Theory of Inference - Notes

Dom Hutchinson

February 24, 2020

Contents

1 Motivation

Remark 1.1 - *General Idea*
Learn something about the world using data & statistical models.

Definition 1.1 - *Statistical Models*
Statistical Models describe the way in which data is generate. They depend upon *unknown* constant parameters, $\boldsymbol{\theta}$, and subsidiary information (known data & parameters).

Definition 1.2 - *Parameteric Statistical Inference*
Parameteric Statistical Inference is the process of taking some data & learning the *unknown* parameters of the model which generated it.
Definition 1.3 - *Parameteric Models*
A *Parameteric Model* is a statistical model whose pdf depends on some unknown parameter.
A *Semi-Parameteric Models* is a statistical models which contains unknown functions, as well as unknown parameters.
A *Non-Parameteric Model* has no parameters and thus makes minimal assumptions about how the data was generated.

Proposition 1.1 - *Inferential Questions*
When performing *Statistical Inference* we wish to answer the following questions

i) *Confidence Intervals & Credible Intervals* - What range of parameter valeus are consistent with the data?

ii) *Hypothesis Testing* - Are some pre-specified valeus (or restrictions) for the parameters consistent with the data?

iii) *Model Checking* - Could our model have generated the data at all?

iv) *Model Selection* - Which of several alternative odels could most plausibly have generated the data?

v) *Statistical Design* - How could we better arrange teh data gathering process to improve the answers to the preceding questions?

1.1 Examples

Example 1.1 - *Mean Annual Temperatures*
Consider a dataset of the mean annual temperature in New Haven, Conneticut.
Suppose we plot it in a histogram & notice that it fits a bell curve, then we may assume the data fits a simple model where each data point is observed independently from a $\mathcal{N}(\mu, \sigma^2)$ distribution with μ, σ^2 unknown.
Then the pdf for each data point, y_i, is

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

The pdf for the whole data set, $\mathbf{y}$, is the joint pdf of each data point since we assume iid

$$f(\mathbf{y}) = \prod_{i=1}^{N} f(y_i)$$

Now suppose we notice that the histogram is *heavy tailed* relative to a normal distribution.
A better model might be

$$\frac{y_i - \mu}{\sigma} \sim t_\alpha$$

where μ, σ^2, α are unknown.
This means the pdf of the whole data set is

$$f(\mathbf{y}) = \prod_{i=1}^{N} \frac{1}{\sigma} f_{t_\alpha}\left(\frac{y_i - \mu}{\sigma}\right)$$

by *standard transformation theory.*

Example 1.2 - *Hourly Air Temperature*
Consider a dataset of the air temperature, a_i, measured at hourly intervals, t_i, over the course of a week.
The temperature is believed to follow a daily cycle, with a long-term dift over the course of the week and to be subject to random autocorrelated depatures from this overall pattern.
A suitable model might be

$$a_i = \underbrace{\theta_0 + \theta_1 t_i}_{\text{Long-Term Drift}} + \underbrace{\theta_2 \sin(2\pi t_i/24) + \theta_3 \cos(2\pi t_i/24)}_{\text{Daily Cycle}} + \underbrace{e_i}_{\text{Auto Correlation}}$$

where $e_{i+1} := \rho r_i + \varepsilon_i$ with $|\rho| < 1$ & $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
This means $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ & $\mathbf{a} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\Sigma_{i,j} = \frac{\rho^{|1-j|}\sigma^2}{1-\rho}$.
Thus, the pdf of the data set, $\mathbf{a}$, is

$$f(\mathbf{a}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(\mathbf{a}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{a}-\boldsymbol{\mu})}$$

Example 1.3 - *Bone Marrow*
Consider a dataset produced 23 patients suffering from non-Hodgkin's Lymphoma are split into two groups, each recieving a different treatment. We wish to test whether one of these treatments is more efficitive than the other.
For each patient the days between treatment & relapse was recorded. We have some *censored data* as the patient had not relapsed by the time of their last appointment.
Consider using an exponential distribution to model the times to relapse with parameters θ_a & θ_b respecitvely. We want to test if $\theta_a = \theta_b$.
We have the follow pdf for patients in group a

$$f_a(t_i) = \begin{cases} \theta_a e^{-\theta_a t_i} & \text{uncensored} \\ \int_{t_i}^{\infty} \theta_a e^{-\theta_a t_i} = e^{-\theta_a t_i} & \text{censored} \end{cases}$$

An equivalent pdf exists for patients in group b, with θ_b swapped in.
Thus the model for the whole data set, $\mathbf{t}$, is

$$f(\mathbf{t}) = \prod_{i=1}^{11} f_a(t_i) \prod_{i=12}^{23} f_b(t_i)$$

when patients $\{1, \dots, 11\}$ are in group a and the rest in group b.

2 Basic Approaches to Inference

Definition 2.1 - *Frequentist Approach*
In the *Frequentist Approach* to inference we assume the model parameters are fixed states, which we wish to estimate. The parameter estimator $\hat{\theta}$ is a random variable which inherits its randomnewss from the data which it is constructed from.

Definition 2.2 - *Bayesian Approach*
In the *Bayesian Approach* to inference model parameters are treated as random variables and we use probability distributions to encode our uncertainty about the parameters. We set a prior distribution, $\mathbb{P}(\theta)$, and then use data to update it and learn a posterior distribution, $\mathbb{P}(\theta|\mathbf{x})$.

Remark 2.1 - *Assumptions*
Often we are required to make assumptions in order to analyse the results these approaches. For the *Frequentist Approach* we often assume we have a large data set, whilst for the *Bayesian Approach* we produce simulations from the posterior.

Example 2.1 - *Comparing Frequentist & Bayesian Approach*
Let $X_1,\dots,X_n \overset{\text{iid}}{\sim} \text{Normal}(\mu,1)$ where μ is an unknown parameter we wish to learn.
Let $\mathbf{x} := \{x_1,\dots,x_n\}$ be a realisation of $\mathbf{X}$.

Frequentist Let's use $\hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i$.
Consider the expectation and variance of $\hat{\mu}$

$$\mathbb{E}(\hat{\mu})=\mathbb{E}(\bar{x})=\frac{1}{n}\sum_{i=1}^n\mathbb{E}(X_i)=\mu \text{ and } \text{Var}(\hat{\mu})=\text{Var}(\bar{x})=\frac{1}{n^2}\sum_{i=1}^n\text{Var}(X_i)=\frac{1}{n}$$

Since $\hat{\mu}$ is a linear transformation of normal random variables it has a normal random variable, thus

$$\hat{\mu}\sim\text{Normal}\left(\mu,\frac{1}{n}\right)$$

Thus $\hat{\mu}$ is an *unbiased* estimator of μ.
By noting that $\sqrt{n}(\hat{\mu}-\mu)\sim\text{Normal}(0,1)$ we can construct *Confidence Intervals* for μ

$$\begin{array}{rcl} 0.95 &=& \mathbb{P}(-1.96<\sqrt{n}(\hat{\mu}-\mu)<1.96)\\ \implies 0.95 &=& \mathbb{P}\left(\hat{\mu}-\frac{1.96}{\sqrt{n}}<\mu<\hat{\mu}+\frac{1.96}{\sqrt{n}}\right)\end{array}$$

Bayesian Here we treat μ as a random variable and thus must choose a distribution for it

$$\mu\sim\text{Normal}(0,\sigma_\mu^2)$$

where σ_μ^2 is a value we set. Generally we choose greater values for the variance when we are less certain.
We want to find $\mathbb{P}(\mu|\mathbf{x})$ and note that *Bayes' Rule* states

$$\mathbb{P}(\mu|\mathbf{x})=\frac{\mathbb{P}(\mathbf{x}|\mu)\mathbb{P}(\mu)}{\mathbb{P}(\mathbf{x})}$$

