Statistics 1 - Notes

Dom Hutchinson

October 2, 2019

Contents

1 Introduction

Remark 1.1 - *Statistics Framework*
All problems in statistics exist within the following framework

- A finite or infinite population of objects;
- A real-value variable associated to each object;
- A quantity of interest;
- A sample of the population; And,
- A dataset of observed values from the variables of sample objects.

1.1 Exploratory Data Analysis

Definition 1.1 - *Exploratory Data Analysis*
Exploratory Data Analysis is a collection technique for an initial data set.
It ensures the following properties hold

- Observations are independent;
- Observations all come from common distributions; And,
- The distribution has a particular type (normal, binomial, etc.).

Remark 1.2 - *Data Set Notation*
We write $\{x_1, \dots, x_n\}$ to denote a data set.
These values are typically in time order st x_i happened before x_{i+1} $\forall$ $i < n$.

Definition 1.2 - *Order Statistic*
A *Order Statistics* is a data set which is arranged in increasing order of value.
We denote the data set as $\{x_{(1)}, \dots, x_{(n)}\}$ st $x_{(i)} \leq x_{(x+1)}$ $\forall$ $i < n$. N.B. We say $x_{(i)}$ has *rank* i.

Remark 1.3 - *Computer Science*
The *Order Statistics* problem was addressed in the *Data Structures & Algorithms, COMS21103.*
See *Subsection 3.3 Order Statistics* in `NotesDSA.pdf`.

Definition 1.3 - *Median*
The *Median* of a data set is the middle value.
Let $n \in \mathbb{N}$ be the number of data points.

- If $\exists$ $m \in \mathbb{N}$ st $n = 2m + 1$ (i.e. n is odd) then the median is $x_{(m+1)}$;
- Else, if $\exists$ $m \in \mathbb{N}$ st $n = 2m$ (i.e. n is even) then the median is $\frac{1}{2}(x_{(m)} + x_{(m+1)})$.

Else, if $\exists$ $m \in \mathbb{N}$ st $n = 2m$ (i.e. n is even) then the median is $\frac{1}{2}(x_{(m)} + x_{(m+1)})$.
Definition 1.4 - *Sample Mean*
The *Sample Mean* of a *Sample* X is defined as

$$\bar{x} := \frac{1}{n} \sum_{1=i}^{n} x_i$$

Remark 1.4 - *Sensitivity of Mean & Median*
The *Mean* is sensitive to extreme values, while the *Median* is much less sensitive as it more

associated with rank than value.

Definition 1.5 - *Trimmed Sample Mean*
The *Trimmed Sample Mean* calculates the mean from a reduced data set, aiming to exclude extreme values.
Trimmed Sample Mean is denoted by $\Delta\%$ for $\Delta \in [0, 100]$.
To calculate the *Trimmed Sample Mean* do the following

i) Set $k = \lfloor \frac{n\Delta}{100} \rfloor$;

ii) Calculate the sample mean of $\{x_{k+1}, \dots, x_{n-k-1}\}$.

1.2 Numerical Measures of Range or Spread of Data

Definition 1.6 - *Sample Variance*
The *Sample Variance* is a measure of spread of data around the mean.
A lower value indicates a lower spread.
The *Sample Variance* of a sample X is defined as

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

N.B. *Standard Deviation* is defined as $s = \sqrt{s^2}$.

Proposition 1.1 - *Easier Sample Variance Equation*
The following is another equation for *Sample Variance* which is generally easier to calculate with by hand

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i^2 - n\bar{x}^2)$$

Remark 1.5 - *Normalisation of Sample Variance*
It is noteworthy that variance is normalised to $n-1$ not n.
This is since given $n-1$ data points from a data set, and its mean we can calculate the final value.
Thus there are effectively only $n-1$ independent values.

Definition 1.7 - *Hinges*
Hinges are values that describe the spread of data values, trying to ignore extreme values.

- The *Lower Hinge* H_1 is the median of the set of data values with rank $\leq$ rank of the sample median;

- The *Upper Hinge* H_3 is the median of the set of data values with rank $\geq$ rank of the sample median;

N.B. When there are an even number of data points we exclude the median from these sets.

Definition 1.8 - *Quartiles*
Quartiles are an alternative to *Hinges*, they describe similar properties.

i) If $\frac{1}{4}(n+1) \in \mathbb{Z}$ then $Q_1 = x_{(\frac{1}{4}(n+1))}$;

ii) If $\frac{3}{4}(n+1) \in \mathbb{Z}$ then $Q_3 = x_{(\frac{3}{4}(n+1))}$; Else,

iii) If these ranks aren't integers then we interpolate values.
Let $k = \lfloor \frac{1}{4}(n+1) \rfloor$ then $Q_1 = x_{(k)} + \left(\frac{n+1}{4} - k\right)[x_{(k+1)} - x_{(k)}]$.

Remark 1.6 - *Quartiles v Hinges*
Generally $H_1 \neq Q_1$ and $H_3 \neq Q_3$, but for large samples we get $H_1 \simeq Q_1$ and $H_3 \simeq Q_3$.
Definition 1.9 - *Five Number Summary*
The *Five Number Summary* for a data set holds the values of the median, lower hinge, upper hinge, minimum value & maximum value.

Definition 1.10 - *Interquartile Range*
The *Interquartile Range* is the difference in value between the upper & lower quartile.

$$IQR = Q_3 - Q_1$$

Definition 1.11 - *Outliers*
Outlier data points are those with values more than $\frac{3}{2}.IQE$ from a hinge.

Definition 1.12 - *Skewed*
A data a set can be skewed in two directions, left & right

- Left Skewed if $H_3 - median > H_1 - median$;
- Right Skewed if $H_3 - median < H_1 - median$;

1.3 Initial Graphical Plots

Remark 1.7 - *Features of Graphical Plots*
When given a graphical plot you should consider the following features

i) Overall pattern of variation within the data (*e.g.* Symmetry, Skewnewss etc.);

ii) Any unusual features within a pattern, or striking deviations from a pattern (*e.g.* Outliers);

iii) Whether any unusual features are random occurrences or are systematic features; And,

iv) Any evidence of clustering.

Definition 1.13 - *Stem-and-Leaf-Plot*
Here is a formal, algorithmic description of a *Stem-and-Leaf-Plot*

i) Truncate or round the data values so that all the variation is the in the last two, or three, significant figures;

ii) Separate each data value into a stem (consisting of all digits except the last) and a leaf (last digit);

iii) Write the stems in a vertical column, smallest to biggest, and draw a vertical line to separate from the right column;

iv) Write each leaf in the row to the right of its corresponding stem, in increasing order;

v) Record any strikingly low or high values separately from the main stem, displaying the individual values in a group above the main stem or below.

Example 1.1 - *Stem-and-Leaf-Plot*
The following is a data set & a representation of it as a *Stem-and-Leaf-Plot.*
The left value is multiples of 100 & values are rounded to nearest 10.

9	30	33	36	38	40	40	44	46	76	82
83	92	99	121	129	139	145	150	157	194	203
209	220	246	263	280	294	304	319	328	335	365
599	638	667	695	710	721	735	736	759	780	832
840	887	937	1336	1354	1617	1901				

0	133444445888902345569
2	01256890234788
4	034566678
6	0470124468
8	3494
10	
12	45
14	
16	2
18	0

Definition 1.14 - *Histogram*
Here is a formal, algorithmic description of a *Histogram*

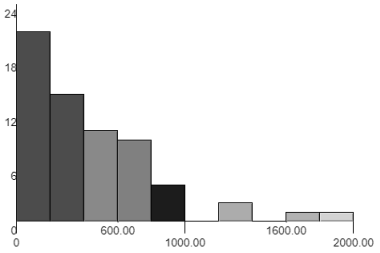
i) Divide the range of data values into K intervals (bins) of equal width.

ii) Counter the frequency of observations falling into each interval.

iii) Display a plot of joined columns above each interval, with the columns height proportional to the count for that interval.

Example 1.2 - *Histogram*
The following is a data set & a representation of it as a *Histogram Plot.*

9	30	33	36	38	40	40	44	46	76	82
83	92	99	121	129	139	145	150	157	194	203
209	220	246	263	280	294	304	319	328	335	365
599	638	667	695	710	721	735	736	759	780	832
840	887	937	1336	1354	1617	1901				



2 Parametric Families & Method of Moments

2.1 Parametric Models

Definition 2.1 - *Probability Density Function*
A *Probability Density Function* is a function of a *Continuous Random Variable.*
It's value at a given point gives the probability having random variable that value.

Definition 2.2 - *Probability Mass Function*
A *Probability Mass Function* is a function of a *Discrete Random Variable*.
It's value at a given point gives the probability of the random variable having that value.

Definition 2.3 - *Parametric Family*
A *Parametric Family* is a collection of distributions of the same type (*e.g.* Normal, Binomial, etc.) and only differ in the value of one, or more, parameter.
N.B. The form of the distribution is a function of these parameters.

Remark 2.1 - *Parametric Family Notation*
Let X represent the parametric family and θ the variable parameter. Write

- $f_X(x;\theta)$ for the probability density function in the parametric family.
  Or, $p_x(X;\theta)$ for the probability mass function if the family is discrete.
- $\mathbb{E}(X;\theta)$ for the expectation of X.
- $\mathbb{P}(X \in A;\theta)$ for the probability that $X \in A$.
  N.B. This is often written as $\mathbb{P}(A;\theta)$ when the random variable is obvious.
- $F_X(x;\theta) = \mathbb{P}(X \in \{y \in \mathbb{R} : y \leq x\};\theta)$ for the cumulative distribution function.

Remark 2.2 - *Estimation of θ*
When trying to estimate the value of θ for a parametric family we could use a sample set, $\hat{\theta}(x_1, \ldots, x_n)$.
However, there is a fundamental flaw in this since the value of $\hat{\theta}(x_1, \ldots, x_n)$ depends on the sample which is taken.
An improvement to this is to use a *Population of Estimators*.

Remark 2.3 - *Population of Estimators for Estimation of θ*
By choosing a population of estimators $(X_1, \ldots, X_n)$ to be distributed according to the parametric model, for the true value of θ, we can use $\hat{\theta}(X_1, \ldots, X_n)$ to estimate θ.
The characterisation of $\hat{\theta}(X_1, \ldots, X_n)$ will show us how close the estimation of an observed sample $\hat{\theta}(x_1, \ldots, x_n)$ is to θ^*.
N.B. θ^* denotes the true value of θ.

Remark 2.4 - *Quantity of Interest & Estimation of Theta*
It may be the case that our *Quantity of Interest* in a problem is not θ itself, but it must be a function of θ, say $\tau(\theta)$.
Once we have calculated an estimated value of $\hat{\theta}$ we can estimate the original *Quantity of Interest* using $\hat{\tau} = \tau(\hat{\theta})$.

Definition 2.4 - *Joint Probability Density of Simple Random Sample*
For a *Simple Random Sample* from a distribution $f_X(x_i;\theta)$ in a parametric family we have

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n;\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)$$

Proof 2.1 - *Joint Probability Density of Simple Random Sample*
Since $X_1, \ldots, X_n$ are independent, their joint probability density function factorises as the product of the margin density functions. Giving

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n;\theta) = f_{X_1}(x_1;\theta) f_{X_2}(x_2;\theta) \ldots f_{X_n}(x_n;\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)$$

2.2 Method of Moments Estimation

Remark 2.5 - *Motivation*
The *Method of Moments Estimation* relies on the idea that if data comes from a simple random sample then the sample values are representative of population values.
This suggests that for $k \in \mathbb{N}$ the k^{th} *Population Moment*$\simeq k^{th}$ *Sample Moment*.

