# CS2 Theory

## Ch-1: Reinsurance

**Definition**

Reinsurance is insurance for the insurance company. By using reinsurance, the insurer seeks to protect itself from large claims. The mean claims amount and variability of claim amounts paid by the insurer both reduces.

**Purpose of buying Reinsurance**

1. Mean claim amount paid by insurer E(Y) falls
2. Variability of claim amounts paid by insurer V(Y) falls.

If variability increases, it becomes difficult to model.

1. Insurer can make use of expertise of reinsurer to model high-risk portfolios

**Asset Quality of Reinsurance**

1. Financial importance not emotional
2. Probability of event occurring against which insurance is taken should be very low.
3. Other Risk Criteria -
   1. Independent Individual risks
   2. Large number of similar risks should be pooled in order to reduce the variability and uncertainty.
   3. The insurer’s liability must be limited.
   4. Moral hazards should be eliminated as far as possible since these are difficult to quantify, result in selection against the insurer and lead unfairness in the treatment of some policyholders.

**Proportional Reinsurance**

Here, the insurer and reinsurer spilt the claim in pre-defined proportions.

**Quota Share**

The proportion is same for all the risks i.e. every individual policy has a say, 60-40 risk share leading to an overall share of the same ratio.

**Surplus**:

Proportion can vary from one risk to the next i.e. claim-wise risk division happens.

**Retained Proportion** (α) where 0<α<=1

Here, Proportion of risk is retained by the insurer. Thus, Y=αX and Z= (1-α) X.

**Non-Proportional Reinsurance**

**Individual Excess of loss**:

Here, the reinsurer covers any claim above a retention limit ‘M’ whereas the insurer pays up to the retention limit.

Thus, Y=min (X, M) and Z=max (0, X-M).

**Stop-loss reinsurance**

Reinsurer is required to make payments if the total claim amount for a specified group of policies exceeds a specified amount (which may be expressed as a percentage of gross premiums).

**Conditional Reinsurance**

Here, the reinsurer is only aware of these claims in which it is involved. This is a special type under Excess of Loss Reinsurance only.

Data for Reinsurer – Truncated as there is complete loss of information of claims below retention limit.

Z = X-M given X>M, Z>0

**Policy Excess or Deductible (E)**

Insurance policies with an excess are common in motor insurance and many other kinds of property and accident insurance.

When a policy excess applies, the policyholder pays for the first part of each loss up to an excess level E. Any amount greater than ‘E’ will be met by the insurer

Here, the insurer’s conditional distribution takes the same form as that of the reinsurer’s conditional distribution.

Y= X-E given X>E

When X follows Pareto (α, λ), Y~Pareto (α, λ+ E)

**Effect of Policy Excess**

Clearly, the premium due on any policy with an excess is less than that on a policy without an excess.

Introducing policy excess increases mean claim amount paid by insurer. This is because small losses (X<E) are met in full by the policyholder. **Under Pareto, mean amount increases by = E / (α-**1)

However, when the mean claim amount increases, the insurer has to deal with fewer claims.

**Risk Premium**

In practice, expenses form a significant part of the insurance cost. So, the presence of an excess might not affect the premium as much as might be expected. A premium calculated ignoring expenses is called a ‘risk premium’.

**Memory Loss Property** of Exponential and Geometric Distribution

These distributions cannot understand “from which point the increment (m) is happening” only how much increment takes place.

Thus, P (X > x + m given X>x) = P (X > 2 x + m given X>2 x)

**Inflation**

Provide the retention limit does not change.

Y = K\*X for K\*X <= M or M for K\*X > M

For smaller claims below M/1.1, increase by 10%

For claims above M, no change – insurer pays only amount M

For claims between M/1.1 and M – increase by less than 10%

As a result the overall mean amount rises by less than 10%.

The burden of inflation is borne by the reinsurer when the retention limit does not change with inflation.

**Comment – Akshay Sir mock 1**

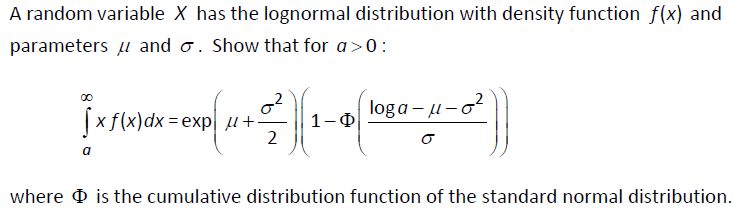
The aggregate excess of loss arrangement results in lower mean aggregate claim amount and a lower standard deviation and hence smaller uncertainty to the company. Therefore, this reinsurance may be preferable for a risk-averse company.

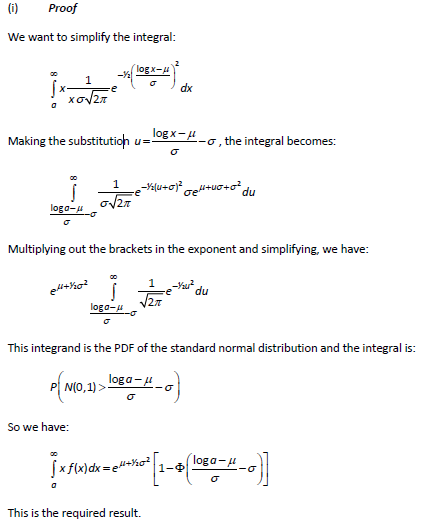
However, we should also consider the cost of each arrangement i.e. reinsurance premium, the risk that the company is prepared to take, and whether such arrangement is possible.

**Comment XOL vs. Proportional Reinsurance – Akshay Sir R mock 2**

* Insurer's mean claim amount = 27.39547 (Proportional) and 30.2943 (XOL)
* Insurer's variability = 1495.392 (Proportional) and 1167.084 (XOL)
* Proportional reinsurance contract provides lower mean payment amount but it has a larger variance
* Choice of contract depends on the current risk appetite of Insurer [Patent comment]
* If they are more concerned with a lower variance of the payments, then they may be inclined to enter into the individual XOL contract despite this leading to a higher expected payment based on the sample
* If they have the ability to deal with the potentially higher payments, then they may instead choose proportional reinsurance to try and benefit from the lower overall expected per claim amount based on the sample
* Reduction in mean claim amount = 15% (Proportional) and 6.005% (XOL)
* Reduction in variability = 27.75% (Proportional) and 43.6% (XOL)

**Proof of Log Normal Distribution**





## Ch-2: Risk Model

**Mixed Random Variable** is a Random Variable which is otherwise continuous but it has a probability mass function (pdf) at one or more points.

**Collective Risk Model**

Aggregate claim amounts may be modelled using a compound distribution. The aggregate claim amount S is the sum of a random number of IID random variables.

* X – Individual claim amount (discrete or continuous)
* N – Number of claims (discrete)
* S = X1 + X2 + … + XN
* If X is discrete, S is discrete and IF X is continuous, S is a mixed RV.

Assumptions-

1. Xi’s are IID random variable.
2. X and N are independent random variable i.e. Cov (X, N) = 0

**Individual Risk Model**

Under this model, a portfolio consisting of a ‘fixed number of risks’ is considered

Assumptions about the whole portfolio-

1. Risks are independent.
2. Claims amounts from these risks are not necessarily identically distributed random variable.
3. The number of risks does not change over the period of insurance cover.

So, here S is defined as the aggregated claim amounts –

S = Y1 + Y2 + … + Yn

#Note- small ‘n’ as the number of policies is fixed.

#Assumptions about each individual risk-

1. No. of claims under the jth risk is either 0 or 1.
2. Probability of claim under the jth risk is qj.

I.e. Nj ~ B (1, qj ) – Bernoulli distribution

If a claim occurs under the jth risk, the claim amount is denoted by Xj with mean μj and variance 𝛿2j

|  |  |  |
| --- | --- | --- |
| **Points** | **Collective** | **Individual** |
| Number of risks | Not specified. | Specified. Remains fixed over the period of insurance cover. |
| Number of risk of each individual risk | No restriction | Restricted. |
| Independence assumption | Individual claim amount are independent | Individual risks are independent. |

**List characteristics that insurable risks usually have**

1. Policyholder has an interest in the risk.
2. Risk is of a financial nature and reasonably qualifiable
3. Independence of risks
4. Probability of event is relatively small
5. Pool large numbers of potentially similar risks
6. ultimate limit on liability of insurer
7. moral hazards eliminated as far as possible
8. claim amount must bear some relationship to financial loss
9. sufficient data to reasonably estimate extent of risk / likelihood of occurrence

**List characteristics of a short-term insurance contract**

1. Fixed Term
2. policy lasts for a relatively short period of time
3. policyholder pays a premium
4. insurer pays claims that arise during the policy term
5. option (but no obligation) to renew policy
6. claim does not bring policy to an end

**Normal vs. Lognormal distribution**

Normal distribution is a symmetrical distribution where majority of the area is covered within μ +/- 3\*sigma.

It has a very short tail as a result of which it undervalues extreme outcomes.

Lognormal distribution has a ‘fat tail’ as a result of which it puts a lot of weightage on extreme outcomes.

**Pareto vs. Exponential** – Pareto has heavier tail than exponential. Under Pareto, upper tail tends to zero as a power of x.

**Give three reasons why a general insurer might use a lognormal distribution to model its losses. (Akshay sir’s mock 3)**

The general insurer might use a log normal distribution to model its losses because in case of general insurance, the number of claims and claim severity increases exponentially and things are more prone to damage. Also, in case of any extreme events the damage is very high resulting in very high claim amounts.

* It takes values from 0 upwards
* It is positively skewed
* It is long-tailed.

**Lognormal distribution as heavier tail than gamma distribution** i.e. we can calculate lower moments about origin in case of lognormal as compared to gamma distribution. Also, when the pdf of lognormal distribution is divided by pdf of gamma distribution; when x tends to infinity, the ratio tends to infinity as well. This indicates that log normal has a heavier tail.

A heavier tail means that the upper tail end of the distribution decays more slowly to zero.

**State simplifications usually made in the basic model for short term insurance contracts.**

Model assumes that mean and standard deviation of claim amounts are KNOWN with certainty.

Model assumes claims are settled without any delays.

No expense and Interest allowance.

E.g. Forms of Short-Term Insurance – Car or content insurance.

**Advantage of Negative Binomial compared to Poisson**

Under Negative Binomial, Variance> Mean. Whereas, under Poisson, Variance = Mean.

Thus, for sample data set having sample variance>sample Mean, Negative Binomial is a better fit.

Example of insurance contract where individual risk model is suitable – Term Assurance and where it is not suitable – Household Contents Insurance

**Why is negative binomial type 2 preferred over negative binomial type 1 for modelling claims?**

Under negative binomial type 1, Number of claims n=k, k +1, k +2…

So, unless there is a specified reason why the number of claims must be at least ‘k’, negative binomial type 1 is not preferred.

**Negative Binomial type 2 vs. Binomial distribution for number of claims**

Under NB2, N can take any non-negative integer value, unlike under Binomial which has an upper limit.

**Other characteristics that most general insurance products share are:**

* + Cover is normally for a fixed period, most commonly one year, after which it has to be renegotiated. There is normally no obligation on the insurer or insured to continue the arrangement thereafter although in most cases a need for continuing cover may be assumed to exist.
  + Claims are not of fixed amounts, and the amount of loss as well as the fact needs to be proved before a claim can be settled.
  + A claim occurring does not bring the policy to an end.

**Coefficient of Skewness –**

* **S ~ Compound Poisson**

Coefficient of Skewness = λ\*m3/ (λ\*m2)1.5

This shows that the distribution of S i.e. aggregate claims is positively skewed because m3 (third moment about zero of non-negative random variable Xi) greater than 0.

Note: distribution of S is positively skewed even if Xi is negatively skewed.

When λ tends to Inf, then coefficient of Skewness tends to 0, thus, S is almost symmetric.

* **S ~ Compound Binomial**

If p>0.5, then the N ~ Binomial is negatively skewed. Here, S is also negatively skewed.

If the claim amounts are fixed, then Coefficient of Skewness of S and N are equal.

## Ch-3 Survival Models

**Life table** is a convenient (by design) computational tool based on a specific survival model. Our task is to process a survival model and the output is the life table.

A Life table is based on unitary method.

**Disadvantage of Life Tables** – It is only defined for integer ages. We can calculate only approximate probability using assumptions for any non-integer ages.

**Use of the expectation of life -** Often used as a measure of the standard of living and health care in a given country

**#Comparison of 3 assumptions -**

1. **Constant Force of Mortality**

* μ\_x + t is a constant μ\_x for 0<=t<1
* Tx ~ Exp(μ\_x)
* Thus, under CFM force of mortality is Constant.
* Survival Function = exp(-t\*μ\_x) i.e. exponentially decreasing on time
* Here, Central mortality rate **mx = μ\_x**

1. **Uniform Distribution of Deaths**

* PDF of Tx is uniform i.e. tpx\*μ\_x+t is a constant.
* Survival Function = t p x is linearly decreasing on time
* μ\_x + t is increasing
* Thus, cancelling out the effect of one another making the product a constant.
* Thus, under UDD force of mortality is increasing.
* This is more acceptable biologically.
* Acceptability - over single-year spans

1. **Balducci Assumption**

* 1/Lx is a linear function.
* Thus, under Balducci force of mortality is decreasing.

**Expected order**

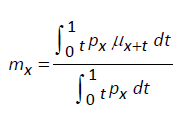
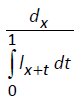
Balducci > CFM > UDD

**Initial Rate of mortality (qx)**

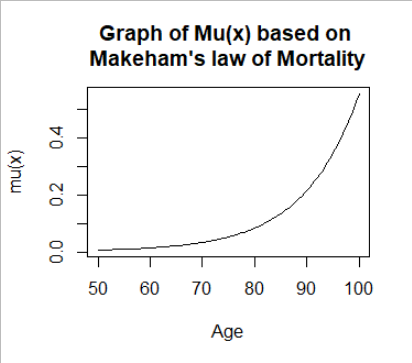
Probability that a life currently aged x will die before exact age x + 1

**Central Rate of Mortality (mx)**

Probability that a life alive between ages x and x + 1, dies.

 or 

**Usefulness** - when the aim is to project number of deaths, given the number of lives alive in that age group. We adjust the definition of mx to incorporate age groups broader than one year.

**Akshay Sir CS2B Mock 1 – Reasonableness Check**

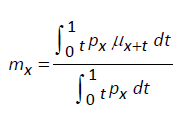
* **Survival probability** - From the graph, we can see that the mortality is increasing over age. So, less people should be able to survive to higher ages and thus the probability 50p50 should be very low. We calculate 50p50=0.285% which is consistent with our expectations.
* **Complete expectation** - The complete expectation of a life aged 50 would be 25 years (if mortality was uniform over 50-100 ages i.e. UDD, equal number of deaths every year). However here, the mortality is increasing thus we expect a value between greater than 25 but less than 50. We calculate E(T\_50) = 26.30195 years which is consistent our expectations.
* **Kurtate expectation** – we know that e\_50 + 0.5 = e-knot\_50.

Here, e\_50 + 0.5 = 26.30384 approximately equal to e-knot\_50 = 26.30195 years.

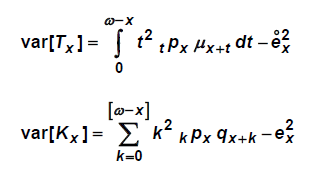
* **Central mortality** – is a measure of the average mortality rate over given age range so we expect it to have similar values as μ\_50.5 and q\_50.

We calculated m\_50 = 0.007424539, μ\_50.5 = 0.007424823 and q\_50 = 0.007399177 – results are consistent with our expectation. Also m\_50 >= q\_50.

**‘mx >= qx always’ True? Answer=Always.**

Denominator of mx is less than 1. So mx >= qx always.

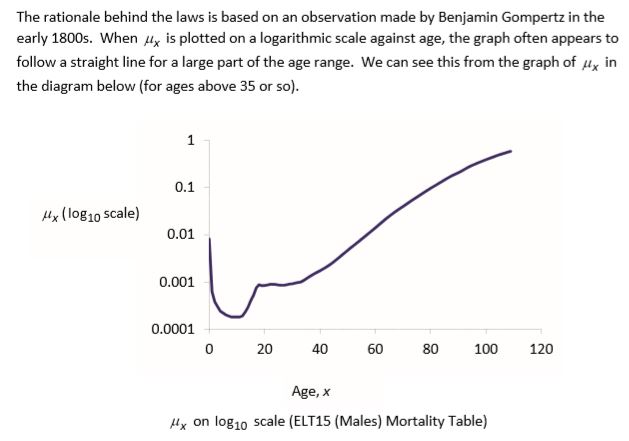
**Future Lifetime Variance**



**Comment on calculating ‘B’ and ‘c’ from given values of 1p50 and 2p50 (UK April 2009)**

* + They estimated values may not satisfy always as they only comply to these specific observations
  + General approach – MLE for fitting more complex models

**Explain meaning for Sx (t):** probability that a life currently aged x lives for at least ‘t’ more years.

**Gompertz and Makeham’s law of mortality**

**Under Gompertz Curve,** μx(t) = a\*exp (-b\*e-c\*t)

Where, a = asymptote/ carry capacity of the system, b = displacement on X axis, c = growth rate.

**Difference between Gompertz Curve & simple Logistic curve**

Under logistic curve, as population increases the growth rate linearly decrease but, under Gompertz curve the growth rate exponentially decreases.

*Gompertz Model is used to model Tumour Cells Growth.*

**Why under Weibull (α, β) survival function is given by αβ\*t (β-1)? (Mock 2)**

* The force of mortality at age x + t is the negative derivative of the log survival function with respect to time ‘t’ i.e. the derivative of α\*t^β
* μ \_x+t = d/dt (α\*t^β) = α\*β\*t^(β-1)

## Ch-4 Estimating the lifetime distribution function

Our task is to estimate F(t) or S(t) empirically i.e. based on data only and without using any statistical distribution.

**Observational Plan** is a framework for any mortality investigation. The aim is to answer three questions –

1. What should we observe?
2. How should we observe?
3. For how long should we observe?

Ideally, we should observe a large cohort of people from different age groups. The method of observation could be maintaining record of say survivors at each integer age. The investigation period can vary as per need of the investigation but should not be too long or else the results become useless.

**Event of interest**

Event around which investigation is set up e.g. recording deaths due to natural causes.

**Censoring Types**

1. **Right Censoring** - Censoring Mechanism cuts short observation in progress. We are only aware of the lower bound of data.

E.g. with respect to insurance –

* Life insurance policyholder surrenders or policy lapses.
* Active lives of pension schemes retire under payment option of lump sum or Annuity paid by 3rd party.
* Endowment policy matures.

E.g. with respect to mortality investigation –

When a subject is removed from the investigation for a reason other than by death, so that we do not know exactly when death will occur, just that it occurs after the time of removal.

1. **Left Censoring** - Censoring mechanism prevents us from knowing when entry into the state that we wish to observe took place. We are only aware of the upper bound of data.

E.g. date of birth is unknown. Exact date of policy entry or becoming sick is unknown.

1. **Interval Censoring** - Observational plan only allows us to say that event of interest falls within some interval of time. Right and left censoring are special cases of interval censoring. E.g. year of birth known but exact age unknown
2. **Random Censoring** – when censoring times are not known in advance. It is a subtype of right censoring.

E.g. with respect to insurance - policy surrender or withdrawal date is not known in advance but policy maturity date is known in advance.

E.g. with respect to mortality investigation - When a subject is removed from the investigation for a reason other than by death, and the timing of the removal can be considered a random variable.

1. **Type 1 Censoring** – when censoring times are known in advance i.e. at a predetermined exact date or duration or age. E.g. 5-year investigation.
2. **Type 2 Censoring** – observation is continued until a predetermined number of deaths have occurred. E.g. medical study continued until 5 deaths occur.
3. **Informative or Non-informative Censoring** – when there is no reason to believe that the risk of death of persons still in cohort and those who had left the cohort are different, then censoring is non-informative.

Under informative censoring, there is a reason to distinguish between risk levels.

E.g. with respect to mortality investigation - When the future mortality of a subject censored from the investigation is likely to be different from those remaining in the investigation.

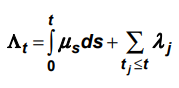
**Nelson- Aalen and Kaplan-Meier (**Assumptions and Notations)

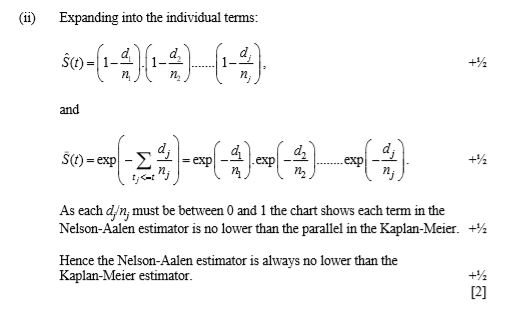
1. Non-informative right censoring is present in population. We observe ‘n’ lives and ‘m’ deaths.
2. Let t1<t2<…<tk be observed times at which deaths are observed. tk denotes kth time some death was observed.
3. We do not assume k=m: more than 1 death can be observed at a single failure time.
4. Suppose dj: number of times at time tj. So, d1 +d2 +…+dk = m
5. Suppose cj: number of lives censored between times tj and tj+1. So, c0+c1+…+ck=n-m. Thus, c0 is the number of lives censored between t0 and t1.
6. Suppose nj: number of lives in a risk group at time tj.
7. Waiting times are independent
8. The group/ sample is homogeneous

**Conventions**

1. Hazard of experiencing the event is 0 at all durations expect those where an event actually happens in our sample.
2. Hazard of experiencing the event is λj = dj/nj at any particular time tj.
3. If any of the individuals are observed to be censored at the same time as one of the deaths, the convention is to treat censoring as if it happened shortly afterwards.

**NOTE –**

* + Because the Kaplan-Meier estimate involves multiplying up survival probabilities, it is sometimes called the **product limit estimate**.
  + Nelson Aalen Estimate is called **Integrated Hazard Function** because  where the integral deals with the continuous part of the distribution and the sum with the discrete part.

**Proof – Nelson-Aalen never lower than Kaplan Meier Estimator**

**Statistically comparing two different hazard rates** –

* + Calculate variance using Greenwood’s Formula.
  + Calculate confidence intervals, Overlap implies equal survival rates and non-overlapping intervals imply that survival rates higher/lower than each other.
  + Other Methods to test – Log-Rank Test and Wilcoxon’s Test.

**Comment – use of proportional Hazard Model:**

* Convenient way of measuring difference in survival rates
* Proportional Hazard model assumes that the hazards are in constant proportion at all durations, which might not be the case as constant proportion implies that graphs do not cross.
* Proportional Model allows for confounding variables.
* Thus, appropriate to use it.

Further tests for proportionality assumption are likely to be required.

**Mr. Bunn’s Bakery – use of Nelson Aalen estimates for future production. Comment (April 2012)**

* + Not a good basis for future production as the analysis is based on one day’s experience.
  + Other Factors-
    - Marketing and Pricing
    - Climate
    - Frequency of Accidental losses (minimise it)
  + Good point of this analysis – accounts for censored data

**Nelson Aalen Estimate vs. Exact Exposed to Risk Estimate (Sept 2018)**

* + Nelson Aalen estimate = 0.38628
  + Exact Exposed to risk estimate = 0.26898
  + Clearly, Nelson Aalen > Exact Exposed to Risk
  + Under Exact Exposed To Risk, we constrain the death rate to be CONSTANT over observation period.
  + Under Nelson Aalen, death rate varies with time according to data.
  + Other Factors –
    - Sample size
    - Inhomogeneous group of people observed.

|  |  |
| --- | --- |
| **Kaplan Meier** | **Nelson Aalen** |
| Non parametric | Non parametric |
| models the cumulative distribution of the lifetime random variable Tx | models the integrated hazard corresponding to the lifetime random variable Tx |
| other decrements are handled via censoring | other decrements are handled via censoring |
| key assumptions –  no heterogeneity (everyone has the same probabilities) | key assumptions –  no heterogeneity (everyone has the same hazard rate)  non-informative censoring |
| Likelihood function – | - |
| fitting the model –  is the estimate of discrete hazard |  |
|  |  |
| Also known as Product limit estimator | Also known as Integrated Hazard Function |

## Ch-5 Proportional Hazard Models

**Covariate** is any quantity recorded in respect of each life, such as age, sex, type of treatment, level of medication, symptom severity and so on.

**3 types of Covariates** –

1. Direct Measurement – age, weight, height
2. Indicator Variable – Gender = 0 if male, 1 if female
3. Quantitative Interpretation of a Qualitative Measurement – Ratings

**Proportional Hazard Model** is called so as hazard of any two lives are in the same proportion irrespective of time.

In a proportional hazards model, the hazard of experiencing an event may be factorised into two components:

* + One depending only on duration since some start event, which is known as baseline hazard,
  + Other depending only on a set of covariates and associated parameters.

The hazards for any two lives are therefore in the same proportion at all durations.

**Semi-parametric Approach**

We are unconcerned with the precise form of Hazard. We can estimate β from the data with the effect of covariates ignoring the shape of Baseline Hazard.

