Extra Workshop on Probability/Probability Distributions

Wednesday 3rd December

2pm – 4pm

Herschel Building, Lecture Theatre 3

Semester 1 nearly over!

This week

- Extra workshop on Wednesday
- CBA3 in practice mode
- Should now be working through assignment 1

Next week

- Assignment 1 due in Wednesday
- CBA3 due in Friday (exam mode)
- Lecture and tutorial as normal – revision worksheet given out in lecture

Monday 5th January 2009

- Last lecture of Semester 1
- A revision booklet will be given out
- Revision tutorials

Semester 2 starts Monday 26th January 2009

Lecture 10

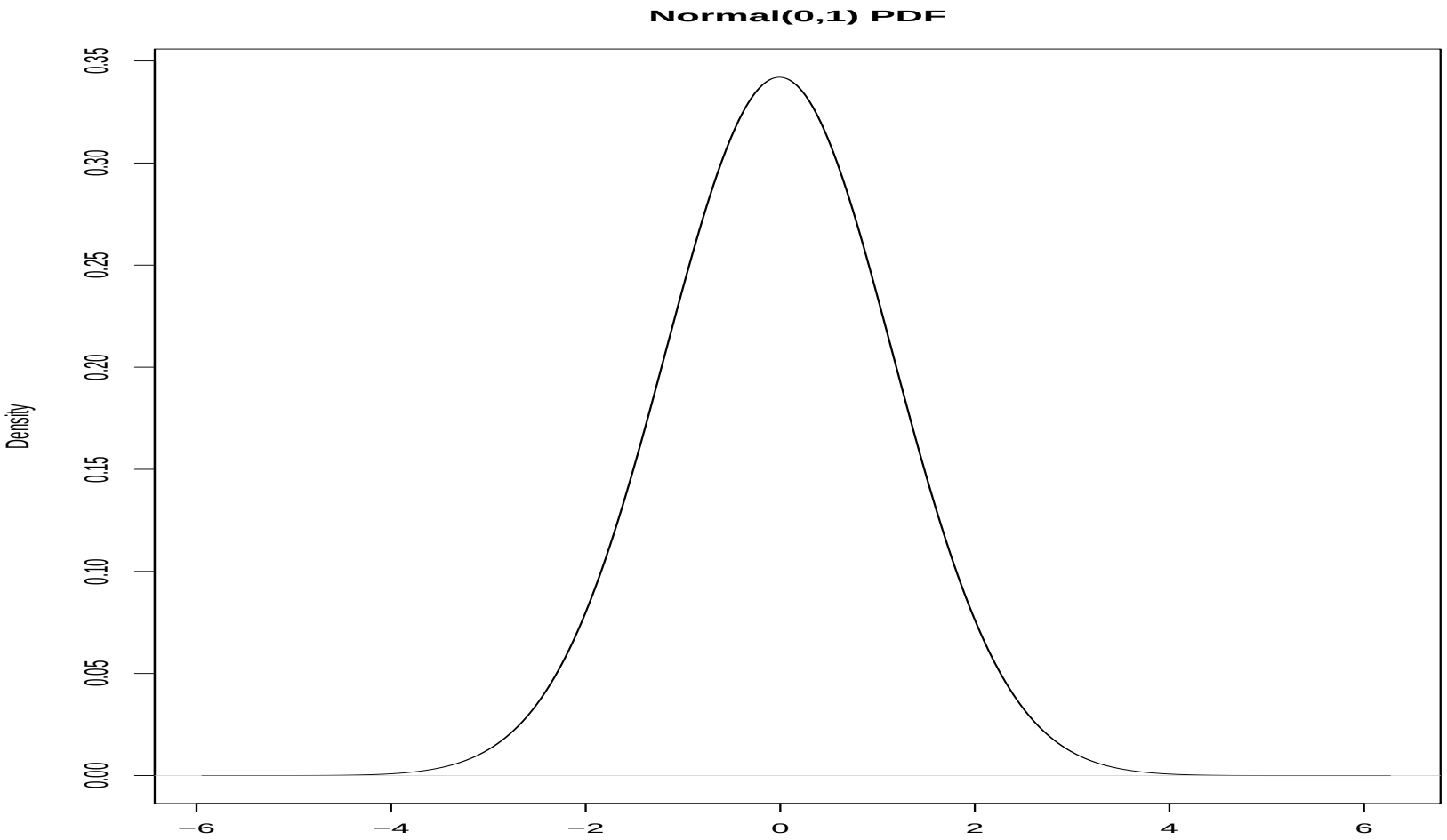
MORE CONTINUOUS PROBABILITY MODELS

Introduction

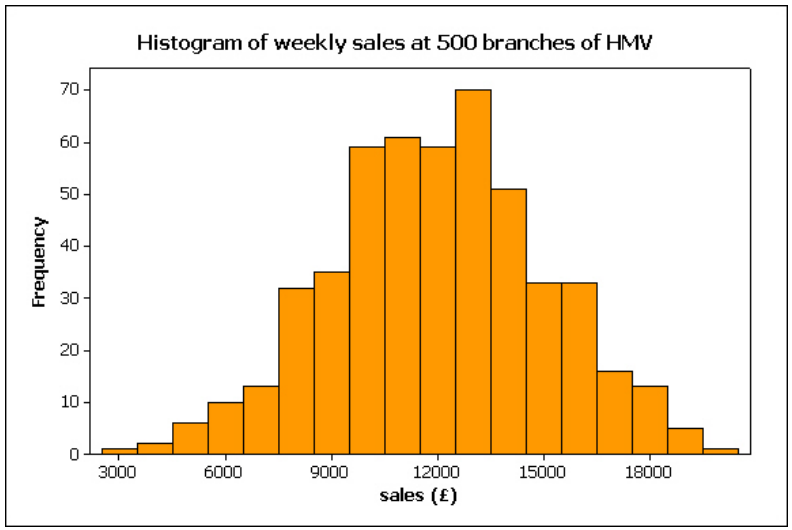
Over the past few weeks we have talked about some "**standard**" probability distributions which can be used to model data. So far, we have looked at:

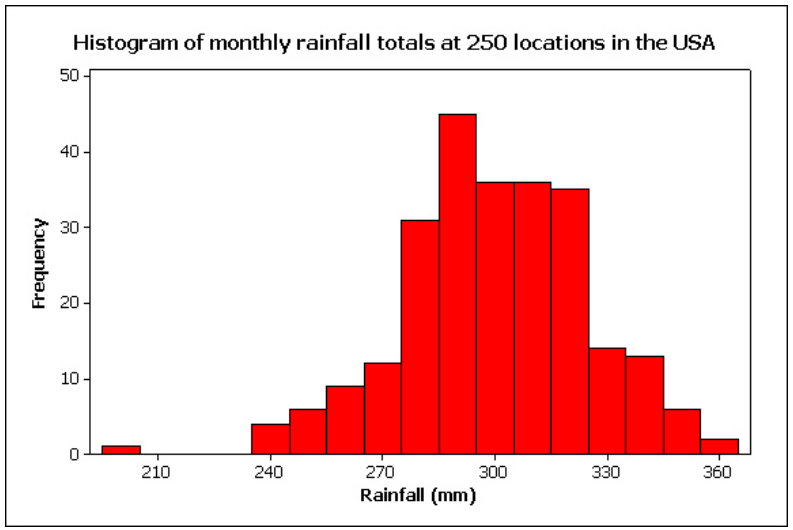
1. **Discrete distributions**
   - The **Binomial** distribution
   - The **Poisson** distribution
2. **Continuous distributions**
   - The **Normal** distribution