Definition 2.5 - *Population Moment*
Assume the population random variable X comes from a parametric family with parameter θ.
For $k \in \mathbb{N}$ we define the k^{th} *Population Moment* as

$$\mathbb{E}(X^k; \theta) := \int_{-\infty}^{\infty} x^k f(x; \theta) dx$$

Definition 2.6 - *Sample Moment*
For a sample with data values $\{x_1, \dots, x_n\}$ for $k \in \mathbb{N}$ we define the k^{th} *Sample Moment* as

$$m_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

N.B. This is the mean value of x^k in the sample.

Proposition 2.1 - *Using Method of Moments Estimation*
Given a sample $\{x_1, \dots, x_n\}$ from a parametric family, if the family has one parameter θ we define the method of moments estimator $\hat{\theta}_{mom}$ to be the solution of

$$\mathbb{E}(X; \hat{\theta}_{mom}) = m_1$$

.
If the family has two parameters $\alpha,\ \beta$ define the moments of estimators $\hat{\alpha}_{mom}$ & $\hat{\beta}_{mom}$ to be the solutions to the simultaneous equations

$$\begin{aligned} \mathbb{E}(X; \hat{\alpha}_{mom}, \hat{\beta}_{mom}) &= m_1 \\ \mathbb{E}(X^2; \hat{\alpha}_{mom}, \hat{\beta}_{mom}) &= m_2 \end{aligned}$$

N.B. For k unknown parameters compare the first k population & sample moments.

Example 2.1 - *Method of Moments, Single Parameter*
Assume $\{x_1, \dots, x_n\}$ come from a simple random sample of the distribution *Exponential*(θ), with θ unknown.
Here there is one unknown parameter, so we use one equation.
For $\theta > 0$ we have $\mathbb{E}(X; \theta) = \frac{1}{\theta}$.
By the method of moments we have $\frac{1}{\theta} = m_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$.
Meaning $\hat{\theta}_{mom} = \frac{1}{m_1}$.

Example 2.2 - *Method of Moments, Two Parameters*
Assume $\{x_1, \dots, x_n\}$ come from a simple random sample of the distribution $\mathcal{N}(\mu, \sigma^2)$, with μ, σ^2 unknown.
We have two unknown parameters, so we need two equations.

$$\begin{aligned} \mathbb{E}(X; \mu, \sigma^2) &= m_1 \\ \mathbb{E}(X^2; \mu, \sigma^2) &= var(X; \mu, \sigma^2) + \mathbb{E}(X; \mu, \sigma^2)^2 \\ &= m_2 \\ &= \sigma^2 + m_1^2 \end{aligned}$$

Thus $\mu = m_1$ & $\sigma^2 = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \approx$ Sample variance.

2.3 Assessing Fit

Remark 2.6 - *Motivation*
When given a random sample $\{x_1, \dots, x_n\}$ from a population whose cumulative frequency function $F_x(x;\theta)$ and probability density function $f_X(x;\theta)$ have parametric forms, and are given an estimate $\hat{theta}$.
We should asses how well our model fits the data by comparing the sample with the values we might expected from $F_X(x;\hat{\theta})$.
If there are striking or systematic differences from what we expect, our assumed model may not be appropriate.

Proposition 2.2 - *Estimation of Cumulative Frequency*
Assuming that $\{x_1, \dots, x_n\}$ come from a simple random sample and that the model $F_X(x;\theta^*)$ is correct. Then

$$\forall\ i, y\ \mathbb{P}(X_i \leq y;\theta^*) = F_X(y;\theta^*)$$

This means that $|\{X_i \leq y\}|\ Bin(n, F_X(y,\theta^*))$.
Then $\mathbb{E}(|\{X_i \leq y\}|) = nF_X(y;\theta^*)$.
Formally we can define a distribution to estimate $F_X(y;\theta^*)$

$$\hat{F}(y) := \frac{1}{n}|\{X_i \leq y\}|$$

Definition 2.7 - *Empirical Distribution Function*
The *Empirical Distribution Function* is defined to be the map

$$y \mapsto \hat{F}(y)$$

Proposition 2.3 - *Assessing Fit*
We compare the *Empirical Distribution Function* ($y \mapsto \hat{F}(y)$) to the function $y \mapsto F_X(y;\theta^*)$ to asses fit.
This is comparing expected values to real values.
Since θ^* is unknown then we use our estimate $\hat{\theta}$.

Definition 2.8 - *Probability Plot*
A *Probability Plot* is a plot of the curve $y \mapsto (F_X(y;\hat{\theta}), \hat{F}(y))$.
If these two functions are the same then the plot will lye on the main diagonal.
N.B. $\hat{F}(x_{(k)}) = \frac{k}{n}$.

Definition 2.9 - *(Q-Q) Plots*
(Q-Q) plots share the same idea as *Probability Plot*, but plot the inverse of each cumulative frequency functions (*i.e.* the map $y \mapsto (F_X^{-1}(y;\hat{\theta}), \hat{F}^{-1}(y))$.
N.B. Since $y \mapsto F_X(y;\theta)$ is continuous & increasing, its inverse always exists.

Proposition 2.4 - *Finding $\hat{F}^{-1}(y)$*
Since $\hat{F}(x_{(k)}) = \frac{k}{n}$ then $\hat{F}^{-1}(\frac{k}{n}) = x_{(k)}$.
We can restrict the (Q-Q) plot to $k \mapsto (y_{k/n} = F_X^{-1}(\frac{k}{n};\hat{\theta}), x_{(k)})$ and then visually assess whether $x_{(k)} \simeq y_{k/n}$ for $k \in [1, n]$.

Proposition 2.5 - *Producing (Q-Q) Plots*
This technique requires an analytic or numerical method for computing values of $F_X^{-1}(x;\theta)$.

i) Compute an estimate $\hat{\theta}$ for θ. (e.g. The method of moments estimate).

ii) Order the observations to obtain the order statistics $x_{(1)}, \dots, x_{(n)}$.

iii) For $k = 1, \dots, n$ compute $F_X^{-1}\left(\frac{k}{n+1}; \hat{\theta}\right)$.
N.B. These are the fitted quantile values.

iv) For $k = 1, \dots, n$ plot the pairs $\left(F_X^{-1}\left(\frac{k}{n+1}; \hat{\theta}\right), x_{(k)}\right)$.

v) Add the line $y = x$ to the plot.

Proposition 2.6 - *Producing Probability Plots*
This process is very similar to that for *(Q-Q) Plots*.
This technique requires an analytic or numerical method for computing values of $F_X(x; \theta)$.

i) Compute an estimate $\hat{\theta}$ for θ. (e.g. The method of moments estimate).

ii) Order the observations to obtain the order statistics $x_{(1)}, \dots, x_{(n)}$.

iii) For $k = 1, \dots, n$ compute $F_X\left(\frac{k}{n+1}; \hat{\theta}\right)$.
N.B. These are the fitted quantile values.

iv) For $k = 1, \dots, n$ plot the pairs $\left(F_X\left(\frac{k}{n+1}; \hat{\theta}\right), x_{(k)}\right)$.

v) Add the line $y = x$ to the plot.

Proposition 2.7 - *Interpreting Quantile Plots*

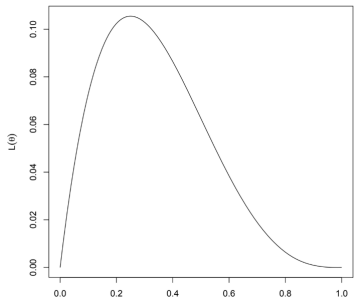
3 Likelihood & Maximum Likelihood Estimation

Definition 3.1 - *Likelihood Function*
Fora a given observation x the *Likelihood Function* is defined as

$$L(\theta; x) := \begin{cases} p(x; \theta) & \text{when } x \text{ is discretevalued} \\ f(x; \theta) & \text{when } x \text{ takes continuous values} \end{cases}$$

Definition 3.2 - *Maximum Likelihood Estimate*
The value of θ that maximises the *likelihood function* $L(\theta; x)$ is called the *Maximum Likelihood Estimate* of θ.

Remark 3.1 - *Motivation*
Consider trying to establish whether a given coin is fair.
Define $\theta := \mathbb{P}(H)$. We can gain information about θ by repeatedly tossing a coin until we get a Head.
Define X to be the number of the toss on which we get our first head.
Assuming each toss is independent then X $Geom(\theta)$ and $p(x; \theta) = (1-\theta)^{x-1}\theta$.
Say we perform the experiment once and get the single observation that $x = 4$.
We get $L(\theta; x) = L(\theta; 4) = p(4; \theta) = (1-\theta)^4\theta$.



The maximum of this graph is the *Maximum Likelihood Estimate* of θ.

Example 3.1 - *Cont. Motivation*
We can find the turning points by setting $\frac{dL}{d\theta} = 0$

$$\begin{aligned} \frac{dL}{d\theta} &= \frac{d}{d\theta}\theta(1-\theta)^3 \\ &= (1-\theta)^3 - 3\theta(1-\theta)^2 \\ &= (1-\theta)^2(1-4\theta) \\ \implies \theta &= 1, \tfrac{1}{4} \end{aligned}$$

To check which is the maximum we calculate $\frac{d^2L}{d\theta^2}$.

$$\begin{aligned} \frac{d^2L}{d\theta^2} &= -2(1-\theta)(1-4\theta) - 4(1-\theta)^2 \\ \theta = \tfrac{1}{4} &\equiv -2.25 < 0 \end{aligned}$$

Making $\frac{1}{4}$ is a maximum $\implies \hat{\theta}_{MLE} = \frac{1}{4}$.

Proposition 3.1 -
If $X_1, \dots, X_n$ is a random sample of size n from a distribution with pmf $p(x;\theta)$ or pdf $f(x;\theta)$ then the X_i are iid and their joint distribution factorises into teh product of marginals. Thus we can define the *Likelihood Function* as

$$L(\theta; x_1, \dots, x_n) := \begin{cases} p(x_1;\theta)\dots p(x_n;\theta) & \text{Discrete} \\ f(x_1;\theta)\dots f(x_n;\theta) & \text{Continuous} \end{cases}$$

Definition 3.3 - *Maximum Likelihood Estimator, Multiple Samples*
For observed values $\{x_1, \dots, x_n\}$ the maximum likelihood estimator $\hat{\theta}_{MLE}(x_1, \dots, x_n)$ is the value of θ which maximises the likelihood function $L(\theta; x_1, \dots, x_n)$.

Example 3.2 - *Multiple Samples*
Suppose we repeat the experiment from the motivation three times and get values $x_1 = 3, x_2 = 5, x_3 = 1$. Then

$$\begin{aligned} L(\theta; x_1, x_2, x_3) &= p_{X_1,X_2,X_3}(x_1, x_2, x_3; \theta) \\ &= (1-\theta)^{x_1-1}\theta(1-\theta)^{x_2-1}\theta(1-\theta)^{x_3-1}\theta \\ &= (1-\theta)^3\theta(1-\theta)^4\theta(1-\theta)^0\theta \\ &= (1-\theta)^7\theta^3 \end{aligned}$$

Maximising $L(\theta)$ we get that $\hat{\theta}_{MLE} = \frac{3}{10}$.

Definition 3.4 - *Log-Likelihood Function*
For observed values $\{x_1, \dots, x_n\}$ and associated likelihood function $L(\theta; x_1, \dots, x_n)$ the *log-likelihood function* is defined as

$$\ell := \ln L(\theta; x_1, \dots, x_n)$$

N.B. Since ln is an increasing function the maximum of $L(\theta)$ & $\ell(\theta)$ is the same.

