Maths Stats Formulas and Definitions

Boyd Kane

Last updated: May 28, 2021

Contents

Chapter 1

Distributions, tricks, and required knowledge

1.1 Common Distributions

1.1.1 Poisson Distribution

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

1.1.2 Negative Binomial Distribution

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

1.1.3 Geometric Distribution

$$X \sim Geo(p) \qquad f_X(x) = pq^{x-1}$$
$$\mathbb{E}[X] = \frac{1}{p} \qquad \mathbb{E}[e^{sX}] = \frac{e^t p}{1 - e^t q}$$
$$\mathbb{V}[X] = \frac{q}{p^2}$$

1.1.4 Binomial Distribution

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

See figure 1.1.

1.1.5 Beta Distribution

$0 < x < 1$

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

See figure 1.2

1.1.6 Normal Distribution

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

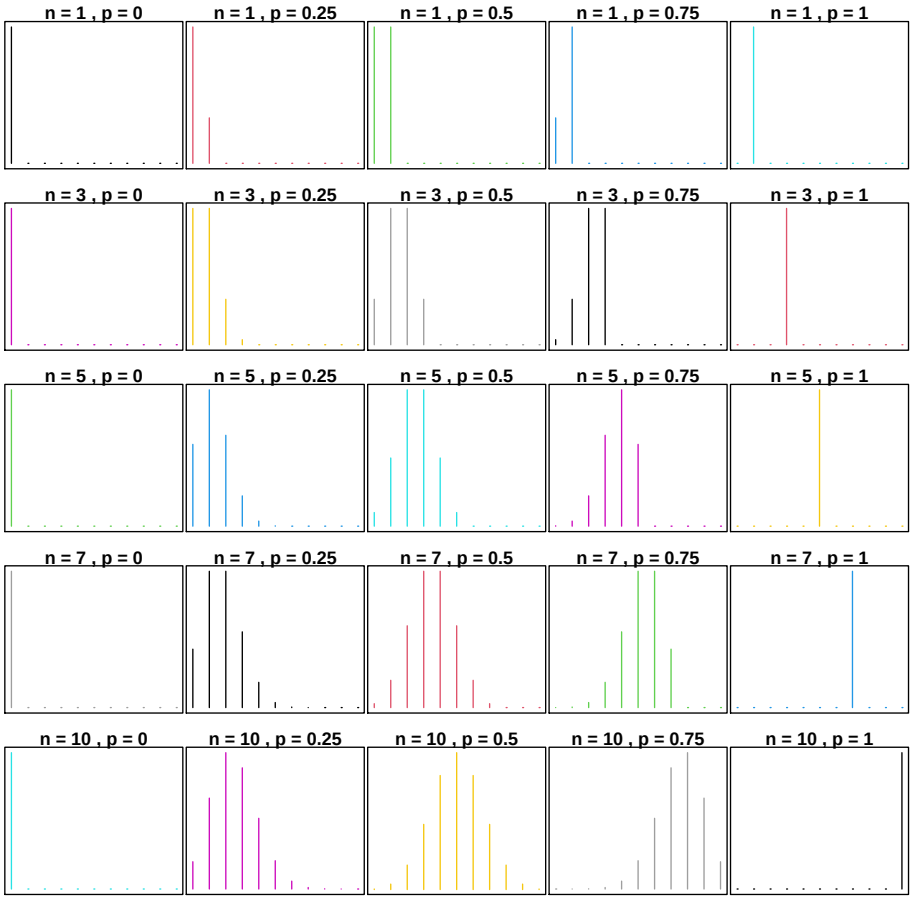

Figure 1.1: The binomial distribution with parameters n and p

1.1.7 Gamma Distribution

$\lambda > 0, \alpha > 0$

$$X \sim Gamma(\lambda, \alpha) \qquad f_X(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{(\alpha-1)!}$$

$$\mathbb{E}[X] = \frac{\alpha}{\lambda} \qquad \mathbb{E}[e^{sX}] = \left(\frac{\lambda}{\lambda - s}\right)^\alpha$$

