 6, ×,×,×,6, 6,×, ×,6,× , or 6,6,×,×,6,×,×,×,6,×, or ……….. all with probability $\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6$

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 $X \sim B(n, p)$

the probability of r successes is

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

Binomial distribution

The binomial *distribution/table* $X \sim B(n, p)$ is shown below.

x	0	1	2	...	r	...	n
$P(X = x)$	${}^{n}C_0\, q^n$	${}^{n}C_1\, pq^{n-1}$	${}^{n}C_2\, p^2q^{n-2}$	...	${}^{n}C_r\, p^rq^{n-r}$	...	${}^{n}C_n\, p^n$

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, or it means a *probability table.*

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 $X \sim B(10, 0{\cdot}73)$, assuming independence of trials.

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

$= P(X = 8) + P(X = 9) + P(X = 10)$

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

$= 0{\cdot}34709235895 = 0{\cdot}347$ 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. 4 D.P. as we are using 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 $X = 12, 13, 14, 15, 16$ or 17 it is worth writing out the numbers

for which values $Y = 8, 7, 6, 5, 4$ or 3 since $X + Y = 20$

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

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

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

Mean and variance of the binomial distribution.

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

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

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

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

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

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

These formulae are proved in the appendix.

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

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

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

and $\sigma^2 = np(1-p) = 100 \times \frac{1}{2} \times \frac{1}{2} = 25$

$\Rightarrow \sigma = \sqrt{25} = 5$

$\mu - 2\sigma = 50 - 10 = 40$,
$\mu + 2\sigma = 50 + 10 = 60$

So we would expect that the probability that the number of heads lies between 40 and 60 inclusive (within 2 standard deviations of the mean) is about 0·95.

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 $X \sim B(n, 0{\cdot}35)$.

The expected mean, $E[x] = \mu = np = 0.35n$

The expected number who like fish and chips exceeds 60

$\Rightarrow \quad 0.35n > 60$

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

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

2 The Poisson distribution

Conditions for a Poisson distribution

Events must occur

1) *singly* – two events cannot occur simultaneously
2) *uniformly* – events occur at a constant rate
3) *independently & randomly* – the occurrence of one event does not influence 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 of some particular length, the probability of *r* occurrences in an interval of the same length is

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

For a proof of these probabilities see the appendix

The Poisson *distribution* (*table*) of $X \sim \mathrm{P_o}(\lambda)$ 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, or it means a *probability table*.

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 $X \sim \mathrm{P_o}(5)$

(*a*) $X \sim \mathrm{P_o}(5)$

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

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

(*b*) $P(\text{at least 3 but fewer than 8}) = P(3 \le X < 8)$

$= P(X \le 7) - P(X \le 2) = 0{\cdot}8666 - 0{\cdot}1247 = 0{\cdot}7419$ 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 **2** minute period.

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

$Y \sim P_o(10)$, 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 $X \sim P_o(5)$

$\Rightarrow$ $P(X > 3) = 1 - P(X \leq 3) = 1 - 0{\cdot}2650 = 0{\cdot}7350$ — from tables

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

is $0{\cdot}7350^2 = 0{\cdot}540225 = 0{\cdot}540$ 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 $X \sim P_o(\lambda)$ then it can be shown that

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

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

This means that if the set of n trials were to be repeated a very large number of times and the number of occurrences recorded each time, x_1, x_2, x_3, x_4 ... …

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

and the variance of x_1, x_2, x_3, x_4 ... … 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 $X \sim \mathrm{P_o}(2.4)$.

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

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

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

$= 0{\cdot}569708746658 = 0{\cdot}570$ 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 in 15 metre lengths is 6.

The Poisson as an approximation to the binomial

Binomial $\mathrm{B}(n, p)$ for small p or q

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

And if p is close to 1 then q will be small, and $Y \sim \mathrm{B}(n, q)$, where $Y = n - X$, which can be approximated by a Poisson distribution $Y' \sim \mathrm{P_o}(nq)$

Notice that when p is *'small'* and n is *'large'*,
the expected variance of $\mathrm{B}(n, p)$ 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.

In practice we use this approximation when p (or q) 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 to the Binomial.

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

(*a*) For $X \sim B(50, 0{\cdot}05)$ 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*) $X \sim B(50, 0{\cdot}05)$ the expected mean $\lambda = np$

$\Rightarrow \lambda = 50 \times 0{\cdot}05 = 2.5 \quad (< 10)$

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/random we can use the Poisson distribution $P_o(7)$ and from the cumulative Poisson tables, taking X as the number of typing errors

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

$\Rightarrow$ $P(X > 10) = 1 - P(X \leq 10) = 1 - 0{\cdot}9015 = 0{\cdot}0985$ 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 we take Y as the number of pages with more than 10 errors. Thus for 5 pages

$Y \sim B(5, 0{\cdot}0985)$

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

$= 0{\cdot}00777$ 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.