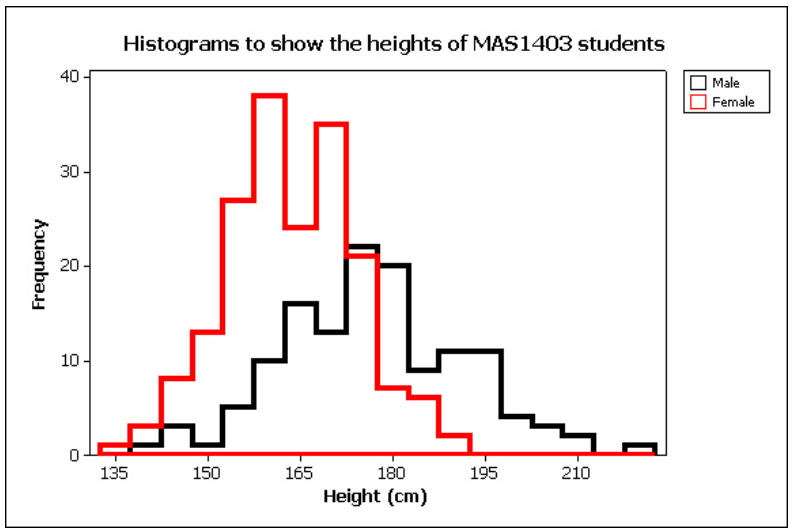
Recall the **probability density function** of the Normal distribution, which is often referred to as a "**bell–shaped curve**":

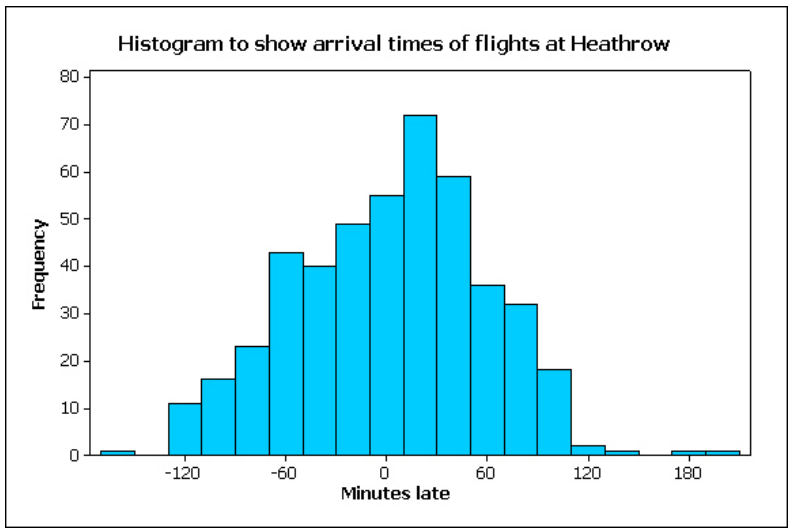


Recall also from last week that many naturally occurring measurements seem to follow this distribution:









But what if we **cannot** assume "Normality" for our data?

Example of "non–Normality"

- You manage a group of Environmental Health Officers and need to decide at what time they should inspect a local hotel
- You decide that any time during the working day (9.00 to 18.00) is okay
- You want to decide the time "randomly"

