UNSW Mathematics Society Presents
MATH2801/2901 Workshop

Presented by John Kameas and Gorden Zhuang

Overview I

1. Probability Theory

Introduction to Probability

Basic Definitions

- An **outcome** is a possible result of an experiment. An example would be rolling a 1 on a six-sided die.
- The **sample space** (Ω) is the set of out all possible outcomes.
- An **event** is a set of outcomes with an assigned probability. An example would be rolling an odd number on a six-sided die.
- A σ-algebra ($\mathcal{F}$) is a collection of all possible events.

Probability Functions

Probability Functions

A probability function ($\mathbb{P}$) is a function that returns the probability of an event. It must satisfy the following properties:

1. All probabilities lie between 0 and 1 inclusive.
2. The probability of the sample space is 1.
3. The probability of any event in the sigma algebra and its complement add up to 1.

Additive Law

- For any two events,

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

- $\mathbb{P}(\emptyset) = 0$, so if A and B are mutually exclusive, then $\mathbb{P}(A \cap B) = 0$.

Extra: Probability Spaces

Definition of Probability Spaces

A **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ is a construct consisting of a sample space, a σ-algebra and a probability function. It forms a formal model to describe a random process.

Conditional Probability

Conditional Probability Formula

The **conditional probability** of A given B is given by:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \mathbb{P}(B) > 0.$$

Multiplicative Law

The above can easily be rearranged to form the multiplicative law:

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

For three events,

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

Independence

Definition of Independence

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

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

Using the conditional probability formula, this would imply

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

Introduction to Probability Question

(MATH2801) S1, 2018 – Q1(a)

Let A and B be two events in some sample space, with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$.

i) Show that, if A and B are independent, they cannot be mutually exclusive.

ii) Show that, if A and B are mutually exclusive, they cannot be independent.

iii) Suppose now that A and B are mutually exclusive. Show that

$$\mathbb{P}(A|A \cup B) = \frac{\mathbb{P}(A)}{\mathbb{P}(A) + \mathbb{P}(B)}.$$

Introduction to Probability Question

i) If A and B are independent, then

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) > 0,$$

since $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. This means $A \cap B \neq \emptyset$ and they cannot be mutually exclusive.

ii) Similarly, if A and B are mutually exclusive, then

$$\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0.$$

Since $\mathbb{P}(A)\mathbb{P}(B) > 0$, $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$, and so they cannot be independent.

iii) Since A and B are mutually exclusive, $A \cap B = \emptyset$ and $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. Then,

$$\mathbb{P}(A|A \cup B) = \frac{\mathbb{P}(A \cap (A \cup B))}{\mathbb{P}(A \cup B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A) + \mathbb{P}(B)}.$$

Independence For Multiple Events

More Independence

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 independent (or mutually independent) if for any subset $\{A_{i1}, A_{i2}, \ldots, A_{im}\}$,

$$\mathbb{P}(A_{i1} \cap A_{i2} \cap \cdots \cap A_{im}) = \prod_{j=1}^{m} \mathbb{P}(A_{ij}).$$

Independence For Multiple Events

More Independence

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 independent (or mutually independent) if for any subset $\{A_{i1}, A_{i2}, \ldots, A_{im}\}$,

$$\mathbb{P}(A_{i1} \cap A_{i2} \cap \cdots \cap A_{im}) = \prod_{j=1}^{m} \mathbb{P}(A_{ij}).$$

Remarks

- Mutual independence is stronger than pairwise independence.
- These definitions are important when proving independence.

Law of Total Probability

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

Bayes' Theorem

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

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

Introduction to Random Variables

Definition of a Random Variable

Formally, a **random variable** is a function that maps from the sample space to the real numbers. Informally, they can be thought of as variables whose values depend on the outcomes of a random experiment.

Notation

Random variables are denoted by capital letters (e.g. X) to distinguish them from deterministic variables.

Cumulative Distribution Functions

Definition of a Cumulative Distribution Function

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

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

Using CDF's

1. As $\{X : X > x\}^c = \{X : X \leq x\}$,

$$\mathbb{P}(X > x) = 1 - F_X(x).$$

2. For any $x < y$,

$$\mathbb{P}(x < X \leq y) = F_X(y) - F_X(x).$$

Properties of Cumulative Distribution Functions

Properties of CDF's

Suppose F is a cumulative distribution function. Then the following apply:

1. F is defined to be between zero and one and

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1;$$

2. F is non-decreasing for all x (i.e. $x \leq y \iff F(x) \leq F(y)$);
3. F is right continuous, meaning $F(x^+) = F(x)$ for all x.

Discrete Random Variables

Definition of a Discrete Random Variable

A random variable is **discrete** if it can take a countable number of values.

Definition of a Probability Mass Function

The **probability mass function** f_X of a discrete random variable X is defined by:

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

It is related to the cumulative distribution function by the following:

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

Continuous Random Variables

Definition of a Continuous Random Variable

A random variable is **continuous** if it can take a continuum of values.

Definition of a Probability Density Function

The **probability density function** (or PDF) f_X of a continuous random variable X is a positive function that satisfies

$$\mathbb{P}(X \in A) = \int_A f_X(y)\,\mathrm{d}y.$$

Importantly,

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y.$$

Introduction to Random Variables Question

Example Question: Symmetric Random Variables

A continuous random variable X is said to be *symmetric* if X and $-X$ have the same cumulative distribution function. On the other hand, a density function f is called *symmetric* if $f(x) = f(-x)$ for all $x \in \mathbb{R}$.

i) Show that $F_{-X}(x) = 1 - F_X(-x)$.

ii) Hence or otherwise, deduce that random variable X is symmetric if and only if the density of X given by f_X is symmetric.

Introduction to Random Variables Question

i) By definition,

$$\begin{aligned} F_{-X}(x) &= \mathbb{P}(-X \leq x) \\ &= \mathbb{P}(X \geq -x) \\ &= 1 - F_X(-x). \end{aligned}$$

(Since X is continuous, the distinction between $\geq$ and $>$ is irrelevant.)

Introduction to Random Variables Question

ii) If X is symmetric, then $F_X(x) = F_{-X}(x)$. Substituting this into the previous equality,

$$\begin{aligned} & F_X(x) = 1 - F_X(-x) \\ \iff\ & \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}(1 - F_X(-x)) && (*) \\ \iff\ & f_X(x) = f_X(-x), \end{aligned}$$

we see that f_X is symmetric. By reversing the steps, the converse is also true. (While reversing step (*) can lead to the two sides differing by a constant, this will not occur since F_X is defined to have a codomain of $[0, 1]$.)

Expectations

Definition of an Expectation

The **expectation** of a random variable X is its mean or average. It is often denoted by

- For discrete random variables,

$$\mathbb{E}(X) = \sum_{\text{all possible } x} x\mathbb{P}(X = x).$$

- For continuous random variables,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x.$$

μ or μ_X is often used to denote the **population mean** of X, which would be $\mathbb{E}(X)$.

Expectations on Transformations

Expectations on Functions of Random Variables

Where $g : \mathbb{R} \to \mathbb{R}$ is a function,

$$\mathbb{E}(g(X)) = \begin{cases} \displaystyle\sum_{\text{all possible } x} g(x)\mathbb{P}(X = x), & \text{for discrete } X; \\ \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x, & \text{for continuous } X. \end{cases}$$

More Expectations

Properties of Expectations

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

1. The expectation of a constant is the constant. i.e.

$$\mathbb{E}(a) = a$$

2. Expectations are linear. i.e.

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

3. IF X and Y are independent,

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

Variance

Definition of Variance

The **variance** of a random variable X is a measure of its spread from its mean. It is defined by

$$\begin{aligned}\mathbb{V}\text{ar}(X) &= \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 .\end{aligned}$$

The **standard deviation** is the square root of the variance.

The **population variance** is often denoted by σ^2 and the standard deviation by σ.

Properties of Variance

Let $a \in \mathbb{R}$ be a constant and X, Y be random variables once more.

Properties of Variance

1. $$\mathbb{V}\mathrm{ar}(aX) = a^2\mathbb{V}\mathrm{ar}(X)$$
2. $$\mathbb{V}\mathrm{ar}(a) = 0$$
3. $$\mathbb{V}\mathrm{ar}(X) = \mathbb{C}\mathrm{ov}(X, X)$$
4. $$\mathbb{V}\mathrm{ar}(X + a) = \mathbb{V}\mathrm{ar}(X)$$
5. $$\mathbb{V}\mathrm{ar}(X + Y) = \mathbb{V}\mathrm{ar}(X) + 2\mathbb{C}\mathrm{ov}(X, Y) + \mathbb{V}\mathrm{ar}(Y)$$

Expectations Question I

Example

Suppose X is a continuous random variable with probability density function

$$f_X(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}}, \quad x > 0.$$

i) Compute the expected value of X.

ii) Compute the variance of X.

iii) Let $Y \sim \mathrm{N}(0,1)$. Compute $\mathbb{E}(|Y|)$.