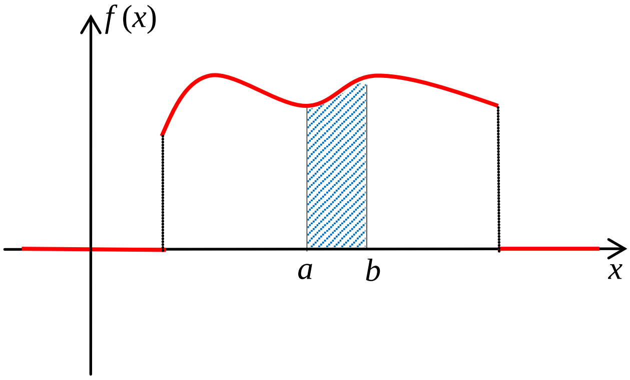
Conditions

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

where

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

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



4. Outside the interval shown, $f(x) = 0$ and this **must** be shown on any sketch – see red lines.
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

$f(x) = kx(4 - x^2)$ for $0 \leq x \leq 2$

$f(x) = 0$ 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

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

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

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

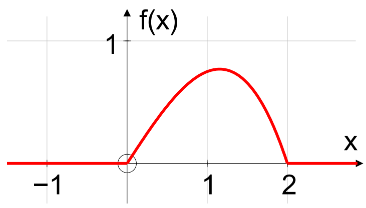
$\Rightarrow k = \frac{1}{4}$

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

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

$$= \frac{1}{4}\left[\left(2 - \tfrac{1}{4}\right) - \left(\tfrac{1}{2} - \tfrac{1}{64}\right)\right] = \frac{81}{256} = 0{\cdot}316$$ 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 $P(x \le X)$

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 $F(X)$, $0 \le F(X) \le 1$.

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

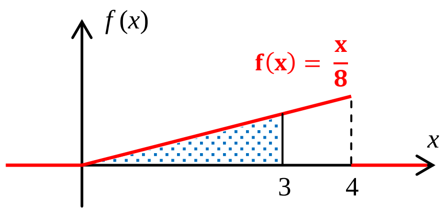
$f(x) = \frac{x}{8}$ for $0 \le x \le 4$,

$f(x) = 0$ otherwise.

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

Solution: We want $F(3) = P(x \le 3)$

$$= \int_0^3 \frac{x}{8}\,dx = \left[\frac{x^2}{16}\right]_0^3 = \frac{9}{16}$$

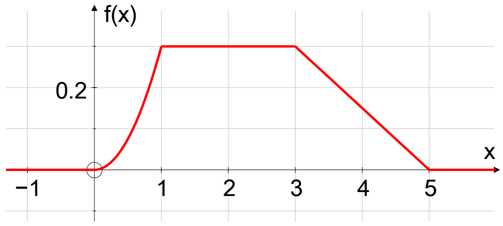


Notice that we could have drawn a sketch and found the area of the triangle

$$P(x \le 3) = \tfrac{1}{2} \times 3 \times \tfrac{3}{8} = \frac{9}{16}$$

Example: A random variable X has a probability density function

$$f(x) = \begin{cases} \frac{3}{10}x^2 & 0 \le x < 1 \\ \frac{3}{10} & 1 \le x < 3 \\ \frac{3}{4} - \frac{3}{20}x & 3 \le x < 5 \\ 0 & \text{otherwise} \end{cases}$$



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

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

Solution:

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

$0 \le x < 1$

$F(x) = \int \frac{3}{10}x^2\ dx = \frac{1}{10}x^3 + c$

$F(smallest\ value) = 0$

$\Rightarrow \quad F(0) = 0 \quad \Rightarrow \quad c = 0$

$\Rightarrow \quad F(x) = \frac{1}{10}x^3$

$1 \le x < 3$

$F(x) = \int \frac{3}{10}\ dx = \frac{3}{10}x + c'$

Finding c' is more complicated here.

The interval $[0, 1)$ ends with $F(1) = \frac{1}{10} \times 1^3 = \frac{1}{10}$,

and so the interval $[1, 3)$ must start with $F(1) = \frac{1}{10}$

$\Rightarrow \quad F(1) = \frac{3}{10} \times 1 + c' = \frac{1}{10} \quad \Rightarrow \quad c' = \frac{-1}{5}$

$\Rightarrow \quad F(x) = \frac{3}{10}x - \frac{1}{5}$

$3 \le x < 5$

$F(x) = \int \frac{3}{4} - \frac{3}{20}x\ dx = \frac{3}{4}x - \frac{3}{40}x^2 + c''$

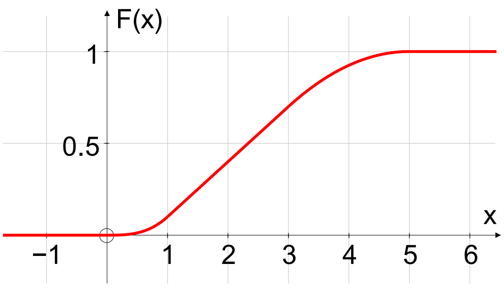
$F(largest\ value) = 1$

$\Rightarrow \quad F(5) = 1 \quad \Rightarrow \quad \frac{3}{4} \times 5 - \frac{3}{40} \times 5^2 + c'' = 1 \quad \Rightarrow \quad c'' = \frac{-7}{8}$

$\Rightarrow \quad F(x) = \frac{3}{4}x - \frac{3}{40}x^2 - \frac{7}{8}$

$$\Rightarrow \quad F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{10}x^3 & 0 \le x < 1 \\ \frac{3}{10}x - \frac{1}{5} & 1 \le x < 3 \\ \frac{3}{4}x - \frac{3}{40}x^2 - \frac{7}{8} & 3 \le x < 5 \\ 1 & x \ge 5 \end{cases}$$

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

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

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

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

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

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

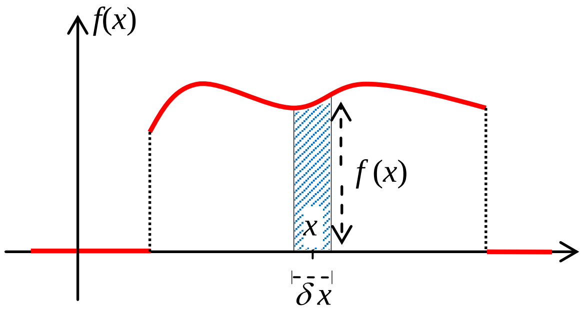
(*b*) $$f(x) = \frac{d(F(x))}{dx} = \frac{d}{dx}\left(\frac{x^2}{625}\right) = \frac{2x}{625}$$

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

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

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

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

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

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

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

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 if necessary.