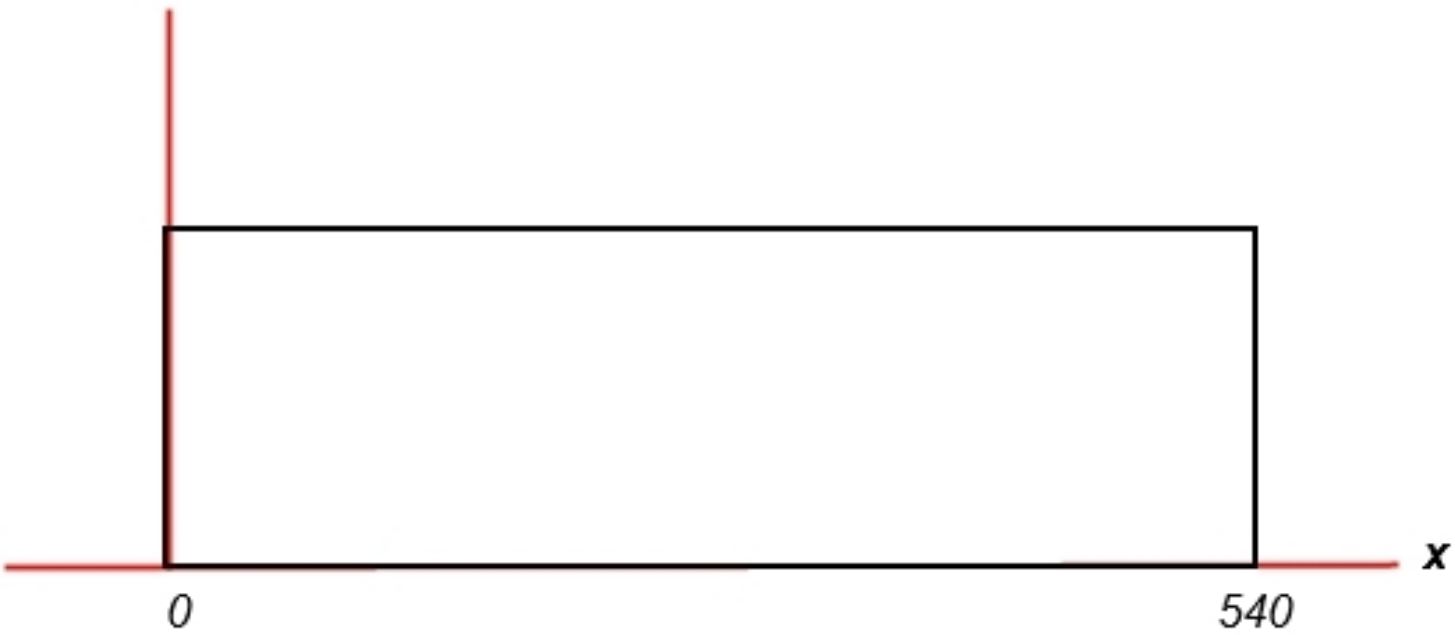
Here, "**randomly**" is a short–hand for

"*a random time, where all times in the working day are equally likely to be chosen*"

The Uniform distribution

Let X be the time to their arrival at the hotel, measured in terms of minutes from the start of the day.

Then X is a **Uniform** random variable between 0 and 540 (page 119):



As with the Normal distribution, the total area (base $\times$ height) under the pdf must equal one.

Therefore, as the base is 540, the height must be $1/540$.

Hence the **probability density function** (pdf) for the continuous random variable X is

$$f(x) = \begin{cases} \dfrac{1}{540} & \text{for } 0 \leq x \leq 540 \\ 0 & \text{otherwise.} \end{cases}$$

In general, we say that a random variable X which is equally likely to take any value between a and b has a **uniform distribution** on the interval **a** to **b**, i.e.

$$X \sim U(a, b).$$

The random variable has **probability density function** (pdf)

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

and probabilities can be calculated using the formula

$$P(X \leq x) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b. \end{cases}$$

Therefore, for example, the **probability that the inspectors visit the hotel in the morning** (within 180 minutes after 9am) is

$$P(X \leq 180) = \frac{180 - 0}{540 - 0} = \frac{1}{3}.$$

The **probability of a visit during the lunch hour** (12.30 to 13.30) is

$$\begin{aligned} P(210 \leq X \leq 270) &= P(X \leq 270) - P(X < 210) \\ &= \frac{270 - 0}{540 - 0} - \frac{210 - 0}{540 - 0} \\ &= \frac{270 - 210}{540} \\ &= \frac{60}{540} \\ &= \frac{1}{9}. \end{aligned}$$

Mean and Variance

Recall that:

- If $X \sim \text{bin}(n, p)$, then
  - $E(X) = n \times p$ and
  - $\text{Var}(X) = n \times p \times (1 - p)$

- If $X \sim \text{Po}(\lambda)$, then
  - $E(X) = \lambda$ and
  - $\text{Var}(X) = \lambda$

We have equivalent formulae for $X \sim U(a, b)$:

$$\begin{aligned} E(X) &= \frac{a+b}{2} \\ \text{Var}(X) &= \frac{(b-a)^2}{12}. \end{aligned}$$

In the above example, we have

$$E(X) = \frac{a+b}{2} = \frac{0+540}{2} = 270,$$

so that the mean arrival of the inspectors is 9am+270 minutes = 13.30.