Theorem 3.1 - *Log-Likelihood Function, Multiple Samples*
For observations from a simple random sample

$$l(\theta; x_1, \dots, x_n) = \begin{cases} \sum_{i=1}^{n} \ln p(x_i; \theta) & \text{Discrete} \\ \sum_{i=1}^{n} \ln f(x_i; \theta) & \text{Continuous} \end{cases}$$

Proof 3.1 - *Log-Likelihood Function, Multiple Samples*
For the continuous case we have

$$\begin{aligned} \ell(\theta) &= \ln L(\theta) \\ &= \ln(f(x_1;\theta) \dots f(x_n;\theta)) \\ &= \ln f(x_1;\theta) + \dots + \ln f(x_n;\theta) \\ &= \textstyle\sum_i \ln f(x_i;\theta) \end{aligned}$$

This is similarly shown for the discrete case.

Example 3.3 - *Log-Likelihood Function*
Let $\{x_1, \dots, x_n\}$ be the sample from $Exp(\theta)$.

$$\begin{aligned} L(\theta) &= (\theta e^{-\theta x_1}) \dots (\theta e^{-\theta x_n}) \\ &= \theta^n e^{-\theta(x_1 + \dots + x_n)} \\ \implies \ell(\theta) &= n \ln \theta - \theta \textstyle\sum_i x_i \end{aligned}$$

Proposition 3.2 - *Calculating $\hat{\theta}_{MLE}$ in Random Sample Case*
Let $f(x;\theta)$ be the pdf a continuous regular distribution.
Here is a technique for calculating $\hat{\theta}_{MLE}$ for a random sample.

i) Calculate $\frac{\partial}{\partial\theta} \ln f(x;\theta)$.

ii) Compute & Simplify $\frac{d}{d\theta} \ln f(x_i;\theta) + \dots + \frac{d}{d\theta} \ln f(x_n;\theta)$.

iii) The *Maximum Likelihood Estimator* is the value of $ii) = 0$.

3.1 Maximum Likelihood Estimate of $\tau(\theta)$

Theorem 3.2 - *Maximum Likelihood Estimate of $\tau(\theta)$*
Suppose our quantity of interest is a function is a continuously differentiable function $\tau(\theta)$, which is either increasing, or decreasing.
We can calculate the *Maximum Likelihood Estimate* of $\tau(\theta)$ by plugging in $\hat{\theta}_{MLE}$

$$\widehat{\tau(\theta)}_{MLE} = \tau(\hat{\theta}_{MLE})$$

Proof 3.2 - *Maximum Likelihood Estimate of $\tau(\theta)$*
This is non-examinable.
Under the new parametrisation we have that $\ell^{new}(\tau(\theta)) = \ell^{old}(\theta)$.
By applying the chain rule to $\ell(\tau(\theta))$ we get

$$\frac{\partial}{\partial\theta}\ell^{old}(\theta) = \frac{\partial}{\partial\theta}\ell^{new}(\tau(\theta)) = \frac{\partial}{\partial\tau(\theta)}\ell^{new}(\tau(\theta)) \times \tau'(\theta)$$

Since $\tau(\theta)$ is increasing or decreasing then $\tau'(\theta) \neq 0$, thus the last expression equals 0 iff $\frac{\partial}{\partial\tau(\theta)}\ell^{new}(\tau(\theta)) \times \tau'(\theta)$.

Example 3.4 -
Let $x_1, \dots, x_n$ be observed values of a simple random variable from $Exp(\theta)$ distribution with θ unknown.
We found previously that $\hat{\theta}_{MLE} = \frac{1}{\bar{x}}$.
Consider the following functional quantities

i) Suppose we are interested in the population variance.
Set $\tau(\theta) = Var(X;\theta) = \frac{1}{\theta^2}$. Then

$$\widehat{\tau(\theta)} = \tau(\hat{\theta}) = \bar{x}^2$$

N.B. This is not the same as the sample variance.

ii) Suppose we are interested in the proportion of the population which have values ≥ 1.
Set $\tau(\theta) = \mathbb{P}(X \geq 1;\theta) = e^{-\theta}$. Then

$$\widehat{\tau(\theta)} = \tau(\hat{\theta}) = e^{-\hat{\theta}} = exp\left(\frac{-1}{\bar{x}}\right)$$

N.B. This is not the same as the sample variance of values ≥ 1.

3.2 Most Likelihood Estimates with Multiple Parameters

Remark 3.2 - *Most Likelihood Estimates with Multiple Parameters with Regular Distributions*
Most Likelihood Estimates can be extended to the case with multiple parameters.
Consider the case with two parameters α & β, we can find $\hat{\alpha}_{MLE}$ & $\hat{\beta}_{MLE}$ by solving the simultaneous solution to the two likelihood equations

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} \ln f(x_i;\alpha,\beta) \quad \& \quad 0 = \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \ln f(x_i;\alpha,\beta)$$

Example 3.5 - *Multiple Parameters, Regular Density*
Consider a simple random sample from $N(\mu,\sigma^2)$, with unknown mean & variance. (*N.B.* The Normal distribution is continuous & regular).
$\hat{\mu}_{MLE}$ & $\hat{\sigma}_{MLE}$ are the simultaneous solutions to the likelihood equations.
Since $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$ we have

$$\begin{array}{rrcl}
 & \ln f(x;\mu,\sigma) & = & -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{(x-\mu)^2}{2\sigma^2} \\
\implies & \frac{\partial}{\partial\mu}\ln f(x;\mu,\sigma^2) & = & \frac{x-\mu}{\sigma^2} \\
\text{Setting} & 0 & = & \sum_{i=1}^{n} \frac{x_i - \hat{\mu}_{mle}}{\hat{\sigma}^2_{mle}} \\
 & & = & \frac{1}{\hat{\sigma}^2_{mle}}\left(-n\hat{\mu}_{mle} + \sum_{i=1}^{n} x_i\right) \\
\implies & \hat{\mu}_{MLE} & = & \frac{1}{n}\sum_{i=1}^{n} x_i \\
 & & = & \bar{x}
\end{array}$$

Also

$$\begin{array}{rrcl}
 & \frac{\partial}{\partial\sigma}\ln f(x;\mu,\sigma^2) & = & -\frac{1}{\sigma} + \frac{x-\hat{\mu}_{mle}}{\sigma^3} \\
\text{Setting} & 0 & = & \sum_{i=1}^{b}\left(-\frac{1}{\hat{\sigma}_{mle}} + \frac{x-\hat{\mu}_{mle}}{\hat{\sigma}^3_{mle}}\right) \\
\implies & \frac{n}{\hat{\sigma}_{mle}} & = & \sum_{i=1}^{n} \frac{(x_i-\hat{\mu}_{mle})^2}{\hat{\sigma}^2_{mle}} \\
\implies & \hat{\sigma}^2_{mle} & = & \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{mle})^2 \\
 & & = & var(X)
\end{array}$$

Remark 3.3 - *Non-Regular Density*
When the density is non-regular we cannot consider the maximum likelihood to be a turning

point. Instead, the likelihood is maximised at one end-point of the interval.
Thus we work with $L(\theta)$ directly.

Example 3.6 - *Non-Regular Density*
Consider a simple random sample $\{x_1, \dots, x_n\}$ from the distribution $U(0, \theta)$.
Here $f(x;\theta) = \begin{cases} \frac{1}{\theta} & if\ 0 \le x \le \theta \\ 0 & otherwise \end{cases}$. Thus

$$L(\theta; x_1, \dots, x_n) = f(x_1;\theta) \dots f(x_n;\theta) = \begin{cases} \frac{1}{\theta^n} & if \theta \ge x_1, \dots, \theta \ge x_n n \\ 0 & otherwise \end{cases}$$

Hence, the likelihood is $\frac{1}{\theta^n}$ if $\theta \ge x_{(n)} = max(x_1, \dots, x_n)$, otherwise it is 0.
This means that the likelihood is maximised for $\hat{\theta}_{MLE} = x_{(n)}$.
N.B. This is easiest to see once you have plotted a graph.

Example 3.7 - *Poisson Data with Unequal Means*
Consider counting photons arriving at a detector in intervals of variable length.
Suppose the arrival is λ per unit time and X_i be the number of arrivals in the i^{th} interval, with known time t_i.
Assume X_i $Poisson(\lambda t_i)$ for $i \in [1, n]$ & that the intervals don't overlap.
The joint probability mass function of $X_1, \dots, X_n$ is

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) &= \mathbb{P}(X_1 = x_1) \dots \mathbb{P}(X_n = x_n) \\
&= \tfrac{1}{x_1!} e^{-\lambda t_1} (\lambda t_1)^{x_1} \times \dots \times \tfrac{1}{x_n!} e^{-\lambda t_n} (\lambda t_n)^{x_n} \\
&= \frac{t_1^{x_1} \dots t_n^{x_n}}{x_1! \dots x_n!} e^{-\lambda(t_1 + \dots + t_n)} \lambda^{x_1 + \dots + x_n} \\
\implies \quad \ell(\lambda) &= -\lambda(t_1 + \dots + t_n) + \ln \lambda (x_1 + \dots + x_n) + non - \lambda\ terms \\
\implies \quad \tfrac{\partial}{\partial \lambda} \ell(\lambda) &= -(t_1 + \dots + t_n) + \tfrac{1}{\lambda}(x_1 + \dots + x_n) \\
\text{Setting} \quad \tfrac{\partial}{\partial \lambda} \ell(\lambda) &= 0 \\
\implies \quad \hat{\lambda}_{mle} &= \tfrac{x_1 + \dots + x_n}{t_1 + \dots + t_n}
\end{aligned}$$

4 Assessing the Performance of Estimators

4.1 Different Methods of Estimation

Remark 4.1 - *Direct Method*
There may be a direct (non-parametric) way to estimate a population value.
i.e. By analysing the sample data.

Example 4.1 - *Different Methods*
Consider estimating the population median of $U[0, \theta]$ with θ unknown.
The parametric method uses the function $\tau(\theta) = \frac{\theta}{2}$.
Considering different methods for this estimation we get

i) The *Method of Moments* finds $\hat{\theta}_{mom} = 2\bar{x} \implies \hat{\tau} = \bar{x}$;

ii) The *Maximum Likelihood Estimates* $\hat{\theta}_{mle} = x_{(n)} \implies \hat{\tau} = \frac{x_{(n)}}{2}$;

iii) The *Non-Parametric Method* gives the sample mean.

N.B. We are given 3 different values for the same quantity, for the same sample.

4.2 Repeated Sampling & Sampling Distributions

Definition 4.1 - *Sampling Distributions*
Let $X_1, \dots, X_n$ be random variables.
The *Sampling Distribution* is the distribution of the estimator $\hat{\theta}(X_1, \dots, X_n)$.

Remark 4.2 - *Good Estimator*
A *Good Estimator* should be one whose distribution is concentrated close to the true value.

Example 4.2 - *Normal Distribution Sampling Distribution*
Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$.
From previous examples, we know $\mu_{mom} = \bar{x} \implies \hat{\mu}_{mom} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Definition 4.2 - *Bias & Mean Square Error*
Let $\hat{\theta}$ be an estimator of θ.
We define two properties

i) $bias(\hat{\theta}; \theta) = \mathbb{E}(\hat{\theta} - \theta; \theta) = \mathbb{E}(\hat{\theta}; \theta) - \theta$.

ii) Mean Square Error, $mse(\hat{\theta}, \theta) = \mathbb{E}[(\hat{\theta} - \theta)^2; \theta]$.

N.B. Bias quantifies systematic error.