In this setting $\mathbb{P}(\mathbf{x})$ is intractable so we use a trick that since $\mathbb{P}(\mathbf{x})$ is a normalising factor we have

$$\mathbb{P}(\mu|\mathbf{x})\propto\mathbb{P}(\mathbf{x}|\mu)\mathbb{P}(\mu)$$

From this proportionality we aim to identity the distribution of $\mathbb{P}(\mu|\mathbf{x})$.

$$\begin{aligned}
\mathbb{P}(\mu|\mathbf{x}) &\propto \exp\left\{-\frac{1}{2\sigma_\mu^2}\sum_{i=1}^n[(x_i-\mu)^2+\mu^2]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(-2n\bar{x}\mu+\frac{\mu^2(n\sigma_\mu^2+1)}{\sigma_\mu^2}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{n\sigma_\mu^2+1}{\sigma_\mu^2}\right)\left(\mu^2-2\bar{x}\mu\frac{n\sigma_\mu^2}{n\sigma_\mu^2+1}\right)\right\}\\
&\propto \exp\Bigg\{-\frac{1}{2}\underbrace{\left(\frac{n\sigma_\mu^2+1}{\sigma_\mu^2}\right)}_{1/\sigma^2}\underbrace{\left(\mu-\bar{x}\frac{n\sigma_\mu^2}{n\sigma_\mu^2+1}\right)^2}_{\mu}\Bigg\} \text{ by completing the square}
\end{aligned}$$

We can produce a *Credible Interval* for μ as

$$\bar{x}\frac{n\sigma_\mu^2}{n\sigma_\mu^2+1}\pm 1.96\frac{\sigma_m u}{\sqrt{n\sigma_\mu^2+1}}$$

If we consider the final distribution from the *Bayesian Approach* as $n\to\infty$ we notice that

$$\mu|\mathbf{x}\to\bar{x}=\hat{\mu}\quad\text{and}\quad\sigma^2|\mathbf{x}\to\frac{1}{n}$$

2.1 Inference by Resampling

Remark 2.2 - *Motivation*
The uncertainty we have about a parameter is inherited from the uncertainty in the data sampling process. Often we have a data set & are unable to repeat the data gathering process, and even if we could we would just combine it into a larger sample rather than split it.

Definition 2.3 - *Resampling*
Let $\mathbf{x}$ be a given data set.
We can *Resample* from $\mathbf{x}$ be sampling values in $\mathbf{x}$ uniformly, with repetition. Since we use repetition the *Resample*'s size is independent of the size of $\mathbf{x}$ (Although it makes little sense for it to be greater than $|\mathbf{x}|$).

Definition 2.4 - *Bootstrapping*
Bootstrapping is the process of generating multiple *Resamples* of a data set & then estimating a parameter value for each of these *resamples*. These estimated values can then be assessed.

Example 2.2 - *Bootstrapping*
The algorithm below describes how to perofrm a *Bootstrapping* operation for the mean of a given data set $\mathbf{x}$. It produces m *resamples* of size n from $\mathbf{x}$ and returns a 95% *Confidence Interval* for the estimated means of these samples.

Algorithm 1: Estimating Mean using Bootstrapping

```
require: x {data set}
μs = {} {resample means}
μs append mean(x)
for i = 0...m do
    x_i ← sample(x, n, replace=TRUE)
    μs append mean(x_i)
return quantile(μs, (0.025, 0.0975)
```

3 Inference for Linear Models

Definition 3.1 - *Linear Model*
A *Linear Model* is a mathematical model where the *response vector*, $\mathbf{y}$, is linear wrt some parameters $\boldsymbol{\beta}$ and zero-mean *random error* $\boldsymbol{\varepsilon}$.

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where X is the *Model Matrix* (*i.e.* observed data).
Usually we assume $\boldsymbol{\varepsilon} \sim \text{Normal}(0, \sigma^2 I)$ although the normality assumption is less important as the *Central Limit Theorem* typically takes care of any issues.

Definition 3.2 - *Model Matrix*
A *Model Matrix*, X, is the set of values observed in a system. Rows are read as a single observation & columns as a single *Predictor Varaible*.
The *Predictor Variables* fulfil one of the following roles

- *Metric* - Quantifable measurement from the system.
- *Factor* - A categorisation. Typically take the a binary value $(0, 1)$ to represent whether an observation fits a given category or not.

Remark 3.1 - *Only the parameters of a Linear model need to be linear. The predictor variables can be composed in any way deemed fit.*
$y = \alpha x^2 + \varepsilon$ is valid but $y = \alpha^2 x + \varepsilon$ is not.

Example 3.1 - *Formulating Linear Model*
The following is a linear model for a system with *Metrics* x_i & z_i and *Factor* g_i.

$$y_i = \gamma_{g_i} + \alpha_1 x_i + \alpha_2 z_i + \alpha_4 z_i^2 + \alpha_4 z_i x_i + \varepsilon_i$$

where γ_{g_i} is the parameter for category represented by g_i.
We can describe the system about in terms of matrices

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & x_1 & z_1 & z_1^2 & z_1 x_1 \\ 0 & 0 & 1 & x_2 & z_2 & z_2^2 & z_2 x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_n & z_n & z_n^2 & z_n x_n \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 + \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

In the above formulation y_1 fulfils category 1, y_2 fulfils 3 and y_n fulfils 2.

Example 3.2 - *Linear Model*
Consider a data set for the stopping *distance* of a car with *Predictor Variable speed* at the point at which the signal to stop is given.
By considering basic physics we can theorise the following model

$$\begin{aligned} distance_i &= \beta_1 speed_i + \beta_2 speed_i^2 + \varepsilon_i \\ &= \text{Thinking} + \text{Loss Kinetic Energy} + \text{Error} \end{aligned}$$

where $\varepsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$.
Suppose we want to test whether to make the model more flexible. We can theorise the following model & test whether $\beta_0 = 0 = \beta_3$ (as expected).

$$distance_i = \beta_0 + \beta_1 speed_i + \beta_2 speed_i^2 + \beta_3 speed_i^3 + \varepsilon_i$$

3.1 Linear Model Estimation & Checking

Proposition 3.1 - *Frequentist Approach*
In the *Frequentist Approach* to *Linear Models* we assume that $\boldsymbol{\beta}$ and σ^2 are fixed states of nature, although they are unknown to us, and all randomness is inherited from the random variability in the data. We want to find a point estimate for $\boldsymbol{\beta}$ which minimises the *Residual Sum of Squares.*

Definition 3.3 - *Residual Sum of Squares*
Let $(X, \mathbf{y})$ be a set of training data & $\boldsymbol{\beta}$ a *Parameter Vector.*
The *Residual Sum of Squares* is the square difference between our estimate for the *Response Variable* and its true value.

$$S := \sum_{i=1}^{n}(y_i - \mu_i)^2 = \|\mathbf{y} - \boldsymbol{\mu}\|^2 \text{ where } \boldsymbol{\mu} = X\boldsymbol{\beta}$$

Proposition 3.2 - *Least Squares for Linear Model*
From the definition of *Residual Sum of Squares* as the *Euclidian Distance* between the response & estimated vectors we note that its value is unchanged if we reflect or rotate $(\mathbf{y} - \boldsymbol{\mu})$.
Next we note that any real matrix, $X \in \mathbb{R}(n \times p)$, can be decomposed into

$$X = \mathcal{Q}\begin{pmatrix} R \\ 0 \end{pmatrix} = QR \text{ note that } \mathcal{Q} \neq Q$$

where $R \in \mathbb{R}(p \times p)$ is an *Upper Triangular Matrix* and $\mathcal{Q} \in \mathbb{R}(n \times n)$ is an *Orthogonal Matrix*, the first p columns of which form Q.
Since $\mathcal{Q}$ is *Orthogonal* we have that $\mathcal{Q}^T\mathcal{Q} = I$.
We can now derive the result that

$$\begin{array}{rcl}
\|\mathbf{y} - X\boldsymbol{\beta}\|^2 & = & \|\mathcal{Q}^T\mathbf{y} - \mathcal{Q}^TX\boldsymbol{\beta}\|^2 \\
& = & \left\|\mathcal{Q}^T\mathbf{y} - \begin{pmatrix} R \\ 0 \end{pmatrix}\boldsymbol{\beta}\right\|^2 \\
& = & \left\|\begin{pmatrix} \mathbf{f} \\ \mathbf{r} \end{pmatrix} - \begin{pmatrix} R \\ 0 \end{pmatrix}\right\|^2 \text{ where } \begin{pmatrix} \mathbf{f} \\ \mathbf{r} \end{pmatrix} \equiv \mathcal{Q}^T\mathbf{y} \\
& = & \|\mathbf{f} - R\boldsymbol{\beta}\|^2 + \|\mathbf{r}\|^2
\end{array}$$