**Why “Partial” Likelihood?**

Here, “*π:(j=1, k)*” acts only for lives for which we have observed “deaths” and “not all lives”. Censored Lives’ contribution is ignored.

**Breslow’s Approximation Assumption**

All deaths happen together at the same time. When calculating contribution to partial likelihood, we include all deaths in denominator for each death.

**Fully Parametric vs. Cox Proportional Hazard Model (Sept 2007)**

* Fully parametric models are good for Homogenous Groups. As, Confidence interval for such fitted parameters help test for difference between these groups. Better than semi and non-parametric approach.
* Object of study – requiring knowledge of form of hazard function then fully parametric is preferred. Under Cox Model, knowledge of ‘Form of Hazard function’ is not required.
* Cox Model – standard feature of many statistical packages for estimating survival model. For this purpose, parametric models are not preferred also, they require additional programming.

**Gompertz Model vs. Semi-Parametric Model (April 2012)**

* Under Gompertz model, Hazard is either monotonically increasing or decreasing.
* When hazard does not increase/decrease monotonically, it is better to avoid Gompertz.
* Here, hazard rises and then falls back to initial position.
* Proportional hazard model allows us to access the impact of each individual risk factor.
* A semi-parametric model allows us to determine the shape of hazard from given data. Thus, better over Gompertz.
* However, Gompertz can be framed as a Cox proportional hazard model.

**Advantage of Cox Regression Model (April 2013) OR Why Cox proportional hazard model is widely used in empirical work**

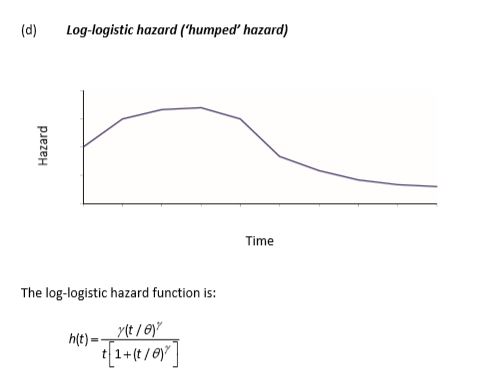
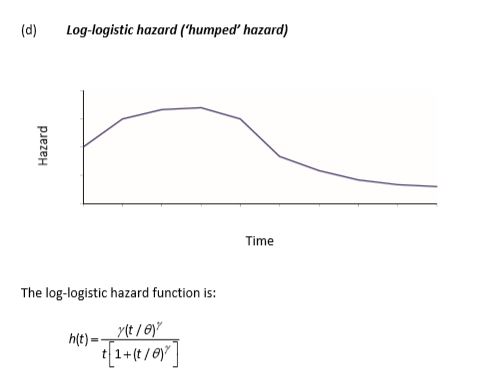
* Exponential function ensures the hazard is always positive.
* The log hazard is linearly related to the covariates.
* We can ignore the shape of hazard and directly calculate the effect of covariates directly from the data.
* We can compare without calculating Baseline Hazard
* Widely available standard computer package.
* Popular and well-established model
* Easy to communicate

**Essential Feature of proportional hazard models**

* In a proportional hazard model, the hazard factorises into a component which depends only on duration [h0(t)] and a component which depends only on the covariates [g(z)].
* As a consequence, the effect of a covariate is to shift the hazard up or down by a same proportional at all durations.

**Extending Current Model for suggested additional factors**

* Perform Log likelihood ratio test
* H0: new parameter has no impact (i.e. βj = 0)
* Calculate the log likelihood of the model as it stands, L\_current.
* For each additional factor individually-
  + Extend the model to include the new parameter
  + Estimate the parameter appropriate to that factor
  + Calculate the log likelihood of the new model L\_new
* TS = -2\*(L\_current – L\_new) ~Chi Square with ‘df’ degree of freedom
* ‘df’ = number of parameters in new model – number of parameters in current model (i.e. p + q – p)
* If, TS > 3.841 we have sufficient evidence to reject the null hypothesis at 5% LS
* Thus, it can be concluded that additional parameter is significant.



**Give an example of a situation in which the hazard function may be expected to follow each of the following distributions:**

* **Exponential** - The constant hazard model (exponential) could reflect the hazard for an individual who remains in good health. The level of hazard would reflect the risk of death from unnatural causes, e.g. accident or murder.
* **Decreasing Weibull**- The decreasing hazard model (decreasing Weibull) could reflect the hazard for patients recovering from major heart surgery. The level of hazard is expected to fall as the time since the operation increases.
* **Gompertz-Makeham** - The exponentially increasing hazard model (Gompertz-Makeham) could reflect the hazard for leukaemia sufferers who are not responding to treatment. The severity of the condition and the level of hazard increase with the survival time. Over longer time periods, the Gompertz-Makeham model could be suitable for describing the increasing chance of death from natural causes as age increases.
* **Log-Logistic**- The humped hazard (log-logistic) could reflect a hazard for patients with a disease that is most likely to cause death during the early stages. As the initial condition becomes more severe, the level of hazard increases. But once patients have survived the period of highest risk, the level of hazard decreases.

**Use of Gompertz Model (April 2012)**

|  |  |
| --- | --- |
| **Month** | **Monthly Hazard** |
| 1 | 0.025 |
| 2 | 0.042 |
| 3 | 0.086 |
| 4 | 0.067 |
| 5 | 0.044 |
| 6 | 0.024 |

* + Taking log of Gompertz Hazard gives us log(λX) = log(B) + x\*log(c)
  + Thus, Gompertz model is useful when hazard is either monotonically (constantly) increasing or decreasing with age. Empirically, this is often a reasonable assumption for middle ages and older ages, which include age range 50-65 years.
  + Here, Hazard initially rises and then falls back to approximately the initial position.
  + To assess the impact of risk factors, a proportional hazard model would be useful as it allows us to assess each individual risk factor.
  + A semi parametric model allows us to determine the shape of hazard from given data. Thus, better over Gompertz.
  + However, Gompertz can be framed as a proportional hazard model like a cox model.

**Why is Gompertz model commonly used in investigations of human mortality?**

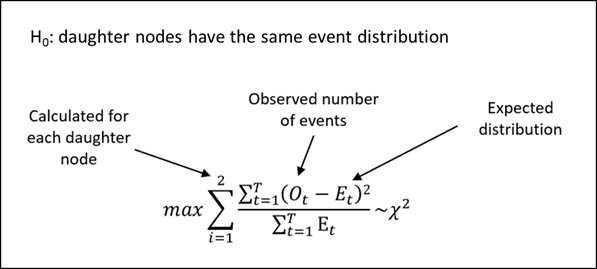
The Gompertz model is simple to understand and to apply, having only two parameters. It also fits human mortality at older ages well (e.g. 30-85 years)

**Extra Stuff – Going beyond Cox regression towards Random Forest Survival Analysis**

Random Forest Modelling has many superlatives, including the ones listed below:

* + It identifies non-linear relations between variables
  + It requires less data transformation than methodologies restricted by their assumptions (e.g. linear, logistics or Cox regression)
  + It accounts for interactions between features
  + It performs well with a high number of features
  + It is a feasible solution in limited timeframes

Using Random forest in Survival Analysis is very similar to Random Forest Models used for classification or regression i.e. many trees are trained based on a subsample of the dataset and the predictions are aggregated based on the trees’ votes. The survival-specific part is that the model bases on e.g. log-Rank test, instead of the Gini Impurity, to evaluate each split according to the difference between observed and expected number of events in each daughter node.

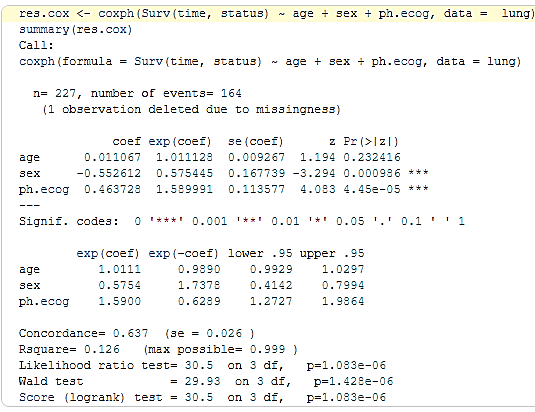


Another interesting alternative algorithm are survival models based on e.g. Penalised Cox Models, Gradient Boosting, Supper Vector Machine, models using Bayesian statistics and many more.

**Survival Analysis in R**

R is specialized in statistical computing and graphics, and therefore it is currently very popular in academia. As survival analysis originates from and is frequently performed in medical research (academic environment), most of the innovation is initiated by this field. Therefore, R offers a wide variety of CRAN packages with plenty of implementations of survival models, incl. alternative algorithms (e.g. randomForestSRC for survival modelling using Random Forest or mboost for Gradient Boosting algorithms), extensions to Cox regression (e.g. glmnet for Cox modelling with regularization or BMA for Bayesian Model Averaging of Cox regression models) or packages for model evaluation (e.g. timeAUC), giving a lot of flexibility to the user. If integration into industrialized production environment is needed, R can be run at scale using solutions like Azure Databricks.

**Comment on Cox Proportional Model fitted – (Lung dataset, model: age+sex+ph.ecog)**



* **β\_sex Coef** = -0.553 (female hazard is lower than male hazard)
* **Hazard ratio of β\_sex = Exp (Coef)** = 0.5750 < 1 - implies that in presence of β\_sex, hazard decreases.

Hazard Ratio is interpretable as multiplicative effect on hazard.

Thus, holding other covariates constant, Being female reduces hazard by a factor of 0.58 or 42%.

Thus, being female is associated with good prognostic.

* **Z** = -3.29 (coef/ se = -0.553/0.167 = -3.29) is the Wald test statistic under null hypothesis β=0. It is a right tailed test.
* **Pr(> |z|)** = 0.0009 – is the p-value. Lower the p-value, more significant is the covariate.

Thus, we have insufficient evidence to reject H0 at 5% LS. β\_sex is a significant parameter.

* **Log-likelihood ratio test** = 10.63 on 1 dof, p = 0.0011 < 0.05
* **Insignificant Covariate Age** – p value = 0.23 > 0.05 indicates it is insignificant. HR ratio = 1.01 with a 95% CI (0.99, 1.03). The Confidence interval includes 1. Thus, these results indicate that age makes a smaller contribution to the difference in the HR after adjusting for ph.ecog values and patient’s sex. For e.g. holding the other covariates constant, an additional year of age induce daily hazard of death by a factor of exp (β\_age) = 1.01 i.e. 1% which is not a significant contribution.
* **Patent Comment -** The likelihood-ratio test, Wald test, and score log rank statistics evaluate the omnibus H0: β=0 i.e. all covariates are insignificant in the model.

These three methods are asymptotically equivalent.

For large enough N, they will give similar results.

For small N, they may differ somewhat. The Likelihood ratio test has better behaviour for small sample sizes, so it is generally preferred.

## Ch-6 Exposed to Risk

**Rate Interval** is a one-year period during which Age Label does not change.

If that one-year period is determined by the person’s birthday, it is called **Life-Year Rate Interval**.

If determined by Policy Purchase date then **policy-year rate interval**.

If determined by Calendar year then **calendar year rate interval**.

**Central Exposed to risk** for a life with age label x, is the time from Date A to Date B where, Date A is the latest of {date of reaching age label x, start of investigation, date of entry} and Date B is the earliest of {date of reaching age label x + 1, end of investigation, date of exit}.

**P\_x, t varies over time because of the following reason**-

P\_x, t falls

* Passing the (x +1) th birthday
* Death
* Censored

P\_x, t rises

* Passing xth birthday
* Joining observation

**Principle of Correspondence**

A life alive at time t should be included in the exposure at age x at time t if and only if, were that life to die immediately, he/she would be counted in the death data dx at age x.

**Sverdrup Rule**

As death is a rare event, data of deaths will be limited so we want to keep in accurate. Thus, we modify the P\_x, t data.

**Assumptions**

* P\_x, t varies “***linearly***” between census dates. Allows us to apply trapezium rule or average between census dates.
* ***Human birthdays are uniformly distributed over the year***. This is needed in order to average the data for 51- and 52-year olds (as per question) or to adjust the Exposed-to-risk from age nearest to age last birthday (as per question).
* 1st January data in year x can be taken as 31 December data in year x− 1. This is needed in order to use start-2015 data as end-2014 data (as per question).

**Sub-Division of Data – necessity**

* To solve the problem of heterogeneity
* The multiple state and Poisson models and analyses are based on the assumption that we can observe groups of identical lives or at least whose mortality characteristics are the same. Such a group is said to be Homogeneous.
* Even with a limited scope of mortality investigation of a specific age/gender there would still be a wide variety, making it heterogeneous. It’s a result of this, we can only estimate an average rate of mortality over the whole group of lives and not for an individual. This is a problem of the Insurance Company.

**Solution to this problem: Sub-Division of data based on factors**

* Sex
* Age
* Policy type
* Smoking status
* Duration in force
* Sales channel
* Policy size
* Known impairments
* Marital status
* **Level of underwriting**

Underwriting is the process by which life insurance companies divide lives into homogeneous risk groups by using the values of certain factors (rating factors) recorded for each life.

* **Occupation of policyholder**
  + Occupation can have several direct effects on mortality and morbidity. Occupation determines a person s environment for 40 or more hours each week. The environment may be rural or urban, the occupation may involve exposure to harmful substances e.g. chemicals, or to potentially dangerous situations e.g. working at heights. Much of this is moderated by health and safety at work regulations.
  + Some occupations are healthier by their very nature e.g. bus drivers have a sedentary and stressful occupation while bus conductors are more active and less stressed. Some work environments e.g. pubs, give exposure to a less healthy lifestyle.
  + Some occupations by their very nature attract more healthy workers. This may be accentuated by health checks made on appointment or by the need to pass regular health checks e.g. airline pilots. Some occupations can attract less healthy workers, for example, former miners who have left the mining industry as a result of ill health and then chosen to sell newspapers. This will inflate the mortality rates of newspaper sellers.
  + A person s occupation largely determines their income, which permits them to adopt a particular lifestyle e.g. content and pattern of diet, quality of housing. This effect can be positive or negative e.g. over-indulgence.
* **Geographical location**
  + Climate and geographical location are closely linked.
  + Levels and patterns of rainfall and temperature lead to an environment which is amicable to certain kinds of diseases e.g. those associated with tropical regions.
  + Effects can also be observed within these broad categories e.g. the differences between rural and urban areas in a geographical region.
  + Some effects may be accentuated or mitigated depending upon the development of an area e.g. industry leading to better roads and communications.
  + Natural disasters (such as tidal waves and famines) will also affect mortality and morbidity rates, and may be correlated to particular climates and geographical locations.
* **Quality of housing**
  + The standard of housing encompasses not only all aspects of the physical quality of housing (e.g. state of repair, type of construction, heating, sanitation) but also the way in which the housing is used e.g. overcrowding and shared cooking.
  + These factors have an important influence on morbidity, particularly that related to infectious diseases (e.g. from tuberculosis and cholera to colds and coughs) and thus on mortality in the longer term.
  + The effect of poor housing is often confounded with the general effects of poverty.

**Consequence to Insurance Company charging same premium over different risks**

* Unstable position if average premium is charged to high and low risks.
* In the Long run, office will tend to loose low risk business and attract more of high risk business. Reason being for high risks, the aggregate premium is lower than actual claim test. This is called Anti Selection leading to short term losses and eventually threatens insolvency.
* The company can avoid this only by changing different premium rates appropriate to different levels of risk presented by the applicants. This is the process of risk classification.

**Problems of Sub-division**

* Sub-division using many factors can result in the numbers in each class being too low.
* We need to strike a balance between maintaining homogeneity and retaining a large group for statistical analysis.
* Sufficient data for sub-division often not available.
* Marketing pressures mean proposal forms are kept to a minimum. Say, the policy is marketed on the strength of simplicity and ease of processing.

**Differentiate: Central and Initial Exposed to Risk (Sept 2007)**

* Central Exposed to risk at age x is observed waiting time in multiple state or Poisson model. It is sum of time spent under observation by each life at age x.
* In aggregate data, Ecx is an estimate of no. of lives exposed to risk at mid-point of the rate interval.
* Initial exposed to risk is the potential waiting time. It requires adjustment for those lives who die, where we require observing until end of rate interval.
* Thus, Ecx + 0.5 dx = Ex

**Example of situation where it is difficult to adhere to principle of correspondence**

* + Problems in adhering to this can arise when the death data and exposed to risk data comes from two different sources where lives are classified differently.
  + E.g. an occupation mortality investigation where deaths data comes from death registers or claims department of office and exposed to risk data from census or based on policies in force which comes from different part of the office.

**Combining Male and Female Life Portfolios – charging same premium based on mixed portfolio – Appropriateness? [2018 September]**

* + Approach will work only if the MIX of males and females remains same – unlikely to be this way.
  + Market Competition – what is the reaction of competitors?
  + Consider mortality improvements and future developments which impact mortality ratio between males and females.
  + Demographic targeted by company – Only Males?
  + Some selection effects are nullified by the fact that all companies are required to charge unisex rates.

**Problem of Heterogeneity**

The multiple state and Poisson models’ analyses are based on the assumption that we can observe groups of identical lives (or at least lives whose mortality characteristics are the same). In practice, this is never possible.

**April 2004 – Same premium in both locations: Gasperton and Great Hawking and 40% more premium for smokers. Comment**

* + As death rates vary geographically it is better to vary premium. Clearly, death rates for smokers and non-smokers are higher in Great Hawking.
  + If the company does not differentiate its prices on the basis of geographical area, it may lose business in Gasperton to a rival company which does differentiate. Conversely, in Great Hawking it may attract new business from rival company, but will under-price the product and hence risk its life assurance fund becoming insolvent.
  + There are relatively little data, so it might be worth adapting a “wait and see” approach.
  + 1.4 times higher death rate will not translate as 1.4 times the premium. The difference may be relatively small.

**President’s Birthday is census date based on astrology-assumptions (April 2013)**

* + Change in P\_x, t is linear
  + Uniform distribution of human birthdays over calendar year
  + Constant force of mortality over 2006-08 period and 2009-10
  + President’s system of census is based on Astrology and his birthday which does not change i.e. is rigid.

**Explain census approximation for calculating Ecx between any two census dates. (Sept 2014)**

* + In survival investigations, population counts will only be available at census dates.
  + Define P\_x, t: number of lives under observation, aged x last birthday, at any time t
  + Suppose that we have the values of P\_x, t only if ‘t’ is a census date.
  + We require the exposed to risk, Ecx, over the interval between the first census and the last.
  + Thus, Ecx = INT (t1, t2): [Px, s ds], where t1 and t2 are the two census dates.
  + To evaluate this, we usually assume that Px, s is linear between census dates.
  + If the censuses are one year apart this leads to the trapezium approximation: Ecx = 0.5\*(Px, t1 + Px, t2)

**Why is it important of divide data into homogeneous classes when undertaking mortality investigation?**

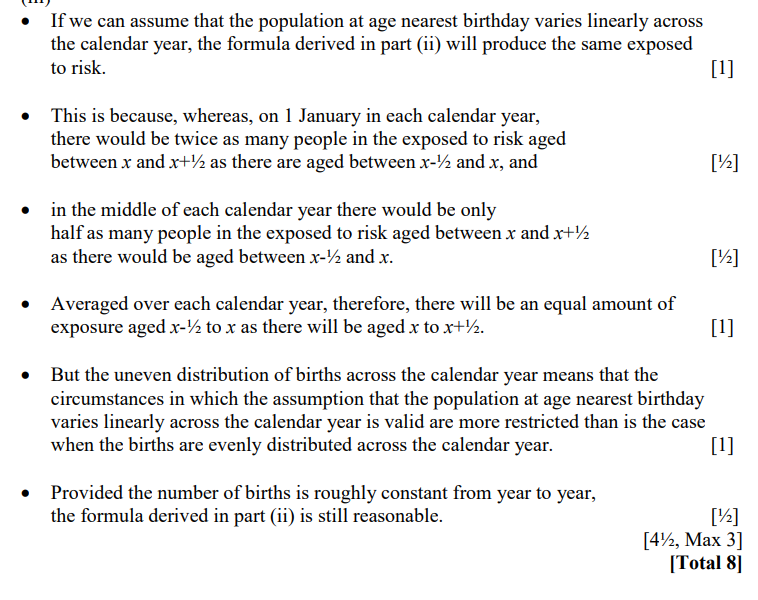
All our models and analyses are based on the assumption that we can observe groups of identical lives or at least, lives whose mortality characteristics are the same.

Although in practise, this is never possible. We can at least subdivide our data according to characteristics known, from experience, to have a significant effect on mortality. This ought to reduce the heterogeneity of each class so formed.

**Assumption – Assignment Question 25**

We are given values of P\_x, t for July of every year. We assume that this figure represents the average of respective years.

Average number of policies in force throughout the year is approximately equal to the number of policies in force at 1st July of every year.

**Implication of formula of P\*x, t (APRIL 2019)**

Given the following data, how to estimate life table?

* + - Date of birth
    - Date of entry into observation
    - Date of death
    - Date of exit from observation while still alive
  + Obtain different estimates for different hazards for each age in respective calendar.
  + If the MLE of Constant Force of Mortality between x and x +1 is μx-hat
  + Then the probability a person at exact aged x will still be alive at age x +1 is just Sx (1). Given the constant force, then
  + Sx (1) = exp (-μ\_x-hat).
  + In general, therefore,
  + S x (m) = m p x = exp (- sum:(j=0, m-1): μ\_x + j -hat)
  + By ‘chaining’ together the probabilities in this way, we can create a life table from our estimates and evaluate probabilities over any relevant age range.

**Appropriateness of charging equal premium for male and female lives by combining the portfolio (2018 Sept)**

* + This approach will only work if the MIX of males and females remains SAME – which is highly unlikely.
  + Market competition – what are the competitor’s doing is important - Other companies may base their rates on a different mix of in force business, or some estimate of future mix.
  + Consider mortality improvements and future developments which affect mortality ratio between males and females.
  + Demography targeted by company – only males?
  + Some selection effects get nullified by charging unisex rates of premium by all companies.
  + The overall mix of business by gender may alter temporarily as those who are likely to lose out by the introduction of the new legislation may make a dash to get cover before the legislation comes into force.

## Ch-7 Extreme Value Theory

**Definition**- Extreme events are outcomes that have a low frequency and high severity. It may arise as result of single cause that has a high financial cost or as an accumulation of events with a related cause.

E.g. with respect to insurance - flood damage to large number of houses, body-injury claim.

E.g. with respect to reinsurance – Claims in the right-hand tail of the claim’s distribution with a low probability but high severity.

**Need** – Because the payments for these claims are large, they can have a significant financial impact on the insurance company. The probabilities for the right-hand tail cannot be calculated accurately using the usual techniques used for the main body of the distribution.

**Q. Why do we need to model extreme events separately?**

* The majority of risk events fall within the main body of the underlying distribution (fitted) and can be used to model the typical events (normal claims) accurately.
* However, there is usually a lack of data of extreme events (Low frequency, high severity). If a distribution is fitted to the entire data set, the parameters estimated will reflect the bulk of data values rather than reflecting the extreme events. Due to which the fitted distribution undervalues the probability of extreme events.

**Generalised Extreme Value distribution (GEV)**

To model extreme events, we can calculate the expression of the form (Xm-αn)/βn that can be approximated by a particular type of distribution called Extreme value distribution.

Whatever, the underlying distribution of data, the distribution of the standardised maximum values will converge to a distribution called the GEV distribution as n increases.

**Parameters** –

α – Location parameter (analogous to mean)

β – Scale parameter (analogous to standard deviation)

γ – Shape parameter (analogous to Skewness)

**GEV Family**

1. **Frechet-type**

Heavy tail, lower bound is known [x > α -β / γ], γ>0, modelling claim amounts or losses

1. **Weibull-type**

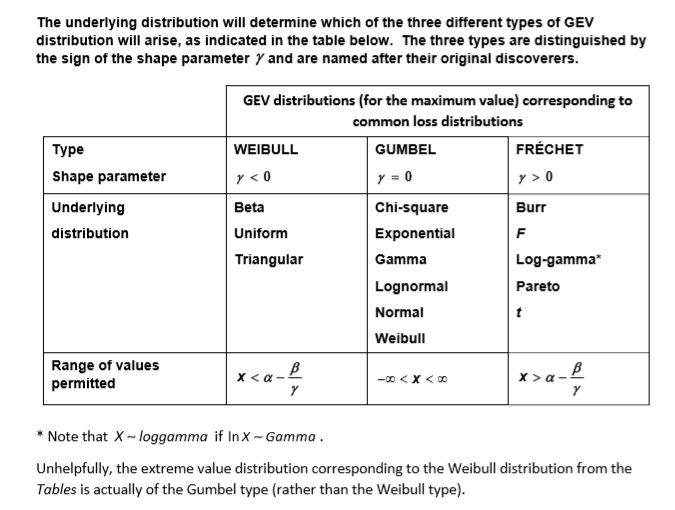
Light tail, Upper bound is known [x < α – β / γ], γ<0, modelling profits and share prices

1. **Gumbel-type**

Unbounded, γ=0

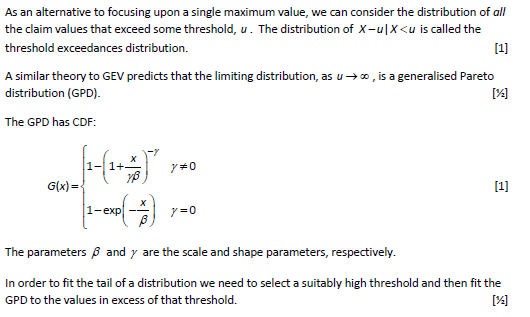
1. **Standard Gumbel**

α = 0, β =1, γ = 0

**Choosing form of GEV Distribution**

**Comparison of Block Maxima and Threshold Exceedance Approach**

* Both assume that the underlying data values are independent and identically distributed.
* Both methods require some subjective decisions about what data values are considered extreme:
  + For the Block Maxima approach, the choice of block size choice is important – too big means too many data values are discarded and too small means more values are retained but many of these are not extreme.
  + For the threshold exceedance approach, the choice of threshold is important – too high means too many data values are discarded and too low means too many non-extreme data values are included.
* In both cases, including non-extreme events will mean the fitted curve will not be the best fit for modelling extreme events.
* By only selecting the maximum in each block, the block maxima approach potentially ignores other extreme events in the same block.