$$\mathbb{V}[X] = \frac{\alpha}{\lambda^2}$$

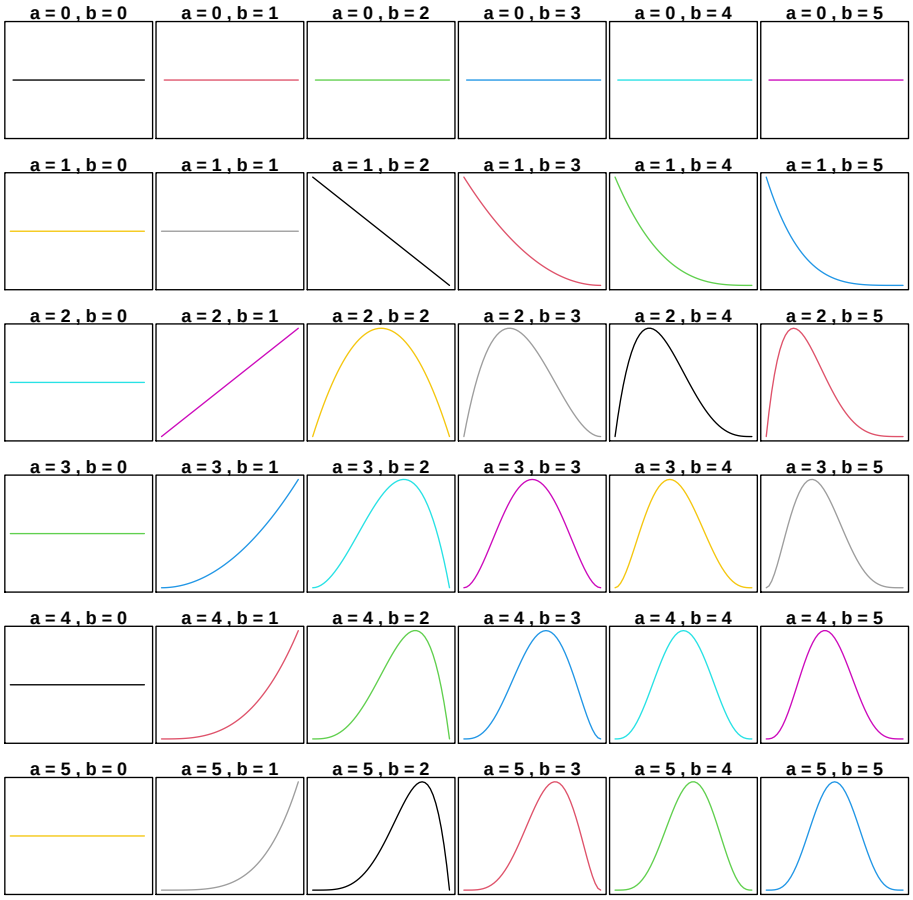

Figure 1.2: The Beta distribution with parameters a and b.

Notes

Sum of independant Exponentials is Gamma distributed:

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

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

1.1.8 Exponential Distribution

$$X \sim Exp(\lambda) \qquad f_X(x) = \lambda e^{-\lambda x} \quad x \geq 0, \lambda > 0$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad F_X(x) = 1 - e^{-\lambda x}$$

$$\mathbb{V}[X] = \frac{1}{\lambda^2} \qquad \mathbb{E}[e^{sX}] = \frac{\lambda}{\lambda - s} \quad s < \lambda$$

$$\hat{\lambda}_{MLE} = \frac{1}{\bar{x}} \qquad \mathbb{P}[Exp(\lambda_1) < Exp(\lambda_2)] = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

1.1.9 Student's t-distribution

Work in progress.

1.1.10 Chi-Squared Distribution

The χ^2_m Distribution is defined as having m degrees of freedom, and is usually defined in terms of a sum of standard normal distributions. ie if

$Z_i \sim N(0,1) \quad \forall i \in 1 \dots k$ and all Z_i independent. then:

$$\begin{aligned}
\sum_{n=1}^{k} (Z_i)^2 &\sim \chi^2_{k-1} \\
\mathbb{E}[\chi^2_k] &= k \\
\mathbb{V}[\chi^2_k] &= 2k \\
\mathbb{M}_{\chi^2_k}(t) &= (1-2t)^{-\frac{k}{2}} \quad \text{for} t < \frac{1}{2} \\
\chi^2_2 &\equiv Exp(\frac{1}{2}) \\
-2\ln U(0,1) &\equiv \chi^2_2 \\
&\equiv Exp(\frac{1}{2})
\end{aligned}$$

And if $X_1 \sim \chi^2_a, X_2 \sim \chi^2_b, X_1 \perp\!\!\!\perp X_2$

$$\begin{aligned}
\frac{X_1}{X_1 + X_2} &\equiv Beta(\frac{a}{2}, \frac{b}{2}) \\
\frac{\frac{\chi^2_a}{a}}{\frac{\chi^2_b}{b}} &\equiv F_{a,b}
\end{aligned}$$

and if X_i are all independent with degrees of freedom f_i

$$\sum_{i=1}^{k} \chi^2_{f_i} \equiv \chi^2_{\sum_{i=1}^{k} f_i}$$

1.2 Generating Functions

1.2.1 Probability Generating Functions

If X is a non-negative, integer-valued, and discrete, then the probability generating function of X is given by:

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

1.2.2 Moment Generating Functions

The Moment Generating Function of X is defined by:

$$\begin{aligned}
M_X(t) : &= \mathbb{E}[e^{tX}] \\
&= \int_{-\infty}^{\infty} e^{tx} \cdot f_X(x)\, dx
\end{aligned}$$

But the MGF might not always exists. We can also define its complex-numbered version, the Characteristic Function, which will always exist:

$$\begin{aligned}
\varphi_(t) : &= \mathbb{E}[e^{itX}] \\
&= \int_{-\infty}^{\infty} e^{itx} \cdot f_X(x)\, dx
\end{aligned}$$

The r-th moment of X can be retrieved by taking the r-th derivative of the MGF and evaluating it at zero. A similar process also works for the

characteristic function:

$$\begin{aligned} r\text{-th Moment of X} &= \mathbb{E}[X^r] \\ &= \frac{d^r}{dX^r} M_X(0) \\ &= (-i)^r \frac{d^r}{dX^r} \varphi_X(0) \end{aligned}$$

1.3 Probability

The Law of Total Probability is given as:

$$\mathbb{P}[A] = \sum_{n=1}^{\infty} \mathbb{P}[A|B_n]\mathbb{P}[B_n]$$

Bayes Theorem:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

1.4 Expectation and Variance

Expectation is always taken with respect to a random variable (a bit like you always differentiate with respect to a certain variable in an equation).