Definition 4.3 - *Unbiased*
We say an estimator $\hat{\theta}$ is *Unbiased* if $bias(\hat{\theta}; \theta) = 0 \ \forall \ \theta$.

Proposition 4.1 - *Mean Square Error Identity*

$$mse(\hat{\theta}; \theta) = Var(\hat{\theta}; \theta) + bias(\hat{\theta}; \theta)^2$$

Proof 4.1 - *Mean Square Error Identity*

$$\begin{aligned} mse(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta))^2; \theta] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] + \mathbb{E}[(\mathbb{E}(\hat{\theta}) - \theta)^2] + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] \\ &= Var(\hat{\theta}) + (\mathbb{E}(\hat{\theta}) - \theta)^2 + 2(\mathbb{E}(\hat{\theta}) - \theta)\mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta})) \\ &= Var(\hat{\theta}) + (\mathbb{E}(\hat{\theta}) - \theta)^2 + 2(\mathbb{E}(\hat{\theta}) - \theta)) \times 0 \\ &= Var(\hat{\theta}) + bias(\hat{\theta})^2 \end{aligned}$$

Example 4.3 -
From the previous example we have $\hat{\mu}_{mom} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
We deduce

$$bias(\hat{\mu}_{mom}; \mu, \sigma^2) = \mathbb{E}(\hat{\mu}_{mom}; \mu, \sigma^2) - \mu = \mu - \mu = 0$$

Thus

$$mse(\hat{mu}_{mom}; \mu, \sigma^2) = Var(\hat{mu}_{mom}; \mu, \sigma^2) + bias(\hat{mu}_{mom}; \mu, \sigma^2)^2 = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n}$$

4.3 Approximating Sampling Distributions by Simulation

Remark 4.3 -
A more general, but empirical, approach for *Approximating a Sampling Distributions* is to use a computer to perform *Simulations*.

Definition 4.4 - *Simulation*
A *Simulation* is the process of artificially generating a data set of independent observations from a given probability distribution.

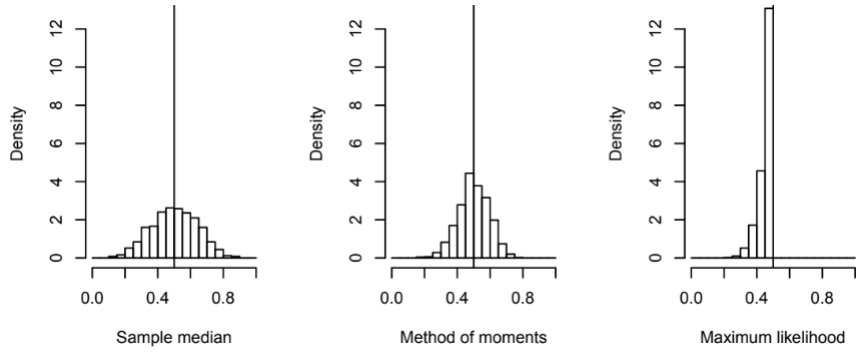
Proposition 4.2 - *R Pseudocode*
The following R pseudocode can be used to generate B data sets of size n from a distribution f and then approximating the sampling distribution

1. Generate nB values from f & group them into B groups of size n.
2. Calculate the estimates for each sample.
3. Analyse the results (numerically or graphically).

N.B. When analysing we can ask questions like *"What is the probability that* $\hat{\theta} \geq 2$ *when* $\theta = 1$*?"*.

Example 4.4 - *Graphical Summaries of Performance-Histograms*
Consider again the problem of estimating the population median for the distribution $U[0, \theta]$.
We can construct histograms for this as follows

1. Simulate 1000 samples of size 10 from $U[0, 1]$.
2. For each sample compute the sample median, the method of moments estimate $\bar{x}$ and the maximum likelihood estimate $max\{x_1, \ldots, x_{10}\}/2$
3. Plot the histograms for the estimate for each sample using each method



Proposition 4.3 - *Numerical Summaries of Performance*
We can sometimes derive explicit analytic expressions for the bias and mean-square-error of an estimator, for any value of θ.
Say we want to estimate $\tau(\theta)$ and that we can simulate $f(\cdot; \theta)$.
This can be done as follows

1. Generate B data sets of size n, made of independent variables st $X_i \sim f(\cdot; \theta)$;
2. Produce B estimates of $\hat{\tau}_i = \tau(\hat{\theta}_i)$.
3. Calculate sample mean $\bar{\tau} = \sum_{i=1}^{B} \frac{\hat{\tau}_i}{B}$ & sample variance $s = \sum_{i=1}^{B} \frac{(\hat{\tau}_i - \bar{\tau})^2}{B-1}$.
4. Compute the average error, $\bar{\tau} - \tau(\theta)$.
5. Estimate the mean-square-error

$$average\ squared - error = \frac{\sum_{i=1}^{B} (\hat{\tau}_i - \tau)^2}{B}$$

$$average\ squared - error \simeq s + (average\ error)^2$$

4.4 Approximation Methods from the Central Limit Theorem

Theorem 4.1 - *Central Limit Theorem*
Let $X_1, \dots, X_n$ be a random sample from a population with mean $\mu = \mathbb{E}(X)$ & variance $\sigma 62 = Var(X)$.
Let $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ be the sample mean.
For large n, whatever the distribution of X

$$\mathbb{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) \simeq \mathbb{P}(N(0,1) \leq x) = \Phi(x)$$

Example 4.5 - *Central Limit Theorem*
Let $X_1, \dots, X_{10}$ be a random sample from $Exp(2)$.
Then $\mu = frac12$ & $\sigma^2 = \frac{1}{4}$.
By the *Central Limit Theorem*

$$\frac{\bar{X}_{10} - \mu}{\sigma/\sqrt{n}} = \frac{\frac{1}{10}(X_1 + \cdots + X_{10}) - \frac{1}{2}}{1/(2\sqrt{10})} \simeq N(0,1)$$

If we want to approximate $\mathbb{P}(X_1 + \cdots + X_{10} \leq 5.2; \lambda = 2)$ then

$$\begin{aligned}
\mathbb{P}(X_1 + \cdots + X_{10} \leq 5.2; \lambda = 2) &= \mathbb{P}(\bar{X}_{10} \leq \tfrac{5.2}{10}; \lambda = 2) \\
&= \mathbb{P}\left(\frac{\bar{X}_{10} - \frac{1}{2}}{\sqrt{1/(4 \times 10)}} \leq \frac{\frac{5.2}{10} - \frac{1}{2}}{\sqrt{1/(4 \times 10)}}; \lambda = 2\right) \\
&\simeq \mathbb{P}\left(N(0,1) \leq \frac{\frac{5.2}{10} - \frac{1}{2}}{\sqrt{1/(4 \times 10)}}\right) \\
&= 0.5503
\end{aligned}$$

Definition 4.5 - *Continuity Correction*
Let X be a random variable of discrete integer values.
Let $T = X_1 + \cdots + X_n$ where X_i are distributed identically & independently to X.
Since T takes integer values *Continuity Correction* is used to include values that would round to target

$$\begin{aligned}
\mathbb{P}(T = x) &\simeq \mathbb{P}\left(x - \tfrac{1}{2} \leq N(n\mu, n\sigma^2) \leq x + \tfrac{1}{2}\right) \\
\mathbb{P}(T \leq x) &\simeq \mathbb{P}\left(N(n\mu, n\sigma^2) \leq x + \tfrac{1}{2}\right)
\end{aligned}$$

N.B. The *Central Limit Theorem* states that $\mathbb{P}(T \leq x) \simeq \mathbb{P}(N(n\mu, n\sigma^2) \leq x)$ but *Continuity Correction* improves the accuracy of our calculations.

Example 4.6 - *Continuity Correction*
Let $X_1, \dots, X_{10}$ be IID $Bernoulli(\frac{1}{4})$.
Then $\mu = \frac{1}{4}$ & $\sigma^2 = \frac{1}{4}(1 - \frac{1}{4}) = \frac{3}{16}$.
Consider $T = X_1 + \cdots + X_{10}$.
The *Central Limit Theorem* suggests $T \simeq N = (n\mu, n\sigma^2) = N\left(\frac{10}{4}, \frac{30}{16}\right)$.
Consider approximating $\mathbb{P}(T \leq 2)$.
We have that the real value is 0.5256.
Using the *Central Limit Theorem* we have

$$\begin{aligned}
\mathbb{P}(T \leq 2) &\simeq \mathbb{P}\left(N(\tfrac{10}{4}, \tfrac{30}{16} \leq 2\right) \\
&= \mathbb{P}\left(\frac{N(\frac{10}{4}, \frac{30}{16}) - \frac{10}{4}}{\sqrt{30/16}} \leq \frac{2 - \frac{10}{4}}{\sqrt{30/16}}\right) \\
&= \mathbb{P}(N(0,1) \leq -0.3651) \\
&= 0.3575
\end{aligned}$$

Using *Continuity Correction* we have

$$\begin{aligned}\mathbb{P}(T \leq 2) &\simeq \mathbb{P}(N(\tfrac{10}{4}, \tfrac{30}{16}) \leq 2.5) \\ &= \mathbb{P}\left(\frac{N(\frac{10}{4}, \frac{30}{16}) - \frac{10}{4}}{\sqrt{30/16}} \leq \frac{2.5 - \frac{10}{4}}{\sqrt{30/16}}\right) \\ &= \mathbb{P}(N(0,1) \leq 0) \\ &= 0.5\end{aligned}$$

Here we can see that *Continuity Correction* produces a better estimation of the true value.

5 Sampling Distributions Related to the Normal Distribution

5.1 Moment Generating Functions

Definition 5.1 - *Moment Generating Function*
Let X be a random variable.
We define the *Moment Generating Function* of X as

$$\mathcal{C}_X(t) := \mathbb{E}(e^{tX}) = \begin{cases} \int e^{tx} f_X(x) dx & \text{Continuous} \\ \sum_x e^{tx} \mathbb{P}(X = x) & \text{Discrete} \end{cases}$$

Proposition 5.1 - *Uniqueness of Moment Generating Function*
The *Moment Generating Function* is unique for all distributions.
This means it can be used to show that two distributions are equivalent.

Proposition 5.2 - *Common Moment Generating Functions*

$$\begin{aligned} X \sim N(\mu, \sigma^2) &\Leftrightarrow \mathcal{M}_X(t) = e^{\mu t + \frac{1}{2}(\sigma^2 t^2)} && t \in \mathbb{R} \\ X \sim Exp(\theta) &\Leftrightarrow \mathcal{M}_X(t) = \frac{\theta}{\theta - t} && t < \theta \\ X \sim Gamma(\alpha, \beta) &\Leftrightarrow \mathcal{M}_X(t) = \frac{\beta^\alpha}{(\beta - t)^\alpha} && t < \beta \end{aligned}$$

Theorem 5.1 - *Linear Expressions & Moment Generating Functions*
Let X & Y be random variables with $Y = aX + b$ for $a, b \in \mathbb{R}$. Then

$$\mathcal{M}_Y(t) = \mathbb{E}(e^{tY} = \mathbb{E}(e^{taX+tb}) = e^{tb}\mathcal{M}_X(ta)$$

Definition 5.2 - *Joint Moment Generating Function*
Let X & Y be random variables. Then

$$\mathcal{M}_{X,Y}(s,t) := \mathbb{E}(e^{sX+tY})$$

Theorem 5.2 - *Marginal Moment Generating Functions*
Let X & Y be random variables. Then

$$\begin{aligned} \mathcal{M}_X(s) &= \mathbb{E}(e^{sX}) &= \mathcal{M}_{X,Y}(s,0) \\ \mathcal{M}_Y(t) &= \mathbb{E}(e^{tX}) &= \mathcal{M}_{X,Y}(0,t) \end{aligned}$$

Theorem 5.3 - *Independent Variables & Moment Generating Function*
Let X & Y be random variables.
X & Y are independent iff

$$\mathcal{M}_{X,Y}(s,t) = \mathcal{M}_X(s)\mathcal{M}_Y(t) = \mathcal{M}_{X,Y}(s,0)\mathcal{M}_{X,Y}(0,t)$$

Remark 5.1 - *Sum of Independent Variables & Moment Generating Function*
Let $X_1, \ldots, X_n$ be iid & define $Y = X_1 + \cdots + X_n$. Then

$$\mathcal{M}_Y(t) = \mathcal{M}_{X_1}(t)\mathcal{M}_{X_2}(t) \ldots \mathcal{M}_{X_n}(t)$$

5.2 Transforming, Adding & Sampling Normals

Theorem 5.4 - *Linear Combinations on Normal Distribution*
Let $X \sim N(\mu, \sigma^2)$. Then

$$\begin{aligned} aX + b &\sim N(a\mu + b, a^2\sigma^2) \\ \frac{X-\mu}{\sigma} &\sim N(0,1) \end{aligned}$$

N.B. The second result is an implementation of the first where $a = \frac{1}{\sigma}$ & $b = \frac{-\mu}{\sigma}$.