**Alternative Approach that can be used in place of GEV distribution to model extreme events –**

**Key Advantage of GPD over alternative solution –**

The GPD method has the advantage that it uses a larger part of the data and models all the large claims above the threshold, not just the single highest value.

**Generalised Pareto Distribution (GPD)**

Whatever be the underlying distribution of data, the distribution of the threshold exceedances will converge to a generalised pareto distribution as the threshold μ increases.

**Parameters –**

* β = scale parameter, β > 0, β increases, tail become heavy

(Provided all else is constant)

* γ = slope parameter, γ = 0, CDF converges to Exp(1/β),

γ not equal to 0, cdf converges to Pareto (γ, γβ)

**Measures of Tail Weights**

If PDF of Xa tends to zero as x tends to infinity, slower than the pdf of Xb, then Xa has heavier tail.

1. **Existence of moments (about origin)**

If we can calculate moments higher than E (Xk), it implies that the distribution has a light tail relatively. Here, k is a positive integer.

1. **Limiting Density Ratio**

Here, we compare the thickness of two distributions by calculating the relative values of the density functions at the far end of the upper tail.

1. **Hazard rate**

If hazard rate increases with X, relatively few data points are left towards the end – corresponding to a light tail.

1. **Mean residual lifetime tail**

Mean residual life time rises, hazard rate falls, and tail is heavy

**Key advantage of GPD over GEV distribution when modelling for losses –**

GPD will make use of all the data in the tail whereas; GEV might exclude some extreme data values because these are not the most extreme within a particular block.

**Disadvantage/Limitations of GPD analysis (Sept 2020)**

* The generalised Pareto distribution is a limiting distribution and the actual distribution of the exceedances over any finite threshold will be different.
* It calls for a need to analyse alternative thresholds before drawing of concrete conclusions.
* If the random variables are not independent and identical, the analysis does not make sense.
* It does not provide an insight into trends in the data, if present.
* Other factors like weather, rules and regulations might affect the results. (as per question)

**Explain why claim amounts from general insurance policies are typically modelled using statistical distributions with heavy tails.**

Compared to other forms of insurance, general insurance claims are positively skewed with long tails. Therefore, they have more extreme claims and so have heavier tails.

**#Kurtosis**

* **Leptokurtic** Financial data sometimes tends to be much more narrowly peaked in the center of the distribution and to have fatter tails than normal distribution. This shape of distribution is known as Leptokurtic. E.g. in modeling of share prices, larger price movements occur more frequently than predicted by normal distribution.
* Leptokurtic is a measure of the kurtosis of a distribution i.e. the fourth standardized central moment given by .
* **Mesokurtic –** All Univariate normal distribution have kurtosis equal to three and are described as Mesokurtic.
* A distribution with kurtosis > 3 i.e. more peaked with flatter tails is called **Leptokurtic.**
* A distribution with kurtosis < 3 i.e. broader peak with more slender tails is called **Platykurtic.**
* One reason for these fat tails is that the volatility of financial variables does not remain constant, but varies stochastically over time. This property is known as **heteroscedasticity**.
* In the case where the volatility is variable, the resultant pdf is more peaked with fatter tails, i.e. it is leptokurtic.

## Ch-8 Stochastic Process

**Stochastic Process**

* A stochastic process is a model for a time-dependent random phenomenon. So, just as a single random variable describes a static random phenomenon, a stochastic process is a collection of random variables X\_t, one for each time t in some set J.
* Process is denoted by {X\_t: t belongs to J}
* The set of values that the random variables X\_t are capable of taking is called the state space of the process S.

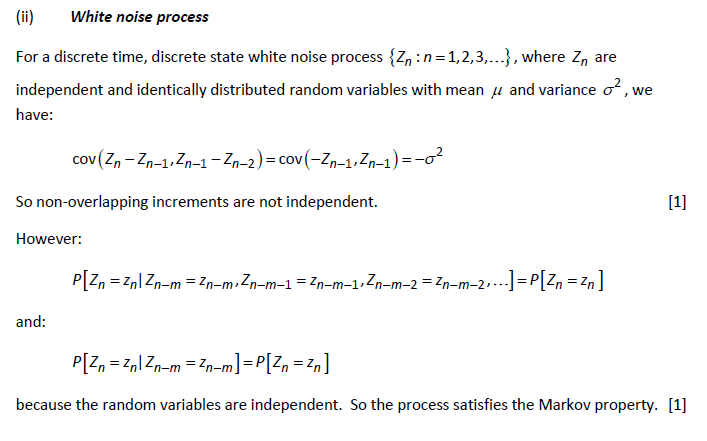
**Differentiate: Stochastic and Deterministic Model (April 2014 and April 2013)**

* A stochastic model is one which recognises the random nature of the input components, whereas a deterministic model does not contain any random components.
* In a stochastic model the output of each run is one value from a distribution. By contrast, in a deterministic model, the output is determined once the set of fixed inputs and the relationships between them have been defined.
* Correlations can be important in stochastic models as they indicate when the behaviour of one variable is associated with that of another.
* Stochastic models are more complex and more difficult to interpret than deterministic models and so require more expertise, expense and computer power.
* In a stochastic model, several independent runs are required for each set of inputs so that statistical theory can be used to help study the implications of a set of inputs. A deterministic model only requires one run.
* Running a stochastic model many times will produce a distribution of results for possible scenarios, whereas a deterministic model will produce results for a single scenario. Thus a deterministic model can be seen as a special case of a stochastic model.
* For many stochastic models, it is necessary to use numerical approximations in order to integrate functions or solve differential equations. The results for a deterministic model can often be obtained by direct calculations.
* Monte Carlo simulation is an example of a stochastic model: a collection of deterministic models each with an associated probability.
* The results of stochastic models often require Monte Carlo simulation, although some stochastic models can have an analytical solution.

**Counting Process**

* A counting process is a stochastic process, X, in discrete or continuous time, whose state space S is the collection of natural numbers {0, 1, 2…} with the property that X (t) is a non-decreasing function of t.
* E.g. number of wickets or balls in a cricket match.
* Discrete State space = collection of natural numbers {0, 1, 2…}
* Discrete or Continuous Time set

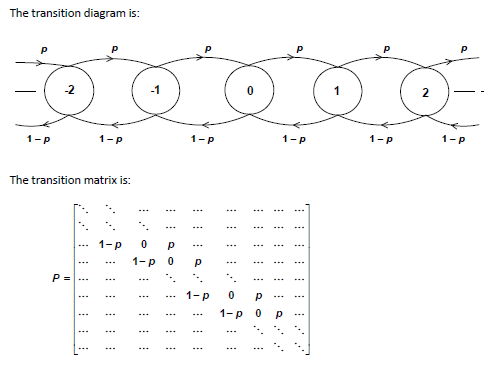
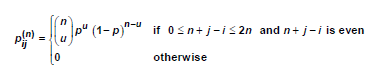
**White Noise Process**

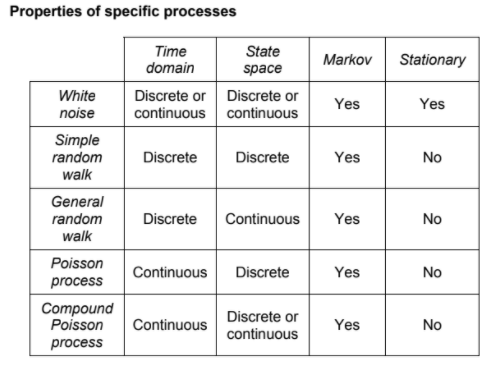
* White noise is a stochastic process that consists of a set of independent and identically distributed random variables. The random variables can either ne discrete or continuous.
* State space and Time Set can be either Continuous or discrete
* Markov Property holds in a trivial way.
* White Noise **does not have independent increments,** the markov definitions is satisfied since the future development of the process is completely independent of its past.
* The process is stationary if and only if all the random variables have the same distribution i.e. IID random variables. Such a process is also called discrete time white noise.
* Non-overlapping increments are not independent but the process satisfies markov property. Proof -

**Increments**

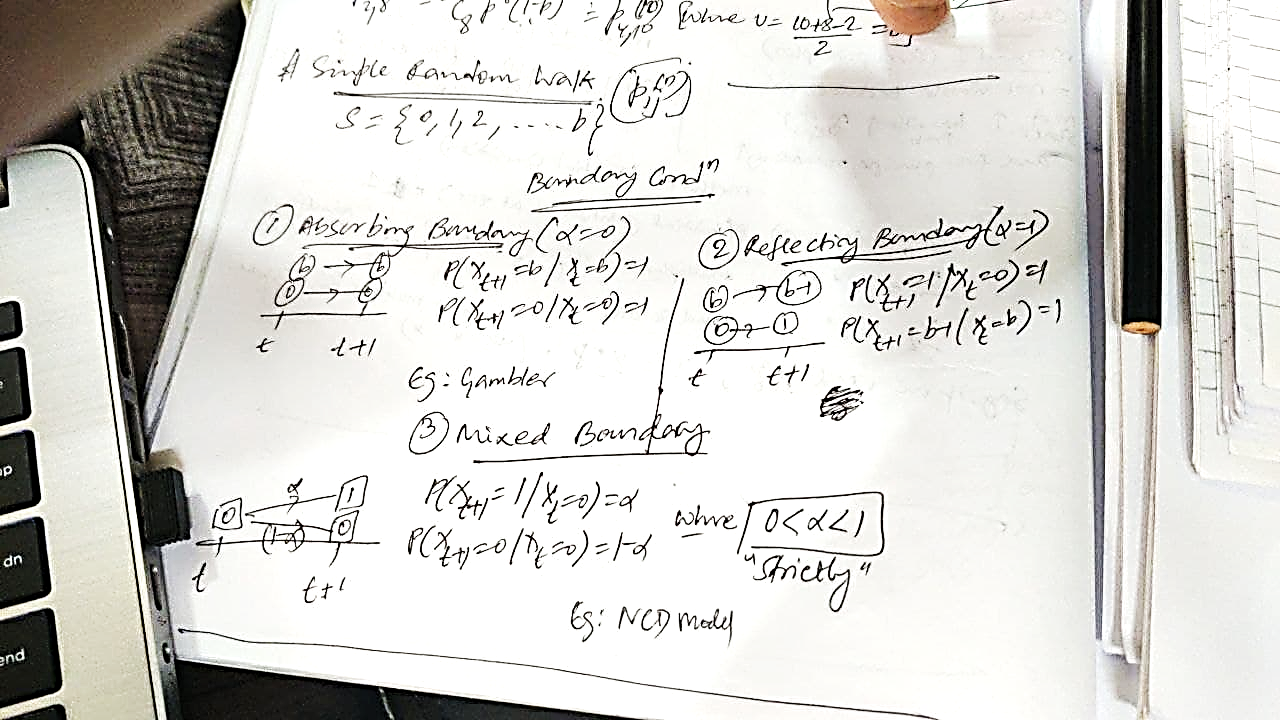
Increment of a process is the amount by which its value changes over a period of time. Given by X\_t+u – X\_t where u>0

**Random Walk**

* Definition: Xn = Xn-1 + Yn where, n = 1,2,3… with X0 = 0
* Stationary Independent Increments i.e. markov property holds
* It is a discrete-time Markov process.
* Not Weakly Stationary as mean and variance are both proportional to time t.
* State Space = Set of Integers = {…-2,-1, 0, 1, 2…}
* Time Domain = {0, 1, 2, 3…}
* When Yj = +/-1 with probability 0.5, the process is called **Simple Symmetric Random Walk** with pij (n) = 0.5
* Chapman Kolmogrov Equation for Simple Random walk = 0.5 (Assignment Question 1)
* In addition to being time homogeneous, a simple random walk is also space-homogeneous: pij (n) = p i + r j + r (n)
* **Irreducible - yes**
* **Period** of every state =2
* **Stationary distribution -** does not exist as infinite state space
* Transition Graph and matrix –
* **Distribution of the process (Sept 2019)**
  + **tCr \* p^(t-r) \* (1-p) ^r**
  + **r = 0, 1, 2, …t**
  + **When sample size tends to infinity,** distribution becomes Normal with parameters mean = N\*(2\*p-1) and Variance=4\*N\*p\*(1-p) where N is sample size
* **Express for pij (n)**
* **Practical Application ­–** e.g. number of cars in a car park controlled by a single entry/exit barrier measured after each time the barrier goes up OR how much a gambler has won or lost if he wins or loses £1 on every bet.



**Boundary Conditions**



**Impact of boundary condition on pij (n)**

* Absorbing boundary – becomes harder to reach state j, pij (n) falls
* Reflecting boundary – pij (n) rises

**Sample Path**

Joint realisation of the random variable Xt for all times ‘t’ in time set ‘J' is called a sample path of the process; this is a function from J to S.

**Strict Stationary**

* A stochastic process is said to be ‘stationery’ or ‘strictly stationery’, if the joint distribution of X\_t 1, X\_t 2, …, X\_t n and X\_t 1+ k, X\_t 2+ k, …, X\_t n + k in J and all integers n.

This means that the statistical properties of the process remain unchanged as time elapses.

* Strict stationery is a stringent requirement - very difficult to test fully and hence Weak Stationery is in use.
* E (Xt) is constant
* Cov (Xt, X t+k) is independent of time ‘t’ and dependent on lag ‘k’
* V (Xt) is constant for lag ‘k’ = 0

**Independent Increments**

A process X\_t is said to have independent increments if for all t and every u > 0, the increment X\_t+u – X\_t is independent of all past of the process {X\_s: 0<= s <=t}

**Markov Property**

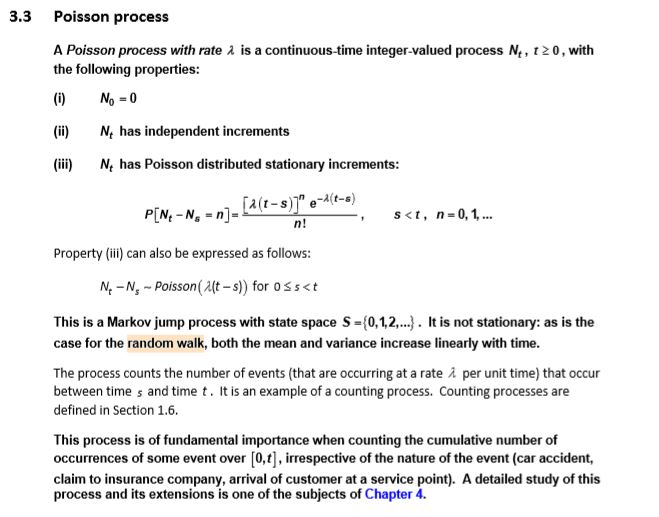


A process is said to have Markov Property if the future evolution of the process can be completely understood by the knowledge of current state alone and all the past information becomes irrelevant.

**Markov Result**

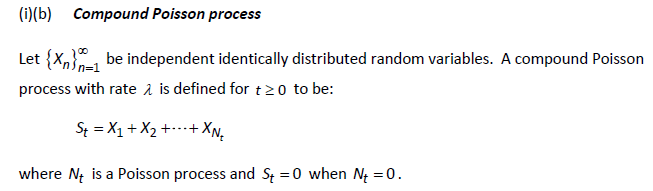
The process with independent increments has the Markov Property. If a process does not have Markov Property, it does not have independent increments.

**Poisson Process**

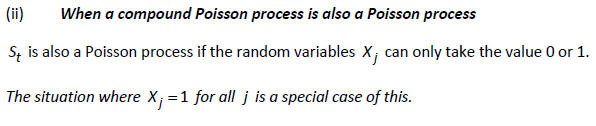


* Markov Jump Process
* State Space = {0, 1, 2…}
* Not Stationary as mean and variance increase linearly with time
* This is an example of counting process

**Compound Poisson Process**

* **Definition -**
* Has independent increments
* Markov property holds
* Serves as a model for cumulative claim amount reaching an insurance company during [0, t]: N\_t is the total number of claims over the period and Yj is the amount of jth claim.
* A common application is in estimating probability of ruin.

**When is a Compound Poisson process also a Poisson process?**



**Compound Poisson process has Markov property? Comment on Stationarity.**

* Having independent increments is a part of the definition of the compound Poisson process, thus, markov property holds.
* E (St) depends on time t, thus the mean of the process is not constant. As a result, it is not weakly stationary.

**IS using Compound Poisson Model for Cumulative claim amount reasonable?**

* This is consistent with insurance claims, since we would only expect the cumulative insurance claims by time *t* to depend on the most recently known value.
* Also, cumulative claims generally increase with time. This fact is reflected in the Compound Poisson model as E(St) depends on time and the process is not stationary. This would be a constant only in the trivial case where the individual claim amounts are £0.

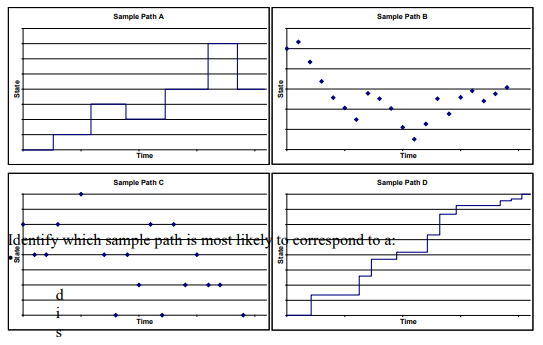
**Examples of compound Poisson process with continuous state space that the estate agent could use include: (Akshay Sir Mock 2)**

* Total cost of all houses sold
* Total estate agent commission collected.

|  |  |  |
| --- | --- | --- |
| **Situation to be modelled** | **Type of Poisson Process** | **Reason (2016 April Q5)** |
| Number of claims for motorcycle accidents received by insurer’s telephone claim line | Time Inhomogeneous | Accidents are more  likely to occur at certain times of year/week and,  likely to occur singly |
| Number of bagels sold | Time Inhomogeneous Compound Poisson | Customer arrival varies with time of day and might follow Poisson Process, then Number of bagels sold will follow compound Poisson as one customer may buy multiple bagels |
| Number of breakdowns of freezers in large supermarket | Time homogeneous | Freezers will be needed to left on continuously and no reason to expect failure at a particular time of day.  Freezers are likely to breakdown individually. |
| Cost of wasted food due to freezer breakdown | Time homogeneous Compound Poisson | If number of failures is a time homogeneous Poisson process, the cost of each failure will vary depending upon food stored in each freezer, repair time and so on.  Hence, cost would follow Compound Poisson |

**Factors considered when deciding – time set J is continuous or discrete**

* Whether outputs from the model are only required at discrete points in time.
* The objectives of the modelling
* The accuracy required
* The nature of input data (which may override the nature of the process).
* The expertise of the analyst.
* Time, cost, IT resources
* The nature of previous model
* If simulation is required it may be easier to make the time step discrete
* Continuous time models are ultimately more flexible than discrete time models.
* Some results for continuous time models cannot be obtained by discrete simulation at all.
* Regulatory requirements.
* The need to explain the model to a non-technical audience.



* Sample Path A – state space **discrete**, time set **continuous**
* Sample Path B – state space **continuous**, time set **discrete**
* Sample Path C – **discrete** state space, **discrete** time set
* Sample Path D – **continuous** state space, **continuous** time set

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of process** | **Statistical Model** | **Problem of Relevance**  **(Food retailer)** | **Problem of Relevance (General Insurer/Actuary)** |
| **S** discrete  **J** discrete | Simple Random Walk,  Counting Process, Markov Chain, Markov Jump Chain | Whether or not particular product out of stock at the end of each day | NCD model |
| **S** discrete  **J** continuous | Counting process, Markov Jump, Poisson Process, Compound Poisson Process | Rate of arrival of customers in shop | Number of claims received monitored continuously  **OR**  HSD Model |
| **S** continuous  **J** discrete | White Noise,  Time Series,  General Random Walk | Value of goods in stock at the end of each day | Total amount insured on a certain type of policy valued at the end of each month  **OR**  Company share price at end of each day or inflation rate |
| **S** continuous  **J** continuous | Compound Poisson Process,  White Noise,  Brownian Motion,  lot Process | Volume/value of trade in shop over a continuous period of time | Value of claims arriving monitored continuously  **OR**  Cumulative claim amount incurred on a portfolio of policies up to time ‘t’ |

## Ch-9 Markov Chain

**Markov chain** is a Markov process with discrete S and J.

**Akshay Sir mock 3 – Why Xn can be modelled as a Markov Chain? (big town and little town)**

* **Define Markov Chain –** A stochastic process that has a discrete state space and a discrete time domain and satisfies Markov Property.
* **State How Process Has Discrete State Space –** state space S = {A ,B} thus, the process has a discrete state space.
* **State How Process Has Discrete Time Domain –** time is measured in terms of journeys. Thus, the time domain is {1,2,3…} which is discrete.
* **State How Markov Property Is Satisfied –** the probability of being in any state depends only on where the transition starts and not on any previous transitions. Thus, with knowledge of current state, we can state the entire future evolution of the chain with respective probability.

**Akshay sir Mock 3 - Explain why this Markov Chain has a unique stationary distribution and why the process will settle down to this distribution in the long run.**

* **Finite state space –** S = {A, B} indicate that the process has a finite state space – thus a stationary distribution exists.
* **Irreducibility –** the chain is irreducible as pij(n) >0 for any state (ij) and integer ‘n’. this indicates that the stationary distribution is unique. **Finite state space + Irreducibility** implies existence of a unique stationary distribution.
* **Aperiodic –** Every state has an arrow back to itself thus, both states and entire chain is aperiodic. **Finite state space + Irreducibility+ Aperiodic** implies that the process will settle down to the unique stationary distribution in the long run.

**Transition Probability Matrix (TPM)**

A square matrix (always) of conditional probability (every row) of markov process from time ‘t’ to ‘t + 1’

**Time homogeneous**

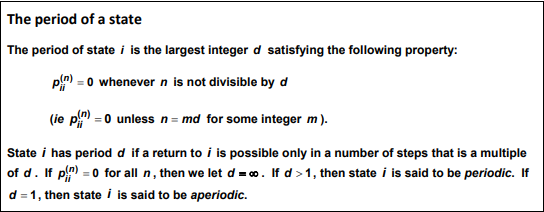
When the probability of the process does not change with time and depends only on lag.

**Irreducibility**

* A markov chain is said to be irreducible if any state ‘j’ can be reached from any state ‘i’.
* In other words, a chain is irreducible if given any pairs of state (i, j) there exists an integer ‘n’ with pij (n) > 0.

**Periodicity**

* If a markov chain is irreducible all its states have the same period.
* A state in a markov chain is periodic with period d>1 if a return to that state is possible only in a number of steps that is a multiple of d. if there is no such d>1, then the state is aperiodic. A Markov chain has period ‘d’ if all states have period ‘d’.



**Stationary distribution**

* Definition – let ‘S’ be the state space. We say that {πj given πj belongs to S} is a stationary probability distribution for a Markov Chain with transition matrix P if the following hold for all j belongs to S:
* πP= π, sum(πj) = 1 and πj >=0 (Assignment Q 31)
* Note - it does not mean that transitions have stopped taking place, it is simply the number of outgoing transitions from a particular state are offset by equal number of transactions incoming to that particular state.
* Explain why π must satisfy**: πM = (1 0 0 0 0 0 0)** where **M =P-I** but with the first column replaced with 1s. **(Y1.1 Assignment)**
  + The 7 simultaneous equations produced from solving πP = π (7x7 TPM), are linearly dependent so that any 6 of them will rearrange to produce the 7th.
  + The equation to discard would be the one with the most non‐zero numbers in it.  Looking at the matrix, this will be the one arising from the first column.
  + The reason the equations are intrinsically linked in this way is that they are arrangements of probabilities, which we know sum to 1.  This gives us the 7th equation:  π1+ π2+ π3+ π4+ π5+ π6+ π7=1
  + Rearranging the above equation, we get: πP-π=0 i.e. π(P-I)=0
  + We can then replace the first column with the πi’s and the first element of the RHS with a 1.  This effectively substitutes in the fact that: π1+ π2+ π3+ π4+ π5+ π6+ π7=1

**Estimating transition probability**

* pij-hat = ij / in = number of transitions from state i to j / number of transitions out of state i

**Differentiate: Time-homogenous and Time-Inhomogeneous Markov Chain**

* In a time-homogeneous Markov chain the transition probabilities are time independent.
* In a time-inhomogeneous Markov chain the transition probabilities depend on the absolute values of time, not just the time difference.
* Examples: time-homogeneous – no claims discount system in which the probability of a claim in each year is constant.
* Time-inhomogeneous – no claims discount system in which accident probabilities reflect changing traffic conditions from one year to the next.