If the random variable is not obvious, then the Expectation of some function g(X) with respect to X (which has probability density function given by $f_X(x)$) is written like $\mathbb{E}_X[g(X)]$ and is defined as:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx$$

So for example, if $g(x) = x^2$ then we can calculate the Expectation of X^2 as:

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\, dx$$

And in the special case where X is non-negative and absolutely continuous, we have the identity:

$$\mathbb{E}[X] = \int_{0}^{\infty} \mathbb{P}[\{X > x\}]\, dx$$

Note that expectation is a linear transformation, such that the following is true (Assuming that both expectations are finite):

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Variance can be calculated as:

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X - \mu_X)^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

1.5 Random Vectors, Covariance, Correlation

So as to not get lost in a sea of brackets, define the following:

$$\begin{aligned}\mu_X &= \mathbb{E}[X] \\ \mu_Y &= \mathbb{E}[Y] \\ \sigma_X^2 &= \mathbb{V}[X] \\ \sigma_Y^2 &= \mathbb{V}[Y] \\ \sigma_{XY} &= \mathbb{C}ov[X, Y] \\ \rho_{XY} &= \mathbb{C}orr[X, Y]\end{aligned}$$

With the prior definitions, the following are useful identities, definitions, or shortcuts relating to Covariance and Correlation:

$$\begin{aligned}\mathbb{C}ov[X, Y] &= \mathbb{E}[(X - \mu_x) \cdot (Y - \mu_Y)] \\ &= \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] \\ \mathbb{C}ov[X, a] &= 0 \quad \text{a } \in \mathbb{R} \\ \mathbb{C}ov[X, Y] &= \mathbb{C}ov[Y, X] \\ \mathbb{C}orr[X, Y] &= \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \\ &= \frac{\mathbb{C}orr[X, Y]}{\sqrt{\mathbb{V}[X] \cdot \mathbb{V}[Y]}} \\ \mathbb{C}ov[X_1, X_2] &= 0 \qquad \text{if } X_1 \text{ is independant of } X_2\end{aligned}$$

1.6 MAM1000W things

$$\begin{aligned} e^x &= \sum_{n=0}^{\infty} \frac{x^n}{n!} \\ &= 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots \\ &= \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n \\ e^{i\theta} &= \cos\theta + i\sin\theta \end{aligned}$$

1.7 Conditioning

The definition of a conditional distribution is given by:

$$f_{X|Y}(x|y) = \frac{f_{xy}(x,y)}{f_Y(y)}$$

And conditional probability by:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Conditional Expectation identities and definitions:

$$\begin{aligned} \mathbb{E}[X|Y=y] &= \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\, dx \\ \mathbb{E}[g(X)|X=x] &= g(x) \\ \mathbb{E}[c|X] &= c \\ \mathbb{E}[aX+bY|Z] &= a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z] \\ \mathbb{E}[g(X,Y)|Y=y] &= \mathbb{E}[g(X,y)|Y=y] \\ \underbrace{\mathbb{E}[g(X_1,X_2)|X_2=x_2]}_{\text{If } X_1 \perp\!\!\!\perp X_2} &= \mathbb{E}[g(X_1,x_2)|X_2=x_2] \end{aligned}$$

Expectation can be re-written as:

$$\begin{aligned} \mathbb{E}_X[X] &= \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]] \\ &= \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y] \cdot f_Y(y)\, dy \end{aligned}$$

Variance can be re-written as:

$$\mathbb{V}_X[X] = \mathbb{E}_Y[\mathbb{V}_{X|Y}[X|Y]] + \mathbb{V}_Y[\mathbb{E}_{X|Y}[X|Y]]$$

1.8 Sum Trickery

$$\sum_{i=n}^{\infty} g(x)^i = \frac{g(x)^n}{1-g(x)} \qquad \text{if} |g(x)| < 1$$

And, if $0 < r < 1$:

$$\sum_{i=1}^{\infty} ar^i = \frac{a}{1-r}$$

$$a\sum_{i=0}^{\infty} ir^i = \frac{ar}{(1-r)^2}$$

$$\sum_{i=0}^{n} i = \frac{1}{2}n(n+1)$$

$$\sum_{i=0}^{n} i^3 = \left(\frac{1}{2}n(n+1)\right)^2$$

Chapter 2

Chains, Processes, and Forecasting

2.1 3041F: Chains

2.1.1 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

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

2.2 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 $\sigma^2 = \mathbb{V}[X]$.

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

2.3 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:

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

2.4 3041F: First Passage Probabilities

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

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

2.5 3041F: Discrete Time

2.5.1 Calculating the Count at a given Time

Forecasting

$$\mathbb{P}[N_{n+k} = a + i | N_n = a] = \binom{k}{i} \cdot p^i \cdot q^{k-i} \tag{2.1}$$

$$(N_{n+k} - N_n) \sim Binomial(k, p) \tag{2.2}$$

$$(N_{n+k} - N_n) \perp\!\!\!\perp N_n \tag{2.3}$$

$$\mathbb{E}[N_{n+k} | N_n = a] = k \cdot p + a \tag{2.4}$$

$$\mathbb{V}[N_{n+k} | N_n = a] = k \cdot p \cdot a \tag{2.5}$$

Backcasting

$$\mathbb{P}[N_n = a | N_{n+k} = a + i] = \frac{\binom{n}{a}\binom{k}{i}}{\binom{n+k}{a+i}} \tag{2.6}$$

$$\mathbb{P}[N_n = a | N_{n+k} = a + i] \sim HyperGeo(n + k, n, a + i) \tag{2.7}$$

$$\mathbb{E}[N_n | N_{n+k} = a + i] = \frac{(a + i) \cdot n}{n + k} \tag{2.8}$$

$$\mathbb{V}[N_n | N_{n+k} = a + i] = \frac{n + k - a - i}{n + k - 1} \cdot \frac{(a + i) \cdot n}{n + k} \cdot (1 - \frac{n}{n + k}) \tag{2.9}$$