Expectations Question I

i)

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^\infty \sqrt{\frac{2}{\pi}} x e^{-\frac{x^2}{2}}\,\mathrm{d}x \\
&= \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-u}\,\mathrm{d}u && \text{(substituting } u = \frac{x^2}{2}\text{)} \\
&= \sqrt{\frac{2}{\pi}} \Big[-e^{-u} \Big]_0^\infty \\
&= \sqrt{\frac{2}{\pi}}
\end{aligned}$$

Expectations Question I

ii) As $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, we must compute $\mathbb{E}(X^2)$:

$$
\begin{aligned}
\mathbb{E}(X^2) &= \int_0^\infty \sqrt{\frac{2}{\pi}} x^2 e^{-\frac{x^2}{2}}\,\mathrm{d}x \\
&= \sqrt{\frac{2}{\pi}} \int_0^\infty x(xe^{-\frac{x^2}{2}})\,\mathrm{d}x \\
&= \sqrt{\frac{2}{\pi}} \left(\Big[-xe^{-\frac{x^2}{2}} \Big]_0^\infty - \int_0^\infty e^{-\frac{x^2}{2}}\,\mathrm{d}x \right) && \text{(IBP)} \\
&= \sqrt{\frac{2}{\pi}} \left(0 + \sqrt{2\pi} \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}\,\mathrm{d}x \right) \\
&= \sqrt{\frac{2}{\pi}} \sqrt{2\pi} \frac{1}{2} \\
&= 1
\end{aligned}
$$

Expectations Question I

ii) (Continued)

$$\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 1 - \frac{2}{\pi}$$

iii) Since $Y \sim \mathrm{N}(0,1)$ is symmetric, f_Y is an even function. Additionally, for $x > 0$, we observe that $f_X(x) = 2f_Y(x)$. Then,

$$\begin{aligned}
\mathbb{E}(|Y|) &= \int_{-\infty}^{\infty} |y| f_Y(y)\,\mathrm{d}y \\
&= 2\int_0^{\infty} y f_Y(y)\,\mathrm{d}y \\
&= \int_0^{\infty} y f_X(y)\,\mathrm{d}y \\
&= \mathbb{E}(X).
\end{aligned}$$

So $\mathbb{E}(|Y|) = \sqrt{\frac{2}{\pi}}$.

Moment Generating Functions

Definition of a Moment Generating Functions

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

$$\mathrm{m_X}(t) = \mathbb{E}(e^{tX}).$$

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

Uses of Moment Generating Functions

Moment generating functions are useful for two things:

1. Finding the non-central moments of a distribution.
2. Identifying distributions, since MGF's are unique to a distribution.

Generating Moments

Moments

The rth (non-central) moment of a random variable is defined as

$$\mathbb{E}(X^r)$$

for $r = 1, 2, \ldots$.

Generating Moments from a MGF

Suppose the moment generating function of X exists. Then for $r = 1, 2, \ldots,$

$$\mathbb{E}(X^r) = \lim_{t \to 0} \left(\frac{\mathrm{d}^r}{\mathrm{d}t^r} \mathrm{m}_X(t) \right).$$

Useful Inequalities

Markov's Inequality (or Chebyshev's First Inequality)

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

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

Chebyshev's Inequality (or Chebyshev's Second Inequality)

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

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

Expectations Question II

Example – Bounding Problem (MATH2801 Notes)

A factory produces 500 machines a day on average. It is subject to a variance of 100. Let X be the amount of machines produced tomorrow. Find a *lower* bound for the probability that between 400 to 600 machines are produced tomorrow.

This can be written as:

$$\mathbb{P}(400 < X < 600) = \mathbb{P}(|X - 500| < 100) = 1 - \mathbb{P}(|X - 500| \geq 100).$$

As $\mu = 500, \sigma = 10$, by Chebyshev's inequality,
$\mathbb{P}(|X - 500| \geq 100) \leq \frac{1}{10^2}$. Then a lower bound is given by

$$1 - \frac{1}{100} = \frac{99}{100}.$$

Extra: Jensen's Inequality

Definition of a Convex Function

A function h is convex if for any $\lambda \in [0, 1]$ and x_1 and x_2 in the domain of h,

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

Jensen's Inequality

If X is a random variable and h is a convex function, then

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

Common Distributions

Formula Sheet

A formula sheet containing details on the important distributions should be available on the Moodle pages of MATH2801 and MATH2901.

Using R

The probabilities associated with these distributions can easily be computed using R. However, for some distributions, the parameters are defined differently.

Discrete Distributions I

Bernoulli Distribution

A **Bernoulli**(p) random variable models whether a trial will be a success (with probability p) or a failure (with probability $q = 1 - p$).

Binomial Distribution

A **Binomial**(n, p) random variable models the number of successes out of n independent trials each with probability of success p.

Discrete Distributions II

Geometric Distribution

A **Geometric**(p) random variable models the number of trials it takes until the first success occurs (with probability p). This is defined to include the first success in R.

The Negative Binomial(k, p) is a generalisation of the Geometric distribution.

Hypergeometric Distribution

A **Hypergeometric**(N, m, n) random variable models the number of "white balls" in n balls randomly drawn out of an urn with m "white balls" and N balls in total. This is similar to the Binomial distribution, except the selections are done without replacement.

This is parameterised differently in R.

Discrete Distributions III

Poisson Distribution

A **Poisson**(λ) random variable models the number of occurrences of a random event that occurs at a **rate** of λ on average. Unlike Binomial random variables, Poisson random variables are not bounded above.

Continuous Distributions I

Uniform Distribution

A **Uniform**(a, b) random variable has a constant density function. It is equally likely to take values in any equally-sized region within $[a, b]$.

Continuous Distributions II

Exponential Distribution

An **Exponential**(β) random variable is often used to model the time it takes for an event to occur. β is the scale parameter and can be thought of as the spread of the distribution.
R takes the rate parameter, which is $\lambda = \frac{1}{\beta}$.

Gamma Distribution

A **Gamma**(α, β) random variable is the sum of α independent exponential(β) random variables.

Continuous Distributions III

Normal Distribution

The **Normal**(μ, σ^2) distribution (or Gaussian distribution) is a very important distribution and is used to approximate many unknown quantities. It is symmetric around its mean μ and has variance σ^2. R uses μ and σ as its parameters.

Standard Normal Distribution

A **standard normal distribution** is just a Normal$(0, 1)$ distribution. If $X \sim \mathrm{N}(\mu, \sigma^2)$, it can be standardised by the transformation:

$$Z := \frac{X - \mu}{\sigma}.$$

Note that a linear transformation of normal random variable will produce another normal random variable.

Continuous Distributions IV

Beta Distribution

A **Beta**(α, β) distribution is used to model the distribution of proportions. Its density function is given by

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}, \quad 0 \le x \le 1,$$

where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Its mean is

$$\frac{\alpha}{\alpha+\beta}.$$

Useful R Commands

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.

Insert the parameters as arguments in place of the `...` . To see how a distribution is parametrised in R, use `help(d___)` . For a list of distributions in R, use `help(distributions)` .

Common Distributions Question I

(MATH2901) T3, 2020 – Q1(i)

Suppose that the probability of hitting a target is $\frac{1}{5}$, and ten arrows are independently fired.

a) What is the probability of the target being hit at least twice?

b) What is the conditional probability that the target is hit at least twice, given that it is hit at least once?

Common Distributions Question I

a) Let X be the number of times the target is hit. Then, $X \sim \text{Binomial}(10, \frac{1}{5})$ and

$$\mathbb{P}(X \geq 2) = 1 - F_X(1) \approx 0.6241904$$

b)

$$\begin{aligned}
\mathbb{P}(X \geq 2 | X \geq 1) &= \frac{\mathbb{P}(X \geq 2, X \geq 1)}{\mathbb{P}(X \geq 1)} \\
&= \frac{\mathbb{P}(X \geq 2)}{\mathbb{P}(X \geq 1)} \\
&\approx \frac{0.6241904}{0.8926258} \\
&\approx 0.6992744
\end{aligned}$$

`pbinom(1, 10, 1/5) = 0.3758096`
`pbinom(0, 10, 1/5) = 0.1073742`.

Common Distributions Question II

Example

A batch of 100 machine parts is inspected by randomly choosing 10 and testing them individually. If at most 1 is faulty, the batch is accepted. Suppose the batch contains 8 faulty parts. Find the probability that the batch is accepted.

Let X be the number of faulty parts found. Since we are selecting 10 parts out of 100 with 8 faulty parts,
$X \sim \text{Hypergeometric}(N = 100, m = 8, n = 10)$. Then,

$$\begin{aligned}\mathbb{P}(X \leq 1) &= F_X(1) \\ &= \frac{\binom{10}{0}\binom{92}{10}}{\binom{100}{10}} + \frac{\binom{10}{1}\binom{92}{9}}{\binom{100}{10}} \\ &\approx 0.8180504.\end{aligned}$$

Q-Q Plots

Quantile Function

The quantile function is the inverse of the cumulative distribution function.

$$q = Q_X(p) \iff F_X(q) = p$$

Quantile-Quantile Plots

Q-Q plots are plots between two quantile functions. Straight lines indicate that the two distributions differ by a linear transform. They are often used a visual check to see if a data set comes from a particular distribution.

Joint Density Functions

Joint Density Functions

The **joint density function** of two continuous random variables determines their joint distribution. It must have the following properties:

1. $f_{X,Y}(x, y) \geq 0$, for all $(x, y) \in \mathbb{R}^2$,
2. $\displaystyle\iint_{\mathbb{R}^2} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y = 1$,
3. $\displaystyle\mathbb{P}(X \in A, Y \in B) = \int_{y \in B} \int_{x \in A} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y$.

Marginal Probability Functions

Marginal Probability Functions

The **marginal density function** is obtained by "integrating out" the other variable from the joint density function:

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

For discrete random variables:

$$f_X(x) = \sum_{\text{all } y} f_{X,Y}(x,y).$$

Independence of Random Variables

Independence of Random Variables

Two random variables are independent if and only if

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

and

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for all possible values of x and y, where $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$.