**Two State TPM – Stationary distribution – form – Akshay Sir Mock 1**

1 2

TPM = 1 p 1-p

2 1-q q

Stationary distribution π = ( , )

**Irreducible and periodicity depends on β (Specimen 2019)**

**A B C**

**A 1-β – β2  β β2**

**TPM = B β 1- 3β 2β**

**C β2 β 1-β – β2**

* If β > 0, then it can reach any other state. So, it is irreducible.
* Every state has an arrow back to itself. thus, aperiodic.
* If β =0, it can never leave its current state – reducible.

**Appropriateness of NCD Model**

1. Marketing and competitive pressures
2. Encourages careful driving to preserve their discount

**Alternative to NCD Model**

* 2 state NCD model based on previous year’s claim
* Constant probability of accident is highly unlikely.
* It is reasonable to assume that probability of having an accident depends on the number of previous accidents.
  + Perhaps more accidents in winter when driving conditions are worse, or in summer when mileage is higher.
* Also, effect on previous accident should wear off eventually. Thus, give more weightage to recent years.
* It is also unreasonable to assume “at most one claim” per month/year
* Use of stationary distribution might not be appropriate.
* Other factors affecting probability –
  + Weather conditions affect chances of accident.
  + Age
  + Sex
  + Car size/age
  + Occupation
  + Geographical location
  + Marital status
  + Location-wise demand
  + Parking space availability for cars
  + Market competition
  + Natural calamity
  + National trend changes
  + Consumer behaviour depending upon discount and purchase volume

**“Protected” NCD Scheme – Making stay at 60% final (Assignment Q 34)**

* The 60% discount level becomes an absorbing state and so it is no longer irreducible.
* However, it is still aperiodic because you cannot get out of the absorbing state 60% and the other states will have no period.
* The process would now be stationary when all drivers are in the absorbing 60% level i.e. π = (0,0,0,1)

**Accident Proneness Model - Comment on assumption that transition rates are time inhomogeneous (Assignment Q 28)**

* Reasonable assumption as number of previous accidents affects the number of future accidents.
* However, the impact must wear off with time.
* **Suggestion:** give more weightage to recent number of accidents.

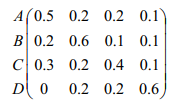
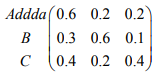
**Credit Rating sum – Downgrade trigger strategy – Practise question**

* Earlier expected number of defaults was 6.26 after 2 years for 100 companies. The downgrade trigger strategy reduced this figure to 5.91. So, there is a fall in the expected number of defaults.
* However, the return on the portfolio will also be a function of the yields on the debt. Companies rated B are likely to have bonds with a higher yield (because of the higher risk), so excluding these may in fact reduce the yield on the portfolio.
* Also, the actual number of defaults may not match the expected number. The return depends on the actual progress of the portfolio, rather than the expected outcome.
* There will also be a cost incurred when buying and selling bonds.

**Farmer Giles’s assumption about customer behaviour based on stationary matrix may be invalid –**

* Drought – could drive down production low enough to be bought by one customer
* Market demand – may change drastically
* Market competition – heavy discounts by opponents may drive down business.
* Customer behaviour – might change as per discount available and purchase volume
* National trend changes
* Field space is limited so constant number of increase is unlikely (60% likelihood of increasing order irrespective of discount level)

**2016 April Q10 – 4 insurance company A, B, C, D – A and D merge to form “Addda” – Assuming that customer base on Company D and A will merge. Comment on this Assumption.**

* Original TPM =  and New TPM = 
* We see that transition probability from D to A is “Zero” i.e. there must exist reasons why Company D customers do not want to use Company A.
* Possible reasons could be Company A’s pricing strategy and/or customer service.
* So, if the new merged company improves on these two grounds, there could be a significant change in probabilities of customers switching to other companies.
* Reduction in competition could encourage new entrant in the sector.

**Sports game involving a ball – markov assumption comment (CS2B)**

The markov property means that the probability from one score to another are independent of the paths through each game. For e.g. probability of going to state (2,1) to (3,1) is ‘q’ regardless of whether state (2,1) was reached from state (2,0) or state (1,1). This assumption is questionable because there may be some kind of momentum for the player who won the previous point. Psychology plays an important role in sports at any level.

Time homogenous assumption says that probability of scoring a point is the same throughout the game. This is assumption is questionable because – (1)Psychology of player is an important factor, (2)different players tire at different rate (3)concentration level (4)true probabilities are influenced by time of year the match is played.

## Ch-10 Two State Model and Poisson Model

**Three Assumptions underlying a simple two state model**

* Assumption 1

The probabilities that a life at any given age will be found in either state at any subsequent age depend only on the ages involved and, on the state, currently occupied. This is the Markov assumption.

* Assumption 2



* Assumption 3



**Differentiate: Two-state and the Poisson model**

* The Poisson model can be considered to be an approximation to the two-state model.
* The Poisson model is not exact since it allows a non-zero probability of more than N deaths, where N is the total number of lives involved in the investigation. However, since this probability is usually negligible, the Poisson model often provides a good approximation.
* The estimation of the transition rates in the two-state model involves the measurement of two random variables – the observed number of decrements and the exposed to risk that gave rise to these decrements.
* The Poisson model assumes that the exposed to risk remains constant and estimation of the transition rates in the model only involves the measurement of the observed number of decrements.

## Ch-11 Graduation

*Graduation is an art.*

**Graduation** refers to the process of using statistical techniques to improve the estimates provided by the crude rates. It refers to smoothing of crude rates. The aim of graduation is to produce a smooth set of rates that are suitable for a particular purpose, to reduce random sampling error and to use the information available from adjacent ages to improve the reliability of estimates.

**Need for graduation?**

The crude estimates calculated from the mortality investigation will contain sampling errors and would be unlikely to progress smoothly with age, whereas the true underlying rates would be expected to be smooth.

If the insurer calculates the premiums using the crude mortality rates, this may produce irregularities in the premium rates and these would be hard to justify in practise.

In addition, the insurer would always consider the suitability of rates it uses for the required purpose.

For e.g. it may wish to take account of changes occurring in mortality since the investigation period, or to build in some caution regarding the expected future experience e.g. so as not to underestimate future mortality for life assurance premium calculated

**Why test graduated rates?**

Every graduation is a trade-off between smoothness and adherence to data. We want the graduated rates to be smooth, but we also want them to be representative of the data from which they were derived. We need to make sure that the graduated rates exhibit both of these characteristics before we can use them in financial calculations.

**Reasons –**

1. **Theoretical argument** – at the heart of our desire to graduate is the intuitive idea that quantities such as qx or μx should be smooth functions of age. There is some evidence from large investigations to support this but it is nevertheless an assumption.
2. **Practical argument** – we use life tables to compute financial quantities (such as premiums and reserves). It is very desirable that such quantities progress smoothly with age. Since, irregularities are hard to justify.

**Disadvantage/Limitations –** What graduation cannot do is remove any bias in the data arising from faulty data collection or otherwise.

**Aim of graduation**

1. Produce smooth set of rates
2. Reduce random sampling error
3. Use information available from adjacent ages

**Desirable features of graduated rates**

1. Smoothness
2. Adherence to data
3. Suitability for the purpose in hand

Note - Smoothness and adherence to data are conflicting requirements perfectly smooth curves/estimates have little to no adherence to data and vice-versa. Thus, we need to strike a balance between the two.

**Concerns for suitability –**

1. **Objective –** 
   1. National rates for social cause – max adherence to data or if already established standard tables are available.
   2. Life insurance companies – loss arises from premature deaths thus, never underestimate mortality.
   3. Pension funds paying annuity – loss arises from delayed deaths thus, never overestimate mortality.
2. **Quantity of data**
3. **Quality of data**
4. **Other Considerations (factual information)**
5. Mortality of males is higher than that of females.
6. Mortality of a person with life insurance policies are lower than that of the population as a whole.
7. Select mortality – mortality of persons who have recently taken out life insurance is lower than that of persons who took the policy a long time ago.
8. Mortality improves over time.

**Why CMI Investigations are based on the number of policies/**

For the pooled mortality investigation, it was decided only to require information based on number of policies because:

* All companies maintain data in this form in order to compile their statutory accounts, so the data are readily available.
* Data based on lives are not required for this purpose, so not all insurers would have the data readily in this form
* When pooling data from different insurance companies, it would be very difficult to identify people who had policies with more than one company.
* If based on lives, many insurers would have to compile data just for the pooled mortality investigation, so might opt out instead – however, the investigators want to maximise the number of participants in order to maximise the credibility of the results.

**Testing for smoothness**

The criteria that will be used to check the smoothness is based on 3rd order of difference which should be –

* Small in magnitude compared with qx 0/ qx s
* Progress with age

**Chi-Square Test – Weakness**

1. There could be a few large deviations offset by a lot of very small deviations. This is in essence as the chi-square statistic summaries a lot of information in a single figure.

**[ISD TEST]**

1. Graduation might be biased above/below the data by a small amount. Chi-square statistic can fail to detect consistent bias if it is small.

**[Signs test and CDT]**

One test that does not require making any assumption while testing for mortality experience – Signs Test **(IAI 2018)**

1. Even if graduation is not biased as a whole, there could be significant groups of consecutive ages – runs or clumps over which it is biased up/down. **[Grouping of signs and Serial Correlation test]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Basis** | **Individual Standardised test** | **Signs Test** | **Cumulative Derivatives Test** | **Grouping of Signs test** | **Serial Correlation** |
| **H0** | Zx ~ N(0,1) | P~B (m, 0.5) |  |  |  |
| **Tail** |  | Two tailed | Two Tailed | Left Tailed | Right Tailed |
| **Weakness** | Not rigidly defined and subjective approach | Ignores magnitude of deviations. So, test can be cleared even by large deviations. It does not look at the pattern of occurrence of deviations i.e. clumping goes undetected. | If the constraint sum:(Oi)  = sum:(Ei) is applied then the test becomes redundant. |  |  |
| **Strength** | Effective against defects A and B | Can detect overall bias and is rigidly defined. | Detects overall bias or long-run deviations/ clumps of the same signs. | Detects Clumping of deviation | Can detect clumping |
| **Note** | 4 points of informal test - overall shape should be symmetrical, middle most 50% data should be between (-0.6745,0.6745), no outliers, symmetry 50% data values are positive | Pal>LS we cannot reject H0  Pal<LS we can reject H0 |  | 3 ways to do the test –  Pal calculation  Critical value from tables  Normal approximation |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Basis** | **Parametric formula** | **Reference to standard tables** | **Graphical graduation** | **Spline Functions** |
| **Use** | Good quality  Good quantity  Precise result | Good quality  Poor quantity  Precise result | Less quantity  Inaccurate result for quick comparison | We need to define multiple formulas for different age-ranges |
| **Step1** | Select graduation formula | Choose appropriate table | Plot crude estimates | Identify ages at which we are choosing knots |
| **Step2** | Estimating parameters | Find simple link to standard table | Mark confidence intervals | Choose appropriate spline function |
| **Step3** | Calculate graduated rates | Estimating parameters | Sketch graph | Estimating parameters |
| **Step4** | Testing | Calculate graduated rates | Read off values | Calculate graduated rates |
| **Step5** |  | Testing | Testing | Testing |
| **Step6** |  |  | Redraw if necessary |  |
| **Step7** |  |  | Hand-polishing the graduated rates |  |

**Degree of Freedom for Chi-square test –**

1. Parametric

– (age groups – number of parameters estimated)

1. Standard Tables

– lose one degree for each fitted parameter and some indeterminate number because of constraints imposed by choice of standard table.

1. Graphical

– lose 2 or 3 degrees of freedom for every 10 or so age groups fitted. (height, slope and curvature)

1. Spline functions

– as per number of parameters estimated.

**Spline Functions**

These are polynomial of a specified degree which is defined on a piece-wise basis across the age range. These pieces join together at knots, where certain continuity conditions for the functions themselves and their derivatives are required.

**Factors when choosing the knots**

1. The knots allow the spline to have a different shape on either side of the join.
2. So knots should be placed near to points on the curve where there is a significant change in the shape.
3. When choosing the number of knots, there is a trade-off between goodness of fit and overfitting.
4. The number of knots should be enough to capture all the different shapes present in the data i.e. to achieve a good fit
5. However, the number of knots should not be excessive, causing the spline to follow random features present in the data i.e. overfitting
6. The number of knots is usually chosen based on past experience of the optimal number determined for similar exercises.
7. A less subjective method is to include the number of knots as a ‘penalty’ in the function used to fit the parameters, in order to obtain the optimal trade-off.

**Conditions at knots –**

1. Continuous function
2. No steep turns i.e. 1st derivative is continuous
3. No change in curvature i.e. 2nd derivative is continuous

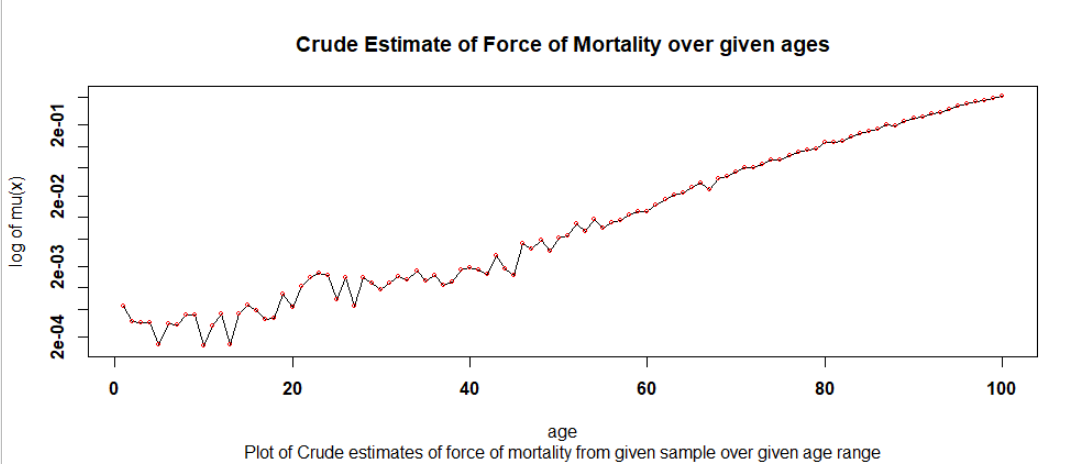
**Determining parameter value**

1. The parameter values can be determined using a least squares approach where the parameters are chosen to minimise the sum of the squared deviations between the values of the spline function and the original data values.
2. Weights can be included to give more emphasis to certain ranges of the function, e.g. to prioritise the ages that have the most financial significance in a graduation exercise.
3. An alternative method is to minimise the value of the chi-square statistic.

**Q. An investigation – mortality rates of males – age under 30 – deprived area of UK – graduated using spline method. Explain (Practise Question)**

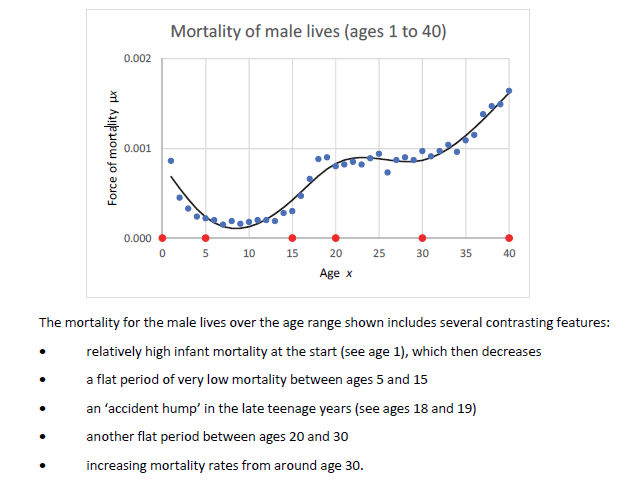
The pattern of mortality rates shows significant changes in shape over the age range of 0 to 25. From birth up to the childhood years the mortality remains low throughout, there is an increase in late teen years (known as the accidental hump), and falls again through the early mid-20s. By choosing appropriate knots, we can produce a spline function that captures these features.

The spline method works well with fairly small datasets as wells as large samples. We are not told the size of the population here, but that is not an issue when using the spline method (as long as the population is not very small).

**Akshay Sir’s R Mock 3 – Graduation – Comment on Graph with respect to known trend in human mortality**

* There is a high initial level of mortality associated with infants. We see that there is a decrease in the mortality rates over the first few ages.
* There appears to be a spike in the mortality rates around the late teens and early twenties, which may relate to the accidental hump.
* The logged mortality rates appear reasonably linear around the middle ages (age 50 onwards), which is consistent with exponential mortality rates in old age
* The crude estimates do not produce a smooth graph as they completely adhere to data
* Thus, we see falls and spike at random ages which could either be due to lack of data or sampling errors
* Overall increasing trend can be seen clearly which conforms to the fact that human mortality increases with age
* The graph also shows that there are a number of changes in shape in some short ranges of ages that the graduation does not appear to capture. This could be indicative of over graduation or sampling error, particularly as there are so few deaths at these younger ages.
* The outcome of serial correlation test at lag 2 was to reject H0, which supports this possibility of over graduation.

**ActEd – ELT15 Males using Spline**

The mortality for the male lives over the age range shown includes several contrasting features:

* Relatively high infant mortality at the start which then decreases
* A flat period of very low mortality between ages 5 and 15
* An ‘accidental hump’ in the late teenage years (ages 18 and 19)
* Another flat period between ages 20 and 30
* Increasing mortality rates from around age 30

**Suitability of 2 methods of graduation used by insurance company for premium calculation –**

* **Reference to standard tables –** depends on the availability of appropriate standard tables. Since we are trying to calculate graduated mortality rates for a group of impaired lives, an appropriate standard table may be difficult to find.
* **Spline function -** this is suitable when dealing with small dataset. However, it is not suitable for very small experiences with scanty data at many ages.

**When is it appropriate to use parametric formula?**

* 1. A suitable mathematical formula can be found that can describe mortality rates adequately over the entire age range of interest.
  2. The expected number of deaths is great enough at all ages to give reliable answers
  3. The data values can be considered to be complete and accurate, and they are adequately subdivided with respect to age, sex and other relevant categories.
  4. An analytic method or computer software that can determine the optimal parameter values is available.

**Advantage of Parametric Formula**

1. Graduated rates will progress smoothly provided number of parameters is small.
2. Good measure to produce standard tables.
3. Can easily be extended to more complex formula, provided optimisation can be achieved.
4. Can fit the same formula to different experiences and compare parameter values to highlight differences between them.

**Disadvantage of Parametric Formula**

1. It can be hard to find a formula to fit well at ages without having lots of parameters.
2. Fit is bound to be best at ages where we have lots of data, and can often be poor at extreme ages.

**Advantage of Reference to Standard Table**

1. It can be used to for relatively small data sets in cases where a suitable standard table exists.
2. The graduated rates should be smooth provided that a simple function is used.
3. The standard table can provide information at extreme ages where data may be scanty.
4. The shape of the table can be used to ‘fill in gaps’ in the data.
5. It can be useful to fit the same table to several experiences with same link function to give insight into how the experience differs over time.

**Advantage Of Graphical Graduation**

1. It can be used for scanty data where no suitable standard table exists.
2. No more sophisticated method is justifiable
3. It enables an experienced analyst to allow for known (or likely) features of the data.
4. It can give a quick initial feel for the rates.

**Considerations - Choosing An Appropriate Standard Table**

* Must be available for all classes of lives, e.g. males and females.
* Must relate to a similar class of lives, e.g. assurances and not annuities in this case
* Must be a ‘benchmark’ table, i.e. generally acceptable
* Should be up-to date i.e. relate to fairly recent experience
* Must cover the age range for which rates are required.
* In addition, it should have the correct pattern of rates by age (not necessarily the correct level of rates though).
* It should not have any special features that are unlikely to be present in the experience being graduated.

**Why an Insurance company - compare results of mortality investigation with previous experience?**

* If the previous experience is the recent experience of the policyholders of a life insurance company, the comparison could be important for pricing life insurance contracts.
* Mortality rates are expected to change over time due to for example, improved medical processes or change in the mix of the population.
* It helps to validate the results of the investigation.
* It is can indicate whether the office’s experience is out of line with the population as a whole.
* Unexpected changes in mortality may have an impact on the underwriting process.
* It is important for the company to know whether the investigation’s results are consistent with published life tables, especially if the company plans to use published tables for any financial calculations.

**Reasons for change in Mortality Experience over time?**

* Advancements in the field of medical sciences, health care provisions and standard of living.
* Change in composition of lives – different mix of nationalities or number of males and females.
* General change in lifestyle or diet
* Change in Underwriting practices e.g. creation of a new category of ‘preferred lives’

**Describe how smoothness is achieved - graduation methods**

* **Parametric formula** - Rates are automatically smooth provided that a formula with sufficiently few parameters is used.
* **Graphical** - Reliance is placed on the skill of the practitioner to draw a sufficiently smooth line through the crude rates. The third differences test for smoothness is useful here. It is usually necessary to make several attempts, and to adjust the results by hand (rather than re-drawing the curve), a process called hand-polishing.
* **With reference to standard tables** - A standard table will already be smooth. Provided a link function is selected with few parameters, this smoothness should be preserved in the graduated rates.

**Dangers of over graduation –**

* Inadequate premium rates – the office may make losses through underestimating mortality for death benefits or overestimating mortality for survival benefits (since the graduated rates do not accurately reflect the true mortality rates at all ages.)
* Excessive premium rates – the reverse occurs where the office may lose business through setting excessively high premium rates.
* Selection – the office may be exposed to selection from other offices whose premium rates more accurately reflect the true mortality rates.
* Reserves – using unbiased rates can also lead to inappropriate levels of reserves being held. Holding insufficient reserves can endanger the company’s solvency, whereas holding excessive reserves will reduce the company’s profitability.

**Dangers of Under Graduation –**

1. Inappropriate premium rates – the office may makes losses or lose business if the premium rates at particular ages have been distorted by random sampling errors that were not smoothed out.
2. Anomalies – the office may lose business or incur unnecessary alteration expenses if the rates do not show a consistent progression from age to age. Policyholders may wait a few years because the rates become cheaper, or they may surrender and take out a new policy to take advantage of an anomaly in the rates at a particular age.

**Patent Comment after graduation sum (Sept 2019)**

* **Conclusion of test** – whether H0 was reject or not i.e. graduated rates are/are not true mortality rates.
* **Working of chi-square test –** is based on squared deviations telling us nothing about the direction of any bias. So H0 may not be rejected in spite of existence of bias. Thus, we perform another test to detect bias.
* **Analysis of data –** (question specific) beyond age 65, number of death falls.
  + This might have been identified by the competitors providing them with first mover advantage due to this information.
  + As mortality is lighter beyond 65, premiums should be reduced otherwise leading to loss of business.
  + 65 could be the age of retirement in the population, and those continuing to work are in better health than those retiring.
* **Lack of data -** The lack of data at older ages could be causing more volatility in the results resulting in the graduated rates diverging from the true rates.

## Ch-12 Markov Jump

**Inhomogeneous** – model parameters are time dependent or change with time.

**Generator Matrix** is a square matrix where each row is differentiation of probability with each row’s sum equal to zero. In other words, (i, j) th entry of A is μ\_ij.

**Meaning of pAA(t) (Akshay Sir Mock 2)**

pAA(t) represents the probability that the process is in state A after another t time units, given that it is now in state A.

**Poisson Process**

An event number process is said to be Poisson process if and only if the following are satisfied-

1. N (0) = 0 and N(S)<=N(T) for s<t

I.e. always start from zero, and is a non-decreasing function

1. P (N(t+h) = r+1 given N(t)=r) = λ\*h + O(h)

P (N(t+h) = r given N(t) = r) = 1 – λ\*h + O(h)

P (N(t+h) > r +1 given N(t) = r) = O(h)

1. Poisson process has independent increments i.e. the number of events in time interval **(s, t]** are independent of events up to time ‘s’

* Markov Jump Process
* State Space = {0, 1, 2…}
* Not Stationary as mean and variance increase linearly with time
* This is an example of counting process
* Distribution of increments: N\_t + s –Ns ~ Poi (λ\*t)

**Perform Serial Correlation test on Poisson process –** to check independence on successive increments/lags (2017 April)

**Problem of Thinning**

When the events in a Poisson process are of different types, each type occurring at random with a certain probability, the events of a particular type form a thinned process. The thinned process is also a Poisson process, with rate equal to the original rate multiplied by the probability for the type of event.