2.5.2 Calculating the Time until a given Count

$$T_a \sim NegBinomial(a, p) \tag{2.10}$$

$$\mathbb{E}[T_a] = \frac{a}{p} \tag{2.11}$$

$$\mathbb{V}[T_a] = \frac{a \cdot q}{p^2} \tag{2.12}$$

Forecasting

$$\mathbb{P}[T_{a+i} = t + j | T_a = t] = \binom{j - 1}{i - 1} \cdot p^i \cdot q^{j-i} \tag{2.13}$$

$$\mathbb{P}[T_{a+i} = t + j | T_a = t] \sim NegBinomial(i, p) \tag{2.14}$$

$$\mathbb{P}[T_{a+i} | T_a = t] = \frac{i}{p} + t \tag{2.15}$$

$$\mathbb{P}[T_{a+i} | T_a = t] = \frac{i \cdot q}{p^2} \tag{2.16}$$

Backcasting

$$\mathbb{P}[T_a = t | T_{a+i} = t + j] = \frac{\binom{t-1}{a-1} \cdot \binom{j-1}{i-1}}{\binom{t+j-1}{a+i-1}} \tag{2.17}$$

$$\mathbb{E}[T_a | T_{a+i} = t + j] = \frac{a(t+j)}{a+i} \tag{2.18}$$

$$\mathbb{V}[T_a | T_{a+i} = t + j] = \frac{ai(t+j)(t+j-a-i)}{(a+i)^2(a+i+1)} \tag{2.19}$$

2.6 3041F: Continuous Time

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

$$X_i \sim Poisson(\lambda t) \tag{2.20}$$

$$Y_n = \sum_{i=1}^{n} X_i \tag{2.21}$$

$$Y_n \sim Poisson(n\lambda t) \tag{2.22}$$

$$Y_{n+k} - Y_n \sim Poisson(k\lambda t) \tag{2.23}$$

$$\mathbb{C}ov[Y_n, Y_m] = \lambda \cdot t \cdot min(n, m) \tag{2.24}$$

$$\mathbb{C}orr[Y_n, Y_m] = \frac{min(n, m)}{\sqrt{nm}} \tag{2.25}$$

2.6.1 Calculating the Count at a given Time

Forecasting

$$\mathbb{P}[Y_{n+k} = a + i | Y_n = a] = \frac{e^{-k\lambda t} \cdot (k\lambda t)^i}{i!} \tag{2.26}$$

$$\mathbb{E}[Y_{n+k} | Y_n = a] = k\lambda t + a \tag{2.27}$$

$$\mathbb{V}[Y_{n+k} | Y_n = a] = k\lambda t \tag{2.28}$$

Backcasting

$$\mathbb{P}[Y_n = a | Y_{n+k} = a+i] = \binom{a+i}{a} \left(\frac{n}{n+k}\right)^a \left(\frac{k}{n+k}\right)^i \tag{2.29}$$

$$\mathbb{P}[Y_n = a | Y_{n+k} = a+i] \sim Binomial\left(a+i, \frac{n}{n+k}\right) \tag{2.30}$$

$$\mathbb{E}[Y_n | Y_{n+k} = a+i] = \frac{(a+i)\cdot n}{n+k} \tag{2.31}$$

$$\mathbb{V}[Y_n | Y_{n+k} = a+i] = \frac{(a+i)\cdot n \cdot k}{(n+k)^2} \tag{2.32}$$

2.6.2 Calculating the Time until a given Count

$$T_1 \sim Exponential\left(\frac{1}{\lambda}\right) \tag{2.33}$$

$$T_a \sim Gamma(a, \lambda) \tag{2.34}$$

$$\mathbb{E}[T_a] = \frac{a}{\lambda} \tag{2.35}$$

$$\mathbb{V}[T_a] = \frac{a}{\lambda^2} \tag{2.36}$$

Forecasting

$$\mathbb{E}[T_{a+i} | T_a = t] = \frac{i}{\lambda} + t \tag{2.37}$$

$$\mathbb{V}[T_{a+i} | T_a = t] = \frac{i}{\lambda^2} \tag{2.38}$$

Backcasting

$$\mathbb{E}[T_a | T_{a+i} = t+j] = \frac{a(t+j)}{a+i} \tag{2.39}$$

$$\mathbb{V}[T_a | T_{a+i} = t+j] = \frac{ai(t+j)^2}{(a+i)^2 \cdot (a+i+1)} \tag{2.40}$$

Chapter 3

Poisson Processes

3.1 3041F: Poisson Processes

3.1.1 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

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

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

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

3.1.2 3041F: Estimation

See page 5 in the notes

3.2 3041F: Basic Poisson Process

3.2.1 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 independant and stationary increments
3. $\mathbb{P}[N(h) = 1] = \lambda h + o(h)$

   So the Probability of one occurance 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 occurance within small timeframe h goes to zero.