Multivariate Distributions Question

(MATH2901) S1, 2015 – Q1(a)

A tetrahedral die has the numbers 1,2,3 and 4 on each of its four faces. Two fair tetrahedral dice are rolled. The number of dice with a 1 on the downward face is recorded as X. The number of dice with a 4 on the downward face is recorded as Y.
The joint distribution of X and Y is shown in the following table.

$$f_{X,Y}(x,y)$$

		y		
		0	1	2
	0	4/16	4/16	1/16
x	1	4/16	2/16	0
	2	1/16	0	0

Multivariate Distributions Question

i) Explain why $f_{X,Y}(1,2) = 0$.

ii) Determine the marginal distribution $f_X(x)$.

iii) Are X and Y independent?

iv) Calculate $\mathbb{E}(X)$ and $\mathbb{V}\mathrm{ar}(X)$.

i) $f_{X,Y}(1,2)$ is the probability that $X = 1$ and $Y = 2$. This would mean getting 3 outcomes from two dice rolls, which is not possible.

ii) By summing the rows of the table,

$$f_X(x) = \begin{cases} \frac{9}{16}, & x = 0, \\ \frac{6}{16}, & x = 1, \\ \frac{1}{16}, & x = 2, \\ 0, & \text{otherwise.} \end{cases}$$

Multivariate Distributions Question

iii) By comparing $f_X(x)$ to the columns of the table, we can see that X and Y are not independent. As proof,

$$f_X(2) = \frac{1}{16} \neq 0 = f_{X|Y}(2|1).$$

iv)

$$\mathbb{E}(X) = \sum_x x f_X(x) = 0\left(\frac{9}{16}\right) + 1\left(\frac{6}{16}\right) + 2\left(\frac{1}{16}\right) = \frac{1}{2}$$

$$\mathbb{E}(X^2) = \sum_x x^2 f_X(x) = 0\left(\frac{9}{16}\right) + 1\left(\frac{6}{16}\right) + 4\left(\frac{1}{16}\right) = \frac{5}{8}$$

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{5}{8} - \left(\frac{1}{2}\right)^2 = \frac{3}{8}$$

Conditional Probability

Computing Conditional Probability

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

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

For discrete random variables, this represents the probability $X = x$ in a situation where $Y = y$. For continuous random variables, this is a density and not a probability.

Conditional Expectations and Variances

Conditional Expectation

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

$$\mathbb{E}(g(X)|Y = y) = \begin{cases} \sum_x g(x) f_{X|Y}(x|y), & \text{(discrete case)} \\ \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\,\mathrm{d}x, & \text{(continuous case).} \end{cases}$$

If Y is not given, this becomes a random variable on Y.

Conditional Variance

The conditional variance of X given $Y = y$ is

$$\mathbb{V}\mathrm{ar}(X|Y = y) = \mathbb{E}(X^2|Y = y) - \mathbb{E}(X|Y = y)^2.$$

Covariance

Definition of Covariance

The **covariance** of the two random variables X and Y is a measure of their joint variability and is given by

$$\begin{aligned}\mathbb{C}\mathrm{ov}(X,Y) &= \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) \\ &= \mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Properties of Covariance

1. If X and Y are independent, $\mathbb{C}\mathrm{ov}(X,Y)=0$. However, the converse is not true.
2. $\mathbb{C}\mathrm{ov}(X,Y)=\mathbb{C}\mathrm{ov}(Y,X)$
3. For any $a,b\in\mathbb{R}$, $\mathbb{C}\mathrm{ov}(aX+bY,Z)=a\,\mathbb{C}\mathrm{ov}(X,Z)+b\,\mathbb{C}\mathrm{ov}(Y,Z)$.

Correlation

Definition Correlation

Correlation is a measure of the strength of the linear relationship between two random variables and is defined by

$$\mathrm{Corr}(X, Y) = \frac{\mathbb{C}\mathrm{ov}(X, Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}}.$$

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 $\mathrm{Corr}(X, Y) = 0$.

Multivariate Gaussian

Multivariate Gaussian Distribution

The **multivariate Gaussian** (or multivariate normal) distribution is the multivariate form of the normal distribution. Its joint density is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu}$ is the vector of means, $\mathbf{\Sigma}$ is the covariance matrix and d is the dimension.

If $\mathbf{X} = (X_1, X_2)$ is multivariate Gaussian, then X_1 and X_2 are normally distributed. However, the converse is not true.

Transformations

Density of a Transformed Variable

Suppose X is a random variable and $Y = h(X)$, where h is monotone over the set $\{x : f(X) > 0\}$. Then the density of Y can be computed as:

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y} \right|.$$

CDF of a Random Variable

For any random variable X whose CDF is strictly increasing,

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

Transformations Question I

(MATH2901) T3, 2020 – Q3(i)

Let Z_i, for $i = 1, 2, \ldots$ be an i.i.d. sequence of random variables and $Z_i \sim \exp(1)$.

a) Compute the distribution of $Y_n := \min(Z_1, \ldots, Z_n)$.

b) Show that the probability density function $Y := (nY_n)^{\frac{1}{k}}$ for $k > 0$ is given by

$$f_Y(y) = ky^{k-1}e^{-y^k}, \quad y \geq 0.$$

c) Compute $\mathbb{E}(Y)$ and $\mathbb{V}\mathrm{ar}(Y)$.

Transformations Question I

a) $Y_n := \min(Z_1, \dots, Z_n)$ implies $Y_n \leq Z_i$ for all $1 \leq i \leq n$. Using this (for $y \geq 0$),

$$\begin{aligned}
F_{Y_n}(y) &= \mathbb{P}(Y_n \leq y) \\
&= 1 - \mathbb{P}(y < Y_n) \\
&= 1 - \mathbb{P}(y < Z_1, y < Z_2, \dots, y < Z_n) \\
&= 1 - (\mathbb{P}(y \leq Z_i))^n \\
&= 1 - (1 - F_Z(y))^n \\
&= 1 - (1 - (1 - e^{-y}))^n \\
&= 1 - e^{-ny}, \qquad y \geq 0.
\end{aligned}$$

Transformations Question I

b) As Y_n is positive, Y must be positive. Then for $y \geq 0$, using the cumulative distribution functions,

$$\begin{aligned} F_Y(y) &= \mathbb{P}((nY_n)^{\frac{1}{k}} \leq y) \\ &= \mathbb{P}(Y_n \leq \frac{1}{n}y^k) \\ &= 1 - e^{-n(\frac{1}{n}y^k)} \\ &= 1 - e^{-y^k}, \qquad y \geq 0. \end{aligned}$$

Then differentiate (using the chain rule) to obtain the probability density function of Y:

$$\begin{aligned} f_Y(y) &= F_Y'(y) \\ &= -\left(\frac{\mathrm{d}}{\mathrm{d}y}(-y^k)\right) e^{-y^k} \\ &= ky^{k-1}e^{-y^k}, \quad y \geq 0. \end{aligned}$$

Transformations Question I

c)

$$\begin{aligned}
\mathbb{E}(Y) &= \int_0^\infty yky^{k-1}e^{-y^k}\,\mathrm{d}y \\
&= \int_0^\infty u^{\frac{1}{k}}e^{-u}\,\mathrm{d}u \\
&= \Gamma(1+\frac{1}{k}) \quad (\text{Using } \Gamma(z) = \int\limits_0^\infty x^{z-1}e^{-x}\,\mathrm{d}x) \\
\mathbb{E}(Y^2) &= \int_0^\infty y^2ky^{k-1}e^{-y^k}\,\mathrm{d}y \\
&= \int_0^\infty u^{\frac{2}{k}}e^{-u}\,\mathrm{d}u \\
&= \Gamma(1+\frac{2}{k}) \\
\mathbb{V}\mathrm{ar}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \Gamma\left(1+\frac{2}{k}\right) - \Gamma\left(1+\frac{1}{k}\right)^2
\end{aligned}$$

Bivariate Transformations I

Using the Jacobian

Suppose random variables X and Y have a joint density $f_{X,Y}(x, y)$ and $(U, V) = (g_1(X, Y), g_2(X, Y))$. Then the joint density of U and V is given by

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

where $\det(J)$ is the determinant of the Jacobian, which would be

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

Even if we're only interested in one variable, say U, we can still apply this method by setting $V = Y$.

Transformations Question II

(MATH2901) S2, 2015 – Q4(a)

Let U and V be two random variables. Suppose $X = U + V$ and $Y = U - V$. If the joint density function of (X, Y) is given by

$$f_{X,Y}(x,y) = \frac{1}{2\sqrt{3}\pi} e^{-\frac{1}{2}\left[\frac{(x-4)^2}{3} + (y-2)^2\right]}, \quad x, y \in \mathbb{R}.$$

i) Compute the joint density function $f_{U,V}(u,v)$.

ii) Compute the marginal density function $f_U(u)$.

Transformations Question II

i) First, find the Jacobian

$$J = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and $|\det(J)| = 2$. So

$$f_{U,V}(u,v) = \frac{1}{\sqrt{3}\pi} e^{-\frac{1}{2}\left[\frac{(u+v-4)^2}{3} + (u-v-2)^2\right]} \quad u, v \in \mathbb{R}.$$

ii)

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{\sqrt{3}\pi} e^{-\frac{1}{2}\left[\frac{1}{3}(4u^2-4uv-20u+4v^2+4v+28)\right]} \\ &= \frac{1}{\sqrt{3}\pi} e^{-\frac{1}{2}\left[\frac{1}{3}((2v-(u-1))^2+3u^2-18u+27)\right]} \end{aligned}$$

Transformations Question II

$$
\begin{aligned}
f_U(u) &= \int_{-\infty}^{\infty} f_{U,V}(u,v)\partial v \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{3}\pi} e^{-\frac{1}{2}\left[\frac{1}{3}((2v-(u-1))^2+3u^2-18u+27)\right]}\partial v \\
&= \frac{e^{-\frac{1}{6}(3u^2-18u+27)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(\frac{3}{4})}} e^{-\frac{1}{2}\left[((v-\frac{u-2}{2})^2/\frac{3}{4}\right]}\partial v \\
&= \frac{e^{-\frac{1}{2}(u^2-6u+9)}}{\sqrt{2\pi}} \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(u-3)^2}
\end{aligned}
$$

Bivariate Transformations II

Using Convolutions

Suppose X and Y are independent random variables and $Z = X + Y$. In the discrete case,

$$f_Z(z) = \sum_{\text{all } y} f_X(z-y) f_Y(y).$$

In the continuous case,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) \, \mathrm{d}y.$$

Transformations Question III

(MATH2901) T3, 2020 – Q2(i)

Let X and Y be two random variables with joint density function

$$f_{X,Y}(x, y) = 2, \quad x \in (0, 1),\, y \in (0, 1),\, x < y.$$

Suppose X and Y represent the length of the base and the length of the height of a right angled triangle respectively. Then

a) Determine the conditional density $f_{Y|X}(y|x)$.

b) Are X and Y independent? Give reasons.

c) By integrating $f_{X,Y}(x, y)$ over an appropriate region of the plane, show that $\mathbb{P}(X + Y < 1) = \frac{1}{2}$.