**Holding Time**

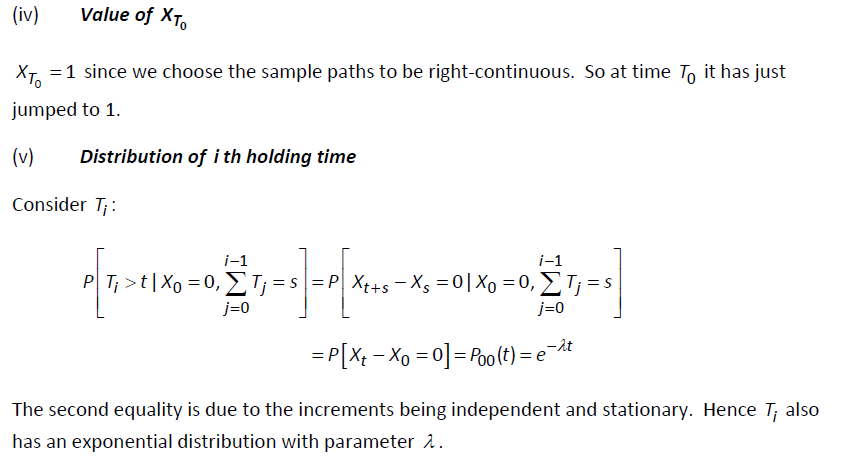
The holding times are inter-event times i.e. the time spent in a particular state between transitions. For the process given, the ith holding time Ti-1 is the time spent in state i-1 before the transition to state i.

**Distribution of first holding time ­**

The first holding time of a time-homogeneous Markov Jump process with transition rates μ\_ij is exponentially distributed with parameter lambda\_i = - μ\_ij

* **P (T\_0> t given X0=0) = P (X0=0 given X0=0) =p00(t)= exp(-λ\*t)**

Ti is a Random Variable denoting the holding time in state i or waiting time to move out of state i.

**Value of X\_T0 and Distribution of ith holding time (From Practise Questions)**

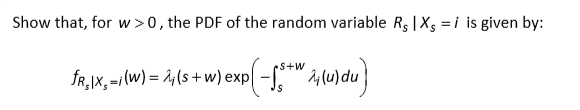
**Advantage of Two-State Model**: can be easily extended to a multiple-state model by breaking the multiple states into numerous 2 state models and then multiplying their likelihoods

**Residual Holding Time**

The residual holding time at time *s* is the random variable representing the remaining time until the next jump.



 This also gives CDF

**PDF of Residual Holding Time**

**Markov Jump Chain –** if a markov jump process is observed only at time of transitions/jump to a different state, then the process is called Markov Jump Chain. The **jump chain** is also sometimes called the **embedded chain**. It is the sequence of states that the process is observed to take. The time spent in each state is ignored.

The **forward equations are more useful** in numerical work for actuarial applications because we usually have an initial condition such as knowing that a policyholder is healthy when a policy is sold, thus we want equations that we can solve forwards in time from that starting point.

**Difference in Outcome – same transition matrix – Markov Jump Chain vs. standard Markov Chain (Sept 2007)**

* The outcome of the jump chain can only differ from that of the standard markov chain if the jump process enters an absorbing state.
* As the jump process will make no further transitions once it enters an absorbing state, the jump chain stops.
* It is possible to model the jump chain as though transitions continue to occur but the chain continues to occupy the same state.

**Markov Jump Process** is a continuous time Markov process with a discrete state space S.

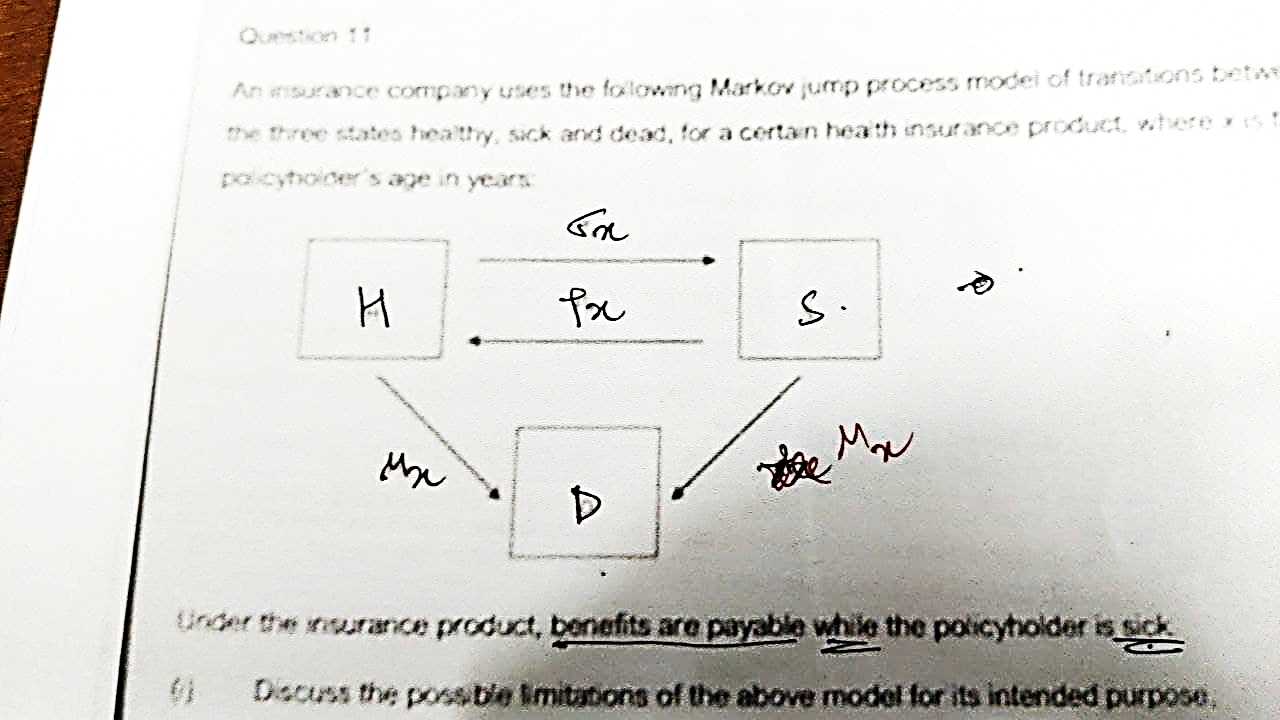
**Explain the condition needed for such a process to be time homogeneous** - In the case where the probabilities P (Xt =j given Xs=i) for i, j in S and 0 <= s <t depend only on the length of time interval t-s, the process is called time-homogeneous.

**ρ(t) should depend on holding time in Sick state (April 2006)**

The more we wait in sick state, chances of transition into dead state increases rather than healthy state. This will make the model more efficient. Empirical evidence suggests likewise. Limit of 6 month pay on sick pay will cause some durational effect around this point.

**Drawbacks** – this increases model complexity as we need higher volume of data and more parameter estimation which also increases cost of the modelling.

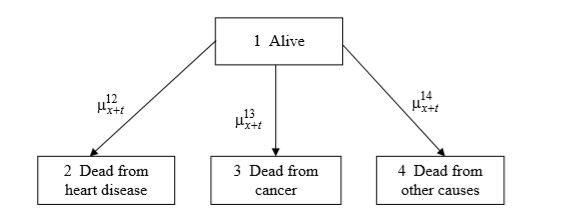
**Disadvantage/Limitation of given HSD model**

****

* The model may not allow easily for some of the policy features e.g. waiting period
* The model assumes the same mortality rate applies for both sick and healthy individuals, this is unrealistic
* The model requires the Markov assumption – that the transition rates depend only on the current state and age of the policyholder. In real life this may not be true.
* Recovery and mortality rates from sick states are likely to depend on the duration in the sick state
* The rates of transition from healthy to sick (sickness inception rates) may depend on how many times a currently healthy person has been sick in the past, and for how long they have been sick.
* Sickness inception and mortality rates of healthy people may depend on how long a policy has been in force.
* All transition rates may depend on other factors, such as sex and smoker status.
* Benefit is only payable while in sick state so we need number of transitions in and out of sick state and waiting time of sick state. Thus, we can ignore data for mortality from healthy.

**Differentiate: Time-homogeneous and Time-Inhomogeneous Poisson Process.**

The probability that an event occurs during the short time interval between t and (t + h) is approximately equal to λ (t)\*h for small h where λ (t) is called the rate of the process. For a time-inhomogeneous process, λ (t) depends on the current time t; for a time-homogeneous process it is independent of time.



**Discuss how you might use this model to analyse the impact of risk factors on the death rate from heart disease and suggest, giving reasons, and a suitable alternative model.**

Using the four-state model, the lives in the investigation would have to be stratified (divided) according to the risk factors and the transition intensities estimated separately for each stratum (division). This is likely to run into **problem of small numbers**.

**Alternative - Using a Cox regression model** with death from heart disease as the event of interest and the risk factors as covariates would avoid this problem. Lives which died from other causes could be treated as censored at the durations when they died.

**September 2009 – Q6 Modelling Complaint arrival using Poisson Model Appropriateness:**

1. Not suitable for the purpose as it gives an average approximation.
2. Does not account of complaints that have been in queue.
3. Only two employees, sickness and vacations become significant factors to be considered.
4. Model assumes claim arrival to be time-homogeneous which might not be the case for companies with seasonal business cycles.
5. Model assumes claim arrival to be independent which might not be the case if there is a problem with quality control or any other aspect affecting the business.
6. Employees can use past experience to develop a standard template to handle complaints of similar nature which affects the average time of response.
7. So, the model could be used to predict the probability of claims not being answered in a stated time i.e. failure to meet service standard instead of claim arrival.

**Suitability of Poisson process for modelling arrival times of ships (Akshay Sir Mock 1)**

* There is an assumption that the process is homogeneous over time
* However, it is likely that boat traffic will vary on a seasonal basis, i.e. by time of year, day of week, and by time of day.
* It is also likely that boat traffic will vary across a broader time scale i.e. it may vary from one year to another e.g. as a result of changes in economy.
* There is an assumption that all arrival times are independent of each other and hence that the markov property applies.
* This is unlikely in practise, particularly for the big and medium sized boats. This is because any that would arrive at more or less the same time would have to wait for each other and take their turn, as there may not be sufficient space to take the ships side by side in the canal.

**Suitability of Poisson process for modelling number of houses sold (Akshay Sir Mock 2)**

* A Poisson process Nt ~ Poi(λ\*t) is a counting process that operates in continuous time.
* This seems sensible as the number of houses sold is also a counting process and a house can be sold at any time.
* According to this model, the rate at which houses sell is a constant ‘λ’.
* This may not be realistic as:
  + Houses may sell at a faster rate in the summer months
  + Houses may sell at a faster rate during periods of economic boom.

A time-inhomogeneous model that incorporates seasonality may be more appropriate to deal with these issues.

**Describe the principle difficulties in modelling using a Markov Jump Process with time inhomogeneous rates.**

* A model with time-inhomogeneous rates has more parameters, and there may not be sufficient data available to estimate these parameters.
* Also, the solutions to Kolmogorov’s equations may not be easy (or even possible) to find analytically.
* Time-inhomogeneous processes are computationally harder to simulate.

**Overcome –**

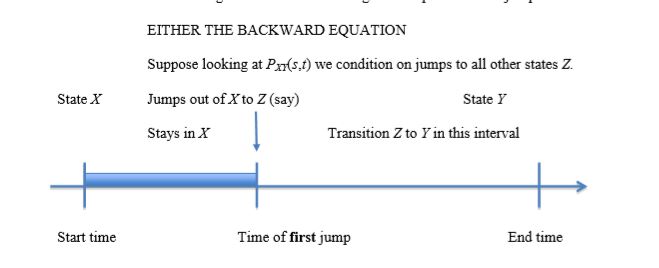
* A possible procedure is to divide the time interval into subintervals, assume that the transition rates are constant over each subinterval, and estimate the transition rates for each subinterval using the procedure described above.
* However, our estimates would be based on a much smaller amount of data, compared to the time-homogeneous case, and would be less reliable.
* Alternatively, we could select an appropriate functional form for μ\_*ij* (*t*) and use the data to estimate the relevant parameters. This is only possible if we have an idea of what kind of formula would be appropriate.

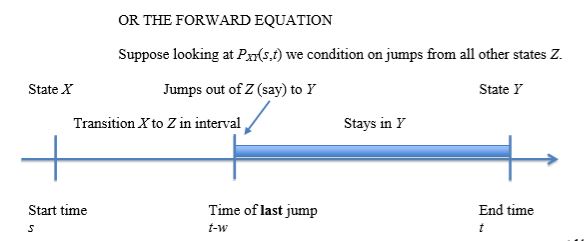
**Similarity and differentiate: Markov Chain, jump chain and Jump Process**

1. All three have discrete state space.
2. Markov chain and Jump chain operate in discrete time but Markov jump operates in continuous time.
3. All three satisfy Markov Property.
4. A Markov Jump Process X is examined only at times of its transitions; the resulting process is called the Jump Chain associated with X.
5. The Jump Chain obeys the Markov Property and behaves as a Markov Chain except when the Jump Chain encounters an absorbing state. From that time, it makes no further transitions, implying that time stops for the Jump Chain.
6. The Jump Chain associated with X takes the same path through the state space as X does. However, questions about the times taken to visit a state are likely to have different answers for X and for the Jump Chain associated with X.
7. The Markov Jump Chain and the Markov Chain are expressed in terms of probabilities whereas the Markov Jump Process is expressed in terms of rates.
8. The Markov Chain can have loops in each state, the Markov Jump process cannot and the Markov Jump Chain only has loops on absorbing states.

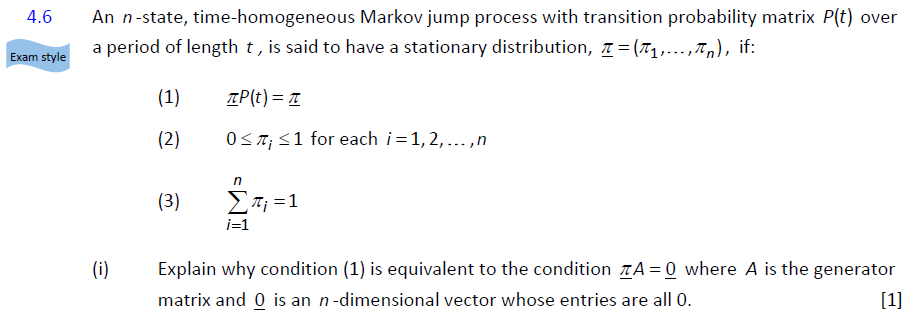
**Describe how integrated Kolmogorov equations can be constructed by conditioning on the first or the last jump, illustrating your answer with a diagram.**

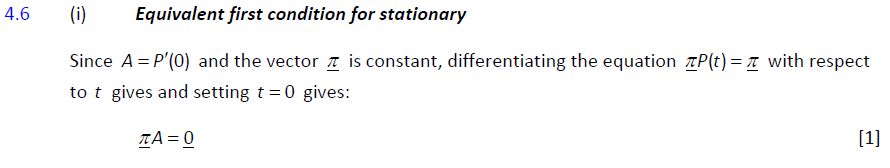
* One method of deriving probabilities for continuous time Markov processes is by integral equations.
* Using the law of total probability, we can consider the full set of possibilities for the first jump from state X or the last jump to state Y.
* For a given time of this first/last jump, the probabilities that the jump was from each state will be in proportion to transition rates at that time.
* By integrating across all possible times for the first/last jump we obtain the overall probability.
* Where a probability for being in the same state at start and end is required, an additional term is needed for the probability of remaining in the same state throughout the period i.e. no jumps.

****

****

**Equivalent first condition for stationary**

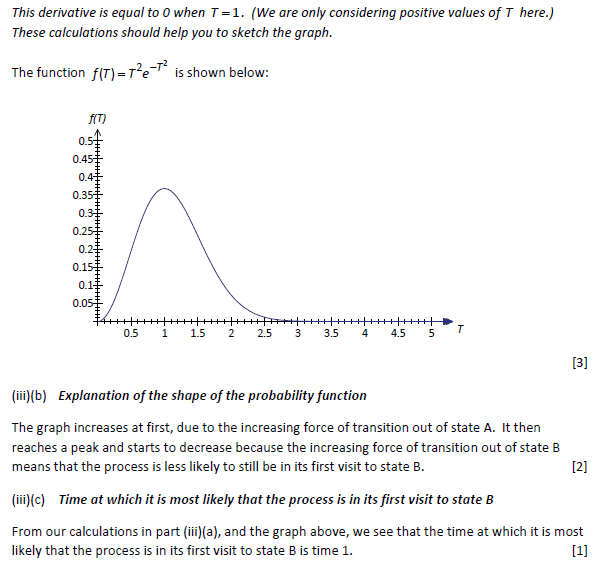




**Outline two methods for simulating the Markov Jump Process, without performing any calculations. (Sept 2007)**

* **Approximate Method**
  + Divide time into very short intervals, h, such that σ ij h is much less than 1.
  + Simulate a discrete-time Markov chain {Yn : n ≥ 0} , with transition probabilities p\*ij(h) = 𝛿ij + h\* σij
  + The jump process, Xt is given by Xt = T\_[t/h]
* **Exact Method** 
  + Simulate the jump chain as a Markov chain, with transition probabilities pij = σij λi
  + Once the path {Xn-hat: n=0,1 …} has been generated, the holding times {Ten: n= 0,1,...} are a sequence of independent exponential random variables, having parameter λ\_Xn-hat .

**Graph and Explanation – Practise Question f(T) = T2 e -T^ 2**



**Transition rate λ(t) = 3 + cos(2\*π\*t) – suitability for motor claims (Sept 2006)**

* Here, we have λ(t) = 4,3,2,3,4 for t=0,0.25,0.5,0.75,1
* Claims are seasonal with 12-month period
* Peak during winter with 4 claims
* Mid of year – lowest point with 2 claims
* Reasonable choice for λ(t) if driving conditions are worse during winter.

**Human Births follow Poisson Process – Comment**

* Unlikely to be entirely realistic
* Occurrence of multiple births i.e. twins or triplets
* Births tend to occur seasonally
* Process might be time inhomogeneous

**Commercial impact on insurance company – home insurance policy in partnership with a bank – more discount to person having more bank accounts (Sept 2020)**

* Thisarrangement could lead to significant volumes of business for insurance company through access to the bank’s customer base if the scheme is popular.
* This arrangement could lead to loss of business if there are a lot of insurance holders with many small bank accounts.
* Market competition could be a factor which the company should take into account.

**Athlete diet monitoring – Omnivore, Vegetarian and Vegan – Time Inhomogeneous Markov Jump model. Improvements –**

* Time Inhomogeneous analysis is good but it could account for (1)individual preferences changing with age and time, (2) generational trends.
* Inclusion of other factors such as (i)country (ii)type of athlete (iii)wealth (iv)food allergies (v)religion and so on...
* more states such as (i)pescatarian (ii)carnivore (iii)paleo-diet and so on...
* P(t) approximation could be more accurate by choosing smaller values of 'h'

**Opinion polls for voting populations political affiliations – Suitability of time in-homogeneous markov jump.**

* The model is critically dependent on given transition rates being accurate and correct. If these are obtained from past data, they might not be necessarily a good guide to the future.
* The methodology or calculations used to find the age-dependent transition rates may be flawed.
* The markov assumption may be questionable here. It means that the probability a voter moves affiliation from one party to another over the next 5 years, say, is independent of their voting history. In fact, you might expect voters who have staunchly remained with one party to continue to do so, or voters have changed their minds a lot in the past to continue to swap affiliations.
* The time-inhomogeneous assumption is probably a very good idea. Party political affiliation is very likely to be influenced in two ways –
  + Voters may show a tendency to move from one party to another as they age e.g. voters have been shown to become more right-wing as they get old.
  + Wildly different ages will represent different generations, who may have different attitudes. Different political parties can appeal to different generations, sometimes changing their policies precisely to do this.
* However, age is probably not the only important factor to consider here. In reality political affiliation will be influenced very strongly by socio-economic background and other factors independent of age. The model does not take this into account.
* change in political affiliation isn’t always gradual, a one-off event could spark a deluge from one party to another, e.g. the way the party in power handles the possibility of war, or the sudden perceived need for a referendum on a particular topic.
* The model assumes that the population of voters remains constant. In reality, people who has previously not voted may start to vote, voters may decide to stop voting, or become extremely ill or die so that they can’t vote. The model does not take this into account.

## Ch-13 Time Series

**Time Series**

Is defined as a stochastic process with continuous state space and discrete time domain

**Univariate time series**

A sequence of observations of a single process taken at a sequence of different times, given by {Xt: t=1, 2, 3…, n}

E.g. AR, MA, ARMA, ARIMA

**Purely In-Deterministic process**

A process is called purely In-deterministic if knowledge of values of X1, X2 …Xn is progressively less useful at predicting the value of XN as N tends to infinity.

When we talk of stationary time series process, we shall mean a weakly stationary purely in-deterministic process.

**Stationary Time Series**

* Atime series is described as “stationary” if its statistical properties do not vary over time.
* For practical purposes, it is sufficient for a series to be weakly stationary, which requires its first two moments to be constant values & the covariance depends only on the lag, not on time t.
* Stationarity is an issue relating only to the autoregressive terms, and is not affected by adding or subtracting constants.

**Invertibility**

A time series Xt is said to be invertible if we can express the white noise terms et as a convergent sum of the Xt terms.

**Integrated Process I (d)**

An I (d)is a process which is not originally stationary but becomes stationary after differencing ‘d’ times.

**Multivariate Time Series**

An ‘m’ dimensional multivariate time series Xt is defined as a column vector with values of ‘m’ different processes at time t.

E.g. VAR

**Co-integrated time series**

Two time series processes X and Y are co-integrated if:

1. X and Y are I (1) processes.
2. There exists a non-zero vector α, β such that α\*X+ β\*Y is stationary.

**When it is reasonable to expect that 2 time series processes are co-integrated:**

1. If one of the processes is driving the other
2. If both are being driven by the same underlying process

**Correlograms**

Plot of the Autocorrelation Function

|  |  |  |
| --- | --- | --- |
|  | **AR (p) Process** | **MA (q) Process** |
| General Form | Xt = μ + α1\*(Xt-1 –μ) + …+ αp\*(Xt-p - μ) + et | Xt = μ + β1\*et-1 + … +βp\*Xt-q + et |
|  | weighted average of the past p observed values | weighted average of the past q white noise terms |
| Number of parameters | p +2 | q +2 |
| YW Equation | γi = α1\*γk-1 + α2\*γk-2 +…+ αp\*γi-p  for k >= p | - |
| Distinguishing property | ACF decays  PACF cuts off  beyond lag ‘p’ | ACF cuts off  PACF decays  beyond lag ‘q’ |
| Stationarity / Invertibility | we check for stationarity  always invertible | we check for Invertibility  always stationary  (Linear combination of white noise process i.e. a sequence of uncorrelated random variables with constant mean, thus weakly stationary) |
| Markov Property | only AR (1) is Markov | Never Markov |

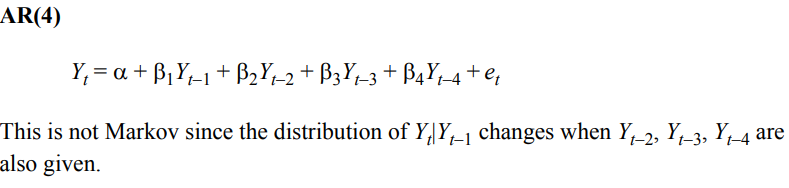
**ARMA (p, q) process**

* Combination of AR (p) and MA (q)
* General form **-** Xt = μ + α1\*(Xt-1 – μ) + … + αp\*(Xt-p – μ) + β1\*et-1 +…+ βp\*et-q + et
* Need to check for stationarity and Invertibility both
* ACF and PACF behaviour **–**
  + Neither ACF nor PACF will cut off, both decay
  + Presence of β1\*et-1 means that PACF will decay exponentially to zero, but will never cut off i.e. remain non-zero. (Sept 2013)
  + Only ARMA (1,0) is Markov

**Method of Moment estimation for ARMA (1,1) process (Q6 Assignment)**

* + In ARMA (1, 1) there are p + q parameters to be estimated.
  + We can calculate the theoretical ACF {ρk} of an ARMA (p, q) process, which will be a function of α’s and β’s.
  + Then the method of moments estimates are those values of α, β such that the theoretical ACF r\_1 …r\_p +q coincides with the observed sample ACF ρ\_1 …ρ\_p +q.
  + This method is easily available for AR (p) models since the corresponding Yule-Walker equations are linear, therefore moment estimation requires solving them with respect to the unknown parameters α\_i.

**Does AR (4) satisfy markov property? (Sept 2006)**



**Three possible causes for non-stationarity –** (non-exhaustive list)

1. A Deterministic Trend (linear or exponential growth)
2. A Deterministic Cycle (Seasonal Cycle)
3. Integrated Time Series

**Variance-stabilising transformation**

Transformations are most commonly used when dependence is suspected between the variance of the residuals and the size of the fitted values.

**Transformation to increase normality**

In certain applications it may be found that most residuals are small and negative, with a few large positive values to offset them. This may be taken to indicate that the distribution of the error terms is non-normal.

