MATH2801/2901 Revision Sheet

UNSW Mathematics Society: Anne Chen, Bruce Chen

We would like to preface this document by saying that this resource is first and foremost meant to be used as a reference and should NOT be used as a replacement for the course resources or lecture recordings. Said resources provided on Moodle are wonderfully written and contain an abundance of fully worked solutions and in depth explanations. Studying for this course using *only* this revision sheet would not be sufficient.

In addition, although the authors have tried their best to include everything essential taught in the course, it was ultimately up to their discretion on whether or not to include results/theorems/definitions etc. Anything that is missing is most definitely a conscious choice made by the authors.

Finally, any and all errors found within this document are most certainly our own. If you have found an error, please contact us via our Facebook page, or give us an email.

Contents

I Probability and Distribution Theory 4

Probability Revision 4
Set Operations 4
Conditional Probability 4
Independence 4
Independence For Multiple Events 4
Probability Laws 4
Bayes' Theorem 4

Random Variables 5
Cumulative Distribution Functions 5
Probability Mass Functions 5
Probability Density Functions 5
Expectation and Moments 5
Standard Deviation and Variance 5
Moment Generating Functions 6
Useful Inequalities 6

Common Distributions 6
Discrete Distributions 6
Continuous Distributions 7
Quantiles and QQ-Plots 7
Distributions in R 7

Bivariate Distributions 8
Joint Density Functions 8
Joint CDFs 8
Marginal Probability/Density 8
Expectation Under Bivariate Distributions 8
Conditional Probability/Density 8
Conditional Expectation 8
Independence 8
Covariance and Correlation 8
Bivariate Gaussian 9

Transformations 10
Monotonic Transformations 10
Probability Integral Transformation 10
Bivariate Transformations 10
Sums of Independent Random Variables 10

Convergence of Random Variables 11
Some Definitions 11
Central Limit Theorem 11
Law of Large Numbers 11
Applications of the CLT 11

Distributions Arising from a Normal Sample 12
χ^2-distribution 12
t-distribution 12
Key Results 12
Quantiles and CDF Values for the t-distribution and χ^2-distribution 12

II Basic Statistical Inference 13

Parameter Estimation 13

Properties of Estimators 13

1. Bias 13
2. Standard Error 13
3. Mean Squared Error 13
4. Consistency 13
5. Asymptotic Normality 13

Methods of Parameter Estimation 13

Method of Moments Estimation 13

Maximum Likelihood Estimation 13

Properties of the MLE 13

Fisher Score and Information 13

Confidence Intervals 15

Confidence Interval for a Normal Random Sample 15

Wald Confidence Intervals 15

Hypothesis Testing 15

Null and Alternative Hypothesis 15

Errors 15

Illustrative Example: Hypothesis Test for the Normal Distribution 15

Wald Tests 16

Likelihood Ratio Tests 16

Power of a Test 16

Part I

Probability and Distribution Theory

Probability Revision

Set Operations

Associative Law

If A, B, and C are sets, then

$$(A \cup B) \cup C = A \cup (B \cup C),$$

and

$$(A \cap B) \cap C = A \cap (B \cap C).$$

Distributive Law

If A, B, and C are sets, then

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C),$$

and

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C).$$

Conditional Probability

The *conditional probability* that an event A occurs, given that an event B has occurred is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ if } \mathbb{P}(B) \neq 0.$$

Independence

Two events A and B are *independent* if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B),$$

of if equivalently,

$$\mathbb{P}(A|B) = \mathbb{P}(A) \text{ and } \mathbb{P}(B|A) = \mathbb{P}(B).$$

Independence For Multiple Events

A set of events $\{A_i\}_{i=1}^n$ is *pairwise independent* if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \text{ for all } i \neq j.$$

A sets of events $\{A_i\}_{i=1}^n$ is *mutually independent* if for any subset $\{A_{i_1}, A_{i_2}, \dots, A_{i_m}\}$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}) = \prod_{j=1}^{m} \mathbb{P}(A_{i_j}).$$

Probability Laws

1. Multiplicative Law:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A) = \mathbb{P}(B|A)\mathbb{P}(A)$$

2. Additive Law:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

3. Law of Total Probability:

   Suppose that $\{A_i\}_{i=1}^k$ forms a partition of the sample space Ω. Then for any event B

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

Bayes' Theorem

Where A can be partitioned into $\{A_i\}_{i=1}^k$, the conditional probability of A given B is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

Random Variables

Cumulative Distribution Functions

The *cumulative distribution function* (CDF) of a random variable X is defined by

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

Using the CDF

1. $\mathbb{P}(X > x) = 1 - F_X(x)$.
2. For any $x < y$, $\mathbb{P}(x < X \le y) = F(y) - F(x)$.