Proof 5.1 - *Linear Combinations on Normal Distribution*
We know that $\mathcal{M}_X(t) = e^{\mu t + \frac{1}{2}(\sigma^2 t^2)}$. Then

$$\begin{aligned} \mathcal{M}_{aX+b}(t) &= e^{bt}\mathcal{M}_X(at) \\ &= e^{bt}e^{\mu at + \frac{1}{2}(\sigma^2 a^2 t^2)} \\ &= e^{(\mu a + b)t + \frac{1}{2}(\sigma^2 a^2 t^2)} \end{aligned}$$

This is the moment generating function for $N(a\mu + b, a^2\sigma^2)$.
For the second part we notice that

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) = \frac{\sqrt{n}}{\sigma}\bar{X} - \frac{\sqrt{n}}{\sigma}\mu \equiv a\bar{X} + b$$

We apply $a\bar{X} + b \sim N(a\mu - b, a^2\frac{\sigma^2}{n})$ we get

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \sim N\left(\frac{\sqrt{n}}{\sigma}\mu - \frac{\sqrt{n}}{\sigma}\mu, \frac{n}{\sigma^2}\frac{\sigma^2}{n^2}\right) = N(0,1)$$

Theorem 5.5 - *Adding Normal Distributions*
Let $X_1, \ldots, X_n$ be iid st $X_i \sim N(\mu_i, \sigma_i^2)$.
Then for any linear combination we have

$$\sum_i a_i X_i \sim N\left(\sum_i a_i\mu_i, \sum_i a_i^2\sigma_i^2\right)$$

Proof 5.2 - *Adding Normal Distributions*

$$\begin{aligned} \mathcal{M}_{\sum a_i X_i}(t) &= \textstyle\prod_{i=1}^n \mathcal{M}_{a_i X_i}(t) \\ &= \textstyle\prod_{i=1}^n \mathcal{M}_{X_i}(a_i t) \\ &= \textstyle\prod_{i=1}^n \exp(\mu_i a_i t + \frac{1}{2}\sigma_i^2 a_i^2 t^2) \\ &= \exp\left((\textstyle\sum_{i=1}^n \mu_i a_i) + \frac{t^2}{2}\left(\sum_{i=1}^n \sigma_i^2 a_i^2\right)\right) \end{aligned}$$

This is the moment generating function of $N(\sum_i a_i\mu_i, \sum_i a_i^2\sigma_i^2)$.

Theorem 5.6 - *Independence of* $\bar{X}$ *and* $\sum_{j=1}^n (X_j - \bar{X})^2$
If $X_1, \ldots, X_n$ are a random sample of size n from $N(\mu, \sigma^2)$ distribution then

$$\bar{X} \text{ and } \sum_{j=1}^n (X_j - \bar{X})^2 \text{ are independent.}$$

N.B. The proof of this is non-examinable.

5.3 χ^2 Distribution

Definition 5.3 - *χ^2 Distribution*
We say that a random variable W has the χ^2 distribution with r degrees of freedom. We write $W \sim \chi^2_r$ if W has moment generating function

$$\mathcal{M}_W(t) = (1-2t)^{-r/2} \quad t < \frac{1}{2}$$

Remark 5.2 - *Symmetry of χ^2*
χ^2 cannot be symmetric around zero as it only takes positive values.

Remark 5.3 - *χ^2 & Gamma Distribution*
By comparison of moment generating functions we have that

$$\chi^2_r \equiv \Gamma\left(\frac{r}{2}, \frac{1}{2}\right)$$

Proposition 5.3 - *Expectation & Variance of χ^2 Distribution*
If $W \sim \chi^2_r$ then

$$\mathbb{E}(W) = \frac{r/2}{1/2} = r \quad Var(W) = \frac{r/2}{(1/2)^2} = 2r$$

Remark 5.4 - *χ^2 is Squared Normal*
If $Z \sim N(0,1)$ then $Y = Z^2 \sim \chi^2_1$.

Proof 5.3 - *χ^2 is Squared Normal*
Below we show that when $t < \frac{1}{2}$ then the squared normal distribution has the same moment generating function as χ^2_1.

$$\begin{aligned}
\mathcal{M}_Y &= \mathbb{E}(e^{tY}) = \mathbb{E}(e^{tZ^2}) \\
&= \int_{-\infty}^{\infty} e^{tz^2}\mathbb{P}(Z=z)dz \\
&= \int_{-\infty}^{\infty} e^{tz^2}\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}z^2}dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} exp(-\tfrac{1}{2}(z^2 - 2tz^2))dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} exp(-\tfrac{1}{2}z^2(1-2t))dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} exp(-\tfrac{1}{2}z^2/\tfrac{1}{1-2t})dz \\
&= \frac{1}{\sqrt{2\pi}} \times \sqrt{2\pi\sigma^2} \qquad \text{Where } \sigma^2 = \tfrac{1}{1-2t} \\
&= \sigma = \frac{1}{\sqrt{1-2t}}
\end{aligned}$$

Theorem 5.7 - *Sum of χ^2 Distributions*
The following hold $\forall\ r, s$

i) If $U \sim \chi^2_r$ & $U \sim \chi^2_s$ are independent, then $U + V \sim \chi^2_{r+s}$.

ii) If $Z_i^2 \sim N(0,1)$ are independent then $\sum_i Z_i^2 \sim \chi^2_n$.

Proof 5.4 - *Sum of χ^2 Distributions*
Consider the moment generating function of $U + V$

$$\begin{aligned}
\mathcal{M}_{U+V}(t) &= \mathcal{M}_U(t) + \mathcal{M}_V(t) \\
&= \frac{1}{(1-2t)^{r/2}} \times \frac{1}{(1-2t)^{s/2}} \\
&= (1-2t)^{-\frac{1}{2}(r+s)}
\end{aligned}$$

So $t \in [0, \frac{1}{2}]$ and we recognise the moment generating function of chi^s_{r+s}

5.4 Normal Sampling Distribution

Theorem 5.8 -
Let $X_1, \dots, X_n$ be a random sample of size n from $N(\mu, \sigma^2)$. Then

i) $\sum_i \frac{1}{\sigma^2}(X_j - \mu)^2 \sim \chi^2_n$; and

ii) $\sum_i \frac{1}{\sigma^2}(X_j - \bar{X})^2 \sim \chi^2_{n-1}$.

Proof 5.5 - *Theorem 5.8*

i) Writing $Y_i = \frac{1}{\sigma}(X_i - \mu)$ we have $Y_i \sim N(0,1)$.
Further Y_i are independent of $\sum_i Y_i^2 \sim \chi^2_n$.
But $\sum_i \frac{1}{\sigma^2}(X_i - \mu)^2 = \sum_i Y_i^2$.

ii) We have

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n (Y_i - \bar{Y} + \bar{Y})^2 = \sum_{i=1}^n \left[(Y_i - \bar{Y})^2 + \bar{Y}^2 + 2\bar{Y}(Y_i - \bar{Y})\right]$$

Note that $\bar{Y}\sum_{i=1}^n (Y_i - \bar{Y}) = 0$.
Let $W_1 = \sum_{i=1}^n (Y_i - \bar{Y})^2$ and $W_2 = nY^2$.
Then $\sum_{i=1}^n Y_i^2 = W_1 + W_2 = W_3$.
We have W_1 & W_2 are independent therefore

$$\mathcal{M}_{W_3}(t) = \mathcal{M}_{W_1}(t)\mathcal{M}_{W_2}(t) \implies \mathcal{M}_{W_1}(t) = \frac{\mathcal{M}_{W_3}(t)}{\mathcal{M}_{W_2}(t)}$$

Notice that $\sqrt{n}\bar{Y} = \sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \sim N(0,1)$.
Thus $W_1 \sim \chi^2_1$.
Therefore $\mathcal{M}_{W_3}(t) = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{\frac{1}{2}(n-1)}$.

5.5 t-Distribution

Definition 5.4 - *t-Distribution*
Let U & V be independent random variables with $U \sim N(0,1)$ & $V \sim \chi^2_r$.
Define $W = \frac{U}{\sqrt{V/r}}$. We say

$$W \sim t_r$$

Theorem 5.9 - *Variance & Expectation of t-Distribution*
Let $W \sim t_r$ then

$$\mathbb{E}(W) = 0 \quad Var(W) = \frac{r}{r-2}$$

Proposition 5.4 - *Properties of t-Distribution*
Let $W \sim t_r$. Then

i) W is symmetric about 0;

ii) W is similar to $N(0,1)$ but with heavier tails; and,

iii) As $r \to \infty$ we have $W \to N(0,1)$.

Theorem 5.10 -
Let $X_1, \dots, X_n$ be a random sample from $N(\mu, \sigma^2)$.
Define $\bar{X} = \frac{1}{n}\sum_i X_i$ & $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$.
We can define

i) $U := \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \sim N(0,1)$;

ii) $V := \frac{1}{\sigma^2}\sum_i (X_i \bar{X})^2$.

Then

1. U & V are independent; and
2. $\frac{\sqrt{n}}{S}(\bar{X} - \mu) \sim t_{n-1}$.

This result allows us to know how far apart we expect μ & $\bar{X}$ to be, even when σ^2 is unknown.

Proof 5.6 - *Theorem 5.10*
We have $\dfrac{S^2}{\sigma^2} = \dfrac{V}{n-1} \sim \dfrac{\chi^2_{n-1}}{n-1}$.
U & V are independent by **Theorem 5.6**.
The result is proved by

$$\frac{\sqrt{n}}{S}(\bar{X} - \mu) = \left(\frac{\sqrt{n}}{S}(\bar{X} - \mu)\right) \frac{1}{\sqrt{S^2/\sigma^2}} = U \times \frac{V}{n-1} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}$$

These terms are independent as required.

5.6 Percentage Points of Distributions

Definition 5.5 - *Percentage Points*
For a given $\alpha \in [0,1]$ we define the *Percentage Point* x_α to be the value where

$$\mathbb{P}(X \geq x_\alpha) = \alpha$$

.
N.B. Usually we use values of the magnitude of$\alpha = 0.1, 0.05, 0.025, \dots$.