Also

$$\text{Var}(X) = \frac{(540-0)^2}{12} = 24300,$$

and therefore $SD(X) = \sqrt{Var(X)} = \sqrt{24300} = 155.9$ minutes.

The Exponential Distribution

The **exponential distribution** is another common distribution that is used to describe continuous random variables.

It is often used to model lifetimes of products and times between "random" events, for example:

- Lifetime of bulbs
- Arrival of customers in a queueing system
- Arrival of orders

The distribution has one parameter, λ. If our random variable X follows an **exponential distribution**, then we say

$$X \sim \exp(\lambda).$$

Its **probability density function** is

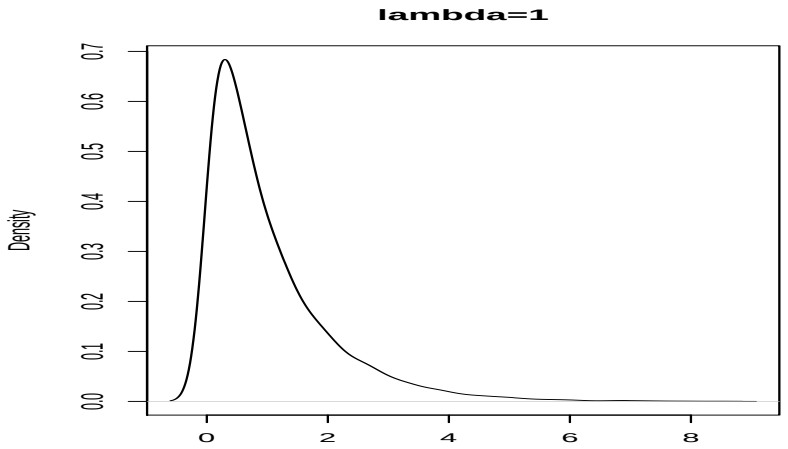
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

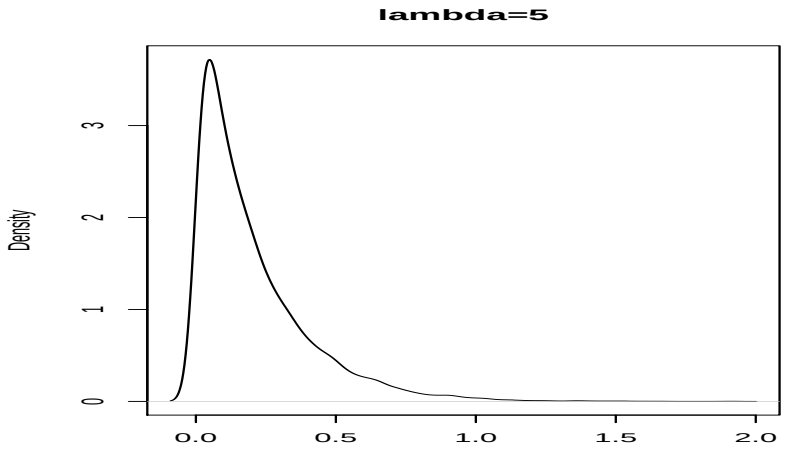
and probabilities can be calculated using