Transformations Question III

a) First find $f_X(x)$:

$$f_X(x) = \int_x^1 2\,\mathrm{d}y = 2 - x$$

Then

$$f_{Y|X} = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{2}{2-x} = \frac{1}{1-x}.$$

b) For X and Y to be independent, $f_Y(y)$ must satisfy

$$f_X(x) f_Y(y) = f_{X,Y}(x,y)$$
$$\iff (2-x) f_Y(y) = 2.$$

As the left side of the equation is a function of x and the right side is not, this equality cannot be satisfied, meaning X and Y are not independent.

Transformations Question III

c) Sketching the region, we can see that the relevant region can be split into two triangles.

$$\begin{aligned}\mathbb{P}(X+Y<1) &= \int_0^{\frac{1}{2}}\int_0^y 2\,\mathrm{d}x\,\mathrm{d}y + \int_{\frac{1}{2}}^1\int_0^{1-y} 2\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{\frac{1}{2}} 2y\,\mathrm{d}y + \int_{\frac{1}{2}}^1 2(1-y)\,\mathrm{d}y \\ &= \Big[y^2\Big]_0^{\frac{1}{2}} + \Big[2y-y^2\Big]_{\frac{1}{2}}^1 \\ &= \frac{1}{4}+\frac{3}{4} \\ &= \frac{1}{2}\end{aligned}$$

Bivariate Tranformations III

Using Moment Generating Functions

Suppose X and Y are independent random variables whose moment generating functions exist. Then,

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

This is useful for identifying the resulting distribution.

Transformations Question IV

(MATH2801) S1, 2017 – Q3(a) (Modified)

Let X and Y be independent and identically distributed exponentially random variables, $X, Y \sim \exp(2)$.

i) Write down the joint density function $f_{X,Y}(x, y)$.

ii) Suppose $W = X + Y$. Determine the moment generating function of W.

iii) Hence determine the distribution of W.

Transformations Question IV

i) Since X and Y are independent,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{4}e^{-\frac{x+y}{2}}.$$

ii) Again, because X and Y are independent,

$$m_W(t) = m_X(t)m_Y(t) = \left(\frac{1}{1-2t}\right)^2.$$

iii) The general form of a Gamma mgf is

$$\left(\frac{1}{1-\beta t}\right)^\alpha.$$

Therefore, $W \sim \text{Gamma}(2,2)$.

Convergence of Random Variables I

Convergence in Distribution

We say a sequence of random variables $X_1, X_2, \ldots$ **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 $X_1, X_2, \ldots$ **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 $X_n \xrightarrow{\mathbb{P}} X$.

Convergence of Random Variables II

Almost Sure Convergence

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

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

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

Convergence in Mean

A sequence of random variables $X_1, X_2, \ldots$ **convergences in** L^p to a random variable X if for $p \geq 1$:

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0.$$

This is often denoted as $X_n \xrightarrow{L^p} X$. We say X_n converges to X in mean square if $p = 2$.

Convergence of Random Variables III

Comparison of Convergence Strengths

- Almost sure convergence is stronger than convergence in probability, which is stronger than convergence in distribution.
- Convergence in L^p is also stronger than convergence in probability, but is not necessarily stronger or weaker than almost sure convergence. Additionally, it is stronger for higher p.

Convergence of Random Variables Question I

Example

Let $X_1, X_2, \dots$ be a sequence of r.v. such that $X_n \sim \text{Bernoulli}(\frac{1}{n})$ for all n. Prove that $X_n \xrightarrow{\mathbb{P}} 0$.

First, consider the case where $0 < \epsilon < 1$:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \epsilon) &= \lim_{n\to\infty} \mathbb{P}(X_n = 1) \\ &= \lim_{n\to\infty} \frac{1}{n} \\ &= 0\end{aligned}$$

Since a Bernoulli random variable can only take values of 0 and 1, for any $\epsilon \geq 1$, $\mathbb{P}(X_n > \epsilon) = 0$. So for all $\epsilon > 0$

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

and therefore, $X_n \xrightarrow{\mathbb{P}} 0$.

Central Limit Theorem

Definition of the Central Limit Theorem

Suppose $X_1, X_2, \dots, X_n$ is a sequence of i.i.d. random variables each with mean $\mathbb{E}(X_i) = \mu$ and finite variance $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2$. Then the sample mean $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ follows a

$$\text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

distribution asymptotically. Alternatively,

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

Convergence of Random Variables Question II

(MATH2801) S1, 2018 – Q3(a)

Let $X_1, X_2, \ldots, X_{48}$ be i.i.d Uniform$(0, 1)$ random variables, and set

$$\bar{X} = \frac{1}{48} \sum_{i=1}^{48} X_i.$$

Using the Central Limit Theorem, compute $\mathbb{P}(\bar{X} > 0.55)$.

Convergence of Random Variables Question II

For a Uniform$(0, 1)$ random variable, $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$. Then by the CLT approximation,

$$\frac{\bar{X} - \frac{1}{2}}{\sqrt{\left(\frac{1}{12}\right)/48}} \sim \mathrm{N}(0, 1).$$

Hence,

$$\begin{aligned}
\mathbb{P}(\bar{X} > 0.55) &\approx \mathbb{P}\left(\frac{\bar{X} - \frac{1}{2}}{\frac{1}{24}} > \frac{0.55 - \frac{1}{2}}{\frac{1}{24}}\right) \\
&= 1 - \mathbb{P}(Z \leq 1.2) \\
&= 0.1150697.
\end{aligned}$$

Law of Large Numbers

Weak Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ 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:

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

Essentially, this means that as we take larger sample sizes, our 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 Central Limit Theorem

Normal approximation to Binomial distribution

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

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

Applications of the Central Limit Theorem

Normal approximation to Binomial distribution

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

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

What is this actually saying?

The binomial distribution is actually the discrete version of the normal distribution!

Applications of the Central Limit Theorem

MATH2901-2020 Q1.ii)

Two theatres compete for the business of 1000 customers. Assume that each customer chooses between the theatres independently (and is indifferent between the two). Let N denote the number of seats in each theatre.

a) Using a binomial model to find a condition, in terms of N, which will guarantee that the probability of a particular theatre turning away a customer (because the theatre is full) is less than 1%.

b) Explain how a binomial distribution can be approximated by a normal distribution.

c) By using the normal approximation, give an approximate value for N, so that the condition obtain in a) is satisfied.

Part a)

First things first, we need to set up appropriate notation.

Part a)

First things first, we need to set up appropriate notation. Let X be the number of people in any particular theatre; then, $X \sim \text{Bin}(1000, 1/2)$ because each customer is indifferent between the two options. This means we require that $\mathbb{P}(X > N) < 0.01$. That is,

$$\mathbb{P}(X > N) = \sum_{x=N+1}^{1000} \binom{1000}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{1000-x} < 0.01$$

which implies that

$$\left(\frac{1}{2}\right)^{1000} \sum_{x=N+1}^{1000} \binom{1000}{x} < 0.01.$$

Part a)

First things first, we need to set up appropriate notation. Let X be the number of people in any particular theatre; then, $X \sim \text{Bin}(1000, 1/2)$ because each customer is indifferent between the two options. This means we require that $\mathbb{P}(X > N) < 0.01$. That is,

$$\mathbb{P}(X > N) = \sum_{x=N+1}^{1000} \binom{1000}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{1000-x} < 0.01$$

which implies that

$$\left(\frac{1}{2}\right)^{1000} \sum_{x=N+1}^{1000} \binom{1000}{x} < 0.01.$$

Here, we can theoretically solve for N, that is, N is the smallest integer that satisfies

$$\left(\frac{1}{2}\right)^{1000} \sum_{x=N+1}^{1000} \binom{1000}{x} < 0.01.$$

Part b)

Note that solving for N in the above expression is not so easy! This is why we're being asked to find an expression involving N (and not being asked to actually solve for it).

Part b)

Note that solving for N in the above expression is not so easy! This is why we're being asked to find an expression involving N (and not being asked to actually solve for it).