**Note** - A further caution when using transformed data involves the final step of turning forecasts for the transformed process into forecasts for the original process, as some transformations introduce a systematic bias.

**Bilinear Model**

It is linear with respect to both Xt and et thus it is called Bilinear.

**The main qualitative difference between the bilinear model and models from the ARMA class** is that many bilinear models exhibit ‘bursty’ behaviour: when the process is far from its mean it tends to exhibit larger fluctuations.

The difference between this model and an ARMA (1,1) process may be seen to lie in the last term on the right-hand side: when Xn-1 is far from μ and en-1 is far from 0 – events which are far from being independent – the final term assumes a much greater significance.

**Threshold Autoregressive Models**

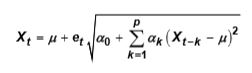


The distinctive feature of some models from the threshold autoregressive class is the limit cycle behaviour. This makes the threshold autoregressive models suitable for the description of ‘cyclic’ phenomena.

In an extreme case we might set α2=0, the Xn follows an autoregressive process until it passes the threshold value d. At this point Xn returns to μ and the process effectively starts again. Thus, we get cyclic behaviour as the process keeps resetting.

**Auto Regressive Models with Conditional heteroscedasticity**

* + 1. Financial assets often display the following behaviour.
    2. After a large change in the asset price there follows a period of high volatility, which can be in either direction.
    3. Following small changes there tend to be further small changes. In other words, the variance of the process is dependent upon the size of the previous value.
    4. This is the property of conditional heteroscedasticity.
    5. The words ‘homoscedastic’ and ‘heteroscedastic’ just mean having equal (i.e. constant) or different variances respectively.
    6. The class of autoregressive models with conditional heteroscedasticity of order p – the ARCH(p) – is defined by the relation

 Where, e is a sequence of independent standard normal random variables.

ARCH (1) Model

The ARCH models have been used for modelling financial time series. If Zt is the price of an asset at the end of the t+h trading day, it is found that the ARCH model can be used to model Xt = ln ( Zt/Zt-1), interpreted as the daily return on day t .

The ARCH family of models captures the feature frequently observed in asset price data that a significant change in the price of an asset is often followed by a period of high volatility.

**Exponential smoothing expected to outperform Box-Jekins?**

**–** When there is a slowly varying trend or multiplicative seasonal variations.

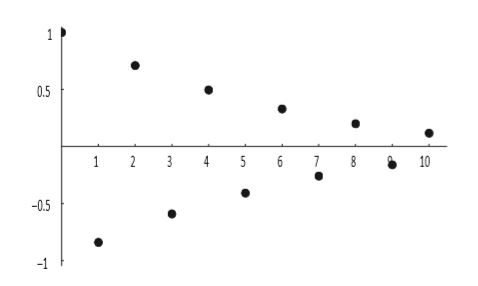
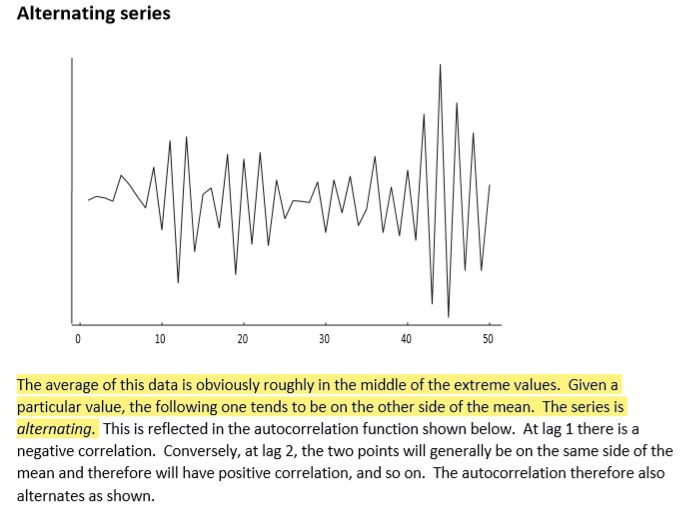
**Exponential smoothing preferred over Box-Jenkins technique?**

**–** When there is a slowly varying trend and there is multiplicative rather than additive, seasonal variation.

**Why CPI and NAEI (national average earnings index) might be cointegrated?**

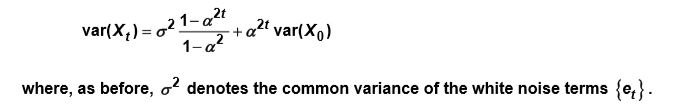
CPI is price inflation, which drives wage inflation (NAEI). So we expect them to ‘move together’. Neither process is stationary – they both have a trend (as prices and wages increase over time), so they may both be I(1) process.

**Alternating Series**

****

Graph of Original Time Series – Graph of Auto Correlation Function

The average of this data is obviously roughly in the middle of the extreme values. Given a particular value, the follow one tends to be on the other side of the mean. The series is alternating. This is reflected in the auto correlation function. At lag 1, there is a negative correlation. Conversely, at lag 2, there is a positive correlation and so on. The autocorrelation therefore also alternates.

**Variance of Xt under AR (1) process**

**Back-forecasting**

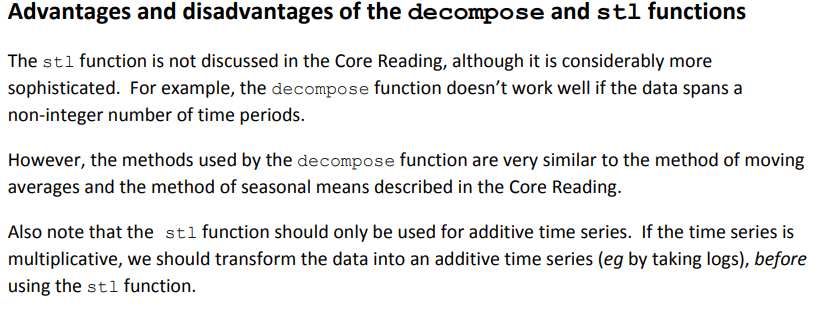
* + 1. Applied when we encounter a difficulty in calculating et from Zt.
    2. Suppose we have an ARMA (1,1) process as et = Zt – α1\*Zt-1 – β1\*et-1; an equation which can be solved iteratively for et as long as some starting value e0 is assumed.
    3. First, we assume that all error terms are zero and estimate αi and βj
    4. Then we use standard forecasting techniques on the time-reversed process {Zn, Zn-1 …Z1} to obtain the predicted values for (e0… eq-1).
    5. These new values can be used as the starting points for another application of the estimation procedure; this continues until the estimates have converged.

**In R, choosing number of ACFs in conducting Ljung Box Test**

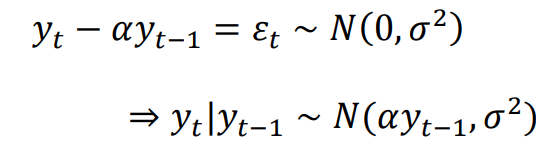
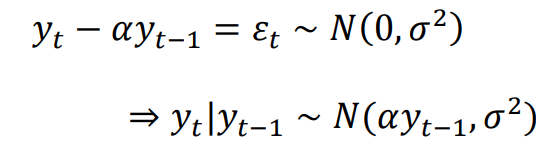
* Under Ljung Box test, H0: residuals are independent and TS ~ Chi-square with degree of freedom = m - (p + q)
* If value of ‘m’ i.e. number of ACFs is large, the degree of freedom is large number. This weakens the test.
* As a result, the TS is more likely to follow Chi Square
* Thus, we need to strike a compromise between these two effects.

**In R, choosing between ‘st’ and ‘decompose’ function**

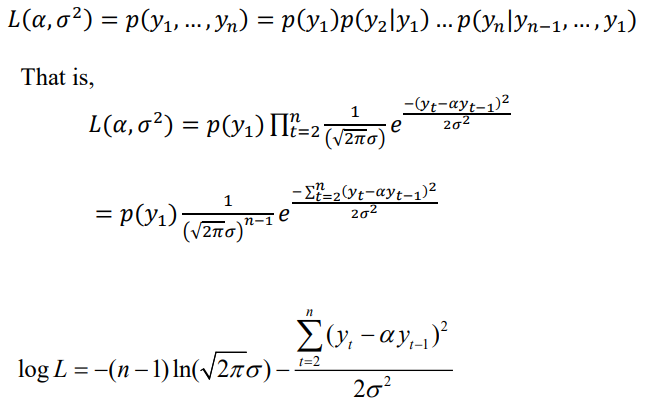
* First plot original time series
* If there is no clear pattern in amplitude of seasonality or volatility is not increasing over time, we use an ‘additive’ model instead of ‘multiplicative’ model.
* If the dataset is monthly observations – not spanning over an integer number of years. We prefer ‘st’ function.
* If data is additive, and does not span over integer time period, we use ‘st’ function



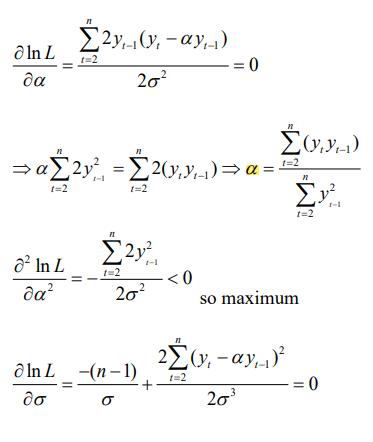
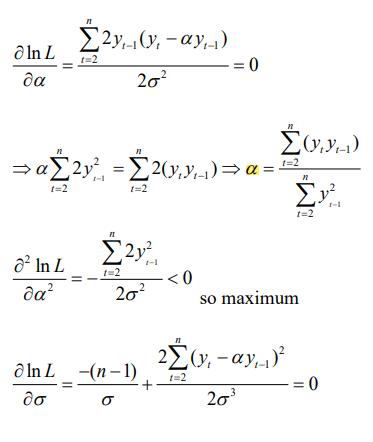
**Conditional Distribution of Yt on Yt-1 for AR(1) Process (2019 Sept)**



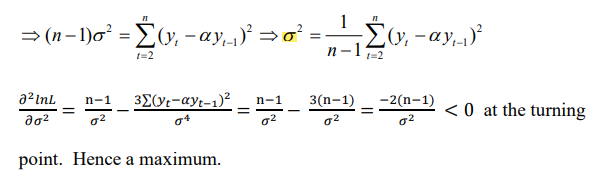
**Likelihood function of AR(1) model (2019 Sept)**



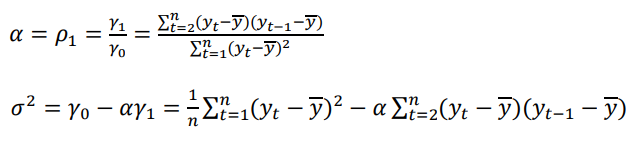
**MLE of α and σ^2**



(i) (ii)

(iii)

**Expression for α and σ^2 using Yule Walker Equation –**



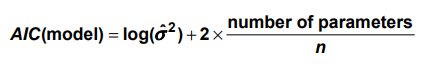
**Comment on difference between MLE and Yule-Walker estimates**

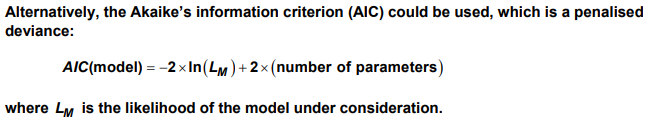
The main difference is that using Yule-Walker, we include the sample mean. This may not have a big effect when average y is small.

**Identification of white noise**

A test for whether a particular sequence of observations forms a standard white noise process may seem of doubtful usefulness, but one of the techniques of residual analysis suggests that the verification of goodness of fit of any model should include a test as to whether the residuals form a white noise process. A suitable test, or portfolio of tests, is therefore a valuable asset.

Clearly the SACF and SPACF of a white noise process are random, being simple functions of the observations. In particular, even if the original process was a perfectly standard white noise the SACF and SPACF would not be identically zero. The question is what scale of deviation from zero is to be expected.

**AIC -** The question of when to stop adding new parameters is addressed by Akaike’s information criterion (AIC), which states that we should only consider adding an extra parameter if this results in a reduction of the residual sum of squares by a factor of at least -2 n e , or alternatively, one can evaluate for each possible model the value of:And then choose as the most appropriate the one corresponding to the lowest such value.

**Updated-**

**Box-Jekin’s Methodology for fitting an ARIMA (p, d, q) process**

1. **Prepare the data**
   1. **Removing linear trend**
      1. Differencing Yt = Xt – Xt-1
      2. Least Square Trend Removal Yt = Xt – (a +bt)
   2. **Removing exponential trend -** Can be removed by taking logs i.e. set Yt = ln (Xt)
   3. **Removing seasonal effect**
      1. Seasonal Differencing - We difference the data by a lag equal to the period. E.g. Yt = Xt – Xt-12 if period is 12.
      2. Method of Seasonal Means - First, we calculate estimates of the seasonal means from the data by finding the average number of people claiming unemployment benefit (as per question) for each month less the overall mean of all months. Overall mean = sum(Xt)/N. Then, we subtract the appropriate seasonal mean from the observed values (or from the de-trended values). For e.g. for each January value, we would subtract January’s seasonal mean.
      3. Method of Moving Averages - The moving average model explains the relationship between the Xt as an indirect effect, arising from the fact that the current value of the process results from the recently passed random error terms as well as the current one. In this sense, Xt is ‘smoothed noise’.
2. **Tentative identification of ARIMA Model**
3. **Estimation of parameters in identified model**
4. **Diagnostic checks**
   1. Inspection of Graph of Residuals
   2. Counting Turning Points
   3. Ljung- Box Test / Portmanteau Test - A portmanteau test is a type of statistical hypothesis test in which the null hypothesis is well specified, but the alternative hypothesis is more loosely specified.
5. **Forecasting** 
   1. Box-Jekins Approach to forecasting time series

Reason for difference between estimated and actual values: The estimated values set the future residuals to be zero. There is some external reason (e.g. change in economic outlook, change in the way figures are recorded and so on).

* 1. Exponential Smoothing

**Relative Merit of using a large/small number of lags ‘m’ by considering how sample size ‘n’ would vary with ‘m’ and if sample acf is not equal to theoretical values. (April 2021)**

* Sample ACF = Theoretical ACF

As ‘m’ increases, the absolute value of test statistic increases. As acf values (ρk) generally decreases with increasing ‘k’ thus, the test statistic increases more slowly than the corresponding chi-square critical value.

There are two potential concerns (1) to make the test powerful, we prefer small values of ‘m’ (2)to ensure meaningful ACF values are included, we prefer large values of ‘m’. thus, we need to strike a balance between the two.

* Sample ACF != Theoretical ACF

If first few values of sample acf > theoretical acf then, small number of lags means small value of TS. Thus, to maximise the power of the test, value of ‘n’ should be higher.

If first few values of sample acf > theoretical acf then, small number of lags means high value of TS. Thus, to maximise the power of test, value of ‘n’ should be small. Here, is ‘n’ is also large, then there could a bias towards rejecting H0.

**Comment on Appropriate value of ‘d’**

* The correct value of ‘d’ should result in differences that appear to be stationary.
* A time series is likely to be non-stationary if the sample autocorrelation decays slowly from 1.
* The sample ACF of original series Xt appears to decay slowly from 1 so it should be differenced.
* The sample ACF of first-order differences ∆Xt appears to decay slowly from 1 so it should be differenced.
* The sample ACF of second-order differences ∆2Xt does not decay slowly from 1. We don’t need to difference the series any more.
* Another way of choosing appropriate value of ‘d’ is by finding the differenced series with the minimum sample Variance. Here, the second-order differenced series has the lowest standard deviation/variance.
* Thus, appropriate value of d is 2.

## Ch-14 Copulas

***Df*** - A copula characterizes the dependence – and only the dependence – between the components of a multivariate distribution; they can be combined with any set of univariate marginal distributions to form a full joint distribution.

***Df*** - A copula is a multivariate CDF whose Univariate marginal distributions are all Uniform(0,1). Copula CY contains all information about dependencies among the components of Y but has no information about the marginal CDFs of Y.

The ***primary financial application*** of copula models is risk assessment and management of portfolios that contain assets which exhibit co-movements in extreme behaviour.

***E.g.*** pair of assets having weakly correlated returns, but their largest losses may tend to occur in the same periods. They are commonly applied to portfolios of loans, bonds and collateralized debt obligations (CDOs).

***Assumption*** that the CDF is Strictly increasing and continuous is reasonable in many financial applications.

**Coefficient of upper tail dependence** indicates how the higher values of X are related with higher values of Y. u

* λu = lim:(u → -1): P(X > F-1x(u) | Y >F-1y(v) )
* λu = lim:(u → -1): (1-2u +C(u, u) )/(1- u)

**Akshay Sir’s Mock 2-**

The coefficient of upper tail dependence considers the probability that a random variable X takes a value in the upper tail of its distribution, given that the random variable Y takes a value in the same-sized upper tail of its distribution.

Specifically, the coefficient of upper tail dependency is the limiting value of this probability as we move further into the upper tail from below.

**Coefficient of lower tail dependence** indicates how the lower values of X are related with lower values of Y. l

* λL = lim:(u → 0+): P(X ≤ F-1x(u) | Y ≤ F-1y(v) )
* λL = lim:(u → 0+): C(u, u)/u

**Copulas Function** – takes marginal CDF as input and gives joint PDF as output.

**Properties of Copulas Function** –

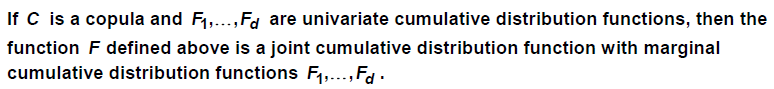
1. Increasing function of its inputs
2. If all inputs of copula function are equal to 1 except one of the marginal cdf then the copula function is equal to that marginal cdf.
3. A copula function gives output between 0 and 1

**Sklar’s Theorem –** let F be a joint cdf with marginal cdf F1,…,Fd. Then there exists a Copula C such that for all x1,…,xd belongs to (-Inf, Inf) F(x1,…xd) = C[F(x1)…F(xd)]

In case of variable that have a continuous distribution the copula is unique.

***In words*** - Sklar’s theorem states that the joint CDF FY can be decomposed into the copula CY, which contains all information about the dependencies among (Y1,…Yd), and the univariate marginal CDFs FY1, …,FYd, which contain all information about the univariate marginal distributions.

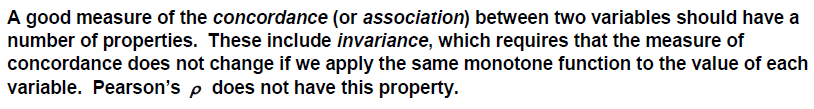
**Converse of Sklar’s Theorem**



**Survival Copula** is given by

* F̅(x,y) = P(X >x, Y >y) = C̅ (F̅(x), F̅(y)) = 1 – F(x) – F(y) + F(x, y)

**Desirable properties of a measure of concordance / association**



**Types of Copula Functions**

1. **Fundamental Copula**
   1. **Independence Copula** – no relation between X and Y thus, no tail dependency

C(u,v) = u\*v where λU = λL = 0

* 1. **Co-monotonic Copula**- perfect positive dependence for all X and Y.

C(u,v) = min(u,v) where λU = λL = 1

* 1. **Counter-monotonic Copula** - perfect negative dependence for all X and Y.

C(u,v) = max(u + v – 1,0) where λU = λL = 0

* Collectively these three are referred to as fundamental copulas. They are a specific case of a more general family of copulas called Frechet-Hoeffding Copulas.
* The Independence and Comonotic copulas can be extended to multiple variables, but ***counter monotonic cannot be extended*** as it is not possible to have three or more variables where each pair has a direct inverse relationship.

1. **Explicit Copula**
   1. **Gumbel** – describes an interdependence structure in which there is upper tail dependence but there is no lower tail dependence.

C(u,v) = exp **(**- **(** (-lnu)α + (-lnv)α **)**1/α **)** where λU = 2 - 21/α and λL = 0

* 1. **Clayton** – no upper tail dependence.

C(u,v) = (u-α + v-α - 1)-1/α where λU = 0, λL = 2-1/α

* 1. **Frank –** no tail dependence.

C(u,v) = (-1/α)\*ln[1 + (e-α\*u -1 )\*(e-α\*v -1)/(e-α -1) ] where, λU = λL = 0

* 1. **Archimedean** - these copula functions are described by specific generator functions. C(u,v) = φ**[-1]**(φ(u) + φ(v))

**Valid Generator Function**

* Decreasing Function
* Convex function
* Zhi (1) = 0

1. **Implicit** – these copulas are based on well-known multivariate distribution but no simplified closed from expression exists for them.

* **Gaussian**

C(u,v) = Φρ [Φ-1(u), Φ-1(v)]

where, Φρ = CDF of Bivariate Normal Distribution with correlation ρ.

* **Student’s T**

C(u,v) = tγ, ρ [t-1γ(u),t-1γ(v)] where t γ, ρ = t-distribution with ‘ρ’ correlation and ‘γ’ dof.

**Akshay Sir Mock 2 – Difficulty in deriving upper tail dependency coefficient for Frank Copula when u =1 – Overcome?**

The numerator and denominator are both 0 when u =1 and so it is undefined.

Overcome – use L’Hopital’s rule; it states that the limit of a ratio of functions is equal to the limit of the ratio of the derivatives of the functions.

**Akshay Sir Mock 2 - Frank copula graph examination – comment on value of coefficient of upper tail dependence**

Looking at a thin strip along the top of the graph, i.e. a large value of ‘v’, we can see that the probability of being in a square in the top right i.e. a large ‘u’ given large ‘v’ is very small. This is because values seem to be spread across the strip. Hence, the coefficient of upper tail dependence is zero.

**The bivariate Gaussian copula can be extended** to the multivariate case incorporating the (d x d) correlation matrix of the individual random variables. So, this is the unique copula that reproduces a joint normal distribution with a specified correlation matrix from the individual marginal distributions. Because it reproduces the joint distribution in this way, it is sometimes called an implicit copula.

**Advantage of Student’s T over Gaussian –** T Copula has an additional parameter γ, the number of degrees of freedom, which can be used to vary the strength of tail dependence. Smaller the value of γ, greater the level of tail dependence

**Suitability of Gumbel Copula**

* Gumbel Copula exhibits non-zero upper-tail dependence, the degree of which can be varied by the single parameter ‘α’. However, it exhibits no lower-tail dependence.
* So, the Gumbel copula, with an appropriate value for the parameter α, might be a suitable copula to use when modelling large general insurance claims resulting from a common underlying cause. (Acted)
* Hence, Gumbel copula is appropriate if we believe that the investments/assets are likely to behave similarly.
* Gumbel Copula is suitable if the portfolio’s returns are closely related in the upper tail i.e. extreme positive returns.
* E.g. modelling aggregate property and liability claims and losses.

**Suitability of Frank Copula**

* Has neither upper-tail nor lower-tail dependency
* E.g. modelling the relationship between equity returns and bond returns as returns are not directly dependent on each other.

**Suitability of Clayton Copula**

* has lower-tail dependence, but no upper-tail dependence
* Only lower-tail dependence makes it suitable to use it for extreme low or negative events are thought to happen together, e.g. returns from a portfolio of shares where poor or negative returns are likely to occur simultaneously on a number of investments e.g. a market crash.

**Note –** the Gumbel and Clayton Copulas are parameterised only with a single variable. This means that there is an implicit assumption that the shape and level of correlation between each bond is assumed to be identical which might not be the case. A wider range of relationships could be described by a two-parameter copula, such as Student’s T-Copula.

**Suitable Copula for Married Couple’s Lifetime (**Q 7 Assignment**)**

* Clayton Copula gives the highest probability (0.108908) of both lives dying within 25 years. This is because Clayton Copula exhibits lower tail dependence. This means that if one life dies early, there is a high probability that the other life will not survive for long.
* Gumbel Copula gives the lowest probability (0.0986) of both lives dying within 25 years. This is because the Gumbel Copula exhibits upper tail dependence. This means that if one life survives for a long time, there is a high probability that the other life will also survive for a long time.
* Studies also suggest that if one member of a married couple dies, this can precipitate the death of the other member (broken heart syndrome). On this basis, we might choose to use a copula function where there is a degree of positive interdependence throughout, e.g. the co-monotonic copula.

**Explain the correlation coefficient and the choice of copula when considering the relationship between the two or more variables.**

* Both are important in describing the overall relationship between the dependent variables.
* The correlation coefficient indicates the overall level of dependence between the bond returns. The higher the value of the coefficient, the greater the degree of dependence
* The copula describes the shape of this relationship i.e. how the level of dependence varies with the level of return on the bonds.

**Describe the coefficient of lower-tail dependence and its relevance to risk modelling.**

Tail dependence are used to describe joint dependence of risk at the extremes of the marginal distributions. For distributions of assets returns, lower tail dependence considers whether very poor returns on one asset class are likely to be associated with very poor returns on another asset class. The coefficient lies in the range [0,1]. A value of zero indicates no dependence in the tail. A value of 1 indicates perfect correlation in the tail. As different copulas exhibit different tail dependence, the presence of tail dependence can help in the selection of an appropriate copula.