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

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

$$f(x) = 0 \qquad \text{otherwise}$$

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

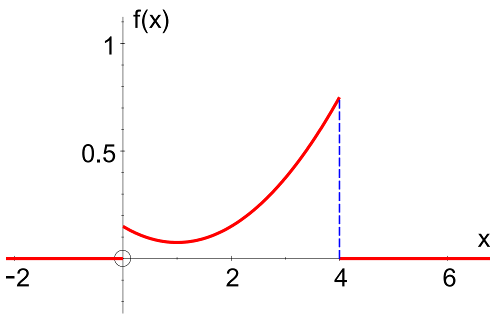
$\Rightarrow$ $\frac{d^2 f}{dx^2} = \frac{6}{40}$, positive for all values of x $\Rightarrow$ **min**imum when $x = 1$

We now look at the whole graph in the interval

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

$\Rightarrow$ 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

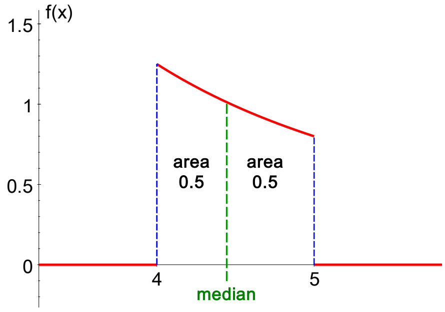
$f(x) = \dfrac{20}{x^2}$ for $4 \le x \le 5$,

$f(x) = 0$ otherwise.

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

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

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



Quartiles

Quartiles are found in the same way as the median.

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

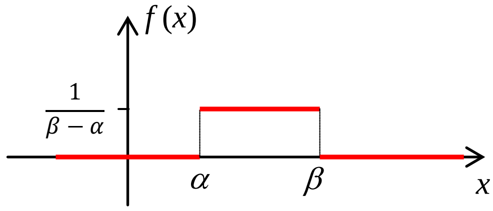
$P(X < Q_3) = \frac{3}{4} \quad \Rightarrow \quad \int_{-\infty}^{Q_3} f(x)\ dx = \frac{3}{4}$.

---

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

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

These formulae are proved in the appendix

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.

For a proof of the Normal Approximation to the Binomial distribution see the appendix.

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.

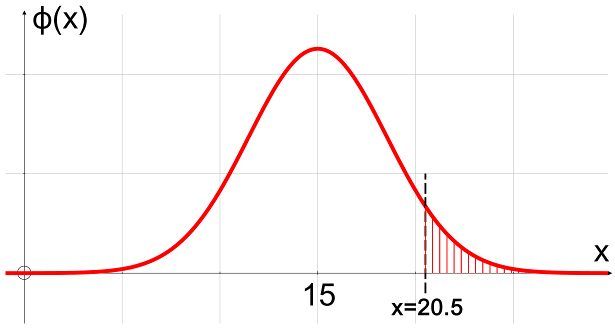
$\mu = np = 90 \times \frac{1}{6} = 15$

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

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

So we use $Y \sim N\left(15, \sqrt{12 \cdot 5}^2\right)$, 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 $N(\lambda, \sqrt{\lambda}^2)$ 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

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

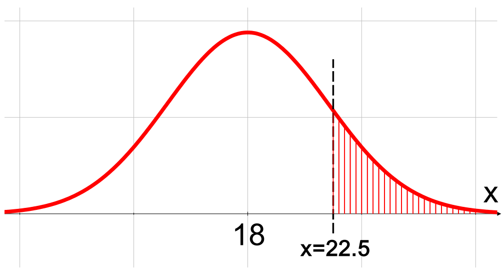
but finding the exact probability would involve much tedious arithmetic.

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.

$\Rightarrow \quad \mu = 18$ 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,



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

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

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

6 Populations and sampling

Words and their meanings

Population A collection of *all* 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, $X_1, X_2, \dots , X_n$, 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 (or any large city)

influence of *subjective choice* where supposedly random selection is affected by personal preferences - e.g. interviewing only people without (fierce) dogs

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.

NOTE: when answering questions on these definitions you may be asked to put your answer in context.

Examples:

The sampling frame is a list of *all amplifiers* and their *serial numbers*.

A sampling unit is one *amplifier*.

The test statistic is the *number voting for Mr. Smith*.

A sample is a random selection of *pupils from the school*.

A census is an investigation in which the *lengths of all rods manufactured are recorded*.

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 (or probability table) 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	$0{\cdot}77^{30}$	$^{30}C_1 0{\cdot}77^{29} \times 0.23$	…	$^{30}C_r \times 0{\cdot}77^{30-r} \times 0.23^r$	…	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 significance 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.