Since, we can't (or more accurately aren't sure) how to find N, we can find an approximation for it, noting that for $X \sim \text{Bin}(n, p)$ we have,

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

Part c)

We will now use this approximation to obtain an answer. We note that in our case, $X \sim \text{Bin}(1000, 1/2)$ and hence, $\mathbb{E}(X) = 1000(1/2) = 500$. Similarly, $\sqrt{\mathbb{V}\text{ar}(X)} = \sqrt{1000(1/2)(1/2)} = \sqrt{250}$.
Then, if we require $\mathbb{P}(X > N) < 0.01$, this is equivalent to

$$\mathbb{P}\left(\frac{X - 500}{\sqrt{250}} > \frac{N - 500}{\sqrt{250}}\right) < 0.01.$$

Using our approximation, we know that

$$\mathbb{P}\left(Z > \frac{N - 500}{\sqrt{250}}\right)$$

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

Part c)-cont

I omitted it from the question stem, but we were also told the following information:

```
> qnorm(0.99,0,1)
[1] 2.326348
> qnorm(0.9,0,1)
[1] 1.281552
> qnorm(0.95,0,1)
[1] 1.644854
```

Part c)-cont

I omitted it from the question stem, but we were also told the following information:

```
> qnorm(0.99,0,1)
[1] 2.326348
> qnorm(0.9,0,1)
[1] 1.281552
> qnorm(0.95,0,1)
[1] 1.644854
```

Since, we want to get a 1% probability, we will be using the top result

Part c)-cont

Continuing on, we have that

$$\begin{aligned} \mathbb{P}(Z > 2.33) \approx 0.99 < 1 &\implies \frac{N - 500}{\sqrt{250}} = 2.33 \\ &\implies N \approx 537. \end{aligned}$$

So to conclude, each theatre should have at least 537 seats.

Slutsky's Theorem

Slutsky's Theorem

Let $(X_n)_{n \in \mathbb{N}_+}$ be a sequence of r.vs converging to X in *distribution* and $(Y_i)_{i \in \mathbb{N}_+}$ is another sequence of r.vs that converges in *probability* to a constant c, then

1. $X_n + Y_n \xrightarrow{\mathrm{d}} X + c$;
2. $X_n Y_n \xrightarrow{\mathrm{d}} Xc$.

A quick example!

Application of Slutsky's Theorem

Suppose that $X_1, X_2, \ldots$ converges in distribution to $X \sim \mathcal{N}(0,1)$, i.e. $X_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$, and suppose that $nY_n \sim \mathrm{Bin}(n, \frac{1}{2})$.
What are the limiting distributions of $X_n + Y_n$ and X_nY_n?

A quick example!

Application of Slutsky's Theorem

Suppose that $X_1, X_2, \ldots$ converges in distribution to $X \sim \mathcal{N}(0,1)$, i.e. $X_n \xrightarrow{\text{d}} \mathcal{N}(0,1)$, and suppose that $nY_n \sim \text{Bin}(n, \frac{1}{2})$.
What are the limiting distributions of $X_n + Y_n$ and X_nY_n?

How can we apply Slutsky's theorem to obtain these distributions? Remember we need **two** sequences of random variables, not one.

A quick example!

Application of Slutsky's Theorem

Suppose that $X_1, X_2, \dots$ converges in distribution to $X \sim \mathcal{N}(0,1)$, i.e. $X_n \xrightarrow{\text{d}} \mathcal{N}(0,1)$, and suppose that $nY_n \sim \text{Bin}(n, \frac{1}{2})$.
What are the limiting distributions of $X_n + Y_n$ and $X_n Y_n$?

How can we apply Slutsky's theorem to obtain these distributions? Remember we need **two** sequences of random variables, not one. We need to recall that any binomial distribution is in fact a sum of n i.i.d Bernoulli random variables.
That is,

$$nY_n = \sum_{i=1}^{n} \text{Bern}(p) \implies Y_n = \frac{1}{n} \sum_{i=1}^{n} \text{Bern}(p).$$

A quick example!

Application of Slutsky's Theorem

Suppose that $X_1, X_2, \ldots$ converges in distribution to $X \sim \mathcal{N}(0,1)$, i.e. $X_n \xrightarrow{\text{d}} \mathcal{N}(0,1)$, and suppose that $nY_n \sim \text{Bin}(n, \frac{1}{2})$.
What are the limiting distributions of $X_n + Y_n$ and $X_n Y_n$?

Hence, we can apply the weak law of large numbers to claim that

$$Y_n \xrightarrow{\mathbb{P}} \mathbb{E}[\text{Bern}(p)] = \frac{1}{2}.$$

Therefore, by Slutsky's theorem,

$$X_n + Y_n \xrightarrow{\text{d}} \mathcal{N}(\frac{1}{2}, 1) \text{ and } X_n Y_n \xrightarrow{\text{d}} \mathcal{N}(0, \frac{1}{4}).$$

The Delta Method

Let $Y_1, Y_2, \dots$ be a sequence of random variables such that

$$\frac{\sqrt{n}(Y_n - \theta)}{\sigma} \sim \mathcal{N}(0, 1).$$

Suppose that g is differentiable in the neighbourhood of θ and $g'(\theta) \neq 0$. Then,

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2[g'(\theta)]^2).$$

What's the point of the delta method? Recall a common problem discussed by Gorden; the situation where we know the distribution of a random variable, but we want to determine the distribution of a function of it. The delta method gives us a very direct route to easily finding (limiting) distributions of a function of a known random variable.

Example of the Delta Method

MATH2901 2015 Q2)(c)

Let X_i, $i = 1, 2, \ldots$, be independent Bernoulli(p) random variables and let $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

i) Show that as $n \to \infty$, $\sqrt{n}(Y_n - p) \xrightarrow{\mathrm{d}} \mathcal{N}(0, p(1-p))$.

ii) Show that for $p \neq 1/2$, as $n \to \infty$, the random variables $Y_n(1 - Y_n)$ satisfies

$$\sqrt{n}(Y_n(1 - Y_n) - p(1-p)) \xrightarrow{\mathrm{d}} \mathcal{N}(0, (1-2p)^2 p(1-p)).$$

Example of the Delta Method

i)

Show that as $n \to \infty$, $\sqrt{n}(Y_n - p) \xrightarrow{\text{d}} \mathcal{N}(0, p(1-p))$.

Example of the Delta Method

i)

Show that as $n \to \infty$, $\sqrt{n}(Y_n - p) \xrightarrow{\text{d}} \mathcal{N}(0, p(1-p))$.

We simply need to use the central limit theorem! As such,

$$\sqrt{n}(Y_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2).$$

In our case, $\mu = p$, and $\sigma^2 = p(1-p)$, and so

$$\sqrt{n}(Y_n - p) \xrightarrow{\text{d}} \mathcal{N}(0, p(1-p)).$$

Example of the Delta Method

ii)

Show that for $p \neq 1/2$, as $n \to \infty$, the random variables $Y_n(1 - Y_n)$ satisfies

$$\sqrt{n}(Y_n(1 - Y_n) - p(1 - p)) \xrightarrow{\text{d}} \mathcal{N}(0, (1 - 2p)^2 p(1 - p)).$$

Example of the Delta Method

ii)

Show that for $p \neq 1/2$, as $n \to \infty$, the random variables $Y_n(1-Y_n)$ satisfies

$$\sqrt{n}(Y_n(1-Y_n) - p(1-p)) \xrightarrow{\text{d}} \mathcal{N}(0, (1-2p)^2 p(1-p)).$$

We can now use the Delta method. We let our differentiable function be $g(x) = x(1-x)$, then $g'(x) = 1-2x$. We also note that $g'(1/2) = 0$ so this is why we need $p \neq 1/2$. So after assuming this, and plugging in to the delta method, we have

$$\begin{aligned} &\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2[g'(\theta)]^2) \\ \implies\ &\sqrt{n}(Y_n(1-Y_n) - p(1-p)) \xrightarrow{\text{d}} \mathcal{N}(0, p(1-p)(1-2p)^2). \end{aligned}$$

2. Statistical Inference

Estimators

Definition of Estimator

Suppose $(X_1, \ldots, X_n) \sim \{f_X(x; \theta), \theta \in \Theta\}$.
An estimator of θ, denoted by $\widehat{\theta}_n$ is any real valued function of $X_1, \ldots, X_n$. That is,

$$\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n) = g(X_1, \ldots, X_n)$$

where $g : \mathbb{R}^n \to \mathbb{R}$.

Estimators

Definition of Estimator

Suppose $(X_1, \ldots, X_n) \sim \{f_X(x;\theta), \theta \in \Theta\}$.
An estimator of θ, denoted by $\widehat{\theta}_n$ is any real valued function of $X_1, \ldots, X_n$. That is,

$$\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n) = g(X_1, \ldots, X_n)$$

where $g : \mathbb{R}^n \to \mathbb{R}$.

Unpacking this Definition

- $X_1, \ldots, X_n$ denote a sample on a random variable X.
- X has pdf $f(x;\theta)$ i.e. f is a function of $x \in \Omega$, with some parameter $\theta \in \Theta$ which we wish to estimate.
- What we call the estimator is simply a function $\widehat{\theta}$ on the sample, which attempts to estimate the value of θ.

Estimators Cont

Properties of the Estimator

- The estimator $\widehat{\theta}$ is a random variable. Why?
- Subsequently, an estimator also has its own probability distribution and can be computed from the distribution of $(X_1, \ldots, X_n)$. Think of an estimator which has a different distribution from the sample data.

Bias of Estimators

Definition of Bias

Let $\widehat{\theta}$ be an estimator of the parameter θ. The bias of the estimator $\widehat{\theta}$ is defined as

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta.$$

Definition of Unbiased Estimator

If

$$\text{Bias}(\widehat{\theta}) = 0 \implies \mathbb{E}(\widehat{\theta}) = \theta,$$

then $\widehat{\theta}$ is said to be an unbiased estimator of θ.

Examples

Example

Suppose X is a discrete random variable with pmf $p(x)$, and that the space of X is finite, say $\mathcal{D} = \{a_1, \ldots, a_m\}$. Suppose that $X_1, \ldots, X_n$ are samples taken from X such that each sample is independent of the other.
Show that

$$\widehat{p}(a_j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_j(X_i)$$

is an unbiased estimator of $p(a_j)$ and compute $\mathbb{V}\mathrm{ar}(\widehat{p})$ if possible.
Recall that

$$\mathbb{I}_j(X_i) := \begin{cases} 1 & \text{if } X_i = a_j, \\ 0 & \text{if } X_i \neq a_j. \end{cases}$$

Examples

To solve these kinds of questions, it is always a good idea to attempt to find the expectation of our estimator and simplify as far as we can go. So applying the expectation on $\widehat{p}$ gives,

$$\mathbb{E}(\widehat{p}(a_j)) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_j(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbb{I}_j(X_i)).$$

Now, we need to figure out the distribution of the indicator function. We can recognise that the indicator function is really just assigning a 1 if an event happens or 0 if it doesn't happen i.e. success or failure. Hence†, the indicator function has a Bernoulli distribution.