Proposition 5.5 - *Percentage Points of Particular Distributions*

RV	Notation	Symmetric around 0
$Z \sim N(0,1)$	$\mathbb{P}(Z \geq z_\alpha) = \alpha$	Yes
$T \sim T_r$	$\mathbb{P}(Z \geq t_{r;\alpha}) = \alpha$	Yes
$W \sim \chi^2_r$	$\mathbb{P}(W \geq \chi^2_{r;\alpha}) = \alpha$	No

Remark 5.5 - *Inversion of Percentage Point*
From these distributions we can deduce that

$$\begin{aligned}(1-\alpha) &= \mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\ (1-\alpha) &= \mathbb{P}(-t_{r;\alpha/2} \leq T \leq t_{r;\alpha/2}) \\ (1-\alpha) &= \mathbb{P}(\chi^2_{r;1-\alpha/2} \leq W \leq \chi^2_{r;\alpha/2})\end{aligned}$$

Theorem 5.11 -
Let $X_1, \dots, X_n$ be a random sample from $Exp(\theta)$. Then

i) $\sum_{i=1}^n X_i \sim \Gamma(n, \theta)$;

ii) $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim \Gamma(n, n\theta)$; and,

iii) $2\theta\sum_{i=1}^n X_i \sim \Gamma(n, \frac{1}{2}) = \chi^2_{2n}$.

6 Confidence Interval

Definition 6.1 - *Confidence Interval*
Let $X_1, \dots, X_n$ be iid random variables with pmf $f_{X_1,\dots,X_n}(\cdot, \theta)$ for some θ.
Choose $\alpha \in (0, 1)$.
A $100(1-\alpha)\%$ *Confidence Interval* for θ is a random interval of the form (C_L, C_U) st

$$\mathbb{P}(C_L \leq \theta^* \leq C_U; \theta) \geq 1 - \alpha$$

N.B. θ^* is the true value of θ.

Definition 6.2 - *Length of a Confidence Interval*
For a *Confidence Interval* (C_L, C_U) the length is defined as the range of these value

$$|C_L - C_U|$$

Proposition 6.1 - *Procedure for Finding Confidence Interval*
This technique depends on some statistics $f(\boldsymbol{X}) = f(X_1, \dots, X_n)$, which is generally given.

i) Treat θ as unknown;

ii) Use facts about the distribution to find an interval depending on θ st

$$\mathbb{P}(g_1(\theta^*) \leq f(\boldsymbol{X}) \leq g_2(\theta^*); \theta) = 1 - \alpha$$

iii) Invert the interval so that it is an inequality about θ

$$\mathbb{P}(C_L(f(\boldsymbol{X})) \leq \theta^* \leq c_U(f(\boldsymbol{X})); \theta) = 1 - \alpha$$

Example 6.1 - *Confidence Interval for μ in $N(\mu, \sigma_0^2)$*
Consider a sample $X_1, \dots, X_n$ from $N(\mu, \sigma_0^2)$.
We know $\bar{X} \sim N(\mu, \frac{1}{n}\sigma^2)$ & $\dfrac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1)$.
We shall calculate a confidence interval for the value of μ with $\alpha = 0.05$.
We know that $z_{0.05/2} = z_{0.025} = 1.96$. Then

$$\mathbb{P}\left(-1,96 \leq \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \leq 1.96; \mu, \sigma_0^2\right) = 0.95$$

But

$$\begin{aligned} & \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} && \leq && 1.96 \\ \implies & \bar{X} - \mu && \leq && 1.96\tfrac{\sigma_0}{\sqrt{n}} \\ \implies & \bar{X} - 1.96\tfrac{\sigma_0}{\sqrt{n}} && \leq && \mu \end{aligned}$$

Similarly

$$\begin{aligned} & \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} && \geq && -1.96 \\ \implies & \bar{X} - \mu && \geq && -1.96\tfrac{\sigma_0}{\sqrt{n}} \\ \implies & \bar{X} + 1.96\tfrac{\sigma_0}{\sqrt{n}} && \geq && \mu \end{aligned}$$

We have the event $\left\{-1.96 \leq \dfrac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}\right\} \equiv \left\{\bar{X} - 1.96\tfrac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\tfrac{\sigma_0}{\sqrt{n}}\right\}$.
Hence

$$\mathbb{P}\left(\bar{X} - 1.96\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma_0}{\sqrt{n}}; \mu, \sigma_0^2\right) = 0.95$$

Proposition 6.2 - *General Confidence Interval*
Consider a more general $100(1-\alpha)\%$ confidence interval for μ using a random sample of $N(\mu, \sigma_0^2)$.

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}; \mu; \sigma_0^2\right) = 1-\alpha$$

Rearranging we find

$$\mathbb{P}\left(\bar{X} - \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}\right) = 1-\alpha$$

Remark 6.1 - *Size of General Confidence Interval*
Consider the confidence interval identified in **Proposition 6.2**, we see that this interval

i) Decreases as sample size n increases;

ii) Increases as population variance σ_0^2 increases;

iii) Increases as confidence level $100(1-\alpha)$ increases.

Proposition 6.3 - *Confidence Interval for μ in $N(\mu, \sigma 62)$ with unknown σ^2*
Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ with σ^2 unknown.
By **Theorem 5.10** we have that

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Let $T \sim t_{n-1}$ with $\mathbb{P}(T \geq t_{n-1;\alpha/2}) = \alpha/2$.
By symmetry $\mathbb{P}(T \leq -t_{n-1;\alpha/2}) = \alpha/2$.
Thus

$$\mathbb{P}\left(-t_{n-1;\alpha/2} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{n-1;\alpha/2}; \mu, \sigma^2\right) = 1-\alpha$$

By rearranging

$$\mathbb{P}\left(\bar{X} - \frac{St_{n-1;\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{St_{n-1;\alpha/2}}{\sqrt{n}}\right) = 1-\alpha$$

N.B. The length of this interval is random, not fixed.

6.1 Confidence Interval for σ^2 for $N(\mu, \sigma^2)$ Data with μ Unknown

Proposition 6.4 - *Confidence Interval for $\sigma 62$ for $N(\mu, \sigma^2)$ Data with μ Unknown*
Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ with σ^2 unknown.
Then

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}(X_j - \bar{X})^2 \sim \chi^2_{n-1}$$

Hence $\forall\ \alpha$

$$\mathbb{P}\left(\chi^2_{n-1;1-\alpha/2} \leq \frac{1}{\sigma^2}\sum_{j=1}^{n}(X_j-\bar{X})^2 \leq \chi^2_{n-1;\alpha/2}; \mu, \sigma^2\right) = 1-\alpha$$

Rearranging we get

$$\mathbb{P}\left(\frac{1}{\chi^2_{n-1;\alpha/2}}\sum_{j=1}^{n}(X_j-\bar{X})^2 \leq \sigma^2 \leq \frac{1}{\chi^2_{n-1;1-\alpha/2}}\sum_{j=1}^{n}(X_j-\bar{X})^2\right) = 1-\alpha$$

6.2 Confidence Interval for θ in $U(0,\theta)$

Proposition 6.5 - *Confidence Interval for θ in $U(0,\theta)$*
Let $X_1,\dots,X_n$ be a random sample from $U(0,\theta)$.
We know that $\hat{\theta}_{MLE} = \max(X_1,\dots,X_n)$ so

$$\begin{aligned} prob(X_{(n)} \le x;\theta) &= \mathbb{P}(X_1,\dots x,\dots,X_n \le x;\theta) \\ &= \mathbb{P}(X_1 \le x;\theta)\dots\mathbb{P}(X_n \le x;\theta) && \text{Independence} \\ &= \mathbb{P}(X \le x;\theta)^n && \text{IdenticallyDistributed} \\ &= \begin{cases} \left(\frac{x}{\theta}\right)^n & \text{if } 1 \le x \le \theta \\ 0 & \text{if } x \le 0 \\ 1 & \text{if } x \ge \theta \end{cases} \end{aligned}$$

We deduce that $\frac{X_{(n)}}{\theta}$ does not depend on θ.

$$\mathbb{P}\left(\frac{X_{(n)}}{\theta} \le \frac{x}{\theta}\right) = \left(\frac{x}{\theta}\right)^n$$

Set $u^{1/n} = \frac{x}{\theta}$

$$\implies \mathbb{P}\left(\frac{X_{(n)}}{\theta} \le u^{1/n}\right) = u$$

Let $u_1, u_2 \in [0,1]$ with $u_1 < u_2$

$$\begin{aligned} \mathbb{P}\left(\tfrac{X_{(n)}}{\theta} \le u_2^{1/n}\right) - \mathbb{P}\left(\tfrac{X_{(n)}}{\theta} \le u_1^{1/n}\right) &= u_2 - u_1 \\ \equiv \quad \mathbb{P}\left(u_1^{1/n} \le \tfrac{X_{(n)}}{\theta} \le u_2^{1/n}\right) &= u_2 - u_1 \end{aligned}$$

Choosing $u_2 = 1-\alpha/2$ & $u_1 = \alpha/2$ we get

$$\mathbb{P}\left(\frac{X_{(n)}}{u_2^{1/n}} \le \theta \le \frac{X_{(n)}}{u_1^{1/n}}\right) = 1-\alpha$$

Remark 6.2 - *The rest of chapter 6 is non-examinable*

7 Hypothesis Testing

Definition 7.1 - *Hypothesis Test*
A *Hypothesis Test* is a procedure for evaluating whether sample data is consistent with one of two contrasting statements about the value of the population parameters.

Remark 7.1 - *Types of Hypothesis Tests*
Suppose we want to test the value of a parameter θ, with *Null Hypothesis* that $\theta = \alpha$.
There are three hypothesis tests we can use

i) $\theta > \alpha$;

ii) $\theta < \alpha$; and,

iii) $\theta \neq \alpha$.

Proposition 7.1 - *Hypothesis Testing Procedure*
A Hypothesis test can be performed by the following procedure

i) State model assumptions,

ii) State null hypothesis & alternative hypotheses,

iii) Choose & calculate the value of an appropriate test statistic,

iv) Compute the resulting p-value.

v) Make conclusions.

Remark 7.2 - *iii) Test Statistic*
We define a suitable test statistics $T(X_1, \dots, X_n)$ which has the following properties

i) Extreme values are unlikely if H_0 is true;

ii) Extreme values are consistent with H_1; and,

iii) When H_0 is true then the distribution of T is known and its distribution function can be easily calculated.

Remark 7.3 - *iv) p-Value*
Let T_{obs} be the observed value of the test statistic.
We calculate the *p-Value* depending upon the *Alternative Hypothesis*

i) $H_1 : \mu > \mu_0 \implies p$-Value$= \mathbb{P}(T \geq t_{obs} | H_0 \text{ true})$;

ii) $H_1 : \mu < \mu_0 \implies p$-Value$= \mathbb{P}(T \leq t_{obs} | H_0 \text{ true})$;

iii) $H_1 : \mu \neq \mu_0 \implies p$-Value$= \mathbb{P}(|T| \geq |t_{obs}| | H_0 \text{ true})$;

Remark 7.4 - *v) Conclusion*
If the *p-Value* is very small we believe that the null hypothesis is false.

