

# Maths Stats Formulas and Definitions

Boyd Kane

May 20, 2021

## Contents

<b>1</b>	<b>Common Distributions</b>	<b>3</b>
1.1	Poisson Distribution . . . . .	3
1.2	Negative Binomial Distribution . . . . .	3
1.3	Geometric Distribution . . . . .	4
1.4	Binomial Distribution . . . . .	4
1.5	Beta Distribution . . . . .	4
1.6	Normal Distribution . . . . .	4
1.7	Gamma Distribution . . . . .	4
1.7.1	Notes . . . . .	5
1.8	Exponential Distribution . . . . .	5
1.9	Chi-Squared Distribution . . . . .	5
<b>2</b>	<b>Generating Functions</b>	<b>5</b>
2.1	Probability Generating Functions . . . . .	5
2.2	Moment Generating Functions . . . . .	6
<b>3</b>	<b>Infinite Sums</b>	<b>6</b>
<b>4</b>	<b>3041F: Chains</b>	<b>6</b>
4.1	Chapman-Kolmogorov Equations . . . . .	6
<b>5</b>	<b>3041F: Branching Processes</b>	<b>6</b>
<b>6</b>	<b>3041F: Probability of Extinction</b>	<b>7</b>
<b>7</b>	<b>3041F: First Passage Probabilities</b>	<b>7</b>

<b>8</b>	<b>3041F: Discrete Time</b>	<b>8</b>
8.1	Calculating the Count at a given Time . . . . .	8
8.1.1	Forecasting . . . . .	8
8.1.2	Backcasting . . . . .	8
8.2	Calculating the Time until a given Count . . . . .	8
8.2.1	Forecasting . . . . .	9
8.2.2	Backcasting . . . . .	9
<b>9</b>	<b>3041F: Continuous Time</b>	<b>9</b>
9.1	Calculating the Count at a given Time . . . . .	10
9.1.1	Forecasting . . . . .	10
9.1.2	Backcasting . . . . .	10
9.2	Calculating the Time until a given Count . . . . .	10
9.2.1	Forecasting . . . . .	11
9.2.2	Backcasting . . . . .	11
<b>10</b>	<b>3041F: Poisson Processes</b>	<b>11</b>
10.1	3041F: Probability Integral Transform . . . . .	11
10.2	3041F: Estimation . . . . .	11
<b>11</b>	<b>3041F: Basic Poisson Process</b>	<b>12</b>
11.1	3041F: Poisson Postulates . . . . .	12
11.2	Poisson Process Information . . . . .	12
11.3	3041F: Modelling Poisson Processes . . . . .	13
11.4	3041F: Estimation of Poisson Processes . . . . .	13
<b>12</b>	<b>3041F: Compound Poisson Processes</b>	<b>14</b>
<b>13</b>	<b>3041F: Non-homogeneous Poisson Processes</b>	<b>14</b>
13.1	3041F: Mean Value Function . . . . .	15
13.2	3041F: Simulation of a non-homogeneous Poisson Process . . .	16
13.2.1	The Thinning algorithm . . . . .	16
13.2.2	Subintervals for simulations . . . . .	17
13.2.3	Combining homogeneous Poisson Processes . . . . .	17
13.3	3041F: Estimation of non-homogeneous Poisson Processes . .	17
13.3.1	Likelihood for non-homogeneous Poisson Processes . .	17

<b>14 3041F: Time Series Analysis</b>	<b>18</b>
14.1 Trend . . . . .	20
14.2 Seasonality . . . . .	20
14.3 Cycles . . . . .	21
14.4 Transformations . . . . .	21
<b>15 3041F: Transformations</b>	<b>21</b>
15.1 3041F: Box-Cox Transformation . . . . .	22
15.2 3041F: Fitting Linear Filters . . . . .	23
15.3 3041F: Moving Averages . . . . .	25
<b>16 Terms and Definitions</b>	<b>27</b>
<b>17 R trickery</b>	<b>29</b>
17.1 runif . . . . .	29
17.2 cumsum . . . . .	30
17.3 plot . . . . .	30
17.4 ts . . . . .	30
<b>18 Notes</b>	<b>31</b>

# 1 Common Distributions

## 1.1 Poisson Distribution

$$\begin{aligned}
X &\sim Poi(\lambda) & f_X(x) &= \frac{\lambda^x e^{-\lambda}}{x!} \\
\mathbb{E}[X] &= \lambda & \mathbb{E}[s^X] &= e^{\lambda(e^s - 1)} \\
\mathbb{V}[X] &= \lambda & \left( \sum_{i=0}^n Poi(\lambda_i) \right) &\sim Poi\left( \sum_{i=0}^n \lambda_i \right)
\end{aligned}$$

## 1.2 Negative Binomial Distribution

$$\begin{aligned}
X &\sim NegBin(p, r) & f_X(x) &= \binom{k-1}{r-1} p^r q^{k-r} \\
\mathbb{E}[X] &= \frac{r}{p} & \mathbb{E}[e^{sX}] &= \left( \frac{pe^s}{1 - qe^s} \right)^r \\
\mathbb{V}[X] &= \frac{rq}{p^2}
\end{aligned}$$

### 1.3 Geometric Distribution

$$\begin{aligned} X &\sim \text{Geo}(p) \\ \mathbb{E}[X] &= \frac{1}{p} & f_X(x) &= pq^{x-1} \\ \mathbb{V}[X] &= \frac{q}{p^2} & \mathbb{E}[e^{sX}] &= \frac{e^tp}{1 - e^tq} \end{aligned}$$

### 1.4 Binomial Distribution

$$\begin{aligned} X &\sim \text{Binomial}(n, p) \\ \mathbb{E}[X] &= np & f_X(x) &= \binom{n}{x} p^x q^{n-x} \\ \mathbb{V}[X] &= npq & \mathbb{E}[e^{sX}] &= (1 - p + pe^s)^n \end{aligned}$$

### 1.5 Beta Distribution

$$0 < x < 1$$

$$\begin{aligned} X &\sim \text{Beta}(a, b) \\ \mathbb{E}[X] &= \frac{a}{a+b} & f_X(x) &= \frac{(a+b-1)!}{(a-1)!(b-1)!} \cdot x^{a-1}(1-x)^{b-1} \\ \mathbb{V}[X] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

### 1.6 Normal Distribution

$$\begin{aligned} X &\sim N(\mu, \sigma^2) \\ \mathbb{E}[X] &= \mu & f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\ \mathbb{V}[X] &= \sigma^2 & \mathbb{E}[e^{sX}] &= e^{\mu s + \frac{1}{2}s^2\sigma^2} \end{aligned}$$