Examples

Thus, $\mathbb{E}(\mathbb{I}_j(X_i)) = \mathbb{P}(X_i = a_j)$ and so,

$$\begin{aligned}
\mathbb{E}(\widehat{p}(a_j)) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\mathbb{I}_j(X_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X_i = a_j) \\
&= \frac{1}{n}\sum_{i=1}^{n} p(a_j) \\
&= p(a_j).
\end{aligned}$$

So we can conclude,

$$\mathrm{Bias}(\widehat{p}(a_j)) = p(a_j) - p(a_j) = 0,$$

which implies that $\widehat{p}(a_j)$ is an unbiased estimator.

Examples

Recall that if $Y \sim \text{Bern}(p)$, then $\mathbb{V}\text{ar}(Y) = p(1-p)$. Hence,

$$\begin{aligned}
\mathbb{V}\text{ar}(\widehat{p}(a_j)) &= \mathbb{V}\text{ar}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_j(X_i)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\text{ar}(\mathbb{I}_j(X_i)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} p(a_j)(1-p(a_j)) \\
&= \frac{p(a_j)[1-p(a_j)]}{n}.
\end{aligned}$$

Which estimators are ‘better’ than others?

2020 Assignment Qn 4

Let $X_1, X_2, \dots, X_n$ be a random sample (i.i.d) with mean μ_X and variance $\sigma_X^2 < \infty$. The usual estimator for μ is $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Assume that $n > 3$. A student investigates an alternative estimator for μ, by ignoring X_{n-1} and X_n and multiplying X_1 by 3, giving

$$\tilde{X}_n = \frac{3X_1 + \sum_{i=1}^{n-2} X_i}{n} = \frac{3X_1 + X_2 + \cdots X_{n-2}}{n}.$$

1. Show that $\tilde{X}_n$ is an unbiased estimator of μ.
2. Determine the mean square error, $\mathrm{MSE}(\tilde{X}_n)$.
3. Show that $\lim_{n\to\infty} \mathrm{MSE}(\tilde{X}_n) = 0$.

Part 1.

1. Once again, we apply the expectation operator on $\tilde{X}_n$ and hope for the best! That is,

$$\mathbb{E}(\tilde{X}_n) = \mathbb{E}\left(\frac{3X_1 + \sum_{i=2}^{n-2} X_i}{n}\right).$$

We can exploit the linearity of $\mathbb{E}$ to get that,

$$\begin{aligned}\mathbb{E}(\tilde{X}_n) &= \frac{1}{n}\left(3\mathbb{E}(X_1) + \sum_{i=2}^{n-2} \mathbb{E}(X_i)\right)\\ &= \frac{3\mu + (n-3)\mu}{n}\\ &= \mu.\end{aligned}$$

So we conclude that $\text{Bias}(\tilde{X}_n) = \mu - \mu = 0$ and hence $\tilde{X}_n$ is an unbiased estimator.

Part 2.

Just in case you need to jog your memory:

Definition of MSE (Mean Squared Error)

We define the MSE as

$$\mathrm{MSE}(\tilde{X}_n) = \mathbb{E}[(\tilde{X}_n - \mu)^2].$$

It is quite straightforward to demonstrate that this definition is equivalent to

$$\mathrm{MSE}(\tilde{X}_n) = \mathrm{Bias}(\tilde{X}_n)^2 + \mathbb{V}\mathrm{ar}(\tilde{X}_n).$$

2. From the previous part, we know that $\mathrm{Bias}(\tilde{X}_n) = 0$ and so it is sufficient to calculate $\mathbb{V}\mathrm{ar}(\tilde{X}_n)$ for the MSE.

Part 2. (Cont)

2-(cont). As such by the usual properties of variance,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\tilde{X}_n) &= \mathbb{V}\mathrm{ar}\left(\frac{3X_1 + \sum_{i=2}^{n-2} X_i}{n}\right) \\
&= \frac{1}{n^2}\left(9\mathbb{V}\mathrm{ar}(X_1) + \sum_{i=2}^{n-2} \mathbb{V}\mathrm{ar}(X_i)\right) \\
&= \frac{9\sigma^2 + \sigma^2(n-3)}{n^2} \\
&= \frac{6\sigma^2}{n^2} + \frac{\sigma^2}{n}.
\end{aligned}$$

Thus, $\mathrm{MSE}(\tilde{X}_n) = \dfrac{\sigma^2}{n^2}(6+n)$.

Part 3

3. We just deduced $\text{MSE}(\tilde{X_n}) = \dfrac{\sigma^2}{n^2}(6+n)$. So by the linearity of the limit operator,

$$\begin{aligned}\lim_{n\to\infty} \text{MSE}(\tilde{X_n}) &= \lim_{n\to\infty} \frac{6\sigma^2}{n^2} + \lim_{n\to\infty} \frac{\sigma^2}{n} \\ &= 0\end{aligned}$$

since σ^2 is constant.

Part 4

Part 4

Both $\tilde{X}_n$ and $\overline{X}_n$ are unbiased estimators for μ for which

$$\lim_{n\to\infty} \mathrm{MSE}(\tilde{X}_n) = \lim_{n\to\infty} \mathrm{MSE}(\overline{X}_n) = 0.$$

Explain, using concepts learned in MATH2901 which of $\tilde{X}_n$ and $\overline{X}_n$ you would consider to be a better estimate of μ.

Part 4

Part 4

Both $\tilde{X}_n$ and $\overline{X}_n$ are unbiased estimators for μ for which

$$\lim_{n\to\infty} \mathrm{MSE}(\tilde{X}_n) = \lim_{n\to\infty} \mathrm{MSE}(\overline{X}_n) = 0.$$

Explain, using concepts learned in MATH2901 which of $\tilde{X}_n$ and $\overline{X}_n$ you would consider to be a better estimate of μ.

Since both estimators are unbiased it is valid to consider one a better estimate of μ if the estimate has a smaller variance than the other. As such, note that

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n} \leq \frac{6\sigma^2}{n^2} + \frac{\sigma^2}{n} = \mathrm{Var}(\tilde{X}_n)$$

since $\frac{6\sigma^2}{n^2} \geq 0$. Hence $\mathrm{Var}(\overline{X}_n) \leq \mathrm{Var}(\tilde{X}_n)$ and so $\overline{X}_n$ is the better estimator.

Asymptotic Properties of the estimator

Consistent Estimator

The estimator $\hat{\theta}$ is a consistent estimator of θ if, as $n \to \infty$,

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

Asymptotically Normal

An estimator $\hat{\theta}_n$ of θ is asymptotically normal if

$$\frac{\hat{\theta}_n}{\mathrm{Se}(\hat{\theta}_n)} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

A very important distribution in inference

Definition of Student t-distribution

A random variable T is said to have t-distribution with degree of freedom ν, if its probability density function

$$f_T(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\Gamma(1/2)}\nu^{-1/2}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \quad x \in (-\infty, \infty).$$

A very important distribution in inference

Definition of Student t-distribution

A random variable T is said to have t-distribution with degree of freedom ν, if its probability density function

$$f_T(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\Gamma(1/2)}\nu^{-1/2}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \quad x \in (-\infty, \infty).$$

Once again, why should you care? Recall that, if $X_1, X_2, \dots$ are i.i.d random samples from $\mathcal{N}(\mu, \sigma^2)$ then

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

However, if we replace σ^2 by $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ then

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

Condidence Intervals

Suppose that we have a random variable of interest X with density $f(x;\theta), \theta \in \Omega$, where θ is unknown. We can estimate θ by a statistic $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ is a sample from the distribution of X. When the sample is drawn, it is unlikely that the value of $\hat{\theta}$ is the *true* value of the parameter. In fact, if $\hat{\theta}$ has a continuous distribution, then $\mathbb{P}(\hat{\theta} = \theta) = 0$.

Condidence Intervals

Suppose that we have a random variable of interest X with density $f(x;\theta), \theta \in \Omega$, where θ is unknown. We can estimate θ by a statistic $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ is a sample from the distribution of X. When the sample is drawn, it is unlikely that the value of $\hat{\theta}$ is the *true* value of the parameter. In fact, if $\hat{\theta}$ has a continuous distribution, then $\mathbb{P}(\hat{\theta} = \theta) = 0$.

How do we reconcile this issue?

What is needed is an estimate of the error of the estimation, i.e., by how much did $\hat{\theta}$ miss θ? This is the purpose of the confidence interval.

Confidence Intervals

Definition of Confidence Interval

Let $X_1, X_2, \ldots, X_n$ be a sample on a random variable X, where X has pdf $f(x; \theta), \theta \in \Omega$. Let $0 < \alpha < 1$ be specified. Let $L = L(X_1, X_2, \ldots, X_n)$ and $U = U(X_1, X_2, \ldots, X_n)$ be two statistics. We say that the interval (L, U) is a $(1 - \alpha)100\%$ confidence interval for θ if

$$1 - \alpha = \mathbb{P}[\theta \in (L, U)].$$

Unpacking this Definition

- We have two functions on the sample L and U which have returned some values, and we construct an interval based on these values, namely, (L, U).
- The probability that $\theta \in (L, U)$ is $1 - \alpha$.
- We call $1 - \alpha$ the **confidence coefficient** of the interval.