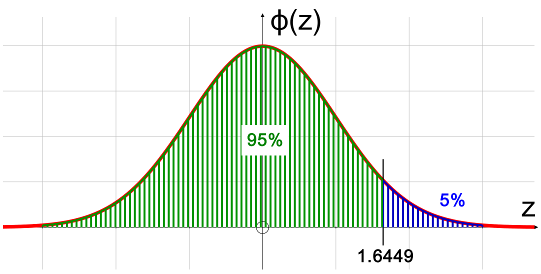
The alternative hypothesis, H_1, is what you conclude if you reject the null hypothesis: it also determines whether you use a one-tail or a two-tail test.

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

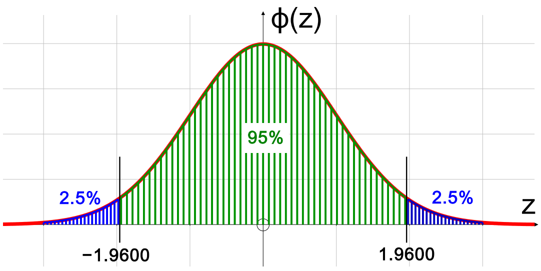
Hypotheses and significance level

From your observed result (test statistic) you decide whether to reject or not to reject the null hypothesis, H_0.

1 *tail test*



2 *tail test*



From the null hypothesis, H_0, we could have a result *anywhere* on the graph – including the small (blue) shaded areas.

If the observed result (test statistic) lies in a small (blue) shaded area, we say that

The test statistic is *significant* at 5%, or that we reject H_0. Thus H_0 could actually be true but we still reject it.

Thus, the *significance level*, 5%, is

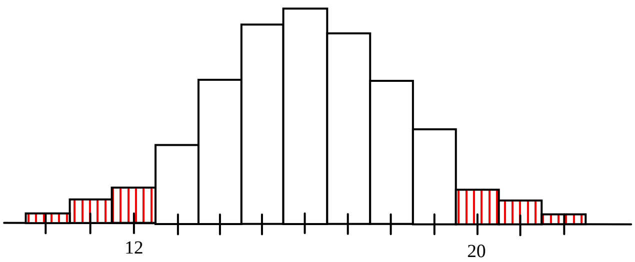
the probability that we reject H_0 when it is in fact true,

or the probability of incorrectly rejecting H_0.

When we reject the null hypothesis, H_0, we use the alternative hypothesis to write the conclusion.

Critical regions and significance levels

Poisson and Binomial



The Poisson and Binomial distributions are discrete, and we look at probability histograms.

In the diagram, the critical region (shown by the shaded areas) is $X \leq 12$ or $X \geq 20$.

We include the whole bar around $X = 12$, and around $X = 20$.

So $P(X \leq 12)$ is the area to the left of 12·5,

and $P(X \geq 20)$) is the area to the right of 19.5.

If $P(X \leq 12) = 0{\cdot}0234$ and $P(X \geq 20) = 0{\cdot}0217$, then

the actual significance level is $0{\cdot}0234 + 0{\cdot}0217 = 0{\cdot}0451 = 4{\cdot}51\%$

Thus the probability of incorrectly rejecting H_0 is 0·0451,

Or, the probability of rejecting H_0 when it is actually true is 0·0451.

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 a is significantly bigger than b.
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** a if is significantly bigger than b **or** if a is significantly less than b.
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 l 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 $X \sim B(40, 0{\cdot}25)$, from H_0, 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: H_0: $p = 0{\cdot}8$

H_1: $p < 0{\cdot}8$

If X is the number of scores from 20 penalties, then $X \sim B(20, 0{\cdot}8)$ – from H_0

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 $Y \sim B(20, 0{\cdot}2)$.

Observed value is $X = 13$

We consider the values $X \le 13$

$\Rightarrow \quad X = \quad 13 \quad 12 \quad 11 \quad 10 \quad \ldots$

$\Rightarrow \quad Y = \quad 7 \quad 8 \quad 9 \quad 10 \quad \ldots$

$\Rightarrow \quad Y \ge 7$

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

$P(X \le 13) = P(Y \ge 7)$

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

$8{\cdot}67\% > 5\%$ (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 not less than 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 the number of ‘4’s in 30 rolls. From H_0 we have 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 from H_0, $X \sim P_o(3)$

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 not arriving* at a *rate greater* than *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 length 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, and using H_0

$X \sim P_o(8)$ — 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

$P(X \geq 14) = 1 - P(X \leq 13) = 1 - 0{\cdot}9658 = 0{\cdot}0342 = 3{\cdot}42\% < 5\%$

$\Rightarrow$ $X = 14$ 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: $p = 0.09$ — new treatment has same recovery rate as the current treatment

H_1: $p > 0.09$ — new treatment is better than the current treatment

$X \sim B(100, 0.09)$, 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 $Y \sim \text{P}_\text{o}(9)$

The test statistic is $X = 16$.

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

$\Rightarrow$ this result 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' 0·5**

so we use a normal approximation.

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

Use the approximation $Y \sim N(20, 4^2)$

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**

$= P\left(Z \geq \frac{25\cdot5-20}{4}\right) = 1 - P(Z \geq 1\cdot375)$ (use $Z = 1\cdot38$)

$= 1 - 0\cdot9162 = 0\cdot0838 > 5\%$

$\Rightarrow$ the result 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. 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.
2. Sometimes it is appropriate to give a mean to 1 or 2 D.P. rather than 3 S.F.
3. When using the tables and doing simple calculations (which do not *need* a calculator), you should give 4 D.P.

General vocabulary

You must include the **context** in your answers, where appropriate. Definitions alone are **not** enough.

Question 1

Explain what you understand by the statistic Y.