### 1.7 Gamma Distribution

$$\lambda > 0, \alpha > 0$$

$$\begin{aligned} X &\sim \text{Gamma}(\lambda, \alpha) \\ \mathbb{E}[X] &= \frac{\alpha}{\lambda} & f_X(x) &= \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{(\alpha-1)!} \\ \mathbb{V}[X] &= \frac{\alpha}{\lambda^2} & \mathbb{E}[e^{sX}] &= \left(\frac{\lambda}{\lambda-s}\right)^\alpha \end{aligned}$$

### 1.7.1 Notes

Sum of independant Exponentials is Gamma distributed:

$$\left( \sum^n Exp(\lambda) \right) \sim Gamma(\lambda, n)$$

$$Exp(\lambda) \equiv Gamma(\lambda, 1)$$

## 1.8 Exponential Distribution

$$\begin{aligned} X &\sim Exp(\lambda) & f_X(x) &= \lambda e^{-\lambda x} \quad x \geq 0, \lambda > 0 \\ \mathbb{E}[X] &= \frac{1}{\lambda} & F_X(x) &= 1 - e^{-\lambda x} \\ \mathbb{V}[X] &= \frac{1}{\lambda^2} & \mathbb{E}[e^{sX}] &= \frac{\lambda}{\lambda - s} \quad s < \lambda \\ \hat{\lambda}_{MLE} &= \frac{1}{\bar{x}} & \mathbb{P}[Exp(\lambda_1) < Exp(\lambda_2)] &= \frac{\lambda_1}{\lambda_1 + \lambda_2} \end{aligned}$$

## 1.9 Chi-Squared Distribution

Work in progress

# 2 Generating Functions

## 2.1 Probability Generating Functions

$$\begin{aligned} \mathbb{G}_X(s) &:= \mathbb{E}[s^X] \\ &= \sum_{x=0}^{\infty} s^x \cdot f_X(x) \\ \mathbb{G}_X(0) &= f_X(0) = \mathbb{P}[X = 0] \\ \mathbb{G}_X(1) &= \sum_{x=0}^{\infty} 1^x \cdot f_X(x) \\ &= 1 \\ \mathbb{E}[X] &= \mathbb{G}'_X(1) \\ \mathbb{V}[X] &= \mathbb{G}''_X(1) + \mathbb{G}'_X(1) - (\mathbb{G}'_X(1))^2 \end{aligned}$$

## 2.2 Moment Generating Functions

Work in progress.

## 3 Infinite Sums

Work in progress.

## 4 3041F: Chains

### 4.1 Chapman-Kolmogorov Equations

Basically, for Markov chains we can just do matrix multiplication to go from time  $i$  to time  $i + 1$ . For  $0 \leq v \leq t$ , and  $p_{ij}(s, t) := \mathbb{P}[X_{s+t} = j | X_s = i]$ :

$$p_{ij}(s, t) = \mathbb{P}[X_{s+t} = j | X_s = i] = \sum_{k \in \mathcal{U}} p_{ik}(s, v) \cdot p_{kj}(s + v, t - v)$$

Or, in matrix notation:

$$\mathbf{P}^{(t)}(s) = \mathbf{P}^{(v)}(s) \cdot \mathbf{P}^{(t-v)}(s + v)$$

And under time homogeneity

$$\mathbf{P}^{(t)} = \mathbf{P}^v \cdot \mathbf{P}^{t-v} = \mathbf{P}^t$$

## 5 3041F: Branching Processes

Let  $Z_m$  be the number of individuals in Generation  $m$ . Let  $X_{im}$  be the number of offspring produced by individual  $i$  in generation  $m$ . Assume all  $X_{im} = X$ , so all  $X_{im}$  are identically distributed. Assume  $Z_0 = 1$  which implies  $Z_1 = X$ .

$$\mathbb{E}[Z_m] = \mathbb{E}[X]^m$$

Let  $\mu = \mathbb{E}[X]$  and  $\sigma^2 = \mathbb{V}[X]$ .

$$\begin{aligned} \mathbb{V}[Z_m] &= m\sigma^2 && \text{if } \mu = 1 \\ \mathbb{V}[Z_m] &= \sigma^2 \cdot \mu^{m-1} \left( \frac{\mu^m - 1}{\mu - 1} \right) && \text{if } \mu \neq 1 \end{aligned}$$

## 6 3041F: Probability of Extinction

To calculate the probability that the process is extinct by generation  $m$ :

$$\mathbb{P}[\text{Extinction by generation } m] = \mathbb{P}[Z_m = 0] = \mathbb{G}_m(0)$$

But to calculate the probability that the process is extinct at exactly generation  $m$ :

$$\mathbb{P}[\text{Extinction at generation } m] = \mathbb{G}_m(0) - \mathbb{G}_{m-1}(0)$$

Define the probability of extinction:

$$\begin{aligned} \eta &:= \mathbb{P}[\text{Eventual Extinction}] = \mathbb{P}\left[\bigcup_{m=1}^{\infty} \{Z_m = 0\}\right] \\ &= \text{smallest non-negative integer root of } \mathbb{G}_X(s) = s \end{aligned}$$

## 7 3041F: First Passage Probabilities

$$\begin{aligned} p_{ij}^{(n)} &:= \mathbb{P}[X_{t+n} = j | X_t = i] \\ P &= [p_{ij}] \\ P_{ij}(s) &:= \sum_{n=0}^{\infty} p_{ij}^{(n)} \cdot s^n \\ \mathbf{P}(s) &:= [P_{ij}(s)] \\ &= (\mathbf{I} - s \cdot P)^{-1} \\ &= (\mathbf{I} - s \cdot [p_{ij}])^{-1} \\ f_{ij}^{(k)} &= \mathbb{P}[\text{a } k\text{-step passage from } i \text{ to } j] \quad f_{ij}^{(0)} = 0 \forall i, j \\ F_{ij}(s) &= \sum_{n=0}^{\infty} f_{ij}^{(n)} \cdot s^n \end{aligned}$$

Don't forget the test comes with identities for converting Probability matrices to first return probabilities.