Properties of the CDF

1. F is bounded between 0 and 1, such that

$$\lim_{x \downarrow -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \uparrow \infty} F(x) = 1.$$

2. F is non-decreasing: if $x < y$, then $F(x) \le F(y)$
3. F is right continuous: $\lim_{t \to x^+} F(t) = F(x)$ for all x.

Probability Mass Functions

The *probability mass function* of a discrete random variable X is the function f_X given by

$$f_X(x) = \mathbb{P}(X = x).$$

It is related to the CDF by the following:

$$F_X(x) = \sum_{y \le x} f_X(y).$$

Properties

1. $f_X(x) \ge 0$ for all $x \in \mathbb{R}$
2. $\sum_{\text{all } x} f_X(x) = 1$

Probability Density Functions

The *probability density function* (PDF) of a continuous random variable X is the function f_X given by

$$f_X(x) = \frac{\partial}{\partial x} F_X(x).$$

Naturally, we can integrate the PDF to find the CDF:

$$F_X(x) = \int_{-\infty}^{x} f_X(t) dt.$$

In addition, for any pair of numbers $a \le b$,

$$\mathbb{P}(a \le X \le b) = \int_a^b f_X(x) dx.$$

Expectation and Moments

The *expected value* of a random variable X is given by

$$\mathbb{E}[X] = \begin{cases} \sum_{\text{all}\, x} x f_X(x), & \text{for discrete } X \\ \int_{-\infty}^{\infty} x f_X(x)\ dx, & \text{for continuous } X. \end{cases}$$

Similarily, the expected value of a function $g(x)$ of a random variable X is

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{\text{all}\, x} g(x) f_X(x), & \text{for discrete } X \\ \int_{-\infty}^{\infty} g(x) f_X(x)\ dx, & \text{for continuous } X \end{cases}$$

The non-central moments of a random variable are

$$\mathbb{E}[X^r], r = 1, 2, \ldots,$$

Properties of the expectation

Let $a, b \in \mathbb{R}$ be constants and X, Y be random variables.

1. Expectation of a constant is constant: $\mathbb{E}(a) = a$.
2. Linearity: $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$
3. If X and Y are independent,

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

Standard Deviation and Variance

Let $\mu = \mathbb{E}[X]$, then the *variance* and *standard deviation* of a random variable X are given by

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

and

$$\text{standard deviation of } X = \sqrt{\text{Var}(X)}.$$

Properties of the variance

1. $\text{Var}(a) = 0$
2. $\text{Var}(aX) = a^2\text{Var}(X)$
3. $\text{Var}(X + b) = \text{Var}(X)$
4. If X and Y are independent,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Moment Generating Functions

The *moment generating function* (MGF) of a random variable X is

$$m_X(u) = \mathbb{E}(e^{uX}).$$

The moment generating function of X exists if there exists a $h > 0$ such that $m_X(u)$ is finite for $u \in [-h, h]$.

Suppose the MGF of X exists, then we can obtain the nth non-central moment of X through differentiation:

$$\mathbb{E}(X^n) = \lim_{u \to 0} \left(\frac{d}{du} m_X^{(n)}(u) \right).$$

(Very useful) properties of MGFs

1. The MGF of a random variable is unique:

$$M_X(u) = M_Y(u) \Rightarrow F_X(x) = F_Y(y)$$

2. Convergence/equality of MGFs implies convergence/equality of CDFs

3. If X and Y are independent, we also have

$$M_{X+Y}(u) = M_X(u)M_Y(u)$$

Useful Inequalities

Markov's Inequality

If X is a non-negative random variable and $a > 0$, then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Chebychev's Inequality

Let X be any random variable with mean μ and variance σ^2. Then for any $k > 0$,

$$\mathbb{P}(|X - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

These two inequalities allow us to find upper and lower bounds for probabilities for variables of any distribution.

Jensen's Inequality

A convex function h is one such that for any $\lambda \in [0, 1]$,

$$h(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda h(x_1) + (1 - \lambda)h(x_2).$$

Jensen's inequality states that if X is a random variable and h is a convex function, then

$$h(\mathbb{E}(X)) \leq \mathbb{E}(h(X)).$$

Common Distributions

Discrete Distributions

Bernoulli distribution

A Bernoulli trial is an experiment which can either succeed (probability p) or fail (probability $(1 - p)$). $X \sim \text{Bernoulli}(p)$ if

$$X = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p. \end{cases}$$

- $\mathbb{E}(X) = np$
- $\mathbb{V}(X) = np(1 - p)$

Binomial distribution

If we have a sequence of n independent Bernoulli trials each with probability of success p, then the total number of successes X is a Binomial random variable and

$$X \sim \text{Bin}(n, p).$$

- $\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0, 1, \ldots, n,$
- $\mathbb{E}(X) = np$
- $\mathbb{V}(X) = np(1 - p)$