3.2.2 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 λ 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 $T_n, n = 1, 2, \ldots$ are iid $\sim Exp(\lambda)$

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

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

The sequence of arrival times $S_i, i = 1, 2, \ldots, n$ 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 independant Poisson processes with parameters λp and $\lambda(1-p)$

Let $\{N^{(1)}(t), t \geq 0\}$ and $\{N^{(2)}(t), t \geq 0\}$ be independant Poisson Processes with rates λ_1 and λ_2. Also define the arrival times $S_n^{(1)}$ and $S_m^{(2)}$ as the arrival of the nth, mth event.

Then:

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

3.2.3 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 $i = 1, \ldots, n$.

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

3.2.4 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 $t_i, \ldots, t_n$. Then the MLE of λ 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 λ with variance $\frac{\lambda}{T}$.

3.3 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 $X_i : i = 1, 2, \dots$:

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

3.4 3041F: Non-homogeneous Poisson Processes

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

parameter λ 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 independant 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 occurance 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.

3.4.1 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:

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

3.4.2 3041F: Simulation of a non-homogeneous Poisson Process

Suppose that events arrive according to a poisson process with rate λ, 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).

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 λ over the interval [0, T] to give the N interarrival times $T_1, \ldots, T_N$. Then select λ_{max} such that

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

Finally, create an indicator function with λ_{max} to select only those events where a Uniform random variable in the range [0, λ_{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.

Subintervals for simulations

The thinning algorithm will regect lots of events if the intensity function $\lambda(t)$ is very different from λ_{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 λ_{max}.

Combining homogeneous Poisson Processes

Since the sum of events for two independant Poisson Processes with rates λ_1 and λ_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.

3.4.3 3041F: Estimation of non-homogeneous Poisson Processess

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

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

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

Note that the probability of the ith 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:

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

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 θ and observed waiting times $w_1, \ldots, w_n$,

2. Calculating the log-likelihood function as

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

3. and then maximising l with respect to θ in order to find the MLEs for θ.

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 θ and observed waiting times $w_1, \ldots, w_n$. 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 θ will give the MLEs of θ.

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.

Chapter 4

Time Series I and Transformations

4.1 3041F: Time Series Analysis

Time Series Video 1

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).

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.

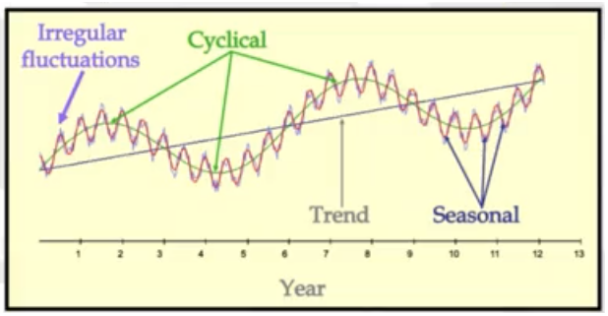

Figure 4.1: Comparison of Seasons vs Cycles

Forecasting is possible through historical data, and the existance 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 Video 2

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

- Trend μ_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

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

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

$$\begin{aligned} X_t &= \mu_t \cdot s_t \cdot c_t \cdot e_t \\ ln(X_t) &= ln(\mu_t) + ln(s_t) + ln(c_t) + ln(e_t) \end{aligned}$$

4.1.1 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 wont' be a good estimate of the true mean.

4.1.2 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.

4.1.3 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.

4.1.4 Transformations

Because of the variablity of the data, we might have to first transform the data such that our additive model is appropriate. Our additive model only

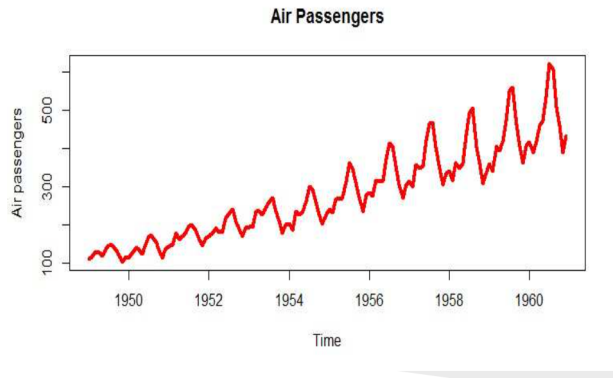

Figure 4.2: Airplane data

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.

4.2 3041F: Transformations

Time Series Video 3

In figure 4.2, you can see the seasonal component in the repeating peaks and troughs. There is an upwards trend in the data and no cyclical component

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 4.3. 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 direcly 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 $-X_t^{-1}$.

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

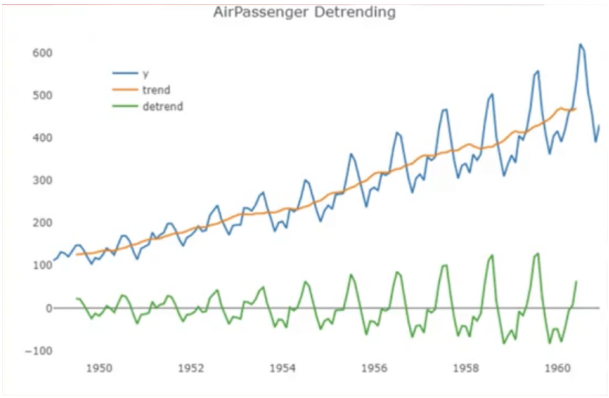

Figure 4.3: Airplane data with the trend subtracted.