## 8 3041F: Discrete Time

### 8.1 Calculating the Count at a given Time

#### 8.1.1 Forecasting

$$\mathbb{P}[N_{n+k} = a + i | N_n = a] = \binom{k}{i} \cdot p^i \cdot q^{k-i} \quad (1)$$

$$(N_{n+k} - N_n) \sim \text{Binomial}(k, p) \quad (2)$$

$$(N_{n+k} - N_n) \perp\!\!\!\perp N_n \quad (3)$$

$$\mathbb{E}[N_{n+k} | N_n = a] = k \cdot p + a \quad (4)$$

$$\mathbb{V}[N_{n+k} | N_n = a] = k \cdot p \cdot a \quad (5)$$

#### 8.1.2 Backcasting

$$\mathbb{P}[N_n = a | N_{n+k} = a + i] = \frac{\binom{n}{a} \binom{k}{i}}{\binom{n+k}{a+i}} \quad (6)$$

$$\mathbb{P}[N_n = a | N_{n+k} = a + i] \sim \text{HyperGeo}(n + k, n, a + i) \quad (7)$$

$$\mathbb{E}[N_n | N_{n+k} = a + i] = \frac{(a + i) \cdot n}{n + k} \quad (8)$$

$$\mathbb{V}[N_n | N_{n+k} = a + i] = \frac{n + k - a - i}{n + k - 1} \cdot \frac{(a + i) \cdot n}{n + k} \cdot \left(1 - \frac{n}{n + k}\right) \quad (9)$$

### 8.2 Calculating the Time until a given Count

$$T_a \sim \text{NegBinomial}(a, p) \quad (10)$$

$$\mathbb{E}[T_a] = \frac{a}{p} \quad (11)$$

$$\mathbb{V}[T_a] = \frac{a \cdot q}{p^2} \quad (12)$$



### 8.2.1 Forecasting

$$\mathbb{P}[T_{a+i} = t + j | T_a = t] = \binom{j-1}{i-1} \cdot p^i \cdot q^{j-i} \quad (13)$$

$$\mathbb{P}[T_{a+i} = t + j | T_a = t] \sim \text{NegBinomial}(i, p) \quad (14)$$

$$\mathbb{P}[T_{a+i} | T_a = t] = \frac{i}{p} + t \quad (15)$$

$$\mathbb{P}[T_{a+i} | T_a = t] = \frac{i \cdot q}{p^2} \quad (16)$$

### 8.2.2 Backcasting

$$\mathbb{P}[T_a = t | T_{a+i} = t + j] = \frac{\binom{t-1}{a-1} \cdot \binom{j-1}{i-1}}{\binom{t+j-1}{a+i-1}} \quad (17)$$

$$\mathbb{E}[T_a | T_{a+i} = t + j] = \frac{a(t+j)}{a+i} \quad (18)$$

$$\mathbb{V}[T_a | T_{a+i} = t + j] = \frac{ai(t+j)(t+j-a-i)}{(a+i)^2(a+i+1)} \quad (19)$$

## 9 3041F: Continuous Time

With the assumptions that  $\lambda_i = \lambda$  and  $t_i = t$  for all  $i$ :

$$X_i \sim \text{Poisson}(\lambda t) \quad (20)$$

$$Y_n = \sum_{i=1}^n X_i \quad (21)$$

$$Y_n \sim \text{Poisson}(n\lambda t) \quad (22)$$

$$Y_{n+k} - Y_n \sim \text{Poisson}(k\lambda t) \quad (23)$$

$$\text{Cov}[Y_n, Y_m] = \lambda \cdot t \cdot \min(n, m) \quad (24)$$

$$\text{Corr}[Y_n, Y_m] = \frac{\min(n, m)}{\sqrt{nm}} \quad (25)$$

## 9.1 Calculating the Count at a given Time

### 9.1.1 Forecasting

$$\mathbb{P}[Y_{n+k} = a + i | Y_n = a] = \frac{e^{-k\lambda t} \cdot (k\lambda t)^i}{i!} \quad (26)$$

$$\mathbb{E}[Y_{n+k} | Y_n = a] = k\lambda t + a \quad (27)$$

$$\mathbb{V}[Y_{n+k} | Y_n = a] = k\lambda t \quad (28)$$

### 9.1.2 Backcasting

$$\mathbb{P}[Y_n = a | Y_{n+k} = a + i] = \binom{a+i}{a} \left(\frac{n}{n+k}\right)^a \left(\frac{k}{n+k}\right)^i \quad (29)$$

$$\mathbb{P}[Y_n = a | Y_{n+k} = a + i] \sim \text{Binomial}\left(a + i, \frac{n}{n+k}\right) \quad (30)$$

$$\mathbb{E}[Y_n | Y_{n+k} = a + i] = \frac{(a + i) \cdot n}{n + k} \quad (31)$$

$$\mathbb{V}[Y_n | Y_{n+k} = a + i] = \frac{(a + i) \cdot n \cdot k}{(n + k)^2} \quad (32)$$

## 9.2 Calculating the Time until a given Count

$$T_1 \sim \text{Exponential}\left(\frac{1}{\lambda}\right) \quad (33)$$

$$T_a \sim \text{Gamma}(a, \lambda) \quad (34)$$

$$\mathbb{E}[T_a] = \frac{a}{\lambda} \quad (35)$$

$$\mathbb{V}[T_a] = \frac{a}{\lambda^2} \quad (36)$$

### 9.2.1 Forecasting

$$\mathbb{E}[T_{a+i}|T_a = t] = \frac{i}{\lambda} + t \quad (37)$$

$$\mathbb{V}[T_{a+i}|T_a = t] = \frac{i}{\lambda^2} \quad (38)$$

### 9.2.2 Backcasting

$$\mathbb{E}[T_a|T_{a+i} = t + j] = \frac{a(t + j)}{a + i} \quad (39)$$

$$\mathbb{V}[T_a|T_{a+i} = t + j] = \frac{ai(t + j)^2}{(a + i)^2 \cdot (a + i + 1)} \quad (40)$$

## 10 3041F: Poisson Processes

### 10.1 3041F: Probability Integral Transform

If we want a random variable,  $X$ , that should have a certain CDF,  $F_X()$ , we can take a uniform random variable  $U \sim U(0, 1)$  and pass it through the inverse of  $F_X()$  in order to get the random variable  $X$ :

$$F_X^{-1}(U) \text{ has CDF } F_X()$$

From this, the exponential distribution can be simulated as

$$-\frac{1}{\lambda} \ln(U) \sim \text{Exp}(\lambda)$$

And the Gamma distribution with  $n$  iid uniform random variables as:

$$-\frac{1}{\lambda} \ln(U_1 \cdot U_2 \cdot \dots \cdot U_n) \sim \text{Gamma}(n, \lambda)$$

### 10.2 3041F: Estimation

See page 5 in the notes

## 11 3041F: Basic Poisson Process

### 11.1 3041F: Poisson Postulates

A counting process  $\{N(t), t \geq 0\}$  is a Poisson process with rate  $\lambda > 0$ , for small  $h$ , iff:

1.  $N(0) = 0$
2. The process has independent and stationary increments
3.  $\mathbb{P}[N(h) = 1] = \lambda h + o(h)$   
So the Probability of one occurrence within small timeframe  $h$  is equal to the rate parameter times  $h$ , plus some terms that go to zero.
4.  $\mathbb{P}[N(h) > 1] = o(h)$   
So the Probability of more than one occurrence within small timeframe  $h$  goes to zero.