Poisson Distribution

For a random variable $X \sim \text{Poisson}(\lambda)$:

- $\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, \ldots$
- $\mathbb{E}(X) = \lambda$
- $\mathbb{V}(X) = \lambda$

Hypergeometric Distribution

For a random variable with hypergeometric distribution with parameters N, m, n, that is, $X \sim \text{Hyp}(n, m, N)$,

- $\mathbb{P}(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, x = 1, \ldots, n$
- $\mathbb{E}(X) = \frac{mn}{N}$

Given a collection of N objects, if m fall into one category and $N - m$ fall into the other, if n are chosen at random, then the number of objects X belonging to the first category satisfies

$$X \sim \text{Hyp}(n, m, N).$$

Continuous Distributions

Gaussian/Normal distribution

A normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ satisfies

- $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$
- $\mathbb{E}(X) = \mu$
- $\mathbb{V}(X) = \sigma^2$

Exponential distribution

An exponential random variable $X \sim \mathrm{Exp}(\lambda)$ satisfies

- $f_x(x) = \frac{1}{\lambda} e^{-\frac{1}{\lambda}x}, x > 0$
- $\mathbb{E}(X) = \lambda$
- $\mathbb{V}(X) = \lambda^2$

Gamma distribution

A Gamma random variable $X \sim \mathrm{Gamma}(\alpha, \beta)$ satisfies

- $f_X(x; \alpha, \beta) = \frac{e^{-\frac{x}{\beta}} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}, x > 0$
- $\mathbb{E}(X) = \alpha\beta$
- $\mathbb{B}(X) = \alpha\beta^2$

Note that the Gamma function $\Gamma(\alpha)$ is given by

$$\Gamma(\alpha) = \int_0^\infty t^{x-1} e^{-t} dt.$$

Beta distribution

A Beta random variable $X \sim \mathrm{Beta}(\alpha, \beta)$ satisfies

- $f_X(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 < x < 1$

Note that the Beta function $B(a, b)$ is given by

$$B(a, b) = \int_0^1 t^{x-1}(1-t)^{y-1} dt.$$

Quantiles and QQ-Plots

Quantiles

Given a continuous random variable X with CDF F_X, the $100k\%$-th *quantile* of X is given by

$$Q_X(k) = F_X^{-1}(k), 0 < k < 1,$$

that is, the level x such that $100k\%$ of random variables following the distribution are less than or equal to x.

Sample quantiles

Given an sample of observations $\mathbf{x} = (x_1, \ldots, x_n)$ we can order the samples in ascending order to obtain

$$\mathbf{x}' = (x_{(1)}, \ldots x_{(n)}).$$

These are the *sample quantiles* of x.

Definition: QQ-Plot

In a QQ plot, we are given a sample of observations and a distribution. We plot the sample quantiles on the y-axis and their corresponding theoretical quantiles on the x-axis. That is, we plot the points

$$(F_X^{-1}(p), x_{(k)}) = \left(F_X^{-1}\left(\frac{k-0.5}{n}\right), x_{(k)}\right)$$

for $k = 1, \ldots, n$, and $p = \frac{k-0.5}{n}$ is a rough measure of where the sample $x_{(k)}$ lies in the data set.

The point is that if the observations follow the distribution, the dots should roughly form a straight line. If not, then the observations most likely don't follow the given distribution.

Distributions in `R`

Each distribution has a family of four commands:

- `d___(x, ...)` gives either the probability mass function or probabiliy density function,
- `p___(q, ...)` gives the cumulative distribution function (i.e. $\mathbb{P}(X \leq q)$),
- `q___(p, ...)` gives the quantile function at p,
- `r___(n, ...)` randomly generates n values according to the distribution.

You will need to be familiar with `qnorm(p,0,1)` in particular.

Bivariate Distributions

Joint Density Functions

The *joint density function* of two continuous random variables X and Y is a bivariate function $f_{X,Y}$ with the following properties:

1. $f_{X,Y}(x,y) \geq 0$, for all $(x,y) \in \mathbb{R}^2$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y) = 1$
3. $\mathbb{P}(X \in A, Y \in B) = \int_{y\in B}\int_{x\in A} f_{X,Y}(x,y)\ dx\ dy$

Similar properties hold for the discrete case.