Example 7.1 - *Hypothesis Testing for Normal Distribution with Known Variance*
Consider a medication that hsa a mean time to recurrence of an illness of $\mu_0 = 53.3$, with $\sigma_0 = 26.4$.
A new medication is trialled on 16 patients. For the sample $\bar{X} = 65.8$.
Assuming the variance is the same for both medications, does the new medication have a longer mean time to recurrence?

i) *Model.*
We assume the recurrence time for the patients of the new medication is distributed $N(\mu, \sigma_0^2) = N(\mu, 26.4^2)$.

ii) *Hypothesis.*
$H_0 : \mu = 53.5 = \mu_0$ & $H_1 : \mu > 53.3 = \mu_0$

iii) *Test Statistics.*
Define $T(X_1, \dots, X_{16}) = \sqrt{n}\left(\dfrac{\bar{X} - \mu_0}{\sigma_0}\right) = \sqrt{16}\left(\dfrac{\bar{X} - 53.5}{26.4}\right)$.
Since $\bar{X} \sim N(\mu, \frac{\sigma_0^2}{n})$ then $T(X_1, \dots, X_{16}) \sim N\left(\sqrt{n}\frac{\mu - \mu_0}{\sigma_0}, 1\right)$.
This means that if H_0 is true then $T \sim N(0, 1)$.
We calculate the observed statistic to be $t_{obs} = \sqrt{16}\left(\frac{65.8 - 53.3}{26.4}\right) = 1.893$.

iv) *p-Value.*
Here $p-value = \mathbb{P}(T \geq t_{obs}; \mu = \mu_0, \sigma = \sigma_0)$.
Since $T \sim N(0,1)$ then

$$\begin{aligned} p-value &= \mathbb{P}(Z \geq t_{obs}; \mu = \mu_0, \sigma = \sigma_0 \\ &= \mathbb{P}(Z \geq 1.893) \\ &= 1 - \mathbb{P}(Z \leq 1.893) \\ &= 1 - \Phi(1.893) \\ &= 0.0292 \end{aligned}$$

v) *Conclusion* This value is sufficiently small, so we reject H_0 in favour of H_1.

Proposition 7.2 - *Confidence Interval for Normal Distribution with Mean & Variance Unknown*

i) Assume $X_1, \ldots, X_n \sim (\mu, \sigma^2)$ with μ & σ^2 unknown.

ii) Define $H_0 : \mu = \mu_0$ & $H_1 = \mu \neq \mu_0$ (Other H_1s are possible).

iii) When H_0 is true $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

iv) Since $H_1 : \mu \neq \mu_0$ we want to check the event $\{|T| \geq |t_{obs}|\}$. Then

$$p-value = \mathbb{P}(|T| \geq |t_{obs}\ |H_0\ is\ true) = \mathbb{P}(|t_{n-1}| \geq |t_{obs}|)$$

7.1 Critical Region

Definition 7.2 - *Critical Region*
the *Critical Region* is the set of values, at a given significance level, which if the test-statistic falls within we reject H_0.

Definition 7.3 - *Critical Value(s)*
*Critical Value*s are values which are the threshold for H_0 being rejected.
N.B. These are denoted c^*.

Example 7.2 - *Critical Region*
Define $H_0 : \mu = \mu_0$ & $H_1 : \mu > \mu_0$.
Let $T := \sqrt{n}\frac{(\bar{X}-\mu}{\sigma_0}$ & α be the significance level.
The the critical region is

$$C := \{x_1, \ldots, x_n\} : T(x_1, \ldots, x_n) \geq c^*) \equiv \{T \geq c^*\}$$

Then

$$\mathbb{P}(H_0\ is\ rejected | H_0\ true) = \mathbb{P}(Type\ 1\ Error) = \alpha$$

for this example this is

$$\mathbb{P}(C | H_0\ is\ True) = \mathbb{P}(T \geq c^* | H_0\ is\ true) = \alpha$$

Since $T \sim N(0,1)$ we have

$$\begin{aligned} \alpha &= \mathbb{P}(T \geq c^* | H_0\ is\ true) \\ &= \mathbb{P}(| \geq c^*) \\ &- 1 - \mathbb{P}(| \leq c^*) \\ &= 1 - \Phi(c^*) \\ \implies c^* &= \phi^{-1}(1-\alpha) \end{aligned}$$

For $\alpha = 0.05 \implies c^* = \phi^{-1}(0.095) = 1.645$.
Thus we reject H_0 if $T \geq c^* = 1.645$.
Equivalently, we reject if $\bar{X} \geq \mu + \frac{1.645\sigma_0}{\sqrt{n}}$ by definition of T.

Remark 7.5 - *Confidence Intervals & Hypothesis Testing*
Consider using a test statistic T which is normally distributed to test $H_0 : \mu = \mu_0$.
If we know $\sigma^2 = \sigma_0^2$, by hypothesis testing, we reject H_0 if

$$z_{\alpha/2} = c^* \leq |T| == \left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0} \right|$$

Rearranging we can find a confidence interval for rejecting H_0

$$\mu_0 \notin \left[\bar{X} - \frac{c^*\sigma_0}{\sqrt{n}}, \bar{X} + \frac{c^*\sigma_0}{\sqrt{n}} \right]$$

8 Comparison of Population Means

Remark 8.1 - *Comparing two groups*
In many real life situations we collected data in order to compare two groups.
There are two possible relationships these two groups can have

i) *Independent Samples.*
Each data set is entirely independent of one another.
In this case each group can be modelled by different population distributions.
e.g. - Patients in different groups given different medication.

ii) *Paired Samples*
Here the data consists of pairs of observations for each member of the population.
e.g. - Patients given drug A & after some time given drug B

Example 8.1 - *Two Sample t-test*
Consider being given two independent samples and asked to test whether they have the same mean.
Let $X_1, \ldots, X_n$ be a random sample of size n from $N(\mu_X, \sigma_X^2)$ & $Y_1, \ldots, Y_m$ be a random sample of size n from $N(\mu_Y, \sigma_Y^2)$.
Define $H_0 : \mu_X - \mu_Y = 0$ & $H_1 : \mu_X - \mu_Y \neq 0$.
Our standard estimators are $\bar{X}$ & $\bar{Y}$ so our analysis will be based on the value of $\bar{X} - \bar{Y}$.
We have $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$ & $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$ so $\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}\right)$.
Let $U = \dfrac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{(\sigma_X^2/n) + (\sigma_Y^2/n)}} \sim N(0,1)$.
If H_0 is true then $U = \dfrac{\bar{X} - \bar{Y}}{\sqrt{(\sigma_X^2/n) + (\sigma_Y^2/n)}} \sim N(0,1)$.
If we know the variance we can compute the p-value & test the hypotheses.
N.B. Generally we have to estimate σ_X^2 & σ_Y^2 with the sample variance.

Remark 8.2 - *Pooled Estimate, S_p^2*
If can assume that two samples (X & Y), have the same variance we can combine the estimators of their variance into a single pooled estimate S_p^2.

$$S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{(n-1) + (n-1)} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{n + m - 2}$$

Then the test statistic is $T = \dfrac{(\bar{X} - \bar{Y})}{S_P\sqrt{(1/n) + (1/m)}} \sim T_{m+n-2}$.
From **Theorem 5.8** $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_X^2} \sim \chi^2_{n-1}$ & $\frac{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{\sigma_Y^2} \sim \chi^2_{n-1}$.
Due to the independence of the samples the sum of these two result of a χ^2_{m+n-2} distribution.
Thus

$$\frac{S_p^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{\sigma^2(n + m - 2} \sim \chi^2_{n+m-2)}$$

N.B. From **Definition 5.4** we deduce that $T \sim t_{n+m-2}$.

9 Linear Regression

Remark 9.1 - *Motivation*
Here we consider a one-dimensional data set x & a linear dependent $Y = ax + b$ with $a, b \in \mathbb{R}$.

Definition 9.1 - *Predictor & Response Variable*
Consider a date set x & linear dependent $Y = ax + b$.
Here x is called the *Predictor Variable* & Y is the *Response Variable*.

Remark 9.2 - *Random Effects*
We need to take account of random variables in the relationship between the *Predictor Variable* & *Response Variable*.
i.e - If we took multiple samples of x we would get different values of Y.

Definition 9.2 - *Linear Regression Model*
The simple *Linear Regression Model* says

$$\mathbb{E}(Y|x) = a + bx \text{ for } a, b \in \mathbb{R}$$

N.B. - If $b = 0$ then x & Y are independent.

Proposition 9.1 - *Model Assumptions*
Let $\{x_1, \ldots, x_n\{$ be observed values of predictor variable x.
Let y_i b the response variable to x_i with $Y_i = a + bx_i + e_i$ where e_i is a random variable & $a, b \in \mathbb{R}$ are unknown.
The assumptions we make are

1. $\mathbb{E}(e_i) = 0$;
2. $Var(e_i) = \sigma^2$ which is unknown; and,
3. $Cov(e_i, e_j) = 0$ for $i \neq j$ (*i.e* Errors are uncorrelated).

Definition 9.3 - *Summary Statistics*
In order to simplify notation we introduce *Summary Statistics* of a data sample

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ & $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$;
- $ss_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}\left(x_i^2\right) - n\bar{x}^2$ & $ss_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left(y_i^2\right) - n\bar{y}^2$;
- $ss_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}$.

9.1 Least Squares Estimates

Definition 9.4 - *Least Squares Estimates*
The *Least Squares Estimate* is the parameter values for

$$\text{argmin}_{a,b} \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

N.B. - These values are denoted as $\hat{a}$ & $\hat{b}$.

Theorem 9.1 - *Finding $\hat{a}$ & $\hat{b}$*
For the simple linear regression model, the least squares estimates $\hat{a}$ & $\hat{b}$ are given by

- $\hat{b} = \frac{ss_{xy}}{ss_{xx}}$; and
- $\hat{a} = \bar{y} - \hat{b}\bar{x}$.

Proof 9.1 - *Theorem 9.1*
We have

$$\begin{aligned} (y_i - (a + bx_i))^2 &= (y_i - \bar{y})^2 + b^2(x_i - \bar{x})^2 + (a - \bar{y} + b\bar{x})^2 \\ &+ 2\big(- b(y_i - \bar{y})(x_i - \bar{x}) - (a - \bar{y} + b\bar{x})[y_i - \bar{y} - b(x_i - \bar{x})]\big) \\ \implies \sum_{i=1}^{n}(y_i - (a + bx_i))^2 &= ss_{yy} + b^2 ss_{xx} + n(a - \bar{y} + b\bar{x})^2 - 2ss_{xy}b + 0 \end{aligned}$$

since $\sum_{i=1}^{n} x_i - \bar{x} = \sum_{i=1}^{n} y_i - \bar{y} = 0$.
Note that $n(a - \bar{y} + b\bar{x})^2 \geq 0$.
Further, for $\hat{b}$ given $\hat{a} = \bar{y} - \hat{b}\bar{x}$ we have $n(a - \bar{y} + b\bar{x})^2 = 0$.
We minimise the rest of the expression $(ss_{yy} + b^2 ss_{xx} - 2ss_{xy}b$ by varying b.
Differentiating & setting to 0 we find

$$2bss_{xx} - 2ss_{xy} = 0 \implies \hat{b} = \frac{ss_{xy}}{ss_{xx}}$$

9.2 Fitted Values, Residuals & Predictions

Definition 9.5 - *Fitted Values*
The *Fitted Values* are the estimated values, under the model, for the observed values

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

Definition 9.6 - *Residual Values*
The *Residual Values* are the difference between observed values & fitted values

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$$

Definition 9.7 - *Residual Sum of Squares, RSS*
We define the *Residual Sum of Squares* as

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2$$

Definition 9.8 - *Best Predictor*
The *Best Predictor* of the value of Y that would be observed for some x value for which we have no data is

$$\hat{y} = \hat{a} + \hat{b}x$$

Proposition 9.2 - *Properties of Residual Sum of Squares*
We can deduce a formula for *Residual Sum of Squares*

$$RSS = ss_{yy} - \frac{ss_{xy}^2}{ss_{xx}}$$

We can estimate variance as $\hat{\sigma}^2 = \frac{1}{n-2}RSS$
Remark 9.3 - *Assessing Fit of a Model*
One way to examine the fit of a model is by examining a plot of the residuals $\hat{e}_1, \dots, \hat{e}_n$ against either: the predictor values $x_1, \dots, x_n$; or, the fitted values $\hat{y}_1, \dots, \hat{y}_n$.
We expect to see a roughly symmetric distribution of the residuals about 0 & very few extreme outliers.