### 11.2 Poisson Process Information

Proof page 7 in the notes.

The number of events in an interval of length  $t$  for a Poisson process with rate parameter  $\lambda$  is a random variable distributed  $Poi(\lambda t)$ .

More explicitly, for  $n$  a non-negative integer:

$$\mathbb{P}[N(t) = n] = \frac{e^{-\lambda t} (\lambda t)^n}{n!}$$

The interarrival times  $T_n, n = 1, 2, \dots$  are iid  $\sim Exp(\lambda)$

Define the sequence of waiting times  $S_n, n = 1, 2, \dots$  as:

$$\left( S_n := \sum_{i=1}^n T_i \right) \sim Gamma(n, \lambda)$$

The sequence of arrival times  $S_i, i = 1, 2, \dots, n$  have the same distribution as an ordered sample of size  $n$  taken from  $U(0, S_n)$

If we have a Poisson process  $\{N(t), t \geq 0\}$  where we classify each event as type 1 with probability  $p$  and type 2 with probability  $1-p$ , then  $\{N_1(t), t \geq 0\}$

and  $\{N_2(t), t \geq 0\}$  are independent Poisson processes with parameters  $\lambda p$  and  $\lambda(1 - p)$

Let  $\{N^{(1)}(t), t \geq 0\}$  and  $\{N^{(2)}(t), t \geq 0\}$  be independent Poisson Processes with rates  $\lambda_1$  and  $\lambda_2$ . Also define the arrival times  $S_n^{(1)}$  and  $S_m^{(2)}$  as the arrival of the  $n$ th,  $m$ th event.

Then:

$$\mathbb{P}[S_n^{(1)} < S_m^{(2)}] = \sum_{i=n}^{n+m-1} \binom{n+m-1}{i} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^i \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n+m-1-i}$$

### 11.3 3041F: Modelling Poisson Processes

There are two forms for modelling a Poisson process. Either you decide on how many events you want and let the events dictate the timeframe, or you decide on the timeframe you want and take note of the events in that timeframe.

**Generating n events** Calculate the events by sampling from a uniform distribution and then calculating  $T_i = -\frac{1}{\lambda} \ln(U_i(0, 1))$  for  $i = 1, \dots, n$ .

**Generating events in timeframe** Define the interarrival times as  $T_1, \dots, T_N$  where all interarrival times are in the interval  $[0, T]$

### 11.4 3041F: Estimation of Poisson Processes

Again, there are two scenarios. Either generate  $n$  events, or generate events within a timeframe  $[0, t]$ .

If the first  $n$  events of a process are observed as the interarrival times  $t_1, \dots, t_n$ . Then the MLE of  $\lambda$  is:

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^n t_i}$$

Alternatively, if there are  $N$  observed events in a given time interval  $[0, T]$ , then the MLE doesn't depend on the observed interarrival times and is:

$$\hat{\lambda}_{MLE} = \frac{N}{T}$$

Also, since  $N \sim Poi(\lambda T)$ ,  $\hat{\lambda}_{MLE}$  is an unbiased estimate of  $\lambda$  with variance  $\frac{\lambda}{T}$ .

## 12 3041F: Compound Poisson Processes

For regular poisson processes, every event is identical and increases the count by a certain amount. The events themselves aren't random, it's only the duration inbetween events which is random.

For Compound Poisson Processes, both the duration between events, and the amount by which our count will be changed is a random variable.

Let  $X_1, \dots, X_N$  be a sequence of iid random variables with PFG  $P_X(s)$  and  $N$  a non-negative integer valued random variable with PGF  $P_N(s)$ .

Then the count  $S_N$  and the PGF thereof are:

$$S_N = \sum_{i=1}^N X_i$$
$$P_{S_N}(s) = P_N(P_X(s))$$

The Expectation and Variance of  $S_N$  are given as a corollary:

$$\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X]$$
$$\mathbb{V}[S_N] = \mathbb{E}[N]\mathbb{V}[X] + (\mathbb{E}[X])^2 \mathbb{V}[N]$$

We formally define a compound Poisson process as a stochastic process  $\{X(t), t > 0\}$  that can be represented as the sum of iid random variables  $X_i$  which are also independant of the Poisson process:

$$X(t) = \sum_{i=1}^{N(t)} X_i$$

From this we can give the moments for a compound poisson process. With  $X(t)$  as the compound poisson process made up of the random variables  $X_i : i = 1, 2, \dots$ :

$$\mathbb{E}[X(t)] = \lambda t \mathbb{E}[X]$$
$$\mathbb{V}[X(t)] = \lambda t \mathbb{E}[X^2]$$

## 13 3041F: Non-homogeneous Poisson Processes

Regular Poisson Processes don't care what time it is. The rate parameter  $\lambda$  is constant, it doesn't change. This means that the frequency of events is constant. Non-homogeneous Poisson processes replace the constant rate

parameter  $\lambda$  with a function that depends on the time, called the intensity function  $\lambda(t)$ .

The non-homogeneous poisson postulates have to be rephrased slightly:

A counting process  $\{N(t), t \geq 0\}$  is a Poisson process with intensity function  $\lambda(t) > 0$  and  $t > 0$  for small  $h$ , iff:

1.  $N(0) = 0$
2. The process has independent increments (we drop the requirement for stationary increments)
3.  $\mathbb{P}[N(t+h) - N(t) = 1] = \lambda(t)h + o(h)$

So the Probability of one occurrence within small timeframe  $t$  to  $t + h$  is equal to the intensity (at time  $t$ ) multiplied by  $h$ , plus some terms that go to zero.

4.  $\mathbb{P}[N(t+h) - N(t) > 1] = o(h)$

So the Probability of more than one occurrence within small timeframe from  $t$  to  $t + h$  goes to zero.