Thus minimising the *Residual Sum of Squares* is reduced to choosing $\boldsymbol{\beta}$ st $R\boldsymbol{\beta} = \mathbf{f}$.
Hence, provided that X and R have full rank

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = R^{-1}\mathbf{f}$$

N.B. After choosing $\boldsymbol{\beta}$ we have that the *Residual Sum of Squares* is just $\|\mathbf{r}\|^2$.
Proposition 3.3 - $\hat{\boldsymbol{\beta}}_{LS}$ *is Unbiased*
We have that

$$\begin{array}{rcl}
\mathbb{E}(\hat{\boldsymbol{\beta}}) & = & \mathbb{E}(\mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}) \\
& = & \mathbf{R}^{-1}\mathbf{Q}^T\mathbb{E}(\mathbf{y}) \\
& = & \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{X}\boldsymbol{\beta} \\
& = & \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{Q}\mathbf{R}\boldsymbol{\beta} \\
& = & \boldsymbol{\beta}
\end{array}$$

Thus $\hat{\boldsymbol{\beta}}_{\text{LS}}$ is unbiased.

Proposition 3.4 - *Variance of* $\hat{\boldsymbol{\beta}}_{LS}$
We have $\Sigma_{\mathbf{y}} = I\sigma^2$.

Thus $\Sigma_{\mathbf{f}} = \mathbf{Q}^T\mathbf{Q}\Sigma_{\mathbf{y}} = \mathbf{Q}^T\mathbf{Q}I\sigma^2 = I\sigma^2$.
Hence
$$\Sigma_{\hat{\beta}} = \mathbf{R}^{-1}\mathbf{R}^{-T}\sigma^2$$

Remark 3.2 - *Checking*
In order to make inferences beyond estimating β we need to check that our assumptions about ε_i still hold.
We can estimate these values as $\hat{\varepsilon}_i = y_i - \hat{\mu}_i$ where $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.
Plotting these estimates, $\hat{\varepsilon}_i$, against fitted values, $\hat{\mu}_i$, allows us to look for systematic patterns in the mean of residuals, which would indicate a violation of the independence assumption

3.2 Gauss-Markov Theorem

Remark 3.3 - *Alternatives to Least-Squares Estimates*

- We may wish to find an estimate of β which is as close to the real value as possible, so minimising $\|\hat{\beta}-\beta\|^2$. However it is possible the data gives a lot of information about β_i but little about β_j, does it make sense to weight these equally.
- We could only allow *unbiased estimators*, ie $\mathbb{E}(\hat{\beta}) = \beta$. And then among those choose the one with least variance.

Theorem 3.1 - *Gauss-Markov Theorem*
Define $\boldsymbol{\mu} := \mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\Sigma_y = \sigma^2 I$.
Let $\tilde{\phi} = \mathbf{c}^T\mathbf{Y}$ be any unbiased linear estimator of $\phi = \mathbf{t}^T\boldsymbol{\beta}$ where $\mathbf{t}$ is an arbitrary vector. Then
$$\text{Var}(\tilde{\phi}) \geq \text{Var}(\hat{\phi}) \text{ where } \hat{\phi} = \mathbf{t}^T\hat{\boldsymbol{\beta}}_{\text{LS}} \ \&\ \hat{\boldsymbol{\beta}}_{\text{LS}} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{Y}$$
Since $\mathbf{t}$ is arbitraru, this implies that each element of $\hat{\boldsymbol{\beta}}$ is a minimum variance unbiased estimator.

Proof 3.1 - *Gauss-Markov Theorem*
Since $\tilde{\phi}$ is a linear transformation of $\mathbf{Y}$, $var(\tilde{\phi}) = \mathbf{c}^T\mathbf{c}\sigma^2$.
To compare the variances of $\hat{\phi}$ and $\tilde{\phi}$ it is useful to express $\text{Var}(\hat{\phi})$ in terms of $\mathbf{c}$.
Because $\tilde{\phi}$ is unbiased we have
$$\begin{array}{rrcl} & \mathbb{E}(\mathbf{c}^T\mathbf{Y}) & = & \mathbf{t}^T\boldsymbol{\beta} \\ \implies & \mathbf{c}^T\mathbb{E}(\mathbf{Y}) & = & \mathbf{t}^T\boldsymbol{\beta} \\ \implies & \mathbf{c}^T\mathbf{X}\boldsymbol{\beta} & = & \mathbf{t}^T\boldsymbol{\beta} \\ \implies & \mathbf{c}^T\mathbf{X} & = & \mathbf{t}^T \end{array}$$
So the variance of $\hat{\phi}$ can be written as
$$\text{Var}(\hat{\phi}) = \text{Var}(\mathbf{t}^T\hat{\boldsymbol{\beta}}) = \text{Var}(\mathbf{c}^T\mathbf{X}\hat{\boldsymbol{\beta}}) = \text{Var}(\mathbf{c}^T\mathbf{Q}\mathbf{R}\hat{\boldsymbol{\beta}})$$
This is the variance of a linear transformation of $\hat{\boldsymbol{\beta}}$ and the covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\mathbf{R}^{-1}\mathbf{R}^{-T}\sigma^2$.
Thus
$$\text{Var}(\hat{\phi}) = \text{Var}(\mathbf{c}^T\mathbf{Q}\mathbf{R}\hat{\boldsymbol{\beta}}) = \mathbf{c}^T\mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{R}^{-T}\mathbf{R}^T\mathbf{Q}^T\mathbf{c}^T\sigma^2 = \mathbf{c}^T\mathbf{Q}\mathbf{Q}^T\mathbf{c}\sigma^2$$
Hence
$$\text{Var}(\tilde{\phi}) - \text{Var}(\hat{\phi}) = \mathbf{c}^T(I - \mathbf{Q}\mathbf{Q}^T)\mathbf{c}\sigma^2$$
Because the columns of $\mathbf{Q}$ are orthogonal, $\mathbf{Q}\mathbf{Q}^T = \mathbf{Q}\mathbf{Q}^T\mathbf{Q}\mathbf{Q}^T$ it follows that
$$\mathbf{c}^T(I - \mathbf{Q}\mathbf{Q}^T)\mathbf{c} = [(I - \mathbf{Q}\mathbf{Q}^T)\mathbf{c}]^T(I - \mathbf{Q}\mathbf{Q}^T)\mathbf{c} \geq 0$$
since this is just the sum of squares of the elements of teh vector $(I - \mathbf{Q}\mathbf{Q}^T)\mathbf{c}$. $\square$

Remark 3.4 - *Least Squares Variance*
Amongst unbiased and linear estimators in $\mathbf{Y}$, least squares estimators have minimum variance. It is still possible that some non-linear estimator might be even better.