4.2.1 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 λ that tries to change reduce the amount the variance changes over time.

The original series $x_1, \ldots, x_n$ is transformed into $y_1(\lambda), \ldots, y_n(\lambda)$ 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 λ has to be estimated, and the resulting transformed data has a more stabilizer variance. The larger value of λ, the stronger the suppressing effect of the transformation. Note that taking the limit as λ trends to zero results in $log(x_t)$, which is where the second case of the piecewise Box-Cox comes in.

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

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

- When $\lambda = -1$: Inverse tranformation

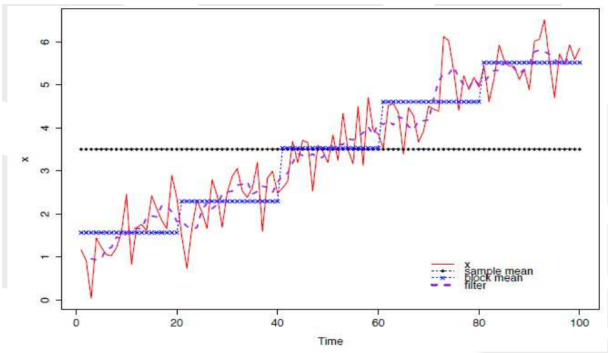

Figure 4.4: An example of moving average, block mean, and sample mean estimates of the trend of the data.

- 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.

4.2.2 3041F: Fitting Linear Filters

Time Series Video 4

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.

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 4.4.

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 4.4.

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 4.4 (labelled 'filter').

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 $y_1, \ldots, y_n$ to a filtered version $\tilde{y_1}, \ldots, \tilde{y_n}$ 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 you're 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 walk-through, which I'm not going to bother writing down.

Time Series Video 5

Time Series Video 6

4.2.3 3041F: Moving Averages

A moving averages is a linear combination of current and past error terms. So we can define the moving average at time t a$x_t = e_{t-n} + \dots + e_t$, 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

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

and the error terms $e_t \sim N(0, \sigma^2)$ are all independent. Also, define the smoothened trend at time t, μ_t as:

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

Then the following properties hold:

$$\mathbb{V}[\hat{\mu}_t] = \sigma^2 \sum_{i=-s}^{s} w_i^2$$

$$\mathbb{C}ov[\hat{\mu}_t, \hat{\mu}_{t+k}] = \sigma^2 \sum_{i=-s}^{s-k} w_i \cdot w_{i+k}$$

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

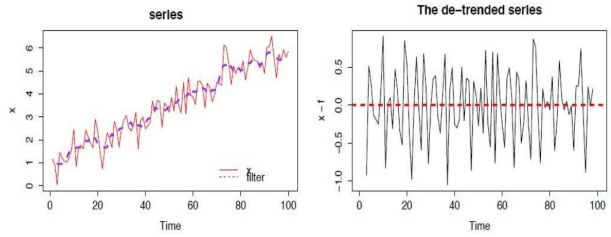

Figure 4.5: The detrended series has an expectation of zero if the estimate of the trend is unbiased.

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 4.5

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 4.6

Some comments about Moving Averages

- They smoothen out a time series (ie they reduce variation)
- They estimate the trend of a time series
- The transformed series will be correlated with itself, as t_ith point in the moving average is a linear combination of the $t_{i-s}, \ldots t_{i+s}$ 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

4.2.4 Exponential Moving Averages

Time Series Video 7

Exponential smoothing is another type of filter. We're effectively taking the average but giving a higher weighting to the most recent points. This filter

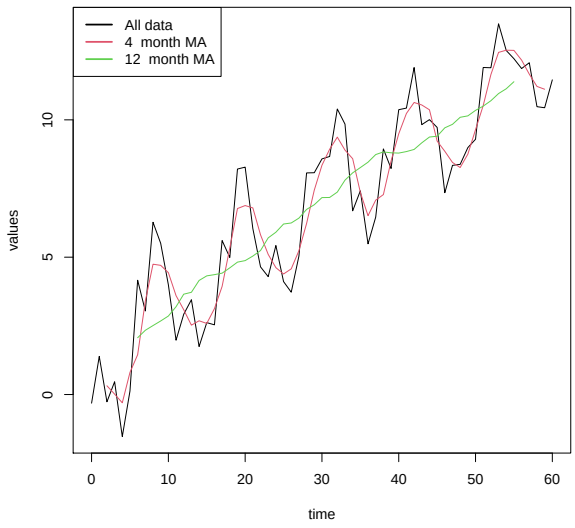

Figure 4.6: The seasonality can be removed by applying a moving average which has a window size of the period of the data.

is an asymmetric infinite filter, applying exponentially decreasing weights to values further back in time.

Explicitly, the weight w_i for the ith point in the data takes a parameter $0 < \alpha < 1$ and is calculated as:

$$w_i = \alpha^i(1 - \alpha)$$

Note that this means the sum of all weights $1 \ldots \infty$ is equal to 1:

$$\sum_{i=0}^{\infty} w_i = 1$$

Single exponential smoothing is the process of using an exponentially weighted moving average. There are other exponential smoothing techniques, but we won't be looking at those.