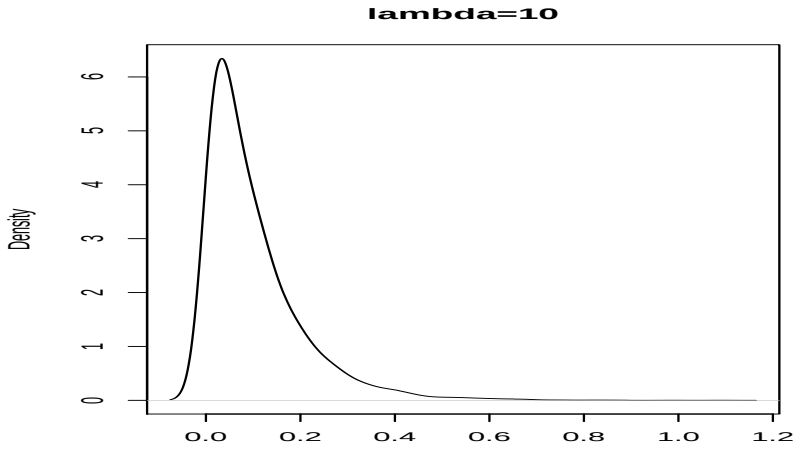
$$P(X \leq x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } x > 0. \end{cases}$$

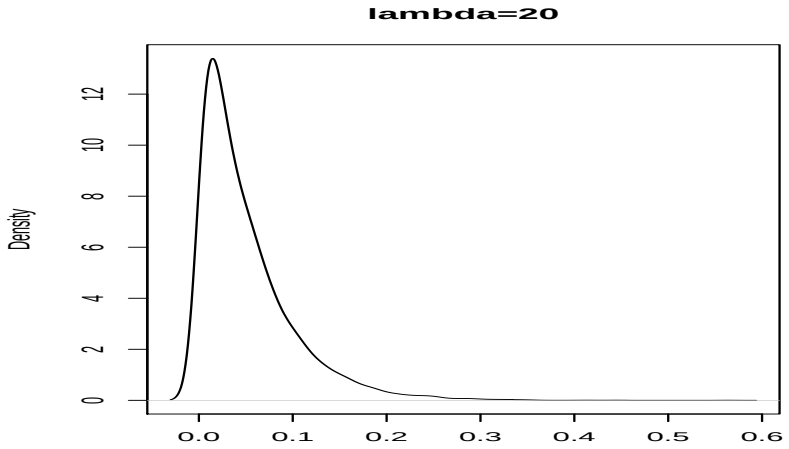
The main **features** of this distribution are:

1. an exponentially distributed random variable can only take **positive** values
2. larger values are **increasingly unlikely** – "**exponential decay**"
3. the value of λ fixes the **rate of decay** – *larger values correspond to more rapid decay*.


lambda=1
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8


lambda=5
Density
0
1
2
3
0.0
0.5
1.0
1.5
2.0


lambda=10
Density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
1.2


lambda=20
Density
0
2
4
6
8
10
12
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Consider an example in which the time (in minutes) between successive users of a pay phone can be modelled by an exponential distribution with $\lambda = 0.3$.

The probability of the gap between phone users being less than 5 minutes is

$$P(X < 5) = 1 - e^{-0.3\times 5} = 1 - 0.223 = 0.777.$$

Also the probability that the gap is more than 10 minutes is

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \left(1 - e^{-0.3\times 10}\right) = e^{-0.3\times 10} = 0.050$$

and the probability that the gap is between 5 and 10 minutes is

$$P(5 < X < 10) = P(X < 10) - P(X \leq 5) = 0.950 - 0.777 = 0.173.$$

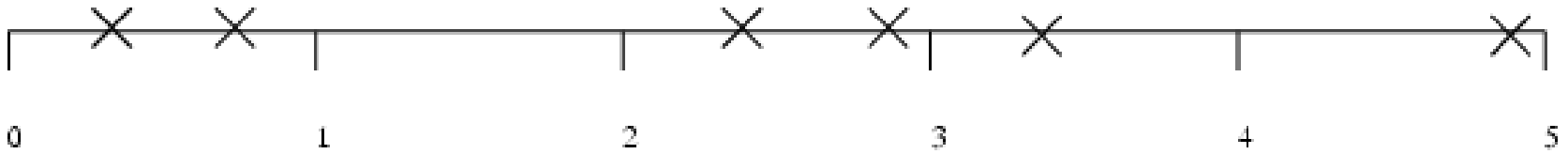
One of the main uses of the exponential distribution is as a model for the **times between events occurring randomly in time**.