Answer

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

Question 2

A random sample is taken of the heights of 20 people living in a small town. The mean height, 163 cm, is calculated. Why can the number 163 be considered as a statistic?

Answer

163 is the **mean height** of the **20 people** in the sample, and can be calculated **only** using the heights in the sample.

Question 3

The number of hurricanes in a two month period follows a Poisson distribution.

Based on the null hypothesis that the mean number of hurricanes in two months is 7, the critical values, at 5% significance level, are 2 and 12. The number of hurricanes, n, in a two month period is then recorded

What is the test statistic, and what is meant by the critical region.

Answer

The test statistic is n, the *number of hurricanes* in a *two month period*, and the critical region, $X \leq 2$ or $X \geq 12$, is the *range of values* of n, *the number of hurricanes* in a two month period, which would lead to the *rejection* of H_0.

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 a *collection* of *all items*

(*b*) A *calculation only* 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*. In this case the sampling distribution is B(100, 0·35)

Skew

Question 1

In a sample of the lengths of 40 worms, the lower quartile was 8 *cm*, the median was 16 *cm* and the upper quartile was 27 *cm*. The length of the shortest worm was 5 *cm*, and the longest was 34 *cm*. Describe the skewness of the sample. Give a reason for your answer.

Answer

$Q_1 = 8$, $Q_2 = 16$ and $Q_3 = 27$, $\Rightarrow$ $Q_3 - Q_2 = 11 > Q_2 - Q_1 = 8$ $\Rightarrow$ *positive skew*.

Question 2

A sample of the weights of apples picked from one tree had mean 103 *g*, median 105 *g* and mode 106 *g*. Describe the skewness of the sample. Give a reason for your answer.

Answer

Mean = 103 < median = 105 < mode = 106 $\Rightarrow$ *negative skew*.

Binomial and Poisson distributions

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 There are exactly 2 outcomes, *faulty bolts* or not *faulty*.
The probability of a *faulty bolt* is constant
Choices of *bolts* are independent.
There is a fixed number of trials, *50 bolts selected*.

Question 2

State two conditions under which a Poisson distribution is a suitable model for the number of misprints on a page.

Answer

*Misprint*s occur *randomly / independently*, *singly* and at a *constant rate* any two of the 3

Question 3

An estate agent sells houses 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

Poisson, $P_o(7)$. *Sales of houses* occur *independently/randomly*, *singly*, at a *constant rate*.

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

The number of daisies in each of several equal size squares was counted. The mean number of daisies per square was 36·9 and the variance was 37·3. Explain why these figures support the choice of a Poisson distribution as a model.

Answer

For a *Poisson* model, *Mean* = *Variance* ; for *these data* $36{\cdot}9 \approx 37{\cdot}3 \Rightarrow$ Poisson

---

Approximations to Poisson and Binomial

Question 1

Over a long period it is known that 2% of articles produced are defective. A sample of 200 articles is taken. What distribution describes this situation, and what is a suitable approximation to estimate the probability that there are exactly 5 defective articles.

Answer

Binomial distribution $X \sim \text{B}(200\ ,\ 0{\cdot}02)$ describes this situation.
$n = 200$ is *large*, $p = 0{\cdot}02$ is *small* so use the Poisson approximation $Y \sim \text{P}_\text{o}\,(np) \Leftrightarrow \text{Po}\,(4)$.

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

n is *large*, $n >$ about 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 $\text{P}_\text{o}(np)$, **or $\text{P}_\text{o}(nq)$**

Question 4

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

Answer

If X ~ B(n,p) and

n is *large* or $n > 10$	or	$np > 5$ **or** $nq > 5$
p is *close to* 0.5	or	$nq > 5$ and $np > 5$

then $X \sim B(n,p)$ can be approximated by $Y \sim N(np, npq)$.

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 by name or number* 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

A bag has a large number of discs numbered 1, 2 or 3. One third of the discs have the number 1, half of the discs have the number 2 and one sixth of the discs have the number 3. Samples of size 3 are taken from the bag. What is meant by the sampling distribution of the median, *M*.?

Answer

Find all possible samples, with their probabilities, of three discs and calculate their medians. Then give all possible values of the median, 1 or 2 or 3, together with their probabilities.

Question 4

11% of pupils in a large school are left handed. Samples of size 20 pupils are taken. Describe the sampling distribution of the number of left handed pupils.

Answer

To find the sampling distribution of the number of left handed pupils, we take all possible values of the number of left handed pupils, 0, 1, 2, 3, …, 20 in a sample, and calculate their probabilities.

This is the binomial distribution, $\text{P}(X) = r = {}^{20}C_r\,(1 - 0{\cdot}11)^{20-r} \times 0{\cdot}11^r$, which can be summarised as B(20, 0·11)

So the sampling distribution of the number of left handed people is B(20, 0·11).