### 13.1 3041F: Mean Value Function

Define the mean value function  $m(t)$  the integral from 0 to  $t$  of  $\lambda(t)$ :

$$m(t) = \int_0^t \lambda(u) du$$

And then the number of events in a time interval  $(t, t + s]$  for a non-homogeneous Poisson process is  $\sim Poi(m(t+s) - m(t))$ , with colours added to make it clear that  $m(t+s) - m(t)$  is just the rate parameter to a Poisson distribution.

In english, the number of events over a given interval is poisson distributed with a rate proportional to the total intensity over that interval. So the probability looks like:

$$\mathbb{P}[N(t+s) - N(t) = n] = \frac{e^{-(m(t+s)-m(t))} \cdot (m(t+s) - m(t))^n}{n!}$$

## 13.2 3041F: Simulation of a non-homogeneous Poisson Process

Suppose that events arrive according to a poisson process with rate  $\lambda$ , but we only consider some of the events, choosing events with probability  $p(t)$ .

This describes a non-homogeneous Poisson process with intensity function  $\lambda(t) = \lambda \cdot p(t)$ . We can therefore simulate a non-homogeneous Poisson process by first using our methods for homogeneous poisson processes, and then discarding events according to the probability function  $p(t)$ .

### 13.2.1 The Thinning algorithm

The thinning algorithm generates a non-homogeneous poisson process by finding an constant valued approximation of the rate function, generating a homogeneous poisson process according to this constant, and then discarding some events such that the approximation of the rate function becomes the actual rate function.

If we want to simulate over a time interval  $[0, T]$  according to intensity function  $\lambda(t) > 0$  first generate events according to a poisson process with rate  $\lambda$  over the interval  $[0, T]$  to give the  $N$  interarrival times  $T_1, \dots, T_N$ . Then select  $\lambda_{max}$  such that

$$\lambda_{max} \geq \lambda(t) \quad \forall t \in [0, T]$$

Finally, create an indicator function with  $\lambda_{max}$  to select only those events where a Uniform random variable in the range  $[0, \lambda_{max})$  is less than  $\lambda(t)$ :

$$U(0, \lambda_{max}) \leq \lambda(t) \quad \text{for } t = \sum_{i=1}^j T_i$$

Then all those selected events make up the required non-homogeneous poisson process.



### 13.2.2 Subintervals for simulations

The thinning algorithm will reject lots of events if the intensity function  $\lambda(t)$  is very different from  $\lambda_{max}$ . We can fix this by dividing up the time interval into multiple smaller subintervals, and then running the thinning algorithm individually on each of these sub-intervals. Each time calculating a new and different value for  $\lambda_{max}$ .

### 13.2.3 Combining homogeneous Poisson Processes

Since the sum of events for two independent Poisson Processes with rates  $\lambda_1$  and  $\lambda_2$  is itself a Poisson Process with rate  $\lambda_1 + \lambda_2$  we can use this to efficiently create piecewise Poisson processes where the intensity function  $\lambda(t)$  is sometimes constant.

## 13.3 3041F: Estimation of non-homogeneous Poisson Processes

The PDF of the  $i$ th interarrival time given the waiting time of the  $(i-1)$ th can be written as:

$$f_{T_i|W_{i-1}}(t) = \lambda(w_{i-1} + t) \cdot e^{-(m(w_{i-1}+t)-m(w_{i-1}))}$$

Note that the arrival times are independent due to the increments being independent.

Note that the probability of the  $i$ th interarrival time being greater than some value  $t$  (given the  $(i-1)$ th arrival time) is equivalent to the probability of there being zero events between the  $(i-1)$ th arrival time and the  $(i-1)$ th arrival time plus  $t$ . This implies:

$$\begin{aligned}\mathbb{P}[T_i > t | W_{i-1} = w_{i-1}] &= \mathbb{P}[N(w_{i-1} + t) - N(w_{i-1}) = 0] \\ &= e^{-(m(w_{i-1}+t)-m(w_{i-1}))}\end{aligned}$$

### 13.3.1 Likelihood for non-homogeneous Poisson Processes

In general, explicit expressions for the MLEs do not exist, and likelihood maximisation must be done numerically. The notes give examples, but they

all boil down to two different cases. The first case is when we have data about the first  $n$  arrivals, and can be calculated as so:

1. taking a given intensity function with unknown parameters  $\theta$  and observed waiting times  $w_1, \dots, w_n$ ,
2. Calculating the log-likelihood function as

$$l(\theta) = \sum_{i=1}^n \ln(\lambda(w_i; \theta)) - (m(w_n; \theta) - m(w_0; \theta))$$

3. and then maximising  $l$  with respect to  $\theta$  in order to find the MLEs for  $\theta$ .

The second case is where we've measured all arrivals in an interval  $[0, T]$ , and can be calculated as shown below. It is different, because we need to account for the probability that we saw no additional events between our last observed event and  $T$ .

1. taking a given intensity function with unknown parameters  $\theta$  and observed waiting times  $w_1, \dots, w_n$ . All waiting times should be within  $[0, T]$ .
2. Calculating the likelihood function as

$$L(\theta) = \prod_{i=1}^N \lambda(w_i; \theta) e^{-m(T; \theta)}$$

3. and then maximising  $L$  with respect to  $\theta$  will give the MLEs of  $\theta$ .

The notes also go into likelihood ratio tests for seeing if a particular Poisson process is non-homogeneous, but I'm not going to write that up here.

## 14 3041F: Time Series Analysis

This course will only look at measurements that are equally spaced in time. We will not be looking at unequal spacing (ie when you measure each event as it comes).

Time  
Series  
Video 1

A time series is denoted as  $X_t, t \in T$  where  $T$  can be discrete or continuous.

The index  $t$  will be discrete in this course.  $X_t$  could be continuous, discrete, or (not in this course) qualitative.

We will often plot the data, try to see if there are any patterns or trends, and then plot summary statistics which could help with forecasting.

There are several objectives of time series analysis:

- Description: Summarizing the data in useful ways like expectation, variance, etc.
- Modelling: Creating a mathematical model of the data that closely predicts what has happened.
- Forecasting: Using the created model to give confidence intervals and estimates for where the process will be at a specified point in the future. Forecasting is possible through historical data, and the existence of a pattern in that data which is expected to continue into the future.
- Control: Deciding if a given process is beyond certain safe operating bounds or if some extra factors need to be applied to the process in order to keep it within those safe operating bounds.