The exponentially smoothed value at time t (s_t) is given by:

$$s_t = y_t(1 - \alpha) + \alpha \cdot s_{t-1}$$

So the smoothed value at time t depends on the smoothed value at time $t-1$. By assuming we've got an infinite number of historical values, we can get a closed expression for s_t.

$$s_t = \sum_{i=0}^{\infty} w_i \cdot y_{t-i}$$

However, in the real world we don't have infinite historical values. Often what we do is set s_0 equal to the first value of the time series, and then all following values $(s_1, s_2, \dots)$ can be calculated off of that starting value.

About the value of alpha:

1. A large α will give lots of preference to prior values, leading to not very much smoothing. The weights decay very quickly.

2. A small α will give lots of preference to prior values, leading to more smoothing. A small α is used for very noisy data.

Basically, the choice of α will be chosen by trial and error, comparing different values according to some metric like Mean Squared Error, Mean Absolute Error, Sum of Squares Error, etc.

Chapter 5

Time Series II

5.1 3041F: Time Series II Intro

Time Series Video 8

We'll be covering:

- Stationarity, strict and weak variants.
- Autocorrelation: Measure the same variable but at different time points, and calculate the correlation between those timepoints
- Autocovariance
- Linear Time series models (moving average, autoregressive process, autoregressive moving average process)

5.2 Stationarity

5.2.1 Strict Stationarity

Strict stationarity is a property of a time series.

- A time series is strictly stationary iff for every offset parameter $m > 0$, all time points $t_1, \ldots, t_n$, the joint distribution of $Z_{t_1}, \ldots, Z_{t_n}$ is equal to that of $Z_{t_1-m}, \ldots, Z_{t_n-m}$.
- We also require that if there are two points, Z_t and Z_{t+m} then the distribution of the random vector (Z_t, Z_{t+m}) must be a function of the **lag** $|m|$.

- Given a timeline of a time series and an offset parameter m
- Consider Random Variables $Z_{t_1}, \ldots, Z_{t_n}$, find the joint distribution of them all, then apply a timeshift by the offset parameter m: $Z_{t_{1-m}}, \ldots, Z_{t_{n-m}}$.
- then if the joint distribution of the first set of Random variables is equal to that the joint distribution of the second set of random variables, then the time series is said to be **Strictly Stationary**.

5.2.2 Autocovariance

Define the **Autocovariance** as the covariance of the same random variable but at different time points:

$$\begin{aligned}\mathbb{C}ov[Z_t, Z_{t+m}] &:= \gamma_{|m|} \\ &= \mathbb{E}[Z_t \cdot Z_{t+m}] - \mathbb{E}[Z_t] \cdot \mathbb{E}[Z_{t+m}] \\ &\text{Also note that:} \\ \gamma_m = \gamma_{-m} \mathbb{V}[Z_t] &= \gamma_0\end{aligned}$$

We can use $|m|$ because Covariance doesn't depend on the order of the random variables given to it.

Define the **Autocorrelation** as the autocovariance divided by the variance of a time series. The autocorrelation (given the offset parameter m) ρ_m is defined by:

$$\rho_m = \frac{\gamma_m}{\gamma_0}$$
$$\rho_0 = \frac{\gamma_0}{\gamma_0} = 1$$

5.2.3 Weak Stationarity

Time Series Video 9

Weak stationarity is defined by orders (order 1, order 2, etc). So to have weak stationarity of order n, the time series must have equal n-th moments. For example, weak stationarity of order 1 requires a constant expectation.

Weak Stationarity of order 2 requires

- A constant expectation.

$$\mathbb{E}[Z_t] = \mathbb{E}[Z_{t-m}] \forall m$$

- the autocovariance to only depend on the offset parameter m. For example, where t and s are valid indices to the time series Z:

$$\mathbb{C}ov[Z_t, Z_{t-m}] = \mathbb{C}ov[Z_s, Z_{s-m}]$$

Therefore Weak Stationary of order 2 implies Weak Stationarity of order 1.

5.2.4 Some examples of Weak stationarity

Time Series Video 10

If we define a time series as $y_t = e_t - \frac{1}{2}e_{t-1}$ with all error terms independant and $e_t \sim N(0, \sigma^2)$ then:

$$\begin{aligned}
\mathbb{E}[y_t] &= \mathbb{E}[e_t - \frac{1}{2}e_{t-1}] = 0 + 0 = 0 \\
\gamma_0 &= \mathbb{E}[\left(e_t - \frac{1}{2}e_{t-1}\right)] - \mathbb{E}[y_t]\mathbb{E}[y_t] \\
&= \dots \\
&= \sigma^2(1 + 0.25) + 0 \cdot 0 \\
&= 1.25\sigma^2 \\
\gamma_1 &= \mathbb{C}ov[y_t, y_{t-1}] \\
&= \mathbb{E}[(y_t)\,(y_{t-1})] - 0 \cdot 0 \\
&= \dots \\
&= -0.5\sigma^2 \\
\gamma_k &= \mathbb{E}[y_t \cdot y_{t-k}] - 0 \cdot 0 \\
&= \mathbb{E}[\left(e_t - \frac{1}{2}e_{t-1}\right)\left(e_{t-k} - \frac{1}{2}e_{t-1-k}\right)] \\
&= 0 - \frac{1}{2} \cdot 0 - \frac{1}{2} \cdot 0 + \frac{1}{4} \cdot 0 \\
&= 0
\end{aligned}$$