3.3 Further Inference on Linear Models

Remark 3.5 - *Requirements*
In order to make further inferences about linear models (*e.g.* confidence intervals & hypothesis testing) we need to make our model completely probabilistic, since these inferences are probabilistic concepts.
This requires us to specify a full distribution for the error $\boldsymbol{\varepsilon}$.
We assume

$$\begin{array}{rcl} \boldsymbol{\varepsilon} & \overset{\text{iid}}{\sim} & \text{Normal}(0, I\sigma^2) \\ \implies \mathbf{y} & \sim & \text{Normal}(\mathbf{X}\beta, I\sigma^2) \\ \implies \hat{\boldsymbol{\beta}} & \sim & \text{Normal}(\boldsymbol{\beta}, \Sigma_{\hat{\beta}}) \\ \text{where } \Sigma_{\hat{\beta}} & = & R^{-1}R^{-T}\sigma^2 \end{array}$$

Theorem 3.2 - $\dfrac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim t_{n-p}$

Proof 3.2 - *Theorem 3.2*
$\boldsymbol{\mathcal{Q}}^T\mathbf{y}$ is a linear transformation of a normal random vector, so is a normal random vector with covariance matrix

$$\Sigma_{\boldsymbol{\mathcal{Q}}^T\mathbf{y}} = \boldsymbol{\mathcal{Q}}^T I \boldsymbol{\mathcal{Q}}\sigma^2 = I\sigma^2$$

The elements of $\boldsymbol{\mathcal{Q}}^T\mathbf{y}$ are mtually independent. Further

$$\begin{array}{rcl} \mathbb{E}\left[\begin{pmatrix}\mathbf{f}\\ \mathbf{r}\end{pmatrix}\right] & = & \mathbb{E}[\boldsymbol{\mathcal{Q}}^T\mathbf{y}) \\ & = & \boldsymbol{\mathcal{Q}}^T\mathbf{X}\boldsymbol{\beta} \\ & = & \begin{pmatrix}\mathbf{R}\\ \mathbf{0}\end{pmatrix}\boldsymbol{\beta} \\ \implies \mathbb{E}(\mathbf{f}) & = & \mathbf{R}\boldsymbol{\beta} \\ \text{and } \mathbb{E}(\mathbf{r}) & = & \mathbf{0} \end{array}$$

Thus

$$\mathbf{f} \sim \text{Normal}(\mathbf{R}\boldsymbol{\beta}, I_p\sigma^2) \text{ and } \mathbf{r} \sim \text{Normal}(0, I_{n-p}\sigma^2)$$

Now we can deduce

$$\begin{array}{rrcl} & r_i & \overset{\text{ind}}{\sim} & \text{Normal}(0, \sigma^2) \\ \implies & \frac{r_i}{\sigma} & \sim & \text{Normal}(0, 1) \\ \implies & \displaystyle\sum_{i=1}^{n-p}\left(\frac{r_i}{\sigma}\right)^2 & \sim & \chi^2_{n-p} \end{array}$$

Since $\mathbb{E}(\chi^2_{n-p}) = n - p$ we have that $\hat{\sigma}^2 = \frac{1}{n-P}\|\mathbf{r}\|^2$ is an unbiased estimator.
Let $\sigma_{\hat{\beta}_i} = \sqrt{\Sigma_{\hat{\beta}_i}(i,i)}$ then $\hat{\sigma}_{\hat{\beta}_i} = \sqrt{\hat{\Sigma}_{\hat{\beta}_i}(i,i)}$ but $\hat{\Sigma}_{\hat{\beta}_i} = \Sigma_{\hat{\beta}_i}\frac{\hat{\sigma}^2}{\sigma^2} \implies \hat{\sigma}_{\hat{\beta}_i}\frac{\hat{\sigma}}{\sigma}$.
Consider

$$\begin{array}{rcl} \dfrac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\beta_i}} & = & \dfrac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}\hat{\sigma}/\sigma} \\ & = & \dfrac{(\hat{\beta}_i - \beta_i)/\sigma_{\hat{\beta}_i}}{\sqrt{\frac{1}{\sigma^2}\|\mathbf{r}\|^2/(n-p)}} \\ & \sim & \dfrac{\text{Normal}(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \\ & \sim & t_{n-p} \end{array}$$

Proposition 3.5 - *Confidence Intervals for β_i*
Supose we want a $(1-2\alpha)100\%$ confidence interval for β_i.
Then

$$\begin{aligned}\mathbb{P}\left(-t_{n-p}(\alpha) < \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} < t_{n-p}(\alpha)\right) &= \mathbb{P}\left(\hat{\beta}_i - t_{n-p}(\alpha)\sigma_{\hat{\beta}_i} < \beta_i < \hat{\beta}_i + t_{n-p}(\alpha)\sigma_{\hat{\beta}_i}\right) \\ &= 1-2\alpha\end{aligned}$$

where $\mathbb{P}(t_{n-p}(\alpha) \geq t_{n-p}) = 1-\alpha$.

3.4 Geometry of Linear Models

Remark 3.6 - *Least Squares Estimation as Geometry*
Least Squares Estimation of linear models is the same as finding the orthogonal projection of the response vector $\mathbf{y} \in \mathbb{R}^n$ onto the p-dimensional linear subspace spanned by the columns of $\mathbf{X} \in \mathbb{R}^{n\times p}$.
By the linear model $\mathbb{E}(\mathbf{y})$ lies in the space spanned by all possible linear combinations of the columns of $\mathbf{X}$ & least squares find the point in that space that is cloests to $\mathbf{y}$ in *Euclidean Distance*.

Remark 3.7 - *Projection Matrix*
Consider the *Projection Matrix* that maps the response data $\mathbf{y}$ to the fitted values $\hat{\boldsymbol{\mu}}$.
We have that

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} = \mathbf{Q}\mathbf{Q}^T\mathbf{y}$$

Thus the projection matrix is $\mathbf{A} = \mathbf{Q}\mathbf{Q}^T$.
N.B. Often $\mathbf{A}$ is referred to as the *Influence Matrix* or *Hat Matrix*.

Proposition 3.6 - *Projection Matrix Idempotent*
Let $\mathbf{A}$ be the *Projection Matrix* of a *Linear Model*.
$\mathbf{A}$ is said to be *Idempotent* since $\mathbf{A} = \mathbf{A}\mathbf{A}$.
This is since the orthogonal projection of $\hat{\boldsymbol{\mu}}$ onto the column space of $\mathbf{X}$ must be $\hat{\boldsymbol{\mu}}$.

3.5 Results in terms of Model Matrix, X

Proposition 3.7 - *Results in terms of Model Matrix,* $\boldsymbol{X}$

$$\begin{aligned}\Sigma_{\hat{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\ &= (\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R})^{-1}\sigma^2 \\ &= (\mathbf{R}^T\mathbf{R}^{-1}\sigma^2 \\ &= \mathbf{R}^{-1}\mathbf{R}^{-T}\sigma^2\end{aligned} \qquad \begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{R}^{-1}\mathbf{R}^{-T}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\ &= \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} \\ &= \mathbf{R}^{-1}\mathbf{f}\end{aligned} \qquad \mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

3.6 Bayesian Analysis

Remark 3.8 - *Bayesian Analysis of Linear Models*
To perfor a full *Bayesian Analysis* of a *Linear Model* we need to define prior distributions for $\boldsymbol{\beta}$ and σ^2. Typically In order to make this problem analytically tractable we use conjugate priors.
Conjugacy can be used for defining

$$\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\beta}_0, \boldsymbol{\psi}^{-1}) \quad \text{and} \quad \tau \sim \Gamma(a,b)$$

where $tau := \frac{1}{\sigma^2}$ is precision measure.
Here $a, b, \boldsymbol{\beta}_0$ and $\boldsymbol{\psi}$ are quantities which we need to define values for, for practical analysis.
This gives us the following distributions

$$\begin{array}{rcl} f(\mathbf{y}, \boldsymbol{\beta}, \tau) & \propto & \tau^{n/2} e^{-\frac{\tau}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2} e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{\psi}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)} e^{-b\tau}\tau^{a-1} \\ f(\tau|\boldsymbol{\beta}, \mathbf{y}) & \propto & \tau^{\frac{n}{2}+a-1} e^{-\tau(b+\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2)} \\ & \sim & \Gamma(\dfrac{n}{2}+a, b+\dfrac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2) \\ f(\boldsymbol{\beta}|\tau, \mathbf{y}) & \propto & e^{-\frac{1}{2}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\tau - 2\beta\mathbf{X}^T\mathbf{y}\tau + \boldsymbol{\beta}^T\boldsymbol{\psi}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\psi}\boldsymbol{\beta}_0)} \\ & \propto & e^{-\frac{1}{2}[\boldsymbol{\beta}-(\mathbf{X}^T\mathbf{X}\tau+\boldsymbol{\psi})^{-1}(\tau\mathbf{X}^T\mathbf{y}+\boldsymbol{\psi}\boldsymbol{\beta}_0)]^T(\mathbf{X}^T\mathbf{X}\tau+\boldsymbol{\psi})\boldsymbol{\beta}-(\mathbf{X}^T\mathbf{X}\tau+\boldsymbol{\psi})^{-1}(\tau\mathbf{X}^T\mathbf{y}+\boldsymbol{\psi}\boldsymbol{\beta}_0)]} \\ & \sim & \text{Normal}[(\mathbf{X}^T\mathbf{X}\tau+\boldsymbol{\psi})^{-1}(\tau\mathbf{X}^T\mathbf{y}+\boldsymbol{\psi}\boldsymbol{\beta}+0), (\mathbf{X}^T\mathbf{X}\tau+\boldsymbol{\psi})^{-1}] \end{array}$$