Joint CDFs

The joint CDF of X and Y is given by

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbb{P}(X \leq x, Y \leq y) \\ &= \begin{cases} \displaystyle\sum_{u\leq x}\sum_{v\leq y} \mathbb{P}(X=u, Y=v), \\ \text{for discrete} \\ \displaystyle\int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\ du\ dv, \\ \text{for continuous.} \end{cases} \end{aligned}$$

Marginal Probability/Density

Given $f_{X,Y}(x,y)$, we can calculate the marginal probability/density function $f_X(x)$ as follows

$$f_X(x) = \begin{cases} \displaystyle\sum_{\text{all } y} f_{X,Y}(x,y), & \text{for discrete} \\ \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy, & \text{for continuous} \end{cases}$$

and similarly for $f_Y(y)$,

$$f_Y(y) = \begin{cases} \displaystyle\sum_{\text{all } x} f_{X,Y}(x,y), & \text{for discrete} \\ \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx, & \text{for continuous.} \end{cases}$$

Expectation Under Bivariate Distributions

For a function $g : \mathbb{R}^2 \to \mathbb{R}$, we have

$$\mathbb{E}[g(X,Y)] = \begin{cases} \displaystyle\sum_{\text{all } x}\sum_{\text{all } y} g(x,y) f_{X,Y}(x,y)\ dx, \\ \text{for discrete} \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\ dy\ dx, \\ \text{for continuous.} \end{cases}$$

Conditional Probability/Density

The conditional probability/density function of X given $Y = y$ is

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

Conditional Expectation

The conditional expectation of $g(X)$ given $Y = y$ is

$$\mathbb{E}(g(X)|Y=y) = \begin{cases} \displaystyle\sum_{\text{all } x} g(x)\mathbb{P}(X=x|Y=y), \\ \text{for discrete} \\ \displaystyle\int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\ dx, \\ \text{for continuous.} \end{cases}$$

Independence

Two random variables are independent if and only if for all x, y

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

or

$$f_{X|Y}(x|y) = f_X(x)$$

or

$$f_{Y|X}(y|x) = f_Y(y).$$

This means that if you can separate the joint density function of two variables X and Y into a product of functions of x and y, then they are independent.

Covariance and Correlation

Covariance

The *covariance* of X and Y is a measure of their joint variability and is given by

$$\begin{aligned} \text{Cov}(X,Y) &= \mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big] \\ &= \mathbb{E}(XY) - \mu_X\mu_Y \end{aligned}$$

where $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$.

Properties

1. If X and Y are independent, then $\text{Cov}(\text{X},\text{Y}) = 0$. **Important:** The converse is not true.
2. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
3. $\text{Cov}(aX + bY, Z) = a\,\text{Cov}(X,Z) + b\,\text{Cov}(Y,Z)$

Relation to variance

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y)$$

Correlation

The *correlation* between X and Y measures the strength of their linear relationship between and is given by

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}.$$

This value will always be between -1 and 1. A value of 1 indicates a perfect positive linear relationship while a value of -1 indicates a perfect negative relationship. X and Y are uncorrelated if $\mathrm{Corr}(X,Y) = 0$.

Bivariate Gaussian

A random vector $\mathbf{X} = (X_1, X_2)$ is Gaussian with $\mu_{\mathbf{X}} = (\mu_{X_1}, \mu_{X_2})$ and covariance matrix V if

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi\sqrt{V}} \exp(-\frac{1}{2}(\mathbf{X} - \mu_{\mathbf{X}})^T V^{-1}(X - \mu_{\mathbf{X}})).$$

The *covariance matrix* is defined with entries

$$V_{i,j} = \mathrm{Cov}(X_i, X_j), i = 1, 2, \ldots, d, j = 1, 2, \ldots, d.$$

Transformations

Monotonic Transformations

For continuous X, if $h(x)$ is monotonic (strictly increasing or decreasing) over the set $\{x : f(x) > 0\}$, then for $Y = h(X)$,

$$\begin{aligned} f_Y(y) &= f_X(x) \left| \frac{dx}{dy} \right| \\ &= f_X\{h^{-1}(y)\} \left| \frac{dx}{dy} \right| \end{aligned}$$

for y such that $f_X\{h^{-1}(y)\} > 0$.

Linear transformations

In the case that $Y = aX + b$ is a linear transformation of X, we have

$$f_Y(y) = \frac{1}{|a|} f_X \left(\frac{y-b}{a} \right)$$

for all y such that $f_X\left(\frac{y-b}{a}\right) > 0$.

Probability Integral Transformation

For any random variable X whose CDF is strictly increasing,

$$Y = F_X(X) \sim \text{Uniform}(0, 1).$$

Bivariate Transformations

If U and V are functions of continuous random variables X and Y, then

$$f_{U,V}(u, v) = f_{X,Y}(x, y) \cdot |J|$$

where J, the Jacobian, is the matrix given by

$$J = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$

Example: This method can be quite abstract, so we've provided an example below to demonstrate how the Jacobian transformation works.

