

Theory of Inference - Notes

Dom Hutchinson

February 3, 2020

Contents

1	Motivation	2
1.1	Examples	2
2	Basic Approaches to Inference	3
2.1	Inference by Resampling	5
3	Inference for Linear Models	6
3.1	Linear Model Estimation & Checking	7
0	Reference	8
0.1	Definitions	8
0.2	Probability	8
0.2.1	Probability Distributions	12

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 generated. They depend upon *unknown* constant parameters, θ , 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 values are consistent with the data?
- ii) *Hypothesis Testing* - Are some pre-specified values (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 models could most plausibly have generated the data?
- v) *Statistical Design* - How could we better arrange the 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, Connecticut.

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 drift over the course of the week and to be subject to random autocorrelated departures 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 $e_{i+1} := \rho e_i + \varepsilon_i$ with $|\rho| < 1$ & $\varepsilon \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

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 receiving a different treatment. We wish to test whether one of these treatments is more effective 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 respectively. 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 randomness 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 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, 1)$ 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

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

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 identify the distribution of $\mathbb{P}(\mu|\mathbf{x})$.

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

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 $n \rightarrow \infty$ we notice that

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

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 perform 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:  $\mathbf{x}$  {data set}
1  $\mu s = \{\}$  {resample means}
2  $\mu s$  append  $mean(\mathbf{x})$ 
3 for  $i = 0 \dots m$  do
4    $x_i \leftarrow sample(\mathbf{x}, n, replace=TRUE)$ 
5    $\mu s$  append  $mean(x_i)$ 
6 return  $quantile(\mu s, (0.025, 0.975))$ 

```

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 Variable*.

The *Predictor Variables* fulfil one of the following roles

- *Metric* - Quantifiable 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

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

where $\varepsilon_i \stackrel{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$.

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).

$$\text{distance}_i = \beta_0 + \beta_1 \text{speed}_i + \beta_2 \text{speed}_i^2 + \beta_3 \text{speed}_i^3 + \varepsilon_i$$

3.1 Linear Model Estimation & Checking

Proposition 3.1 - Frequentist Approach

In the *Frequentist Approach* to *Linear Models* we assume that β 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 β which minimises the *Residual Sum of Squares*.

Definition 3.3 - Residual Sum of Squares

Let (X, \mathbf{y}) be a set of training data & β a *Parameter Vector*.

The *Residual Sum of Squares* is the square difference between our estimate for the *Response Variable* and its true value.

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

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

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

Thus minimising the *Residual Sum of Squares* is reduced to choosing β st $R\beta = \mathbf{f}$.

Hence, provided that X and R have full rank

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

N.B. After choosing β we have that the *Residual Sum of Squares* is just $\|\mathbf{r}\|^2$.

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*

0.2 Probability

Definition 0.5 - *Random Variable*

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

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

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 0.6 - *Cummulative Distribution*

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

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

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*

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

Definition 0.7 - *Quantile Function*

The *Quantile Function* of a *Random Variable* is the inverse function of the *Cummulative Distribution*.

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

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 0.8 - (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 closer this line is to $y = x$ the more likely it is the data was generated by this *Cumulative Distribution*.

N.B. AKA *Quantile-Quantile Plot*

Definition 0.9 - Probability Mass Function

A *Probability Mass Function* returns the probability of a discrete random variable taking a particular value.

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

Definition 0.10 - Probability Density Function

Since the probability of a *Continuous Random Variable* taking a specific value is zero we cannot use the *Probability Mass Function*.

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

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

Definition 0.11 - Joint Probability Density Function

Let X & Y be *Random Variables*.

The *Joint Probability Density Function* of X and Y is the function $f_{X,Y}(x, y)$ st

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

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

Definition 0.12 - 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 0.13 - Expected Value, \mathbb{E}

The *Expected Value* of a *Random Variable*, X , is its mean value.

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

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

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

Definition 0.14 - 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

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

Definition 0.15 - 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 0.16 - 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

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

Definition 0.17 - 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.

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

Proposition 0.1 - Conditional Distributions with Three Random Variables

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

Definition 0.18 - 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

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

Theorem 0.1 - Bayes' Theorem

Let X & Y be *Random Variables*.

Bayes' Theorem states that

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

Definition 0.19 - 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

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

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 0.20 - β -Distribution

Let $X \sim \text{Beta}(\alpha, \beta)$.

A *continuous* random variable with shape parameters $\alpha, \beta > 0$. Then

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

Definition 0.21 - Bernoulli Distribution

Let $X \sim \text{Bernoulli}(p)$.

A *discrete* random variable which takes 1 with probability p & 0 with probability $(1 - p)$. Then

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

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

Definition 0.22 - 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 .

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

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

Definition 0.23 - Categorical Distribution

Let $X \sim \text{Categorical}(\mathbf{p})$.

A *discrete* random variable where probability vector \mathbf{p} for a set of events $\{1, \dots, m\}$.

$$f_X(i) = p_i$$

Definition 0.24 - χ^2 Distribution

Let $X \sim \chi_r^2$.

A *continuous* random variable modelled by the χ^2 Distribution with r degrees of freedom. Then

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

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

Definition 0.25 - Exponential Distribution

Let $X \sim \text{Exponential}(\lambda)$.

A *continuous* random variable modelled by a *Exponential Distribution* with rate-parameter λ . Then

$$\begin{aligned} f_X(x) &= \mathbb{1}\{t \geq 0\} \cdot \lambda e^{-\lambda x} \\ F_X(x) &= \mathbb{1}\{t \geq 0\} \cdot (1 - e^{-\lambda x}) \\ \mathbb{E}(X) &= \frac{1}{\lambda} \\ \text{Var}(X) &= \frac{1}{\lambda^2} \\ \mathcal{M}_X(t) &= \mathbb{1}\{t < \lambda\} \frac{\lambda}{\lambda - t} \end{aligned}$$

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

Definition 0.26 - 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

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

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

Definition 0.27 - Multinomial Distribution

Let $\mathbf{X} \sim \text{Multinomial}(n, \mathbf{p})$.

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

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

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

Definition 0.28 - Normal Distribution

Let $X \sim \text{Normal}(\mu, \sigma^2)$.

A *continuous* random variable with mean μ & variance σ^2 .

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

Definition 0.29 - 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

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

Definition 0.30 - Poisson Distribution

Let $X \sim \text{Poisson}(\lambda)$.

A *discrete* random variable modelled by a *Poisson Distribution* with rate parameter λ . Then

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

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

Definition 0.31 - t -Distribution

Let $X \sim t_r$.

A *continuous* random variable with r degrees of freedom. Then

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

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

Definition 0.32 - Uniform Distribution - Uniform

Let $X \sim \text{Uniform}(a, b)$.

A *continuous* random variable with lower bound a & upper bound b . Then

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