If either the sample size tends to infinity (*i.e.* $n \to \infty$) or the prior precision matrix tends to the zero matrix then

$$f(\boldsymbol{\beta}|\tau, \mathbf{y}) \overset{\to}{\sim} \text{Normal}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

N.B. We have not produced the joint distribution $\boldsymbol{\beta}, \tau|\mathbf{y}$ but just two conditionals.

Remark 3.9 - *Proceeding from Conditionals*
There are a few options to proceed from the results in **Remark 3.8**

i) Iteratively find the posteior modes of $\boldsymbol{\beta}$ given teh estiamte mode of τ and the posterior mode of τ given the estimated modes of $\boldsymbol{\beta}$ until the mode of τ connverges.
Then plug this into the conditional density of $\boldsymbol{\beta}$.

ii) Integrate $\boldsymbol{\beta}$ out of $f(\tau|\boldsymbol{\beta}, \mathbf{y})$ to obtain the marginal likelihood $f(\tau|\mathbf{y})$ which can be maximised to find $\hat{\tau}$.
$\hat{\tau}$ can be plugged into $f(\boldsymbol{\beta}|\tau, \mathbf{y})$. *N.B.* Also known as *Empirical Bayes.*

iii) Alternate simulate of $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}|\tau, \mathbf{y})$ given tau with simulation from $f(\tau|\boldsymbol{\beta}, \mathbf{y})$, given the last simulated $\boldsymbol{\beta}$, to generate joint draws of τ & $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}, \tau|\mathbf{y})$.
N.B. Also known as *Gibbs Sampling.*

4 Causality, Confounding & Randomisation

Definition 4.1 - *Causality*
Causality is a problem in statistical inference where we wish to find out which variables affect a particular variable, and are mearly correlated. This is more difficult that other forms of inference, but is useful in many real world scenarios especially in science & economics.

Example 4.1 - *Causation & Correlation*
There is an observed correlation between birth rates in Europe & stork populations. There is no causation between the two, however it is likely that increased industrialisation led to the decrease in both since it lead to more healthcare for humans & less habitats for storks.

4.1 Controlled Experiments and Randomisation

Definition 4.2 - *Hidden Variables*
Hidden Variables are variables which likely effect a system but which we can/do not observed.

Definition 4.3 - *Randomisation*
Randomisation is the process of splitting subjects into different groups. Typically a control &

an active group. This is meant to break correlation between observed & hidden variables.

Remark 4.1 - *Hidden Variables*
Consider the scenario where we wish to test whether exercise influences fat mass. It is likely that ther are lots of other factors. These factors will correlate with both exercise & fat mass. By splitting subjects into a control & exercise groups at random we break the correlation of these other features but not that between fat mass & exercise. The other factors are now random error.

Proposition 4.1 - *Formalisation of Hidden Variables*
Consider the true model matrix (X, H) where X is the observed variables & H is the hidden variables. We assume that the columns of H have mean 0.
We have

$$\tilde{\beta}_X = (X^TX)^{-1}X^T\mathbf{y} \text{ for assumed model } y = X\beta_X + \varepsilon$$

If we knew H then we would have

$$\begin{pmatrix}\hat{\beta}_X\\ \hat{\beta}_H\end{pmatrix} = \begin{pmatrix}X^X & X^TH\\ H^TX & H^TH\end{pmatrix}^{-1}\begin{pmatrix}X^T\\ H^T\end{pmatrix}\mathbf{y} \text{ for true model } \mathbf{y} = X\beta_X + H\beta_H + \varepsilon$$

Since $X^TH \neq 0 \implies \tilde{\beta}_X \neq \hat{\beta}_X$.
The randomised allocation to groups is used to try and make $X^TH = 0$.

Remark 4.2 - *Randomised Tests*
Sometimes it is frowned upon (ethically) to perform random tests. Such as testing if high levels of alcohol consumtion is correlated with heart disease.
N.B. There is a reason China is becoming such an advanced country.

4.2 Instrumental Variables

Definition 4.4 - *Instrumental Variable*
An *Instrumental Variable*, Z, is used in regression analysis when there are *hidden variables* in the model. *Instrumental Variables* are correlated with the *Explanatory Variables*, X but uncorrelated with the error term $\mathbf{e}$ in the model $\mathbf{y} = X\beta + \mathbf{e}$.

Proposition 4.2 - *Without Instrumental Variables*
Consider the true model $\mathbf{y} = X\beta_X + H\beta_H + \boldsymbol{\varepsilon}$ with H being the hidden variables with columns centred at 0.
Suppose we wish to fit the model $\mathbf{y} = X\beta_X + \mathbf{e}$.
In this case $\mathbf{e} = H\beta_H + \varepsilon$ and likely does not fulfil the criteria of linear model random error.
We have

$$\begin{array}{rcl}\mathbb{E}(\hat{\beta}_X) & = & (X^TX)^{-1}X^T\begin{pmatrix}X & H\end{pmatrix}\begin{pmatrix}\beta_X\\ \beta_H\end{pmatrix}\\ & = & \beta_X + (X^TX)^{-1}X^TH\beta_H\\ & \neq & \beta_x \text{ since } X \perp \mathbf{e} \text{ and thus } X^TH \neq 0\end{array}$$

.

Proposition 4.3 - *With Instrumental Variables*
Let Z be an instrumental variable (*i.e.* it is correlated with X but not with H).
Assume that $\text{rank}(Z) \geq \text{rank}(X)$ and Z's columns are centred around 0.
Project X onto column space of Z

$$X \mapsto A_ZY \text{ where } \underbrace{A_Z = Z(Z^TZ)^{-1}Z^T}_{\text{Projection Matrix}}$$

Now use $A_Z X$ as the model matrix

$$\begin{aligned}
\hat{\beta}_X &= (X^T A_Z X)^{-1} X^T A_z \mathbf{y} \\
\mathbb{E}(\mathbf{y}) &= \begin{pmatrix} X & H \end{pmatrix} \begin{pmatrix} \beta_X \\ \beta_H \end{pmatrix} \\
\mathbb{E}(\hat{\beta}_X) &= (X^T A_Z X)^{-1} X^T A_Z X \beta_X + (X^T A_Z X)^{-1} X^T \underbrace{A_Z H}_{\approx 0} \beta_H \\
&= \beta_X + 0 \\
&= \beta
\end{aligned}$$

Thus this $\hat{\beta}_X$ is unbiased.
N.B. $A_Z H \approx 0$ since Z and H are uncorrelated.

5 Maximum Likelihood Estimation

Definition 5.1 - *Maximum Likelihood Estimation*
A *Maximum Likelihood Estimate* is the estimated value of a parameter which maximises some likelihood function, wrt observed data.