Suppose that we have X and Y such that

$$f_{X,Y}(x, y) = \exp(-x - y), x, y > 0,$$

and that we want to find the joint density of

$$U = X + Y, V = \frac{X}{X + Y}.$$

Then, to find $\frac{\partial x}{\partial u}, \frac{\partial x}{\partial v}, \frac{\partial y}{\partial u}$ and $\frac{\partial y}{\partial v}$, we need to find X and Y in terms of U and V. In this case,

$$X = UV, \text{ and } Y = U - UV.$$

We also need to find the range of U and V. Since $X, Y > 0$ is the only restriction, $U > 0$ and $0 < Y < 1$.

So we have that

$$J = \begin{bmatrix} v & u \\ 1 - v & -u \end{bmatrix},$$

and hence

$$|J| = | - uv - u(1 - v)| = u.$$

So the joint density is

$$\begin{aligned} f_{U,V}(u, v) &= \exp(-x - y)u \\ &= u \exp(-u), u > 0, 0 < v < 1. \end{aligned}$$

Alternate approach

We can also find $f_U(u)$ by first calculating the CDF $F_U(u) = \mathbb{P}(U \leq u)$ and then differentiating.

Sums of Independent Random Variables

Convolution formula

Suppose X and Y are independent continuous variables with density functions $f_X(x)$ and $f_Y(y)$. Then $Z = X + Y$ has the density

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx.$$

MGF approach

Recall that if X and Y independent random variables with moment generating functions m_X and m_Y. Then

$$m_{X+Y}(u) = m_X(u) m_Y(u).$$

This is often useful for deriving the distributions for sums of independent random variables due to the uniqueness of moment generating functions.

Useful properties of common distributions

Many common distributions are additive in a way. Suppose $(X_i)_{i=1,2,\ldots,n}$ is an independent sequence of random variables. Take $Y = \sum_{i=1}^n X_i$ their sum.

- If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then $Y \sim \mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.
- If $X_i \sim \text{Exp}(\lambda)$, then $Y \sim \text{Gamma}(n, \lambda)$.
- If $X_i \sim \text{Gamma}(\alpha_i, \beta)$, then $Y \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.
- If $X_i \sim \text{Poisson}(\lambda_i)$, then $Y \sim \text{Poisson}(\sum_{i=1}^n \lambda_i)$.
- If $X_i \sim \text{Bernoulli}(p)$, then $Y \sim \text{Binomial}(n, p)$.
- If $X_i \sim \text{Binomial}(n_i, p)$, then $Y \sim \text{Binomial}(\sum_{i=1}^n n_i, p)$.

Convergence of Random Variables

Some Definitions

Convergence in distribution

We say a sequence of random variables $X_1, X_2, \dots$ *convergences in distribution* to a random variable X if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x),$$

for every x. This is often denoted as $X_n \xrightarrow{d} X$.

Convergence in probability

A sequence of random variables $X_1, X_2, \dots$ *convergences in probability* to a random variable X if, for all $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

This is often denoted as $X_n \xrightarrow{\mathbb{P}} X$.

Almost sure convergence

A sequence of random variables $X_1, X_2, \dots$ *convergences almost surely* to a random variable X if:

$$\mathbb{P}\left(\lim_{n \to \infty} X_n = X\right) = 1.$$

This is often denoted as $X_n \xrightarrow{\text{a.s.}} X$.

Convergence implications

The types of convergence are related as follows:

$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X.$$

Central Limit Theorem

Suppose that $X_1, \dots, X_n$ are i.i.d. random variables with a common mean $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = \text{Var}(X_i) < \infty$. For each $n \geq 1$, let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1).$$

Law of Large Numbers

Weak Law of Large Numbers

Let $X_1, X_2, \dots, X_n$ be a sequence of independent random variables each with mean μ and finite variance σ^2. Then the sample mean will converge in probability to the true mean:

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mu.$$

This implies that as the sample size increases, the sample mean will more likely be closer to the true mean.

Strong Law of Large Numbers

The strong law of large numbers is the same but stricter, as the convergence happens almost surely. i.e.

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu.$$

Applications of the CLT

General applications

We can often use the CLT to estimate probabilities associated with the sample mean of a distribution.

- Given i.i.d. random variables $X_1, \dots, X_n$ X with mean μ and variance σ^2, write $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ for $Z \sim \mathcal{N}(0, 1)$.
- Then write

$$\begin{aligned} \mathbb{P}(\bar{X} < r) &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{r - \mu}{\sigma/\sqrt{n}}\right) \\ &= \mathbb{P}\left(Z < \frac{r - \mu}{\sigma/\sqrt{n}}\right). \end{aligned}$$

- You can then use the normal probability tables or `R` to estimate $\mathbb{P}(Z < r)$ for any r.