---

9 Appendix

Mean and variance of B(n, p)

Proof of formulae

For a *single trial* with probabilities of success, p, and failure, q, and $p + q = 1$

number of success	probability		
y	p	yp	y^2p
0	q	0	0
1	p	p	p

$\mathrm{E}[Y] = \sum yp = p$

$\mathrm{Var}[Y] = \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2 = p - p^2 = p(1-p) = pq$

For n *independent single trials*, $X_1, X_2, X_3, \ldots, X_n$,

$\mathrm{E}[X_i] = p$

$\mathrm{Var}[X_i] = pq$, for $i = 1, 2, 3, \ldots, n$

Defined $X = X_1 + X_2 + X_3 + \ldots + X_n$,

then $X \sim \mathrm{B}(n, p)$

$\mu = \mathrm{E}[X] = \mathrm{E}[X_1 + X_2 + X_3 + \ldots + X_n] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \mathrm{E}[X_3] + \ldots + \mathrm{E}[X_n] = np$

$\sigma^2 = \mathrm{Var}[X] = \mathrm{Var}[X_1 + X_2 + X_3 + \ldots + X_n]$

$= \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \mathrm{Var}[X_3] + \ldots + \mathrm{Var}[X_n] = npq$ ……**I**

Thus for $X \sim \mathrm{B}(n, p)$

$\mu = \mathrm{E}[X] = np$

$\sigma^2 = \mathrm{Var}[X] = npq = np(1-p)$.

Note that the result **I** requires all X_i to be independent – see S3.

Mean and variance for a continuous uniform distribution

Proof of formulae

(*a*) Expected mean

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

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

or by symmetry.

(*b*) Expected variance

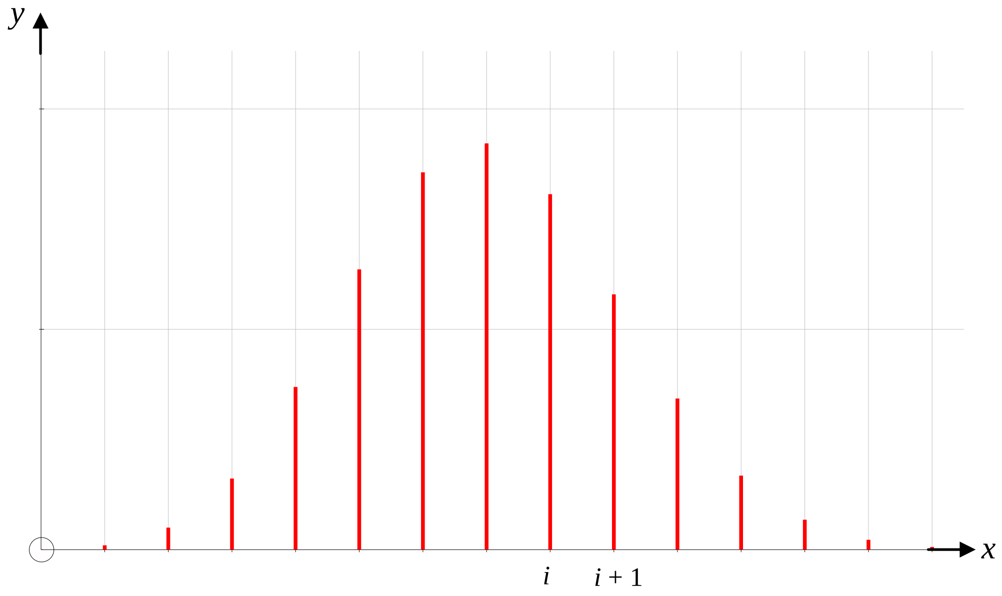
$$\mathrm{Var}[X] = \int x^2 f(x)\,dx - \mu^2 = \int_{\alpha}^{\beta} x^2\frac{1}{\beta-\alpha}\,dx - \left(\frac{\beta+\alpha}{2}\right)^2$$

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

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

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

Normal approximation to the Binomial



Let $X \sim B(n, p)$, shown in the above graph.

Successive values of x are $i, i+1 \Rightarrow \delta x = 1$.

Let corresponding values of y be y_i and y_{i+1}

$$\Rightarrow \qquad y_i = {}^nC_i\, p^i q^{n-i} = \frac{n!}{(n-i)!(i)!} p^i q^{n-i}$$

and

$$y_{i+1} = {}^nC_{i+1}\, p^{i+1} q^{n-i-1} = \frac{n!}{(n-i-1)!(i+1)!} p^{i+1} q^{n-i-1}$$

$$= \frac{n!}{(n-i)!(i)!} p^i q^{n-i} \times \frac{n-i}{i+1} \times \frac{p}{q}$$

$$= y_i \times \frac{n-i}{i+1} \times \frac{p}{q}$$

and so the difference in successive values of y is $\delta y = y_{i+1} - y_i$

$$\Rightarrow \qquad \delta y = y_i \left(\frac{n-i}{i+1} \times \frac{p}{q} - 1\right)$$

We shall be allowing n to become infinitely large, so we change the variables to keep the graph on the page.

Let $X = \frac{x-np}{\sqrt{npq}} = \frac{i-np}{\sqrt{npq}}$

$\Rightarrow \overline{X} = 0$, and the variance of X is 1

$\Rightarrow \quad i = np + X\sqrt{npq}$

and since $\delta x = 1$, $\delta X = \frac{1}{\sqrt{npq}}$.