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \text{argmax}_\theta \ell(\boldsymbol{\theta}) \text{ where } \ell(\cdot) := \ln f(\mathbf{y}|\boldsymbol{\theta})$$

Remark 5.1 - *MLEs are only unbiased for large sample sizes*

Proposition 5.1 - *MLE - Frequentist Approach*
In the *Frequentist Approach* parameters are fixed states of natire and teh uncertainty comes from our estimates of these parameters.
We define the *Likelihood Function* to be the probability of observing certain values given the parameters have certain values

$$L(\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta}) \text{ where } \mathbf{y} \text{ is fixed}$$

Note that the natural log of the likelihood function is increasing wrt it so they have the same maximum.
Thus we define the *Maximum Likelihood Estimate* to be the set of parameters which maximise the *Log-Likelihood Function*

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \text{argmax}_\theta \ell(\boldsymbol{\theta}) \text{ where } \ell(\cdot) := \ln f(\mathbf{y}|\boldsymbol{\theta})$$

Remark 5.2 - *Effectiveness of MLE*
We have that

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} \underset{n\to\infty}{\sim} \text{Normal}(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}) \text{ where } \boldsymbol{\mathcal{I}}^{-1} := -\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right)$$

This is the best that can be achieved for an unbiased estimator.
Definition 5.2 - *Nested Models*
Two models are said to be *Nested Models* if one can be expressed as the other model subject to some restrictions, $\mathbf{R}$, on its parameters, $\boldsymbol{\theta}$, which are written as $R(\boldsymbol{\theta}) = \mathbf{0}$.

Proposition 5.2 -
Consider wishing to compare two nested models.
Let $\hat{\boldsymbol{\theta}}_0$ denote the MLE of $\boldsymbol{\theta}$ under the restrictions of the nested model.
We want to test $H_0 : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$.
If this null hypothesis is true then

$$2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0] \sim \chi^2_r \text{ where } r = |\text{restrictions}|$$

Example 5.1 -
Consider the following model with a single parameter β

$$y_i \sim \text{Poisson}(e^{\beta x_i})$$

Let $\{(x_1, y_1), \dots, (x_n, y_n)\}$.
For our model we have $f(y_i) = \dfrac{(e^{\beta x_i})^{y_i} e^{-e^{\beta x_i}}}{y_i!}$.
Thus

$$L(\beta) = \prod_{i=1}^{n} \frac{(e^{\beta x_i})^{y_i} e^{-e^{\beta x_i}}}{y_i!}$$

0 Reference

0.1 Definitions

Definition 0.1 - *Heavy Tailed*

Definition 0.2 - *Censored Data*

Definition 0.3 - *Upper Triangular Matrix*

Definition 0.4 - *Orthogonal Matrix*

Definition 0.5 - *p-Value*

Definition 0.6 - *Euclidean Distance*

0.2 Probability

Definition 0.7 - *Random Variable*
A *Random Variable* is a function from the sample space to the reals.

$$X : \Omega \to \mathbb{R}$$

Random Variables take a different value each time they are observed and thus we define distributions for the probability of them taking particular values.
Random Variables form the basis of models.

Definition 0.8 - *Cummulative Distribution*
The *Cummulative Distribution* function of a *Random Variable*, X, is the function $F_X(\cdot)$ st

$$\begin{array}{rcccl} F_X(\cdot) & : & \mathbb{R} \to [0,1] & & \\ F_X(x) & := & \mathbb{P}(X \leq x) & = & \displaystyle\sum_{i=-\infty}^{x} \mathbb{P}(X = i) \\ & & & = & \displaystyle\int_{-\infty}^{x} f_X(x)dx \end{array}$$

The *Cummulative Distribution* is a monotonic function.

Remark 0.1 - *Continuous Cummulative Distribution*
If a *Cummulative Distribution* is *continuous* then $F_X(X) \sim \text{Uniform}[0,1]$.

Proof 0.1 - *Remark 2.1*

$$\begin{array}{rcl} F(X) & = & \mathbb{P}(X \leq x) \\ & = & \mathbb{P}(F(X) \leq F(x)) \\ \implies \mathbb{P}(F(X) \leq u) & = & u \text{ if } F \text{ is continuous} \end{array}$$

Definition 0.9 - *Quantile Function*
The *Quantile Function* of a *Random Variable* is the inverse function of the *Cumulative Distribution.*

$$\begin{array}{rcl} F_X^-(\cdot) & : & [0,1] \to \mathbb{R} \\ F_X^-(u) & := & \min\{x : F(x) \geq u\} \end{array}$$

If a distribution has a computable *Quantile Function* then we are able to generate random variable values by sampling from a uniform distribution & then passing that value into the *Quantile Function*.

Definition 0.10 - *(Q-Q) Plot*
Consider a data set $\{x_1, \dots, x_n\}$.
A *(Q-Q) Plot* of this data set plots the ordered data set, $\{x_{(1)}, \dots, x_{(n)}\}$, against the theoretical quantiles $F^-\left(\frac{i-.5}{n}\right)$.
The close this line is to $y = x$ the more likely it is the data was generated by this *Cummulative Distribtion*.
N.B. AKA *Quantile-Quantile Plot*
Definition 0.11 - *Probabiltiy Mass Function*
A *Probability Mass Function* returns the probabiltiy of a discrete random variable taking a particular value.

$$\begin{array}{rcl} f_X(\cdot) & : & \mathbb{R} \to [0,1] \\ f_X(x) & := & \mathbb{P}(X = x) \end{array}$$

Definition 0.12 - *Probability Density Function*
Since the probabiltiy of a *Continuous Random Variable* taking a specific value is zero we cannot use the *Probability Mass Function.*

$$\begin{array}{rcl} f_X(\cdot) & : & \mathbb{R} \to [0,1] \\ \mathbb{P}(a \leq X \leq b) & = & \displaystyle\int_a^b f(x)dx \end{array}$$

N.B. $F_X'(x) = f(x)$ when $F_X'(\cdot)$ exists.

Definition 0.13 - *Joint Probabiltiy Density Function*
Let X &Y be *Random Variables.*
The *Joint Probabiltiy Density Function* of X and Y is the function $f_{X,Y}(x,y)$ st

$$\mathbb{P}((X,Y) \in \Omega) = \iint_\Omega f_{X,Y}(x,y)dxdy$$

N.B. This can be seen as evaluation Ω in the $X - Y$ plane.

Definition 0.14 - *Marginal Distribution*
Let X & Y be *Random Variables* with *Joint Probability Density* $f_{X,Y}(\cdot,\cdot)$.
We can find the *Marginal Distribution* of X by evaluating the $f_{X,Y}$ at each value wrt Y.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

Definition 0.15 - *Expected Value,* $\mathbb{E}$
The *Expected Value* of a *Random Variable*, X, is its mean value.

$$\begin{array}{rcll} \mathbb{E}(X) & := & \displaystyle\int_{-\infty}^{\infty} xf(x)dx & \text{[Continuous]} \\ \mathbb{E}(g(X)) & := & \displaystyle\int_{-\infty}^{\infty} g(x)f(x)dx & \\ & & & \\ \mathbb{E}(X) & := & \displaystyle\sum_{-\infty}^{\infty} xf(x) & \text{[Discrete]} \\ \mathbb{E}(g(X)) & := & \displaystyle\sum_{-\infty}^{\infty} g(x)f(x) & \end{array}$$

Remark 0.2 - *Linear Transformations of Expected Value*

$$\mathbb{E}(a+bX)=a+b\mathbb{E}(X) \text{ where } a,b\in\mathbb{R}$$

Remark 0.3 - *Expected Value of Composed Random Variables*
Let X & Y be *Random Variables*. Then

$$\mathbb{E}(X+Y)=\mathbb{E}(X)+\mathbb{E}(Y)$$

If X & Y are *independent*. Then

$$\mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y)$$

Proof 0.2 - *Remark 2.3*

$$\begin{array}{rcl}\mathbb{E}(X+Y) & = & \int(x+y)f_{X,Y}(x,y)dxdy\\ & = & \int xf_{X,Y}(x,y)dxdy+\int yf_{X,Y}(x,y)dxdy\\ & = & \mathbb{E}(X)+\mathbb{E}(Y)\\ \mathbb{E}(XY) & = & \int xyf_{X,Y}(x,y)dxdy\\ & = & \int xf_X(x)yf_Y(y)dxdy \text{ by independence}\\ & = & \int xf_X(x)dx\int yf_Y(y)dy\\ & = & \mathbb{E}(X)\mathbb{E}(Y)\end{array}$$