Normal approximation to the Binomial distribution

Suppose $X \sim \text{Bin}(n, p)$, then

$$\frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{\text{d}} \mathcal{N}(0, 1).$$

Delta Method

The Delta Method gives us a way to find the distribution of any function of an MLE.
If

$$\frac{X_n - \theta}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1),$$

and g is a differentiable function in a neighbourhood of θ, and $g'(\theta) \neq 0$, then

$$\frac{g(X_n) - g(\theta)}{\sigma g'(\theta)/\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1),$$

Extended Delta Method

The Delta Method doesn't work when $g'(\theta) = 0$. In that case we find the first k such that $g^{(k)}(\theta) \neq 0$. Then

$$\frac{g(X_n) - g(\theta)}{(\sigma/\sqrt{n})^k} \xrightarrow{\text{d}} \frac{1}{k!} g^{(k)}(\theta) Z^k$$

for $Z \sim \mathcal{N}(0, 1)$.

Distributions Arising from a Normal Sample

χ^2-distribution

For independent random variables $(X_i)_{i=1,2,\ldots,n} \sim \mathcal{N}(0,1)$,

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n),$$

referred to as a χ^2 distribution with n degrees of freedom.

t-distribution

For independent $Y, Z \sim \mathcal{N}(0,1)$ and $Z \sim \chi^2_\nu$,

$$\frac{Y}{\sqrt{Z/\nu}} \sim t_{\nu-1},$$

referred to as a t-distribution with $\nu - 1$ degrees of freedom.

Key Results

Let $X_1, \ldots, X_n$ be a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Then, where

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

is the *sample mean*, and

$$S_X = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

is the *sample standard deviation*, we have that

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

and

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Quantiles and CDF Values for the t-distribution and χ^2-distribution

Quantiles and CDF values for the t-distribution and χ^2-distribution can be found using the quantile and CDF tables for each of the distributions, which are generally formatted the same as normal probability and quantile tables.

To use the t-distribution tables, you can use the fact that the t-distribution is symmetric. Note that the χ^2-distribution is not symmetric.

Part II

Basic Statistical Inference

Parameter Estimation

Let $X_1, \dots, X_n \sim f_X(x;\theta)$, then an *estimator* $\hat{\theta}$ for θ is a real valued function of $X_1, \dots, X_n$. Since the estimator is a function of random variables, it is a random variable itself with density function $f_{\hat{\theta}}$.

Properties of Estimators

1. Bias

The bias of $\hat{\theta}$ is given by

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

If $\mathbb{E}(\hat{\theta}) = \theta$ then $\hat{\theta}$ is said to be unbiased.

2. Standard Error

The standard error of $\hat{\theta}$ is the standard deviation

$$\text{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

3. Mean Squared Error

The mean squared error allows us to combine the bias and standard error into a single measure that gives us an indication of the quality of an estimator:

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}\{(\hat{\theta} - \theta)^2\} \\ &= \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})\end{aligned}$$

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ, then $\hat{\theta}_1$ is better than $\hat{\theta}_2$ (with respect to the MSE) if

$$\text{MSE}(\hat{\theta}_1) < \text{MSE}(\hat{\theta}_2).$$

4. Consistency

An estimator $\hat{\theta}$ is consistent if $\hat{\theta}_n$ converges to θ as the sample size n increases, i.e.

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta.$$

An easy way to check that an unbiased estimator is consistent is to show that its variance decreases to 0 as $n \to \infty$.

5. Asymptotic Normality

The estimator $\hat{\theta}$ is asymptotically normal if

$$\frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Methods of Parameter Estimation

Method of Moments Estimation

Let $x_1 \dots, x_n$ be observations from the model $f(x; \theta_1, \dots, \theta_k)$ containing k parameters.

1. Form a system of k equations that equates the moments of f_X
2. Solve simultaneously to obtain the estimators

Maximum Likelihood Estimation

Let $x_1 \dots, x_n$ be observations from PDF $f(x) = f(x;\theta)$ depending on a parameter θ.

The likelihood function $\mathcal{L}$ is a function of θ given by

$$\mathcal{L}(\theta) = f(x_1;\theta) \cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

and the log-likelihood function of θ is

$$\ell(\theta) = \ln\{\mathcal{L}(\theta)\}.$$

The *maximum likelihood estimate* of θ is the choice

$$\hat{\theta} = \theta \text{ that maximises } \ell(\theta).$$

It can usually be determined by setting the derivative of the log-likelihood function to zero and solving as in the case of a univariate optimisation problem.