Time series are often modelled additively as a sequence of components:

- Trend  $\mu_t$ : The underlying expectation at time  $t$  of the time series. Often modelled as a simple function.
- Seasonality  $s_t$ : When there are repetitions in the data, with peaks and troughs.
- Cycles  $c_t$ : which is excluded in this course.
- Noise  $e_t$ : which represents the random fluctuations of the data.

Not every component will be included in every model. This can be modelled as

$$X_t = \mu_t + s_t + c_t + e_t$$

Time  
Series  
Video 2

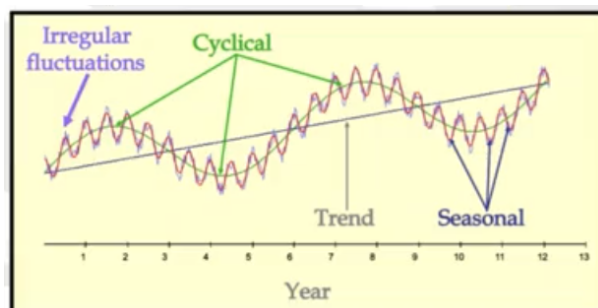


Figure 1: Comparison of Seasons vs Cycles

Or multiplicatively, although the multiplicative model will be log-transformed and then we'll just treat it as an additive model.

$$X_t = \mu_t \cdot s_t \cdot c_t \cdot e_t$$

$$\ln(X_t) = \ln(\mu_t) + \ln(s_t) + \ln(c_t) + \ln(e_t)$$

## 14.1 Trend

The long term growth/decline of a time series, modelled with a simple function. Basically just the expectation.

We can also look at the local trend, which just takes into account  $n$  prior points of the time series instead of all the data.

Some examples of Trend:

- $X_t = \epsilon_t$  with  $\epsilon_t \sim N(0, a)$  all iid. Since model's expectation  $\mu_t = \mathbb{E}[X_t] = 0$  doesn't depend on time, we can use the sample mean to estimate the population mean as all samples are from the same distribution.
- $X_t = t + \epsilon_t$  with  $\epsilon_t \sim N(0, a)$  all iid. This model's expectation  $\mu_t = \mathbb{E}[X_t] = t$  does depend on time, so the sample mean won't be a good estimate of the true mean.

## 14.2 Seasonality

When you have repeating trends in the data such as temperature increasing every summer or business flights decreasing over the weekend.

The amplitude of the seasonality is the difference between the peaks and troughs of the data. This amplitude might change over time. The amplitude will affect the variance of the overall, as a high amplitude will increase the variance.

### 14.3 Cycles

Basically seasonality on top of the existing seasonality. If seasonality repeats every year, then cycles repeat over decades.

Cycles are often used to refer to less regular oscillations in the data, such as bull and bear stock market trends which are less predictable than the annual trends produced by tax deadlines and such. We won't be covering the cyclical component in STA3041F.

### 14.4 Transformations

Because of the variability of the data, we might have to first transform the data such that our additive model is appropriate. Our additive model only works with data that changes linearly in time, so if the seasonality amplitude increases exponentially then we'll have to transform the data to get rid of that exponential growth. The family of transformations we'll be using is the Box-Cox transformations.

## 15 3041F: Transformations

In figure 2, you can see the seasonal component in the repeating peaks and troughs. There is an upwards trend in the data and no cyclical component

Time  
Series  
Video 3

There is an increasing amount of variance as time continues as shown by how the difference between the peaks and troughs. This implies that a multiplicative model might work well.

See figure 3. We can now subtract the long term trend (orange) from the root data (blue) in order to isolate just the error and seasonal components (assuming no cyclical component):

Sometimes you'll want to take the root data  $X_t$  and apply a transformation directly to it in an attempt to end up with linear data. Transformations might also be used to make the variance equal across the series. Possible transformations might be  $\sqrt{X_t}$ ,  $\sqrt[3]{X_t}$ ,  $\log(X_t)$ , or  $-X_t^{-1}$ .

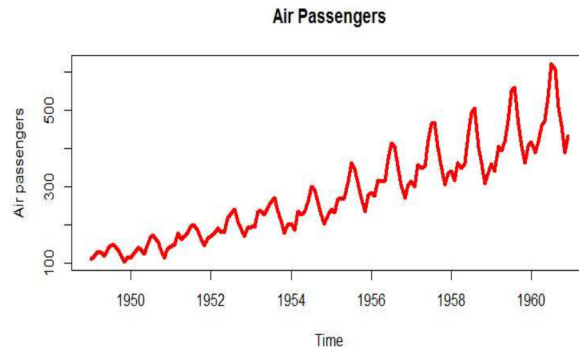


Figure 2: Airplane data

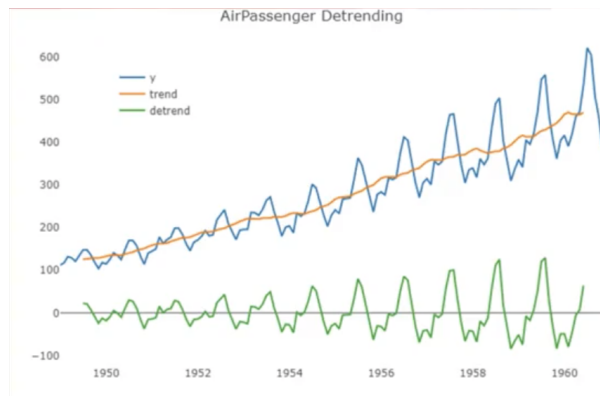


Figure 3: Airplane data with the trend subtracted.

However, the Box-Cox transformation can make our lives easier:

## 15.1 3041F: Box-Cox Transformation

If our variance increases over time, then we want to account for this in our model. However, linear methods don't really like dealing with a changing variance. The solution is to take the original data, transform it in some way to remove/reduce how much the variance changes over time, and then proceed to use linear models on the transformed data.

Box-Cox is a class of transformations with parameter  $\lambda$  that tries to change reduce the amount the variance changes over time.

The original series  $x_1, \dots, x_n$  is transformed into  $y_1(\lambda), \dots, y_n(\lambda)$  as follows:

$$y_t(\lambda) = \begin{cases} \frac{x_t^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x_t) & \text{if } \lambda = 0 \end{cases}$$

Where  $\lambda$  has to be estimated, and the resulting transformed data has a more stabilizer variance. The larger value of  $\lambda$ , the stronger the suppressing effect of the transformation. Note that taking the limit as  $\lambda$  trends to zero results in  $\log(x_t)$ , which is where the second case of the piecewise Box-Cox comes in.