Definition 0.16 - *Variance,* σ^2
The *Variance* of a *Random Variable*, X, is a measure of its spread around its expected value.

$$\text{Var}(X)=\mathbb{E}[(X-\mathbb{E}(X))^2]=\mathbb{E}(X^2)-\mathbb{E}(X)^2$$

Remark 0.4 - *Linear Transformations of Variance*

$$\text{Var}(a+bX)=b^2\text{Var}(X) \text{ where } a,b\in\mathbb{R}$$

Proof 0.3 - *Remark 2.4*

$$\begin{array}{rcl}\text{Var}(a+bX) & = & \mathbb{E}[((a+bX)-(a-b\mu))^2]\\ & = & \mathbb{E}[b^2(X-\mu)^2]\\ & = & b^2\mathbb{E}[(X-\mu)^2]\\ & = & b^2\text{Var}(X)\end{array}$$

Definition 0.17 - *Co-Variance*
Co-Variance is a measure of the joint variability of two *Random Variables.*

$$\text{Cov}(X,Y):=\mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$$

N.B. If X & Y are independent then $\text{Cov}(X,Y)=0$ since $\mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y)$.
N.B. $\text{Cov}(X,Y)=\text{Cov}(Y,X)$.

Definition 0.18 - *Co-Variance Matrix,* Σ
Let $\mathbf{X}:=\{X_1,\dots,X_n\}$ be a set of random variables.
A *Co-Variance Matrix* describes the *Variance* & *Co-Variance* of each combination of *Random Variables* in $\mathbf{X}$.

$$\Sigma:=\mathbb{E}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T]$$

N.B. $\Sigma_{ii}=\text{Var}(X_i)$ & $\Sigma_{ij}=\text{Cov}(X_i,X_j)$ for $i\neq j$. Σ is symmetric.

Remark 0.5 - *Linear Transformation of Covariance*

$$\Sigma_{AX+b} = A\Sigma A^T$$

Proof 0.4 - *Remark 2.5*

$$\begin{array}{rcl}\Sigma_{AX+b} & = & \mathbb{E}[(AX+\mathbf{b}-A\boldsymbol{\mu}-\mathbf{b})(AX+\mathbf{b}-A\boldsymbol{\mu}-\mathbf{b})^T] \\ & = & \mathbb{E}[(AX-A\boldsymbol{\mu})(AX-A\boldsymbol{\mu})^T] \\ & = & A\mathbb{E}[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^T]A^T \\ & = & A\Sigma A^T\end{array}$$

Definition 0.19 - *Conditional Distribution*
Let X & Y be *Random Variables* with *Joint Probability Density* $f_{X,Y}(\cdot,\cdot)$.
Suppose we know that Y takes the value y_0 & we wish to establish the probability of X taking the value x.

$$f(X=x|Y=y_0)=\frac{f_{X,Y}(x,y_0)}{f_Y(y_0)}$$

assuming $f(y_0)>0$.

Proof 0.5 - *Conditional Distribution*
We expect $f(X=x|Y=y_0)=kf_{X,Y}(x,y_0)$ for some constant k.
We know that for $kf_{X,Y}(x,y_0)$ to be a valid distribution it must integrate to one.

$$\begin{array}{rrcl} & k\int_{-\infty}^{\infty}f_{X,Y}(x,y_0)dx & = & 1 \\ \implies & kf_Y(y_0) & = & 1 \\ \implies & k & = & \frac{1}{f_Y(y_0)} \\ \implies & f(X=x|Y=y_0) & = & \frac{f_{X,Y}(x,y_0)}{f_Y(y_0)}\end{array}$$

Proposition 0.1 - *Conditional Distributions with Three Random Variables*

$$\begin{array}{rcl} f(x,z|y) & = & f(x|z,y)f(z|y) \\ f(x,y,z) & = & f(x|y,z)f(z|y)f(y) \\ & = & f(x|y,z)f(y,z)\end{array}$$

Definition 0.20 - *Independent Random Variables*
Let X & Y be random variables.
X & Y are said to be *Statistically Independent* if the *Conditional Distribution* $f(x|y)$ is independent of y.
Thus

$$f(x)=\int_{-\infty}^{\infty}f(x,y)dy$$

$$\begin{array}{rcl} & = & \int_{-\infty}^{\infty}f(x|y)f(y)dy \\ & = & f(x|y)\int_{-\infty}^{\infty}f(y)dy \\ & = & f(x|y) \\ \implies f(x,y) & = & f(x|y)f_Y(y)=f_X(x)f_Y(y)\end{array}$$

Theorem 0.1 - *Bayes' Theorem*
Let X & Y be *Random Variables.*
Bayes' Theorem states that

$$f(X|Y)=\frac{f(Y|X)x(X)}{f(Y)}$$

Definition 0.21 - *First Order Markov Property*
Let $\mathbf{X} := \{X_1, \dots, X_n\}$ be a set of *Random Variables.*
The set $\mathbf{X}$ is said to have the *First Order Markov Property* if

$$f(X_i|\mathbf{X}_{\neg i}) = f(X_i|X_{i-1}) \text{ where } \mathbf{X}_{\neg i} := \mathbf{X}/\{X_i\}$$

Thus we can infer the *marginal distribution*

$$f(\mathbf{X}) = f(X_1)\prod_{i=2}^{N} f(X_i|X_{i-1})$$

0.2.1 Probability Distributions

Definition 0.22 - *β-Distribution*
Let $X \sim \text{Beta}(\alpha, \beta)$.
A *continuous* random variable with shape parameters $\alpha, \beta > 0$. Then

$$\begin{array}{rcl} f_X(x) & \propto & x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}\{x \in [0,1]\} \\ \mathbb{E}(X) & = & \dfrac{\alpha}{\alpha+\beta} \\ \text{Var}(X) & = & \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ \mathcal{M}_X(t) & = & 1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\dfrac{\alpha+r}{\alpha+\beta+r}\right)\dfrac{t^k}{k!} \end{array}$$

Definition 0.23 - *Bernoulli Distribution*
Let $X \sim \text{Bernoulli}(p)$.
A *discrete* random variable which takes 1 with probability p & 0 with probability $(1-p)$. Then

$$\begin{array}{rcl} p_X(k) & = & \begin{cases} 1-p & \text{if } k=0 \\ p & \text{if } k=1 \\ 0 & \text{otherwise} \end{cases} \\ P_X(k) & = & \begin{cases} 0 & \text{if } k<0 \\ 1-p & \text{if } k \in [0,1) \\ 1 & \text{otherwise} \end{cases} \\ \mathbb{E}(X) & = & p \\ \text{Var}(X) & = & p(1-p) \\ \mathcal{M}_X(t) & = & (1-p)+pe^t \end{array}$$

N.B. Often we define $q := 1-p$ for simplicity.

Definition 0.24 - *Binomial Distribution*
Let $X \sim \text{Binomial}(n, p)$.
A *discrete* random variable modelled by a *Binomial Distribution* on n independent events and rate of success p.

$$\begin{array}{rcl} p_X(k) & = & \binom{n}{k}p^k(1-p)^{n-k} \\ P_X(k) & = & \sum_{i=1}^{k}\binom{n}{i}p^i(1-p)^{n-i} \\ \mathbb{E}(X) & = & np \\ \text{Var}(X) & = & np(1-p) \\ \mathcal{M}_X(t) & = & [(1-p)+pe^t]^n \end{array}$$

N.B. If $Y := \sum_{i=1}^n X_i$ where $\mathbf{X} \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ then $Y \sim \text{Binomial}(n, p)$.