Properties of the MLE

1. **Consistency**: $\hat{\theta} \xrightarrow{\mathbb{P}} \theta$.
2. **Invariance**: If g is a continuous and injective function, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
3. **Asymptotic normality**:

$$\frac{\hat{\theta}_n - \theta}{\text{se}(\hat{\theta}_n)} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Fisher Score and Information

The Fisher score is defined as

$$S_n(\theta) = \ell_n'(\theta)$$

and the Fisher information is defined as

$$I_n(\theta) = -\mathbb{E}_\theta \ell_n''(\theta).$$

1. $\mathbb{E}_\theta S_n(\theta) = 0$
2. $\text{Var}_\theta S_n(\theta) = I_n(\theta)$

Variance of the MLE

For $(X_1, X_2, ..., X_n)$ a random sample and $\hat{\theta_n}$ the MLE of θ,

$$I_n(\theta)\mathrm{Var}_\theta(\hat{\theta_n}) \xrightarrow{\mathbb{P}} 1.$$

So for large enough n,

$$\sqrt{\mathrm{Var}_\theta(\hat{\theta})} = \mathrm{se}(\hat{\theta}) \approx (I_n(\theta))^{-\frac{1}{2}}.$$

Cramer Rao Lower Bound

If $\tilde{\theta}_n = g(X_1, \dots, X_n)$ is an unbiased estimator of θ, then

$$\mathrm{Var}_\theta(\tilde{\theta}) \geq \frac{1}{I_n(\theta)}.$$

Thus,

$$\frac{1}{I_n(\theta)}$$

is called the Cramer-Rao lower bound on the variance of an unbiased estimator.

Since the inverse of the Fisher information is the asymptotic variance of the MLE, a consequence of this is that the MLE is asymptotically optimal.

Delta Method for MLEs

The Delta Method gives us a way to find the distribution of any function of an MLE.

If

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbb{V}(\hat{\theta})}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1),$$

and g is a differentiable function in a neighbourhood of θ, and $g'(\theta) \neq 0$, then

$$\frac{g(\hat{\theta_n}) - g(\theta)}{\sqrt{\mathbb{V}(g(\hat{\theta}_n)}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1),$$

where in the case of an MLE

$$\mathbb{V}(g(\hat{\theta})) \approx (g'(\theta))^2 I_n^{-1}(\theta).$$

Example: For any MLE,

$$\frac{\hat{\theta_n}^2 - \theta^2}{\sqrt{\mathbb{V}(g(\hat{\theta}_n))}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1),$$

where $\mathbb{V}(\hat{\theta_n}) \approx 2\theta I_n^{-1}(\theta) \approx 2\hat{\theta_n} I_n^{-1}(\hat{\theta_n})$.

Multivariate MLEs

For a model with multiple parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, the Fisher information matrix is given by

$$I_n(\boldsymbol{\theta}) = -\mathbb{E}(H) = -\begin{pmatrix} \mathbb{E}(H_{11}) & \dots & \mathbb{E}(H_{1n}) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(H_{n1}) & \dots & \mathbb{E}(H_{nn}) \end{pmatrix},$$

where H is the Hessian matrix (for those who have studied MATH2011/2111) and

$$H_{ij} = \frac{\partial^2}{\partial i \partial j} l(\boldsymbol{\theta}; x_1, \dots, x_n).$$

Multivariate Delta Method

Let g be a function of the parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. The MLE of $g(\boldsymbol{\theta})$ is $g(\hat{\boldsymbol{\theta}})$. Then

$$\frac{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})}{\widehat{se}(g(\hat{\boldsymbol{\theta}}))} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1),$$

where

$$\widehat{se}(g(\hat{\boldsymbol{\theta}})) \approx \sqrt{\nabla(\hat{\boldsymbol{\theta}}) I_n^{-1}(\hat{\boldsymbol{\theta}}) \nabla g(\hat{\boldsymbol{\theta}})}.$$

Note that

$$\nabla(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} g(\hat{\boldsymbol{\theta}}) \\ \vdots \\ \frac{\partial}{\partial \theta_k} g(\hat{\boldsymbol{\theta}}) \end{pmatrix}.$$

Confidence Intervals

A $100(1-\alpha)\%$ confidence interval for some unknown parameter θ is an interval $[L, U]$ such that

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha.$$

We can construct a confidence interval whenever we have an estimator of θ whose distribution is known.

Confidence Interval for a Normal Random Sample

Let $X_1, \ldots, X_n$ be a random sample from the $\mathcal{N}(\mu, \sigma^2)$ distribution. Then a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Wald Confidence Intervals

The Wald confidence interval gives a confidence interval for a parameter θ using its MLE $\hat{\theta}$. The $100(1-\alpha)\%$ confidence interval is

$$\left(\hat{\theta}_n - z_{1-\alpha/2}\,\mathrm{se}(\hat{\theta}), \hat{\theta}_n + z_{1-\alpha/2}\,\mathrm{se}(\hat{\theta})\right),$$

or, rewritten,

$$\left(\hat{\theta}_n - z_{1-\alpha/2}\,(I_n(\theta))^{-\frac{1}{2}}, \hat{\theta}_n + z_{1-\alpha/2}\,(I_n(\theta))^{-\frac{1}{2}}\right).$$

Hypothesis Testing

Null and Alternative Hypothesis

A hypothesis test is a way of testing a hypothesis about the value of a parameter θ.