**An insurer uses copulas to model the dependencies between various types of claims. A statistical analysis of the insurer’s claims shows that a Gumbel copula is a better representation of historic claims data than a Gaussian Copula. Discuss why the use of a Gaussian Copula in a claims model could result in solvency issues for insurer. [September 2020]**

The Gaussian copula has zero upper-tail dependence (for ρ < 1) whereas the Gumbel copula has positive upper-tail dependence (for α > 1). Given that the analysis suggests that the Gumbel copula is a better fit by using the Gaussian copula, the insurer is likely to be underestimating the probability/risk of multiple extreme claims occurring simultaneously/in tandem. The insurer’s solvency could be threatened if this occurs in practice and the insurer has not made adequate provision for such events.

**Explain in words, the meaning of C(u1,u2,u3) [April 2019]**

This gives the probability that RV1 is in the bottom u1 percentile, and RV2 is in the bottom u2 percentile, and RV3 is in the bottom u3 percentile.

This is a multivariate copula with inputs as CDFs of three random variables. This will give us the joint probability of these three random variables as output.

## Ch-15 Machine Learning

**What is machine learning?**

Machine learning is a set of techniques that are being developed to identify patterns in data for situations where we can’t use conventional mathematical models.

**A problem will be categorised as a Machine Learning problem if –**

1. A lot of data is present
2. There exists a pattern in our data
3. We cannot pin down a mathematical equation for it.

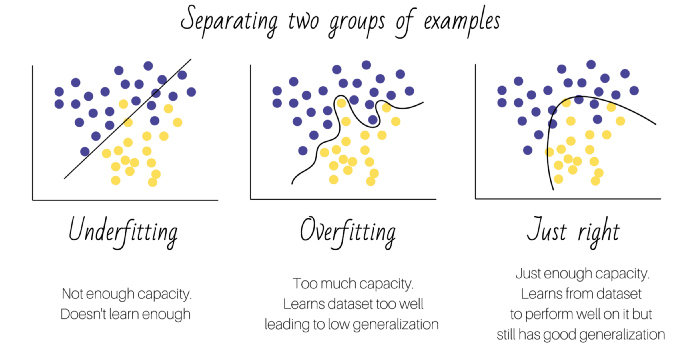
**Examples in our daily life** – Google search engine, filter out unwanted email, face and fingerprint recognition, suitable match on dating apps.

**Examples for Actuarial Work** – Risk Classification, identifying fraudulent claims and tax declarations

**Under mathematical model**, we use the SAME DATASET for modelling and testing, whereas under machine learning, the dataset is divided into groups for training, validation and testing.

**Machine learning disadvantage** –

* In statistics, ‘fit’ refers to how well the target function is approximated.
* *Over fitting* refers to learning the training data detail and noise which leads to poor generalisation. It can be limited by using resampling and defining a validation dataset. Here, the algorithm looks for a pattern that does not exist in real life
* *Under fitting* refers to poor inductive learning from training and poor generalisation.



**Need for Scaling data (CS2B Specimen 2019) –**

Clustering algorithms require a measure of the distance between each observation on each feature. If the data are not scaled, the weight given to a feature by the algorithm will depend on the units of measurement, as the absolute distances between the observations will vary more for some features than others.

E.g. if one variable varies from 0.03 to 219.66 and the other varies 0.007 to 0.77 only, we need to scale the data.

**#Types of Machine Learning–**

1. **Supervised Learning** – if the computer is told what the target of the analysis is, then the problem will be categorised as supervised machine learning.

E.g. if we want to predict claim amount or mortality rate then these are specific targets. The outcome could be numerical or categorical. If the outcome is numerical, the problem is categorised as a Regression Analysis. If the outcome is categorical, the problem is categorised as classification problem.

E.g. Generalised linear models, Naive Bayes classification, decision trees, prediction of future lifetime, neural networks, prediction of claims on certain classes of insurance, defaulting of loan, regression models, logistic regression, Probit models, discriminant analysis, perceptron, support vector machines.

1. **Unsupervised Learning** – if the algorithm is not trying to predict an outcome instead it is trying to create a set of categories. Then the problem is categorised as unsupervised learning. The idea is that the data in any one category should be homogeneous as possible and each category should be as different as possible.

E.g. K Means Clustering, Principal components analysis, Apriori algorithm, market basket analysis, text analysis, neural networks

Examples of unsupervised learning techniques include cluster analysis, and the use of association rules such as the apriori algorithm.

The **apriori algorithm** is a machine learning technique that identifies combinations of data values that frequently occur together in a data set, E.g. where users of a music website will tend to download items by the same artist or items of the same genre. It can be used by online retailers as the basis for the ‘Other customers also bought …’ recommendations or for promoting bundles of items that are frequently bought together.

1. **Semi-Supervised Learning** – Mixture of unsupervised and supervised learning. The algorithm might first try to create a set of categories and then predict a particular outcome.

In supervised learning, labelled data are used to build a model. However, labelling the

training data for some real-world applications can be difficult and time consuming. In some cases, there are implicit costs associated with obtaining these labels from domain experts. Semi-supervised learning attempts to address this inherent challenge by allowing the model to integrate part or all of the available unlabelled data in its supervised learning.

E.g. identifying people in photos where there are some photos with people already tagged by a user and a large number of untagged photos. A semi-supervised approach uses both the labelled (tagged photos) and unlabelled (untagged photos) data to try and identify people in photos.

1. **Reinforcement Learning** – in this type of machine learning problem we don’t tell the computer what the target of the analysis is, it should be trying to find a course of action that minimises a reward function using trial and error method.

It is the process by which the learner updates the probabilities of taking particular actions on the basis of past rewards received. The learner updates their strategy based on past actions and the associated outcomes.

**Examples** –

* Finding ***strategies for games***. The algorithm can use a representation of the current state of the game as a basis for defining a reward function and look for strategies that improve the chances of winning.
* This process resembles the idea of ***Actuarial Control Cycle*** where adjustments are made periodically based on feedback from past experience to ensure that a particular strategy remains on track. However, with reinforcement learning, it is the machine that updates the strategy based on feedback, without direct human intervention.
* Another popular machine learning technique that involves reinforcement is ***genetic algorithms***, which are based on the idea of selective breeding from biology. These create successive generations of possible solutions. Small random variations (‘mutations’) are introduced into each solution to create the next generation. The solutions are then tested and the ones that perform best are selected to continue to the next generation. After a few generations, a good solution may have been discovered amongst the surviving solutions.

**#Differentiate: Supervised and Unsupervised Learning**

1. Paragraphs above
2. The difference between these lies not (as one might think) in the level of involvement of the human researcher in the development of the algorithm, or in the supervision of the machine. Instead, it lies in the extent to which the machine is given an instruction as to the endpoint (or target) of the analysis

**#Stages of Machine Learning-**

1. Collect data
2. Explore and prepare data (data cleaning)
3. Scaling
4. Split the data (Train 60%, Validate 20%, Testing 20%)
5. Train the model
6. Validate the outputs (Sensibility of outcome)
7. Evaluate performance (Testing)
8. Improve performance
9. Reproducibility (same output for same inputs every time)

**Cleaning The Data**- replacing missing values, and checking the data for obvious errors is an important stage of any analysis, including machine learning.

**#Discuss the Train-Validation-Test Approach (2019 September)**

In machine learning, the convention is to divide the data into two parts. One part is used to train the algorithm and the other part is used to test the output for training.

The train-validation-test approach uses three data sets as follows:

* A training data set which is the sample of data used to fit the model that is, to train the algorithm to choose the most appropriate hypothesis
* A validation data set which is the sample of data used to provide an unbiased evaluation of model fit on the training dataset while adjusting the hyper-parameters these hyper-parameters are often specified in advance and then adjusted/optimised according to the performance of the model on the validation data.
* A test data set which is the sample of data used to provide an unbiased evaluation of the final model fit on the training data set. Under machine learning the results of the modelling exercise are applied to data which was not used to develop the algorithm, so the test data should be representative of the data on which the algorithm is to be used.

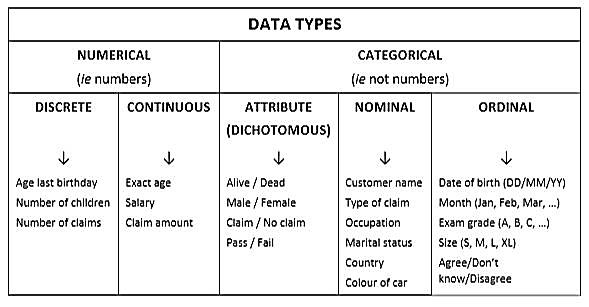
A typical split of data is 60% for training, 20% for validation and 20% for testing the principle being that enough data must be selected for the validation and testing sets, with the remainder used for the training set.

**E.g. of classification, regression and clustering problem -**

* An example of a classification problem is a spam filter that classifies emails into the two categories ‘Safe’ or ‘Suspicious’.
* An example of a regression problem is a health awareness app that predicts the user’s life expectancy.
* An example of a clustering problem is a system that groups together postcode areas that tend to have a similar experience of insurance claims.

**#AIC and BIC**

* The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) both impose a penalty for additional parameters. If the sample size is N and the number of parameters is J:
* AIC deviance + 2\*J BIC deviance + J\*ln(N)
* Where, the deviance is a measure of the model’s goodness of fit. By minimising the AIC or the BIC, we can achieve a trade-off between obtaining a good fit to the data and minimising the number of parameters in the model.

**#Types Of Data**

* **Attribute (or dichotomous) data** refers to variables whose values consist of just two categories.
* **Ordinal variables** take values that can be ordered in a natural way, whereas the values for **nominal variables** cannot.

**#Bias-Variance Trade-off:**

* In supervised learning, the prediction error ‘e’ is composed of the bias, the variance and the irreducible part.
* Bias refers to simplifying assumptions made to learn the target function easily.
* Variance refers to sensitivity of the model to changes in the training data.
* The goal of parameterisation is to achieve low bias (underlying pattern not too simplified) and low variance (not sensitive to specificities of the training data) trade-off.
* It is also important for data analysis to be reproducible and well-documented.

**Reproducibility of research –**

* Data used should be fully described and easily accessible to others
* Any modification to data should be clearly defined with computer codes. Under machine learning, this is called “features engineering”, whereby combination of features are used to create something more meaningful.
* Algorithm selection and model development process should be described with computer codes. This should include model parameters’ selection process and reason of selection.

**How can we achieve a good balance between bias and variance? - Regularisation**

* A method that can use all the features to choose the final hypothesis g , but will prevent it becoming too complex so that generalisation is poor.
* It is called regularisation or penalisation. This approach exacts a penalty for having too many parameters.
* Let the loss function be L\*(w1, w2, …Wj) . Then the hypothesis ‘g’ will be chosen to be the hypothesis with a set of weights which minimises L\*(w1, w2, …Wj).
* The idea of regularisation, or penalisation, is to add to L\* a cost for model complexity.

**#Penalty Function**

The more we miscalculate μ, higher until the value of penalty function. Thus, our aim is to minimise the penalty function.

**Why Prefer of Penalty function?**

* The basic MLE can sometimes lead to unrealistic results.
* The estimated values of the parameters can be very sensitive to the sample data and can vary wildly.
* This is most likely to happen when the sample size is small or likelihood function is very flat so that changes in the parameter values make very little difference to the log-likelihood.
* Applying a penalty function encourages the method to produce parameter estimates that are close to the values that would be expected from prior expectations.

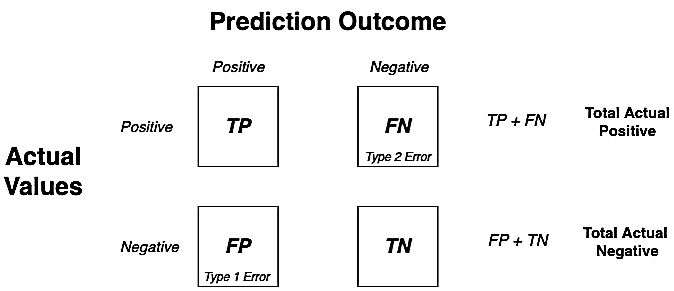
**Why lambda\*(μ-a) is used as penalty function?**

* In real life, dataset must contain such values as to suggest that μ is equal to a.
* Thus, by above given penalty function we check the deviation from target values we are expecting.

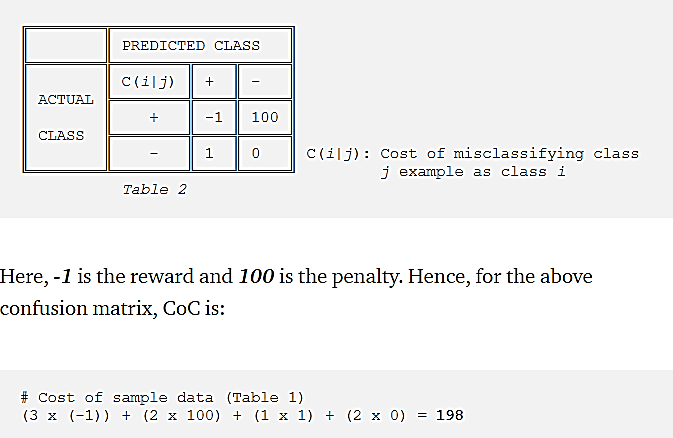
**Influence of Regularisation Parameter – λ**

* The regularity parameter lambda will be assigned a non-negative value. Otherwise, it would correspond to a reward rather than a penalty.
* If lambda was set to 0, there would be no-penalty and the method would reduce the maximum likelihood estimation.
* If lambda was given a very high value, the penalty dominates the calculations.[Then calculate for lambda tends to infinity and state the value of the estimate].

**#Evaluation Metrics In Machine Learning** <https://towardsdatascience.com/evaluation-metrics-18db24a91223>

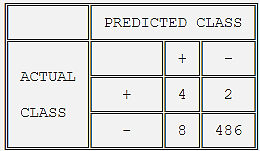
*Confusion Matrix* 

* **Cost of classification –** is a measure of computing cost for classification models. We assign weights to the confusion matrix i.e. we grant *reward* if classification is correct and *penalise* if it is wrong.

Lower the cost, better the model.

* **Accuracy –** ratio of correct prediction over total predicted values.

**Choice Between Accuracy And Cost Of Classification –** it depends on the aim of modelling. If the aim is to predict medical disorder, we want few false positives (type1 error) but we can afford to have some false negatives (type2 errors). Thus, we aim for higher accuracy. If the model is related to manufacturing company, we can afford a few errors but the cost is of the essence, thus a model with lower cost.

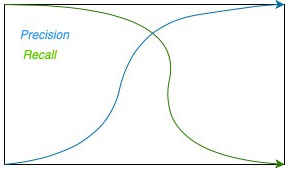
**Drawback Of Accuracy Is A Measure** - if the data is unbalanced, accuracy will be very high, but its predictive capability of the model may be very bad. Thus, accuracy is not sufficient for this type of classification. It is just not robust enough. To overcome this, we use other metrics. here, accuracy is 98% but predictive ability is very poor.

* **True positive rate –** value lies between 0 and 1. The higher the value of TP, the better.
* **False negative rate -** value lies between 0 and 1. The lower the value of TP, the better.
* **True Negative Rate and False Positive Rate** (formula above)
* F1 Score – is the harmonic mean of precision and recall. It eliminates the dilemma of prioritizing precision or recall. It reaches a maximum when precision is equal to recall.

*False Positive = Type 1 Error False Negative = Type 2 Error*

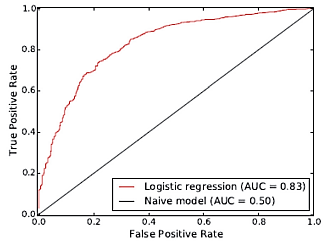
**Akshay Sir CS2B Mock 1 – Take Up Rate**

* Actual take up rate = (TP+FN)/total observation
* Predicted take up rate = (TP+FP)/total observation

Precision –vs.- Recall

**#Thresholding –** When we have a model that is predicting probabilities of classes and not classes themselves, we can set a *threshold value to classify probabilities as 0 or 1.*

**#Receiver Operating Characteristic Curve** **(ROC)**



* **Trade off** - ROC curve illustrates the trade-off between *recall* and the *false positive rate.*
* **Measurement** - The area under the ROC provides a single-figure measure of the efficacy of the model. The further away from the diagonal is the ROC, the greater the area under the curve and the better the model is at correctly classifying the cases.
* **Diagonal line** corresponds to a neutral ‘zero-sum’ test where there is a simple trade-off with any improvement in the true positive rate being matched by an equal deterioration in the false positive rate.
* **Points near the top left** of the graph correspond to a good test where the true positive rate is high and the false positive rate is low.
* **Area of the triangle below****the diagonal** is 0.5 and the area of the whole rectangle is 1 (the maximum possible score for the ROC).

**Uses** - This type of graph is most useful when the test involves a threshold of some kind. E.g. medical test for measuring chemical concentration.

**#Comment on values of Precision, Recall and F1 score (APRIL’19)**

* **Precision values** show that ­­­­­­­\_\_\_\_ (with higher value) is more effective at correctly identifying individuals who do have the disease/feature.
* **Recall values** show that \_\_\_\_ (with higher value) is much better at identifying individuals who do have the disease/feature.
* Whether recall or precision are chosen as measures will depend on whether it is most important to identify all persons who have the disease/feature, or not to unduly worry and treat people who are disease free.
* **As disease/feature is serious** it is best to maximise true positive and minimise false negative. For this we prefer …
* **F1 Score** show that the overall performance of \_\_\_\_\_ (with higher value) is much better than the \_\_\_\_ (with lower value).
  + The F1 score, however, is reasonably robust to the situation where most people do not have the disease, as its calculation does not involve the true negatives.
* **False positive** rates {for both are quite low or clinical>questionnaire} indicating that only a small proportion of individuals who do not have the disease/feature are incorrectly flagged as having it.
* **Patent comment**
  + **As the sample size** is relatively small, the test should be re-performed on a larger population before drawing any conclusions.
  + **Questionnaire/algorithm** is likely to be **easier and cheaper to administrate** and thus may be a **good short-term substitute** until clinical procedure can be widely established.
  + **F1 score** lacks interpretability and hence it should be used in combination with other evaluation metrics.
  + **Performance of well-curated algorithm** also depends on class distribution of target variable, cost of misclassification, and size of training and test sets.

**#K Means Clustering**

**Uses of K mean clustering –** recommending products based on similar customers, anomaly detection in customer behaviour.

**What does value of k represent?**

‘k’ is a hyper-parameter specifying the number of clusters the algorithm should aim to produce.

**Steps –**

* Step 1 – machine will randomly allocate each data value to a group
* Step 2 – calculate centroid for each cluster
* Step 3 – reassigns if another data value is closer to the centroid until further reassignment cannot be done and distinct categories are created.

**‘The k means clustering algorithm converged after 3 iterations.’ Meaning of convergence?**

The algorithm involves repeatedly finding the centroid of the data points that have been allocated to each cluster and then reallocating the points to the cluster whose centroid they are nearest to. When this process reaches a stage where no further changes are made, the algorithm has converged to the solution.

**Sept 2020 CS2B – Kmeans Clustering Comment**

* Graph of data values based on assigned labels shows that the algorithm reasonably identified the two sets of clusters
* Inspection of the table might suggest that the clusters are unlikely to change much.
  + 70 remain in label A and 97 remain in label B
  + 3 move to label A and 30 moves to label B
* Patent Comments –
  + The analyst could potentially improve the results by updating the centres of the clusters, re-calculating the distances, updating the labels and repeating this process until convergence (i.e. until the labels remain constant).
  + The analyst could implement the full kmeans algorithm to ensure convergence of the final clusters.
  + The analyst may want to apply feature scaling / data normalisation to the values of x\_1 and x\_2 so that each of them contributes approximately proportionately to the Euclidean distances and then re-run the analysis.
  + Techniques to measure within-cluster homogeneity could be used to assess the validity of the results.

**Advantage or strength of K means-**

* Uses simple principles for identifying clusters which can be explained in non-statistical terms
* Highly flexible and can be adapted to address nearly all its shortcomings with simple adjustments
* Fairly efficient and performs well
* No learning phase as it works with the entire training set.
* Robust to noisy data, no need to filter outliers

**Disadvantage or Weakness of K means-**

* Less sophisticated than more recent clustering algorithms
* Not guaranteed to find the optimal set of clusters because it incorporates a random element
* Requires a reasonable guess as to how many clusters naturally exist in the data
* Results are sensitive to units of measurement used
* Clusters may have no natural interpretation
* It can’t be used unless the data have a natural numerical order

**Within-Cluster Sum of Squares by Cluster (faithful dataset in R)**

For any particular cluster, this gives the sum of squared distances from each point to the cluster centre.  where

* Wij: waiting time of ith point in cluster ‘j’
* Eij: eruption time of ith point in cluster ‘j’
* Wj-bar: average waiting time of points in cluster ‘j’
* Ej-bar: average eruption time of points in cluster ‘j’
* nj: number of points in cluster ‘j’

This is the Euclidean distance of ith point from centre in cluster ‘j’.

It is a measure of within-cluster homogeneity.

Smaller the figure, more tightly packed is each cluster.

The ratio between the cluster sum of squares to the total sum of squares is given by

**(between\_SS/ total\_SS = \_\_\_ )**

**Between Cluster Sum Of Squares:**

where, W-bar and E-bar are the overall average of all points and ‘k’ is the number of clusters.

**Total Sum Of Squares**:



The ratio is a measure of how much total sum of squares has been explained by different clusters. Thus, similar to idea of R2 under linear regression.

The closer the ratio is to 1, the greater heterogeneity between clusters and greater the homogeneity within clusters.

Note: We can change the ratio by considering different number of clusters (thus, k is a hyper parameter). Indeed we can force the ratio to equal 1 by simply having each individual point as its own cluster of size 1.

**#Decision Tree**

**Greedy splitting**

* Creating a binary decision tree is a process of diving up the input space. A ‘greedy’ approach is used to divide the space – called ***recursive binary spitting***.

***Description*** - This is a numerical procedure where all the values are lined up and different split points are tried and tested using a cost function. The split with the lowest cost is selected. All input variables and all possible split points are chosen in a greedy manner i.e. the very best split is chosen each time.

* ***Greedy*** means that at each stage, we just choose the split that appears to be the most effective at separating the remaining elements, without thinking ahead of the consequences this might have on latter splits.
* ***For regression problem*** – we minimise cost function given by sum squared error = Σ(yi – yi-hat)2 i.e. sum:(actual-predicted)2.
* ***For classification problem*** –we minimise Gini Index function given by Gi = Σ pk\*(1-pk)

**Gini Index**

The Gini index is a measure of inequality of a distribution that was introduced by the Italian statistician Corrado Gini.

The Gini Index is a measure of the ‘impurity’ of the nodes in the decision trees i.e. the extent to which the final nodes contain a mixture of different data types.

* G = Σ(range of k): pk (1-pk)
* G = 2p1\*p2
* Where, pk is the proportion of training instances with ‘class k’ in the rectangle of interest.
* *For Binary Classification Range***: 0** (perfect class purity) to **0.5** (worst purity)
* *For ‘m’ Categories Range***: 0** (perfect class purity) to **1-(1/m)** (even split of each category, i.e. the proportion of each category in the node is 1/m).
* *LOWER the value of Gini Index, the BETTER.*
* **Process –** when constructing a tree, the Gini score is used at each step to decide how to divide the data. The score is calculated for the candidate splits and when using the greedy approach, the split with the lowest score is chosen. This is the split which gives the *purest overall child nodes*.

*This does not guarantee the best overall classification tree.*

**Identity** 

**Proof** 

**How this identity can be used to calculate a measure of effectiveness of a proposed split point when constructing a decision tree?**

We can measure the effectiveness of a proposed split point by examining the ‘purity’ of the data in each of the child nodes and then calculating an overall measure of the purity of the split.

This can be done by multiplying the proportion of items of type k at each child node by the proportions for each other type j k ≠ and summing. These values are then weighted by the number of items at that node to calculate an overall measure called the Gini index. Using the identity above leads to the following formula:where the sum is taken over all the child nodes of the split and node n is the number of items in each child node.

**Comment on Gini Index value**

* In case of a binary classification problem, the Gini Index can take any value between 0 and 0.5.
* If the points at each node are alike, the value of the index is 0.
* In this case the value of Gini Index is 0.3170.
* **Patent Comment -** So, the algorithm is not very effective at distinguishing between the different types of data points.

**Stopping Criterion**

The recursive binary splitting procedure needs to know when to stop splitting as it works its way down the tree with the training data.

Most common – using a ***minimum count*** on the ***number of training members*** (sample size) assigned to each leaf node. If count is less than the minimum, then the split is not accepted and the node is taken as a final lead node. This is another example of a hyper-parameter. This defines how specific to the training data the tree will be. Too specific (e.g. a count of 1) and the tree will overfit the training data and perform poorly with the test data.