We have previously considered events which occur at random points in time in connection with the **Poisson distribution**.

The Poisson distribution describes probabilities for the number of events taking place in a given time period.

The exponential distribution describes probabilities for the times between events. Both of these concern events occurring randomly in time (at a constant average rate, say λ). This is known as a **Poisson process**.

Consider a series of randomly occurring events such as calls at a credit card call centre. The times of calls might look like



We can view these data in **two ways**:

- The number of calls in each minute (here 2, 0, 2, 1 and 1)
- the times between successive calls

For the **Poisson process**,

- the number of calls has a **Poisson** distribution with parameter λ, and
- the time between successive calls has an **exponential** distribution with parameter λ.

Mean and Variance

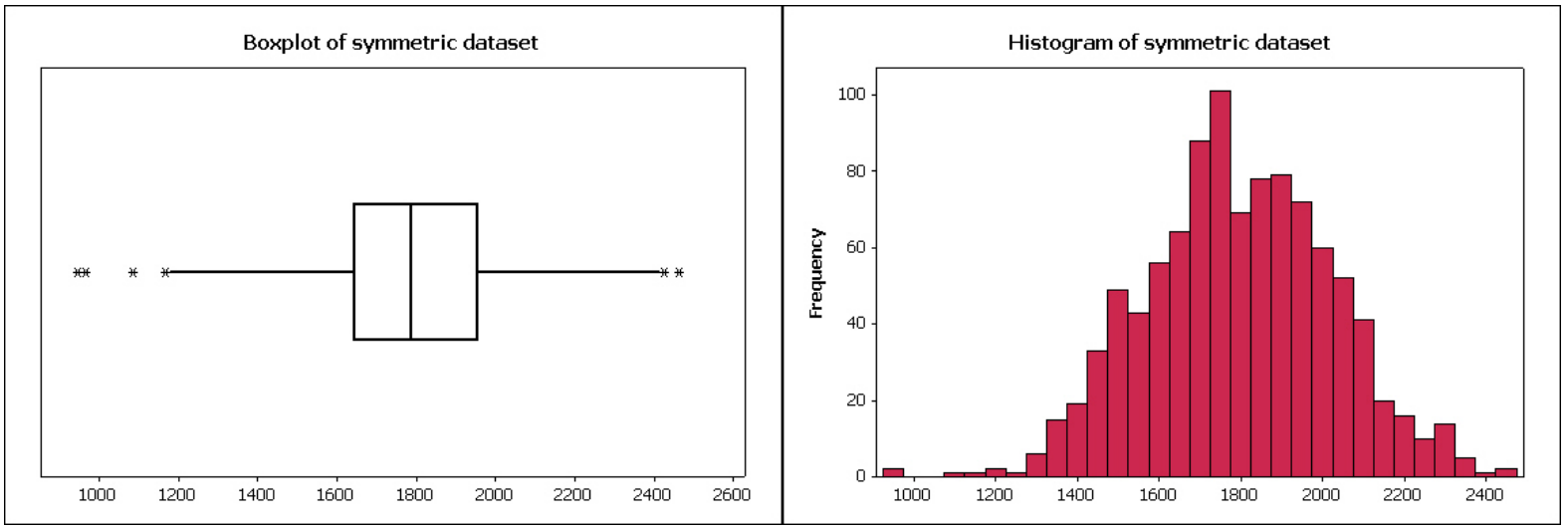
The mean and variance of the exponential distribution can be shown to be

$$E(X) = \frac{1}{\lambda}, \qquad \text{Var}(X) = \frac{1}{\lambda^2}.$$

Commenting on graphs

Is your graph **symmetric** or **asymmetric**?

Are there any **outliers**?



Commenting on graphs

When comparing groups...

- Highest/lowest?
- Overlap/completely separate?
- which has greater **variability**?