To keep the area as 1, let $Y_i = y_i\sqrt{npq}$

$$\Rightarrow \quad \delta Y = \delta y\sqrt{npq} = y_i\left(\frac{n-i}{i+1}\times\frac{p}{q}-1\right)\sqrt{npq}$$

$$= y_i\left(\frac{np-ip-qi-q}{(i+1)q}\right)\sqrt{npq}$$

$$= y_i\left(\frac{np-i-q}{(i+1)q}\right)\sqrt{npq} \qquad \text{since } p+q=1$$

$$= y_i\left(\frac{np-np-X\sqrt{npq}-q}{(np+X\sqrt{npq}+1)q}\right)\sqrt{npq} \qquad \text{since } i = np + X\sqrt{npq}$$

$$= y_i\left(\frac{-npqX-q\sqrt{npq}}{(np+X\sqrt{npq}+1)q}\right)$$

$$\cong y_i\left(\frac{-npqX}{(np)q}\right) \qquad \text{when } n \text{ is large}$$

$$\Rightarrow \quad \delta Y \cong \frac{Y_i}{\sqrt{npq}}(-X) \qquad \text{since } Y_i = y_i\sqrt{npq}$$

$$\Rightarrow \quad \delta Y \cong -XY\ \delta X \qquad \text{since } \delta X = \frac{1}{\sqrt{npq}}$$

Let $\delta X \to 0$

$$\Rightarrow \quad \frac{dY}{dX} = -XY$$

$$\Rightarrow \quad \int \frac{1}{Y}\,dY = \int -X\,dX$$

$$\Rightarrow \quad \ln Y = -\frac{1}{2}X^2 + \ln A$$

$$\Rightarrow \quad Y = A\,e^{-\frac{1}{2}X^2}$$

Because the area under the binomial histogram was 1, the area under the curve is also 1 (the transformations ensured that the area did not change)

$$\Rightarrow \quad \int_{-\infty}^{\infty} A\,e^{-\frac{1}{2}X^2}\,dX = 1$$

We can show, see next page, that

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}X^2}\,dX = \sqrt{2\pi} \qquad \Rightarrow \qquad A = \frac{1}{\sqrt{2\pi}}$$

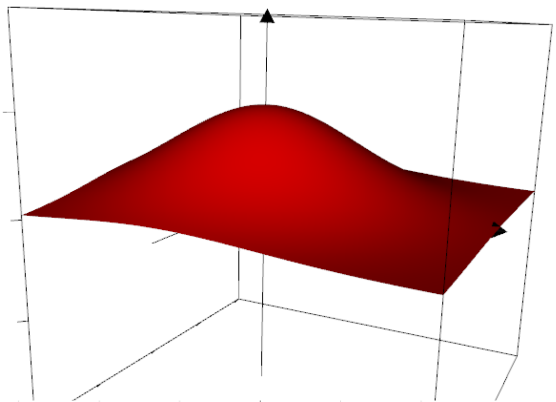
$\Rightarrow \quad Y = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}X^2}$, which is the equation of the normal distribution $N(0, 1^2)$.

Integral of e^(-0.5x^2)

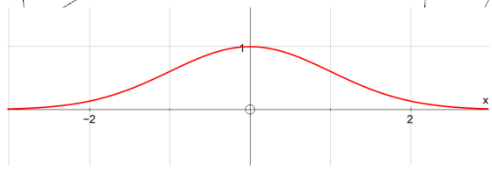
The surface

$z = e^{-0.5(x^2+y^2)}$, in Cartesian coordinates can also be written as

$z = e^{-0.5r^2}$ in cylindrical polar coordinates (use polar coordinates in the *xy*-plane and then move up *z*).



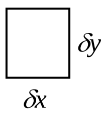
This surface can be thought of as the curve $y = e^{-0.5x^2}$ rotated through 2π about the *z*-axis.



To find the volume, we take a small region in the *xy*-plane, or in the $r\theta$-plane, multiply its area by *z* to find the volume of a vertical 'pillar', and then do a double summation to find the total volume.

In the *xy*-plane

area of rectangle is $\delta x \times \delta y$



$\Rightarrow$ volume of pillar is $z\, \delta x\, \delta y$

$\Rightarrow$ total volume is $\sum_x \sum_y z\, \delta x\, \delta y$

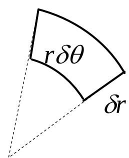
$$\to \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-0.5(x^2+y^2)}\, dx\, dy$$

$$= \int_{-\infty}^{\infty} e^{-0.5x^2} dx \int_{-\infty}^{\infty} e^{-0.5y^2}\, dy$$

$$= \left(\int_{-\infty}^{\infty} e^{-0.5x^2} dx\right)^2$$

In the $r\theta$-plane

area of element is approx $r\ \delta\theta\, \delta r$



$\Rightarrow$ volume of pillar is $z\ r\ \delta\theta\, \delta r$

$\Rightarrow$ total volume is $\sum_r \sum_\theta z\ r\ \delta\theta\, \delta r$

$$\to \int_0^{\infty}\int_0^{2\pi} re^{-0.5r^2}\, dr\, d\theta$$

$$= \int_0^{\infty} re^{-0.5r^2} dr \int_0^{2\pi} d\theta$$

$$= \left[-e^{-0.5r^2}\right]_0^{\infty} \times \left[\theta\ \right]_0^{2\pi}$$

$$= 2\pi$$

$$\Rightarrow \left(\int_{-\infty}^{\infty} e^{-0.5x^2} dx\right)^2 = 2\pi$$

$$\Rightarrow \int_{-\infty}^{\infty} e^{-0.5x^2} dx = \sqrt{2\pi}$$

Poisson probabilities from first principles

Preliminary result

In the Poisson distribution events must occur *uniformly* (or *at a constant rate*), *singly, randomly and independently*.

In a Poisson distribution let μ be the mean number of occurrences in an interval of length 1, then in an interval of length t the mean number of occurrences will be μt.

Divide the interval of length t into n very small intervals, each of length δt, then $t = n\,\delta t$.

As events occur singly, the chance of one event occurring in an interval δt is small, and the chance of two events is negligible.