**Pruning the tree**

* Definition – enhancing the performance of a tree by removing those sections which provide very little power to classify data.
* The *simplest form of pruning* is Reduced Error Pruning: starting at the leaves, each node is replaced with its most popular class. If the prediction accuracy is not affected, then the change is kept.
* The *fastest and simplest pruning method* is to work through each leaf node in the tree and evaluate the effect of removing it using a hold-out test set (another data set that wasn’t used in training the model). Leaf nodes are removed only if it results in a drop in the overall cost function on the entire test. You stop removing nodes when no further improvement can be made.
* *More sophisticated pruning methods* can be used such as cost complexity pruning (also called ‘weakest link pruning’) where a learning parameter (alpha) is used to weigh whether nodes can be removed based on a combination of - the size of the sub-tree and classification performance of resulting reduced tree.

**Advantages of Pruning –**

* Easy to interpret
* No overfitting with pruning
* Works for both classification and regression problems
* Can take any type of variables without modifications
* Does not require any data preparation

**Improve Decision tree – (Akshay sir CS2B mock 1)**

* We could include left out covariate (input variables) – it could potentially improve the tree. However, we will need to access the impact of any additional input variable. If it improves overall performance, we can include.
* We could produce a range of different tree models using different orders to make the splits and using different split levels for the age variables, to see if there is a different model that performs better.
* In most applications, by aggregating many decision trees, using methods like bagging, random forests, and boosting, the predictive performance of decision trees can be substantially improved.

**Advantage Of Decision Tree –**

* Easy to explain
* Closely mirror human-decision making compared to other regression and classification approaches.
* Easy to display graphically
* They can easily handle qualitative predictors without the need to create dummy variables.
* Does not require any pre-processing of data like normalization or standardisation of features – thus it is invariant to data scaling as features are processed separately.
* Decision tree works well with features that are on different scales or a mix of binary and continuous features.
* They are able to generate understandable rules
* They perform classification without requiring much computation.

**Disadvantage Of Decision Tree –**

* They have a lower level of predictive ability than other approaches since they aren’t quite robust because a small change in data can cause a large change in the final decision tree.
* Even with use of pre-pruning, they tend to over-fit and provide poor generalisation performance.
* Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
* Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.

**#Naive Bayes**

* Naive Bayes is a Supervised Machine Learning algorithm based on the Bayes Theorem that is used to solve classification problems by following a probabilistic approach.
* It is based on the idea that the predictor variables in a Machine Learning model are independent of each other i.e. the outcome of a model depends on a set of independent variables that have nothing to do with each other.
* A Generative model first models the joint distribution, P(x1,x2,…y) and then uses Bayes’ rule to estimate conditional probabilities P(y|x1, x2,…y), which can then be used to predict the category given the covariates.

**Problem** – number of separate probabilities to be computed increases exponentially with the number of covariates.

**Solution** – by assuming the covariates are independent.

**Naive Bayes Classifier**

**Example –** Article Classification using binary word presence, Email spam detection using a similar technique

**But why is Naive Bayes called ‘Naive’?**

In real-world problems, predictor variables aren’t always independent of each other; there are always some correlations between them. Since Naive Bayes considers each predictor variable to be independent of any other variable in the model, it is called ‘Naive’.

**#Advantage of Naive-Bayes Approach –**

* + Fast because only probabilities need to be calculated.
  + Easy to apply
  + Requires very little data
  + Good for few category variables
  + If naive assumption work can converge quicker than other models. It can be used for smaller training data.

**#Disadvantage of Naive-Bayes Approach –** Assumes that the conditional probabilities are independent which can be a poor approximation when the variables are correlated.

**#Hyper-Parameters**

They cannot be estimated from the data – indeed they must often be defined before an algorithm can be implemented. Hyper-parameters are external to the model and their values cannot be estimated from the data. They are typically specified by the practitioner and may be set using heuristic guidelines. Nevertheless, they are critical to the predictive success of a model.

‘Heuristic’ means that there are no hard and fast rules for these. They are determined using rough guidelines and past experience of what works well, combined with experimentation.

**#Differentiate: Parameters and Hyper-parameters (Akshay Sir’s Mock 3)**

* Parameters are variables internal to a model. They are either estimated or learned from the data. Their values are used when calculating predictions from the model.

E.g. the linear regression model:

Yi = β0 + β1\*x1i + β2\*x2i+ …+ βj\*xij +ei

It has parameters β0, β1, β2 …βj and σ^2

* Hyper parameters are external to the model. Their values cannot be estimated. Their values often have to be set prior to running an algorithm.

E.g. number of covariates J to include in a regression model, the number of categories in a classification exercise, or the rate at which the model should learn from the data.

**#Advantage of large no. of parameters**

It can improve the accuracy of the model and predictions, because a model with, more parameters will fit the data more closely than one with fewer parameters.

**#Disadvantage of large no. of parameters**

There is a risk of over-fitting where the estimates from the model will reflect idiosyncratic characteristics of the “training” data set rather than characteristics which apply to the whole data set. This may lead to the analyst identifying patterns which do not exist. If too many parameters are used the model can become complex and computationally expensive to run. Using too many parameters may lead to model stability issues.

**#Hyper-Parameters Relevant To Actuaries**

* **Life insurance** – Premium calculation and say we need to define the limit of cigarette consumption to be categorised as heavy smoker.
* **Graduation** - If we are graduating mortality data using a Gompertz-Makeham formula and we need to determine the order of the two polynomials.
* **Time Series** – determining values of p, d ,q for ARIMA process
* **GLM** - If we are applying a generalised linear model, we need to decide on the form of the link function to use.
* **Reinsurance** - If we are modelling large claims in general insurance, we need to specify the cut-off point for a claim to count as ‘large’.
* **Motor insurance** - If we are using geographical area as a rating factor in motor insurance, we need to decide on how many areas to use and which locations these cover.
* **Health** - If we are using a patient’s body mass index (BMI) as a predictor for the outcome of a medical procedure, we may need to specify the dividing lines between weight bands such as underweight (<20), normal (20–25), overweight (25–30) and obese (30+).

**#Generalisation Error**

An upper bound can be determined for the magnitude of out-of-sample errors. This shows that, with a large enough training set, the out-of-sample error can be made as small as desired.

**#Cross-validation**

* To assess how the predictive ability of the model changes as the number of parameters / features increases we can withhold a portion of the ‘training’ data and use it to validate models with different numbers of parameters / features.
* One approach is to divide the training data into, say, s slices, and to ‘train’ the model ‘s’ times, using a different slice for validation each time. This is called **S-Fold Cross-Validation.**

**#Supervised Learning Techniques -**

* + - 1. **Discriminant Analysis –** here, we model conditional probabilities P(y given x1,x2…) directly. The model builds decision boundaries between classes by directly modelling the conditional probabilities P(y given x1, x2…) rather than modelling full joint distribution P(x1,x2… y)
      2. **Binary Logistic Regression** – can be used to predict a binary outcome (e.g. Yes-No). The output here are estimates of conditional probabilities P(Yes given x1, x2 …) and P(No given x1, x2…). This is a discriminant approach as we don’t model joint probabilities P(x1,x2… Yes) and P(x1,x2,…No).

**Note -** Logistic Regression is based on **Logistic Function** which converts an input value (-Inf < x< Inf) to an output value on a continuous scale between 0 and 1.

If we interpret the output as probability, we can convert it to a categorical output by saying that the values exceeding a specified value of probability corresponds to Yes, while smaller values correspond to No.

* + - 1. **Multinomial Logistic Regression** – is an extension to binary logistic regression for outcomes that can take more than two values.
      2. **Perceptrons** and **neural networks** - Use interconnected layers of artificial neurons that can be activated or deactivated in a way that mimics the behaviour of the neurons in animal brains.
      3. **Probit Models -** It (short for ‘probability unit’) produces outputs that can only take one of two values, e.g. Yes / No or 0 / 1.
      4. **Support Vector Machine** - Here, a hyper-plane is selected to separate the points in the input variable space by their class, with the largest margin. The closest data points (defining the margin) are called the support vectors. But the real data cannot be perfectly separated, that is why a ‘C’ defines the amount of violation of the margin allowed. The lower C, the more sensitive SVM is to training data.

**#Advantage of Support Vector Machine -**

* Allows non-linear separation with non-linear Kernels
* Works good in high dimensional space
* Robust to multi-collinearity and overfitting

**#Semi-Supervised Learning**

**For example -** a system that aims to *identify pickpockets* operating in a busy shopping street might first identify people who appear several times throughout the day with the same clothing. These people’s faces could then be matched against a database of known offenders.

**Other Examples** – Face detection from images, Target audience classification from tweets

**A semi-supervised learning approach could be carried out as follows:**

(i) Create a first model ‘M1’ using ‘data1’ alone.

(ii) Predict the target values in ‘data2’ using model ‘M1’.

(iii) Create an improved model using the observed target values from ‘data1’ and the predicted target values for ‘data2’ obtained in (ii) as if they were the known target values for the unlabelled dataset.

**Auto encoder** - An auto encoder compresses the raw data by focusing on features that appear to be significant, e.g. it might identify the different types of object that appear in a photo, even though it doesn’t know what they actually are. It can considerably speed up any future modelling analysis.

**#CART algorithm provides a foundation for important algorithms like:**

**Bagged decision trees** - In bagged decision trees, we create random sub-samples of our data with replacement, train a CART model on each sample, and (given new data) calculate the average prediction from each model.

**Random forest** - Random forests apply a method based on averaging a number of randomly generated decision trees.

**Boosted decision trees** - Boosting here refers to a method of repeatedly making small adjustments to improve the effectiveness of a model by reducing the residual error.

**Parametric -vs- Non-parametric Methods in Machine Learning**

**Parametric Models -** In parametric methods, we typically make an assumption with regards to the form of the function ‘f’. for e.g. the unknown function ‘f’ could be linear i.e. of the form f(X) = β0 + β1\*x1 + β2\*x2 + … + βp\*xp. Then we select a model that aligns with this assumption.

Thus, under parametric method in machine learning, we take a model-based approach where we make an assumption with respect to the form of function to be estimated and then we select a suitable model based on this assumption to estimate the set of parameters.

**Disadvantage** – assumption made may not be true. Thus, these methods involves less flexible algorithms and are usually used for less complex problems.

**Advantage** – they are quite fast and require significantly less data than non-parametric models. As they are less flexible and suitable for less complex problems, they are more interpretable.

E.g. Linear Discriminant Analysis, Naive Bayes, Perceptrons.

**Non-Parametric Models -** Non-parametric methods refer to a set of algorithms that do not make any underlying assumptions with respect to the form of the function to be estimated.

**Advantage** - They tend to be more accurate as they seek to best fit the data points. They are quite flexible and can lead to better model performance since no assumptions are being made about the underlying function.

**Disadvantage** - But they require a very large number of observations. Additionally, these methods tend to be less efficient when it comes to training the models. Furthermore, they may sometimes introduce overfitting. These algorithms tend to be more flexible, they may sometimes learn the errors and noise in a way that they cannot generalise well to new, unseen data points.

E.g. Support Vector Machines, K-Nearest Neighbours.

**Sept 2021 – R comment on confusion matrix**

* Confusion matrix → . Precision = 40% and Recall is 100%.
* Recall is the % of defaults that the model managed to identify. Here the model has performed well and has identified all 20 actual defaults.
* Precision is the % of predicted defaults that are in fact actual defaults. Here the model isn’t very precise.
* The model has predicted far more defaults than was actually the case. Hence, if this model had been used to approve the loans of these 100 customers, 30% of them would have not been approved for a lone even though they did not actually default.
* The model is therefore ***not commercially optimised***.
* This is in line with our conclusions from part (iv) i.e. the specimen probability of default which is unrealistically too high.
* The Actuary could ***refine the parameters*** a, b, c used to construct the specimen data and ***re-train*** the model to improve it.
* Better still, further ***data*** could be ***gathered from the public domain***, if available, and used to train the model to improve it.
* Additionally, ***other decision tree models*** e.g. bagged decision trees, random forests, boosted decision trees should be investigated to see if a better fit can be obtained.
* Alternatively, ***other classification machine learning models*** like Naive Bayes should be investigated to see if a better fit can be obtained.
* Actuary could change the approach from classification of loan default/not default to a probability of loan default approach and use some form of ***regression machine learning algorithm to predict probability***.

## Ch-16 Mortality Projection

Mortality projection models are used by actuaries to take account of number of variables that can play an important role in prediction of future mortality. Mainly these variables are AGE, TIME PERIOD and COHORT.

* **2-factor Age-Period Model** – these involve the model of mx, t i.e. mortality rate of people attaining age x in time t. Here, we assume that the year in which mortality occurs is a factor in determining the mortality.
* **2-factor Age-Cohort Model** – these involve the model of mx, c i.e. mortality rate of people attaining age x in cohort ‘c’. Here, we assume that the year of birth is a factor in determining the mortality.
* **3-factor Age-Cohort-Time Model** – here we assume that both time and cohort affect mortality of life aged x.

**mx, t = ax + bx\*ft**where ft is function of time, ax, bx are functions of age

**mx, c = ax + bx\*fc**where fc is function of cohort

**mx, c = ax + b1x\*fc**+ **b2x\*ft**

**Methods of projecting Mortality Rates**

1. **Methods of expectation** – they depend on Expert Opinion and Judgement as to how mortality is likely to change in future.

* **Rx, t = αx + (1-αx) \*(1-fn, x) t / n**
  + Rx, t- reduction factor,
  + αx - ultimate reduction factor,
  + fn, x - proportion of total decline expected to occur in ‘n’ years.

It can also be judged that mortality rates are likely to be reduced in future and a model can be constructed that projects a gradual reduction in mortality rates from the current level towards their assumed minimum value.

**Properties –**

* For **t = 0**, Rx, t = 1 i.e. no reduction in current year.
* For **t = n**, Rx, t = 1– (1–αx) \*fn, x i.e. proportion of maximum possible value of reduction factor.
* For **t →Inf**, Rx, t = αx i.e. minimum reduction

**Advantage** – widely used in the past, straightforward and easy to implement.

**Disadvantage –** It can tend to underestimate mortality when experts are unable to judge the changing lifestyle trends and medical advancements leading to improvement in mortality.

There are theoretical problems, too, with targeting. The setting of the ‘target’ is a forecast, which implies that the method is circular; and setting a target leads to an underestimation of the true level of uncertainty around the forecast.

1. **Extrapolation based method** – it uses past mortality-based experience to produce a model of future mortality experience.

**Step 1** – choose a mathematical model

**Step 2** – Parameter Estimation

**Step 3** – Use model to predict future mortality rates

**Lee-Carter Model**

* Log (mx, t) = ax +bx\*kt + ex, t
* mx, t

– is the central rate of mortality

* bx

– is the extent to which time period affects the mortality rate at age x – captures the differences in relative rates of change by age.

* kt

– factor that depends only on year t

– reflects year-to-year changes in the general level of mortality.

* ex, t

– stochastic error terms which are assumed to be iid rv for all x, t

* ax

– mean of time-averaged logarithms of mx, t

– captures the differences in mortality by age.

**Why log –** log-transformation is applied to reflect the fact the mortality rates historically increase exponentially with increasing age.

**Assumption about kt** : is a time series model. It can also be a regression model. Thus, there is an element of judgement and expert opinion here as well.

**Model Ignores –** <https://youtu.be/NF3poHz40Sc>

* Correlation structure of errors across ages
* Long term trends or rotation of relative rate of decline by age
* Some shocks like HIV AIDS or 1918 Influenza pandemic that affect ages differently.
* Any cohort-specific factors, this is a period model driven by whatever happens in that period.

**Constraints –** Σbx = 1 and Σkt = 0

* We assume mortality is improving over time then time trend factor kt would reduce with increasing time. Thus, assuming linearly over time
* bx is the relative impact time trend has on mortality rates at age x. evidence suggests that trends are less apparent at older ages. So, absolute value of bx reduces with age x.

**How the *bx* parameter affects the projected time trend**

1. When bx-hat =1, the projected change in mortality over time directly reflects the changes in the time trend function kt-hat over the specified time period.
2. When bx-hat is positive, the change in mortality over time is the same direction as the time trend function.
3. When bx-hat is negative, the trend in mortality assumed at that age is in the opposite direction to the time trend function in the model.
4. When 0< |bx-hat| <1, the change in mortality over time is smaller in absolute terms than the change in the time trend function.

When |bx-hat|>1, the change in mortality over time is greater in absolute terms than the change in the time trend function.

1. **Explanation based method** **–** it projects mortality rates separately by cause of death and combines them to produce overall projected mortality rates.

**#Cause-deleted life table approach**

**Briefly describe the various sources of uncertainty involved in the mortality rates.**

* **Errors in the estimation of the graduated (base) mortality rates** 
  + Random error in the observed experience data (on which the base-year graduated rates were based)
  + Errors in fitting the model for the mortality rates to the experience data in the graduation process
* **Errors in the projection of the base mortality rates into the future** 
  + ***Model error***: the formula for, Rx, t will not have exactly the right functional form to produce a realistic projection of future mortality rates, regardless of the parameter values used.
  + ***Parameter error***: the parameter values will be incorrect, ei the values chosen for αx and, fx, n for each x will be subject to error.
  + ***Random error***: the actual mortality rates will differ from the projected rates due to random fluctuations in the mortality experience.
  + ***Changes to the class of lives*** involved between past and future, e.g. if we are projecting pensioner mortality, the type of people becoming pensioners in future may differ from those who were pensioners when the base mortality rates were calculated. This may be because of changing types of pension arrangements or alternative options available, changes in the standard of living amongst pensioners, changes in the mix of pensioner lives by nationality, etc.

**Advantage of Lee-Carter Model**

* Once the parameters have been estimated, forecasting is straightforward and can proceed using standard time-series methods, the statistical properties of which are well known.
* The degree of uncertainty in parameter estimates, and hence the extent of random error in mortality forecasts, can be assessed.
* The Lee-Carter model can also be extended and adapted to suit particular contexts.

**Disadvantage of Lee-Carter Model**

* Future estimates of mortality at different ages are heavily dependent on the original estimates of the parameters ax and bx. The forecasting assumes that these remain constant into the future. This will incorporate any roughness in the past data. They may be distorted by past period events which affected different ages to different degrees. If the estimated bx values show variability from age to age, it is possible for the forecast age-specific mortality rates to ‘cross over’.

This can be avoided by smoothing the estimates of ax and bx.

* There is a tendency for Lee-Carter forecasts to become increasingly rough over time.
* The model assumes that the underlying rates of mortality change are constant over time across all ages, when there is empirical evidence that this is not so.
* The Lee-Carter model does not include a cohort term, whereas there is evidence from UK population mortality experience that certain cohorts exhibit higher improvements than others.
* Unless observed rates are used for the forecasting, it can produce ‘jump-off’ effects (i.e. an implausible jump between the most recent observed data and the forecast for the first future period).
* **Updated** - When the random walk is used to forecast kt within the Lee-Carter model, the resulting ratio of the rates of mortality change at different ages remains constant over time, when there is empirical evidence that this is not so

**Age-period-cohort models have substantial disadvantages:**

1. The identification problem: any one factor is linearly dependent on the other two. Various solutions to this problem have been proposed, including the use of three-way classification in data collection, and the imposition of constraints on model parameters.
2. Models incorporating cohort effects impose heavy data demands. To observe the mortality of a cohort over all ages requires around 100 years of data. Cohort data truncated at the current age of the cohort can be used, but a model will be needed to estimate the experience of the cohort at older ages.
3. Major Difficulty present here that is not found in either age-period or age-cohort – the three-factor model has the “logical problem” that each factor is linearly dependent on other two. So we need to endure that the three arguments of the function work together in a consistent way in the formula. (Practise Questions)

**Advantages of the p-spline approach**

* natural extension of methods of graduation and smoothing
* relatively straightforward to implement in R

**Disadvantages** **of the p-spline approach**

* When applied to ages separately, mortality at different ages is forecast independently. So there is a danger that there will be roughness between adjacent ages.

This can be overcome by fitting the model and forecasting in two dimensions (age and time) simultaneously.

* There is no explanatory element to the projection.
* p-splines tend to be over-responsive to an extra year of data (though this can be ameliorated by increasing the knot spacing).

This means that if we fit the model to ‘n’ years of data, and fit it again to ‘n + 1’ years of data (e.g. because we’ve just gathered another complete year of observed experience), the model changes more dramatically than we would normally expect (e.g. compared to the case where we are fitting a standard mathematical formula like the Gompertz model).

**Difficulties of the method based on explanation approach include**:

* Forecasting future changes in the risk factors / disease states
* Allowing for the lag between changes in the risk factors and their effect on mortality
* Difficulties in identifying and categorising the cause of death

**Sources of error in mortality forecasting –**

* Model Misspecification
* Parameter Uncertainty
* Incorrect Judgement Or Prior Knowledge
* Random Variation, Including Seasonal Effects
* Data Errors.

**Reasons for inadequate fit and how it could be improved by cubic spline function**

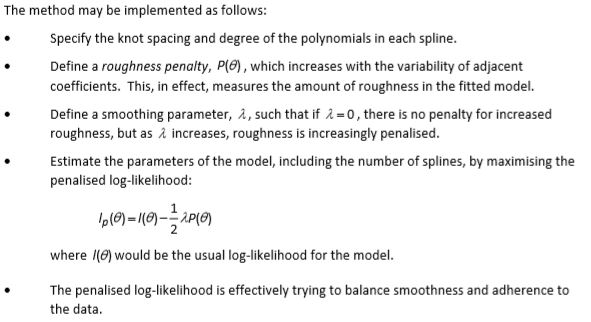
* The trend in mortality over time is unlikely to follow a quadratic function, even after it has been log-transformed, as in this model, because the progression of predicted values is likely to be too smooth.
* There may be significant variations in the trends in the past data that may be relevant to future projections and which we would therefore like the model to take into account.
* Spline functions are very flexible models in terms of the shape of the function being fitted.
* Adherence to data can be improved both by increasing the number of knots used, and by placing the knots in locations where the greatest changes in curvature of the trend line occur.
* However, some smoothing is still a requirement, and using cubic splines generally produces the smoothest result (compared to using splines of higher orders).

**Disadvantages of using p-splines**

* When applied to ages separately, mortality at different ages is forecast independently so there is a danger that there will be roughness between adjacent ages.
* There is no explanatory element to the projection (in the way that time series methods use a structure for mortality and an identifiable time series for projection).
* P-splines tend to be over-responsive to adding an extra year of data.

**Use of p-splines**

* The problem with splines is that they can be too flexible, and may cause the model to include historical trend variations that are either short-term or past-specific, and which are not expected to recur in future.
* To include these features in the model may then be inappropriate or unhelpful when we attempt to use the model for forecasting purposes.
* One symptom of this over-adherence, or roughness, in the model, is that the sequence of estimated parameters θ1-hat, θ2-hat, ..., θJ-hat may form an uneven progression, and smoothing this progression can help reduce the roughness in the predicted values from the model.
* The method of p-splines attempts to find an optimal model by introducing a penalty for models which have excessive roughness.

**P-splines implementation as follows -**

**Government of small country – mortality over the next 20 years based on past 10 years standard tables – suggestion of exponential curve to the time trend.**

* The use of an exponential curve is attractive as there is evidence that age-specific mortality has declined exponentially in some past periods.
* The approach is simple to understand and easy to implement.
* However, fitting separate curves at each age risks the projected future mortality rates in any given year not progressing smoothly with age (and even decreasing with age in age ranges where this is implausible) i.e. under-graduated rates.
* This problem could be overcome by graduating the projected rates or by using an alternative method/model in the first place.
* The approach assumes that developments in medical technology, lifestyle, etc. in the future will progress steadily as they have in the past 10 years.
* The appropriateness of this projection method may depend on whether the past history displays an exponential change over time.
* Using cohorts to project mortality instead of time period may lead to improvements in the reliability of the projection.
* It could be argued that 10 years of historic Life tables may not be sufficient to provide a reliable projection of future mortality.

**Alternative Model -** Lee-Carter model, age-cohort model, age-period-cohort model, penalised splines, decomposition of mortality by cause of death / Explanatory method; OR Adjust rates using projected rates from a similar country.

**Suggest three types of information source which could be used in recommending parameters to use in an actuarial model. (Sept 2018)**

* Internal data for an old model performing the same/similar function.
* Internal data for a model performing a different function
* Market observable yields or rates.
* Expert opinion.
* Industry data, for example a standard table or surveys.
* Regulations set out by regulatory authorities.
* Government statistical data.

**Comment on a practical difficulty which could arise with using each type of information source. (Sept 2018)**

* Internal data for an old model performing the same/similar function: Past experience may not be representative of future experience.
* Internal data on a model performing a different function: Subjective adjustments may be needed.
* Market observable yields or rates: May be a time delay before they become available, OR different sources may give slightly different rates for the same item.
* Expert opinion: May be hard or expensive to find a relevant expert, OR the expert’s advice may be theoretical and hard to adapt into a pragmatic model.
* Industry data, for example a standard table or surveys: May not be directly relevant to the situation to be modelled, OR survey may be expensive OR the experience likely to differ by firm due to distribution approaches/target markets, etc., OR Firms for whom the assumption is insignificant may take a high-level approach and this may not be readily apparent.
* Regulations set out by regulatory authorities: Current regulations may change in the future
* Government statistical data will tend to apply to the population as a whole, and the model may apply to a non-representative subset of the population.