Confidence Intervals

Confidence Intervals for a Normal Random Sample

Let $X_1, X_2, \ldots, X_n$ be a sample from the $\mathcal{N}(\mu, \sigma^2)$. Then a $100(1-\alpha)\%$ confidence interval for μ is

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

Unpacking this: it's just a formula!

- $\overline{X}$ is just the in-sample mean.
- t_{n-1} is just the t distribution with $n-1$ degrees of freedom.
- S is just the in-sample variance.
- $t_{n-1,1-\alpha/2}$ is just the $(1-\alpha/2)$th quantile of the t_{n-1} dist. If we have a large sample ($n \to \infty$), and we want to know the 95th percentile, we can write in R, `abs(qt(0.95, Inf))`.

Example of constructing a Confidence Interval

MATH2901 Assignment Q5

The density of a lognormal random variable X is

$$f_X(x;\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x)^2}{2\sigma^2}},\ x > 0$$

where $\sigma > 0$ is some constant. Note that

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i)^2},$$

is an estimator for σ.

- Compute the distribution of the estimator $\frac{\hat{\sigma}^2}{\sigma^2}$ and compute, using the data set below, construct a two sided 95% confidence interval for σ^2.

MATH2901 Assignment Q5

Squaring both sides, then dividing both sides by σ^2, we have,

$$\frac{\hat{\sigma}^2}{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\ln X_i}{\sigma}\right)^2.$$

We now make the following observations:

- $\ln X \sim \mathcal{N}(0, \sigma^2)$.
- Standardising this normal variable gives us $\frac{\ln X}{\sigma} \sim Z = \mathcal{N}(0, 1)$.
- The sum of n normal random variables squared is a chi-squared distribution with n degrees of freedom, i.e. $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$.
- The chi-squared distribution is a special case of the gamma distribution: $\chi_n^2 \sim \text{Gamma}(\frac{n}{2}, 2)$.
- If some random variable $X \sim \text{Gamma}(\alpha, \beta)$, then $\frac{X}{n} \sim \text{Gamma}(\alpha, \frac{\beta}{n})$.

Hence,

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \text{Gamma}\left(\frac{n}{2}, \frac{2}{n}\right).$$

Hence, we require that

$$g_{0.025} < \frac{\hat{\sigma}^2}{\sigma^2} < g_{0.975} \iff \frac{\hat{\sigma}^2}{g_{0.975}} < \sigma^2 < \frac{\hat{\sigma}^2}{g_{0.025}}.$$

We can then use the below R script to obtain these values with some data-set provided:

```
x <- c(SOME NUMBERS)
n <- length(x)
estim <- sqrt(1/n * sum(log(x)^2))
g0.025 <- qgamma(0.025, n/2, 2/n)
g0.975 <- qgamma(0.975, n/2, 2/n)
lower <- estim^2 / g0.975
higher <- estim^2 /g0.025
```

Likelihood Estimator

Definition of Likelihood Estimator

Let $x_1, \ldots, x_n$ be observations from the pdf f where

$$f(x) = f(x;\theta)$$

for some $\theta \in \Theta$. The likelihood function $\mathcal{L}$ (which is a function of θ), is

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

and the log-likelihood function of θ is,

$$\ell(\theta) = \ln\{\mathcal{L}(\theta)\} = \sum_i \ln\{f(x_i;\theta)\}.$$

Maximum Likelihood Estimator (MLE)

Definition of Maximum Likelihood Estimator (MLE)

Let $x_1, \ldots, x_n$ be observations from the pdf f where

$$f(x) = f(x; \theta)$$

for some $\theta \in \Theta$. The maximum likelihood estimate of θ is the choice

$$\hat{\theta} = \theta \text{ that maximises } \mathcal{L}(\theta) \text{ over } \theta \in \Theta.$$

An Important Result

Result

The point at which $\mathcal{L}(\theta)$ attains its maximum over $\theta \in \Theta$ is also where

$$\ell(\theta) = \ln\{\mathcal{L}(\theta)\} = \sum_i \ln\{f(x_i; \theta)\}$$

attains its maximum. Therefore, the maximum likelihood estimate of θ is

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

MLE

2020 MATH2901 Assignment Q5 (again)

The density of a lognormal random variable X is

$$f_X(x;\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x)^2}{2\sigma^2}}, \; x > 0$$

where $\sigma > 0$ is some constant.

1. Let $X_1, \ldots, X_n$ be a random sample of size n from the parametric family $f_X(x;\sigma)$, show that the maximum likelihood estimator of σ is given by
$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i)^2}.$$

MLE

Denote the likelihood estimator by the following

$$\mathcal{L}_n(x) = \prod_{i=1}^{n} \frac{1}{x_i \sigma \sqrt{2\pi}} e^{-\frac{(\ln x_i)^2}{2\sigma^2}}$$

and denote the log-likelihood estimator with

$$\ell_n(x) = \ln(\mathcal{L}_n(x)) = \ln \left(\prod_{i=1}^{n} \frac{1}{x_i \sigma \sqrt{2\pi}} e^{-\frac{(\ln x_i)^2}{2\sigma^2}} \right).$$

Note that maximising $\mathcal{L}_n(x)$ is equivalent to maximising $\ell_n(x)$.

MLE

Now it follows from the usual properties of the logarithm that

$$\begin{aligned}\ell_n(x) &= \sum_{i=1}^{n}\left[\ln\left(\frac{1}{x_i\sigma\sqrt{2\pi}}\right) - \frac{(\ln x_i)^2}{2\sigma^2}\right] \\ &= \sum_{i=1}^{n}\left[-\ln(\sigma) - \ln(x_i) - \ln(\sqrt{2\pi}) - \frac{(\ln x_i)^2}{2\sigma^2}\right] \\ &= -n\ln(\sigma) - n\ln\sqrt{2\pi} - \sum_{i=1}^{n}\ln(x_i) - \sum_{i=1}^{n}\frac{(\ln x_i)^2}{2\sigma^2}.\end{aligned}$$

MLE

We can then compute:

$$\frac{\partial}{\partial \sigma}\ell_n(x) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(\ln x_i)^2$$

and by setting $\frac{\partial}{\partial \sigma}\ell_n(x) = 0$ we can potentially maximise $\mathcal{L}_n(x)$.
Therefore,

$$-\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(\ln x_i)^2 = 0$$

and by solving for $\hat{\sigma}$ we find that

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i)^2}.$$

MLE

We need to check that this stationary point is actually a maximum, by confirming that $\dfrac{\partial^2}{\partial^2 \sigma}\ell_n(x) < 0$ at $\sigma = \hat{\sigma}$. Indeed,

$$\frac{\partial^2}{\partial^2 \sigma}\ell_n(x) = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(\ln x_i)^2$$

and subbing in $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i)^2}$, we have

$$\begin{aligned}
\frac{\partial^2}{\partial^2 \sigma}\ell_n(x) &= \frac{n^2}{\sum_{i=1}^{n}(\ln x_i)^2} - \frac{3n^2}{\sum_{i=1}^{n}(\ln x_i)^2} \\
&= -\frac{2n^2}{\sum_{i=1}^{n}(\ln x_i)^2} \\
&< 0.
\end{aligned}$$

MLE

Hence, confirming that

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i)^2}$$

is indeed the maximum likelihood estimator of σ.

Properties of the MLE

Properties of the MLE

Suppose that $\hat{\theta}$ is the MLE of θ given some random sample $(X_1, \dots, X_n)$.

1. **Consistency**: $\hat{\theta} \xrightarrow{\mathbb{P}} \theta$.
2. **Equivalence**: If g is a 'nice' function then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
3. **Asymptotic Normality**:

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

The Fisher Score & Information

Definition of Fisher Score & Fisher Information

- The Fisher score is defined to be

$$S_n(\theta) := \partial_\theta I(\theta; X_1, \ldots, X_n)$$

  where $I(\theta; X_1, \ldots, X_n)$ is the log likelihood.
- The Fisher Information given by $X_1, \ldots, X_n$ is defined to be

$$\begin{aligned} I_n(\theta) &:= -\mathbb{E}(\partial_\theta^2 I(\theta; X_1, \ldots, X_n) \\ &= -\int_{\mathbb{R}^n} \partial_\theta^2 I(\theta; x_1, \ldots, x_n) \prod_{i=1}^n f(x_i; \theta) \, \mathrm{d}x_i. \end{aligned}$$

Why would we create such a thing?

The Fisher Score & Information

Properties

- $\mathbb{E}_\theta S_n(\theta) = 0$.
- $I_n(\theta) = \mathbb{E}_\theta[\ell_n'(\theta)]^2 = \mathbb{V}\text{ar}_\theta(S_n(\theta))$.