5.2.5 More examples, introducing the autoregressive process

Time Series Video 11

We'll introduce the autoregressive (AR) process as having parameter $|\phi| < 1$ and being a time series like:

$$y_t = \phi y_{t-1} + e_t$$

Which implies that the expectation can be calculated as:

$$\mathbb{E}[y_t] = \cdots = \phi\mathbb{E}[y_{t-1}]$$

So now we've got a self-referencial statement where the expectation at time t depends on the expectation at time $t-1$. This can collapse to become:

$$\begin{aligned}\mathbb{E}[y_t] &= \lim_{h\to\infty} \phi^h \mathbb{E}[y_{t-h}] \\ &= 0 \cdot \mathbb{E}[y_{t-h}] \\ &= 0\end{aligned}$$

And the variance will similarily be self-referencial, and collapse to give us:

$$\mathbb{V}[y_t] = \cdots = \phi^{2h}\mathbb{V}[y_{t-h}] + \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}$$

And so taking the limit as $h \to \infty$:

$$\begin{aligned}\mathbb{V}[y_t] &= \sigma^2 \sum_{i=0}^{h-1} \phi^{2i} \\ &= \frac{\sigma^2}{1-\phi^2}\end{aligned}$$

5.3 Autocorrelation

Given a certain time series like $X_t = (x_1, \ldots, x_n)$ we can calculate the autocorrelation for offset m like so:

$$\hat{\rho}_m = \mathbb{C}orr[X_t, X_{t-m}] = \frac{\sum_{t=m+1}^{n} \left(X_t - \bar{X}\right)\left(X_{t-m} - \bar{X}\right)}{\sum_{t=1}^{n} \left(X_t - \bar{X}\right)^2}$$

A **Correlogram** is a plot with m on the x axis and the m-th lagged autocorrelations on the y axis.

5.4 Significance testing of Autocorrelation

Time Series Video 12

Note that when you calculate autocorrelation with an offset of m, you'll lose exactly m data points. For example, in calculating the autocorrelation with $m = 1$, there is no Z_{t_0} data point to match up with the Z_{t_1} data point.

Today we look at the statistical significance of each of the m autocorrelations. If the underlying process is a moving average process, then the number of statistically significant autocorrelations is equal to the order of that moving average process.

Recall that the m-th autocorrelation is defined as:

$$\hat{\rho}_m = \frac{\sum_{t=m+1}^{n} \left(X_t - \bar{X}\right)\left(X_{t-m} - \bar{X}\right)}{\sum_{t=1}^{n} \left(X_t - \bar{X}\right)^2}$$

Now we can form a standard normal random variable by subtracting the mean and dividing by the sample variation. We have to find the distribution of $\hat{\rho}_m$ numerically, by simulating the histogram many many times (No easy analytical solution exists). After doing this, we find:

$$\hat{\rho}_m \sim N(0, \frac{1}{n})$$

And so we can construct a standard normal like so:

$$Z = \frac{\hat{\rho}_m - 0}{\frac{1}{\sqrt{n}}}$$

5.4.1 Hypothesis tests with Autocorrelation

We know that $\hat{\rho}_m \sim N(0, \frac{1}{n})$ (as $n \to \infty$). Thus if $|\hat{\rho}_m| > \frac{t_{\alpha/2}}{\sqrt{n}}$ we would reject $H_0 : \rho_m = 0$ at the α significance level (where $t_{\alpha/2}$ is the usual value from student's t-distribution).

The **Bartlett's Test** is the same as above, except we reject H_0 if $|\hat{\rho}_m| > \frac{2}{\sqrt{n}}$ because fuck accuracy.

The **Portmanteau test** allows us to also test more complicated hypotheses such as:

$$\begin{aligned} H_0 &: \hat{\rho}_i = 0 \quad \forall i \in \{1, 2, \ldots, k\} \\ H_1 &: \hat{\rho}_i \neq 0 \quad \text{For at least one} i \in \{1, 2, \ldots, k\} \end{aligned}$$

Recalling that the sum of squared independent standard normals is chi-square distributed, we define the Portmanteau test for offset m as:

$$\underbrace{Q(m) = n \sum_{i=1}^{m} \hat{\rho}_i^2}_{\sim \chi^2_m}$$

So now we've got a chi-squared distribution, we can look up the value of $Q(m)$ in the tables.

5.4.2 The Backshift Operator

We define the **Backshift Operator** as:

$$\underbrace{B(x_t)}_{\text{The backshift operator}} = x_{t-1}$$

Time Series Video 13

Chapter 6

Terms, Definitions, R trickery, and side-notes

6.1 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 itselve, 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, evey state communicates with itself.

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

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 recurrant 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 recurrant state in a chain with infinite total states and therefore the mean recurrance time infinite.

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

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

Recurrant An equivalence class with 1 or more recurrant 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.

6.2 R trickery

Useful functions:

6.2.1 runif

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

6.2.2 cumsum

cumsum(sequence): cumulative sum of the given 'sequence'

6.2.3 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'.

6.2.4 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.

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