Remark 9.4 - *If Model seems Bad*
If it appears that a model does not fit well we may change the model to allow the error variance to depend on x or could change the formula to be non-linear.
i.e. $\mathbb{E}(Y|x) = a + bx + cx^2$.

10 Linear Regression: Confidence Intervals & Hypothesis Tests

10.1 Simple Normal Linear Regression

Remark 10.1 - *Adjusting Linear Regression Model for Testing*
In order to use a linear regression model to perform hypothesis tests & find confidence intervals we need to make an extra assumption about the distribution of the residuals $e_1, e_2, \dots$.

Definition 10.1 -
Let $x_1, \dots, x_n$ be values for predictor variable X & assume that hte vlaue of y_i is the response variable for x_i, from random variable Y_i.
We assume that

$$Y_i = a + bx_i + e_i$$

where e_i are IID with $s_i \sim N(0, \sigma^2)$ and a, b, σ^2 are unknown.

10.2 Properties of $\hat{a}, \hat{b}$ & $\hat{\sigma}^2$

Theorem 10.1 - *Distributions of $\hat{a}$ & $\hat{b}$*
If $e_1, \dots, e_n$ are normally distributed then

i) $\hat{b} \sim N(b, \sigma^2 \frac{1}{ss_{xx}})$

ii) $\hat{a} \sim N(a, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{ss_{xx}}))$

N.B. These means & variances hold without assuming that $e_1, \dots, e_n$a re normally distributed.

Proof 10.1 - *Theorem 10.1*
We first establish that $\hat{b}$ is normally distributed.
We know that $\hat{b} = \dfrac{ss_{xy}}{ss_{xx}} = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{x})}{ss_{xx}}$.
Since $\sum_{i=1}^n (x_i - \bar{x}) = 0$ then $\sum_{i=1}^n (x_i - \bar{x})(-\bar{y}) = 0$.
As a result $\hat{b} = \sum_{i=1}^n \dfrac{(x_i - \bar{x})}{ss_{xx}} y_i = \sum_{i=1}^n b_i Y_i$ where b_i is some constant.

From **Theorem 5.5** we know that $\sum_{i=1}^n b_i Y_i$ is normally distributed.
Now

$$\begin{aligned}
\mathbb{E}(\hat{b}) &= \mathbb{E}\left(\sum_{i=1}^n \frac{1}{ss_{xx}}(x_i-\bar{x})Y_i\right) \\
&= \sum_{i=1}^n \left(\mathbb{E}\frac{1}{ss_{xx}}(x_i-\bar{x})Y_i\right) \\
&= \sum_{i=1}^n \frac{1}{ss_{xx}}(x_i-\bar{x})\mathbb{E}(Y_i) \\
&- \sum_{i=1}^n \frac{1}{ss_{xx}}(x_i-\bar{x})(a+bx_i) \\
&= b\sum_{i=1}^n \frac{1}{ss_{xx}}(x_i-\bar{x})x_i \\
&= b\sum_{i=1}^n \left(\frac{(x_i-\bar{x})(x_i-\bar{x})}{ss_{xx}} + \frac{(x_i-\bar{x})\bar{x}}{ss_{xx}}\right) \\
&= b(1+0) \\
&= b
\end{aligned}$$

10.3 t-Distributions for $\hat{\alpha}$ & $\hat{\beta}$

Theorem 10.2 - *Estimating $\hat{\sigma}^2$*
We have that

$$\frac{1}{\sigma^2}(n-2)\hat{\sigma}^2 \sim \chi^2_{n-2} \iff \frac{1}{\sigma^2}\sum_{i=1}^n \hat{e}_i^2 \sim \chi^2_{n-2}$$

Note that this is independent of $\hat{\alpha}$ & $\hat{\beta}$.
Thus $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$ & $Var(\hat{\sigma}^2) = \frac{2}{n-2}\sigma^4$.

Theorem 10.3 - *t-Distributions for $\hat{\alpha}$ & $\hat{\beta}$*
Define $s_{\hat{\alpha}} = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{ss_{xx}}\right)}$ to estimate the standard deviation of $\hat{\alpha}$.
Define $s_{\hat{\beta}} = \sqrt{\hat{\sigma}^2/ss_{xx}}$ to estimate the standard deviation of $\hat{\beta}$.
Then

$$\frac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}} \sim t_{n-2} \quad \frac{\hat{\beta}-\beta}{s_{\hat{\beta}}} \sim t_{n-2}$$

Proof 10.2 - *Theorem 10.3*
First consider the distribution $\hat{\alpha}$.
By **Theorem 10.1** we have

$$\frac{\hat{\alpha}-\alpha}{\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{ss_{xx}}}} \sim N(0,1)$$

Further, **Theorem 10.2** states

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-2}}{n-2}$$

By the definition for the t-distribution we know that

$$\frac{\hat{\alpha}-\alpha}{s_{\hat{\alpha}}} = \frac{\hat{\alpha}-\alpha}{\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{ss_{xx}}}}\frac{1}{\sqrt{\hat{\sigma}^2/\sigma^2}} \sim \frac{N(0,1)}{\sqrt{\frac{1}{n-2}\chi^2_{n-2}}} = t_{n-2}$$

Now consider the distribution $\hat{\beta}$.
By **Theorem 10.1** we have

$$\frac{\hat{\beta}-\beta}{\sigma\sqrt{1/ss_{xx}}} \sim N(0,1)$$

By the definition for the t-distribution

$$\frac{\hat{\beta}-\beta}{s_{\hat{\beta}}} = \frac{\hat{\beta}-\beta}{\sigma\sqrt{1/ss_{xx}}}\frac{1}{\sqrt{\hat{\sigma}^2/\sigma^2}} \sim \frac{N(0,1)}{\sqrt{\frac{1}{n-2}\chi^2_{n-2}}} = t_{n-2}$$

10.4 Confidence Intervals for α & β

Proposition 10.1 - *Confidence Intervals for α & β*
Let γ be our significance level.
We want to find an interval (CL_α, CU_α) st $\mathbb{P}(CL_\alpha \leq \alpha \leq CU_\alpha) = 1-\gamma$.
Using **Theorem 10.3**

$$\begin{aligned} 1-\gamma &= \mathbb{P}\left(-t_{n-2,\gamma/2} \leq \frac{\hat{\alpha}-\alpha}{s_\alpha} \leq t_{n-2,\gamma/2}\right) \\ &= \mathbb{P}(-s_\alpha t_{n-2,\gamma/2} - \hat{\alpha} \leq -\alpha \leq s_\alpha t_{n-2,\gamma/2} + \hat{\alpha}) \\ &= \mathbb{P}(s_\alpha t_{n-2,\gamma/2} + \hat{\alpha} \geq \alpha \geq -s_\alpha t_{n-2,\gamma/2} + \hat{\alpha}) \end{aligned}$$

So $CL_\alpha = \hat{\alpha} - s_\alpha t_{n-2,\gamma/2}$ & $CU_\alpha = \hat{\alpha} + s_\alpha t_{n-2,\gamma/2}$.
For β we have $CL_\beta = \hat{\beta} - s_\beta t_{n-2,\gamma/2}$ & $CU_\beta = \hat{\beta} + s_\beta t_{n-2,\gamma/2}$.

10.5 Hypothesis Tests for β

Remark 10.2 - *Assumptions*
Here we assume that e_i are IID with $e_i \sim N(0, \sigma^2)$.

Example 10.1 - *Hypothesis Tests for β*
Model Assumptions - We make the model assumptions of **Definition 10.1**.
Hypothesis - We test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$.
Test Statistic - We define test statistic $T = \hat{\beta}/s_{\hat{\beta}}$.
When H_0 is true then $T \sim t_{n-2}$.
p-value - Since we are considering a two-sided alternative with $t_{obs} = \hat{\beta}/s_{\hat{\beta}}$.

$$p - value = \mathbb{P}(|T| \geq |t_{obs}| \mid H_0 \; true) = \mathbb{P}(|t_{n-2}| \geq |t_{obs}|) = 2(1 - \mathbb{P}(t_{n-2} \leq |t_{obs}|))$$

0 Reference

0.1 Definitions

Definition 0.1 - *Alternative Hypothesis*
The *Alternative Hypothesis* is the hypothesis used in hypothesis testing that is contrary to the *Null Hypothesis*.

Definition 0.2 - *Critical Region*
The *Critical Region* is the set of values which lead you to reject the *Null Hypothesis* at a given *Significance Level*.

Definition 0.3 - *Cumulative Frequency Function*
The *Cumulative Frequency Function* for a random variable X returns the probability of a value being less than the given value

$$F_X(x) = \mathbb{P}(x \leq X)$$

N.B. This is also know as the *distribution function*.
Definition 0.4 - *Null Hypothesis*
The *Null Hypothesis* is a default position that there is no relationship between two phenomena.

Definition 0.5 - *Power*
The *Power* of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true.

Definition 0.6 - *p-Value*
The *p-value* is the probability for a given statistical model given the sampled value, assuming the *Null Hypothesis* is true.

Definition 0.7 - *Significance Level*
The *Significance Level* of a hypothesis test is the probability of rejecting a null hypothesis when it is true
Definition 0.8 - *Type 1 Error*
A *Type 1 Error* occurs when the null hypothesis is true, but is rejected.

Definition 0.9 - *Type 2 Error*
A *Type 2 Error* occurs when the null hypothesis is false, but is accepted.

0.2 Notation

Notation 0.1 - *Data Set*
A data set with n elements is denoted by

$$\{x_1, \dots, x_n\}$$

Typically x_i happens before x_{i+1} $\forall$ $i < n$.

Notation 0.2 - *Hypothesis Tests*
Hypothesis Tests are denoted by H_i where $i \in \mathbb{N}$.
The *Null Hypothesis* is denoted by H_0.

Notation 0.3 - *Order Statistic*
An order statistics with N elements is denoted by

$$\{x_{(1)}, \dots, x_{(n)}\}$$

$x_{(i)} \leq x_{(i+1)} \ \forall \ i < n$.

Notation 0.4 - *Probability Density Function*
The *Probability Density Function* of a continuous random variable X is denoted by $f_X(x)$.

Notation 0.5 - *Probability Density Function*
For a continuous random variable X is denoted by $f_X(x)$.

Notation 0.6 - *Probability Mass Function*
The *Probability Mass Function* of a discrete random variable X is denoted by $p_X(x)$.

Notation 0.7 - *Parametric Family, Estimation*
For a parametric family that depends upon the variable θ an estimation for the value of θ is denoted by $\hat{\theta}(x_1, \dots, x_n)$.
N.B. This is often abbreviated to $\hat{\theta}$.
Notation 0.8 - *True Value*
An asterisk is used to denote the true value of a parameter, such as

$$\theta^*$$

N.B. This is a constant.

0.3 Identities

Theorem 0.1 - *Normal Distribution*

$$\mathbb{E}(X^2) = Var(X) + \mathbb{E}(X)^2$$

0.4 R

Definition 0.10 - *Combine*
The *Combine* command takes in several variables and combine into a single vector.
This command is denoted as $c(x_1, \dots, x_n)$.

Definition 0.11 - *Variable Assignment*
To assign a value to a variable we use the command $l \leftarrow r$.
This assigns the value r to variable name l.