Definition 0.25 - *Categorical Distribution*
Let $X \sim \text{Categorical}(\mathbf{p})$.
A *discrete* random varaible where probability vector $\mathbf{p}$ for a set of events $\{1, \dots, m\}$.

$$f_X(i) = p_i$$

Definition 0.26 - χ^2 *Distribution*
Let $X \sim \chi^2_r$.
A *continuous* random variable modelled by the χ^2 *Distribution* with r degrees of freedom. Then

$$\begin{array}{rcl} f_X(x) & = & \dfrac{1}{2^{r/2}\Gamma(r/2)}x^{\frac{r}{2}-1}e^{-\frac{x}{2}} \\ F_X(x) & = & \dfrac{1}{\Gamma(k/2)}\gamma\left(\frac{r}{2},\frac{x}{2}\right) \\ \mathbb{E}(X) & = & r \\ \text{Var}(X) & = & 2r \\ \mathcal{M}_X(t) & = & \mathbb{1}\{t<\frac{1}{2}\}(1-2t)^{-\frac{r}{2}} \end{array}$$

N.B. If $Y := \sum_{i=1}^k Z_i^2$ with $\mathbf{Z} \overset{\text{iid}}{\sim} \text{Normal}(0,1)$ then $Y \sim \chi^2_k$.

Definition 0.27 - *Exponential Distribution*
Let $X \sim \text{Exponential}(\lambda)$.
A *continuous* random variable modelled by a *Exponential Distribution* with rate-parameter λ. Then

$$\begin{array}{rcl} f_X(x) & = & \mathbb{1}\{t \geq 0\}.\lambda e^{-\lambda x} \\ F_X(x) & = & \mathbb{1}\{t \geq 0\}.\left(1-e^{-\lambda x}\right) \\ \mathbb{E}(X) & = & \dfrac{1}{\lambda} \\ \text{Var}(X) & = & \dfrac{1}{\lambda^2} \\ \mathcal{M}_X(t) & = & \mathbb{1}\{t<\lambda\}\dfrac{\lambda}{\lambda-t} \end{array}$$

N.B. Exponential Distribution is used to model the wait time between decays of a radioactive source.

Definition 0.28 - *Gamma Distribution*
Let $X \sim \Gamma(\alpha,\beta)$.
A *continuous* random variable modelled by a *Gamma Distribution* with shape parameter $\alpha > 0$ & rate parameter β. Then

$$\begin{array}{rcl} f_X(x) & = & \dfrac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}e^{-\beta x} \\ F_X(x) & = & \dfrac{\Gamma(\alpha)}{\gamma}(\alpha,\beta x) \\ \mathbb{E}(X) & = & \dfrac{\alpha}{\beta} \\ \text{Var}(X) & = & \dfrac{\alpha}{\beta^2} \\ \mathcal{M}_X(t) & = & \mathbb{1}\{t<\beta\}\left(1-\frac{t}{\beta}\right)^{-\alpha} \end{array}$$

N.B. There is an equivalent definition of a *Gamma Distribution* in terms of a shape & <u>scale</u> parameter. The scale parameter is 1 over the rate parameter in this definition.

Definition 0.29 - *Multinomial Distribution*
Let $\mathbf{X} \sim \text{Multinomial}(n,\mathbf{p})$.

A *discrete* random varible which models n events with probability vector $\mathbf{p}$ for events $\{1,\dots,m\}$.

$$\begin{array}{rcl}
f_{\mathbf{X}}(\mathbf{x}) &=& \mathbb{1}\left\{\sum_{i=1}^m x_i \equiv m\right\}\frac{n!}{x_1!\cdot\dots\cdot x_n!}\prod_{i=1}^n p_i^{x_i}\\
\mathbb{E}(X_i) &=& np_i\\
\text{Var}(X_i) &=& np_i(1-p_i)\\
\text{Cov}(X_i,x_j) &=& -np_ip_j \text{ for } i\neq j\\
\mathcal{M}_{X_i}(\theta_i) &=& \left(\sum_{i=1}^m p_ie^{\theta_i}\right)^n
\end{array}$$

N.B. In a realisation $\mathbf{x}$ of $\mathbf{X}$, x_i is the number of times event i has occured.

Definition 0.30 - *Normal Distribution*
Let $X\sim\text{Normal}(\mu,\sigma^2)$.
A *continuous* random variable with mean μ & variance σ^2.

$$\begin{array}{rcl}
f_X(x) &=& \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\\
F_X(x) &=& \frac{1}{\sqrt{2\pi\sigma^2}}\int\limits_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}}dy\\
\mathbb{E}(X) &=& \mu\\
\text{Var}(X) &=& \sigma^2\\
\mathcal{M}_X(\theta) &=& e^{\mu\theta+\sigma^2\theta^2(1/2)}
\end{array}$$

Definition 0.31 - *Pareto Distribution*
Let $X\sim\text{Pareto}(x_0,\theta)$.
A *continuous* random variable modelled by a *Pareto Distribution* with minimum value x_0 & shape parameter $\alpha>0$. Then

$$\begin{array}{rcl}
f_X(x) &=& \frac{\alpha x_0^\alpha}{x^{\alpha+1}}\\
F_X(x) &=& 1-\left(\frac{x_0}{x}\right)^\alpha\\
\mathbb{E}(X) &=& \begin{cases}\infty & \alpha\leq 1\\ \frac{\alpha x_0}{\alpha-1} & \alpha>1\end{cases}\\
\text{Var}(X) &=& \begin{cases}\infty & \alpha\leq 2\\ \frac{x_0^2\alpha}{(\alpha-1)^2(\alpha-2)} & \alpha>2\end{cases}\\
\mathcal{M}_X(t) &=& \mathbb{1}\{t<0\}\alpha(-x_0t)^\alpha\Gamma(-\alpha,-x_0t)
\end{array}$$

Definition 0.32 - *Poisson Distribution*
Let $X\sim\text{Poisson}(\lambda)$.
A *discrete* random variable modelled by a *Poisson Distribution* with rate parameter λ. Then

$$\begin{array}{rcl}
p_X(k) &=& \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{for } k\in\mathbb{N}_0\\
P_X(k) &=& e^{-\lambda}\sum_{i=1}^k\frac{\lambda^i}{i!}\\
\mathbb{E}(X) &=& \lambda\\
\text{Var}(X) &=& \lambda\\
\mathcal{M}_X(t) &=& e^{\lambda(e^t-1)}
\end{array}$$

N.B. Poisson Distribution is used to model the number of radioactive decays in a time period.

Definition 0.33 - *t-Distribution*
Let $X \sim t_r$.
A *continuous* random variable with r degrees of freedom. Then

$$\begin{array}{rcl} f_X(k) &=& \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \\ \mathbb{E}(X) &=& \begin{cases} 0 & \text{if } \nu > 1 \\ \text{undefined} & \text{otherwise} \end{cases} \\ \text{Var}(X) &=& \begin{cases} \frac{\nu}{\nu-2} & \text{if } \nu > 2 \\ \infty & 1 < \nu \leq 2 \\ \text{undefined} & \text{otherwise} \end{cases} \\ \mathcal{M}_X(t) &=& \text{undefined} \end{array}$$

N.B. Let $Y \sim \text{Normal}(0,1)$ & $Z \sim \chi^2_r$ be independent random variables then $X := \dfrac{Y}{\sqrt{Z/r}} \sim t_r$.

Definition 0.34 - *Uniform Distribution - Uniform*
Let $X \sim \text{Uniform}(a,b)$.
A *continuous* random variable with lower bound a & upper bound b. Then

$$\begin{array}{rcl} f_X(x) &=& \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases} \\ F_X(x) &=& \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & \text{otherwise} \end{cases} \\ \mathbb{E}(X) &=& \frac{1}{2}(a+b) \\ \text{Var}(X) &=& \frac{1}{12}(b-a)^2 \\ \mathcal{M}_X(t) &=& \begin{cases} \dfrac{e^{tb}-e^{ta}}{t(b-a)} & t \neq 0 \\ 1 & t = 0 \end{cases} \end{array}$$