Note that the Box-Cox tranformation is equivalent to many others, depending on the value of  $\lambda$ :

- When  $\lambda = -1$ : Inverse transformation
- When  $\lambda = 0$ : Natural Log transformation.
- When  $\lambda = \frac{1}{2}$ : (basically) Square root transformation.
- When  $\lambda = \frac{1}{3}$ : (basically) Cube root transformation.
- When  $\lambda = 1$ : (basically) No transformation.

Desmos  
Inter-  
active  
graph  
of the  
 $\lambda \neq 0$   
part of  
Box-  
Cox

## 15.2 3041F: Fitting Linear Filters

In time series analysis, we often need to remove the noise present in the data. Filtering the data is the processes of removing noise from the data, and there are different methods of filtering.

Time  
Series  
Video 4

The sample mean of the data is often used to estimate the trend of that data. Although this assumes that the trend does not change over time (ie it's not good for increasing/decreasing data). An example of the sample mean is shown with the black +’s in Figure 4.

A better way of estimating the trend is to divide the data into smaller blocks, and calculate the sample mean for each of those blocks. Then use these sample means to be the trend, although this still isn’t great. An example of the block mean is shown with the blue x’s in Figure 4.

An even better way is via linear filters, one of which is the weighted average, an example of which is shown with the purple dashed line in Figure 4 (labelled ‘filter’).

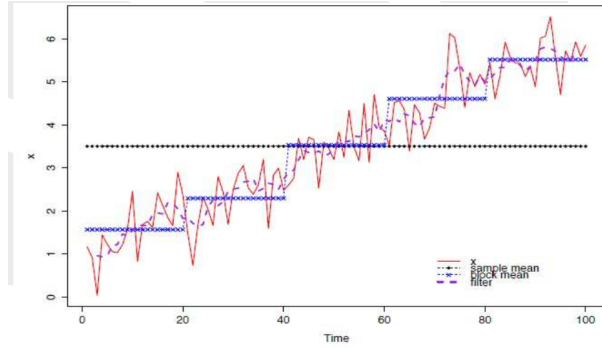


Figure 4: An example of moving average, block mean, and sample mean estimates of the trend of the data.

You can see that the weighted average stays the closest to the data, while still ignoring some of the jagged peaks. You can think of it as smoothing the data while remaining close to it.

Linear filters can also be used to estimate seasonal components by varying various parameters that the filters take. These parameters change how smooth the filter makes the data, or how much attention it pays to local changes compared to larger scale changes.

We can transform our original time series data  $y_1, \dots, y_n$  to a filtered version  $\tilde{y}_1, \dots, \tilde{y}_n$  by iterating over some of the items in  $y$  and multiplying them by a weight  $w$ .

We'll look at all the  $q$  items before the  $y_i$  we care about, as well as all the  $s$  items after that same  $y_i$ . Then we'll multiply them by a weight  $w_i$  and add them up:

$$\tilde{y}_t = \sum_{i=-q}^s w_i \cdot y_{t+i}$$

Note that the  $w_i$ 's all have to sum to 1 (so the sample mean stays the same), and usually we set  $s = q$  (so we look at the same number of items behind us as we do ahead of us).

We can also show that if  $\mathbb{E}[y_t] = \alpha + \beta t$  then the linear filter preserves this and  $\mathbb{E}[\tilde{y}_t] = \alpha + \beta t$ .

The linear filter won't always preserve a quadratic trend unless we get funky with how we choose our weights. For example, if we have an equation for  $y_t$  in the form:



$$y_t = \alpha + \beta t + \phi t^2 + e_t$$

Then the expectation of our linear filter version of  $y_t$  is given by

$$\mathbb{E}[\tilde{y}_t] = \alpha + \beta t + \phi t^2 + \phi \sum_i w_i i^2$$

So in order to preserve a quadratic trend, we'd need to choose the weights  $w_i$  such that the final term is equal to zero, ie:

$$0 = \sum_i w_i i^2$$

Estimating the weights of the linear filter can be done as follows:

1. Decide on how many weights you'll use. This is called the window size
2. Decide if your trend equation is linear, quadratic, cubic, etc.
3. Estimate the coefficients of the trend equation  $y_t = \alpha + \beta t + \dots$  using least squares.
4. We can center the window such that the current  $y_t$  is at zero, which simplifies the calculations a bit.

**Time Series Video 5** This video was a lot of math walkthroughs, which I'm not going to bother writing down.

Time  
Series  
Video 5  
Time  
Series  
Video 6

## 15.3 3041F: Moving Averages

A moving average is a linear combination of current and past error terms. So we can define the moving average at time  $t$  as  $\bar{x}_t = e_{t-n} + \dots + e_t$ , where we call  $n$  the order of the moving average, often written as MA( $n$ ).

There are certain useful properties, but in order for them to hold we need the time series to not have a seasonal component and be represented by

$$y_t = \mu_t + e_t \quad \text{where } \mu_t \text{ is the mean at time } t$$

and the error terms  $e_t \sim N(0, \sigma^2)$  are all independent. Also, define the smoothed trend at time  $t$ ,  $\hat{\mu}_t$  as:

$$\hat{\mu}_t = \sum_{i=-s}^s w_i \cdot y_{t+i}$$

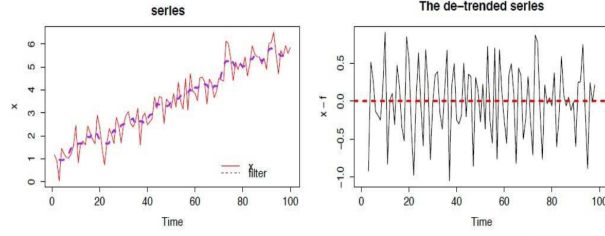


Figure 5: The detrended series has an expectation of zero if the estimate of the trend is unbiased.

Then the following properties hold:

$$\begin{aligned}\mathbb{V}[\hat{\mu}_t] &= \sigma^2 \sum_{i=-s}^s w_i^2 \\ \text{Cov}[\hat{\mu}_t, \hat{\mu}_{t+k}] &= \sigma^2 \sum_{i=-s}^{s-k} w_i \cdot w_{i+k} \\ \text{Corr}[\hat{\mu}_t, \hat{\mu}_{t+k}] &= \frac{\text{Cov}[\hat{\mu}_t, \hat{\mu}_{t+k}]}{\mathbb{V}[\hat{\mu}_t, \hat{\mu}_{t+k}]} \\ &= \frac{\sum_{i=-s}^{s-k} w_i \cdot w_{i+k}}{\sum_{i=-s}^s w_i^2}\end{aligned}$$

It can also be proven that the variance of the moving average is less than the variance of the original data.