We can approximate this situation by saying that the probability of one event in an interval δt is p, and the probability of more than one event is 0. Thus the probability of 0 events in an interval δt is $1 - p$.

Now we have a binomial distribution $B(n, p)$ which gives the probabilities (approximately) of i events in an interval of $t = n\,\delta t$, where $i = 0, 1, 2, \ldots, n$.

The mean in an interval of t, considered as $B(n, p)$ is np, and, considered as a Poisson distribution the mean is μt.

$\Rightarrow \quad np \cong \mu t$, and together with $t = n\,\delta t$,

$\Rightarrow \quad p \cong \mu\,\delta t$.

In the following we take

P(1 event in an interval δt) = $\mu\,\delta t$ and P(0 events in an interval δt) = $(1 - \mu\,\delta t)$.
We discount the possibility of more than 1 event in an interval δt.

Deriving the Poisson probabilities

Let $p_i(t)$ = probability of i events in an interval of length t.

Then $\quad p_0(\delta t) = 1 - \mu\,\delta t, \quad p_1(\delta t) = \mu\,\delta t$

and $\quad p_0(t + \delta t) = $ P(0 event in t **and** 0 events in δt)

$\quad = p_0(t) \times p_0(\delta t)$ $\quad$ since events are independent

$\Rightarrow \quad p_0(t + \delta t) = p_0(t) \times (1 - \mu\,\delta t)$

$$\Rightarrow \quad \frac{p_0(t+\delta t) - p_0(t)}{\delta t} = -\mu\, p_0(t)$$

as $\delta t \to 0$, $\frac{dp_0}{dt} = -\mu\, p_0$

$$\Rightarrow \quad \int \frac{1}{p_0}\, dp_0 = \int -\mu\, dt \quad \Rightarrow \quad p_0(t) = Ae^{-\mu t}$$

But $p_0(0) = 1 \Rightarrow A = 1$ $\quad$ since the probability of 0 events in an interval of length 0 is 1

$\Rightarrow \quad p_0(t) = e^{-\mu t}$

Now $p_1(t + \delta t)$ = P({1 event in t **and** 0 events in δt} **or** {0 events in t **and** 1 event in δt})

$= p_1(t) \times p_0(\delta t) + p_0(t) \times p_1(\delta t)$ since events are independent

$\Rightarrow \quad p_1(t + \delta t) = p_1(t) \times (1 - \mu\,\delta t) + e^{-\mu t} \times \mu\,\delta t$

$\Rightarrow \quad \frac{p_1(t+\delta t) - p_1(t)}{\delta t} = -\mu\, p_1(t) + \mu\, e^{-\mu t}$

as $\delta t \to 0$, $\frac{dp_1}{dt} = -\mu\, p_1 + \mu\, e^{-\mu t}$

$\Rightarrow \quad \frac{dp_1}{dt} + \mu\, p_1 = \mu\, e^{-\mu t}$ integrating factor is $e^{\int \mu\, dt} = e^{\mu t}$

$\Rightarrow \quad \frac{d}{dt}(e^{\mu t} p_1) = \mu$

$\Rightarrow \quad p_1(t) = \mu t e^{-\mu t} + c\, e^{-\mu t}$

$\Rightarrow \quad p_1(t) = \mu t e^{-\mu t}$ since the probability of 1 event in an interval of length 0 is 0

Moving up $p_2(t + \delta t)$ = P({2 events in t **and** 0 events in δt} **or** {1 event in t **and** 1 event in δt})

Note that our model does not allow more than 1 event in an interval δt.

$\Rightarrow \quad p_2(t + \delta t) = p_2(t) \times p_0(\delta t) + p_1(t) \times p_1(\delta t)$ since events are independent

$\Rightarrow \quad p_2(t + \delta t) = p_2(t) \times (1 - \mu\,\delta t) + \mu t e^{-\mu t} \times \mu\,\delta t$

$\Rightarrow \quad \frac{p_2(t+\delta t) - p_2(t)}{\delta t} = -\mu\, p_2(t) + \mu^2 t e^{-\mu t}$

as $\delta t \to 0$, $\frac{dp_2}{dt} = -\mu\, p_2 + \mu^2 t e^{-\mu t}$

$\Rightarrow \quad \frac{dp_2}{dt} + \mu\, p_2 = \mu^2 t e^{-\mu t}$ integrating factor is $e^{\int \mu\, dt} = e^{\mu t}$

$\Rightarrow \quad \frac{d}{dt}(e^{\mu t} p_2) = \mu^2 t$

$\Rightarrow \quad p_2(t) = \frac{\mu^2 t^2 e^{-\mu t}}{2} + c\, e^{-\mu t}$

$\Rightarrow \quad p_2(t) = \frac{\mu^2 t^2 e^{-\mu t}}{2}$ since the probability of 2 events in an interval of length 0 is 0

Continuing in the same way we get

$p_3(t) = \frac{\mu^3 t^3 e^{-\mu t}}{3!}$, etc.

Notice that μt is the mean number of occurrences in an interval of length t. If we write $\lambda = \mu t$

we have $p_0 = e^{-\lambda}$, $p_1 = \lambda\, e^{-\lambda}$, $p_2 = \frac{\lambda^2 e^{-\lambda}}{2!}$, $p_3 = \frac{\lambda^3 e^{-\lambda}}{3!}$, ….. , $p_r = \frac{\lambda^r e^{-\lambda}}{r!}$,

which are the Poisson probabilities for an interval with mean number of occurrences λ.

Note that the proof can be formalised by using proof by induction.

Index