Result

Let $X_1, \ldots, X_n$ be random variables with common density function f depending on the parameter θ, and let $\hat{\theta}_n$ be the MLE of θ. Then, as $n \to \infty$

$$I_n(\theta)\mathbb{V}\text{ar}(\hat{\theta}_n) \xrightarrow{\mathbb{P}} 1.$$

Hence,

$$\text{se}(\hat{\theta}) \approx \frac{1}{\sqrt{I_n(\hat{\theta}_n)}}.$$

How to use Asymptotic Normality

Example

Suppose that $X_1, \ldots, X_n \sim f$, where $f(x;\theta) = 2\theta x e^{-\theta x^2}, x \geq 0; \theta > 0$. Find the estimated standard error of $\hat{\theta}$, and the approximate distribution of $\hat{\theta}$ if it is known that

$$I_n(\theta) = n/\theta^2.$$

We know that

$$\widehat{\mathrm{Se}}(\hat{\theta}) \approx \frac{1}{\sqrt{I_n(\hat{\theta})}} = \frac{\hat{\theta}}{\sqrt{n}}.$$

Thus,

$$\frac{\hat{\theta} - \theta}{\theta/\sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \implies \hat{\theta} \overset{\text{app.}}{\sim} \mathcal{N}(\theta, \theta^2/n).$$

MATH2901 Q5 Assignment Q

Q5 part 3

Show that the Fisher information for σ, given $n = 1$, is

$$I_1(\sigma) = \frac{2}{\sigma^2}.$$

Recall that

$$\frac{\partial^2}{\partial^2 \sigma} \ell_n(x) = \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{n} (\ln x_i)^2,$$

and if $n = 1$, then

$$\frac{\partial^2}{\partial \sigma^2} \ell_1(x) = \frac{1}{\sigma^2} - \frac{3}{\sigma^4} (\ln x_1)^2.$$

In addition, we know that $\ln X \sim \mathcal{N}(0, \sigma^2)$ and therefore,

$$\mathbb{E}((\ln X)^2) = \mathbb{V}\text{ar}((\ln X)^2) - \mathbb{E}((\ln X)^2)^2 = \sigma^2 - 0^2 = \sigma^2.$$

Q5 part 3

Show that the Fisher information for σ, given $n = 1$, is

$$I_1(\sigma) = \frac{2}{\sigma^2}.$$

Thus, we have

$$\begin{aligned} I_1(\sigma) = -\mathbb{E}\left(\frac{\partial^2}{\partial\sigma^2}\ell_1(x)\right) &= -\left(\frac{1}{\sigma^2} - \frac{3\sigma^2}{\sigma^4}\right) \\ &= \frac{2}{\sigma^2}. \end{aligned}$$

Likelihood Based Confidence Intervals

Definition of Wald Confidence Intervals

Let $X_1, \ldots, X_n$ be random variables with common density function f, where

$$f(x) = f(x; \theta), \quad \theta \in \Theta$$

and let $\hat{\theta}$ be the MLE of θ. Under the conditions for which θ is asymptotically normal,

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

is an approximate $1 - \alpha$ confidence interval for θ for large n, where $\mathrm{Se}(\hat{\theta}) = 1/\sqrt{I_n(\theta)}$.

Cramer-Rao lower bound

Cramer-Rao lower bound

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

$$\mathbb{V}\mathrm{ar}_\theta(\tilde{\theta}) \geq \frac{1}{nI_1(\theta)}.$$

Multi-parameter Maximum Likelihood Inference

Fisher information matrix

Let $\theta = (\theta_1, \ldots, \theta_k)$ be the vector of parameters in a multi-parameter model. The Fisher information matrix is given by

$$I_n(\theta) = -\begin{pmatrix} \mathbb{E}(H_{11}) & \mathbb{E}(H_{12}) & \cdots & \mathbb{E}(H_{1k}) \\ \mathbb{E}(H_{21}) & \mathbb{E}(H_{22}) & \cdots & \mathbb{E}(H_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(H_{k1}) & \mathbb{E}(H_{k2}) & \cdots & \mathbb{E}(H_{kk}) \end{pmatrix}$$

where

$$H_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta).$$

Multi-parameter Maximum Likelihood Inference

Asymptotoic Normality

Let $\tau = g(\theta)$ be a real-valued function of $\theta = (\theta, \ldots, \theta_k)$, with maximum likelihood estimate $\hat{\theta}$ and $\hat{\tau} = g(\hat{\theta})$. Under appropriate regularity conditions, including the existence of all second order partial derivatives of $\mathcal{L}(\theta)$ and first order partial derivatives of g, as $n \to \infty$

$$\frac{\hat{\tau} - \tau}{\mathrm{Se}(\hat{\tau})} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1)$$

where

$$\mathrm{Se}(\hat{\tau}) = \sqrt{\nabla g(\theta)^\top I_n(\theta)^{-1} \nabla g(\theta)}.$$

What exactly is a Hypothesis Test?

Definition

The null hypothesis, labelled H_0, is a claim that a parameter of interest to us (θ) takes a particular value $(\theta)_0$. Hence, H_0 has the form $\theta = \theta_0$ for some pre-specified value θ_0.

The alternative hypothesis, labelled H_1, is a more general hypothesis about the parameter of interest to us, which we will accept to be true if the evidence against the null hypothesis is strong enough. The form of H_1 tends to be one of the following:

$$H_1 : \theta \neq \theta_0;$$
$$H_1 : \theta > \theta_0;$$
$$H_1 : \theta < \theta_0.$$

In a hypothesis test, we use our data to test H_0, by measuring how much evidence our data offer against H_0 in favour of H_1.

Mythbusters

Mythbusters Example

The Mythbusters were testing whether or not toast lands butter side down more often than butter side up.
In 24 trials, they found that 14 slices of bread landed butter side down.
Is this evidence that toast lands butter-side down more often than butter side up?

How to conduct a Hypothesis Test

Method

A hypothesis test has the following steps.

1. State the null hypothesis (H_0) and the alternative hypothesis (H_1). By convention, the null hypothesis is the more specific of the two hypotheses.
2. We use our data to answer the question: "How much evidence is there against the null hypothesis?"
   2.1 Find a test statistic that measures how "far" our data are from what is expected under the null hypothesis.
   2.2 Calculate a P–value, a probability that measures how much evidence there is against the null hypothesis, for the data we observed.
3. Write a conclusion.

Doing the Mythbusters Example

1. Let
$$p = \mathbb{P}(\text{Toast lands butter side down}).$$
Then,
$$H_0 : p = 1/2 \quad \text{versus} \quad H_1 : p > 1/2.$$

2.1 Let $\hat{p}$ be the sample proportion, and in particular we will look at the test statistic
$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1).$$
Under the H_0,
$$Z = \frac{\hat{p} - 0.5}{\sqrt{0.5(1-0.5)/n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1).$$

Doing the Mythbusters Example

2.2 We need to ascertain whether $\hat{p} = 14/24$ is unusually large, if $p = 0.5$. That is we're asking what is the value of

$$\mathbb{P}(\hat{p} \geq 14/24)?$$

Doing the Mythbusters Example

2.2 We need to ascertain whether $\hat{p} = 14/24$ is unusually large, if $p = 0.5$. That is we're asking what is the value of

$$\mathbb{P}(\hat{p} \geq 14/24)?$$

We can figure that out by observing,

$$\begin{aligned}
\mathbb{P}(\hat{p} \geq 14/24) &= \mathbb{P}(\hat{p} - 0.5 \geq 14/24 - 0.5) \\
&= \mathbb{P}\left(\frac{\hat{p} - p}{\sqrt{0.5(1-0.5)/24}} \geq \frac{14/24 - p}{\sqrt{0.5(1-0.5)/24}}\right) \\
&= \mathbb{P}\left(Z \geq \frac{14/24 - p}{\sqrt{0.5(1-0.5)/24}}\right) \\
&\approx \mathbb{P}(Z > 0.82) \approx 0.2071.
\end{aligned}$$

Which begs the question, how significant is 0.2071 as a probability?

P-values

Range of P-value	Conclusion
P-value ≥ 0.1	little or no evidence against H_0
$0.01 \leq P-$value < 0.1	some, but inconclusive evidence of H_0
$0.001 \leq P-$value < 0.01	evidence against H_0
P-value < 0.001	strong evidence against H_0

Conclusion

3. Since $0.2071 > 0.1$, we conclude that we have no evidence against the claim that $p = 0.5$ – because our data is consistent with this hypothesis.

Normal Samples

In the situation where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, we can use the fact that

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

to test any of the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{verses any of } \begin{cases} H_1 : \mu < \mu_0; \\ H_1 : \mu \neq \mu_0; \\ H_1 : \mu > \mu_0. \end{cases}$$

Example

Normal Sample Example

Before the installation of new machinery, the daily yield of fertilizer produced by a chemical plant had a mean $\mu = 880$ tonnes. Some new machinery was installed, and we would like to know if the new machinery is more efficent.

During the first $n = 50$ days of operation of the new machinery, the yield of fertilizer was recorded. The sample mean was $\overline{x} = 888$ with a standard deviation of $s = 21$.

Is there evidence that the new machinery is more efficient? Use a hypothesis test to answer this question, assuming that yield is approximately normal.

Solution Method 1)

1. We have $H_0 : \mu = 880$, $H_1 : \mu > 880$.
2. 2.1 We use the test statistic

   $$T = \frac{\overline{X} - 880}{S/\sqrt{n}} \sim t_{n-1}.$$

   2.2 We would like to find,

   $$\mathbb{P}\left(T > \frac{\overline{x} - 880}{s/\sqrt{n}}\right) = \mathbb{P}\left(T > \frac{888 - 880}{21/\sqrt{50}}\right) = \mathbb{P}(T > 2.69)$$

   where $T \sim t_{49}$. Using R, we find that this probability is $0.0049 < 0.005$.
3. This tells us that, if H_0 were true, we would be highly unlikely to observe a T statistic as large as 2.7. We have strong evidence against the claim that $\mu = 880$.

Solution Method 2)

Another way to tackle this problem, is to find a rejection region for a test of size 0.05.

Definition of Rejection Region

The rejection region is the set of values of the *test statistic* for which H_0 is rejected in favour of H_1. To determine a rejection region, we first choose a size or significance level for the test, this is typically 0.05.

- Since our test statistic $T \sim t_{49}$, then we can use R to find some value c such that $\mathbb{P}(T > c) \approx 0.05$. Using `qt(0.95,49)` gives us 1.676. Hence, our rejection region is $T > 1.676$. That is, if $T > 1.676$, we will reject H_0 in favour of H_1, else, we will retain H_0.
- Our observed value of T was 2.69 which is in our rejection region.
- As such, we reject H_0 and conclude that there is evidence that $\mu > 880$.