If we have a time series described by  $y_t = \mu_t + e_t$  and the estimate for the trend is given by  $\tilde{y}_t$ , then we define the de-trended series as the difference  $y_t - \tilde{y}_t$ . If  $\tilde{y}_t$  is an unbiased estimate of the trend then  $\mathbb{E}[y_t - \tilde{y}_t] = 0$ .

An example of the original series and the de-trended series is shown in figure 5

Note that the seasonal component of a series can be removed using a moving average with a window size equal to the period of the season. This can be seen by looking at the MA(12) line in figure 6

Some comments about Moving Averages

- They smoothen out a time series (ie they reduce variation)
- They estimate the trend of a time series

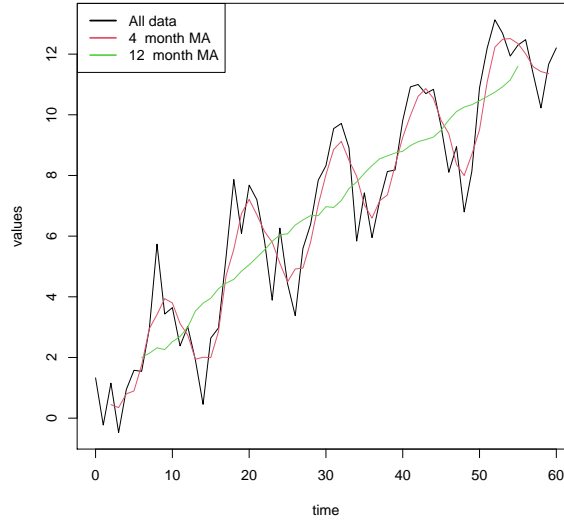


Figure 6: The seasonality can be removed by applying a moving average which has a window size of the period of the data.

- The transformed series will be correlated with itself, as  $t_i$ th point in the moving average is a linear combination of the  $t_{i-s}, \dots, t_{i+s}$  points in the original data.
- The above point means that even uncorrelated, completely random data will end up with correlations.
- We are unable to estimate the end points, as there are no data points in our window
- 

## 16 Terms and Definitions

**Absorbing** Once you're in an absorbing state, it's impossible to leave.

**Accessible** State  $j$  is accessible from state  $i$  if there exists  $n$  greater than 0 such that the probability of going from  $j$  to  $i$  in  $n$  steps is greater than zero.

**Autocovariance** Covariance of a given random variable with itself, but at different points in time.

**Closed Sets** A subset of the set of all possible states, which entered cannot be exited.

**Communicate** States  $i$  and  $j$  communicate if they are both accessible from each other. By convention, every state communicates with itself.

**Cross-sectional Data** All features are collected at a single period in time. All measurements are independent

**Decomposition Theorem** The state space can be partitioned uniquely into one set of transient states (although this need not be an equivalence class), and several closed irreducible sets of recurrent states.

**Equivalence Class** A set of states in which all states communicate.

**Ergodic** A class that's both aperiodic and positively persistent.

**Interarrival Time** The duration between event  $i$  and  $i+1$ , or the sequence of all such events.

**Irreducible Chain** When the state space of the chain is an irreducible class.

**Irreducible Class** A set that's both closed and an equivalence class

**Limiting Distribution** If a Markov Chain has a limiting distribution, then after enough time steps the chain will end up at this distribution regardless of starting state.

**Periodic** A state  $i$  is periodic with period  $k$  if the probability of first passage back to  $i$  in some number of steps  $n$  is equal to zero for all  $n \% k \neq 0$ .

**Persistent, null** A recurrent state in a chain with infinite total states and therefore the mean recurrence time infinite.

**Persistent, positively** A recurrent state in a chain with finite total states and therefore the mean recurrence time is less than infinity.

**Recurrent State** Starting at a recurrent state  $i$ , and given infinite time, you will return to state  $i$  with probability 1.

**Recurrent** An equivalence class with 1 or more recurrent states.

**Regular Chain** Given enough steps, you can visit every state, regardless of which state you start out at. This implies  $W^n$  only has positive entries.

**Reversibility** A stochastic matrix is reversible with respect to a given distribution if stepping forwards or backwards in time has no effect on the distribution.

**Stationary Distribution** When the marginal distribution doesn't change over time.

**Time Series Data** Features are measured at multiple points in time. Measurements in the future are often dependant on the measurements in the past.

**Time Series** A particular realisation of a stochastic process.

**Transient State** Starting at transient state  $i$ , the average time until first passage back to  $i$  is infinite, and we say the passage is uncertain.

## 17 R trickery

Useful functions:

### 17.1 runif

`runif(n)`: list of 'n' Random uniform distribution variables

## 17.2 cumsum

`cumsum(sequence)`: cumulative sum of the given ‘sequence’

## 17.3 plot

`plot(x y)`: Plot a plot of x vs y

- `type=""`, type of plot should be drawn. Possible types are
  - `"p"` for `*p*`oints,
  - `"l"` for `*l*`ines,
  - `"b"` for `*b*`oth,
  - `"c"` for the lines part alone of `"b"`,
  - `"o"` for both `*o*`verplotted',
  - `"h"` for `*h*`istogram' like (or 'high-density') vertical lines,
  - `"s"` for stair `*s*`teps,
  - `"S"` for other `*s*`teps, see 'Details' below,
  - `"n"` for no plotting.
- `'main'` an overall title for the plot: see 'title'.
- `'sub'` a sub title for the plot: see 'title'.
- `'xlab'` a title for the x axis: see 'title'.
- `'ylab'` a title for the y axis: see 'title'.
- `'asp'` the y/x aspect ratio, see 'plot.window'.

## 17.4 ts

`ts(data = NA, start = 1, end = numeric(), frequency = 1, deltat = 1, ts.eps = getOption("ts.eps"), class = , names = )`

The function `'ts'` is used to create time-series objects. `'as.ts'` and `'is.ts'` coerce an object to a time-series and test whether an object is a time series.

## 18 Notes

These should eventually all be categorised, but for now they're all lumped together here.

An irreducible chain implies it has a unique stationary distribution.

Trace of a matrix is equal to the sum of it's eigenvalues

A Markov Matrix always has 1 as an eigenvalue

Diagonalization:  $D = PAP^{-1}$