In a hypothesis test we formulate two hypotheses, the *null hypothesis* H_0 and the *alternative hypothesis*, which we test against each other.

If the entire parameter space is Θ (e.g. $\mathbb{R}^+$), then we have

$$H_0 : \theta \in \Theta_0$$

and

$$H_1 : \theta \in \Theta_1$$

for subsets $\Theta_0, \Theta_1 \subseteq \Theta$.

Errors

- Type I error corresponds to rejection of the null hypothesis when it is really true.
- Type II error corresponds to acceptance of the null hypothesis when it is really false.

Significance level

We want to control the probability of a Type I error to a level of precision we require.

The probability of making a Type I error is called the *level of the test*, or the α-level.

Test statistic

A *test statistic* T is a function of the observations in a hypothesis test, often an estimator of the parameter.

Using the test statistic, we conduct our hypothesis using one of the following methods:

- **Rejection region**: We observe the value of the *test statistic* and if it lies in the *rejection region*, we reject the null hypothesis. The rejection region R needs to satisfy

$$\mathbb{P}(T \in R) = \alpha.$$

- **p-value**: We observe the p-value of the test. The p-value is the probability of obtaining a value of T more extreme than the value observed.

Illustrative Example: Hypothesis Test for the Normal Distribution

Suppose $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where μ and σ are both unknown. Our test statistic is

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

- **One-sided test**: Two options –
  - $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$
    - In this case, since our alternative hypothesis is that μ is large, the test statistic being *more extreme* than the observed value means greater than the observed value.
    - The rejection region is
$$R = \left\{ \mathbf{x} : \frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{1-\alpha,n-1} \right\},$$
where $t_{1-\alpha,n-1}$ is the $100(1-\alpha)\%$ percentile of the t_{n-1} distribution.
    - The p-value for a observed value of the test statistic $T = t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, where $\bar{x}$ and s are the observed mean and standard deviation respectively, is
$$p = \mathbb{P}\left(T > \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right) \text{ for } T \sim t_{n-1},$$
which can be determined from the t-distribution tables.
  - $H_0 : \mu = \mu_0$ and $H_1 : \mu < \mu_0$
    - The rejection region is
$$R = \left\{ \mathbf{x} : \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < t_{\alpha,n-1} \right\}.$$
    - The p-value is
$$p = \mathbb{P}\left(T < \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right) \text{ for } T \sim t_{n-1}.$$
- **Two-sided test**:
  - $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$
    - In a two-sided test, the test statistic being *more extreme* than the observed value means greater in magnitude than the observed value.
    - The rejection region is
$$R = \left\{ \mathbf{x} : \frac{|\bar{X} - \mu_0}{S/\sqrt{n}}| > t_{1-\alpha/2,n-1} \right\}.$$
    - The p-value is
$$p = \mathbb{P}\left(|T| > |\frac{\bar{x} - \mu_0}{s/\sqrt{n}}| \right) = 2\mathbb{P}(T > |t|)$$
since the t-distribution is symmetric.

Wald Tests

The *Wald test statistic* for a parameter θ with MLE $\hat{\theta}$ is given by
$$W = \frac{\hat{\theta} - \theta}{\hat{se}(\hat{\theta})} \sim \mathcal{N}(0,1).$$

- **One-sided test**:
  - When
$$H_0 : \theta = \theta_0, H_1 : \theta > \theta_0,$$
the rejection region is
$$R = \left\{ \mathbf{x} : \frac{\hat{\theta} - \theta_0}{\hat{se}(\hat{\theta})} > Z_{1-\alpha} \right\}.$$
  - When
$$H_0 : \theta = \theta_0, H_1 : \theta < \theta_0,$$
the rejection region is
$$R = \left\{ \mathbf{x} : \frac{\hat{\theta} - \theta_0}{\hat{se}(\hat{\theta})} < Z_{\alpha} \right\}.$$
- **Two-sided test**:
  - We have
$$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0.$$
  - The rejection region is
$$R = \{ \mathbf{x} : |\frac{\hat{\theta} - \theta_0}{\hat{se}(\hat{\theta})}| > Z_{1-\alpha/2} \},$$
where $Z_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of $\mathcal{N}(0,1)$.

Likelihood Ratio Tests

The generalised likelihood test for a hypothesis test
$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta - \Theta_0$$
has rejection region $\left\{ T(x) := \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} < c \right\}$, where c depends on the desired size of the test.

Power of a Test

Suppose that we are given the real value of a parameter θ. The *power of a test* is the probability of rejecting the null hypothesis. We have that
$$\beta(\theta) = \mathbb{P}(X \in R; \theta).$$

The significance level, or size, of a test α is given by
$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$